EconomEtica

Centro interuniversitario per l'etica economica
e la responsabilità sociale di impresa
promosso dalla Fondazione Italiana Accenture

N.73 August 2019

Compliance with socially responsible norms of behavior: reputation vs. conformity

Virginia Cecchini Manara,
Lorenzo Sacconi

Compliance with socially responsible norms of behavior: reputation vs. conformity

Virginia CECCHINI MANARA[1] *and* Lorenzo SACCONI [2, 3]

[1] Department of Economics and Management, University of Trento

[2] Department of Public and Supranational Law, University of Milan

[3] EconomEtica, University Milano Bicocca, University of Milan – Bicocca

1. Self regulation and the problem of compliance

The Social Responsibility of Business typically involves self-regulation, which entails spontaneous compliance with social norms or standards that are not imposed by hard law.

In the debate on Corporate Social Responsibility, its voluntary basis has been stressed both in some official documents[1] and, more significantly, by the major scholars of the stakeholder approach.

In particular, the effectiveness of social responsible behavior requires a change in beliefs and motivations rather than in formal rules. For example, Margaret Blair and Lynn Stout (1999) argue that the internal hierarchy of public corporations should play a mediating role among all the team members (stakeholders of the firm) that does not require a change in the corporate law, but a shift from the dominant paradigm of shareholder primacy towards a team production model. Also in the stakeholder theory (Freeman et al. 2010), the importance of modifying the current mindsets about business is highlighted as a determinant factor for "putting business and ethics together" in a way that is implementable in the real world.

In order to understand the mechanisms that lead economic agents to comply with socially responsible norms that are not legally enforced, and do not coincide with profit, or self-interest, maximization, our starting point is the analysis of Corporate Social Responsibility as an *institution*, in the sense of Aoki (2001), who conceptualizes institutions as equilibria, i.e. self-enforcing systems of beliefs and behaviors, where individual and collective cognitions and motivations are intertwined. In this interpretation, the organizational process may be considered as a game among agents with different

[1] For example the 2001 Green Paper by the European Commission defined CSR as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders *on a voluntary basis*"; and the same was reaffirmed in the 2002 Communication, "Corporate Social Responsibility: A business contribution to Sustainable Development". Also the International Labour Organization (ILO) defines CSR as a "voluntary, enterprise-driven initiative" that goes beyond legal obigations.

knowledge, beliefs and motivations; an institution refers to that portion of agents' equilibrium beliefs common to (almost) all of them regarding how the game is actually played.

2. Cognitions and motivations

In this work we investigate if and how self-regulation can bring to an effective implementation of social responsibility within a framework where Corporate Social Responsibility (CSR) is understood as a corporate governance model based on duties towards several stakeholders.

In particular, we ground our proposal on the definition of CSR as a multi-stakeholder governance model whereby those who run a firm (entrepreneurs, directors, managers) have responsibilities that range from fulfillment of fiduciary duties toward the owners to fulfillment of analogous – even if not identical – fiduciary duties toward all the firm's stakeholders (Sacconi, 2010a).

The relevant perspective is that of an institution in Aoki's sense (a self-sustaining system of shared beliefs about a salient way in which a game is repeatedly played) completed with the idea of a **social contract** reached, in a Rawlsian hypothetical situation, "under a veil of ignorance" between the firm and its stakeholders (Sacconi, 2010c).

The definition of an institution in Aoki's work is depicted in *Figure 1*: cognitive components (i.e., beliefs deriving from compressed mental representations of salient aspects of ongoing equilibrium play) and behavioral components (i.e., the iterated play of a given set of equilibrium strategies) are interlocked in a recursive scheme. The starting point is cognitive, and it consists in pattern recognition whereby given situations of interaction are framed as games of a certain form wherein players are expected to reason in a given quasi-symmetrical way. At step two, this framing of the situation induces players to entertain quasi-converging beliefs about a certain mode of playing the game. Thus, at step three, on passing from beliefs to the players' actual behavior, each player adopts a tentative strategy based on the belief that others will also adopt strategies consistent with the aforementioned mode of behavior. Hence, in step four, strategies clash and some of them prove to be more successful and based on a better prediction. By trial and error, therefore, strategies converge towards an equilibrium of the game.

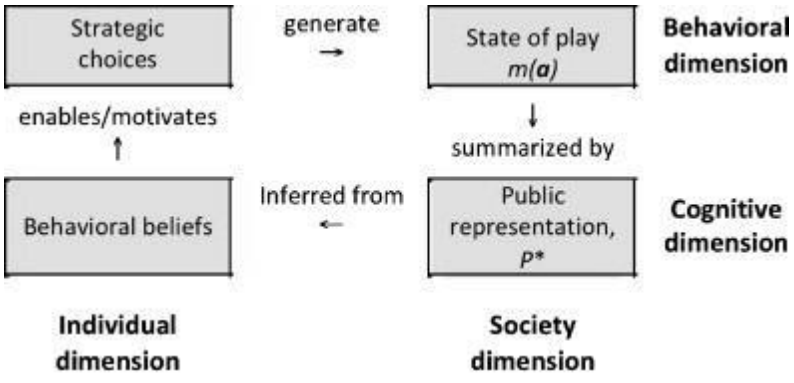

Figure 1 – The mediating role of institutions in substantive form (source: Aoki, 2011)

This may be construed as an evolutionary result because the mode of playing attracts more and more players through iterated adaptation to the other players' aggregate behaviors in the long run. At each repetition, however, this evolving equilibrium is summarily represented in its salient features by a

compressed mental model resident in the players mind so the fifth step concluding the circle is again cognitive. Aoki (2011) suggests that «there ought to be some *public representation* that mediates between the equilibrium play of a societal game and individual belief formation». He refers to an «external media» or artifact that linguistically represent salient features of equilibrium plays (such as norms, rules organizations of known types, laws).

We draw on some recent works (see for example Sacconi, 2012), where a modified version of Aoki's account of institutions is presented, completed with the idea of a Rawlsian **social contract**, reached in a hypothetical situation, "under a veil of ignorance", between the firm and its stakeholders (Sacconi, 2010c).

The introduction of an *agreement*[2] among actors has in our view a double effect: on one side it is able to activate in their mind a peculiar way of reasoning that generates shared representations leading to a fair outcome, on the other side it enriches their motives to act with a new motivation, based on the *sense of justice* that, once developed, overcomes incentives to cheat and transforms fair behavior into each participant's best response to the other individuals' behaviors.

Our work highlights the interplay of cognitive and motivational processes[3] and the fact that "human beings are biologically adapted for participating in collaborative activities involving shared goals and socially coordinated action plans"[4], by recognizing that the (real or hypothetical) participation in an agreement is able to impact on both cognition and motivation.

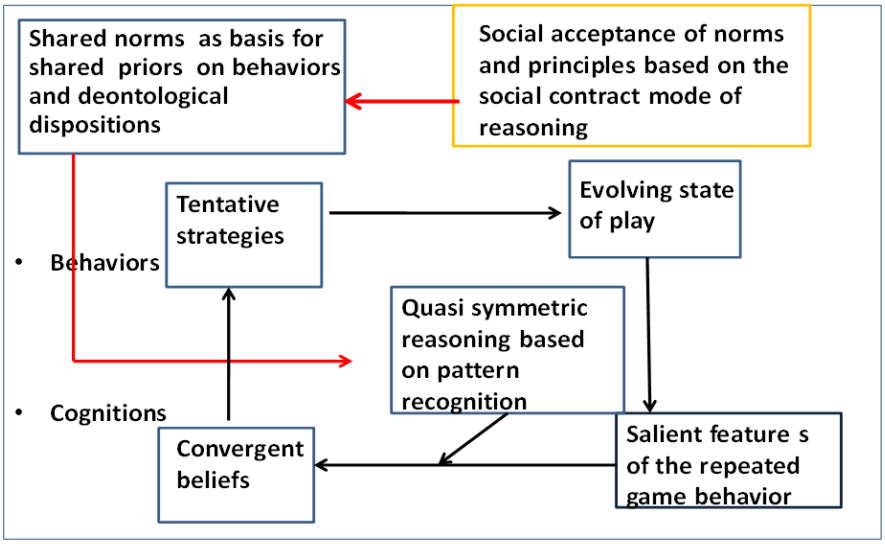

Figure 2 – Aoki's modified diagram representing the recursive process of institution formation

Aoki's recursive model can be reformulated (see *Figure 2*), adding a social norm that derives from social contract reasoning employed by players in order to agree on basic principles and norms when equilibrium institutions are not already established.

[2] In the tradition of social contract theory: see Hobbes, 1651; Buchanan, 1975; Gauthier, 1986; Rawls, 1971.
[3] See for example Kruglanski et al., 2002.
[4] Tomasello et al., 2005.

The normative meaning of norms does not depend on knowledge about the ongoing behavior of other players. Instead, norms are able to justify and give first-place reasons for shared acceptance of a mode of behavior addressing all the participants in a given interaction domain before it has been established as an equilibrium point. A norm gives intentional reasons to act independently on the evolutionary benefits of adaptation in the long run because when an individual or a group of agents in a given action domain initiate an institutional change, it cannot stem from the pressure of evolutionary forces, which unfold their attraction only in the long run. Instead, a norm enters the players' shared mental model (Denzau and North, 1994) of how the game should be played, shapes the players' reciprocal disposition to act and their default beliefs about common behaviors, and hence becomes the basis for their first coordination on a specific equilibrium. In other words, it works as the first move in a process of equilibrium selection that activates the recursive process outlined by Aoki.

The presence of a peculiar cognitive mechanism linked to social contract theory has been studied in cognitive psychology by scholars who have shown how the human mind is evolutionary shaped to reason about social contracts[5]. Other works have developed the model of *conformist preferences*[6], in order to give a formal treatment of the motivating effect of an egalitarian social contract.

Some recent experimental works[7] have also shown how the participation in *ex ante* agreements has a strong impact on beliefs and motivations in the *ex post* behavior in games of cooperation.

3. The Trust Game.

We suggest that a proper game-theoretic representation of the interplay between non-controlling stakeholders and the entrepreneur/manager of a firm is given by the **Trust Game**, illustrated by David Kreps (1990) as a one-sided version of the prisoners' dilemma for the study of authority relations in hierarchical contexts (see also Sacconi, 1997).

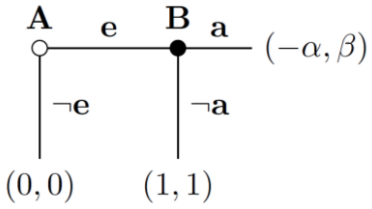

Figure 3 – The Trust Game in Extensive Form

Figure 3 represents the extensive form of this game: Player A (a stakeholder in our case, for example an employee with a specific investment at stake) must choose whether or not to trust player B (the controlling stakeholder B, i.e. the entrepreneur or the manager).

By entering the relationship, the trustor (player A) accepts (trusts) the authority of the trustee (player B). On the contrary, by not entering, he refuses to take a subordinate position in the relationship with B. Moreover by entering A invests idiosyncratically in the relationship.

[5] Cosmides and Tooby, 1989.
[6] Grimalda and Sacconi, 2005.
[7] Sacconi and Faillo, 2010; Sacconi et al., 2011; Faillo et al., 2015.

If he does not elect trust to B (strategy ¬*e*), then they both get nothing (0,0). Otherwise (strategy *e*), B is made aware of A's trust and he has the option to honor it or to abuse it. The trustee is an authority who can abuse some discretionary power. In the first case (strategy ¬*a*), players equitably share the benefits of their relation and they both get a positive payoff (1,1); in the second case (strategy *a*), B takes advantage from A's trust getting a high benefit ($\beta > 1$) and leaves player A with a negative payoff ($-\alpha < 0$). *Figure 4* shows the normal form of the game.

		B	
		a	**¬a**
A	**e**	$-\alpha, \beta$	$1, 1$
	¬e	$0, 0$	$0, 0$

Figure 4 – The Trust Game in Normal Form

As the first player anticipates that B will abuse, he will choose not to enter and payoffs will be zero for both. This is the unique equilibrium outcome of the game, played once and with individuals motivated only by monetary payoffs, and it is clearly inefficient since both players would be better off if A had trusted and B had honored his trust.

As already noted by Kreps (1990, p. 101), "one thing that the two transacting parties might do is to sign at the outset a contract that bins B to honor. Note that ex ante each will willingly sign such a contract as long as it is enforceable, because without it each will net nothing". Such a contract, or agreement, is *internally rational* (Gauthier, 1986, p. 118) since each player would agree on it, but it is not *externally rational*, since for player B it would not be rational to act on that agreement, once A has entered. B will always have incentive to violate the agreement.

Typically, this relation does not take place only once, but it happens many times: the interaction between the firm and stakeholders is better described in terms of a repeated Trust Game, and therefore the usual Folk Theorems for repeated games apply[8]: if players are sufficiently patient, then any feasible, individually rational payoff can be enforced by an equilibrium. Figure 3 represents the payoff space for the repeated game, when $\alpha = 1$; $\beta = 3$. All the points in the gray region are possible equilibria of the repeated game, and we now face a problem of multiplicity and selection among multiple equilibria.

[8] See Fudenberg and Tirole, 1990, chapter 9 for a textbook treatment.

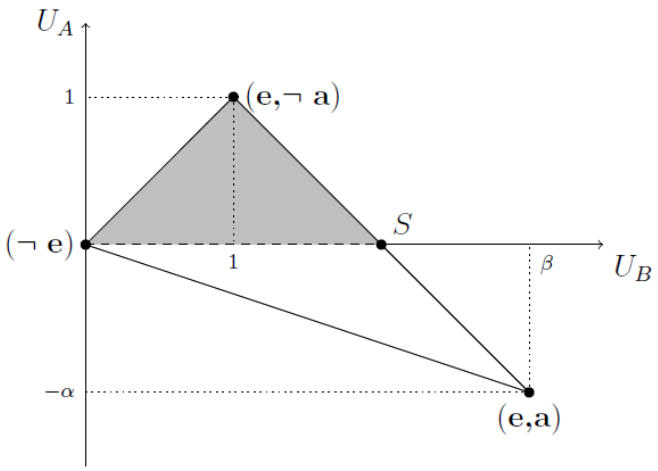

Figure 5 – Feasible and individually rational outcomes in the repeated Trust Game

Given so many equilibria, many possible conventions can emerge from reciprocal coordination. We consider two particular solutions: the *Nash Bargaining Solution (NBS)*, by which B equally shares the surplus, is a social norm of fair treatment where the firm is run to the fair reciprocal advantage of both stakeholders. By contrast, a model of Corporate Governance consistent with a purely shareholder-value maximization approach would justify the equilibrium corresponding to the *Stackelberg solution (S)*.

In fact the equilibrium chosen by the Firm would not be the fully fair and cooperative one, but rather the one whereby the managers acquire a reputation for abusing the trust of stakeholders – but only to the extent that makes them indifferent between maintaining their relations of cooperation and withdrawing from them.

This behavior can be analyzed with the lens of reputation effects (Fudenberg and Levine, 1989, 1992): if we consider this relation not as a repetition among two fixed players (a Firm and a Stakeholder), but among a single long-run player that plays against a sequence of several short-run players who play only once, but observe all previous play, then this setting allows for the long-lived Firm to build a reputation.

An important result in this kind of games is that the Firm can obtain the payoffs he prefers, taking a commitment to repeatedly play a given strategy. In our example of the Trust Game represented in *Figure 5*, the point S in the graph represents the preferred equilibrium by the Firm (the Stackelberg equilibrium), that corresponds to the strategy of abusing and not abusing with a given frequency, that makes the Stakeholder slightly prefer entering than not. This happens when $U_A(e) > U_A(\neg e)$, which requires the probability of abuse p to be less than $1/(1+\alpha)$. If the Firm abuses $[1/(1+\alpha)]+\varepsilon$ times, and does not abuse for the remaining times, than she will be able to get her Stackelberg payoff while convincing the Stakeholder to enter every time.

On the other side, Binmore (2005) has shown that in a context of constitutional choice where agents confront one another in a state of nature, a Rawlsian Social Contract (Rawls, 1971) is able to solve the normative equilibrium selection problem, i.e. to choose a constitutional order through a decision procedure that satisfies elementary conditions of impersonality, impartiality, and empathy.

Application of the Binmore-Rawls theory of equilibrium selection based on the *ex ante* social contract is starkly simple in this case (Sacconi, 2010b): it requires to consider the intersection subset of the original payoff space X_{AB} and its symmetric translation X_{BA} with respect to the player utility axes U_A and U_B. The symmetrical intersection subset line segment (along the bisector) consists of all the egalitarian distributions; adding basic strong Pareto Optimality (i.e., agreeing on solutions that permit mutual improvements for all, if available) directly leads to choosing the equilibrium point consistent with the Nash Bargaining Solution of the original game, which is also its egalitarian (and maximin) solution, as shown in *Figure 6*.

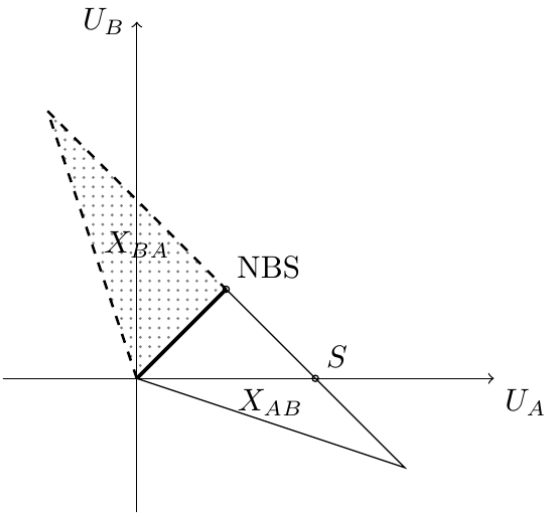

Figure 6 – Application of Binmore's egalitarian solution in the repeated trust game

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 definitive reason to be carried out, and thus the one corresponding to the ex ante agreement need not have any incentive effect on compliance.

Indeed, although an ex ante social contract would be able to justify the choice of a fair equilibrium, ex post we are 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", where they play according to the entire set of their preferences and motivations to act.

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.

We suggest that the drivers of compliance with the norm that has been agreed from an ex ante perspective are of two kinds: on one side, the normative social contract elicits a frame supporting the fair solution also *ex post*. In order to do this, an additional cognitive psychology assumption is needed: because the players have cognitive limitations, they do not consider all the logical possibilities in the *ex post* game, they continue to conceive their interactions within the 'frame' in which they entered when assuming for normative reasons the perspective the original position. In particular, this frame assumes that they are equal and interchangeable and it delimits the information that an agent may

consider as relevant (within the frame). Hence the only information to which the agent pays attention is the subset consistent with the frame itself.

On the other side, the fact that a norm has been agreed upon may affect the motivational force exerted by different equilibria in a game. This leads to a restriction on the number of equilibrium points that have motivational force over the players' behavior. In other words, norms can refine the equilibrium set of a game in terms of the motivational strength of certain equilibria over other equilibria.

4. The role of an agreement on cognitions

The presence of a peculiar cognitive mechanism linked to social contract theory has been studied in cognitive psychology by scholars who have shown how the human mind is evolutionary shaped to reason about social contracts (Cosmides and Tooby, 1989) deriving implications about the structure of the mental algorithms regulating reasoning about the domain of social exchange.

Some recent experimental works (Sacconi and Faillo, 2010; Faillo et al. 2015; Tammi, 2011) have also shown how the participation in ex ante agreements has a strong impact on beliefs and motivations in the ex post behavior in games of cooperation.

In a recent work[9], we show how an agreement among players is able to activate a peculiar way of reasoning that is compatible with bounded rationality and is also able to generate shared representations leading to a fair outcome. At the cognitive level, the role of an agreement beyond the veil of ignorance is to activate a "symmetric" mental model: symmetry becomes focal and only symmetric frames are conceived. In a public discussion under veil of ignorance, the best social outcome would be detected, and the framed game for which it is an equilibrium.

Agents may enter the interaction with their own frame in mind, due to previous interactions, transfer mechanisms, education, prototypes and many other factors. But the social contract might play a key role in this. The precondition for activating a symmetric frame can be envisaged in the Rawlsian veil of ignorance, a procedure that allows to enrich the mental representations by focusing on impartiality and impersonality.

This role of the social contract is explained through its main characteristics: impersonality, impartiality and prescriptivity.

The first step is the application of the principle of impersonality, which is able to broaden the number of strategies that are taken into consideration: for any conceivable action, this might be thought as possible for any player – we have an expansion of the considered subset of strategies. Nonetheless, this allows agents to create a summary representation of the game that considers the diagonal, because their cognitive frame, acquired through the veil of ignorance, forces them to adjust their model in order to admit symmetry in actions.

Then, prescriptivity comes into the picture. When an agent is in front of a very big set of strategies, instead of focusing on the ones he already used before in his previous interactions (a kind of path dependence in mental models), he might partition this big set of actions into subsets, that we will call categories. The steps involved in this process are three: first, divide the space into small subsets;

[9] Cecchini Manara and Sacconi, 2019.

second, give a label to each of them; and third, choose on which to concentrate. Categorization can be driven by different factors, and we want to suggest that social contract reasoning might induce a categorization driven by fairness considerations (Rawls, 1971). The prescriptivity of social contract reasoning might help agents in choosing a particular subset of actions that can be described (or labeled) as fair for the joint production in the context of a social contract under veil of ignorance.

5. The role of an agreement on motivations

An agreement on a norm does not only contribute to create a focal point and a shared mental model, as suggested in the previous section, but it also adds a motivational force, because it helps to categorize the available actions in a certain way adding a value for the agent.

The model explained and applied in this section has been developed in order to give a formal treatment of this motivating effect: the key elements of Rawls's analysis have been incorporated into the Conformist Preference Model (Grimalda and Sacconi, 2005), grounded on the literature on psychological games (Geanakoplos, Pearce and Stacchetti, 1989) and reciprocity (Rabin, 1993). The main features of the model are summarized in the Appendix.

According to this model, a player characterized by conformity preferences complies with an agreement on a principle that dictates a choice in contrast with his self-interest if:

i) he participates in the *ex ante* agreement on the principle,
ii) he expects that other players who have contributed to choosing the principle will comply,
iii) he expects that others will expect that he will comply.

The behavioral hypotheses underlying this model have been tested through several experiments that provide support to the theory.

Starting off with the standard Trust Game described above, where the strategic interaction is specified only in terms of material payoffs, we derive a *psychological game* in which the payoff functions are redefined so as to reflect an intrinsic component for norm compliance (in particular it reflects the Rawlsian sense of justice), which is not unilateral and unconditioned, but depends on players' beliefs about expected and reciprocal conformity.

To begin, let us illustrate the conformist preference model with reference to its application to the one shot Trust Game.

To calculate conformist psychological payoffs and equilibria, let's consider the game matrix of the stage game, reported in *Figure 4*. There are four possible states of affairs σ coinciding with the cells of the normal form matrix: $(\neg e, \neg a)$, $(\neg e, a)$ with material payoffs (0,0); (e, a) with material payoffs (-α,β); and $(e, \neg a)$, with material payoffs (1,1).

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 we mean a principle of justice for the distribution of material utilities coinciding with the *ex ante* social contract.

Let us assume that players have agreed on a social contract concerning the principle of justice that should govern cooperation in society and that it prescribes to apply 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 – (0,0) – which is the assurance level that player *A* can grant herself for whatever player B's choice, included the case that he doesn't start 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: the Nash bargaining product calculated for each pure strategy combination needed to measure the consistency of each state with respect to the principle T and the players' relevant degrees of conditional and expected reciprocal conformity for each state (*Figure 7*), and the overall pay-offs resulting from the addition of the psychological conformist preference to the material pay-offs where this addition is appropriate (*Figure 8*).

σ	T
(e, a)	$-\alpha\beta$
$(e, \neg a)$	1
$(\neg e, a)$	0
$(\neg e, \neg a)$	0

Figure 7 – T values for every strategy combination in the TG

		B	
		a	**¬a**
A	**e**	$-\alpha, \beta$	$1+\lambda, 1+\lambda$
	¬e	λ, λ	$0, 0$

Figure 8 – Payoffs including conformist preferences

Summing up, the normal form of the game is modified as in Figure 4, and the combination of strategies where the Stakeholder enters and the Firm does not abuse becomes an equilibrium of the game, provided that the motivational weight is high enough to counterbalance the temptation of the Firm to abuse. More precisely, what is required is that the material utility from cooperation (that in the proposed representation was equal to 1) with the addition of the psychological parameter λ is bigger than the benefits that the Firm would obtain by abusing.

Proposition 1. In the one-shot Trust Game where players have conformist preferences with λ>β-1, the following strategy profiles can be supported as Nash equilibria of the game: $(e, \neg a)$ and $(\neg e, a)$.

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 unique. In the following part, we analyze the impact of conformity preferences on mixed strategies, and then apply the results to the repeated Trust Game, where also many standard Nash

equilibria are possible, but we will show that conformist preferences with an *ex ante* agreed principle of justice will simplify the equilibrium selection problem.

Now let us consider the repeated Trust Game. 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. 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 that employs alternatively the two player's *i* pure strategies of the stage game with a given frequency, generating the three stage-game outcomes, 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 each of the three stage-game payoffs a given number of times out of the total number of times defining the cycle (granted, of course, that during the game each repeated strategy pairs used by any player repeatedly enters a cycle with the same pattern of outcomes and the same average payoff value for the player that adopts it). It is thus simple to see that a Firm's mixed strategy that employs the two pure strategies ***a*** and ***¬a*** with probability *p* and *(1-p)*, respectively, against – to keep things simple – the stakeholder's pure entry strategy ***e***, is equal to the average values attached to a repeated strategy whereby the Firm plays the stage-game strategy ***a*** *p* per cent of the time and the stage-game strategy ***¬a*** (1-p) 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 repeated psychological Trust Game in pure and mixed strategies has the same material pay-off space as the repeated 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 (under the psychological extension) to the mixed strategy equilibrium points of the corresponding standard repeated Trust Game. In the case of the repeated game, the psychological payoff space is partly translated, because of psychological utility deriving from conformity to the ideal, as shown in Figures *Figure 9, Figure 10* and *Figure 11*.

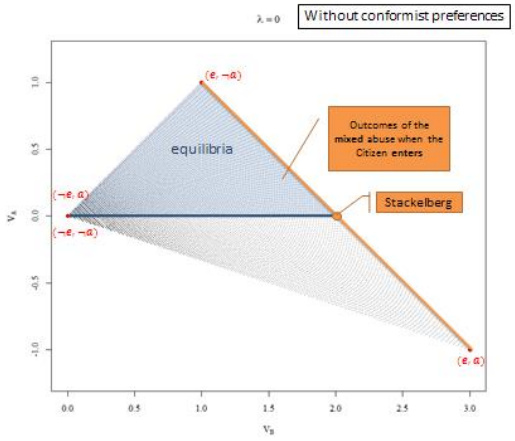

Figure 9

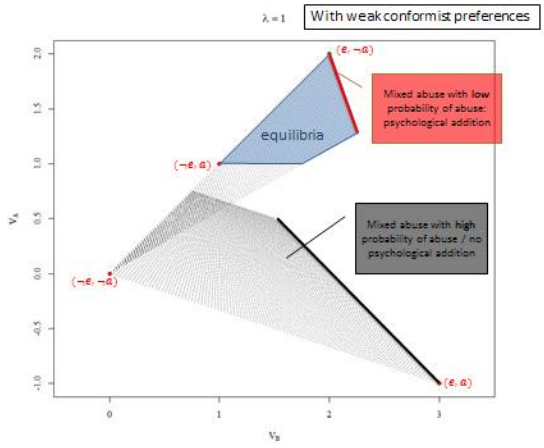

Figure 10

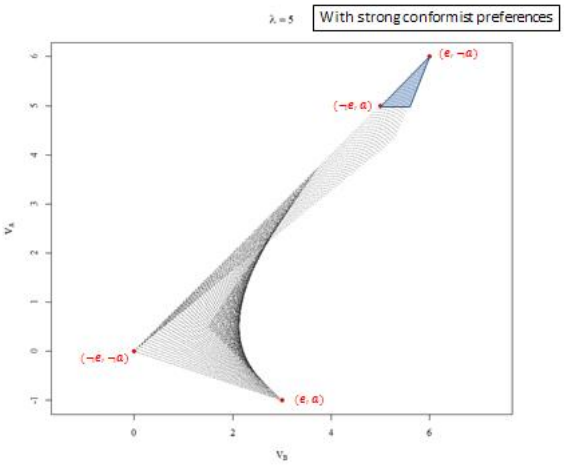

Figure 11

The main results can be summarized in the following statements:

Proposition 2. In the Repeated Game where the Firm is allowed to mix over his pure strategies while the Stakeholder only plays pure strategies, when players have conformist preferences

with λ>β-1, the Stackelberg Equilibrium is ruled out as it does not gain any psychological utility component.

Moreover, in most cases A's "giving in" is not a best reply to a mixed strategy. The threshold that allows mixed strategies to gain some positive support from psychological conformist utility is reached at the mixed strategy p=p*. Here the expected value of T is zero for every A's choice, so A is equally conformist by choosing either *e* or ¬*e*. By playing the mixed strategy player B is partially conformist, because given A's entrance the T value would be minimized by playing *a*. but if A stays out, player B is turned to be completely conformist, as no unconformity could be ascribed to him. Thus adding just a bit of psychological utility (due to B's partial reciprocal conformity) does not mean that B's mixed strategy induces "enter" as A's psychological best response The player A's overall payoff gained from ***¬¬e*** is still higher than the overall payoff from giving in to player B's mixed strategy. The reason is that B's conformity is higher when A stays out, while A's conformity is the same in both cases (*a* and ¬*a)* and the *overall index of conformity* is higher with ***¬¬e***.

In general the B's most valuable best reply to "giving in" is no abuse: assume player B has been able to accumulate a reputation that for the first time induces player A to enter, and assume that player A enter strategy is unconditional, then he (B) immediately would recognize the incentive to switch to a strategy that employs the strategy ¬*a* with the highest frequency. It follows that player B's best reply to player's A entry is to switch from any mixed strategy to ¬*a*.

Proposition 3. Given a repeated TG with pure and mixed strategies, whereby a psychological game with conformist preferences is defined, so that the motivational exogenous parameter λ is great enough to guarantee the existence of a pure strategies psychological equilibrium in correspondence to (*e*, ¬*a*), then: there are still two psychological equilibria in repeated pure strategies (*e*, ¬*a*) and (¬e, *a*), but could player B act as a "psychological leader", he would not choose the typical Stackelberg equilibrium but the equilibrium mutually preferred by both the players (*e* | *conditional on never abusing)* , ¬*a* | *conditional on always entering*)

All the remaining mixed strategy equilibria are Pareto ranked, and hence dominated by the single psychological equilibrium (***e***, ¬***a***) the equilibrium set is substantially reduced with respect to the non psychological case. As shown in *Figure 12*, the set of admissible equilibria changes for growing values of λ and it tends to shrinks as the parameter λ increase, reducing in the limit to the point (***e***, ¬***a***), which is the Pareto dominating equilibria.

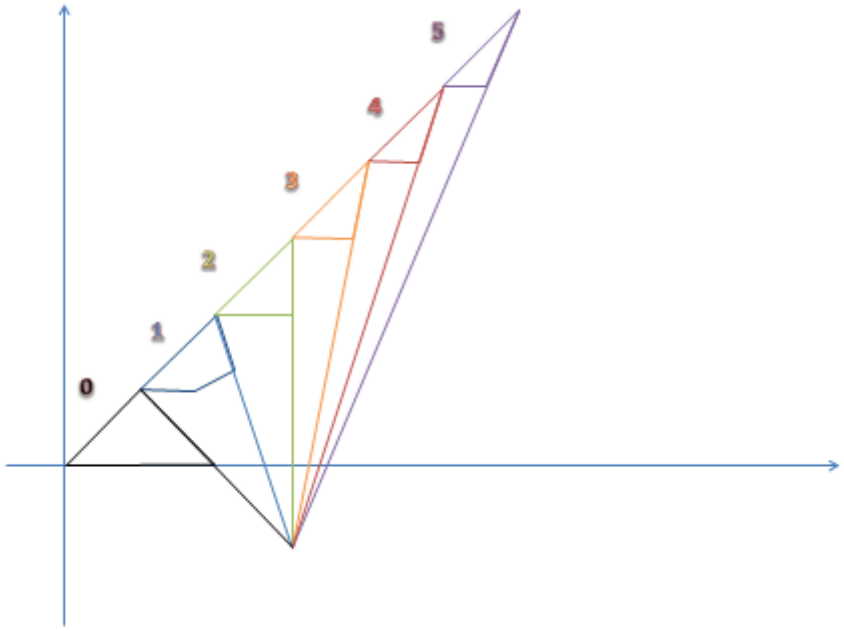

Figure 12 – Sets of admissible equilibria for growing values of λ

6. Conclusions and application to the realm of a "revised social responsibility of business"

In this work we have considered mechanism for norm compliance alternative to reputation: conformity and reciprocity that derive from an impartial agreement among stakeholders. We analysed the role of an agreement on cognitions and motivations, grounding on insights from psychology, game theory and experimental findings.

The main result of this paper is that an egalitarian social contract matters because it shapes preferences. If players have agreed rationally on a certain constitution, this affects their preferences, admitted that they expect reciprocity from others in complying with the same institution. Under expectations of reciprocity, preferences incorporate a desire to conform with agreed principles, which is a direct function of the extent to which an agent is conforming given his expectation of others' behavior, and the extent to which he expects reciprocal conformity by others given their belief in his own action. Impartial agreement on a principle lays the bases; then mutual expectations are essential in shaping a preference for conformity. Hence the social contract eminently affects our preference for compliance; and the role for normativity of the social contract is adequately rescued.

7. Appendix – The Conformist Preference Model

Players have *two* kinds of preferences defined over states of affairs resulting from 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.* 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 we mean a principle of justice for the distribution of material utilities coinciding with the *ex ante* social contract.

Let us assume that players have agreed on a social contract concerning the principle of justice that should govern cooperation in society. 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. 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

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

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 a situation σ where material utilities are distributed in such a way that they are mostly consistent with the distributive principle T within the available set of 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 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:

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

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 i's conditional conformity index:

$$1 + f_i(\sigma_{ik}, b_i^1) \tag{3}$$

This index takes its values as a function of $\boldsymbol{f_i}$ 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 i's deviation degree:

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

where b_i^1 is player i's belief concerning player j's action, $T^{MAX}(b_i^1)$ 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, $T^{MIN}(b_i^1)$ 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 $T(\sigma_{ik}, b_i^1)$ 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 is expected to contribute to a fair payoff distribution in terms of the principle T, given what he is expected to expect 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 j's reciprocal expected conformity index:

$$1 + \tilde{f}_j(b_i^2, b_i^1) \tag{5}$$

In fact it is as well a function of $\tilde{f}_j$, 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 deviation degree:

$$\tilde{f}_i(b_i^1, b_i^2) = \frac{T(b_i^1, b_i^2) - T^{MAX}(b_i^2)}{T^{MAX}(b_i^2) - T^{MIN}(b_i^2)} \tag{6}$$

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 believes about the action adopted by player i, while $T^{MAX}(b_i^2)$ and $T^{MIN}(b_i^2)$ are defined as above but in relation to player i's second order belief.

Fourth, there is an exogenous parameter λ ($\lambda \geq 0$) 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 parameterλ, 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

$$\lambda_i\left[1 + \tilde{f}_j(b_i^2, b_i^1)\right][1 + f_i(\sigma_{ik}, b_i^1)] \qquad (7)$$

which reduces to the following cases:

$$\lambda[(1-x)(1-y)] = \lambda$$

i) $\lambda[(1-x)(1-y)] = \lambda$, 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) $\lambda[(1-x)(1-y)] = \alpha\lambda < \lambda$, where $\alpha < 1$ since $0 < x < -1$ and/or $0 < y < -1$, if player i partially deviates and/or expects player j to partially deviate from complete conformity;

iii) $\lambda[(1-x)(1-y)] = \alpha\lambda = 0$, since 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.

8. References

Aoki, M. (2001) Toward a Comparative Institutional Analysis, Cambridge, MA: MIT Press.

Aoki, M. (2011), "Institutions as Cognitive Media between Strategic Interactions and Individual Beliefs", Journal of Economic Behavior & Organization, 79(1–2), pp. 20-34.

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

Blair, M., and L. Stout (1999) 'A Team Production Theory of Corporate Law', Virginia Law Review, vol. 85, pp. 247–328.

Buchanan, J. (1975), The Limits of Liberty: Between Anarchy and Leviathan, Library of Economics and Liberty [Online] available from http://www.econlib.org/library/Buchanan/buchCv7.html.

Cecchini Manara, V. and Sacconi, L. (2019), Institutions, Frames and Social Contract Reasoning, mimeo.

Tooby, J. & Cosmides, L. (1989). Evolutionary psychology and the generation of culture, Part I. Theoretical considerations. Ethology & Sociobiology, 10, 29-49.

Denzau and North, 1994) Denzau, A. and North, D. (1994), "Shared Mental Models: Ideologies and Institutions", Kyklos, 47(1), pp. 3-31.

Faillo M. Sacconi, L. and Ottone, S. (2015), The social contract in the laboratory. An experimental analysis of self-enforcing impartial agreements, Public Choice (2015), 163:225–246.

Freeman et al. 2010 Freeman, R. E., Harrison, J. S., Wicks, A. C., Parmar, B. L., & De Colle, S. (2010). Stakeholder theory: The state of the art. Cambridge University Press.

Fudenberg, D. and D.K. Levine (1989) 'Reputation and Equilibrium Selection in Games with a Patient Player', Econometrica, vol. 57(4), 759–78.

Fudenberg, D. and D.K. Levine (1992) 'Maintaining Reputation when Strategies are Imperfectly Observed', Review of Economic Studies, vol. 59, 561–79.

Fudenberg D. and Jean Tirole (1991) Game Theory. Cambridge, MA: MIT Press.

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

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. 249–276.

Hobbes, T. (1651). Leviathan, ed. CB Macpherson. In London: Pengui.

Kreps, D. (1990) 'Corporate Culture and Economic Theory', in J.E. Alt and K.A. Shepsle (eds), Perspectives on Positive Political Economy. Cambridge: Cambridge University Press.

Kruglanski, A. W., Shah, J. Y., Friedman, R., Fishbach, A., Chun, W. Y., & Sleeth-Keppler, D. 2002. A theory of goal systems. Advances in Experimental Social Psychology, 34: 331–378.

Rabin, M. (1993) 'Incorporating Fairness into Game Theory', American Economic Review, 83 (5), pp. 1281–1302.

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

Sacconi, L. (2012), "Ethics, Economic Organisation and the Social Contract", EconomEtica working paper, 41

Sacconi L. and M. Faillo (2010) 'Conformity, Reciprocity and the Sense of Justice. How Social Contract-based Preferences and Beliefs Explain Norm Compliance: the Experimental Evidence', Constitutional Political Economy, 21 (2), pp. 171–201.

Sacconi, L., Blair M., Freeman E. and Vercelli, A. (2011), Corporate social responsibility and corporate governance, Palgrave Macmillan.

Sacconi, L. (1997) Economia, etica, organizzazione. Bari: Laterza.

Sacconi, L. (2010a), "A Rawlsian View of CSR and the Game Theory of Its Implementation (Part I): The Multistakeholder Model of Corporate Governance." In Lorenzo Sacconi, Margaret Blair, R. Edward Freeman, and Alessandro Vercelli, eds., Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines, 157□193. Basingstoke: Palgrave Macmillan.

Sacconi, L. (2010b), "A Rawlsian View of CSR and the Game Theory of Its Implementation (Part II): Fairness and Equilibrium." In Lorenzo Sacconi, Margaret Blair, R. Edward Freeman, and Alessandro Vercelli, eds., Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines. 194□125. Basingstoke: Palgrave Macmillan.

Sacconi L. (2010c), "A Rawlsian View of CRS and the Game of its Implementation (Part III): Conformism and Equilibrium Selection", in L. Sacconi and G. Degli Antoni (ed.), Social Capital, Corporate Social Responsibility, Economic Behavior and Performance" Palgrave London

Tammi, T. (2011). Contractual preferences and moral biases: social identity and procedural fairness in the exclusion game experiment. Constitutional Political Economy, 22(4), 737-397.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll H. (2005), Understanding and sharing intentions: The origins of cultural cognition, Behavioral and Brain Sciences, 28, pp. 675-735.