Journal of Business Ethics (2007) 75:77–96
DOI 10.1007/s10551-006-9239-6

A Social Contract Account for CSR as an Extended Model of Corporate Governance (II): Compliance, Reputation and Reciprocity

Lorenzo Sacconi

ABSTRACT. This essay seeks to give a contractarian foundation to the concept of Corporate Social Responsibility (CSR), meant as an extended model of corporate governance of the firm. Whereas, justificatory issues have been discussed in a related paper (Sacconi, L.: 2006b, this journal), in this essay I focus on the implementation of and compliance with this normative model. The theory of reputation games, with reference to the basic game of trust, is introduced in order to make sense of self-regulation as a way to implement the social contract on the multi-fiduciary model of corporate governance. This affords understanding of why self-regulation, meant as mere recourse to a long-run strategy in a repeated trust game, fails. Two basic problems for the functioning of the reputation mechanism are examined: the cognitive fragility problem, and the motivational problem. As regards the cognitive fragilities of reputation (which result from the impact of unforeseen contingencies and from bounded rationality), the paper develops the logic and the structure that self-regulatory norms must satisfy if they are to serve as gap-filling tools with which to remedy cognitive limitations in the reputation mechanism. The motivation problem then arises from the possibility of sophisticated abuse by the firm. Developed in this case is an entirely new application of the theory of conformism-and-reciprocity-based preferences, the result of which is that the stakeholders refuse to acquiesce to sophisticated abuse on the part of the firm.

KEY WORDS: self-regulation, ethical norms, reputation games, unforeseen contingencies, fuzzy logic and default reasoning, reciprocity and fairness, conformist preferences

Lorenzo Sacconi is professor of economics and Unicredit Chair in economic ethics and corporate social responsibility at the Department of Economics of the University of Trento, where he leads the LaSER - Laboratory of research in Social responsibility, Ethics and Rationality, and head of the graduate program (laurea magistralis) in ''economic decisions, enterprise and corporate social responsibility''. He is also director of EconomEtica, the interuniversity centre for economic ethics and corporate social responsibility joining over 20 Italian Universities placed at the Milano-Bicocca University. Past president of tthe Italian Business Ethics Network and past member of the EBEN executive committtee, currrently he is a member of the executive committee of the Italian chapter of EBEN (EBEN Italy). On related subjects, he is author of the book: The social contract of the firm, Springer, 2000

The compliance problem

In a paper related to this essay (see Sacconi, 2006b) I have provided a contractarian justification for the Corporate Social Responsibility (CSR) model of corporate governance. Here I move from justification to implementation of the model – that is, to the problem of why the social contract on the CSR model of corporate governance should be complied with. The question asked is whether the social contract is also able to produce endogenous incentives and motivations, which may be strong enough to induce individual behaviour to conform with the normative model of extended fiduciary duties. In other words, the question is whether the institutional model of the firm governed in the interests of all its stakeholders may be self-enforcing in the sense that compliance with it does not have to be enforced by an external authority – the authority of the law – or at most requires only a mild external imposition which can be considered residual.

Justifications in themselves do not answer questions about compliance with and implementation of a CSR normative model of corporate governance. This is because the agent's standpoint in the justification context is neutral i.e. detached from the particular personal perspective of each concrete

agent (be this an individual or an artificial actor like the corporation or its board of directors). In the implementation context, reasons for action are instead agent-relative (Nagel, 1986). They reflect intentions, motivational drives and preferences which the agent holds simply because he is *that* particular agent in *that* particular decision position. This simple condition of realism suggests that the effectiveness of a norm consists in that, by implementing the norm, the agent will also pursue his preferences in a rational manner (in the sense of coherence amongst preferences and between preferences and actions). It admits both the view that complying can be a means to fulfil preferences (instrumental view) and that the norm itself may influence preference formation (intrinsic view).

Taking a contractarian standpoint in the justificatory domain, of course, greatly simplifies accomplishment of the implementation task. Impartiality within a contractarian framework, in fact, amounts to no more than a condition of invariant individual rational acceptance of a given bargaining outcome (under the permutation of personal standpoints allowing the impartial decision-maker to take each player's point of view in turn). Thus impartiality is no more than invariance in a class of agent-relative reasons for action.

Nevertheless, the bulk of the task is still to be accomplished. Implementation is the typical sphere where non-cooperative games are relevant and *ex post* rationality is required; whereas invariance in individual decisions to accept a norm imposing a joint strategy on all the players concerns *ex ante* rationality alone. In the implementation stage, instead, separate but interdependent strategies are under consideration, and the players are always able to say whether or not they want to implement the joint strategy. It follows that the main problem to be solved in the implementation context is how a CSR norm subscribed voluntarily can also generate motivational causal forces strong enough to induce the execution of the norm in situations where it may require a counter-interested behaviour of the agent at least in the immediate term. Clearly, this would be the case of corporate directors, managers or proprietors were a CSR model of governance to require – as it is likely to do – sharing of the firm's rent or surplus with other stakeholders.

In the long-standing debate on the relationship between rationality and morality, some authors have sought to revise the notion of instrumental rationality to include rational choice of dispositions.[1] A disposition would constrain later choices, so that the agent can disregard local incentives even if these imply that there are local advantages too in deviating from the action plan corresponding to the disposition. Given that the disposition allows the agent to abide by a plan disregarding local incentives to deviate from it, one can show that having a disposition which corresponds to a conditional cooperation plan is beneficial. It enables the decision-maker to gain higher overall utility when he meets (and recognises) another symmetrically disposed decision-maker, whereas his utility equates that of a non-disposed agent when such a non-disposed agent is met (and recognised). Since this is the case in general, any rational actor should decide on the basis of an instrumental rationality calculation to undertake the disposition which enables him to abide by a plan of conditional cooperation which also allows for locally counter-interested actions.

I will not follow this line of reasoning, however. This revision of the notion of instrumental rationality seems in fact to presuppose what it should demonstrate. This approach seeks to reduce morality to rationality by showing that abiding by a moral norm is rational. But in doing so it must presume that moral dispositions are 'out there' and endowed with all their disciplining force independently of rational choice. And whilst dispositions are taken to be choices at our disposal – we can decide whether to develop them or not – they are also presumed to command our later behaviours, being immune from opportunistic changes when these seem profitable, as if these choices were beyond our control.

The most natural reply to the question concerning compatibility between compliance with a norm (for example a CSR ethical norm for the governance and management of a firm) and the rational pursuit of personal preferences is therefore still the one based on reputation. Reputation – seen as a means to gain personal advantage – is an incentive in so far as it is instrumental to trust relationships between the firm and its stakeholders conducive to better and low-cost preferences-fulfilling transactions. Conformity with a norm which *per se* is not conducive to personal interest proves to be in the agent's best interest because it affects reputation that fosters trust, and this makes mutually beneficial transactions possible. This

opens the way for a solution to the compliance problem in terms of equilibrium analysis. To conform to the social contract, or to a norm derived from a hypothetical social contract amongst the firm's stakeholders, to each player in the implementation game (those in a position of authority and the non-controlling stakeholders) consists of nothing more than implementing his or her reciprocal best-response strategies, given his or her conjecture concerning the other player's behaviour (different authors understand this way social norms, see for example Lewis (1969) Posner (2000) Binmore (2005) but see also the definition given by Petit (1990), including not only common knowledge of the strategies but also common knowledge of normative acceptance). Answering the question about the effectiveness of a norm thus requires careful consideration of the conditions under which the reputation mechanism can work properly. I therefore divide the problems to be addressed within the implementation domain into two subclasses concerning how reputation effects can be effective in the field of business ethics.

1. *The cognitive problem*: Economic agents are endowed with bounded rationality (Kreps 1998). Hence the supposition that reputation may depend on commitments defined conditionally on any possible state of the world is unrealistic. Reputation can be obstructed if the firm does not know how to make itself recognisable, or when it does not know which benchmark can be used to appraise its honest behaviour when unforeseen contingencies emerge such that traditional commitments are mute. It is here that the cognitive role of explicit yet voluntary business ethics norms, such as a code of ethics or a CSR management standard, comes into play. From the implementation standpoint, answering the question about the cognitive role of explicit business ethics norms within the reputation mechanism also meets the criticism of lack of prescriptivism and univocality (Jensen, 2001): it would be no longer true that a boundedly rational manager can resort only to 'shareholders' value maximisation' because of the simplicity of the rule. Complying with a CSR governance and management standard may be much more consistent with Simon's view of procedural rationality.
2. *The motivation problem*: Once the cognitive problem has solved, reputation will activate incentives to comply with a voluntary norm prescribing the CSR model of corporate governance. No exception is needed to the standard model of selfish economic man. However, reputations can be of many kinds. A company endowed with strong market power, and which establishes idiosyncratic relationships with its stakeholders, may develop a reputation for abusing the trust of its employees, customers, suppliers, and capital-lenders only to the extent that they are indifferent between maintaining their relations with the firm and withdrawing from them. Then a company, by making a minor concession to the stakeholders, may attempt to acquire their acquiescence to its substantive non-compliance. This is not the case in practice, however. Stakeholders, or at least those who engage in stakeholder activism, refuse to acquiesce and actively countervail hypocritical corporate conduct. How does the social contract approach account for these apparently irrational and unselfish actions? Recent behavioural theories of the economic agent's motivational complexity offer interesting explanations, concentrating variously on intrinsic value, social preferences and inequity version, reciprocity and intentional kindness.[2] In this essay, however, I assume a related but original view of *deontological motivations* closely connected with the idea that motivations are driven by coherence with a principle or ideal. This is a contractarian view of conformist preferences and reciprocity in that non-selfish utilities derive from the desire to conform with an ideal of fairness, assuming that this ideal can be derived from a hypothetical contract, and that the other parties to the social contract are also expected to reciprocate conformity with the same ideal of fairness.

The article proceeds as follows. The next section briefly discusses the alternative between legal enforcement and self-regulation. The third section 'The reputation mechanism and its fragilities' intro-

duces the reputation game, which is the basic piece of game theory needed to make sense of the preceding discussion of implementation, compliance and self-regulation. It is also necessary if the cognitive fragilities of the reputation mechanism and the impact upon it of unforeseen contingencies are to be properly understood. The fourth section elaborates on the logic and structure that self-regulatory norms must possess if they are to serve as gap-filling tools to remedy the cognitive limitations to the reputation mechanism. This section outlines a reputation mechanism grounded on an ethical decision procedure employing fuzzy logic and default reasoning. Finally, the fifth section develops an entirely new application of the theory of conformist preferences to the problem of the motivational role performed by business ethics norms in activating stakeholder activism. This section answers the 'real life' question raised in point (2) of this introduction – at the price, however, of the more technical language of the section. It was impossible, in fact, to assume conformist preferences as known (hence *Appendix 1* sets out the mathematical model) and, moreover, the result derives straightforwardly from calculation of stakeholders' overall utilities when the hypotheses of conformism and reciprocity are introduced.

Views of self-regulation

From the practical standpoint – at the cost of forgoing contractarianism as a self-contained theory – I could claim that I am not in need to provide an equilibrium solution of the compliance problem, given that whatever is justified by the normative model can also be made enforceable by law (which in this case would originate externally to the players interacting in the social contract model). This is not the case however. Although welfare state regulations, labour-market laws and environmental regulations establish a general legal framework, they cannot regulate every detail of firms' decisions. They may lay down compulsory conditions, but in many settings their application requires interpretation of a 'grey' zone; or else compliance with them may not be observable. Moreover, even when management decisions closely affect stakeholders, the law cannot regulate those decisions in every respect: the decision whether or not to restructure or downsize a firm is always a business decision notwithstanding the requirements of the law in regard to the protection of third parties or employees. Regulations intended to dictate how such decisions should be taken in every circumstance would inevitably be inefficient (as demonstrated by command economies). Nor do contracts provide a solution. Indeed, it is the fact that contracts are generally incomplete which has prompted the current discussion of CSR. On my hypothesis, the social contract amongst the stakeholders is the 'hypothetical contract' which furnishes default or 'gap-filling' rules which complete incomplete contracts (Coleman, 1992). The obligation to fulfil explicit but incomplete contracts hence does not guarantee compliance with the obligations deriving from the hypothetical contract, and this precisely because the explicit contract does not cover matters that the hypothetical contract helps to clarify.

This brings us to the role of self-regulation see (Capzggi (2006) in print). We may distinguish between two approaches to CSR self-regulation. The first I call the *discretionary approach* and it is discussed in the Section 'The reputation mechanism and its fragilities'. I call it *unfit* in so far as it pertains to the mere sphere of consequentialist expected utility maximising choice of predictable (or at least foreseeable) actions and consequences. It does not imply any explicit principle or rule-constrained behaviour in general, but at most a commitment to a given strategy. Its basic tenet is that there is no reason to add any further specification or constraint apart from the *enlightened* self-interest of those who run the firm for their own advantage. Respect for the stakeholders' claims will come about through free choice, or through the firm's free exercise of discretion. Enlightened self-interest would thus be an endogenous force able to induce self-discipline once it were assumed that it requires acting for one's own personal interest in the *long run*. By virtue of the long run, the firm, as it pursues the simple goal of profit maximisation, is induced to respect the fiduciary relation with the stakeholders and make due consideration of their well-being. On this view, self-regulation is nothing more than *the consistent adherence to a strategy* whereby the firm does not behave in a manner such to abuse the trust that stakeholders have placed in it. The firm does not self-impose any formal system of rules or adopt any explicit management system imposing compliance with standards

or norms, even if voluntary – this is self-discipline *without* explicit rules.

It would be too easy to discredit this thesis by saying that the only argument in its favour is the one based on the obviously unrealistic hypothesis of perfect competition and the 'invisible hand' of the market. If competitions were perfect, no firms would exist, in the sense that they are alternative institutions of governance arisen in order to minimise transaction costs. It is therefore obvious that the argument cannot rest on the ideal world in which the 'invisible hand' operates. I shall instead take this thesis at its best, although even in this case it fails.

At best, the thesis maintains that enlightened self-interest induces respect for the trust of stakeholders, and therefore prevents their abuse, in that the firm recognises the importance of safeguarding and enhancing its reputation, which depends on non-abuse of the stakeholders. Reputation is one of the most valuable, albeit intangible, of the firm's assets. It is reputation that induces the stakeholders to trust the firm and consequently to cooperate with it, so that transactions come about at low costs of control or bargaining. Moreover, reputation activates some sort of self-fulfilling virtuous circle which leads to spontaneous compliance with social norms – that is, it makes them self-enforcing. In fact, compliance with the norms *creates* reputation; reputation *induces* a cooperative response from the stakeholders; those who abide by the norms are thus *offered* a benefit; and this benefit acts as an *incentive* for complying with the norms. Unfortunately, however, reputation does not support the discretionary approach to self-regulation, one that, without imposing any constraint or explicit rule upon the firm behaviour, would allow it to decide unilaterally what actions to undertake. To understand this, we must delve at least to a certain depth into the theory of reputation games.[3]

The reputation mechanism and its fragilities

I will illustrate the reputation mechanism by means of a simple interactive situation (called *the trust game*, see Figure 1) representing a transaction based on the fiduciary relation between a stakeholder A and the firm B.

The stakeholder must decide whether not to place his trust in the firm by entering or otherwise to trust

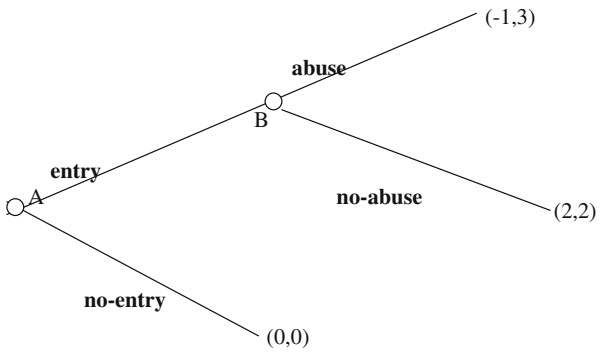

Figure 1. The trust game.

and having an exchange relation with it (assume that if he does so, he must necessarily make a specific investment (which costs -1)). The firm then decides between abusing and not abusing. If, after the stakeholder has entered the exchange relation, the firm does not abuse his trust, there will be a reasonably good outcome for both of them. However, if the stakeholder places trust in the firm, the latter has an interest in abusing that trust, because in the current game this is the most remunerative (dominant) option. Consequently, the stakeholder will not grant his trust and the transaction will not take place.

The underlying idea of a reputation game is that there is an alternative solution which permits the transaction between the two parties to take place if the basic game is infinitely repeated, and if an incentive is thus created for the firm to protect its reputation (Fudenberg and Levine, 1989). We thus have an infinitely repeated game (which expresses the long-run idea) whose stage-game is the trust game already defined. The players in the game are on the one hand an infinite series of stakeholders, called A_i (where i denotes the order of entry into the game), each of them lasting only for the stage-game in which they decide whether to enter or not to enter (and are therefore short-run players), and on the other a firm (B), the long-run player, which lasts throughout all repetitions of the game.

Information is crucial for the players: each A_i is uncertain about the *type* of B – in that B may be a *type* that never abuses trust or a *type* that always abuses it, or even a *type* that abuses with a certain probability and does not abuse with the residual probability. Let us assume that the various stakeholders believe that there exist only *pure types*, which

always abuse or do not abuse trust (and at most only two types of minimum deviations from these pure types, which represent the possibility of a *mistake* by the pure types). Hence the *types* can be understood as commitments to the stereotyped use of a given basic game strategy (unless a mistake occurs). For every player A_i, all these types of B have some *a priori* positive probability (and, in particular, the type that never abuses – which for simplicity I may call 'honest' – is assigned positive, though very low probability). At each stage of the game the current A_i player changes his beliefs (the probability assigned to types) according to what he has learned from the previous stage of game. In each of these stages, the conditioned probabilities of the types are updated on the basis of the evidence furnished by the manner in which B has played the previous game.

Player B's reputation is the probability assigned by each player A_i at the current stage to the various types of player B. Player B's reputation of being a certain type increases as evidence is gathered which confirms that *type* (the probability of a *type* increases with the observation of those actions whose likelihood is positive, given the type), but it diminishes dramatically if a single observation is made that falsifies the type (if, at any stage, an action by B whose likelihood given its type is 0 is observed, the conditioned probability of that type is nullified as well). Player B, on the other hand, is perfectly rational and informed, so that his strategic reasoning scheme also includes awareness of the limitedly informed reasoning performed by players A_i. In particular, his reasoning scheme enables player B to predict about the mechanism by which players A_i learn and update beliefs about B's types. The players' interests are such that each stakeholder A_i maximises its benefits in the current game (i.e. is short-sighted) while B is interested in long-run benefit. B may therefore be more or less far-sighted on the basis of a discount rate on future utilities which, in each period, increasingly reduces (though at a marginally decreasing rate) the payoffs associated with further outcomes of the repeated game.

These factors influence calculation of the players' rational choices. On the basis of its calculation of expected utility, each A_i chooses between **entry** and **no-entry** in the light of the current conditioned probability of the types of B. Obviously, in the first stage-game, the probability of types is such that the first player A_i will usually not place her trust in B, because the expected utility of **entry** is less than that of the alternative. Sooner or later, however, some A_i will decide to trust B if they have observed a series of **non-abuse** as a result of which the conditioned probability of the honest type has increased sufficiently to give the **entry** choice an expected utility greater than **no-entry**. There is always a calculable number of learning periods (in which learning must never contradict the hypothesis that B is an honest type) necessary for the probability of the type in question to reach the critical threshold $p^\star$ above which a short-run player A_i will for the first time rationally decide to trust B.

Analysing player B's choices requires consideration of the equilibrium strategies of the iterated game. At first sight, B might opt for the equilibrium strategy of each stage-game, namely **abuse**, which is certainly the best response to the choices made by the players A_i in the first periods. This strategy by player B gives rise to an equilibrium profile in the iterated game whose outcome consists of an infinite series of (**no-entry, abuse**) outcomes. However, player B has a different strategy available, which consists in exploitation of his knowledge of the mechanism by which the beliefs of the various A_i are updated. He may choose to simulate the behaviour of the 'honest' type until the stage occurs in which the conditioned probability of this type reaches the critical level $p^\star$ at which the first A_i will enter. At this point, B calculates whether to play the **no-abuse** action and consequently induce the players A_i to enter again, or to profit from the first opportunity to defect by choosing **abuse**, thereby gaining an unilateral advantage from the (−1, 3) outcome on the first occasion, but thereafter condemning itself to an infinite series of (0, 0) outcomes. If B is not impatient, and if the discount rate of future utilities does not excessively reduce the value of the future prospects of cooperation, infinite outcomes of future cooperation (which begin once the first A_i has entered) are able to offset the cost of the initial series of null outcomes (in which no A_i enters but B does not abuse), and to thwart the incentive to take advantage of an individual stakeholder as he enters. One possible rational strategy for B, therefore, is to sustain its reputation and to induce the sequence of stakeholders to trust it. The best response to this strategy by stakeholders A_i, from the time when the first of

them places its trust in B onwards, is to continue to be trustful until they observe a period in which B abuses. This gives rise to a game equilibrium profile in which a series of cooperative outcomes (**entry, no-abuse**) are observed from a certain point onwards (Fudenberg and Levine, 1989).[4] Hence, the long-run search for reputation induces the firm to behave as if it wants to fulfil its fiduciary duties towards the stakeholders.

However, it is essential to understand the conditions under which this result holds:

(a) *Signalling the types*: the firm must be able to signal the possibility that it is an honest type which does not abuse trust;
(b) *Quasi-simultaneity*: the firm and the stakeholder observe the result of each game simultaneously, for if the stakeholder acts first, the firm would have no reason to reveal its choice had the stakeholder not entered, so that there is no basis for *learning*;
(c) *Observability of the results*: at the end of each stage-game, the stakeholder must be able to observe the outcome of the firm's choice without ambiguity, and it must be able to determine without ambiguity whether the firm has behaved according to a *type*. Since types can also be viewed as commitments (to a certain game action), the essential condition is that at the end of each stage-game each stakeholder (the current one) should be able to observe that 'what had to be done has been done';
(d) *Shared knowledge amongst stakeholders*: each stakeholder must be able to transmit what he has learnt in a given period to the stakeholder that comes next. In other words, all the stakeholders in succession must have the same judgement on the firm's fulfilment of its commitments;
(e) *Absence of optimal mixed-strategies types*: were the firm able to calculate the probability of *non-abuse* at which the stakeholder would be indifferent between entering and not entering, then it would abide by its *type* that refrains from abusing only at that minimum indispensable level of probability. This type would obviously not induce an equilibrium in which the firm does not abuse, but instead one in which it abuses with the maximum possible probability compatible with maintaining stakeholder indifference between entry and non-entry. Hence mixed equilibrium strategies must be excluded for the virtuous effect of reputation to emerge.

In general, these conditions as a whole *are not* spontaneously fulfilled in situations relevant to the purposes of CSR, the consequence being that discretionary self-discipline (based on simple enlightened self-interest) normally fails. Putting aside the last assumption, which will be returned to in the section 'The motivational role of explicit ethical norms' the main reason why discretionary self-discipline fails is the *cognitive fragility* of reputation. This is evinced by conditions (a), (b), (c) and (d) above – all of which refer to the knowledge that the players must possess if the model is to hold true. Accumulating reputation may be prohibitively difficult if, in order to show that a commitment has been maintained, it is necessary to enable *each* stakeholder to observe that *concrete* actions have been undertaken, or that the *concrete* results have been obtained, so that they match their description established *ex ante* in a commitment announced by the firm (a possible *type*).

Typically, CSR is involved in *incomplete contract* situations where a contract does not contain clauses covering unforeseen contingencies, so that there is no concrete benchmark against which claims of renegotiation can be assessed when unforeseen events occur. Moreover, consider *unobservable quality*, such that the customer may not be able to verify a commitment to a quality level of a good or service on the basis of the information made available to him by inspection or experience. Or consider *organisational authority* where the 'boss' takes genuinely discretionary decisions with regard to tasks given to the employees. Finally, in *collusion situations* information about illicit agreements is not disclosed, so that those not present when a bribe is negotiated are unable to determine whether a commitment not to bribe has been breached. These are all settings in which information or knowledge about the firm's action is incomplete or highly asymmetric. Either commitments have not been defined in relation to unforeseen events, and therefore cannot be verified, or their fulfilment is not observable. The problem is that

incomplete information makes it impossible to determine whether 'what had to be done has been done': either it was not established *ex ante*, so that there is nothing to verify, or it is impossible to observe results by which it can at least be inferred whether the commitment has been respected (since the result coincides with at least one of the possible results contemplated *ex ante*). Activation of the reputation mechanism is obstructed by a cognitive gap.[5]

Filling the gap in the reputation mechanism: a cognitive role for explicit ethical norms

Although reputation may not be effective in supporting the firm in fulfilling its simple commitments on strategies, may we nevertheless continue to say that extended fiduciary duties can be self-enforcing? The answer is yes. This can be shown by substituting self-regulation in the *proper sense* – as a set of voluntary but nevertheless *explicit* and *standardised* ethical norms which are used to fill the cognitive gap – for the discretionary approach based on the mere farsighted self-interest considered so far. Instead of being concerned with the enforcement of legal norms by an external authority, self-regulation concerns the need to create the cognitive and informational bases that enable the social mechanism of reputation – with its endogenous rewards and punishments – to function properly. This comes about through the voluntarily-taken decision to accept *explicit norms with an appropriate structure* established by the firm in the light of a multi-stakeholder social dialogue such to induce impartial acceptability. For this reason, self-regulation is a voluntary but *not* discretionary approach. Voluntariness resides in the decision to endorse an *explicitly announced ethical standard* for the firm's management system and governance, which is *ex ante* shared among the firm and its stakeholders. This standard sets out general principles, whose contents are such to elicit stakeholder consensus, as well as explicit commitments to compliance with principles and rules, which are known *ex ante* by stakeholders. It is clear that stakeholder consensus can be more easily obtained if the standard relative to the strategic management system, intended to ensure CSR, is established by the firm through explicit dialogue with the stakeholders.

However, explicit statement and dialogue do not detract from the voluntary nature of the standards; nor does it preclude that compliance may then be obtained via the self-enforcement of the fiduciary duties established. How this occurs can once again be explained by referring to the reputation mechanism. The standard, and the procedures ensuring compliance with it, are announced *ex ante*; and it is *on these* – not in relation to particular (unforeseen) events or to particular (unobservable) actions or outcomes – that firm and stakeholders pass homogeneous judgement on *ex post* compliance with them. These duties assert – in the proper form – what is to be expected of the firm in *unexpected* situations as well, or in ones where the results of actions *are not observable*. Once the gap has been filled, it is possible to reactivate the reputation reward and punishment mechanism which generates endogenous incentives to comply with the standard itself. Everything rotates around the gap-filling function performed by the CSR self-regulatory norms and standards whereby the firm's fiduciary duties towards its stakeholders are made explicit and announced. In order for this to work properly, three conditions must be met:[6]

(i) *General and abstract principles*: Principles define the *vision* of the social contract that each firm proposes to its stakeholders (which must therefore be completely identified). These principles must offer terms of treatment, which each stakeholder accepts as fair. Their form is abstract and general, so that they apply to a wide variety of events, including those which cannot be predicted or described beforehand. Consequently, their application does not require an *ex ante detailed description* of any possible situation; all that is necessary is *ex post* recognition of the presence of certain abstract features which reflect a pattern established at the outset.[7]

(ii) *Precautionary rules of behaviour*: Definition of principles allows identification of areas of potential opportunism where interactions between stakeholders and firm put those principles at risk (in the intuitive sense of the term, without its probabilistic specification). Given each of these risky areas, precautionary rules of behaviour can be

established, which assure the relevant stakeholder that a particular form of opportunism has been prevented. The distinctive feature of these rules is that their implementation is not conditional on the actual occurrence of concrete foreseen situations. Operationally, they are applied when the extent to which the occurring situation belongs to the domain in which a principle is breached exceeds a pre-announced threshold. Hence, the conditions of their implementation can be established *ex ante* by the firm, and on these the stakeholder may legitimately form expectations about the firm's behaviour. Their application constitutes evidence that no principle has been intentionally breached, and consequently that the firm's reputation is well-deserved.

(iii) *Communication and dialogue with the stakeholders*: Stakeholders base their assessments on the match amongst principles or rules announced *ex ante*, level of membership in the principles domain by any event which has occurred and the behaviour adopted. Dialogue and communication generate a common understanding between the firm and its stakeholders about the former's principles and commitments and the latter's expectations. Dialogue enables symmetric interpretation of critical situations by the parties, so that no serious divergence arises amongst them about the contingencies where rules of conduct must be implemented.

As regards the first condition, general principles identify moral properties associated with abstract and universal characteristics not necessarily bound to a complete description of every concrete contingency that may occur in all the states of affairs. An abstract and general principle corresponds to a domain of application (a set of states) and membership of that domain is a matter of degree. My suggestion is that a fuzzy membership function, with values in the real interval [0,1], can be defined for each unforeseen state that occurs (or proves to be possible *ex post*); which implies that the domain of application of a principle is a *fuzzy set*.[8] Foreseen states of the world will belong to it or otherwise in a clear-cut way, for they will or will not exhibit the properties assumed to be associated with the given ethical principle. Unforeseen contingencies will instead define a vague domain of application of the principle, namely the set of states of the world whose possession of the descriptive characteristics associated with the principle is a *matter of vagueness*. Although their belonging may be vague, unforeseen contingencies will nevertheless always belong to some extent to the domain of application of a general principle. It is precisely the abstractness of principles that makes their application to *every* situation possible, even if this situation is *ex ante* unforeseen, whereas concrete rules, which are contingent upon a detailed state description, would be simply *mute*.

Once we have accepted that there are general and abstract properties to which unforeseen states adapt at least imperfectly, we may resort to a default inference rule such as the following: "an unforeseen state with a degree of membership in the principle's domain at least equal to α (for $1 < \alpha < 0$) is 'normally' an exemplar of the principle".[9] Hence using the same mode of inference I conclude that even if I do not have complete proof that this is the case, a rule of behaviour conforming to the principle is required. The logic at work here is *default reasoning*: even if it is obviously fallible, default reasoning is the best that we can rely upon in order to cope with unforeseen contingencies using limited rationality (Ginsberg, 1987; Reiter, 1980).[10]

What really matters is that now we are in a much better situation as far as the possibility of undertaking commitment amidst unforeseen contingencies is concerned. The principles and rules of conduct couple allows *ex ante* specification to be made of the *conditions* under which a certain procedure must be carried out *ex post*, *without the requirement* of giving *ex ante* any detailed concrete description of these states of affairs (because they may be entirely unforeseen). These conditions essentially state that a situation, whatever it may be, must belong *at least to some extent* to the domain of application of a principle. It is thus possible *ex ante* to *undertake* commitments and generate expectations about future behaviours.[11]

Such preventive rules of conduct, of course, will not induce utility maximisation in every state. Whilst the non-monotonicity of default reasoning allows one to infer from vague information that

'normally' cases 'such and such' must be managed according to a certain procedure, it also entails that in the presence of additional information those conclusions may no longer hold.[12] The main point, however, is that principles and rules of conduct allow for at least provisional 'completion of the contract'. This completion is made possible by the specification of what the stakeholder may expect in terms of commitments under certain *ex ante conditions*. The *ex post* execution of rules of conduct will then provide a reliable base for deciding whether 'what had to be done has been done'. This allows recourse, even under unforeseen contingencies, to the reputation effects mechanism. The 'honest' type of the firm must be replaced by the type which conforms with a rule of conduct when an ethical principle in the CSR management standard requires it: that is, in all the contexts where the *ex ante* announced conditions on principles and rules are fulfilled.

Summing up, somewhat paradoxically, Jensen's (2001) criticism that, by requiring multiple objectives and multiple obligations, CSR would make managerial decisions excessively complex, given that decision-makers are endowed with only bounded rationality, retorts on itself. The multi-fiduciary model of governance is, in fact, implemented through a voluntary but explicit set of CSR principles and rules of conduct able to overcome the problem of undertaking commitments in a world of bounded rationality and unforeseen events. A principle-and-rule-following decision model, conditional on requisites specified in order to cope with vagueness and unforeseen contingencies, fits the idea of bounded rationality. Those who concede that the multi-stakeholder approach would over-complicate decision processes have conceded too much to its adversaries (on this 'excess' of concession, see Phillips et al., 2003).

The motivational role of explicit ethical norms

The problem of refined abuse

Compliance with voluntary but explicit CSR norms (codes of ethics, management system standards etc.) can be effective once they work as parameters against which a firm's reputation can be assessed. Yet a firm which has acquired a reputation as a relatively mild abuser of its stakeholders' trust can nonetheless acquire their cooperation if the incomplete fulfilment of its duties generates an expected payoff for the stakeholder not less than that promised by his alternative choice of withdrawing from all transactions. In this case, the firm would fulfil its duties only to the minimum extent necessary to dissuade the stakeholder from exiting the relation (although it is rather unfair). Moreover, if this strategy is available to the firm, and if the firm is able to acquire the corresponding reputation, it will be an equilibrium strategy; and one that the firm will certainly try to adopt, given that under this equilibrium the stakeholder will acquiesce by relinquishing to the firm the largest part of the surplus and obtaining practically nil for himself. I call this behaviour by the firm a strategy of *refined abuse*.

This introduces the need to consider the motivational force of explicit but self-imposed ethical norms (or CSR management standards). Given that these entail ethical principles and duties of conduct – and are therefore couched in the language of *deontology* – they enable recourse to stakeholder's motivations, which extend beyond the mere material advantage deriving from transactions with the firm. Many stakeholders, in fact, have motives to act that are not purely self-interested or geared to material advantages (consequences). These stakeholders also place importance on the firm's fulfilment of CSR duties deriving from the social contract, especially if the firm enunciates these duties in codes of ethics and communicates them externally. Hence, any deviation from the CSR standard, or from the firm's ethical commitments, may be subjected to harsher punishment than would be the case if simple material interests were concerned.

Let us assume that the firm is a single player with an extremely low level of intrinsic desire to conform with CSR fiduciary duties, although it has an ethical code and a standardised CSR management system that may constitute the terms of reference for behaviour ideally conforming with those duties. Contrary to the case of the firm, though, assume that the stakeholders have a strong conformist orientation. If the firm has adopted a standard and a code of ethics which affirm an ideal of fairness – and if the stakeholders expect the firm to comply with these

principles – the stakeholders will associate high-intrinsic utility with the fact that the firm is behaving consistently with the code of ethics and principles of mutual advantage and fairness. By contrast, if the stakeholders behave cooperatively and expect the firm to know that they are doing so, any deviation by the firm from its principles will affect them not only materially but also because it contradicts their conformist principles. (This seems to be the case of the ethical investors or responsible consumers who expect that the respect for human rights and concern for environmental impact currently required in their countries – and in regard to which companies are usually considered to be good corporate citizens at the national level – should be shown by those same companies wherever they operate.)

This prompts the question of whether these hypotheses are such to prevent the firm's refined abuse strategy. By hypothesis, the firm will adopt this strategy only if it is able to induce the stakeholder to enter (see again the trust game illustrated in Figure 1). However, it can be shown that if the stakeholder is conformist, it will refuse to enter, with the consequence that the firm's refined abuse strategy no longer induces equilibrium.

Conformist preferences and reciprocity

In this section I summarise briefly the ideas underlying the conformist preferences model (Grimalda and Sacconi, 2002, 2005; Sacconi, 2004; Sacconi and Grimalda, 2006), before applying it to the problem raised at the end of the previous section (see *Appendix 1* for the mathematical model). Let me assume that stakeholders have not only self-interested motives of preference but also ideological motives of preference, and that their accepted ideology coincides with the social contract of the firm – i.e. the guiding principle of an extended corporate governance system. I suggest that these two classes of motives can be accounted for by two types of preferences of the Self and by their relative mathematical representation in the corresponding utility function.

To begin with, we should consider that strategic interaction generates states of affairs, which can be differently described according to their characteristics. A first description of states views them as *consequences*. Consequences may be described as attributed only to the acting self – what happens to the decision-maker in any state. This description is the basis of self-interest: the Self has preferences for consequences which are self-referred. By contrast, consequences may be attributed to all persons (extended consequences) in so far as they can be viewed as what happens to any whatever individual. This makes a case for some sort of impartial consequentialist ethic like utilitarianism or altruism. In general, if a player defines his preferences only on states *described as consequences*, then he has *consequentialist personal preferences*.

The second type of preferences is comprisesd of what I call *conformist personal preferences*. States description is no less important here, but states are now described as sets of interdependent actions characterised in terms of whether or not they conform with a given abstract principle or ideal. This ideal can be captured by a function of individual first-type utilities attached to states, which measures the fairness of welfare distribution within each state of the world. A pattern of behaviours (a vector of strategies) is fixed and defined as perfectly *deontological* if it is fully consistent with the abstract principle of fairness – that is, if it maximises the function just defined. I shall call such a state the *ideal*. In the context of my definition of CSR it is particularly appropriate to assume that the principle is contractarian. I accordingly assume that its technical form will be the same as the Nash Bargaining Solution applied in the ideal bargaining game amongst all the firm's stakeholders (the owner and manager included). This is the same normative principle that I have used to derive extended fiduciary duties within the context of justification (see Sacconi, 2006a, b). Thus a state (a strategy-combination in the reference game) will be completely consistent with the principle if this state maximises the Nash Bargaining Function with respect to the state space of the game. We may also view it as the Nash Social Welfare Function (N.S.W.F.), which induces a definite welfare-distribution ordering over the state space.

There is a hierarchical relation within this model between the two kinds of preferences – consequentialist and conformist. In order to define fairness, in fact, I inspect the distributions of the payoffs derived from *first type of preference* – i.e. material utilities. But this does not reduce second-type preferences to first-type ones. First-type utilities are

no more than *rough materials* for the definition of second-type preferences. What matters for description of the latter are not consequences or material payoffs as such, but a distributive property defined over payoffs, which is expressed by the function representing a principle of fairness.

Thus characterisation of second-type preferences accords more with deontology than consequentialism. The more an expected state of affairs (a combination of actions) conforms with the ideal, the more it is preferred by a player (through the measure of expected reciprocal conformity – which we will shortly see is the basis of each player's conformist preferences for states). Moreover, there is no reason to link deontology with a belief that there is some objective source of value possessing ontological reality 'out there' (independently of the decision-maker's affections)[13]. In fact, while conformist preferences depend on degrees of deontology, deontology itself may be understood, as I do, simply as individual compliance with a fair distribution principle *that players could have rationally agreed upon in an ex ante hypothetical bargaining situation*.

Recall, however, that we are in the compliance and implementation context, so that motivation does not simply consist in consistency with a normative principle. Hence, conformist preferences do not directly depend on the characterisation of the state of affairs in terms of the principles as such. We have to explain how individual *preferences* for conformity may be produced. The idea here is that a player who has at least hypothetically agreed to a fair principle will have a motive to prefer acting in accordance with that principle (not necessarily an overwhelming reason, but nevertheless one endowed with some causal force) *if* (i) it is within the scope of his *responsibility* (i.e. it is an agent-relative reason to act) to contribute significantly to fulfilling the ideal, conditional on his expectation of the other individual's action; and if at the same time (ii) he expects that, conditional on what is expected from himself, the other player will *reciprocate* by contributing significantly to fulfilment of the same (agreed) ideal. These conditions can be more clearly distinguished as follows: (*i'*) *conditional conformity*, the *level* at which the agent himself *conforms* with the ideal, given his belief about the other party's choice; (*ii'*) *expected reciprocal conformity*, the *belief* concerning the reciprocity level at which each of the other players will replicate conformity with the ideal, given what they believe about the first party's choice.

Hence, a degree of conformity with the ideal may be appended to each individual strategy choice by seeing whether the *ideal* comes about through this choice, given what the first party believes about the other parties' choice. This is a measure of the extent to which conformity with the ideal can be attributed to the *responsibility* of each player in *carrying out his duties*, conditionally on the expected action of the other players.

Once these concepts have been translated into a formal model, a player's *comprehensive utility function* consisting of two parts (which they assume to be separable), i.e. the representations of consequentialist and conformist preferences, can be defined as

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

where U_i is the traditional player i's 'consequentialist' utility for state σ (a given strategy combination), the weight λ_i – which may be any positive real number – is an exogenous psychological parameter expressing how important the conformist component is within the motivational system of player i (we may call it player i's 'maximum disposition to act according to conformist reasons', granted that certain conditions apply), and F is a function representing reciprocal conformity with the principle T which in turn is a function taking a value for each state σ. Note that T will be defined as the contractarian principle of fairness (i.e. it is the N.S.W.F.) and that F (a function of the principle) will take the form of two combined indexes expressing respectively:

(a) (*conditional conformity*) or the extent to which player i contributes fulfilling the ideal by carrying out a state as much consistent with the maximal approximation to the ideal as it is affordable given what he believes about player j's choice;
(b) (*reciprocal conformity*) the extent to which player j contributes to the fulfilment of the ideal (as seen through player i's beliefs) by carrying out a state as much consistent with the ideal as it is affordable given what player j believes that player i will do.

(For complete treatment of these indexes and the overall utility function see *Appendix 1*).

How conformist preferences prevent refined abuse

In this section I will provide a proof of the following proposition: *if we assume that stakeholders have (inter alia) conformist preferences, then the firm's opportunistic use of a mixed-strategy of minimal compliance with a (contractarian) code of ethics will no longer be a rational strategy for the firm.*

Let us first return to the infinitely repeated reputation game between a long-run firm B and one short-run stakeholder at time A_i whose stage-game is the trust game (see again Figure 1). First consider the player B type that adopts a mixed-strategy (*2/3* ***a***,1/3 ***no-a***) i.e. B may try to develop a reputation for being this type by playing the two pure strategies with the attached probability throughout all the repetitions of the game (let us assume that these mixed-strategy types can be understood and learned by the A_i players). This type, once stakeholders believe it with probability one, renders each player A_i indifferent amongst all his pure or mixed-strategies, since in every case he obtains no more than 0. Typically, player A_i's best response is to remain indifferent between her pure strategies, so that he will use the mixed-strategy (1/2 ***e***, 1/2 ***no-e***), which also gives A_i the expected payoff 0, while granting player B an expected payoff of 1.33.

Of course, this is not an equilibrium of the stage-game, because player B's best response to any probability of entrance by players A_i is to play abuse. However, we are here considering the reputation game, and assuming that mixed-strategy types of player B can accumulate reputation by repeatedly playing their characteristic strategy. This suggests the following repeated game equilibrium: B can improve his stage-game payoff above the Stackelberg pure strategies payoff of 2 by simply resorting to a commitment on the mixed-strategy ($2/3-\varepsilon$ ***a***, $1/3+\varepsilon$ ***no-a***) (with ε as small as possible.) In fact, if player B is able to acquire the reputation that he is this type, player A_i necessarily enters in order to get the expected positive payoff of 3ε (as in the original) (which, owing to the infinitesimal ε, is practically nil) and gives player B a stage-game expected payoff of $2.66-\varepsilon$. Note that player B's best response is not to deviate to stable defection (**abuse**), for this would change his reputation until a player A_i will consider only a nearly complete abusive type to be possible, so that no further A_i players will enter thereafter. This suggests that there is a mixed-strategy equilibrium in which, also by not complying entirely, player B (the firm) can acquire a reputation such that stakeholders enter and acquiesce to a firm appropriating large part of the surplus. But this implies that for the most of time the firm will not fulfil its fiduciary duties towards the stakeholders.

This pessimistic result obtains when no role is played by conformist preferences. Let us now see how the more complex representation of the stakeholders' system of preferences may change the picture. I will be as parsimonious as possible in introducing motivations that facilitate compliance with fiduciary duties by the firm. Thus, as before, I assume that only stakeholders (for example employees or consumers, or investors) cleave to the ideal of the socially responsible firm as defined by a code of ethics or a voluntary CSR management system expressing the ideal model of a social contract between the firm and its stakeholders. This presupposes that the firm has at least deliberated and signalled a commitment to such a code, but *not necessarily* that its management has developed the kind of attitude that I call conformist preference for reciprocal compliance with it. Hence, I set to 0 the λ_B parameter in the manager's or entrepreneur's utility function with which I capture the weight of conformity within the utility function of player B, the 'firm'. Stakeholder A for his part has a positive weight λ_A and an *overall utility function* combining both consequentialist (self-interested) motives and conformist motives to act. Hence, the stakeholder's ideal of socially responsible governance of the firm may be calculated by means of the N.S.W.F. Owing to the very simplistic representation of the basic trust game, this welfare function is defined only for the firm's payoffs and the payoffs of the sole stakeholder taking part in the current stage-game, without considering that at any time there will be many stakeholders involved, or that what the firm decides at one time may be relevant to other stakeholders at later times. I am confident that accounting for these more complex interactions would only reinforce the result of this section, in so far as it would imply that further stakeholders' conformist preferences would push the outcome towards the same direction.

Conformity indexes (both conditional and reciprocal) require beliefs concerning the other

player's actions. Player A's relevant first-order and second-order beliefs in these exercises are

$$b^1_{\mathrm{A}} = (2/3\,\boldsymbol{a}, 1/3\,\boldsymbol{no\text{-}a}), \quad \text{in short } (2/3, 1/3)$$
$$b^2_{\mathrm{A}} = \boldsymbol{e}, \quad b^2_{\mathrm{A}} = \boldsymbol{no\text{-}e}$$

In other words, I define player A's overall utility function for a situation in which she believes that player B will abuse with probability 2/3 and not abuse with probability 1/3, while she has the second-order belief that player B predicts that she (player A) will enter. Admittedly, this is not the effective mixed-strategy equilibrium of the repeated game without conformism, for this would be the case only for an infinitesimal ε added to the probability of the 'no-abuse' strategy. However, we may disregard the small ε as it makes almost no change to the two players' payoffs, so that we may proceed as if the effective mixed-strategy equilibrium were (***e***, *(2/3,1/3)*). Hereafter, I will contrast this utility value with the *alternative* utility value for the case that player A believes that player B will play the 'equilibrium' mixed-strategy *(2/3,1/3)*, but she will not play the entry strategy, so that her second-order belief is that she herself *does not enter* and the firm predicts that she will not enter (formally $b^2_{\mathrm{A}} = \boldsymbol{no\text{-}e}$). In order to calculate player A's overall utility for the two alternatives (***e*** and ***no-e***), we must consider two couples of indexes:

- player A's conditional conformity index (case A*) for *strategy* ***e*** when she chooses ***e*** given that she believes player B is choosing *(2/3,1/3)*, combined with player B's reciprocal conformity index (case B*) based on player A's second-order beliefs, i.e. the index annexed to B's choosing *(2/3,1/3)* (and player A believing it), given that he believes that she chooses ***e*** (and A believes that B believes it).
- player A's conditional conformity index for *strategy* ***no-e*** (case A**) when she chooses ***no-e*** given that she believes that B plays *(2/3,1/3)*, combined with player B's conformity index (case B**) based on player A's second-order beliefs, i.e. the conformity index that results from B's choosing *(2/3,1/3)* (and A believing it) given that he believes that she chooses ***no-e*** (and player A believes that B believes it).

Let me anticipate (from Appendix 1) the formula for player A's conditional conformity index (varying from 0 to −1)

$$f_{\mathrm{A}}\left(\sigma_{\mathrm{A}k}, b^1_{\mathrm{A}}\right) = \frac{T\left(\sigma_{\mathrm{A}k}, b^1_{\mathrm{A}}\right) - T^{\mathrm{MAX}}\left(b^1_{\mathrm{A}}\right)}{T^{\mathrm{MAX}}\left(b^1_{\mathrm{A}}\right) - T^{\mathrm{MIN}}\left(b^1_{\mathrm{A}}\right)} \qquad (2)$$

where b^1_{A} is the (first order) belief of player A concerning player B's action; $T^{\mathrm{MAX}}\left(b^1_{\mathrm{A}}\right)$ is the maximum attainable by the N.S.W.F. T given A's belief; $T^{\mathrm{MIN}}\left(b^1_{\mathrm{B}}\right)$ is the minimum attainable by the N.S.W.F. T given *i*'s belief; and $T\left(\sigma_{\mathrm{A}k}, b^1_i\right)$ is the effective level attained by the N.S.W.F. T when the player A adopts his strategy σ_k, given his belief b^1_{A} about the other player's behaviour. Let us now use formula (2) to compute player A's conditional conformity indexes for the two cases considered.

Case A*: player A's strategy ***e***, given beliefs *(2/3,1/3)*

$$\frac{T(\boldsymbol{e}, (2/3, 1/3)) - T^{\mathrm{MAX}}(2/3, 1/3)}{T^{\mathrm{MAX}}(2/3, 1/3) - T^{\mathrm{MIN}}(2/3, 1/3)} = 0$$

Case A**: player A's strategy ***no-e***, given beliefs *(2/3,1/3)*

$$\frac{T(\boldsymbol{no\text{-}e}, (\boldsymbol{2/3}, \boldsymbol{1/3})) - \boldsymbol{T}^{\mathrm{MAX}}(\boldsymbol{2/3}, \boldsymbol{1/3})}{T^{\mathrm{MAX}}(2/3, 1/3) - T^{\mathrm{MIN}}(2/3, 1/3)} = 0$$

In fact, the value of the *N.S.W.F.* in case A* is *T(e, (1/3,2/3))* = 0 × 2.66, and it is *T*(***no-e***,*(2/3,1/3)*) = 0 × 0 in case A**. At the same time, T^{MAX}*(2/3,1/3)* and T^{MIN}*(2/3,1/3)* are both 0 because player A's expected payoff is always 0 under player B's strategy *(2/3,1/3)*. The meaning of these 0-levels of the conformity index is better understood by interpreting them as degrees of deviation from complete compliance with the ideal, conditional on the other player's expected choice. In both the A* and A** cases, player B's expected mixed-strategy *(2/3,1/3)* nullifies any effort that player A might make to enhance the level of ideal attainment. Whatever player A does, in fact, the level of T is always 0. Thus A has no

responsibility for any deviation from the maximum feasible level of T, given B's choice. Formally, in fact, the level of T is at its maximum (which is also its minimum in this case) given player B's opportunistic strategy which always reduces player A's payoff to 0 (or practically nil).

Now consider (again anticipating Appendix 1) the formula for player B's (expected reciprocal) conformity index (which also varies from 0 to −1) based on player A's first- and second-order beliefs, where b_{A}^{1} represents player A's belief about player B's strategy choice (i.e. an expected strategy of player B) and b_{A}^{2} represents player A's second-order belief about player B's belief about player A's choice (that is, a second-order expected strategy of player A).

$$\tilde{f}_B\left(b_{\mathrm{A}}^{1}, b_{\mathrm{A}}^{2}\right) = \frac{T\left(b_{\mathrm{A}}^{1}, b_{\mathrm{A}}^{2}\right) - T^{\mathrm{MAX}}\left(b_{\mathrm{A}}^{2}\right)}{T^{\mathrm{MAX}}\left(b_{\mathrm{A}}^{2}\right) - T^{\mathrm{MIN}}\left(b_{\mathrm{A}}^{2}\right)} \quad (3)$$

According to formula (3), this index, takes the following values in the two cases considered, i.e. when B uses the mixed-strategy against ***e*** or ***no-e***.

Case B*: strategy *(2/3,1/3)* used by player B given his belief that A chooses ***e***

$$\frac{T((2/3,1/3),\boldsymbol{e}) - T^{\mathrm{MAX}}(\boldsymbol{no\text{-}a},\boldsymbol{e})}{T^{\mathrm{MAX}}(\boldsymbol{no\text{-}a},\boldsymbol{e}) - T^{\mathrm{MIN}}(\boldsymbol{a},\boldsymbol{e})} = -\frac{4}{7} = -0.57$$

$T^{\mathrm{MAX}} = 2 \times 2$ is in fact the maximum value of the *Nash product* that ensues when B does not abuse, given that A enters, while $T^{\mathrm{MIN}} = -1 \times 3$ is its minimum value ensuing when B abuses given that A enters. The ideal's value if B plays the mixed-strategy when A enters is nevertheless zero because $T = 0 \times 2.66$. Note that the index contrasts the T value produced by player B's opportunistic mixed-strategy if player A acquiesces with the T^{MAX} value that player B could attain given the same player A choice. But this implies a marked deviation from maximal conformity conditional on A's behaviour which can be *imputed entirely to player* B's decision to play his mixed-strategy instead of his ***no-a*** strategy. In this case, player B does not conform with the ideal at a significant level, and this results in the negative value assumed by his conformity index.

Case B**: strategy *(2/3,1/3)* used by player B when he believes that A chooses ***no-e***, and player A believes that B believes it

$$\frac{T((2/3,1/3),\boldsymbol{no\text{-}e}) - T^{\mathrm{MAX}}(\boldsymbol{no\text{-}a},\boldsymbol{no\text{-}e})}{T^{\mathrm{MAX}}(\boldsymbol{no\text{-}a},\boldsymbol{no\text{-}e}) - T^{\mathrm{MIN}}(\boldsymbol{a},\boldsymbol{no\text{-}e})} = 0$$

In fact, for whatever player B choice given the expected ***no-e*** by player A, T^{MAX} as well as T^{MIN} are both zero, and this is also true of the single payoff obtained by player A under *(2/3,1/3)*. Given his belief ***no-e***, player B cannot significantly deviate from the ideal, for even if he plays his mixed-strategy he is not accountable for a deviation from the maximal ideal's value given **no entry** by player A. Comparing B* and B** shows that the *intention* to exploit player A's acquiescence *implies that B has a significant responsibility* for a deviation from (non-conformity with) the ideal only conditional on the expectation that in effect player A will acquiesce, for it is precisely in this case that he does not reciprocate player A's conformity.

Finally (according to formula (5) of *Appendix 1*), I can calculate player A's overall utility values for the two alternative strategies ***e*** and ***no-e***, respectively, given that player A predicts that player B will use strategy *(2/3,1/3)*, under the assumption that player B predicts player A's choice, i.e. he believes (and she believes that he believes) that player A uses either strategy ***e*** or strategy ***non-e***, respectively. The material payoff to player A if she plays 'enter' under the equilibrium mixed-strategy is 0 (even accounting for an additional infinitesimal probability ε attached to ***no-a***, the payoff is still practically nil), whereas her conformist utility is based on the indexes A* and B*. Thus player A's overall utility for strategy ***e*** is

$$V_{\mathrm{A}}(\boldsymbol{e}, b_{\mathrm{A}}^{1}, b_{\mathrm{A}}^{2}) = 0 + \lambda_{\mathrm{A}}(1 + (-0.57))(1 + 0) = 0.43\lambda_{\mathrm{A}}$$

On the other hand, player A's conformist utility for strategy ***no-e*** is given by the indexes A** and B**, whereas his material payoff is again 0. Thus A's overall utility for strategy ***no-e*** is

$$V_{\mathrm{A}}(\boldsymbol{no\text{-}e}, \boldsymbol{b_{\mathrm{A}}^{1}}, \boldsymbol{b_{\mathrm{A}}^{2}}) = \boldsymbol{0} + \lambda_{\mathrm{A}}(\boldsymbol{1} + \boldsymbol{0})(\boldsymbol{1} + \boldsymbol{0}) = \lambda_{\mathrm{A}}$$

A conclusion follows straightforwardly: player A with conformist preferences refuses to play the

(interval) mixed equilibrium strategy of the repeated material trust game. In so far as player B's conformity index annexed to the equilibrium mixed-strategy is negative and, *ceteris paribus*, the weight λ_A that the stakeholder attaches to conformity is positive, the logic of strategic choice under conformist preferences reverses the result of standard strategic calculation in a repeated trust game with mixed strategies.

Granted that λ_A is positive, this result typically follows from the opportunistic nature of player B's mixed-strategy type. In fact, he endeavours to minimise (to nil) the stakeholders' payoffs, whereas he is nevertheless convincing them to enter so that the firm can appropriate a part of the surplus larger than the 'cooperative' payoff (2) in each stage-game. Of course, were B willing to concede more to A, for example by raising the additional probability ε of 'no abuse' to a substantial level, then A's firmness in rejecting 'entry' would be lessened. However, straightforward calculation shows that, *ceteris paribus*, A stops rejecting 'entry' only when $T((2/3 - \varepsilon, 1/3 + \varepsilon), \boldsymbol{e})$ takes value 4, which means that the probability of 'no abuse' is 1 – i.e. player B's mixed-strategy type degenerates to his 'honest' type.

Conclusion

This essay has examined compliance with, and implementation of, the social contract agreed upon the fiduciary-duties institutional framework of the socially responsible corporation. It has discussed endogenous incentives and motivations to comply with the CSR corporate governance structure in the domain of reputation games, on the assumption that the firm may benefit from reputation when it fulfils its fiduciary duties to its stakeholders. However, it has been shown that use of the standard models of reputation is obstructed by two factors: first, cognitive limitations jeopardise the undertaking and specification of commitments with respect to unforeseen contingencies; second, a firm's sophisticated opportunism may take advantage of the other's self-interest, inducing the stakeholder to acquiesce to abuse by the firm instead of inducing the firm to conform with the social contract.

A solution to the former problem has been based on the cognitive role of explicit norms of ethics and CSR management and governance standards. These fill the gaps because of their logical structure, which requires general and abstract principles of ethics, precautionary rules of behaviour and dialogue with the stakeholder in order to achieve common understanding on the extent to which states of affairs belong to the domain of application of the principle and rules couple. It is to be noted that the use of explicit norms as decision devices able to overcome cognitive limitations in managing reputation introduces a marked change in the economic agent's logic of decision because fuzzy logic and default reasoning form the basis of what elsewhere is called procedural rationality.

This is closely connected with the underlying contractarian nature of business ethics principles and norms, which are fruitful essentially because of their abstractness and generality. Although they are vague, they have the virtue of extending from one context to another. Such principles can be typically understood as chosen in a hypothetical decision situation. The decision-maker adopts a counterfactual mode of reasoning to detach himself from the details and concrete characteristics of any particular situation in order to reach agreement on universalisable principles acceptable to all and applicable to very wide classes of situations, even if they are unforeseen *ex ante*. The core of this theory is a model of reasoning that makes it possible to undertake principle-based commitments with regard to unforeseen contingencies. (Sacconi 1991, Sacconi 2000). (For a quite different view about the role of general and abstract principles of ethics in front of bounded rationality see Donaldson and Dunfee (1995))

The same contractarian approach to fairness principles has provided a remedy for the second drawback of reputation theory, which concerns the motivational role of contractarian business ethics norms. In fact, the contractarian principle of fairness (the Nash bargaining solution of an ideal bargaining game) is a building block of the model of conformist preferences. This model enables formalisation of the drive to act out of a desire to conform with an ideal, granted that the same ideal is agreed upon by other players, and also that these players are expected to reciprocate conformity with the same ideal.

This new model of preference yields the following noteworthy result: if stakeholders share the ideal of a multi-stakeholder firm (that is, the contractarian principles of a corporate code of ethics), and if they earn ideal utility from reciprocally expected conformity with that code, *then* a mixed-strategy in the repeated trust game which induces stakeholders to endure exploitation can no longer be an equilibrium.

This depends on the nature itself of conformist preferences as they ensue from reciprocity in conformity. If one examines the significance (or the intention) of the stakeholder's decision whether or not to enter, given the hypothesis that the firm wants to adopt the sophisticated abuse mixed-strategy, one finds that the stakeholder is entirely unable to alter the situation's proximity to the social contract ideal. However, when the stakeholder considers the firm's behaviour, given that the latter expects that in equilibrium the former will want to enter, the significance of the firm's choice is a marked deviation from the social contract ideal, because the intention is to expropriate the stakeholder precisely when he adopts a cooperative behaviour. This therefore negatively affects the conformist preferences of the stakeholder, which will be doubly dissatisfied with the firm's behaviour. The inclusion of a negative conformity index in the overall utility function entails that the stakeholder will not enter the relation with the firm. Indeed, the stakeholder will punish the firm more harshly than the damage that the firm's calculated abuse would warrant in the absence of conformist preferences.

Notes

[1] As in Gauthier's *constrained maximisation* theory (1986, 1990, 1996) and McClennen's *resolute choice* theory (1990, 1993) The typical alternative game theoretical route to understanding endogeneous compliance with the social contract-an alternative that however I don not follow for I'm interested in the weakness of cognitive mechanisms of human rationality, but not in putting it completely aside, is *evolutionary game theory*, see Binmore 1997, 2005, Sugden 1986, Skyrms, 2004.

[2] Among the studies suggesting to go beyond the mere selfish representation of human preferences see in particular Bernheim (1994), Rabin (1993), Chareness and Rabin (2002), Sugden (1998), Frey (1997), Falck and Fishbaker (2000), Fehr and Schmidt (2001), Falk et al. (2003). However, in this essay I shall refer to my own contributions to the field (see Sacconi, 2004; Grimalda and Sacconi, 2002, 2005; Sacconi and Grimalda, 2006). For these references see the bibliography at the end of Part II.

[3] Basic reference are Kreps et al. (1982), Kreps and Wilson (1982); for a general presentation of the subject see Fudenberg and Tirole (1991) Ch. 9, see also Fudenberg (1991).

[4] For a simple illustration of this model and result see Sacconi (2000).

[5] Kreps made this point first (see Kreps 1990).

[6] This theory of reputation under unforeseen contingencies and incomplete contract is fully developed in Sacconi (2000, 2005a). For the design of a CSR management standard corresponding to these requirements see the Q-RES model in Sacconi et al. (2003), another example is Clarkson Centre for Business Ethics (2002), *Principles of Stakeholder Management*.

[7] I am here elaborating on a suggestion by David Kreps' theory of ''corporate culture'' (Kreps, 1990). However, my modelling of principles is completely different from the one attempted by Kreps, which used a focal points approach, a method that I deem not very useful in this context (see Sacconi, 2000 ch.4).

[8] Fuzzy sets were introduced by Zadeh (1965). For a partial survey of the burgeoning literature on fuzzy sets see Zimmerman (1991); for simple definitions useful for my application see Sacconi (2000).

[9] Note that when degrees are 0 or 1 we are in a situation of non-vagueness concerning the truth of statements about whether a case belongs to the domain of a principle or not. But I assume that this is true only for foreseen states of affairs, which are *ex ante* described with exactly the same language and state of knowledge where we also put forward the principles *ex ante*.

[10] For the application of default reasoning in contact theory and reputation theory see Sacconi (2000, 2005a).

[11] See Sacconi (2000) chap. 8 and Sacconi (2005a).

[12] On *Non-monotonic logic*, see Ginsberg (1987).

[13] See Gauthier (1986) on this point.

Appendix 1

This appendix sets out the explicit form of the overall utility function, which is expressed only in compressed form in the main text by the formula

$$V_i(\sigma) = U_i(\sigma) + \lambda_i F[T(\sigma)] \qquad \text{(A1)}$$

where U_i is player i's material utility for the state σ (a combination of individual strategies); λ_i is a weight that may be any positive (perhaps infinite) number; T is a fairness (to be specified) principle defined for the state σ; F is a function (to be specified) of the fairness principle expressing both the agent's conditioned conformity and other individuals' expected reciprocal conformity to T.

First to be specified is a form of the fairness-function T which represents the ideal formally. This must be a mapping from the set of states (and first-order utilities attached to them) to a fairness ordering ranging over states. A characterisation in contractarian terms of the ideal principle T is given by the Nash bargaining solution i.e. the *N.S.W.F.* $T(\sigma) = \prod_{i=1}^{N}(U_i - d_i)$ where d_i represents the reservation utility that agents can obtain when the bargaining process breaks down. In this case d_i coincides with a covering of the costs of each player's specific investments, which means that fair bargaining on the surplus may only start if parties are assured that they will end up at least with reimbursement of the cost they must bear in order to participate in cooperation.

Then let us define the two personal indexes of conformity which are compounded in the measure F of mutual expected conformity and enter the utility function of the players. In this construction, I take the point of view of player i (any other player j's perspective is symmetrical).

A. Player i's personal index of conditional conformity

This is player i's degree of deviation from the ideal principle T (which varies from 0 to -1) due to player i's choice, given her expectation about player j's behaviour. It is normalised by the magnitude of the difference between players' full conformity and no conformity at all, conditional on player j's choice

$$f_i(\sigma_{ik}, b_i^1) = \frac{T(\sigma_{ik}, b_i^1) - T^{\mathrm{MAX}}(b_i^1)}{T^{\mathrm{MAX}}(b_i^1) - T^{\mathrm{MIN}}(b_i^1)} \qquad \text{(A2)}$$

where b_i^1 is player i's belief concerning player j's action, $T^{\mathrm{MAX}}(b_i^1)$ is the maximum attainable by the function T given i's belief, $T^{\mathrm{MIN}}(b_i^1)$ is the minimum attainable by the function T given i's belief, and $T(\sigma_{ik}, b_i^1)$ is the effective level attained by T when the player i adopts his strategy σ_k (where the index k means that player i's strategy is chosen within a set where k may vary from 1 to N), given his belief about the other player's behaviour.

B. Estimation of the second player index of reciprocal conformity

This is player j's degree of deviation from the ideal principle T (which also varies from 0 to -1), as seen through player i's beliefs – also normalised by the magnitude of the difference between player j's full conformity and no conformity at all, given what j believes (and player i believes that he believes) about player i's choice.

$$\tilde{f}_j(b_i^1, b_i^2) = \frac{T(b_i^1, b_i^2) - T^{\mathrm{MAX}}(b_i^2)}{T^{\mathrm{MAX}}(b_i^2) - T^{\mathrm{MIN}}(b_i^2)} \qquad \text{(A3)}$$

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 concerning player j's belief about the action adopted by player i (i.e. formally identical to a player i strategy predicted by player j).

These indexes are used to construct the following ideal component of the utility function

$$\lambda_i\left[1 + \tilde{f}_j(b_i^2, b_i^1)\right]\left[1 + f_i(\sigma_{ii}, b_i^1)\right] \qquad \text{(A4)}$$

where the weight λ_i is an exogenous psychological parameter that expresses, prior to any consideration of reciprocity, the extent of the disposition to act according to conformist considerations within the motivational system of player i. The formula states the following: if player i perfectly conforms with the ideal, given her expectation, while player j is also expected to perfectly conform, then the two individual indexes take value zero, so that the resulting utility value due to conformism is *(1)(1)* λ_i. Thus the maximum conformist utility value is λ_i. By contrast, if a player does not entirely conform, while not expecting the other player entirely to conform either, then the two indexes take negative values (possibly -1). Thus the utility calculation for conformist reasons reduces to $(1-x)\,(1-y)$ (possibly both equal to zero) multiplied by the weight λ_i and gives

less than λ_i (possibly zero) as the conformist utility value.

The overall utility function V_i is the linear combination of the two components

$$V_i(\sigma_i, b_i^1, b_i^2) = U_i(\sigma_i, b_i^1) + \lambda_i \left[1 + \tilde{f}_j(b_i^2, b_i^1)\right]\left[1 + f_i(\sigma_{ii}, b_i^1)\right] \quad \text{(A5)}$$

This suggests that if a player predicts reciprocal conformism (as it enters the utility function), in so far as weight λ_i is high, it is then possible that the overall utility of a strategy choice reverses the effect of player *i*'s simple consequentialist preferences represented by $U_i(\sigma_i, b_i)$. For example, it may induce the player to select strategies that he would never choose if he relied on material utility only.

When overall utility functions are employed in game theoretical contexts, they require appropriate definition of the players' best-response choices. Grimalda and Sacconi 2005 and Sacconi and Grimalda (2006) elaborate on Rabin (1993) in order to define a new model of reciprocity. Hence, as for Rabin, inclusion of beliefs in the arguments of the utility functions calls for extension from the standard concept of the Nash equilibrium to that of the (PNE) as defined by Genakoplos et al. (1989). The idea behind PNE is that, in equilibrium, the beliefs of rational players must be coherent with the strategies that are being played: that is, beliefs of any level predict lower level beliefs and actions. Hence also the result of fourth section is given in terms of the Psychological Nash Equilibria of the relevant game.

References

Bernheim, B.: 1994, 'A Theory of Conformity', *Journal of Political Economy* **102**(5), 841–877.

Binmore, K.: 1997, *Just playing* (MIT Press, Cambridge, Mass).

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

Camerer, C. and E. Fehr: 2002, Measuring Social Norms and Preferences Using Experimental Games: a Guide for Social Scientists (Institute for Empirical Research in Economics, University of Zurich), WP N.1424–0459.

Capezggi, F.(ed.) 2006, *Reforming Self-regulation in European Private Law*, Kluwer Law International, London, (in print).

Charness, G. and M. Rabin: 2002, 'Understanding Social Preferences with Simple Tests', *The Quarterly Journal of Economics*, August, 818–869.

Clarkson Centre for Business Ethics: 2002, 'Principles of Stakeholder Management', *Business Ethics Quarterly* **12**(2), 257–264.

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

Donaldson, T. and T. W. Dunfee: 1995, 'Integrative Social Contracts Theory', *Economics and Philosophy* **11**, 85–112.

Falk, A., E. Fehr and U. Fishbacker: 2003, 'On the Nature of Fair Behaviour', *Economic Inquiry* **41**(1), 20–26.

Falk, A. and U. Fischbacher: 2000, *A Theory of Reciprocity* (Institute for Empirical Research in Economics, University of Zurich), WP N.6.

Fehr, E. and K. Schmidt: K. 2001, *Theories of Fairness and Reciprocity – Evidence and Economic Applications* (Institute for Empirical Research in Economics, University of Zurich), WP N.75.

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

Fudenberg, D.: 1991, 'Explaining Cooperation and Commitment in Repeated Games', in J. J. Laffont (ed.), *Advances in Economic Theory, 6th World Congress* (Cambridge University Press, Cambridge).

Fudenberg, D. and D. Levine: 1989, 'Reputation and Equilibrium Selection in Games with a Patient Player', *Econometrica* **57**, 759–778.

Fudenberg, D. and J. Tirole: 1991, *Game Theory* (MIT Press, Cambridge, Mass).

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

Gauthier, D.: 1990, 'Economic Man and the Rational Reasoner', in J. Nichols and C. Wright (eds.), *From Political Economy to Economics and Back?* (ICS Press, San Francisco).

Gauthier, D.: 1996, 'Commitment and Choice: An Essay on the Rationality of Plans', in F. Farina, S. Vannucci and F. Hahn (eds.), *Ethics, Rationality, Economic Behaviour* (Oxford U.P., Oxford), pp. 12–14.

Geanakoplos, J., D. Pearce and E. Stacchetti: 1989, 'Psychological Games and Sequential Rationality', *Games and Economic Behavior* **1**, 60–79.

Ginsberg, M. L.: 1987, *Reading in Nonmonotonic Reasoning* (Morgan Kaufmann Publisher Inc, Los Altos, CA).

Grimalda, G. and L. Sacconi: 2002, *The Constitution of the Non-profit Enterprise: Ideals, Conformity and Reciprocity*, LIUC paper n.110 (Catellanza, Varese).

Grimalda, G. and L. Sacconi (2005), 'The Constitution of the Not-For-Profit Organisation: Reciprocal Conformity to Morality' *Constitutional Political Economy* **16**(3), 249–276.

Jensen, M. C.: 2001, 'Value Maximization, Stakeholder Theory, and the Corporate Objective Function' *Journal of Applied Corporate Finance*, **14**(3), 8–21.

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

Kreps, D.: 1998, 'Bounded Rationality', in *The New Palgrave Dictionary of Economics and Law*, (McMillan, London).

Kreps, D. and R. Wilson: 1982, 'Reputation and Imperfect Information', *Journal of Economic Theory* **27**, 257–279.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson: 1982, 'Rational Cooperation in the Finitely Repeated Prisoner's Dilemma', *Journal of Economic Theory* **27**, 245–252.

Lewis, D.: 1969, *Convention, A Philosophical Study* (Harvard University Press, Cambridge, Mass).

McClennen, E.: 1990, 'Foundational Exploration for a Normative Theory of Political Economy', *Constitutional Political Economy* **1**, 67–99.

McClennen, E.: 1993, 'Rationality, Constitutions and the Ethics of Rules', *Constitutional Political Economy* **4**, 173–210.

McDermott, D. and J. Doyle: 1980, 'Nonmonotonic Logic I', *Artificial Intelligence* **13**, 41–72.

Nagel, T.: 1986, *The view from nowhere* (Oxford University Press, Oxford).

Pettit, P.: 1990, 'Virtus normativa. Rational Choice Perspectives', *Ethics* **100**, 725–755.

Phillips, R., E. Freeman and A. C. Wicks: 2003, 'What Stakeholder Theory is Not', *Business Ethics quarterly* **13**(4), 479–502.

Posner, E. A.: 2000, *Law and Social Norms* (Harvard UP., Cambridge, Mass)..

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

Reiter, R.: 1980, 'A Logic for Default Reasoning', *Artificial Intelligence* **13**, 81–132.

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

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

Sacconi, L.: 2004, 'The Efficiency of the Non-Profit Enterprise: Constitutional Ideology, Conformist Preferences and Reputation', in B. Hogdson. (ed.), *The Invisible Hand and the Common Good* (Springer Verlag, Berlin).

Sacconi, L.: 2006a, '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* (Routledge, London) (in print).

Sacconi, L.: 2006b, 'A Social Contract Account For CSR as an Extended Model of Corporate Governance (I): Rational Bargaining and Justification' in *Journal of Business Ethics*, **68** (3), 259–281.

Sacconi, L., S. DeColle and E. Baldin: 2003, 'The Q-RES Project: The Quality of Social and Ethical Responsibility of Corporations', in J. Wieland. (ed.), *Standards and Audits for Ethics Management Systems, The European Perspective* (Springer Verlag, Berlin), pp. 60–117.

Sacconi, L. and G. Grimalda: 2006, 'Ideals, Conformism and Reciprocity: A Model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case', in L. Bruni and P. L. Porta (eds.), *Handbook on the Economics of Happiness*, (Edward Elgar, Brookfield) (in print).

Skyrms, S.: 2004, *The Stag-Hunt and the Evolution of the Social Structure* (Cambridge UP, Cambridge)..

Sugden, R.: 1986, *The Economics of Rights, Co-operation and Welfare* (Basil Blackwell, London).

Sugden, R.: 1998, 'Normative Expectations: the Simultaneous Evolution of Institutions and Norms', in A. Ben-Ner and L. Putterman (eds.), *Economics, Values, and Organization* (Cambridge University Press, Cambridge), pp. 73–100.

Zadeh, L. A.: 1965, 'Fuzzy Sets', *Information and Control* **8**, 338–353.

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

Lorenzo Sacconi
Department of Economics,
University of Trento,
Trento, Italy

and

EconomEtica, interuniversity centre of research,
Milano-Bicocca University,
Milano, Italy
E-mail: lorenzo.sacconi@economia.unitn.it