A Rawlsian View of CSR and the Game Theory of its Implementation (III): Conformism and Equilibrium Selection[1]

Lorenzo Sacconi

Department of Economics - University of Trento and EconomEtica, Inter university centre of research University Milan- Bicocca, Italy

1 Introduction

This is the third part of a comprehensive essay on the Rawlsian view of corporate social responsibility (CSR) seen as an extended model of corporate governance and the corresponding firm's objective function (for part I and II see Sacconi 2010a,b). In the first part of this essay, I provided the following definition of CSR as a multi-stakeholder governance model (see also Sacconi 2004/2007, 2006, 2009b):

CSR is a model of extended corporate governance whereby those who run a firm (entrepreneurs, directors, managers) have responsibilities that range from fulfillment of fiduciary duties towards the owners to fulfillment of analogous – even if not identical - fiduciary duties towards all the firm's stakeholders.

This definition has been articulated and defended as an institutional model of corporate governance implementable through explicitly expressed norms of self-regulation based on company/stakeholders social dialog – which means that CSR is neither a matter of managerial discretion nor one of external regulation enforced though statutory laws. The basic idea is that such a model of self-regulation, provided it is not obstructed by statutory company law which imposes a single-stakeholder fiduciary model and objective function on companies, is self sustaining. Hence the relevant perspective from which to understand the normative nature of CSR is that of an institution in Aoki's sense (see Aoki 2002, and Sacconi 2010a) . Let us summarize Aoki's definition:

An institution is a self-sustaining system of shared beliefs about a salient way in which a game is repeatedly played; it is based on a summary representation of compressed information about the equilibrium strategy combination which is currently being played in the repeated game characteristic of a given social domain (cfr. Aoki 2001).

[1] to appear in Social Capital, *Corporate Social Responsibility, Economic Behaviour and Performance* edited by Lorenzo Sacconi and Giacomo Degli Antoni , Palgrave London (in print)

However, the addition of a social contract perspective essentially completes the definition of 'institution' (Sacconi 2010a). The aim of this addition is to account for the crucial role that not just regularities of behavior and descriptive beliefs but also of norms and normative beliefs play as inherent parts of the beliefs system characterizing an institution as an equilibrium supported by a consistent system of expectations. To explain the role of the social contract on explicitly expressed self-regulatory norms of corporate governance, I take the game theoretic perspective of a repeated game between the firm - or those who occupy positions of authority within the hierarchical control structure of the firm - and the series of its stakeholders as the typical game in the 'corporate governance domain' (Aoki 2002).

Within this context, four roles played by a Rawlsian social contract have been identified in the first part of this essay in determining the equilibrium institution that satisfies the normative requirement of CSR. They are at the same time able to meet the main game theoretical challenges for the emergence of such an institution.

- The *cognitive-constructive role,* which answers the question on *how* the firm *works out* the *set* of commitments that it *can* undertake with respect to generic states of the world 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 entertain expectations about them;

- The *normative role*, which answers the question on what (if any) pattern of interaction the firm and its stakeholders must a priori *select* from the set of possible equilibria to be carried out *ex post* (according to the answer given to the first question), if they adopt an *ex ante* standpoint enabling an agreement to be reached impartially;

- The *motivational role*, which answers the question on *what* and *how many* equilibrium patterns of behavior, amongst those that may 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 the second question;

- The *cognitive-predictive role* concerning how the ex ante agreement on 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 each 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 the third question). The question to be answered by this function is thus '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 first two roles have been examined at length in part I and II respectively. In particular, it was seen in part II (see Sacconi 2010b) that, from the ex ante perspective, a Rawlsian social contract is able to solve the *normative* equilibrium selection problem, i.e. to choose a governance structure through a decision procedure that satisfies elementary conditions of impersonality, impartiality, and empathy. At the same time, it resulted in the egalitarian solution, consistent with the Rawlsian maximin principle, not just because of those ethical assumptions, but precisely because it internalizes the requirement of self-sustainability and implementation in equilibrium. This takes us to the typical Rawlsian maximization of the worst-off participant seen as a criterion for the constitutional choice of the firm's governance structure basically consistent with both justification and realistic implementation. Nevertheless, roles three and four still need to be explained. In fact, although the social contract is able to select ex ante a reasonable equilibrium, ex post we are again faced with the problem of the incentives to which players will respond when they exit from the original-position-and-veil-of-ignorance thought experiment and return to 'the game of life' (Binmore 2005) where they play according to the entire set of their preferences and motivations to act. This requires discussion of the equilibrium selection problem from the *ex post* perspective.

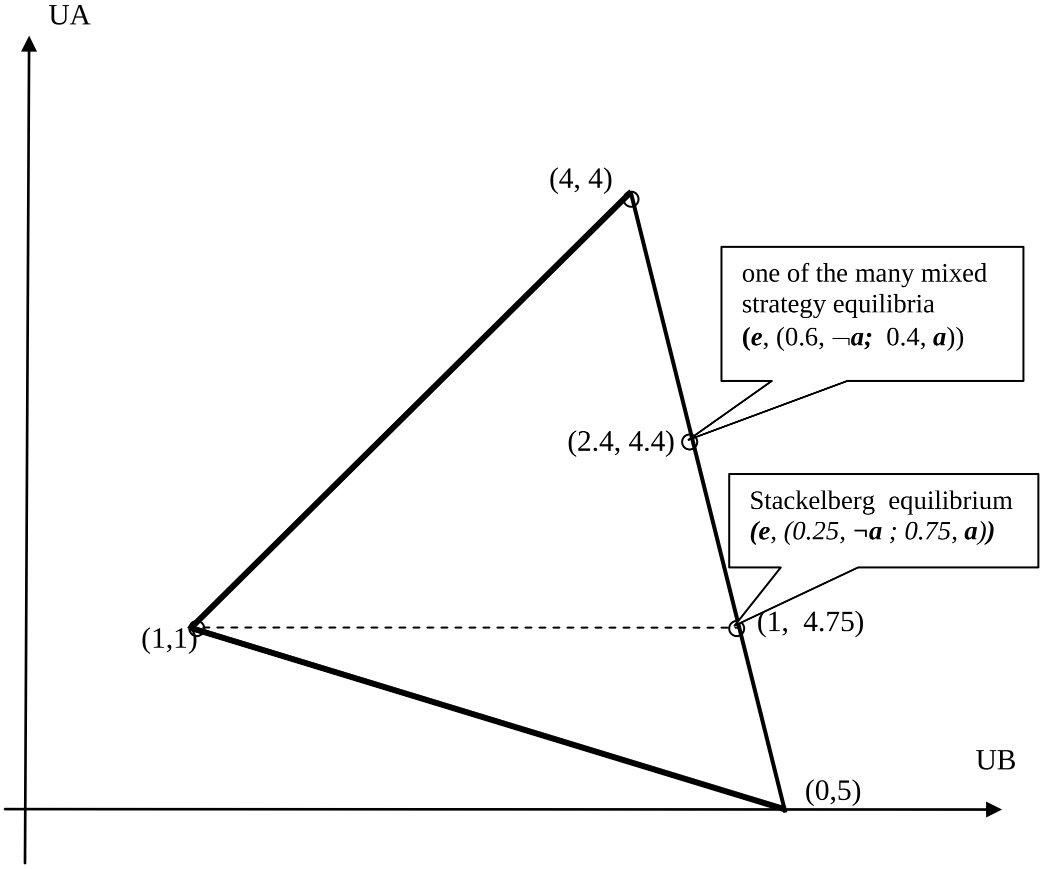
To gain better understanding of where we stand, consider that the appropriate game representation of the firm/stakeholders interaction is the iterated Trust Game, with the following stage game:

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

*Fig. 1 O*n- shot Trust Game

In the case of a one-shot game, player A (the stakeholder) will enter (or not) by trusting (or not) player B (who runs the firm) and by carrying out a specific investment. Player B decides whether to appropriate player A's investment by abusing or not. If he chooses non-abuse the surplus is shared in an equitable way. Otherwise the stakeholder is deprived of any benefit from entrance (including the endowment that s/he would possess if s/he did not invest), while the party who runs the firm gains a large profit. Note, however, that this mode of interaction is intuitively understood as socially inefficient in a utilitarian sense - that is, admitted utility comparability, the firm still prefers individually to abuse, but the fair sharing in the case of non-abuse would yield a larger amount of interpersonal social welfare. However, notwithstanding any consideration of social efficiency, the only Nash equilibrium is the strategy pair such that B abuses and A stays out. The mutually beneficial outcome (4, 4) cannot be sustained in equilibrium as long as the game is played one shot.

But now consider the equilibrium set of the repeated Trust Game between the long-run firm B and the 'average' stakeholder (call him/her again A because this is useful for considering the average payoff of an infinite series of short-run stakeholders that enter or otherwise the position of the one-shot A player at each repetition), who enters each stage game (or refuses to enter). Under the usual assumptions for reputation games (see part I), the repeated trust game will display a convex payoff space (constituted by all the average discounted payoff vectors obtainable from pairs of repeated strategies) coinciding with the convex envelope of the one-stage pure payoff vectors (see sec. 4 for more details).



one of the many mixed strategy equilibria (*e*, (0.6, ¬***a***; 0.4, ***a***))

Stackelberg equilibrium ***(e**, (0.25, ¬**a** ; 0.75, **a**))*

Fig. 2 Repeated Trust Game between the long-run firm B and the 'average' stakeholder A

Within this payoff space, every point above the dotted line corresponds to an equilibrium strategy profile such that player A "enters" with a given frequency and player B abuses or not with the appropriate probability mixture (Fudenberg and Levine 1989, Fudenberg 1991). Of course, the most relevant equilibria are the one where player A never enters because player B will always abuse, with average discounted payoffs (1,1), and the equilibrium with average discounted payoffs (4,4) where player 2 never abuses and hence player 1 enters each time. But also remarkable is the *Stackelberg equilibrium*, where the firm B seems to makes a commitment on the mixed strategy *(0.75**a**, 0.25 **no-a**)*. In fact B may develop a reputation for being this type by playing the two pure strategies with the attached probability throughout all the repetitions of the game. Thus each stakeholder in the role of player A necessarily enters, since his/her payoff is the same as staying put (1) – i.e. s/he is indifferent between entering and staying put (if player B were to give him/her an infinitesimal additional positive utility ε by reducing his/her abuse probability correspondingly, "entrance" would be certain). This gives B an average expected payoff of 4.75, which is the best payoff that player B can obtain in equilibrium. Then player B's best response is to stick to this type/commitment whenever s/he is able to convince player A that s/he is this type so that s/he responds with his/her best response to this type's mixed strategy (see also Andreozzi 2010, for a discussion of the relevance of this fact in the game theoretical explanation of CSR).

There is some evidence of this behavior in real life relationships between companies and their stakeholders. An example is provided by companies that claim to be socially accountable because they publish a social report and announce a code of ethics, but nevertheless are not accurate in reporting all the relevant social and environmental impacts of their conduct on all the concerned stakeholders and comply in only few cases, or to a minimal extent, with the declared code. Thus a company may acquire a reputation for abusing the trust of its employees, customers, suppliers, investors, capital-lenders and local communities wherein it operates – but only to the extent that makes them indifferent between maintaining their relations with the firm and withdrawing from them.

However, there is also evidence of stakeholder activism that refuses to acquiesce and actively countervails such hypocritical corporate conduct. In fact, stakeholder activism is a growing component of market behavior. Examples are phenomena such as responsible consumerism, socially responsible finance, human rights advocacy through active participation in shareholders meetings, brand boycotts in the case of environmental disasters, allegations of human rights violations or discrimination against employees by companies (especially when operating plants relocated to developing countries). Further examples of the same behaviors are corporate bankruptcies decreed by investors through the mass liquidation of stocks after ethical scandals (as in the case of Arthur Andersen after the Enron scandal). These companies - evidently responsible for intentional breaches of their ethical commitments - are doomed by their shareholders to collapse more dramatically than would be 'rational' according to those shareholders' self-interest (i.e. their share-value-maximization). All these examples illustrate behaviors by active stakeholders that cannot be captured in terms of their mere self-interest and cannot be understood as mere defense of their own material interest.

Admittedly some of these behaviors can be understood as reflecting a concern for other stakeholders' well-being, rather than the well-being of the active stakeholders themselves. More exactly, however, they express the stakeholders' attachment to impersonal principles of justice, i.e. a desire to conform with socially accepted norms of fair treatment - even when such conformity concerns not so much the active stakeholder itself but mostly the well-being of third parties. Hence only disinterested (from the egoistic point of view) motivations may be of relevance in explaining such action. A proper understanding of these third-parties-concerned non-egoistic behaviors in terms of norm compliance based on conformist preferences has been the focus of previous works on this topic (see Grimalda and Sacconi 2005; Sacconi and Grimalda 2007). Here I shall try to make sense of the evidence by focusing on the basic firm/stakeholders bilateral strategic relationships. This perspective is also a basis for extending the explanation to larger firms/stakeholders networks, where the creation of social capital and support for non-egoistically profitable trust relationships is at stake (see Sacconi and Degli Antoni 2009, Degli Antoni and Sacconi 2010, *infra*).

How does the social contract approach account for these apparently 'irrational' but unselfish actions, given that acquiescence would be the stakeholder's best response? In part II (Sacconi 2010b) the focus was on the ex ante agreement on CSR norms and standards of behaviour as a useful collective decision device for the unique selection of an equilibrium point. The concern now is with how stakeholders react to the discovery that in the game of life the firm has strong incentives to behave in a way quite different from strict compliance with the ex ante agreed equilibrium, and *de facto it prefers to* deviate from it. As a consequence it seeks to develop a reputation of being a type of player who systematically adopts a sophisticated abuse behaviour that if it was taken for granted would induce stakeholders to abandon the ex ante agreed equilibrium point and adapt to the less than fully compliant equilibrium profile.

This can be understood as a struggle for the *ex post* equilibrium selection amongst the many still possible. What we are in fact facing are two tightly connected but nevertheless distinct game theoretical problems. Firstly, *ex ante* equilibrium selection by agreement does not necessarily work well as an *ex post* equilibrium selection mechanism. Even though it ensures that the decision taken ‘behind the veil of ignorance’ could be self-enforceable if there were a system of expectations that predicted that decision as the effective ex post behavior of the parties, it does not ensure that these expectations will *de facto* emerge, and therefore that selection will be *ex post* effective. There is no logical *necessity* linking ex ante equilibrium selection to the emergence of the shared knowledge condition required for the unicity of the solution in the ex post perspective. But, secondly, this also raises the compliance problem again. Given multiple ex post equilibria, why should the player comply with the agreement by carrying out exactly the equilibrium chosen under the veil of ignorance? The problem is that, in the presence of multiple equilibria, each with some motivating force conditional on existence of a system of expectations consistent with it, no particular equilibria has any reason to be carried out, and thus the one corresponding to the ex ante agreement need not have any incentive effect on compliance.

A different answer could be given if the ex ante selective function of an impartial agreement by itself performed a *causal* role in changing incentives and beliefs on the set of admissible equilibria of the game of life relevant in the ex post perspective. This can happen along two routes. The first is a behavioral mechanism according to which the agreed equilibrium carries additional motivational (i.e. preferential) force precisely because it has been selected ‘behind the veil of ignorance’. The second is (again) a psychological mechanism according to which agreeing ‘behind the veil’ (as a *matter of fact about reasoning*, but without logical necessity) also influences beliefs about other parties’ behavior ex post: that is, it induces a state of shared beliefs whereby what was chosen behind the veil will be also implemented ex post. These two behavioral hypotheses are interlocked (i.e. beliefs formation must be granted in order to introduce the psychological preferences). Some empirical evidence for them can be found in related experimental works (Sacconi and Faillo 2008, Faillo, Ottone, Sacconi 2008). We discuss the first hypothesis in the next few sections by introducing a Rawlsian idea of the sense of justice and the corresponding model of conformist preferences. The latter hypothesis will be addressed in sections 5 and 6.

2 The true Rawlsian theory of norm compliance

An original approach to the institutional compliance problem was suggested by John Rawls in the *Theory of Justice* (1971), where he proposed the “sense of justice” as a solution for the stability problem of a well-ordered society - i.e. a society whose institutions are arranged according to the principles of justice (norms in our sense) chosen under a ‘veil of ignorance’. This solution, however, was for long overlooked by economists and game theorists because it was at odds with the methodology of rational choice in that it resorted to socio-psychological assumptions common in theories on moral learning.

However, given the behaviorist turn in microeconomics, it is time to reconsider this neglected solution and to acknowledge that it may suggest an illuminating explanation of why (sometimes) some of us comply with just institutions even if we have some direct material incentive not to do so. The rest of this section thus summarizes Rawls’ argument about how a sense of justice is engendered in a well-ordered society, and finally suggests the relevant features of Rawls’ theory captured in the conformist preferences model.

Justice as fairness, Rawls says, understood as the set of principles of justice chosen ‘under a veil of ignorance’ – once the principles are assumed to shape the institutions of a well-ordered society – provides its own support to the stability of just institutions. In fact when institutions are just (here it is clear that we are taking the ex post perspective, i.e. once the constitutional decision from the ex ante position has already been taken and for some reason has been successful), those who take part in the arrangement develop a sense of justice that carries with it the desire to support and maintain that arrangement. The idea is that motives to act are now enriched with a new motivation able to overcome the counteracting tendency to injustice. Note that instability is clearly seen in term of a PD-like situation: institutions may be unstable because complying with them may not result in the best response of each participant to other members’ behavior. However, the sense of justice, once developed, overcomes incentives to cheat and transforms fair behavior into each participant’s best response to the other individuals’ behaviors.

To understand how this is possible, it is necessary to consider the definition of ‘sense of justice’. Although it presupposes the development of lower-level moral sentiments of love and trust, understood as feelings of attachment to lower-level institutions (families and just associations), if these institutions are perceived to be just, it is noticeable that the sense of justice is a desire to act upon general and abstract principles of justice as such, once they have been chosen under a veil of ignorance as the shaping principles of institutions, and hence have proved beneficial to ourselves in practice. Note that it is not the case that we act upon the principles insofar as they are beneficial only to concrete persons with whom we have direct links and emotional involvements. Once the level of a morality of principles has been reached, our desire to act upon the principles does not depend on other people’s approbation or on other contingent facts such as satisfaction of the interests of some particular concrete person. On the contrary, it is the system of principles of justice in itself that constitutes the object of the sense of justice.

The question to be answered thus becomes how it is possible that principles themselves are capable of influencing our affections - that is, of generating the sense of justice as a relatively self-contained "desire to conform with the principles". The answer is twofold.

First, the sense of justice is not independent of the *content* of principles. These are principles that we could have decided to agree upon under a veil of ignorance as expressions of our rationality as free and equal moral persons. These principles are mutually advantageous and hence impartially acceptable by a rational choice, even if it is made from an impartial perspective, for they promote our interests and hence have some relation with our affections (preferences). Thus, in order for a sense of justice to develop, principles cannot be arbitrary. They must be those principles that would have been chosen by a rational impartial agreement.

Second, despite the intellectual effect of recognizing that principles are rationally acceptable, the basic fact about the sense of justice is that it is by nature a moral sentiment inherently connected to natural attitudes. Moral sentiments are systems of dispositions interlocked with the human capability to realize natural attitudes. Thus moral liability for lacking moral sentiments has a direct counterpart in the lack of certain natural attitudes which results in affective responses like a sense of guilt, indignation or shame. Hence, even though the thought experiment of a decision under the veil of ignorance merely aids us in the *intellectual* recognition of principles acceptability, the sense of justice retains a motivational force on its own, which can be only traced back to its nature as a moral sentiment or desire not entirely reducible to the experience of its intellectual justification.

The proper functioning of the sense of justice can be understood, however, as the third level of a process of moral learning which in its first two steps already cultivates moral sentiments of love for parents and trust and friendship vis-à-vis the members of just associations in which the individual already takes part - and which s/he re-elaborates on those pre-existing sentiments. "*Given that a person's capacity for fellow feeling has been realized by forming attachment in accordance with the first two …[levels] and given that a society's institutions are just and are publicly known to be just, then this person acquires the correspondent sense of justice as he recognized that he and those for whom he cares are the beneficiaries of these arrangements*" (Rawls 1971, p.491.)

As seems clear, reciprocity is a basic element in this definition. In fact reciprocity is understood as a deep-lying psychological fact of human nature amounting to the tendency to "answer in kind". The sense of justice "*arises from the manifest intention of other persons to act for our good. Because they recognize they wish us well we care for their well being in return. Thus we acquire attachment to persons and institutions according to how we perceive our good to be affected by them. The basic idea is one of reciprocity, a tendency to answer in kind*" (p. 494). Two aspects are to be noted concerning the other person's "manifest intention" which elicits the tendency to "answer in kind". We recognize the caring for our good deriving from other people acting consistently with the principles of justice. Hence reciprocity is elicited not from the mere coherence of institutions with the principles of justice, but from the fact that other people make our good by acting intentionally upon those principles. What matters is not just reciprocity in accepting the principles, but the intention displayed by other players' concretely acting upon the principles for our well-being. Secondly, this intention cannot be a direct intention from concrete person toward us as particular persons. By complying with principles, our good is pursued in an unconditional way - that is, impersonally and not conditionally on any particular description of us based on contingent characteristics or positions.

It also makes immediately evident that the sense of justice is a force that typically emerges and stabilizes a well-ordered society only *ex post*, when institutions are already "out there" operating through some level of compliance by the members of society. Thus the question arises of where compliance with principles arise from at the very first step of their implementation, when it cannot be said that there is an history of well-ordered society institutions already operating.

Important here are the following elements taken from Rawls's analysis and incorporated into the model of conformist preference explained in the next section.

i) First, there is an exogenous disposition in our motivational system of drives to action – the capacity of a desire to act upon principles or the agent's duties. This derives from learning about the justice of lower-level institutions (family, associations) or the widespread operating of the institutions of a well-ordered society (such that if these conditions are not fully satisfied this exogenous motivational factor cannot be assumed to have an overwhelming force in general, and thus must balance with other motivational drives).

ii) Second, the foregoing element defines as just a capacity for the sense of justice, but its proper formation depends upon conditions relative only to the principles of justice and their compliance, as follows

a. agents construe and justify norms as the result of an impartial agreement under the 'veil of ignorance', i.e. before considering conformity, the principles of different states of affairs resulting from compliant or non-compliant actions must be assessed in term of their consistency with the fair principles - compliance is not arbitrary;

b. each agent knows that also others justify the norm and assess compliance decisions in a similar way;

c. we know, or have the reasoned belief that other agents are effectively playing their part in carrying out the principles, and this behavior , because of the content of the principles it conforms with, expresses an intention to be beneficial to us in impartial terms. Thus by playing our part in compliance we may be understood as reciprocating other agents' intentions - i.e. our compliance is conditional on theirs;

d. owing to the hypothesis of public knowledge, also other agents are predicted as having (and we know that they have) the reasoned belief that we do our part in benefiting them in an impartial manner by acting upon the principles, and thus they may be seen as reciprocating our intention expressed by our compliance with the principles – hence our compliance is conditional on their reciprocity as well.

e. When these conditions are satisfied, our capacity to form a "sense of justice" becomes effective and translates into a motivational force able to counteract incentives to act unjustly in situation like the PD game – i.e. a psychological preference for complying overcomes the preference for personal advantages gained by not complying and opportunistically exploiting other agents' cooperation.

What we will see in the next section is how conformist preferences derived from the Rawlsian idea of a sense of justice may affect compliance with the social contract amongst the firm and its stakeholders. Preferences incorporating the sense of justice will affect compliance by selecting as admissible the only subset of equilibria which are compatible with compliance with the agreed principles.

3 The motivational role of social contract: conformist preferences in the trust game

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. One may wonder whether the fact that a norm has been agreed from an ex ante (pre-play) perspective and exhibits various 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, one may ask 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 constraining the firm's discretion in the firm /stakeholder interaction, would in fact perform a motivational function. It would restrict 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 pay-offs 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 – that is, 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 pay-off 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.

Whatever the case, if the norm generates a modification in the players' pay-offs in favor of situations in which no significant deviation from reciprocal conformity occurs, then it may be 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 – the psychological Nash equilibrium (Genakoplos, et al., 1986) – based on conformist preferences (Grimalda and Sacconi 2005; Sacconi and Grimalda 2007)[1].

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 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[2].

To begin, let us illustrate the conformist preference model with reference to its application to the one shot (stage) Trust Game (TG) involving a firm (player *B*) and its stakeholder (player *A*) (see *fig.1*). However, stakeholder and firm now have *two* kinds of preferences defined over states of affairs resulting form their interaction, which are both capable of motivating their actions. On one hand (more basic), the first kind of preferences is based on the description of states of affairs σ brought about by their interaction *as consequences*, and their preferences regarding consequences are called *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 affairs σ resulting from their interaction but described as just *combination of actions.* (To be clear the typical Trust Game – see again fig. 1.1 - identifies four possible states σ coinciding with cells of its normal form, where pairs of strategies are represented – (***e, ¬a), (e, a), (¬e, ¬a), (¬e, a)*** - before attaching payoff over them.) When these states of affairs are qualified in terms of their consistency with an ex ante agreed ethical norm preference over them are *conformist* - where 'consistency' is defined as 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 justice for the distribution of material utilities coinciding with the stakeholders' social contract of the firm.

Let us assume that players have just agreed on 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 simply complying with the principle, if the same stakeholder expects that in doing so she will be able to contribute to fulfilling the distributive principle, and taking into account that she expects the other stakeholders (or the firm) also to contribute to fulfilling the same principle, given their expectations.

A complete measure of the player preferences is an overall utility function combining material utility, derived from her consequentialist preferences, with the representation of her conformist preferences represented by the conformist-psychological component of her utility function (see Grimalda and Sacconi, 2005). 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

$$V_i(\sigma) = U_i(\sigma) + \lambda_i F[T(\sigma)] \qquad (1)$$

where

i. U_i is player *i*'s material utility for the state σ;
ii. λ_i is an exogenous parameter $\lambda_i \leq 0$;
iii. *T* is a fairness principle defined for the state σ;
iv. *F* is a compounded index expressing the agent *i*'s conditional conformity and her expectation of reciprocal by any other player *j* with respect to the principle *T* for each state σ

Let's concentrate on the conformist part of the utility function. *First* (as it can be seen within the most internal brackets), there is a norm *T*, a social welfare function that establishes a distributive principle of material utilities. Players adopt *T* 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 pay-off (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. As we are looking for a contractarian principle of welfare distribution, let us assume - according to what I have argued in part II sec. 7- that *T* coincides with the Nash bargaining function taking the stay out outcome of the trust game as the *status quo*

(agreed principle of fair welfare distribution T)

$$T(\sigma) = N(U_1,...,U_n) = \prod_{i=1}^{n} (U_i\text{-}d_i) \qquad (2)$$

Second, a measure of the extent to which, given the other agents' expected actions, the first player by her strategy choice contributes to a fully fair distribution of material pay-offs 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 (she expects that) the other player will do. It is a *conditional 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/her 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) with the following form

(player's i conditional conformity index)

$$\left[1+\ f_i\left(\sigma_{ik},b_i^1\right)\right] \qquad (3)$$

This index takes its values as a function of f_i which in turn varies from 0 to -1 and measures player *i'*s *deviation degree* from the ideal principle *T* by making her choice conditional on her expectation about player *j*'s behavior

(player's i deviation degree)

$$f_i\left(\sigma_{ik},b_i^1\right)=\frac{T\left(\sigma_{ik},b_i^1\right)-T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right)-T^{MIN}\left(b_i^1\right)} \qquad (4)$$

where b_i^1 is player *i'*s belief concerning player *j*'s action, $T^{MAX}\left(b_i^1\right)$ is the maximum value of the function *T* due to whatever feasible strategy player *i* may choose given her belief about player *j*'s choice, $T^{MIN}\left(b_i^1\right)$ is the minimum value of the function *T* due to whatever feasible strategy player *i* may choose given her belief about player's *j* choice, and $T\left(\sigma_{ik},b_i^1\right)$ is the actual value of *T* due to player *i* adoption of her k-ary strategy σ_{ik} given her belief about player *j's* choice.

Third, a measure of the extent to which the *other* player (respectively the stakeholder or the firm) is expected to contribute to a fair payoff 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 for generating a fair allocation of the surplus, given what he (is believed to) believes. This measure consists of a *reciprocally expected 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). It is formally very similar to the conditional conformity index of the first player, i.e.

(player's j reciprocal expected conformity index)

$$\left[1 + \tilde{f}_j\left(b_i^2, b_i^1\right)\right]$$

In fact it is as well a function of $\tilde{\tilde{f}}$, the *expected player j 's degree of deviation* from the ideal principle *T*, which also varies from 0 to – 1 as is also normalized by the magnitude of the difference between player *j'*s full conformity and no conformity at all, given what he believes (and player *i* believes that he believes) about player *i*'s choice, i.e.

(*expected player j 's degree of deviation)*

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

where b_i^1 is player *i*'s *first order* belief about player *j*'s action (i.e. formally identical to a strategy of player j), b_i^2 is player *i*'s *second order* belief about what player *j*'s believes about the action adopted by player *i* , while $T^{MAX}\left(b_i^2\right)$ and $T^{MIN}\left(b_i^2\right)$ are defined as above but in relation to second player *i's* second order belief.

Fourth, there is an exogenous parameter λ (λ≥ 0) representing the motivational force of the agent's psychological disposition to act on the motive of reciprocal conformity with an agreed norm. This is a psychological parameter representing how strong the *sense of justice* or the "desire to be just" has grown up for an individual in a given population; it may be taken as dependent on exogenous variables like as the development of the affective capacity to act upon one's principles and duties that comes from lower level domain of interaction (as in Rawls' theory of moral development, the family and the circle of friends and small scale associations). Notice however that in the model it doesn't operates as such but as only once the agreement over T is given and as it is weighted by the measure of reciprocal conformity.

In fact 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* on the parameter λ, deciding whether it will actually affect or not (and, if so, to what extent) the player's pay-offs. Thus the complete psychological component of the utility function representing conformist preferences is

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

which reduces to the following cases:

(i) *λ[(1-x) × (1-y)]* = λ since both *x* and *y* are 0, if player *i* doesn't deviate and expects that player *j* doesn't deviate at all from complete conformity; (ii) *λ[(1-x) × (1-y)] = aλ < λ,* where *a*<1 since at least one (or both) of *x* and *y* are *0< x <-1* and *0< y <-1,* if player *i* partially deviates and /or expects player *j* partially deviates from complete conformity*;* (iii) $a\lambda$ =0 since in the above expression at least one (or both) of *x* or *y* are -1, if player *i* does not conform at all and/or expects that player *j* doesn't conform at all.

Summing up the effect of the different components, if a stakeholder expects that the firm (or vice versa) is reciprocally responsible for the maximal value of *T*, given what the firm expects about that stakeholder's behaviour, and the former is also responsible for a maximal value of *T* given the firm's (expected) behaviour, then the motivational weight of conformity λ will entirely enter the stakeholder's utility function. In other words, in the player's preference system λ will show all the force of the disposition to conform to agreed norms, so that complying with the principle will yield full conformist utility (in the psychological sense) in addition to the material pay-off of the same strategy. In the one shot Trust Game, this happens at its best in the state of affairs where the stakeholder enters, the firm does not abuse, and they mutually predict these strategy choices.

Note that if a player cannot do anything better to improve the 'collective' value of the principle *T* with respect to the *status quo* by means of his unilateral decision given the expected strategy choice of the other player, then he will be considered completely compliant by choosing to keep the status quo (no deviation from maximal conformity can be ascribed to his responsibility since her choice cannot do any better to maximize *T* than keeping to the status quo). This feature of the model depends on considering compliance in a non-cooperative ex post context wherein players are able to deviate unilaterally from an agreed norm, and secondly by considering conformity as conditional on the other player's expect level of compliance. Hence, in cases like the Trust Game, if the firm is expected to abuse, the stakeholder cannot do anything to improve the value of *T* on the status quo and therefore the stakeholder will be considered fully compliant with the principle by deciding to stay out. (As a matter of fact she could only worsen the *T* value by entering.) At the same time, the firm predicting that the stakeholder will stay out – given his prediction of the firm's abuse – cannot modify the value of *T*. Thus whatever 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, the conformity weight λ adds to the players' pay-offs. Under this respect, there is no difference between the case (*no-entry, abuse*) and the case of the stakeholder entering because she predicts that the firm is going not to be abusive and the firm refraining from being abusive because it predicts that the stakeholder will enter - which is obviously the case in which both players maximize *T* given the expected behaviour of the other player and hence necessarily the weight λ enters their payoffs as they are full compliant.

By contrast, if the stakeholder enters when the firm is unilaterally predicted to abuse, she would minimize *T* with reference to the alternative choice open to her of not entering, which scores a higher level of *T*. At the same time, the firm misses the opportunity to maximize *T* given the stakeholder's decision to enter, and hence the latter will be considered as not complying at all. This implies that when the firm unilaterally and successfully abuses its stakeholder, none of the conformist preferences can add value to the players' material pay-offs.

Lastly, if the firm chooses a mixed strategy whereby the stakeholder's decision between entry or non-entry has no influence on the *T* value, the stakeholder, whether she decides to enter or not, would be unable to improve the value of *T*. Therefore, by staying out she maximizes *T* as well. If, however, 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 pay-offs. On the contrary, were the stakeholder willing to enter when the firm adopts the mixed strategy (so that by entering she is equally compliant as when staying out), the firm would become responsible for a sharp deviation from full compliance, for it could have chosen not to abuse at all. In that case, it would not have maximized the value of *T* as it possibly could have. This may not be the minimum value for *T*, but it has nonetheless 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.

What has been said till now is by no means conclusive about the existence of psychological equilibria based on conformist preferences in the one shot Trust Game. It simply helps to give an intuition of how the psychological payoffs behave under different strategic and beliefs configurations. However to calculate just pure strategy psychological Nash equilibria let's start by considering the game matrix (a) (that replicates fig.1.1 for the reader convenience). Strategies combinations (state of affairs) and the relative material payoffs vectors are (*no-entry, abuse*) and (*no-entry, no-abuse*) with material pay-offs (1,1); (*entry, abuse*) with material pay-offs (0,5); and (*entry, no-abuse*) with material pay-offs (4,4). This is helpful in understanding 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 this particular case, the *status quo* coincides with the outcome of the no-entry strategy – (1,1) – which is the assurance level that player *A* can grant himself whatever player B’s choice not starting any trust based interaction. This pay-off must then be subtracted from whatever pay-off is used in the calculation of the Nash product annexed to any state of affair (strategy combination). The two further matrices (see below) show respectively (b) the Nash bargaining product calculated for each pure strategy combination needed to measure the consistency of each state with the respect to the principle *T* and the players’ relevant degrees of conditional and expected reciprocal conformity for each state, and (c) the overall pay-offs resulting from the addition of the psychological conformist preference weight $\lambda = 2$ to the material pay-offs where this addition is appropriate.

	¬ a	***a***
e	4,4	0,5
¬ e	1,1	1,1

Matrix (a):TG normal form

	¬ a	***a***
e	(4-1)(4-1) = 9	(0-1)(5-1) = - 4
¬ e	(1-1)(1-1) = 0	(1-1)(1-1) = 0

Matrix (b): *T* values at each state

	¬a	***a***
e	$(4+\lambda) = 6$, $(4+\lambda) = 6$	0, 5
¬ 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 ¬***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 pay-off. On the other hand, if the firm is predicted to play ¬***a***, then the stakeholder maximizes *T* by choosing ***e***. If this choice is also predicted by the firm, its choice for maximizing *T* is ¬***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). Lastly, 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) is sufficient to nullify the overall level of conformity. Moreover, when this is the case, no psychological conformity weights are implied in the players' pay-offs.

Summing up, given the value λ = 2, we may see that, as far as only pure strategies are concerned, two Nash psychological equilibria do exists (***e, ¬a***) and (¬***e, a***). Thus even in the one shot game, the situation is ameliorated for not only the 'bad' equilibrium is now possible, but from the point of view of the solution determinateness the situation is also worsened as it isn't any unique. I don't bother here the reader with the existence of mixed-strategy-psychological-Nash equilibria in the one-shot Trust Game as they are mostly relevant to out argument in the context of the repeated Trust Game considered in the next section (where also many standard Nash equilibria are possible). It is within the perspective of the repeated Trust Game that we have to verify whether conformist preferences with an ex ante agreed principle of justice will simplify the equilibrium selection problem.

4 Mixed strategies and refinement of the equilibrium set in the iterated trust game

4.1. Mixed strategies

Now let us consider the repeated Trust Game (TG). Recall that its pay-off space in terms of material utilities is the convex hull of all the linear (probability) combinations of the three pay-off vectors generated out of the pure strategy pairs of the basic Trust Game (see *fig. 2*). This is the same as representing the expected pay-offs 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 pay-off for a mixed strategy is formally the same as the *average pay-off* 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 pay-off of this repeated strategy, given a certain pure (or mixed) response by the second player, can be equated to the average pay-off of a cycle along which player *i* gets, in each of the three stage-game pay-offs, a given proportion of times out of the total number of times in the cycle (granted, of course, that during the game each repeated strategy pairs of the two players repeatedly enters a cycle with the same pattern of outcomes and the same average pay-off value for each player). It is thus 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 pay-offs of (0.25×4+0.75×5 = 4.75) and (0.25×4+0.75×0 = 1), respectively. This is equal to the average values attached to a repeated strategy whereby the firm plays the stage-game strategy ¬***a*** 75 per cent of the time and the stage-game strategy ***a*** 25 per cent of the time, assuming – to keep things simple again – that the stakeholder always responds with the stage-game strategy ***e***. It is obvious to see that in the one-shot Trust Game, no mixed strategy exists as a best response for the firm. In the repeated Trust Game, however, one knows that this is no longer true. In fact, the firm may create a reputation (along, for example, 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.

One must, however, consider the pay-off space of the psychological game, which can be generated from that of the Trust Game when all of the expected pay-offs of mixed strategy pairs are accounted for. This psychological Trust Game in pure and mixed strategies has the same pay-off space as the repeated psychological TG wherein the average pay-offs 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, one may ask what happens to the mixed strategy equilibrium points of the corresponding standard repeated Trust Game.

Before answering that question, one must define a way to calculate the expected psychological utility of any mixed strategy. Let us 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* – (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 0.6×9 + 0.4×(-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 to any other pure or mixed strategy by *A*, whereas $\neg$***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 $\sigma_B{}^{0.6}$ is a somewhat intermediate value 0.61. But assuming that *B* believes that player *A* will not enter by $\neg$***e***, then *B*'s mixed strategy $\sigma_B{}^{0.6}$ will maximize *T* no less than any other strategy by *B*. *B*'s conformity index under this hypothesis is thus 1. To conclude the example, consider *A*'s respective expected material pay-offs from playing ***e*** or $\neg$***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 B's expected material pay-offs 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, that is, by a value (1)(0.61)λ. Given $\lambda = 2$, the weight of the conformist motivation is 1.22, and the overall utility pay-offs of players *A* and *B* are 3.62 and 5.62, respectively.

In the repeated psychological Trust Game, these pay-offs correspond to the following pair of player *B* and player *A*'s repeated strategies: player *B* employs his pure strategies $\neg$***a*** and ***a*** repeatedly with frequency 0.6 and 0.4 respectively. By this repeated strategy, he tries to convince player *A* (or the sequence of short-run players who participate in the repeated game in the position of *A*) that he will stick to this frequency forever. Player *A* decides to play repeatedly her entry strategy ***e*** as long as she does not see player *B* employing *abuse* with a frequency higher than 0.4, but if this frequency is exceeded she will switch to '$\neg$***e*** forever'.

Since player *A*'s threat seems 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 with outcomes (***e, a***) and stage games with outcome (***e, ¬a***). The average pay-offs 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 I will come back to this point a little later.

Following the method mentioned above, under the hypothesis λ = 2, it is in fact possible to account for the entire pay-off space of the psychological Trust Game, including mixed strategies as well (see *Figure 3*).

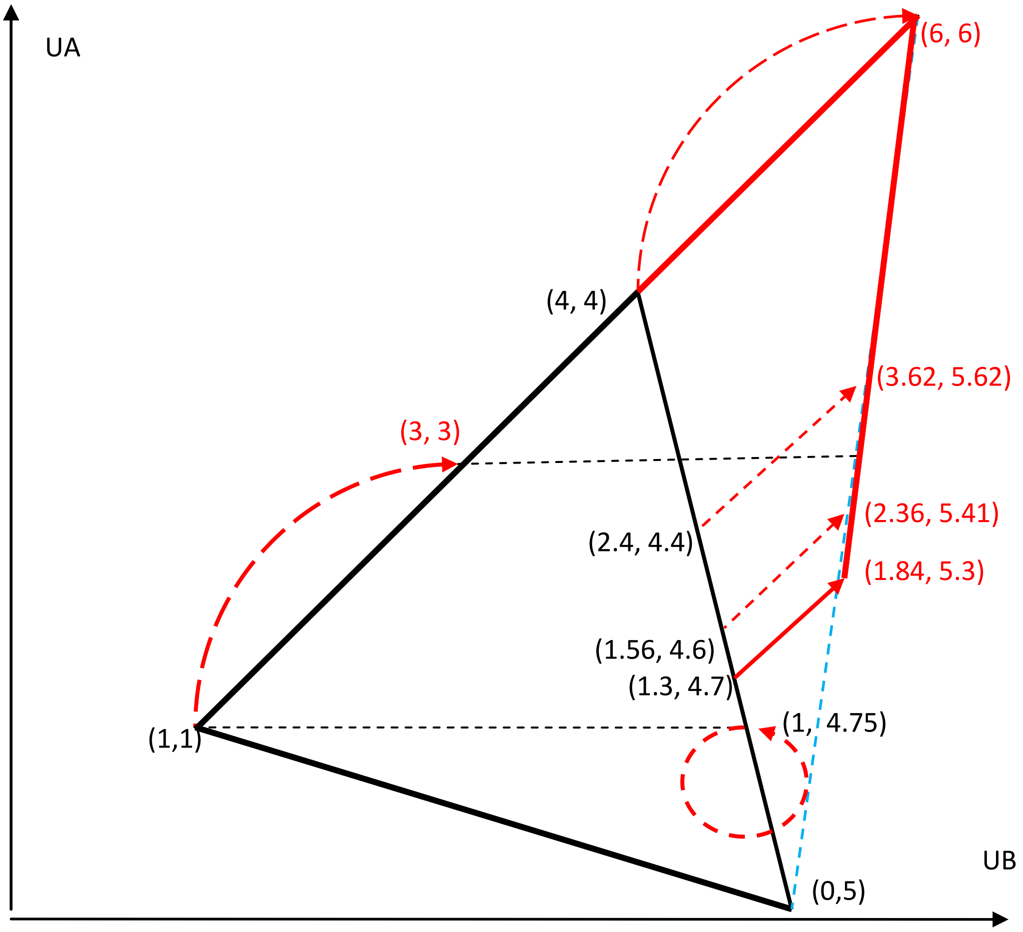


Fig. 3. The payoff space of the iterated psychological TG. Payoffs of pure and mixed strategies are represented and their translations into the psychological game payoff space. Up to the mixed strategy $\sigma_B^{0.39}$ no psychological utilities accrue to players and hence a region of the basic TG payoff space does not translates into the psychological payoff space.

First, let us 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 in the northeast direction 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 in the northeast direction to the point (6,6), which is also a psychological equilibrium. Let us recall that both these psychological equilibria correspond to Nash equilibria of the repeated Trust Game, so that these two equilibria are sure to be preserved under the pay-off 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 – the expected value of *T* associated with any particular probability mixture of the two pure strategies $\neg$***a*** and ***a*** – is no longer 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*** (-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 pay-offs do not add any values to the material pay-offs 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 pay-offs can be added to their material pay-offs.

This does not means that psychological utilities are not at work for these mixed strategies. Simply, the psychological component adds to the pay-offs of strategy pairs such as (*no entry*, *mixed strategy*), which is the same as for the strategy pair (*no entry*, *abuse*), or (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 (the *T* value) is negative (−0.75) for the pair ($\boldsymbol{e}$, $\sigma_B^{0.25}$), whereas *T* is

0 if player *A* chooses ¬***e***. It is thus obvious that *A* maximizes *T* by choosing ¬***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 pay-off 1, which – due to the probability mixture provided by $\sigma_B^{0.25}$ – is no different from the *material* pay-off of staying out. By staying out with ¬***e***, however, he gets an *overall* pay-off 3, because the psychological conformist weight 2 now adds to this strategic material pay-off. 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 pay-off equal to its material pay-off 4.75**.** When player *A* does not enter against $\sigma_B^{0.25}$, *B*'s pay-off benefits from the psychological conformist component (becoming 3) as well as from any other choice by *B* when he knows that *A* will play 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 Trust Game, would correspond to the preferred (by the firm) equilibrium strategy of the repeated Trust Game. 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 pay-off is high. Given strategy $\sigma_B^{0.25}$, neither is player *A*'s best response ***e***, nor is player *B*'s material pay-off 4.75 sufficient to make the strategy $\sigma_B^{0.25}$ the preferred than ***a*** choice when *A* plays ***e*,** simply because, due to a sufficiently high λ associated with the psychological equilibrium in pure strategies (*entry, no-abuse*), playing ¬***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, ¬***a***); (0.693, ***a***)}. 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 ¬***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 pay-offs (1.3, 4.7) generating an overall pay-off vector (1.84, 5.31). It is important to note, however, 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 pay-off to *A*'s entry strategy, player *A*'s overall pay-off from no-entry (3) is still higher than the overall pay-off (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***. In fact *B*'s full conformity would be reached by the strategy ¬***a***, whereas $\sigma_B{}^{0.307}$ scores only the modest conformity index 0.31. This affects the psychological conformist component of player *A*'s overall pay-off for strategy ***e***, which is lower than for ¬***e***.

Now let us consider mixed strategy $\sigma_B{}^{0.39}$ ={(0. 0.39, ¬***a***); (0. 61, ***a***)}. With this small increase in the probability of strategy ¬***a***, things finally seem to change. Player *A* with overall pay-off 2.36 benefits substantially from the psychological conformist utility of his entry strategy ***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 ¬***e*** given $\sigma_B{}^{0.39}$ would minimize the value of *T* in respect to the alternative entry strategy (and also any other mixed strategy). Hence, player *A*'s overall utility for the no-entry strategy ¬***e*** also dramatically drops to 1 (the material pay-off only). Moreover, for the pair (***e***, $\sigma_B{}^{0.39}$) , player *B*'s overall pay-off contains a substantial psychological conformist component such that his overall pay-off now reaches 5.41. If player *A* were to choose ¬***e***, however, player *B*'s pay-off would be reduced just to his material pay-off 1, since the conformity index of player *A*'s strategy ¬***e*** is zero (though *B*'s index remains positive). Note, nonetheless, that this does not imply that one has reached 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 ¬***a*** would do better in terms of conformity index, scoring an overall pay-off 6 higher than the mixed strategy.

This suggests a general fact about the model. Let us consider again the mixed strategy

$\sigma_B{}^{0.6}$ ={(0.6, ¬***a***); (0.4, ***a***)}.

As we know, player *A*'s conformity index if she uses strategy ***e*** against $\sigma_B{}^{0.6}$ is 1, whereas the mixed strategy's conformity index is 0.61. The annexed overall pay-offs are (3.62, 5.62), respectively. Even though high psychological conformist utility enters both the players' pay-offs, this is not enough to define reciprocal best responses at (***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 pay-off 6.

4.2 Equilibrium set of the psychological repeated Trust Game

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

- *Case 1*, $\forall\sigma_B{}^n$ with $n \geq 0$ s.t. $E^{n|e}(\Pi_{A,B}) < 0$, such that the pure strategy $\neg$***e*** induces $\Pi_{A,B} = 0 > E(\Pi_{A,B})^n$, the pure strategy ***e*** does not add any psychological conformist utility to player *A*'s material pay-off, whereas the pure strategy $\neg$***e*** adds the psychological conformity weight λ to the 'status quo' material pay-off. Hence player *A*'s best reply is $\neg$***e*** whereby *any* mixed strategy in this case is as good as strategy ***a*** to player *B*. The equilibrium for this case is the psychological equilibrium point ($\neg$**e**, ***a***). This equilibrium is weak since every mixed strategy in this case gives player *B* the same overall pay-off of ***a.***
- *Case 2*, $\forall\sigma_B{}^n$ with 0<n<1 s.t. $E^{n|e}(\Pi_{A,B}) > 0$, such that the pure strategy $\neg$***e*** induces $\Pi_{A,B} = 0 < E(\Pi_{A,B})^n$. Each pair (***e,*** $\sigma_B{}^n$) adds some psychological conformist utility to both players' material pay-offs, whereas the pure strategy $\neg$***e*** reduces player *A* to the 'status quo' material pay-off. This follows from the minimal conformity index of strategy $\neg$***e*** , while in this case mixed strategies $\sigma_B{}^n$ 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 ***e.*** Nevertheless, against strategy **e**, player *B*'s best is $\neg$***a***. In other words, as little as player *B*'s psychological conformist utility of a mixed strategy $\sigma_B{}^n$ is positive, player *B*'s pure strategy $\neg$***a*** against ***e*** (or whatever mixed strategy by player *A*) induces a psychological conformist pay-off higher than $\sigma_B{}^n$, so that player *B* has an incentive to deviate from $\sigma_B{}^n$ to $\neg$***a***. When this occurs, player *A* obviously has no reason to change her choice, and the equilibrium point is (**e**, $\neg$***a***).
- *Case 3*, for a single 0<*n*<1 $\exists\ \sigma_B{}^n$ such that $E^{n|e}(\Pi_{A,B}) = 0$, such that the pure strategy $\neg$***e*** induces $\Pi_{A,B} = 0 = E^{n|e}(\Pi_{A,B})$. In this case, both the strategy pairs (***e***, $\sigma_B{}^n$) and ($\neg$***e***, $\sigma_B{}^n$) add positive psychological conformist utility to the material pay-offs of both the players *A* and *B*. Nevertheless, player *A*'s overall pay-off gained from ($\neg$***e***, $\sigma_B{}^n$) strictly dominates her overall pay-off gained from (***e***, $\sigma_B{}^n$) since, whereas the two pure strategies ***e*** and $\neg$***e*** score the same conformity index, the case of player *B*'s conformity indexes is different. Player *B* against $\neg$***e*** cannot do any better than play $\sigma_B{}^n$ with conformity index 1, but given

e the strategy $\sigma_B{}^n$ conformity index is strictly less than 1, which is the conformity index of his pure strategy ¬***a***. Since the strictly less than 1 conformity index of strategy $\sigma_B{}^n$ directly depends on the required probability value *n*, which also affects the expected material utility of player *A* for (***e***, $\sigma_B{}^n$), this correlation is crucial in this case. It turns out that the greater player *A*'s pay-off gained from (***e***, ¬ ***a***) is, the smaller the probability required for the $\Pi_{A,B}$ indifference, but also the smaller the resulting player *B* conformity index for $\sigma_B{}^n$. Thus, player B's 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 pay-off – 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 the material pay-offs 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* in relation to player *A*, the *T* expected value of the mixed strategy pair (***e***, $\sigma_B{}^n$) cannot exceed that of player *A*'s staying out. This is not just because *A* is dissatisfied with her material outcome, but because of the insufficient conformity index of such mixed strategies. When a mixed strategy $\sigma_B{}^n$ 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 pay-off to *B*. At this point, however, 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* will always prefer the pure strategy of full conformity.

This also depends, of course, largely 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 regards to the basic TG and the repeated TG. If, however, λ is greater than player *B*'s pay-off difference between abusing and not abusing (given player *A*'s entry), its motivational effectiveness necessarily becomes maximal for the strategy of full conformity. In general, it biases the game towards excluding mixed strategies from giving rise to psychological equilibria. A look at the pay-off space reveals a single northeast vertex where both payers have highest pay-offs than anywhere on the eastern frontier where all the expected pay-offs generated by mixed strategies lie. In short, given its overall pay-offs, the pair (***e***, ¬***a***) strictly dominates any other strategy pair involving a mixed strategy $\sigma_B{}^n$ and player *A*'s entry strategy ***e.*** We have argued enough to state the following

PROPOSITION I

> *Given a Trust Game 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, none of player B's mixed strategies is the best reply to player A's pure entry strategy **e**, even if the entry strategy **e** is player A's best reply to player B's mixed strategy.*

From this proposition comes the following

COROLLARY

In the repeated psychological Trust Game, psychological equilibria 'refine' the equilibrium set of the corresponding repeated TG in a discontinuous way as a function of the increase in the motivational exogenous parameter λ.

- Given any λ such that in the one-shot psychological TG, there is no psychological equilibrium in correspondence with the pair (e, ¬a), the psychological equilibrium set is the same as the equilibrium set of the repeated TG due to the sole effect of material pay-offs (see northeast boundary X in Figure 4).
- If the value of λ is such that in the one-shot psychological TG player B's overall pay-off derived from the strategy combination (e, ¬a) is no different from the overall pay-off derived by B from the strategy combination (e, a) – so that a weak psychological equilibrium exists for (e, ¬a) – then in the corresponding psychological repeated TG the psychological equilibria constituted by a mixed strategy σBn and the pure strategy e have all the same player B expected pay-offs, and thus they are all weak equilibria. Given the continuity of the probability mixture set over the two pure strategies ¬a and a, the value of λ such that this is true is unique (see northeast boundary Y in Figure 4).
- If λ is such that in the psychological one-shot TG in correspondence to the pair (e, ¬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, ¬a) and (¬e, a). (See northeast boundary Z of Figure 4).

The corollary is important, because it is in this context that we see our result. As far as the pay-off space of a one-shot basic TG is concerned, 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 pay-off space is seen (as in the corollary) as the convex set of all the average pay-offs for repeated strategies in a repeated TG, then represented within this space may be the average pay-offs of player B's repeated strategies mixing the two pure strategies a and ¬a according to some pre-established frequencies.

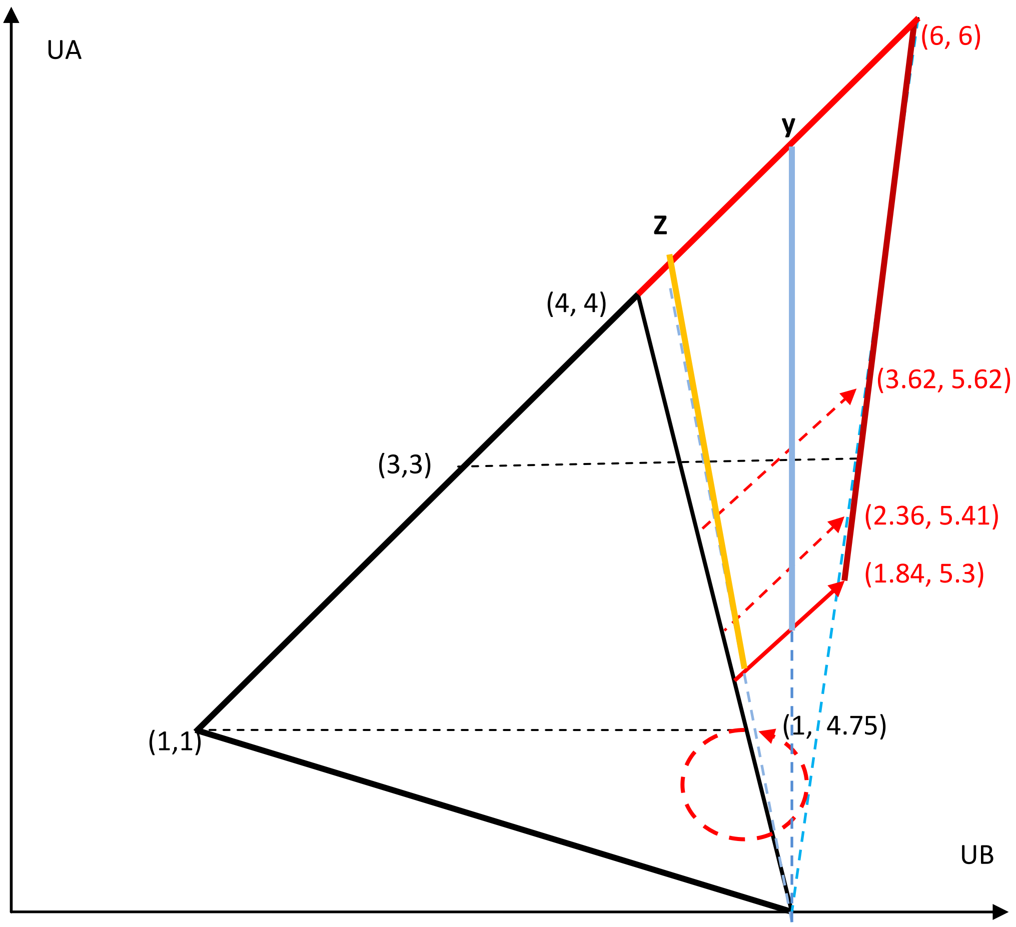


Fig. 4. Payoff spaces of the repeated psychological TG under three values of the parameter λ
λ <1 implies the NE frontier Z
λ = 1 implies the NE frontier Y
λ =2 implies the NE frontier X

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 if I find 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* per cent 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.

The situation changes significantly when the repeated psychological TG is considered, however. In this case, a pay-off space identical to the convex hull of all the pay-off pairs deriving from pure strategy combinations in the one-shot psychological TG is generated by taking the set of all the *average* pay-off 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 that induces player *A* to enter for the first time, then he immediately recognizes the incentive to switch to a strategy that employs ¬***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 her 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 *A*'s entry is to deviate from any mixed strategy σ_B^n to ¬***a***. If, however, player *B* deviates to a strategy more concessive to him, *A* 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 us to determine new equilibrium points (in other words, 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, transforming the pay-off 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?

5 Social contract-based ex ante beliefs

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 (see 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 strategy 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 the other players also 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. In other words, 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 – 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 one has ex ante decided that a principle corresponding to an equilibrium is enough to create this general expectation? It could, but it is important to realize that there is no necessity in this inference. What one decides to do in order to be impartial in the ex ante perspective is not necessarily what one 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. Let us recall that the motivational force of conformist preference – driving players to conform with an ex ante agreed principle – also operates conditionally on the previous expectations that 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-tiered answer. At a first stage, it is suggested that if each player has actually adopted a 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[3].

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 behavioural 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 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 (that is, no future behavior can be outguessed without a mental model of an agent performing the corresponding behavior). Let us 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 in them all conforming to 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 whatever 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***, ***¬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 also be 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. Moreover, this is 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 follows. If a player has agreed on a fairness principle it *normally* has a mental model of an agent who carries out the corresponding commitment, for this is the behavioral content of the principle he has agreed to. Moreover, nothing in his knowledge base (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 [4]. Considered separately, these mental models allow for 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 (that is, by a conclusion in terms of what is 'normally true') given our base of knowledge and given our two contrasting defaults – rules of implication – that an agent will 'normally' conform or not to the agreed principle. There is some uncertainty about whether the game situation he is playing either belongs to the situation set sketched by the one *frame* or by the other. This suggests 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 models 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 which entails that they will derive their beliefs over there from the same common prior probability distribution (this does not imply that the prior probability must be uniform; that will depend on the degree of vividness of each shared model).

A probability distribution over two pure strategy equilibria does not, however, 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* consists in assuming that the 'common prior' distribution generated by the two plausible mental models is taken as the starting point for revising beliefs, 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 reduces to the following plausible

PROPOSITION II

> *Agreeing impartially on a fair principle will give the shared mental model of a rational agent's (who conform to the principle) sufficient vividness to say that both players will in fact start their belief revision process from a common prior probability distribution wherefrom they will necessarily converge to a point such that they equally will completely believe that the solution of the game consists in the psychological equilibrium of full compliance.*

This proposition is in part empirical and its verification is left to future research (but see Sacconi and Faillo 2008, and Faillo, Ottone and Sacconi 2008 for related experimental results). However it has also a logical content in the statement that from a reasonably wide range of common probability distributions there exists a beliefs convergence process necessarily reaching the unique psychological equilibrium of full compliance. The remaining of this chapter is devoted to the proof of this second part of proposition II.

6 Eductive equilibrium selection from ex ante social-contract-based beliefs

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's best reply to the current state of his beliefs increases according to the probability of that player's strategy with respect to its prior probability.

Along this mental process of simulation, players never carry out a decision until uncertainty vanishes[5].

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 reply 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 give an idea of the *tracing procedure*, let us consider a thought process that takes place in a sort of 'reasoning time', which by construction starts from a stage of complete uncertainty $t° = 0$ and 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 $t° = 0$ 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 pay-off 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 $p°(\sigma_{ij}) = p°_{ij}$ revisions are generated by the following simple algorithm:

- $1\text{-}t^n(p°_{ij}) + t^n$ 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 i's best reply;
- $1\text{-}t^n(p°_{i,k\neq j})$ 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 Trust Game - which are well defined only for states of knowledge for which first- and second-level expectations 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 that the two mutually consistent expectation systems supporting each of them are thought to be possible as well. The player is thus uncertain about which of the two expectation systems 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 of the other player's uncertainty about his own equilibrium choices (thus, for each player, *first*- and *second*-order expectations about the other player's choices and beliefs are derived from a common prior and its revisions).

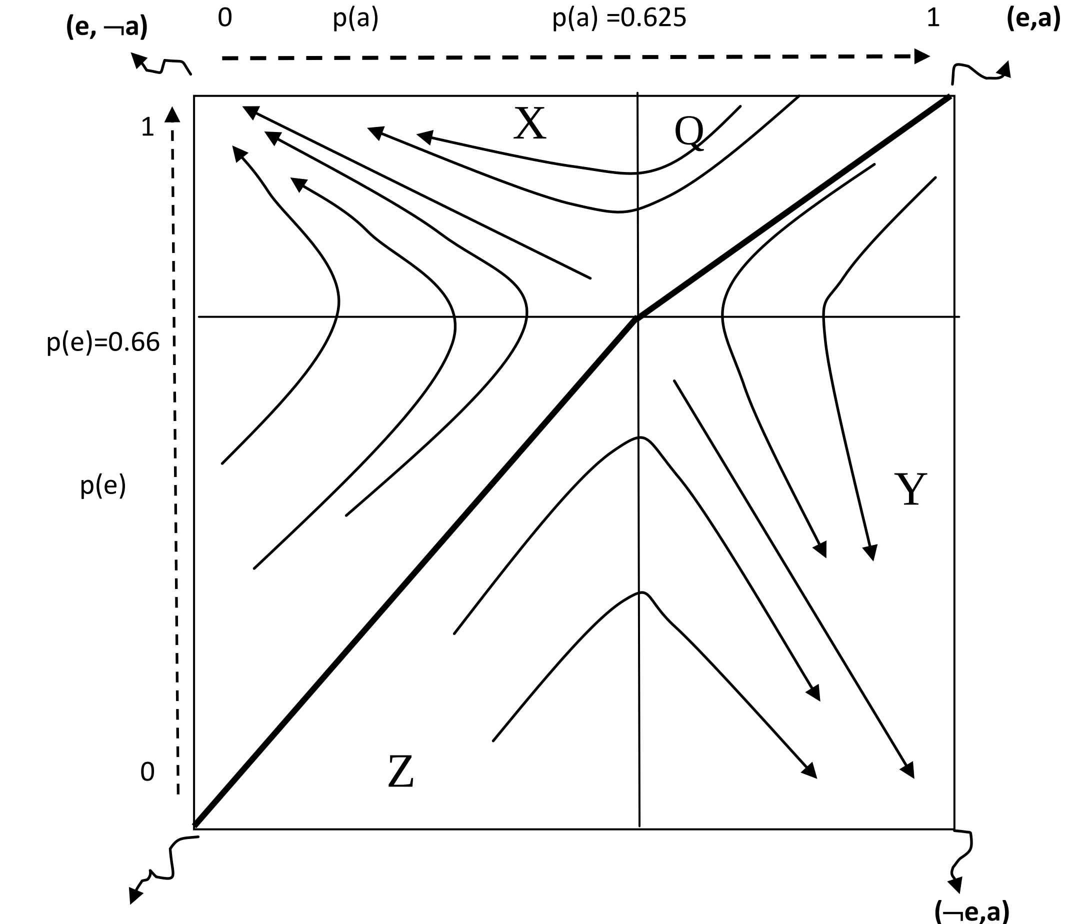


Fig. 5 The Tracing Procedure represented in a phase diagram with two basins of attraction

Consider the phase diagram of Fig. 5. It is a *probability square* representing all the possible discrete probability distributions over player A and 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. The probability of player B's strategy ***a***, on the other hand, 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 player A and B's strategies ***e*** and ***a***, respectively (as well as the probabilities assigned to the alternative strategies of both the players ¬***e***, ¬***a***). Corners represent pure strategy pairs when they are perfectly predicted (with probability 1) as indicated hereafter:

top-left:	(***e,*** ¬ ***a***);
top-right:	(***e*** , ***a***);
bottom-left:	**(**¬***e***, ¬***a***);
bottom-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 Figure 6.1 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 that is the attractor of the basin wherein the path has started.

Concerning paths starting from outside any basin of attraction, the procedure tends to induce a ‘change of mind’ in the players. Note that in the phase diagram of Figure 6.1 all paths traced by the procedure tend to move away from a non-equilibrium corner towards another non-equilibrium corner. From the region Z, paths generated by the tracing procedure move towards the northeast, that is, towards the non-equilibrium outcome (***e, a***), while from the region Q, paths move towards the southwest, that is, the non-equilibrium outcome ***(¬e, a***). Along these paths players make choices in accordance 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 pay-off. Hence he is under increasing pressure to change his choice to ***¬e***. At the same time, player B sees probabilities for drawing closer to corner (***e, a***), where he gets only the overall pay-off 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 towards the relevant equilibrium corner, which is the attractor within the intercepted basin of attraction.

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 Figure 6.1 will reach (***e,*** ¬***a***). These paths 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 move towards the corner (***e,*** ¬***a***) when they cross the boundary of the basin of attraction X. Under such circumstances, player A – who until that moment would have chosen the strategy ***a*** as his best reply within the dynamics process – changes his best reply as he is at risk of reaching the non-equilibrium outcome ($\neg$***e***, $\neg$ ***a***). Above the boldface diagonal this happens necessarily before an analogous incentive pushes player A to switch from strategy $\neg$ ***a*** to strategy ***a.***

It is also true, however, that the largest part of the probability square gives rise to paths, those starting at the point beneath the boldface broken diagonal, converging to the equilibrium corner ($\neg$***e***, ***a***). This means that the tracing procedure does not allow by itself a unique prediction that the equilibrium that is fully conformist with the CSR norm will be selected.
We must here resort again to the first step in our two-tiered answer. The ex ante agreement on a principle of fairness by default allows for 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 a commitment to carry out a certain behavior later on.

Although 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's 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 great *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. Stated another way, the thought experiment of putting players under a 'veil of ignorance' allows them to abstract themselves from a concrete context of distrust and to frame the situation as one of 'constitutional choice' whereby they choose *from the beginning* the rule for entering a new interaction. This allows them to make default inferences abstracted 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.

There is some reliable evidence gathered by experimental studies about the formation of conformist preferences in favor of this hypothesis (see 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 (that is, 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 turns 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 believe other participants in the agreement (even if it is completely anonymous) will also conform and believe others will conform. The only difference between the players who decide before participating in 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 acts consistently with the behavioral content of the agreement, so that they rationally reply using the equilibrium strategy of full conformity to the principle.

7 Concluding remarks

This concludes the explanation of the initially suggested four roles of voluntary, yet explicit CSR norms based on a Rawlsian social contract. These norms make it possible to describe strategies and equilibrium points, even when the equilibria are multiple, in a game played under unforeseen contingencies among the firm and its stakeholders (see part I, Sacconi 2010a). A CSR norm allows for the ex ante selection of the equilibrium point that meets the requirements of an impartial choice (see part II, Sacconi 2010b). An explicit agreement on a contractarian norm is, moreover, a way of introducing psychological conformist equilibria and, surprisingly, of deriving the important result that mixed strategy equilibria are absent from a psychological repeated Trust Game (see section 4). 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 mostly consistent with the assumption that players have agreed on an ex ante norm affecting their *de facto* mutual expectations, entails 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 strategy components in this equilibrium. The game theory of endogenous implementation of the normative model of multi-stakeholder fiduciary duties is thus complete.

Notes

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

[2] 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 thus a peculiar and somewhat surprising result of the conformist preferences model within the Trust Game context.

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

[4]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'.

[5] On evolutionary equilibrium selection mechanisms with learning through repeated plays, see Young (1998). The distinction between 'eductive' vs. 'evolutionary' equilibrium selection dynamics is provided by Binmore (1987).

References

Andreozzi L (2010), "When Reputation is not Enough: Justifying Corporate Social Responsibility", in L. Sacconi, M. Blair, E. Freeman and A. Vercelli (ed.) '*Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*', Palgrave London , in print

Aoki, M. (1984), *The Cooperative Game Theory of the Firm*, Cambridge: Cambridge University Press.

Aoki, M. (2001), *Toward a Comparative Institutional Analysis,* Cambridge, MA: 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, May 24–25, 2007.

Aoki, M. (2007), 'Endogenizing Institutions and Institutional Change,' *Journal of Institutional Economics*, **3**, pp. 1–39.

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**, pp. 7–48.

Bacharach, M. (2006), *Beyond Individual Choice, Teams and Frames in Game Theory*, edited by N. Gold and R. Sugden, Princeton, N.J.: Princeton University Press.

Binmore, K. (1991), 'Game theory and the social contract' in R. Selten (ed.), *Game Equilibrium Models II, Methods, Morals, Markets*, Berlin: Springer Verlag.

Binmore, K. (1987), 'Modeling rational players', *Economics and Philosophy*, **1** (3), pp. 9–55 and **2** (4), pp. 179–214.

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

Chareness, S. and M. Dufenberg (2006), 'Promises and Partnership', *Econometrica*, **74** (6), pp. 1579–601.

Degli Antoni G. and L.Sacconi (2010), infra

Dezau, A. and D. North (1994), ‘Shared mental models: Ideologies and institutions’, *KIKLOS*, **47** (1), 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.

Faillo, M. and L. Sacconi (2007), ‘Norm Compliance: The contribution of Behavioral Economics models’, in A. Innocenti and P. Sbriglia (eds), *Games, Rationality and Behavior*, London: Palgrave Macmillan.

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; online available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1151245.

Frey, B. (1997), *Not Just for the Money*, Cheltenham: Edward Elgar.

Fudenberg, D. (1991), ‘Explaining cooperation and commitment in repeated games’, in J.J. Laffont (ed.), *Advances in Economic Theory, 6th World Congress*, Cambridge: Cambridge University Press.

Fudenberg, D. and D. Levine (1989), ‘Reputation and equilibrium selection in games with a patient player’, *Econometrica*, **57**, pp. 759–78.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989), ‘Psychological games and sequential for Non-cooperative Games’, *International Journal of Game Theory*, **5** (1975), pp. 61–94.

Grimalda, G. and L. Sacconi (2005), ‘The constitution of the not-for-profit organisation: reciprocal conformity to morality’, *Constitutional Political Economy*, **16** (3), pp. 249–76.

Harsanyi, J.C and R. Selten, (1988), ‘A General Theory of Equilibrium Selection’, Cambridge, MA: MIT Press.

Harsanyi, J.C. (1975), ‘The Tracing Procedure. A Bayesian Approach to Defining a Solution’, International Journal of Game Theory, **4** (2), pp. 61-94.

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

Johnson-Laird, P.N. and R.M.J. Byrne (1991), *Deduction*, Hove and London: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. (1986), *Mental Models Towards a cognitive science of language, inference and Consciousness*, Cambridge: Cambridge University Press.

Nash, J. (1950), ‘The Bargaining Problem’, *Econometrica*, **18**, pp. 155–62.

Posner, E.A. (2000), *Law and Social Norms*, Cambridge, MA: Harvard University Press.

Rabin, M. (1993), ‘Incorporating fairness into game theory’, *American Economic Review*, **83** (5), pp. 1281–302.

Rawls, J. (1971), *A Theory of Justice*, Oxford: Oxford University Press.

Reiter, R. (1980), ‘A Logic for Default Reasoning’, *Artificial Intelligence*, **13**, pp. 81–132.

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 No. 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*, Volume 68, Number 3 / October, 2006, pp.259-281

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*, **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. and M. Faillo (2005), ‘Conformity and Reciprocity in the ‘Exclusion Game’: An Experimental Investigation’, Discussion paper del Dipartimento di economia dell’Università di Trento, No. 12, 2005; 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*, Cheltenham: Edward Elgar.

Sacconi, L. and S. Moretti (2008), ‘A Fuzzy Logic and Default Reasoning Model of Social Norms and Equilibrium Selection in Games under Unforeseen Contingencies’, *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, **16** (1), pp. 59–81.

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’, Constitutional political economy

Sacconi, L. (2009), ‘Corporate Social Responsibility: Implementing a Contractarian Model of Multi-stakeholder Corporate Governance trough Game Theory’ in J.P. Touffut and R. Solow (ed.), *Does Company Ownership Matter?*, Centre for economic Studies Series, Edward Elgar Publishing Ltd., London.

Sacconi, L. and G. Degli Antoni (2009), ………………..

Sacconi, L. (2010a), “A Rawlsian view of CSR and the Game Theory of its Implementation (Part I): The Multistakeholder Model of Corporate Governance”, in L. Sacconi, M. Blair, E. Freeman and A. Vercelli (ed.) ‘*Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*’, Palgrave London, in print. .

Sacconi L. (2010b), “A Rawlsian view of CSR and the Game Theory of its Implementation (Part II): Fairness and Equilibrium”, in L. Sacconi, M. Blair, E. Freeman and A. Vercelli (ed.) ‘*Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*’, Palgrave London, in print..

Segal, U. and J. Sobel, (2007), ‘Tit for tat: Foundations of preferences for reciprocity in strategic settings’, *Journal of Economic Theory*, **136**, pp. 197– 216.

Van Damme, E. (1987), *Stability and Perfection of Nash Equilibria*, Berlin: Springer Verlag.

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