THE EFFICIENCY OF THE NON-PROFIT ENTERPRISE: CONSTITUTIONAL IDEOLOGY, CONFORMIST PREFERENCES AND REPUTATION(*)

Lorenzo Sacconi•

1. Introduction and motivation

Which is the distinctive comparative advantage, if any, of the non profit enterprise? One line of argument, which I want to pursue in this paper, is that the non profit enterprise (in short NPE) is able to attract ideological entrepreneurs and workers (Rose-Ackerman 1996). "Ideologues" are agents committed the principles of a given philosophy of service, a "mission" in a field of provision and distribution of some welfare good or service. In terms of the idea of the basic institutions of society (Rawls 1971), a "mission" may be meant as typically providing primary goods such as health, education, culture, basic income (that is the basis of self-respect) and ideologues are individuals committed to the institutional mission of providing the "primary good" and its fair distribution.

Within the ideologues' approach the organisational *mission* is seen as a value *per se,* rather than being instrumental to achieving a further good or satisfying other interests, for example maximizing profit. The ideological entrepreneur gets directly satisfaction from conformity to the mission or the philosophy of service of the organisation he runs, even if it may be detrimental to his/her material personal interest. It is easy to understand, then, why ideologues prefer a form of enterprise where there is not an owner claiming the residual. Otherwise the mission would be instrumental to another goal, the owner wealth. The non profit enterprise corresponds quite well to this motivational structure, as far as the residual distribution constraint debars from the outset that the entrepreneur, once the production inputs have been remunerated, must devolve the residual to satisfy an interest external to the organisational mission.

(*)Paper presented at the Conference, "The Invisible Hand and the Common Good", Trent University, Peterborough, Ontario, 13-16 June 2002. To appear in B.Hodgson (ed.) The Invisible Hand and the Common Good, Proceedings of the SEEP Conference, Springer-Verlag, Seires in Economic Ethics and Philosophy. This paper is connected to a joint larger research project with Gianluca Grimalda, and deeper formulations and generalisations of its propositions are included in our joint work (see Grimalda and Sacconi 2002). Early versions of this paper have beneficiated from comments by Carlo Borzaga, Fabrizio Cafaggi, Ferruccio Marzano, Luigino Bruni and Johnny Dotti. Support by the MIUR within a national research project on Law and Economics is also gratefully acknowledged.

• Department of Economics, University of Trento (Italy) and Centre for Ethics, Law & Economics, University Cattaneo-Liuc Castellanza (Italy) - lorenzosacconi@iol.it

The "ideologues" thesis in this sense is complementary, not alternative, to the "property rights" explanation of the non profit enterprise (Hansmann 1980, 1987, 1988). It in fact says that, due to radical asymmetry of information between welfare services providers (agents) and their sponsors or beneficiaries (principals), any design of the contractual relation between the agent and the principal that gives the agent control on the firm, would result in serious transaction costs borne by beneficiaries, while nevertheless the principals are not efficient in exercising control. The distribution constraint on the contrary reduces the agent's incentive to embrace opportunistic behaviour toward the beneficiaries, thus making possible the production of services and goods requiring fiduciary relationships between principals and agents. This explanation tells us that the non profits create a more congenial (less costly, from the transaction costs perspective) institutional environment for the efficient production of welfare goods. However, it should be taken for granted that the non distribution constraint is not a *panacea*, as far as there are many other ways in which a rent could be appropriated by agents (managers, entrepreneurs, etc.) other than the explicit appropriation of the residual in the legal form of profit. Thus the "ideologues" thesis provides an explanation of why nonetheless there are people motivationally ready to refrain form exploiting these opportunity, who want to stick to their "mission".

Anyway, it is evident that an ideological commitment can mean advantages and disadvantages as well to the firm producing welfare goods (let me call it simply the *Social Enterprise* hereafter). The obvious disadvantage is less flexibility and less adaptation to changes that may create business opportunities in fields other than those in which the original mission has been defined. Why advantage?

Here the main thesis of the paper can be suitably anticipated. It is convenient to split it in four propositions

- *Proposition I:* ideologues, both entrepreneurs and workers, share a principle of justice seen as the constitutional ideology of the NPE, giving the *ab origine* justification of its existence and mission. This is an idea of the constitutional contract of the NPE from which they derive fiduciary duties toward the beneficiaries, that is duties that they rationally accept – i.e. a deontology (see sec. 3).
- *Proposition II*: the constitutional principle provides an independent source of motivation (a source of utility) of the players in the "social enterprise game", in so far as they believe in the reciprocity of expected conformity to the ideology by all the participants. I call this conformity-based utility "ideological", and I see it as the representation of a preference for expected conformity to the given constitutional principle (i.e. the preference to conform given the expectation of a deontological mode of behaviour pursued by all the participant in the game). In order to differentiate it from the utility meant as a function of classical preference, I call the classical concept "material" utility (see sec. 3 and 4).
- *Proposition III*: the conjunction of propositions I and II makes possible to overcome personal incentives to embrace opportunistic behaviour within the functioning of the social enterprise, so that the proper Non-profit Enterprise emerges as it can be proved that in the "social enterprise game" amongst the member of the organisation there exists an organisational equilibrium minimising transaction costs to the beneficiaries (see. sec. 5).

- *Proposition IV:* At last, this equilibrium rests on the emergence of an expectations system of reciprocal conformity to the constitutional ideology. Not just because expectations support non opportunistic equilibrium strategies, but simply because beliefs about reciprocal conformity enter the players' preferences and, by changing the payoff structure, they do "create" the equilibrium (see sec. 5).

The last proposition makes inherently fragile the organisational equilibrium that minimises transaction costs to the beneficiaries. As the existence - not even the selection – of the internal organizational equilibrium rests heavily on the existence of the appropriate system of reciprocal expectations, the problem of how we can justify the emergence of the appropriate system of beliefs must be underlined. Here is where the explicit moral codes of the Non profit Enterprise enters the scene. I see the code of ethics – i.e. a set of general and abstract principles with annexed more concrete precautionary rules of behaviour – as the building block for deriving a reputation equilibrium between the NPE as a whole and its external stakeholders within a repeated game (see. sec.6. *proposition V* in particular), whose stage-game is the typical game of trust played under incomplete knowledge and unforeseen contingencies (Kreps 1990). Due to the reputation equilibrium, the expectation that the ideology is conformed to by the internal members of the NPE is justified. Thus in this paper I explore at the same time three basic roles of ethics in the NPE: (i) the *justificatory* role driven by the constitutional ideology, (ii) the *motivational* role, driven by conformist preferences, and (iii) the *cognitive* role, driven by the code of ethics, which is the basis for reasonably defining expectations on the carrying out of commitments in the presence of unforeseen contingencies (Sacconi 2000, 2001). At last, I find that they virtuously play interdependent and mutually supporting roles in the emergence of the efficient NPE (see sec.7, *proposition VI* in particular).

My results are clearly indebted to that strand of economic literature that sees expectations and beliefs about strategic behaviour as directly entering the utility functions of the players (utilities depend strictly on what the players believe about the conformity of other players to given strategy combinations) so that beliefs contribute to create an entirely new set of equilibrium points of the relevant game (Geanakoplos et al. 1989, Rabin 1993, Bernheim 1994, Sugden 1998a, 1998b). A first difference with many of these contributions can be seen in that I do not attach normative force to common mutual expectations *per se*. On the contrary I characterize directly the normative principle or constitutional ideology of the NPE in terms of contactarian ethics. It is because the organisational members play an hypothetical bargaining game, where they rationally agree on the ideal constitution of the firm, that they then will use ideology to identifying the real game behaviours to whom they attach ideological utility. In other word, it is because expected strategies comply with an independent ethical principle that they get additional ideological utility, whereas the simple fact that players may commonly expect one another to follow a given rule of behaviour (mostly equilibrium conventions) doesn't imply *per se* any additional source of conformist motivation. In this sense my approach may be see as *moral conformism*, not *natural* conformism (as can be understood normative expectations according Sugden's approach).

Secondly, I try to work out the philosophical underpinnings of this reform of the players' utility functions in the NPE game (see sec.4). The notion of ideological utility is based on conformist preferences. These are preferences for those actions that are part of states of affairs described in terms of interdependent actions conforming to an abstract norm or principle, which become effective once the preferences' holder does expect that the other players do they part in that state of affairs and they do expect that himself do his part in the same state of affairs. What result is that a player's ideological utility depends on the expectation of deontological modes of behaviour followed by all the participants, himself included. True, this type of preferences will result nonetheless based on a function defined over the material utilities of the players (the *Nash Bargaining Function*). However, the form of this function establishes not a goal but only a *fairness* criterion, which implies a distributive pattern of utilities depending on an abstract principle of justice (what players can rationally agree upon under ideally symmetrical bargaining conditions). Moreover what counts for defining the ideological utility is that the distance between the ideal state of affairs and the states of affairs ensuing from actions actually undertaken, be minimized. That is, what counts is that the players' actions conform to an ideal of behaviour. The level of preference changes not according to how large is the slice of the pie that the player gets, but according the to the distance (conformity level) of the expected actions to that ideal.

Once this not secondary reform is accepted[1], my explanation of the NPE - centred as it is on the role of ideology, conformism and moral codes - results nevertheless consistent with methodological individualism and other typical rational choices explanations. Here, as in any other individualist methodological explanation, economic agents maximise the utility function of the Self and their social behaviour can be predicted in terms of some equilibrium points of the relevant game representing their strategic interaction.

2. The social enterprise game: what would happen without ideology?

Let begin the analysis in a standard economic setting, without any change in the motivational system of the participants. I suggest that in the social enterprise (SE in short) then will take place a strategic interaction, which I model in a stylised way by a non-cooperative game with three players: an entrepreneur, a worker and an external beneficiary (who consumes the welfare good produced by the firm). The collective decision problem they face is how to allocate a surplus, for example the result of a fund raising campaign or the residual resulting after costs from the previous accounting period. This surplus must be allocated amongst different uses: covering extra administrative-cost that the management claims have been incurred or are to be incurred in the next future, paying higher wage to the worker as extra-compensation for the previous period or to induce him to exert extra-effort in the next period, improving the quality and quantity of services to be provided to the external beneficiary. The game could be seen as a bargaining game within which the three players attempt to agree over a conjoint strategy in order to solve both a cooperative problem and a distributive problem. In fact on one hand they must agree in order to make possible their conjoint contribution to the production and the consumption of the welfare good, but on the other hand each would stay in the joint venture only if he gets what he seeks from the SE (I will give an account of this hypothetical bargaining game in the next section).

However, I assume that the actual game played in SE is a non-cooperative game in which the players do not have to bargain in the proper sense over the allocation of the surplus. This reflect the assumption that there are not external institutions or rule that may enforce any bargain the three players can agree upon, so that if they decide to follow some pattern of conjoint behaviour, its implementation can only rest on their individual interdependent strategy choices. It is also intended to reflect a situation like the following: the entrepreneur, due to his hierarchical position of an authority in the firm, simply can announce a higher or lower level of administration-costs in order to obtain that a larger o lower share of the surplus be devoted to what he declares. At the same time the worker is entitled to claim higher or lower wage because he may give-up any gift of labour (or effort) to the SE, which he gave in the foregoing periods, or continue to give some labour or extra-effort to the firm on a purely voluntary (not paid) basis. This can be understood as if the worker might ask for the market wage level or lower wage level, in a situation in which labour has some market-wide bargaining force (so that the market level of wages in the industry already permits the worker to appropriate a rent) and the management can not bargain over the decision of the worker to claim what the market in general would offers to him. In this game the consumer or beneficiary doesn’t have any real influence on the result, as far as it is subject to an internal allocation decision between the members of the firm. However he receives the effect of the other players’ strategic decisions, so that we can consider his payoffs together those of the other players.

Let me illustrate the strategy set of each player, as it is represented by the matrix game of fig.1.

- The worker’s strategy set: given a level of effort, the worker may choose strategy LW (claiming “Low” wage) or strategy HW (claiming “High” –i.e. market – wage)
- The entrepreneur's strategy set: given an amount of the firm’s output, the entrepreneur may choose strategy LC (asking for covering “Low”- i.e. true - administrative cost) or choose strategy HC (pretending “High” administrative cost – where high costs represent a level of rent appropriation by the entrepreneur)
- The beneficiary’s strategy set: it is empty, because the beneficiary is a “dummy” player with no direct influence over the outcomes of the game. Her payoffs (see payoffs within brackets) however are determined by the other two players’ interdependent choices .

From the outcomes depicted in the matrix game of fig.1 it results that if the players claim high wage and high costs they appropriate all the surplus and nothing is left to the beneficiary. If they both moderate their claims, on the contrary, resources are allowed for higher quality or increased quantity of welfare goods to the beneficiary.

	LC	HC
LW	2, 2, (6)	2, 6, (1)
HW	6, 2, (1)	4.5, 4.5, (0)

Fig.1 the SE game

Unilaterally giving up the high claim by one of the two players admits very low utility to the beneficiary and facilitates the counterpart in reaping his maximum payoff. Notice that all of the four outcomes are included in the Pareto set of the game, but quite evidently the total amount of benefits distributed when both the active players restrain their claims is higher than in the case they claim their highest payoffs, and also higher than in the case one player take advantage of the counterpart's moderation. Even though Pareto efficiency can not aid in discriminating amongst outcomes, there are senses in which we may recognise that by claiming the highest payoffs the players generate an inefficiency of the SE. First of all the SE becomes inefficient in terms of total production of the welfare goods provided to the beneficiary, which in fact gets nil. Secondly, transaction costs borne by the beneficiary when the two active players ask for high wage and high costs are higher than the total transaction costs they would face whether they renounce to claim such payoffs. In terms of Kaldor -Hicks efficiency concept, it is possible to construe by an hypothetical bargaining that the beneficiary could bribe the two active players in order to convince them to abandon the "High/High" strategy pair and making acceptable to them the outcome that allows maximum benefit to the beneficiary. It would be enough for the beneficiary to give up a value of 4 of its payoff under the outcome (LW,LC) in order to compensate the active players, while maintaining nevertheless a surplus share of 2. On the other part, there is no possible bargain by which the active players would be able to convince the beneficiary to pass from the "Low/Low" strategy pair to the "High/High" one. Even though they were ready to reduce themselves to the same payoff they would reap under (LW, LC), the total amount of the bribe (4) would not compensate the beneficiary for surrendering the outcome where she get the payoff 6. Thus the outcome (HW,HC) is dominated by the outcome (LW, LC) according to Kaldor-Hicks efficiency.

Of course the reason for the NPE is an interesting subject is that such a Coasian contract between the beneficiary and the producers of the welfare good is not possible. The beneficiary may have nor information, nor rationality enough to contract over the allocation of the surplus amongst alternative uses internal to the organisation. If he were to try, transaction costs would dissipate all the surplus, so that it is recommended to establish a firm able to implement a fiduciary relationship between the beneficiary and the producers. This means that the entrepreneur should exercise the authority to manage the firm in the best interest of the fiduciary (i.e. the beneficiary)[2].

It is apparent from the matrix game's payoffs, that the only Nash equilibrium of this game is in dominant strategies and coincides with the strategy combination (HW, HC). Individual rationality will consequently push the players to act opportunistically and to claim high wage and high administrative costs. Notice that as far as the game is analysed only according to payoffs of the active players, the equilibrium is also Pareto dominant in the two-person small society of one entrepreneur and one worker. Thus, according to this restricted view of the game, there is not any principle of social efficiency that, contrasted with the principle of individual rationality, would give rise to conflicting prescriptions - what on the contrary typically happens when opportunism is at work. The effect, however, within the enlarged society of three players, is that high transaction costs are borne by the beneficiary (her rent is 0).

What we have seen is how the SE (that is simply an enterprise that produces social welfare goods within a fiduciary relation to the beneficiaries) degenerates to a for profit: all the surplus is appropriated by the producers and no part of it is devoted to bettering quality or increasing quantity of the services provided to the beneficiary. So many the ways are in which the non distribution constraint may be circumvented that there is no reason however to expect that the SE will necessarily take the legal form of a for profit. My game exemplifies just one of them, as far as the entrepreneur can pretend that administrative costs are higher than they actually are, and he can do that by abusing of his formal position of authority in the organization. Let it be as it might, the provision of welfare goods will be undersized. The prevailing "residual-appropriation-seeking" behaviour implies an organisational failure, which is the counterpart of the typical market failures in the industry of welfare goods.

3. Hypothetical game, constitutional ideology and its motivational force

Why does the SE escapes the failure described at the end of the previous section and how does it definitely assume the character of an efficient NPE? In other words why a NPE emerges such that the internal members of the organisation accomplish fiduciary duties to the beneficiary? My answer is that both the entrepreneur and the worker are "ideologues". I 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 I want to capture in terms of contractarian ethics. The second is the motivational role, which I will model by conformist preferences. It is a basic tenet of this paper that these two roles must be considered as both indispensable but irreducible one to the other, so that both should be squarely faced by any intellectually honest endeavour to explain how morality can play a role in economic organisations[3].

Hp.1: The NPE's internal players stick to an ideology. It states that the NPE is based on an hypothetical "social contract" amongst all the players - the beneficiary included - affirming an ethical principle of fairness .

The situation has to be understood as if, before playing the actual game, an 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 that 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. To say it differently, 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 "anybody" - 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. Thus the “social contract” works as a “Constitutional” ideology legitimating the enterprise as an institution *ab ovo*.

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 beneficiary 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 players in the actual game. On the contrary it considers the ex ante perspective in which also the beneficiary would have a voice about the terms of agreement on the cooperative venture in which the beneficiary essentially contributes, as he is the consumer of the organisation’s output. Rational agreement in this hypothetical game thus requires 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 should pick up the distribution maximizing the product of the three players’ payoffs net of the status quo (Nash 1950). Note that Nash Bargaining Solution selects always 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 our simple game maximising Nash product implies choosing the outcome where both the internal players choose the Low strategy allowing the most part of the surplus to go to the beneficiary[4]. In sum, I resort to the Nash bargaining solution as a normative criterion for defining a moral preference over the outcomes of the original game. I will use it as a sort of “Social Welfare Function” that orders outcomes according to “distributive justice” (remember that all our outcomes are already situated on the Pareto frontier of our decision problem)[5] .

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 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 payoff structure of the actual non-cooperative game, we know that the players *do not have* the appropriate incentives to put in practice the precepts of the constitutional ideology asking them to leave the most part of the surplus to the beneficiary. Why then the active players, the entrepreneur and the worker, do comply with their constitutional ideology?

Here comes in my 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.*

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. In other words, if the worker is acting in conformity with the ideology (that is he chooses LW) and if he believes that also the entrepreneur is acting in conformity to the ideology (i.e. he chooses LC), then he gets additional utility from acting in such a way, which adds to the utility that he gains from the material outcome of his choice (which depends on the other party choice). Symmetrically the entrepreneur gains additional utility form acting according to the ideology if he believes that the worker does the same and (he believes that) the worker also believes that he acts according to the ideology - that is by choosing LC whether he believes that (LW,LC) is the current outcome. Following the theory of psychological games (Geanakoplos et al. 1989), I hypothesise that there exists a component of the utility functions of the players, defined on their strategy choices, which depends on their beliefs about their reciprocal choices. This is an additional component to be considered separate from the utility they gain from the material payoffs associated to any outcome of the game.

Hence, a player who adopts his dominated strategy (LW or LC), which may obtain only outcomes with low payoffs, in the event that he believes that the counterpart would choose reciprocally the dominated strategy, would gain an additional amount of utility as far as the result is an outcome conforming to the ideology. If this effect is strong enough to overcome the effect of the material payoffs, that is if in the balance the ideologically based utility prevails, it can be predicted that the strategy choice of the two players will conform to the ideology. In this case the reciprocal behaviour of the players confirms their reciprocal expectation, so that they will not have any reason to revise ex post their expectations. Absence of reasons to revise expectations characterises this outcome as an equilibrium of the psychological game, that is a system of mutually consistent expectations inducing a strategy choices profile such that expectations are confirmed in practice (Geanakoplos et al. 1989).

4. Two distinct concepts of preferences of the Self

It should be made clear here what concepts of preferences and utility I am implicitly employing. This somewhat long section is intended to discuss the philosophical underpinnings of the reform introduced by Hp.2, which will be embodied in the formal model of the game in sec.5.

On the one hand, we have in fact preferences and utilities defined over the outcomes of the players' interaction, that is preferences over what happen to a player under the outcomes depicted by the matrix game of fig.1. On the other hand, we have preferences and utilities meant as a function of the beliefs that players entertain about their reciprocal conformity to an abstract principle or a solution concept for a wide class of games. In the first case a player would get utility form the consequences of the outcomes (what happen to him because of the interaction result), whereas in the second case he gets utility form beliefs about a mode of behaviour jointly put forward by all the players, which is viewed in so much as it conforms to an abstract norm or principle. The second source of utility does not follows from the usual strain of consequentialist reasoning, whilst it introduces at the basis of utility a typically deontological argument. It is an *intrinsic* characteristic of a set of actions (to be precise a combination of each player's action and his expectation over the other players actions, upon which it is contingent) what gives raise to the kind of preference under consideration.

As far as a norm is simply rationally agreed upon in the *ex ante* hypothetical bargaining game, it is not yet a source of utility. It gets its motivational force once the player has developed the expectation that the norm is also reciprocally conformed to by every players in the game, him included. This kind of preference may appropriately be called *conformism*, as it expresses a desire to see those norms that all have rationally agreed upon to be complied with by everybody. Moreover, it should be better understood as *moral conformism,* because the relevant preference is developed only with reference to a principle of fairness or an ideology, which is the result of ex ante unanimous, impartial and rational choices.[6]

In effect what I am defining are two distinct concepts of personal preferences, i.e. two types of *preferences of the Self*, wherein self-interest or egoism is only a particular case. Therefore, within the whole model of individual preference-based rationality there is room for different kinds of interest of the Self.

4.1. Consequentilistic preferences of the Self

First of all, there are preferences of the Self defined over consequences. Consequences are meant as *what happens* to some individual under a given outcome of interaction. These preferences can be defined over consequences concerning the Self alone and affecting only him-self without any regard to what happens to any other individual. In this case the Self is self-interested because his preferences would depend only on the consequences happening to him-self in each state of the world. Quite different would be the case if the Self would take his preferences not only over consequences occurring to him, i.e. *self-referred consequences*, but also over consequences concerning any other individual.

Actually, strategic interaction generates states of affairs which can be differently described according to their different characteristics. Such characteristics can be seen as *what happens* to the *decision maker* in a state – i.e. as the consequences *to the decision maker* - or what happens to any subset of individuals or to *every* individuals – that is the consequence *to everybody* in the same state. In the first case the characteristics would be attributes of the single agent him-self (his wealth, leisure, effort, his power exercise, the manifestation of his creativity and the like) and they result out 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 subset of individuals or whichever individual, 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.

If a decision maker defines his utility as a representation of the consequences that concern only him-self, we will have a utility function that represent his self-interest. I call the underlying preferences of this utility function *personal self-referred consequentialist preferences.* However, if the preferences of the Self are defined over descriptions of the states of affairs in terms of consequences concerning any subset of individuals or every individuals, then the Self is considering *extended consequences (impartially* extended in the latter case). This seems the natural way of accounting for a moral preference of the Self in consequentialist sense, which will be represented by some social welfare function. Of course this is not yet a complete account of utilitarianism. Utilitarianism asks for considering, as the basis for an impartial preference judgment, not only the consequences to whichever individual, but also the evaluation of such consequences from the very point of view of each individual's preferences – which asks at least implicitly for interpersonal utility comparisons. Nevertheless evaluating a state according to the consequences that it attaches to every individuals means to take an impartial perspective over consequences, whichever the individual be to whom that consequences might happen. This is what I call *consequentialist personal moral preference.* From a purely formal point of view there is no difference in taking as the basis for a utility representation the preference ordering defined over the first kind of descriptions of the state's characteristics, or the preference ordering defined over the second enlarged kind of descriptions of the same state's characteristics, even if the moral meaning is quite different in the two cases.[7]

4.2. Conformist preference of the Self

Let come now to the second type of preferences of the Self, which I call *Personal Conformist Preferences.* As well as the first type of preferences, also conformist references are defined over states of affairs, but these are not described in terms of consequences occurring to any individual whatsoever, but as patterns of collective, interdependent or conjoint behaviours, and as beliefs about such modes of behaviour. The elements to be considered here are in sequence: a) the relevant description of states of affairs constituting the basis for defining the new type of preferences, b) the preference ordering over the states of affairs as it depends upon the relevant description of states of affairs, c) the induced preferences ordering over the actions set of each individual player, d) the numerical representation of such preferences by an utility index that I call *ideological utility*.

(i) The relevant description of the states of affairs. At this stage of the argument, states are primarily characterised as set of interdependent actions, conforming or not to a given abstract principle. What we are looking after in this description are modes of deontological collective behaviours maintained by the players. I fix a pattern of behaviours (a vector of strategies) that I define as perfectly *deontological* because it fully conform to an abstract principle of fairness or to a fair criterion of benefits distribution amongst the interested parties. Call such a state the *ideal*. Then I look after the degree of conformity to the ideal displayed by each state of affairs resulting out of the individual choices actually performed by all the players. I accomplish this task by seeing whether the *ideal* comes about through the actual individual choice carried out by each player, given the choice (he expects form) any other parties. This in fact helps us not only to say whether the actual state, appropriately described, conforms to the deontological ideal, but also to impute to each player's action the cause for any deviation from the ideal. Thus, what we describe are states of affairs in terms of combinations of actions and their proximity to, or their deviation form, the ideal.

Moreover, notice the importance of beliefs in the relevant description of the states of affairs. To say "given" within strategic interaction asks for adding "according to the player's beliefs", i.e. we look after states of affairs resulting from the choice of each player given his beliefs about other players' actions, which in turn is based on what the player believe the other players believe about the first player choice. Hence, by describing states of affairs we describe how far a vector of actual strategy choices, contingent upon the vector of individual beliefs justifying these choices, is faraway from the vector of strategy choices defined as the ideal. However in equilibrium - Nash equilibriums and also their extensions as the psychological equilibria - beliefs are confirmed by the actual actions. So we can understand equilibrium states of affairs as directly identifying the set of actually occurring strategy choices (associated with the "probability one" beliefs justifying them).

But remember that in order to define fairness we have to look at the distributions of payoffs, that is distributions of utilities based on the first type of preferences – i.e. material utilities based on personal consequentialist preferences. 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 corresponds to the ideal distribution defined according to an abstract principle. The Nash's SWF (remember what I have said about SWF in note 3) will describe each state according to the fairness principle. Therefore we will be enabled to see whether the occurring vector of strategies in any states determines a payoffs distribution such that a multiplicative function defined over material payoffs is maximised or not. 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 - i.e. how large it is the product of the payoffs multiplication net of the status quo.

Nothing does imply that a state of affairs under this description may be seen as a consequence which will happen to any particular individual, to all the concerned individuals, let alone the typical utilitarian fictitious mean individual, who gets 1/n of the sum of the individuals' payoffs. Under this description there is no individual to whom the relevant state of affairs happens as a *consequence*. We simply have a distribution saying the *ratio* according to which an efficient pie is partitioned amongst different players. The ratio, the partition, or the distribution as such are not consequences to any player, although they give a criterion (a formal property of the distribution) according to which the players share the pie - from which they may calculate the slice that will accrue to each of them as a consequence. In fact the relevant description of the state of affairs - based on the underlying description of the payoff gained by each players – is no more then an abstract formal property of the utility distribution (how large the utilities' product is under different outomes). I take this property to be meaningful as far as rational impartial acceptability of an agreement is concerned, or as far as I want to know whether the distribution is *fair* because of the proportionality of shares to relative needs or to relative marginal variation in the material utilities of players (Brock 1979, Sacconi 2000).

(ii) Conformist preference ordering over states. Any player makes choices within his strategy set. This means that preferences must be defined over his set of feasible actions. These preference however are to be derived from the preference ordering defined over states of affairs as described so far. Remember that preferences over states of affairs are not defined directly over consequences, but over *acts* because of their conformity to an abstract norm, i.e. a distributive principle. It is apparent that under these

descriptions acts are not taken in isolation but as sets of reciprocal acts (strategy vectors). It is also 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 states of affairs. The more a state of affairs conforms 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. This is the characteristic that I assume players take as endowed with moral value in order to say how desirable a state is. In this sense at the basis of conformist preferences lies a measure of how much deontology there is in the pattern of behaviour displayed by all the players in each state.

We may consider nonetheless preferences over states of affairs as ultimately based on subjective affections of the players (Gauthier 1986). In fact there is no reason to think that the preference criterion should be based on some objective value having an ontological reality out there, completely independent on the affections, the decision making activity or the judgement of those who are asked to express their preference. Note that, while conformist preferences depends on degrees of conformity, that is 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. At last rationally agreed principles of fair distribution are simply 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.[8] Each participant in the bargain seeks to gain as large utility as it is compatible with the symmetrical rationality of other bargainers. A fair principle of distribution follows form taking seriously the idea of an agreement acceptable by each player under this assumption of reciprocally expected rationality, an agreement that has to be recognised rational from whichever player's point of view. It is the idea of rational agreement - grounded on rational maximisation of the first type of utility under perfectly symmetrical bargaining conditions - the basis for deriving the principle of fairness. Therefore, rational agreement defines the value, not the value the reason for the agreement.

To be clear, let state the hierarchy within which the different pieces of the argument should be understood. First of all, for each player I take for granted the existence of some first order utility defined on possible agreements, which are initially described in terms of the consequence that each player gets from them. Second, players accept some terms of agreement concerning the surplus distribution. This agreement is worked out according to the fundamentally subjective notion of rational choice under ideally symmetrical bargaining conditions (this is drawn from the underlying idea that before playing the actual game, players will participate in a hypothetical bargaining game solved according to the Nash Bargaining Solution). Third, this agreement defines a norm for distributing benefits in any game situation of the kind under consideration. Fourth, I adopt this principle 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 introduce my deontological assessment of states of affairs. As from this step, a preference is no more merely a subjective attitude toward consequences, but a binary relationship giving rise to an ordering of states of affairs according to an objective measure of conformity.

The result is a preference ordering defined over states of affairs, which we hold not just because of our primitive psychological desires for material utility or preferred consequences, but *because* it conforms to a rationally agreed abstract principle. The fact that conformist preferences are based on a fairness 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 activity. Duties are simply those we have rationally agreed upon in a hypothetical bargaining situation.

(iii) Conformist preference over actions of a single player. At the end what really counts for determining the result of the game are any single active player's preferences over his own actions. As consequentialist preferences of the Self induce personal preferences over the actions' sets of every players, so much must also be true for conformist preferences of the Self. Simply these are induced by the conformist preference ordering over states described so far. If a player observes that a strategy combination conforming to the principle of fairness is the currently most probable state of affairs, then he will prefer the action that conforms to the duty – call it the deontological action – exactly *because* it contributes to the materialisation of a state of affairs conforming to the duty.

To state it a bit formally, agent A conformistically prefers action X_1 more than 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 action X_1 more than together action X_2.

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

To give again a definition a bit formal, agent A prefers action X_1 more than action X_2 if he believes that the other 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 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 natural explaining personal conformist preferences of agent A as resting on the existence of a hierarchy of mutual beliefs, within which any layer of beliefs is justified by a higher order layer of beliefs:[9]

Player A will prefer action X_1 more than action X_2 if he believes that

(a) player's A action X_1, together action Y, that A believes will be chosen by the other player B, brings about a state of affaires conforming to principle P more than action X_2;

(b) player's A action X_1 (which the other player B believes A will adopt) together action Y that (player B believes) A believes will be chosen by the other player B, brings about a state of affaires conforming to principle P more than action X_2;

(c) player's A action X_1 (which A believes the other player B believes A will adopt), together action Y that (A believes that players B believes) A believes will be chosen by the other player B, brings about a state of affaires conforming to principle P more than action X_2,

-
-
-

…*and so on* (the reasoning can be iterated on expectations about expectations of every order).

Because of course also these preferences are two place relationships, by assuming that they satisfy the usual assumptions of completeness and transitivity, we can derive an usual preference ordering over the strategy set of player A.[10]

(iv) Ideological utility. There is no reason for this preference ordering should not be represented by an individual utility function. I call it individual *ideological utility* of actions as it is based on the individual's conformist preference ordering on actions. It is derived in turn, *first,* from how much the individual believes that other players will conform with a principle, believing that every other players will also conform to the principle. And, *second*, from the fact that together the expected actions by the other players, the individual's action under consideration will bring about a state of affairs that will conform to the principle more than some other state of affairs brought about by another individual's action. Given that the conformist preference changes continuously with expected conformity of actions combinations to a principle, the Nash social welfare function (SWF) is a suitable basis for deriving a quantitative measure of conformity. Ideological utility must be a direct function of the conformity measure, in terms of the distance between any strategy combination and the state where the principle results fully satisfied.

At last, we must ask for what the overall utility function of a player is, which leads as a whole the decision making of each active player in the game. It should be the joint representation of the two types of preference, such that, assuming a plausible condition of decomposability, it may be reduced to an additive weighted combination of the utility representation defined over consequentialistic preferences and the utility representation defined over conformist preferences.

5. The NPE Game with mixed preferences

Formally the measure of the conformist component of the players' utility functions has to be based on a distance. In order to keep thing simple I will look for each outcome directly at the value of the Nash's SWF and I will express conformist utility for each player as a direct function of the distance between the value of SWF materialised in each outcome and its maximum value in the whole game. As it will be seen, this reduces to entering the Nash's SWF values within the utility function of the players[11].

Let $\mathbf{x}$ be the maximum value of the Nash product calculated over the outcomes of the game. Let moreover y_1,y_2,y_3,y_4 respectively be the values that Nash product takes at each outcome of the game (see respectively the top/left, top/right, bottom/left and bottom/right payoffs vectors in the matrix game of Fig.1). Then, we have the basic information in order to measure how faraway it is the result of any combination of actions from the ideal. Describing a "state of affairs" according to its conformity to the ideal can now be made to depend on the distance

$$\Delta = (\mathbf{x} - y_i) \qquad (i = 1,2,3,4)$$

The idea is simply to make the utility function of any player a monotone decreasing function of this distance for each conditional outcome of the game- i.e. for each action of a player given each of the expected action of the other player. That is the lesser the distance of the actual outcome form the ideal value, the greater the additional conformist component in the players' utility functions. This is provided by considering the following very simple ideological utility function

$$u_i = \mathbf{x} - (\mathbf{x} - y_i)$$

which satisfies the condition that the numerical representation (coinciding with the numerical values of Nash SWF) of the conformist preference ordering must preserve the property that "a state of affair" is more preferred than another if it is less faraway from the ideal. Then if conformity is complete (both the players conform, and the distance is nil) the additional component is maximal in players' utility functions. If nobody conforms (the distance is maximal), no additional component figures in the utility functions. For intermediate level of conformity – when only one of the players conforms – there will be an intermediate additional component in the utility functions of both the players (this allows that I can get some utility form your conformity without doing my part).

Let now take an outcome - for example (LW,LC) that is what I mean as "state of affairs" - and describe it according to both its consequences to the player i and its conformity to the ideal. Next look at how player i summarizes all this information through the two types of preference he holds – his consequentialist self-refereed personal preferences and his conformist personal preferences. Following the decomposability assumption introduced at the end of sec.4, a natural way to represent all that through player i utility function is simply taking the weighted additive combination of the two numerical representation of the two pieces of player's i preference system for that outcome

$$U_i(LW,LC) = \pi_i(LW,LC) + \lambda\,[\mathbf{x} - (\mathbf{x} - y_1)]$$

which I call player's *i overall utility function* for this "state of affairs", where

- π_i is player *i material payoff* for the self-referred consequences of the strategy vector (LW,LC),
- $\mathbf{x}$ is the value $\text{Max}\Pi_i(U_i - d_i)$ of the Nash SWF defined over the material *payoff* space of the game, that is what I call the *ideal,* taking (0,0) as the *status quo,*
- y_1 is the value of the Nash SWF for the outcome (LW,LC) - i.e. in our example $y_1 = 2\times2\times5$ (see fig. 1).
- $(x - y_1)$ is the measure of the distance of the actual state of affairs form the ideal;
- $\mathbf{x} - (\mathbf{x} - y_1) = u_i$ is the ideological utility attached to state y_1 by player *i*;
- Last, λ is a weight ($\lambda \geq 0$) which says how much conformist considerations count within player *i* overall preferences system.

I do not discard the hypothesis that the weight λ may go to infinitum, so certainly overcoming the material utility of payoffs. But I maintain for most of the time that in fact it takes its values in the interval [0,1] (indeed I do not need to assume more than so to illustrate the present example). Notice that at outcome (LW,LC) there is the maximum value of the Nash SWF in the game at hand. Thus $\mathbf{x} - y_1 = 0$ and the whole value **x,** weighted by λ, enters the utility function of player *i.*[12]

How do the players choose a strategy or another? Only a slight adaption in the rationality rule is required as they will maximise their overall expected utility. Thus for example player 1 (the worker) will choose LW if

$$EU_1(LW) = p\{\pi_1(LW,LC) + \lambda[\mathbf{x} - (\mathbf{x} - y_1)]\} + (1\text{-}p)\{\pi_1(LW,HC) + \lambda\,[\mathbf{x} - (\mathbf{x} - y_2)]\}$$
$$\geq$$
$$EU_1(HS) = p\{\pi_1(HW,LC) + \lambda[\mathbf{x} - (\mathbf{x} - y_3)]\,\} + (1\text{-}p)\{\pi_1(HW,HC) + \lambda\,[\mathbf{x} - (\mathbf{x} - y_4)]\}$$

where p is the probability representing player 1 beliefs over player's 2 LC choice, and 1-p is his belief over player's 2 HC choice. To simplify notation write the four outcomes (LW,LC), (LW,HC), (HW,LC), (HW,HC) as s1, s2, s3, s4 respectively. We know that for player 1

$$\pi_1(s3) > \pi_1(s4) > \pi_1(s1) = \pi_1(s2)$$

This in fact follows from the utility representation of outcomes (state of affairs) according to their consequences to the player 1 and to the consequentialist preference ordering settled by player 1 over them. Moreover we know that

$$[\mathbf{x} - (\mathbf{x} - y_1)] > [\mathbf{x} - (\mathbf{x} - y_3)] = [\mathbf{x} - (\mathbf{x} - y_2)] > [\mathbf{x} - (\mathbf{x} - y_4)]$$

which is the ideological utility ordering of states of affairs, derived form conformist preferences, which in turn is based on the distance of any outcome from the ideal. For given material payoffs and probabilities, representing the player's beliefs on the reciprocal choice of the other player, we are always able to derive the condition such that a player is indifferent between his two strategy choices. For example LW gives player 1 no less overall expected utility than HW if

$$\lambda = \frac{p[\,\pi_1(s3) - \pi_1(s1)] + (1\text{–}p)\,[\,\pi_1(s4) - \pi_1(s2)]}{p\,\{[\,\mathbf{x} - (\mathbf{x} - y_1)] - [\mathbf{x} - (\mathbf{x} - y_3)]\} + (1\text{–}p)\,\{[\,\mathbf{x} - (\mathbf{x} - y_2)] - [\mathbf{x} - (\,\mathbf{x} - y_4)\}}$$

It is telling that the strategic variable is the weight λ to be given to conformist preferences, so that if this weight is large enough, the worker will be ready to choose Low salary even if this strategy is dominated in terms of material payoffs. Notice the complete symmetry of the game, so that identical consideration applies to Player 2 (the entrepreneur). Moreover, to keep things simple, I ignore mixed strategies and I also assume that λ is identical amongst the two players.

Now let see whether new equilibriums do exist due to these moderate changes in the utility functions and rationality criterion. In equilibrium (put aside for the moment the question whether these are in fact equilibriums - the answer will follow from the argument - a player will face a precise choice by the counterpart, which he predicts with certainty (remember that I disregard mixed strategy equilibriums). How do the players reply optimally to any precise prediction of the other's action? Admitting that player 2 chooses the LC strategy (and that player 1 predict it), player 1 will respond optimally by choosing LW if

$$U_1(LS|LC) \geq U_1(HS|LC)$$

which is true when

$$\lambda \geq \frac{\pi_1(s3) - \pi_1(s1)}{[\mathbf{x} - (\mathbf{x} - y_1)] - [\mathbf{x} - (\mathbf{x} - y_3)]} = \lambda^*$$

This is symmetrically true to player 2. For the given values of parameters, at $\lambda^* = 0.33$ an equilibrium exits where both the players choose their "Low" strategy, if they predict that the other player is also choosing his "Low" strategy. At this level of λ the value of the ideological utility (which is at its maximum) overcomes material payoffs that for both the players would be higher under the "High" strategy then under the "Low" strategy. Lower level of λ however implies that HW (and symmetrically HC) remains the player's best reply strategy given the prediction that player 2 chooses strategy LC.

The conditions of player 1 best replies when he predicts (assuming to be in equilibrium) that player 2 is choosing his strategy HC can be calculated similarly. Then he will choose strategy LW if

$$U_1(LW|HC) \geq U_1(HW|HC),$$

that needs

$$\lambda \geq \frac{\pi_1(s4) - \pi_1(s2)}{[\mathbf{x} - (\mathbf{x} - y_2)] - [\mathbf{x} - (\mathbf{x} - y_4)]} = \lambda^{**}$$

The critical weight in this case is $\lambda^{**} = 0.21$. Summing up, when λ exceeds level 0.21, player's 1 strategy of claiming low wage is his best reply against an entrepreneur pretending high costs (symmetrically for players 2 claiming low costs) When eventually λ grows up to 0.33 and over, low salary becomes also the player's 1 best reply against players 2 asking for low costs, that is starting form level $\lambda = 0.33$ the "Low" strategy becomes the dominant strategy of player 1 (symmetrically for player 2).

By considering simultaneously players 1's and player 2's best replies, the equilibria in pure strategies of the game under mixed preferences are computed. The equilibrium set of the game, under the given parameters values, is summarized in Fig. 2.

	LC	HC
LW	*s1 is an equilibrium as from* $\lambda \geq 0.33$	*s2 is an equilibrium as from* $0.33 > \lambda \geq 021$
HW	*s3 is an equilibrium as from* $0.33 > \lambda \geq 0.21$	*s4 is an equilibrium as from* $\lambda < 0.21$

Fig.2 The equilibrium set of the NPE game
(in each box of the matrix is reported the condition over λ such that the corresponding outcome in the matrix of the normal form game of fig.1 is an equilibrium when the utility functions of the players are meant as their overall utility functions, including the conformist component)

To different level of λ several equilibriums come about due to conformist preferences. Low λ (see. Fig.2 bottom/right box) implies that only consequentialist preferences count, therefore maximising the material payoff is still the only equilibrium of the game. High λ implies (see box top/left) that conformism strongly counts throughout the players' overall preferences, thus both players comply with the ideology.

However there are also “strange” equilibria where one player conforms if the other does not conforms (see boxes bottom/left and top/right). This apparently strange result depends on how overall utility functions are defined. They admit that a player gains unilaterally ideological utility even if he conforms against an adversary that does not conforms. At the same time, a player can benefit from his ideological utility in the event that the other conforms even if he doesn’t conform, because in a sense he profits also from the other party level of adherence to the common ideology. This exhibits the possibility that when a player can benefit form his unilateral conformity, due to his ideological utility, this same fact may also push the other party to *free ride* the first party conformity. It may seem realistic that “ideologues” can be exploited by opportunists. Nevertheless it should be excluded in my account for conformist preference. It proceeds in fact not from the idea that conformity is a value *as such*, but from the quite different hypothesis that a player will get conformist satisfaction, i.e. ideological utility, from believing that conformity is the behaviour generally spread over the relevant group of interacting players - that is from believing that when he conforms the other player also conforms.

The “strange” equilibriums can be avoided by making λ an endogenous variable of the model and in particular making it to depend on the level of expectation of reciprocal conformity by each player. Then I assume that λ will jump beyond an upper threshold when the player does in fact believe (i.e. believes an hypothesis more than the opposite hypothesis) that the other player will conform to the ideology, while λ will remain below a lower threshold unless the player does in fact believe that the other party will conform to the ideology.[13]

Hp 3: if player *i* (for $i = 1,2$) believes that the other will conform (probability $p > 0.5$, i.e. more than indifference) then λ is higher than a (upper threshold) level λ^*, if player *i* does not believe that the other will conform (probability $p \leq 0.5$, no higher than indifference*)* then λ is lower that a (lower threshold) levelλ^{**}.

Notice that levels λ^*, and λ^{**} are the thresholds I have found as the conditions, *first*, for the “Low” strategy of a player becomes his best reply against the symmetrical Low strategy of the counterpart and, *second*, for the “Low” strategy be the best reply against the asymmetrical “High” strategy of the counterpart. This means that if a player does not in fact believe that the counterpart is conforming, he does not attribute very strong weight to conformist considerations in his preference system, whereas if the does in fact believe that the counterpart will conform then he feels very strongly the importance of conformist considerations when settling his systems of preferences.

For example, taking the point of view of player 1, I can assume

$$p(LW) > 0.5 \Rightarrow \lambda \geq 0.33$$

$$p(LW) \leq 0.5 \Rightarrow \lambda < 0.21$$

For symmetrical conditions apply to player 2, the values of the parameter λ implying that s2 and s3 are equilibriums will never take place and the equilibrium set shrinks to the only two “plausible” elements s1 and s4. It should be clear that my hypothesis Hp3 is not *ad hoc*. On the contrary it is coherent with the very idea of conformist preference meant as a preference over states of affairs according to the expected “reciprocity in conformity” attached to them. This can only be seen through the expectations that players have on the mutual conformity of the other party when the player himself is planning to conform to the ideology. Therefore, making the weight a player assign to conformism in his preference system resting upon what he expects from the other party, simply express the very idea of conformist preference as preference for expected mutual conformity to the constitutional ideology.

Let summarise the results I have obtained so far and their limitation. Under high enough level of λ an organisational equilibrium exist that induces the worker and the entrepreneur to claim low wage and costs. It allows high level of provision of welfare goods to the advantage of the beneficiary and minimises her transaction costs. Under such an equilibrium the NPE is entirely trustworthy to the beneficiary. It emerges as the efficient NPE as it devolves the most part of the surplus to benefit the beneficiary. However also a second equilibrium would exists if λ were low, where both the players do not conform to the ideology but appropriate opportunistically all the surplus. Our final result in fact is that the equilibrium set of the game includes both an equilibrium of mutual conformity and an equilibrium of mutual deviance from the ideology.

Moreover this is not the usual multiple equilibriums result, because the two equilibriums are not simultaneously possible as they both rest on an appropriate level of λ, which changes with the level of belief the players entertain about reciprocal conformity. Therefore if a value of λ occurs that admits the deviation equilibrium, there is no multiplicity build in the situation, because this is the only equilibrium possible. Everything depends upon mutual beliefs of the players. If they mutually expect high levels of conformity from one another, they both come to assign high weight λ to conformity within their preference system, and in turn this will generate the corresponding equilibrium and will eliminate the competing one. The second equilibrium on the contrary will emerge as the unique solution in the event that players do not expect one another high level of conformity to the ideology. Typically in this situation expectations tends to be self-fulfilling. At the end introducing a constitutional ideology results a necessary but not sufficient condition for the emergence of a socially beneficial equilibrium. Sufficiency rest upon the emergence of the appropriate system of reciprocal expectations.

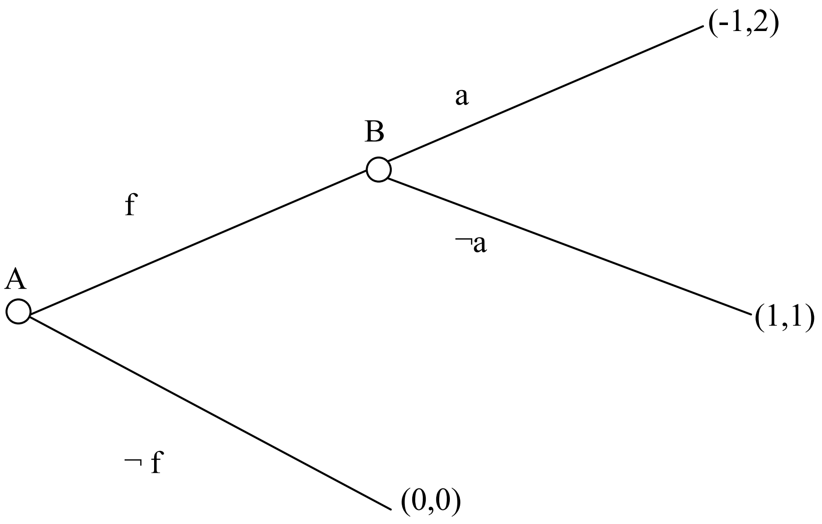
6. Explicit moral code, reputation and the endogenous development of trust

In this section I suggest that a way out of the non standard multiplicity problem posed in the foregoing section is provided by a reputation game between the NPE as a whole and its external stakeholders, that is donors and beneficiaries. I make the statement that in order to establish trust and reputation a code of ethics must be introduced, which is the linkage with the theory of the NPE ideology put forward in the previous part of the paper.

Let start from the hypothesis that the donors and the beneficiaries must decide whether to establish a relationship of trust with the NPE and that at first glance it is not possible for them to trust it. This is the necessary starting point in order to analytically account for how a self-imposed code of ethics might affect the endogenous development of trust. The analysis I'm going to propose will justify the fifth proposition of the paper after those stated in the introduction.

- *Proposition V*: Even in absence of complete reliability of ideological motivations of the internal members of the NPE, donors' and beneficiaries' trust develops, in an endogenous way, by virtue of the cognitive function of the code of ethics that allows the setting up of a game of reputation. It follows that the code of ethics is a self-enforcing set of principles and rules thanks to reputation effects it generates.

In order to make sense of a problem of trust it is necessary to go beyond the simplest market transactions, characterised by complete information among parties, and parametric choices. The problem is well illustrated by the "game of trust" in fig. 3 (Kreps, 1990). In this game player A, for example a consumer, can decide whether to trust *f* or not ¬ *f* player B, a firm. Player B can decide whether to abuse *a* or not to abuse ¬ *a* the trust of player A. From the game tree it is understandable that player's B dominant strategy is to abuse, so that player A's best reply is not to trust. The equilibrium of this simple game is the strategy pair *(¬ f, a),* resulting in the payoff vector (0,0). This means that no transaction takes place.



(Fig. 3 the game of trust)

Even though this is a basic and very widespread situation, actually transactions take place normally. If, in fact, the game of trust is played repeatedly, reputation effects support a result where the transaction takes place without abuse. First of all, it is necessary to assume that the game repeats for an infinite number of times and that there is an infinite series of short-run players $A_1,\dots,A_n$ (where n goes to infinitum), each one taking part to a single stage-game, while player B is a long-run player taking part to every repetition of the basic stage-game. Player's B choice set consists of strategies of the repeated game, which are rules for choosing actions in each stage-game as a function of each history of the game until any stage at which player B must choose. The payoff function are the following: any short-run player (A_i) only receive the payoff of the stage game he takes part in. On the contrary, the long-run player's (B) payoff is meant as the infinite sum of the payoffs he gets each stage. A crucial assumption is that player B is far-sighted, that is he does not discount too much the payoffs of future stages compared to current stage (i.e. player's B discount rate δ for future utilities is close to 1).[14]

Players' beliefs characterise reputation games. The long-run player is perfectly rational (from a strategic point of view) and perfectly informed about the game, the utilities and modes of reasoning of the short-run players. Short-run players, on the contrary, are not perfectly informed on the "type" of player B. By "type" I mean a rule of behaviour that the player B idiosyncratically follows in each stage game according to what the A_i believes about B. Thus a "type" is a commitment to choose a given action under the different contingencies of the game – according to what players A_i believes about B. A "type" is strictly related to the action he chooses in each stage-game. Each player A_i takes as possible different player's B types. Besides the "rational" type, who always chooses the dominant strategy of the stage-game, the type who never chooses to abuse and some types combining abuse and non-abuse randomly are also possible (for example the type that mixes *a* and $\neg a$ with probabilities 0.75 and 0.25 respectively). Thus a positive prior probability must be attached to all these types by any player A_i (in particular the type who never abuses, called "the absolutely honest", has a positive probability, even if very small).[15] Probabilities are updated according to Bayes rule: each stage conditional probabilities of types change as a function of the evidence produced by how the past stage-games have been played by the long-run player.

Thus Player B's reputation is the probability assigned by each player A_i at every stage to player B's different types. Reputation to be a certain type grows as evidences confirming the "type" are collected, but it is drastically lost with a single observation incompatible with the type (this is obviously not true for the probabilistic types). Beliefs affects players' strategy choice. Each Player A_i chooses, according to the expected utility reasoning, between *f* and $\neg f$ in the light of the conditioned probability of B's types. During the first stages of the game the first players A_i does not trust B because expected utility of strategy *f* is lower than the alternative. Eventually (say after N periods), however, a short run player (say A_{N+1}) may begin to give his trust if a series of A_i before him have observed $\neg a$ so many time that the conditioned probability of the "absolutely honest" type results updated to the level p* where the expected utility of *f* becomes higher than $\neg f$.

Player's B optimal choices bring to consider the equilibrium strategies of the repeated game. First of all, player B can decide to choose always the equilibrium strategy of the stage game, that is always *a*. To such strategy the best reply of each player A_i is to go on playing action $\neg f$ at any stage. This leads to a repeated game equilibrium, because nobody has the incentive to deviate from such choices for the whole length of the game. This in fact constitutes the lower boundary of the set of equilibriums of the game (Fudenberg e Levine 1989). Player B, however, has a different strategy at his disposal, consisting of exploiting his awareness of the updating mechanism followed by players $A_1,\ldots,A_n$. He can choose to simulate the behaviour of the "absolutely honest" type until its conditioned probability reaches the critical level p*. At this point he can calculate if it is better to him to continue playing $\neg a$ - consequently inducing over and over choices *f* from players A_i after A_{N+1} -or to defect by choosing *a*, whose result is 2, on the first time, followed by a series of payoffs 0. If δ is close to 1 (that is player B is not short-sighted), then the infinite sum of payoffs 1, even if discounted, will more than counterbalance a single chance to win payoff 2.

The best reply to such strategy by each short-run player will be exactly the one foreseen by player B: from period N + 1, once the type has reached probability p*, they will trust and will continue to do so until they see a defection. For δ close to 1, simulating the "absolutely honest" type leads to an equilibrium, which is the upper boundary of the equilibrium set of the game (Fudenberg and Levine 1989). Thus there is an equilibrium profile within which the long-run player B can build an adequate level of reputation such that a fiduciary relationship emerges between the long run player and the short-run players after N periods spent to accumulate reputation..

Let sum up the hypothesis needed for this result be true. Besides player's B far-sightedness (δ close to 1) it is necessary that (i) each stage-game ends up with a couple of observable actions by the players, that is B chooses his action in each stage game even when A_i does not trust him (i.e. the stage game must be considered as a simultaneous move game). (ii) Each player A_i must be able to observe and learn, without any ambiguity, the result of the stage-game in which he takes part; he must be able to update the conditioned probabilities of the types and to communicated that probability to the adiacent player A_{i+1}. This must be true in particular for every pair of adjacent players A_i e A_{i+1}. Whether each player A_i were not able to observe the result or incapable to infer from the observed result the meaning of the action made by player B (that is to recognize whether action *a* or ¬*a* has been played), then reputation would not be up-dated and the mechanism would break-down.

Transactions take place in many markets. This suggests that the problem of trust between parties can be "spontaneously" solved. Even in the presence of implicit contracts, transactions can be supported by the mechanism of reputation if such economic situations satisfy the hypothesis mentioned in the model. In these cases trust develops without the settlement of fiduciary duties, codes of ethics and the like. Fiduciary duties, codes of ethics and deontologies, on the contrary, points out cases in which fiduciary relationships are not spontaneously supported by the simple mechanism of reputation (Flannigan 1989, Frankel 1999).

Social and welfare goods fall within the just mentioned contexts. Let consider, for example, the case where player's A position in the game of trust is taken by a donor, who needs to trust the firm producing welfare goods in favour of beneficiaries that are completely separated from the donor himself. In this case, it is obvious that the donor does not observe the result of the producer's activity, that is to say he does not observe the result of each stage of the game because the consequences of the choice (abusing or non-abusing of the donor's trust by an effective use - or not - of the available resources) falls on a third party - a case of "credence goods" (Darby and Karni 1973) . Therefore condition (ii) is violated.

The problem is however much wider than so, as contract incompleteness in inherent to the provision of welfare goods[16]. By contract incompleteness I mean that the firm's (player B) action and its expected result are not ex ante specified, neither explicitly nor implicitly, by contingent clauses on each ex post possible but ex ante unforeseen state of the world. This is due to the fact that some of these events are genuinely unforeseen by the parties. This means that, in unforeseen states of the world, firm's commitments have not even been specified and, consequently, the contract in these cases is simply "mute". In terms of the reputation model, commitments corresponding to the various types are not ex ante specified.

Actually, this means that by observing ex post the outcomes, even if this is possible, players can not understand if player B chooses action a or ¬a. This is not due to simple statistic uncertainty but to the very fact that the meaning of these acts in those states of the world was not ex-ante specified. The meaning of an act depends in fact not only on the description of the payoffs, but also on the description of the state upon which it is contingent. Take for example an action that contingently upon an ex ante known state means "abuse". When seen as depending on unforeseen states of the world, the sense of the same action may become ambiguous. Even though you observe the same payoff as before, it does not mean that the action is "abuse" if the situation differs completely and some features are genuinely unexpected.[17]

Generally speaking, the reputation mechanism depends on the fact that each player Ai can say about each type if "what had to be done, has been done" at every stage of the game (Kreps 1990). That is the player needs to understand what the commitment requires at each state, and he must be able to verify from the outcome at each stage whether the commitment has been complied with. It is clear however that ex ante nothing is specified by an incomplete contract as far as unforeseen states are concerned. Thus, neither the ex post description of the outcome nor of mere labelling of the action might tell him for each type whether in an unforeseen state of the world "what had to be done, has been done". In this case the simple mechanism of reputation does not apply.

My thesis is that a code of ethics can fill the gap in the players' expectations: the code of ethics generates expectations where contract incompleteness makes commitments mute[18]. A code is a cognitive and deliberative device which is build up out of two parts: (i) general and abstract principles; (ii) preventive rules of conducts. Concerning the first, general principles identify moral properties associated to abstract and universalizable characteristics which are not necessarily bound to a complete description of every concrete contingencies that might occur in all the states. Hence, in order to find out the characteristic identified by the principle, we do not need to look after a complete and detailed description of all the characteristics of the possible states of the world. Moreover unforeseen contingencies will always belong to some degree to the domain of application of a general principle, although their belonging can be a matter of vagueness. It is exactly the abstractness of principles that makes possible their application to every situations, even if these are ex ante unforeseen, while concrete rules, contingent upon detailed state description, would be simply mute.

Vagueness of course is the price to be paid for being able "to say something" with respect also to the unforeseen states of the world. But it is worth to be paid. The resulting situation is in fact completely different from the one occurring when expectations concerning the unexpected state are completely undetermined. In this case, we can try to manage vagueness by a measure of membership of each state into the domain of application of the general principle. I suggest that a fuzzy membership function, taking it values in the real interval [0,1] can be defined for each unforeseen state that will occur (or reveals to be possible ex post), which implies that the domain of application of a principle must be understood as a *fuzzy set*[19],

Summing up, whether our language contains only concrete descriptions of actions contingent upon ex ante known states, then the occurrence of unforeseen states would make impossible to say what commitments would require at these states. Quite differently, we can use general and abstract properties to whom unforeseen states adapt at least imperfectly. In that case, by default reasoning, we may infer that an unforeseen state with a degree of membership to the principle at least equal to α (for $1<\alpha<0$) is "normally" treated as an exemplar of the principle. Thus we conclude by the same mode of reasoning, even if we have not a complete proof that this is the case, that a rule of behaviour conforming to the principle must be carried out. The logic at work here is "default reasoning" as based on fuzzy semantics. It is the better we can reasonably do in order to face unforeseen contingencies that testify how limited human rationality is (Reiter 1980, Geffner 1993, Sacconi 2000, 2001).

The second part of the cognitive device here has just entered the scene. If a state belongs to the domain of a principle at least to degree α, then we can conclude, by default, that "normally" in such situations a given practical rule of conduct does apply. As it is simply a procedure (not an outcome), this practical rule can be ex ante described without being contingent upon ay ex ante predictions of states or the related outcomes. Therefore we have certain standardized characteristics, which we can ex post say whether have been carried out or not in actual behaviour. What really matters is that a code allows to specify ex ante the conditions under which a certain procedure must be carried out ex post, without any reference to a concrete description of the details of any state of the world. Such conditions in fact consist essentially in that whichever situation (even if ex ante unforeseen) must belongs at least to a certain degree to the domain of application of a principle. Thus it is possible ex ante to undertake commitments and generate expectations on future behaviours without any detailed knowledge or forecast of future states of the world, which may be left at least in part unforeseen. [20]

No doubt, such preventive rules of conduct will not permit maximizing utilities in every states. It is implicit in the non-monotonicity of default reasoning that from vague information we are allowed to conclude that "normally" cases "such and such" are to be managed according to a certain procedure, but also that in the presence of more information that conclusion may have to be revised. [21] The main aspect however is that principles allow us at least provisorly to "complete the contract" by the specification of what we may expect with respect to commitments that are ex ante determined in terms of compliance with procedures, given certain conditions on the degree of membership into the fuzzy domain of a general principle. Rules of conduct will then give a reliable base in order to decide if "what had to be done has been done". This permits to apply again the mechanism of reputation effects.

The "absolutely honest" type of the firm has to be replaced by the type who conforms to a rule of conduct when the code of ethic asks to do that, that is in all the contexts where the ex ante announced conditions occur. Donors and beneficiaries can observe what happens (aside from the procedure being optimal in the case in point) and then evaluate the reputation of the organization. Despite contract incompleteness, a strategy of compliance with the code of ethics makes to the NPE possible to accumulate reputation.

7. The virtuous interplay between internal conformist motivations and external reputation of the NPE

I can sum up my overall account for the role of ideology in the efficiency of the NPE. It results from the interplay between the cognitive role of the code of ethics in supporting the fiduciary relationship of the organization with its external stakeholders and the motivational role of ideology in generating conformist preferences amongst active agents inside the organization itself. The following proposition summarizes it:

- *Proposition VI*: Assume that the same "social constract" model is the basis for the ideology and the code of ethics of the NPE. Then, on one hand, the NPE constitutional ideology, by means of conformist motivations, allows to predict the existence of an efficient organizational equilibrium in the production of a welfare good. On the other hand, the code of ethics allows predicting the existence of a reputation equilibrium in the interaction between the NPE and its external stakeholders. These two elements mutually support one another as the one equilibrium contributes to raising the conditions under which the second equilibrium can be proved to exist and *vice versa*. Reputation equilibrium from outside favours the development of mutual beliefs among the inside members of the organization and bring about the system of expectations such that conformist motivations become effective (i.e. make effective the organizational equilibrium). The other way round, the reputation equilibrium between the firm and its stakeholders is facilitated by the existence of conformist preferences of the internal NPE members, because at least they justify positive, even if small, prior probabilities attached to the "ethical" type of NPE.

I split the argument in six steps. First of all, we have seen that the NPE is the most suitable institutional form in order to attract ideological entrepreneurs and workers. That is the internal members of the NPE at least hypothetically agree on a set of constitutional principles (the ideology) able to elicit the rational ex-ante acceptance by all the NPE's members.

Second, such hypothetical agreement produces a conformist motivation among the internal active members of the organization, in that the agents develop a preference in favour of mutual compliance with the ideology they are ready to accept. The weight of such conformist preference however depends on mutual expectations, and ex ante we can not say whether an expectations system such that conformist preferences will be strong enough to bring the organization on the path of the conformist equilibrium will develop.

Third, even if the previous step is not sufficient in order to predict which equilibrium is going to emerge from the internal interaction, conformist utility can grant that any reasonable player assigns at least a small positive probability to the fact that the firm as a whole will conform to its constitutional ideology. In fact this is simply a function of the positive probability that each of the internal (active) players will also conform.[22]

Fourth, the donors' and beneficiaries' level of trust towards the organization reflects the fact that the constitutional ideology is not sufficient by itself to assure that the equilibrium behaviour of the firm will be consistent with ideology. External stakeholders are uncertain as far as the type of the organization is concerned (if conformist, or non conformist). Nevertheless step three tells us that the existence of a psychological primitive motivation in favour of mutual conformity to ideology would justify that external stakeholders assign a positive, even if small, probability to the conformist type of organization in a repeated game, notwithstanding that in the stage game abusing the donor's and the beneficiary's trust is the dominant strategy. Such a positive probability is the basic condition for starting the reputation effects mechanism.

Fifth, an explicit code of ethics works as a substitute for concrete commitments in contexts characterised by contract incompleteness and unforeseen contingencies - situations typically requiring fiduciary relationships between the welfare good's producers and their stakeholders. An explicit code of ethics aids to classify unforeseen contingency according to their fuzzy membership relation to general principles taken as an abstract term of reference (pattern recognition). These principles are the same principles of fairness we have found at the basis of the constitutional ideology of the NPE. The code also allows judging whether, in the presence of an unforeseen event, some preventive behavioural standard has been complied with. Such rules of behaviour are seen by default as obligatory once it is recognized that a state belongs to the fuzzy set defining the application dominion of the principle up to a certain degree (threshold α). At last the NPE's reputation amount to no more than saying whether the NPE each stage of a repeated game sticks to its code of ethics.

Sixth, the reputation mechanism implies that an equilibrium of trust between the organization and its stakeholders does exist, as it is self-enforced by reputation effects themselves. Thus, under a code of ethics, the organization has the appropriate incentive to assure compliance to its own code of ethics, in order to support donors' and beneficiaries' trust. Seen from inside, this fact will mean that all the internal members of the organisation, or most of them, will strive to assure the appropriate level of reputation toward the stakeholders by up-holding the organisation's code of ethics. As a matter of consequence, general conformity to the code will foster the beliefs that any members entertain about the other member's conformity to the ideology expressed into the code of ethics.

Summing up, different parts of the theory mutually support each other. The model of internal organisational relationships (allowing for the existence of a conformist equilibrium contingent upon beliefs) is exploited as a justification of the initial conditions for the reputation game model of external relationships between the organization and its stakeholders. At the same time the result of the reputation model of relationships between the NPE and the external stakeholders is used to complete the solution of the internal interaction, by moving the conformist equilibrium from a merely "virtual" existence to actual effectivity.

The reputation model in fact usually suffers the arbitrariness of the assumption that players believe "types" that carry out idiosyncratic behaviours, which are not even compatible with the current theory of strategic rationality. In my context, however there are double justifications for that. On one side, members of the NPE have a constitutional ideology embodied by the code of ethics and this justifies that an "ideological type" be represented in the mind of donors and beneficiaries as a possible model of behaviour for the organisation. On the other side, the mere existence of conformist preferences gives some psychological basis to the incomplete "salience" of the ideological type also in the eyes of the external stakeholders.[23] Therefore, it is no more arbitrary that reputation game models assume that players will update their conditioned probabilities starting from "a priori" positive probabilities over "types" of the NPE. Thereafter the reputation effects mechanism explains how happens that donors and beneficiaries develop an "effective trust" towards the NPE because the NPE does comply with its ideology in order to support reputation towards stakeholders.

But now the equilibrium in the relationships among NPE and its external stakeholder reflects also inside the NPE. As the organization is going to comply with its code of ethics, because of the reputation effects, then any player can also predict that players inside the organisation, or at least most of them, are going to conform to the ideology. From the assumption that the organization as a whole chooses the "do not abuse" strategy, then it follows that the probability of conformist actions on the part of any internal active player grows up to a certain threshold. For example, in the case of the entrepreneur, $p(LC|\neg a) \geq \lambda^*$ - where λ^* is the critical value for the existence of the equilibrium in conformist strategies in the game among internal members of the NPE.

The same reality, the same observable behaviour of the NPE, can be explained from two coexistent and compatible points of view. As far as external interactions are concerned, the organization complies with its code of ethics in order to support reputation effects toward donors and beneficiaries. In the coexistent perspective of the internal game (where beneficiaries are dummy players), organization members conform to the ideology, which is also the content of the organisational code of ethics, because of their expectations over their mutual behaviour and the associated conformist preferences. That the organisation as a whole seeks its reputation on one hand, and the organisation members act according to their internal motivational force, given their expectation of mutual conformity to an ideology, on the other hand, are two explanations not contradictory but mutually supporting one another. This should be seen as a success in enlarging our understanding and accounting for phenomena usually seen as falling short of the scope of rational choice models.

References

AKERLOF G. A. (1984), “Gift Exchange and Efficiency-wage Theory”, in *American Economic Review, vol.74, n.2,* Papers and proceedings.

BEN-NER A. (1986), "Non-profit Organisations: Why Do They Exist in Market Economies?", in ROSE-ACKERMAN, S. , *The Economics of Non-profit Institutions*, chapter 5, pp. 94-113, Oxford University Press, New York.

BEN-NER A., L.PUTTERNAM (eds.) (1998), *Economics, Values and Organization*, Cambridge U.P., Cambridge.

BERNHEIM D. (1994), “A Theory of Conformity”, *Journal of Political Economy*, 102., 5, pp.841-877.

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

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

BROOME J. (1999), *Ethics out of Economics*, Cambridge U.P., Cambridge

COLEMAN J. (1995), *Risks and Wrongs*, Cambridge U.P., Cambridge.

DA COSTA WERLANG S.R., TAN T.C. (1988), The Bayesian Foundations of Solution Concepts of Games, *Journal of Economic Theory* 45(2), 21

DARBY M.R., KARNI E. (1973), “Free competition and the optimal amount of fraud*”, Journal of Law and Economics*, 16, pp.67-88

FLANNIGAN R. (1989), “The Fiduciary Obligation”*, Oxford Journal of Legal Studies*, 9, pp.285-294.

FRANKEL T. (1998), “Fiduciary Duties”, *The New Palgrave Dictionary of Economics and the Law,* (P.Newman ed.), McMillan, London, vol. II, pp.127-132.

FREY B. S. (1997), *Not Just for the money, An economic theory of personal motivation*, Edward Elgar.

FUDENBERG D., LEVINE D (1989), “Reputation and equilibrium selection in games with a patient player”, *Econometrica*, 57, 759-778.

FUDENBERG D., TIROLE J. (1991), *Game theory*, MIT Press, Cambridge Mass.

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

GEANAKOPLOS J., PEARCE D. , STACCHETTI E.(1989), “Psychological Games and Sequential Rationality”, *Games and Economic Behaviours*, 1,

GEFFNER H. (1992), *Default Reasoning, Causal and Conditional Theories*, MIT Press, Cambridge Mass.

GRIMALDA G., SACCONI L.(2002), *The Constitution of the Non profit Enterprise*, LIUC papers, Serie etica, diritto ed economia (in stampa).

GROSSMAN S. E , HART O. (1986), “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration”, *Journal of political Economy*, 94, pp.691-719.

HANSMANN H. B., (1980), "The Role of Non-profit Enterprise", *The Yale Law Journal*, vol. 89, n. 5, 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*

HARSANYI J.C. (1977), *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*, Cambridge U.P., Cambridge.

HART O. (1993), "An Economist View of Fiduciary Duty", *University of Toronto Law Journal*, XLIII, 3, Summer, pp.299-314.

HART O. (1995) *Firm, Contracts and Financial Structure*, Clarendon Press, Oxford,.

HART O., MOORE J. (1999), "Foundation of Incomplete Contracts", *Review of Economics Studies*, 66, pp.115-138.

KRASHINSKY M. (1997), "Stakeholder Theories of the Non-profit Sector: One Cut at the Economic Literature", *Voluntas*, 8, 2, pp.149-161.

KREPS D. MILGROM P. ROBERT J., WILSON R. (1982), "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", *Journal of Economic Theory,* 27, pp.245-252.

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

KREPS D.M. (1997) "Intrinsic Motivation and Extrinsic Incentives" , *Amer. Economic Review, AEA Papers and Proceedings,* vol.87, N.2, pp.359-364.

LEWIS D. (1969), *Convention, A Philosophical Study*, Cambridge, Mass. 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

NASH J. (1950) "The Bargaining Problem" *Econometrica*, 18, pp.155-162.

ORTMANN A, SCHLESINGER M. (1997), Trust, Repute and the Role of Non-pofit Enterprise, *Voluntas*, 8,2, pp.97-119.

PESTOFF V. (1998), *Beyond the Market and the State, Social Enterprises and Civil Democracy in a Welfare Society*, Adelsrhot & Brookfield, Ashgate.

RABIN M. (1983) "Incorporating Fairness into game theory and Economics" in *American Economic Review*, 83, 5.

RABIN M. (1998), "Psychology and Economics", *Journal of Economic Literature* vol.36, March, pp.11-46.

RAWLS J. (1971), *A Theory of Justice*, Oxford U.P

REITER R. (1980), "A Logic for Default Reasoning", *Artificial Intelligence*, 13, pp.81-132.

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

ROSE-ACKERMAN S. (1996), "Altruism, Nonprofits, and Economic Theory", in *Journal of Economic Literature*, vol. XXXIV, n. 2, pp. 701-728.

ROSE-AKERMAN S., (1986), *The Economics of Non-profit Institutions*, Oxford University Press, New York.

SACCONI L. (1995) "Considerazioni sulla possibilità del vincolo morale razionale" in S.Veca e S.Maffettone (eds.) *Filosofa , Politica e Società, Annali di etica pubblica,* Donzelli, Roma.

SACCONI L. (2000), *The social contract of the firm*, Springer, Berlin-Heidelberg.

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

SACCONI L.(2001), *Incomplete Contracts and Corporate Ethics: A Game Theoretical Model under Fuzzy Information*, LIUC Papers, n. 91, Serie etica,diritto ed economia.

SACCONI L., MORETTI S (2002), *Fuzzy Norms, Default Reasoning and Equilibrium Selection in Games under Unforeseen Contingencies*, LIUC Papers, n.104, Serie etica, diritto ed economia.

SUGDEN R. (1984), "Reciprocity: the supply of public goods through voluntary contributions", *Economic Journal*, 94, 376, pp.772-87.

SUGDEN R. (1998a) “Normative expectations: the simultaneous evolution of institutions and norms” in A. Ben-Ner and L. Putternam, *Economics, Values and Organization*, Cambridge U.P.

SUGDEN R. (1998b), *The motivating power of expectations, UEA,* School of Economic and Social Studies, MIMEO.

SUGDEN R. (2001), *Sociality and the correspondence of sentiments, UEA,* School of Economic and Social Studies, MIMEO.

TIROLE J., MASKIN E. (1999), “Unforeseen Coontingiencies and Incomplete Contracts”, *Review of Economics Studies*, 66, pp.83-114.

WEISBROD B. A., (1988), *The Non-profit Economy*, Harvard University Press, Cambridge.

ZIMMERMAN H.J. (1991) *Fuzzy Set Theory and its Applications*, Kluwer, Academic Press, Dordrecht – Boston.

Notes

[1] I think that it is essentially coherent with the *formal* notion of utility as a representation of a general *betterness* relation, suggested by John Broome (Broome 1999), which is much more wider than both the traditional notions of preference as coherent desire or as revealed preference, and permits to appreciate its formal conditions of coherence without tying up them with a particular and questionable substantial interpretation of the good.

[2] On fiduciary relationships and duties see Flannigan (1989) and Frankel (1998).

[3] This is a point I draw from David Gauthier (Gauthier 1986), as 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, which he attempts to solve by his "constrained maximisation" theory. While I agree with the contractarian approach (see Sacconi 1991, Sacconi 2000), I do not believe that Gauthier's constrained maximisation is completely successful (see Sacconi 1995) in explaining rational compliance. This is a reason for looking after a theory of conformist preference, at least in order to explain certain class of economic behaviours like the comparatively successful performance of NPE within the provision of welfare goods, as it asks for a restraint on the players' individual self-interest. Nevertheless, I agree with Gauthier that compliance – what in my approach should be called conformity – is a distinct problem that must be considered separately from the rationally justificatory force of the social contract as seen in the ex ante perspective taken by who have to decide whether he would join to a prospective cooperative social venture.

[4] According to Nash Product, if d_i is the status quo from which a generic party i may enter the agreement and u_i is his utility for any given agreement, then the rational bargaining solution is the unique point on the convex frontier of the convex-compact payoff space where the net payoffs of the players $(u_i - d_i)$ are such that the value $\mathrm{Max}\Pi^N_i(u_i - d_i)$ is obtained (where N is the number of players), see (Nash 1950, Harsanyi 1977). Notice that in hour game the Nash product is maximised at (LW,LC), while at (HW,HC) it is nil, due to the component zero in the corresponding payoffs vector.

[5] 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 does exist 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 himself. This is natural given the underlying contractarian account of the Nash Bargaining Solution.

[6] Here it should become clear the main difference between my approach (see also Grimalda and Sacconi 2002) and Sugden's approach (see Sugden 1998a). According to Sugden in fact there is no any independent normative condition for what he calls normative expectations seen as source of additional utility deriving from the common reciprocal expectation of conformity to the same rule of behaviour by a set of players. In a game may exist multiple regularities of behaviour to whom are associated coordination equilibriums, which can be seen as convections. For each of them can develop a conformist source of utility such that it stabilizes the convention itself, also against mistaken deviations by one or a minority of the players' population members. As far as there are common reciprocal expectation that players will follow them, and nobody sees its utility to decline with respect to the expected utility in the case of complete compliance, each rule of behaviour develops its supporting normative expectations (Sugden 1998b), also – I remark - those who are morally repugnant. In my model, which under this respect is more akin to Rabin (1983) on the contrary conformity is an additional source of utility only with reference to an abstract norm of fairness that has been hypothetically agreed upon by the players. Rational ex ante acceptation is not sufficient as such to assure motivational force able to overcome *akrasia.* Nevertheless, it is a necessary condition for a motivation to act resting on mutual conformity to a norm does arise. In other word, we do not gain any additional source of motivation by seeing ex post that a norm has got general conformity and by adhering to it if that norm has not been ex ante accepted through an impartial agreement.

[7] Harsanyi's "preference utilitarianism" (Harsanyi 1977) fundamentally shows that the Utilitarian Social Welfare Function, meant as an impersonal moral judgment of an individual, is no more than an extension of the typical vNM personal utility function, expressing self-referred consequentialist preference, which, under suitable additional assumptions, is defined over the extended set of consequences.

[8] 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.

[9] Hierarchies of beliefs are typical game theoretical constructions after David Lewis' 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).

[10] As far as conformist preferences can be 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. Conformist preference as *betterness* relationship (which is a quite formal and non interpreted notion) is coherent. Therefore 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. .

[11] For a more precise approach to modelling how a measure of conformity enters the utility function of each player, see Grimalda and Sacconi (2002). There we develop the idea that each player will consider the distance between the ideal state (where the SWF is at a maximum) and the result separately determined by each player (himself and the second player) by choosing any strategy available to him given his belief on the strategy that he expects is used by the other player (the second player and himself in turn). If both the players will conform to a certain positive level (which means that each player, when making his choice, does conform and expects that the other conforms as well) the value of the index resulting form these calculation will be positive. If just one player does not conform at all the index degenerate to nil. If according to what is envisaged by a player, both the players conform to the highest level (expecting the same behaviour on the other party) the index has value 1, which at the end means that everything rests on a weight λ independently assigned to conformity by each player. In this case the SWF does not enter as such the utility function. Moreover we can express the interplay between what a player does, given what he expects, and what a player believes the other party will do, given what he believes on his own (as seen from the first party). This is a better modelling of the idea of mutual expectation of reciprocity in conformity. However, the formalism become more complex, while the results and the basic intuition remains the same. Thus, in this paper I keep to the simpler formalisation, which seems appropriate to the task of simply illustrating an example. The inelegance of entering directly the value of Nash Product within the utility function of a player (one may wonder what could mean introducing the product of the utility of different players within the utility function representing each player preferences) could be avoided by taking x^{Max} and x^{Min} to be the maximum and minimum value of the Nash SWF in the game and y the value of the same function defined for generic outcomes and letting the representation of the dependence of ideological utility upon the distance from complete conformity, to take the form

$$u_i = \frac{x^{Max} - y}{x^{Max} - x^{Min}}$$

Then, in the case of complete conformity ideological utility is 1, and 0 in the case of no conformity at all. Therefore under this modelling all would rest on the weight $\lambda \geq 0$ attached to conformist preference in the player’s overall utility function.

[12] Similarly can be written the overall utility of the other outcomes of the game. For example

$$U_i(HW,HC) = \pi_i\,(HW,HC) + \lambda\,[\mathbf{x} - (\mathbf{x} - y_4)]$$

is the overall utility of player i in the state of affairs where conformity is minimal. The distance from the ideal in this case is at its maximum, in fact $y_4 = 4.5 \times 4.5 \times 0 = 0$ and $(\mathbf{x} - y_4) = \mathbf{x}$. As a result no additional value enters the utility function .

[13] Upper and lower thresholds shall not in general coincide. In correspondence of the jump λ is a continuous but not differentiable function of the probability assignment of the other player’ strategies, an inelegance that at this primitive level of formalisation can be excused.

[14] For a synthesis of the theory of games of reputation, see Fudenberg e Tirole (1991) chapter 9.

[15] Luciano Andreozzi (private communication) remarked that the result depends heavily on the *types* admitted and that without restrictions on the type set the Stakelberg strategy of player B is a mixed strategy combining *abuse* and *not abuse*. Therefore, I make clear in the following the hypothesis that I’m a implicitly assuming in the argument here and elsewhere (see Sacconi 2000, 2001). Even if every player’s B types (i.e. every probability mixture of the two pure strategies) could be possible in principle, I do not see any reason for the players A_i thinking about the player’s B idiosyncratic modes of play, should in fact account for all these mathematical constructions. Only *some* of these should reasonably be considered, that is those who “nearly” adopt strategy *a* and those who “nearly” adopt strategy ¬*a*. (Equivalently this point can be made by assuming that only mixed types where the probability mass is nearly all concentrated on *a* or ¬*a* have positive prior probabilities according to the players $A_1,\ldots,A_n$). This implies that player’s B Stackelberg equilibrium strategy will not coincide with a mixed strategy that assigns substantial probability to both the two player’s B pure strategies (*a, ¬a*), for example the mixed stategy (2/3,1/3). On the contrary it will coincide with strategy ¬*a* or some strategies giving very high probability to ¬*a*. In fact in the example in which the only mixed type is (0.75, 0.25), the Stackelberg equilibrium strategy of players B is ¬*a*. The reason for skipping more uniform mixtures of the two strategies is that types must represent *commitments*. Player A sees B as a player sticking to the rule of behaviour derived from the “rational” solution of the stage-game, or nearly so, but also admits on the other hand that he may be *committed* to not abusing at all, or nearly so. While these rules of behaviour are understandable “commitments” it seems to me a “non sense” the utterance that “player B is committed to act or not to act, with nearly the same probability” or - to say - with probability (2/3, 1/3). This should be better understood as avoiding of committing oneself at all and remaining free to act arbitrarily.

[16] On the theory of incomplete contracts see Grossman and Hart (1986), Hart and Moore (1999), see also Sacconi (2000).

[17] This suggests that genuine unforeseen contingencies can not be handled by means of the hypothesis of "state neutrality "of payoffs, as it is assumed in the criticism of incomplete contracts theory put forward by Tirole and Maskin (1999).

[18] This argument was implicitly suggested by Kreps (1990) and developed in Sacconi (2000, 2001)

[19] On *fuzzy set* see Zimmerman (1991).

[20] See Sacconi 2000 chap.8 and Sacconi (2001).

[21] On *Non-monotonic logic* see Ginsberg (1987). For an application to game theory and particularly on problems of equilibrium selection , see Sacconi and Moretti (2002).

[22] Possibly someone would find here the risk of a composition failure. But I do not see why if an external observer would assigns some, even if small, positive probability to the fact that each internal member of the NPE will conform, then he should not also assign a positive, even if possibly smaller, probability to the fact that the NPE as a whole will also conform to its code of ethics, which is based on the same "constitutional contract" ideology.

[23] Let argue this point a little bit more in depth. The weight λ is not fixed because it depends on reciprocal beliefs between the internal players. I want to suggest that they are to be conditioned on the simultaneous development of the external game amongst the NPE and its stakeholders. Were λ fixed from the outset, the equilibrium would definitely coincide with the conformist one or the non conformist one (the situation would reduce to any other game of reputation). But external stakeholders do not need to know (nor could they intrinsically know) the reciprocal beliefs of the internal players. They can only elicit a subjective probability distribution from seeing that the NPE has an ideology, given that this implies conformist preferences, aside of the actual weight the internal players attach to this component of their utility function. In fact I do not ask that the external stakeholders are convinced from the outset that this additional component will be strong enough to dictate full compliance with the ideology by the internal players. I simply state that ideology and conformist preference explains why the stakeholders' prior on compliant the type of NPE may be positive.