CSR AS CONTRACTARIAN MODEL OF MULTI-STAKEHOLDER CORPORATE GOVERNANCE AND THE GAME-THEORY OF ITS IMPLEMENTATION ♦

by

Lorenzo Sacconi

Department of Economics, University of Trento and

EconomEtica, Interuniversity centre of research at University Milano-Bicocca

ABSTRACT

Corporate Social Responsibility (CSR) is here defined as a multi-stakeholder model of corporate governance and fiduciary duties naturally emerging from a critical assessment of the incomplete contracts view of the firm based on concepts like as authority and residual rights of control. As far as the normative point of view is concerned, multi-stakeholder fiduciary duties are deduced from a theory of the firm's stakeholders Social Contract. This provide for a clear cut and calculable objective function, a criterion for governance and strategic management no less able to set a bottom-line to the firm management than the profit maximization principle. The theory of co-operative bargaining games, and the Nash bargaining solution in particular, provides the key concepts. By the way this also answers some criticisms raised by Michael Jensen (2001) against the notion of stakeholders value.

As far as implementation of the normative model is concerned, four roles of voluntary but explicit CSR norms or social standard are presented in terms of a non-cooperative game theory of implementation. It is shown that they allow the description of strategies and equilibria, even if multiple, in a game played under unforeseen contingencies. Secondly, a CSR norm permits the ex ante selection of the equilibrium point that meets the requirements of an impartial choice. An explicit agreement on a contractarian norm is moreover a way to introduce psychological conformist equilibria, and quite surprisingly to derive the significant result that mixed strategy equilibria are absent in a psychological repeated Trust Game. Lastly, a cognitive and predictive role is played by an agreed CSR norm as the appropriate starting point for an equilibrium selection mechanism that, from a state of predictive uncertainty about possible equilibria, generates a state of mutually consistent expectations consistent with the prediction that all players will converge on the psychological equilibrium fully conforming with the norm as the effective solution of the game.

Keywords: corporate social responsibility,stakeholders, incomplete contracts, social norms, reputations, psychological games, conformity, equilibrium selection

♦ I'm grateful to Masahiko Aoki for valuable comments to the first version of this paper that suggested me to provide the substantial extension of the second part. Usual disclaims apply.
This paper is part of a larger research project (PRIN) on "CSR and corporate governance" supported by the MIUR.

1. Introduction

Aim of this essay is pointing out a full fledged theory of corporate social responsibility (CSR) seen as both normatively convincing and implementable model of multi-stakeholder corporate governance. My use of 'normative' is here in the same vein of welfare economics and economic ethics -as it is concerned with proposals of desirable reforms of economic institutions that improve social wellbeing and consistency with social justice and fairness according to an ethical assessment based on the idea of impartial *ex ante* unanimous agreement (i.e. the *social contract*)[1]. Questions about implementation of a given normative model, like as what kind of rules should provide for it, what incentives and individual motivations support it, whether a legal enforcement is required or not, and whether it may rest on self-enforceability - are important as well. I will come to a theory of implementation in the second part of this essay, where I develop a multiple attack to the problem of endogenity and self-sustainably of a CSR norm or social standard understood according to the normative model. But, of course, the definition of the normative model must come first.

In this perspective the main questions that a CSR theory is to answer are whether the proposed normative model is uniquely defined and sets out a clear-cut normative meaning, how can it solve possible clashes amongst legitimate claims and interests that risk to make its implications ambiguous, whether it is impartially justified and has the capability to induce efficient level of economic investments, what its impacts might be on values like as economic welfare, distributive justice and social stability.

As long as this essay argues in favor of CSR as a normative model of corporate governance in the multi-stakeholder and multi-fiduciary perspective, there are however also some specific and well know challenges that must be squarely faced – which as a whole I refer to as Jensen's challenges to the stakeholder model (see Jensen 2001). First of all, it is a commonly accepted prejudice to say that, due to multidimensionality in the objectives pursued by a company adhering to the stakeholder approach, its objective-function must be necessarily ill defined and incapable to endow managers and directors with a clear normative goal able of directing their conduct or providing for a definite bottom line whereby their performance becomes assessable. Moreover, the lack of a unique goal would also improperly enlarge managerial discretion and make the board's fiduciary duties and accountability owed to shareholders devoid of any precise

[1] On the two main, somewhat alternative, understanding of the contractarian program in contemporary political philosophy cf. Rawls (1971) and Gauthier (1986)

content, whereby it may be concluded that the indefiniteness of a multiple objective-function opens the route to managerial opportunistic manipulation and self-dealing.

On the contrary, I will show that the CSR model of corporate governance rests on a well defined company objective-function, at least as well precise as the 'profit maximisation' goal that microeconomic theory traditionally attributes to the firm. It is worthwhile noting that such a statement is made without giving up the methodological individualist tenet typical of economic theory that the corporate goal must not be defined as an attribute of the corporation seen as a holistic and collectivist entity on its own, but should be reduced to goals and interests of its constituencies -that is in my view the corporate stakeholders. Moreover, in so far as the CSR model derives the corporate objective-function from a social contract theory of the firm, managers and directors cannot behave with arbitrary discretion. On the contrary they result constrained by the principles of contractarian ethics, which allows deriving fiduciary duties owed to each company stakeholders. The social contract model hence helps curbing both managerial slacks and abuse of authority that those who are in position to run the firm may carry out against the legitimate interests of all the non-controlling corporate stakeholders who are also under many respects under-protected by normally incomplete contracts.

What is most important, however, is that the social contract model counteracts the presumption that CSR, while expressing laudable concerns for social issues, would be unable to account for the proper nature of the firm as an economic institution. As it will be shown, the CSR model answer questions concerning the very nature of the firm and provide for both explanation and justification of how the corporation might emerge as an economic institution run to the mutual advantage of its stakeholders.

Once the normative model will have been developed, however, any economic-minded reader will wonder whether it is not just wishful thinking, i.e. whether the model may accords with real world incentives and interests that drive economic interaction amongst real life economics agents. This amounts to require a basic change in the perspective of argumentation, from the normative - where impartial and universalizable reasons to act must take the precedence - to the implementation one, a step that I will undertake in the second part. [2]

Implementation is a domain wherein we have to look for mechanisms of endogenous motivation, based on a realistic account of preference and beliefs of economic agents as they are,

[2] I have presented the ideas of a normative model of CSR also in some previous works (see Sacconi 2004, 2006). Its implementation theory here exposed on the contrary is novel except for the treatment of unforeseen contingencies already given in (Sacconi 2000 and 2007).

that would make the model implementation self-sustainable. First of all it requires verifying whether the model could stand up by itself without a strong imposition of sanctions and inducement from the outside its own basic system of interaction (the corporate realm of the interaction amongst firms and their stakeholders). No doubts, this does not exclude supports that could be given by a proper regulation or - as long as some legal frameworks have been developed to support a different normative model of the corporation - by a change in regulation, or in general supports that may come from social and legal institutions based in different social subsystems (i.e. the legal system, or the civil society)[3]. But to start an implementation theory it is needed first of all to see whether the model can be sustained through the (equilibrium) rational choices carried out by the agents participating within the social interaction context basically relevant to the model (i.e. that involving interactions amongst companies and stakeholders - which is a domain wide enough to allow for a large array of situations) and the social norm and institutions they are able to develop within this context by themselves.

The idea of self-sustainability immediately leads to think that the implementation of the CSR normative model should be a matter of self-regulation and voluntariness. This is also a basic tenet of the second part of this essays. But it should be clear that there is a large gulf between two ways in which voluntariness and self-regulation may be understood, and they must not be confused. On the one hand a view again masterly represented by professor Jensen (Jensen 2001) thinks that it coincides with shareholder value maximization in the long run, a sort of re-elaboration of the standard selfish view of the proprietary's goal (even if sometime ownership is well diffused throughout the stock market) whereby the firm should be run. According to this view CSR is only a matter of wise strategic management that endeavors to long term shareholder value maximization with the appropriate means (including stakeholders claims satisfaction). On the other hand there is an idea, that I actively purport, that voluntariness must coincide with the development and voluntary adhesion to social norms and social standards explicitly formulated through a process of social dialog amongst companies and their stakeholders, which simulates the social contract model, such that these explicit norms are able to generate by them-selves the incentives and motivations that permits them to be largely endogenously self-enforced.

I therefore will give an articulate game-theoretic account of how an agreed CSR social norm or standard helps solving many problems in the realm of implementation understood as a non-cooperative game wherein the firm is a player endowed with his own preferences and interests strategically interacting with its stakeholders. The game of reference is the Trust Game, a game

[3] For the idea of institutional complementarities that may be used here see Aoki (2001)

where the possibility that a firm abuses its stakeholders' trust - i.e. it doesn't comply with a model of fair cooperation with them - is consubstantial. In this context an explicit CSR norm or social standard, incorporating the idea of multi-stakeholder governance, is shown to be helpful under many respects. It makes possible describing the game so that several types of reputations based on the respect of the CSR model itself may be developed even if unforeseen contingencies are involved (se also Sacconi 2000, 2007). It allows impartially selecting just *one* fair reputation equilibrium amongst the many possible. Elaborating on Binmore's *Natural justice* (2005) this task is accomplished again from the ex ante (under the 'veil of ignorance') point of view, but in a way that allows to find out a unique course of action that satisfies the requirement of incentive compatibility (i.e. a Nash equilibrium). Further, an agreed CSR social norm aids reducing to *just two* the candidate reputation equilibria that ex post, in the real world interaction taking place beyond the "veil of ignorance", may be played after an agreement (maybe seen as cheap-talk and not-binding) over a general principle of fairness has been reached by the firm and its stakeholders. These equilibria are defined not as traditional Nash equilibria, but as psychological equilibria according to the theory of conformist preferences (Grimalda and Sacconi 2005) developed along the lines of other behavioral game models (Genakoplos et al. 1986, Rabin 1993). Last, given the psychological equilibria that remain candidate as possible results of the game, it admits to identify and to make credible the initial players' beliefs over the possible game solutions wherefrom an equilibrium selection dynamic (representing the revision process of mutual expectation) singles out the game solution effectively carried out (my favorite equilibrium selection dynamics is the Harsanyi's *tracing procedure* – see Harsanyi and Selten 1988). For a large array of situations, that are cognitively the most reliable in case the players have ex ante agreed on a social norm or standard (even if the agreement is not binding), the process selects an equilibrium corresponding to the normative model of multi-stakeholder fiduciary duties.

PART I

2. A definition of CSR

According to many views, *Corporate Social Responsibility* (CSR thereafter) is a form of corporate strategic management that sets corporate standards of conduct at a level higher than mandatory legal constraints, and envisages itself as a system for the *governance* of transactions between a firm

and its stakeholders.[4] It is clear that here 'governance' is no longer the set of rules simply allocating property rights and defining the owners' control over the company's management. Instead it resembles the neo-institutional view whereby the firm, like the contract and other institutional forms, is a 'governance system' which establishes diverse rights and obligations in order to reduce 'transaction costs' and the negative externalities due to economic transactions.

I therefore propose the following definition of CSR: *a model of extended corporate governance whereby who runs a firm (entrepreneurs, directors, managers) have responsibilities that range from fulfilment of their fiduciary duties towards the owners to fulfilment of analogous fiduciary duties towards all the firm's stakeholders.*

Two terms must be defined for the foregoing proposition may be clearly understood:

a) Fiduciary duties. It is assumed that a subject has a legitimate interest but is unable to make the relevant decisions, in the sense that s/he does not know what goals to pursue, what alternative to choose, or how to deploy his/her resources in order to satisfy his/her interest. S/he, the *trustor*, therefore delegates decisions to a *trustee* empowered to choose actions and goals. The trustee may thus use the trustor's resources and select the appropriate course of action. For a fiduciary relationship – this being the basis of the trustee's authority *vis-à-vis* the trustor – to arise, the latter must possess a claim (right) towards the former. In other words, the trustee directs actions and uses the resources made over to him/her so that results are obtained which satisfy (to the best extent possible) the trustor's interests. These claims (i.e. the trustor's *rights*) impose fiduciary duties on the agent who is entitled with authority (the trustee), which s/he is obliged to fulfil. The fiduciary relation applies in a wide variety of instances: tutor/minor and teacher/pupil

[4] This view is consistent with the UE Commission initial definition of CSR "By stating their social responsibility and voluntarily taking on commitments which go beyond common regulatory and conventional requirements, which they would have to respect in any case, companies endeavor to raise the standards of social development, environmental protection and respect of fundamental rights and embrace an *open governance, reconciling interests of various stakeholders in an overall approach of quality and sustainability*" (*Promoting a European Framework for Corporate Social Responsibility,* Green Paper, p.4, Brussels, 18.7.2001, emphasis added). Moreover it is consistent with the way CSR is understood in many European multistakeholder project of CSR management systems like as in the UK the *Accountability 1000* standard for ethical and social audit and reporting has been developed, followed by the *Sigma Project* supported by the Blair government. A *Values Management System* has been developed in Germany at the University of Konstanz (Wieland 2003), while CSR standards modeled on the example of quality management systems have been proposed by independent bodies of standardization in Spain (*Aenor*) and France (*Afnor*). As far Italy is concerned, may be quoted the *GBS* standard for social reporting issued in the spring of 2001, and the *Q-RES Project* aiming to define a quality standard of management systems for ethical and social responsibility of firms. The latter initiative led in October 2001 to issuing the *Q-RES Management Guidelines*, see (Sacconi et al. 2003), and more recently to the setting of the *Q-RES Norm for the Improvement of Corporate Ethical-Social Performances of Organisations* (March 2003) see (Sacconi, deColle, Baldin, Oakley, Wieland and Zadek 2003)
More in general, even if some authors would not subscribe to CSR as the proper nickname, this view is consistent with how corporate responsibilities and accountability are seen in the stakeholder approach (Freeman 1984, Freeman and Evans 1989, Donaldson and Preston 1995, Clarkson 1999, Freeman, McVea 2004, Freeman, Velamury, 2006).

relationships, and (in the corporate domain) the relation between the board of a trust and its beneficiaries, or according to the predominant opinion, between the board of directors of a joint-stock company and its shareholders and then more generally between management and owners (if the latter do not run the enterprise themselves). By the term 'fiduciary duty', therefore, is meant the duty (or responsibility) to exercise authority for the good of those who have granted that authority and are therefore subject to it.[5]

b) Stakeholders. This term denotes individuals or groups with a major stake in the running of the firm and who are able to influence it significantly (Freeman and McVea 2002). However, a distinction should be drawn between the following two categories:

(i) *Stakeholders in the strict sense*: those who have an interest at stake because they have made specific investments in the firm (in the form of human capital, financial capital, social capital or trust, physical or environmental capital, or for the development of dedicated technologies, etc.) – that is, investments which may significantly increase the total value generated by the firm (net of the costs sustained for that purpose) and which are made specifically in relation to *that* firm (and not in any other) so that their value is idiosyncratically related to the completion of the transactions carried out by or in relation to that firm. These stakeholders are reciprocally dependent on the firm because they influence its value but at the same time – given the specificity of their investment – depend largely upon it for satisfaction of their well-being prospects (lock-in effect).

(ii) *Stakeholders in the broad sense*: those individuals or groups whose interest is involved because they *undergo* the 'external effects', positive or negative, of the transactions performed by the firm, even if they do not directly participate in the transaction, so that they do not contribute to, nor directly receive value from the firm.

It is evident that these two categories cannot be sharply separated. For example, a manufacturer in a developing country who supplies a component for an industrial good assembled in a Western European country is essentially dependent on his contract; and with his low labour costs (due to the customer's market power) he makes a crucial contribution to the European firm's profits. At the same time, however, if a mature technology is used, he is easily replaceable by the European firm, whose dependence on the supplier is therefore limited (in short, the reciprocal dependence relation is not symmetric). Likewise, a local community may not be party to the transactions performed by a company with a plant on its territory, but it is

[5] On fiduciary duties see Flannigan (1989).

nevertheless subject to that plant's environmental and social externalities. However, if the community has representative institutions with the power to grant or withhold a 'licence to operate', it is able to influence the company's creation of value and negotiate a reduction in the negative externalities. These decisions – connected as they are with the furnishing of infrastructures – may be viewed as investments intended to select and attract production activities whose positive externalities outweigh their negative ones.

We are now able to appreciate the scope of CSR defined as an extended form of governance: it extends the concept of fiduciary duty from a mono-stakeholder setting (where the sole stakeholder relevant to identification of fiduciary duties is the owner of the firm) to a multi-stakeholder one in which the firm owes fiduciary duties to *all* its stakeholders (the owners included). It is obvious that classification of stakeholders on the basis of the nature of their relationship with the firm must be regarded as important in gauging these further fiduciary duties.[6]

3. 'Abuse of authority': the economic basis of extended fiduciary duties owed to corporate stakeholders

Let us now inquire whether economic theory provides support for the thesis that the firm has 'further' responsibilities towards its stakeholders. According to neo-institutional theory (Williamson 1975, 1986; Grossman and Hart 1986; Hart and Moore 1990; Hart 1995; Hansmann 1996), the firm emerges as an institutional form of 'unified transactions governance' intended to remedy imperfections in the contracts that regulate exchange relations among subjects endowed with diverse assets (capital, labour, instrumental goods, consumption decisions, and so on). These assets, if used jointly, are able to generate a surplus over the cost of their use that is higher than in the case of their separate use by each asset-holder. However, contracts by which these asset-holders regulate their exchanges are incomplete: they do not include provisos covering unforeseen events, owing to the costs of drafting them, or because the cognitive limits of the human mind make it impossible to predict all possible states of the world. Yet for these assets to

[6] At first sight, it might be objected that many stakeholders, in both the 'strict' and 'broad' senses, do not have relations with a firm such that they formally delegate authority to those who run it (for example, they do not vote), with the consequence that the fiduciary duties as defined earlier do not apply to them. However, in the model of the social contract as a hypothetical explanation of the origin of the firm – see section 5.2 – all the stakeholders participate in the "firm's second social contract", with the consequence that their trust constitutes the authority of the firm's owner and manager. This also explains how the authority of the latter may be accepted by these subjects. Moreover, the hypothetical social contract is typically used to explain how authority – that is, legitimate power – may come about at both the political and organizational levels: cf. Green (1990), Raz (1985), Watt (1982). For a discussion of managerial authority see MacMahon (1989) and Sacconi (1991).

be used in the best manner possible, specific investments must be made: investments undertaken with a view to the value that they may produce within a idiosyncratic contractual relation. This entails that the surplus generated with respect to the costs sustained by each party to the exchange is determined by the undertaking of *specific* activities with *specific* counterparts (suppliers, customers, employees, financiers, etc.). Let us assume that parties behave opportunistically (that is, they are egoists who act with astuteness). Thus, once the investments have been made, contractual incompleteness means that the terms of the contract can be renegotiated, so that the party in a stronger *ex post* position is able to appropriate the entire surplus, thereby expropriating the other stakeholders. But if agents expect to be expropriated, they will have no incentive to undertake their investments at the optimal level. This expectation of unfair treatment gives rise to a loss of efficiency at the social level.

The firm responds to this problem by bringing the various transactions under control of a hierarchical authority – the authority, that is, of the party which owns the firm and through ownership is entitled to make decisions over the contingencies that were not ex ante contractible. Unified governance supplements incomplete contracts with authority relations through the vertical and horizontal integration of the units that previously made separate contributions. The firm is therefore a special contractual form: when contracts lack provisos contingent upon unforeseen events, they can be 'completed' with the 'residual right of control' which entitles its holder to decide what should be done about decisions not *ex ante* contractible– that is, decisions 'left over' from the original contract and that become available only when unforeseen situations occur.

The residual right of control underpins authority: those parties entitled with residual right of control may threaten the other parties to the contract with exclusion from the physical assets of the firm, thereby ensuring that *ex ante* non-contracted decisions are taken *ex post* to their own advantage. They are thus safeguarded against opportunism by the other stakeholders, and they are able to protect the expected value of their investments in situations where contract incompleteness provides margins of discretion when residual decisions have to be taken. There is therefore an efficiency rationale for the idea of the firm as 'unified governance' of transactions: if one party (a class of stakeholders) has made a specific investment of greater importance than those made by the others at risk, or if its exercise of 'unified governance' discourages opportunism by the others to appropriate the surplus, then that party should be granted the property right and with it the right to take 'residual' decisions. This is also the basis for regulation of authority delegation from the owners to directors or managers by corporate governance rules,

when the owners themselves are not able of directly exercising the entire residual right of control. Fiduciary duties owed to the owners must guarantee that delegated exercise of residual rights of control by the board of directors or managers will maintain or improve the efficiency of their original allocation to the selected class of stakeholders.

However, one should not underestimate the risks of the firm *qua* unified governance. There is not just one single stakeholder at risk because of contract incompleteness; it is usually the case that multiple stakeholders undertake specific investments (investments in human capital, investments of trust by consumers, investments of financial capital, investments by suppliers in raw materials, technologies and instrumental goods). Contracts with these stakeholders are also incomplete.

Yet if a firm brings its contracts with certain stakeholders (labour contracts, obligations towards and relations with minority shareholders) under the authority of a party to whom is allocated control over residual decisions (for example, the controlling shareholder group) – and more generally if a party is enabled by its *de facto* power to exercise discretion over *ex ante* non-contractible decisions concerning implicit or explicit contractual relations with the other stakeholders (consumers, customers, suppliers, creditors, etc.) – what, one may ask, is there to ensure protection of investments and interests other than those of the controlling stakeholder? It is evident that if fiduciary duties attach *only* to ownership, those stakeholders *without* residual right of control will *not* be protected by the fiduciary duties of those who run the firm.

The inherent risk, therefore, is an abuse of authority (Sacconi 2000). Those wielding authority may use it to expropriate the specific investments of others by exploiting 'gaps' in contracts – which persist even under unified governance (in fact it simply allocates to only one stakeholder the right to 'fill' those gaps with its discretionary decisions). Those in a position of authority, in fact, are able to threaten the other stakeholders with exclusion from access to physical assets of the firm, or from the benefits of the contract, to the point that those other stakeholders become indifferent between accepting the expropriation and forgoing the value of their investments by withdrawing from the relation. Thus the entire surplus, included that part of it imputable to efforts and investments made by the non controlling stakeholders, will be appropriated by the controlling party. Again forward-looking stakeholders will be deterred form entering the hierarchical transaction with the controlling party. In general, this will produce a internal *crisis of legitimacy* between firm and stakeholders (a crisis in the relationships between the organizational authorities and participants in the organization) and an external *crisis of trust* (in relationships with stakeholders that have entered into contractual or external relations with the organization).

Various stakeholders will *ex ante* have a reduced incentive to invest (if they foresee the risk of abuse), while *ex post* they will resort to conflicting or disloyal behavior (typically possible when asymmetry of information is inherent in the execution of some subordinate activity) in the belief that they are being subjected to abuse of authority. In the economist's jargon, this is a 'second best' state of affairs (less than optimum): all governance solutions based on the allocation of property rights to a single party may approximate social efficiency, but they can *never* fully achieve it. This much is acknowledged by the theoreticians of contractual incompleteness when they point out that the allocation of the residual right of control induces the party protected by that right to over-invest, while those not so protected are induced to under-invest, with a consequent shortfall with regard to the social optimum (Grossman and Hart 1986; Hart 1995).

On the other hand, if the stakeholder category entitled to exercise ownership (the double right of controlling residual decisions and claiming residual revenue; cf. Hansmann 1987, 1996) is selected on the basis of its ability to minimize total costs deriving form the summation of contractual costs borne by of various stakeholders and costs of exercising authority, it is by no means certain that a solution will be found which reduces *each* of those costs to the minimum (that is, reduces opportunism suffered by each stakeholder to the minimum). Sufficient for this solution to emerge is, for example, that the governance costs of one class (the capital-holders, for example) are low enough to counterbalance a relative increase in the contractual costs borne by another class (the workers, for example) compared to alternative cases (for instance the case in which there is no centralized governance, or the one in which it is a sub-set of workers that governs, or the consumers). In this case, too, some incentives are nullified, which distances the real-world solution from complete (Paretian) social efficiency. The fact is that the relative (in)efficiency depends on manifest or simply expected unfairness: separation between efficiency and fairness (a myth of neoclassical economics) is no longer feasible when we face the real-life problem of working out acceptable solution for the governance of transactions.

My suggestion is therefore that when CSR is viewed as 'extended governance', it completes the firm as an institution of transactions governance (cf. Sacconi 2000). The firm's legitimacy deficit (whatever category of stakeholders is placed in control of it) is remedied if the residual control right is accompanied by further fiduciary duties towards the subjects at risk of abuse of authority and deprived of the residual control right. At the same time, this is a move towards greater social efficiency because it reduces the disincentives and social costs generated by the abuse of authority. From this perspective, 'extended governance' should comprise:

- *the residual control right* (ownership) allocated to the stakeholder with the largest investments at risk and with relatively low governance costs, as well as the right to delegate authority to professional directors and management;

- *the fiduciary duties* of those who effectively run the firm (administrators and managers) towards the owners, given that these have delegated control to them;

- *the fiduciary duties of those in a position of authority in the firm (the owner or the managers) towards the non-controlling stakeholders*: the obligation, that is, to run the firm in a manner such that these stakeholders are not deprived of their fair shares of the surplus produced from their specific investments, and that they are not subject to negative externalities.[7]

A number of recent economic and legal models of governance support this view of CSR. For example, the firm can be seen as a 'nexus' of specific investments regulated by incomplete contracts, rather than as a nexus of simple contracts, and therefore as a team of actors cooperating to produce a surplus from those specific investments (Rajan and Zingales 2000). Based on a similar view which combines different theories of the firm – the theory of incomplete contracts with that of team production – is the model of multi-stakeholder governance developed by Margaret Blair and Lynn Stout, and which sees the purpose of corporate governance structures as being prevention of opportunistic behavior among the N members of the team that make specific investments. When applied to a public company, this model translates into a board of directors acting as a mediating hierarchy: an authority system charged with the task of finding the appropriate balance in the protection of diverse interests (cf. Blair, Stout 1999, 2006). The (controversial) legal basis for this form of "impartial governance" exercised by the board of directors and by management in the US joint-stock company is the 'business judgment doctrine': the manager's use of a standard of professional conduct which insulates his/her choices against claims by shareholders (cf. Blair, Stout 1999, but also see Meese 2002).

However, a number of unanswered questions remain which the proponent of CSR as 'extended governance' must necessarily address. Does there exist a criterion with which to give more precise specification to these extended duties, and from which it is possible to derive a strategic management standard of sufficient clarity such that the 'extended governance' model cannot be accused to entail higher governance costs than the traditional 'narrow' corporate governance view? What norms are effective for the implementation of CSR? What is the role of company

[7] I have proposed in a previous work (cf. Sacconi 1991) a view of managerial ethics based on a similar analysis of the theory of firm, as well as on the cooperative game theory of the firm put forward by Mashairo Aoki (Aoki 1984) and the notion of extended fiduciary duties (cf. Sacconi 2000).

law with respect to other parts of the law that impose constraints on corporate behaviour? And what role can be played by self-regulation?

4. The Social Contract as a governance criterion for the fair balance of stakeholders' interests

If a firm is a team of participants with specific investments, then the metaphor of a 'bargaining cooperative game' among multiple stakeholders can be applied. These stakeholders must agree on a shared action plan (a joint strategy) which allocates tasks among the members of the team so that the contribution of each of them is efficient (because it produces the maximum surplus net of each stakeholder's costs). The 'bargaining cooperative game' played by the stakeholders is typically one of mixed interests. Although it is in their common interest to cooperate, because this enables them to produce a surplus that would otherwise be impossible, conflict nevertheless persists among the stakeholders over the distribution of the value created. 'Governance' and strategic management consequently consist in the solution of two problems:

a. Identifying the joint strategy that the stakeholders (as the players in the cooperative game) may utilize to coordinate themselves, in that they accept it *ex ante* as a voluntary agreement to cooperate – so that strategic management can reduce bargaining costs (time, conflict, etc.) and the costs of gathering information on the alternatives available and on the intentions of each players about cooperation.

b. Ensuring *ex post* that each member of the team complies with the agreement on the joint strategy selected and does not act as a free rider with regard to the others.

Choosing the joint strategy (point a) is equivalent to select a bargaining equilibrium. It must therefore answer the question of what is due to each stakeholder and what each of them can expect from the firm in exchange for its contribution, so that each stakeholder may agree on that joint strategy. The question thus arises as to how the stakeholders' interests can be balanced against each other, and what claims on the firm should be considered the appropriate basis for the management's fiduciary duties. 'Stakeholder', in fact, is a descriptive term. It reminds us that a variety of classes of individuals have interests at stake in the running of the firm, and that they may sometimes advance conflicting claims. The use of the term 'stakeholder', however, does not provide a *criterion* with which to balance claims when they are mutually conflicting.

To answer the question we consequently need a *criterion* able to identify the balance that *any whatever stakeholder* would accept as the basis for its voluntary cooperation with the firm: that is, an *impartial* criterion. It is here that ethics – understood as a set of *impartial criteria* for collective choice-making – come into play as part of the firm's governance and strategic management.

As an ethical criterion, therefore, I suggest the 'social contract' among the stakeholders of the firm (Sacconi 1991, 2000). By 'social contract' I mean not any whatever real-life bargain but a 'touchstone' from which point of view to assess the diverse outcomes of day by day practical running of the firm. In other words, the social contract is the agreement that would be reached by the representatives of all the firm's stakeholders in a hypothetical situation of impartial choice.[8] Corresponding to the notion of 'social contract' is the following multi-stage deliberative procedure which generates impartially acceptable agreements.

(i) Force, fraud and manipulation must be set aside.

(ii) Each party comes to the bargaining table with only its capacity to contribute and its assessment of the utility of each agreement or non-agreement proposed (dispensing with any form of threat other than its possible refusal to agree).

(iii) The bargaining *status quo* must be set at a level such that each stakeholder results immune against the cost of its specific investments – that is, each stakeholder must obtain from the social contract at least reimbursement of the cost of the specific investment with which it has contributed to the surplus (otherwise the bargaining process would permit opportunistic exploitation of the counterparty's lock-in situation). The distribution of the surplus is regulated by the social contract – and by the corresponding deliberative procedure – on the basis of 'initial endowments' thus defined.

(iv) Each party in turn puts itself in the position of all the others, and in the position of each of them he can accept or reject the contractual alternatives proposed. [9]

[8] It is quite evident the debt of this contractarian view on the theory of firm to the works of both John Rawls (1971, 1993) and David Gauthier (1986). For the first formulation of the theory of the corporate social contract, based the revision of neo-institutionalist theory of firm and with reference to the problem of the abuse of authority *vis-à-vis* stakeholders, see however Sacconi (1991), and latterly Sacconi (2000). For a formulation external to economic theory see Dunfee and Donaldson (1995).

[9] This step amounts to requiring that the decision maker puts himself under a 'veil of ignorance' concerning his identity and hence he takes each individual position in turn in order to discover what he would accept in case he were each particular stakeholder who seats at the bargaining table. This veil of ignorance is thinner than Rawls' one, for by this permutation of the individuals' point of views the decision maker may appreciate each player's preference, whereas the Rawlsian veil of ignorance makes impossible to learn any individual characteristics that would permit reconstructing the individuals' life plans, so that all the different players are all perfectly indistinguishable from the point of view of the decision maker under the veil of ignorance (see Rawls 1971).

(v) If solutions are found which are acceptable to some stakeholders but not to others, these solutions must be discarded and the procedure repeated (which reflects the assumption that cooperation by all stakeholders is recognized as necessary).

(vi) The terms of the agreement reached are therefore those that each stakeholder is willing to accept from its particular point of view: that is, the non-empty intersection of the joint strategies and relative distributions acceptable to each of them.

Note that this intersection is necessarily *non-empty*, for otherwise the game would not allow a cooperative surplus. That is to say, it would not be the case that joint action by the parties may produce something more than their separate action and that at least one surplus distribution proves to be reciprocally advantageous (if it *must* be so, then there exists at least one agreement acceptable to all).

5. Social contract and uniquely defined objective-function of the firm

The main objection brought against CSR is that the multi-stakeholder approach to firm's governance leaves the management without a clearly-stated and uniquely defined 'bottom line', to be used as a benchmark against which to evaluate its success or failure (Jensen 2001). The consequence, the argument runs, is that the management exploits this situation to pursue its personal interests. It comes up with every possible device to conceal its essentially self-dealing behaviour behind the interests of some or other stakeholder. Whereas, the critics of CSR maintain, it is easy to check the managerial strategy (among the alternatives available at any particular time) against the criterion of increasing as much as possible the firm's profits, and with which the management can be straightforwardly made to comply, this is not the case of 'stakeholder value', since this consists of numerous dimensions to maximize simultaneously (the interests of the various stakeholders). Consequently, stakeholder value contains an intrinsic contradiction – the pursuit of conflicting, or at any rate divergent, goals at the same time – so that the choice of which strategy to adopt is ultimately left to the mere managerial discretion.[10]

It is evident, however, that this objection does not apply to the model of the social contract of the firm proposed here – which by no means ignores the existence of a distributive conflict, but

[10] This danger is also stressed by Tirole (2001), who however recognizes the relevance of the stakeholder approach to corporate governance.

instead resolves it by identifying a bargaining equilibrium that permits mutual cooperation among the members of the team.

To be precise, I would add that counterpart to the philosophical contractarian model is a mathematical model of bargaining whose solution is as exactly computable as the profit function in microeconomic theory. Hence we can simply substitute profit maximization with maximization of the function which assigns the solution to the bargaining game and assume this as the firm's (perfectly computable) objective-function.[11] This solution is simultaneously an answer to both the problem of cooperation and that of distributive conflict among the stakeholders. Moreover (for those able to appreciate the marvels of mathematics), if the bargaining set is well defined (and if one accepts the Nash, Harsanyi and Zeuthen postulates of bargaining theory), the solution is defined uniquely, so that the set of admissible solutions is reduced to one single alternative. Hence the best pursuit of the interest of the stakeholder controlling the firm is equivalent to the solution of the bargaining problem among all the stakeholders. (In any event, various theories of bargaining yield solutions which quite closely resemble each other – cf. Gauthier (1986) and Kalai and Smordinski (1975) as they are slight changes of the basic Nash's solution; and for the purposes of this study, identifying a set of 'close' solutions compatible with the idea of rational bargaining seems good enough).

What does this solution say? On the assumption that the external effects on third parties with no influence on the transactions are minimized by selection of the forms of cooperation which damage them least, one may proceed as follows. Net of the initial (pre-bargaining) position or *status quo*, in which coverage of the costs of specific investments by each stakeholder must be included, calculation is made of the value for each stakeholder of each cooperative outcome from joint action plans. The set of these values (or better, the set of these vectors of values) is the outcomes space associating to each joint action plan (joint strategy) an allocation of the cooperative surplus (positive or nil) to the players. Rational bargaining takes place within this space among players endeavoring to obtain a share as high as possible of the cooperative surplus for themselves, to the detriment of the others' shares - once it is taken for granted the need for reciprocal agreement, for in its absence they would be unable to obtain anything more than the initial position (*status quo*). Calculable within this space is the Nash bargaining function, the product of the utilities of the various stakeholders with specific investments, that is, an aggregative function of their utilities. Where the product (the aggregation) is maximum, there is a

[11] For the theory of bargaining games see Harsanyi (1977). Mashaiko Aoki uses the Nash bargaining solution in his theory of the firm, which envisages impartial governance by managers (Aoki 1984).

bargaining equilibrium (i.e. a rational agreement among the participants in the social contract) corresponding to the Nash bargaining solution (Nash 1950; Harsanyi 1977). (see Fig. 1 for an example with two players).[12]

This equilibrium does not need operational interpersonal comparisons of utility (which operationally are very problematic) in order to be calculated (interpersonal comparisons can be confined to the interpretive level[13]). It obeys, in fact, simple axioms of individual rationality in bargaining – like the decision to grant a concession according to the expected personal utility given the probability that the counterparty will accept or refuse it, or that a player will not make a concession that he or she would not expect the counterparty will make in a similar situation – and conditions of mutually expected rationality, like expecting that the willingness of acceptance by the counterparty depends on a symmetric probabilistic assessment of the first party behaviour, and not to expect the counterparty to accept something that oneself would not accept, and so on (cf. Harsanyi 1977)**]**.

Of course, if these postulates are taken literally, they can be criticised as unrealistic; and it is likely that in the real world agents are unable to maximize or to estimate probabilities coherently, or to

---

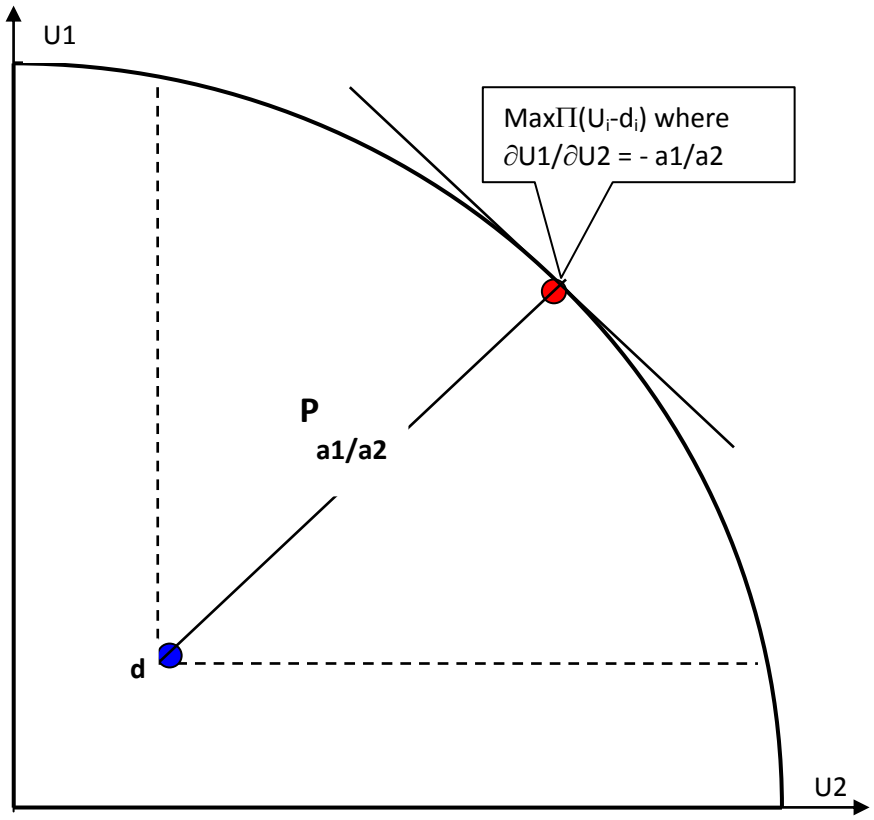
[12] Specifically, let us consider a case with two players, 1 and 2, and let us assume that the solution is a point in space R^2 enclosed between the positive Cartesian axes U1 and U2, each of which measures the utility for a player of the outcomes of the cooperative game (see Figure 1). The space therefore represents the outcomes subject to bargaining in terms of their value in utility for the players (i.e. their payoffs). The standard analytical assumption is that the payoff space is convex and compact. The payoff space **P** therefore has an efficient frontier (in the upper-right positive ortant) which represents the set of outcomes for which the players' utilities cannot be increased by an alternative agreement without reducing the utility of at least one other player. Below this frontier are agreements with respect to which gains are still possible for all; above it are outcomes unfeasible by any agreement or joint plan of action. All points in the space represents different possible values of the coalition among the two players. In fact, only when all of them agree on the solution of the game can they leave the *status quo* **d**, which is represented by a point interior to the space, so that they may benefit form cooperation. The characteristic function of the coalition among all the players is therefore super-additive (it is better to agree than not to agree). Obviously, of interest are only those agreements for which there is an efficient allocation. But in what point among those on the frontier should the agreement fall? The Nash solution states that the players will agree on the joint strategy corresponding to the point of the frontier where the following holds:

$$Max\ \Pi_i\ (U_i - d_i) \qquad \text{(i=1,2 denotes the various participants in the bargaining)}$$

where U_i is the utility of the generic stakeholder *i* for the cooperative transaction that it undertakes with the firm, and d_i is the cost of the specific investments made by *i* in order to participate in the joint action plan (that is, *i* always at least recoups the cost of its specific investment). The solution assumes that bargaining should provide each player with at least a small net advantage, which is the difference between the share of the surplus received and the *status quo* value. As a consequence of additional rationality postulates, these net individual advantages can be identified as such that the *product* of all of them is the maximum among those in the set of the possible outcomes of the cooperation. We may say that this is the collective choice function adopted by the members of the coalition, in light of their bargaining, to resolve the problem of their joint action. It is coherent with the proportionality of the remunerations with the relative utilities, because the ratio in which the shares of the surplus a1/a2 are distributed is proportional to the ratio between the marginal variations in the players' utilities $\partial U1/\partial U2 = - a1/a2$ (Brock 1979, Sacconi 1991, 1997). On the basis of Nash's postulates (1950) and those of the Zeuthen-Harsanyi (cf. Harsanyi 1977), this solution expresses a bargaining equilibrium based on individual rationality of the players.

[13] See Brock (1979), and Sacconi (1991).

make accurate forecasts about the rational behaviour of others. But what matters for my purposes here is that these postulates are a good approximation of rational behaviour in a hypothetical (ideal) bargaining situation among stakeholders (for it is a *normative* model being developed here – one no less normative than that of profit maximization).



(Fig.1. A cooperative bargaining game with two players and the Nash solution in a symmetrical case)

Moreover the outcome of the bargaining game can be interpreted as the solution which is coherent with a notion of distributive justice. On the assumption that it is possible to unify the units measuring utility, we find that the bargaining solution computed always distributes the advantages proportionally to 'relative needs' or to the relative marginal variations in the intensity of personal utilities (Brock 1979; Sacconi 1991, 1997). Because it is located on the upper-right frontier of the space of the bargaining outcomes, it fulfils the requirement of social efficiency (no advantage from cooperation is lost) and at the same time corresponds to an intuitive notion of fairness.

6. The emergence of extended fiduciary duties from the social contract of the firm

Thus far, the social contract has been presented as a normative deliberative procedure by which to identify the terms of an agreement that would be acceptable from an impartial standpoint – that is, from the point of view of any whatever stakeholder – so that it can be adopted as a standard of behavior by, for example, the mediating hierarchy proposed by Blair and Stout. However, the social contract can also furnish a reconstruction – understood as a 'potential explanation' – of how bargaining has given rise to a firm with *both* fiduciary duties towards the owners *and* social responsibility (i.e. further fiduciary duties) towards all the stakeholders.

Consider a 'state of nature' prior to the creation of the firm. Bilateral transactions among stakeholders regulated by incomplete contracts are subject to reciprocal opportunistic behaviour, with the consequence that prohibitive bargaining costs render them inefficient. At the same time, the parties to those transactions are entirely unconcerned about the negative external effects of their transactions on other agents, who although they do not participate, are nevertheless affected. This is a Hobbesian scenario in which the life of economic transactions among agents is "solitary, poor, nasty, brutish, and short".[14] The stakeholders thus address the problem of creating an association whereby all their transactions can be undertaken in accordance with agreed-to rules and are therefore not subject to contract-costs, while at the same time the negative effects on those who do not participate in the benefits from the transactions are reduced to the minimum. The 'First Social Contract' of the firm (*pactum unionis*) is nothing other than the agreement which the stakeholders reach *among themselves* to set up this association (the '*just* firm'). They negotiate on the association's constitution, which consists in a common plan of action (joint strategy) to which each of them contributes either by carrying out a positive effort or by simply refraining from applying his/her veto. This *first social contract of the firm* stipulates as follows:

a. rejection of shared plans of action which generate negative externalities for those not participating in the cooperative venture or, if these negative externalities are essential for the production of the cooperative surplus, a compensation of third parties so that they are rendered neutral;

b. production of the maximum surplus possible (difference between the value of the product for its consumers, who belong to the association, and the costs sustained by each stakeholder to produce it);

[14] Cf. Hobbes, *Leviathan*, (1651), part 1, chapter 13.

c. a distribution of the surplus which is 'fair', or rationally acceptable to each stakeholder in a bargaining process free from force or fraud and based on an equitable *status quo*, that is, considering the surplus net of the specific investments.

However, if an attempt is made to reach this form of an ideal association (the 'just firm') which eliminates all the participants' contract-costs, they arrive in practice to an organisational form which is found to be inefficient from the point of view of its governance costs. The stakeholders discover, for example, that the general assembly of all members is unable to take coherent decisions in a reasonable amount of time. In the absence of a monitoring system, once the members of the association have established fair shares of the surplus to be distributed among them, they have an incentive to act opportunistically and not to play their part. Coordination problems arise on how the joint strategy can be implemented under changing circumstances, which may alter beliefs and reciprocal expectations asymmetrically. The stakeholders consequently draw up a *second social contract* of the firm (*pactum subjections*)[15] by which they constitute, in the proper sense of the term, a governance structure for the association. It is only now that the association becomes a hierarchical structure.

The second social contract provides that authority should be delegated to the stakeholder most efficient in performing governance functions (the taking of residual decisions, devising coordination solutions as circumstances change, monitoring, the enactment of sanctions, excluding potential free riders, etc.). For this reason, it can also be seen as a contract *between the stakeholders and* those who is given control over the firm (social contract *with the firm*). After comparative examination of the governance costs of each stakeholder, the one with the lowest costs is selected and assigned ownership, and is therefore the one to which the right of governing the association is delegated (Hansmann 1996). This class, which is remunerated with the *residual* is authorised to delegate some discretionary decisions in regard to running the firm to professional director and managers, and to appoint those who are in the authority position of running the firm. *Prima facie*, their authority will be effectively constituted – that is, the delegation will remain valid – as long as they comply with what I call

Narrow fiduciary proviso: the owners are remunerated with the maximum residual revenue possible (in forms compatible with the diverse nature of the controlling stakeholder: profits, returns, discounts, improved conditions of service, improved conditions of employment, and so on) in the light of conditions obtaining in the firm's specific market.

[15] Interestingly, also Blair and Stout (1999) adopt the analogy between the firm and the two social contracts typical of the social contract tradition.

However, it is evident that this proviso entails that the positions of the other stakeholders change (from the "just firm" to *just a firm*). Formerly co-equal members of the association, they are now subject in various ways to the discretionary decisions taken by the stakeholder entitled with authority, and by the administrators that it has appointed. Unlike in the standard economic theory of the firm, in the social contract theory the risk of the abuse of authority can squarely be faced. The *second social contract* is therefore conceived in a manner such that this cost of hierarchy is forestalled as well. Hence, under the second social contract, the stakeholders agree to submit to authority, thereby rendering it effective, if the contract contains the which stipulates that the firm's new governance structure must comply with *fiduciary duties* towards all the stakeholders (owners and non-owners).

Extended fiduciary proviso:

- *Towards the non-owners*

– The firm must abstain from activities which impose negative external effects on stakeholders not party to transactions, or compensate them so that they remain neutral;

– The firm must remunerate the stakeholders participating in the firm's transactions with pay-offs (monetary or of other kinds, for example in terms of the quantity, quality and prices of goods, services, working conditions, etc.) which, taken for granted a fair status quo, must contain a part tied to the firm's economic performance such to approximate fair/efficient shares of the surplus (assuming that this is positive) as envisaged by the first social contract.[16]

- *Towards the owners*: The firm must remunerate the owners with the maximum residual compatible with fair remuneration – as defined by the first social contract – of the efficient contributions made by all the other stakeholders.

What does this hypothetical explanation yield? It yields a definition of the 'corporate interest' of the company – that is, the interest that the manager acting in the name of the company must serve – which is consistent with the contractarian model. According to this reconstruction, in fact, the manager (appointed through the second social contract) has a special fiduciary duty towards the owners (or the 'residual claimant') that has delegated authority to him/her (*via* narrow fiduciary proviso). This duty applies, however, only under the constraint that the *general*

[16] Note that meant here is remuneration in utility and not necessarily in money. Put in economic parlance, this remuneration consists of the consumer rent, the producer rent, the worker rent and so on, accruing to each of them from the firm's transactions. This means that some stakeholders may not want to receive monetary benefits from the firm, but rather improvements in working conditions or in purchasing power, in the quality of goods and services, of contractual conditions, etc., to which the shares of the surplus are in any case devoted.

fiduciary duties are fulfilled towards *all* the stakeholders – which is defined *via* the extended fiduciary *proviso*. We may thus construct the corporate interest by means of a hierarchical decision-making procedure which moves from the most general conditions to the most specific ones:

- *First step*: minimize the negative externalities affecting stakeholders in the broad sense (perhaps by paying suitable compensation);

- *Second step*: identify the agreements compatible with the maximization of the joint surplus and its simultaneous fair distribution, as established by the impartial cooperative agreement among the stakeholders in the strict sense;

- *Third step*: if more than one option is available in the above defined feasible se, choose the one that maximizes the *residual* allocated to the owner (for example, the shareholder).

Hence, the narrow corporate interest (the one usually advocated by supporters of the "shareholder value" view) results from a series of steps which select the admissible ways in which this interest can be satisfied – that is, those that are consistent with the various constraints imposed by the first social contract on the owner's behaviour. It should be emphasized that this concept cannot be reduced to that of value maximization for the 'residual claimant' (the owners) once constraints imposed by positive contractual obligations have been fulfilled. This is because we recognize all contracts are incomplete, and they are always susceptible to opportunism (even by those who run the firm), so that it is the entire hierarchical decision procedure which provides the basis for satisfying the corporate interest – i.e. the social contract identifies the goals or the internal (not merely external) moral constraints that channels managerial discretion. It results from satisfaction in sequence of the three requirements set out above and which can be summarized as follows: maximize the value for the residual claimant under the constraint of complying with the social contract between firm and stakeholders which defines the 'stakeholder value'.

PART II

7. Implementing the model: regulation and self-regulation

This and the following sections address the different problem of how the normative model of CSR based on the stakeholders' social contract of the firm may be implemented through norms

supported by endogenous incentives and motivations. The idea is that the normative model of multistakeholder corporate governance and objective function is supported by strong endogenous incentives and motivations, so that its implementation may rest largely on voluntary self-regulatory norms deliberated by companies, and on their decisions to comply effectively with extended fiduciary duties owed to their stakeholders.

Of course, effective CSR self regulation is a viable option only within an institutional and legal environment that does not obstruct it. Such obstruction would occur in the case of too narrow definitions of the firm's objective-function such as that prescribing shareholder value maximization as the company's only goal – as nowadays to be found in many company laws at European level. Whenever maximizing the joint stakeholder value conflicted even in the very short run with immediate shareholder value maximization, these laws would prevent the board from deciding to balance stakeholders' interests according to the social contract view which implies a constrained maximization view (that is, constraining shareholder value maximization with the condition of the simultaneous maximization of other stakeholders' utility according to a bargaining solution). The recent 2006 company law reform in the UK is an example of how the corporation's goals may be enlarged by means of a general and abstract principle in order to legitimate the exercise of some balancing decision amongst different stakeholders as well giving relevance to the notion of reputation in the long run.[17] This is not at all a concrete norm prescribing a precise balancing criterion. On the contrary, it is a very general principle that enables the company board and governance structure to trade some interests off against others at least in the short run. If complemented with an accountability requirement concerning how the board of directors will account for its carrying out implementation of the balancing decisions, a regulation like this effectively opens the door to a self-regulatory CSR standard that more precisely specifies principles and guidelines whereby the CSR model of governance must be implemented. On being asked to account for their balancing decisions, boards would appeal to

[17] The 2006 UK company law reform, Art. 172, states: "Duty to promote the success of the company:
(1) A director of a company must act in the way he considers, in good faith, would be most likely to promote the success of the company for the benefit of its members as a whole, and in doing so have regard (amongst other matters) to—
(a) the likely consequences of any decision in the long term,
(b) the interests of the company's employees,
(c) the need to foster the company's business relationships with suppliers, customers and others,
(d) the impact of the company's operations on the community and the environment,
(e) the desirability of the company maintaining a reputation for high standards of business conduct, and
(f) the need to act fairly as between members of the company.
(2) Where or to the extent that the purposes of the company consist of or include purposes other than the benefit of its members, subsection (1) has effect as if the reference to promoting the success of the company for the benefit of its members were to achieving those purposes"

such principles or criteria in order to justify their behavior to those stakeholders that may be disadvantaged by any particular balancing decision. This desirable complementariness between legal regulation and soft law or social standard-based self-regulation can be seen as very useful for the purpose of implementing the CSR social contract model and multistakeholder corporate governance.

Nevertheless the thrust of my argument is that, once company law does not obstruct proper self-regulation, the endogenous beliefs, motivations and preferences of economic agents (companies and stakeholders) are the essential forces driving the implementation of the CSR model of multistakeholder governance. To put it in game theoretical terms, the normative model is implementable *in equilibrium*. The rest of this paper aims to give substance to this statement.

First, however, it must be said that this position should not be confused with the standard economist's view that if CSR is to emerge as equilibrium behavior from endogenous incentives, its driving force must be simply enlightened self-interest in the long run. According to this view, a self-interested entrepreneur who cares only for his self-interest in the long run (or cares for the self interest of the firm's owners in the long run – i.e. the shareholders) would adopt behavior that spontaneously satisfies the company stakeholders' interests with no need to single out a principle of fairness nor to agree on any social contract principles with stakeholders in order to state explicitly that the firm owes a fiduciary duty to them. Self interest in the long run would simply guarantee that the treatment of corporate stakeholders spontaneously simulates a behavior fulfilling extended fiduciary duties, thus making any explicit statement of these duties superfluous. The social contract would be useless as well, since in the long run there is nothing like a conflictual "state of nature" amongst the company's stakeholder. On the contrary, we would observe harmony of interest amongst them, so that seeking the shareholder interest in the long run would also coincide with fulfillment at any time of the stakeholders' interest. As a consequence, the only goal that must be specified as the proper constraint on managerial and entrepreneurial discretion in the management of the firm is the coherent pursuit of shareholder value in the long run. Satisfaction of the stakeholders' legitimate interests is seen as simply a side-effect of this main goal, because they are related to it through a mean-end relation. If one strives to achieve the end of maximizing shareholder value, in the long run one will necessarily choose *as the means* those strategies that will also satisfy the stakeholders' interests. Hence whilst

stakeholders are to be accounted for within the dimension of corporate means, only shareholders are recognized as sources for corporate ends. [18]

This view does not recognize any need for a norm that explicitly states a principle of fair balancing amongst stakeholders. Note that this norm is excluded not only from mandatory law but also from principles of business ethics or self-regulatory standards or codes of ethics, in so far as these social norms may be seen as underhand attempts to modify or integrate the ends or the bottom line whereby the manager must account for his conduct.

This self-interest-in-the-long-run view is untenable. First of all, without the explicit statement of a CSR norm - be it worked out autonomously by the board of directors or through a social multistakeholder dialogue - based on the hypothetical agreement of the company stakeholders, a long run self-interested corporate strategy unintentionally simulating a behavior pursuing stakeholder value may simply not exist (or become something that the firm cannot be aware of at all). Moreover, even if such a behavior in the long run could be worked out as something of which the firm might be aware, nevertheless other behaviors in the long run could hence be worked out by the company, so that they provided very limited and minimal satisfaction of the stakeholders' claim of fair treatment. These further behaviors would not only be preferable to the firm's owners, they would also command a certain acquiescence on the part of the stakeholders – which could be made indifferent between the prospects of giving in to these firm's opportunistic strategies or staying out of any relationship with it. We must conclude that the simple self-interest in the long run view, translated into the shareholder value in the long run doctrine, would imply a large amount of violation of stakeholders' legitimate claims and abuse of ownership-based authority.

By contrast, the self regulatory view defended here requires the putting in place of explicit norms arrived at by social dialogue and multistakeholder agreements, and taking the form of CSR governance codes or management standards, voluntarily accepted by firms because they contain and specify the terms of the ideal and fair social contract between the firm and its stakeholders. They are explicitly formulated in language (written or oral) and their utterances contain the statements of extended fiduciary duties and obligations that the firm owes to its stakeholders. At the same time they are voluntarily adhered to, and as far as enforcement is concerned, they are

[18] This is probably the opinion of Jensen when he says "Indeed, it is a basic principle of enlightened value maximization that *we cannot maximize the long-term market value of an organization if we ignore or mistreat any important constituency*. We cannot create value without good relations with customers, employees, financial backers, suppliers, regulators, and communities. But having said that, we can now use the value criterion for choosing among those competing interests. I say "competing" interests because no constituency can be given full satisfaction if the firm is to flourish and survive." (Jensen 2001). See also Sternberg (1999).

not imposed by external legal sanction but instead through endogenous social and economic sanctions and incentives. In this sense they are self-enforceable explicit norms, put into practice essentially by means of endogenous economic and social forces such as reputation effects and conformity.

To understand why an explicit utterance of the stakeholders' social contract of the firm by means of an explicit voluntary CSR norm is so essential for the endogeneity and self-sustainability of the model's implementation, we must consider the *four roles* performed by explicit and voluntarily agreed norms:

a) the *cognitive-constructive role*, which answers the question about *how* the firm *works out* the *set* of commitments that it *can* undertake with respect to future events it is aware of not being able to predict in any detail, and therefore *what* types of *possible* equilibrium behavior the firm can work out so that stakeholders may expect them from the firm;

b) The *normative role*, which answers the question about what (if any) pattern of behavior and interaction the firm and its stakeholders must *select* from the set of possible equilibrium patterns to carry out *ex post* (according to the answer given to question a), if they put themselves in the *ex ante* position enabling an agreement to be made from an impartial point of view;

c) The *motivational role*, which answers the question about *what* and *how many* equilibrium patterns of behaviors, amongst those that could emerge ex post from the interaction between firm and stakeholder, would retain *their motivational force* if firm and stakeholder were able to agree in an ex ante perspective on a CSR norm along the lines of question (b);

d) The *cognitive-predictive role* concerning how a CSR norm *affects* the beliefs formation process whereby a firm and its stakeholders cognitively converge on a system of mutually consistent expectations such that they reciprocally predict from one another the execution of a given equilibrium in their *ex post* interaction (given that more than one equilibrium point still retains motivational force according to the answer to question (c)). Does the norm shape the expectation formation process so that in the end it will coincide with what the *ex ante* agreed principle would require of firm and stakeholders?

The reference that one must keep in mind concerning these four roles of explicit voluntary CSR norms and standards are repeated reputation games, and in particular the repeated Trust Game that I will introduce in the next section.

8. Cognitive /constructive role of a CSR norm: filling the gaps in the game form.

One basic idea in the domain of CSR implementation is that the incentives related to the formation of a reputation can play a key role in a firm's decision to endorse and respect the model of extended fiduciary duties owed to its stakeholders. If the firm wants to induce its stakeholders to enter into cooperative relations, it develops a reputation. Stakeholders will decide to cooperate with the firm if they trust that it will not abuse them. For a reputation to be developed, stakeholders must verify that the firm behaves according to a "cooperative" type of behaviour and that it does not abuse them when they decide to trust it (i.e. that the firm respects what we understand as the fair term of their social contract). In essence, stakeholders must believe that the firm is an "honest type" (i.e. a type which does not abuse its authority). A firm's reputation for being a certain type increases if evidence is gathered that confirms it is that type. A firm that wants to induce the stakeholders' cooperation must act so that its behaviour becomes indistinguishable from the voluntary discharge of its fiduciary duties towards its stakeholders. Otherwise, if the stakeholders observe an opportunistic behaviour, the firm's reputation will suffer a dramatic setback.

One way to illustrate the functioning of the reputation mechanism is a simple interactive situation representing a transaction based on the fiduciary relation (the Trust Game) between a stakeholder A and a firm B (see fig.2). A stakeholder must decide whether or not to trust the firm and enter or otherwise into an exchange relation with it. The firm decides whether or not to abuse the stakeholder's trust. As is well known, this game played one shot has just one Nash equilibrium, the strategy pair (no-entry, abuse) = ($\neg$***e, a)***.

(insert fig. 2 about here)

Things change however when the Trust Game is repeatedly played for an indefinite number of times between an infinite series of short-run players $A_1,\ldots,A_n$ (where n goes to infinitum), each one taking part in a single stage-game, and a long run player B taking part in every repetition of the basic stage-game. Player B's strategies in the repeated game are hence behavioural rules for choosing actions at each stage-game as a function of each history of the game until any stage at which player B must choose. Payoff functions must rearrange accordingly. Whereas any short-run player (A_i) is only interested in the payoff of the stage-game in which he takes part, the long-run player's (B) payoff is the infinite sum of the payoffs he gets at each stage. A crucial assumption is

that player B is more or less far-sighted (i.e. player B's discount rate δ for future utilities is more or less close to 1).

As already mentioned, the idea that players entertain beliefs about the possible types of the counterparty characterises reputation games. The long-run player is perfectly rational (from a strategic point of view) and perfectly informed about the game, but short-run players are not perfectly informed and hence are uncertain about the "type" of player B. By a "type" is meant a commitment to choose a given action under the different contingencies of the game - i.e. the action that player B chooses in each stage-game. Hence "types" reflect the idea that player B may be idiosyncratically committed to some rule of behaviour, even if he is uncertain which it is.

The different types of player B taken as possible by Ai are the following: i) the stage-game "rational" type, who always chooses the dominant strategy of the stage-game, ii) the type who never chooses to abuse (the honest type), iii) types who variously combine abuse and non-abuse randomly by means of different mixed strategies. Player B's reputation is the probability assigned by each player A_i at every stage to different types of player B. Probabilities are updated according to the Bayes rule: at each stage, the 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.

Each Player A_i chooses, according to the expected utility reasoning, between ***e*** and ¬***e***. During the first stages of the repeated game players A_i necessarily do not trust . Eventually (say after N periods), however, a short run player (say A_{N+1}) may begin to trust player B if a series of A_i before him have observed ¬***a*** so many times that the conditioned probability of the "absolutely honest" type is updated to the level p* where the expected utility of ***e*** becomes higher than ¬***e***. Player B's optimal choices follow from this feature of the model, and they define the equilibrium set of the repeated game. First of all, player B can decide always to choose the equilibrium strategy of the stage game, that is, always ***a,*** which induces A_i not to enter for ever. This leads to a repeated game equilibrium, because nobody has the incentive to deviate from such choices for the entire duration of the game. But Player B has a different strategy at his disposal. This consists 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 the conditioned probability reaches the critical level p*. At this point he calculates whether it is better for him to continue playing ¬***a*** - consequently over and over again inducing choices ***e*** from players A_i after A_{N+1} - or to defect by choosing ***a***. If δ is close to 1 (that is, player B is not short-sighted), then the infinite sum of payoffs 2, even if discounted, will more than counterbalance a single chance of winning payoff 3 (cf. Fudenberg e Levine 1989, 1991).

However, stakeholders can form their beliefs on the firm's behaviour (and, consequently, the firm can accumulate its reputation of being an "honest" type), only if they can observe without ambiguity whether the firm has behaved according to a type. A problem with regard to this condition arises when the relations between the firm and its stakeholders take place in a setting where information or knowledge about the firm's action is incomplete. Typically, a firm and its stakeholders are involved in incomplete contracts situations where a contract does not contain provisos covering unforeseen contingencies, so that there is no concrete benchmark against which claims of renegotiation can be assessed when unforeseen events occur. Because of incomplete knowledge, the stakeholders cannot verify whether the firm has actually behaved honestly according to the terms of the contract – in fact, the contract is mute when unforeseen events occur.

Contract incompleteness is not just a curiosity. Repeated reputation games are games that firms can play with their stakeholders only in the knowledge context appropriate for the existence of the firm itself. According to the neo-institutional thesis adopted throughout this paper, the context suited to the firm is one of incompleteness of contracts wherein contractual commitments are badly specified or not existent at all with reference to unforeseen states of the world. If one takes strategies that the firm will pursue and behavioral types that it will display throughout the repetition of a reputation game as coincident with contractual commitments (assuming that they are endorsed only if fair), incompleteness simply implies that *commitments* are unspecified in regard to unforeseen states. But a reputation is the probability that player B will comply with given commitments. Hence neither inductive learning about compliance can work nor reputation can be accumulated in such states.

Voluntary CSR norms, admitted that they are explicitly formulated through general and abstract ethics principles of fair treatment and precautionary rules of behavior, state the commitments that a reputation mechanism takes as reference point in its working but that a conditional contract cannot anticipate in regard to unforeseen events.[19] In other words, norms, principles and precautionary standards of behavior state the strategies that the firm may pursue in whatever state of the world, and they can be taken as the reference point for the formation, confirmation or refutation of the stakeholders' expectations concerning the company types. A key point concerning how abstract principles and precautionary rules of behavior may state commitments referred to unforeseen state of the world is the way in which they help manage *vagueness* and

[19] This is an idea basically very akin to that of "principles of a corporate culture" suggested by Kreps (1990), with the important difference, however, that I defend universalistic corporate ethics, not the relative-to-context notion of "culture" (see Sacconi 2000)

ambiguity. In fact their statement does not require a complete forecast of a state of the world in all its concrete details. They simply require *fuzzy pattern recognition* of a state in terms of some abstract characteristic related to a general principle, and the following activation of precautionary rules of behavior – which are not conditional on ex ante complete descriptions of (unforeseen) states of the world and can be activated just by default (Sacconi 2000, 2007) .

To see how, assume that a general principle of ethics defines as its domain of application the set of those situations that have a certain abstract and universalizable characteristic. In order to belong to the domain of a general principle, any given state of the world does not need to be completely and clearly described ex ante – which would be implausible for unforeseen states. In fact the description of unforeseen contingencies is necessarily mute about many concrete proprieties that characterize the ex ante representation of foreseen state of the world as these properties simply do not occur in an unforeseen state (this is what makes it 'unforeseeable' in the strict sense). But as far as abstract, general e universalizable characteristics identified as relevant by a principle of ethics are concerned the situation may be well different. In this case the description may be not simply mute, but *vague* (*vagueness* is the typical trade-off we may face for the *comprehensiveness* of a general and abstract universalizable principle). This means however that unforeseen states of the world will exhibit a vague but *quantifiable membership* in relation to the set identifying the domain of application of the principle (an intermediate degree between 0 and 1 in terms of a *fuzzy membership function*). Indeed, unforeseen states are what make the domain of a general principle or norm a *fuzzy set*.

So far so good, but what about the firm's commitments? Assume that a precautionary rule of behavior is stated in order to preempt the occurrence of principle violations. This is not defined conditionally on a complete or clear ex ante description of any state wherein it must be conditionally implemented. On the contrary it is defined with reference to membership in the domain of the general principle, and the commitment to carry out the rule is undertaken conditionally on *just* an *ex ante required degree of membership* of any state, foreseen or not, of the *fuzzy set* that defines its application domain. The firm can undertake this commitment without knowing all the possible states ex ante, simply by establishing a *membership threshold* condition which is communicable ex ante to the stakeholders. This only requires assuming that whatever state (foreseen or not) may occur, both the players will be able to employ the fuzzy pattern recognition model of reasoning just illustrated. Satisfaction of the condition on the fuzzy membership threshold is shared knowledge *ex post* because the degree of vagueness of any state in regard to the principle can be commonly understood as a *matter of fact* concerning the ex post state of group

knowledge. But if this were not the case, the ex post degree of vagueness to be assumed as relevant by the firm is that expressed by an impartial and non malevolent observer who takes the point of view of every stakeholder in turn. Hence the firm can undertake ex ante its commitments and to be confident that ex post it will be able to account for them according to the stakeholders' vague judgment capacity. Note that the inference from satisfaction of the required threshold condition to implementation of the stated precautionary rule is *default inference*. It requires one to reason according to the format: "even if information on the case in point x is *incomplete* and it is *no verified truth* that this state belongs to the domain of the principle P….it is not *inconsistent* with the knowledge base that x∈P", or "given satisfaction of the threshold condition, *normally* a situation like x is such that x∈P" *so that* the required rule of behavior must be carried out[20].

In order to avoid the consequences of incomplete information on the formation of reputation, the firm will therefore explicitly announce and subscribe to a set of CSR abstract and general principles whose contents are such to elicit stakeholder consensus, as well as to the explicit commitment of conforming with pre-established precautionary rules when they are put at risk (under a *vagueness* interpretation of *risk*), which are both known *ex ante* (before the occurrence of unforeseen events) by stakeholders. Thus it is the CSR norm that enables the cognitive mechanism of reputation to function properly. In the absence of such a reference point, stakeholders could not develop trust because they would not be able to check whether the firm respects whatever commitments.

A cognitive role is played in the working of a CSR norm as a *gap filling device* (Coleman 1992) that states the types of behaviors that stakeholders can expect from the firm in situations where contracts fail owing to the absence of conditional provisos constraining residual decisions and abuse of authority. This cognitive function is primarily *constructive*. The *game form* (Aoki 2007) is badly specified under unforeseen situations because contingent strategies are unspecified over such states. But norms nevertheless allow by default the inference of how the honest type of the firm will behave under these circumstances. In fact these "strategies" are not defined contingently on states of the world that the parties are unable to write down in the contract or are even unable to foresee. Explicit norms complete the description of the game form by substituting default rules of behavior for conditional strategies. These rules are based on the satisfaction of a membership condition in the domain of abstract and general ethical principles that are ex ante

[20] On *logic default reasoning* see Reiter (1981); on its uses in game theory see Bacharach (1994), Sacconi (2000, 2007) Sacconi and Moretti (2008)

known and always ex post verifiable through a shared understanding of the inherent vagueness of the unforeseen contingencies. Once these norms have been stated ex ante in terms of precautionary standards of behavior, we are able to say how the firm is expected to behave in whatever unforeseen state that may put a general principle at risk, until a contrary proof is given that the principle does not apply to the new situation. In other words, the firm types implementing or otherwise strategies of conformity to norms are described. What is involved here is not inductive learning about the probability of an already given set of possible but uncertain set of types, but the very conception of the type set itself that contributes to an (approximate) description of what may occur in the future. Accordingly, their role is *constructive*: through the agreed statement of voluntary norms, firms and stakeholders *construct* an approximate model of the game form that they will play in states of the world that they are unable to describe ex ante in every detail.

The cognitive (and constructive) function of norms leaves us only half-way in our argument. A well conceived game form allows definition of the players' strategy combinations and the reputation equilibria wherein the firm may be described as acting in support of its reputation, so that after some time stakeholders begin to trust it. Under the usual condition of the long run player's non myopia, these equilibrium combinations include the firm's continuing not to abuse the stakeholders and the stakeholders continuing to enter the relation with the firm. However, in general, this will be *just one* of the many possible reputation equilibria of the game. Other equilibria will consist in strategy combinations including random compliance with the norm by the firm (a mixed strategy) such that the best response of stakeholders is to yield to the firm's strategy. These equilibria (see fig.3, the Stackelberg equilibrium in particular) admit that a firm has been able to accumulate a reputation for mixed levels of abuse that leaves stakeholder indifferent between entering or not entering the relation with the firm. In other words, when a repeated reputation game is constructively defined in terms of strategies abiding or not abiding with the ex ante agreed CSR norm, it will have too many equilibrium points, not only the preferable equilibrium where the firm abstains from abusing stakeholders by strictly complying with the norm and cooperates with them at any time (see fig. 3, where the equilibrium set X is depicted as the dashed area). The typical game theoretical problem of *multiple equilibria* arises. But here it turns out that CSR norms have three other roles to play in facilitating an equilibrium selection consistent with the normative model.

(insert fig.3 about here)

9. The normative role in the selection of an impartial equilibrium.

By normative role I mean the function of a fairness principle to give impartial reasons for singling out a unique equilibrium solution amongst the many possible. This will be a particular equilibrium point coinciding with an outcome rationally acceptable by all the participants in an ex ante decision on equilibria to be implemented later. Note that here the normative principle is used to choose an equilibrium point within the equilibrium set of the game to be played in the implementation phase. The perspective is still that of an *ex ante* impartial choice, but it now concerns equilibria, i.e. game solutions that are self-enforceable. Modeling this impartial decision simply requires understanding rational acceptability as equilibrium solution invariance in regard to symmetrical player replacement (i.e. invariance of the agreed solution under exchange of the players' positions – again a way to give operability to the 'veil of ignorance' idea)

To keep things simple, assume that there are only two players. Consider a standard representation of the equilibrium set of a repeated game coinciding with its (convex and compact) payoff space. Moreover, take a symmetric permutation of each player's place with respect to the set of equilibrium solutions of the game, i.e. the symmetric translation of the payoff space Y (see fig. 4) with respect to the Cartesian axes representing the players' utility functions and payoffs. Hence, for every equilibrium point in the original outcome set Y, whatever the payoff equilibrium point afforded to player A in the initial representation, the same payoff will be afforded to player B under the translated outcome set X, and vice versa.

(Insert fig. 4 about here)

Impartiality simply requires that the accepted solution is invariant under this payoff space translation, because the accepted solution must not depend on the strategic position that a player occupies in the game. An impartial solution is an equilibrium point that allows each player to achieve an outcome which is invariant whatever position a player happens to occupy. By contrast, a solution (given a particular representation of the game payoff space) is said to *depend* on the strategic position that players hold in the game if implementing the corresponding equilibrium point yields players a payoff that they could not obtain if the same equilibrium point were implemented under the symmetric translation of the payoff space - that is, under the symmetric

replacement of players with respect to the set of equilibrium outcomes. We require this translation invariance to be satisfied in order for the equilibrium point selected to be normatively considered *the solution*.

This can be achieved in two ways. *First*, assume that a player would ex ante accept (under a given representation of the payoff space) a certain equilibrium point as the solution, but under the payoff space translation this equilibrium point translates into a *different* point in the payoff space. Once the player places have been exchanged, the payoff space translation identifies a symmetric point corresponding to the same equilibrium, but this point (a payoffs vector) does not afford each player the same payoff as before (simply because it replaces the payoff of the "fortunate" player with that of the "unfortunate" one, and vice versa). Under the symmetric translation of the payoff space, on the whole players would agree on the *same* equilibrium point, but payoffs would be the symmetric translation of the initial ones, and hence cannot be invariant.

This fact does not imply that the solution cannot be invariant under the players' replacement. Invariance can be obtained by assuming that each player randomizes between the two payoffs that he can expect at the same equilibrium point when he takes the place of player A and that of player B in turn. An expected payoff reflects the required independence of the player's choice from any particular strategic position he may hold in the game in so far as the player is able to take both the positions (if they are only two) with the *same* probability.

Thus an equally randomized solution is consistent with impartiality.[21] But in the present exercise of impartial choice it raises a question of *feasibility* and implementation. This intuition was the starting point of Binmore's contribution to the theory of social contract (cf. Binmore 1991). Equally randomizing between the original representation of an equilibrium point on the one side, and its symmetrical translation on the other, identifies a mid-point between two equilibrium payoffs vectors, the first belonging to the initial set of equilibrium outcomes, the second to its symmetrical translation (all the points generated like this will lie on the bisector out coming from the origin, see fig. 5).

(Insert fig. 5 about here)

[21] The idea of solution invariance under the permutation of the individuals' point of views and the implied one of equally probable combinations of the two payoffs a player gets from a solution and its symmetrical translation in Binmore theory captures the idea of Rawlsian 'veil of ignorance', which is however thinner than the Rawls' model itself (1971). Cf. note 9 above.

This may be an equilibrium point on its own (for example in mixed strategies) or it may not be. If in correspondence to this mid-point there is an equilibrium point formed by strategies (pure or mixed) that in practice the players may carry out in the ex post game, then that equilibrium can be selected in order to generate an impartial solution. The case is different if the "mid-point" results only from the convex combination (joint randomization) of two points each belonging to one payoff space, whereas the probability combination falls outside both the original spaces and their intersection (see again fig.5). Joint randomization, in fact, is an admissible operation within the context of cooperative games, where joint strategies (plan of actions) can be randomized by an interpersonally valid random mechanism without fear that individual players will act according to separate mixed strategies in practice. But the game that we are considering is non-cooperative. Thus we are not allowed to generate from the original outcome space and its symmetric translation the convex hull of all their components. Save in the case where the initial payoffs space is also symmetric by itself, which is a very special case, it follows that the symmetric translation of payoffs associated with a given equilibrium point will carry that point into a new point that in general will not belong to the original space. Moreover, the equal-probability combination of the two points (a mid point) will not belong to the intersection of the two spaces. This means that there is no ex post equilibrium point to which the players may resort in order to implement their ex ante agreement. Such an equilibrium is *not feasible*.

Nevertheless a *second* way to satisfy the condition of replacement invariance is still available. It amounts to the restriction of selecting the acceptable solution only within the intersection of the original outcome space and its symmetric translation. Any selection within this set does not create the feasibility problem just considered because any point in the intersection set corresponds to an equilibrium point which is always existent as long as it belongs to both the original and the translated outcome sets. As before, invariance may be obtained for each player by taking as a solution the equally-probable combination of his payoff derivable from an equilibrium point and its symmetric translation. Such a mid-point will necessarily identify one equilibrium that any player can ex post achieve by a feasible pure or mixed strategy as long as it belongs to the intersection set. Binmore convincingly argued that equally probable combinations imply the egalitarian solution within the intersection set, and by simply adding the obvious Pareto efficiency condition it leads to the symmetric Nash bargaining solution (see fig. 5 again). This solution is identical to singling out the maximin solution with respect to the original payoff space - which amounts to vindicating Rawls (1971) in the apparently alien context of a game theoretic social contract (Binmore 1991, 1998, 2005).

However, to confine our treatment to the context of the repeated Trust Game, we are in a particularly favorable situation as far as our ex ante equilibrium selection exercise is concerned. With reference to the payoff space of the repeated Trust Game, the simple requirement of selecting a solution within the intersection of the basic payoff space and its symmetric translation is sufficient for singling out a unique solution, at least if the obvious Pareto condition is granted (see fig.6).

(Insert fig. 6 about here)

Thus we do not have to bother with the ethical intuitions underlying the egalitarian solution and the 'veil of ignorance' idea that justifies that of equally probable combinations as expressions of fairness. Simply, if the solution has to be invariant under the players' symmetric position exchanges with respect to the outcome set, plus the PO condition, then it cannot be but the symmetric Nash bargaining solution of the original game, In fact the intersection set coincides with the north-west boundary of the payoff space, which lies on the 45° line from the origin (see again fig.6). Because it is reduced to a segment of the 45° line (the bisector), the solution cannot but be the only point of this line segment belonging to the Pareto efficient set, i.e. the symmetric Nash bargaining solution. Note that the key point for this conclusion is simply that an impartial exercise of choice (replacement invariance) must select an equilibrium point *within* the intersection set, that is, an equilibrium point which must exist in any case and hence be implementable by each player whatever the position he occupies in the ex post perspective. A stability condition (the solution must lie in the set of those points that correspond to ex post implementable equilibria) joined to the weak fairness condition of invariance to players replacement is sufficient to derive the egalitarian solution. Students of *corporate governance* may be struck by the simplicity of this result, which contradicts much of the subject's credos[22]:

PROPOSITION:

> In order to select an institutional form of corporate governance under the constraint of being incentive compatible – i.e. implementable by an equilibrium point – do not bother with wealth maximization; instead look at the "egalitarian solution" (in the qualified sense of being the Nash symmetric solution within the intersection set resulting from the symmetrical translation of the outcome equilibrium set).

[22] For a detailed exposition of how the dogmas of welfare maximization and efficiency may permeate all the economics of institutions see Kaplow and Shavell (2002)

Thus an explicit normative model including a specific method of impartial reasoning helps resolve the multiplicity problem from the ex ante perspective. But we must not overemphasize this result. What would effectively solve the multiplicity problem would be an equilibrium selection theory able to predict the *ex post* game equilibrium solution so that it is consistent with the *ex ante* solution identified. In other words, selection is ex post effective only if it gives reasons to act that fit the ex post reasoning context. Ex post, only common knowledge of the solution, i.e. a system of mutually consistent expectations converging on the prediction of a uniquely determined equilibrium point, conveys to each player the appropriate reason to act, because choosing an equilibrium strategy amongst many others requires having a clear prediction of other players' behavior and beliefs. But given the fact that in the ex ante perspective a solution is invariant to the players' position replacement, there is no logical reason to conclude that it will be effectively implemented. The reason that justifies a particular decision in the ex post game is knowledge of what the players will effectively do. Moreover, this knowledge about the other players' decisions must be consistent with their being symmetrically able to predict our behavior and to choose their best response to these predictions. Therefore it is not the impartial selection of a desirable ex ante solution but the knowledge of other players' *de facto* behaviors that provides the proper reason for acting in the ex post context. And there is no logical implication from what is fair ex ante selection (even if it falls on an equilibrium point) as to what other players will actually do. Maybe they will act in accordance with the principle, maybe not. The fair ex ante agreement, or impartial choice, does not gives us common knowledge of the ex post behavior of players. But if we actually do not know how other players' will behave, we have no reason to play a given strategy, even though the fair solution is part of an equilibrium point.

This is not to say that the ex ante agreement on an impartial solution does not provide any reason to believe that players will act according to the same principle in the ex post interaction. This is simply a matter of fact or of cognitive psychology, not a matter of logic. Common knowledge, on the contrary, is a matter of epistemic logic: this means recursive group knowledge of what everybody knows to be true (a *truism*).[23] It may be the case that a given equilibrium is known to be played only if each player has many layers of knowledge about every other player's action, beliefs, beliefs about beliefs, etc., that are consistent and justify the prediction that this equilibrium will be played. This state of knowledge can be approximated by a theory of belief

[23] The ex post rationality of Nash equilibrium - implied by the notion of common knowledge- was already clear in Lewis (1968), who also suggested that an agreement could give just an empirical explanation of how a state of common knowledge could emerge. He however focused on the different cognitive phenomenon of salience. On game theoretic definition of common knowledge see Binmore, Brandeburger (1990) and Kreps (1990); on the epistemic logic of common knowledge see Fagin, Halpern, Moses and Vardi (1996).

formation that at last leads us to a stable prediction of any other player's equilibrium choice and belief (see the predictive function of norms, sec. 11, *infra*). Ex ante selection on the contrary does not predict how we will actually decide; it only answers the question about what equilibrium should be chosen because it is invariant under the individuals' position replacement. The step from an answer to the question of what equilibrium is fair to an answer to the different question of how players will actually behave is a default inference that some player may in fact make (see again sec. 11), but there is nothing necessary about it. Thus in the perspective of the ex post game much has still to be done in order to say that the multiplicity problem has been solved.

10. Motivational role. Are all the equilibria equally capable to provide motivations?

Any equilibrium point exerts a (limited) motivational force able to command actual behavior, which is effective in so far as each player believes that other players will play their strategy components of the same equilibrium. It may be wondered whether the facts that a norm has been agreed from an ex ante (pre-play) perspective, and exhibits different levels of consistency with different equilibria, may affect the motivational force exerted by different equilibria in a game. A positive answer would amount to a restriction on the number of equilibrium points that have motivational force over the players' behavior. In other words, it may be asked whether norms can "refine" the equilibrium set of a game in terms of the motivational strength of certain equilibria over other equilibria.

A voluntary CSR norm would in fact perform a motivational function in the restriction of the admissible equilibrium set in the event that – having been chosen via a unanimous impartial agreement and granted that players expect reciprocal compliance with the norm – it generates an additional utility weight to be introduced into the payoffs of the players. The conjecture is that a preference for equilibrium strategies may in part depend not just on their outcomes but also on the level of conformity that any equilibrium exhibits in regard to an agreed norm. A conformity level must be understood as conditional on beliefs – i.e. conformity depends on one player's compliance given his beliefs about the other players' behaviors and about other players reciprocity in compliance, given their beliefs. It follows that the additional psychological payoff involved by a given level of conformity is not just an exogenous parameter reflecting the absolute motivational force of the desire to be consistent with an agreed norm. The exogenous component is also conditioned by a function of beliefs concerning reciprocal behaviors.

Anyway, if the norm generates a modification in the players' payoffs in favor of those situations wherein no significant deviation from reciprocal conformity occurs, then it may be the case that the overall motivational strength reinforcing an equilibrium behavior may be integrated (relatively augmented or reduced) by an additional motivational factor that in the end confines overall motivational strength only to those equilibria that exhibit significant compliance levels with the norm.

The reference is of course to a different notion of equilibrium, i.e. the psychological Nash equilibrium (Genakoplos et al. 1986) based on conformist preferences (Grimalda and Sacconi 2005, Sacconi and Grimalda 2007)[24]. This results from a modification of the players' utility functions through integration of preferences with an intrinsic component for norm-compliance, seen not as unilateral and unconditioned but as conditioned by beliefs about other players' reciprocal conformity. The "refinement effect" on the admissible equilibria that this change in the equilibrium notion entails is quite surprising (and unexpected). As we will see, the equilibrium set of the repeated Trust Games under this revision of the utility function shrinks dramatically to the pure strategy equilibria of the repeated psychological Trust Game.[25]

10.1. The conformist preferences model and the TG

To begin with, let us informally restate the conformist preference model with reference to the Trust Game involving a firm (player B) and its stakeholder (player A). Stakeholder and firm have *two* kinds of preferences, both able to motivate their action. On one hand (more basic), they describe the outcomes of their interaction as consequences, and their preferences regarding consequences are defined as *consequentialist*. These may be not only typical self-interested preferences but also altruistic ones.

This part of the argument is by no means new. The new part instead concerns *conformist preferences*. Players also have preferences defined over states of the world resulting from their interaction, which are described in terms of their consistency with an ex ante agreed ethical norm - where "consistency" is how far the players' strategy choices (jointly a state) are from the set of actions that would completely fulfil the agreed ethical norm of equity. By norm I mean a principle of

[24] Relevant literature on psychological games and reciprocity also includes Rabin (1993), Chareness and Dufenberg (2006), Segal and Sobel (2007).

[25] The extensive literature on equilibrium refinements (see van Damme 1987) may be seen as an indirect approach to equilibrium selection in the sense that by specifying additional requirements on the solution concept it reduces admissible elements of the Nash equilibria set. By contrast, psychological games are not usually seen as "refinements" for they seem to enlarge the equilibrium set with reference to the Nash equilibrium set. This refinement effect is hence a peculiar and somewhat surprising result of conformist preferences model within the TG context.

justice for the distribution of material utilities coinciding with the stakeholders' social contract of the firm.

Let us assume that players have just agreed upon a social contract concerning the principle of justice that should govern as a norm the distribution of the social surplus produced by means of their cooperation through the firm. Conformist preferences may now enter the picture. Intuitively speaking, a stakeholder will gain intrinsic utility from the simple fact of complying with the principle, if the same stakeholder expects that in this way she/he will be able to contribute to fulfilling the distributive principle, admitted that she/he expects the other stakeholders (or the firm) also to contribute to fulfilling the same principle, given their expectations.

A complete measure of conformist preferences consists in the combination of the following four elements through the conformist-psychological component of a player utility function (see Grimalda and Sacconi 2005):

First, a principle T, which is a social welfare function that establishes a distributive criterion of material utilities. Players adopt T (the norm) by agreement in a pre-play phase, and employ it in the generation of a consistency ordering over the set of possible states σ, each seen as a combination of individual strategies. The highest value of T is reached in situations σ where material utilities are distributed in such a way that they are mostly consistent with the distributive principle T within the available alternatives. Note that what matters to T is not "who gets how much" material payoff (the principle T is neutral with respect to individual positions), but how utilities are distributed across players. Satisfaction of the distributional property is the basis for conformist preferences. Let us assume that T coincides with the Nash bargaining function.

Second, a measure of the extent to which, given the other agents' expected actions, the first player by her/his strategy choice contributes to a fair distribution of material payoffs in terms of the principle T. This may also be put in terms of the extent to which the first player is *responsible* for a fair distribution, given what (he expects that) the other player will do. It reduces to a conformity index assuming values from 0 (no conformity at all, when the first player chooses a strategy that minimizes the value of T given his expectation about the other strategy choice) to 1 (full

conformity, when the first player chooses a strategy that maximizes the value of T given the other player's expected strategy choice).[26]

Third, a measure of the extent to which the *other* player (respectively the stakeholder or the firm) is expected to contribute to a fair distribution in terms of the principle T, given what he (is expected to) expects from the first player's behaviour. This may also be put in terms of the (expected) *responsibility* of the other player (i.e. the firm) for generation of a fair allocation of the surplus, given what it (is believed to) believes. This reduces to a reciprocal conformity index assuming values from 0 (no conformity at all, when the other player is expected to choose a strategy that minimizes T given what he expects from the first player) to 1 (full conformity, when the other player is expected to maximize the value of T given what he expects from the first players) formally identical to the conditional conformity index of the first player .

Fourth, an exogenous parameter λ representing the motivational force of the agent's psychological disposition to act on the motive of reciprocal conformity with an agreed norm.

Steps two and three coalesce in defining an overall index F of conditional and expected reciprocal conformity for each player in each state of the game. This index operates as a weight (again between 0 and 1) on the exogenous parameter λ deciding whether λ will actually affect or not (and, if so, to what extent) the player's payoffs. [27]

---

26 Player *i's* personal index of conformity is

$$\left[1 + f_i\left(\sigma_{ik}, b_i^1\right)\right]$$

where the function f_i (which varies from 0 to –1) measures player *i*'s deviation degree from the ideal principle *T* due to player *i*'s k-ary choice, given her expectation about player *j*'s behavior, and is the following

$$f_i\left(\sigma_{ik}, b_i^1\right) = \frac{T\left(\sigma_{ik}, b_i^1\right) - T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right) - T^{MIN}\left(b_i^1\right)} \qquad (2)$$

where b_i^1 is player *i*'s belief concerning player *j*'s action, $T^{MAX}\left(b_i^1\right)$ is the maximum attainable by the function *T* due to whatever feasible strategy player *i* may choose given *i*'s belief, $T^{MIN}\left(b_i^1\right)$ is the minimum attainable by the function *T* due to whatever player *i*'s feasible choice is given *i*'s belief, and $T\left(\sigma_{ik}, b_i^1\right)$ is the effective level attained by *T* when player *i* adopts his k-ary strategy σ_{ik} given his belief about player *j's* action.

[27] The overall utility function of player *i* with reference to the state σ (understood as a strategy combination of player *i* strategy σ_i and the other players strategies σ_{-i} , is the following

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

where

i. U_i is player *i*'s material utility for the state σ;

Summing up the effect of the different components, if a stakeholder expects that the firm (or vice versa) is responsible for the maximal value of T, given what the firm expects about his/her behaviour, and he/she also is responsible for a maximal value of T, given the firm's (expected) behaviour, then the motivational weight of conformity λ will enter his/her utility function. That is, in the player's preference system it will show all the force of the disposition to conform to agreed norms, so that complying with the principle will yield utility (in the psychological sense) additional to the material payoff of the same strategy. In the Trust Game this clearly happens at best in the state where the stakeholder enters, the firm does not abuse, and they mutually predict these strategy choices.

Note that if a player cannot improve on the value of T simply by means of his unilateral choice, given the expected strategy choice of the other player, then he will be considered completely compliant (no deviation from the maximum value of T reachable by his choice can be ascribed to his responsibility). This feature of the model depends on the fact that we are considering compliance in a non-cooperative ex post context wherein players must be able to deviate from an agreed norm. Hence, in cases like the Trust Game, if the firm is expected to abuse, the stakeholder's decision to stay out cannot improve on the value of T and hence the stakeholder, being incapable to improve over the status quo, will be considered fully compliant with the principle. At the same time, when the stakeholder is predicted to stay out, given his prediction of the firm's abuse, by the firm by abusing cannot modify the value of T. Thus whichever the firm's strategy choice, it is fully compliant in this case. The result is that in the (no-entry, abuse) equilibrium point of the basic Trust Game we also see the conformity weight λ adding to the players' payoffs. Under this respect thre is no difference with the case where the stakeholder enters predicting that the firm is not abusing, while vice versa the firm refrains from abusing given its prediction that the stakeholder will enter.

By contrast, when the firm is unilaterally predicted to abuse, if the stakeholder enters he would minimize T with reference to the alternative choice open to him (no-entry), which scores a higher level of T. At the same time the firm misses the opportunity to maximize T given the

---

ii. λ_i is an exogenous parameter that may be any positive number and expresses the motivational force of the disposition to comply with an agreed principle or norm;

iii. *T is* a fairness principle (assumed to be a *social welfare function* with the specific form of NBS), whose value here is defined for the state σ;

iv. *F* is a compounded index expressing both the agent *i*'s conditional conformity and the other individuals' expected reciprocal conformity with principle *T* in state σ , given player *i*'s beliefs of first and second order (i.e. beliefs about other players' first order beliefs) predicting that state σ is in fact the case.

stakeholder's decision to entry, and hence he will be considered as not complying at all. This implies that when the firm unilaterally and successfully abuses its stakeholder no conformist preference adds any value to the players' material payoffs.

Last, if the firm chooses a mixed strategy whereby the stakeholder's decision between entry or non-entry is not influential on the T value, the stakeholder on making any choice would be unable to improve the value of T by both entering or not entering - hence by staying out he maximises T as well. But if the stakeholder still stays out, no other firm's strategy can do any better in maximizing T than the one just described, and thus the firm is also completely compliant when it abuses. Hence a firm's equilibrium mixed strategy responded to by the stakeholder's no entry strategy implies that conformist weights are added to the player's payoffs. On the contrary, were the stakeholder willing to enter when the firm adopts the mixed strategy (so that by entering he is equally compliant as when staying out), the firm becomes responsible for a sharp deviation from full compliance, for it could have chosen not to abuse at all. In fact it has not maximized the value of T as it could have done. Maybe this is not the minimum value for T, but nevertheless it has produced a significant deviation from full compliance (proportional to the distance from the maximum value of T conditional on the stakeholder's choice). Thus, in this case the motivational weight of conformity cannot enter the utility functions of both players in all its strength.

Focusing for the moment only on the choice by the stakeholder, this may dramatically change his overall utility calculation in order to decide whether or not to surrender to a company's strategy of sophisticated opportunism which manages to keep compliance with the CSR norm to its minimum compatible with the stakeholders' incentive to entry. Whereas in the iterated Trust Game the mixed strategy may reward stakeholders to a sufficient extent to convince them to play entry, this is not the case in its psychological game. A stakeholder will refuse to surrender and will prefer a more resolute 'hard-nosed' approach by quitting the relationship (as we have seen, this will also set to zero the firm's responsibility for any deviation from compliance).

However, the conformist preference model has much more to say about equilibrium "refinement" when we consider a case of reciprocal conformism, that is, the case when conformist preferences are present on both sides. Consider a small change in the Trust Game payoffs completely non influential on its basic logic. The four pure strategy combinations are now (*no-entry, abuse*) and (*no-entry, no-abuse*) with material payoffs (1,1); (*entry, abuse*) with material payoffs (0,5); (*entry, no-abuse*) with material payoffs (4,4). This is simply useful to understand what is meant by calculating the level of conformity in the different states by applying the Nash

bargaining solution, which requires maximizing the product of individual surpluses net of the *status quo*. In our case the *status quo* coincides with the outcome of the no-entry strategy i.e. (1,1), which in fact is the assurance level that player A can grant himself whatever player B's choice. This payoff must then be subtracted from whatever payoff is used in the calculation of the Nash product annexed to any state (strategy combination). The three matrices (see below) then show (a) the new TG in normal form, (b) the Nash bargaining product calculated for each pure strategy combination needed to measure the consistency of each state with the combination, and (c) the overall payoffs resulting from the addition of the psychological conformist preference weight $\lambda = 2$ to the material payoffs where this addition is appropriate.

	$\neg a$	a
e	4,4	0,5
$\neg e$	1,1	1,1

Matrix (a):TG normal form

	$\neg a$	a
e	(4-1)(4-1) = 9	(0-1)(5-1) = - 4
$\neg e$	(1-1)(1-1) = 0	(1-1)(1-1) = 0

Matrix (b): T values at each state

	$\neg a$	a
e	$(4+\lambda) = 6, (4+\lambda) = 6$	0, 5
$\neg e$	1,1	$(1+\lambda) = 3, (1+\lambda) = 3$

Matrix (c) : psychological TG with conformist utilities included with $\lambda = 2$

Inspection of matrix (b) shows that if the firm is predicted to play strategy ***a***, the stakeholder maximizes T by playing strategy $\neg e$. If this is known, the firm also maximizes T by playing ***a***, since neither strategy is better or worse than ***a*** in order to maximize T from the firm's point of view. Hence in the bottom right cell of matrix (c) the psychological weight λ adds to each player's material payoff. On the other hand if the firm is predicted to play $\neg a$, then the stakeholder maximizes T by choosing ***e***, and if this choice is predicted by the firm, its choice maximizing T is $\neg a$ as well. Consequently in the top left cell of matrix (c) psychological weights λ are also present. If the firm plays ***a*** the stakeholder will minimize T by **e**, which is also true if the same

result is seen the other way round (given **e**, the firm minimizes T by abusing with ***a***). No weights must be added in the top right cell of matrix (c). Last, if the firm is predicted as not abusing, the stakeholder minimizes T by staying out with ¬***e***. Even though the firm is maximizing T when it plays ¬***a*** , a zero index of individual conformity (the stakeholder's one) is sufficient to nullify the overall level of conformity. Also when this is the case no psychological conformity weights are implied in the players' payoffs.

10.2 Mixed strategies and the repeated psychological TG

Now consider the repeated Trust Game (TG). Recall that its payoff space is the convex hull of all the linear (probability) combinations of the three payoffs vectors generated out of the pure strategy pairs of the basic Trust Game. This is the same as representing the expected payoffs of every possible pair of pure and mixed strategies of the two players in the basic Trust Game. In fact the player's *i* expected payoff for a mixed strategy is formally the same as the *average payoff* of the player's *i* repeated strategy of the repeated game that employs alternatively the two player's *i* pure strategies of the stage game with a given frequency, generating the three stage game outcomes (1,1,), (4,4), (5,5) according to the frequency of the two players' choices. The cumulative payoff of this repeated strategy, given a certain pure (or mixed) response by the second player, can be equated to the average payoff of a cycle along which player *i* gets each of the three stage-game payoffs a given proportion of times out of the total number of times in the cycle (granted, of course, that along the game each repeated strategies pair of the two players repeatedly enters a cycle with the same pattern of outcomes and the same average payoff value for each player). Hence it is simple to see that a firm's mixed strategy that employs the two pure strategies ¬***a*** and ***a*** with probability 0.25 and 0.75 respectively against - to keep things simple - the stakeholder's pure entry strategy ***e***, affords the firm and the stakeholder expected payoffs (0.25×4+0.75×5 = 4.75) and (0.25×4+0.75×0 = 1) respectively, equal to the average values attached to a repeated strategy whereby the firm plays the stage-games strategy ***¬a*** the 75% of the times and the stage-games strategy ***a*** 25% of the times assuming – to keep things simple again – that the stakeholder always responds with the stage-game strategy ***e***. Obviously it is true that in the one-shot TG no mixed strategy is a best response for the firm. But in the repeated TG we know this is no longer true. In fact, the firm may create a reputation (along, say, the first N repetitions of the game) to be a *type* that uses *the strategies* ¬***a*** and ***a*** in a given frequency, such that the stakeholder's best response is "always ***e"*** until by repeated observations he realizes that the frequency is respected, but sanctioning by "¬***e*** forever" were it to become clear that the

frequency is not respected. This induces the firm to stick to its repeated strategy, mixing ***a*** and ¬***a*** according to the given frequency.

But then we must consider the payoff space of the psychological game that can be generated from that of the Trust Game when all the expected payoffs of mixed strategy pairs are accounted for. This psychological TG in pure and mixed strategies has the same payoff space as the repeated psychological TG wherein the average payoffs of each repeated strategy – which employs the pure strategies of a player in a given frequency – is identical to the expected utility of the mixed strategy using the corresponding probability mixtures. Hence we may ask what happens to the mixed strategy equilibrium points of the corresponding standard repeated TG.

Before answering this question, we must define a way to calculate the expected psychological utility of any mixed strategy. Take the point of view of the stakeholder (call him A) when he predicts the firm (call it B) will choose a mixed strategy, for example

$$\sigma_B^{0.6} = \{(0.6, \neg \boldsymbol{a}); (0.4, \boldsymbol{a})\}$$

A believes that, if he enters by playing the pure strategy ***e***, two states (***e***, ¬***a***) and (***e***, ***a***) may occur, so that two different values of the principle T – i.e. (9) and (-4) – can arise, each of them weighted with the probabilities 0.6 and 0.4. of the respective states. Hence the expected Nash bargaining product generated by B's mixed strategy $\sigma_B^{0.6}$, given A's entrance, is $0.6\times9 + 0.4\times(-4) = 3.9$ whereas if A does not enter, the expected T value is 0 as usual. Given $\sigma_B^{0.6}$, player A's strategy ***e*** maximizes T in respect of any other pure or mixed strategy by A, whereas ¬***e*** minimizes it. It turns out that player A's conformity indexes are 1 and 0 for his pure strategies respectively.

On the other hand, player B's conformity indexes are the following. Assuming that B believes A will enter, B does not maximize T by playing the strategy $\sigma_B^{0.6}$, because it is obvious that no-abuse would do better in terms of T. Nor does playing the mixed strategy minimize T, which in fact would happen by playing ***a***. As a result, B's conformity index for strategy $\sigma_B^{0.6}$ actually is a somewhat intermediate value 0.61. But assuming that B believes that player A will not enter by ¬***e***, then B's mixed strategy $\sigma_B^{0.6}$ will maximize T no less than any other B's strategy. B's conformity index under this hypothesis is hence 1. To conclude the example, consider A's respective expected material payoffs from playing ***e*** or ¬***e*** against the mixed strategy $\sigma_B^{0.6}$

$$EU_A(\boldsymbol{e}, \sigma_B^{0.6}) = 2.4, \quad EU_A(\neg \boldsymbol{e}, \sigma_B^{0.6}) = 1$$

Similarly, player's B expected material payoffs from playing the mixed strategy against the two pure strategies of player A are

$EU_B(e, \sigma_B^{0.6}) = 4.4, \quad EU_B(\neg e, \sigma_B^{0.6}) = 1$

Since the conformity indexes of players A and B for the strategy pair $(\boldsymbol{e}, \sigma_B^{0.6})$ are 1 and 0.61 respectively, the psychological conformity weight λ will enter the players' utility functions accordingly. - i.e. by a value $(1)(0.61)\lambda$. Given $\lambda = 2$, the weight of the conformist motivation is 1.22 and the overall utility payoffs of players A and B are 3.62 and 5.62 respectively.

In the repeated psychological TG, these payoffs correspond to the following pair of player B's and player A's repeated strategies: player B employs his pure strategies $\neg\boldsymbol{a}$ and $\boldsymbol{a}$ repeatedly with frequency 0.6 and 0.4 respectively, and by this repeated strategy tries to convince player A (or the sequence of short run players that participate in the repeated game in the position of A) that he will stick to this frequency forever. Player A decides to play repeatedly his entry strategy $\boldsymbol{e}$ as long as he does not see player B employing *abuse* with a frequency higher than 0.4, but if this frequency is exceeded he will switch to "$\neg\boldsymbol{e}$ forever". Since player A's threat seems quite convincing, player B plays *ad infinitum* his above-defined mixed repeated strategy. Assume that exactly 100 times are sufficient to say that the required frequency has been verified so that - if the players adopt the pair of repeated strategies described above - 100 times is a cycle that repeats more and more along the repeated game with always the same proportion of stage-games having outcomes $(\boldsymbol{e, a})$ and stage games having outcome $(\boldsymbol{e, \neg a})$. The average payoffs for this pair of repeated strategies - including the psychological component – is the vector (3.62, 5.62). It would seem to be a good incentive for player A to yield to player B's mixed abuse strategy, but before saying anything about equilibria we have sill to wait for a while.

Following the method mentioned above, under the hypothesis $\lambda = 2$, it is in fact possible to account for the entire payoff space of the psychological Trust Game including mixed strategies as well (see fig. 7).

(Insert fig. 7 about here)

First note that the status quo point (1,1) – the only Nash equilibrium of the *basic one-shot* TG and moreover an equilibrium of the *repeated* TG – is translated toward North-East along the bisector

to a point with overall utilities (3,3), which is also a psychological equilibrium of the new game. At the same time, thanks to the motivational conformist weights $\lambda = 2$, the outcome (4,4) where the Nash bargaining product is maximized translates toward North-East to the point (6,6), which is also a psychological equilibrium. Recall that both these psychological equilibria correspond to Nash equilibria of the repeated TG, so that these two equilibria are certainly preserved under the payoff change provided by conformist preferences.

In regard to player B's mixed strategies, it can be seen that the entry strategy ***e*** of player A cannot be rewarded with any additional psychological conformist utility until the expected Nash Bargaining product - i.e. the expected value of T associated with any particular probability mixture of the two pure strategies $\neg$***a*** and ***a*** - is not positive, granted player A uses ***e***. This necessarily happens until a mixed strategy associates the pure strategy $\neg$***a*** with a probability high enough to give the respective T value (9) a weight able to counterbalance the T value of ***a*** (i.e. -4), so that the T expected value exceeds the T level fixed by the "status quo" no-entry strategy (which is 0). Hence, within player B's continuous set of probability mixtures of two pure strategies $\neg$***a*** and ***a,*** the relevant threshold is fixed by player B's mixed strategy that scores an expected Nash product no different from the T value of staying out. As long as this threshold is not exceeded, psychological payoffs do not add any values to the material payoffs of both players A and B, because entering by ***e*** minimizes the T value and exhibits zero conformity level. This is true also when player B adopts a mixed strategy that makes him partially, and hence positively, compliant. In fact until player A's choice to enter by ***e*** exhibits a zero conformity index, the overall conformity level is also nil for both players and no psychological payoffs may add to their material payoff.

This does not means that psychological utilities are not at work for these mixed strategies. Simply, the psychological component adds to the payoffs of strategy pairs like (*no entry*, *mixed strategy*), which is the same as for the strategy pair (*no entry*, *abuse*), i.e. (3,3). This means that the best responses for these cases is $\neg$***e***, which gives player A an overall payoff 3 whereby player B's mixed strategies and the pure strategy ***a*** become indifferent as they both give B the same overall payoff 3.

As an example, consider the mixed strategy $\sigma_B^{0.25} = \{(0.25, \neg\boldsymbol{a}); (0.75, \boldsymbol{a})\}$. The expected Nash bargaining product (i.e. the T value) is negative (-0.75) for the pair ($\boldsymbol{e}$, $\sigma_B^{0.25}$), whereas T is 0 if player A chooses $\neg$***e***. Hence it is obvious that A maximizes T by choosing $\neg\boldsymbol{e}$ **,** with conformity index 1, whereas the conformity index for choosing ***e*** is 0. As a result, by entering with ***e***, player

A can only get the expected overall payoff 1, which - due to the probability mixture provided by $\sigma_B^{0.25}$ - is no different from the *material* payoff of staying out. But by staying out with $\neg$**e** he gets an *overall* payoff 3, because the psychological conformist weight 2 now adds to this strategy material payoff. Thus A's best response is obviously to stay out. As far as player B is concerned, the mixed strategy $\sigma_B^{0.25}$ against ***e*** gives a payoff equal to its material payoff 4.75**.** When player A does not enter against $\sigma_B^{0.25}$, B's payoff benefits from the psychological conformist component (becoming 3) as well as from any other choice by B when he knows that A plays no-entry.

Note the importance of the mixed strategy $\sigma_B^{0.25}$. This is player B's Stackelberg mixed strategy that, from the one-shot TG, would correspond to the preferred (by the firm) equilibrium strategy of the repeated TG. It identifies exactly the equilibrium point of the repeated TG which would be the most obvious choice from the point of view of player B were he able to select the solution of the game by himself. It is noticeable, however, that the pair (**e**, $\sigma_B^{0.25}$) is not an equilibrium in the psychological TG even if player B's material payoff is quite high. Given strategy $\sigma_B^{0.25}$ neither is player A's best response **e**, nor is player B's material payoff 4.75 sufficient to make the strategy $\sigma_B^{0.25}$ preferable to ***a*** when A plays ***e,*** simply because, due to a sufficiently high λ associated with the psychological equilibrium in pure strategies (*entry, no-abuse*), playing $\neg$***a*** pays B more (6).

The threshold that allows mixed strategies to gain support from psychological conformist utility is reached at the mixed strategy $\sigma_B^{0.307} = \{(0.307, \neg\boldsymbol{a}); (0.693, \boldsymbol{a})\}$. Actually, given this mixed strategy, the expected value of T is zero for any strategy choice by A, so that A is fully conformist by choosing either ***e*** or $\neg$***e.*** At the same time playing the mixed strategy is partially conformist also for player B, because the minimum T value given A's entrance would be obtained by playing *a*. Hence, under the pair (***e***, $\sigma_B^{0.307}$), psychological utilities add to both the players' material payoffs (1.3, 4.7) generating an overall payoff vector (1.84, 5.31). But it is important to note that adding a bit of psychological utility does not mean that this strategy combination becomes a psychological equilibrium. Although it is true that player B's mixed strategy $\sigma_B^{0.307}$ grants a positive overall payoff to A's entry strategy, the player A's overall payoff from no-entry (3) is still higher than the overall payoff (1.84) from giving in to player B's mixed strategy. This is due to the incomplete conformity level of strategy $\sigma_B^{0.307}$ when player A chooses ***e***. B's full conformity would be reached by the strategy $\neg$***a,*** whereas $\sigma_B^{0.307}$ scores only the conformity index 0.31. This affects the psychological conformist component of player A's overall payoff for strategy ***e***, which is lower than for ***e***.

Now consider mixed strategy $\sigma_B^{0.39} = \{(0.\ 0.39, \neg\boldsymbol{a}); (0.\ 61, \boldsymbol{a})\}$. With this small increase in the probability of strategy $\neg\boldsymbol{a}$ things finally seem to change. Player A with overall payoff 2.36 benefits substantially from the psychological conformist utility of his entry strategy $\boldsymbol{e}$. At the same time, as typically happens when a pure strategy is surpassed in its conformity index, player A's conformity index of no-entry drops to zero since choosing $\neg\boldsymbol{e}$ given $\sigma_B^{0.39}$ would minimize the value of T in respect of the alternative entry strategy (and also any other mixed strategy). Hence also player A's overall utility for the no-entry strategy $\neg\boldsymbol{e}$ dramatically drops to 1 (just the material payoff). Moreover, for the pair $(\boldsymbol{e}, \sigma_B^{0.39})$ the overall payoff of player B contains a substantial psychological conformist component such that his overall payoff now reaches 5.41, whereas if player A were to choose $\neg\boldsymbol{e}$ player B's payoff would be reduced just to his material payoff 1, since the conformity index of player A's strategy $\neg\boldsymbol{e}$ is zero (though B's index remains positive). But note that this does not imply that we are at an equilibrium point. Even though entry is player A's best reply to player B's mixed strategy $\sigma_B^{0.39}$, this strategy is not reciprocally player B's best response. The perfectly compliant strategy $\neg\boldsymbol{a}$ would do better in terms of conformity index, so that it scores an overall payoff 6 higher than the mixed strategy.

This suggest a general fact about the model. In fact, consider again the mixed strategy

$$\sigma_B^{0.6} = \{(0.6, \neg\boldsymbol{a}); (0.4, \boldsymbol{a})\}$$

As we know, player A's conformity index if he uses strategy $\boldsymbol{e}$ against $\sigma_B^{0.6}$ is 1, whereas the mixed strategy's conformity index is 0.61. The annexed overall payoffs are (3.62, 5.62) respectively. Even though high psychological conformist utility enters both the players' payoffs this is not enough to define reciprocal best responses at $(\boldsymbol{e}, \sigma_B^{0.6})$ since, given player A's entry strategy, player B's best reply is again no-abuse at all with its overall payoff 6.

10.3 Equilibrium set of the psychological repeated TG

In order to give a general assessment of the two players' best reply sets in the psychological TG assume that λ is high enough for the pure strategy equilibrium $(\boldsymbol{e},\neg\boldsymbol{a})$ to exist. Call $E^{n|e}(\Pi_{A,B})$ the expected Nash Bargaining Product corresponding to player B's n-ary mixed strategy σ_B^n (where the index n corresponds to the probability weight assigned to the pure strategy $\neg\boldsymbol{a}$) given player A's strategy $\boldsymbol{e}$. Hence let $\Pi_{A,B}$ denote a generic Nash bargaining product. Last, call "status quo" the material payoff granted by A's pure strategy $\neg\boldsymbol{e}$. The relevant facts about the psychological TG are the following:

- *Case 1*, $\forall\sigma_B^n$ with $n \geq 0$ s.t. $E^{n|e}(\Pi_{A,B}) < 0$, i.e. such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 > E(\Pi_{A,B})^n$, the pure strategy $\boldsymbol{e}$ does not add any psychological conformist utility to player A's material payoff, whereas the pure strategy $\neg\boldsymbol{e}$ adds the psychological conformity weight λ to the "status quo" material payoff. Hence player A's best reply is $\neg\boldsymbol{e}$ whereby *any* mixed strategy in this case is as good as strategy $\boldsymbol{a}$ to player B. The equilibrium for this case is the psychological equilibrium point ($\neg$**e**, $\boldsymbol{a}$). This equilibrium is weak since every mixed strategy in this case gives player B the same overall payoff as $\boldsymbol{a}$.

- *Case 2*, $\forall\sigma_B^n$ with $0<n<1$ s.t. $E^{n|e}(\Pi_{A,B}) > 0$, i.e. such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 < E(\Pi_{A,B})^n$, each pair $(\boldsymbol{e}, \sigma_B^n)$ adds some psychological conformist utility to both players' material payoffs, whereas the pure strategy $\neg\boldsymbol{e}$ reduces player A to the "status quo" material payoff. This follows from the minimal conformity index of strategy $\neg\boldsymbol{e}$, while in this case mixed strategies σ_B^n have positive conformity indexes strictly less than 1. Thus for both players A and B there is an intermediate overall index F of conditional and expected reciprocal conformity. In this case player A's best reply is strategy $\boldsymbol{e}$. Nevertheless, against strategy $\boldsymbol{e}$ player B's best is $\neg\boldsymbol{a}$. In other words, as little as player B's psychological conformist utility of a mixed strategy σ_B^n is positive, player B's pure strategy $\neg\boldsymbol{a}$ against $\boldsymbol{e}$ (or whatever mixed strategy by player A) induces a psychological conformist payoff higher than σ_B^n, so that player B has an incentive to deviate from σ_B^n to $\neg\boldsymbol{a}$. When this occurs, obviously player A has no reason to change his choice and the equilibrium point is (**e**, $\neg\boldsymbol{a}$).

- *Case 3*, for a single $0<n<1$ $\exists\,\sigma_B^n$ such that $E^{n|e}(\Pi_{A,B}) = 0$, i.e. such that the pure strategy $\neg\boldsymbol{e}$ induces $\Pi_{A,B} = 0 = E^{n|e}(\Pi_{A,B})$. In this case both the strategy pairs $(\boldsymbol{e}, \sigma_B^n)$ and $(\neg\boldsymbol{e}, \sigma_B^n)$ add positive psychological conformist utility to the material payoffs of both the players A and B. Nevertheless, player A's overall payoff gained from $(\neg\boldsymbol{e}, \sigma_B^n)$ strictly dominates his overall payoff gained from $(\boldsymbol{e}, \sigma_B^n)$ since, whereas the two pure strategies $\boldsymbol{e}$ and $\neg\boldsymbol{e}$ score the same conformity index, the case of player B' conformity indexes is quite different. Player B against $\neg\boldsymbol{e}$ cannot do any better than play σ_B^n with conformity index 1, but given $\boldsymbol{e}$ the strategy σ_B^n conformity index is strictly less than 1, which is the conformity index of his pure strategy $\neg\boldsymbol{a}$. Since the strictly less than 1 conformity index of strategy σ_B^n directly depends on the required probability value n, which also affects the expected material utility of player A for $(\boldsymbol{e}, \sigma_B^n)$, this correlation is crucial in this case. It turns out that the greater player's A payoff gained from $(\boldsymbol{e}, \neg\,\boldsymbol{a})$, the smaller the probability required for the $\Pi_{A,B}$ indifference, but also the

smaller the resulting player B conformity index for σ_B^n. Thus a player B' small conformity index at the same time affects negatively (via a small probability) player A's material expected utility - since a small probability of (***e***, ¬ ***a***) will counterbalance its high payoff - and also makes the strategy ***e*** psychological utility increasingly lower than the strictly dominant psychological utility of strategy ¬**e**. The resulting equilibrium point of this case is still (¬**e**, ***a***).

Boundaries between the three cases are established by the distribution of material payoffs associated with any mixed strategy, and in particular how much surplus it assigns to player A. As long as a mixed strategy overwhelmingly advantages player B over player A, the T expected value of the mixed strategy pair (***e***, σ_B^n) cannot exceed that of player A's staying out. This is so not just because A is dissatisfied with his material outcome, but because of the insufficient conformity index of such mixed strategies. When a mixed strategy σ_B^n instead offers a substantial share of the material surplus to player A, it becomes the most conformist solution, and then provides psychological utility to both the players against a loss of material payoff to B. But at this point player B is able to compare the psychological utility of incomplete conformity against that of full conformity. It is evident that if the parameter λ is high enough to guarantee the existence of the psychological equilibrium in pure strategies, then it is also true that player B always prefers the pure strategy of full conformity.

Of course, much also depends on the λ exogenous parameter of the two players (granted they are symmetric, which is not necessarily true). Were λ too low, the situation would not change in regard to the basic TG and the repeated TG. But if λ is greater than player B's payoff difference between abusing and not abusing (given player's A entry), its motivational effectiveness necessarily becomes maximal for the strategy of full conformity. In general it biases the game towards excluding that mixed strategies can give rise to psychological equilibria. Actually, a look at the payoff space reveals a single North-East vertex where both payers have highest payoffs than anywhere on the eastern frontier where all the expected payoffs generated by mixed strategies lie. In short, given its overall payoffs, the pair (***e***, ¬***a***) strictly dominates any other strategy pair involving a mixed strategy σ_B^n and player A's entry strategy ***e.*** We have argued enough to state the following

PROPOSITION:

> Given a TG with pure and mixed strategies, whereby a psychological game with conformist preferences is defined, so that the motivational exogenous parameter λ is great enough to guarantee the existence of a psychological equilibrium in correspondence to (***e***, ¬***a***), the game's psychological equilibria are only the two in pure strategy (***e***, ¬***a***) and (¬**e**, ***a***), and no equilibrium

points in mixed strategies exist. In particular no player B mixed strategy is the best reply to player's A pure entry strategy ***e***, even if the entry strategy ***e*** is player A's best reply to the mixed strategy of player B.

From this proposition follows the

COROLLARY:

In the repeated psychological TG, psychological equilibria "refine" the equilibrium set of corresponding repeated TG in a discontinuous way in function of the increase of the motivational exogenous parameter λ, so that

- Given any λ such that in the one-shot psychological TG there is no psychological equilibrium in correspondence to the pair (***e***, $\neg$***a***), then the psychological equilibrium set is the same as the equilibrium set of the repeated TG due to the sole effect of material payoffs (see fig.8 north-east boundary X).
- If the value of λ is such that in the one-shot psychological TG player B's overall payoff derived from the strategy combination (***e***, $\neg$***a***) is no different from the overall payoff derived by B from the strategy combination (***e, a***) - so that a weak psychological equilibrium exists for (***e***, $\neg$***a***), then in the corresponding psychological repeated TG the psychological equilibria constituted by a mixed strategy σ_B^n and the pure strategy ***e*** have all the same player B expected payoffs, and thus they are all weak equilibria. Given the continuity of the probability mixture set over the two pure strategies $\neg$***a*** and ***a,*** the value of λ such that this is true is unique (see fig.8 north-east boundary Y).
- If λ is such that in the psychological one-shot TG in correspondence to the pair (***e***, $\neg$***a***) there is a strong psychological equilibrium, then in the repeated psychological TG there are no psychological equilibria in mixed strategies and the psychological equilibrium set dramatically shrinks to the only two pure strategy equilibrium points (***e***, $\neg$***a***) and ($\neg$**e**, ***a***). (See fig. 8 north-east boundary Z).

(Insert fig. 8 about here)

The corollary is important because it is in this context that we see our result. As far as the payoff space of a one-shot basic TG is concerned, also mixed strategies are not equilibria. If B adopts a mixed strategy that induces A to enter, B immediately has an incentive to deviate to the abuse strategy since the mixed strategy is not the best reply to A's choice to enter. On the contrary, if the payoff space is seen (as in the corollary) as the convex set of all the average payoffs for repeated strategies in a repeated TG, then represented within this space may be the average payoffs of player B's repeated strategies mixing the two pure strategies ***a*** and $\neg$***a*** according to some pre-established frequencies. Thus, if player B is able to accumulate a reputation of being a player that unfailingly plays one such strategy, he will have no reason to deviate if player A adopts a conditioned strategy of entrance like "as long as my observations are compatible with

the hypothesis that B is playing ***a*** and ¬***a*** according to the given pre-established frequency I will continue to enter by **e,** but were I to realize that my observations are incompatible with that frequency, I will switch to ¬**e** forever". In fact, given player A's conditioned entrance strategy, player B verifies that maintaining his reputation of being the type of player who uses the repeated strategy "abuse no more than x% of the time, and no abuse for the rest of the time" is profitable since it allows him to gain a certain portion of the surplus. Summing up, player B has the incentive to keep abuses at a certain frequency in order to support his reputation of being the relevant type.

Quite different however is the situation when the repeated psychological TG is considered. In this case, a payoff space identical to the convex hull of all the payoff pairs deriving from pure strategy combinations in the one-shot psychological TG is generated by taking the set of all the *average* payoffs pairs given by combinations of the two players' (pure and mixed) repeated strategies. What happens is that if player B has chosen a repeated mixed strategy whereby he has been able to accumulate a positive reputation which induces player A to enter for the first time, then he immediately recognizes the incentive to switch to a strategy that employs the strategy ¬***a*** with higher frequency. This feature of the repeated psychological TG completely changes the best response structure with regard to the standard repeated TG. In the standard case, player B has a clear incentive to maintain his strategy once he has been able to build up a reputation for being a mixed *type*, since abusing less would give away a larger part of the surplus to player A, while abusing more would induce player A to carry out his sanction. At the same time, player A has a strong incentive to monitor and sanction the relevant possible deviation by player B. In the repeated psychological TG, by contrast, player B's best reply to player's A entry is to deviate from any mixed strategy σ_B^n to ¬***a***. But if player B deviates to a strategy more concessive to him, A certainly does not have any reason to punish him. Thus the repeated mixed strategy equilibrium of the basic repeated TG is destabilized. Summing up, any mixed strategy by player B that induces player A to enter, according to player B's point of view is dominated by the pure strategy "always ¬***a***", so that a rational player B would never strive after a reputation such as being committed to the mixed strategy σ_B^n. From the outset he would prefer to develop the dominant reputation of being an "always ¬***a"*** player***.***

From this the conclusion follows that even though generating a psychological game from a basic Trust Game enables determination of new equilibrium points (i.e. to pass from only one equilibrium to at least two), when the change involves a step from the one-shot TG to the

repeated TG, then transforming the payoff space by means of conformist preferences has a powerful effect in reducing the psychological equilibria to a subset of the Nash equilibria. It remains, however, that the equilibria are two. Which of the two is to be selected?

11. The cognitive/predictive function of norms and ex post equilibrium selection

It is a somewhat disturbing truth in the foundation of game theory that the existence of "one sole" Nash equilibrium point, even if it is in dominant strategies, does not assure sufficient conditions for deducing the rational solution of the game (cf. Bacharach 1987). In order to predict that rational players will carry out their equilibrium strategies, something more is needed: the system of reciprocally consistent expectations that justify the prediction that players will adopt exactly *that* combination of equilibrium strategies. A player rationally chooses an equilibrium strategy only when he has formed the backing expectation that the other players will also play the equilibrium strategies components of the same equilibrium point, so that his choice is rationally justified as his best response to them. Moreover, this backing expectation must be consistent with the assumption that also the other players act with similar backing expectations. Hence, in order to be considered as a *solution* that each player will *rationally* play, an equilibrium point even if unique needs *previously* to be predicted as the set of strategies that every player will play, i.e. it must be *previously known* by each player as the description of strategies that all the other players will effectively carry out, given that they all expect exactly these strategies from one another (this amounts to the somewhat circular statement that a Nash equilibrium is a solution as far as the solution – i.e. the equilibrium point to be the solution – is common knowledge).

Where can this *previous knowledge* come from? The simple existence of an equilibrium does not entail that it will be played since, again, in order to infer that it will actually be put into practice a player needs some reason to believe that other players besides himself have already formed the expectation that everybody will play it. In other words, a process of expectation formation converging on this mutually consistent system of beliefs and prediction must be worked out even in the apparently simple case that "one sole" equilibrium point exists. Indubitably, therefore, a more pressing problem of expectations formation exists if the possible equilibrium points are many. Without answering the question as to which of them is mutually expected by players to be the actual solution of the game, there is no way to say that players have any incentive to play a particular strategy combination, even if it is an equilibrium point of the game.

To return to our context, recall that the foregoing section concluded that *at most two* Nash psychological equilibria remain as solution candidates once the game has been transformed into a psychological game through the ex ante agreement on a CSR norm and the introduction of conformist motivations. *Two*, however, are enough to create significant uncertainty about the actual solution. Though one of these equilibria properly corresponds to the *ex ante* agreement on a fairness principle (the Nash Bargaining Solution is maximized by the outcome (4,4), this is not enough to say that it is the predicted solution of the ex post game.

In order to solve the problem, the ex ante "should-be" agreed solution should also be known as the ex post *de facto* implemented set of strategy choices. Any player knows that a strategy combination is implemented only if this knowledge is consistent with the prediction that any other player also believes that everybody will in fact play that equilibrium. Could the fact that we have ex ante decided a principle corresponding to an equilibrium be enough to create this general expectation? It could, but it is important to realize that there is no necessity in this inference. What we decide to do in order to be impartial in the ex ante perspective is not necessarily what we will actually do in the ex post perspective. Moreover, it is not necessarily what other players will do in the ex post situation. This inference would be unwarranted from a logical point of view. Recall in fact that also the motivational force of conformist preference - driving players to conform with an ex ante agreed principle – operates conditionally on the previous expectations that also the counterparty will reciprocate compliance. Hence the existence of a previous system of mutual expectations must also be granted in the context of psychological equilibria.

Here one appreciates the role that norms play in a cognitive process of belief formation converging on the mutual prediction across players that a given psychological equilibrium will *be de facto* executed. This role consists in a two-tier answer. At a first stage it is suggested that if each player has actually adopted an unanimous impartial agreement in the ex ante perspective, then he will get to hold at least one *mental model* of a decision maker (at least *himself*) who plans at a moment in time to act in accordance with the terms of the agreed course of action.[28] Notwithstanding the genuineness of the intention, agreeing on a set of actions to be carried out later in fact implies making a plan on some ensuing action which is simply the behavioral content of the statement of agreement. In order to stipulate that "we will act in a certain way later on" – which may be seen as the content of a generic agreement – each player at least must have in mind the mental model of an agent *who will act in that certain way later on*, where the "way" is the one *signed* in the agreement. What could otherwise be meant by finding a strategy combination that is

[28] On mental models see Johnson-Laird (1983), Johnson-Laird and Byrne (1991), Dezau and North (1994).

an equilibrium point invariant under the players' position replacement, but having in mind a model of an agent who, without going against his incentives, behaves ex post exactly in the *same* way whatever his position in the game?

This is not a reason to say that if this mental model is admitted then it follows that the player will actually carry out the correspondent action, nor is it a reason to say that if the existence of such a mental model is true for other players, then they will in fact carry out the corresponding actions. This is a matter of *approximate* and *default* reasoning, not one of pure logic or necessity (Reiter 1980, Bacharach 1994, Sacconi and Moretti 2008). The model is derived from introspection, because the player himself is a rational agent who has been able to plan action in accordance with the behavioral content of the statement of agreement. The paradigmatic case whereby the model is derived by generalization is that of the agent himself. Let us therefore simply state that a player holds in his mind the mental model of a rational agent (himself) who acts according to the behavioral content of the statement which is the term of agreement.

Assume, moreover, that mental models are necessarily used in order to figure out possible situations and predict them (i.e. no future behavior can be outguessed without a mental model of an agent performing the corresponding behavior). And hypothesize that at a point in time no further mental model of a rational agent comes to the mind of our players but that of an agent who *will act in a certain way later on*. If no contrary evidence is thus far forthcoming about the actual behavior of other players, the only way that an agent can simulate the other players' choice is to resort by default to his own mental model of a rational agent. By default, then, the same mental model is used to simulate every players' reasoning and behavior. This simulation can be recursive, so that a player uses his mental model not only to predict another player's behavior but also in order to simulate the other player's reasoning and beliefs, so that a *shared mental model* of all the rational agents results wherein they all conform with the terms of agreement.

This explains, if not justifies, why the agent may categorize or recognize this situation (until proof of the contrary) as an element of the class wherein agents conform to the norm. It produces, as a matter of description of how players *de facto* reason not as a matter of deduction from whatsoever absolute logical principle, the state of reciprocal beliefs that justifies the decision of any player to carry out the strategies consistent with the psychological equilibrium of full conformity to the principle T. In the Trust Game the pair $(e, \neg a)$.

Of course, it is also possible that a player may have a mental model of an agent who does not comply with an agreement, and until proof to the contrary, this model can be also assigned by

default to other players in order to simulate their choice. If generalized, such a mental model would generate a state of mutual beliefs such that the predicted equilibrium point is the one where no player respects the norm, and hence the firm abuses and the stakeholder plays no-entry. Note that this equilibrium is also compatible with conformist preferences, for when a player predicts that the other will abuse, his psychological best response most compliant with the principle is no-entry. This is also the prediction that would induce the other player to abuse also on the basis of his conformist preferences.

To be consistent with the idea of *default reasoning* we may proceed as follow. If a player has agreed on a fairness principle it *normally* has a mental model of an agent who carry out the corresponding commitment, for this is the behavioral content of the principle he has agreed to. Moreover nothing in his base of knowledge (until proof or evidence to the contrary) contradicts that an agent who subscribed to an agreement on the principle will carry out the corresponding commitment (assume this is provisionally true). At the same time it may be the case that it comes to the player's mind that an agent may also not comply with the agreed principle and (assume that) nothing in the player's base of knowledge contradicts that proposition. Thus to the player's mind come *two* mental models, that are both *contingently* true according to two different incomparable mental *framings* of the situation.[29] Each considered by itself these mental models allow a default inference in the format "it is not inconsistent with the base of knowledge that…..". But taken together they are inconsistent. Thus, we cannot conclude by default reasoning (i.e. by a conclusions in term of what is "normally true") given our base of knowledge and given our two contrasting defaults (i.e. rules of implication) that an agent will "normally" conform or not with the agreed principle. There is some uncertainty about whether the state wherein we are either belongs to the situations set sketched by a one *frame* or by another. This admits that players express through a subjective probability distribution their beliefs about the two possible equilibrium points corresponding to the generalization of the two mental models. Now consider that the players are again just two. Since the *same* mental model may come to the mind of both the players with exactly the same *vividness,* they share the same uncertainty about the same shared mental models (what does not imply that the prior probability must be uniform - this will depend on the degree of vividness of each shared model.)

[29] The idea that different mental models, according to different *framing* of the situation, may "come to the player's' mind" is taken from Bacharach (2006), even if I do not discuss here the interpretation that the model within which the agent is seen as compliant with the agreement can be interpreted as a consequence of what Bacharach called "we thinking".

But a probability distribution over two pure strategy equilibria does not guarantee a consistent prediction of an equilibrium solution, and it allows for inconsistent best replies chosen by the players. The second step in the *cognitive-predictive function* then consists in assuming that the common prior generated by the two plausible mental models is taken as the starting point for a revision dynamics of expectations, such that for a reasonable range of prior beliefs the equilibrium point of full conformity is selected as the outcome of the revision dynamics. This step therefore actually reduces to the plausible hypothesis (experimentally testable) that agreeing impartially on a fair principle will give the shared mental model of a rational agent (who conform to the principle) sufficient vividness to say that both players will in fact start their belief revision dynamics at a prior wherefrom they will necessarily converge to a point where they will completely believe that the solution of the game will consist in the psychological equilibrium of full compliance. The remaining work is left to the proper operation of an appropriate dynamic of belief revision.

To this end I adopt the *tracing procedure* (Harsanyi 1975, Harsanyi and Selten 1988) which is an eductive equilibrium selection dynamics whereby the prior probabilities distributed over a pair of feasible equilibrium strategies for each of two players are continuously modified as a result of a repeated mental simulation of both players' best reply calculations given the current state of each player's beliefs. Each simulation that identifies a player' best reply to the current state of his beliefs augments according to the probability of that player's strategy with respect to its prior probability.

Along this mental process of simulation, players never actually carry out a decision until uncertainty vanishes.[30] They simply repeatedly calculate their best reply given a revised prior, and these priors are revised on the basis of the best replies just calculated at the previous stage of the process. At any step the simulated best eply of the second player nurtures the change in the first player's beliefs by assigning additional probabilities to the simulated choice, thus affecting the recalculation of the first player's best reply, and hence inducing also a further change in the second player's expectation. Only at the end of the process, when the players have both reached mutually compatible predictions concentrated on a particular equilibrium point, do they actually carry out their strategy choices.

To gain an idea of the *tracing procedure* consider a thought process that takes place in a sort of "reasoning time" which by construction starts from a stage of complete uncertainty $t^\circ = 0$ and

[30] On evolutionary equilibrium selection mechanisms with learning through repeated plays see Young (1998). The distinction of 'eductive' VS. 'evolutionary' equilibrium selection dynamics is given by Binmore (1987).

continues until a stage of perfect predictability $t^1 = 1$ is reached. Time is a continuous parameter t that varies from 0 to 1, so that for example its realization t^n is identical to the number $0 \leq n \leq 1$. Assume that at time $t^\circ = 0$ players A and B think that just two equilibrium points are possible. Given a prior p° that assigns probabilities over the two possible pure equilibrium strategies (indexed 1 or 2) of the two players A and B, each of them separately maximizes his expected payoff by choosing a pure strategy σ_{ij} (for i =A,B, j = 1,2). At any further time t^n the prior probability of each equilibrium strategy σ_{ij} for each player is revised in consideration of whether at the previous point in time t^m (where, granted $m<n$, m is taken as near as possible to n) that strategy is calculated to be the best response of a player to his current expectations or otherwise. Given for each equilibrium strategy σ_{ij} the prior probability $p^\circ(\sigma_{ij}) = p^\circ_{ij}$ revisions are generated by the following simple algorithm

- $1\text{-}t^n(p^\circ_{ij}) + t^n$ is the probability at time t^n of the player's i equilibrium strategy σ_{ij} whenever at time t^m σ_i is calculated to be player's i best reply;
- $1\text{-}t^n(p^\circ_{i,k\neq j})$ is the probability at time t^n of any other equilibrium strategy $\sigma_{i,k\neq j}$ that at t^m is not calculated to be the player's best reply.

As time passes, the tracing procedure entails that the prior p° loses more and more of its initial weight, whereas the probability derived from a strategy being recursively predicted to be the player's best response tends to 1.

The *tracing procedure* is a dynamic that simulates the formation process of mutually consistent expectations. Thus it also seems appropriate for the study of psychological equilibrium selection, such as (***e, ¬a***) or (***¬e, a***) in the psychological TG - which are well defined only for states of knowledge wherein expectations of first and second level are consistent with the prediction of a particular equilibrium. Until these systems of mutually consistent expectations have been formed, a player cannot act on the basis of his conformist preferences and therefore remains naturally involved in an outguessing process. A player thinks that two equilibria − (**e, ¬*a***) or (***¬e , a***) – are possible, and hence the two mutually consistent expectations systems supporting each of them are thought to be possible as well. Then the player is uncertain about which of the two expectations system is actually the case. Indeed, a common prior (and any revision of it) represents not only a player's uncertainty about the adversary's two equilibrium choices but also his prediction on the other player's uncertainty about his own equilibrium choices (thus, for each player, at least expectations of *first* and *second* order about the other player's choices and beliefs are derived from a common prior and its revisions).

(Insert fig. 9 about here)

Consider the phase diagram of fig.9. It is a *probability square* representing all the possible discrete probability distributions over player A' and player B's pairs of the equilibrium strategies (***e, ¬ e***) and (***a, ¬ a***). From each player's point of view the square can be seen as representing his own uncertainty (prior and revised according to the procedure) about the other player's two possible equilibrium strategies and his prediction about the other player's uncertainty (prior and revised) concerning his own two possible choices - both derived from a common prior and its revisions. The probability of player A's strategy ***e*** varies from 0 to 1 (the reverse for strategy ***¬e***) moving upward along the vertical sides of the square. On the other hand the probability of player B's strategy ***a*** varies from 0 to 1 (the reverse for strategy ***¬a***) moving rightward along the horizontal sides of the square. Thus each point within the square represents a pair of probabilities assigned to players' A and B strategies ***e*** and ***a*** respectively (and taking the complement also the probabilities assigned to the alternative strategies of both the players ***¬e***, ***¬a***). Corners represent pure strategies pairs when they are perfectly predicted (with probability 1) as indicated thereafter :

top-left: (***e, ¬ a***);
top-right: (***e , a***) ,
down-left: (***¬e, ¬a***);
down-right: (***¬e, a***)

Starting from each inner point within the square, the tracing procedure plots a single and uniquely determined path whence forward beliefs change until a corner is reached (this happens by construction because at time t^1 uncertainty necessarily vanishes and each player comes out with a "probability one" prediction of the pure strategies they are both playing).

Each equilibrium has in effect a basin of attraction defined by all the starting points wherefrom a path begins and evolves through the tracing procedure until it reaches the corner corresponding to a given equilibrium. Equilibrium basins of attraction are indexed in the phase diagram of fig. 9 by X for corner (***e, ¬ a***) and Y for corner (***¬e, a***). From every inner point within one of these basins the dynamics tend to converge through continuous belief revisions to the relevant

attractor (pure strategy equilibrium). Along these paths players always select as best replies against their current expectations a pair of strategies that jointly compound an equilibrium pair, so that their choices never approximate a result where their incentives become incompatible. Players will continue to play the relevant pair of equilibrium strategies until they reach a point where they predict with "probability one" that each of them will play exactly the equilibrium which is the attractor of the basin wherein the path has started.

As far as paths starting from outside any basin of attraction are concerned, the procedure tends to induce a "change of mind" in the players. Note that in the phase diagram of fig.9 all paths traced by the procedure tend to move away from a non equilibrium corner toward another non equilibrium corner. Actually, from the region Z paths generated by the tracing procedure move toward North-Right, i.e. toward the non equilibrium outcome (***e, a***), while from the region Q paths move toward South-Left i.e. the non equilibrium outcome ***(¬e, a***). Along these paths players make choices that accord with their current expectations, but also increase step by step the probability of reaching a non equilibrium outcome that progressively reduces both players' calculated expected utilities for the ongoing best replies. Player A, for example, along a path starting from a point in Z, is afraid to reach the corner (***e, a***) where he gets only 0 in terms of overall payoff. Hence he is under increasing pressure to change his choice to ***¬e***. At the same time player B see probabilities of drawing closer to corner (***e, a***), where he gets only the overall payoff 5 instead of 6, which he would get in (***e,*** ¬ ***a***). Hence he is under pressure to change his choice to ¬ ***a*** . The effect of the increasing probability of the disequilibrium outcome, however, eventually induces one player to change his choice before the other. This happens when they reach a switching point where the path intersects the boundary of an equilibrium basin of attraction. At that point, paths switch from the current trajectories and turn toward the relevant equilibrium corner which is the attractor within the intercepted basin of attraction.

It is noticeable that the tracing procedure admits a large range of situations wherefrom the dynamics selects the equilibrium (***e,*** ¬***a***). Specifically not only all the paths starting from inner points within the basin of attraction X , but also all the paths starting at points in the region Z above the boldface broken diagonal depicted in fig. 8 will reach (***e,*** ¬***a***). These paths in fact will eventually reach a switching point at the boundary of the basin of attraction X, where the tracing procedure makes sure that player B for the first time, and before player A's incentive to change his choice becomes too intense, changes his choice and starts playing the alternative equilibrium strategy ¬ ***a***. Moreover, all the paths starting from region Q above the boldface diagonal will

switch toward the corner (***e,*** ¬***a***) when they cross the boundary of the basic of attraction X. In that case player A, who until that moment would have been choosing the strategy ***a*** as his best reply within the dynamics process, changes his best reply as he is under pressure of the risk to reach the non equilibrium outcome (¬***e,*** ¬ ***a***). Over the boldface diagonal this happens necessarily before than an analogous incentive pushes player A to switch from strategy ¬ ***a*** to strategy ***a.***

However it is also true that the largest part of the probability square gives rise to paths, those staring at point beneath the boldface broken diagonal, converging to the equilibrium corner (**¬*e,*** ***a***). This means that the tracing procedure does not allows by itself an unique prediction that the equilibrium fully conformist with CSR norm will be selected.

We must here resort again to the first step in our two-tier answer. The ex ante agreement on a principle of fairness by default allows the formation of a prior belief favorable to the propositional content of the mental model representing an agent discharging the commitments of his agreement. Just after the agreement there is no evidence that any player will not conform, whereas there is the intuitive evidence of the mental representation of an agent who agrees to a principle and hence expresses at least at that point in time the commitment to carry out a certain behavior later on.

Despite it would be excessive to say that this completely resolves the players prior uncertainty, it justifies the assumption that, after an agreement on the CSR norm amongst the firm management and stakeholders has been worked out – as far as it is understood as a constitutional, fair, initial (*ab origine*) agreement under the 'veil of ignorance' - the model of a compliant agent 'comes to their mind' with most *vividness*. This implies that by an impartial, voluntarily devised behind an hypothetical 'veil of ignorance' agreement over a principle of fairness, players can escape from a real life context of mutual distrust, in case they are already living within it. To say it differently, the thought experiment of putting players under a 'veil of ignorance' allows them to abstract from a concrete context of distrust and to frame the situation as one of 'constitutional choice' whereby they *from the beginning* choose the rule for entering a new interaction. This permits them to make default inferences abstracting from their previous experience within non constitutional situations and to reason solely on what is appropriate in such a perspective.

If this hypothesis is tenable, the starting point of the tracing procedure will be set at a place above the boldface broken diagonal of our phase diagram, and then the tracing procedure will carry it to converge to the fully conformist psychological equilibrium.[31]

This concludes our discussion of the four roles of voluntary but explicit CSR norms. They allow the description of strategies and equilibrium points, even if the equilibria are multiple, in a game played under unforeseen contingencies. Secondly, a CSR norm permits the ex ante selection of the equilibrium point that meets the requirements of an impartial choice. An explicit agreement on a contractarian norm is moreover a way to introduce psychological conformist equilibria, and quite surprisingly to derive the significant result that mixed strategy equilibria are absent in a psychological repeated Trust Game. Lastly, a cognitive and predictive role is played by an equilibrium selection mechanism that, from a state of predictive uncertainty about possible equilibrium points, generates a state of mutually consistent expectations (equilibrium expectations). An extensive range of prior probabilities, which are largely plausible and consistent with the assumption that players have agreed on an ex ante norm affecting their *de facto* mutual expectations, are consistent with the prediction that players will converge on believing that the solution of the psychological game is the (*entry, no abuse)* equilibrium, so that they will actually play their strategies component in this equilibrium. The theory of the endogenous implementation of the normative model of multistakeholder fiduciary duties is thus complete.

[31] In favor of this hypothesis there is some reliable laboratory's evidence gathered by experimental studies about the formation of conformist preference (Sacconi and Faillo 2005, Faillo and Sacconi 2007, Sacconi and Faillo 2008, Faillo , Ottone and Sacconi , 2008). Experimental subjects in an apparently cheap-talk, pre-play collective choice situation are given the opportunity to agree impartially (i.e. under a "veil of ignorance") on a principle of fair division they will be in the position to implement ex post in a non cooperative game they will successively play - wherein they do not have any material incentive to comply with the principle. It comes out however that most of the experimental subjects conform with the principle and, what is most compelling, they conform against their material interests just because they believes other participants in the agreement (even if it is completely anonymous) will also conform and believe others will conform. Only difference between the players who decide before making the experience of a fair, impartial, anonymous agreement and those who decide in the game after having participated in the pre play fair agreement, is the agreement itself. Hence, we conclude that the decisional experience of a fair, impartial anonymous agreement under the veil of ignorance is by itself able to generate the frame of mind such that the mental model solely comes to their mind , or it comes with the maximal relative vividness, such that an agent act consistently with the behavioral content of the agreement, so that they rationally reply by the equilibrium strategy of full conformity to the principle.

REFERENCES

AA.VV. (1993) "The Corporate Stakeholder Conference", *University of Toronto Law Journal*, XLII, n.3, pp. 297-793

AOKI M. (1984), *The Cooperative Game Theory of the Firm*, Cambridge (Cambridge UP).

AOKI M. (2001), *Toward a Comparative Institutional Analysis,* Cambridge, Mass.: MIT Press.

AOKI M. (2007), "Three-Level Approach to the Rules of the Societal Game: Generic, Substantive and Operational" paper presented at the conference on "Changing Institutions (in developed countries): Economics, Politics and Welfare" Paris on May 24-25, 2007

AOKI, M. (2007), "Endogenizing Institutions and Institutional Change," *Journal of Institutional Economics* **3**:1-39

BACHARACH M. (2006) *Beyond individual choice*, eds. by Gold N,. and R.Sugden, Princteton UP, Princeton

BACHARACH (1987), A Theory of Rational Decisions in Games, *Erkenntnis* 27, pp.17-55

BACHARACH M (1994), "The Epistemic Structure of a Game", *Theory and Decisions* 37,7-48.

BINMORE K. AND A. BRANDEBURGER (1990), "Common Knowledge and Game Theory" in K.Binmore ed. *Essays in the Foundation of Game Theory,* Blackwell, Oxford,

BINMORE (1991), "Game theory and the social contract" in Selten R. editor, *Game Equilibrium Models II, Methods, Morals, Markets*, Springer Verlag, Berlin

BINMORE K. (1987) "Modeling rational players", *Economics and Philosophy*, Part I, n.3, pp.9-55, Part II, n.4, pp.179-214

BINMORE K. (1998), *Just playing*, MIT press, Cambridge MA,

BINMORE K. (2005), *Natural Justice*, Oxford: Oxford University Press.

BLAIR. M AND L. STOUT (1999) **"**A Team Production Theory of Corporate Law", *Virginia Law Review,* Vol. 85, No. 2

BLAIR. M AND L. STOUT (2006) "Specific Investment: Explaining Anomalies in Corporate Law", Journal of Corporation Law, n. 31**,** pp. 719-44

BROCK H.W. (1979), "A Game Theoretical Account of Social Justice", *Theory and Decision*, 11, pp. 239-265.

CHARENESS S. G. AND M. DUFENBERG, (2006) "Promises and Partnership", *Econometrica*, Vol. 74, No. 6, November, 1579–1601

CLARKSON M. B.E (1999), *Principles of Stakeholder Management*, Clarkson Center for Business, ethics, Toronto

COLEMAN J. (1992), *Risks and Wrongs*, Cambridge (Cambridge University Press).

DEZAU AND NORTH (1994) , "Shared mental models: Ideologies and institutions", KIKLOS, 47, pp.1-31

DONALDSON T. AND L.E. PRESTON (1995) "Stakeholder theory and the Corporation: concepts evidence and implication", *Academy of Management Review*, 20, 1, pp.65-91

DUNFEE T. AND T. DONALDSON (1995) "Contractarian Business Ethics", *Business Ethics Quarterly*, 5, pp.167-172

FAGIN R., J.Y.HALPERN, Y.MOSES AND M.Y.VARDI (1996), *Reasoning about Knowledge*, Cambridge, Mass.:MIT Press.

FAILLO M. AND L. SACCONI (2007), "Norm Compliance: The contribution of Behavioral Economics models" In Innocenti A. and Sbriglia P.(eds). *Games, Rationality and Behavior*. LONDON: Palgrave.

FAILLO M., S. OTTONE AND L. SACCONI (2008) *Compliance by Believing: An Experimental Exploration on Social Norms and Impartial Agreements*, University of Trento - Department of Economics Working paper; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1151245

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

for Non-cooperative Games", *International Journal of Game Theory* 5 (1975) 61-94.

FREEMAN R.E AND S. RAMAKISHNA VELAMURI, (2006) "A New approach to CSR Company Stakeholder Resposnability*"*, in Kakabadse and Morsine (eds.) *Corporate social responsibility reconciling aspiration and application*, Palgrave Mcmillan, London

FREEMAN R.E. (1984), *Strategic Management: A Stakeholder Approach*, Pitman, Boston.

FREEMAN R.E. AND P. EVANS ,1989, "Stakeholder Management and the Modern Corporation: Kantian Capitalism", in Beuchamp and Bowie (eds.) *Ethical Theory and Business*, 3rd ed., Prentice Hall, Englewood Cliffs,N.J.

FREEMAN R.E. AND J. MCVEA (2002) *A stakeholder approach to Strategic management,* working paper n.01-02 , Darden Graduate School of Business Administration

FREY B. (1997), *Not Just for the Money*, Edward Elgar.

FUDENBERG D. (1991), "Explaining cooperation and commitement in repeated games", in J.J.Laffont (ed.) *Advances in Economic Theory, 6th World Congress, Cambridge* U.P., Cambridge.

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

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

GEANAKOPLOS, J., D. PEARCE AND E. STACCHETTI (1989), 'Psychological games and sequential

GREEN L. (1990) *The Authority of the State*, Oxford, (Clarendon Press).

GRIMALDA G. AND L. SACCONI (2005), 'The constitution of the not-for-profit organisation: reciprocal conformity to morality', *Constitutional Political Economy*, **16** (3), September, 249–76

GROSSMAN S. AND O.HART (1986): "The Costs and Benefit of Ownership: A Theory of Vertical and Lateral Integration", *Journal of Political Economy*, 94, pp. 691-719.

HANSMANN H. (1996) *The Ownership of the Enterprise*, Harvard University Press, Cambridge Mass.

HARE R. M. (1981): *Moral Thinking*, Oxford (Clarendon Press).

HARSANYI J.C AND R.SELTEN , (1988) A General Theory of Equilibrium Selection", (MIT Press, 1988)

HARSANYI J.C. (1975) The Tracing Procedure. A Bayesian Approach to Defining a Solution

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

HART O. (1995) *Firms, Contracts and Financial Structure*, Oxford (Clarendon press).

HART O. AND J.MOORE (1990): "Property Rights and the Nature of the Firm", *Journal of Political Economy*, 98, pp.1119-1158.

HOBBES T. (1651), *Leviatano*, (trad. it.), Firenze (La nuova Italia) , 1975.

JENSEN M.C. **(**2001) "Value Maximization, Stakeholder Theory, and the Corporate Objective Function**"** *Journal of Applied Corporate Finance,*Vol. 14, No 3, Fall.

JOHNSON LAIRD P.N. AND BYRNE (1991), *Deduction*, Lawrence Erlbaum Associated Publ., Hove and London

JOHNSON LAIRD P.N. (1986), *Mental Models Towards a cognitive science of language, inference and Consciousness*, Cambridge UP, Cambridge,

KALAI AND SMORDINSKI (1975), "Other Solution to Nash's Bargaining Problem", *Econometrica*, vol.43, 3, pp.880-895.

KAPLOW L. AND S. SHAVELL (2002), *Fairness and Welfare*, Harvard University Press, Cambrdige Mass.

KREPS D. (1990): "Corporate Culture and Economic Theory" J.Alt and K.Shepsle (eds.*), Perspectives on Positive Political Economy*, Cambridge (Cambridge University Press).

KREPS D.(1990), *Games and Economic Modelling*, (Oxford U.P., Oxford,).

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

MCMAHON C. (1989) "Managerial Authority", *Ethics*, October , 1989.

MEESE A.L. (2002), *The Team Production Theory of Corporate Law: A critical Assessment*, mimeo

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

POSNER E. A. (2000) Law and Social Norms, Cambridge Mass (Harvard UP)

RABIN M. (1993), 'Incorporating fairness into game theory', *American Economic Review*,

RAJAN R. AND L.ZINGALES (2000)"The Governance of the New Enterprise", in Xavier Vives (ed.) *Corporate Governance, Theoretical and Empirical Perspective*, Cambridge (Cambridge UP)

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

RAWLS J. (1993), *Political Liberalism*, New Your (Columbia UP).

RAZ J. (1985), "Authority and Justification", *Philosophy and Public Affairs*, 1985, pp.3-29

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

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

SACCONI L. (1995) Equilibrio e giustizia, parte II: la selezione del contratto sociale", *Il Giorale dell'economista,* **83** (5), 1281–302.

SACCONI L. (2000): *The Social Contract of the Firm. Economics, Ethics and Organisation,* ,Springer Verlag, Berlin.

SACCONI L (2004) "CSR as a model of extended corporate governance, an explanation based on the economic theory of social contract, reputation and reciprocal conformism", LIUC paper n.142, LIUC University Cattaneo of Castellanza, (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=514522)

SACCONI L. (2006) "A Social Contract Account For CSR as Extended Model of Corporate Governance (Part I): Rational Bargaining and Justification" *Journal of Business Ethics*, ,

SACCONI L. (2007) "A Social Contract Account for CSR as Extended Model of Corporate Governance (Part II): Compliance, Reputation and Reciprocity" *Journal of Business Ethics,*. N. 75 pp.77-96

SACCONI L. (2007), "Incomplete Contracts and Corporate Ethics: A Game Theoretical Model under Fuzzy Information", in F.Cafaggi, A. Nicita and U. Pagano (eds.), *Legal Orderings and economic institutions*, London (Routledge)

SACCONI L., DECOLLE AND E.BALDIN (2003), "The Q-RES Project: The Quality of Social and Ethical Responsibility of Corporations" (with S.), in Wieland, Josef (ed.), *Standards and Audits for Ethics Management Systems , The European Perspective* „Springer Verlag, Berlin, pp.60-117

SACCONI L. AND M. FAILLO (2005), "Conformity and Reciprocity in the 'Exclusion Game': An Experimental Investigation", Discussion paper del Dipartimento di economia dell'Università di Trento, N. 12, 2005, pp.36; (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=755745)

SACCONI L. AND G. GRIMALDA (2007). "Ideals, conformism and reciprocity: A model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case" in: P.L. Porta and L.Bruni eds. . *Handbook of Happiness in Economics*, Edward Elgar. London

SACCONI L. AND S.MORETTI (2008). "A Fuzzy Logic and Default Reasoning Model of Social Norms and Eqiulibrium Selection in Games under Unforeseen Contingencies". *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, Vol. 16, No. 1

SACCONI L. AND M. FAILLO (2008), *Conformity, Reciprocity and the Sense of Justice, How Social Contract-based Preferences and Beliefs Explain Norm Compliance: the Experimental Evidence,* University of Trento - Department of Economics Working paper

SEGAL U. AND J. SOBEL, (2007) , "Tit for tat: Foundations of preferences for reciprocity in strategic settings" *Journal of Economic Theory* 136 197 – 216

SEN A. (1993): *Moral Codes and Economic Success*, London (LSE, ST-ICERD discussion papers n.49).

STERNBERG E. (1999) , The stakeholder Concept: a Mistake Doctrine, Foundation for Business Responsibility , Issue Paper No. 4, November 1999. (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=263144)

STOUT L. (2006), Social Norms and Other-Regarding Preferences**,** *NORMS AND THE LAW, John N. Drobak, ed., Cambridge University Press, 2006*

TIROLE J. (2001), "Coporate Governance", *Econometrica*, vol 69, n.1, pp.1-35

VAN DAMME E. (1987) "Stability and Perfection of Nash Equilibria", Springer Verlag, Berlin

WATT E.D. (1982) *Authority*, London (Croom Helm)

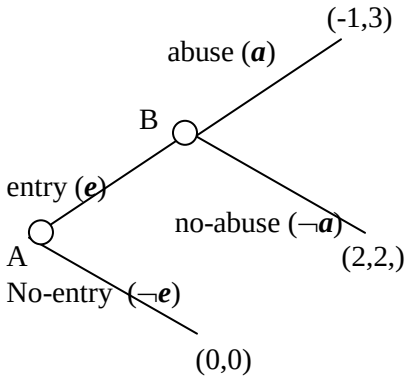
WIELAND J. (2003) (ed.) *Standards and Audits for Ethics Management Systems, The Europen Perspective*, Berlin (Springer),

WILLAMSON O.(1975), *Market and Hierarchies*, The Free Press

WILLAMSON O.(1986), *The Economic Institutions of Capitalism*, New York (The Free Press)

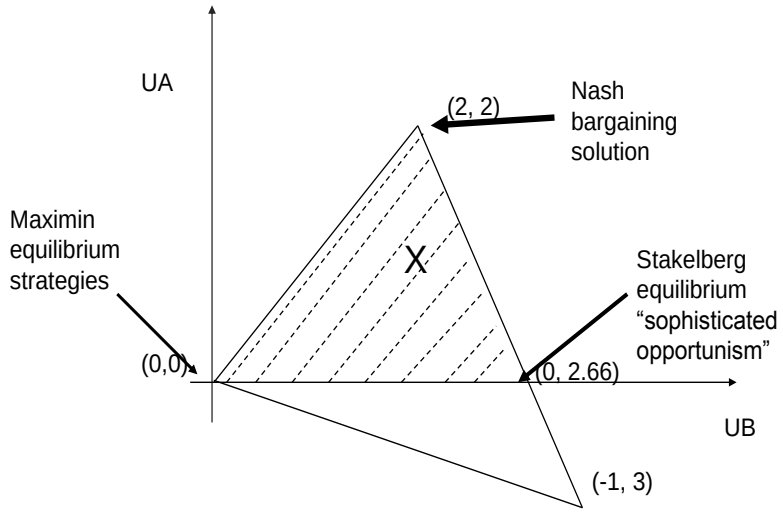
YOUNG, H.P. (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions,* Princeton, NJ: Princeton University Press

ZIMMERMAN H.J. (1991): *Fuzzy Set Theory and Its Applications,* 2nd revised ed., Dordrecht-Boston, (Kluwer Academic Press).



(fig.2, one shot Trust Game in extensive form)

(Fig. 3, Equilibrium set X of the repeated TG)



9

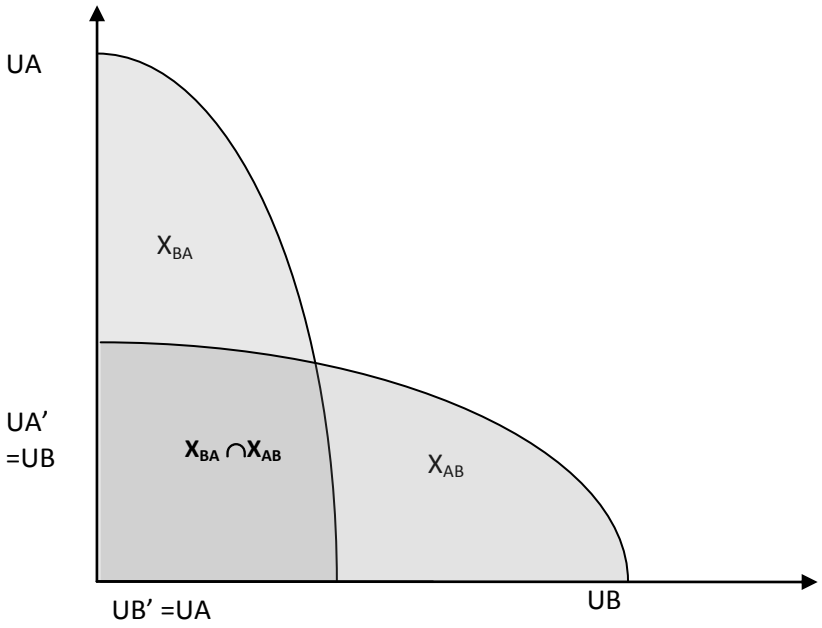


Fig.4 Symmetric translation of the payoff space X_{AB} with respect to the utility axes, so that the utility function U_A is replaced by $U_A' = U_B$ and vice versa

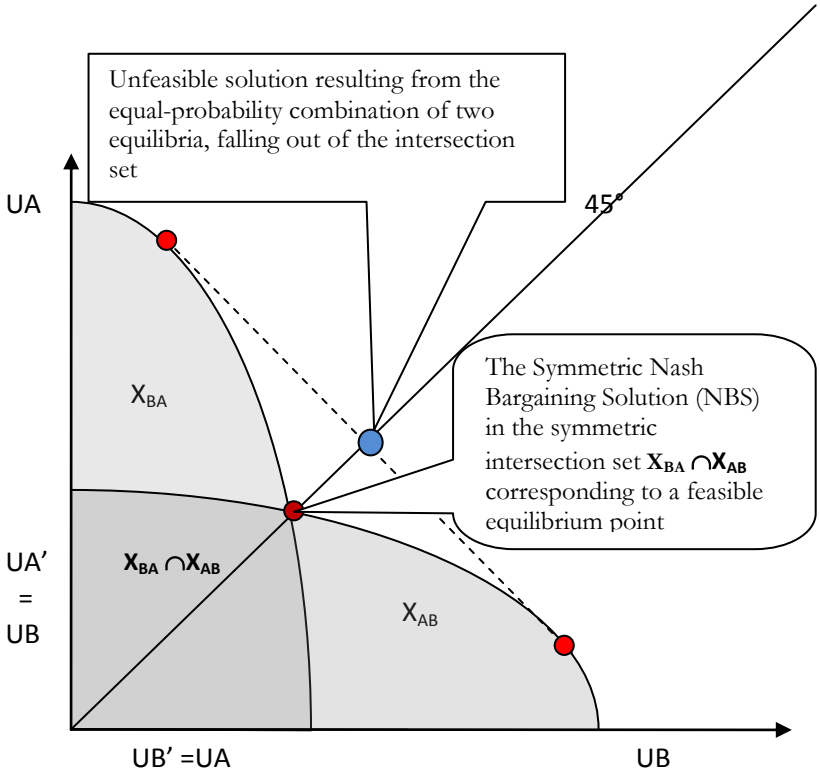


Fig. 5 Egalitarian feasible solution and efficient unfeasible solution

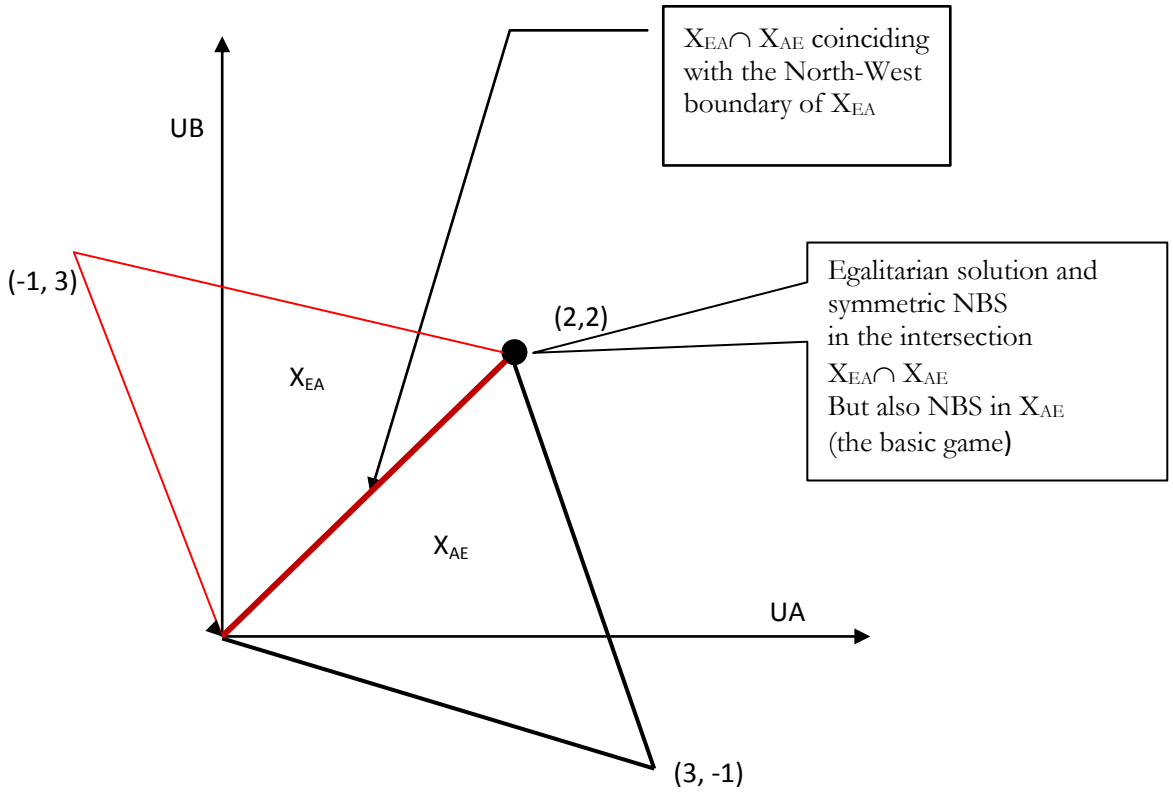


fig. 6, Symmetric translation of the repeated TG payoff space and its 'intersection solution'

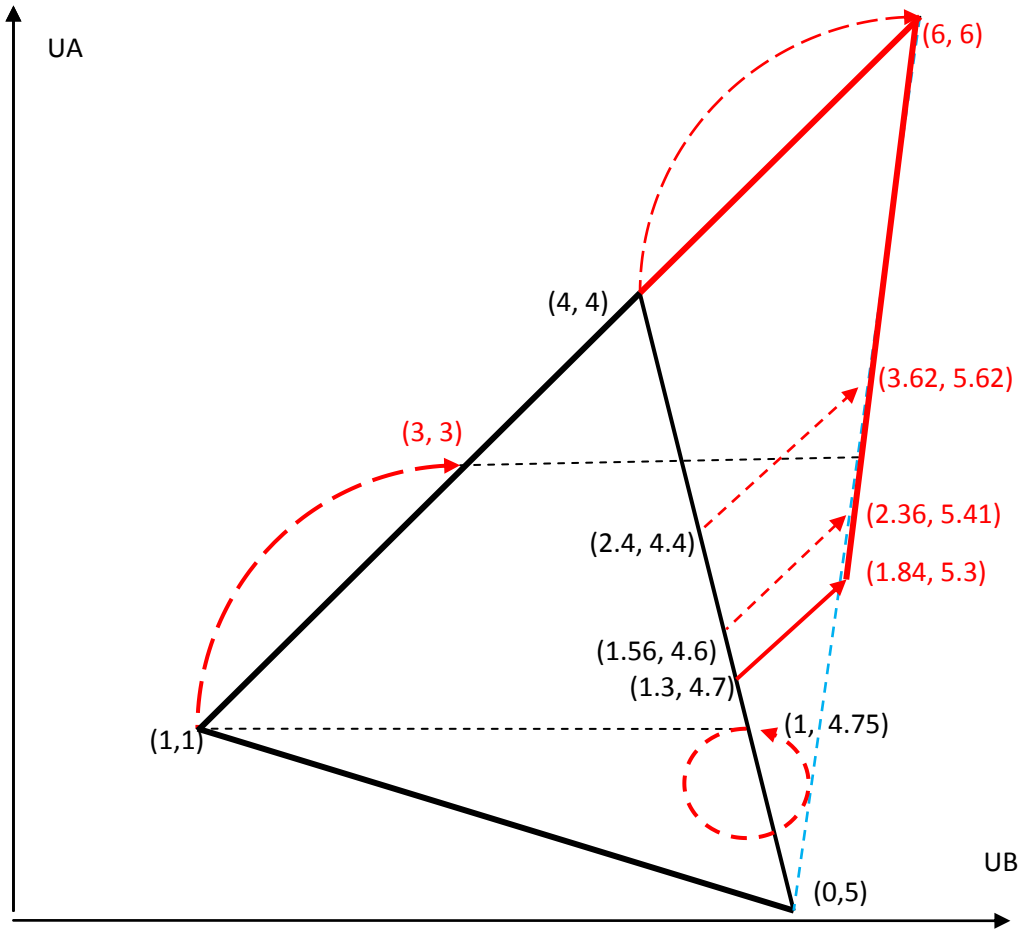


fig. 7, The payoff space of the iterated psychological TG. Payoffs of pure and mixed strategies are represented and their translations into the psychological game payoff space. Up to the mixed strategy $\sigma_B^{0.39}$ no psychological utilities accrue to players and hence a region of the basic TG payoff space does not translates into the psychological payoff space.

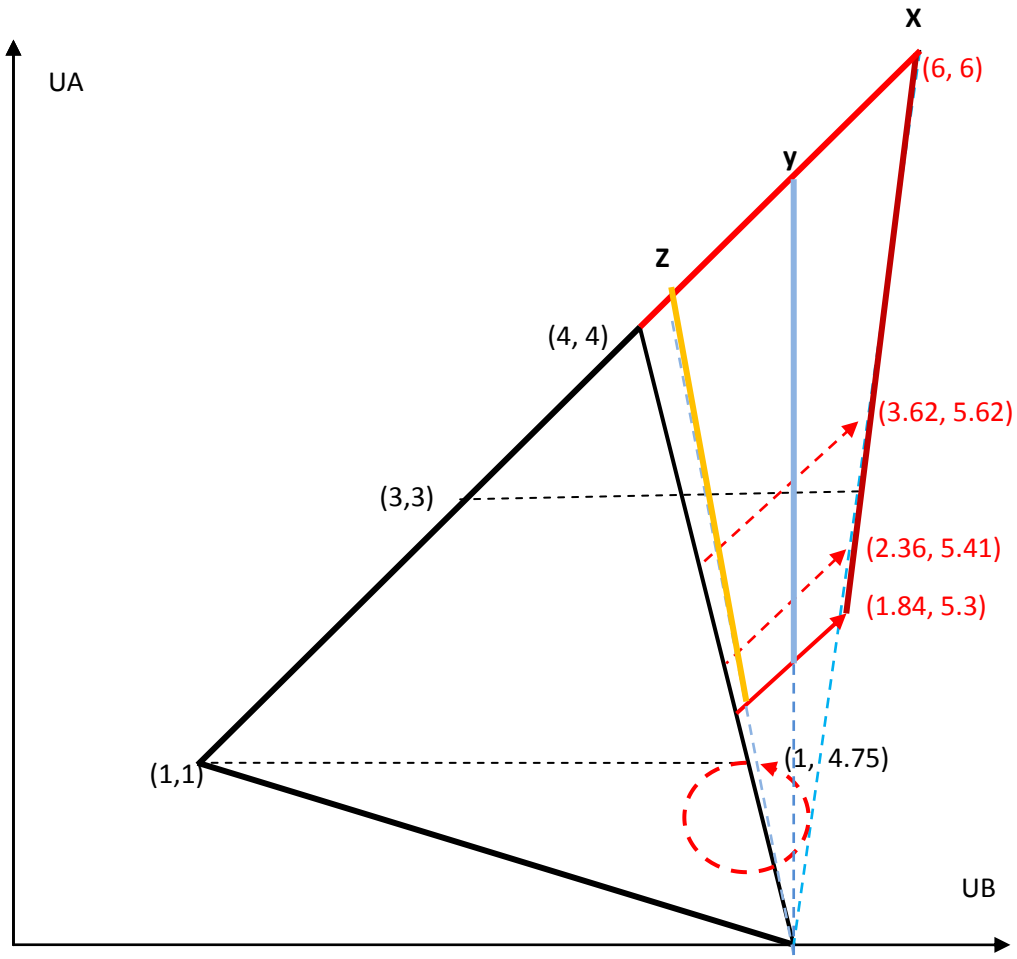


fig. 8, payoff spaces of the repeated psychological TG under three values of the parameter λ
λ <1 implies the NE frontier Z
λ = 1 implies the NE frontier Y
λ =2 implies the NE frontier X

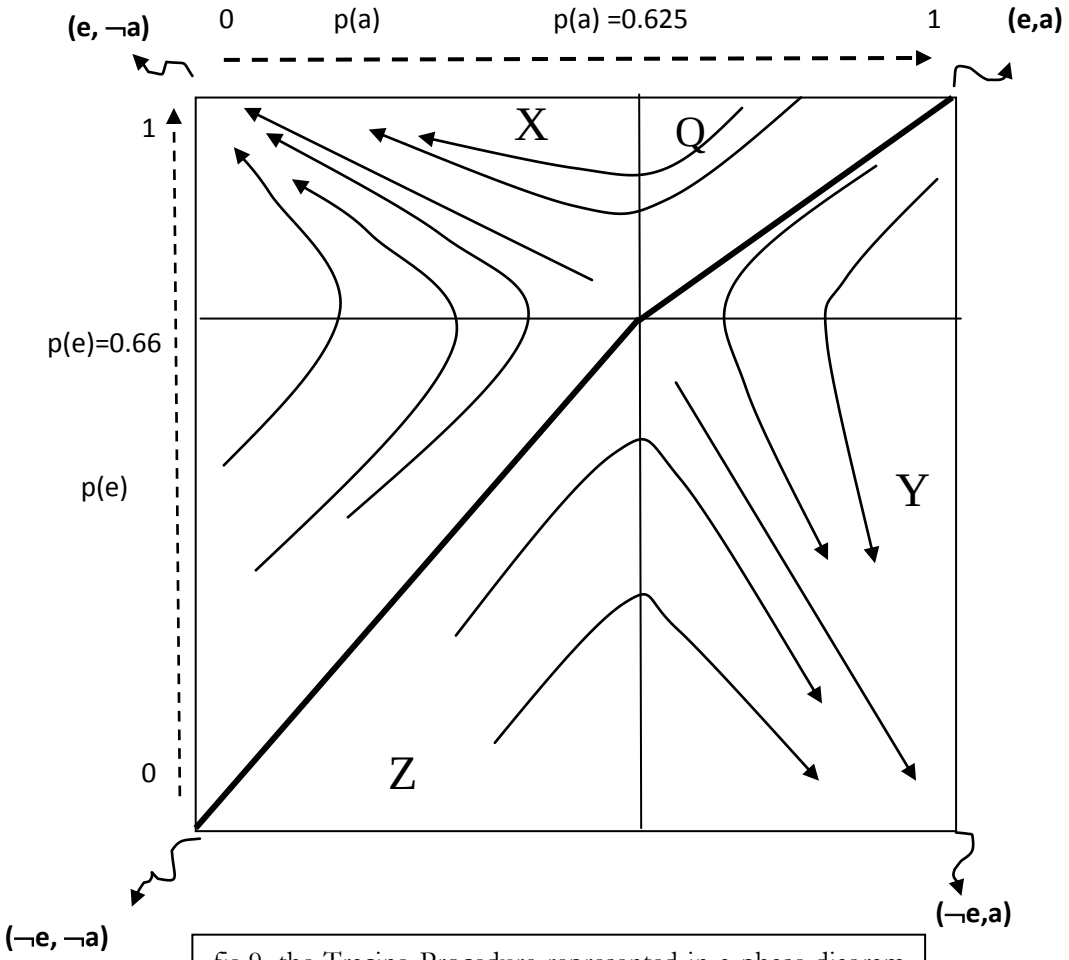


fig.9, the Tracing Procedure represented in a phase diagram with two basins of attraction