Analyse & Kritik 01/2011 (© Lucius & Lucius, Stuttgart) S. 273–310

Lorenzo Sacconi, Marco Faillo and Stefania Ottone

Contractarian Compliance and the 'Sense of Justice': A Behavioral Conformity Model and Its Experimental Support

Abstract: The social contract approach to the study if institutions aims at providing a solution to the problem of compliance with rational agreements in situations characterized by a conflict between individual rationality and social optimality. After a short discussion of some attempts to deal with this problem from a rational choice perspective, we focus on John Rawls's idea of 'sense of justice' and its application to the explanation of the stability of a well-ordered society. We show how the relevant features of Rawls's theory can be captured by a behavioral game theory model of beliefs-dependent dispositions to comply, and we present the results of two experimental studies that provide support to the theory.

1. Introduction and Motivations

1.1 The Compliance Problem

The social contract perspective seems appropriate for appreciating the problem of norm (and institutions) compliance. In fact, the social contract approach maintains that norms and institutions must be based on the consensus and voluntary compliance of those regulated by the norm itself. The idea of a rational agreement (i.e. the social contract) must thus be simultaneously developed in two directions: on the one hand, it must work as a justification by giving reasons for agreeing on the norm or institutional rule from an impartial and impersonal standpoint; on the other, the same idea must have direct implications for personal incentives and motivations to comply with the norm in practice. In a 'state of nature', namely a situation of pre-institutional strategic interaction, the norm can be implemented only if the agreement is self-enforceable. In other words, the social contract can resort to no other means of implementation than those which the agreement is able to induce by itself.

David Gauthier clarifies that these two directions pose two separate choice problems, with clearly not concomitant rationality tests, that a contractarian explanation of norms and institutions must nevertheless overcome simultaneously and consistently (see Gauthier 1986, 116–8):

a) *The entry into agreement problem (internal rationality)*: all individuals, when deciding whether to enter an agreement on the institution of a norm, perform a rationality assessment on whether the norm will enable them to escape from a reciprocally unprofitable interaction and permit them to initiate a mode of mutually beneficial cooperative interaction. This point of view requires *internal* rationality because it views the agreement from *within* the perspective of cooperative bargaining—which takes it *for granted* that if an agreement is reached, it will be implemented to the mutual advantage of the participants. Hence no *ex post* decision (after entrance) is relevant here. On the contrary, this case requires that entering the agreement *ex ante* may be recognized as mutually beneficial. Rational bargaining takes place in situations where there is some feasible surplus to be distributed amongst the individual participants, granted that they are able to reach an agreement. But there are too many agreements possible: some of them preferred by one party, others by another. A bargaining game is a way to solve this coordination problem before the cooperative game is played whereby the agreed joint strategy is executed in order to produce and allocate the cooperative surplus.

b) *The compliance problem (external rationality)*: when we move from the *ex ante* to the *ex post* perspective, we ask whether an agreement reached can also be complied with by the same players who agreed on it. This is a different problem because the game-logic of compliance differs from that of entering a bargain in a cooperative game. It is instead the logic of an *ex post* non-cooperative game in which the players decide separately but interdependently whether or not to comply with the *ex ante* agreed contract. From this perspective, the question is not so much whether the contract provides reasonably high joint benefits and distributes them in an acceptably fair way; rather, the question is mainly whether there are incentives for cheating on the counterparty to the agreement, given the expectation that s/he will abide by the contract. Thus, according to Gauthier, the search for external rationality must address the problem of a potential divorce between individual rationality (expected personal utility maximisation) and social optimality (i.e. Pareto efficiency), a divorce which is instantiated by the typical Prisoner's Dilemma game

Impartiality within a contractarian framework amounts to no more than a condition of invariance for the *ex ante* acceptance of a given bargaining outcome, which means acceptance from the viewpoint of each and all (under the permutation of personal standpoints allowing the impartial decision-maker to take each player's point of view in turn). By contrast, compliance is the typical sphere in which *ex post* (personal, *not* impersonal and impartial) rationality is required. In the compliance problem, separate but interdependent strategy choices are under consideration, and the players are always able to say whether or not they want to implement the agreement given their prediction of the other player's decision whether or not to comply with it. It follows that the main problem to be solved in the compliance context is how a norm can also generate motivational causal

forces strong enough to induce the execution of the norm in situations where it may require a *prima facie* counter-interested behaviour by the agent at least in the immediate term.

Justifications in themselves do not answer questions about compliance. 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 the fact 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).

Assume, however, that preferences are typically consistent with the model of self-interested rational individual choice and that, following a large part of the rational-choice literature on the emergence of institutions (Gauthier included), the Hobbesian 'state of nature' is modelled like a Prisoners' Dilemma (PD) game. Assuming a PD-like state of nature seems natural if one's goal is to make evident the need for the social contract. It results in a pragmatic necessity in so far as it is a rational means for exiting a mutually destructive 'state of nature' which inevitably entails a suboptimal solution of the rational and egoistic non-cooperative interaction amongst players. Since the individually rational non-cooperative mode of interaction necessarily engenders a suboptimal outcome, in order to escape from this outcome it is a pragmatical necessity explicitly to agree on a different (mutually cooperative) mode of behaviour and outcome. The Hobbesian 'state of nature' is perfectly captured in the PD-logic. The necessary suboptimal solution of the game, consisting in the unique equilibrium outcome in dominant strategies, corresponds to the idea of a necessary 'war of all against all'.

However, this argument on the logical consequentiality of the state of nature suboptimal outcome is also what makes the compliance problem in the social contract perspective apparently insolvable. If the game within which players must decide to comply with the social contract is a PD, as seems unavoidable if the 'state of nature' wherein they are embedded is a PD, then the social contract will obviously not be complied with, since non-compliance (naturally corresponding to mutual defection) is the only equilibrium point. And this difficulty cannot be overcome by proceeding to the second step in the argument, where the social contract includes delegation of authority to one party in the agreement to enforce the contract. Agreeing and complying with the social contract, and hence accepting to conform with any authority directives (which is the content of the contract and consists in legitimizing authority by agreement) is again a deal that should be struck and respected within a PD situation. But according to the PD-logic, compliance with this authority-delegation agreement

(i.e. cooperating in the PD) cannot be consistent with individual incentives to act. Thus players may agree *ex ante* in 'cheap talk' to relinquish their complete freedom and submit to an authority by accepting its directives. But as a matter of fact they will not transfer their 'natural' capability to act to the delegated authority—i.e. they will not fulfil the cheap-talk promise to comply with the authority directives. Thus, once the implementation stage has been reached, authority will vanish.

Here the compliance problem takes a particularly marked form which we call 'contractarian compliance impossibility'. The problem can be stated as follows: assume that a PD-theory of the 'state of nature' provides the complete theory of the social contract, i.e. a complete description of the choice situation wherein a social contract is to be complied with. Thus, keeping the social contract cannot be consistent with the premises concerning a rational solution of this type of game. In other words, social contract compliance is not a theorem that can be deduced from a PD-theory of the state of nature. Then, by freely rephrasing the famous Gödel incompleteness theorem, if the PD theory of social contract is complete it must be inconsistent. On the contrary, let us require that the social contract (including compliance with it) is consistent with the premises on rational action in the choice situation (the state of nature). In other words, it should be a theorem of the theory designed to give a solution to the game understood as the formal model of the 'state of nature' situation. Hence the PD cannot be a complete description of the choice situation. Some additional features of the game must be added to the representation of the state of nature, which by itself necessitates a suboptimal solution of the non-cooperative interaction played by selfish and rational players.

But of course if the game description and assumptions are wider than a PD, the necessity of a suboptimal solution of the state of nature interaction that highlights the need for a social contract is no longer guaranteed. Otherwise, it should be shown *how*—given the non-cooperative PD game that players are doomed to play in the 'state of nature' if they avoid entering the social contract—having considered the cooperative decision to enter the social contract (an *ex ante* justificatory decision), this mental experiment of justification *will change* the nature itself of the *ex post* compliance problem and the structure of the corresponding game (as we will consider in this article).

1.2 Ways Out of the Compliance Problem: Nash Equilibria

The difficulty of the compliance problem is significantly reduced if norms and institutions are viewed from the conventionalist perspective (Hume 2000[1740]; Lewis 1969; Sugden 1986; Aoki 2001). Conventions are regularities of behavior that give solution to coordination games repeatedly played by (at least) pairs of players in a given population. A regularity of behavior is a convention if players, who are all following the regularity, are characterized by a system of mutually consistent expectations regarding actions and beliefs such that (i) each player correctly expects that any other player will follow the regularity, and also expects that other players correctly expect that s/he will follow the regularity; (ii)

these expectations are common knowledge—i.e. everybody knows that everybody knows... that everybody have the expectations that all players will follow the regularity; (iii) given these expectations each player prefers to follow the regularity of behavior (since it is the solution of a coordination problem), and this interest is also common knowledge.

Under these conditions, compliance reduces to a trivial problem since conventions are Nash equilibria, and intrinsic to the definition of a Nash equilibrium is that it is a strategy vector such that no player having a component in it has any incentive to change his/her strategy in order to deviate from the behavior consistent with the strategy vector. Moreover, conventions are Nash equilibria endowed with the conditions that make it individually rational for each player to play exactly his/her strategy pertaining to a given equilibrium. In fact, common knowledge of the game solution entails that the solution cannot but be a Nash equilibrium (otherwise, by knowing the solution every players would deviate from it, thus destroying the solution itself: see Luce and Raiffa 1957). Thus, in so far as players commonly know that a given Nash equilibrium is the solution of the game, each of them cannot play any strategy other than his/her component of the given equilibrium. But conventions are exactly situations where agents commonly know that a given regularity, which is a coordination equilibrium, is being followed by all the agents. Thus nobody has any incentive to deviate from it.

An important aspect of conventions is the idea of a many-level systems of mutually concordant expectations, which is considered to be one of main the reasons why agents comply with a rule. According to such a system, players expect reciprocation of compliance with a rule, and they also expect that reciprocation is mutually expected. But, as clarified by Lewis's definition, mutual beliefs of reciprocation of the same behavior are not *per se* reasons for compliance. An independent reason in term of instrumental rational choice is needed. Under mutual expectations of reciprocity, compliance must consist in a best response in terms of independently defined individual payoffs. It is true that after a convention has been working for some time as a successful coordination mechanism of the players' choices, a reason *per se* for continuing reciprocation may spring from the simple awareness of mutual expectation of compliance (Sugden 1986; 1998b; Bicchieri 2006). But this comes only after a convention has emerged, and it presupposes that all players basically know (or at least have known in the past) that playing according to the rule of behavior is an *instrumental best response* given their mutually consistent expectations and belief systems.

How far do conventions extend their domain of application? There is no reason for limiting the latter only to cases of pure coordination. It can be conveniently extended to cover mixed motives coordination problems (Sudgen 1998). But it is certainly beyond the reach of convention theory to give a solution to the compliance problem as it was modelled in the social contract perspective, i.e. to assure compliance with agreements capable of solving PD-like interactions. The reason is simple. The aim of the social contract perspective is to provide a solution to compliance problems in situations—like the Prisoners' Dilemma or the Trust Game (TG)—in which there is a divorce between individual self-

interested rationality and social optimality. In a TG, for example, the trustor cannot trust the trustee in equilibrium, because after the former has entered into a relation with the latter, the latter's best response is to abuse the former's trust by defecting from the mutually beneficial behavior. If in such games (PDs or TGs) conventions were possible so as to support a significant level of cooperation, it would be simple to suggest agreement on these rules of behaviors. But such games simply have no conventions at all.

It is certainly true that also these games take the form required for the existence of conventions when they are infinitely repeated so that folk-theorems apply, and consequently numerous equilibria come to existence, some of them inducing substantial cooperation. However, whereas convention theory says too little about the solution of the original one-shot games, now it says too much about these repeated models. Here there is no longer the typical failure of individual rationality that would entail a suboptimal outcome of individual self-interested non-cooperative interaction. On the contrary, there are numerous equilibria that realize a considerable level of coordination to the mutual advantage of players. All these equilibria can be interpreted as 'spontaneous orders' that may emerge from tacit adaptive interactions among non explicitly cooperating players. In principle, a social contract is not necessary for the emergence of such a spontaneous order. Evolution, trial and error, 'liberation' or cumulative learning processes (Bayesian or otherwise), or mere contextual cognitive *salience*, may work as well, without entailing the *ex ante* exchange on the mutual promises and obligations that are typical of an agreement (as Hume pointed out).

However, a serious problem also concerns repeated PD or TG games where many (or infinite) conventions are in principle possible. The multiplicity problem amounts to complete uncertainty about which of the possible conventions will emerge from the non-cooperative interaction among players considered as individual agents making their choices given their personal preferences and expectations on other players' choices in a non-cooperative context. Until this uncertainty is removed, neither can the compliance problem be solved. In fact, only when a commonly known system of mutually consistent expectations is achieved, one which converges on the prediction of a specific equilibrium as the game solution, one can assume that each player has a uniquely determined best response consisting in abiding by the strategy pertaining to the given equilibrium convention. Until this common knowledge state is reached, there is no stability and mutual consistency in the decisions whereby players try to achieve some possible outcome of the game corresponding to the social contract. Assume that some *ex ante* justification has been established concerning the impartial acceptability of a particular outcome, but that *ex post* the multiplicity problem has still not been solved. Thus each player remains uncertain about which equilibrium combination will be *de facto* played by all the players. And there is no univocal answer to the problem of compliance with the agreement in terms of reciprocal best responses. This can be seen as the main potential failure of individual non-cooperative rational choice.

Since the condition for a convention to exist is the formation of the required system of mutually consistent expectations on actions and beliefs of whatever level, the question is now how such a system can form in the minds of all the interacting players, so that each of them will be able to predict that a given norm is in fact the regularity of behavior currently being played in the resolution of the given game. A natural suggestion is thus to consider the social contract as one, if not the main, tool with which to solve the multiplicity problem; that is, to consider the agreement as an equilibrium selection device. By solving the *ex ante* decision problem (entry into the agreement) consistently with the requirement of selecting just one equilibrium, the agreement may render the *ex post* decision problem trivial. It can immediately induce compliance with the selected-by-agreement equilibrium point that players predict as the game solution.

In fact, some authors (Hampton 1986; Sacconi 1991; 1993a,b; Skyrms 1996; Binmore 1989; 1994; 1998; 2005) have suggested a similar way to extend the conventionalist perspective so as to transform it into an allied to social contract theory. Indeed, they all reduce the social contract to an equilibrium selection procedure. Thus the compliance problem is greatly simplified by assuming that the agreement space is confined to a subspace of the possible *ex ante* outcomes—i.e. is a subset coinciding with the set of the game equilibrium points (or some subset properly defined of the equilibrium set). Thus, when the agreement selects an equilibrium, compliance naturally follows.

The simplest example of this line of reasoning is the Stag Hunt game as the proper formal representation of social contract problem (Hampton 1986). The social optimal is hunting a stag with coordinated players' efforts, but lack of trust among hunters will induce them to hunt one hare each with separate efforts, which is the risk-dominant strategy for each player (i.e. there is no risk of failure in hunting hares individually). This equilibrium, reached though non-coordinated actions, can be seen as the 'state of nature' suboptimal outcome. But now assume that reaching an agreement on hunting a stag is an effective way to create a complete state of trust among the players about the hypothesis that if one tries to hunt the stag, the others will do the same. This agreement helps reach the optimal outcome. Since it is an equilibrium supported by full trust, equivalent to having a system of mutually consistent expectations converging on the prediction that the stag will be hunted by all, it also assures full compliance.

However, the stag hunt is too simplified a situation to be considered a proper representation of the multiplicity problem. Let us then consider Binmore's much more elaborate joint solution of the social contract selection and compliance problems (Binmore 2005). The underlying situation is the 'game of life', namely a repeated evolutionary PD-like game played by agents taking asymmetrically powerful social roles (Adam and Eve). The set of all the evolutionary histories of this repeated game defines a set of possible outcomes composed of all the equilibria of the repeated game (according to some version of the *folk theorem*). The outcome space itself is asymmetrical because it reflects asymmetry in the strategic opportunities available to the players, each occupying the role of Adam or Eve. Taking the set of possible repeated equilibria as given suggests consid-

ering the matter from an *ex ante* perspective, as if it were possible to decide once for all the evolutionary equilibrium path that the two players will follow among the many possible. Thus the *ex ante* agreement works as an equilibrium selection device.

What is remarkable in Binmore's view is that he models this equilibrium selection procedure as an *ex ante* social contract in the proper sense: that is, as a model of justifiable *ex ante* choice able to the agreement to be entered from an impersonal and impartial standpoint or (in other words) from behind a 'veil of ignorance'. Impersonality is captured in the model by allowing the players to exchange their roles (Adam and Eve) so that each player may take both the roles and the relevant payoffs. The original equilibrium space is thus complemented with another representation of the same equilibrium space resulting form its symmetrical translation with respect to the Cartesian axes whereon each players' utility function and payoffs are represented. Thus each outcome that in the first space gives to a player the payoff associated with the advantaged role of Adam is also represented in the translated outcome space by a symmetrical outcome giving to the same player the payoff associated with the disadvantaged role of Eve, and *vice versa*. Impartiality is captured in the requirement that only equally probable convex combinations of pairs of symmetrical outcomes—the one belonging to the original equilibrium set and the other belonging to its symmetrical translation—must be considered as candidate agreements. In other words, an outcome can be considered a candidate for agreement only under the condition of also considering the same outcome with the players' roles reversed as a candidate for agreement with equal probability. Once an equilibrium outcome has been selected for agreement, each of its payoffs can be assigned to both the players with equal probability. Thus the agreement may only fall on the bisector of the equilibrium spaces.

However, the convexity of the outcome space engendered by taking together the original space and its symmetrical translation plus all the convex combinations of any pair of outcomes belonging to the two spaces is not admitted. In fact, this would amount to assuming that we have exited the state of nature so that agreements can be reached on whatever probability combination of any possible outcomes, because any outcome of an agreement can be enforced by an external authority. But this is not the case. In the state of nature only equilibrium points can be agreed on, whereas many convex combinations of outcomes belonging to the two symmetrical outcome spaces do not belong to the initial equilibrium space. This amounts to restricting the set of possible agreements to the symmetrical intersection of the original equilibrium set and its symmetrical translation with respect to the utility axes. Any probability combination of two points belonging to this symmetrical outcome space corresponds also to an equilibrium point of the original outcome space. Furthermore, this correspondence is assured for all the equally probable combinations of symmetrical outcomes, because they lie on the bisector of the symmetrical subset resulting from the intersection of the two outcome spaces. Having restricted the possible set of agreements to the symmetrical intersection subset, and especially to its bisector, it is natural to think that the social contract will coincide with the maximal Nash

bargaining product within this symmetric outcome space. In other words, the contractarian solution is the egalitarian Nash bargaining solution pertaining to the symmetric intersection set, which is also the maximin solution with respect to the original outcome space.

Summing up, under the conditions of a veil of ignorance (impersonality, impartiality and empathetic preferences) and self-sustainability through equilibrium choices (which entails no convexity of the feasible outcome space taken as agreement domain) the solution is Rawlsian. Equilibrium selection is accomplished through two steps: the first refines the equilibrium set by confining it to the symmetrical subset of the original space (i.e. the symmetrical intersection set); then the Nash bargaining solution selects a unique optimal agreement on the bisector of this symmetrical set which corresponds to the Rawlsian maximin.

However Binmore's results should not be overemphasized as far as the equilibrium selection problem is concerned. What would effectively solve the multiplicity problem is an equilibrium selection theory able to predict the *ex post* game equilibrium solution so that it is consistent with the *ex ante* solution identified. In other words, selection is *ex post* effective only if it gives reasons to act that fit the *ex post* reasoning context. *Ex post*, only common knowledge of the solution—that is, a system of mutually consistent expectations converging on the prediction of a uniquely determined equilibrium point—conveys to each player the appropriate reason to act, because choosing an equilibrium strategy amongst many others requires having a clear prediction of other players' behaviors and beliefs. However, there is no logical reason to conclude from the fact that in the *ex ante* perspective a solution is invariant to the players' position replacement that that solution will be effectively implemented *ex post*. The reason that explains a particular decision in the *ex post* game is knowledge of what the players will effectively do. Moreover, this knowledge about the other players' decisions must be consistent with their being symmetrically able to predict the others' behavior and to choose their best response to those predictions. Therefore, it is not the impartial selection of a desirable *ex ante* solution, but the knowledge of other players' *de facto* behaviors, that provides the proper reason for acting in the *ex post* context. Moreover, there is no logical implication from what is fair *ex ante* selection (even if it falls on an equilibrium point) as to what other players will actually do. Maybe they will act in accordance with the principle, maybe they will not. The fair *ex ante* agreement, or impartial choice, does not give common knowledge of the *ex post* behavior of players. If, however, one does not know how other players will behave, one has no reason to play a given strategy, even though the fair solution is part of an equilibrium point.

This is not to say that the *ex ante* agreement on an impartial solution does not provide any reason to believe that players will act according to the same principle in the *ex post* interaction. But this is simply a matter of fact, or of cognitive psychology; it is not a matter of logic. Common knowledge, on the contrary, is a matter of epistemic logic: which means recursive group knowledge of what everybody knows to be true (a *truism*).[1] It is the case that a given

[1] The *ex post* rationality of the Nash equilibrium—implied by the notion of common knowledge—was already clear in Lewis 1969, who also suggested that an agreement could

equilibrium is *commonly known* to be played only if each player has many layers of knowledge about every other player's action, beliefs, beliefs about beliefs, and so on, that are consistent and justify the prediction that this equilibrium will be played. This state of knowledge can be approximated by a theory of belief formation that at last leads to a stable prediction of any other player's equilibrium choice and belief. *Ex ante* selection, by contrast, does not predict how one will actually decide; it only answers the question of what equilibrium *should* be chosen, because it is invariant under the individuals' position replacement. The step from an answer to the question of which equilibrium is *fair* to an answer to the question of how players will *actually* behave is a *default inference* that some player may in fact make; but this is only a possibility. Thus, from the perspective of the *ex post* game, there is still much to do before the multiplicity problem is solved.

Let us add some exemplifications of how the *ex ante* social contract could not effectively work as an *ex post* equilibrium selection device and consequently also as a solution of the compliance problem. Assume that the game of life is a repeated TG, so that the player in the role of Adam is the trustee, while the player in the role of Eve is the trustor (who typically occupies a disadvantaged role in this game). The equilibrium set of this game includes an entire outcome region resulting from mixed strategies of the trustee, who combines the abusing and not abusing strategies, and from the strategy 'to enter' used by the trustor. The latter acquiesces to the trustee's abuse because of the minimal positive payoff that these mixed strategy equilibria give to him. Amongst the outcomes of these equilibria there is also the one that gives Stackelberg payoffs to players. According to these payoffs, the trustor is reduced to complete indifference between entering and staying out, while the trustee reaps practically all the surplus. This is clearly the preferred equilibrium from the trustor's perspective. It can be simply shown that under a veil or ignorance the egalitarian Nash bargaining solution selects the mutually advantageous outcome (Sacconi 2010b), which is also the perfectly symmetrical equilibrium of the repeated game. But consider the possibility that what has been agreed under the veil of ignorance is not enough to convince one player that everybody else is going to play such an egalitarian solution, and, moreover, that it is not sufficient to convince all players that others will expect such an equilibrium to be the game solution. Thus the trustee is quite uncertain about the equilibrium that will be selected by the trustor in response to his entrance. He may guess that, since the trustee prefers the Stackelberg payoff so intensely, he will put as much effort as possible into selection of the corresponding equilibrium. Since this reasoning can be replicated in the trustor's mind, there is at least a reasonable line of thought that seems to induce the players to conclude that the more probable equilibrium that

---

give an empirical explanation of how a state of common knowledge could emerge. However, Lewis focused on the different cognitive phenomenon of salience. On the game theoretic definition of common knowledge, see Binmore/Brandenburger 1990; Kreps 1990; on the epistemic logic of common knowledge, see Fagin/Halpern/Moses/Vardi 1996. On the selection of Nash equilibria based on common knowledge of the unique solution see Harsanyi/Selten 1988.

the trustee will select after the trustor's entrance is the one giving Stackelberg payoffs (Sacconi 2011; Andreozzi 2010).

Similarly, take the Stag Hunt as the game to be considered under the hypothesis that an *ex ante* agreement is not able to engender complete certainty about the fact that both the players will play the optimal equilibrium consisting in hunting the stag. But only a little uncertainty may in this case have a dramatic consequence. If a player is uncertain about the prediction concerning the optimal solution, then the risk dominant argument regains its force, and it is able *ex post* to induce players to shift to the hare-hunting equilibrium. Obviously, this would destroy the entire result from the *ex post* perspective. Thus this section suggests that Nash equilibria are not sufficient in order to give a convincing solution to the social-contract-compliance problem.

1.3 Ways Out of the Compliance Problem: Psychological Dispositions

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 (Gauthier 1986; 1990; 1994; McClennen 1990a,b; 1993). A disposition would constrain later choices, so that the agent can disregard local incentives even if these imply that there are local advantages to deviating from the action plan corresponding to the disposition.

These attempts to overcome the compliance problem seem not to have been successful. On the one hand, the revision of the instrumental rationality required for a theory of disposition choice seems to presuppose what it should demonstrate. The choice of a disposition seems to be very similar to the decision to undertake a conditional binding commitment, which is obviously problematic in that the compliance problem is assumed to have a PD-like structure that prevents assuming that such binding commitments are possible. If binding commitments are allowed, of course, the proposed line of argument is not a reform of instrumental rationality at all—it only amounts to a perhaps reasonable change of the game considered. It *seems* to reduce morality to instrumental rationality by showing that abiding by a norm of conditional cooperation is rational. But in doing so, it must *presume* that 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 or not to develop them—they are also presumed to command our later behaviours, being immune to opportunistic changes when these seem profitable, as if these choices were beyond our control.

On the other hand, the situation becomes quite problematic if we try to explain how developing a conditional disposition to abide by a norm of cooperation may be reduced to a question of instrumental rationality and practical deliberation. This amounts to demonstrating that it is 'rational' to decide to be that kind of person who acts according to a conditional disposition to comply with the norm, even before the disposition is capable of constraining our behaviour and even if we could also devise dispositions able to cheat other players simi-

larly involved in cultivating conditionally cooperative dispositions. An example is a deceitful disposition that continues to dispose the player, who undertakes it, to conditionally cooperate until another player interacting with him 'reads' the disposition itself, but then changes the disposition in order to allow the player exploiting the second player's disposition to cooperate (Danielson 1992). In short, this line of reasoning seems bound to produce many sorts of contradictions (see Binmore 1994)

What seems mistaken in this approach, however, is not the idea of analyzing moral dispositions but the idea that undertaking moral dispositions may be a matter of practical reasoning and sophisticated instrumental decision calculus, whereas it could be a matter of developing a moral sentiment (the 'desire' to be just) endowed with some motivational force on its own, and capable of generating additional motivational drives to act that can be introduced into the players' preference systems—under proper conditions to be defined. If this simple idea is accepted, the desire to comply could be an input to the compliance decision, not the output from a reform of the decision theoretic machine, and we would only need to understand how this desire may be engendered and how it is connected to the social contract. This could enable us to discover other—quite different—causal connections between the decision to comply and the rationality of an *ex ante* agreement.

A similar approach to the compliance problem was suggested by John Rawls in the *Theory of Justice* (1971), where he proposed the 'sense of justice' as a solution for the stability problem of a well-ordered society—i.e. a society whose institutions are arranged according to principles of justice (norms in our sense) chosen under a 'veil of ignorance'. This solution, however, was for long overlooked by economists and game theorists because it was at odds with the methodology of rational choice in that it resorted to socio-psychological assumptions common in theories of moral learning.[2]

[2] Rather ironically, we have seen that Rawls's theory of the social contract has been vindicated by its important translation in game theoretical terms, with the proof that, given a set of non-cooperative equilibria resulting from the natural evolution of society, the only selection compatible with both the feasibility condition of equilibria and the 'the veil of ignorance' (invariance under the personal position symmetric replacement) is Rawlsian maximin or the egalitarian solution (Binmore 2005). This means that if one wants to implement a choice under the veil of ignorance through an equilibrium point that guarantees incentive compatibility, then one must focus on equality or the maximin solution. This is ironic, because if Rawlsian principles for institutions were stable in the Nash equilibrium sense—that is, if they provided the only equilibrium selection mechanism under the veil of ignorance, whereas other principles, like utilitarianism, would be unfeasible and not implementable—they would be complied with only for self-interested reasons and would dispense with the emergence of the sense of justice. But this argument only proves that Rawls's proposal is superior to any other social contract solution under the veil of ignorance if what is required is making a selection *within* the set of equilibria emerging because of natural evolution of the game of life. Rawls's general argument, however, could be understood as not imposing *ex ante* the constraint that a fair agreement should be confined *within* the naturally evolved set of equilibrium points since some fairer outcome should result that is beyond the reach of the *ex ante* agreement. McClennen (1990a) took this part of Rawls's theory as the basis for his own approach to the stability of the constitutional contract. Similarly to Rawls and Gauthier, McClennen thinks that compliance is a disposition to cooperate conditionally on other players' cooperation. However, he finds Rawls's solution weak in so far as it postulates the sense of justice as an attitude which is

However, given the behavioral turn in microeconomics, it is time to reconsider this neglected solution and to acknowledge that it may suggest an illuminating explanation of why some of us comply with just institutions even if we have some direct material incentive not to do so.

1.4 Plan of the Paper

The *next section* summarizes Rawls' argument about how a sense of justice is engendered by the awareness that principles of justice followed in a 'well ordered society' have been agreed on under a veil of ignorance and how this attitude offers endogenous support to the well-ordered society's stability. It then suggests that the relevant features of Rawls's theory captured in the behavioral model of contractarian conformist preference that we will follow in this paper (Grimalda/Sacconi 2005; Sacconi/Grimalda 2007; Faillo/Sacconi 2007; Sacconi/Faillo 2010). This model is set out in *section 3*. It explains how a disposition to conform with agreed principles of justice, conditional on beliefs concerning other agents' choices and their expected reciprocity, may enter the preference system of a player by assigning psychological payoffs—additional to material payoffs—to choices that approximate an ideal of justice given the other players' choices and their level of conformity. Since psychological payoffs are a function of the players' reciprocal beliefs of first and second order about their levels of conformity, the model is based on psychological game theory (Genakoplos et al. 1989; Rabin 1993). But it introduces the completely new idea that conformity preferences depend on an *ex ante* impartial agreement on a principle. Then in *section 4* the contractarian conformist theory is applied to a special game, the *exclusion game*, explicitly devised for the purpose of representing a

---

engendered by the creation of the well-ordered society's institutions, chosen under a veil of ignorance, but it is not a matter of rational choice over dispositions as such. According to McClennen, Rawls's approach is exogenous with respect to the mechanism of rational choice, and he seeks to endogenize the sense of justice through his theory of resolute choice extended to the context of non-cooperative games (McClennen 1993). Resolute choice in these contexts means that a player undertakes by a decision a disposition that commits himself to forgoing, at some later decision node in the game tree, opportunities for defection which are locally advantageous (so that locally defecting can be dominant). The reason for doing so is a requirement of consistency with an initial plan, which—when followed by all—permits players to achieve collectively higher payoffs and to fare better. Of course this idea cannot work if players have the effective option of defecting at a later decision node where they find that it is locally rational to do so because of local incentives (for example in a last stage having the Prisoners' Dilemma structure without reputation effects) Hence McClennen suggests reform of the theory of rational decisions in games, admitting that in order to allow this kind of choice among effective dispositions able effectively to commit players, we should relinquish hypotheses such as the game tree's separability into its sub-games, and consequently renounce the possibility of truncating sub-trees and substituting them with their local solutions (when available). He concludes by discarding backward induction (also when it could provide uniquely determined solutions). Our opinion is that this reform of rationality criteria is too costly to game theory, whereas Rawls's perspective is endogenous enough for the endeavor to explain the emergence of a sense of justice as a set of attitudes governed by a disposition having motivational force on its own (the force of a desire) because it is grounded on the *ex ante* decision under a veil of ignorance and it influences the cognitive mechanisms of expectations formation and motivations formation, leading to a preference capable of commanding a decision behavior.

social situation wherein only adhesion to impartial distributive justice principles may allow escape from an egoistic equilibrium of the devision problem on a given surplus such that some players are completely discriminated against in the distribution. This section shows how an agreement under the veil of ignorance not only selects an impartial principle about the sharing of a surplus, but may also activate contractarian conformist preferences. These transform the payoff structure of the game by engendering psychological equilibria that represent the sense of justice effectuality in inducing endogenous compliance with principles of justice selected under the veil of ignorance.

Let us recall that the theory is not confined to the kind of games wherein the mutual advantage of decision-makers cannot support a fair social contract. Elsewhere (Sacconi 2011) it has been proved that, in the repeated TG, conformist preferences are effective in refining the equilibrium set up to only two psychological equilibria. However, the game that we discuss in this paper seems more akin to the Rawlsian perspective, in so far as in this game no Nash equilibria may support a fair social contract that allows inclusion of all the players in the sharing of a social surplus. On the contrary, psychological equilibria based on conformist preferences—with which we formally represent the 'sense of justice'—provide an endogenous explanation of social contract compliance.

The remaining sections of the paper are devoted to reporting experimental support for the conformist contractarian theory. The experiments considered (see also Sacconi/Faillo 2010; Faillo/Ottone/Sacconi 2008) are naturally situated in the ever-growing literature on experimental games aimed at investigating non-selfish economic behavior, social preferences, reciprocity, and the importance of social norms in understanding players' preferences and behaviors (for example: Rabin 1993; Fehr/Schmidt 1999; Falk/Fischbacher 2006; Levine 1998). Compared with existing experimental studies on social preferences and reciprocity, the experiments reported here consider the relevance of an aspect that no other study has to date considered: the importance of an impartial agreement under the veil of ignorance in inducing a strong change in the players' behavior with respect to conformity with principles of justice that have been agreed, although without assuming binding commitments, and not presupposing either externally enforcing mechanisms or the possibility of reputation effects.

The first experiment reported was based on an intra-group design that experimentally replicated the exclusion game logic. A group of experimental subjects anonymously participated first in an experimental version of the exclusion game with assigned roles. They were then allowed to participate in a voting procedure under a veil of ignorance (i.e. ignoring their role in the game). The aim was to reach unanimous agreement on principles of division.

Last, if the subjects had been able to agree, they could play the exclusion game again with roles reassigned. The second experiment was based on an intergroup design where distinct groups of subjects were confronted (i) with the exclusion game as such (with no opportunity for *ex ante* agreement), and (ii) with a treatment where they were allowed to participate in a pre-play stage of impartial agreement on division principles under a veil of ignorance about their roles (voting procedure), before—admitted that they had agreed on some

principle—they had the opportunity to play the exclusion game with assigned roles. A third treatment was introduced into this experimental design in order to compare levels of conformity with principles in two cases: (a) when principles had been agreed under a veil of ignorance by a group of anonymous players who were then asked to play the exclusion game after the role assignation; (b) when after the agreement the composition of groups was partially changed so that each group of strong players was matched with an outsider who as far as they knew might had agreed on a different principle. What this experimental design sought to investigate was whether it is the impartial agreement on principles, plus expectations about reciprocal conformity by others, or the simple expectations of others about our behavior even if we have not accepted the principle on our own, that is the main source of conformist preferences and norm compliance. An essential aspect of studying this matter is elicitation of first- and second-order beliefs about others' choices, as well as elicitation of normative judgments and second-order normative expectations (what one thinks that others normatively expect from one).

2. The 'Sense of Justice'

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

To understand how this is possible, it is necessary to consider the definition of 'sense of justice'. Although the latter presupposes the development of lower-level moral sentiments of love and trust, understood as feelings of attachment to lower-level institutions (families and just associations) perceived to be just, the sense of justice is a desire to act upon general and abstract principles of justice as such, once they have been chosen under a veil of ignorance as the shaping principles of institutions, and hence have proved beneficial to ourselves in practice. Note that it is not the case that we act upon the principles insofar as they are beneficial only to the concrete persons with whom we have direct links and emotional involvements. Once the level of a morality of principles has been reached, our desire to act upon the principles does not depend on other

people's approbation or on other contingent factors such as satisfaction of the interests of some particular concrete person. On the contrary, it is the system of principles of justice in itself that constitutes the object of the sense of justice.

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

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

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

The proper functioning of the sense of justice can be understood, however, as the third level of a process of moral learning which in its first two steps already cultivates moral sentiments of love for parents and trust and friendship vis-à-vis the members of just associations in which the individual already takes part—and which s/he re-elaborates on those pre-existing sentiments.

> "Given that a person's capacity for fellow feeling has been realized by forming attachment in accordance with the first two [...] [levels] and given that a society's institutions are just and are publicly known to be just, then this person acquires the correspondent sense of justice as he recognized that he and those for whom he cares are the beneficiaries of these arrangements." (Rawls 1971, 491)

As seems clear, reciprocity is a basic element in this definition. In fact, reciprocity is understood as a deep-lying psychological fact of human nature amounting to the tendency to "answer in kind". The sense of justice

> "arises from the manifest intention of other persons to act for our good. Because they recognize they wish us well we care for their

> well being in return. Thus we acquire attachment to persons and institutions according to how we perceive our good to be affected by them. The basic idea is one of reciprocity, a tendency to answer in kind." (494)

Two aspects are to be noted concerning the other person's 'manifest intention' which elicits the tendency to 'answer in kind'. We recognize an *unconditional* caring for our good deriving from other people acting consistently with the principles of justice. Hence reciprocity is elicited not from the mere coherence of institutions with the principles of justice, but from the fact that other people make our good by acting intentionally upon those principles. What matters is not just reciprocity in accepting the principles, but the intention displayed by other players concretely acting upon the principles for our well-being. Secondly, this intention cannot be a direct intention toward us as particular persons. By complying with principles, our good is pursued in an unconditional way—that is, impartially and not conditionally on any particular description of us based on contingent characteristics or positions.

Summing up, we may reconstruct the hypotheses that according to Rawls must be satisfied in order for a sense of justice to evolve:

a) lower level moral sentiments must have fostered our capacity for a sense of justice; they are exogenous factors pertaining to the psychological make-up of the person and affecting his/her emotional capacity;

b) we recognize that ongoing institutions (norms) are just because we are able to justify then in terms of their acceptability under a veil of ignorance agreement;

c) it is public knowledge that institutions are just, which seems to mean not only that we know that they are justified, and we know that also other individuals know that they are justified, but also that we publicly know that they effectively operate for the most of the time in accordance with the principle of justice;

d) from the facts that we publicly know that institution are just, and that others know that they are just and work according to the principles of justice, it follows that other individuals conform with the principles and hence are our beneficiaries in an unconditional way, and we know that they are;

e) under the foregoing conditions, everybody is driven by a deep psychological tendency to answer in kind, which means replicating conformity with the principles, given that conformity with principles by others expresses an intention to be beneficial to us in an unconditional and impartial manner.

When these premises are satisfied, the sense of justice develops, and becomes an integral part of our conception of the good. That is to say, it becomes an integral part of what we see as our good, part of the final ends that we pursue with our intentional behavior.

Clearly, some points left inexplicit in Rawls's text have been completed by interpretation in our reconstruction. It also makes immediately evident that the sense of justice is a force that typically emerges and stabilizes a well-ordered society only *ex post*, when institutions are already 'out there' operating through some level of compliance by the members of society. Thus the question arises of from where compliance with principles stems at the very first step of their implementation, when it cannot be said that there is an history of well-ordered society institutions already operating.

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

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

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

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

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

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

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

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

An alternative interpretation could assume that simply because all individuals know that institutions are just in terms of the principles, any particular individual develops the desire to comply with them. But in this case it would be entirely unclear how an individual is able to understand that other agents' behaviors are expressing the intention to benefit him/her by following the principle of justice, which seems a necessary condition for saying that by complying with the principle s/he 'responds in kind'. If his/her response in kind does not simply amount to intellectual acceptance of the principles but also consists in complying with them, it is necessary that other agents do not simply accept or recognize intellectually that institutions are just; they must also be seen as acting upon the principles in practice. Only in this case can compliance be a response in kind: compliance in return for compliance. Thus the sense of justice not only depends on the direct assessment of any decision in terms of its coherence with principles but is also conditional on beliefs concerning the effective compliance by other agents given what they themselves believe. Even if this seems to be the correct understanding of Rawls, we call it a weak version (conditional and reciprocity based) of Rawls's sense of justice.

3. Conformist Preferences

The theory of conformist preference (Grimalda/Sacconi 2002; 2005; Sacconi/ Grimalda 2007) initially devised in order to explain nonprofit behaviors and organizations, proves entirely consistent with (but more precisely testable than) the general idea of norm compliance derived from Rawls.[3]

Assume that two or more players are involved in a typical non-cooperative game where Nash equilibria are suboptimal, or a non-cooperative division game exists such that the Nash equilibrium is so defined that at least some players are completely excluded from any sharing of the surplus, and hence equilibria are not mutually beneficial (think of such a game as a Dictator Game with two interacting dictators and one passive receiver). Of course, such games do not have mutually beneficial equilibrium solutions including all the players in some sharing of the pie at stake (more on this game in *section 4*). Before such a game

[3] We do not say that the theory is entirely Rawlsian, since it assumes that in the *ex ante* decision a social contract is subscribed on the Nash bargaining solution of the relevant game, whereas Rawls would have suggested the maximin solution. What we simply say is that the solution given to the norm compliance problem through conformist preferences is strictly consistent with Rawls' idea of the sense of justice. However, consider that the Nash bargaining solution in a symmetric bargaining situation implies the egalitarian solution which is also consistent with Rawls' maximin (see the introduction to this paper and Binmore 2005).

is played, it is assumed that a pre-play communication stage occurs wherein, by an impartial ('behind the veil of ignorance') decision, players may agree on a principle of distributive justice (a norm) assigning a solution to the ensuing game (even though this solution will not necessarily coincide with an equilibrium point, i.e. it may not be incentive incompatible). In classical game theory terms, this pre-play communication phase is simply 'cheap talk' in that agreements reached in this phase are not binding commitments and hence do not constrain or restrict the strategy space of the ensuing game in any way. Since pre-play agreement is just cheap talk, it should not prevent players from choosing in the game's prosecution the strategy that, given their prediction of other players' strategies, maximizes their own material payoff independently of the content of the agreed norm.

Nevertheless, at this stage, players put themselves in the hypothetical situation of an *ex ante* potential agreement. They perform the collective thought experiment of playing a bargaining game under a 'veil of ignorance', each of them concealing from the others his/her identity and role as a player in the ensuing actual game. By this stage they can agree on a principle of justice able to determine a solution for the ensuing game from a normative point of view. The theory of conformist preference explains why this pre-play communication stage can result in effective decisions to comply with the agreed principle through an endogenous engendering of a preference favorable to compliance with the agreed principle that may counterbalance the material incentives represented in the initial description of the game.

The idea is that economic agents are motivated both by consequentialist (and mainly self-interested) and 'conformist' preferences—that is, the intrinsic motivation to act according to an agreed principle if complied with reciprocally by other interacting agents as well. Thus, the utility maximization model of a rational economic man can be considerably revised, extending its explanatory and normative power at a substantive level by representing these different kinds of preferences in the corresponding part of a *comprehensive* utility function.

The model assumes what we call a *state description-relative* viewpoint of preferences. The same states of affairs generated by the players' strategic decisions can be described in different ways according to their relevant characteristics. A first description of states views them as consequences: what happens to any particular participant, or only to the decision-maker, because of a given course of action. In general, if a player defines his/her preferences only on states *described as consequences*, then s/he has *consequentialist personal preferences*. These preferences are accounted for by the typical utility function of a player, U_i which for convenience will be called the material *component of the utility function*.

But secondly, states can be described as sets of interdependent actions and then characterized in terms of whether or not they are consistent with a given abstract principle of distributive justice seen as resulting from a (possibly hypothetical) *ex-ante* agreement between the players involved in the interaction. The utility function component representing these preferences will be called the 'conformist utility' of a player and it must be defined so as to give a consistent representation of the deontological motive to act that underlies this preference.

According to the conformist preference theory, this motivation is not absolute but contingent on expectations about the other agent's action and level of reciprocity in conformity given his/her expectation on the first player's action. In other words, a player's disposition to conform depends on how far s/he can approximate the ideal level of principle satisfaction given his/her expectation about the other agent's action, and on how much s/he believes the other player can contribute to the approximation of the ideal level of principle satisfaction given his/her beliefs in the first player.

In other words, intuitively speaking, a player will gain intrinsic utility from the simple fact of acting in accordance with a principle, if s/he expects that in this way s/he will be able to contribute to fulfilling the distributive principle, admitted that s/he expects the other players also to contribute to fulfilling the same principle, given their expectations.

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

First, a principle T, which is a social welfare function that establishes a distributive criterion of material utilities. Players adopt T (the norm) by agreement in a pre-play phase, and employ it in the generation of a consistency ordering over the set of possible states σ, each seen as a combination of individual strategies. The highest value of T is reached in situations σ where material utilities are distributed in such a way that they are most consistent with the distributive principle T among the alternatives available. Note that what matters to T is not 'who gets how much' material payoff (the principle T is neutral with respect to individual positions), but how utilities are distributed across players. Satisfaction of the distributional property is the basis for conformist preferences. We assume that T coincides with the Nash bargaining product (NBP)—i.e. $\sum_i (s_i - d_i)$ where s_i and d_i are player i's payoff and reservation utility respectively.[4]

Second, a measure of the extent to which, given the other agents' expected actions, the first player by her/his strategy choice contributes to a fair distribution of material payoffs in terms of the principle T. This may also be put in terms of the extent to which the first player is *responsible* for a fair distribution, given what (s/he expects that) the other player will do. To put it differently, it is a measure f_1 of the extent to which, given player 2's expected actions, the first player contributes with his/her choice to the realization of the state of affairs in which the social welfare function T is maximized. It reduces to a conformity index assuming values between -1 (no conformity at all—or maximum distance from the maximum value of T) and 0 (full conformity—or minimum distance from the maximum value of T).

Third, a measure of the extent to which the *other* player is expected to contribute to a fair distribution in terms of the principle T, given what s/he (is expected to) expects from the first player's behaviour. This may also be

[4] The Nash Bargaining Solution may be understood as a formal model for the 'social contract' that players would agree in an *ex ante* (possibly hypothetical) collective decision on the rules that should constrain (at least as a matter of 'ought') the allocation of surpluses arising form their interactions (see Sacconi 2000; Binmore 1998; 2005).

put in terms of the (expected) *responsibility* of the other player for generating a fair allocation of the surplus, given what s/he (is believed to) believes. Put otherwise, it is a measure $\widetilde{f}_2$ of the extent to which the other player is expected to contribute to the realization of the state in which T is maximized, given what s/he is believed to expect about the first player's action. Again, this reduces to a reciprocal conformity index assuming values between -1 (no conformity at all, as in the case in which the expected action chosen by the other player minimizes the value of T, given his/her expectations about the first player's choice) and 0 (full conformity, as in the case in which the expected action chosen by the player maximizes the value of T, given his/her expectations about the first players' choice).

Fourth, an exogenous parameter λ representing the motivational force of the agent's psychological disposition to act on the motive of reciprocal conformity with an agreed norm.[5]

Steps two and three combine to define an overall index F of conditional and expected reciprocal conformity for each player in each state of the game. This index operates as a weight (between 0 and 1) on the exogenous parameter λ determining whether or not λ will actually affect (and, if so, to what extent) the player's payoffs. To sum up the effect of the different components, if a player expects that the other player will be responsible for the maximal value of T, given what the other player expects about his/her behaviour, and s/he is also responsible for a maximal value of T, given the other player's (expected) behaviour, then the motivational weight of conformity λ will fully enter his/her utility function. That is, the player's preference system will show all the force of the disposition to conform with agreed norms, so that complying with the principle will yield utility (in the psychological sense) additional to the material payoff of the same strategy.

As a consequence, the overall utility function of player i with reference to the state σ (understood as a strategy combination of player i's strategy σ_i and the other players' strategies σ_{-i}), is the following

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

where

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

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

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

iv. $F = (1 + f_1)(1 + \widetilde{f}_2)$ is a compound index expressing both agent i's conditional conformity and the other individuals' expected reciprocal conformity

[5] This assumption corresponds to Rawls's assumption of a capacity to form a sense of justice derivable from the lower-level moral sentiments.

with principle T in state σ, given player i's beliefs of first and second order (i.e. beliefs about the other players' first-order beliefs) predicting that state σ is in fact the case.

When agent 1 does not comply with the norm and s/he does not expect conformity by agent 2—i.e. the indexes of conditional (f_1) and reciprocal ($\widetilde{f_2}$) are equal to -1—F is equal to zero. F is equal to 1 in the case of full conformity by agent 1 and expected full conformity by agent 2—i.e. when both the indexes of conditional (f_1) and reciprocal ($\widetilde{f_2}$) conformity are equal to 0. λ_i is zero when agent i does not have any desire to conform with a shared norm of distributive justice. When either λ_i or F is equal to zero, the conformist/deontological component of the utility function is not active and the player will choose the strategy that maximizes only his/her monetary payoff.

To sum up, if a player expects that the other player will contribute to the maximization of T, given what the other player expects about his/her choice, and s/he also contributes to the maximization of T, given the other player's expected choice, than the disposition index λ will fully enter his/her utility function, because F is equal to 1. That is, the motivational force to conform with the norm will be maximum, and complying with the norm will yield additional psychological utility that sums with the material utility deriving from the consequentialist/material component of the utility function.

4. A Game of Reference: The Exclusion Game

The relevance and implications of conformist preference in strategic interactions have been empirically tested by Sacconi/Faillo (2010) and by Faillo/Ottone/Sacconi (2008) in two experimental studies based on the so-called Exclusion Game. The game reproduces a situation in which the interaction among a set of individuals makes a social surplus affordable. But only some of these individuals—the strong players—are in charge of taking decisions concerning the allocation of the surplus. The other individuals—the weak players—have no voice and their income depends completely on the strong players' decision. Strong players then have the power to decide whether to include the weak ones in the sharing of the surplus or whether to appropriate the whole of it.

The exclusion game poses the problem of distributive justice possibility in its purest way. It explores whether it may happen that active players include a weak player in the sharing of a given surplus such that nobody can claim to have contributed more to its production than anyone else, but nonetheless there is no mutual advantage for the active players in taking the inclusive decision, given that they are not afraid of losing their reputation, and the weak player does not have the threat power that may coerce them to include him/her.

In the basic three-player version of the Exclusion Game two strong players (Strong1 and Strong2) must decide how to allocate a sum of money (S) between themselves and a weak player (Weak) who has no decisional power and whose payoff depends on strong players' choices. In particular, Strong1 and Strong2 have to declare—simultaneously and independently of each other—how much of

		Strong 2		
		Ask for 3	Ask for 4	Ask for 6
Strong1	Ask for 3	3,3,6	3,4,5	3,6,3
	Ask for 4	4,3,5	4,4,4	4,6,2
	Ask for 6	6,3,3	6,4,2	6,6,0

Figure 1: A three-player Exclusion Game. Payoff matrix

the sum S they want for themselves choosing among three options—1/4 of S, 1/3 of S or 1/2 of S. The final payoff of each strong player corresponds to the amount asked for her/himself, the remaining sum is assigned to Weak. *Figure 1* reports the payoff matrix of the game when S=12; in each cell the first two numbers are Strong1's and Strong2's payoff respectively, the third number is Weak's payoff.

Given the choice options of the strong players, if both Strong1 and Strong2 choose to ask for half of the surplus (cell Ask for 6, Ask for 6 of the payoff matrix), their payoff is 6 while Weak's payoff is 0. This is the case in which the weak player is completely excluded from the sharing of S. Inclusion of Weak implies that at least one of the strong players asks for less than a half of the surplus. An equal split is obtained when both Strong1 and Strong2 decide to ask only for 1/3 of S (cell Ask for 4, Ask for 4 in figure 1).

If we assume that strong players are motivated only by the desire to maximize their material payoff, the only Nash equilibrium in dominant strategies of this game is the one in which both Strong1 and Strong2 ask for half of the surplus, leaving nothing to Weak.

4.1 The Exclusion Game Played by Agent with Conformist Preferences

Let us now consider how they Exclusion Game of figure 1 would be played by agents endowed with conformist preferences and who have reached a pre-play agreement on a norm of distributive justice corresponding to the maximization of Nash bargaining product.

In order to find the equilibria of the game, the first step consists in ranking the possible outcomes of the game (states of affairs) on the basis of corresponding values of the social welfare function T.

In the case of the game of *figure 1*, assuming that players have a reservation utility of zero, the ranking is the following:

$$T(4,4,4) = 64 > T(3,4,5) = T(4,3,5) = 60 > T(3,6,3) = T(6,3,3) = T(3,3,6) = 54 > T(6,4,2) = T(4,6,2) = 48 > T(6,6,0) = 0$$

The Nash product is maximized when both the strong players ask for 4, and it is minimized when both the active players ask for 6.

Then, if each player i believes that his/her opponent j will choose 'Ask for 4', and if i believes that j expects that i will choose 'Ask for 4', the outcome (Ask for 4, Ask for 4), will be an equilibrium of the game.[6] Note, however, that (Ask for 6, Ask for 6) is an equilibrium of the game when player i believes that his/her opponent j will choose 'Ask for 6' and i believes that j expects that i will choose 'Ask for 6'.

We start with a game in which only monetary payoff is taken into account and in which there is a unique Nash equilibrium (Ask for 6, Ask for 6). Then, by introducing conformist preferences, we move to a new game, where payoffs are characterized also by a deontological/conformist component and in which there are two possible equilibria depending on the players' reciprocal beliefs. One corresponds to the maximum degree of conformity with the principle of distributive justice, while the other is the one in which the degree of conformity is zero.

The solution of the game depends also on the value of the parameter λ. Given the existence of beliefs coherent with reciprocal conformity, the stronger is the players' disposition toward conformity with a shared principle of distributive justice, the greater is λ and the higher is the probability that the outcome (Ask for 4, Ask for 4) will be selected as the solution of the game.

5. The Experimental Evidence

In their experimental studies, Sacconi/Faillo (2010) (SF from now on) and Faillo/Ottone/Sacconi (2008) (FOS from now on) compare the case in which the Exclusion game is played without any further interaction between the subjects with the case in which, before the subjects play the Exclusion Game, they are given the possibility to agree on a *non-binding* fairness rule concerning the division of a surplus between strong and weak players. The subjects know that after the agreement they will play the Exclusion Game, but they do not know with which roles they will do so. In substance, the subjects have to agree, under a veil of ignorance, on what is the right way to play the game. The key question that the two studies address is whether subjects, after having chosen a rule, implement it in the game even when it prescribes a choice in contrast with the pursuit of their material self-interest.

5.1 The SF Experiment

In the SF experiment participants played the basic version of the three-player Exclusion Game (*Figure 1*) under two different scenarios—before and after the agreement on a division rule.[7]

[6] In this context, the appropriate notion of equilibrium is that of Psychological Nash Equilibrium (Geanakoplos et al. 1989), which is an extension of the Nash equilibrium for situations in which expectations enter the player's utility function.

[7] In this synthesis of the two studies we will deliberately avoid discussing some of the more technical features of the experiment. Interested readers will find all the details in the original papers.

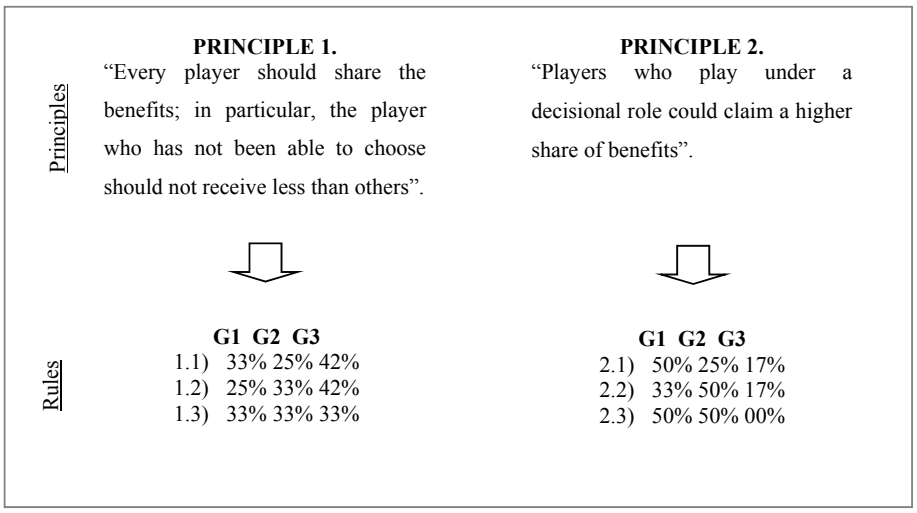

Figure 2: Second phase. Principles and rules (Sacconi/Faillo 2010)

The experiment consisted of three stages. In the first stage subjects were divided into groups of three and they were asked to play the Exclusion Game with S=12 euros. Within each group, the roles of Strong1, Strong2 and Weak were randomly assigned.[8] Strong1 and Strong2 were invited to play the Exclusion Game deciding how many of the 12 euros to ask for themselves and how many to leave to Weak. In particular, the two strong players were able to choose among three options: ask for 3 euros (1/4 of S), ask for 4 euros (1/3 of S) and ask for 6 euros (1/2 of S).

Choices were made simultaneously.

Subjects played the game three times, in three different rounds. At the beginning of each round, the three roles were randomly re-assigned to the members of the group. Each participant was able to take each of the three roles in turn. At the end of the experiment, only one of these rounds was randomly selected, and subjects' earnings for phase 1 were determined according to the outcome of that round. The game was played anonymously and subjects were not aware of the outcomes of previous rounds. This procedure produced two observations for each player in this phase: his/her choice in the Strong1 role and his/her choice in the Strong2 role.

In phase 2, subjects were assigned to new groups consisting of three members, and they were invited to agree, by means of a voting procedure, upon a hypothetical rule for the allocation of a sum between two strong players and a weak player. The agreement was to be reached by repeatedly voting until unanimity was reached, within a limited number of trials. No explicit communication or

[8] In the experiment, the roles were given the names of G1, G2 and G3 and a more neutral wording was used, distinguishing between 'active' and 'non-active' players instead of 'strong' and 'weak' players.

mutual identification was allowed among the players of any given group. At the beginning of this phase, subjects were not assigned any specific role and were informed that in the following phase they would again play the game played in the first phase. They were requested to vote for one of two general principles and for one among some more specific rules deduced from the selected general principle (*Figure 2*). Subjects knew that the rule would not be binding.

Groups had to reach a unanimous agreement by voting for the same principle within five trials before they could pass to the vote on the specific allocation rule, upon which the groups had to agree within ten trials. A lack of unanimity after the last of the trials would have prevented subjects from entering the third phase.

The third and last phase started with the composition of the group unchanged and with the random assignment of Strong1, Strong2 and Strong3 roles to the members of each group that had agreed upon a given rule.
Subjects played the same game that they had played in the first phase, but now strong players had the additional option of choosing between implementing the rule that they had agreed in the second phase or choosing one of the alternative strategies. Immediately after their choice, strong players were asked to express their expectations about the opponent's willingness to implement the rule by guessing the outcome of the game.[9]

To summarize, in the first phase the three-player version of the Exclusion Game was played. In the second phase, groups were rematched and subjects voted for the rule. In the third phase, the veil of ignorance was removed and strong players had to decide whether or not to comply with the rule. The main comparison to be made in this experiment was that between the behaviour of subjects in the first phase and the behaviour of the *same* subjects in the third phase.

5.2 The FOS Experiment

The FOS experiment was based on a different design—a so-called 'between-subject design' as opposed to the 'within-subject' design adopted in the SF experiment. The comparison in this case was made between the behaviours of *different* subjects taking part in *different* treatments. In particular, the experiment consisted of three treatments: the *Baseline Treatment*, the *Agreement Treatment* and the *Outsider Treatment*.

All the treatments were based on a four-player version of the Exclusion Game with a surplus S of 60 units of an experimental currency called 'tokens'. Subjects interacted in groups of four, and in each group there were three strong players and one weak player. Strong players had to decide, simultaneously and independently of each other, how much of S to ask by choosing one of the three options available: ask for 15 tokens (25% of S), ask for 18 tokens (30% of S), or ask for 20 tokens (33% of S). In this version of the Exclusion Game the weak player was completely excluded from the sharing of the surplus when all the three strong players chose to ask for 20 tokens, while S was equally divided among all the

[9] Subjects were asked to indicate the cell of the payoff matrix in which s/he thought the game would end.

players (strong and weak) when all the strong players decided to ask only for 15 tokens.

In the *Baseline Treatment* participants simply played the basic four-player version of the Exclusion Game.
In the *Agreement Treatment* participants were involved in a two-stage game. In the first stage, in each group each player, without knowing her/his role in the game, was invited to vote for a specific allocation rule by choosing among three options. This stage corresponded to the voting procedure under the veil of ignorance, and it had the same characteristics as the procedure adopted in the SF experiment. In this case, subjects had to vote for one of the following three options (type A participants A1, A2 and A3 are the strong players of the group, while participant B is the weak one):

> Rule 1:
> The division of the 60 tokens depends on the role of the participants. In particular, Type A participants will receive a number of tokens which is three times the number of tokens assigned to Type B participants. This corresponds to the following division of the 60 tokens: 18 tokens for A1, 18 tokens for A2, 18 tokens for A3, 6 tokens for B.
>
> Rule 2:
> The division of the 60 tokens depends on the role of the participants. In particular, the three Type A participants will receive the same number of tokens; the Type B participant will receive zero tokens. This corresponds to the following division of the 60 tokens: 20 tokens for A1, 20 tokens for A2, 20 tokens for A3, 0 tokens for B.
>
> Rule 3:
> The division of the 60 tokens does not depend on the role of the participants. Each participant will receive the same number of tokens. This corresponds to the following division of the 60 tokens: 15 tokens for A1, 15 tokens for A2, 15 tokens for A3, 0 tokens for B.

Players had to reach a unanimous agreement on the rule within 10 trials. The rule was not binding, but only groups that unanimously voted for the same rule in this first stage would participate in the second stage.

In the second stage, without changing the composition of the groups, roles were randomly assigned and the subjects played the four-player Exclusion Game. Strong players could either decide to implement the rule selected in stage 1 or choose one of the alternative allocations.

In the *Outsider Treatment*, the first stage, as well as the rule on entering the second stage, were the same as in the *Agreement Treatment.* But at the beginning of the second stage, after the subjects had been assigned their role, the groups were partially re-matched. In particular, a strong player for each group ('the outsider') was reassigned to a different group and informed about the rule voted by her/his new group. The other members of the group hosting

the outsider did not know what rule the outsider's previous group had adopted. After the re-matching, the subjects played the four-player Exclusion Game.

In all the treatments, at the end of the game and before players were informed about the other strong players' choices, first-order and second-order normative and empirical expectations were elicited by means of a brief questionnaire.[10] In particular, in each group each strong player was asked to make statements concerning:

1. the probabilities of each possible choice by co-strong players (First-Order Empirical Expectations);

2. the probability of each co-strong player's possible judgement about his/her own choice—what the other strong player believed that s/he had chosen (Second-Order Empirical Expectations);

3. the choice that s/he considered ought to be made in that particular situation (First-Order Normative Expectations);

4. the choice that co-players considered ought to be made (Second-Order Normative Expectations).

In the Outsider Treatment, expectations about the behaviour and beliefs of partners and outsiders were elicited separately.

5.3 Empirical Hypotheses

A set of empirical hypotheses can be derived by assuming that the motivations of the subjects who participated in these two experiments can be described in terms of conformist preferences.

Firstly, in regard to both experiments, we can expect that when the subjects do not have the opportunity to agree on a rule for the division of the surplus (first phase of SF and Baseline Treatment of FOS) the ideal component of the utility function is not active, since there is not a shared ideal to conform with. Strong subjects have not reason to believe that the other strong subjects will conform to any particular fairness principle. Hence the following hypothesis can be put forward.

Hypothesis 1. When subjects have not the possibility to agree on a fairness rule on how to play the Exclusion Game—as in the first phase of SF and in the Baseline Treatment of FOS—there is no reason to expect that they will play differently from purely self-interested individuals. Consequently, in both experiments they will choose to ask for the maximum—6 euros in SF and 20 tokens in FOS.

[10] Only good guesses on Empirical Expectations were rewarded on the basis of a mechanism known as the quadratic scoring rule (Davis/Holt 1993) which assigns a payoff that depends positively on the probability assigned to the event that actually occurs.

With regard to the agreement phase, in both the treatments, when the subjects have to choose the rule under the veil of ignorance, it could be expected that the majority of them will opt for the rule that prescribes the equal division of the surplus. The voting phase mimics a typical constitutional choice of a fairness principle which, according to a contractarian approach, will induce participants to assume an impartial perspective. They will judge the outcomes of the game from the point of view of each role, and they will choose a rule acceptable from whichever point of view. This implies a solution invariant to the permutation of the individual point of view, like the one in which the surplus is equally divided. Furthermore, in this setting, the equal division of the surplus is also the intuitively obvious choice—the most salient one. Given that the agreement is a necessary condition for accessing to the next phase, players may vote for the most salient rule in order to coordinate within the maximum number of trials. Hypothesis 2 follows.

Hypothesis 2. In both the experiments, in the pre-play agreement phase—second phase of SF and first phase of the Agreement and Outsider Treatments of FOS—the majority of subjects will choose the rule that prescribes the equal division of the surplus.

According to conformist preferences, the subjects will comply if i) they are part of the group that has chosen the rule; ii) they believe that other members of their group will comply (First Order Empirical Expectations compatible with the choice dictated by the rule) *and* if iii) they believe that other members of the group expect that they will comply (Second Order Empirical Expectations compatible with the choice dictated by the rule). In FS one should expect to find that strong subjects complying with the rule are those who expect that the game will end in the cell of the payoff matrix compatible with mutual compliance with the rule. In FOS, in the Agreement Treatment, one would expect strong subjects to comply if they believe both that other strong subjects will comply (first order empirical expectations) and that the other strong players will expect them to comply (second order empirical expectations). In the Outsider Treatment, groups playing the Exclusion Game are formed by both insiders—those who voted the rule—and outsiders—those who participated in the voting procedure in another group. One should expect to find a lower level of compliance in the Outsider Treatment with respect to the Agreement Treatment because outsiders are not part of the group which has chosen the rule and insiders do not have any reason to believe that outsiders will comply with the rule, hence they will not comply. The following hypothesis can be put forward

Hypothesis 3. In both the experiments, strong players comply with the agreed rule only if they expect that the other strong players of their group comply as well.

5.4 Empirical Results[11]

Overall, 366 undergraduate students took part in the experiments (150 in the FS and 216 in the FOS).[12] In the SF experiment, during the first stage, all participants played twice as strong players, while in the third stage only 3/4 of the subjects played as strong players (the remaining 1/4 were selected as weak players and they did not have any decisional power). Consequently, we observed choices of 150 subjects in the first stage (2 observations for each player) and of 100 subjects in the third stage. All 150 players participated in the voting procedure.

In the FOS experiment, 56 players were recruited for the *Baseline Treatment*, 72 for the *Agreement Treatment*, and 88 for the *Outsider Treatment*. We had observations of 42 strong players in the *Baseline Treatment*, 54 in the *Agreement Treatment*, and 66 in the *Outsider Treatment*. The preferred rule was signalled by all 216 players.

Result 1. In both the experiments, in the absence of an ex-ante impartial agreement most subjects behaved in a self-interested way.

In SF, in the first stage, 59.3% of players always asked for the maximum amount, while 26.7% chose the self-interested strategy at least once. In FOS, in the *Baseline Treatment*, 73.8% of strong players chose to ask for the highest amount of tokens (20).

These findings are in line with *Hypothesis 1.*

Result 2. In both the experiments, when agreement was possible, it was reached by all groups. Moreover, almost all groups agreed on the equal division rule.

As we expected (*Hypothesis 2*), in both the experiments agreement was always reached. This is not surprising since agreement was not binding but failure to reach it was costly, because it prevented access to the following phase. However, the interesting point is that, as we expected, the egalitarian distribution rule seems to have been a focal point. In particular, in SF, 64% of subjects chose the egalitarian rule (32 out of 50 groups), while in FOS, 17 groups out of 18 in the *Agreement Treatment* and 20 out of 22 in the *Outsider Treatment* chose the fair-division rule.

---

[11] All these results have been tested with statistical and econometric techniques. For details see Sacconi/Faillo 2010 and Faillo/Ottone/Sacconi 2008.

[12] 150 subjects participated in Experiment 1 and 216 in Experiment 2. All 10 sessions of Experiment 1 were run in Trento (CEEL—University of Trento), while Experiment 2 was run both in Milan (EELAB—University of Milan Bicocca) and Trento. In particular, 3 sessions were run for the Baseline Treatment (1 in Milan and 2 in Trento), 4 sessions for the Agreement Treatment (2 in Milan and 2 in Trento), 5 sessions for the Outsider Treatment (3 in Milan and 2 in Trento).

Result 3. In both experiments, we observe a high degree of compliance with the selected rule. In addition, strong players complied if they expected compliance also by the other strong players of their group.

When we analyse players' choices after the agreement stage, we find a high degree of compliance with the non-binding rule selected. In particular, in SF, 77% of subjects decided to comply with the chosen rule when playing the Exclusion Game, while in Experiment 2, 50% of players in the *Agreement Treatment* and 39.4% in the *Outsider Treatment* did so.

In accordance with *Hypothesis 3*, strong subjects who complied with the rule expected that the other strong subjects of their group would comply and would expect them to comply.

In addition, in both the experiments, for a significant percentage of subjects, agreement on a fairness rule seems to have been a sufficient condition for the emergence of expectations of reciprocal conformity. This can be considered a pure empirical result. Conformist preferences theory predicts compliance when the weight of the ideal component (lambda) is sufficiently high and reciprocal expectations of compliance exist. But the theory is silent with regard to the origin of these expectations. The experimental evidence seem to suggest that the impartial agreement is itself the source of reciprocal beliefs of compliance.

In SF and in the Agreement Treatment of FOS, strong players who were parts of groups that had chosen a particular rule tended to comply because, since they had reached an impartial agreement on that specific rule, they believed that the other strong players would comply with the rule and would expect them to comply. Since in both experiments most of the groups agreed on the equal division rule, this resulted in a significantly smaller percentage of selfish choices in the cases in which an impartial agreement had been reached.

A detailed account of this process can be provided for the FOS experiment, where the data on expectations are more accurate. On comparing empirical expectations in the *Baseline Treatment* and in the *Agreement Treatment* of FOS, it is clear that the agreement influenced the players' beliefs. In the *Baseline Treatment* 88% of subjects expected that the other strong players in the group would ask for the maximum amount (20 tokens). In the *Agreement Treatment*, 17 groups out of 18 chose the fair rule, according to which each strong player should ask for 15 tokens. On analysing the subjects' expectations, we find that in the *Agreement Treatment* there is a significant decrease of subjects who thought that the other members of their group had asked for 20 tokens—only 40%.

At this point, a more sophisticated process emerges. The relationship between agreement and choice can be described in two steps. Step 1: the agreement influenced the players' empirical expectations. Step 2: empirical expectations defined the subjects' choices. This means that the difference between the *Baseline Treatment* and the *Agreement Treatment* is a consequence of the impact of the agreement on players' beliefs and preferences.

To sum up, we can say that the agreement leads to a convergence of empirical expectations on the fair rule and these expectations influence subjects' decisions.[13]

Result 4. In the Outsider Treatment of FOS a lower percentage of players complied with the chosen rule with respect to the Agreement Treatment.

In FOS, when the Exclusion Game was played in groups where one subject was an 'outsider' (in the *Outsider Treatment*), a lower percentage of players complied with the chosen rule.

This result seems to confirm that having reached an impartial agreement on the rule with the members of her/his own group was a necessary condition for a subject to comply with that rule. When groups were rematched and one of the strong players (the outsider) was assigned to a new group, the members of his/her new group (the insiders) did not expect compliance from him/her, and consequently they did not comply. The outsider seemed to acknowledge this, and, on expecting non-compliance by the insiders, s/he did not comply.

The same two-step process described above can explain the evidence on the *Outsider Treatment*. When we analyse subjects' expectation of compliance, it turns out that this value is higher in the *Agreement Treatment* than in the *Outsider Treatment*—46% against 27%. This means that, once again, the difference between the *Agreement Treatment* and the *Outsider Treatment* is a consequence of the impact of the outsider on players' beliefs. Step 1: the introduction of an outsider influences the players' empirical expectations. Step 2: empirical expectations define the subjects' choices. If we analyse expectations and choices in both treatments, it turns out that in the *Outsider Treatment* subjects were more likely to expect deviation by the co-players from the chosen rule and these expectations of compliance influence subjects' decisions.

6. Conclusion

To conclude, let us explain how our experiment is consistent with a Rawlsian perspective on norm compliance (see conditions at the end of section 2), and how we are also able to explain the reciprocal belief formation necessary for selection of the 'compliance-with-the-principle' behavior (a point that we discussed in the introduction on the equilibrium selection approach to the compliance problem, but which is relevant here in terms of selection of the psychological equilibrium inducing full conformity with the principles of justice).

The participants in the experiment were young people (students at the University of Trento and Milan), who had grown up in the context of a nearly well-ordered society, wherein they had experienced the functioning of at least some lower-level nearly-just institutions like the family or associations. Through these

[13] When analysing the correlation between choices and expectations, it turns out that actions and beliefs are in line. Moreover, a more sophisticated econometric analysis confirms both steps.

experiences they had developed—to some extent—the capacity to form a sense of justice (in our model represented by the extent of the parameter λ). Nonetheless, when confronted with the exclusion game without the opportunity to choose the principles of division impartially, they still did not perform the mental experiment of an impartial justification of whatever solution. Hence they acted according to their self-interest. When the agreement in both the experiments was to be reached, however, the subjects performed the experiment of being put under a veil of ignorance in order to agree on a principle of distributive justice; that concerns then not because they are assured about the enforceability of the agreement, but simply because in agreeing they share the simple intent of giving each of them the chance to participate in the following non-cooperative game. Agreeing on a principle/rule is similar to taking part in a constitutional decision. The constitution mattered to the participants not because it provided binding commitments but simply because it gave them a chance to participate in a potentially beneficial game only if they reached an anonymous agreement on the principle/rule. Once the agreement had been stipulated, they were entirely free to violate it, but they were also in the position to benefit from each other in so far as the agreement gave an opportunity to play the exclusion game.

In the experiments, when the subjects entered the exclusion game stage after having agreed on some principle of distributive justice, they found themselves in a situation that Rawls would have recognized as a well-ordered society in which, because they had developed the capacity to form a sense of justice, and because they realized that principles had been chosen under the veil of ignorance, they should have been able to comply with principles of justice. We can check whether the experiment conveyed evidence favorable to the idea of a sense of justice or—in other words—whether the experiment satisfied Rawls's hypothesis (see *section 2*) and engendered behavior consistent with the idea of the emergence of a sense of justice.

Consider the players who agreed on the egalitarian principle/rule. All of them, within their own group, knew that the others had made the same choice, i.e. under a Rawlsian interpretation that they all justified the same course of action under a veil of ignorance. In fact, they took no more than a very few voting rounds to agree on that principle/rule in stage two of the first experiment and in the agreement treatment of the second experiment.

What about the shared knowledge that the opponents' effective behavior was beneficial to each single member of the group and which is necessary in order to elicit reciprocity? At this point players could not rely on the evidence of a long past history of norm compliance. Nevertheless, most of those who agreed on an egalitarian principle/rule also believed that their opponents would conform (as far as the agreement treatment is concerned). This suggests that the agreement under a veil of ignorance may by itself have a strong causal effect on shaping reciprocal expectations.

This is not implicit in the conformist preference model, but is a natural consequence of the veil of ignorance reasoning format, which accords with the idea of default reasoning, and receives surprisingly strong evidence from the experiment. In order to make sense of this fact, it is important to realize that

there is no logical necessity in the inference from the *ex ante* agreement to the expectation of *de facto* compliance by other participants in the stage two agreement. On the contrary, this involves a cognitive mechanism known as *default* reasoning (Reiter 1980; Bacharach 1994; Sacconi/Moretti 2008). The idea is simply that if each player has actually adopted an unanimous impartial agreement in the *ex ante* perspective, then s/he will acquire at least the *mental model* of a decision maker who acts in accordance with a plan whose content coincides with the terms of the agreed course of action. Agreeing on a set of actions to be carried out later implies having a mental representation of an agent carrying out a plan of action—which is simply the content of the statement of agreement.

A normally rational agent cannot fail to have this mental model because it is derived from introspection, and because the player him/herself is an exemplar of an agent who has planned to act in accordance with the content of the statement of agreement later on. But then consider that mental models are necessarily used in order to figure out possible situations and predict them. And hypothesize that at this point in time no *framing* of an agent 'comes to the players' mind' (Bacharach 2006) other than the mental model of an agent who *will act according to the content of the agreement.* If no contrary evidence is forthcoming, the only way an agent can simulate the other players' choice is to resort by default to his/her own mental model of a rational agent. By default, then, the same mental model is used to simulate every players' reasoning and behavior. This simulation may be recursive, so that a player uses his/her mental model not only to predict another player's behavior but also in order to simulate the other player's reasoning and beliefs, so that a *shared mental model* of all the rational agents is formed such that they are all expected to conform with the terms of agreement. The treatment with outsiders confirms this line of reasoning. When players were informed that there were members of the group playing the exclusion game who had not agreed on the principle, a different frame of a rational player came to the players' mind—i.e. a frame such that an agent may not act according to the content of the agreement because *she has not taken part in agreeing on it.* In fact, in this case the first-order and second-order beliefs about other players' conformity break down, and as a consequence also the level of behavioral conformity is sharply reduced.

This explains—if not logically justifies—why the agent (as long as there is no proof to the contrary) may frame the case as a situation wherein agents conform with the norm. The *ex ante* agreement on a principle of fairness allows by default the formation of a prior belief that the propositional content of the mental model representing an agent discharging his/her commitments to an agreement is true. Just after the agreement there is no evidence that any player will not conform, whereas there is the intuitive evidence of the mental representation of an agent who agrees to a principle and hence expresses at least at that point in time the commitment to carry out a certain behavior later on. Although it would be excessive to say that this completely resolves the players' prior uncertainty, it explains how, after an agreement has been worked out—in so far as it is understood as being a constitutional, fair, initial (*ab origine*) agreement under

the 'veil of ignorance'—the model of a compliant agent 'comes to their minds' with most *vividness*.

The result is that, in both of the experiments considered, also the fourth condition for a 'sense of justice' was satisfied in the case of the groups choosing the egalitarian rule of division (the 'agreement treatment' in the second experiment): not only were their members exogenously capable of it, they agreed on a principle of justice under the veil of ignorance and had shared knowledge that they all agreed, but they also had the shared belief that they were all behaving in a way that was impartially beneficial to each other. It follows that, if Rawls is right, those subjects who satisfied these assumptions—those who belonged to the group choosing the egalitarian rule in the first experiment and those who did the same in the *agreement treatment* of the second—should show the formation of a sense of justice sufficiently strong to induce them to comply with the rule chosen. Which in fact was verified by our experimental evidence.

Bibliography

Andreozzi L. (2010), When Reputation Is Not Enough: Justifying Corporate Social Responsibility, in: Sacconi, L./M. Blair/R. E. Freeman/A. Vercelli (eds.), *Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*, Basingstoke, 253–271

Aoki M. (2001), *Toward a Comparative Institutional Analysis*, Cambridge/MA

Bacharach, M. (1994), The Epistemic Structure of a Game, in: *Theory and Decisions* 37, 7–48

— (2006), *Beyond Individual Choice: Teams and Frames in Game Theory*, Princeton

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge

Binmore K. (1989), The Social Contract: Harsanyi and Rawls, in: *Economic Journal* 99, 84–106

— /A. Brandenburger (1990), Common Knowledge and Game Theory, in: *Essays on the Foundation of Game Theory* (ed. by K. Binmore), Oxford, 105–150

— (1994), *Game Theory and the Social Contract, vol. I, Playing Fair*, Cambridge/MA

— (1997), *Game Theory and the Social Contract, vol. II, Just Playing*, Cambridge/MA

— (2005), *Natural Justice*, Oxford

Bolton, G. E. (1991), A Comparative Model of Bargaining: Theory and Evidence, in: *American Economic Review* 81, 1096–1136

— /A. Ockenfels (2000), A Theory of Equity, Reciprocity and Competition, in: *American Economic Review* 100, 166–193

Buchanan J. (1975), *The Limits of Liberty*, Chicago

Cubitt, R./C. Starmer/R. Sugden (1998), On the Validity of the Random Lottery Incentive System, in: *Experimental Economics* 1, 115–131

Danielson, P. (1992), *Artificial Morality, Virtuous Robot for Virtual Games*, London

Davis, D. D./C. A. Holt (1993), *Experimental Economics*, Princeton

Fagin, R./J. Y. Halpern/Y. Moses/M. Y. Vardi (1996), *Reasoning about Knowledge*, Cambridge/MA

Faillo, M./L. Sacconi (2007), Norm Compliance: The Contribution of Behavioral Economics Theories, in: Innocenti, A./P. Sbriglia (eds.), *Games, Rationality and Behaviour, Essays in Behavioural Game Theory and Experiments*, London, 101–133

— /S. Ottone/L. Sacconi (2008), *Compliance by Believing: An Experimental Exploration on Social Norms and Impartial Agreements*, working paper N. 10/2008 del Dipartimento di economia dell'Università di Trento, URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1151245

Falk, A./U. Fischbacher (2001), *A Theory of Reciprocity*, Institute for Empirical Research in Economics Working Paper No. 6

Fehr, E./K. M. Schmidt (1999), A Theory of Fairness, Competition and Co-operation, in: *Quarterly Journal of Economics* 114, 817–868

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

— (1990), Economic Man and the Rational Reasoner, in: Nicols, J. H./C. Wright (eds.), *From Political Economy to Economics – and Back?*, San Francisco, 105–131

— (1994), Commitment and Choice: An Essay on the Rationality of Plans, in: Farina, F./S. Vannucci/F. Hahn (eds.), *Ethics, Rationality, Economic Behavior*, Oxford, 217–245

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

Grimalda, G./L. Sacconi (2002), *The Constitution of the No Profit Enterprise, Ideals, Conformism and Reciprocity*, University Carlo Cattaneo—LIUC paper n. 155

— / — (2005), The Constitution of the Not-for-Profit Organization: Reciprocal Conformity to Morality, in: *Constitutional Political Economy* 16(3), 249–276

Hampton J. (1986), *Hobbes and the Social Contract Tradition*, Cambridge

Harsanyi, J. C./R. Selten (1988), *A General Theory of Equilibrium Selection*, Cambridge/MA

Hume D. (2000[1740]), *A Treatise of Human Nature*, Oxford

Kreps D. (1990), *Games and Economic Modelling*, Cambridge

Levine, D. K. (1998), Modelling Altruism and Spitefulness in Experiments, in: *Review of Economic Dynamics* 1, 593–622

Lewis D. (1969), *Conventions. A Philosophical Study*, Cambridge/MA

Luce R. D./H. Raiffa (1957), *Games and Decisions*, New York

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

— (1990b), *Rationality and Dynamic Choice*, Cambridge

— (1993), Rationality Constitutions and the Ethics of Rules, in: *Constitutional Political Economy* 4(2), 94–118

Nagel, T. (1986), *A View from Nowhere*, Oxford

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

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

— (1993), *Political Liberalism*, Cambridge/MA

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

Sacconi, L./M. Faillo (2005), *Conformity and Reciprocity in the 'Exclusion Game': An Experimental Investigation*, Discussion Paper No. 12/05, Department of Economics University of Trento

— / — (2010), Conformity, Reciprocity and the Sense of Justice. How Social Contract-based Preferences and Beliefs Explain Norm Compliance: The Experimental Evidence, in: *Constitutional Political Economy* 21(2), 171–201

— /G. Grimalda (2007), Ideals, Conformism and Reciprocity: A Model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case, in: Bruni, L./P. L. Porta (eds.), *Handbook of Happiness in Economics*, London

— /S. Moretti (2008), A Fuzzy Logic and Default Reasoning Model of Social Norms and Equilibrium Selection in Games under Unforeseen Contingencies, in: *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 16(1), 59–81
— (1991), *Etica degli affari*, Milano
— (1993a), Equilibrio e giustizia (I): la stabilità del contratto sociale, in: *Il giornale degli economisti* 52(10-12), 479–528
— (1993b), Equilibrio e giustizia (II): la selezione del contratto sociale, in: *Il giornale degli economisti* 52(10-12), 529–575
— (2000), *The Social Contract of the Firm*, Berlin
— (2010), A Rawlsian View of CSR and the Game Theory of Its Implementation (Part II): Fairness and Equilibrium, in: Sacconi, L./M. Blair/R. E. Freeman/A. Vercelli (eds.), *Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*, Basingstoke, 194–252
— (2011), A Rawlsian View of CSR and the Game Theory of its Implementation (Part III): Conformism, Equilibrium Refinements and Selection, in: *Social Capital, Corporate Social Responsibility, Economic Behavior and Performance*, edited by L. Sacconi/G. Degli Antoni, Basingstoke, 42–79
Skyrms B. (1996), *Evolution of the Social Contract*, Cambridge
Starmer, C./R. Sugden (1991), Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation, in: *American Economic Review* 81, 971–978
Sugden, R. (1986), *The Economics of Rights, Cooperation and Welfare*, London–Blackwell
— (1998), Conventions, in: *The Palgrave Dictionary of Economics and the Law*, (ed. by P. Newman), Basingstoke
— (1998b), Normative Expectation, in: Ben Ner, A./L. Putterman (eds.), *Economics, Values and Organizations*, Cambridge