The social contract in the lab.
An experimental analysis of *self-enforcing* impartial agreements.

Marco Faillo[a, *], Stefania Ottone[b,c] Lorenzo Sacconi[a,c]

Abstract

Theories of social contract are based on the idea of the "consent of the governed", according to which norms, rules and institutions, a constitutional in particular, must be based on the general consensus (or unanimous consent) of the individuals who are subject to the regulation.

The paper reports the results of an experiment aimed at identifying the conditions for the emergence of a self-enforcing social contract in the laboratory. Our main result is that spontaneous compliance with a non-self-interested norm of distribution is likely to occur if individuals have been part of the *same* process of *ex-ante* agreement on the distributive norm under a 'veil of ignorance', to which is also related the emergence of reciprocal expectations of conformity. This is in line with Rawls's idea of an endogenous 'sense of justice' stabilizing *ex-post* institutions that would have been *ex-ante* chosen in the original position.

a. Department of Economics and Management. University of Trento. Via Inama, 5 38122 Trento, Italy.
b. Department of Economics, Quantitative Methods and Business Strategy, University of Milano-Bicocca.
Piazza dell'Ateneo,1, 20126 Milan, Italy
c. Econometica. Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy.

***** corresponding author.

Email addresses: marco.faillo@unitn.it; stefania.ottone@unimib.it, lorenzo.sacconi@unitn.it.

JEL Classification: C73;C91; D02; D63.

Keywords: Social contract; social norms; distributive justice; beliefs elicitation; dictator games.

1. Introduction

Social contract theory is one of the best-known theoretical accounts of the origins and legitimacy of institutions. The renewal of the classical approach due to scholars like John Rawls, James Buchanan and David Gauthier (but see also, to cite authors who have studied the social contract using game theoretical models, Jane Hampton 1986; Brian Skyrms 1996; Ken Binmore 1994, 1997, 2005), has made the theory accessible to economists and other social scientists committed to the use of rational choice and game theory language and methodology.

One can identify several and diverse theories of social contract, but all of them are based on some version of the idea of the 'consent of the governed', according to which norms, rules, institutions, and constitutions in particular, must be based on the general consensus (or unanimous consent) and voluntary compliance of who will be bound by the agreement. This implies that individuals must have good reasons to agree on a norm or institution, but they *also* must have the effective incentives and motivations to comply with it.

Whilst the decision to enter an agreement on a distributive justice principle has been explored extensively, the problem of the *ex-post* compliance with the contract is still far from being settled. In this paper we focus on the 'compliance problem', and we report the results of an experiment aimed at identifying the conditions for the emergence of a self-enforcing social contract in the laboratory. Our main finding is that, in a context in which players are involved in one-shot interactions, spontaneous compliance with a distributive principle is observed if expectations of reciprocal conformity emerge, and this happens only if individuals have previously taken part in an *ex-ante* fair agreement on the same distributive principle. Thus, individuals who enter a group that adopted a principle by agreement on which they have had no part, are less likely to comply with that principle, even if they come from another group that has adopted the same principle.

In the experiment reported in this paper we analyze how the agreement induces the convergence of expectations of different degree and nature. In particular, we investigated three types of expectations of a generic player i: a) First Order Empirical Expectations (FOEE): player i's beliefs about the other players' choices; b) Second Order Empirical Expectations (SOEE): player i's beliefs about the other players' beliefs about his/her choice; c) Normative Expectations (NE): player i's beliefs about what the other players consider to be the right choice in a particular situation.

The rest of the paper is organized as follows. Section 2 presents the problem of the ex-post stability of social contract. In Section 3 we posit a 'conditional compliance' hypothesis as explanation of the stability of the contract and we review the related literature. The experiment is presented in Section 4. Results are discussed in Section 5. Section 6 concludes the paper.

2. Social contract and the compliance problem

In *Moral by Agreement* (1986) David Gauthier discusses the logical distinction between the decision to enter an (*ex-ante*) agreement and the decision to comply (*ex-post*) with it. The move from the former to the latter implies a fundamental change of perspective. The decision to agree on a particular norm or institution is taken under an *ex-ante* perspective. Individuals must assess whether a particular agreement will enable them to reach a mutually beneficial solution. From this perspective, they view the problem as a cooperative bargaining game aimed at solving the problem of deciding what agreement should be chosen among the many ones possible, on the assumption that if the agreement is reached, it automatically will be implemented.

When we ask if the agreement actually will be implemented, we move to an *ex post* perspective. The game logic is now that of a non-cooperative game in which individuals choose, separately but interdependently, whether or not to comply with the social contract. What matters now for the implementation of the agreement is the risk of a conflict between one's own self-interest and the attainment of a socially beneficial outcome. From this perspective, the main problem is assessing whether the agreed norm or institution will also generate the motivational forces able to induce the adoption of potentially counter-interested behavior.

A different, but nevertheless very perceptive, way to emphasize the logical distinction between the two perspectives is to consider the former as a decision taken 'under a veil of ignorance' (Rawls 1971) and the latter as a decision to comply with what has already been agreed but is to be taken when the veil of ignorance has been lifted.

If we see the *ex-post* decision problem as a game played by self-interested individuals interacting in a Prisoner's Dilemma-like setting, then a social contract implying the adoption of a dominated strategy of cooperation will obviously not be complied with. *Ex ante*, individuals, given that they reason cooperatively, have the incentive to reach an agreement to escape the suboptimal equilibrium. But *ex post*, each single agent, who now acts as a separate entity, whenever the others were expected to implement the social contract, has the incentive to cheat on them. Of course, this problem vanishes if the non-cooperative game played in the state of nature is seen as a non-cooperative game endowed with equilibria, some of which are mutually advantageous, as typically happens in repeated games (but also other simple games like the 'stag hunt' or the 'battle of the sexes'). Viewing the state of nature like this is equivalent to thinking of institutions as social conventions emerging as 'spontaneous orders', for which compliance poses no problems (Hume 1740/1978; Lewis 1969; Sugden 1986; Hardin 1999, Aoki 2001).

However, shifting from a social contract to a social convention perspective does not provide a complete solution. First of all, in social dilemmas like the (one-shot) Prisoner's Dilemma, Trust Games and Public Goods Games, there are no conventions. Secondly, if the interaction is repeated over a time horizon of indefinite length, folk theorems apply, so that equilibria implying cooperation are also included among the many ones possible. But solving the problem of how one reasonably desirable convention may be selected among the many possible is not trivial. In fact, the emergence of a particular convention implies the convergence of individuals' expectations on it. How can these convergent expectations emerge? Interestingly, the social contract again becomes the natural candidate for solution of this problem, now seen as an equilibrium selection device.

Binmore's game theoretic approach to the social contract (Binmore 1984, 1989, 1997, 2005) is probably the best-known example of application of the view that the social contract is an equilibrium selection device (but see also Hampton 1986; Sacconi 1993a, b; Skyrms 1996).

> "When an appeal (to justice) is made the players disappear behind the veil of ignorance where they negotiate in ignorance of their current and future identities about what equilibrium in the game of morals should be operated in the future (…) So a fair social contract is an equilibrium in the game of morals but it must never be forgotten

> that it is also an equilibrium in the game of life (…) Indeed the game of morals is nothing more than a coordination device for selecting one of the equilibria of the game of life" (Binmore 2005, p.172)

The situation of reference is the 'game of life', a repeated Prisoner's Dilemma-like game with two agents, Adam and Eve, playing two asymmetrical roles, with Adam in an advantageous position. In order to select one of the many possible equilibria of the 'game of life', however, they enter the 'game of moral' that they play under a 'veil of ignorance'. The social contract takes the form of an *ex ante* agreement taken in such game but ultimately aimed at solving Adam and Eve interaction. By assessing 'under a veil of ignorance' the expected payoff that s/he may get from any agreement, each player will consider him/herself as having equal probability of obtaining the payoffs resulting from any given outcome under the symmetric replacement of players' roles. Moreover, players know that the solution must identify an outcome belonging to the original equilibrium space, to which they return when they go 'beyond the veil' in the *ex post* perspective. Given the 'veil of ignorance' assumption, this entails that the agreement must coincide with an outcome belonging to the equilibrium subset resulting from the intersection of the original equilibrium space and its symmetrical translation generated by exchanging the players' positions – i.e., from the intersection of two representations of the payoff space that a player consider as equally possible by taking both the perspectives of Adam and Eve. Within this space any bargaining solution necessarily falls on the bisector, which is the geometrical locus of egalitarian solutions.

This solution vindicates the 'Rawlsian maximin': once the possible set of agreements is restricted to the symmetrical intersection subset, the social contract will coincide with the symmetric Nash bargaining solution, which corresponds also to the maximin solution with respect to the original (asymmetric) outcome space. But two problems also exist with Binmore's approach. First, even if in many cases interaction in a state of nature can be interpreted as a repeated game, there are situations in which players are involved in interactions that approximate one-shot games. In these cases, we are back to the problem of *ex-post* compliance discussed above. Second, and more importantly, the *ex-ante* agreement will be implemented only if, *ex-post*, each player has good reasons to believe that his/her opponent will not deviate from it. Compliance, then, can be expected only if the *impartial* agreement implies the common knowledge that players will converge, *ex-post,* on the equilibrium selected *ex-ante.* But this convergence cannot be *logically* deduced from the fact that players have reached an *ex-ante* agreement. The crucial question then becomes if, and under what conditions, the agreement is the *cause* of the emergence of expectations of reciprocal compliance. This is a question that should be primarily investigated empirically, as a matter of psychology of reasoning or preference formation.[1]

A solution to both problems can be found by taking a step further in rediscovery of Rawls's argument, and in particular by looking at his concept of 'sense of justice'. According to Rawls (1971), institutions are recognized to be just if we justify them as acceptable under the 'veil of ignorance' (this coincides with the *ex-ante* acceptance of them). *Ex post,* if there is public knowledge that institutions can be justified, and moreover if a public awareness has emerged that agents entertain reciprocal expectations of conformity, then agents develop a psychological attitude of conformity with justified institutions (the 'sense of justice') that can effectively counteract the self-interested incentive to act against institutions that could have been chosen in an 'original position' under the veil of ignorance.[2]

[1] To be sure, the last point is most compelling in a classical game theory context, wherein players may resort to the 'thought experiment' of putting themselves under a veil of ignorance to resolve an equilibrium selection problem (Binmore 1984, 1989). This point seems somewhat less compelling in an evolutionary game setting (like the one embraced in Binmore 2005) wherein myopic best responses will take care of the ex post convergence to an equilibrium. Nevertheless, if the veil of ignorance must be understood as the trigger of an evolutionary equilibrium selection dynamic, convergence is guaranteed only if it may causally affect the initial conditions of the system, wherefrom myopic best responses converge on a particular equilibrium. Again this not a matter of logic but an empirical fact concerning whether the ex-ante agreement may shape players' beliefs and attitudes so that they fall within the required basin of attraction (we thank an anonymous referee for having raised this point).

[2] It is noticeable that the 'sense of justice' does not work under the veil of ignorance since in the original position parties are modeled as 'mutually disinterested', instrumentally rational but completely ignorant about their personal identity and their individual plans of life (see Rawls 1971, pp. 146-147). Hence, under the veil, agents seek to advance their individual interests as much as possible, even though they do not know what these interests are. By contrast, the 'sense of justice' works beyond the veil when agents *ex post* – in a 'well-ordered society' (Rawls 1971, p.454) – once again know their personal identities. At that time, having public knowledge that the principles of current institutions have been impartially justified, and also having public knowledge of other participants' reciprocal conformity with those principles, they develop an attitude of reciprocity 'to answer

3. A conditional compliance hypothesis

3.1 This paper's intuition

Imagine a situation in which a group of individuals have to decide how to divide a sum of money – to the production of which they have by no means contributed – among themselves, and to this end they can agree on a distributive principle before knowing their actual roles and powers in their real-life interaction. Suppose that they do not know their actual *ex-post* probability of appropriating a certain share of the sum. We thus have a typical *ex-ante* choice of a distributive norm under a veil of ignorance, followed by the *ex-post* decision whether or not to comply with the chosen norm once the veil is lifted and players become aware, for instance, that they can earn higher payoffs by not conforming with the norm.

Intuition, together with a body of literature on the theory of justice (see, in particular, Rawls 1971; Harsanyi 1977; Barry 1989; Binmore 1984, 1989, 1997), suggests that if *ex-ante* – behind the veil of ignorance – agents cannot identify with any one of the possible roles that they may assume *ex-post*, then they are induced to reason impersonally and impartially, so that the principle accepted by agreement must be 'fair'.

The main question that we want to address is this: under what conditions may we expect to observe individuals implementing the agreement? Consistent with our intuition that the *ex-ante* agreement on the division rule will be 'fair', this question is to be answered by resorting to the Rawlsian 'sense of justice'. The idea of 'sense of justice', as applied to this situation, can be translated into a more operational 'conditional compliance' hypothesis according to which, in a strategic interaction among *N* players who agree on a principle of distributive justice (which may dictate a choice in contrast with their material self-interest), each player's decision to comply with the principle is conditional on the fact that (i) the player has taken part in the agreement (s/he has accepted the principle under the veil of ignorance) and (ii) because of the agreement, s/he believes that the other players will share the same beliefs and will comply.

in kind' (Rawls 1971, pp. 491, 494) – i.e., a desire to conform that stabilize justified institutions (Rawls 1971, pp. 454-456).

In Section 4 we report an experiment in a stylized social contract setting conducted to test the above-formulated conditional compliance hypothesis and its alternative interpretations. Our aim is to test whether participation in an impartial agreement over a fair principle of distribution explains the *ex-post* decision of complying with it and whether such an explanation involves each agent's reciprocal expectations of compliance, expectations held by the agents apparently because they reached the impartial agreement. Before describing the experiment, however, a short review of the relevant literature is in order.

3.2 Background literature

We can find a few attempts in the recent behavioral and psychological games literature to model the decision to comply with shared norms or principles by referring to motivations and psychological processes clearly linked with our conditional compliance hypothesis.

A first example is the theory of conformity preferences put forward by Grimalda and Sacconi (2005)[3] and based on psychological game theory (Geanakoplos et al. 1989; Dufwenberg 2008). The authors consider a situation in which players take part in an *ex-ante* agreement on a principle of distributive justice behind a veil of ignorance in which one can expect convergence on egalitarian principles. The players are characterized by a utility function that consists of both a material component – which depends on monetary payoffs – and a psychological component – which depends on players' expected degree of reciprocal conformity with the principle. The latter component is activated, and can overcome the material one, only if expectations of reciprocal conformity emerge. Also introduced into the model is the further (default reasoning) hypothesis that the agreement itself, in the absence of contrary evidence, is a sufficient condition for the emergence of such expectations (Sacconi and Faillo, 2010).

According to this model, a player characterized by conformity preferences complies with an agreement on a principle that dictates a choice in contrast with his/her self-interest if i) s/he participates in the *ex-ante* agreement on the principle, ii) s/he expects that other players who have contributed to choosing the principle will comply, and iii) s/he expects that others will expect that s/he

[3] See the Appendix.

will comply. Experimental tests of this theory have been conducted by Sacconi and Faillo (2010) and Tammi (2011).

Other theories of preference for compliance can be applied to the study of *ex-post* stability of the social contract, even if they are not centered explicitly on this goal.

Bicchieri (2006) argues that a player's compliance with a social norm is observed when s/he is aware of the existence of the norm and believes that a sufficiently large number of people comply with the norm (empirical expectations) and either a sufficiently large number of people think that s/he ought to conform or a sufficiently large number of people are ready to sanction him/her for not conforming (normative expectations). According to this approach, agreement on the norm is not a necessary condition for compliance, and it is replaced by a general idea of awareness of the existence of the norm (its salience) in the community of reference. In this context, one may say that the agreement is only one of the possible ways in which a norm becomes salient. Bicchieri translated her hypothesis into a formal model in which the player's utility depends both on his/her sensitivity to the norm, which in turn depends on normative expectations that are elicited from the interaction context, and on the number of norm-deviators. Bicchieri and Xiao (2007) showed, however, that when normative expectations and empirical expectations contradict one another, subjects choose according to the latter.

López-Pérez (2008) proposed a model of aversion to norm-breaking that is close to Bicchieri's and in which, in the absence of punishment threats, a player will obey a norm if s/he has internalized it with enough intensity – i.e., s/he suffers a psychological cost if s/he deviates – and s/he believes that a sufficient number of other players will comply with the norm. In his explanation of the experimental evidence, López-Pérez assumes that subjects bring into the laboratory a specific distributive norm (the E-norm) that dictates a choice compatible with maximization of a welfare function that combines efficiency and equality.[4]

[4] The idea of a psychological cost associated with deviation from the norm and/or from what the others expect us to do is similar to that of guilt aversion (Charness and Dufwenberg 2006; Battigalli and Dufwenberg 2007; see also Vanberg 2008). According to guilt aversion theory, people care about what others expect them to do and feel guilty if they do not fulfill what they think others' expectations are. Even if there is no explicit reference to

In all these models, players are characterized by mixed motives. They care about their material payoffs but they also have a certain degree of sensitivity to the norm. However, there are substantial differences concerning the *source* of this sensitivity. Whilst in explicitly contractarian theories this sensitivity emerges endogenously from the agreement, which is a part of the model and the experimental design (Grimalda and Sacconi 2005; Sacconi and Faillo 2010), in other models this sensitivity seems to depend on normative expectations simply elicited from the interaction context (Bicchieri 2006), or it is given exogenously by the experimenter (Lopez- Perez 2008). Note, however, that the decision to comply that we are investigating is conditional on participation in the impartial agreement – which is a necessary condition for the emergence of the sense of justice – and we consider reciprocal expectations of compliance to be also conditional on participation in the agreement and hence perform their explanatory function in conjunction with the agreement. Thus our hypothesis is more in line with the first model discussed above, and it also differs from the conditional cooperation hypothesis studied, for example, in experiments on the provision of public goods (Fischbacher et al. 2001).

3.3 Further related literature

Our conditional compliance hypothesis is compatible with results on the efficiency-enhancing effect of pre-play communication, for example in common pool resource experiments. As shown by Ostrom et al. (1992) self-regulation is possible if by engaging in preplay cheap talk people are able to agree on a joint cooperative strategy which implies the convergence of subjects' expectations and choices on a specific course of action. A similar convergence process is discussed in the experiment conducted by Walker et al. (2000) in which the subjects were able to propose and vote for allocation rules at each round of a common pool resource game. Note, however, that the process that induced subjects to coordinate on efficient outcomes in this experiment – and also, at least partially, in Ostrom et al.'s (1992) – is different from the process at the basis of our conditional compliance hypothesis. In Walker et al., for example, proposals and votes provide information to the group members, who, according to

shared norms or principles in this theory, one can easily see the existence of a norm as the source of second order beliefs.

the authors, learn how to coordinate on efficient outcomes that, if reached, are beneficial to the group. In our setting, efficiency concerns are ruled out, and the fair agreement induces a convergence of expectations on a strategy that might be inconsistent with some subjects' self-interest. There are also important differences with respect to Ostrom's experiments on cheap talk and ex-ante communication. The use of face-to-face communication does not allow control of the communication process, and it introduces a large number of potential variables into the explanation. The explanation based on preferences for conformity attributes great significance to an experimentally carefully controlled agreement process carried out under anonymity conditions. But what is more important is that, in our approach, no affective relationships based on personal mutual exchanges and personal identification are established among the parties, since these are intentionally omitted from the experimental design through the condition of anonymity.

In a broader sense, the hypothesis of conditional compliance underlying the experimental design and results of Section 4 can also be related to attempts in the public choice literature to study the emergence of self-enforcing norms and institutions from collective choices taken in anarchic interaction situations investigated by both experiments and historical case studies. As far as laboratory experiments are concerned, we refer to conjectures inspired by the theories of Thomas Hobbes, James Buchanan and Robert Nozick about the formation of institutions and the emergence of stable patterns of cooperative interaction in the absence of formal authorities (Powell and Wilson 2008; Powel and Stringham 2009; Smith et al. 2012). More interesting, however, are insights from historical case studies of anarchic societies which until relatively recent or even contemporary times lived as 'fringe groups' in relative isolation from the surrounding institutions because they could not resort to existing formal governments in order to define and enforce their internal rules of cooperation. Hence such 'fringe societies' can be studied as cases where governance rules emerged from their anarchic endogenous interactions without the external enforcement provided by pre-existing states. The eighteenth-century internal governance rules of pirate crews, the gypsy laws named *Romaniya*, and the *Leges Marchiarum* are some examples of such stateless societies. The emergence of self-governance rules in each of these 'fringe societies' has been explained by resorting to different economic models: Buchanan and Tullock's (1962) 'calculus of consent'; the efficient incentive effect of superstitions seen as payoff changes that allow the internalization of law-breaking costs and favor self-enforcement of norms by increasing the costs of free riding on the collective decision to ostracize culpable members of a 'fringe society'; and the folk theorem showing the existence of cooperative equilibria in repeated games (Leeson 2009a, 2009b, 2013).

Even if in broad sense all these studies, both experimental and historical, provide various approaches to the same general problem of self-regulation, self-enforceability and endogenous compliance with norms and institutions that we also investigate, none of them adopts our Rawlsian perspective. Moreover, none of them explains the apparently counter-interested decision to comply with a norm simply as a function of previous participation in an impartial agreement properly modelled as a rational decision under a veil of ignorance.

4. The experiment[5]

The experiment design is based on the Exclusion Game (Sacconi and Faillo 2010). This is a sort of 'triple mini-dictator game' in which three subjects – players A (A1, A2 and A3 respectively) – must decide how to allocate a sum S among themselves and a fourth subject – player B – who has no decisional power. In particular, A1, A2 and A3 have to decide separately and independently the amount to ask for themselves, choosing one of three possible strategies: asking for 25%, 30% or 33% of S. The payoff for players A is exactly the sum requested for themselves (a1, a2 and a3 respectively), while the payoff for player B is the remaining sum (S – a1 – a2 – a3). Each group is given 60 tokens – each token corresponded to € 0.50 – and each player A's strategies were: "Ask for 15 tokens", "Ask for 18 tokens", "Ask for 20 tokens".

The experiment consists of four treatments: the Baseline Treatment (BT), the Agreement Treatment (AT), the Outsider Treatment with One-sided information (OTO), and the Outsider Treatment with Two-sided information (OTT).

In the Baseline Treatment, participants are matched into groups of four and play the Exclusion Game.

[5] A summary of the experimental design and the questionnaire used to elicit beliefs are available at http://www.econometica.it/allegati/FOS_PUCH_Supplementary_material.pdf

In the Agreement Treatment, the participants are randomly matched into groups of four players and are told about the stages of the experiment and about the Exclusion Game. In the first stage, without knowing their role in the game, they take part in a voting procedure. In each group, they are invited to vote for one of three alternative rules (the fourth number is the type-B player's payoff): {15,15,15,15},{18,18,18,6}, {20,20,20,0}. The first rule assigns the same payoff to every member of the group; the second rule corresponds to a partial inclusion of player B in the share-out of the money; the third rule implies the total exclusion of the type-B player. The players have to reach unanimous agreement on the rule within ten trials. Voting is computerized and completely anonymous. The agreement is not binding, but only groups who reach agreement in this first stage can participate in the second stage. In the second stage, the composition of the groups is unchanged and roles are randomly assigned to play the Exclusion Game. Player A can either decide to implement the rule selected or choose one of the alternative allocations. Players who do not enter the second stage wait until the end of the session. Their payoff is the show-up fee.

In the Outsider Treatment with One-sided information (OTO) the first stage, as well as the rule on entering the second stage, are the same as in the Agreement Treatment (AT) . At the beginning of the second stage, the players are informed about their role, and the groups are re-matched. In particular, a player A for each group (hence called the outsider) is reassigned to a different group and told about the rule chosen by the new group, while the other members of the group (the insiders) do not know what rule the outsider's previous group has adopted. After the re-matching, the subjects play the Exclusion Game.

In the Outsider Treatment with Two-sided information (OTT), the design is exactly the same as in the OTO with the sole difference that now also the insider is informed of the rule chosen by the outsider's original group.

4.1 Experimental procedures

The experiment was run in both Milan (EELAB – University of Milan Bicocca) and Trento (CEEL – University of Trento). We ran three sessions for the BT (one in Milan and two in Trento), four sessions for the AT (two in Milan and one in Trento), five sessions for the OTO (three in Milan and two in Trento) and six sessions for the OTT (three in Milan and three in Trento). Overall, 332 undergraduate students – 164 in Milan and 168 in Trento – participated in the experiment. Fifty-six players were recruited for the BT, 72 for the AT, 88 for the OTO and 116 for the OTT. We have observations on 42 subjects A in the BT, 54 in the AT, 66 in the OTO and 87 in the OTT. The experiment was programmed and conducted using the z-Tree software (Fischbacher 2007). The instructions were read by the participants on their computer screens while an experimenter read them out loud.[6]

After the instructions had been read, and before the subjects were invited to make their decisions, some control questions were asked in order to ensure that the players had understood the rules of the game. At the end of each session, subjects were asked to fill in a questionnaire for the collection of socio-demographic data. On average, 55% of players A were male and the average age was 21.

Players were given a show-up fee of three euros.

4.2 Beliefs elicitation

In all the treatments, at the end of the game and before the players were informed about the decisions made during the *Exclusion Game* by the other co-players, first- and second order expectations (both normative and descriptive) were elicited by means of a brief questionnaire. In particular, in each group each player made statements concerning:

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

2. the probabilities of each possible co-player’s expectations about his/her own choice (Second Order Empirical Expectations - SOEE);

3. the choice that co-players considered to be the ‘right’ one (Normative Expectations - NE).

[6] The English translation of the instructions is available at http://www.econometica.it/allegati/FOS_PUCH_Supplementary_material.pdf

In both Outsider Treatments, guesses about the behavior and beliefs of insiders and outsiders were elicited separately. Only good guesses of the Empirical Expectations were rewarded on the basis of a quadratic scoring rule (Davis and Holt 1993).[7]

4.3 Hypotheses

Taking into account the design of the experiment, the generic 'conditional compliance' hypothesis, and the theoretical models presented in Section 3, we put forward the following empirical hypotheses.

Hypothesis 1: In the Baseline, the majority of active players should leave nothing to player B.

The Baseline treatment serves as a benchmark to assess the role of the agreement and group re-matching (see the hypotheses below). In this treatment subjects do not participate in the impartial agreement procedure. This implies that, even if they have conformity preferences (see the Appendix), they have no reason to expect that the other active players will choose any rule which is in contrast with the pursuit of their self-interest.

Hypothesis 2: In treatments AT, OTT and OTO, agreement should be reached by all groups.

In these treatments the single player has 0.25 probability of being player B (the dummy player). In this case, if the group agrees on the 20-20-20-0 rule, his/her payoff would be equal to the show-up fee of 3 euros. This is the same payoff that s/he would obtain if the group fails to reach the agreement. However, if there is a non-zero probability that the agreement will not be on the 20-20-20-0 rule and that the active players will respect it, the player selected as the dummy may expect a payoff from participation in the agreement greater than the show-up fee.

On the other hand, with probability 0.75 s/he will not be the dummy. In this case there is a non-zero probability that the group will agree on a rule that, if respected, gives a payoff greater that the show-

[7] We used the following scoring rule:

$$Q(p) = a - b\sum_{k=1}^{N}(I_k - p_k)^2,$$

where I_k takes value 1 if the event realized is event k and 0 otherwise; p_k is the probability associated with event k. The maximum score is a, and the minimum score is a-$2b$. We chose $a=2$ and $b=1$.

up fee to the active players. In any case, after the agreement s/he will have the possibility of asking for what s/he wants, obtaining a payoff greater than the show-up fee. All that considered, not contributing to the group's effort to reach agreement is a weakly dominated strategy.

Hypothesis 3: *The majority of the groups will agree on rule 15-15-15-15.*

The choice of the subjects under the veil of ignorance can be described as an impartial bargaining procedure whose solution corresponds to the selection of the rule that would be accepted from whichever point of view. In a bargaining situation with symmetrical *status quo* (the show-up payoff of three euros) and symmetrical strategic resources for each participant to make offers and counteroffers on distribution rules, the agreement consistent with Nash bargaining theory is the egalitarian one.

Moreover, under the veil of ignorance where each subject is able to assume any role in the game, there is no information about characteristics possessed by particular subjects that might reasonably suggest treating some of them differently. Hence the egalitarian rule 15-15-15-15 seems an obvious way to treat all of them impartially.

In this regard, consider our intuition in Section 3.1 whereby the solution of this agreement problem must be 'fair'. But what does 'fairness' mean in the context of the simple agreement depicted by this experimental design? In the particular situation studied here, what players know about the roles that they could undertake in the *ex-post* perspective does not give them any reason in the *ex-ante* choice of the rule to accept any asymmetry into the distribution. For example, since the amount of money to be divided is not produced by them and effort is not involved, it may appear arbitrary *ex ante* to introduce any inequality into the distribution rule.

Hypothesis 4: In AT, OTO and OTT active players will comply with the rule agreed under the veil of ignorance if i) they believe that other members of their group[8] *will comply (First Order Empirical Expectations compatible with the choice dictated by the rule)* and *ii) they believe that other members*

[8] In OTT and OTO, beliefs concern the choices made by the members of the group, including the outsider, *after* the rematching.

of the group expect that they will comply (Second Order Empirical Expectations compatible with the choice dictated by the rule).

This hypothesis derives from the application of conformist preferences theory to this experimental setting.

The introduction of treatments with rematching enables us to understand what drives compliance. In particular, by considering the comparison among AT, OTO and OTT, we can assess the relative importance, in term of expected compliance, of the *statement* of the norm compared with participation in the process of impartial agreement that reached the norm.

Subjects' experiences in OTO and OTT have the same characteristics as the experiences of people entering a new group or a society. In the former, the hosts know the constitutional rules of the newcomer's society of origin; in the latter, they are instead uncertain about it.

We introduce two alternative hypotheses about the roles of these two variables.

Hypothesis 5a: If active players agreed on the same rule, even if in different groups, they should comply with that rule and expect the others to do the same.

If groups are rematched but the outsider's original group agreed on the same rule chosen by the insiders, rematching should not affect the degree of compliance. This implies that the degree of compliance in AT should not be significantly different from that in OTT and both should be higher than that observed in OTO, in which insiders are not sure about the rule chosen by the outsider.

With regard to the last point, note that if Hypothesis 3 is true, then we should observe very few groups not agreeing on the 15-15-15-15 rule. If we restrict ourselves to comparison between AT and OTT, this would prevent us from reaching a meaningful conclusion, because we would have very few cases in which the outsider comes from a group that agreed on a rule different from the one chosen by the insiders. Comparison between OTT and OTO enables us to solve this problem of interpretation.[9] According to Hypothesis 5a, we should observe less compliance in OTO because insiders are not sure about the rule chosen by the outsider's original group.

[9] We thank an anonymous referee for suggesting this solution.

We put forward also an alternative hypothesis.

Hypothesis 5b: If active players agreed on the same rule, but in different groups, then they should not comply with that rule and expect non-compliance by the others.

According to this hypothesis, from the point of view of an active player who has to decide whether or not to comply with the agreed rule, it is not enough to know that other active players have chosen the same rule. S/he has also to be sure that the rule was selected by the same agreement process wherein other active participants acted exactly in the same way as those (anonymous) with whom s/he has reached the agreement (so that they cannot be differentiated from them).

This implies that the degree of compliance in AT should be higher than that observed both in OTT and OTO, while we should not expect any significant difference between OTT and OTO.

5. Results[10]

In analyzing the results of the experiment, we first consider the relation between beliefs and behavior. We then test whether and how different scenarios influence beliefs and, consequently, people's decisions.

Result 1. Subjects' choices tend to be in line with both their Empirical and Normative Expectations, but Empirical Expectations have a central role in the explanation of subjects' decisions.

The first point that we want to check is whether choices are in line with expectations. If this is the case, we want to determine what kinds of expectations are most closely correlated with our subjects' choices. In carrying out this analysis we will also distinguish between subjects who comply with the voted rule and subjects who deviate from it.

First of all, we want to investigate whether subjects show any differences among the different kinds and levels of expectations. In particular:

[10] Additional tables, figures and details on the regression analysis are available at http://www.econometica.it/allegati/FOS_PUCH_Supplementary_material.pdf

a) We have to check whether first-order (FOEE) and second-order (SOEE) Empirical Expectations converge on the same action profile. When we collected data on FOEE, we asked subjects to provide their subjective probabilities for all the possible choices of the other players in the group. In other words, we asked subjects to predict the probability that the other players in the group would ask for 15, 18 or 20 tokens. We call these probabilities FOEE15, FOEE18 and FOEE20 respectively. The same procedure was implemented to elicit SOEE – SOEE15, SOEE18 and SOEE20. Consequently, the first step of our analysis focuses on a comparison between FOEE and SOEE for each possible choice. The null hypotheses that we want to test for each treatment through a series of Wilcoxon tests are:

1) H_0_1: FOEE15 = SOEE15; 2) H_0_2: FOEE18 = SOEE18; 3) H_0_3: FOEE20 = SOEE20[11]

It turns out that no significant difference at the 5% level emerges when comparing FOEE and SOEE. In other words, in all the treatments, Empirical Expectations of first and second order tend to converge on the same profile of action. This means that, generally, subjects believed that the other players in the group would make the same choice that they thought they would make.

In terms of compliance this implies that most of subjects who participated in a treatment with the voting stage – 89%, 88% and 90% in the AT, OTO and OTT respectively – had either a reciprocal expectation of compliance or a reciprocal expectation of non-compliance.

b) The second step consists of comparing Empirical Expectations and Normative Expectations (NE). When we elicited NE, we asked participants to declare what they thought the others thought a player A *should* do. In the OTO and in the OTT, insiders were asked to declare both the number of tokens the other insider thought an insider should ask for (NE_I1) and the number of tokens the outsider thought an insider should ask for (NE_I2). Outsiders were asked to declare what they thought the others thought that the outsider should do (NE_O). Consequently, if we want to compare Empirical Expectations – reported as subjective probabilities – with NE, we have to work on the former in order

[11] In order to take account of the possible dependence of observations within the groups, we perform a robustness check in our econometric analysis running several specifications with clustered errors at the group level. The results and conclusions do not change when switching from the specifications without correction to the regressions with clustered errors.

to make the two variables comparable. The most natural solution is to compare the number of tokens a player thought the others thought a player A *should* ask for (NE) with the number of tokens a player thought was more likely to be selected (FIRST_FOEE). In other words, for each player FIRST_FOEE reports the option that obtains the highest probability when eliciting FOEE. Both in the OT and in the OTT, we will disentangle for insiders the number of tokens they thought that the other insider was more likely to ask for (FIRST_FOEE_I) and the number of tokens the outsider was more likely to ask for (FIRST_FOEE_O). Obviously, for outsiders we have data on FIRST_FOEE_I only.

In order to compare FIRST_FOEE and NE, we run a series of Fisher-exact tests[12] and Spearman's correlation tests. The former allows testing of the null hypothesis of independence of the two variables; the latter provides a measure of the intensity of the association between the two variables. Generally, we can conclude that NE are in line with Empirical Expectations.

c) The third step is to look at the relation between subjects' choices and beliefs. This allows us to identify different types of subjects.

Table 1 reports the number of subjects whose choice was coherent with different sets of beliefs, distinguishing between subjects who complied with the voted rule and subjects who did not comply in the last three treatments. In the first column, for example, there are the subjects whose choice was coherent with First Order (FIRST_FOEE), Second Order (FIRST_SOEE) Empirical Expectations and Normative Expectations (NE). Thus, the 16 subjects in the first row chose X, believed that the other members of their groups were more likely to choose X, believed that the other members of their group were more likely to believe that they would choose X, and believed that the other members of their group thought that a generic active player should choose X.

[TABLE 1]

It is worth noting that, in the case of compliance with the norm, generally the most represented category of subjects (58%) was that of those whose choice is coherent with FIRST_FOEE, FIRST_SOEE and NE. At the same time, 86% of subjects who tended to choose only according to their empirical beliefs were non-compliant.

[12] We run Fisher-exact tests since expected values for more than one cell are less than five.

Table 1 shows that Empirical Expectations have a central role in explaining the choice made by our subjects. Moreover – as in Bicchieri and Xiao (2007) – *if we limit our analysis to the cases in which Normative Expectations contradict Empirical Expectations*, we observe that the latter play a more important role in the players' decision-making and are significantly correlated with the subjects' choices in all cases (Spearman's coefficient > 0.52, p < 0.024), while Normative Expectations are not correlated with choices (Spearman's test, p > 0.12).[13]

Given this evidence, in what follows the analysis – and in particular the proofs of Results 3 and 4 – will focus on Empirical Expectations.

Result 2. When agreement is possible, it is reached by all groups. Moreover, almost all groups agree on the 15-15-15-15 rule.

This result is perfectly in line with Hypothesis 2 and Hypothesis 3. When agreement is possible, it is reached by all groups. Sixty-three groups out of 69 chose the 15-15-15-15 rule, while four groups chose the 18-18-18-6 rule and two groups the 20-20-20-0 one. In all treatments where the voting procedure was implemented, the first choice of more than 70% of players was the 15-15-15-15 rule. On running a binomial test (choosing the 15-15-15-15 rule against choosing another rule), we find that these values are significant (p = 0.000 in all treatments).[14] On average, the 15-15-15-15 rule was reached after 2.3 trials, while the 18-18-18-6 rule and the 20-20-20-0 rule after 2.6 and 4.7 trials, respectively.

Result 3. In the BT and in the AT, subjects' expectations differ significantly. They lead subjects to choose more frequently 20 tokens in the BT and 15 tokens in the AT.

This result is in line with Hypothesis 1 and Hypothesis 4. In the BT around 74% of the players asked for 20, while in the AT only 37% of the participants asked for the maximum. At the same time, less

[13] Test run only on observations for which NE and FIRST_FOEE differ.

[14] Note that the distributions of the first choices and of the voted rule in the AT, in the OTO and in the OTT are almost identical. A Pearson's Chi-squared test on the former (p = 0.941, Cramer's V = 0.043) and a Fisher-exact test on the latter (p = 0.642, Cramer's V = 0.081) did not reject the null hypothesis of independence between first choices and the treatment and between voted rules and the treatment respectively.

than 5% in the BT and 46% in the AT asked for 15 tokens. How can we explain this difference? First of all, we use an ordered probit regression to check what factors affected subjects' choices. The specification is:

$$CHOICE_i = \gamma AT_i + \beta_1 AGE_i +_i \beta_2 MALE_i + PAST\_PART_i + \varepsilon_i \qquad \text{(R1)},$$

where *CHOICE*$_i$ is equal to the number of tokens that subject *i* with the role of player A asks for him/herself. It can be equal to 20, 18 or 15. *AT*$_i$ is equal to 1 if the observation comes from a subject who plays in the AT; 0 otherwise. *PAST_PART*$_i$ is equal to 1 if the observation comes from a subject who participates in at least one other experiment before; 0 otherwise. *AGE*$_i$ and *MALE*$_i$ are demographic variables reporting the age and the gender of the players.

According to the results from (R1), participants in the AT are 49 percentage points less likely to ask for 20 tokens and 38 percentage points more likely to ask for 15 tokens. At this point, the second step is to check whether subjects' Empirical Expectations change when switching from the BT to the AT. Again, we run an ordered probit regression where the dependent variable is *FIRST_FOEE*$_i$.[15] The specification is:

$$FIRST\_FOEE_i = \alpha AT_i + \beta_1 AGE_i +_i \beta_2 MALE_i + PAST\_PART_i + \varepsilon_i \qquad \text{(R2)},$$

where *FIRST_FOEEi* is equal to the number of tokens that player A thinks the other players in the group are more likely to ask for themselves. It can be equal to 20, 18 or 15.[16]

It turns out that in the AT players are 65 percentage points less likely to think that the other players in the group will ask for 20 tokens while they are 46 percentage points more likely to think that the other players in the group will ask for 15 tokens.

[15] SOEE are omitted because of perfect collinearity with FOEE.

[16] We decided to focus our analysis on the option that was considered as the most probable because we think it is more informative that the probability *per se*. For instance, saying that subject *i* thinks that the other player asks for 20 tokens with a probability of 40%, provides partial information. In fact, it may mean that player *i* assigns a 60% probability to another option as well as that she assigns probabilities of 35% and 25% to the remaining two options. In the latter case, the 20-token option is believed to be the most probable, while in the former case it is not. We think that this information cannot be simply ignored.

Consequently, we again run the ordered probit regression on people's choices including two new control variables: *FIRST_FOEE20* (a dummy variable equal to 1 if player *i* believes that the others will ask for 20 tokens) and *FIRST_FOEE15* (a dummy variable equal to 1 if player *i* believes that the others will ask for 15 tokens). The new specification is:

$$\begin{aligned} CHOICE_i &= \delta_1 FIRST\_FOEE20_{\mathrm{i}} + \delta_2 FIRST\_FOEE15_{\mathrm{i}} + \gamma AT_i + \\ &+ \beta_1 AGE_i +_i \beta_2 MALE_i + PAST\_PART_i + \varepsilon_i \end{aligned} \quad \text{(R3).}$$

From (R3) it emerges that, once we include Empirical Expectations in the regression, the AT coefficient no longer is significant .

We then perform a further analysis to test the robustness of our results through a series of recursive bivariate probit regressions[17] where subjects' choices are the dependent variable of the structural equations with *FIRST_FOEE* as explanatory factors, while *FIRST_FOEE* are the dependent variable of the reduced-form equations. We run a regression for each of the two most selected options – 20 and 15. Thus:

$$\begin{aligned} CHOICE\_20_i &= \delta_1 FIRST\_FOEE20_i + \beta_1 AT_i + \beta_2 MALE_i + \beta_3 AGE_i + \varepsilon_{1i} \\ FIRST\_FOEE20_i &= \alpha AT_i + \beta_4 PAST\_PART_i + \beta_5 MALE_i + \beta_6 AGE_i + \varepsilon_{2i} \end{aligned} \quad \text{(R4),}$$

where *CHOICE_20$_i$* is equal to 1 if subject *i* chooses 20 tokens; 0 otherwise. And:

$$\begin{aligned} CHOICE\_15_i &= \delta_1 FIRST\_FOEE15_i + \beta_1 AT_i + \beta_2 MALE_i + \beta_3 AGE_i + \varepsilon_{1i} \\ FIRST\_FOEE15_i &= \alpha AT_i + \beta_4 PAST\_PART_i + \beta_5 MALE_i + \beta_6 AGE_i + \varepsilon_{2i} \end{aligned} \quad \text{(R5),}$$

where *CHOICE_15$_i$* is equal to 1 if subject *i* chooses 15 tokens; 0 otherwise.

The advantage of this model is that it makes it possible to check for the recursive nature of the decisional process – the treatment influences Empirical Expectations, and Empirical Expectations influence choices.[18] It shows that:

[17] A variation of the analysis run by Di Novi (2007).

[18] The error terms are assumed to be independently and identically distributed as bivariate normal

1) the agreement influences Empirical Expectations. In fact, in the AT it is less likely that subjects think that the other members of their group will ask for 20 tokens (*AT* in the reduced-form equation in (R4), β_1 = -1.901, p = 0.000). This is not surprising if we consider that in the AT, 17 groups out of 18 chose the 15-15-15-15 rule, and 1 chose the 18-18-18-6 one. At the same time, there is an increase in the probability that subjects think that the others will ask for 15 tokens (*AT* in the reduced-form equation in (R5), β_1 = 2.507, p = 0.000).

2) Empirical Expectations influence subjects' decisions. From the structural equations, it turns out that players who expect that the other members of the group will ask for 20 tokens are more likely to behave selfishly when playing the Exclusion game (*FIRST_FOEE20* in (R4), δ_1 = 2.424, p = 0.034), while those who believe that the others will ask for 15 tokens are more likely to ask for 15 tokens as well (*FIRST_FOEE15* in (R5), δ_1 = 2.505, p = 0.001).

3) the causality is unidirectional. The value of ρ is not significantly different from 0 in both cases (p = 0.971 in (R4) and p = 0.652 in (R5)).

Result 4. Expectations of compliance are higher when the composition of the groups remains unchanged. This induces more compliance in the AT with respect to the two treatments with rematching (OTO and OTT).

$$\begin{pmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{pmatrix} \approx IIDN\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

This model allows a test of exogeneity to be run. In particular, the exogeneity condition is stated in terms of the correlation coefficient ρ. When ρ = 0, $FOEE\_20_i^*$ and ε_{1i} are uncorrelated and $FOEE\_20_i^*$ is exogenous for the first equation in (R4). This would imply that by means of this model we can detect not only the relation among agreement, beliefs and choices but also whether the causality between beliefs and choices is unidirectional. In fact, if choices influenced beliefs, $FOEE\_20_i^*$ and ε_{1i} would be correlated and ρ would be significantly different from 0. The same procedure applies to (R5). The usual VIF procedure is used to ensure that no multicollinearity problem occurs. Moreover, we follow Maddala (1983), according to whom at least one of the reduced-form exogenous variables should not be included in the structural equation as explanatory variables.

About half of the subjects complied with the voted rule in the AT. When we rematched the groups (in the OTO and in the OTT), a smaller percentage of players complied with the chosen rule (39% and 44% respectively). Also in this case we want to understand whether variations in players' expectations may explain this difference in their behaviour. This means that, once again, we want to show that the difference between AT and OTO and AT and OTT is a consequence of the impact of the outsider on players' beliefs. Also in this case, the most suitable econometric model is a recursive bivariate probit model where the subjects' decision to comply (*COMPLIANCE*) is the dependent variable of the structural equation with the expectation of compliance as explanatory factor, while the expectation of compliance is the dependent variable of the reduced-form equation. We provide results from three specifications. In (R6) we compare the three treatments. Thus:

$$COMPLIANCE_i = \delta_4 EXP\_COMPLIANCE_i + \varphi_1 OTO_i + \varphi_2 OTT_i + \varphi_3 FIRST\_RULE_i + \\ + \varphi_4 MALE_i + \varphi_5 AGE_i + \varphi_6 OUTSIDER_i + v_{1i}$$

$$EXP\_COMPLIANCE_i = \omega_1 OTO_i + \omega_2 OTT_i + \varphi_7 TRIALS_i + \varphi_8 OTO * TRIALS_i + \varphi_9 OTT * TRIALS_i + \\ + \varphi_{10} PAST\_PART_i + \varphi_{11} MALE_i + \varphi_{12} AGE_i + \varphi_{13} OUTSIDER_i + v_{2i} \qquad \text{(R6),}$$

where $COMPLIANCE_i$ is equal to 1 if subject *i* chooses to comply with the voted rule when playing the Exclusion Game; 0 otherwise. $EXP\_COMPLIANCE_i$ is equal to 1 if a subject expects that the most probable option is that others comply with the voted rule; 0 otherwise. OTO_i is equal to 1 if the observation comes from a subject who plays in the OTO; 0 otherwise. $FIRST\_RULE_i$ is equal to 1 if the rule chosen by the group is the first voted rule: 0 otherwise. $TRIALS_i$ reports the number of trials that subjects played before reaching an agreement. $OTO*TRIALS_i$ reports the number of trials that subjects played before reaching agreement when the observation comes from a subject who plays in the OTO; 0 otherwise. OTT_i is equal to 1 if the observation comes from a subject who plays in the OTT; 0 otherwise. $OTT*TRIALS_i$ reports the number of trials subjects played before reaching agreement when the observation comes from a subject who plays in the OTT; 0 otherwise. $OUTSIDER_i$ is equal to 1 if the observation comes from an outsider; 0 otherwise.

Again, we can check: 1) the relation among agreement, beliefs and choices; 2) whether the eventual causality between beliefs and choices is unidirectional.

We find that:

1) the agreement influences Empirical Expectations. In fact, both in the OTO and in the OTT it is less probable that subjects think that the other members of their group will comply (*OTO* and *OTT* in the reduced-form equation, $\varphi_1 = -1.165$, p = 0.014 and $\varphi_2 = -1.071$, p = 0.027).

2) Empirical Expectations influence subjects' decisions. From the structural equation, it turns out that players who expect that the other members of the group will comply are more likely to comply when playing the Exclusion game (*EXP_COMPLIANCE*, $\delta_4 = 2.623$, p = 0.000).

3) the causality is unidirectional. The value of ρ is not significantly different from 0 (p = 0.415).

Result 5. Expectation of compliance and, consequently, the level of compliance in the OTO and in the OTT are not significantly different.

In order to check whether the lack of information about the rule chosen by the outsider in his/her original group affects subjects' expectation of compliance and, consequently, their decision to comply, we run a small variation of the recursive bivariate probit model (R6) that we will call (R7). In this new regression, the reference treatment is OTO and *AT* and *OTT* are the two dummy variables we insert to test the treatment effect.

It turns out that expected compliance, and consequently the level of compliance, does not significantly differ when we compare OTO and OTT (*OTT* in the reduced-form equation, $\varphi_2 = 0.094$, p = 0.844).

Results 4 and 5 are in line with Hypothesis 5b.

6. Discussion and conclusions

In this article we have contributed to the long-standing debate on the ex-post stability of the social contract using the toolboxes of behavioral game theory and experimental economics. Our interest in particular has been in the motivations at the basis of subjects' choices to conform with a norm chosen through an impartial agreement and in the absence of legal enforcement, where self-interest cannot help in supporting compliance.

By comparing the behavior of subjects in the Baseline Treatment with that observed in other treatments, we could assess the effect of the agreement on the subjects' expectations and choices in the Exclusion Game. Comparison among the choices made in the three treatments with the agreement enabled us to assess the relative importance, in terms of expected compliance, of the *statement* of the norm compared with participation in the process of impartial agreement.

To summarize our main findings, in the Agreement Treatment we observed that all groups reached an agreement, that the large majority of groups agreed on the equal division rule, and that a high percentage of subjects chose to comply with the rule believing that other members of their group would do the same. In addition, on considering the relation among agreement, expectations and actual choices, we can conclude that the agreement 'under the veil' induced the convergence of subjects' beliefs of reciprocal compliance, and consequently activated a preference to act in accordance with fairly agreed principles conditionally on reciprocal compliance beliefs.

The evidence on the Outsider Treatments suggests that participation in the agreement induces convergence of empirical expectations, which in turn induces the decision to comply. Specifically, there is evidence that outsiders and insiders, even if they agreed on the same norm in their previous step of agreement (but as members of different groups), reduce their degree of compliance when they are matched with each other in an Outsider Treatment, no matter whether the information flow is one-sided or two-sided. The fact that the rate of compliance is the same in both Outsiders Treatments and smaller than in the Agreement Treatment implies that what really matters is that, when players have to decide whether or not to comply, they know for sure that the other active players have participated in the *same* agreement procedure. This in turn might be explained by considering that subjects who have taken part in the same procedure know what happened during the voting stage: for example, what rules were proposed during the process. We can conclude that what drives compliance in the ex-post stage of play is not the sameness of the statement of the particular rule achieved by agreement but the awareness and public knowledge that also other participants have taken part in the same and equally impartial and impersonal agreement process – so that the parties could judge its outcome as 'fair'.

We make no reference here to the force of personal direct relationships established during the agreement stage as explanation of this preference for compliance with members of the same group, even when the outsider is known to be characterized by an identical rule. In fact, it is obvious that no personal direct relationships can be established amongst subjects in our process of agreement under the veil of ignorance, which is specifically designed to incorporate maximal impersonality. Nevertheless, some sort of solidarity among those fellow-members of the group who have equally participated in an impartial and impersonal procedure of agreement – whereby its outcome (whatever it will be) can considered as treating each of them 'fairly'– can be conjectured. Hence, the evidence on the two treatments with re-matching can aid understanding of some fundamental facts that characterize the migration of people across countries, organizations and groups, and the bias in favour of fellow-citizens characterised by their belonging to the same impartial, constitutional collective choice processes for establishing principles and norms.

More generally, our results are pertinent to many domains of application. They contribute to definition of the conditions under which inclusion and cooperation can emerge as self-sustainable forms of interactions. In this regard, our research can be seen as an example of empirical institutional economics inspired by the work of Elinor Ostrom and colleagues on the role of pre-play communication.

At the same time, by adopting a genuine behavioral perspective, we have attempted to provide a normative theory (the social contract) with some realistic psychological support. This suggests that, first (from the perspective of a 'quasi-empirical' test of normative theories), in an appropriately designed pre-play situation resembling an impartial agreement, real-life agents would justify institutions consistent with the prescriptions of social contract theory. Second (from a policy-oriented viewpoint), by appropriately designing a process of *ex-ante* agreement in order to reach shared consensus on a norm, the *ex-post* problem of compliance can easily be solved, since agents will be ready to comply voluntarily with that norm. Last (from a historical-explanatory viewpoint), endogenous *compliance* with some real-life social and legal institutions – even if they are not construable as working in the immediate self-interest of the parties involved – can be nonetheless explained on the basis of the simple fact that agents have reasonably *accepted* them. In other words, if during the process of formation of an institution, a situation of pre-play impartial and impersonal agreement can be retrieved as a salient part of the process that led to the institution, then there is a satisfactory basis for understanding why agents comply with it, even though compliance is not in their immediate best self-interest.

Appendix. The conformity preferences model.

The conformity preferences model (Grimalda and Sacconi 2005) is a variation of Rabin's (1993) theory of reciprocity in which the kindness functions are replaced by conformity indexes. In a normal form game with two players (*i* and *j*), a first index f_i is defined as a measure of the extent to which, by choosing a specific strategy, player *i* contributes to the implementation of the agreed principle of fairness *T* – i.e., maximizing the Nash Bargaining Product as a principle of fair distribution accepted by an ex-ante agreement through pre-play communication – given his/her belief about player *j*'s strategy. The second index $\tilde{f}_j$ measures player *i*'s expectation of the extent to which player *j* is contributing to the implementation of the same principle given *i*'s beliefs about *j*'s beliefs about *i*'s choice. Each index, given a conjecture on the other player's behavior, measures the distance between the outcome that a player contributes (or is believed to contribute) to realizing and the ideal outcome – the one in which the value of the social welfare function is maximum. Both f_i and $\tilde{f}_j$ take values between -1 (zero conformity) and 0 (full conformity). An overall index of conformity $F= (1+ f_i)(1+ \tilde{f}_j)$ is then defined. Hence the utility function V_i of the generic player *i* is given by

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

where U_i is player i's material utility associated with the outcome σ; $\lambda_i > 0$ is the weight of the deontological motivation (to conform with the principle) in the utility function; *T* is the Nash Bargaining Product defined over σ; and *F* is defined as above. Thus, when agent *i* does not comply with the norm or s/he does not expect any reciprocal conformity by agent *j* – i.e., either the index of conditional conformity (f_i) or reciprocal expected conformity $(\tilde{f}_j)$ are equal to -1 – *F* reduces to zero.

To be precise, player *i*'s personal index of conditional conformity has the following form:

$$f_i\left(\sigma_i,b_i^1\right)=\frac{T\left(\sigma_i,b_i^1\right)-T^{MAX}\left(b_i^1\right)}{T^{MAX}\left(b_i^1\right)-T^{MIN}\left(b_i^1\right)}$$

where b_i^1 is player *i*'s belief concerning player *j*'s action (i.e., a prediction formally identical to a strategy of player *j*) , $T^{MAX}\left(b_i^1\right)$ is the maximum attainable by the function T given *i*'s belief, $T^{MIN}\left(b_i^1\right)$ is the minimum attainable by the function T given *i*'s belief, $T\left(\sigma_i,b_i^1\right)$ is the effective level attained by T when the player adopts strategy σ_i, given his/her belief about the other player's behavior. On the other hand, the second index represents player's *i* assessment of player's *j* reciprocal conformity and has the following form:

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

where b_i^1 is player *i*'s first order belief about player *j*’s action (i.e., player *j*’s predicted strategy), b_i^2 is player *i*'s second order belief about player *j*’s belief about the action adopted by player *i* (i.e., formally identical to a player ’s strategy as predicted by player *j*).

References

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

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton, NJ: Princeton University Press.

Barry, B. (1989). *Theories of Justice*. London: Harvester-Wheatsheaf

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review: Papers and Proceedings,* 97(2), 170-176.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge/New York: Cambridge University Press.

Bicchieri, C., & Xiao E., (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, 22(2), 191-208.

Binmore, K.G. (1984). *Game Theory and The Social Contract*. STICERD working paper, 84/108

Binmore, K.G. (1989). The social contract: Harsanyi and Rawls. *The Economic Journal,* 99(395), 84-106.

Binmore, K.G. (1994). *Playing Fair.* Cambridge MA: MIT Press.

Binmore, K, G., (1997). *Just Playing.* Cambridge MA: MIT Press

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

Buchanan, J.M., and Tullock G. (1962). *The Calculus of Consent*. Ann Arbor: Michigan University Press.

Buchanan, J.M. (1975). *The Limits of liberty*. Chicago: Chicago University Press.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74 (6) , 1579-1601.

Davis, D. D.,& Holt, C.A. (1993). *Experimental Economics.* New Jersey: Princeton University Press.

Di Novi C. (2007). An Economic Evaluation of Life-Style and Air-pollution-related Damages: Results from the BRFSS. *JEPS Working Papers 07-001, JEPS.*

Dufwenberg, M. (2008). Psychological Games. In Durlauf, S.N. & Blume, L. E. (Eds.). *The New Palgrave Dictionary of Economics* (2nd Edition). Basingstoke, Hampshire / New York: Palgrave Macmillan.

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics.* 10(2) , 171-178.

Fischbacher, U., Gachter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters.* 71(3), 397-404.

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

Geanakoplos, J., Pearce, D. & Stacchetti E. (1989) Psychological Games and Sequential Rationality. *Games and Economic Behavior*, 1(1), 60-79.

Grimalda, G., & Sacconi E. (2005). The Constitution of the Not-for-Profit Organisation: Reciprocal Conformity to Morality. *Constitutional Political Economy*. 16 (3), 249-276.

Hardin, R. 1999. *Liberalism, Constitutionalism, and Democracy*. Oxford: Oxford University Press.

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

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

Hume, David. [1739-40] 1978. A Treatise of Human Nature. Oxford: Oxford University Press.

Leeson, P.T. (2009a). The Calculus of Practical Consent: the Myth of the Myth of Social Contract. *Public Choice*, 139 (3-4), 443-459.

Leeson, P.T. (2009b). The Laws of lawlessness. *The Journal of Legal Studies*, . 38 (2), 471-503.

Leeson, P.T. (2013). The Gypsy law. *Public Choice*, 155 (3-4), 273-292.

López-Pérez, R. (2008). Aversion to norm-breaking: A mode. *Games and Economic Behavior*, 64(1), 237-267.

Lewis D. (1969). *Conventions. A Philosophical Study*. Cambridge MA: Harvard University Press.

Maddala G.S., (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Ostrom, E., Walker, J.M., & Gardner, R. (1992). Covenants with and without swords: self-governance is possible. *American Political Science Review*. 86(2), 404-417.

Powell, B., & Wilson, B. J. (2008). An experimental investigation of Hobbesian Jungles. *Journal of Economic Behavior & Organization*. 66(3), 669-686.

Powell, B., & Stringham, E. P. (2009) .Public choice and the economic analysis of anarchy: a survey. *Public Choice*. 140(3-4), 503-538.

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

Rawls, J. (1971). *A Theory of Justice*. Cambridge MA: Harvard University Press.

Sacconi, L., (1993a). Equilibrio e giustizia (I): la stabilità del contratto sociale. *Il giornale degli economisti e annali di economia*, LII (10/12), 479-528

Sacconi, L., (1993b). Equilibrio e giustizia (II): la selezione del contratto sociale. *Il giornale degli economisti e annali di economia.* LII (10/12), 529-575

Sacconi, L., & Faillo, M. (2010). Conformity, reciprocity and the sense of justice. How social contract-based preferences and beliefs explain norm compliance: the experimental evidence. *Constitutional Political Economy*. 21(2), 171-201.

Smith, A. C., Skarbek, D. B., & Wilson, B. J. (2012). Anarchy, groups, and conflict: an experiment on the emergence of protective associations. *Social Choice and Welfare*. 38(2), 325-353.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge U.P.

Sugden, R. (1986). *The Economics of Rights, Cooperation and Welfare*. London: Blackwell.

Tammi, T. (2011). Contractual preferences and moral biases: social identity and procedural fairness in the exclusion game experiment. *Constitutional Political Economy*. 22(4), 373-397.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations”, *Econometrica*. 76(6), 1467-1480.

Walker, J.M, Gardner, R., Herr, A., & Ostrom, E. (2000). Collective choice in the commons: experimental results on proposed allocation rules and votes. *The Economic Journal,* 110(460), 212-234.

Tables

Table 1. Subjects’ choices and beliefs

	Choices are in line with …			
	FIRST_FOEE + FIRST_SOEE + NE	**FIRST_FOEE + FIRST_SOEE**	**OTHER**	**TOTAL**
BT	16	19	7	42
AT _ COMPLIANT SUBJECTS	16	5	6	27
AT _ NON-COMPLIANT SUBJECTS	8	13	6	27
OTO _ COMPLIANT INSIDERS	8	1	8	17
OTO_ NON-COMPLIANT INSIDERS	9	13	5	27
OTO _ COMPLIANT OUTSIDERS	5	1	3	9
OTO _ NON-COMPLIANT OUTSIDERS	5	6	2	13
OTT _ COMPLIANT INSIDERS	19	\	10	29
OTT_ NON-COMPLIANT INSIDERS	15	6	8	29
OTT _ COMPLIANT OUTSIDERS	5	\	4	9
OTT _ NON-COMPLIANT OUTSIDERS	9	7	4	20