Liberal egalitarianism in distribution of a common output. The role of agreement under ignorance and mutual expectations for selection and compliance.

Giacomo Degli Antoni[1,4] Marco Faillo[2*]., Pedro Francés-Gómez[3]. Lorenzo Sacconi[2,4]

Abstract

Drawing on the theoretical and experimental literature on distributive justice, we provide evidence on the influence that explicit agreement under the veil of ignorance may have on individuals' conception of justice and its implementation in a context in which income is not 'manna from heaven' but it is earned by working. One crucial characteristic of our experiment is that subjects were randomly assigned unequal endowments of working time. Thus, their work naturally generated unequal levels of earnings due to pure luck.

The main result is that the agreement under a veil of ignorance induces the subjects to accept a liberal egalitarian division rule not only in the ex-ante agreement but also in the actual implementation of the distribution of income, even if this contradicted their self-interest and some common economic assumptions about reciprocal expectations of rationality. In addition, our results show that deliberating through open discussion increases the level of ex-post compliance. This works therefore essentially contributes to the research on the psychological realism of theories of justice, both concerning the "realistic" choice of principles of justice ex ante, and the motivational force of compliance with principles of justice ex post.

Keywords: Distributive justice, liberal egalitarianism, dictator game, social contract, normative realism.

JEL Classification: C72,C91, D02, D63

*Corresponding author

1..Department of Law, University of Parma, via Università 12, Parma, Italy
2. Department of Economics and Management, University of Trento, via Inama 5, 38122, Trento, Italy.
3.Department of Philosophy I, Faculty of Philosophy, Campus de la Cartuja 18071 Granada, Spain.
4. EconomEtica, c/o University Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milan, Italy

Liberal egalitarianism in distribution of a common output. The role of agreement under ignorance and mutual expectations for selection and compliance.

1. Introduction

Liberal egalitarianism (henceforth LE), occupies a prominent place among theories of distributive justice (Rawls 1971, 1993; Dworkin 1981a, 1981b; Roemer 1986, 1996 Sen 2009). Liberal egalitarians embrace a view that usually involves a set of principles rather than a simple formula. Paradigmatically, according to Rawls (1971) primary socially goods should be in general distributed equally. However, with the exception of liberties and careers, which should be open to all according to "fair equality of opportunity", inequalities can be admitted for reasons of incentive, but only in so far as they are to the "greatest benefit of the least advantaged" (Rawls, 1971: 46). Other authors stress in similar ways two basic components of LE: equal respect – usually implying equal distribution of some basic goods– and some form of recognition of personal merit, responsibility, contribution. The philosophical argument behind LE tends to be highly speculative, often relying on mental experiments. Consequently, a frequent criticism of liberal egalitarian theories is that they lack realism, or that they are not actually applicable in most social contexts.

Shortly after Rawls's theory became popular, experimental research sought to establish whether people would in fact favor the distribution apparently demanded by the principles of justice. This line of inquiry has yielded scarce and mixed results so far (see section 3 below). Even if people might choose liberal egalitarian distributive principles as normative ideals, the abundant experimental research on what motivates people to act non-egoistically in certain contexts has not yet established whether considerations of justice are motivationally effective. It has been alleged that: "the extant literature that addresses preferences for redistribution suffers in part because people are not observed actually agreeing for their own income to be redistributed" (Esarey et al. 2012: 605). The results reflect what people think, rather than what they prefer or what they are willing to do.

This paper contributes to filling this gap by testing the *realism* of LE at two levels: principle acceptance and motivational force. It does so first by checking whether individuals situated behind a veil of ignorance would in fact choose LE as the legitimate principle for the distribution of a common output; and second by testing whether individuals would behave as they have agreed they should. This double test is conducted by means of an experiment in which pairs of players perform a productive task. In our main treatments, before the subjects start that task, and before knowing whether they would be favored or disadvantaged by chance, they must agree on the principle that should govern the distribution of the common output. After the task has been performed –and the initial inequality therefore revealed– each participant decides how to distribute what the pair has produced. This decision is entirely uncoerced. There are no further consequences (no reputation effect, no implication for the next rounds of the game, no possibility of punishment, etc.) if the previously agreed principle is not applied. We observe that subjects easily converge on LE as the legitimate distributive principle if they have the opportunity to reflect on it *ex-ante*, and they also comply with LE *ex-post*. Our data reveal that both *ex-ante* convergence and *ex-post* choice consistent with LE crucially depend on the possibility to agree impartially and on mutual expectations.

The rest of the paper is organized as follows. Section 2 defines LE, suggests a contractarian argument in its favor and illustrates our research questions and main hypothesis. Section 3 reviews the existing experimental literature on distributive justice and how this paper relates to it. Section 4 describes the experimental design and procedures, and presents the different treatments applied. Section 5 illustrates our results with respect to the empirical hypotheses that we derive from our general conjectures on the normative realism of LE. Section 6 concludes.

2. Theoretical background and research questions

We focus on selection and compliance with agreed principles for the distribution of the common output of a production activity carried out in an experimental game involving pairs of agents. In the benchmark situation, after the production task has been performed, both agents' outputs are pooled, and each agent chooses a rule of distribution or a share of the total output for him/herself. Then, the rule chosen by one of the agents is randomly selected and implemented, determining both agents' payoffs. The thrust of the paper, however, lies in an alternative situation that we consider: before carrying out their separate tasks – that is, *ex-ante*, and behind a veil of ignorance about the means with which they are relatively endowed to carry out the activity – agents must agree unanimously on a rule of distribution, and only if an agreement is reached are they allowed to proceed to the productive stage of the game. Afterwards they face a second decision (entirely similar to that faced in the benchmark case) in which each of them chooses whether to implement the agreed rule or another distribution. Then, again, random selection decides the individual choice to implement.

Each pair's total product is treated as a "common output". For the purposes of this study the details of joint production, technological complementarities in agents' resources or super-additivity of the production function, are omitted. "Common output" is simply understood in the sense that agents possess the output of their work in common, and in order to access any portion of it (or establishing a "property right") they have to choose a distributive rule - so that the payoff obtained by each agent depends on the rule/share chosen by both members of the pair.

One key feature of this situation is the random unequal allocation of agents' basic endowments with production means. This random allocation may be taken in consideration *ex-ante*, when agents agree on the distribution rule before the task is performed. The 'veil of ignorance' is an important feature of this *ex-ante* assessment, since when the parties are asked to agree they do not yet know who will happen to be favored by the random allocation of production means. In designing the distribution rule, agents may anticipate the effects that brute luck in endowments allocation may have on final payoffs. So agents may also agree on, and eventually implement, a rule that redresses inequalities due to brute luck. However, they cannot simply cancel the 'social and natural lottery' concerning productive endowments.

Liberal egalitarianism

We conjecture that the most reasonable distributive rule to be chosen in the *ex-ante* agreement on the distribution of a common output is Liberal Egalitarianism (LE). According to Rawls, any person, seen as a free and equal participant in a scheme of social cooperation, should have equal opportunity to take advantage of social cooperation. However, to be substantially fair, and not merely formal, 'equality of opportunity' requires consideration of the conditions whereby a person enters social cooperation and that involuntarily affect him/her opportunity to profit from it. This is called the 'basic structure' determining each person's endowment in terms of production means, resources (the term used by Dworkin), capabilities (Sen) or primary goods (Rawls). All these terms denote endowments that are instrumental to social production and hence affect the actual opportunity that persons have to profit from cooperation. Taking equal consideration and respect as the baseline principle entails that any inequality in the distribution of such endowments must be justified – brute social or natural luck is not self-justifying. Only the effects of the voluntary and intentional exercise of agency, counting as desert, merit, or a contribution to the creation of the conditions for the scheme of social cooperation, could be considered relevant. It is evident, however, that the amount of resources (or "endowments") that can be allotted to each person as the pre-condition for him/her productive life cannot be directly ascribed to her personal agency, and are not caused by him/her, precisely because they are (part of) the pre-condition for the exercise of his/her agency. Thus, equal consideration and respect entails that such resources must be distributed equally.[1]

However, LE does not deny that, once endowments have been fairly distributed, and the arbitrary effects of natural and social differences have been neutralized, agency and voluntary contribution may make the difference. An individual at this point can be held individually responsible for the outcome of her/his work, and the claim to be remunerated in proportion to her/his active contribution to the output for which he is responsible is justified. The final distribution should account for that, and depend on it.

Summing up, we understand LE as primarily a two-step distributive rule:

> **LE1:** endowments (practical opportunities, tools of production, resources, capabilities or primary goods) which are necessary pre-conditions for participation in the production of a common output, should be distributed equally;
>
> **LE2:** the final distribution of a common output should be proportional to the individual's contribution to production if and only if obtained by voluntary exercise of agency through the use of fairly allotted endowments as production means.

Moreover, we consider a realistic situation in which luck generates differences in the subjects' endowments – that in our study is *working time* – hence affecting their production. This takes us to the case in which differences in individual outputs do not reflect voluntary agency, but mainly arbitrary differences in endowments. In this case LE needs to be completed by a redress rule:

> **LE3**: if voluntary contributions are given by means of arbitrarily allocated endowments, the difference in agents' outputs should not translate into an identical difference in their final outcome distributions; on the contrary, the differential output obtained by use of arbitrary allocation of differential endowments should be distributed in equal parts.

A contractarian argument

[1] Note that according to LE, possession of special natural talents does not generate *per se* justifiable claims to a higher portion of resources. In this study, however, given the simple nature of the task performed by the subjects, considerations about the initial distribution of talents can be set aside.

Let us now briefly outline the social contract argument that justifies LE. Two widely held ideas of justice have intuitive force: moral equality and just desert. 'Moral equality' requires that each person deserves equal consideration and respect; hence prospective participants in a joint venture should be given equal rights over productive means. 'Just desert' instead focuses on responsible agency, and maintains that distribution should be proportional to contribution. These two broad notions fit quite naturally with two contexts that can be dubbed 'manna from heaven', where control on given resources is distributed amongst subjects that have not produced them, and 'non-manna from heaven', in which the resources to be distributed must be earned through production. In the former kind of situation, insofar as there is no relevant difference amongst individuals, equality seems to be the obvious principle of distribution.[2] In the latter situation, distribution according to relative contribution by personal agency is seen as just.[3]

The social contract theory enables us to organize these two contexts hierarchically according to the two-tier model of constitutional and post-constitutional contracts (Buchanan, 1975, Brock, 1979, Sacconi, 2006, 2011). By "constitutional contract" is meant an agreement prior to any production and contribution (viz. in a "manna context") and endorsed under a veil of ignorance. This agreement establishes the basic "rules of the game" and agents' rights on basic endowments. Essential among these are means of production and capabilities to be used in the next stage of social cooperation. Thereafter, specific cooperative ventures ("non-manna" contexts) are established. Agents endowed with an amount of rights and capabilities agree by a "post-constitutional agreement" on how to distribute the surplus deriving from cooperation in order to set up mutually acceptable cooperative ventures within the ongoing constitution. In the first stage, the egalitarian rule seems to be the only focal point for rational agreement. But in the second stage, distribution of the final payoff proportional to contribution is a rational agreement.[4]

But in the real world endowments are normally not distributed equally. This means that the constitution may have failed. Under this arbitrary condition, co-operative productive ventures yield arbitrary distributions – even if they have followed the proportionality to contribution rule. At this

[2] A distribution proportional to relative needs – because needs are independent from contributions – may also fit the "manna from heaven" context. However, the constitutional choice is taken under a veil of ignorance. This prevents knowledge of the details of different individual life plans, and hence who needs what differential means in order to pursue his/her life-plan.

[3] See Dworkin (1981a,b), Roemer (1986, 1996), Cohen (1989), Fleurbaey (2009).

[4] This can be modelled as a coalitional game in which any player's expected payoff is calculable by the Shapley value (Brock 1979), while the constitutional contract is a symmetric Nash Bargaining Game played under the veil of ignorance (see Binmore 2005).

point, individuals situated behind a veil of ignorance would recognize the violation of the constitutional principle, and deliberate a redistribution rule of final payoffs redressing the arbitrarily disadvantaged party. This completes the equivalence to LE as we understand it.

Psychological realism of normative judgments

Since the aim of this study is *not* to investigate normative issues raised by an ideal conception of justice, the simplified normative argument for LE presented above suffices for our purposes. We are instead concerned with the psychological realism of normative judgments and motivations associated with LE[5]. Hence we first inquire whether agents, in expressing their normative judgments about principles of just distribution, actually agree on LE as the rule to be used for the distribution of a common output, affected by prior random allocation of endowments. Before of any further consideration, we here only stress that this is a *positive* question concerning a *normative* fact (how agents express their normative judgments) that may be empirically answered by studying how *de facto* subjects form and express their judgment about distributive justice when they take a normative stance. To answer this question, a device for studying people's *normative* judgment is required.

Rawls suggested such a device with his idea of a "thought experiment" characterized as the "original position", a decision model "under a veil of ignorance". In the original position agents are characterized as "reasonable" in that the veil constraints them to reason impartially. But they are simply "rational in the prudential sense" and "non tuist" with respect to their generic interest in having more rather than less primary goods. Neither a metaphysical idea of the good nor a substantive adhesion to moral equality is presumed to work under the veil of ignorance, beyond the impersonal and impartial viewpoint implicit in it.

We implement an experimental simulation of the Rawlsian ideal thought experiment. In our operationalization of the original position we limited ignorance to only some variables – in particular the random allocation of means of production – which are nevertheless the aspects relevant to ensuring that agents express judgments impartially with the least inducement on the part of the experimenter.

There is a reason for questioning whether LE would realistically emerge from the *ex-ante* choice, LE is a demanding principle. Under certain circumstances it requires agents to give up part of the

[5] See Frances, Sacconi , Faillo (2015)

payoff that they could gain by simply applying what may seem to be a *prima facie* legitimate principle, i.e. remuneration according to contribution (). Agents must reason so that they regard themselves able to act against their self-interest if it turns out that they are the lucky player under the endowment lottery. Hence it requires maintaining an impartial perspective balancing this case with that of being unlucky, not giving less weight to this latter. Most importantly, they must understand the three-stage reasoning implicit in LE. Clearly, there are rules simpler than this, like straightforward egalitarianism or distribution according to contribution. Hence, LE is demanding on two counts: (i) it requires not considering it irrational to give up the results of pure luck; (ii) it involves a multistep reasoning about justice. It seems hence reasonable to question whether it is likely that ignorance about the endowment lottery will elicit LE acceptance. Our main conjecture translates into the hypothesis that impartial judgment, engendered by the veil, would indeed induce subjects to choose LE.

Realism of normative motivation.

The main question concerning the *psychological realism* of LE, however, is whether it would be *ex-post* complied with by actual behaviors. This is the motivational question. Assume that subjects in a pre-play stage have reached an agreement on LE. Will they comply with this rule when the veil has been lifted and they can choose individually the rule to be actually implemented? This question regards the motivational effect that the agreement has on compliance when agents move from the *ex-ante* agreement stage to the *ex-post* choice one. Given the new information possessed by agents in the *ex-post* decision, it is no longer true that only impartial reasons count. Self-interest may resume a strong motivational force. The question is therefore whether motivations and beliefs engendered by the *ex-ante* agreement, rather than some independent motives, can explain conformity *ex-post.*

Our conjecture is that reasoning behind the veil leads subjects to endorse an agreement on distributive principles, and this constitutes a *justified* joint commitment. Hence agents express the joint intention to implement a distributive rule thereafter; they believe that they all are ready to act upon it; and given such belief they develop the preference to acting upon it.[6] Thus, what counts in engendering preferences for conformity is (i) participation in the impartial agreement, and (ii) that the agreement elicits the mental model of an agent who - having agreed – simply intends to carry out the agreed action and by default believes that the other agreeing party will also comply.

[6] On joint commitment, see Gilbert (2014)

This hypothesis on compliance is suggested by the idea of a "sense of justice" (Rawls 1971, chapter 8). According to Rawls, once principles of justice under the veil of ignorance have been unanimously selected and this is publicly known, if there is also shared knowledge that each agent expects that the other will comply with the principles, then s/he develops an attitude of reciprocity in compliance with the principles such that incentives to defection will be overridden and the principles stabilized. Note that the "sense of justice" is an attitude which shows its effectiveness after the veil of ignorance has been lifted.[7]

In sum, our study is aimed at answering two questions about normative realism of LE: whether the agreement under a veil of ignorance univocally converges toward egalitarian principles; and whether agreement on a LE distributive rule is self-sustained by the sense of justice of subjects when they are in a context where usual incentives are at play.

3. Relation with existing experimental literature.

Here, we mainly focus on how our study relates to the experimental literature on distributive justice. Our first result, concerning the *ex-ante* tendency to opt for LE concurs with Konow's accountability principle (Konow, 1996, 2000, 2001, 2003, 2005) according to which "fair allocations are proportional to the contributions agents control (called "discretionary" variables) but do not adjust for factors they cannot influence (called "exogenous" variables)" (Konow, 2005: 378). Konow states that in a productive context, if a worker produces more than his/her colleague, a distribution assigning a higher payment to the more productive worker would be deemed fair *only* if the difference in personal outputs were due *only* to different effort. If the difference were due to variables beyond the direct control of the two workers, the same distribution would be judged as unfair (see also Cappelen et al. 2014 and Mollerstrom et al. 2015). Differently from Konow, however, we observed subjects who were not impartial spectators, but were directly affected by the allocation outcome, acting to remove the exogenous source of inequality.[8] Our study adds also an

[7] A reformulation of economic rationality consistent with the sense of justice has been suggested – also with the support of early experimental tests – by the theory of *conformist preferences* (see Grimalda and Sacconi 2005, Sacconi and Faillo 2010, Sacconi et al. 2011, Faillo et al. 2015).

[8] A distinction made by Konow is between stakeholders and spectators, and he shows that stakeholders are affected by self-serving biases in allocations while the spectators are not. We have the same result, when subjects make their decisions without being put behind a veil of ignorance. But we also find that the stakeholders themselves, once they have the opportunity to deliberate under the veil, do not show a self-serving bias. This is a main difference in modelling impartiality. We investigate the explicit agreement among "stakeholders" under the veil of ignorance, whereas Konow works with third-person impartial spectators – i.e. we work in the contractarian tradition whereas he does so from the impartial spectator perspective. This allows us to say what the stakeholders themselves would do when positioned

explicit justification – missing in Konow's definition – for why equality is fair in the absence of any discretionary factor that justifies unequal distributions.

Other existing works inspired by Rawls's principles of justice focus almost exclusively either on the relevance of the difference principle (Brickman, 1977; Yaari and Bar-Hillel, 1984; Frohlich et al. 1987; Frohlich and Oppenheimer 1990 and 1992; Bond and Park, 1991; Lissowski, Tyszka, Okrasa, 1991; Jackson and Hill; 1995; Michelbach et al. 2003; De la Cruz-Doña and Martina, 2000) or on the effects of the choice behind the veil of ignorance on stated preferences for redistribution (Anderson and Lyttkens 1999; Traub et al. 2005; Herne and Suojanen, 2004; Herne and Mard, 2008; Schildberg-Hörisch, 2010; Durante, Putterman and van der Weele, 2014). No study (except Faillo et al. 2015) has considered Rawls's concept of "sense of justice" and its role in the solution of the problem of *ex-post* compliance with a principle chosen behind the veil.

In the typical experiment, once the principle has been chosen, it is automatically implemented. Considering the subset of studies that explicitly implement the choice among alternative principles, participants usually choose individually as they are confronted with hypothetical scenarios (Schokkaert, and Lagrou, 1983; Konow, 1996, 2001, 2003; Scott et al., 2001; Fong, 2001, Favarelli, 2007). To our knowledge, only Frohlich et al. (1987), and Faillo et al. (2015) have implemented agreement among participants as a way to choose the principles. Here we make a significant additional step forward by ascertaining whether subjects *ex-post* comply – and why – with a relatively extended array of distributive justice principles, including the liberal egalitarian one.[9]

behind the veil, and moreover whether they comply *ex-post*, which cannot be done by taking the third party spectator perspective.

[9] Our results are quite remote from the experimental literature on promise keeping (Charness and Dufwenberg 2006, Vanberg 2008, Ellingsen et al. 2010). With regard to the role of obligations in explaining promise keeping, we would agree with Vanberg, who shows that arbitrary second order expectations do not have much explanatory force. However, in the case of agreements, the concern for one's and obligations necessarily involves mutual beliefs in the very understanding of them. Following Margaret Gilbert (2014), an agreement constitutes a joint commitment - i.e. a commitment undertaken by two or more agents to espouse a common goal "as a single body" - since it expresses the parties' readiness to act as if they were a single unit of action. By entering such a commitment, parties believe by default that they are espousing the goal as a single body. Should any contrary evidence arise, the joint commitment would vanish. Thus joint commitment holds only insofar as at no level of reciprocal beliefs do agents expect defection. Joint commitment entails each participant's obligation to other participants in the agreement (cf. Gilbert 2014, pp. 34-35). These are direct performance obligations, simultaneous and interdependent. Gilbert analyses many examples of promise exchanges and concludes that none of them satisfies the requirements for the existence of obligations like that engendered by agreements (Gilbert 1993, pp.634-43). In line with her analysis, no exchange of promises of the kind considered in Vanberg's article may make sense of agreements and obligations in our work. It could, however, be asked whether subjects – especially in the chat treatment (see sec 4.3) - exchanged promises in addition to making

4. Experimental design and procedures.

The experiment consisted of three treatments: Noveil, Bargaining and Chat. In all the treatments subjects were matched in pairs, and asked to perform a task. In each pair one subject was randomly given six minutes and the other ten minutes to perform the task. Indeed, the endowment that our experiment distributed unequally was time. Subjects generated an amount of money that depended on the outcome of the task. At the end of the task each subject was asked how to divide the total amount of money produced by the pair through the task. They could answer either by reporting the percentage to be assigned to each member of the pair or by choosing a division corresponding to one of five rules proposed by the experimenter and described below. Henceforth, we will call this latter decision the "*ex-post* choice" to distinguish it from the former or "*ex-ante* choice", which is material only to the Bargaining and Chat treatments. In these treatments, in fact, before performing the task and before knowing who would have six minutes and who ten minutes to perform it, the members of each pair had to agree on one of the five division rules mentioned below. We call this phase also as the "*ex-ante* agreement". For the sake of comparability with the other treatments, we will refer to the division choice as the "*ex-post* choice" also in the Noveil treatment, even if in this treatment there was no "*ex-ante* agreement" phase. Once the subjects had decided, one of the two members was randomly selected and his/her choice was implemented.

In all the treatments, the subjects knew about all the phases of the experiment from the beginning, before they made their very first choice.

Detailed descriptions of the treatments follow.[10]

4.1 Noveil treatment

In the Noveil treatment subjects were randomly matched in pairs. The treatment consisted of three phases in the following sequence: a practice phase, a task phase, and a division phase. We describe the three phases following the order used in the instructions, in which subjects first learned about the task and the division phases, and then about the practice phase.

agreements. Our answer is 'no'. We found that 38 out of 76 subjects argued in favor of the principle on which they were agreeing, assuming that it would have been implemented literally as it was agreed (as in a joint committment). Only 6 subjects expressed sentences containing explicit personal promises or assurance. 4 seemed strangely to undertake the joint commitment that, whatever the principle they agrees, they would afterwards choose the egalitarian distribution. The rest expressed just preferences without arguments, and made no promises.

[10] Instructions are included in section III of the Supplementary Online Materials.

The task

The task consisted in encoding words. In each pair, one of the subjects was given a total time of ten minutes to perform the task, while the other was given only six minutes. The assignment was random. Information about the time limits was given just before the task. A sequence of words appeared on the subjects' screens, and using a conversion table they had to convert the words into sequences of numbers. A new word appeared only after a code (either correct or mistaken) was written for the current word. The remaining time was shown through a countdown on the computer screen. The total production (i.e. the number of tokens - one token=0.15 euros) generated in the task corresponded to the number of words correctly encoded by the two subjects.

At the end of the task, the following data were provided to the subjects: the total production (total number of words correctly encoded) of the pair, individual productions of the two members of the pair, productivity (words/minute) of each member, production and productivity of the subject with the ten minutes both in the first six minutes and in the second four minutes.

The division phase and the rules.

In the division phase (or "*ex-post* choice") each member of the pair was asked to choose how to divide the total income generated by the pair in the task phase. S/he could do this either by choosing a percentage from 0 to 100%[11] of the total income to ask for him/herself or by choosing a division corresponding to the application of one of the five division rules. Subjects saw on their screens the final payoffs corresponding to the application of each of the five rules.

The rules were the following:

1) *Rule 1 – Equal split*: each subject obtains exactly half of the total product generated through the activity performed by the two subjects.

2) Rule 2 – One gets all: one subject obtains all the total product. A random draw selects the subject who gets 100% of the total product. Both subjects have a 50% probability of being selected.

3) Rule3 – One subject gets what s/he has produced: each subject obtains exactly what s/he has produced through his/her activity.

[11] The option of free percentages characterized also the *ex-post* division choice in the two treatments with the agreement (see below). The possibility to choose a free percentage put compliance with the rule agreed behind the veil of ignorance in the worst condition to be realized. In fact, free percentages ensured that no subjects complied with the agreement because of the lack of alternatives.

4) *Rule 4 – Time independent division*: each subject obtains what s/he has produced through his/her activity during the first 6 minutes; what is produced by the subject who has 10 minutes of time in the last 4 minutes is divided 50% between the two subjects.
5) *Rule 5 – Divide according to productivity*: if the ratio between the productivity (words per minute) of A and B is *x*, then A's payoff should be *x* times the payoff of B, subject to the constraint that the sum of the two payoffs is equal to the total income produced by the pair.

Subjects could read the text of the rules, and they were also shown the payoff which they would obtain if that rule was applied, given the outcome of the task.
Once both the members of the pair had made their individual decisions, by opting for a division consistent with one out of the five rules or for a percentage, one of them was randomly selected and his/her decision was implemented.

The practice phase.
Before starting the task, the subjects could practice with the rules, individually, by using a simulation platform that replicated the actual division choice screen of the third phase. They could read the five rules on their screen and choose one of them. They could also insert the number of words encoded by the person with six minutes and by the one with ten minutes both in the first six minutes and in the remaining four minutes, and they could decide the person (the one with six minutes or the one with ten minutes) whose final choice would be selected. They could play with the platform for five minutes, changing the parameters and checking the resulting outcomes. Figure 1a reports the exact sequence of the phases.

4.2 Bargaining treatment

In the Bargaining treatment, the practice phase, the task and the division phase (or "*ex-post* choice") were the same as in the Noveil treatment, but the task and the division phases were preceded by a stage in which the members of the pair, before knowing the allocation of time, could reach an *ex-ante* agreement on one of the same five rules through a bargaining procedure – the agreement did not concern the choice of a percentage from 0 to 100% of the total production (see Figure 1b for the exact sequence of the phases). The procedure consisted of a maximum of thirteen rounds. In the first six rounds, subjects simultaneously chose one of the rules, proposing it for the final division of the total product generated through the task. They could choose the rule using a choice screen similar to the final division choice screen. At the end of each round, they were informed about the rule chosen by their partner, and if they had chosen the same rule, this was an agreement. Pairs unable to reach an agreement on one of the rules (by choosing the same rule) in the first six rounds accessed a second bargaining stage of four sequential choices. Each sequential choice consisted of an offer and, if the receiver refused it, of a counter-offer. At the beginning of each of the two sequential choices, one of the two members of the pair was randomly selected to make the first offer. The other member, once s/he had received the offer, decided whether to accept or refuse it. If s/he rejected the offer, then s/he had to make a counter-offer that might be accepted or refused by the counterpart. Pairs that failed to reaching an agreement also in this second stage moved to a final sequence of three further simultaneous choices.[12] The subjects knew that the rule was not going to be enforced, but they also knew that they could proceed to the experiment's next phase (the task) only if they reached an agreement. If they failed, they would be excluded from the experiment and they would be asked to fill in a questionnaire not related with the experiment. In this case, their earning would be equal to the show up fee of 3 euros.

The agreement phase was preceded by the practice phase, which allowed subjects to become familiar with the choice interface and to the consequences, in terms of final payoffs, of their decisions.

In the ex-post choice, subjects were reminded of the rule chosen by their pair (the rule appeared also with a different background color) in the *ex-ante* agreement, and they could choose either a percentage of the total product to ask for themselves, a division of the total product corresponding to application of the agreed rule, or a division corresponding to the application of a different rule. As in the Noveil treatment, the final payoffs corresponding to the application of each of the five rules were reported on the subjects' computer screens.

4.3 Chat treatment

The Chat treatment was very similar to the Bargaining treatment. Subjects had to reach an agreement on one of the five division rules in order to access the task and the *ex-post* division phase. However, in this treatment the *ex-ante* agreement procedure was based on a chat. Subjects were given five minutes for discussion. The chat was anonymous. Communication of personal

[12] The first sequence of simultaneous proposals was introduced to capture the simultaneous nature of the bargaining. The second sequential bargaining phase was introduced to help break possible non-coordination cycles in the simultaneous choices. The last simultaneous choices phase was intended to prevent agreements reached in the sequential bargaining phase from suffering the typical hold-up problem that characterizes finite sequential bargaining, in which the second to last mover has an advantage over the last mover. Note that only two pairs failed to reach an agreement within the first sequence of simultaneous choices.

information, PC number, threats, promises of side payments and the use of offensive language were prohibited. Once the two members of the pair had reached an agreement, they had to choose the same rule on a choice screen similar to the final division choice screen. The choice of the same rule could be made at any time within the limit of 5 minutes available to discuss through the chat function. Thus, selecting the same rule on the screen after having agreed to it by the chat was a way to make clear that the agreement had been actually reached and that there was no misunderstanding about it.[13] All the pairs succeeded in choosing the same rule (it took on average 3.75 minutes). As in the Bargaining treatment, they knew that the agreement was not going to be enforced in the later stage of the game, but if they failed to reach the agreement they would be excluded from the experiment and asked to fill in a general questionnaire not related with the experiment. See Figure 1c for the exact sequence of the phases.

As in the Bargaining treatment, in the *ex-post* choice subjects were reminded of the rule chosen by their pair and they could choose separately either a free percentage of the total product to ask for themselves, or a division corresponding to the application of the agreed rule, or a division corresponding to the application of a different rule.

4.4. The difference in the endowment of time.

It is now clear that the endowment unequally distributed was time. The rule LE that we tested with our experiment established the following: *assign each member of the pair what s/he has produced in the first six minutes; then distribute what the member who had ten minutes produced in the last four minutes equally among both.* Accordingly a subject endowed with more time, to exploit this opportunity, was implicitly required to spend an additional effort for which s/he was not repaid. LE entailed thus that those who accepted it *as a matter of fact* discounted that such additional effort cost was not perfectly repaid, and admittedly this feature made acceptance of the LE rule a little awkward. However, LE fits intuitively situations like our experiment, where the difference - four minutes of quite simple work - was practically negligible; it could be difficult to measure and repay precisely it and so such a difference should not matter very much to subjects. Moreover, concern for this lack of precision could be overridden by consideration of the focal unfairness in the situation - i.e. the sharp inequality of the initial allocation of practical opportunity to work. In any case, the

[13] If a pair reached an agreement on a rule during the chat, but one of the members chose the wrong rule on the screen, a warning message about the “mistake” appeared and the subject could make another choice. Only one mistake was allowed.

imperfection in covering extra effort costs introduces some *attrition* that worked against - and made it bolder – the conjecture we aimed to corroborate by this experimental design.

4.5.Beliefs and questionnaire

In all the treatments, at the end of the ex-post choice, before a subject knew if his/her choice had been selected for payment, first- and second-order beliefs were elicited by asking what s/he believed the other member of the pair had chosen (either one of the five rules or a percentage of total product) and what s/he believed the other member believed was his/her choice. Correct guesses were rewarded with one euro. Participants were also asked to fill in a questionnaire containing both socio-demographic questions and questions about trust, risk attitude and happiness.[14] In each treatment, in two sessions the questionnaire was administered at the beginning of the experiment, before the instructions about the phases of the experiment were read; and in two sessions it was administered at the very end of the experiment, just before the payment (note that our main empirical results are virtually unchanged when we consider this distinction).

4.6 Sessions and procedures

The experiment was programmed by using z-Tree (Fischbacher, 2007) and conducted at the EGEO laboratory of the University of Granada. Subjects were paid a show-up fee of 3 euros. We adopted a between-subject design. No individual participated in more than one session.

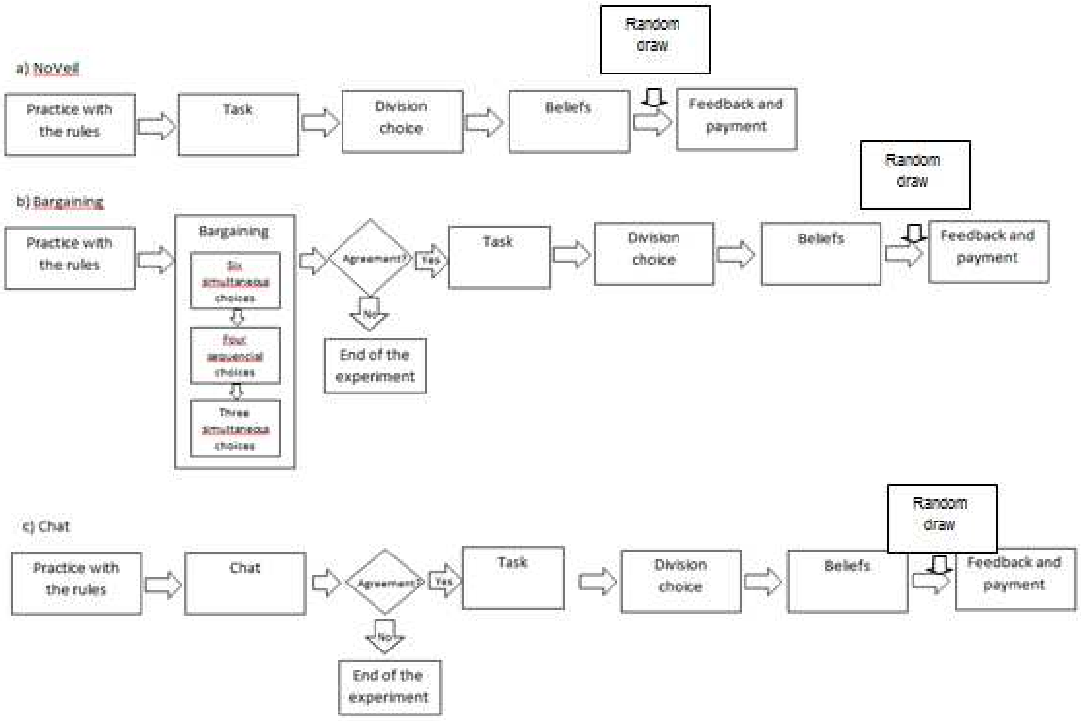
The average payment per participant was 9.80 € (including the show-up fee) and the sessions lasted approximately one hour.

In all the treatments, at the beginning of each session, participants were welcomed and asked to draw lots. They were then randomly assigned to terminals. The instructions were handed to them in written form and were read aloud by the experimenter. The participants had to answer several control questions, and we did not proceed with the actual experiment until all participants had answered all questions correctly.

A total of 236 students participated in the experiment between May 2014 and March 2015. We ran four sessions of 20 subjects each for the Noveil and the Bargaining treatments, and four sessions, three with 20 participants and one with 16 participants, for the Chat treatment.

Figure 1. A synthesis of the structures of the three treatments

[14] The questionnaire is included in the SOM.



4.Result

In what follows we report the experimental results, discussing their coherence with the theory developed in Section 1 and 2.[15]

Result 1–LE is preferred in the *ex-ante* agreement

In the ex-ante agreement characterizing the Chat and Bargaining treatment the majority of pairs agreed on Liberal egalitarian rule (Rule 4)

This is in line with our conjecture that the immediate reasoning accomplished by subjects under the veil of ignorance favors the liberal egalitarian rule, even if the cognitive task implied by it is relatively more cumbersome than those for the alternatives.

Table 1. The rule chosen across treatments (percentage values)

Rule	**Noveil**	**Bargaining**		**Chat**	
		Ex-ante agreement	**Ex-post choice**	**Ex-ante agreement**	**Ex-post choice**
Rule 1. Pure Equal Split	22.5	12.5	17.5	15.79	21.05
Rule 2.	13.75	2.5	3.75	0	7.89

[15] We consider the *ex-post* choice of opting for a percentage equal to 50% or to 100% as equivalent to the *ex-post* choice of Rule 1 or 2, respectively. In fact, in terms of the *ex-post* division of the total production, opting for Rule 1 (Rule 2) in the *ex-post* choice is equivalent to the choice of opting for the 50% (100%). Results are virtually unchanged if we do not merge subjects who opted for the previous percentages in their division choices with subjects who opted for Rule 1 or 2. When differences emerge in the econometric estimates, they are reported in the text or footnotes.

One gets all					
Rule 3. One gets what one has produced	22.5	15	22.5	5.26	6.58
Rule 4. Time independent division	16.25	57.5	40	57.89	40.79
Rule 5. Divide according to productivity	10	12.5	10	21.05	17.11
Rule 6. Percentage	15	*Option not available*	6.25	*Option not available*	6.58

Table 1 shows the percentages of subjects who chose the various rules across treatments. At a first glance, it is evident that Rule 4 was chosen by the great majority of subjects in the *ex-ante* agreement – 57.5% of subjects in the Bargaining treatment and 57.89% in the Chat treatment. The other rules chosen with greater frequency in the *ex-ante* agreement were Rule 3 and Rule 5 in the Bargaining and in the Chat treatment, respectively. A test of proportions revealed that both in the Chat and in the Bargaining treatment Rule 4 was chosen by a proportion of subjects significantly greater than 40%.[16] Conversely, the same test revealed that the other rules were agreed by proportions of subjects equal to or lower than 20%.[17]

Result 2 –*Ex-post* compliance in with ex-ante agreement

The majority of members of the pairs who have reached an agreement behind the veil of ignorance complied with it. This result holds also for subjects who ex-ante have agreed on the liberal egalitarian rule (Rule 4)[18]

[16] One-sample proportion test, p= proportion of subjects who agreed on Rule 4; H_0 =0.4: Bargaining treatment, Ha: p != 0.4, Pr(|Z| > |z|) = 0.0014; Ha: p > 0.4, Pr(Z > z) = 0.0007; Ha: p < 0.4, Pr(Z > z) = 0.9993; Chat treatment, Ha: p != 0.4, Pr(|Z| > |z|) = 0.0015; Ha: p > 0.4, Pr(Z > z) = 0.0007; Ha: p < 0.4, Pr(Z > z) = 0.9993.

[17] One-sample proportion test, p= proportion of subjects who agreed on the different rules; H_0 =0.2: Rule 1: Bargaining treatment, Ha: p != 0.2, Pr(|Z| > |z|) =0.0935; Ha: p < 0.2, Pr(Z > z) =0.0468; Ha: p > 0.2, Pr(Z > z) = 0.9532; Chat treatment, Ha: p != 0.2, Pr(|Z| > |z|) =0.3588; Ha: p < 0.2, Pr(Z > z) =0.1794; Ha: p > 0.2, Pr(Z > z) = 0.8206; Rule 2: Bargaining treatment, Ha: p != 0.2, Pr(|Z| > |z|) = 0.0001; Ha: p < 0.2, Pr(Z > z) =0.0000; Ha: p > 0.2, Pr(Z > z) = 1.000; Chat treatment: no observations; Rule 3: Bargaining treatment, Ha: p != 0.2, Pr(|Z| > |z|) =0.2636; Ha: p < 0.2, Pr(Z > z) =0.1318; Ha: p > 0.2, Pr(Z > z) = 0.8682; Chat treatment, Ha: p != 0.2, Pr(|Z| > |z|) =0.0013; Ha: p < 0.2, Pr(Z > z) =0.0007; Ha: p > 0.2, Pr(Z > z) = 0.9993; Rule 5: Bargaining treatment, Ha: p != 0.2, Pr(|Z| > |z|) =0.0935; Ha: p < 0.2, Pr(Z > z) =0.0468; Ha: p > 0.2, Pr(Z > z) = 0.9532; Chat treatment, Ha: p != 0.2, Pr(|Z| > |z|) = 0.8185; Ha: p < 0.2, Pr(Z > z) =0.5907; Ha: p > 0.2, Pr(Z > z) =0.4093.

[18]Due to space limitations, all the results of the estimations regarding the subsample of subjects who agreed, ex-ante, on rule 4 (see in particular results 3 and 4) will be made available on request.

This result is in line with the conjecture - derived from the Rawlsian idea of an effective “sense of justice”, widely discussed in Section 2, and the model of conformity preference(see note 5) - that the veil of ignorance activates an impartial viewpoint that affects beliefs, ex-post preferences and choice and is

Overall, 60% of subjects complied with the rule agreed in the *ex-ante* phase. Table 2 shows the level of compliance across treatments and rules chosen in the agreement. The highest percentage compliance is observed for subjects who agreed on Rule 1 in the Chat treatment (91.67%). The lowest percentage concerns Rule 5 in the Bargaining treatment (20%). In the Bargaining and the Chat treatment, the percentage of subjects who opted for Rule 4 in the *ex-ante* agreement and complied with the agreement is equal to 54.35% and 68.18% respectively.

Table 2. Subjects who complied with the rule chosen in the *ex-ante* agreement – percentage values (absolute values in parenthesis).

	Rule 1	**Rule 2**	**Rule 3**	**Rule 4**	**Rule 5**
Bargaining	50 (5)	50 (1)	50 (6)	54.35 (25)	20 (2)
Chat	91.67 (11)	*No observations*	50 (2)	68.18 (30)	68.75 (11)

From results 1 and 2 it follows that the combination between the *ex-ante* and the *ex-post* choices in the two treatments with agreement induces a higher frequency of choices of Rule 4 in the ex-post phase of Chat and Bargaining treatment with respect to the Noveil treatment is observed. 40.79%, 40% and 16.25% of subjects selected Rule 4 in the *ex-post* choice of the Chat, Bargaining and Noveil treatment respectively (Table 1). The percentage of subjects who chose divisions consistent with Rule 4 was significantly lower in the Noveil treatment than in the Bargaining (Pearson chi2(1), Pr=0.001) and the Chat (Pearson chi2(1), Pr=0.001) treatment; conversely, no difference emerges between the Bargaining and Chat treatments (Pearson chi2(1), Pr=0.920).

In order to check for the significance of the treatment effect on the *ex-post* division choice consistent with Rule 4, we ran Logit regressions (Table 3). The dependent variable was a binary indicator (*Rule_4_ex-post*) which took value 1 if subjects opted for a division consistent with Rule 4 in their *ex-post* choice. The independent variables of main interest were the two dummies identifying the treatment in which subjects were involved, i.e. *Chat* and *Bargaining*. Estimates included socio-demographic characteristics - i.e. age, sex, income, the propensity to take financial risk, religious orientation, the propensity to trust unknown others - controls connected with the experimental conditions - i.e. the number of words encrypted in the task, the number of words encrypted per minute - and the fact of having already taken part in Lab experiment[19] (see section I ofthe SOM for description and descriptive statistics of all variables included in the estimates)[20].

The last line of the Table reports Wald-tests useful for comparing subjects' behavior in the Chat and the Bargaining treatments.

The division consistent with Rule 4 was more likely to be chosen in the *ex-post* choice both in the Chat and the Bargaining treatment than in the Noveil treatment,[21] while no difference emerges between Chat and Bargaining (Column 1).[22] These results show that Rule 4 is chosen significantly more in the treatment characterized by the agreement than in the Noveil treatment.

Being involved in the Chat (Bargaining) treatment increases by 27.2% (26.4%) the probability of opting for the division associated with Rule 4 in the ex-post choice with respect to the Noveil treatment.

Added in column 2 of Table 3 is a dummy variable (*Rule_agr_4*) equal to 1 if subjects involved in the Chat or in the Bargaining treatment opted for Rule 4 in the *ex-ante* agreement. This variable captured the role of the agreement in affecting subjects' *ex-post* choice. This variable significantly affected the decision to select a division consistent with that rule in the *ex-post* choice. Moreover, it entirely explains the propensity to opt in the *ex-post* choice for a division consistent with Rule 4 more frequently in the Chat and in the Bargaining treatment than in the Noveil one.[23]

Table 3. Determinants of choice of the rule

	(1)	(2)
	Logit	Logit
Dependent variable:	*Rule_4_ex-post* - DV=1 if a division consistent with Rule 4 is	

[19] Two tailed Kruskal-Wallis tests run for gender (p=0.0067), age (p=0.0026) and income (p=0.0698) revealed that the three sub-samples of subjects involved in the different treatments were not perfectly balanced with respect to these variables. We replicated all the estimates reported in the following tables by controlling for these differences when significant. In particular, we included in our regressions interaction terms (when statistically significant) between the two treatment variables *Chat* and *Bargaining* and the three variables *Female*, *Age* and *Income*. We report in the footnotes the main differences emerging when interaction terms are considered.

[20] These control variables have been excluded from the Tables for reasons of space. As for all the following Tables, full estimates results are reported in section II of the SOM.

[21] When we consider interaction terms (see footnote 16), we find that: a) the level of significance disappears for Men with respect to the Bargaining treatment; b) the level of significance decreases for Women, even though it remains within the 10% level (7.4%).

[22] When the interaction terms are considered, the difference between Chat and Bargaining emerges for Men, who opt for Rule 4 more in the Chat than in the bargaining treatment.

[23] When we consider interaction terms, this result is in general confirmed. Moreover, with specific respect to the Bargaining treatment, when controlling for *Rule_agr_4* it turns out that Men choose Rule 4 less than in the Noveil. This further highlights the importance of the agreement in favouring the decision to opt for Rule 4 in the *ex-post* choice.

	selected in the ex-post choice	
	Whole sample	
Chat	1.231***	-0.344
	(0.408)	(0.545)
Bargaining	1.201***	-0.326
	(0.417)	(0.553)
Rule_agr_4		2.485***
		(0.448)
Constant	13.12*	14.80*
	(7.184)	(8.684)
Control variables	YES	YES
Observations	236	236
Pseudo R^2	0.0904	0.2262
Chat-Bargaining	0.030	-0.018
	(0.352)	(0.405)

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Result 3a–Consistency of compliance with beliefs and agreement.

In the ex-post choice of Chat and Bargaining treatment the majority of participants who complied with the agreed rule believed that: i) their counterpart chose the rule and agreed; ii) their counterpart believed that they would have done the same (alignment of beliefs and choice).This result holds also for subjects who ex-ante agreed on the liberal egalitarian rule (Rule 4)

This result is in line with the Rawlsian theory of *ex-post* stability of distributive principle of justice (see Section 2) the model of conformist preferences (note 5) and, with regard to Rule 4, to the contractarian justification of the adoption of the liberal egalitarian rule .

In the two treatments characterized by the agreement, 70.51% of subjects believed that the other player in the pair was going to comply. The percentage increases to 78.95% when we consider the Chat treatment and decreases to 62.50% in the Bargaining treatment. This difference is statistically significant (Pearson chi2(1), Pr=0.024). When we consider second-order beliefs, we find that 73.72% of subjects believed that the other player in their pair believed that they were going to comply. The percentage increases to 86.84% when the Chat treatment is considered and decreases to 61.25% in the Bargaining treatment. This difference is statistically significant (Pearson chi2(1), Pr=0.000).

Among those who complied, 77.42% believed that the counterpart would comply as well. This percentage increases when we look at the Chat treatment (87.04%) and decreases for the Bargaining treatment (64.10%), generating a statistically significant difference between the two treatments (Pearson chi^2(1), Pr=0.009). As regards second-order beliefs, 90.31% of subjects who complied believed that the counterpart believed that they were going to comply. Also in this case, the percentage is significantly larger (Fisher's exact=0.032) in the Chat (96.30%) than in the Bargaining (82.05%).

Finally, 73.12% of subjects who complied had aligned first-order and second-order beliefs: that is, they believed that the counterpart would comply and believed that the counterpart believed that they would do the same. Also in this case, the percentage is significantly larger (Pearson chi2(1), Pr=0.000) in the Chat (87.04%) than in the Bargaining (53.85%) treatment. It should be noted that, in the Chat, all subjects who complied and believed that the counterpart was going to comply, also had the second-order belief aligned with compliance. In Table 4 we analyze the relation between the decision to comply with the agreement and the reciprocal alignment of beliefs. Estimates consider only subjects involved in the Chat and Bargaining treatment. With respect to the estimates presented in Table 3, we added the payoff associated with the rule agreed in the *ex-ante* agreement (*Payment_agreement*) and the variable *Belief_aligned_compliance* that takes the value of 1 for subjects who believed that the counterpart was going to comply (first-order belief) and, at the same time, believed that the counterpart believed that they would comply (second-order belief). The significance of the latter variable (at 1% level) in the regression presented in Table 4 – column 1, in which the dependent variable is the dummy taking the value of 1 for subjects who complied with the agreement, shows a strict connection between compliance and first-order and second-order beliefs concerning compliance. Moreover, we find that the alignment of beliefs, despite the differences characterizing the Chat and the Bargaining treatments, is correlated with compliance also when we consider separately the sub-sample of subjects involved in each of these two treatments (Table 4, columns 2 and 3).[24].

Table 4. Compliance and Beliefs

	(1)	(2)	(3)
Method	Logit	Logit	Logit
Dependent variable		*Compliance*	
	All subjects	Sub-sample of	Sub-sample of

[24] When we do not merge subjects who opted for percentages equal to 50% and 100% with subjects who opted for Rule 1 or Rule 2, respectively, (see footnote 13) the *Belief_aligned_compliance* variable becomes significant at 10% in column 3.

	involved in the Chat and Bargaining treatment	subjects involved in the Chat	subjects involved in the Bargaining
Chat	0.601		
	(0.421)		
Belief_aligned_compliance	1.894***	4.449***	1.130**
	(0.431)	(1.169)	(0.558)
Payment_agreement	0.0434*	0.0438	0.0184
	(0.0244)	(0.0508)	(0.0285)
Constant	8.253	4.454	16.04
	(16.15)	(33.82)	(28.41)
Control variables	YES	YES	YES
Observations	156	76	80
Pseudo R^2	0.2260	0.4651	0.2054

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Result 3b – Ex-post beliefs are more aligned in case of agreement.

In the ex-post choice of treatment Noveil the number of participants having beliefs (of first and second order) aligned with their choice is lower than in both Bargaining and Chat treatments. This result holds also for subjects who ex-ante have agreed on the liberal egalitarian rule (Rule 4)

In discussing this result we start by showing that the alignment of beliefs with the rule actually chosen in the *ex-post* choice is more likely to be observed in the two treatments characterized by an *ex-ante* agreement than in the Noveil treatment. Second, we will show that this is due to the subjects whose first-order and second-order beliefs were aligned with the rule chosen in the ex-ante agreement. Therefore, we conclude that the higher probability of observing beliefs reciprocally aligned (and consistent with compliance) in the two treatments with the *ex-ante* agreement stems exactly from the agreement itself, which generates aligned beliefs concerning compliance with the ex-ante agreement.

As regards the first step, we note that only 17.5% of subjects had first-order and second-order beliefs aligned with the *ex-post* choice in the Noveil treatment. The percentage increases to 27.5% and to 63.16% with respect to the *ex-post* choice in the Bargaining and Chat treatment respectively. Overall, the difference in the alignment of beliefs between the treatments with the agreement and the Noveil treatment is statistically significant (Pearson chi2(1), Pr=0.000). However, note that the significance is mainly due to the subjects involved in the Chat treatment. In fact, when we compare the Bargaining and the Noveil treatment we do not find a statistically significant difference characterizing the alignment of belief (Pearson chi2(1), Pr=0.130).

Table 5, column 1, analyses the determinants of a dummy variable (*Belief_aligned_division*) capturing the alignment of belief with the *ex-post* choice (i.e. this variable assumes the value of 1 when first-order belief, second-order belief and *ex-post* division choice indicate the same rule).[25] The negative and statistically significant coefficient of the *No_veil* variable in column 1. Column 2 confirms that this result is mainly due to subjects involved in the Chat treatment. In fact, we find that beliefs of subjects involved in the Bargaining treatment are significantly less aligned than beliefs of subjects involved in the Chat, and no differences characterize the alignment of beliefs of subjects in the Bargaining and Noveil treatments (Wald test on the null hypothesis that the coefficient of *No_veil* and *Bargaining* is equal to zero: p=0.103).[26]

In order to investigate if the alignment of belief in the Chat and Bargaining treatment is due to the role of the agreement in inducing first-order and second-order beliefs concerning compliance, we include in the estimates the variable *Belief_aligned_compliance*: that is, the previously described variable which identifies subjects who believed that the counterpart was going to comply with the agreement and, at the same time, believed that the counterpart believed that they would comply.[27] Column 3 shows that, when we consider this variable, the relation between the different treatments and the alignment of beliefs completely changes. In particular, when *Belief_aligned_compliance* is included in the analysis, the coefficient of *No_veil* (column 3, Table 5) becomes positive and significant.[28]This reveals that the higher probability of observing the alignment of beliefs with the ex-post choice in the two treatments characterized by the agreement was entirely due to subjects who had beliefs aligned with compliance with the agreement. When we control for the effect of the compliance with the agreement on the alignment of belief with the *ex-post* choice through a specific dummy variable (i.e., *Belief_aligned_compliance*), it turns out that the subjects were less likely to have beliefs aligned with the rule chosen in the *ex-post* choice in the Chat and Bargaining treatments than in the Noveil. Indeed, we conclude that the difference in the alignment of beliefs

[25] Note that no subjects who chose the percentage in the *ex-post* division had first-order and second-order beliefs aligned with their choice.

[26] The difference becomes significant at 10% level when we do not merge subjects who opted for percentages equal to 50% and 100% with subjects who opted for Rule 1 or Rule 2, respectively.

[27] Note that in Table 4 *Belief_aligned_compliance* only concerns subjects involved in the two treatments with the agreement. In Table 5, this variable takes the value of zero for all subjects involved in the Noveil treatment.

[28] The significance of *Noveil* is slightly lower (6.9%) when we do not merge subjects who opted for percentages equal to 50% and 100% with subjects who opted for Rule 1 or Rule 2, respectively

which emerges between the treatments with the agreement and the Noveil is due to the *ex-ante* agreement and the beliefs coherent with compliance with the agreement itself.

Finally, when we distinguish between the Chat and Bargaining treatments (Column 4), we find that the decisive role of the agreement in generating the closer alignment of beliefs (captured by including the variable *Belief_aligned_compliance* in the regression) is confirmed for both the treatments: a) beliefs are more aligned in the Noveil treatment than in the Bargaining treatment (Wald test on the null hypothesis that the coefficient of *No_veil* and *Bargaining* is equal to zero: p=0.003)[29]; b) no differences in the alignment of belief emerge between subjects involved in the Noveil and in the Chat.

Table 5. Ex-post division choice and Beliefs

	(1)	(2)	(3)	(4)
Dependent variable		*Belief_aligned_division*		
No_veil	-1.495***	-2.100***	1.910**	1.306
	(0.370)	(0.401)	(0.800)	(0.822)
Bargaining		-1.408***		-1.289**
		(0.363)		(0.516)
Belief_aligned_compliance			4.857***	4.863***
			(0.787)	(0.812)
Constant	6.176	6.412	6.401	5.368
	(7.318)	(7.185)	(9.582)	(8.979)
Control variables	YES	YES	YES	YES
Observations	236	236	236	236
Pseudo R^2	0.0969	0.1487	0.4116	0.4328

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Result 4 Chat is more effective than Bargaining in inducing compliance.

The Chat treatment is more effective in inducing compliance than the Bargaining treatment.

A deviation from result 4 is observed in the case of Rule 4, for which, as shown in Table 2, the gap between the rates of compliance in the two treatments with agreement is higher than that observed

[29] P=0.011 if subjects who opted for percentages equal to 50% and 100% are not merged with subjects who opted for Rule 1 or Rule 2, respectively.

for rule 3 but lower than that observed for rule 1 and 5. This seems to be driven mainly by the higher rate of compliance observed in the bargaining treatment for Rule 4.

Chat is also more effective in affecting beliefs in the sense that first-order and second-order beliefs concerning reciprocal compliance with the agreement are mostly aligned in the Chat treatment, and this explains why subjects comply with the agreement more in Chat than in Bargaining

Table 6 shows the econometric analysis related to the determinants of compliance. Estimates consider only subjects involved in the Chat and Bargaining treatment. With respect to Table 4, we added to the control variables the rule chosen in the *ex-ante* agreement (*Rule_agr_1, Rule_agr_2, Rule_agr_3, Rule_agr_5*) - the residual category is represented by subjects who agreed on Rule 4. In column 2 we include two dummy variables aimed at capturing the role of first-order and second-order belief, *Belief_first=agr* (DV=1 if the subject believes that the other player in the pair is going to comply), *Belief_second=agr* (DV=1 if the subject believes that the other player believes that s/he is going to comply).

Table 6 shows that: the level of compliance is higher in the Chat than in the Bargaining treatment (column 1).[30] However, this effect is entirely explained by the role of beliefs. In fact, when the latter are included in the estimate (column 2), the difference in the level of compliance between subjects involved in the Chat and in the Bargaining treatment is no longer significant.[31] Moreover, we observe that second-order beliefs positively affect the decision to comply, while no effect emerges for first-order beliefs once both first-order and second-order beliefs are included in the regressions

Table 6. The determinants of compliance

	(1)	(2)
Method	Logit	Logit
Dependent variable	*Compliance*	
	All subjects involved in the Chat and Bargaining treatment	
Chat	0.973**	0.553
	(0.407)	(0.470)
Rule_agr_1	1.720**	2.045**
	(0.731)	(0.837)
Rule_agr_2	-2.393	-1.463
	(2.045)	(2.430)

[30] When we consider possible differences between men and women (see footnote 16), we find that this result holds only for Men.

[31] In this case, when the analysis takes specific account of differences between men and women, we find that men still comply more in the Chat than in the Bargaining treatment.

Rule_agr_3	0.280	-0.124
	(0.670)	(0.730)
Rule_agr_5	0.0659	0.0575
	(0.638)	(0.718)
Payment_agreement	0.0648*	0.0567
	(0.0354)	(0.0389)
Belief_first=agr		0.597
		(0.492)
Belief_second=agr		2.310***
		(0.530)
Constant	1.145	0.847
	(14.66)	(17.08)
Control variables	YES	YES
Observations	156	156
Pseudo R^2	0.1596	0.2930

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

6. Conclusions

Our findings provide support for what we call "psychological realism" for the justification and motivational force of LE; But they also furnish operational insights for decision makers. Starting from the latter, an implication of our results is that when policy makers take distributive decisions concerning the division of an output generated by a group, they should be aware that the liberal egalitarian decision is the one that would be approved by the majority of people when put in the condition to take an impartial decision. Moreover, they should consider that the involvement of subjects in deliberative procedures concerning the agreement on distributive rules not only leads to the adoption of this class of principles, but positively affects the motivational force of agents in complying with the agreement, even against their material self-interest.

With regard to the support for the LE principle, our results on the ex-ante choice are especially remarkable insofar as the solution on which the majority of the subjects agreed in this case required – as in most real cases – a normatively complex reasoning (that the subjects showed themselves capable of) regarding arbitrary initial endowments, individual effort, and the possibility to dictate final distributive decisions. In other words, what we are dealing with is not an obvious case of "salience" of the egalitarian distribution.

This seems to confirm the two-tier constitutional/post-constitutional approach to justice (see Section 2) entailing a descending hierarchy between the constitutional principle of the equal distribution of production endowments and the post-constitutional principle of distribution according to contribution. Even if the subjects had to choose a single principle of distributive justice that would regulate the outcome division after production had already taken place, they did not forget the requirement concerning the constitutional stage. The subjects claimed redress for the initial injustice of the endowment allocation, asking for the fruit of unequal endowments to be redistributed equally. And they agreed that only equal endowments may be used as the basis for a legitimate distribution according to contribution.

It is intuitive that the veil of ignorance, by covering any individual bias, facilitates agreement on egalitarian distributive rules. The device of the veil of ignorance in the experiment made the subjects aware of the arbitrary character of one crucial factor in determining individual production, i.e. the endowment of time; and they easily agreed to redress for that arbitrariness, thus restoring the egalitarian intuition, even if the productive nature of the experiment seemed to cue for a solution based exclusively on individual production. Hence our result supports Rawls's view that pure initial luck is not a rational basis for special claims on the total collective output.

Our second main result concerning the realism of LE is that the subjects behaved according to agreed principles – in particular, according to the LE principle. This is a crucial contribution to experimental justice since the experiment was designed, so to speak, *against* compliance: in the *ex-post* choice, the subjects decided in the role of dictator and there was no "second round" in which reputation effect could have an impact. Virtually any instrumentally and self-interested rational motivation to comply was removed. According to our hypotheses, compliance can only be explained because the subjects possessed a "sense of justice" that was activated by the agreement behind the veil. The *ex-ante* agreement was taken as a rational commitment that held *ex-post.* Subjects who believed that their counterpart would act on the agreement were motivated to do so, contrary to the predictions of rational choice theory. And most of the subjects who especially agreed after the chat procedure believed that their counterpart would comply – even if this belief was not in accordance with the hypothesis of rational self-interest. Apparently, mutual trust emerges from deliberative agreement, which somehow elicits both beliefs in the counterpart's compliance and a (non-self-interested) preference for compliance, which we identify with the attitude that Rawls called the "sense of justice".

Acknowledgement: This research was funded by a grant from the Spanish MINECO Project BENEB (code FFI2011-29005).

References

Anderson, F., and Lyttkens, C.H. (1999). “Preferences for equity in health behind a veil of ignorance”. *Health Economics,* 8(5): 369-78.

Barry, B. (1989). *Theories of justice*. Berkeley: University of California Press.

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

Bond, D., and Park, J. (1991). “An Empirical Test of Rawls’s Theory of Justice: A Second Approach, in Korea and the United States”. *Simulation Gaming*, 22(4): 443-462.

Brickman, P. (1977). “Preference for Inequality”. *Sociometry*, 40(4): 303-310.

Brock, H.W. (1979). “A Game Theoretical Account of Social Justice”. *Theory and Decision*, 11: 239-265.

Buchanan, J.M. (1975). *The Limits of Liberty, Between Anarchy and Leviathan*. Chicago: The University of Chicago Press.

Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., and Tungodden, B. (2014). “Just Luck: An Experimental Study of Risk Taking and Fairness”. *American Economic Review*, *124*(4), 1398–1413.

Charenss G, N. Dufwemberg (2006), “Promises and Partnership”, *Econometrica*, Vol 74, No.6 pp.1579-1601

Choen, G.A. (1989). “On the Currency of Egalitarian Justice”. *Ethics,* 99: 906-944.

De La Cruz-Dona, R., and Martina, A. (2000). “Diverse groups agreeing on a system of justice in distribution: Evidence from the Philippines”. *Journal of Interdisciplinary Economics*, 11: 35-76.

Durante, R., Putterman, L., and Van der Weele, J. (2014). “Preferences for Redistribution and Perception of Fairness: An Experimental Study”. *Journal of the European Economic Association,* 12(4): 1059-1086.

Dworkin, R. (1981a.) “What is Equality? Part 1: Equality of Welfare”. *Philosophy and Public Affairs,* 10(3): 185-246.

Dworkin, R. (1981b). “What is Equality? Part 2: Equality of Resources”. *Philosophy and Public Affairs,* 10(4): 283-345.

Ellingsen, T. Johannesson, M. Tjøtta, S. and Torsvik, G. (2010), "Testing guilt aversion", Games and Economic Behavior, 68(1): 95-107

Esarey, J., Salmon, T. and Barrilleaux, C. (2012) “What Motivates Political Preferences? Self-Interest, Ideology, and Fairness in a Laboratory Democracy”, *Economic Inquiry*,Vol. 50, No. 3, July 2012: 604–624.

Faillo, M., Ottone, S. and Sacconi, L. (2015). "The social contract in the laboratory. An experimental analysis of self-enforcing impartial agreements". *Public Choice*, 163 (3-4): 225-246.

Faravelli, Marco (2007) "How context matters: A survey based experiment on distributive justice". *Journal of Public Economics*, 91 7-8: 1399-1422.

Fischbacher, Urs (2007). " z-Tree: Zurich Toolbox for Ready-Made Economic Experiments". *Experimental Economics* 10): 171–78.

Fleurbaey, M. (2009), *Fairness, responsibility and welfare,* Oxford: Oxford University Press.

Fong, C. (2001). "Social preferences, self-interest, and the demand for redistribution". *Journal of Public Economics*, *82*(2), 225–246.

Francés-Gómez, P. , Sacconi, L., and Faillo, M. (2015), "Experimental economics as a method for normative business ethics", *Business Ethics: A European Review,* 24 (S1): 41-53;

Frohlich, N., and Oppenheimer, J.A. (1990), "Choosing Justice in Experimental Democracies with Production". *American Political Science Review*, 84(2): 461-477.

Frohlich, N., and Oppenheimer, J.A. (1992). *Choosing Justice: An Experimental Approach to Ethical Theory.* Berkeley: University of California Press.

Frohlich, N., Oppenheimer, J.A., and Eavey, C.L. (1987). "Choices of Principles of Distributive Justice in Experimental Groups". *American Journal of Political Science*, 31(3): 606-636.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). "Psychological Games and Sequential for Non-Cooperative Games". *International Journal of Game Theory,* 5: 61–94.

Gilbert, M. (1993) , "Is an Agreement an Exchange of Promises?", *The Journal of Philosophy*, vol. 90, No. 12, pp.627-649

Gilbert, M. (2014*), Joint Commitments*, Oxford U.P.

Gilbert, M. (2014), *Joint Commitment: How We Make the Social World*, Oxford University Press,

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

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge, MA: MIT Press.

Herne, K., and Mard, T. (2008). "Three Versions of Impartiality: An Experimental Investigation." *Homo Oeconomicus*, 25(1): 27-53.

Herne, K., and Suojanen, M. (2004). "The Role of Information in Choices Over Income Distributions". *Journal of Conflict Resolution*, 48(2): 173-193.

Jackson, M. and Hill, P. (1995). "A Fair Share", *Journal of Theoretical Politics,* 7 (2), 169-180.

Konow, J. (1996). “A Positive Theory of Economic Fairness”. *Journal of Economic Behavior and Organization*, 31(1): 13-35.

Konow, J. (2000). “Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions”. *American Economic Review*, vol. 90(4), pages 1072-1091.

Konow, J. (2001). “Fair and Square: The Four Sides of Distributive Justice”. *Journal of Economic Behavior and Organization*, 46(2): 137-164.

Konow, J. (2003). "Which Is the Fairest One of All? A Positive Analysis of Justice Theories.” *Journal of Economic Literature, American Economic Association*, vol. 41(4), pages 1188-1239.

Lissowski, G., Tyszka, T., and Okrasa, W. (1991). “Principles of Distributive Justice”. *Journal of Conflict Resolution*, 35(1): 98-119.

Michelbach, P.A., Scott, J.T., Matland, R.E., and Bornstein, B.H. (2003). “Doing Rawls Justice: An Experimental Study of Income Distribution Norms”. *American Journal of Political Science*, 47(3): 523-539.

Mollerstrom, J., Reme, B.-A., and Sørensen, E. Ø. (2015). “Luck, choice and responsibility — An experimental study of fairness views”. *Journal of Public Economics*, *131*, 33–40.

Nash, J. (1950). “The Bargaining Problem”. *Econometrica*, 18: 155-162.

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

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

Roemer, J. (1986) “The mismarriage of bargaining theory and distributive justice”, *Ethics* 97 (1):88-110 (1986)

Roemer, J. (1996). *Theories of Distributive Justice*. Cambridge Mass: Harvard University Press.

Sacconi, L. (2006). “A Social Contract Account For CSR as Extended Model of Corporate Governance (I): Rational Bargaining andJustification”. *Journal of business ethics*, vol. 68, p. 258-281.

Sacconi, L. (2011). “A Rawlsian View of CSR and the Game Theory of Its Implementation (Part II): Fairness and Equilibrium”. In L. Sacconi, M. Blair, E. Freeman and A. Vercelli (Eds.), *Corporate Social Responsibility and Corporate Governance: The Contribution of Economic Theory and Related Disciplines*. Basingstoke: Palgrave Macmillan.

Sacconi, L., and 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.

Scanlon (1998) *What we Owe to Each* Other. Cambridge (Mass.): Harvard University Press.

Schildberg-Hörisch, H. (2010). “Is the veil of ignorance only a concept about risk? An experiment.” *Journal of Public Economics*, 94(11-12): 1062-1066.

Schokkaert, E., and Lagrou, L. (1983). “An empirical approach to distributive justice”. *Journal of Public Economics*, *21*(1), 33–52.

Scott, J.T., Matland, R.E., Michelbach, P.A., and Bornstein, B.H. (2001). “Just Deserts: An Experimental Study of Distributive Justice Norms”. *American Journal of Political Science*, 45(4): 749-767

Sen, A. (2009). *The Idea of Justice*, Harvard University Press.

Traub, S., Seidl, C., Schmidt, U., Levati, M. (2005). “Friedman, Harsanyi, Rawls, Boulding – or Somebody Else? An Experimental Investigation of Distributive Justice”. *Social Choice and Welfare,* 24(2): 283-309.

Vamberg,, C. (2008), Why Do keep promises? An Experimental test of two explanations”, *Econometrica*, vol. 76, No. 6, pp.1467-1480

Yaari, M.E., and Bar-Hillel, M. (1984). “On Dividing Justly”. *Social Choice and Welfare,* 1, 1-24.

APPENDIX 1. Theoretical views that underpin the result on Consistency of compliance with beliefs and agreement

Result 3b may conflict with standard economic rationality in many senses. An ex-ante unenforced agreement would not be complied with unless it is consistent with self-interested ex-post incentives – which is not the case in this experiment. Moreover, under the assumption of symmetrical economic rationality, a subject would not expect others to conform with an agreement on rule 4, and would not believe that s/he is expected to comply with it. Thus, what justifies observations concerning 3b?

First, the intentional explanation of action (Searle 2005) suggests in our case that the content of the ex-ante agreement, i.e. dividing an outcome according to a rule, is a commitment to dividing the outcome according to the rule later on. Hence it is an intention to act. The agreement means undertaking a commitment. Having accepted (for some reason, such as impartiality, fairness, etc.) to subscribe to an agreement amounts to having a reason to act upon the corresponding commitment. A commitment is an intentional state for an action, which is not a desire-based, but a commitment-based reason to act (Searle 2005). By no means is this reason to act the only logically possible intention explaining action. Nevertheless, such an intentional state may translate into a preference to act upon the commitment and hence may be effective - among other intentions - in causing action. Only a free deliberation may pick this out of the admissible set of reasons to act, thus 'filling the gap' (Searle 2005) and letting it produce actual conduct.

Second, this intentional explanation constitutes a mental model (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991, Holoyak and Spellman, 1993, Legrenzi et. al, 1993). Subjects, having agreed ex ante, may hold this model as the basis for interpreting and predicting ex-post actions of other subjects, consistently with the evidence that they have agreed on a rule. By no means is this the only logically possible interpretation of the situation. But agents do not have enough thinking resources to consider all the logically possible state of affairs; and as a matter of fact this is the mostly immediate intentional interpretation of their behavior elicited by the content of the agreement to which they have subscribed. Moreover, it fits the situation: it makes sense of the behavior of both the self and other agents by giving an intentional interpretation of their action.

Third, the cognitive mechanism at work is *framing* (Bacharach 2006): because of the agreement, it comes to the subject's mind the *frame* of an agent who acts upon a commitment, and hence has the intention to carry out the commitment. A frame delimits the ways in which a subject may 'see' or understand a given situation. In this case the frame coming to the agent's mind is that subjects are intentional agents acting upon the undertaken commitment, and as long as the situation is framed this way, there is no room for explaining subjects as agents pursuing their self-interest in the ex-post decision.

Fourth, a frame defines the (necessarily incomplete) delimited base of knowledge whereby any default, fallible but nevertheless reasonable, prediction of the subjects' behavior must be drawn by inference (Reiter 1980, Bacharach 1994, Sacconi 2000, Sacconi and Moretti 2008). A default inference works as follows: as long as there is no evidence contrary to the assumption that subjects satisfy the model of an intentional agent acting upon commitments, nothing contradicts that, if an agent has the commitment-based intention to act according an agreed rule, s/he will in fact carry out the rule. Whence a default reasoner derives the prediction that subjects will act according to their commitment. It may be wrong, of course. But this is the simplest intentional explanation and the only one consistent with the framed mental model of an intentional agent that delimits the 'base of knowledge' held by subjects.

Summing up, moral reasoning behind the veil of ignorance leads mostly to an agreement on the liberal egalitarian rule, which is a commitment to redressing an unequal allocation of endowments ex post. Such a commitment provides a basis for a reason to act that may translate into a preference, so that the agent acts not on a desire-based intention but on a commitment-based intention that may engender a preference (a reason-to-act-based preference). At the same time the commitment-based intentional explanation constitutes a model for understanding other agents' behavior. Well-known cognitive constraints on reasoning, however, explain why this model rules out other in principle possible predictions of other agents' behavior. Thus, as long as no contradictory evidence unfolds – i.e. by default – subjects expect that because they have agreed on a rule (and hence committed themselves to carrying out the rule) they will act accordingly. But this completes the picture about the emergence of the conditional desire to comply. Since the agent expects mutual conformity, the commitment-based intention is selected as the one effectively determining his/her choice, and hence how *de facto* s/he desires to behave. This is not only consistent with the 'sense of justice' idea but also explains why so many subjects behaved consistently with this idea in our experiment.

References

Holyoak, K.J., & Spellman, B.A. (1993). "Thinking". In L. W. Porter & M.R. Rosenzweig (Eds.), *Annual review of psychology,* 44: 265-315. San Diego, CA: Academic Press.

Legrenzi P., Girotto V., & Johnson-Laird, P.N. (1993). "Focussing in Reasoning and Decision Making". *Cognition*, 49: 37-66.

Johnson-Laird, P.N., (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Harvard University Press, 1983

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Searle, J.R. (2005). *Rationality in Action.* Cambridge Mass: The MIT press.

Tversky, A. & Kahneman, D. (1981). "The framing of decisions and the psychology of choice". *Science,* 211.4481: 453-458.

Appendix 2. On Agreement Compliance vs. Promise Keeping

One could wonder whether there is any relation between our results and the experimental literature on promise keeping. The relation with our work is quite remote, since our concern is *not* unilateral promises but compliance with impartial agreements about principles accepted behind a veil of ignorance.

The literature on promise keeping is centered on two alternative explanations: a direct concern for promise keeping, or the motivational force of second order expectations (see Charness and Dufwenberg 2006, Vanberg 2008, Ellingsen et al. 2010). In so far as this literature regards the direct role of commitments and obligations in explaining behavior with respect to the role of relatively arbitrary second order beliefs (my prediction of others parties' expectations on me), we position ourselves on Vanberg's side of the debate who claims that promise keeping follows from a normative reason to act – i.e. the promisor's obligation; whereas the prediction of others' beliefs – i.e. my prediction of their description of the probable behavior that I am going to adopt – does not give us a reason to act. In the end – if we well understand the meaning of Vanberg's results – he restates the naturalistic fallacy: 'an *ought* does not follow from an *is*'.

However, in case of agreements, the concern for one's commitments and obligations necessarily involves *mutual* beliefs of first, second, and maybe higher order: in order to understand why people respect agreements, we must refer to reciprocal expectations and knowledge, since agreements are not unilateral but *joint* commitments involving reciprocal beliefs in their condition of validity.

In fact, following Margaret Gilbert (2014), an agreement constitutes a *joint commitment* i.e. a commitment undertaken by two or more agents to espouse a common goal "as a single body", since an agreement expresses the parties' readiness to act as if they were a single unit of action (cf. Gilbert 2014, pp. 31-33). This is very similar to the reasoning that players adopt in cooperative bargaining games, where agents agree on a joint strategy on the presumption that, whatever the agreement they reach, it will be implemented by a unit of agency (namely a coalition) having the joint strategy as its goal. Hence in the case of joint commitment agents reason 'as if' they were a part of a coalition in a cooperative game.

This does not mean that thereafter they will not face a situation in which they may separately decide whether or not to comply. It simply means that the very fact of genuinely agreeing entails that they express their readiness (intention) to commit themselves to implementing a common plan of action or a goal (or a principle) 'as if' they were a single unit of agency - or 'as if' they were part of a coalition that has an autonomous capacity to act as a single unit of agency. In terms of game theory, this means that they assume that commitments are *binding*, not because of an external enforcing mechanism but for some internal reason whereby commitments bind their later behavior.

What counts for us here, however, is the *validity condition* for joint commitments: by entering a joint commitment, parties believe by default (until proof to contrary) that they are espousing the goal as a single body. This refers to a state of belief engendered by the mutual explicit expression, through agreement, of their readiness to enter the commitment to espouse the joint plan as a single body. This state of belief, entered by default, ceases if contrary evidence unfolds. So should any contrary evidence arise, the joint commitment would vanish – i.e. should any party receive the information that his/her fellow partner in the agreement is not behaving "as a single body" according to the agreement, s/he would be completely freed from the commitment (cf. Gilbert 2014, pp. 41-42).

The default clause is obviously defined also for second order beliefs: until I believe that others believe that I am doing my part in the joint plan, and thus do not feel freed from the commitment and stop doing their parts, *ceteris paribus* I have no reason to free myself and not to do my part. Thus, a constitutive requirement for saying that a joint commitment holds is that at no layer of reciprocal beliefs do parties expect defection.

A commitment gives reasons to act. In the case of individual commitment, an agent who changes his/her mind is required to account to him/herself that his reasons are overridden by some further reasons to act otherwise. Moreover, a joint commitment gives participants a shared reason to act because it was they who agreed and entered the joint commitment. Hence, were any of them to decide to withdraw from the commitment, s/he would be rationally accountable to all about why that shared reason should no longer hold. As long as there is no reason to change their mind, and no reason to believe that the joint commitment does not hold anymore, *ceteris paribus* the commitment continues to motivate behavior.

From this it follows that a joint commitment engenders obligations. An individual decision commits me to doing something in a later moment; hence, because of an internal consistency requirement, it entails the *duty* that I account to myself for any deviation from the plan. Similarly, a joint commitment entails the expression of the intention or readiness to endorse a joint plan of action as a single body. This is a reason to act for the participants, with respect to which they are accountable to each other. Thus a joint commitment entails each participant's *obligation* to respond to other participants in the agreement for the performance of (or withdrawal from) the act entailed by the commitment to espouse a joint plan of action as if? they were a single body (cf. Gilbert 2014, pp. 34-35).

To Gilbert's analysis we add – following Rawls – that an impartial agreement gives to these reasons to act the additional support of an impartial and impersonal justification derived from the deliberation behind the veil. Thus we have *justified joint commitment* and the resulting obligations are *obligations of justice*. The "sense of justice" is a psychological motivation expressed by a positive attitude (and preferences manifested through an attitude) that naturally relates to these obligations: the "sense" of obligation to act justly (i.e. according to impartially agreed principles). Also the sense of justice, however, entails an essential reference to a state of shared knowledge concerning ongoing reciprocal conformity to principles.

We may now argue, again following Gilbert (1993), that obligations giving concern for promise keeping are not the same as obligations deriving from agreements through joint commitments. *First*, obligations deriving from agreements are *direct performance obligations* - i.e. an agreement is a sufficient condition for activating an actual obligation to perform an act (Gilbert, 1993, p.630). To be sure, instructions that appear as the internal content of an agreement can be contingent clauses - conditional on the occurrence of certain events (like contingent contracts). But the (external) obligation to follow such a (maybe contingent) plan is directly undertaken by a valid agreement. The performance obligation itself cannot be conditional on some further behavior of the other party that the agent is waiting to be completed? before his/her performance obligation becomes actual. In this case s/he would be not actually committed.

Secondly, these obligations are *simultaneous* - no one may be obliged by an agreement *before* anyone else? - and *thirdly*, they are *interdependent* - one cannot be obliged unless one believes ‘until proof to the contrary’ that one’s fellow partner in the agreement is obliged as well (Gilbert 1993, p. 631-32).

In fact, the obligation to carry out a joint plan “as a single body” holds by default as long as no evidence unfolds that other parties to the agreement do not intend to carry out their part: in such a case, each participant’s commitment would be nullified as the violation of the default clause makes the expression “as a single body” void since by him/herself s/he could not be able to satisfy it. Should this evidence unfold, the joint commitment would cease to exist and the obligation would expire.

Agreements cannot be reduced to exchanges of promises if these cannot satisfy the three conditions stated above. Take for example a pair of *unconditional* promises that might be exchanged by agents A and B:

A: “I promise to do my part X in the joint plan (X, Y)”,

B: “I promise to do my part Y in the joint plan (X,Y)”

They do not satisfy both the conditions of *simultaneity* – by promising, A enters an obligation prior to B – and *interdependence* – if B does not go ahead with his/her promise, A continues to be obliged by his/her promise, even though s/he knows about B’s failure in accomplishing his/her task. Consider then an exchange of *conditional* promises:

A: On condition that B unconditionally promises to do Y, I promise: “I will do X in the joint plan (X,Y)”

B: On condition that A unconditionally promises to do X , I promise: “I will do Y in the joint plan (X,Y)”

Clearly this pair of promises does not constitute any *performance* obligation for the agents, since neither A nor B makes the required unconditional promise (both externally condition their promise to that of the counterparty). Hence, neither of them enters any actual obligation. Finally consider the following pair of *conditional* promises:

> A: On condition that B makes the same conditional promise (replacing Y for X) that I'm doing, I promise "I do action X if B does action Y"
>
> B: On condition that A makes the same conditional promise (replacing X for Y) that I'm doing I promise "I do action Y if A does action X"

Here neither party is obliged before the other one is. They make symmetrical conditional promises, so that they are simultaneously (conditionally) obliged. However, the symmetrical conditionality of the promises entails that these obligations never activate actual behaviors, since each has to wait for the other to perform his/her action before s/he does his/her own. Symmetry of the conditionals means that the content of promises never starts to be enacted, since neither party? has any reason to act before the other agent's action has been carried out. Hence it is doubtful that they are under an (actual) *performance* obligation to do X and Y respectively. Such obligations have no pragmatic effectivity.

Moreover, assume that B makes the promise but then fails to accomplish his/her task Y. In an agreement, such evidence of B's withdrawal from the joint commitment to do (X,Y) would violate the "until proof to the contrary" condition for the validity of A's obligation. In this promise exchange, however, A enters his/her promise since B satisfies the conditional premise, but then there is no default condition that make A's obligation *interdependent* with B's actual compliance. A's promise to do X is internally conditional ("if B does Y") and it is not contradicted by the knowledge that B has *still not* accomplished his/her part. The conditional obligation continues to exist along with A's waiting for B eventually to do his/her part Y. Therefore A is not freed from his/her obligation, even though such conditional obligation does not require him/her to become active before B does his/her part. Hence, neither A nor B does anything, but they remain in a state of conditional obligation waiting for the other party's decision to satisfy the condition. This situation violates the interdependence condition for agreement obligations, which requires that a party continues to be obliged only as long as s/he does not learn that other participants are not accomplishing their part in the agreement. So, the situation seems somewhat the opposite with respect to what would happen under an agreement: (i) the parties simultaneously enter the joint commitment to act on the joint plan 'as a single body', and hence become active in its implementation until proof to the contrary about other parties' compliance. But (ii) were it the case that A received information about B's lack of compliance, s/he would be completely freed from any obligation.

Gilbert analyses these and other examples of promise exchanges and concludes that none of them is able to satisfy the requirements for the existence of obligations like that engendered by agreements (Gilbert 1993, pp.634-43). We subscribe to her analysis, and simply add that as far as it is accepted, no exchange of promises of the kind considered in Vanberg's paper may make sense for agreements and obligations in our work.

A last question can nevertheless be asked. It may be possible that subjects in our experiment had exchanged promises *in addition* to making agreements that entailed joint commitments. If they had done so, we should disentangle the effects of joint commitments from that of promises. This is particularly relevant to the chat treatment, where subjects could use any kind of arguments in order to convince each other before reaching an agreement. So did they also exchange promises? We controlled for this aspect by conducting an analysis of the chats content. It was carried out by one of us who was perfectly able to appreciate the linguistic nuances of meaning in the expressions used by subjects during their chats (in Spanish). Our conclusion is that, when arguing behind the veil, subjects tried to convince each other to agree on a principle on the basis of its consequences or properties, understood as a plan that would be *carried out as it was agreed* - as would happen if they were actually jointly committed to doing so. The reference to joint commitments is implicit in the very fact of making the agreement. Hence they did not bother about other consequences different from that of a literal implementation of the agreement. But they *did not* exchange individual promises about compliance.

Phrases used during the chats in order to argue in favor or against the choice of any rule were classified in two broad categories. The first included any expression aimed at illustrating the properties of the principle if properly implemented, with reference to the consequences for both of the parties ("we"), or for any of them ("me", "you"), under any kind of argumentation, including arguments based on personal or collective benefit, convenience, but also in terms of fairness of the principle itself ("it is more just"), or fairness of the distribution that would result for the parties in case of implementation of that particular rule. What is characteristic of this broad category is that subjects argued about the properties of the principle under the assumption that if a rule were agreed, it would be directly (automatically) implemented. Hence they could consider its direct results to the parties: "if we agree on this rule, what will happen is that the distribution will be…" followed by an estimation of the distribution or a qualification of it. Generally speaking, these subject assessed principles under the assumption that agreeing on one of them would mean that the course of action implementing the principle would be carried out by some automatism, which is identical to the assumption that by agreeing they would undertake the joint commitment to implement the principle "as a single body" or a unit of action (i.e. without considering the opportunity that they may make an individual, separate choice at the moment when actual implementation would be at stake).

A second category used to classify the subjects' phrases included expressions of explicit promise or assurance, such as "believe me…", "trust me…", "you have my word…." , "if we make this agreement don't be afraid that…." etc. This category contained any expression of explicit assurance or commitment, and could also contain cases where commitments were undertaken in the name of "we" - and hence not so much as a *personal promise*. Hence, we are ready to discard some potential cases of joint commitment, in order not to reduce the amount of promising artificially. The underlying idea, however, is that joint commitments are largely implicit in arguing in favor of an agreement by assessing its content 'as if' its implementation were the direct effect of the same act of agreeing. Or, to be more precise, as if the implementation were the direct effect of a commitment that the parties implicitly undertake through the agreement by expressing their readiness to put a course of action into practice - as they were "a single body" - having the agreement content as its result. So we disentangle these cases of joint commitments from any case of explicit promise or assurance (that certainly includes all the cases of personal promising or exchange promises) .

76 subjects participated in the chat treatment). Of these, 28 uttered phrases that were impossible to classify simply because they expressed a preference without any argumentation (for example, "let's choose rule 4" or "I like rule 3 more"). A majority of 38 subjects talked about the principle on which they were agreeing, taking for granted that it would have been implemented literally as it was agreed (without second thoughts). This is compatible with the interpretation that they argued about the agreement on a principle under the presumption that, if they agreed, they were undertaking a joint commitment to act according to the principle "as a single body". Only 6 expressed sentences containing promises or assurance, while 4 seemed to undertake a joint commitment to afterwards choosing the egalitarian distribution once the first step of agreeing on a principle had been accomplished. We conclude that subjects involved in impartial agreements do not exchange promises but undertake joint commitments.

References

G. Charenss and N. Dufwemberg (2006), “Promises and Partnership”, *Econometrica*, Vol 74, No.6 pp.1579-1601

M. Gilbert (2014*), Joint Commitments*, Oxford U.P.

M. Gilbert (1993) , “Is an Agreement an Exchange of Promises?”, *The Journal of Philosophy*, vol. 90, No. 12, pp.627-649

C. Vamberg (2008), Why Do keep promises? An Experimental test of two explanations”, *Econometrica*, vol. 76, No. 6, pp.1467-1480

Ellingsen, T. Johannesson, M. Tjøtta, S. and   Torsvik, G. (2010), "Testing guilt aversion", Games and Economic Behavior, 68(1): 95-107