Liberal-egalitarianism as a *fair* joint commitment:

Insights from normative agreement and compliance in an experimental setting

PhD Student

Laura Marcon

Directors

Professor Pedro Francés-Gómez

Professor Marco Faillo

Doctoral Program in Philosophy

Department of Philosophy I

University of Granada

In co-tutelle with:

Doctoral Program in Economics and Management

Doctoral School of Social Sciences

University of Trento

Editor: Universidad de Granada. Tesis Doctorales
Autor: Laura Marcon
ISBN: 978-84-1306-583-0
URI: http://hdl.handle.net/10481/63453

τῶι μὲν θεῶι καλὰ πάντα καὶ ἀγαθὰ καὶ δίκαια,
ἄνθρωποι δὲ ἃ μὲν ἄδικαὑπειλήφασιν ἃ δὲ δίκαια

(Eraclito – fr.102 Diels-Kranz)

Table of Contents:

Acknowledgments

This thesis was financed by the FFI2017-87953-R project, BENEB3: A new social contract. Business Ethics: Normativity and Economic Behavior III, which allowed me a whole year to concentrate and devote myself to the final drafting of this dissertation, without the burden of having to carry out other jobs.

Its development took place both at the Department of Philosophy I, at the University of Granada (Spain) and at the Department of Economics and Management, at the University of Trento (Italy), thanks to a co-tutelle agreement between these two institutions.

A special thanks goes to my two thesis directors: to Professor Pedro Francés-Gómez, for his support, for his in-depth knowledge of the social contract theory, and for his philosophical insights and suggestions; to professor Marco Faillo, for his huge help in the statistical analysis of data and for having taught me how to design an experiment using zTree program.

I am grateful to Professor Cristina Bicchieri, to her book *The Grammar of Society: The Nature and Dynamics of Social Norms thanks* to which my interest for social norms arose. This is a double thank-you note because I had the opportunity to spend my stay abroad at her Department, the PPE (Philosophy, Politics, and Economics program), at the University of Pennsylvania, Philadelphia, where I met new colleagues and friends.

To Professor Lorenzo Sacconi for the great inspiration he gave me with his philosophical thoughts.

A last thanks goes to my family and to several friends and colleagues with which I have discussed some of the topics it addresses – Elisa, Sara, Klaudijo, Luis, Federico, Francesco, Giuseppe e Gabriele.

Summary

This inquiry is devoted to the study of practical reasoning, in particular, in its declination of moral reasoning. This implies considering humans as capable of choice and free will: moral action is not the result of deterministic laws that regulate nature, but the result of reasoning, deliberation, and intention to act in accordance with what previously decided. What reasons do I have for preferring one moral action to another? What does it mean to act morally? The spectrum of questions of this nature can be greatly extended, so it is necessary to circumscribe the field of investigation to which this research is directed. This research is part of normative ethics, which studies the formation of moral judgments and whether such judgments are able to motivate people to act in accordance with them. The gap between the dimension of the common good and the private life of each citizen has led to questioning whether there are moral norms whose content may constitute, *per se*, a sufficient reason for action. Studying behavioural dynamics in the laboratory has revealed how other-regarding preferences and concerns for fairness influence people's decision-making. Well-established social norms are another feature that leads to conformist behaviour within some reference group: that is, one specific norm would require a certain type of behaviour, triggering a series of reciprocal empirical and normative expectations that motivate people to comply with what the norm claims. There remains the issue of understanding under what conditions a norm becomes binding enough to cause compliance. More specifically, when a norm, collectively chosen and shared, succeeds in self-imposing without any intervention of external authority.

Social norms, distinct from legal codes, include a series of collective behaviours ranging from conventions to fashions, up to those phenomena (precisely social norms) that need individuals' empirical and normative expectations to be established and collectively abided by. A particular group of social norms are the norms of fairness: in this work, I will refer to them also as norms of justice because it is assumed that justice is explained as fairness – as presented by John Rawls.

This means that the main element to define some principles of justice that can have a practical value is their being fair.

The interest for this theme arises from the urgency of defining conditions under which a group of interdependent agents, can organize themselves to obtain long-term collective benefits, find in collective action, a motivation, a commitment and a responsibility that decrease the temptation of opportunistic behaviour in distribution contexts.

The main goal of this inquiry is trying to propose a normative solution of a problem of distributive justice in the following terms of: how can a norm generate a motivational causal force that induces compliance with what it asserts, in contexts where selfish rather than prosocial behaviour would be expected? This problem, approached from different perspectives, economic, psychological, sociological, also requires a philosophical reflection. Within this framework, the issue of compliance might be reread as a motivational problem: thus, the aim of this thesis would be to try to clarify the relationship between an impartial ethical point of view and what kind of real motivations people have to act in accordance with some specific ethical principles – namely, what reasons people have for acting in alignment with principles of distributive justice. It is worth noting that this purpose fits into a metaethical framework, in particular challenging the problem of the so-called moral point of view. Ethical theories have traditionally been divided into advocates of the impersonal, neutral or third-person point of view or in those holding relative-agent, first-person perspective. What point of view to adopt assumes fundamental importance when we have to deal with distributive justice: what can be deemed as fair for one, can be unfair for another, all things being equal. This is why Frohlich and Oppenheimer argue that “the key to understanding distributive justice is impartial reasoning” (1992, p.3): the ethical point of view, impartial by its very definition, would be the only one capable of exceeding interests and personal views in the name of what should be done. Through both speculative and experimental approach,

this research tries to analyse a circumscribed problem of distributive justice by proposing that, given the context presented in the experiments, a distributive criterion that appeals to liberal egalitarianism would work on both an ethical and psychological level only if it is collectively decided as the norm to be followed. That is, the impartial reasoning should be consistent with psychological individual motivation to act according to what moral principles prescribe.

Before explaining what is meant by 'collectively decided as the norm to be followed', it must be clarify what the expression 'liberal egalitarianism' refers to. Three aspects come up: 1) as ideal of fairness, morally understood; 2) as an expression used by Degli Antoni et al (2016) to indicate a specific distributive criterion; 3) a normative solution to a normative problem of distributive justice – where point 2 and 3 are closely linked to each other.

Liberal egalitarianism as ideal of fairness is considered as a criterion for solving redistributive problems. In this research, which follows Degli Antoni and colleagues' work, it is also used as a principle of (distributive) justice, with a direct reference to John Rawls's theory. Rawlsian is also the result of a constructivist procedure that pinpoints the parties actively engaged in deciding which norm to be followed. As will be seen in the fourth chapter, liberal egalitarianism – an ideal insofar as it provides a solution of how a total product should be (equally) divided – differs from other moral intuitions on how to deem fairness. Experimentally, individuals find themselves having to decide on a distributive criterion, each of which summarizes a moral intuition: the selfishness of *homo oeconomicus* (one takes the whole cake and leaves the other without anything); pure egalitarianism (everyone gets half the cake); egalitarianism based on merit (everyone gets the corresponding slice of her/his product, on the basis of distinct variables such as the time available or the productivity calculated in relation to what was produced given time available). Liberal egalitarianism, therefore, is included as another distributive principle that has, as its main purpose, that of redressing the unfair and unjustified differences caused by mere luck, regardless of personal merit. Read in these terms, this principle is very close to the Rawlsian

difference principle, according to which the only type of socio-economic inequalities of a well-ordered society should benefit the worst-off.

So, the most specific objective is to determine how a liberal egalitarian principle is perceived and followed as a standard of conduct that guides action under the necessary condition that it is the result of a unanimous agreement under the veil of ignorance.

The focus of this research is both philosophical and methodological. Starting from the philosophical side, it tries to face Kantian constructivism as metaethics with normative implications. There are two main metaethical aspects: the moral point of view and the realism of moral principles/commands. The first asks what is the moral point of view to be taken so that an utterance can count as moral, while the second asks how far a moral principle – which satisfies the conditions of universalizability, prescriptivism and overridingness – can constitute a real source of moral motivation. In other words, should the moral perspective, to be such, concern only the impersonal and impartial point of view (the eye of the external observer, the third-person perspective) or, instead, it should also take into account the first-person perspective? This question is intertwined with the other issue on the realism of moral arguments. It should be emphasized that, by using the term 'realism', I mean the influence that a moral utterance may or may not arouse for the purposes of action. To what extent can – or can not – a moral command influence someone's action? 'Realism', therefore, refers to how much a moral principle or argument can be put into practice, to what extent, on its own, it can create a motivational source in accordance with the principle. The two aspects are tied: Rawls's Kantian constructivism might constitute a method that allows to integrate the impartial point of view with relative-agents perspectives. In the case of justice as fairness, the principles of justice will be fair as the product of a method that justifies them within their construction procedure, keeping impartial perspective and individual

psychologies together, hence by creating a motivational force such that those principles of justice are not only considered right in an abstract and ideal way, but also in practice.

Secondly, this work would like to highlight a problem concerning fairness norms: are they moral or social?

A well-known model of social norms will be later exposed (see chapter II, section 3): the one formulated by Bicchieri (2006). According to it, norms of fairness fall into the category of social norms, because what is fair would depend on the context and on the kind of interdependent behaviours individuals sustain to comply with. For instance, I can conform because I believe that what the norm asserts should be done, and/or I conform because following the norm promotes my personal interests. I can also conform because I expect others to conform: these kinds of expectations presuppose a set of conditional preferences on others' behaviour. Roughly, following a fairness norm would be the result of the emergence of conditional preferences: once such a norm is collectively recognized within a group, individuals would have conditional preferences to abide by it (conditional to others with whom we interact). However, what can be considered fair by an individual, can be completely unfair for another one. So how could one determine what is fair or what should be done given the circumstances of the environment in which we find ourselves? Conditional preferences of conforming to what others do, do not guarantee the *rightness* of a norm. For instance, you should pay for the tram ticket in Rome, but you see the relevant others do not pay, so you prefer to conform to the tacit non-payment norm instead of going against it (in fact, you should be seen as the black sheep!). Conditions for following either a new or a well-established, even if bad, social norm are met. Nonetheless, there is a sense, there is a common understanding, for which this is not what people should do, because it is not fair and it is likely to damage the common good – for example, it lowers the quality and the maintenance of public services. Moral norms would come into play, understood as commands, as precepts, which direct our action *super partes*: as universalizable, moral norms seem to be a benchmark for evaluating a social norm as fair – or not. Universalizability would make it possible by abstracting from specific

situations to find the widest application. This characteristic would act as a guide, as a heuristic, for individual and collective behaviours in dilemmatic situations. Moreover, it may constitute that parameter through which a critical reflective equilibrium on social norms can be brought about: without a moral norm, universalizable and potentially sharable by all, people would not be able to modify unjust social practices or norms (however slow and difficult the change process is). It would follow that norms of fairness cannot be exclusively delimited within the category of social norms. How would a wrong social norm change if there were no moral principles/arguments that define what one *ought* to do?

To narrow the difference between social and moral norms in the type of behaviour (interdependent vs independent) and in the different kind of individual preferences that derive from it (conditional vs unconditional) could risk emptying moral norms from their original function: to guide human *ethos*. This is a Greek term, from which precisely ethics, that manages to give us back a complexity of meanings around the value of moral arguments: it means at the same time costume, habit, but also character, personality. *Ethos* as a stay of being, such as the historical contingency in which we are thrown and in which we find ourselves relating. This would imply the importance of moral arguments, by helping people to exercise critical reflection and to be actively capable of changing those social norms that make institutions unfair. Without this sort of heuristic benchmark, that allows us to judge whether or not a social norm is fair, changing a bad social norm would seem to be stuck in an impasse. When it comes to justice, there should be a unanimous agreement on what we mean by *fair* and, in doing so, we can also leave room for criticisms and revisions of a practice or an institution that, over time, proves to be unjust. So, if we think that people are responsible for their actions and for the society where they want to live in, moral arguments about what we judge fair to the point of complying with, should be considered as motivational factors. In order to pursue this view, the conjecture here is that Kantian constructivism could be considered as a metaethical methodology that leads people to the same conception of justice in a specific means' production context. Kantian constructivism, indeed,

would be the most suitable candidate to demonstrate how a moral argument can be realistic – where, by 'realistic', I intend a moral argument that can be put into practice. Given the experimental context (explained mostly in the fourth chapter), the second thesis' main conjecture is that liberal egalitarianism is that moral argument capable of eliciting motivational support sufficiently strong to implement prescribed action (i.e., a moral principle) into practical behaviour. Understanding to what extent the liberal egalitarian principle, as a moral argument to solve a distributive justice's problem, is realistic depends on how far its rational justification and deriving motivations are tied together. Kantian constructivism would be a method such as to justify principle of justice, by giving, at the same time, motivations to conform to them: the procedure justifies principles' content (on which individuals agree) and, in turn, the justification gives reason to put them into practice. Furthermore, the fact of collectively agreeing on those principle has a psychological force: it motivates the agents because it is the result of a joint commitment – to be understood in Gilbert's terms. I say 'psychological force' because reaching an agreement on which principles of justice people would like to adopt for institutions of their society does not imply, logically, any causal link with real people's behaviour.

This gap is what John Rawls presented as the stability problem in *A Theory of Justice* (1971, 1999). According to Rawls, the element that allows one to behave in accordance with what is agreed upon is a sense of justice. This is considered a psychological disposition that intervenes in the process of deliberation and gives reasons for acting fairly. However, in the absence of external authority that sanctions those who do not respect the covenant, verifying whether the agreement could be self-enforcing is still a challenge.

Hence, the main research question is: under what conditions could an agreement reached behind the veil of ignorance give real motivation to act? Can that procedure normatively bind agents' willingness?

The attempt to answer this question requires a series of more specific sub-questions:

1. Could the Rawlsian veil of ignorance be a moral device able to bind agents' willingness towards fair behaviour?

2. If (1) above is verified, could the veil of ignorance affect individuals even without an agreement amongst parties?

3. Could the impartial agreement provide reasons to choose a liberal egalitarian distributive criterion, ensuring compliance?

4. Could this special kind of agreement, intended as Gilbert's joint commitment, be a procedure that holds together ethical impartiality and agent-relative perspective?

Thus, the dissertation provides evidence from laboratory experiments that supports John Rawls's Kantian constructivism as a method that allows subjects to evaluate different distribution criteria in production situations, making them reach an agreement on the liberal egalitarian principle. The Rawlsian method would rationally justify the ex-ante collective choice on that principle and it gives real motivations to comply ex-post, as it provides the conditions for creating a rationally justified joint commitment.

Two things will be proved:

- In line with the literature on norms of justice, the conception of fairness that individuals internally have does not seem to coincide with a principle of liberal egalitarianism. This conception changes when it is the result of a collective deliberation procedure, which presupposes a plural subject with a collective intentionality that cannot be reduced to the sum of many individuals' intentionality.
- In order for a principle, directly derived from liberal egalitarianism, to constitute a social norm to which people conform, the elicitation of mutual expectations remains a determining factor for the effective realization of compliance. However, an individual

commitment accordingly to one own's previous decision gives a weaker reason to abide by the chose norm, compared to a joint commitment.

The three main chapters that constitute the thesis fit into this framework. Chapters III and IV report experimental evidence collected in lab experiments at the University of Granada and at the University of Trento, within the research project BENEB (Business Ethics: Normativity and Economic Behaviour), funded by the Spanish Ministry of Economy (Directorate General of Scientific and Technical Research; National Plan of I+D+I Project FFI2011-29005).

The first one *Does impartial reasoning matter in economic decisions? An experimental result about distributive (un)fairness in a production context* presents a first attempt to test the veil of ignorance as a moral device to elicit fairer behaviour in a distributive context. The experimental designs presented here and in the second article try to recreate what Rawls called the original position. The basic idea is to put the subjects in a position from which they do not know who will be the most or the least advantaged and, under this condition of ignorance, they must deliberate about which distribution should be chosen.

In line with Rawls, talking about just society means thinking about a society where inequalities, if any, have to benefit the worst off. One aspect both articles have in common, in their experimental section, is the fact that subjects should think about how they would like to share a certain endowment gained through an effort/labour, knowing that everyone has a 50% chance of being more or less advantaged during the activity. In the first article, against our expectations, the subjects were not influenced by reflection behind the veil of ignorance, on the contrary they behaved in a more selfish manner, compared to the treatment without the veil. This result has challenged the function of the veil of ignorance as an individual thought experiment.

In other words, the veil of ignorance would work as moral device because it intervenes at an individual level but within a community. Unanimous consent is required to reach an impartial

agreement on what principles of justice should be applied in a society. Therefore, considering the Rawlsian construction of a theory of justice, the results of the first article seem to invalidate the moral learning effect of the veil of ignorance.

However, moving forward, *Distributive justice in the lab: testing the binding role of the agreement*, the problem remains to understand why the veil of ignorance, individually, has not had any effect towards a more just behaviour - where the notion of fairness is treated in Rawlsian terms. Unlike the previous chapter, in this one the subjects are asked to reach an agreement: the results show that the agreement procedure not only modifies the conception of what is considered fair on the distribution of a surplus, but affects the level of compliance, which increases compared to an individual decision. A distributive criterion like liberal egalitarianism is the one that is preferred in the presence of the agreement, while it does not appear to be the most chosen when the decision is individual. What can be inferred from these results is that the procedure, thus reaching a unanimous agreement on a principle/rule of distributive justice, creates obligations. It is binding to the extent that it leads to a greater level of compliance. According to some theories on compliance, the motivational gap between norm and action is sanctioned by reciprocal empirical (what others do) and normative (what others expect I should do) expectations. So, the presence or absence of the procedure should not affect the result. That is, the veil of ignorance alone, as a moral device, and the common knowledge about what the relevant other chooses, should induce a certain type of reasoning for which each put herself in the shoes of the other, considering what is the most appropriate thing to do, given the context.

However, although mutual expectations are fundamental for inducing compliance on one rule rather than another, only through the agreement it was observed the highest level of collective choice on liberal egalitarian criterion and the corresponding degree of compliance.

This would indicate, as my third chapter *A Rational Justification for Gilbert's joint commitment by using the Rawlsian constructivist method* tries to explain, two things:

- The commitment that derives from a collective intentionality to do something together is not reducible to an aggregate of individual intentions to do something together. People would perceive themselves as a plural subject and this feeling would create a *different* type of commitment and responsibility able to provide motivation to act.
- Agents' deliberation on the liberal egalitarian principle and relative compliance occurs because a special kind of agreement is reached: an agreement which, through impartiality and impersonality, rationally justifies that principle of justice – hence, a moral principle - and therefore, it justifies the formation of a joint commitment and corresponding obligations to to act as collectively decided.

So, to see to what extent the oughtness of the liberal egalitarianism would give real motivations to act in production situations as the one described in this project, a prior agreement, that satisfies conditions of impartiality and impersonality, would be a necessary determinant for collective conformity.

Resumen

Esta investigación se centra en el estudio de la razón práctica, en particular, en su declinación moral. Eso implica considerar a los seres humanos como capaces de elección y libre albedrío: la acción moral no es el resultado de leyes deterministas que regulan la naturaleza, sino el resultado del razonamiento, de la deliberación y de la intención de actuar de acuerdo con lo que se decidió previamente. ¿Qué razones tengo para preferir una acción moral a otra? ¿Qué significa actuar moralmente? El espectro de ese tipo de preguntas puede extenderse mucho, por lo que es necesario circunscribir el campo de investigación al que se dirige esta investigación.

Este trabajo se inscribe en el marco de la ética normativa, disciplina que estudia la formación de juicios morales y la motivación resultante para la conducta conforme a los mismos. El desfase entre la dimensión del bien común y de la vida privada de cada ciudadano ha llevado a cuestionarse si existen normas morales cuyo contenido pueda constituir, per se, una razón suficiente para la acción. El estudio de la dinámica del comportamiento en el laboratorio ha revelado cómo las preferencias altruistas o pro-sociales (other-regarding preferences) y consideraciones de justicia influyen en la toma de decisiones. Además, las normas sociales ampliamente aceptadas representan un elemento más que conduce a un comportamiento conformista dentro de un grupo de referencia: es decir, una norma específica requeriría un cierto tipo de comportamiento, provocando una serie de expectativas recíprocas – empíricas y normativas – que motivan que las personas cumplan mayoritariamente con lo que la norma prescribe.

Resta la cuestión de comprender bajo qué condiciones logra una norma poseer la fuerza vinculante suficiente como para ser cumplida. Más específicamente, y por las razones que se explicarán en detalle a lo largo de la tesis, en este estudio se formula la pregunta en estos términos: ¿cuándo logra una norma, elegida y compartida colectivamente, auto-imponerse sin intervención de ninguna autoridad externa? Las normas sociales incluyen una serie de comportamientos colectivos

que incluyen las convenciones, las modas, y otros, entre los que destacan aquellos fenómenos (normas sociales en sentido propio) cuyo establecimiento y cumplimiento requiere expectativas empíricas y normativas de los individuos que forman el grupo social donde la norma impera. Un grupo particular de normas sociales son las normas de equidad (norms of fairness): en este trabajo, me referiré a ellas también como normas de justicia, adoptando en líneas generales la teoría de la justicia de John Rawls que, como es sabido, defendió una idea de "justicia como equidad" o "imparcialidad" como aquella descripción más básica y abarcante del concepto. Eso significa que el elemento principal para definir algunos principios distributivos que pueden tener un valor práctico es que sean justos o imparciales (fair).

El interés por este tema surge de la urgencia de definir las condiciones para que un grupo de agentes interdependientes pueda organizarse para obtener beneficios colectivos a largo plazo; para encontrar en la acción colectiva una motivación, un compromiso y una responsabilidad que disminuyan la tentación de comportamientos oportunistas en contextos de distribución.

El objetivo principal de esta investigación es tratar de proponer una solución normativa a un problema de justicia distributiva en los siguientes términos: ¿cómo puede una norma generar una fuerza causal, motivacional, que induzca el cumplimiento de lo que prescribe, en contextos donde el comportamiento egoísta sería el esperado en lugar de la sumisión a la norma? Este problema, abordado desde diferentes perspectivas –económica, psicológica, sociológica– requiere también una reflexión filosófica. En este marco, el problema del cumplimiento podría releerse como un problema motivacional: por lo tanto, el objetivo sería tratar de aclarar la relación entre un punto de vista ético imparcial y qué tipo de motivaciones reales tienen las personas para actuar de acuerdo con algunos principios éticos específicos, a saber, qué razones tienen las personas para alinear su conducta con los principios de la justicia distributiva.

Vale la pena señalar que este propósito encaja en un marco metaético, en particular desafiando el problema del punto de vista moral (moral point of view). Tradicionalmente, las teorías éticas se

han dividido en defensores del punto de vista impersonal, neutral o en tercera persona o en quienes apoyan una perspectiva desde la primera persona, relativo al agente (agent-relative). Qué punto de vista adoptar tiene una importancia fundamental cuando tenemos que hablar de justicia distributiva: lo que puede considerarse justo para uno, puede ser injusto para otro, ceteris paribus. Esta es la razón por la cual Frohlich y Oppenheimer (1992, p.3) sostienen que "the key to understanding distributive justice is impartial reasoning": el punto de vista ético, imparcial por su propia definición, sería el único capaz de exceder los intereses personales en nombre de lo que debe hacerse, de modo obligatorio (ought to do).

A través de un enfoque tanto especulativo como experimental, este proyecto intenta analizar un problema circunscrito de justicia distributiva al proponer que, dado el contexto presentado en los experimentos, un criterio distributivo que apela al igualitarismo liberal (liberal egalitarianism) funcionaría tanto a nivel ético como psicológico sólo si se decide colectivamente que esa es la norma que hay que cumplir. Es decir, debe existir coherencia entre el resultado del razonamiento imparcial realizado colectivamente y la motivación psicológica individual para actuar de acuerdo con lo que prescriben los principios morales.

Antes de explicar qué se entiende por decisión colectiva sobre la norma qua ha de seguirse, debe aclararse a qué se refiere la expresión 'igualitarismo liberal'. Esta expresión tiene tres sentidos: 1) como ideal de equidad, moralmente entendido; 2) como una expresión utilizada por Degli Antoni et al (2016) para indicar una regla distributiva específica; 3) como una solución normativa a un problema de justicia distributiva, donde los puntos 2 y 3 están estrechamente relacionados.

El igualitarismo liberal, como ideal de equidad, se considera un criterio para resolver problemas redistributivos. En esta investigación, que sigue el trabajo de Degli Antoni y sus colegas, se usa también como principio de justicia (distributiva), con referencia directa a la teoría de John Rawls. Rawlsiano es también el resultado del procedimiento constructivista, que determina la decisión sobre qué norma adoptarán las partes. Como se verá en el capítulo IV, el igualitarismo liberal, un ideal en la medida en que proporciona una solución a cómo se debe dividir (equitativamente) un

producto total, difiere de otras intuiciones morales sobre cómo considerar la justicia. En referencia a los experimentos concretos que se describen en esta tesis los individuos tienen que decidir sobre una regla distributiva de entre un conjunto de reglas, cada una de las cuales sintetiza una intuición moral: el egoísmo del homo oeconomicus (un participante se lleva todo y el otro se queda sin nada); el puro igualitarismo (cada uno reciben la mitad del producto común); el igualitarismo meritocrático (cada jugador obtiene una porción que depende de distintas variables como el tiempo disponible o su productividad). El igualitarismo liberal, por lo tanto, se incluye como otro principio distributivo que tiene como su objetivo principal, corregir las diferencias injustificadas causadas por mera suerte, independientemente del mérito personal. Leído en estos términos, este principio está muy cerca del principio de diferencia Rawlsiano, según el cual el único tipo permitido de desigualdades socio-económicas de una sociedad bien-ordenada deberían ser aquellas que sirviesen para beneficiar a los más desfavorecidos.

El objetivo más específico es determinar cómo se concibe y se sigue un principio liberal igualitario como un estándar de conducta que guía la acción bajo la condición necesaria de que sea el resultado de un pacto unánime, bajo el velo de la ignorancia.

El enfoque de esta investigación es filosófico y metodológico. Partiendo desde la perspectiva filosófica, se trata de presentar el constructivismo kantiano como una metaética que tiene implicaciones normativas. Hay dos aspectos metaeticos principales: el punto de vista moral y el realismo de los principios/comandos morales. El primero pregunta cuál es el punto de vista moral que debe adoptarse para que un enunciado pueda valer como moral, mientras que el segundo pregunta hasta qué punto un principio moral, que satisface las condiciones de universalización, prescriptivismo y superioridad (overridingness), puede constituir una fuente concreta de motivación moral. En otras palabras, ¿la perspectiva moral, para ser tal, tendría que referirse solo al punto de vista impersonal e imparcial (el ojo del observador externo, la perspectiva de la tercera persona) o, en cambio, debería también tener en cuenta la perspectiva desde la primera persona?

Esta pregunta se entrelaza con la cuestión sobre el realismo de los argumentos morales. Es preciso destacar que, al usar el término 'realismo', me refiero a la influencia que un precepto moral puede o no tener sobre la conducta. ¿Hasta qué punto puede una máxima moral influir en el comportamiento de las personas? 'Realismo', por lo tanto, se refiere a cuánto puede un principio o argumento moral motivar una conducta en conformidad con el principio mismo. Los dos aspectos están vinculados: el constructivismo kantiano para Rawls podría constituir un método que permita integrar el punto de vista imparcial con las perspectivas relativas a los agentes. En el caso de la justicia como equidad, la justicia de los principios se establece como resultado de un método que los justifique dentro de su procedimiento de construcción, manteniendo unidas la perspectiva imparcial y las psicologías individuales, creando así una fuerza motivadora de tal manera que esos principios sean considerado justos no solo idealmente, sino también prácticamente.

En segundo lugar, este trabajo quisiera destacar un problema relacionado con las normas de justicia: ¿se trata de normas morales o sociales? Más adelante se expondrá un modelo bien conocido de normas sociales (capítulo II, sección 3), como es el formulado por Bicchieri (2006). Según este modeo, las normas de justicia entran en la categoría de normas sociales, porque lo que es justo dependería del contexto y del tipo de comportamientos interdependientes que los individuos realizan cuando cumplen una regla de justicia distributiva. Por ejemplo, puedo obrar en conformidad con una norma porque creo que lo que la norma afirma debe hacerse, y/o porque seguir la norma promueve mis intereses personales. También puedo conformarme porque espero que otros lo hagan: este tipo de expectativas presuponen un conjunto de preferencias condicionales sobre el comportamiento de los demás.

Seguir una norma de justicia sería el resultado de preferencias condicionales: una vez que una norma sea reconocida colectivamente como relevante, los individuos tendrían preferencias condicionales para cumplir (condicionadas a las preferencias de los demás con quienes interactuamos). Sin embargo, lo que puede considerarse justo para un individuo, puede ser

completamente injusto para otro. Entonces, ¿cómo podría uno determinar qué es justo o qué se debe hacer dadas las circunstancias del entorno en el que nos encontramos? Las preferencias condicionales de ajustarse a lo que los demás hacen, no garantizan la justicia (rightness) de una norma. Por ejemplo, hay que pagar el billete del tranvía en Roma, pero ves que los demás no lo pagan, razón suficiente para seguir la norma tácita de no pagar en lugar de ir en contra de ella (¡de hecho, podrían verte como la oveja negra si lo haces!). Hay razones (condicionales) para seguir una norma social nueva o bien establecida, incluso si es colectivamente perjudicial. En estos casos, puede haber un sentido, una convicción común, de que esto no es lo que la gente debería hacer, porque no es justo y es probable que dañe el bien común; por ejemplo, disminuye la calidad y el mantenimiento de los servicios públicos. Las normas morales entrarían aquí en juego, entendidas como mandatos, como preceptos, que dirigen nuestra acción super partes: como normas universalizables, las normas morales parecen ser un punto de referencia para evaluar una norma social como justa, o no. La universalización lo haría posible al abstraerse de situaciones específicas para encontrar la aplicación más amplia. Esta característica actuaría como guía, como heurística, para comportamientos individuales y colectivos en situaciones dilemáticas. Además, puede constituir un parámetro a través del cual se puede lograr un equilibrio reflexivo (reflective equilibrium) crítico sobre las normas sociales: sin una norma moral, universalizable y potencialmente compartible por todos, las personas no serían capaces de modificar prácticas o normas sociales injustas (por lento y difícil que sea el procedimiento de cambio). Se sigue de esto que las normas de justicia no se pueden delimitar exclusivamente dentro de la categoría de normas sociales. ¿Cómo cambiaría una norma social injusta si no hubiera principios/argumentos morales que definan lo que moralmente se debe hacer (ought to do)?

Suprimir la diferencia entre las normas sociales y morales, en vez de diferenciarlas por el tipo de comportamiento (interdependiente versus independiente) y el tipo diferente de preferencias individuales de las que se derivan (condicionales versus incondicionales), supondría vaciar las normas morales de su función original: guiar el ethos humano. Éste es un término griego, del cual

precisamente procede la palabra "ética", que logra devolvernos una complejidad de significados en torno al valor de los argumentos morales: significa al mismo tiempo disfraz, hábito, pero también carácter, personalidad. El ethos como una permanencia del ser, como la contingencia histórica en la que somos arrojados y en la que nos encontramos. He ahí la importancia de los argumentos morales, al ayudar a las personas a ejercer una reflexión crítica y ser activamente capaces de cambiar las normas sociales cuando hacen que las instituciones sean injustas. Sin este tipo de punto de referencia heurístico, que nos permita juzgar si una norma social es justa o no, cambiar una mala norma social parecería un camino sin salida. Cuando se trata de justicia, tendría que haber un acuerdo unánime sobre que entendemos por justicia y, al hacerlo, podemos también dejar espacio para críticas y revisiones de una práctica o una institución que, con el tiempo, demuestra ser injusta. Por lo tanto, si pensamos que las personas son responsables de sus acciones y de la sociedad en la que quieren vivir, los argumentos morales sobre lo que consideramos justo deberían considerarse como factores de motivación para cumplir con aquellos principios. Entonces, este trabajo parte de suponer que el constructivismo kantiano podría considerarse como una metodología metaética que lleva a las personas a la misma concepción de la justicia en un contexto específico de medios de producción. El constructivismo kantiano, de hecho, sería el candidato más adecuado para demostrar cómo un argumento moral puede ser realista, donde, por 'realista', se entiende un argumento moral que pueda ponerse en práctica. Dado el contexto experimental (explicado sobre todo en el capítulo IV), la segunda hipótesis principal de esta tesis es que el igualitarismo liberal sería un argumento moral capaz de obtener un apoyo motivacional lo suficientemente fuerte como para implementar la acción prescrita (es decir, un principio moral) en el comportamiento práctico. Comprender en qué medida el principio liberal igualitario, como argumento moral para resolver un problema de justicia distributiva, es realista depende de hasta qué punto su justificación racional y sus motivaciones derivadas estén unidas. Sería un método –el constructivismo kantiano– para justificar los principios de justicia, proporcionando, al mismo tiempo, motivaciones para conformarse a ellos: el procedimiento justifica el contenido de los

principios (a través del acuerdo) y, a su vez, su justificación (o sea el porqué se han elegido ciertos principios en vez de otros) da razones para ponerlos en práctica. Además, el hecho de acordar colectivamente esos principios tiene una fuerza psicológica: motiva a los agentes porque es el resultado de un compromiso conjunto (joint commitment), que debe entenderse en los términos de Gilbert. Digo 'fuerza psicológica' porque llegar a un acuerdo sobre los principios de justicia que a las personas les gustaría adoptar para las instituciones de su sociedad no implica, lógicamente, ningún vínculo causal con el comportamiento real de las personas.

Ese "gap" es lo que John Rawls presentó como el problema de la estabilidad en A Theory of Justice (1971, 1999). Según Rawls, el elemento que permitiría a las personas actuar en conformidad con el contenido del pacto es un sentido de justicia (sense of justice). Esto se considera una disposición psicológica que interviene en el proceso de deliberación y da razones para actuar de manera justa. Sin embargo, en ausencia de una autoridad externa que sancione a aquellos que no respetan el pacto, verificar si el acuerdo podría ser auto-impuesto (self-enforcing) sigue siendo un reto.

Por lo tanto, la principal pregunta de investigación es: ¿bajo qué condiciones podría un acuerdo alcanzado tras el velo de ignorancia proporcionar una motivación real para actuar? ¿Puede ese procedimiento obligar normativamente la voluntad de los agentes?

El intento de responder a esta pregunta requiere una serie de sub-preguntas más específicas:

1. ¿Podría el velo de ignorancia Rawlsiano ser un dispositivo moral capaz de vincular la voluntad de los agentes hacia un comportamiento justo?
2. Si se verifica (1), ¿podría el velo de ignorancia afectar a las personas incluso sin un acuerdo explícito previo entre las partes?
3. ¿Podría el acuerdo imparcial proporcionar razones para elegir un criterio distributivo igualitario liberal que garantice el cumplimiento (compliance)?

4. ¿Podría este tipo especial de acuerdo, concebido como un compromiso conjunto de Gilbert, ser un procedimiento que unifique la imparcialidad ética y la perspectiva relativa al agente?

Esta investigación proporciona evidencia de experimentos de laboratorio que apoyan el constructivismo kantiano por John Rawls como un método que permite a los sujetos evaluar diferentes criterios de distribución en situaciones de producción, haciendo que lleguen a un acuerdo sobre la regla liberal igualitaria. El método rawlsiano justificaría racionalmente la elección colectiva ex-ante sobre ese principio y daría motivaciones reales para cumplir ex-post, ya que aporta las condiciones para crear un compromiso conjunto racionalmente justificado.

Se intentará comprobar dos cosas:

- En línea con la literatura sobre las normas de justicia, la concepción de justicia que los individuos tienen internamente no parece coincidir con un principio de igualitarismo liberal. Esta concepción cambia cuando es el resultado de un procedimiento de deliberación colectiva, que presupone un sujeto plural (plural subject) con una intencionalidad colectiva que no puede reducirse a la suma de la intencionalidad de muchos individuos.
- Para que un principio, directamente derivado del igualitarismo liberal, constituya una norma social a la que las personas se conforman, un factor determinante es que se susciten expectativas mutuas de cumplimiento. Sin embargo, un compromiso individual con la decisión anterior que cada uno tomó, da razones más débiles para cumplir con la norma elegida, en comparación con un compromiso conjunto.

Los tres capítulos principales que constituyen la tesis desarrollan estas cuestiones en ese marco metaético y metodológico. El primero Does impartial reasoning matter in economic decisions? An experimental result about distributive (un)fairness in a production context presenta un primer intento de poner a prueba el velo de la ignorancia como un dispositivo moral para provocar un

comportamiento más justo en un contexto distributivo. Los diseños experimentales presentados aquí y en el capitulo V intentan recrear lo que Rawls llamó la posición original. La idea básica es colocar a los sujetos en una posición en la que no sepan quién será el más o el menos favorecido y, bajo esta condición de ignorancia, deliberar sobre qué distribución elegir –qué regla distributiva aplicar para distribuir el resultado común de una tarea productiva.

En línea con Rawls, hablar de una sociedad justa significa pensar en una sociedad donde las desigualdades se admiten sólo en beneficio de los más desfavorecidos. Un aspecto que los dos capitulos experimentales tienen en común es el hecho de que los sujetos deben pensar en cómo les gustaría compartir un determinado output obtenido a través de un esfuerzo/trabajo, sabiendo que todos tienen un 50% de probabilidad de ser más o menos favorecidos en la cantidad de tiempo para desarrollar la tarea (o sea, el trabajo). En este capítulo, contrariamente a nuestras expectativas, los sujetos no fueron influenciados por la reflexión individual trás del velo de la ignorancia. Al contrario, se comportaron de una manera más egoísta, frente al tratamiento sin velo. Este resultado disputa la función del velo de la ignorancia como un experimento mental individual. En otras palabras, el velo de la ignorancia funcionaría como un dispositivo moral porque interviene a nivel individual, pero sólo dentro de una comunidad. Se requiere el consenso unánime fáctico para llegar a un acuerdo imparcial sobre qué principios de justicia deberían aplicarse en una sociedad. Por lo tanto, considerando la construcción rawlsiana de una teoría de la justicia, los resultados del capítulo IV parecen invalidar el efecto de aprendizaje moral del velo de la ignorancia.

El siguiente capítulo experimental Distributive justice in the lab: testing the binding role of the agreement, siéndose plantea comprender por qué el velo de la ignorancia, individualmente, no tuvo ningún efecto hacia un comportamiento más justo. A diferencia del capítulo precedente, en estese pide a los sujetos que lleguen a un acuerdo: los resultados muestran que el procedimiento

del pacto no solo modifica la concepción de lo que se considera justo en la distribución, sino que afecta el nivel de cumplimiento, con respecto a una decisión individual.

Un estándar distributivo como el igualitarismo liberal es el que se prefiere en presencia del pacto, mientras que no es el más elegido cuando la decisión es individual. Lo que se puede inferir de estos resultados es que el procedimiento mismo –llegar efectivamente a un acuerdo unánime sobre un principio/regla distributivo– crea obligaciones. El acuerdo es vinculante en la medida en que conduce a un mayor nivel de cumplimiento. Según algunas teorías sobre el cumplimiento, el gap motivacional entre la norma y la conducta está mediado por expectativas recíprocas empíricas (lo que los demás hacen) y normativas (lo que los demás esperan que yo tendría que hacer). Por lo tanto, la presencia o ausencia del procedimiento no debería afectar el resultado. Es decir, el velo de la ignorancia, solo, como dispositivo moral, y el conocimiento común sobre lo que elige la otra persona en nuestra pareja de juego, debería inducir un cierto tipo de razonamiento por el cual cada uno se pone en el lugar del otro, considerando cuál regla distributiva sería la más apropiada, dado el contexto, y generando así cierta expectativa recíproca. A pesare de esto, aunque las expectativas mutuas son fundamentales para inducir el nivel de cumplimiento sobre una regla en lugar de otra, solo a través del acuerdo real se observa el nivel más alto de elección y conformidad colectivas por el criterio igualitario liberal.

Esto indicaría, como el capítulo V, A Rational Justification for Gilbert's joint commitment by using the Rawlsian constructivist method, intenta explicar, dos cosas:

- El compromiso derivado de una intencionalidad colectiva de hacer algo juntos no es reducible a un conjunto de intenciones individuales de hacer algo juntos. Las personas se percibirían a sí mismas como un sujeto plural y este sentimiento crearía un tipo diferente de compromiso y responsabilidad capaz de proporcionar motivación para actuar.

- La deliberación de los agentes sobre el principio igualitarista liberal y el cumplimiento relativo se produce porque se llega a un tipo especial de acuerdo: un acuerdo que, a través de la imparcialidad y la personificación, justifica racionalmente ese principio de justicia, por lo tanto,

un principio moral, y por lo tanto, justifica la formación de un compromiso conjunto y las obligaciones correspondientes de actuar según lo decidido colectivamente.

Entonces, para ver hasta qué punto la obligatoriedad (oughtness) del igualitarismo liberal daría motivaciones reales para actuar en situaciones de producción como la descrita en este proyecto, un acuerdo previo explicito – que satisfaga las condiciones de imparcialidad e impersonalidad – sería un determinante necesario para la conformidad colectiva.

Esta conclusión me permitirá defender que, al menos sobre la base de la evidencia empírica obtenida con los experimentos aquí presentados, se puede reivindicar desde el punto de vista meta-ético el procedimiento constructivista rawlsiano: el proceso de acuerdo tras un velo de ignorancia efectivamente genera un consenso sobre un principio liberal igualitarista de distribución, de un modo que la mera reflexión individual desde una perspectiva de tercera persona no lo hace. Además el acuerdo real genera una motivación para cumplir con esa regla distributiva de un modo más eficaz que el uso del acuerdo como mero experimento mental. Finalmente, que esto sólo puede explicarse porque los sujetos adoptan en el caso del acuerdo real, pero no así en el caso del razonamiento individual, una perspectiva conjunta, en la que se razona como un sujeto plural o colectivo, cuyas intenciones son interiorizadas por los individuos que lo forman, según la explicación de Gilbert. Esta explicación es la única consistente con los resultados experimentales, y por tanto los mismos cuestionan la explicación del cumplimiento de normas de justicia basada en la teoría de las normas sociales de Bicchieri. En definitiva, esta tesis muestra que el método experimentar puede arrojar luz sobre cuestiones fundamentales de ética normativa y meta-ética; lo que representa una línea de trabajo filosófica que merece la pena desarrollar en el futuro, a pesar de las limitaciones que posee, de las que soy consciente, y de las que se dará cuenta a lo largo del texto.

I. Introduction

This doctoral dissertation is consistent with a line of research that seeks to study the relationship between rationality and normativity. Both these concepts are held together by a particular type of reason, that philosophy defines as practical reason, understood as "the general human capacity for resolving, through reflection, the question of what one is to do". This implies a twofold concern for the understanding of whether deliberation in itself may directly imply action and for the content of the deliberation, i.e. which rules for assessment of action are found to be binding (Wallace, 2014). Practical reason mainly concerns the sphere of moral decisions, that is when there is a conflict between what a moral argument prescribes – what should be done – and the action that is carried out. Such behaviour may conform to moral principles or may not. It follows the human capacity to define and judge an act as moral, immoral or amoral. Two different perspectives emerge: on the one hand, the impartiality of ethics, of reason – understood in Kantian terms – which prescribes what actions should be undertaken; on the other, there are people's partial and personal points of view that must choose one course of action. These are two sides of the same coin, thus, the problem is to see how they may interact with each other. For example, concerns for fairness and other-regarding preferences are considered motivational factors[1] that induce people to act in an altruistic way, contrary to what one would expect from the classical rational choice theory (Fehr and Gächter 2000, 2002; Dawes et

[1] Over the past 50 years, many data collected in laboratory experiments have shown systematic behavioural deviations from the typical self-interest predicted by rational choice theory (Camerer 2003). A new interest in what were the motivational factors responsible for these deviations emerged. Scholars in this area usually identify the various theories in two groups, those based on so-called social preferences (Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Andreoni and Miller 2002) and those based on the notion of reciprocity (Rabin 1993; Charness and Rabin 2002). The main difference consists in the motivation that induces the player to choose: for the first group of theories, the choice depends on a focus on the distributive consequences; for the second, the choice is also the result of the beliefs that each player respects mutual kindness.

al 2007.) Nevertheless, there is the urgency to redefine a concept of rationality that is not only bounded, as highlighted by Simon (1955), but also moral, considering the moral intuitions that inevitably affect human life. It should be clear that this is not the place for a metaphysical discussion about moral norms, rather it is a matter of recognizing that, on a practical level, moral intuitions and beliefs have a decisive weight within the decision-making process. Trying to study collective behaviour in production situations about the distribution of material goods is the horizon within which this project takes place, motive for which it may be summarize as an interdisciplinary project straddling moral philosophy and experimental economics.

This thesis is indeed about the interaction between the impersonal point of view and the personal perspectives, when people face a problem of distributive justice. More precisely, the general objective is to understand, *ceteris paribus*, whether people comply with the fairness norm they have chosen, and to what extent their behaviour changes depending on how a norm of fairness is implemented – i.e., as the result of a collective action or as an individual choice. As Frohlich and Oppenheimer (1992, p. 3) say: "reasoning premised on setting aside one's particular interests and perspectives and giving balanced weight to the interests of all. It is reasoning not from one's own narrow viewpoint but from the broadest possible perspective". It might be said that this is the traditional task of ethics: to be such, moral arguments must be universalizable, they must be applicable to as many situations as possible, abstracting from particular cases. Then, they must be prescriptive: a moral principle should offer a guidance to action. Finally, a moral principle, to be effective, must override other alternatives: for instance, a moral judgment is overriding against competitive others if it is the one we refer to implement actual behaviour. However, the gap between 'ought' and 'is', between impartial reasoning and personal (and therefore partial) points of view, remains and it needs be recognized in order to

understand how these two elements – both present in the decision-making process – are tied one to the other.

By using both the empirical methodology of experimental economics and the theoretical methodology of ethical reflection, a twofold contribution follows. In the attempt to decrease the distance between 'ought' and 'is', "an empirical approach based on laboratory experiments is required to discover what impartial individual would do" (Frohlich and Oppenheimer, 1992, p.3). Experiments would approach the psychological level where reasons for action emerge. Then, it might be thought that philosophical inquiry no longer matters, since the main interest is to observe how people behave when a conception of fairness is required. Instead, the philosophical contribution is very relevant: empirical evidence describes states of affairs, but it would exclude that normative component present in human reasoning and decision-making. Someone's behaviour can be normative, that is, the reasons she has for doing something can be traced back to a moral principle/argument she upholds. In the case of fairness norms, as this research will show, impartial reasoning must be considered as a device (i.e., the veil of ignorance), in order to find a normative solution[2] to a problem of distributive justice.

The goal is to study which conditions are required to produce impartial reasoning, that is, understanding what conditions must be met to lead individuals to put in practice those principles stemmed from the impartial reasoning procedure. More specifically, it is about what conditions must be satisfied by agents to consider rational both the choice to apply and the conduct to comply with a liberal egalitarian rule in a production context[3].

[2] When I argue that it is a 'normative solution,' I mean that the solution is a norm (of fairness), prescribing what one ought to do, by eliciting obligations of doing so.

[3] See Chapter IV for an adequate explanation of what 'liberal egalitarianism' and 'liberal egalitarian rule' mean in this research.

The thesis consists of five chapters, where the third, the fourth and the fifth develop the central part of the research, preceded by a chapter regarding the state of the art in which the project takes place. The last chapter sums up general conclusions and thesis' implications.

The three central chapters are written in the form of scientific articles: the first two (III and IV) are based on laboratory experiments, while the third one (V) is specifically philosophical, the aim of which is to theoretically systematize the collected empirical evidence. Each chapter presents its own conclusions.

The state of the art is articulated in four sections: (1) moral reasoning; (2) Kantian constructivism; (3) social norms and norms of fairness; (4) methodology. I will start by presenting the main problem of theories of moral motivation, that is what reasons people actually should act morally. Then I will present the Kantian constructivism of John Rawls. His philosophical approach is ground-breaking for several reasons: firstly, for his conception of justice as fairness; secondly, for not keeping ethics and moral psychology as two separate field of research; thirdly, for the concept of a moral person that he introduced. Then, I will present how social norms influence and guide our behaviour, especially regarding to norms of justice. Finally, I will deal with the great impact that the behavioural turn had within economics and how experimental economics, as methodology, may be helpful for moral philosophers (and vice versa).

II. State-of-the-art

1. Moral reasoning

When we speak about moral judgment, deliberation, moral action, we refer to a specific kind of reasoning underlying those processes, namely what has been called moral reasoning: "an agent's first-personal (individual or collective) practical reasoning about what, morally, they ought to do" (Richardson 2018). Moral reasoning is considered an essential element when we look at how moral judgments are formed: the field dealing with moral judgments' formation focuses both on cognitive mechanisms and on judgments concerning decisions, facts, that call into question socially accepted and shared norms (and relevant behaviours) belonging to the domain of morality. Starting from the model of moral development theorized by Piaget (1932), Lawrence Kohlberg (1974) proposes an explanation of moral judgment which is mainly based on the fundamental role of reason. According to him, the sphere of morality includes feelings, thoughts and actions but it is reasoning that qualifies actions as specifically moral: it focuses on normative judgments prescribing what one should do or what is the right behaviour. Kohlberg's theory describes the personal development of deontic reasoning, that is, a reasoning that focuses on the problem of norms and justice (in terms of rights and duties), affirming that universal problems of this kind constitute the core of morality. From this, it can be said that Kohlberg followed the Kantian tradition, favouring reasoning and by promoting the categorical imperative: a behaviour, for being considered categorically moral, must be universalizable. According to this perspective, emotions can be stimuli for reasoning processes, but moral emotions have never been the direct cause of moral judgments.

Opposite to the rationalist view, we find Humean empiricism, that considers passions – including feelings, desires, emotions, etc – and not reasoning the ultimate source to influence human willingness (to perform some moral action). However, such a strict opposition has limitations in both extremes. As Korsgaard said:

"On an empiricist view, to be practically rational is to be caused to act in a certain way-specifically, to have motives which are caused by the recognition of certain truths which are made relevant to action by one's pre-existing motives. On a rationalist view, by contrast, to be rational is to deliberately conform one's will to certain rational truths, or truths about reasons, which exist independently of the will" (Korsgaard 1996, p. 219).

This gave rise to the most recent debate on the source of motivation that separated philosophers between internalists and externalists (Falk 1947; Frankena 1958; Stevenson 1937; Smith 1994). The former argue that moral judgments necessarily motivate action, both in the strong sense that motivation is contained in moral judgments and in its weaker version that supports a necessary connection between motivation and moral judgment. Externalists, instead, deny internalists' claim because they see "moral motivation being mediated by desires and feelings that exist prior to the moral judgment and that explain instrumentally why moral judgments normally have motivating force. For example, if what is just promotes what we antecedently care about, the fact that something is just gives us an instrumental reason to care about it" (Campbell 2007, pp. 324-325).

It is not my intention here to analyse these positions in detail. My aim is rather to clarify in what way motivational elements would shape moral reasoning. How can moral reasoning linked with

motivationally psychological states induce causal effect? How can moral reasoning lead people to do what they actually do?

These questions find their roots within the traditional philosophical debate between facts and values, between 'is' and 'ought' (namely the naturalistic fallacy)[4]. However, it happens that our actions are often guided by rules, by precepts, which tell us what should be done, namely normative reasons that influence final behaviours. For instance, how can we judge a set of institutions as fair? How can we ensure that the society in which we live provides adequate opportunities and primary goods to allow each citizen to live their lives with dignity? Under what conditions do the hypothetical principles of justice lead to *real* actions to act in accordance with them?

The motivation theory by John Rawls and his constructivist method might indicate the conditions for overcoming the fracture between 'is' and 'ought', linking together moral reasoning and *real* (and therefore psychological) reasons for action.

Before going into the Rawlsian theory, it should be noted that there are four main ongoing theories of moral motivation, whose aim is to clarify the relationship between reasons that are recognized as normative – i.e. stealing is wrong – and reasons that actually affect the action. What kind of correlation is at stake? One can theorize from the armchair on the normativity of moral judgments, on the type of obligation that stems from them, but if there is a real interest in understanding what drives a person (or a collectivity) to behave, the speculation should be integrated with empirical evidence – without diminishing the importance of a profound and critical philosophical reflection.

[4]For an overview on the dichotomy between normative and positive, values and facts, see Hands, D.W. (2012). The Positive-Normative Dichotomy and Economics, in Uskali Mäki (ed.), *Philosophy of Economics*, Elsevier, pp. 219-39.

Indeed, as postulated by Schroeder and colleagues "motivation is not merely an abstraction, and that it plays a causal role in the production of action" (Schroeder et al 2010, p.72).

This study systematizes those four main approaches regarding the source of moral motivation: the instrumentalist, the cognitivist, the sentimentalist, and the personalist. According to the instrumentalists, the motivation to perform an action arises as a satisfaction of an intrinsic desire: when an individual forms the belief that, by doing a certain action, she will fulfil her desire, then she is motivated to perform that action. Sentimentalists instead identify emotions as source of motivation: they argue that an action, to be morally valid, must be motivated, ultimately by emotions of some kind, for example compassion. Unlike these two previous models, that identify in a mental state – i.e., desire and sentiment – the source of moral motivation, personalists focus on the good character, proposing a holistic view. Referring to the Greek conception of virtue as of that exercise which, if practised consistently, forges the good character of a person, supporters of this model believe that an action is morally worthy if it comes from a good character. However, it is the cognitivist model that I lean towards:

"The cognitivist holds the view that moral motivation begins with occurrent belief. In particular, it begins with beliefs about what actions would be right. The cognitivist holds that, at least in cases of morally worthy action, such beliefs lead to motivation to perform those actions, quite independently of any antecedent desires. The cognitivist is happy to call this motivational state "a desire," but thinks of it as entirely dependent upon the moral belief that created it" ((Schroeder et al 2010, p.76).

Although not counted among the names of philosophers who have adopted a cognitive model to establish the theory of moral motivation, John Rawls would have proposed a very congruent view.

He speaks of principle-dependent desires "whose objects cannot be characterized without reference to some rational or moral principle […] for instance, conceiving of oneself as a citizen, one may desire to bear one's fair share of society's burdens" (Richardson 2018). Despite the term desire, the Rawlsian approach might be considered cognitive, in the sense described above: moral action might be considered worthy when one believes that this action is consistent with an objective and universalizable content, a command of reason that can be publicly recognized and judged. The introduction of this moral motive is the result of Rawls's conception of what morality is. It could be said that he regards morality as a social and psychological phenomenon. The Rawlsian meta-ethics depends on some assumptions that make it possible to justify justice as fairness, making it the best alternative for a normative ethics. As we will see in detail in the next section, dedicated to Kantian constructivism, Rawls holds the *a priori* assumption on the nature of moral agents (Cremaschi 2004). Proposing the principle-dependent desires as the motivational force for moral action could solve the antagonism between 'is' and 'ought', showing how, *de facto*, a theory of justice, to be practical, cannot separate normative foundations from moral psychology.

2. Kantian constructivism

What reasons do people have to act morally? Kant answered this question resolutely with the formulation of the moral law: the motive for moral action stems from pure practical reason, since it is the only one, unlike desires and passions, which can guarantee objectivity and universality. Kant's position inaugurated one of the most influential theories of normative ethics in the history of philosophy: deontology. Traditionally, deontological theories deal with propositions claiming that a certain type of action would be right (or wrong), under specific circumstances, regardless of its consequences, while teleological or consequentialist theories affirm that an action is right or wrong depending on its tendency to produce certain outputs, respectively good or bad. Deontologism, teleologism and ethics of virtue are the three main theories that constitute the ground of normative ethics, whose object is justifying moral actions based on corresponding moral norms and judgments (Da Re 2010). The most successful teleological doctrine is the utilitarianism (Mill 1859; Bentham 1907), a theory according to which an action is right if the consequences produce good. So, in broad terms, teleologism identifies in the consequences the parameter for judging an action as morally valid. Deontologism, on the other hand, is focused on the intention underlying the action. The ethics of virtues differs from the two previous theories in that it focuses more on the agent who performs the action instead of on the act itself.

For the present purpose, I deal with deontologism, in particular in the renewed form of Kantian constructivism proposed by John Rawls. Kant argued that the moral law inherent in each of us guides behaviour, prescribing what should or should not be done: the will can be determined according to the moral law, but also according to the desires and sensitive inclinations. At this point, the will must choose whether or not to follow the moral law, choice that is not automatic, the result of a constraint,

namely the command of the reason to will (Landucci 1993). The categorical imperative directs human will which, being free, is not determined as the laws of nature, but is self-determined: in other words, it is the person who freely and consciously decides how to direct their own will. As we see below, the role played by the will is also central to Rawls. To understand it, we must begin by explaining why Rawls calls his constructivism method 'Kantian', the background from which it emerged, and what assumptions about the nature of the moral agent are supported. First, the adjective 'Kantian' is used by analogy: unlike for Kant[5], practical reason is not totally autonomous for Rawls. He admits some motivational factors that Kant would have defined as heteronomous, for instance the sense of justice as moral sentiment. Although there are several aspects shared with the philosopher of Königsberg (a deontic ethics, the priority of the right over the good, the concept of rationality, the concept of the person), the historical period in which Rawls operates clearly differs. So, a responsible critical philosophical reflection must take into account the different historical periods in which the two lived. The Rawlsian question arises from the recognition of a situation of long impasse within democratic thought, namely the inability to agree on what should be the structure of basic social institutions that conceives "freedom and equality of citizens as moral persons". Indeed:

"The social role of a conception of justice is to enable all members of society to make mutually acceptable to one another their shared institutions and basic arrangements, by citing what are publicly recognized as

[5]The problem of motivation assumes in Kant the terms of how the moral law can move human will, therefore how it can play the role of subjective principle of determination. On a human scale, it is the question of how pure reason can actually be practical. For pure reason, being practical means that it motivates the will, not through sensitive/external motives, but through the representation of the moral law. The latter can effectively exercise a causal action on the human will thanks to the awareness, the consciousness (Gewissen) of the subject. This work of freedom for self-determination finds its positive outcome in the concept of moral respect: "Therefore respect for the moral law is a feeling that is brought about by an intellectual basis, and this feeling is the only one that we recognize completely a priori and the necessity of which we can have insight into" (Kant, I. (1788), *Critique of Practical Reason*, Translated by Werner S. Pluhar, 2002, p.97.

sufficient reasons, as identified by that conception. To succeed in doing this, a conception must specify admissible social institutions and their possible arrangements into one system, so that they can be justified to all citizens, whatever their social position or more particular interests" (Rawls 1980, p. 517).

What Kantian moral philosophy would hold is a conception of justice together with a conception of persons as both free and equal, as capable of acting both reasonably and rationally, and therefore as capable of taking part in social cooperation. A first characteristic that Rawls retrieves from Kant is the concept of autonomy, by distinguishing a rational autonomy of the parties into the original position and a full autonomy of the citizens in the society. The first type of autonomy would see a mixture of autonomous and heteronomous factors underling choice and conformity with the principles of justice. In other words, reasons for selecting and coordinating on certain principles of justice would be prudential (i.e., Kantian hypothetical imperatives). Full autonomy is a different matter: citizens decide which principles of justice they would like to live by in their society. This means, reasons for conforming to the agreed principles would not depend on external elements, but on some endogenous force: they would be self-enforcing. For Rawls, as for Kant, when the will decides to follow a rationally justified moral norm, it determines itself, that is, it finds in itself reason for acting. In turn, this power of the will is possible because a specific conception of person is sustained.

Rawls introduces a concept of a person that proves to be the pivotal point to justify his theory of procedural justice: "we take moral persons to be characterized by two moral powers and by two corresponding highest-order interests in realizing and exercising these powers. The first power is the capacity for an effective sense of justice, that is, the capacity to understand, to apply and to act from

(and not merely in accordance with) the principles of justice. The second moral power is the capacity to form, to revise, and rationally to pursue a conception of the good" (Rawls 1980, p.525).

In so far as we are moral persons, how can we deliberate being both rational and, at the same time, adopting a moral point of view?

To answer this question, it must be clear the difference between rationality and reasonableness. Because moral persons are rational, they are interested in defining and achieving their own personal life plans. In order to do this, they realize that it is necessary to live in a society that guarantees equal rights and opportunities, mostly in social-economic terms. Then, reasonableness comes, guiding people in social life, that is guiding their life plans in relation to other people. Establishing principles of justice throughout an agreement is the first step to ensure a society in which everyone can pursue their individual projects while allowing others around them to do the same. How is it possible to guarantee social justice and personal interest in the realization of one's life? Here, Rawls resorts to what he calls the original position, the main justification strategy of *A Theory of Justice* (1971). As it will be said later in *Kantian Constructivism in Moral Theory*, "the original position is a third mediating model-conception: its role is to establish the connection between the model-conception of a moral person and the principles of justice that characterize the relations of citizens in the model-conception of a well-ordered society" (Rawls 1980, p.520). Often compared with the classical state of nature, it is a situation of impartiality in which individuals, as free and equal persons, would reach a unanimous agreement on the principles of justice, which would have the dual purpose of improving both the structure of political institutions and the socio-economic model of distributive justice. Through this thought experiment, agents would look at different conceptions of justice, therefore different types of moral reasoning, to unanimously lean towards justice as fairness. The impartiality of the procedure would be supported both by the veil of ignorance and the sense of justice, distinctive

of Rawlsian moral agents. To "nullify the effects of specific contingencies which put men at odds and tempt them to exploit social and natural circumstances to their own advantage", Rawls assumes that "the parties are situated behind a veil of ignorance. They do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations" (Rawls 1999, p. 118). Rawls assumes that introducing the veil of ignorance would justify the emergence of procedural justice: the principles of justice that would result from the agreement between the parties, placed behind the veil, in the original position, can only be just. The veil of ignorance implies that the parties do not know certain facts:

"First of all, no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like. Nor, again, does anyone know his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism […] the parties do not know the particular circumstances of their own society. That is, they do not know its economic or political situation, or the level of civilization and culture it has been able to achieve" (Rawls 1999, p.118).

The veil of ignorance makes Rawls's social contract theory different from the contractarian tradition[6]. In fact, it could be defined as a moral device that lead the parties to agree on the principles of justice by adopting a moral point of view (hence impartial and impersonal): ignorance on those facts given the ignorance of their destinies and the choice in the original position, the parties may not negotiate

[6] In philosophy, " 'Contractarianism' names both a political theory of the legitimacy of political authority and a moral theory about the origin or legitimate content of moral norms" (Ann Cudd and Seena Eftekhari 2017, Stanford Encyclopedia of Philosophy). Contractarian theories are based on Hobbes' thinking, while the contractualist ones follow Kant's philosophy.

the terms of the agreement with each other. Without those restrictions "we would not be able to work out any definite theory of justice at all" (Rawls 1999, p. 121).

According to Rawls, in order to ensure a valid rational justification with the ethical principles that underpin the basic structure of a well-ordered society, it is fundamental to sketch a theory that is able not only to justify reasons people have to choose certain principles of justice, but also to legitimate motivating reasons people have to conform to those principles. Ethical principles would be justified through the procedure itself: it is the process of deliberation, where free and equal persons would agree on the principles of justice, that become the rational justification, legitimized by the consent of the members of society. To be clear, a distinction should be made between the content of Rawls's theory, namely justice as fairness, and the method he uses to justify its content, namely the Kantian constructivism. It follows that one could say that the principles of justice are the content of Rawls's normative ethics, while his Kantian constructivism is the metaethical assumption – to be able to identify and rationally justify the principles of justice. By reasoning behind the veil, individuals "do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations" (Rawls, 1999, p.142). The original position is based on the idea that fair procedures could ensure the fairness of the chosen principles. The parties in the original position, although unaware of much information, know that in the real society the principle of limited sympathy and that of scarcity of resources hold. So, an agreement is required on what criteria must be adopted in order to determine the cooperative distribution of costs and benefit. Individuals, faced with that problem, have to choose among different conceptions of justice, through the reflective equilibrium because "from the standpoint of moral theory, the best account of a person's sense of justice is not the one which fits his judgments prior to his examining

any conception of justice, but rather the one which matches his judgments in reflective equilibrium" (Rawls, 1999, p.43).

It should be noted that Rawls denies that the parties act in an original position driven by altruistic considerations, however the individuals behind the veil of ignorance are reasonable and possess a capacity for a sense of justice. The reasonableness capacity together with the capacity of rationality ability define practical reason. To be effective, practical reason must be capable of protecting and pursuing the good and the right. The good is the object of rationality, as each person tries to pursue the means that allows them to achieve their own good, understood as a life plan that everyone would like to achieve – hence, to rationally deliberate on the conditions that, if satisfied , would allow everyone to achieve their own good without harming others for such satisfaction or suffer damage for the achievement of their own good.

A different object, although of a moral nature, is the one proper to reasonableness: it is the concept of right, "which includes individual moral duties and moral requirements of right and justice applying to institutions and society" (Freeman 2014).

Being capable of reasonableness means having a sense of justice: Rawls considers the sense of justice as a psychological disposition, "a condition for human sociability" (Rawls, 1999, p.43). At the basis of this assumption stands the idea that people, insofar as they are capable of moral powers, are not driven solely by selfishness, as Hobbes claimed, but that willingness to behave rightly is founded on the human capacity to have a sense of justice by which they determine what is right and acting accordingly. What is right is defined by moral principles and it is possible because people have two moral powers: "the first power is the capacity for an effective sense of justice, that is, the capacity to understand, to apply and to act from (and not merely in accordance with) the principles of justice. The second moral power is the capacity to form, to revise, and rationally to pursue a conception of

the good" (Rawls,1980, p.525). Rawls assumes a priori that people are capable of these moral powers: it follows that, from these powers, two "highest-order interests" derive, whose realization becomes motive for the will: "... these interests are supremely regulative as well as effective. This implies that, whenever circumstances are relevant to their fulfillment, these interests govern deliberation and conduct" (Rawls 1980, p.525). The sense of justice is introduced by Rawls to justify the stability of his theory of justice as fairness. This psychological disposition has a twofold implication: if the institutions are right and are publicly recognized as such, then the person who benefits from these institutions develops an adequate sense of justice. On the other hand, assessing whether and to what extent the institutions of society in which one lives are right depends on the capacity, we could say an intrinsic capacity, to possess and exercise a sense of justice. In other words, the sense of justice is an attitude that emerges through the contractual procedure to define the principles of justice on which the institutions of a well-ordered basis would be founded.

This awareness of being capable of a sense of justice would also allow to exert a critical point of view through which it would be possible to change institutions if they proved to be unjust. This ability to review institutional justice is made possible by the reflective equilibrium – the latest strategy Rawls used to justify his theory of justice as fairness, along with the original position and sense of justice. It is worth saying, to conclude the Rawlsian methodology, how reflective equilibrium presupposes a balance between the principles of justice (built in the original position) and the considered judgments that each person owns.

"When a person is presented with an intuitively appealing account of his sense of justice (one, say, which embodies various reasonable and natural presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly […] From the standpoint of moral theory, the best account of a person's sense of justice is not the one which fits

his judgments prior to his examining any conception of justice, but rather the one which matches his judgments in reflective equilibrium" (Rawls 1999, pp. 42-43).

The Rawlsian method, therefore, proposes a conception of the human being as a moral person, who, through his rationality and reasonableness, is capable of constructing those ethical principles which, as justified, are motivating for the will. It could be argued that Rawls's Kantian constructivism presents itself as a Socratic maieutic in that it does not dis-veil external moral principles, taken as grounded (position supported for example by intuitionists), but they are the result of a construction. That is, the principles of justice are justified because their content is constructed through a deliberation procedure.

The method itself would link the justification of the content of the principles with the motivation to comply with. If we endogenously justify which principles an institution must respect to be fair, then these principles might be considered realistic: it would be unreasonable to decide and agree on some moral principles starting from the idea that these principles are not feasible. The motivation to put these principles into practice could derive from the fact that the principles are self-determined by the subjects, *via* agreement, and not imposed from external authority.

This procedure leads Rawls to formulate the two principles of justice:

"The first statement of the two principles reads as follows. First: each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others.

Second: social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all" (Rawls 1999, p. 53).

The second part of the second principle is an equal opportunity principle which constitutes a standard element from a liberal perspective. It is defined as the difference principle focusing on the socio-economic effects of a distribution of primary goods. Rawls would enunciate this principle to institutionally exclude the results of moral arbitrariness: being talented from birth, for instance, is not a problem, *per se*, but it becomes so if used by social institutions to distribute primary goods. According to Rawls, principles of justice refer to "the basic structure of society, the arrangement of major social institutions into one scheme of cooperation" (Rawls 1999, p.47), considering equality not as a reward for bad luck but as a condition for making and maintaining a fair and stable liberal democratic society over time. The Rawlsian principles of justice apply to institutions, because institutions of a society should equally distribute primary goods, by guaranteeing equal resources and opportunities. Although Rawls considers the element of luck in human lives, his theory cannot be considered within that of luck egalitarianists. The latter claim that "the idea of the moral equality of persons requires that each person take responsibility for her choices and assume the costs of these choices. Conversely, it holds that no one should be worse off just because of bad luck" (Kok-Chor 2008, p. 665). This discussion applies at a grounding level, with the aim of answering why distributive justice matters. Justice as fairness, instead, would also want to deal with ethical justification, by proposing a method that a well-ordered society should undertake to fairly distribute primary goods, resources, opportunities to its citizens. Fairness is therefore a condition for holding stability within a liberal-democratic society. It should be stressed that, from the Rawlsian point of view, the moral principles that determine individuals' economic and social choices differ from those on which institutions should be based. The difference principle, for instance, is addressed to the basic structure, hence to institutions, and only indirectly to individuals. From this perspective, institutions should be

able to create general rules capable of guiding citizens' lives, while still maintaining people's pluralistic view, thoughts, beliefs.

In reaching the agreement, behind a veil of ignorance, on principles of justice, the specific element that Rawls singles out, as seen above, is the sense of justice. Before moving on to the next section, it should be underlined how the sense of justice can constitute a "theory of norm compliance" (Sacconi and Faillo 2008, p.1). Distinction must be made on why the sense of justice can be considered an element for norm compliance and, on the other hand, how it would be introduced in the model of conformist preferences by Grimalda and Sacconi (2002, 2005; Sacconi and Grimalda 2007). In fact, mostly in the Chapter IV, I assume this conformity model, by questioning the social norms' model by Bicchieri (explained in the next section). In particular, "Grimalda and Sacconi's model sees compliance as the consequence both of agents' participation in choosing the norm in a social contract setting 'under a veil of ignorance' and of the existence of expectations about reciprocal willingness to conform. Agents are characterized both by a standard *consequentialist* motivation, and by a *conditional willingness to conform with an ideal* normative *principle* of justice" (Sacconi and Faillo 2008, pp. 3-4).

Unlike Bicchieri's model, conformist preferences model underlines the intrinsic determinant for compliance, which is supported by the contractarian argument. An ideal normative principle of justice, deliberated *via* agreement, might provide intrinsic motivations for conformity: it would be a moral principle that has not only a normative value – what ought to do – but that becomes, because realistic, psychologically binding. By expressing their unanimous consent, agents undertake to respect the agreement because they have rationally justified the choice of one specific ideal normative principle over relevant others. Moral arguments, when realistic, in the sense that they can be brought into practice, might be considered as an intrinsic motivational force. Moral norms, therefore, should

be included as motivational factors leading to conformist behaviour when they are rationally justified, determining the conditionality of people's preferences.

3. Social norms and norms of justice

The problem of explaining normative behaviour, understood as acting for a reason, being guided by norms, emerges in those conformist behaviours in which following a norm means making a costly choice in terms of material selfishness and where rewards or sanctions are not incentives for compliance. As demonstrated by a large collection of experimental data, other-regarding preferences, sensitivity to established social norms, concerns for fairness, emotions and psychological dispositions intervene in the decision-making process (Sober and Wilson 1999; Sally 1995; Houser 2008). Among the various issues that this field of study raises, moral philosophy should be interested in "what determines individuals' agreement with the 'ought component' of norms" (Fehr & Schurtenberger 2018, p. 466), to see to what extent and under what conditions some moral principles (i.e., normative reasons) might constitute *real* reasons, hence motivations, for acting in accordance with what they prescribe.

John Rawls's Kantian constructivism might be a method through which the liberal egalitarian principle turns out to be a normative solution to a problem of distributive justice. That is, the method would determine the oughtness of liberal egalitarianism. As explained in the previous section, Rawls's objective is to propose principles of justice that can motivate and impose themselves on individuals' will, then compliance with principles of justice would occur because people are endowed with moral powers and a sense of justice and legitimized by the collective procedure of choice (i.e., the agreement). The norms of justice to which we can appeal, each corresponding to a certain ideal of fairness, are different, but all share a characteristic: they are context-dependent. On this aspect both Bicchieri and Elster agree in their respective attempt to distinguish a social norm from a moral one.

According to Elster (2006), the norms of justice can be considered as social, quasi-moral or as moral. This flexibility, on the other hand, is not admitted by Bicchieri (2006, 2010), according to whom the rules of justice are exclusively social (Dubreuil and Grégoire 2013). Bicchieri's conclusion seems to be too strict and it would exclude the critical capacity to review principles of justice (when they turn out to be unjust within a society).

It should be emphasized that terms such as 'norms', 'principles' and 'rules', regarding justice, are used interchangeably within this dissertation. In fact, they are considered in their broadest sense as informal rules that guide people's behaviour in observable social phenomena. According to Young (2007) there are three mechanisms of norm enforcement: the first is a reason for pure coordination as in the case of conventions established in each society – convention is here to be understood according to the definition by Lewis (1969)[7]. The second reason derives from the threat of social disapproval or punishment for norm violations. The third mechanism, the most complex, arises through the internalization of norms of proper conduct.

Focusing on this kind of norm enforcement and drawing from Lewis's (1969) and Ullmann-Margalit's (1978) thinking, Bicchieri's model of social norms holds that the motivation for abiding by norms of justice lies in a conditional preference most of people would have because they "acknowledge that the normative expectations are legitimate and should therefore be satisfied […] There is nothing inherently good in our fairness norms, above and beyond their role in regulating our ways of allocating and distributing goods and privileges according to the basic structure of our

[7] "Our final definition is therefore: A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P, (1) almost everyone conforms to R; (2) almost everyone expects almost everyone else to conform to R; (3) almost everyone has approximately the same preferences regarding all possible combinations of actions; (4) almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R; (5) almost everyone would prefer that any one more conform to R', on condition that almost everyone conform to R', where Rf is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of 5 among members of P could conform both to Rf and to R" (Lewis 1969, p.78).

society" (Bicchieri 2006, p.21). However, the content stated by a norm, to have practical implications, must be recognized and approved collectively by the members of a society. To function, to guide people's behaviour towards compliance, such a norm must have a content that binds people to comply. To be effective, norms of justice should hold together moral principles, thus guaranteeing the moral point of view, and on the other, capable of promoting the personal interest of individuals. It follows that excluding *tout court* morality from the definition of what norms of justice are, might undermine an understanding of why individuals, collectively, should be concerned about ideal fairness: in a nutshell, this is the objective of the research, which will be explained later with specific reference to the main chapters.

In this section, I will start presenting Bicchieri's model, then I will point out some criticisms and the need to reconsider why people comply with norms of justice.

As already anticipated above one of the major problems in defining what a norm is, is trying to state the conditions under which a norm may be considered moral or social (Konow 2003; Elster 2004, 2009; Bicchieri 2006; Nichols 2010; Dubreuil and Grégoire 2013; Chung and Rimal 2016). Among the various positions that have been developed, I focus on the one held by Cristina Bicchieri. According to her, what clearly differentiates moral and social norms is the type of behaviour they imply. The former would be the result of an independent behaviour while the latter of an interdependent behaviour. In other words, people would be moved to act according to a moral norm regardless of what others do or deem appropriate to do in a given circumstance. So, it is said that they have unconditional preferences for acting in accordance with their moral principle. For social norms, instead, the motivation would derive from the complex of conditional preferences related to what others do and expect me to do.

Beyond the potential problems that this model may raise on a philosophical level – note that Bicchieri's theory has a sociological aim – one of the aspects I am most interested in is why people should conform to a collectively recognized social norm[8]. At this point, the theory of categorization, schemata and scripts is introduced: Bicchieri assumes that "fairness, reciprocity, trust, and so on, are local concepts, in the sense that their interpretations and expectations and prescriptions that surround them with objects, people, and situations to which they apply" (Bicchieri 2006, p. 76). More specifically, "A given situational cue may not just be a helper a norm of fairness, but also a very specific interpretation of fairness. However, even when shared, fairness criteria such as 'give according to merit', or how to contribute to is, and so on" (Bicchieri 2006, p. 76). The argument lies in norm's activation as context-dependent. The categorization is that the cognitive process allows for "interpreting the social world, as it activates schemata (or scripts) that may be linked to personal stories of the social way" (Bicchieri 2006, p. 81). At the cognitive level, there is an established tendency to categorize people and events that have similar characteristics. Through this process, each person accumulates their own background of experience which will be fundamental to recognizing and predicting similar situations. It is a strategy to conform to what other people do and expect within the relevant group. Being able to recognize a certain type of social situation given the elements available should help us to choose the most appropriate behaviour.

This mechanism defines the theory of schemata, considered as "cognitive structures that represent stored knowledge about people, events, and roles" (Bicchieri 2006, p. 93). Since schemata referring to events are called scripts, Bicchieri's conclusion is that "social norms are embedded into scripts" (Bicchieri 2006, p. 94). How does this theory relate to the problem of compliance? Within this treatment of social norms, the adequacy of behaviour to what the norm establishes would occur

[8] For an overview on compliance of human cooperation see Fehr and Schurtenberger (2018).

because, given a recognized categorization of certain external stimuli, a particular script would be activated, leading people to react by default. It should be emphasized that Bicchieri does not deny the possibility of an autonomous and conscious deliberation (I consciously choose to adhere to a specific social norm) but she holds that, given past experiences and cognitive habit, people often conform to a social norm without questioning why they do it[9].

In reading Bicchieri's model what may be lacking is the identification of the role of motivation in deliberation: what reasons do people have for acting following a social norm? Her answer would be that, due to conditional preferences, normative and empirical expectations, conforming to a social norm is cognitively the default behaviour.

The evaluation of how a certain social norm should be considered has ultimate roots in morality, or at least in the consideration of what is right or wrong. To formulate moral judgments is connected with reasons for action, but the main issue is whether or not some moral principle might give sufficient and real reasons, that psychologically lead individuals not only to share (to give consent) to the rightness of a certain principle but also to comply with what the principle in itself prescribes.

Going back to the main research question of this dissertation, my argument is that, in order for subjects to be driven towards compliance with the liberal egalitarian principle in situations of production, an explicit priori agreement that rationally justifies the principle is necessary. I purport to show that it is the agreement – through the constructivist procedure – what gives agents a reason to act in conformity with the agreed principle. I will focus on a source of motivation that is overlooked

[9] In psychology the brain 'division' between system 1 and system 2 is defined: the first would provide impulsive, fast and automatic answers (for example because a scheme is activated, it is likely that I will behave as I used to in the past, without asking what reason I have). The second, on the other hand, is the one associated with cognitive calculus, where alternatives, possible consequences, and reasons are considered before choosing a course of action (Kahneman and Tversky 2012).

by Bicchieri, that is the moral obligation stemmed from a special kind of commitments, namely joint commitments.

4. Methodology

The Encyclopedia Britannica (1991, p. 395) presents this view: “Economists are sometimes confronted with the charge that their discipline is not a science. Human behavior, it is said, cannot be analyzed with the same objectivity as the behavior of atoms and molecules. Value judgements, philosophical preconceptions, and ideological biases must interfere with the attempt to derive conclusions.” This statement, reported in a footnote in the first chapter of *Experimental Economics* (1993) by Davis and Holt, brings up the criticism in considering economics as an experimental science. The lack of objectivity of its subject of study would preclude the experimental approach to a science that should devote itself to speculation. Unlike what is enunciated by the Encyclopedia Britannica, the fact that value judgments, philosophical preconceptions, and ideological biases intervene in the decision-making process makes relevant the use of experimental methods also in Economics. The behavioural turn has strongly shown how psychological factors (therefore beliefs, cognitive biases, moral values, norms) have a decisive weight in the formation of intentions and in the corresponding actions.

Experimental economics sees in laboratory experiments a tool for examining “the willingness of individuals to overcome collective action problems” (Ostrom 2000, p.139) and its approach consists of three main strands in literature: the first one deals with market experiments focusing on the predictive value of the neoclassical price theory; the second focuses on game experiments, run in a setting similar to those of natural markets; the third focuses on individual decision-making experiments, in which “the only uncertainty is due to exogenous events, as opposed to the decisions of the other agents” (Davis and Holt 1993, p.5). In particular, “interest in individual decision-making experiments grew from a desire to examine behavioral content of the axioms of expected utility

theory" (Davis and Holt 1993, p.5). Along with this third trend of literature, we find a more specific range of experiments concerning distributive justice, within which this project is situated. When questions are asked about how people form a certain ideal of fairness and what they deem as such, it would be necessary to mutually draw between theories and facts to see if the theories actually predict how people will behave and, on the other hand, to check whether the multitude of data collected might favour an interpretation rather than another. A large group of game experiments arose with the intention of identifying what moves agents to act in contexts of economic choice, testing the validity of the expected utility theory. The results showed a problem in that theory, insofar as the real behaviour deviated from the predictions. After the forerunner games – ultimatum game, prisoner dilemma, dictator game – several collective variants were proposed such as trust game, public goods game, exclusion game, where the main objective was to analyze how individuals decided if faced to social dilemmas – conflict between private and public interests. From the quantity of data collected emerged the need to propose models that would include other factors, in addition to the classic utility, to be able to propose a function that had real predictive value for human behaviour.

This is not the place for a discussion on the pros and cons of the experimental economics' methodology (for this reference to Guala 2005), but I would like to justify its use for the purposes of this project. In fact, it presents a part of experimental evidence, developed to show how an interdisciplinary approach is required when the object in question is human behaviour, or rather, how people choose under certain circumstances. We are increasingly noticing how a set of social behaviours can be defined as normative: that is, they are behaviours that are guided by the existence of norms, in a broad and informal sense, such norms would give good reasons to act. This aspect is worth addressing a little more, because in it lies the justification of using the methodology of the experimental economy for a research that has a philosophical objective. When we say that there are

norms that give reason to act, on the one hand we refer to the so called normative reasons, those precepts that require an impartial point of view (as an external observer) and, on the other, there are motivating reasons, those that give people real reasons to act (one speaks therefore of first-person perspective).

One purpose of ethics, among others, is to define which moral norms can actually influence our behaviour and whether they can really bind human will to perform certain action (see state of the art, section I). However, questioning and providing plausible answers is not enough: to see what moves people towards specific choices and related actions, we must also investigate, on a psychological, level what cognitively happens. This framework expands and becomes more interesting when, instead of just stopping at the individual decision-making, one moves towards problems of collective action. In order to solve these problems, the parties in question will have to collectively choose a course of action – thus forming a collective intention to do something together – and then they need reasons to behave in conformity with the collective intention. It is within this question that the experiments presented in this project are situated: how subjects choose individually and how, instead, they collectively agree on a distributive principle; towards which ideal of fairness (i.e., pure egalitarianism, libertarian, liberal egalitarianism) they favour; what reasons they have to comply with the chosen norm; whether or not the joint deliberation procedure changes the taste for fairness. These are some of the objectives that the project proposes, picking out the methodology of the experimental economics as a useful tool for controlling variables and corroborating (or rejecting) hypotheses.

As Stevens (2018, para. 722) says "The assumption that agents are motivated by social norms has implications for both positive and normative theory. In particular, social norm preferences can easily be added to the utility function of economic agents to both describe what they actually do and prescribe what they should do in an economic setting".

Therefore, a research that takes seriously the reasons why people consider a norm to be true in this respect cannot be limited to a descriptive analysis of behaviour. If a norm – which says what one ought to do – is at stake, it is likely that it is considered as a salient feature in the reasoning. Then the data, however descriptive, bear a strong normative element which remains silent. Identifying how and why certain type of behaviour can be called normative requires different but complementary approaches such as both the methodology of the experimental economics and the criticism of philosophical reflection.

III. *Does impartial reasoning matter in economic decisions? An experimental result about distributive (un)fairness in a production context*[10]

1. Introduction

This paper reports an unexpected experimental result and contributes to the debate about other-regarding vs self-interested behaviours in noncooperative games.

The experimental approach was intended to check the impact of 'reasoning behind a veil of ignorance' on an individual's distributive decision. By making subjects think on the distributive decision they would be asked to make *before* they knew the one relevant difference that the experimenter was going to introduce between the two persons that should split the earnings, this experiment was designed as a partial test of this feature of Rawls's theory. The veil, intended as a moral cue, should have induced a reflection from an impartial perspective, leading subjects to put themselves in the shoes of the least advantaged person once the veil would have been removed. Drawing on the philosophical assumptions of Kantian constructivism, we supposed that the mere conception of a distributive choice behind the veil of ignorance should create a normative stance towards the actual distributive decision, even if experimental subjects were not acting within the rules of any known institution. And we further supposed that subjects' behaviour, exposed to a clear moral cue in distributing a common output, would adjust in line with their normative conclusion.

In fact, although the Rawlsian original position and the veil mechanism are abstractions, not traceable in everyday life, they are useful tools for studying individual behaviour, in situations of potential

[10] This chapter reproduces a paper published in THEORIA, An International Journal for Theory, History and Foundations of Science, DOI: http://dx.doi.org/10.1387/theoria.21011 .

conflict between personal interest and the common good (see Faillo et al 2015). It is a question of understanding whether there is a correspondence between Rawls's theoretical apparatus on distributive justice and how, *de facto*, people behave in distributive contexts. The first empirical studies on Rawls's theory (Frohlich et al, 1987; Frohlich and Oppenheimer 1990, 1992; Lissowski et al, 1991) showed how the maximin principle defended by Rawls has no counterpart in the laboratory: subjects prefer to maximize the income after having established a floor constrain. Other results followed from questions on how subjects perceive the principles of allocation: Scott et al (2001) and Michelbach et al (2003) distinguished four principles of allocation - equality, efficiency, need, and merit - demonstrating how these principles often intervene simultaneously and in an interdependent manner so subjects can formulate judgments on distributive justice.

In parallel with this literature on how different allocation principles intervene to form judgments on distributive justice, there exists an ever-growing experimental literature on non-cooperative games. These experiments mostly investigate underlying motivations in giving behaviour. The assumption in experimental economics is that to give away money is costly for the individual and therefore it should be expected only within a framework of social, institutional or moral obligations coercively imposed. However, in some economic games, such as the Dictator, people seem to share for no reason whatsoever.

Some factors have been detected to explain why a fair distribution got the drop on the rational choice theory predictions: normative and empirical expectations established by social norms amongst players (Hoffman et al, 1996; Bicchieri 2006; Krupka and Weber, 2013); the reputation effect – subjects feel observed, hence judged, and they do not want to contradict their self-image (Servátka, 2009); the frame itself, conveyed by the game. In relation to the frame effect: if the taking option is also provided in a Dictator's game, subjects may perceive choices as conflicting messages, deciding

to take from the other person because it is one of the available alternatives (List, 2007; Levitt and List, 2007; Barsdley, 2008; Franzen and Pointner, 2012). Rigdon and colleagues (2009) used even a minimal social cue of a visual type, called watching-eyes configuration, demonstrating how its introduction had an impact in terms of increasing generosity. Another very important element is the endowment, or better, how it accrues to the subjects: if it is "manna from heaven", namely offered by experimenters at the beginning of the game, or whether it is the product of subjects' effort, behaviour on the final distribution changes (Faillo et al, 2019).

Given this experimental literature, the actual game, explained in detail below, was a variation of the Dictator Game[11] (DG) that was initially used to question the assumptions of economic rationality. On average, people give between 30% and 40%, which seems to show that ordinary people are not as selfish as economic theories assume. The introduction of Dictator with Taking changed this view, since in this version of the game, money is given to *both* players, and then only one of them (the dictator) has to decide whether to keep her portion, give part of it to the other player, or take some from the other. On average, subjects in this game *take* from their pairs, as pointed above, what seems to be a very selfish behaviour.

In our case, the experimental currency was earned by both players through a real-effort task rather than simply given to them by the experimenter, and each member's contribution to the pair's total output was common knowledge between them. The experiment introduces an asymmetry among players so that one member of the pair suffers an unjustified 'disadvantage' relative to the other. We expected our subjects to be aware of the need to redress the unjustified inequality. Our prediction was

[11]An experimental situation in which a subject is asked to split an amount of money, generally given by the experimenter, between himself and a second subject, who plays no role in the game, but simply receives what the 'dictator' decides.

that whatever the average distribution was in the base treatment, introducing 'reasoning behind the veil' would move the average distribution towards a more 'liberal egalitarian' pattern.

In this framework, the liberal egalitarian perspective consists in a type of distribution that should compensate for the initial disadvantage – in our case, a shorter time limit to perform the task, which is a mere random element. The assumption is that those with more time available could produce more and therefore claim a 'right' to a greater income. So, to repeat, the main objective is to verify the effectiveness of the veil as a moral cue: without a veil we expect that subjects would have chosen a distribution based on merit (everyone gets what she/he has produced); while introducing the veil, we expect that subjects, realizing that the advantageous position will not be the result of merit but of mere fortune, should seek to level this inequality during the distribution stage, by introducing some re-distribution from the advantaged party to the disadvantaged one. In fact, reasoning behind the veil should lead each person to think of the possibility that she is the disadvantaged party and, wishing to protect herself from undeserved bad luck, establish a redress mechanism.

By isolating this element, we expected to gather empirical support for the Rawlsian contractual argument: if distribution changes after introducing deliberation behind the veil, this would imply that the original position story does track the constitution of our moral intuitions and norms regarding distributive justice. We had two arguments to support our working hypotheses: first, reasoning behind the veil of ignorance should direct our subjects to a generally egalitarian split of the cake; second, the fact that the thought experiment implied a normative reflection should make the rules of property and merit that are part of common morality even more salient.

We thought that the reasoning behind the veil meant introducing a moral cue within the decision-making process, given the previous experiments in moral psychology and behavioural ethics. A strong line of research in this area has worked particularly on the underlying dynamics about people's

honesty and dishonesty. Some results show that trying to behave honestly is perceived as an effort, as a practice that is not automatic but requires strength of will and commitment (Aquino and Reed 2002). Others point out that, on the one hand, people tend to justify their dishonest behaviour (Shalvi et al 2015), but, in many situations, the self-concept of being and perceiving themselves as moral persons is a motivation that leads people to be more honest (Mazar *et a.*, 2008). Ayal and colleagues (2015) have shown how the use of certain moral cues have an effect on human actions even in conditions of anonymity. One of the moral cues they used is called 'reminding', which, as confirmed by data, has been salient in influencing behaviour "utilizing principles of right and wrong, specific examples of morals and 'do's and don'ts', and even slogan" (Ayal et al 2015, p. 739).

So, we expected the thought experiment to measurably move whatever result we obtained in a baseline design (with no veil of ignorance, and no moral cue) towards a more equal split. However, the predicted move towards egalitarianism was not observed. This left us puzzled. We revised our procedures, method and assumptions –see discussion sections below– and found no significant flaw. We conjecture a philosophical explanation for our result, namely, that the *mere* use of a moral theory as a thought experiment is not enough for eliciting the behaviour prescribed by the theory. This is a major amendment to Social Contract theory –and in general to moral theories that rely on counter-factual reasoning. But we need to be careful with our conclusion; since this result is so opposed to literature on moral cues, further research is surely required.

The paper is structured as follows: the next section describes the experimental design and hypotheses in detail; results are presented in section three; section four summarizes and discusses our findings and explanatory conjecture.

2. Experimental design and hypotheses

In this study we compare two treatments, one called NOVEIL, which constitutes our baseline, and a second one named VEIL. In both treatments, (i) participants are grouped in pairs (ii) the endowment was earned (by the pair) through a task; (iii) each member of the pair was randomly assigned different time limits (10 or 6 minutes) to perform their task, therefore making almost sure that their 'contribution' would be different, due in great part to a chance event (whether they have 6 minutes or 10 minutes) that happens *before* the task begins and the earnings are collected; (iv) participants played a DG in which they can just keep their earned endowment, give part of their endowment to the other or take a part of the other's endowment; in other words, they can distribute the pair's total earnings as they wish. The underlying assumption behind condition (iii) is that the person with more available minutes has an advantage, *and* a corresponding responsibility; the foreseeable larger contribution of the person with ten minutes would not simply be the effect of chance, but the combined effect of chance *and* additional work on her part. This situation purports to represent the most common social distributive problems –those that are solved through liberal-egalitarian principles.

Let us briefly note that the use of time as a basic resource to represent initial inequality has no precedent in experimental literature. The reason may be that more time may imply an extra effort, and therefore it may be taken to represent a disadvantage rather than an advantage. In this experiment this is not the case, since the task is easy enough, and the working time short enough in any case. We are sure –in part from the debriefing questions registered after the experiment– that having more time was invariably interpreted by the subjects as having an advantage. Other forms of representing initial inequality could have been used; but we found that working time required less intervention in the

design of the rules of the game and was easily identified as a difference that happened *before* the task itself began.

In our baseline treatment (NOVEIL), subjects were randomly assigned computer-cubicles and they were anonymously paired with someone else in the room. They were informed that they were going to perform a task –which was coding words, translating them in numbers by using a conversion table– for which one member of the pair would have six minutes, and the other member ten minutes. They were informed that right before the task began, they would know whether they have six or ten minutes: this would show on their monitors. When they were done, they were informed about each member's performance (total number of words coded and productivity measured as words *per minute*). The same list of words was presented to every participant. Subjects were paid in experimental currency called token. At the end of the experiments tokens were converted in Euros at the exchange rate of 1 token = €0.20. They received one token for each word correctly coded. At this point, each member of the pair had to decide how to divide the total output of the pair by claiming any percentage for herself. Once both members of the pair made their choices, one member was randomly selected and her choice was implemented. So, each participant decided under the expectation that her choice had a 50% probability of being her real final payoff. There were no further rounds, and each subject participated just once. Once the member with six minutes consumed her task-time, she would be playing a game unrelated with the experiment and with no effect on her payoff.

The VEIL treatment was the same as the NOVEIL except that, before the time limit was assigned and the subjects proceeded to the task and decision phases, they were asked how they think the total output should be divided, by stating how much (what percentage) they should claim for themselves. At this moment they knew the details of the game, but they ignored whether they will be given six or ten

minutes, so they decided behind a veil covering the information about their labour time. Finally, the subjects knew that their choice at this time had no effect on their final choice after the task. This *ex-ante* phase lasted for two minutes, which we calculated is plenty of time. We deliberately gave them time to spare. The goal was to make them think; we made clear that they could change their option any time until the questionnaire closed. Their choice was recorded, and then they proceeded as in the previous treatment: they were informed about the time assigned to each of them and they worked on the task. After the task, they could confirm their choice behind the veil, or change it.

In both treatments, instructions were read aloud by one the experimenters and a set of control questions were proposed to make sure that participants understood the instructions. Participants were students recruited at the University of Granada in May 2013. 50.6% of the participants were females, the average age was of 22.27 years, 96% were Spanish, with an average number of previous experiments of 2.13 (max=9); 29% were enrolled in the Economics program and 61% in the Management program; 10 were enrolled in different programs (see table 1 for the data on the two treatments' samples). We conducted 4 sessions of 20 participants for each treatment, for a total of 80 participants. The average of payments was of €11.30 (included a show-up fee of €3). The experiments were run at the Egeo Lab (University of Granada). The experiment was programmed by using the z-Tree platform (Fischbacher, 2007). We used a between-subjects design; no subject participated in more than one treatment.

Table 1. The characteristics of the samples (standard deviations in parenthesis)

	NOVEIL	VEIL
Age	22.3	22.3

	(1.83)	(2.02)
Gender (% of female)	46.2	55
Nationality (% of spanish)	96.2	97.5
Major (% of economics and management students)	87.5	92.5
Number of experiments in which the subject took part.	2.15 (1.46)	2.12 (1.42)

We were interested in testing the two following hypotheses, respectively related to treatments:

Hypothesis 1: Agents in treatment NOVEIL should choose a distribution that tracks individual earnings.

It should be a cost, in moral terms, to steal from another who has earned a certain amount of money by working – namely the taking option. In addition, if social norms (about work, effort, and desert) carry over to the laboratory, they should weigh towards this distribution. In other words, even if agents find themselves in complete anonymity and may fear no punishment, the design of the experiment seems to call for them to keep what they have earned and leave to their partner what (s)he has earned.

Hypothesis 2: Agents in the VEIL treatment should choose a distribution that approximately tracks a liberal-egalitarian principle.

Agents in this treatment are subject to the thought experiment of the veil of ignorance. This provides an impersonal and impartial point of view that elicits a fairer and less selfish behaviour: each individual is drawn to think as if (s)he could have more or less endowment, with no reason (randomly). From this each participant is aware that the design involves an unjustified inequality; and they have the power to 'correct' it by choosing to distribute the common output in a more egalitarian way –even if, since they have data about individual productivity, we never expected convergence on pure egalitarian distributions. From a moral reasoning perspective, our hypothesis 2 means that the Rawlsian assumption about the moral capacities of people prevail over the Hobbesian assumption; in other words, the hypothesis implies that moral reasoning and moral conclusions would have an effective power to shape behaviour even in absence of external public authority.

3. Results.

Starting from the evidence on the task, we run a two-way ANOVA to assess the effect of treatment and time on the total number of words encoded (table 2). We do not observe any significant difference between treatments in terms of amount of words encoded ($F(1,156)=0.00$, $p = 0.97$) while, as expected, the production of the participants with ten minutes is higher than that of participants with six minutes in both the treatments ($F = 333,3$, $p < 0.01$; interaction of treatment and time: $F = 0.00$, $p = 0.96$). Participants received one token for each word correctly encoded, so the numbers in table 2 correspond also to participants' average earnings. The level of productivity (words per minute) is 5.15 for the participants with 6 minutes and 5.26 for those with ten minutes in both the treatments (time: $F = 0.60$, $p = 0.44$; treatment: $F = 0.00$, $p = 0.97$; interaction: $F = 0.00$, $p = 0.99$).

Table 2: Production (correctly encoded words).

		Time	
		6 minutes	10 minutes
Treatment	NOVEIL	30.90	52.62
		(5.82)	(8.45)
	VEIL	30.92	52.67
		(5.37)	(9.62)

Means, standard deviations in parentheses.

We can then put forward our first result.

Result 1: Production and productivity.

The production of participants with the same endowment in terms of time is the same across treatments. In both treatments, the level of production of participants with ten minutes is higher than that of participants with six minutes. There are no differences in the level of productivity, neither between treatments nor between subjects with different time groups.

Table 3. Percentage and number of tokens (percentage x total production) claimed after the task.

		Percentage		Tokens	
		6 minutes	10 minutes	6 minutes	10 minutes
Treatment	NOVEIL	74.50%	77.50%	62.29	64.35
		(19.20)	(17.05)	(19.10)	(15.17)
	VEIL	74.00%	85.75%	61.59	71.87
		(19.18)	(13.93)	(18.12)	(16.10)

Means, standard deviations in parentheses.

Table 3 (columns 3 and 4) reports the number of tokens claimed by the participants after the task, obtained by multiplying the percentage claimed by the total production of the pair. In the NOVEIL, the amount of token claimed is significantly higher than the amount of tokens earned with the task, for both participants with six minutes and participants with ten minutes (Wilcoxon signed-rank test, participants with six minutes: $z = 5.51$, $p < 0.01$; participants with ten minutes: $z = 3.99$, $p < 0.01$).

Result 2: Individual production and claims in the NOVEIL treatment.

In NOVEIL treatment participants' claims are significantly higher than their individual production, independently on the time available for the task.

This result is related to hypothesis 1. We certainly expected a better fit between working-time and claim (claims from subjects with six minutes should be approximately 60% of the claim of subjects with ten minutes), and less distance between production and claim –or less 'stealing' from partners. However, the results are not wholly unexpected. They do support our hypothesis 1.

As for hypothesis 2, the implementation of a liberal egalitarian principle, aimed at reducing the inequality due to pure luck, would have resulted in participants with six minutes in the VEIL treatment asking more than those in the NOVEIL, and participants with ten minutes in the VEIL treatment asking less than those in the NOVEIL.

Looking at the data, we observe that claims by participants with six minutes in the two conditions are not significantly different (Wilcoxon rank-sum - Mann-Whitney test, on percentages: $z = 0.19$, $p = 0.85$; on number of tokens: $z = 0.10$, $p = 0.91$). As for participants with ten minutes, those in the VEIL

ask *even more* than those in the NOVEIL (on percentages: z = 2.19, p = 0.03; on number of tokens: z = 2.11, p = 0.03).

Result 3: Ex post claims across treatments.

Participants with six minutes in the VEIL treatment make the same claim of those in the NOVEIL. Participants with ten minutes in the VEIL treatment ask for more than those in the NOVEIL.

Based on result 3 we reject hypothesis 2.

Table 4. Percentage claimed before the task in the VEIL treatment

	Tokens
Six minutes	73.50
	(17.76)
Ten minutes	79.50
	(16.78)

Means, standard deviations in parentheses.

In the VEIL treatment, before knowing the distribution of time within the pairs, subjects were asked to choose a percentage of the total production they may claim after the task. Table 4 reports the average percentage chosen by the subjects, distinguishing between those who later would have been assigned six minutes and those who would have been assigned ten minutes. We observe that while *ex post* claimed percentages by participants with six minutes (table 3, raw 2, column 1) are not statistically different from their claims *ex ante* (Wilcoxon signed-rank test, participants with six minutes: z = 0.26, p = 0.79), *ex post* percentage claims of participants with ten minutes (table 3, raw 2, column 2) are even higher than their claims *ex ante.* (z = 2.97, p = 0.03).

Result 4: Ex ante and ex post claims in the VEIL treatment.

In the ex post phase of the VEIL treatment, participants with six minutes confirm their choice ex ante, while participants with ten minutes claim more than what they have judged as the right claim ex ante.

Table 5 reports the results of an OLS estimation in which we control for socio-demographic characteristics and participants' experience with experiments. The results reported in column 2 confirm that claims of the subjects with ten minutes in the VEIL are slightly higher than those of subjects with 10 minutes in the NOVEIL. Interestingly, females claim less than male but only in the VEIL treatment, and subjects with less experience with experiments claim less than more experienced ones only in the NOVEIL treatment. The results in column 2 confirm the correlation between claim ex-ante and claim ex-post in the VEIL treatment, which is weaker for the subjects with ten minutes who tend to ask more than the ex-ante claim with respect to the subjects with six minutes.

Table 5. Determinants of individual claims after the task

	(1) OVERALL	(2) VEIL TREATMENT
AGE	-0.161	0.138
	(0.720)	(0.611)
GENDER	-0.902	-4.185*
	(3.890)	(2.485)
NATIONALITY	6.256	-2.880
	(8.206)	(8.343)
MAJOR	-2.779	0.891
	(2.313)	(2.269)
N. OF EXPERIMENTS	3.647***	0.318
	(1.328)	(0.735)
TEN	1.400	30.08***
	(3.882)	(11.35)

VEIL	10.98*	
	(5.806)	
TEN*VEIL	10.31*	
	(5.531)	
GENDER*VEIL	-7.932	
	(5.554)	
N. OF EXPERIMENTS*VEIL	-3.605**	
	(1.755)	
CLAIM EX-ANTE		0.876***
		(0.101)
CLAIM EX-ANTE * TEN		-0.289**
		(0.143)
CONSTANT	70.41***	9.028
	(19.88)	(17.72)
Observations	160	80
R-squared	0.157	0.681

OLS regression.
The dependent variable is the percentage claimed after the task (in the case of the VEIL treatment it corresponds to the ex-post choice)
Results in column 2 refer only to the subsample of subjects who took part in the VEIL treatment.
GENDER takes value 1 if the subject is a female and zero otherwise.
NATIONALITY takes value 1 if the subject is Spanish and zero otherwise.
MAJOR takes value 1 if the subject is enrolled in an economics or management program and zero otherwise.
N. OF EXPERIMENTS is the number of previous experiments in which the subject took part.
TEN takes value 1 if the subject is endowed a time of ten minutes and zero otherwise.
VEIL takes value 1 if the treatmentis VEIL and zero otherwise.
CLAIM EX-ANTE is the percentage claimed in the ex-ante phase of the VEIL treatment.

Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.10

In conclusion, our data support only hypothesis 1.

IV. Summary and discussion

In the experiment reported here a majority of young students stole from an anonymous peer. They did this while knowing that the earnings that each had contributed to the total to be split was obtained through a work they performed –a work that while not particularly hard, was not entirely effortless. Even though the game was anonymous (they did not know from whom they were taking the money), it was a very homogeneous population of students. This fact would have made us predict a higher degree of empathy. According to Binmore (2005), this kind of preferences are used to find the fair solution to coordination problems. In Seabright's (2005) reading of *Natural Justice*, we find condensed into a very effective phrase the function that empathetic preferences should play: "individuals must have empathetic preferences – when imagining themselves in the situation of others, not their own". Furthermore, "morality serves to coordinate between the possible equilibria of social life", so the moral rules internalized by evolution should help us to resolve potential situations of social inequality. The introduction of the veil, applied to an individual rather than a collective choice, would have to direct the behaviour to a fair split, taking into account the initial fortuitous disparities. Considering the role that moral principles we have internalized usually have in the process of deliberation, we hypothesized the prevalence of the social norms about effort and merit, the unwritten rules about earnings and possession; we hypothesized further that a reflection towards fairness, reinforced by the individual position of ignorance at the time of the thought-experiment, would help bending individual decisions towards equality. We were not particularly optimistic about the effect of these rules on the average behaviour of our experimental population, but we definitely expected at least some departure from standard results in DG with taking, reflecting the role of effort and the workings of the veil of ignorance. To our dismay, the only significant difference between

baseline treatment and treatment VEIL was contrary to our hypothesis: subjects with ten minutes-time asked, on average, more than their counterparts in the baseline treatment. The rejection of our hypotheses left us baffled and wanting an explanation.

Let us note first that this result may be related to other experiments in which the most advantaged subjects were reluctant to give. These experiments usually involve effort and luck (Erkal et al 2011, Rey-Biel et al 2018; Konow 2000). However, in our design, having more time to perform the task does not imply merit, effort or recognizing some specific abilities the player has. The veil of ignorance should have functioned to highlight that the initial disparity was not due to a difference in ability amongst subjects but to a merely external and random element. As stated by Erkal et al (2011, 3332) "in real life, earnings are determined not only by effort, but also by luck. Evidence suggests that giving behavior might be different when luck affects earnings and that people are more likely to receive support when they have been negatively affected by luck (Christina M. Fong 2007)". Nevertheless, when subjects believed that they have earned a certain reward resulting from an effort, considerations on fairness ideals and on how material goods should be redistributed would not have a significant impact.

We have reached the conclusion that the setting in which the game was made – marked by anonymity and non-iteration– created a context in which there was a lack of motivation to act in other-regarding ways. As Eckel and Grossman emphasize:

"The decision makers cannot identify each other, nor do they have enough information to know if their partner is poor or otherwise deserving of their generosity; thus there is little or no basis for altruism to play a part in the decision. Furthermore, as subjects cannot be identified by either the experimenter or other subjects, there is no role for social esteem to affect the decision. Only self esteem (or warm glow) remains. With little motivation for other-regarding behavior, it is

not surprising that the subjects' behavior closely approximates the game-theoretic predictions for noncooperative, non repeated games with selfish, payoff-maximizing subjects" (Eckel et al 1996, p. 182).

However, some features of our design could make us think of fairer behaviours: the effort needed for production –as opposed to the 'manna from heaven' situation common in other experiments– and the veil of ignorance at the individual level proved completely irrelevant. The veil of ignorance should have had an effect in the final distribution given that subjects were put in a condition of reflecting on the different possible scenarios at the end of the game. If everyone had an effective sensitivity to moral cues, the fact of allowing subjects to think in advance what percentage of their product to bring home and which one to offer should have fostered a fairer behaviour in equity terms.

A first explanation could be related to a misunderstanding of the experiment by the subjects. However, this possibility was ruled out from the start: not only we did run check questions after reading the instructions, but the subjects were well versed in the workings of experimental sessions, and most of them had some training in economics or management. If anything, our subjects' 'mistake' was to grasp the experiment all too well. They saw the true nature of the choice at their disposal and were not influenced by a setting that involved working in pairs, or pooling together earnings, or making a distributive choice. They understood that at the end of the day, what they were going to get was determined simply by their choice –multiplied by the 0.5 chance of being actually selected for final payoff– and they responded wisely –in economic terms. They were, on average, very good monetary payoff maximizers.

The fact that our subjects were students, belonging to the same group of reference, could have caused gamesmen behaviour, because the context would not be perceived as a situation in which real interests are at stake. This interpretation is in line with the results of Hoffman et al (1996): they saw that "when

a double blind procedure, intended to guarantee the complete social isolation of the individual's decision (no one including the experimenter or any subsequent observer of the data could possibly know any subject's decision), was used, 64 percent of the offers were $0 with only 8 percent offering $4 or more" (Hoffman et al, 1996, pp. 653-54).

On a related line, Fehr et al (2006) found that students attending majors in Economics and Management are less sensitive to egalitarian concerns. However, this would not explain the surprising result that our ten-minute subjects claimed on average *more* in the treatment with a veil. We were aware of this possible effect, so we were not surprised by the rate of 'stealing' in the baseline treatment. But the null effect of the individual reflection behind the veil still wants an explanation.

Another possible reason for the observed proficiency in maximizing behaviour combined with their utter disregard for social norms or lack of pro-social preferences might be that the subjects were recruited over a platform Orsee: people included in the platform were 'used' to do economic experiments, so it may be thought that they are sophisticated economic choice-makers. However, we have excluded also this explanation because the recruitment of experimental subjects through registration on platforms is a widespread and established method to conduct experiments in the laboratory.

One possible explanation might have to do with the problem of stability of intentions, as it applied to the stability of the principles chosen behind the veil. The subjects might have been sensitive to the moral cue, and correctly concluded that the final distribution had to be re-distributive, and still fail to follow this conclusion in practice (cfr. McClennen 1989, Klosko 1994, Barry 1995). There is no logical implication that choosing a distribution criterion behind the veil will bring about a corresponding behaviour. The problem of stability can be read as the problem of compliance: why

should subjects act according to a certain ideal of fairness –even if it is an ideal they chose as their preferred option before–, when they could obtain a greater gain by deviating?

There are two reasons why this line of reflection is not wholly satisfactory as an explanation. First, our design excludes the interpretation of the *ex ante* choice as a plan. Our subjects know from the beginning that their final choice is up to them and that they will decide only after they now all the facts (whether they had six or ten minutes, how much the pair produced, etc.). So there is nothing in the design implying that their first choice is a kind of plan they have to follow or rule they have to abide by. The initial choice should work as a mere moral cue, that is, a as a reminder of the fact that there was an unjustified inequality, and that they had the opportunity to re-dress for it.

Second, the issue of stability of principles is quite complex, involving psychological conditions for the stability of intentions and conceptions of rationality and morality. The contribution of this paper is to clarify one aspect of that dynamic, and the experiment abstracts as much as possible of the deeper philosophical issues related to the problem of stability. Our hypothesis tried to relate the veil of ignorance as moral cue with an adjustment of individual behaviour in a prosocial and egalitarian direction.

The difficulties with this explanation do not imply that our result is irrelevant to the debate of stability. After all, the experiment shows that whatever thoughts are derived from reasoning behind the veil do not carry over to the real choice situation. But from our data we cannot conclude whether this is a problem of stability or a problem of irrelevance: we cannot know whether our subjects correctly concluded that initial inequality called for re-dress but then failed to make a re-distributive choice, or they, also correctly, found that reasoning behind the veil was irrelevant in the situation.

Upon reflection, the puzzling fact of stealing from peers should not be surprising. Note that the final distribution of money (how much one gets) depends on three factors: effort (own, and one's pair), decision, and chance (being selected for final payoff). After all, this is what a lottery is: working people voluntarily bet part of their legitimate earnings and subject themselves to a chance event. They know that the most likely outcome is that someone else will get that money. And in the event that they win, they experience no remorse in taking money from other working people, maybe poorer than them. And this is because the whole practice is voluntary in the first place and based on free choices. Our subjects might have approached the experiment in this 'gambling mood' and this would explain the result.

The problem is that this explanation would be applicable to virtually all economic experiments involving populations of students with some previous experiments in their records.

It might be the case that, by making the experiment one-shot and anonymous dictator's choice –no chance of reciprocity, no need to regard the other member of the pair's attitudes, believes or dispositions– with the conspicuous absence of punishment by any kind of authority, we gave our subjects a true Gige's ring to act as they pleased[12]. And our finding is that they behaved as Trasymachus and Hobbes would have predicted, rather than the way Socrates, Kant or Rawls would.

In conclusion, two interesting elements emerge from this study: the effectiveness of the veil of ignorance and the actual behavioural adherence to moral principles given the social dimension in which we are immersed. We wish to comment on each in turn.

Regarding the first point, this study shows that the veil of ignorance, as moral device, has a low effect on individual choice when used as counter-factual *ex ante* thought experiment. This is in contrast with experiments where the choice behind the veil was the result of an actual agreement among parties

[12]Cfr. Plato, The Republic 359d.

(Sacconi and Faillo, 2010; Schildberg-Hörisch, 2010; Faillo et al, 2015). In these cases, impartiality and impersonality – conveyed from the veil – had an impact on participants' behaviour. So, it would be interesting to study the conditions under which the veil might work, inducing pro-social behaviour, on individual choice – there are many occasions in our everyday life where we have to decide how to behave, morally or not, without prior agreements. In general, moral philosophy tends to make use of counter-factual reasoning. It may be argued that counter-factual arm-chair philosophy is not intended to *motivate* action, but to justify, or explain. However, criticism about the lack of motivating power was taken very seriously by leading contractarian philosophers, like J. Rawls or D. Gauthier. While acknowledging that their theories, being rationalistic and laid out in analytic language, where not particularly enticing, they did defend that, properly understood and once the distance between ideal theory and non-ideal social conditions are taken into account, they should give people reasons to act as they prescribe. We assumed that reasoning behind the veil is a kind of toy moral theory for a particular case. Given the setting of the experiment, rational people should get to the conclusion that equal –or approximately equal– split would be more justified. They generally did reach this conclusion. Now, the fact that they did not take this conclusion as a sufficient reason, while actual agreements reached behind the veil do seem to create new normative reasons for action, is enlightening. It is a result that has potential interest in institutional settings. It speaks of the relevance of the social in inducing real cognitive and motivational changes in individual agents.

Secondly, these results might contribute in defying how distinguish social and moral norms, when such norms are collectively recognized, what they include and how they are perceived by the decision makers. Although far from our original intent, what emerged seems to be that moral reflections are not so binding on behaviour. Observability, and the consequent awareness resulting from a choice made in public, is that fundamental feature present in a social dynamic, but lacking in a moral one,

in which the risk of social punishment or exclusion is high. Not only that, observability allows each subject to show himself in a certain way – appearing right rather than being right – because in social contexts we often want to give a certain image of ourselves and then do not betray it, neither in the eyes of others nor in the ones of ourselves. This would be a confirmation of how social norms are behavioural rules that arise from conditional preferences in certain contexts. Empirical and normative expectations are essential ingredients for deciding to follow a social norm and to comply with it, even when a decision is the result of an individual deliberation, not influenced by any agreements. Without a social environment in which these conditions are created, moral motivations, potentially existing in *foro interno*, do not apply externally (Bicchieri, 2006).

Following this line, we may say that a moral cue of this type results to be much less effective than a minimal social cue as the one proposed by Rigdon and colleagues (2009). In that case, presenting a very simple watching-eyes configuration had an impact on giving behaviour, while our findings show how inducing a reflection from an impartial point of view is not enough under conditions of total anonymity and of individual choice. It seems that moral cues have an effect on human behaviour only if they are already proposed as ethical principles. For example, reading the ten Commandments before an experiment in total anonymity had a real weight in leading subjects' choices towards honest behaviour (Ariely, 2012). Unlike this type of reminder, our findings show how providing time to make subjects reflect from an impartial point of view does not have the same cognitive and motivational force: it seems much easier to take care of moral principles when someone else makes us remember them, instead of finding the ethical answer by our own.

Despite our hypotheses were rejected, this study might contribute to better understand how moral cues can intervene in the decision-making process and to remember how important the distinction

among social and moral norms is, both in the study of prosocial behaviour and in the relationship between motivation and action.

Appendix: Experimental Instructions

NOVEIL Treatment

Good morning, thank you for participating in this activity. You are taking part into a study on economic decisions. During the activity, you can, depending on your decisions and on other participants' decisions, earn an amount of money in addition to the 3 euros you will receive anyway. The answers you give and the choices you make will be totally anonymous. The experimenters will not be able neither is their intention to associate your choices and your answers to your name. During the activity you cannot communicate with other participants and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. You can find a copy of the instruction on your desk. You can check them in any moment during the activity. If you have any questions, please ask the experimenters.

Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0.15 euros.

At the end of the activity, you will be asked to fill a short questionnaire; afterwards, we will proceed with the payment, that will occur in cash. During the activity you are paired with another participant. You will not be informed of other's identity, and the other will not be informed about your identity.

The activity consists of two stages

STAGE 1

In the first stage you (sometimes we use “you” other times we use “the two participants”. Could this create some confusion?) will be asked to perform a task. The task is the same for all the participants and it determines the earnings of stage 1.

You will be presented a series words and you will be asked to encode the words by substituting the letters of alphabet with numbers, using the table 1.

For example, if the word that appears on your screen is “HOLA” you must enter the numbers 24 for “H”, 21 for “O”, 25 for “L” and 6 for “A”.

For each word you encode you will receive 1 token.

You will be given a time limit and within this limit you must encode as many words as you can.

In each pair the two participants are given two different time limits. One will have 10 minutes at his/her disposal, while the other will have 6 minutes. The assignment of time limits is random, each of the participant of the pair has a probability of 50% of getting 10 minutes and 50% of getting 6 minutes.

The participant with the 6 minute limits will be asked, at the end of the task, to answer a few general questions, not related with the activity, for the remaining three minutes. I think that the activity that we ask to the participant who has less time to encode words should be (and should be perceived as) not very stressful. If the effort asked during the three minutes left is “high” one could perceive as very unfair to receive less money when s/he made a big “unproductive” effort. I think that we should think carefully to this point that could modify the experimental results (I would not be surprised to find out different results when the activity asked to the subject who has less time to encode words is watching TV instead of answering mathematical questions.

STAGE 2.

At the end of the task the members of each pair will be informed about their earnings and the total earnings obtained by the pair, corresponding to the sum of the two members' earnings:

Total earnings = member 1's earnings from the task + member 2's earnings from the task

At this point the two participants will be asked to decide how to divide the total earnings by choosing a combination of percentages from table 2. The first column corresponds to the percentage of the total earnings assigned to him/herself and the second to the percentage assigned to the other participant.

The software will extract at random one of the two participants and the percentage selected by him/her will be used for the final division of the earnings. The probability of being selected is 50%.

For example, suppose that the total earnings of your pair is 80 tokens, you choose the percentage 60% for and 40% for the other participant, while the other participant chooses 70% for him/her and 30% for you. If you are extracted, then your final earnings will be 60% of 80 = 48 tokens and the other participant's final earnings will be 40% of 80 = 32 tokens. If the other participant is extracted, then your final earning will be 30% of 80=21 tokens and the others will be 70% of 80 =56 tokens.

SUMMARY OF THE STAGES

Stage 1: participants are informed about their time limits, perform the task and are paid according to the number of words encoded. Total earnings are computed.

Stage 2: each participant select the percentage to use to divide the total earnings. One of the participant is extracted and his/her decision is implemented. Final earnings are computed

CONTROL QUESTIONS.

You encode a total of 21 words, the other participant will encode a total of 20 words:

Your earnings from the task is of …….. tokens.

The other participant's earnings is of . …….tokens.

The total earnings of your pair is 70. You have chosen to divide it by choosing the percentage 35% for you and 65% for the other participant. The other participant has chosen the percentages: 55% for him/her and 45% for you.

Maybe 0, 5%, 10%, 15% etc. is too much? Could be enough to give the opportunity to choose 0, 10%, 20%...?

If you are extracted, your earnings is of ……. Tokens and the other participant's earnings is of …..tokens.

If the other participant is extracted, your earnings is of ……. and the other participant's earnings is of …..tokens.

ONLYVEIL Treatment

Good morning, thank you for participating in this activity. You are taking part into a study on economic decisions. During the activity, you can, depending on your decisions and on other participants' decisions, earn an amount of money in addition to the 3 euros you will receive anyway. The answers you give and the choices you make will be totally anonymous. The researchers will not be able neither is their intention to associate your choices and your answers to your name. During the

activity you cannot communicate with other participants and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. You can find a copy of the instruction on your desk. You can check them in any moment during the activity. If you have any questions, please ask the researchers. Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0.15 euros.

At the end of the activity, you will be asked to fill a short questionnaire; afterwards, we will proceed with the payment, that will occur in cash and privately.

During the activity you are paired with another participant. You will not be informed of other's identity, and the other will not be informed about your identity.

The experiment consists of three stages. We will tell you about Stage 1 in a minute, but we begin by presenting first Stage 2 because your choices in Stage 1 depend on the knowledge of Stage 2 procedure.

STAGE 2

In the STAGE 2 you will be asked to perform a task. The task is the same for all the participants and it determines the earnings of stage 3.

You will be presented a series words and you will be asked to encode the words by substituting the letters of alphabet with numbers, using the table 1.

For example, if the word that appears on your screen is "HOLA" you must enter the numbers 24 for "H", 21 for "O", 25 for "L" and 6 for "A".

For each word you encode you will receive 1 token. The words are the same for all the participants.

You will be given a time limit and within this limit you must encode as many words as you can.

You and the other participant are given two different time limits. One of you will have 10 minutes at his/her disposal, while the other will have 6 minutes. The assignment of time limits is random and it is made by the software without any intervention by the experimenter. Thus you have a probability of 50% of getting 10 minutes and 50% of getting 6 minutes.

If you are the participant with the 6 minute limits you will be asked, at the end of the task, to answer a few general questions, not related with the activity, for the remaining 4 minutes. This activity do not produce any earnings, and it is introduced only with the aim of not allowing the identification of people with lower limits.

At the end of the task you and the other participant will be informed about your earnings and the total earnings obtained by your pair, corresponding to the sum of your earnings:

Total earnings = your earnings from the task + other participant's earnings from the task.

STAGE 1

In stage 1 you will be asked to choose a combination of percentages (table 2), that you think it should be applied to divide the total earnings produced in in stage 2 (as described above). The first column of table 2 corresponds to the percentage of the total earnings assigned to you and the second to the percentage assigned to the other participant. To access the second stage you have to take this decision in no less than 1 minute and no more than 3 minutes.

Remember that you take this decision before knowing the time limits assigned to you. The other participant will be asked to do the same.

STAGE 3

After having selected the percentages (decided in stage 1) , being informed about your time limits and performed the task (stage 2), both you and the other participant will be asked if you want to divide the total earnings using the percentage combination chosen in stage 1 or to select a different one.

At this point the software will choose at random you or the other participant. The percentage selected by the extracted person will be implemented for the final division of the earnings. The probability of being extracted is 50%.

For example, suppose that the total earnings of your pair is 80 tokens: in stage 1, you choose a combination that assigns a percentage of 60% of the total income to you and 40% to the other participant, while the other participant chooses a division that assigns 70% to him/her and 30% to you. Suppose also that both you and the other participant, at the end of the task, confirm the choices made in stage 1.

If you are extracted, then your final earnings will be 60% of 80 = 48 tokens and the other participant's final earnings will be 40% of 80 = 32 tokens. If the other participant is extracted, then your final earning will be 30% of 80=21 tokens and the others will be 70% of 80 =56 tokens.

Otherwise, suppose that, after the task, you decide to select a combination of percentages different from the one selected in stage 1: if you are extracted, then these new percentages will be applied, otherwise the final earnings will depend on the other participant's decisions.

SUMMARY OF THE STAGES

Stage 1: Both you and the other participant select a percentage combination for the division of total earnings (without knowing the time limit within which you will be asked to accomplish the task)
Stage 2: Both you and the other participant are informed about your time limits, perform the task and informed about the number of tokens obtained with the task by your pair.

Stage 3: Both you and the other participant decide whether to confirm or not the percentages chosen in stage 1. One of you is extracted and his/her decision is implemented. Final earnings are computed.

CONTROL QUESTIONS.

You encode a total of 21 words, the other participant will encode a total of 20 words.

Your earnings from the task is of …….. tokens

The other participant's earnings is of . …….tokens.

The total earnings of your pair is 70. In stage 3, after the task, you have choose to divide it by selecting the percentage 35% for you and 65% for the other participant. The other participant has chosen the percentages: 55% for him/her and 45% for you.

If you are extracted, your earnings is of ……. tokens and the other participant's earnings is of …..tokens.

If the other participant's is extracted, your earnings is of ……. and the other participant's earnings is of …..tokens.

Table 1.

Letter	Number
A	6
B	26
C	13
D	3
E	14
F	19
G	10
H	24
I	2
J	20
K	5
L	25
M	9
N	17
O	21
P	1
Q	11
R	8
S	4
T	18
U	22
V	12
W	16
X	7
Y	23
Z	15

Table 2

You	Other
0%	100%
5%	95.00%
10%	90.00%
15%	85.00%
20%	80.00%
25%	75.00%
30%	70.00%
35%	65.00%
40%	60.00%
45%	55.00%
50%	50.00%
55%	45.00%
60%	40.00%
65%	35.00%
70%	30.00%
75%	25.00%
80%	20.00%
85%	15.00%
90%	10.00%
95%	5.00%
100%	0.00%

IV. *Distributive justice in the lab: testing the binding role of the agreement.*

1. Introduction

Distributive justice deals with allocation problems. . It can be said that a theory of distributive justice is "concerned with what rules, procedures, or mechanisms a society or group should use to allocate its scarce resources, commodities, and necessary burdens to individuals with competing needs and claims" (Oleson 2001, p.13). To understand how principles of distributive justice might provide moral guidance for social, economic and political structures, an empirical approach focuses on individuals' moral psychology; it explores whether and to what extent people are motivated by considerations of justice. Some interesting results have been obtained by introducing reasoning behind the veil of ignorance (Voigt 2015; Huang et al 2019). Some of these studies have shown that agreement in conditions of ignorance may play a role in determining both individual views about justice and individual motivation to act justly.

Relying on Degli Antoni et al's (2016)[13] work, the purpose of the experiment reported in this chapter is to test the robustness of the agreement, in particular we would like to verify whether the collective unanimous choice (the impartial agreement) could create the conditions to convey a conception of distributive justice consistent with what we would call 'liberal egalitarianism'. In broad sense, under the label of 'liberal egalitarianism', we mean those theories that deal with concerns for fair distribution of resources and opportunities and individual merit-based reward. In particular, by liberal

[13] In this paper we refer both to Degli Antoni, G., Faillo, M., Francés-Gómez, P., and Sacconi, L. (2016), Distributive Justice with Production and the Social Contract. An Experimental Study*, EconomEtica*, 60, and to a revised version of it (forthcoming).

egalitarian rule we mean a distributive rule that holds two principles together. On the one hand, a principle of equality in resources and, on the other, an allocation principle proportional to contribution. Indeed, "according to the theory of the social contract the rule chosen under the veil of ignorance should be the liberal egalitarian one: inequalities should be caused only by the subjects' differential use of an equal endowment of time, whereas arbitrary inequalities of endowments should be neutralized – because, in fact, according to the rule its output is re-distributed equally" (Degli Antoni et al 2016, p. 8).

The experiment unfolds on two levels: the choice behind the veil of ignorance (ex-ante) and the actual compliance with the relevant choice (ex-post). In the aforementioned study, the main characteristic of the experimental design is that "subjects are assigned unequal endowments for which they are not responsible; the assignment is random. At the same time, their work naturally generates unequal levels of earnings" (Degli Anotoni et al 2016, p.1). This would have allowed to empirically test "the idea of an agreement behind a veil of ignorance, followed by implementation of the agreed distributive rule(s) in a context of decision in the absence of coercive authority" (Degli Anotoni et al 2016, p.1). Given the presumed different type of commitment stemming from an individual decision or from a collective deliberation, their main hypothesis was that the agreement, and the joint commitment derived from it, justified greater ex-post compliance with what was collectively decided ex-ante. Moreover, the contractarian argument would have guided subjects towards a liberal egalitarian rule both ex-ante and ex-post.

In order to reinforce Degli Antoni et al's findings and strengthen the contractarian argument, we have maintained the overall experimental design, by adding two new treatments. Our results reveal that, when reasoning behind the veil is applied individually, the liberal egalitarian rule is neither chosen nor implemented at the same frequency observed in the treatment with the agreement: the

motivational force to comply with that rule proves to be related to an impartial agreement and we conjecture that this distributive standard, once agreed upon, is perceived by our subjects as the fairest one. Thus, we suggest that our results support that an explicit prior agreement might be a necessary condition for compliance on a liberal egalitarian rule in production situations, insofar as a collective deliberation would motivationally commit subjects in a stronger way than compliance arising from mutual individual expectations.

The remainder of this paper is structured into four sections as follows: the theoretical background; the experimental design; the results and the final discussion.

I. Theoretical background

Fairness ideals might provide sufficient reasons to act. Let us suppose that principles of distributive justice are rationally chosen. What reasons do individuals have to comply, *de facto*, with the content of those principles?

Defining what is meant by fairness in production contexts is one of the primary objectives of the different theories of distributive justice (Cappelen et al 2007). Among those, "…equal opportunity theories of distributive justice (Rawls, 1971; Dworkin, 1981; Roemer, 1998),that combine an egalitarian commitment with a concern for individual responsibility" (Cappelen et al 2005, p. 2) remain at the centre of the contemporary debate of normative ethical theories. By following this tradition, we would like to focus on the normative question of what people should deem fair in distributive contexts and the positive question of how people actually behave when faced with distributive dilemmas. The gap from the normative to the positive is under discussion. It can be

questioned why people should conform to a distributive shared norm, when they can increase their earnings by deviating. This is the question of whether additional incentives are required for people to behave according to a shared normative standard. Our data allow us to contribute an interesting answer to this question.

In the normative literature of fairness in production situations, the method proposed by John Rawls (1971) holds a characteristic role as a method capable to redress the initial unjustified inequalities. For our purpose, John Rawls is a reference in two ways: firstly, we adopt the difference principle as a criterion that "secures for all a guaranteed minimum of the all-purpose means (including income and wealth) that individuals need to pursue their interests and to maintain their self-respect as free and equal persons." (Freeman 2019)[14]. This is the egalitarian part of a distribution principle that allows differences under certain conditions. This egalitarian part is particularly relevant for our study since the background norms that are usually activated in production contexts are *not* egalitarian; they are commonly based on individual entitlements, individual merit/effort, and freedom of contract. Secondly, Rawls is relevant because we adopt, and try to operationalize the contractual procedure. We hypothesize that reaching an agreement behind the veil of ignorance would lead individuals to choose a rule –the "liberal egalitarian" rule that will be defined below– that closely tracks Rawls's ideals of fairness. This rule would be chosen as the normative solution to a problem of distributive justice, *and* would play a role in inducing positive compliance (subjects' real behaviour) with it. The intuition would be that reaching an agreement behind the veil of ignorance would mean reasoning from an impartial perspective, hence inducing a normative perspective on subjects – that is what subjects *ought* to do. This is related with the Rawlsian idea that people have a "sense of justice."

[14]Entry on "Original Position", *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), URL: <https://plato.stanford.edu/entries/original-position/>.

In line with Degli Antoni et al (2016), we want to examine two main questions: whether the contractarian argument guides subjects to choose a distributive rule of liberal egalitarianism, and to what extent people are willing to comply with that rule ex-post. Regarding the liberal egalitarian rule, it takes up Konow's accountability principle (Konow2000, 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, p.378). Liberal egalitarian theories as the ones held by Rawls (1971) and Dworkin (1981b) consider individuals as "morally equal persons deserving equal consideration and respect […] the introduction of any difference among them must be morally justified in terms of outcomes that they can be responsible for because of their agency and independently of the results of social and natural lottery" (Degli Antoni et al 2016, p.7). Nonetheless, it should be said that Rawls and Dworkin are located in two different strands on the issue of distributive justice: the first belongs to what is called 'democratic equality' and the second to 'luck egalitarianism' (Kok-Chor 2008; Brown 2005; Anderson 1999). However, the liberal egalitarian rule we use would like to be a "substantive principle of distributive equality" that "specifies how to distribute what" (Kok-Chor 2008, p. 674). So understood, the liberal egalitarian rule of our experiment recalls, on one side, the Rawlsian difference principle, "that specifies how to distribute (that is, choose that arrangement that maximizes the situation of the worst off) and presumes a common metric of equality (that is, primary goods like income and wealth)" (Kok-Chor 2008, p. 674). On the other, it refers to Dworkin's idea insofar as it tries to "mitigate the effects of luck on the social distribution of goods and resources among persons" (Kok-Chor 2008, p. 665).

Regarding ex-post compliance, the question arises because subjects, ex-post, are faced with a dictator game. In this game, their foreseeable conduct would be self-interested – in the absence of incentives

and social bonds. However, Degli Antoni and colleagues found that the liberal egalitarian rule is preferred in the ex-ante agreement, and also that it is followed ex-post. This result supports the Rawlsian conception of the sense of justice and the theory of conformist preferences by Sacconi and Grimalda (Sacconi 2006, 2011, Grimalda and Sacconi 2005). However, the question remains whether the agreement itself has the double key role that Degli Antoni et al suggest it has –as a deliberative procedure ex-ante that actually binds parties ex post, hence inducing high levels of unexpected compliance with the liberal-egalitarian rule. Our question focuses on the role of "reasoning behind the veil of ignorance" in the absence of agreement. That is, in the absence of an agreement, would the subjects choose the liberal egalitarian rule ex-ante? Would they comply with? Would the sense of justice emerge in any case?

The hypothesis here is that if the contractarian argument is correct, the impartial agreement behind the veil of ignorance is a fundamental condition for leading individuals to choose and comply with the liberal egalitarian rule. If the choice of liberal egalitarianism were not related with impartial agreement, we might observe the choice of this rule through mere individual moral reflection as applied to the distributive situation created in the lab. According to Bicchieri (2006), social norms (including fairness norms) emerge from the context and are activated by external cues to which people are sensitive. So, given the information provided by the context (in our case experimental instructions), the liberal egalitarian rule, although demanding, should be the focal one even without reaching the agreement. In our experiment we compare the treatment with agreement (fully based on the contractarian argument) with two new treatments that substitute individual decisions. The two "individual" treatments differ in that in one of them ex-ante choice is common knowledge, while in other it is not. This is purported to test reciprocity as a determinant for compliance with the liberal-egalitarian rule. These individual treatments will help us know whether subjects' individual choice

and commitment can lead to the choice of liberal egalitarianism and motivate compliance as effectively as impartial agreement. If this were the case, the argument for the contractarian procedure would be undermined. In the individual treatments we study two situations that we hypothesize should influence compliance in different ways: in the first one the veil of ignorance is used as a moral cue individually applied; in the second one the elicitation of mutual expectations between subjects is added to the veil of ignorance as moral cue. (In our design, the code name of these treatments are "Individual Choice" and "Rule Other" respectively).

To recap, we would like to see whether subjects are led to choose the liberal egalitarian principle and to comply with it, even in the absence of an impartial agreement.

II. Experimental design

In this study we compared three treatments, one called Agreement[15], which constitutes our baseline, a second one named Individual Choice, and a third one called Rule Other. All the treatments consist of three stages: in the first one, subjects had to choose how they would like to split the final total income; in the second, subjects performed a task on which depends the final outcome; in the third phase, subjects could decide amongst three options: (i) confirm the choice made in the first stage; (ii) change the previous choice by clicking on a different distributive criterion; or (iii) select a percentage corresponding to how much of the final amount they wanted to obtain. Subjects could decide how to distribute the final income by choosing one of the five distributive rules/principles (norms in broader sense) that track different fairness ideals. Hereafter are the five rules:

[15]Replication of "Bargaining treatment" (Degli Antoni et al 2016, pp. 13-14).

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

Example: subject A produces X in 10 minutes; subject Y in 6 minutes. Each one obtains [X+Y]/2.

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

Example: subject A produces X in 10 minutes; subject B produces Y in 6 minutes. The subject who is randomly selected (50% probability of being selected) obtains X+Y, the other subject obtains 0.

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

Example: subject A produces X in 10 minutes; subject B produces Y in 6 minutes. Subject A obtains X; subject B obtains Y.

4) Rule 4 – Time independent division: each subject obtains what s/he has produced through her/his activity during the first 6 minutes; for the subject who has 10 minutes, the product of herlast 4 minutes work is divided at 50% between the two subjects.

Example: subject A produces X in the first 6 minutes and K in the last 4 minutes; subject B produces Y in 6 minutes. Subject A obtains X+(K/2) and subject B obtains Y+(K/2).

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 time 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.

Example: subject A produces 60 words in 10 minutes, subject B produces 40 in 6 minutes. The ratio between A's and B's productivity is 6/6.66= 0.90. The payoff of A should be 0.90 times the payoff

of B, and the sum of the two payoffs should be 60+40=100 tokens. A's payoff is 47.4 tokens and B's payoff is 52.6 tokens.

These five distributive rules did not change across treatments and they follow the main moral intuitions (Haidt 2007; Sinnott-Armstrong et al 2010) about ideals of fairness. Rule 4 tracks liberal egalitarianism (understood as explained in the theoretical background). Then, rule 1 recalls "pure egalitarianism (the total product is distributed equally)"; rule 3 is a merit rule "based on contribution or entitlement (each subject gets –is entitled to– what she/he has individually produced)". Rules 2 and 5 "reflect views that are typical in economic contexts: self-interest (each subject claims the entire product of the pair), and distribution strictly proportional to productivity (so that the person with less endowment may actually get more if s/he has been more productive per minute)" (Degli Antoni et al 2016, p.7).

In all treatments, (i) participants are grouped in pairs (ii) the endowment was earned (by the pair) through a task; (iii) each member of the pair was randomly assigned different time limits (10 or 6 minutes) to perform their task; (iv) in the third stage, participants played a dictator game, in which the software randomly assigned the dictator and responder roles. The underlying assumption behind condition (iii) is that the person with more available minutes has an advantage, *and* a corresponding responsibility; the foreseeable larger contribution of the person with ten minutes would not simply be the effect of chance, but the combined effect of chance *and* additional work on her part. This situation purports to represent the most common social distributive problems –those that are solved through liberal-egalitarian principles.

The task was coding words, by using a conversion table, and it was the same across all the treatments. Before starting the task, participants saw on their monitors how much time (six or ten minutes) they would receive to complete the task. At the end of the task, they were informed about each member's performance (total number of coded words and productivity measured as words *per minute*). The same list of words was presented to every participant. Subjects were paid in experimental currency called token. At the end of the experiments tokens were converted in Euros at the exchange rate of 1 token = €0.15. They received one token for each word correctly coded. At this point, each member of the pair had to decide how to divide the total output of the pair. After this final choice, one member was randomly selected and her choice was implemented. So, each participant decided knowing that her choice had a 50% probability of being implemented as the real final payoff.

The three treatments differed in what follows:

Agreement

It was the baseline treatment and it reproduced the *Bargaining treatment* by Degli Antoni et al (2016)[16]. In the first phase of the game, subjects had 13 total rounds available in order to reach an agreement on which principle to choose, before moving on to the next phases of the game. The first 6 rounds were simultaneous, then there were 4 sequential offer and counter-offer rounds. The sequential process stopped when the rule proposed by one player was accepted by the other. For example, suppose that player A proposed the rule 4: player B could accept it, so the agreement was

[16]"…the task and the division phases were preceded by a stage in which the members of the pairs, before knowing the allocation of the time for the task, 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" (Degli Antoni et al, 2016, p.14).

reached and they could start the task phase. If player B did not, she couldmake a counter-offer by proposing a different rule. Player A could accept and they had an agreement, otherwise they had another sequential round as such. If subjects failed to agree, they had 3 additional simultaneous rounds. If the agreement was not reached within the 13 rounds, the subjects were excluded from the experiment, they then filled out a short questionnaire and remained in the laboratory until the end of the experiment.

Individual Choice and Rule Other

However, one might consider reasoning behind the veil of ignorance as a salient cue, that is a reason to activate a specific norm and to comply with it (Bicchieri 2006, 2008). If it were the case, a prior agreement among the parties would not be a necessary condition for making a specific distributive standard focal (i.e., Rule 4 in our experiment).

In both treatments, the subjects did not have to reach an agreement, rather they were asked to individually choose, behind the veil of ignorance, one rule before the task. The task remained the same even across these treatments and constituted the second phase.

The third phase, however, differed in the two treatments. In Individual Choice treatment, participants were asked to confirm the previously chosen rule, change it with another or select a percentage. In Rule Other treatment, subjects were informed about their partner's choice ex ante.[17]. They could then decide whether to confirm their first preference, change it with another rule or ask for a percentage of the product. Knowing the other person's judgment behind the veil should create a set of mutual

[17] By using the expression ex-ante choice regarding to individual treatments, we want to establish a parallelism with the baseline. However, given the absence of the agreement, it must be remembered that ex-ante choice is to be understood as a choice made behind a veil of ignorance.

expectations through which the two players could try to coordinate: both players prefer to conform to a shared principle of distributive justice rather than deviate at the risk of getting nothing –that is, if both choose selfishly, one of them will get nothing for sure. If, behind the veil of ignorance, the players choose Rule 4, and this rule becomes common knowledge, then they should act accordingly (given the emergence of empirical and normative expectations).

Individual Choice treatment and Rule Other treatment were designed to test the strength of the agreement. For our experimental design, merit should not influence the subjects' decision as it is a merely random element – namely, time limits are not assigned because one subject has achieved a higher score in some performance than her/his partner. The implicit assumption is that, without a collective choice, the subjects overshadow the random assignment of time. After the task, the effort to codify words becomes the focal point. If someone works for ten minutes is entitled of what she has produced in that time. The same is true for the person with six minutes. . Attention to the fact that the assignment of time is random would reveal a disposition to identify fairness with our liberal-egalitarian principle. Given the theoretical background explained above, a set of hypotheses can be put forward.

H1. The procedure via agreement behind the veil of ignorance leads subjects:

(i) To adopt a distributive criterion in line with liberal egalitarianism (Rule 4 in our experimental design);

(ii) To comply with Rule 4.

This hypothesis is in line with Degli Antoni et al's hypotheses and evidence on the effectiveness of the agreement in inducing the convergence, both ex-ante and ex-post, on the liberal egalitarian principle of distributive justice.

H2. Individual reasoning behind the veil has a smaller impact on the convergence on the liberal egalitarian principle with respect to the agreement, both in terms of ex-ante choice and in the ex-post choice.

This hypothesis has to do with the idea that a justified joint commitment is stronger than an individual one in inducing the adoption and the implementation of a principle of justice in general and of the liberal egalitarian in particular. The justification of the hypothesis relies on Rawls' "sense of justice"[18] and on the evidence about the agreement as a precondition for its emergence (Sacconi and Grimalda 2007; Faillo and Sacconi 2007; Sacconi and Faillo 2008, 2010; Sacconi, Faillo and Ottone 2011).

H3. Information about the rule chosen by the other person (and common knowledge about this fact) induces both a convergence ex ante on the liberal egalitarian rule and ex-post compliance with it.

Knowing that the other participant will be informed about my ex-ante choice (and vice versa) could create a condition, proper of common knowledge, for the elicitation of mutual expectations: hence, the Rule Other treatment would find in the elicitation of mutual expectations the potential justification for compliance. Within the experimental literature dealing with conformity, Bicchieri's (2006) theory

[18] "If we answered love with hate, or came to dislike those who acted fairly toward us, or were averse to activities that furthered our good, a community would soon dissolve […] A capacity for a sense of justice built up by responses in kind would appear to be a condition of human sociability" (Rawls, J. (1999). *A Theory of Justice. Revised ed*., p.433).

locates the necessary conditions for compliance with a social norm in mutual expectations. In fact, Bicchieri identifies three reasons for an agent to decide to comply with a standard: to avoid a negative social sanction, to promote one's own desire to please others; to accept others' normative expectations as well founded. She says "If I recognize your expectations as reasonable, I have reason to fulfil them. I may still be tempted to do something contrary to your expectations, but then I would have to justify (if only to myself) my choice by offering alternative good reasons and show how they trump your reasons"[19]. People have good reasons to comply with others' expectations, therefore conforming, when the norm at stake is made salient by external cues – related to the context where they find themselves. These cues are interpreted as scripts "embedded in the norms", that would be responsible to activate norms, on which conformity is required, through the cognitive process of categorization. Roughly, given people's background of past experiences, they see the relevant others of the group – for instance, the community where they live – act according to a certain behavioural pattern, under specific circumstances, so each of them would form expectations on how most of people behave (empirical expectations) and how others think s/he should behave (normative expectations). According to Bicchieri, these kinds of expectations are legitimate and constitute a justified reason for compliance. Reading the instructions of the Rule Other treatment, the participants were informed of the initial unjustified inequality in the assignment of the endowment, and they were told that the ex-ante choice of each player would be communicated to the other player (after the task phase and before the final ex-post choice). Subjects were also informed, from the very beginning, that the real division of the income would take place by implementing one choice among the ones of the two players. By these external cues, subjects should have activated a script for which, not knowing who would have

[19] Bicchieri, C. (2006). *The Grammar of Society. The Nature and Dynamics of Social norm*, Cambridge University Press, pp. 23-24.

a larger endowment (and therefore an opportunity to produce a larger contribution), rule 4 turned out to be the focal ex-ante distributive rule. Giving empirical expectations and conditional preferences, if the context was properly understood, participant should have been sensitive to rule 4 even without an explicit agreement (H3).

In all treatments, instructions[20] were read aloud by one of the experimenters, a set of control questions were proposed to make sure that participants understood the instructions and they were paid a fixed show-up fee of €3. The experiment was programmed by using zTree (Fischbacher, 2007) and conducted at the Cognitive and Experimental Economics Laboratory (CEEL) at the University of Trento. A total of 175 students participated in the experiment between March 2018 and March 2019. Two sessions of 18 subjects and one with 20 participants were run for the Agreement treatment, three sessions of 20 subjects each for the Individual Choice treatment, two sessions of 20 subjects and one with 18 participants for the Rule Other Information treatment.

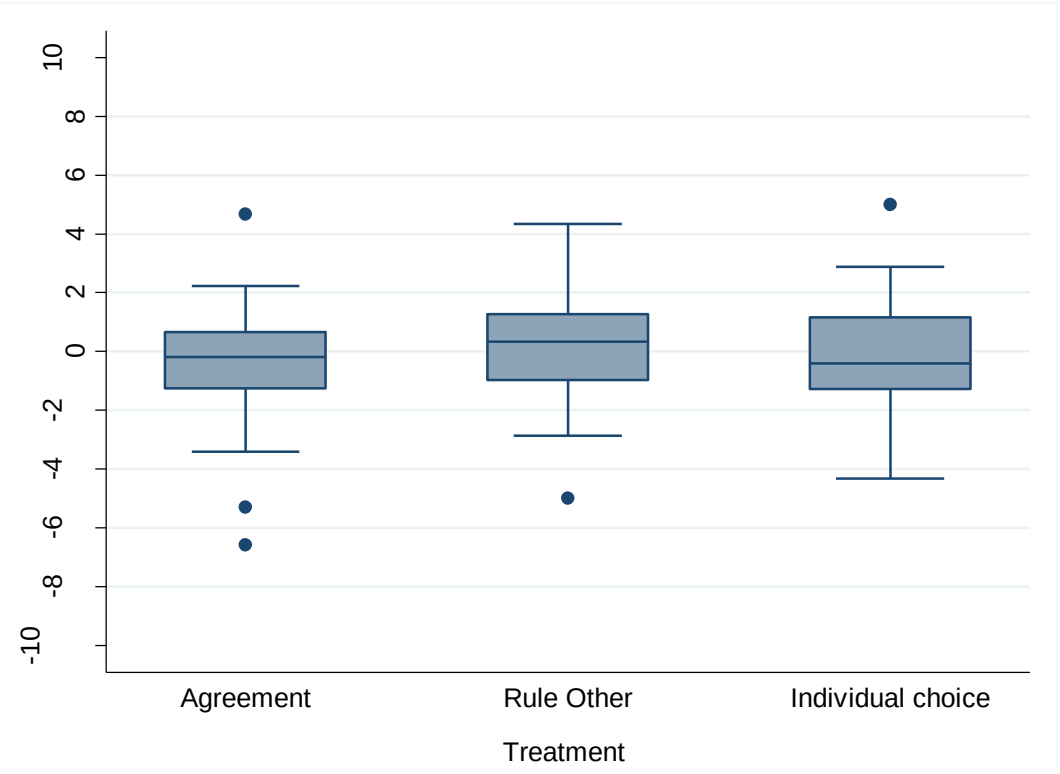
[20] See Appendix at the end of chapter.

III. Results

Task

Looking at subjects' performance in the task, we do not observe significant differences in productivity, measured as words per minute, within the pairs. The median difference between the productivity of the two members of the pair is very close to zero in all the treatment (Figure 1). This supports the idea, at the basis of the original design by Degli Antoni et. al (2016), that different abilities have only a marginal role in explaining differences in performance in the task, and the main source of difference are the different time limits assigned to the two categories of subjects.

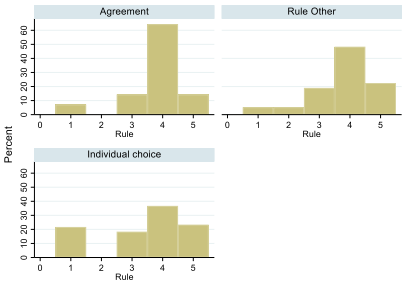
Figure 1. Difference in productivity (words per minute) within the pair



Choice ex-ante

Figure 2 reports the choices made by participants in the ex-ante phase of the experiment, before starting the task and before knowing the time assigned to complete it.

Figure 2. Choice ex-ante



In all the treatments Rule 4 is the most frequently chosen rule. A proportion test reveals that Rule 4 is chosen more frequently in the Agreement than both in the Individual choice (z=2.97, p=0.003) and in the Rule Other treatments, but in the latter case the difference is significant only at the 10% (z= 1.72, p=0.08). The first result is confirmed by a probit estimation in which we control for subject's age, gender and experience with the experiments (Table 1).

The number of subjects choosing other rules is too small to perform a detailed analysis. Notice however that the frequency of choice of Rule 1 is significantly higher in the Individual treatment than both in Rule Other and in the Agreement treatment (Individual choice vs. Rule Other: z= -2.61, p=0.01; Individual choice vs. Agreement z= 2.21, p=0.02).

Table 1. Determinants of ex- ante choice of Rule 4

Dep. Variable: Rule 4 ex ante	(1) Probit
Agreement	0.742*** (0.242)
Rule Other	0.311 (0.236)
Age	-0.0277 (0.0394)
Gender	0.177 (0.196)
Experiments	-0.00751 (0.0129)
Constant	0.234 (0.856)
Agreement – Rule Other	0.43 (0.240)
Observations	174
Pseudo R^2	0.04
Log Likelihood	-115.011

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
The dependent variable is equal to 1 if the subject chooses Rule 4 before the task and 0 otherwise
Agreement is equal to 1 if treatment is the Agreement treatment and 0 otherwise
Rule Other is equal to 1 if treatment is the Rule Other treatment and 0 otherwise
Age is the age of the subjects, Gender is equal to one if the subject is female and 0 otherwise, Experiment is the number of experiment the subject as taken part in the past.

We can then put forward the following results.

Result 1.

In the ex-ante choice of Agreement treatment, the choice of Rule 4 is more frequent than in the Individual choice treatments.

Result 2.

The frequency of ex-ante choices of Rule 4 in Rule Other treatment is not statistically smaller than in Agreement treatment, but it is not statistically greater than in the Individual Choice treatment.

The two results support our hypotheses H1 and H2 with regard to the ex-ante choice. As for H3 hypothesis, using the Individual choice treatment as a benchmark, the convergence of Rule other treatment's results towards the results obtained in the Agreement treatment is only partial.

Choice ex-post

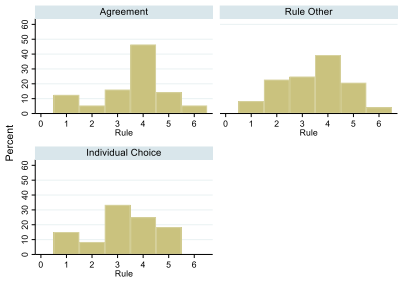
After the real effort task, subjects can choose either one of the five rules or a free percentage. Subjects opting for the percentage were six in the Agreement treatment, four in Individual Choice and two in the Rule Other treatment. We decided to consider the two choice of a percentage of 50% as equivalent to the choice of Rule 1, and two choices of a percentage of 100% as equivalent to the choice of Rule 2.[21]

As for the choice ex-post (Figure 3), Rule 4 is the most chosen in the Agreement and in the Rule Other treatments. Rule 3 prevails in the ex-post decision of Individual choice treatment. Rule 4 is chosen more frequently in the Agreement than in the Individual choice treatment ($z=2.41$, $p=0.01$),

[21] Two subjects choose a percentage of 1%.

Rule 3 is chosen more frequently in the Individual choice treatment than in the Agreement treatment (z= 2.14, p=0.031), and Rule 2 is chosen more frequently in the Individual choice treatment than both in the Baseline (z=2.15, p=0.03) and in the Rule Other treatments (z=1.69, p=0.09), even if the latter difference is significant at only the 10%.

Figure 3. Choice ex-post



This evidence is confirmed by probit estimations (Table 2, columns 1 and 2). Table 2 show also that subjects with ten minutes are less likely to choose Rule 4 ex-post[22] than subjects with six minutes. The third column of Table 2 reports the results of a probit estimation limited to the choice made by the subjects in the Rule Other treatment. We observe that the knowing that the other subject in the pair chose Rule 4 ex-ante (Rule other 4) has no effect on the choice of Rule 4 ex-post.

[22]We checked for difference between treatments by using interactions between the dummy variable Ten and treatment dummies Agreement and Rule Other. None of the coefficients was different from zero.

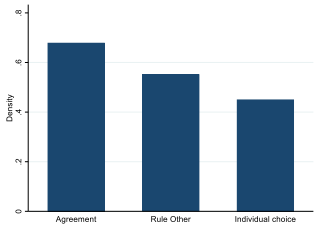
Table 2. Determinants of the ex-post choice of Rule 4

	(1)	(2)	(3)
Dependent variable: Rule 4 ex-post	Probit Full sample	Probit Full sample	Probit Rule Other only
Agreement	0.671*** (0.252)	0.713*** (0.258)	
Rule Other	0.324 (0.255)	0.352 (0.261)	
Ten		-0.577*** (0.207)	-0.536*** (0.205)
Age	0.0141 (0.041)	0.0157 (0.042)	0.0195 (0.0410)
Gender	0.602*** (0.206)	0.586*** (0.210)	(0.204) -0.0178
Experiment	-0.0190 (0.014)	-0.0179 (0.014)	(0.0145)
Rule other 4			-0.0792 (0.291)
Constant	-1.175 (0.910)	-0.969 (0.934)	-0.655 (0.896)
Agreement – Rule Other	0.360 (0.250)	0.339 (0.470)	
Observations	174	174	58
Pseudo R	0.07	0.11	0.08
Log Likelihood	-103.89	-99.960	-33.817

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
The dependent variable is equal to 1 the subject chooses Rule 4 after the task and 0 otherwise
Agreement is equal to 1 if treatment is the Agreement treatment and 0 otherwise
Rule Other is equal to 1 if treatment is the Rule Other treatment and 0 otherwise
Ten is equal to 1 the subject has 10 minutes and zero otherwise.
Rule Other 4 is equal to 1 if the other subject in the pair chose Rule 4 ex-ante.
Age is the age of the subjects, Gender is equal to one if the subject is female and 0 otherwise, Experiment is the number of experiment the subject as taken part in the past.

Independently of the rule chosen, the frequency of compliance (ex-post choice confirming the choice ex-ante) is significantly higher in the Agreement than in the Individual choice treatment (z=2.48 p=0.013) (Figure 4).

Figure 4. Compliance across treatments



Looking at the frequency of compliance across rules and treatments (Table 3), we also see that compliance with rule 4 is significantly higher in the Agreement than both in the Individual (proportion test: z=3.66, p=0.002) and in the Rule Other treatment (z=2.10, p=0.03).

Table 3. Proportion of compliant subjects across treatments and rules

	Treatment		
Rule	Agreement	Individual choice	Rule other
1	4/4 (100%)	1/3 (33,3%)	5/13 (38,4%)
2	0/0	2/3 (66,6%)	0/0
3	5/8 (63,5%)	8/11 (72,7%)	5/11 (45,5%)
4	25/36 (69,4%)	15/28 (53,6%)	11/22 (50%)
5	4/8 (50%)	6/13 (46,1%)	6/14 (42,8%)
	38/56 (67,8%)	32/58 (55,2%)	27/60 (45%)

This result is confirmed by the probit estimation of Table 4 in which we also observe that the likelihood of compliance with Rule 4 for subjects with 10 minutes is lower than that for subjects with 6 minutes[23].

Table 4. Determinants of compliance with Rule 4

Dep. Variable: Compliance with Rule 4	(1) Probit
Agreement	0.873*** (0.264)
Rule Other	0.304 (0.322)
Rule_Other 4	0.102 (0.380)
Ten	-0.464** (0.216)
Age	0.00385 (0.0441)
Gender	0.494** (0.214) -0.00759
Experiments	-0.00759 (0.0150)
Constant	-1.018 (0.972)
Agreement – Rule Other	0.56* (0.31)
Observations	174
Pseudo R^2	0.10
Log Likelihood	-94.70

[23] We replicated the same estimation by adding the interaction term Agreement X Ten to check for differences in the response to the manipulation by participants with different time limits, but the coefficient of the term is statistically not significantly different from zero (coefficiet = 0.26, SE = 0.45, z = 0.58, p = 0.57) and there is no impanct on the coefficients of the other variables.

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
The dependent variable is equal to 1 if the subject chooses Rule 4 both ex-ante and ex-post.
Agreement is equal to 1 if treatment is the Agreement treatment and 0 otherwise
Rule Other is equal to 1 if treatment is the Rule Other treatment and 0 otherwise
Rule Other 4 is equal to 1 if the other subject in the pair chose Rule 4 ex-ante.
Ten is equal to 1 the subject has 10 minutes and zero otherwise.
Age is the age of the subjects, Gender is equal to one if the subject is female and 0 otherwise, Experiment is the number of experiment the subject as taken part in the past.

We summarize the evidence on ex-post choice with an additional set of results.

Result 3

The choice of Rule 4 in the ex-post phase is more frequent in the Agreement treatment than in the Individual choice treatment.

Result 4

The frequency of Rule 4 ex-post choices in the Rule Other treatment is not statistically smaller than that observed in the Agreement treatment and it is not statistically greater than in the Individual Choice treatment.

Result 5

Compliance with Rule 4 in the Agreement treatment is higher than that observed in both Individual choice and Rule other treatments.

These results support hypothesis H1 and H2. H3 is only partially supported.

IV. Conclusions and further research

Even if the extent of these data is moderate, they would help to prove the robustness of the agreement both ex-ante and ex-post. Regarding the strength of the agreement behind the veil of ignorance, the question was about the effect that the impartiality elicited by the veil could have on the agents. The role of agreement gets confirmed because, in its absence and given the context, rule 4 was not individually chosen. The veil of ignorance, therefore, would seem to be a moral device only if applied to collective deliberative processes. Concerning ex-post compliance, the collective decision (namely, the agreement) would justify a joint commitment to put the liberal egalitarian rule into practice. This normative feature of the agreement is not found when the ex-ante decision and ex-post compliance are relegated to the individual sphere. This would suggest that the contractarian argument, leading to liberal egalitarianism, is not as ideal as is commonly deemed. The bond created by an explicit prior agreement between the parties would create a commitment such as to put into practice that specific distributive rule: liberal egalitarianism, as it is exposed here, might be a realistic moral argument, namely it can be relevant for human conduct.

The hypothesis H1 was confirmed, in line with the previous results of Degli Antoni and colleagues: the impartial agreement between the parties can be considered as a necessary condition (1) to induce subjects towards a distributive criterion resembling the liberal egalitarian one, and (2) to ensure compliance with the agreed norm.

This condition is reinforced by evidence in support of hypothesis H2: the individual choice behind a veil of ignorance does not constitute a moral learning device neither for the ex-ante choice nor for the

ex-post compliance. The veil of ignorance, if individually applied, fails its moral function, challenging the effectiveness of moral principles as motivational source in *foro interno*. The results obtained align with other data on classic bargaining games, in conditions of anonymity, non-iteration and absence of external negative sanction and/or positive reward (Hoffman et al, 1996; Rodriguez-Lara and Moreno-Garrido 2012). In the Individual Choice treatment, in fact, rule 1, which reflects pure egalitarianism, was more frequently chosen compared to the other two treatments. This result could be interpreted as an easy choice (half a cake each one), pointing out subjects' lack of interest in nullifying that kind of inequalities produced by luck –or inequalities in initial endowment. Once the activity is performed, the pure egalitarian choice is replaced by rule 3 (each gets what has produced). Although merit is not elicited in our experimental design, since the advantage position (having more time to perform the activity) is not the result of a tournament/competition/skill test, it would seem that the probability of being more or less favoured, solely by chance, is taken as legitimate. Subjects in individual condition focus on their personal production: they choose to take what they know is the product of their effort coding words. . Unlike a strand of literature that tries to isolate the elements of luck and effort, in relation to giving/taking behaviour (Konow 2000; Fong 2007; Erkal et al 2011; Rey-Biel et al 2018; Charness et al 2018), our experimental design does not link effort and merit: the fact of getting extra minutes to produce, increasing one's income, does not result from a pre-selection based on merit nor on particular abilities. However it was observed that the subjects who got the highest income were more reluctant to give a part of their earnings to the others, as if being the beneficiary of a random event was a form of legitimizing the difference.

We also observe that the likelihood of compliance with Rule 4 for subjects with 10 minutes is lower than that for subjects with 6 minutes. This result is interesting because it would confirm the idea that those who, *de facto*, were in a situation of less advantage ones, expressed their willingness to

implement a rule that allows them to tackle the luck effect. The more advantaged, instead, adopted a more self-interested behaviour. The richest, in their leading position, do not care much about helping the most disadvantaged subjects (Cabrales et al 2011). Unlike some studies in which external moral cues (Leavitt et al 2016) affected the moral identity of the participants, the non-effectiveness of the veil of ignorance in the Individual Choice treatment confirms the experiment reported i the previous chapter. Recall that, in that experiment, one treatment, the ONLYVEIL, presents the same structure of the Individual Choice –except there is one difference that might be cognitively interesting: in the ONLYVEIL treatment, subjects were asked to indicate, behind the veil of ignorance, a *percentage* that they would like to apply to divide the final product, rather than a "rule" from a menu of rules. The experiment followed likewise, but in the last stage they could confirm the percentage indicated behind the veil of ignorance or asked for a different amount. It could be said that, when faced with a numerical choice, subjects would be influenced by the quantity (of money) they can earn, while, with written rules, they are somehow forced to stop and think about what implies applying one rule instead of another. This could explain the more selfish behaviour in the ONLYVEIL treatment: after having produced their own income, the most advantaged subjects asked for a percentage higher compared to their production. By having to decide which rule to implement, however, selfishness is reduced. In Individual Choice, subjects can choose whether to implement a rule or ask for a percentage of the product, the introduction of written rules seem to stem pure selfishness – at least they don't steal – but their reflection behind the veil of ignorance has a null effect.

Hypothesis H3, concerning the Rule Other treatment, was only partially confirmed. The subjects were told that, before their final choice on how dividing the total income, they would be informed about the choice made behind the veil (ex-ante) by their partner . This element should affect mutual empirical expectations (what others do) even behind the veil of ignorance, insofar as rule 4 is chosen

almost with the same frequency observed in the Agreement treatment. Once the veil is removed, the degree of compliance with rule 4 is greater than the one in the Individual Choice treatment, but it is lower compared to the Agreement. In relation to these results, we want to emphasize the importance of mutual expectations in inducing compliance with a social norm, and we propose some interpretations in support of the main argument: an explicit prior agreement is a fundamental condition in leading subjects to choose the liberal egalitarian rule and to comply with it. Compliance with fairness norms, as said above, would depend on the context and on the information that individuals have available. Assuming that, in our experimental design, choosing and conforming to a distributive principle rather than another could be understood as a preference for a specific social norm, the Rule Other treatment should have made rule 4 focal behind the veil of ignorance, inducing a corresponding degree of compliance. Social norms are not only coordination rules, as by Lewis[24], but mainly rules of interdependent behaviour. People prefer to conform because they have empirical expectations (they believe that relevant others do conform) and normative expectations (they believe that relevant others think they should conform too). In other terms, social norms are followed because we believe others follow them and we believe those others think we should follow them too. If a social norm is recognized by the parties, the latter should be motivated to conform given the mutual expectations, both empirical and normative:

Providing information on the choice made by the other player behind the veil should have created a set of reciprocal expectations, at least empirical, about how the other wanted to divide the final product. The data might reveal how reciprocal empirical expectations have an effect: many subjects choose Rule 4 behind the veil of ignorance and the degree of compliance is greater than in the

[24] "Social norms are customary rules of behavior that coordinate our interactions with others. Once a particular way of doing things becomes established as a rule, it continues in force because we prefer to conform to the rule given the expectation that others are going to conform (Lewis, 1969)" (Young 2007, p.2).

Individual Choice treatment, where no information were provided. So, on the one hand, knowing that the other person will be informed of my decision behind the veil (and vice versa) would elicit a preference for rule 4, since the other could expect I would consider the initial random distribution of endowment as unfair. On the other hand, the absence of agreement lowers compliance, as compared with our baseline. Empirical and normative expectations have a decisive role in promoting conformity to a rule, but they would not be sufficient to lead towards liberal egalitarianism. The gap between the decision to choose rule 4 and having reasons to comply with it is closed by the contractual procedure: the agreement would constitute a motivating mechanism for which the subjects would feel jointly committed to complying with what was collectively decided.

The fact to feel jointly committed underlines what doing something together, as a single body (Gilbert 1989, 2014), means, that cannot be reduced to some aggregate of personal commitments to do X (knowing that maybe others will do the same thing). As Gilbert says, joint commitments, unlike personal ones, create obligations. Obliging, in this context, is understood in general terms. Our results seem to give us reasons to endorse Gilbert's view on obligations derived from joint commitments. The idea is that, by having an obligation, agents have sufficient reason to act: "the obligations of joint commitment are powerful inputs to reasoning about what to do […] they trump inclinations and self-interest as such, and they are recalcitrant to a person's own fiat" (Gilbert 2006, p. 158).

The contractarian argument and the scheme we designed to model it in the lab through the agreement behind a veil of ignorance, would fix the conditions for the emergence of a joint commitment capable of motivating agents to put the agreed content into practice. Given the decision-making conditions of the deliberative procedure – impartiality and impersonality –the agreement leads to consensus on the liberal egalitarian rule. Our interpretation is that, the procedure is not only related to the agreement on liberal egalitarianism, but also that it creates a form of commitment that is ontologically different

from a mere sum of personal commitments. This form of commitment it obliges the parties to act as agreed. An explicit prior agreement, therefore, would be a *conditio sine qua non* for choosing and conforming to rule 4. Mutual expectations, individually created from the context plus personal sensitivity towards the involved ideal of fairness, would not be sufficient to guarantee the ex-ante choice of the liberal egalitarian rule or to demonstrate its stability ex-post, that is, the individual effective disposition to act according to the rule. The ethical justification of the ex-ante choice through explicit agreement would provide psychologically valid reasons for applying rule 4 also ex-post, making the content of this rule binding on the subjects and. In one sense, this evidence lends realism to the moral contractarian argument. It makes us confident that the problem of "stability" as formulated by Rawls, can be solved: an agreement under impartial circumstances can be actually binding for ordinary people. This in turn makes liberal egalitarianism a justified principle, not only in theory but also –with the appropriate limitations– in practice. According to Bicchieri, conformity to norms of fairness does not need an explicit prior agreement, since schemata and categorization process would constitute the benchmark for complying with one norm instead of another. However, the data showed how liberal egalitarianism requires an explicit prior agreement between the parties to achieve collective decision and corresponding joint motivation to act accordingly. If the explanation of compliance with a fair distributive norm were captured by Bicchieri's theory, the design in Individual Choice should have provided the minimum elements for moral action. Reasoning behind the veil in the context of the laboratory decision situation would have constituted a moral cue and it would have made rule 4 the ex-ante and ex-post salient distribution rule.

Even more clearly, in Rule Other treatment, if I know that the other knows what rule I have chosen behind the veil (and vice versa), at least empirical expectations should arise: both players know that the experiment consists of some distribution of a common output; moreover, they know that, at certain

point, they will be informed about the other's ex-ante choice. It is therefore reasonable to think that ex-ante choices express players' conditional preferences: I choose a certain rule because I expect you to stick the same one and I believe that you expect I should do the same. The salience of one distribution rule over another would be determined by external cues, by the scripts we activate given the context and the categorization process. If this were the case, rule 4 should be the focal rule, by meeting the conditions for individuals to comply, and an explicit prior agreement would not be necessary.

However, H2 and H3 are confirmed. The conclusion is therefore that an explicit agreement between the parties is necessary in order for the parties to reach the liberal egalitarian rule after deliberation, and then embrace a joint commitment to implement it.

From a philosophical perspective, the data could reinforce Kantian constructivism as a method of procedural justice. The deliberation process behind the veil of ignorance would have allowed subjects to put themselves in the shoes of others. Unlike the other two treatments, the veil of ignorance would work as a moral learning device, but under some conditions: impartiality and unanimous consensus. In motivational terms, the agreement would enable individuals to believe that the right thing to do is to behave in line with the chosen distributive principle. So, the content of the agreement gives a *real*, psychologically motivating reason to act in accordance.

The procedure would be the source that would justify the choice of the liberal egalitarian principle. The devised procedure makes subjects reflect impartially and, given the characteristics of the procedure, such principle is considered fair in a binding way for action. The outcome is made just *by* the procedure, providing reasons to comply even if, ex post, it is a dictator game –the rules of the game allow self-serving actions. The content of the agreement constrains subjects' will, making them

ready to the conforming action. How? Because the procedure self-imposes the fairness of its outcome. It is not the case that liberal egalitarianism is the fairest ideal independently (like an external principle needed to be justified), but its fairness is justified through the procedure and this is why, without the pure proceduralism, it is not considered the fairest criterion.

So, from a philosophical perspective, our experimental design tried to highlight how an ideal of fairness such as the liberal egalitarianism would be feasible as a result of a pure proceduralism, using Rawls's idea: an impartial deliberation that detects in the agreement the guarantee to uphold concerns for fairness, by fostering, at the same time one's preferences to behave accordingly.

Our results have limitations. We are perhaps jumping to philosophical conclusions from results that are circumscribed to a very particular design. New developments of our experimental design are required for sure. However, the use of actual agreement behind a veil of ignorance in this experimental design might be a source of insight to reinforce the argument that, where initial unjustified inequalities are at stake, a prior agreement is actually a necessary condition for compliance because it would generate the motivational forces able to induce the adoption of potentially counter-interested behaviour. The impartial agreement behind the veil of ignorance could constitute a normative solution to a compliance problem in terms of motivational gap: it provides a normative solution (the liberal egalitarian criterion) on a normative problem (how should subjects distribute a surplus, ceteris paribus?), binding subjects to actually behave accordingly to the agreed principle.

Appendix: Experimental Instructions

AGREEMENT TREATMENT

Good morning, thank you for participating in this activity. You are taking part into a study on economic decisions. During the activity, you can, depending on your decisions and on other participants' decisions, earn an amount of money in addition to the 3 euros you will receive anyway.

The answers you give and the choices you make will be totally anonymous. The researchers will not be able neither is their intention to associate your choices and your answers to your name.

At the end of the experiment, we ask you to fill a short questionnaire, after which we will proceed with the payment, that will occur in cash and privately. During the activity you are paired with another participant. You will not be informed of other's identity, and the other will not be informed about your identity

During the activity you cannot communicate with other participants (under penalty of exclusion from the experiment) and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. You can find a copy of the instruction on your desk. You can check them in any moment during the activity. If you have any questions, please ask the researchers.

Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0.15 euros.

The activity consists of three stages. Stage 1 will be explained to you subsequently, after presenting Stage 2, because for the decisions to be made in Stage 1, it is required to know the procedure of Stage 2.

Stage 2:

In Stage 2, you will be asked to perform a task. The task is the same for all the participants and it determines the earnings of Stage 1. You will be presented a series words and you will be asked to produce words by substituting the letters of alphabet with numbers, using the Table 1. For example, if the word that appears on your screen is "HELLO" you must enter the numbers 24 for "H", 14 for "E", 25 for "L", 25 for "L" and 21 for "O".

For each word produced through a correct encoding you will receive 1 token. The words are the same for all the participants. You will be given a time limit and within this limit you can produce as many words as you can. You and the other participant are given two different time limits. One of you will have 10 minutes at his/her disposal, while the other will have 6 minutes. The assignment of time limits is random and it is made by the software without any intervention by the experimenter. Thus, you have a probability of 50% of getting 10 minutes and 50% of getting 6 minutes.

If you are the participant with the 6 minutes limits, a game will appear on your monitor at the end of the task. This activity is a simple entertainment, not related to the experiment, and it does not produce

any earnings. It is introduced only with the aim of not allowing the identification of people with lower limits.

At the end of the task you and the other participant will be informed about the words coded by the person who had ten minutes during the first six minutes and the words coded during the last four minutes; the productivity of the person who had 10 minutes, calculated in words per minute, in the first six minutes and in the last four minutes; words coded by the person who had six minutes; productivity, calculated in words per minute, of the person who had 6 minutes; and the total product of the pair, corresponding to the sum of your both productions:

total product = your product in the task + other participant's product in the task

(Remember that each word produced through the encoding activity corresponds to 1 token)

Stage 1:

In Stage 1, both you and the other participant will be asked to agree on a distributive criterion to be applied for sharing the total product, achieved in Stage 2 (the stage described above).

The agreement procedure consists of three phases of choice of one of the five rules reported on the last page of the instructions.

The first phase consists of six rounds of simultaneous choices. During these six rounds, both you and the other participant choose a rule simultaneously, without knowing the other's choice. If, during this period, both you and the other participant choose the same rule, you reach an agreement on that rule, the agreement procedure finishes, and Stage 2 of the experiment begins. If, during these six rounds, an agreement is not reached, the procedure continues into the second phase.

Four sequential rounds are presented as follows. You or the other participant is randomly selected by the software:

If the selected person is A and the other person is B, A proposes one rule to B; B can accept or reject that rule.

If B accepts, then the agreement is reached, the procedure finishes, and Stage 2 of the experiment begins.

If B rejects, then B is asked to make a counteroffer, by proposing a rule different from the one that A suggested.

The sequence is repeated four times, so that A can make two offers and B can also make two offers, so each one has two opportunities to accept or reject the other's offer.

If during the fourth round, where B make the second counteroffer, A accepts the proposed rule, so the agreement is reached, the procedure finishes, and Stage 2 of the experiment begins.

If A rejects that rule, the sequential rounds are terminated, and both you and the other participant go to the last phase of the procedure, consisting of three rounds working as the first six rounds described above.

If the agreement is not reached at the end of these last three rounds, both you and the other participant are excluded from the subsequent stages of the experiment. Participants that are excluded from the experiment have to wait up to the end of the experimental session to be paid (for participation).

Stage 3:

At this point you both will be asked to decide how to divide the total product (and the corresponding earning) generated through the activity. You can decide how to divide the total product by confirming the rule agreed in Stage 1 through the bargaining procedure, by choosing a different rule among the

five reported on the last page of the instructions or by a indicating any combination of percentages (only integers from 1 to 100) indicating a division of the product between you and the other participant.

The software will extract at random you or the other participant and the division decision made by the extracted person will be used for the final division of the product. The probability of being extracted is 50%.

SUMMARY OF STAGES

Stage 1: Decision procedure for choosing a division criterion. You can select one of the five rules described on the last page of the instructions. Reaching an agreement on one of the five rules within the 13 rounds (simultaneous and sequential) is required to enter the Stage 2.

Stage 2: The task. Both you and the other participant are informed about your time limits, perform the task and are informed about the number of words produced through the task by your pair.

Stage 3: The division. Both you and the other participant choose how to divide the total product. You or the other participant is extracted and the decision of the extracted person is implemented. Final earnings are computed by associating 1 token to each word and you are paid.

INDIVIDUAL CHOICE TREATMENT

Good morning, thank you for participating in this activity. You are taking part into a study on economic decisions. During the activity, you can, depending on your decisions and on other participants' decisions, earn an amount of money in addition to the 3 euros you will receive anyway.

The answers you give and the choices you make will be totally anonymous. The researchers will not be able neither is their intention to associate your choices and your answers to your name.

At the end of the experiment, we ask you to fill a short questionnaire, after which we will proceed with the payment, that will occur in cash and privately. During the activity you are paired with another participant. You will not be informed of other's identity, and the other will not be informed about your identity

During the activity you cannot communicate with other participants (under penalty of exclusion from the experiment) and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. You can find a copy of the instruction on your desk. You can check them in any moment during the activity. If you have any questions, please ask the researchers.

Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0.15 euros.

The activity consists of three stages. Stage 1 will be explained to you subsequently, after presenting Stage 2, because for the decisions to be made in Stage 1, it is required to know the procedure of Stage 2.

Stage 2:

In Stage 2, you will be asked to perform a task. The task is the same for all the participants and it determines the earnings of Stage 1. You will be presented a series words and you will be asked to produce words by substituting the letters of alphabet with numbers, using the Table 1. For example, if the word that appears on your screen is "HELLO" you must enter the numbers 24 for "H", 14 for "E", 25 for "L", 25 for "L" and 21 for "O".

For each word produced through a correct encoding you will receive 1 token. The words are the same for all the participants. You will be given a time limit and within this limit you can produce as many words as you can. You and the other participant are given two different time limits. One of you will have 10 minutes at his/her disposal, while the other will have 6 minutes. The assignment of time limits is random and it is made by the software without any intervention by the experimenter. Thus, you have a probability of 50% of getting 10 minutes and 50% of getting 6 minutes.

If you are the participant with the 6 minutes limits, a game will appear on your monitor at the end of the task. This activity is a simple entertainment, not related to the experiment, and it does not produce any earnings. It is introduced only with the aim of not allowing the identification of people with lower limits.

At the end of the task you and the other participant will be informed about the words coded by the person who had ten minutes during the first six minutes and the words coded during the last four

minutes; the productivity of the person who had 10 minutes, calculated in words per minute, in the first six minutes and in the last four minutes; words coded by the person who had six minutes; productivity, calculated in words per minute, of the person who had 6 minutes; and the total product of the pair, corresponding to the sum of your both productions:

total product = your product in the task + other participant's product in the task

(Remember that each word produced through the encoding activity corresponds to 1 token)

Stage 1:

In Stage 1, you will be asked to choose a rule to be applied for sharing the total product, achieved in Stage 2 (the stage described above). The choice refers to one of the five rules you can find on the last page of the instructions, which define a distribution of the total product.

Stage 3:

At this point you both will be asked to decide how to divide the total product (and the corresponding earning) generated through the activity. You can decide how to divide the total product by confirming the same rule chosen in Stage 1, by choosing a different rule among the five reported on the last page of the instructions or by a indicating any combination of percentages (only integers from 1 to 100) indicating a division of the product between you and the other participant.

The software will extract at random you or the other participant and the division decision made by the extracted person will be used for the final division of the product. The probability of being extracted is 50%.

SUMMARY OF STAGES

Stage 1: Decision procedure for choosing a division criterion. You can select one of the five rules described on the last page of the instructions.

Stage 2: The task. Both you and the other participant are informed about your time limits, perform the task and are informed about the number of words produced through the task by your pair.

Stage 3: The division. Both you and the other participant choose how to divide the total product. You or the other participant is extracted and the decision of the extracted person is implemented. Final earnings are computed by associating 1 token to each word and you are paid.

RULE OTHER INFORMATION TREATMENT

Good morning, thank you for participating in this activity. You are taking part into a study on economic decisions. During the activity, you can, depending on your decisions and on other participants' decisions, earn an amount of money in addition to the 3 euros you will receive anyway.

The answers you give and the choices you make will be totally anonymous. The researchers will not be able neither is their intention to associate your choices and your answers to your name.

At the end of the experiment, we ask you to fill a short questionnaire, after which we will proceed with the payment, that will occur in cash and privately. During the activity you are paired with another participant. You will not be informed of other's identity, and the other will not be informed about your identity

During the activity you cannot communicate with other participants (under penalty of exclusion from the experiment) and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. You can find a copy of the instruction on your desk. You can check them in any moment during the activity. If you have any questions, please ask the researchers.

Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0.15 euros.

The activity consists of three stages. Stage 1 will be explained to you subsequently, after presenting Stage 2, because for the decisions to be made in Stage 1, it is required to know the procedure of Stage 2.

Stage 2:

In Stage 2, you will be asked to perform a task. The task is the same for all the participants and it determines the earnings of Stage 1. You will be presented a series words and you will be asked to produce words by substituting the letters of alphabet with numbers, using the Table 1. For example, if the word that appears on your screen is "HELLO" you must enter the numbers 24 for "H", 14 for "E", 25 for "L", 25 for "L" and 21 for "O".

For each word produced through a correct encoding you will receive 1 token. The words are the same for all the participants. You will be given a time limit and within this limit you can produce as many words as you can. You and the other participant are given two different time limits. One of you will have 10 minutes at his/her disposal, while the other will have 6 minutes. The assignment of time limits is random and it is made by the software without any intervention by the experimenter. Thus, you have a probability of 50% of getting 10 minutes and 50% of getting 6 minutes.

If you are the participant with the 6 minutes limits, a game will appear on your monitor at the end of the task. This activity is a simple entertainment, not related to the experiment, and it does not produce any earnings. It is introduced only with the aim of not allowing the identification of people with lower limits.

At the end of the task you and the other participant will be informed about the words coded by the person who had ten minutes during the first six minutes and the words coded during the last four

minutes; the productivity of the person who had 10 minutes, calculated in words per minute, in the first six minutes and in the last four minutes; words coded by the person who had six minutes; productivity, calculated in words per minute, of the person who had 6 minutes; and the total product of the pair, corresponding to the sum of your both productions:

total product = your product in the task + other participant's product in the task

(Remember that each word produced through the encoding activity corresponds to 1 token)

Stage 1:

In Stage 1, you will be asked to choose a rule to be applied for sharing the total product, achieved in Stage 2 (the stage described above). The choice refers to one of the five rules you can find on the last page of the instructions, which define a distribution of the total product.

Stage 3:

At this point you both will be asked to decide how to divide the total product (and the corresponding earning) generated through the activity. Before continuing, you will be informed of the choice made by the other participant in Stage 1 and the other person will be informed of the choice you made in Stage 1 as well. You can decide how to divide the total product by confirming the same rule chosen in Stage 1, by choosing a different rule among the five reported on the last page of the instructions or by a indicating any combination of percentages (only integers from 1 to 100) indicating a division of the product between you and the other participant.

The software will extract at random you or the other participant and the division decision made by the extracted person will be used for the final division of the product. The probability of being extracted is 50%.

SUMMARY OF STAGES

Stage 1: Decision procedure for choosing a division criterion. You can select one of the five rules described on the last page of the instructions.

Stage 2: The task. Both you and the other participant are informed about your time limits, perform the task and are informed about the number of words produced through the task by your pair.

Stage 3: The division. Both you and the other participant are informed of the reciprocal choices made in Stage 1. Both you and the other participant choose how to divide the total product. You or the other participant is extracted and the decision of the extracted person is implemented. Final earnings are computed by associating 1 token to each word and you are paid.

In a few minutes, we will ask you to answer a few control questions. They will help you to verify whether the instructions are clear to you. Before the control questions, you have the opportunity to practice with the five rules that can be chosen (along with any combination of percentage) to divide the product (as previously explained).

In this simulation you can introduce the following parameters: the words coded by the person who had 10 minutes during the first 6 minutes; the words coded by the person who had 10 minutes during the last 4 minutes (the sum of these parameters corresponds to the total words encoded by the person who had 10 minutes); the words coded by the person who had 6 minutes; the person who is selected (the one who had 10 or the one who had 6 minutes). Furthermore, you can choose what rule to apply by clicking on it. After introducing these parameters and choosing one rule, you can click on "calculate" to see how the total product is divided in each case. We invite you to try the result of different rules.

CONTROL QUESTIONS

1. You produce a total of 21 words, the other participant produces a total of 20 words:

Your earning from the task is of … tokens.

The other participant's earning is of … tokens.

2. The total product of your pair is 70 words. You have chosen the first rule to divide it. The other participant has chosen the second rule.

If you are extracted:

You obtain a part of the total product of … and your earning is of … tokens;

The other participant obtains a part of the total product of … and his/her earning is of … tokens.

If the other participant is selected and the second rule is used to divide the total products:

If you are extracted, according to the procedure of the second rule, you obtain a part of the total product of ... and your earning is of ... tokens;

If the other participant is extracted, according to the procedure of the second rule, s/he obtains a part of the total product of ... and his/her earning is of ... tokens.

3. You produced 60 words and the other participant produced 40 words. Both you and the other participant have chosen the third rule to divide it:
You obtain a part of the total product of ... and your earning is of … tokens;

The other participant obtains a part of the total product of … and his/her earning is of … tokens.

4. You had 10 minutes to perform the task and the other participant 6 minutes. You produced 50 words in the first 6 minutes and 20 words in the remaining 4 minutes. The other participant produced 30 words in his/her 6 minutes. According to the rule number 4:
You obtain a part of the total product of … and your earning is of … tokens;

The other participant obtains a part of the total product of … and his/her earning is of … tokens.

5. The total product of your pair is 70 words. Both you and the other participant have chosen rule 5 to divide it.
If you are the more productive during the first 6 minutes you obtain a part of the total product of … and your earning is of … tokens;

If the other participant is the more productive during the first 6 minutes s/he obtains a part of the total product of … and his/her earning is of … tokens.

Table 1.

Letter	Number
A	6
B	26
C	13
D	3
E	14
F	19
G	10
H	24
I	2
J	20
K	5
L	25
M	9
N	17
O	21
P	1
Q	11
R	8
S	4
T	18
U	22
V	12
W	16
X	7
Y	23
Z	15

V. *A Rational Justification for Gilbert's joint commitment by using the Rawlsian constructivist method*

1. Introduction

In the previous chapters, I have exposed the results of two experiments linked by the conjecture that an explicit prior agreement between the parties turns out to be a fundamental condition to both collectively deliberate and conform to the liberal egalitarian rule.

In this chapter I would like to explore the empirical results by broadening the notion of joint commitment introduced by Margaret Gilbert. In particular, I would like to argue that an explicit prior agreement jointly commits subjects to decide and apply the rule of liberal egalitarianism, while that same rule would not be the rational choice in the absence of agreement. The procedure behind a veil of ignorance would bring out the conditions of impartiality and impersonality, under which the agreement takes place. This would mean that persons involved in a deliberative process behind a veil of ignorance would find it rational to choose the liberal egalitarian rule, as the result of a collective pondering of different conceptions of justice –here represented by the appropriate distributive rules. The deliberative procedure finds in the explicit agreement the necessary condition for the emergence of a joint commitment which is an expression of mutual willingness and readiness to comply with what has been collectively decided (i.e. the content of the agreement itself).

The notion of joint commitment plays a central role in the structure of social phenomena, the main object in Gilbert's social ontology. This notion is such that it can include and define all those

collective actions brought about by two or more people. Gilbert often expresses that social phenomena are to be understood in terms of everyday language and practices: the representation of how a commitment between several parties is perceived by themselves is different if compared with the same action carried out simultaneously from many agents.

It is a perception that people *actually* have when they perform a collective action together with other people. According to her argument, an explicit prior agreement is not a necessary condition for collective compliance. The joint commitment, *per se*, determines conformity. It does not matter where that joint commitment comes from. Gilbertian theory would not be completely clear about the relationship between joint commitment and explicit prior agreements, especially if the context creates dissent among people, such as how to behave in production situations where there is neither unique way to intend what is fair nor a single shared conception of fairness (Konow 2000). Studying this dynamic can no longer be limited to an exclusive philosophical speculation, but it must also take advantage from empirical evidence that actually observes how people behave and what their actions are motivated by. This is why the empirical results will be provided to enlarge Gilbert's notion of joint commitments by using the constructivist method by John Rawls.

In this chapter, I would like to argue that an explicit prior agreement would be a fundamental condition for compliance with liberal egalitarianism as a feasible method to propose a normative solution for allocation problems: the liberal egalitarian rule could be considered as the normative solution of an allocation problem because it nullifies the arbitrary effect of luck, bringing about a conception of justice as fairness that recognizes people as free and equal, capable of moral powers. By 'liberal egalitarianism', I refer to the concept of distributive justice set out in the previous chapters. The attempt is to claim how the liberal egalitarian distributive rule works, how it might be a moral realistic argument, when resulting from an explicit prior agreement. The deliberation process between

the parties would guarantee the impartiality and impersonality of the result – that is, the agreement on liberal egalitarianism as the 'winning' conception of distributive justice among those proposed. Collective deliberation and the corresponding consensus would create the conditions for the parties to perceive themselves, in Gilbert's terms, as a plural subject. This perception, added to the mutual awareness that the agreed content is not imposed from the outside but it is the achievement of a collective decision, would motivate individuals not only to commit themselves to respect it, but to comply with it, *de facto*.

Gilbert identifies the notion of joint commitment as the basic element of all social phenomena. Explicit agreements would be just one type of phenomenon through which joint commitments may be brought about. Against this view, I would like to argue that a joint commitment such as to motivate individuals to comply with the liberal egalitarian distributive rule in production contexts might be possible only if preceded by an explicit agreement behind a veil of ignorance. This means that the agreement must respect the conditions elicited by the veil of ignorance: impartiality and impersonality. Further compliance would be motivated not only because people feel jointly committed, but also because they have rationally justified the decision to be jointly committed on the agreed rule. In situations in which cooperation is costly, an explicit prior agreement behind the veil of ignorance could give a basis to overcome certain impulses (what Gilbert calls 'personal inclinations'), as source of motivation –*pacta servanda sunt*. The contractual procedure theorized by John Rawls would be vindicated under a new conceptualization: the impartial agreement as a justified joint commitment phenomenon.

To deploy my argument, I will present the grounding concept of Margaret Gilbert's plural subject theory: that is, the notion of joint commitment. Secondly, I will present how Gilbert conceives agreements and what kind of obligations derive from them. Thirdly, I will refer to some experimental

results (Chapter IV of this thesis) that would reveal how an explicit prior agreement behind the veil of ignorance would be determinant for the emergence of a joint commitment and its relative compliance. Given the experimental results, I will try to show how the Rawlsian contractarian argument would not remain so ideal as it has been criticized, but it might have practical implications on what people consider fair in distributive contexts, providing actual reasons for behaving accordingly. Some conclusions and suggestions for further research will follow.

2. Joint commitment and plural subject theory.

Margaret Gilbert, a pioneer in the field of social ontology (Epstein 2016, 2018), argues that to understand social phenomena there is a need for an analysis of the structure on which they are based: the foundational atom of human social behaviour is the joint commitment. This statement represents the heart of her plural subject theory and deserves to be explored within Gilbert's framework. First of all, her research focuses on collective actions, defending the position of the collective intentions' irreducibility to a mere sum of individual intentions. There would be a different mental state, something that belongs exclusively to social phenomena, hence collective, which allows subjects to feel like members of a community rather than separate individuals performing certain action simultaneously. A collectivity is defined in terms of plural subject, whose members cannot participate in a genuinely social action if they do not feel themselves as parts of that subject. It should be underlined that Gilbert's plural subject theory was presented, in its first delineation, in *On Social Facts* (1989), where the reference to Emile Durkheim's sociological theory is made explicit: "The title is meant to recall Durkheim's view that the phenomena particularly apt for the label 'social' are

significantly special" (Gilbert 1989, p.2). Hence, Gilbert's holistic notion of plural subject theory is of direct Durkheimian derivation: "If, as is granted, this *sui generis* synthesis which constitutes every society gives rise to new phenomena, different from those which occur in consciousnesses in isolation, one is forced to admit that these specific facts reside in the society itself that produces them, and not in its parts, that is to say, its members. In this sense therefore they lie outside individual consciousnesses (*consciences individuelles*) considered as such" (Gilbert 1989, p. 250).

However, Gilbert's social ontology distances from Durkheim's view for two main reasons. Firstly, in Durkheim's refusal to use everyday concepts: according to Gilbert, using this type of concept does not mean lacking scientificity, rather allowing a deeper understanding of social phenomena. Secondly, in how to understand the notion of holism.

Durkheim, following Schäffle, underlines the idea that there is a radical difference between the life of an organism and that of society. While the life of an organism is mechanically regulated, society is united not by a material relationship, but by bonds of ideas. So, society has its own and specific characteristics that are distinct from those of the individuals who compose it. According to Schäffle, and so for Durkheim, "society is not a simple aggregate of individuals, but has an existence that precedes those who make it up and such that will survive them; society affects them more than it is affected by them and has its own life and conscience, its own interests and destiny" (Giddens 1971, p. 113, English translation by the author).

Durkheim is interested in analysing political dynamics, especially focused on the division of labour, disparity between social classes, individualistic culture. Society, in his perspective, assumes a twofold connotation depending on whether it manifests a mechanical or an organic solidarity. Organic solidarity does not derive from the mere acknowledgement of a set of common beliefs and feelings, but from functional interdependence in the division of labour. Whereas mechanical solidarity is the

main basis of social cohesion, and therefore presupposes the similarity between individuals; organic solidarity presupposes difference in beliefs and behaviours amongst individuals. The development of organic solidarity and the expansion of the division of labour are therefore accompanied by the growth of individualism (Giddens 1971, p. 123, English translation by the author).Undoubtedly, Durkheim considered social phenomena as a synthesis, as inherent in special social groups that cannot be reduced to the mere aggregation of separate individuals, but sometime his idea of a collective consciousness would suggest that "in social situations, it is not the individual who decides and acts, but rather the collective consciousness who determines the course of action, and acts through the individual" (Schweikard and Schmid 2013). According to Gilbert's holistic view, instead, there is no ontological difference in the type of social cohesion – depending on the type of solidarity. In both mechanic and organic forms of solidarity, she identifies the conditions for the appearance of a plural subject, that is, a group of two or more people sharing an intention to do X that collectively undertake to achieve it. Durkheim is more interested in studying the moral, social, political, economic consequences of a society in which different types of solidarity can coexist, while Gilbert wants to answer the ontological question "under what conditions do we count a set of human beings as a collectivity, or social group?" (Gilbert 1989, p.2).

In order to do something together, the constitutive members of a plural subject must share a common goal, publicly recognized by the parties. This shared aim becomes the motivational source that allows members' individual actions to converge: having a common goal modifies individual behaviours, since each member of the plural subject, in the relevant contexts, will act in accordance with the common goal (Tossut 2018). This kind of goal presupposes a specific willingness: collective intention as a requirement to enter the mental state of doing something together. As Gilbert says: "Persons X

and Y collectively intend to perform action A (for short, to do A) if and only if they are jointly committed to intend as a body to do A" (Gilbert 2006, p.5).

It is important to underline, in order to avoid misunderstandings, that the expression 'as a body' – elsewhere we find 'as a single body' or 'as a unit'– is to be understood in a specific way. Gilbert does not claim that collective action gives rise to a sort of super-entity, a subject that eliminates the different individual characteristics. These types of expressions support her holistic idea that collective actions are not reducible to the decisions and actions of individuals. Note that saying that subjects act 'as if' they were a single body would refer to that particular condition in which people find themselves when they partake of a collective action. Gilbert insists on how doing something together implies a different perception by those who make up the group. In this sense her holistic perspective is a point of view that highlights that special 'glue' people have and feel when they are part of a plural subject. The term that creates ambiguity is 'collective', around which different positions have emerged on how to intend a collective intention. This is not the place for a discussion of the different modalities dealing with that adjective, just remember that Margaret Gilbert defends the position that collective intentionality is deemed to be a subject, while the other two major currents maintain that intentionality refers to a content (Bratman 1999) or to a mode (Tuomela and Miller 1988; Tuomela 2007). For Gilbert, the main feature lies in the type of perception that people feel when they consciously form a social group. The fact that people perceive themselves as a group and not as a mere bunch of separate individualities, would allow them to see the group itself as the subject of intentionality. The group intends to do such-and-such: this means that the group's members are aware that their joint action is the action they intend to perform as a group. Thus, the subject would emerge from mutual awareness and readiness to perceive themselves as a group intending to act as determined. Potential ambiguities and misunderstandings are inherent in expressions such as 'as a body'. Gilbert is advocating a position

very close to common sense: recognizing that I am part of a group presupposes a type of perception that differs from considering myself as a single person who acts together with another individuals, at the same time, but moved by different intentions. Persons X and Y collectively intend to perform action A, not only if they feel 'as a body' but under the necessary condition of being jointly committed to intend A as a body.

In Gibert's theory, acting together is based on the concept of the joint commitment. To underline the relevance that this notion assumes within the theory, it is worth following the steps that the author presents up to the formulation of the jointness of the commitment. Whether they are joint or personal commitments, their specific object is human willingness: "the commitment is of the will, by the will" (Gilbert 2014, p.84). The centrality of the role played by intentionality, therefore by the will, is clearly expressed in *On Social Facts*, in which it is said that every person "must make clear his willingness that the goal in question be accepted by himself and the other, jointly […] One might say that each must manifest his willingness to constitute with the other a plural subject of the goal that they travel in one another's company" (Gilbert 1989, p. 163).

The centrality of the will leads to understanding why a commitment should be able to constrain individuals. Gilbert identifies two binding aspects of the will: reason and persistence. The intention to act constrains our will to the point of behaving – so from mental state to action – only if we have a reason, so the action is motivated. This is also the case when the decision is the result of a collective intention: each member of the group has reasons to act in accordance with what is jointly deliberated. The idea is that if a collective decision on how to act in a certain situation reaches unanimous consent, each individual, as a deliberative part of the group, has reason to comply.

Gilbert uses the term 'reason' in a broad and intuitive sense: commitment, regardless of being personal or joint, is an exercise of will. If, for instance, I decide to become a vegetarian, what is

rational to do, given my prior decision, is not to eat animal-derived foods. I create a commitment with myself, I express to myself the readiness to pursue this goal, so the intention to become a vegetarian will have a real outcome if and only if I will follow a vegetarian diet. The second aspect is the permanence in time: this aspect concerns above all those decisions that may take some time before implementation. If the reason for which I have committed myself is still valid over time, I will keep the decision to behave in a certain way. If it is no longer valid – for example, given the same circumstance I changed my preferences – I withdraw from the commitment. Deciding to leave a commitment implies a lack or a lower level of persistence of will. Returning to the example, the decision to become vegetarian implies not only a change in my diet, but also its maintenance over time. Suppose I decide to follow vegetarianism for a month, then, I create a commitment with myself not to eat food of animal origin for one month. This personal commitment becomes effective if and only if I exercise my will to achieve the goal without rescinding the commitment.

Gilbert proposes many definitions of joint commitment. The following, however, strongly underline its intrinsic aspects: "[...] all joint commitments are joint commitments to do something as a body, where "doing something" is construed broadly so as to include such psychological states as belief, the acceptance of a rule or principle of action, and so on. The phrase "as a body" is not sacrosanct. One might use other phrases such as "as a unit" or "as one". […] The relevant joint commitment is an instruction to the parties to see to it that they act in such a way as to emulate as best they can a single body with the goal in question" (Gilbert 2014, pp. 32-33).

The very binding nature of personal commitments, characterized by reason and persistence, also constitutes the obliging nature of joint commitments. These ones replicate the mechanism created on a personal level with a substantial difference: in the case of the formation of a joint commitment, the subject is plural, therefore "each party must make clear to the others that he is ready to be jointly

committed [...] Thus each must express a certain condition of his will" (Gilbert 2014, p. 86).The fact that the subject is plural implies that the parts of which it is composed publicly recognize their readiness to act as decided, until proof to the contrary. So, unlike the binding nature of personal commitments, that of joint commitments, according to Gilbert, "has sufficient reason to conform to the commitment subject to the appropriate exercise of both his own will and that of the other parties [...] each is obliged to the other to conform to the commitment and has a right against the other to his conformity" (Gilbert 2014, p. 88).

According to Gilbert, personal commitments have a binding nature, in a weaker sense, because they do not oblige. If I am going to do something, then I express to myself the intention of doing that particular action. One could say that I make a commitment with myself to put into practice what I have mentally decided. Having an intention to take a certain course of action puts me in a position to find a way to reach it, so I will uphold myself. This type of commitment is personal because, as much as I would feel bound to do what I told myself to do, I can always change my mind, by failing to my intentions. Personal commitments bind because it is supposed that, broadly speaking, if someone has the means to do something that she has decided to do, it would be rational to put it into practice. But she is not obliged towards herself: obligations arise when there is at least one other person with whom she commits herself. The difference between personal and joint commitments lies precisely in how to consider the (collective) subject and in what kind of (joint) intention emerges. According to Gilbert, in fact, when the commitments are joint, their nature is not only binding but also obliging. Situations where the subject is plural would be ontologically different, because there are at least two people – so two intentions –who want to do something as if they were one single subject. Gilbert calls this view intentionalism, that is: "the view that according to our everyday collectivity concepts, individual human beings must see themselves in a particular way in order to constitute a collectivity. In other

words, intentions (broadly construed) are logically prior to collectivities […] Human beings appear to be in an important sense powered by their ideas and views of their situation" (1989, p 12).

Remember that when we talk about commitments, Gilbert always refers to the everyday meaning: intuitively we understand that if we commit ourselves to intend to X, this commitment has a different level of obligation compared to a commitment such that it does not include other people. It is a matter of perception: you and I can walk along the Tiber, without knowing each other, without realizing that we are walking nearby. Someone who sees us might think that we decided to take a walk together, but, actually, that is not the case, because you and I are only fulfilling some personal commitments to take a walk along the Tiber. My motivation could be to get out for some fresh air, your motivation could be coming home from work. As far as it may seem to external eyes that we have decided to walk together given, suppose, our closeness, there is neither collective intention nor commitment, there is no 'we'. I can decide to give up on my initial intention of walking 10 km and after 2 km comeback. I broke-up my personal commitment to myself. If you and I, however, decide to meet at noon and take a walk together along the Tiber, subject's intention refers to the collectivity made up of you and me. If we express our (*we*-)intention to meet each other, then we have a different type of motivation, namely we have an obligation to fulfil the joint commitment. We perceive that otherness, the fact of feeling you and me as 'we', affects how we should behave.

It could be said that the joint commitment creates a motivational source such that the parties are motivated to act with what was established by the commitment itself. Such statements may create misunderstandings if we do not take into account Gilbert's framework. Even if she talks about will, obligations, rights and compliance, her goal is not to solve the contractual problem, neither how to

get out of the state of nature, –a problem according to which mutual promises[25] become sources of obligation only if a coercive power is established among those who make the contract. Her broader objective, although she also deals with the theory of political obligation, is to define, through the concept of joint commitment, the distinctive and exclusive ontological status of social phenomena.

The fact that, once a joint commitment is created, everyone has reason to comply with the previous collective decision, could work on a theoretical level, but would seem to encounter obstacles in practice. If, for instance, we define social dilemmas as "situations in which each member of a group has a clear and unambiguous incentive to make a choice that when made by all members provides poorer outcomes for all than they would have received if none had made the choice "(Dawes and Messick 2000, p.111), then the collective intention to do anything may prove motivationally weak when the time comes for each individual to behave as intended. In the classic case of the prisoner's dilemma (Gauthier 1986), if agents have the opportunity to agree on how to act, they will prefer to coordinate. When they must decide whether to comply with what was previously decided, the subjects are in a position to choose individually. In these terms, the prisoner's dilemma (Gauthier 1986) will result in a non-cooperative game, in which deviating can benefit the individual over the group. It is a question whether reaching an ex-ante agreement generates a type of collective intentionality that binds the subjects to comply with the agreed content. As it will be explained in the next section, the Rawlsian approach sees in the contractarian argument (i.e., the explicit prior agreement behind the veil of ignorance) the grounding condition in generating the "motivational forces able to induce the adoption of potentially counter-interested behaviour" (Faillo et al 2015).

[25] On agreements vs. promise-exchange see Gilbert, M. (1993). Is an agreement an exchange of promises?, *Journal of Philosophy*, 60 (12), pp. 627-649.

Before proposing a reading of some experimental data consistent with the Rawlsian method, it is worth exploring the connection that Gilbert proposes between joint commitment phenomena and agreements. With Gilbert, I argue that agreements are a primary source of obligation, however, against her view, I would sustain that agreements do not emerge "from intuitive understandings as opposed to empirical observations or moral judgments" (Gilbert 2006, p.57). It is true that Gilbert, as she says several times, intends to analyse the everyday concept of what agreements are, but in doing so, she also refers to some well-known social contract theories, such as the Rawlsian one.I would like to show how, by virtue of the joint commitment's definition, *per se*, the impartial agreement behind the veil of ignorance would give sufficient reason to conform. Furthermore, arising from a collective evaluation and deliberation on what is just, given the circumstances, the agreement would stem from a moral judgment. To achieve it, I begin by presenting Gilbert'sview about agreements.

3. Agreement and obligations.

As already anticipated, when Gilbert speaks of 'agreements' she refers to the ordinary use of the term. The examples she proposes are taken from everyday life, like two people who decide to walk together – in the variants of explicit, implicit or null agreement – or situations in which we agree that you will prepare the dinner and I will wash the dishes. Since it is an acting together, the agreement gives rise to a series of obligations. Notice that Gilbert assumes the term 'obligation/s' broadly, in contrast with the position by H. L. A. Hart (1955), according to which this term should be used more narrowly. From Gilbert's perspective, having an obligation is a sufficient reason for one to act. From this it follows that the agreements, as joint commitment phenomena, oblige likewise. They therefore

provide direct obligations, establishing what the parties must do simultaneously: each one owes to the other the agreed actions, or each owes the other the expression (the readiness) of doing her part. This type of obligation would belong only to agreements: promises do not engage, simultaneously and directly, to perform any action.

Promises have singular, not plural subject. I can fulfil the promise I made to you, but this does not imply any kind of action on your part. If I break the promise, without saying anything, you will expect my action and you will feel disappointed if the promised act does not occur. The subject, however, remains a single individual who has exposed himself by making a promise, without being able to bring it about. Gilbert (1989, p.380) claims:

"Promises are made by one person to another (or to several others at once). The promiser in effect binds himself to another. The binding, meanwhile, is unconditional and achievable independently of anyone else's vows. If I make my promise first, I am then obliged to keep it, regardless of the fact that you make a promise in exchange or not. Agreements meanwhile are specifically devices whereby a set of persons (minimum two) can achieve the result that all are bound simultaneously and interdependently to enter upon a certain course of action".

Promise-exchange indeed lacks three characteristics that the agreements have. If two people decide to agree, they both will have a mutual obligation to perform the established action, a common Gilbert's example is agreeing that you prepare the dinner and I wash the dishes. Entering into this type of agreement generates obligations – reason to act – that the promise-exchange does not imply. Firstly, these obligations are unconditional in form and, secondly, they "are arrived at simultaneously" (Gilbert 2000, p. 63). The promisor can always fail because it is a matter of personal decision (whether or not fulfil the promise). The promisee will figure it out and, therefore, she will

not do her part. On the contrary, an agreement implies simultaneity in doing one's part: insofar as everyone agrees to act as decided, no one may be obliged to do such-and-such before another.

The uncontitionality of the obligations deriving from the agreement would emerge simultaneously. In the case of the exchange of promises, on the contrary, the promisee will comply *on condition* that the promisor fulfils his promise. Thirdly, these obligations must be interdependent: if one of the two parties of an agreement fails to comply with the agreed content – therefore without doing her part– the other is exempt from any obligations to perform his action as it was decided and who rescinded the pact has no rights on the other person.

So, agreements are considered as a sufficient condition for activating actual obligations to perform an act: it is about "a *joint decision*, involving a joint commitment to uphold as a body the decision in question" (Gilbert 2000, p. 59). The joint commitment account offers a solid base for the dynamics of collective actions. However, it remains unclear whether the parties to a shared decision would really feel motivated to comply with what was previously decided. In other words, Gilbert (2003, p. 47), assumes that the commitment has "a normative force." It constitutes the source of motivation to do (to intend, to believe, etc.) what was established. It would induce the parties to be ready to act as decided. But the following question arises: What would be the difference between personal decisions' bonds of will and a joint commitment's obligations?

It may be the case that if we talk about agreements referring to structurally everyday situations, a prior explicit agreement is not necessary for the establishment of a collective action. However, without an explicit prior agreement, the obligations arising from the joint commitment would appear to be binding as much as personal decisions. Parties of a joint commitment reached by agreement might withdraw from the joint action as much as individuals can drop off from their personal decisions. Of course, it can be unpleasant when, from one day to another, my partner stops walking

with me with no reason or excuse, but it has no consequences for the society in which both of us live. Totally different issue would be a situation in which the institutions of the city where I live, suddenly become unjust – as judged by an ample majority of citizens – by raising taxing and lowering the salaries. To prevent glaring inequalities, members of a society must not only have the intention to make their institutions fair, but they should also converge on what they deem fair, applying a principle of justice to regulate society.

As highlighted by Dewey (1922), morals enter into everyday individual life and human actions involve decisions about what considering right and what wrong. Not only at the individual level, but also within the social sphere: to solve the allocation problems, for instance, people should first share the same ideal of fairness that they want to use as a criterion. Then, it should be observed whether people behave in accordance with the norm. In this regard, I would like to illustrate how the impartial agreement à la Rawls allows subjects to agree on a specific distributive principle which, intuitively, recalls the liberal egalitarianism and, as a joint commitment between the parties, it gives reasons to act in conformity. An agreement like the one theorized by Rawls would seem to be much more 'actual' than Gilbert thinks. The fact of allowing subjects to impartially reflect about how to split – *as if* they could be the least advantage – would enable the creation of a collective intention to respect both normative and motivating reasons. Referring to this relationship between the Rawlsian agreement and Gilbert's joint commitment, the purpose is to support how the Rawlsian contractarian argument would provide justification to collectively choose and comply with liberal egalitarianism. The conditions for compliance are impartiality and impersonality elicited by the agreement behind the veil of ignorance. This type of agreement might be integrated to Gilbert's notion of joint commitment, in order to provide a rational justification for it.

4. The Rawlsian approach: an integration

The contractarian argument and the notion of joint commitment differ mainly because they apply to two different fields of investigation. In the previous chapters of this dissertation, the contractarian argument has been showed not to be 'ideal'. It was showed that ordinary people, if placed behind a veil of ignorance, would choose and comply with a liberal egalitarian rule. However, in order to be directed towards that specific rule of distribution, choice and conformist conduct require an explicit prior agreement, from which the committing motivations emerge. According to Gilbert, on the other hand, joint commitments would always emerge when there is a plural subject that collectively shares the intention of doing something as if they were a single body. The motivational source for conformity would be constituted by the perception of feeling jointly committed and this would happen even without an explicit prior agreement. The different assumptions and aims between Rawls's and Gilbert's approaches are evident. Distributive justice deals with defining principles that prescribe how to solve allocation problems. The agreement behind a veil of ignorance is a strategy for understanding which principles would be chosen under specific conditions. Or rather, which conception of justice, made up of moral principles, would be preferred by the subjects through deliberative and unanimous consent.

Gilbert's joint commitment would be enlarged by Rawls's contractarian argument insofar as it would be rationally justified through the agreement's procedure behind the veil of ignorance. According to Gilbert, in fact, what is rational to do means, in an intuitive sense, to abide by a decision that has been taken and "the obligations of joint commitment are powerful inputs to reasoning about what to do" insofar as "they trump inclinations and self-interest as such, and they are recalcitrant to a person's

own fiat" (Gilbert 2006, p. 158). Nonetheless, there are many social phenomena where collective obligations do not trump personal inclinations, for instance, when different conceptions of justice emerge and where the incentive to deviate from a well-established norm of fairness, without any kind of social punishment, is a feasible option. One could say that self-interest prevails; that people generally act according to their personal ideal of justice, trying to get, where they can, greater profit/social power/fame. So, how could members of a society feel jointly committed to behave fairly if they do not agree on what fairness requires? Remember that saying that norms of fairness are context-dependent (Bicchieri 2006) does not mean that principles of justice, understood as moral arguments, are context-dependent as well. On the contrary, the principles of justice have an intrinsic moral value: thanks to their universalizability, prescriptivity and overridingness, they would allow to produce that critical exercise – in Rawlsian terms, the reflective equilibrium– to be able to judge when a society, or its institutions, are unjust .

Gilbert's joint commitment, as a motivation source that leads people to comply, would emerge *via* the agreement procedure and it would produce moral obligations towards subjects. The conditions through which the parties reach an agreement on, for example, the liberal egalitarian rule, rationally justify the resultant joint commitment. The unanimous consent would make the contracting parties a plural subject: in other words, they feel as if they were a single body to decide and, therefore, they feel the responsibility to conform. The conditions of impersonality and impartiality allow to evaluate, to give a moral judgment among the various conceptions of justice, by adhering to the liberal egalitarian rule as the best one. From a deontological perspective, evaluating one conception of justice as better than another means considering it as the most just, given the priority of the right over the good. The obligation to comply with what the agreement prescribes, might be considered moral precisely because the agreed content (i.e., principles of justice) has been morally evaluated, through

an impartial reasoning – a primary characteristic for a rational justification of ethical principles. Thus, to the extent that agreements are joint commitment phenomena, their content is of the following form "«act as a body» in a specified way, where «acting» is taken in broad sense. Thus, people may jointly commit to deciding as a body, to accepting a certain goal as a body, to intending ad a body, to believing as a body a certain proposition" (Gilbert 2006, p. 51).

To strengthen the psychological realism of the contractarian argument, experimental evidence would show that, when decisions and relative commitments are personal, the liberal egalitarianism is not the most frequently chosen distributive rule[26]. To succeed in the purpose of being a fundamental condition to couch a joint commitment, the agreement should be reached from an impartial and impersonal point of view, that is, individuals should agree by considering both their personal inclinations and what they ought to do, given the circumstances. To explain this point, I refer to the method of pure procedural justice by John Rawls, also known as Kantian constructivism. The joint decision reached through the agreement would represent the procedure through which the principles of justice are chosen. Deliberating by giving one's own consent would mean expressing one's own readiness to evaluate one principle as the fairest over the others, namely a moral evaluation. Then, by providing a collective judgment that recognizes the priority of one specific principle, the belief about parties' conformity would arise. The motivational force to behave in accordance with the collective decision would not just be supported by the formation of a joint commitment, binding by definition, but also by the method through which the relevant principle is chosen: not as some external imposition but as the result of a collective deliberation, made from an impartial perspective.

If two people must agree on how to distribute a surplus it would seem necessary to share the same conception of justice, otherwise how would the agreement on one distributive rule be justified? In

[26] Individual Choice treatment in the previous Chapter.

addition, assuming that the subjects agree behind the veil of ignorance on a liberal egalitarian principle, will they conform to it when the veil is lifted? What reason should they have for so doing? As Gilbert holds, indeed, there is a huge difference between "doing one's share" and "doing one's *fair* share"[27] and, apparently, there is no motive why subjects should decide and implement a principle like that.

The problem is twofold: firstly, how an agreement about one distributive norm of justice would be justified without a moral shared ground on what is *fair*; and, secondly why the liberal egalitarian rule should be the most frequently ex-ante chosen and ex-post implemented. Answers to both questions rely on the normative role of the agreement and the procedure it implies: rational justification of the principles of justice. The agreement, being publicly recognized, and directly engaging the parties, affects our psychology. It would make people feel part of something and motivates them to action – even for the mere fact of feeling part of something and of owing some action to their partners. Now, there are social phenomena which, *de facto*, are the result of a normative reflection: in productive contexts, the question of what one 'ought to do' emerges spontaneously. It is that component of oughtness that moral principles possess and that provide heuristics to people in their behaviour: I know what it would be the right thing to do on a certain occasion, but I decide to behave differently. This simple reasoning implies a moral consideration – a judgment – about what that situation would behaviourally require.

In relation to the most significant results: (i) the liberal egalitarian principle is the most frequently ex-ante chosen and ex-post complied principle; (ii) ex-ante collective choice on a distributive criterion,

[27] "One can distinguish between "doing one's part" or "doing one's share" and "doing one's fair share" in the sense of "doing what is fair. If people understand themselves to be party to a joint commitment to accept certain rules as a body, then obeying the rules, absent special circumstances, appears to be simply a matter of doing one's part", (Gilbert 2003, pp. 63-64).

although different from the liberal egalitarian one, is held by the fact of reaching an agreement; (iii) people prefer a merit-based criterion to solve the allocation problem, under the same circumstances but without an explicit prior agreement.

Starting from the latter result: it was observed that without the possibility of reaching an agreement, all else being equal, the veil of ignorance was not sufficiently strong neither ex-ante nor ex-post in motivating participants to choose the liberal egalitarian solution. Given the same context of initial unjustified inequality and the effort made, however small, to obtain one's own income, the possibility to reach an agreement behind the veil of ignorance changed subjects' conception of justice, also increasing the overall level of compliance. If an explicit prior agreement were not a necessary condition for binding people to act in accordance with the ex-ante choice, knowing that they were participating in a bargaining game with another person should be a sufficient condition to a joint commitment's creation[28]. However, this was not the case. In the treatments without the possibility of agreement, the liberal egalitarian rule was not chosen behind the veil of ignorance as the preferred one nor implemented once the veil was lifted. Individual preferences appeared to be those for a merit-based rule. It might be that, in *foro interno*, individuals have personal preferences for other kind of distributive criteria, but, once unanimous agreement is jointly reached (agreement expressing collective preferences), individual preferences do not affect conduct. Although extremely simplified, the veil of ignorance was thought as a moral device: the underlying idea was that not knowing which position will be occupied by the subject, hence the condition of ignorance, it should induce a moral point of view from which subjects try to consider what is the right thing to do both for themselves and for the others. As the instructions and the experimental design were presented, even if without

[28] If an explicit prior agreement were not necessary, there would still be the conditions for the formation of a joint commitment arising perhaps from a tacit agreement (given that the subjects receive the same information on the situation they faced with).

the agreement, the idea to fairly distribute the product as the common goal could emerge: if so, there could be conditions for the formation of a joint commitment by implicit or null agreement. Instead, data would show how only a prior agreement behind the veil of ignorance on which criterion to apply would create obligations towards subjects. It may be deduced that the veil of ignorance would have an effect as a moral device only if collectively applied: the parties, in order to agree, must evaluate different circumstance of justice, hence expressing a moral judgment. Since the agreement is the result of the collective deliberation, namely the mutual readiness's expression to behave according to the established norm, the parties would feel obliged to converge.

The second main result would sustain Gilbert's joint commitment account, according to which reaching an agreement is the most common way to create a joint commitment that obligates individuals as if they were a single body, so to act accordingly.

However, this account would not explain the first result data provide. Mostly when subjects were allowed to make the ex-ante agreement via communication, the liberal egalitarian rule was the preferred one. Of course, communication enables participants to share opinions, beliefs and mutual readiness to do something together, but it is not just a matter of a positive fact: the underlying reasoning is normative. That is, collective deliberation would emerge from a normative question such as what should we do, all else being equal? And, as such, normative reasoning needs a rational justification to provide a normative solution that can actually motivate peoples' willingness to act. So, an impartial and impersonal[29] agreement would identify in this principle the normative solution to a problem of distributive justice and it would give reason for conforming to it.

Unlike what Gilbert claims, a moral evaluation would be presupposed and, therefore, the deliberative procedure behind the veil of ignorance becomes a method through which the agreed choice is a

[29] Conditions provided by the procedure behind the veil of ignorance.

normative decision about what ought to do, ceteris paribus. It follows that the liberal egalitarian rule could be considered as the normative solution of the distributive problem[30] because it nullifies the arbitrary effect of luck, bringing about a conception of justice *as fairness* that recognizes people as free and equal, capable of moral powers. The method is constructivist: there is an intention to choose those principles of justice that make the institutions of a society just, such as to nullify inequalities resulting from bad luck in the distribution of primary goods, resources and opportunities. The formation of a plural subject, whose parts express their mutual readiness to opt for a conception of justice *as fairness*, would stem from the deliberative process. Once the distributive norm – namely, what in the previous chapters is called 'liberal egalitarian rule'– has been chosen, compliance with it would follow thanks to the intrinsic motivation arising from the joint commitment. This commitment obliges – in the sense of giving reasons to comply – because the agreed content was rationally justified by the impartial and impersonal procedure behind the veil. In other words, by rationally justifying the choice, the parties would undertake to respect what they agreed on. Justification occurs because subjects express unanimous consent resulting from a moral (i.e., impartial and impersonal) kind of reasoning: the justifying procedure might lead subjects to feel sufficiently motivated to ensure corresponding compliance. If the agreed content (i.e., the collective decision behind the veil of ignorance) is justified, then the emerging joint commitment might become a motivational source for the parties: that is, a normative reason (the liberal egalitarian standard) also becomes a motivating reason. Such an agreement would be the necessary condition that allows the parties to express their willingness and readiness to comply with what was established.

[30] Explained in Chapter IV of this thesis.

5. Final remarks and further research

To summarize, a rational justification of joint commitments is what would seem to be lacking within Gilbert's theory: an explicit prior agreement would constitute a necessary condition to justify why people chose one distributive principle instead of another and why the agreed content becomes binding for compliance. Joint commitments' obligations would motivate people to act as agreed, but without an ethical justification on why people should feel committed there is no reason to follow that obligations. The procedure is such that the subjects, evaluating different criteria of justice and being capable of moral powers, share the liberal egalitarian principle, all other thing being equal. Such an impartial and impersonal agreement would not only be the parties' expression of their willingness to conform, but it would also generate the belief, in the absence of proof to the contrary, that everyone will do her *fair* share. This belief would arise if we consider an impartial agreement as a joint commitment, able to express collective preferences for one ideal of fairness among the competing others. As Gilbert sustains, joint commitment trumps personal inclinations, but this would be possible, relative to distributive justice, only if there is a prior agreement behind the veil of ignorance on what should be considered fair. What is considered fair is the result of a normative procedure that allows people to weigh and evaluate the different conceptions of justice up to one that is collectively approved. This method constitutes the core of John Rawls' Kantian constructivism, according to which the first principles of morality are the outcome of a construction procedure.

Empirical findings about the normative role exerted by the agreement show two things: (1) the agreement creates a joint commitment on any distribution rule; (2) the fact that the procedure, leading to the agreement, meets the requirements of impartiality and impersonality, conditions for the creation of a joint commitment in distributing the surplus according to the liberal egalitarian principle.

Regarding the first point, it was observed that the mere fact of agreeing increases the level of compliance with its content, regardless of what the situation could require as fair. As Gilbert says, the fact of agreeing, no matter on what, constitutes a psychological matter which is binding in its own nature: obligations, as was mentioned, play a role of primary importance since many social phenomena involve obligations of one person to another. According to Gilbert, to make a (personal or collective) decision to do something is a psychological matter, while one's being committed is a normative matter. However, the fact that a joint commitment obligates the parties one to the other to act in accordance with the commitment does not demand moral requirements nor engage in any specifically moral judgments or arguments: "To appeal to obligations of joint commitment, then, does not commit one to moralism. In particular, it is not to say that one is morally required to conform to one's joint commitment - though one may be, all else being equal" (Gilbert 2014, p. 8).

I partially agree with Gilbert that, up to this point, there is no moral requirement needed for the joint commitment account broadly intended: the formation of unconditional, simultaneous and interdependent mutual obligations are sufficient reason to feel part of something in which each has corresponding obligations and rights. This account would be complete enough to include the greater number of social phenomena in which someone owes to the other some action. This gap between the psychological and the normative would be bridged by the joint commitment, by obligating the parties to comply, but what justifies the liberal egalitarian principle as the agreement's content and the resulting obligations?

In situations where concerns for fairness and conceptions of justice have a weight, justifying the content of the agreement would allow to increase the level of compliance and decrease free-riding behaviours. As data revealed, if the agreement satisfies conditions of impartiality and impersonality (i.e., the original position as one of the arguments that Rawls provides for justifying his theory of

justice as fairness), the emergence of the corresponding joint commitment on the liberal egalitarian principle is rationally justified, ensuring the ex-post compliance. This could mean that the liberal egalitarian principle might represent a normative solution to a problem of distributive justice, insofar as it is a moral principle, rationally justified *via* deliberative procedure, such that compliance is held by joint commitment's motivations.

In line with Gilbert, I argued that joint commitments, unlike personal ones, imply a special type of obligations that belongs to agreements. In this regard, she says "a joint commitment trumps a personal commitment, as such, from the point of view of what reason requires" (Gilbert 2003, p. 58). But what is the point of view of practical reasoning? The impartial and impersonal perspective that, through by constructing moral principles and obligations of justice, does not constitute only a normative reason but also a motivating reason for the plural subject who has decided which conception of justice adopt. The reason for this conception would derive from the procedure, from the agreement, which rationally justifies the commitment between the parties, trumping personal inclinations.

Therefore, a normative reason[31]– the liberal egalitarian principle – would become a motivating reason through the agreement procedure that justifies the joint commitment, thanks to the impartial and impersonal deliberation, giving real reason to ex-post conformity. Justification behind the veil of ignorance makes the parties a plural subject, in Gilbertian terms, which express, *via* agreement, their mutual willingness and readiness to comply with what collectively established. But the procedure is brought about as if parties did not know their socio-economic positions: it follows that the liberal

[31] "A normative reason is a reason (for someone) to act—in T. M. Scanlon's phrase, "a consideration that counts in favour of" someone's acting in a certain way (1998 and 2004). A motivating reason is a reason for which someone does something, a reason that, in the agent's eyes, counts in favour of her acting in a certain way. When an agent acts motivated by a reason, she acts "in light of that reason" and the reason will be a premise in the practical reasoning, if any, that leads to the action" (Alvarez, M. (2016). Reasons for Action: Justification, Motivation, Explanation, https://plato.stanford.edu/entries/reasons-just-vs-expl/).

egalitarian principle is chosen by evaluating different conceptions of justice, so, the resulting agreement on it is founded on a moral judgment. Further compliance would be motivated not only because people feel only jointly committed, but also because they feel that the decision to be jointly committed on the liberal egalitarian standard is rationally justified. By offering a reading of these experimental results, Gilbert's theory could be expanded, arguing that the obligations emerging from the joint commitment are moral due to the conditions of impartiality and impersonality typical of the Rawlsian contractarian argument: Rawls's constructivist method provides the moral point of view, namely collective reasoning behind the veil of ignorance, due to impartiality and impersonality, and, at the same time, it gives real motivations to conform because the resulting commitment is justified The envisaged agreement, therefore, would constitute a necessary condition for a rational justification of joint commitments and for compliance with a liberal egalitarian principle.

VI. Conclusions

The experimental designs presented in chapters III and IV present some problematic aspects. Although the experimenters have tried to present instructions in the clearest and most intuitive way, also providing examples of how the rules work, it is possible that some rules are more intuitive than others. This could suggest that another treatment where participants can communicate on what would be the behaviour they want to assume, might be significant.

Regarding the experimental design proposed in the first article some criticism follow. Firstly, it should be clarified that the veil of ignorance is considered as moral cue, as moral learning device, not as anchor. Starting from the NOVEIL treatment, it is good to repeat that the participants knew that two different time limits would be randomly assigned, but they were not asked to express how they would deem fair to distribute the total income. Being this treatment the baseline, the idea was to compare a Dictator game without a veil of ignorance with one in which the veil is introduced: in the latter case, participants had couple of minutes to reflect on how they would distribute the surplus, *ceteris paribus*, hence forming an ex-ante judgment. The aim was to observe whether or not the introduction of the veil of ignorance plays a role not only in the impartial judgment but in the real distribution of the final income. Furthermore, the veil of ignorance should lead participants towards a liberal egalitarian direction (remember that this distribution criterion would allow, if chosen, to nullify the initial unjustified inequalities between the participants insofar as each has 50% probability to be the disadvantaged). Against expectations, the moral cue's function failed because the subjects with the most time available, in the final stage of the game stole from their partners. Another problematic point concerns the reason why subjects should care about justice. Indeed, participants were not asked to express their opinion/belief on how the surplus should be divided, even when the

written rules were presented to them. The device of the veil of ignorance is supposed to elicit norms of fairness because of its implicit impartiality. The case of so-called 'adversarial ethics' should be no different than an economic distributive game. In the case of a play of cards, for example, it is right that the norms within the game imply there is no concern for justice; but the veil may be recalled when a doubt arises about how to solve in a fair way an unexpected circumstance. Depending on your position on the game you may have different opinions *about what is fair* in this case.

The veil of ignorance is introduced as a device that could be able to resolve a conflictual situation of interests, appealing to a moral (i.e., impartial and impersonal) point of view. If it were effective, given that the circumstances are clear and that the norms of fairness are local and context-dependent, reflection on how to split the surplus should help. In other words, the veil of ignorance could be considered as a tool capable of resolving conflicts of personal interests, but, as the results show, this purpose would be possible only through a rational procedure (the impartial agreement), in which the subjects consider the different conceptions of justice, agreeing on a specific one. Its aim would be to put people in the same condition, inducing them to think from an impartial perspective on what it would be rational to do under those circumstances. The underlying idea was that the veil would work by activating some previous idea of fairness –or some sense of justice– that subjects should possess. Another issue is the difference in types (six minutes vs. ten minutes) which could be considered institutional, and therefore, it does not exactly represent the kind of ignorance that Rawlsian original position is supposed to constitute. The unjustified difference in minutes to perform the task would reproduce a mere element of luck –like inheritance or social environment in real life– since luck affects real life situations independently of what people would deserve. Within the stream of experimental literature that tries to isolate the elements of effort and luck, subjects are usually involved in a real-task effort, equal for all, and afterwards, they are subject to what the literature calls

a “random shock”, a random event favouring some players and penalizing others. A problem of introducing arbitrary inequality by a random shock could be to distort individuals’ perceptions. In fact, as highlighted by Erkal et al (2011), the subjects who work harder, reaching the highest earning rank, are more reluctant to give (both in general, and after a random shock that favours them). This problem could be linked to a cognitive bias for which when subjects make an effort, however minimal, good luck that increases their income is seen as deserved. Nonetheless, the presented experimental design tries to avoid, if possible, that effect. The fact of playing with more or less minutes is a purely random factor and, as such, players should put themselves in the other’s shoes, bearing in mind that it was luck that benefited one participant rather than the other, but seeing this as wholly undeserved, since this would happen before the task began. As it happens, our subjects with ten minutes (good luck) claimed a share larger (in proportion) than six-minute types. This is in line with a stream of literature working on determinants in stated-effort and real-effort (Charness et al, 2018) and on how income is generated, by effort and/or by random shock (Rey-Biel et al, 2018; Erkal et al, 2011). Although this was not expected, it confirms at least that our design does capture the idea of good-luck as perceived by subjects. It is interesting that subjects favoured by fortune tend to see their situation as deserved. This may explain the fact that, behind the veil, most subjects imagine themselves in the disadvantaged position. But when the veil is lifted, only those who find themselves disadvantaged behave coherently, while those who are favours by fortune tend to raise their claims.

The different allocation of time is not intended at an institutional level; rather it tries to trace the characteristics of the original position in which individuals do not know the social and economic role that they will fulfil. Once the veil has been lifted, ‘society’ is represented by this working scheme in which some persons are disadvantaged and others advantaged, regardless of their abilities.

Given these issues, a first general conclusion is that the veil of ignorance individually applied does not lead subjects to apply – nor ex-ante neither ex-post – the liberal egalitarian principle.

As mentioned at the end of the second article, the fact that subjects are told to choose a percentage instead of a written rule, as a distribution criterion, might have a cognitive impact. It would appear that the percentage, as a number, leads subjects to consider the situation in a more utilitarian way. The direct consequence is observed on subjects' behaviour, most of them appeared to forget the initial unjustified inequality, by acting selfishly. The importance associated with the way in which people form an idea of what would be the right thing to do, could be strongly linked to the type of goal frame at stake.

A second conclusion of this dissertation would confirm the influence that goals have in priming human conceptions of justice. To this regard, it should be noticed how the provided evidence could support Lindenberg's overarching goal frame theory. According to it, there would be three main goal frames that coexist: depending on the situation, one goal is made salient and the corresponding frame overwhelms the others: "They are the hedonic goal 'to feel better right now,' the gain goal 'to guard and improve one's resources,' and the normative goal 'to act appropriately'" (Lindenberg and Steg 2007, p. 119).

The most unstable goal frame is the normative one, as immediate personal satisfaction or greater gain have a strong motivational force. The authors say to this regard:

"*A priori*, the three goal frames are not likely to be equally strong. The hedonic goal frame, being related to need satisfaction and thus the most basic, is very likely to be *a priori* the strongest of the three goal frames. In other words, it probably needs the least support from the social surroundings of the individual. As Weber

(1946) has shown, the gain goal frame needs institutions (such as religion and/or secure property rights) that allow the individual to act on behalf of a reasonably well-established future self. The normative goal frame is even more dependent on external support, be it through institutions, moralization (see Lindenberg, 1983), or explicit disapproval for not following the norm (see Tangney & Dearing, 2002)" (Lindenberg and Steg 2007, p. 122).

The main problem with the normative goal frame is that, in contexts such as the ones related to distributive justice, the norms of fairness are very abstract and it is not clear whether people have the same conception of justice, all other things being equal.

This research would show the impact that a prior rational agreement, reached behind a veil of ignorance, would have to define what the common goal is – here, the distribution of a surplus in an imperfect situation of arbitrary allocation of endowment. Being able to reach an agreement, by taking the moral point of view and, at the same time, expressing the readiness to comply with the agreed choice, creates the conditions to form a collective intentionality that draws motivational force from the procedure itself. During the decision-making process, namely which criterion to apply, an element of intrinsic oughtness emerges within the procedure: it might be that special type of obligations arising from joint commitments is such that it binds the parties in a substantially and ontologically different way, compared to individual commitments.

By studying the deliberation process on how to distribute a surplus, the agreement behind the veil would make the shared goal focal, producing intrinsic motivational force to comply with it. Without agreement, on the contrary, it is the gain goal frame that takes over.

This relationship between the goal frames and the procedure would deserve further research within the context of cognitive sociology and moral psychology, to see to what extent collective deliberation

behind a veil of ignorance can actually constitute a real motivation for action. This would have social implications, because it would put citizens in a position to decide, to feel collectively committed, to feel responsible. Although external factors play an important role in the dynamics of compliance (Fehr and Fishbacher 2018), the crucial point would remain the one of identifying a cognitive process that gives intrinsic motivation to act, as a way of recovering attention for the common good, increasingly fragile. The planet, for instance, is giving ever stronger signs of how urgent climate change is, an issue from which we cannot ignore by de-responsibilizing. Proposing a solution that might help people to empower, would be the ultimate goal of a project that necessarily needs a further in-depth analysis, taking into account both empirical data on what are people's behaviours and beliefs, and ethical theories able to motivate people in do the right thing, by reducing the gap between abstract norms and individual motivations. Unlike laws, social norms present a deeper and more articulated dynamic: they are not respected only for fear of external sanction. There is of course the psychological cost of deviating, but social norms would seem to be bearers of a different type of obligations. To this regard, Lindenberg and Steg argue that "a feeling of oughtness" belongs to social norms, a feeling that has three features:

"The first component is a sense of importance, meaning that the particular norm is not to be taken lightly in comparison to other considerations (such as mood, preference gain). Second, oughtness implies that one disapproves of others' transgressing the norm (i.e., the core motivation for informal enforcement derives from the oughtness of the norm itself). Third, oughtness implies that one feels obliged to follow the norm oneself (i.e., the core motivation for informal enforcement is also directed toward oneself). The stronger the sense of oughtness, the more directly relevant a social norm will be for action (Sripada & Stich, 2006). The degree of activation of a norm is tantamount to the strength of the sense of oughtness with regard to this norm at a given moment." (Lindenberg and Steg 2013, p. 38).

A third conclusion would include, in addition to these components, a further determinant: the feeling of oughtness, deriving from the norm *per se*, would arise because a joint commitment between the parties emerges, a readiness to feel part of something that sees in the achievement of such a norm the plural subject's goal. The source of motivation might result from a type of collective intentionality that is not comparable to a bunch of individual intentions, since people alone can always withdraw from their private and personal decision, let us say *more easily*. Things are different when the decision is the result of a deliberation, of a collective consensus, of a shared expression of what is ought to do. For this reason, in the last section, the issue of a joint commitment has been addressed as it is exposed by Margaret Gilbert. She gets the credit of having identified a mechanism common to social phenomena, a 'glue' that makes collective actions perceived as ontologically distinct from the individual ones. Not only that, she also argues and clarifies the difference among obligations arising from personal commitments and from the collective ones – with direct consequences on reason for acting and on the corresponding responsibility to act. Her main interest is to define the subject of collective intentionality, a subject that cannot be reduced to the mere sum of individuals and which, as plural, implies an ontological difference.

In support of this irreducibility, the data showed that only through an agreement behind the veil of ignorance (i.e., result of a collective deliberation), the liberal egalitarian criterion is chosen ex-ante and implemented ex-post. This might underline how a collective choice, in which the subject is constitutive plural, produces a commitment that is a motivational source for action: a normative reason (the liberal egalitarian principle) would become also motivating. Gilbert, however, excludes the relevance of the procedure, namely the method, by which unanimous consent is reached. So, one challenge this research tried to solve is a demonstration of how the procedure is a necessary condition

for joint commitments and obligations of justice. The liberal egalitarian principle is identified as a normative solution to a normative problem of distributive justice because a prior impartial agreement is reached. In matters such as those relating to social justice, in which the ideals of fairness can be several, and where personal interests may be in conflict, a prior agreement on which conception of justice to embrace becomes a *condition sine qua non* for compliance.

From a purely philosophical perspective, a relevant conclusion would be to have highlighted which conditions must be met so that a normative ethics is not only abstract and utopian, but also *real*, by providing real motivations for agents to behave in accordance with its principles.

The term utopian refers directly to Thomas Nagel who had exposed the limits of a political theory precisely in its bad utopian essence. One of the major problems he identifies is the problem inherent in the idea of political justification, because it leaves undefined the relationship between ethical and motivational elements. In terms of motivation theory, this is the problem between normative and motivating reasons: it is known, in fact, that people can know what it would be right to do in a certain situation, but this does not mean that people will behave as they should.

If any interest, inclination, or strong personal reason intervenes in the decision-making process, the normative reason remains in the background: it can constitute a heuristic, a guide for behaviour, but it would not be strong enough to motivate. It could also be said that a moral principle is a sort of ideal, which is recognized by individuals in *foro interno*, but under what conditions does this ideal become binding for action?

Nagel replies with the dual justification's requirement: "justification in political theory must address itself to people twice – first as occupants of the impersonal standpoint and second as occupants of particular roles within an impersonally acceptable system. This is not capitulation to human badness

or weakness, but a necessary acknowledgment of human complexity. To ignore the second task is to risk utopianism in the bad sense" (Nagel 1989, p. 914).

Nagel supports an ethical conception that he calls normative realism (1986). In this perspective, the propositions regarding reasons to act can be true or false regardless of the individual perspective of who formulates them. This is possible because each person is able to transcend the partial, personal point of view, by considering it only one point of view within a broader horizon that includes other points of view. So, people would be motivated by 'neutral' (objective and impartial) reasons and agent-relative reasons. The latter are shaped on real life necessities, desires, self-interested aims. Nagel's attempt would be to focus on the person, on the one who acts and on the human ability to go beyond one's individuality, embracing the relevant others in our society. This is the power of ethics, a critical reflection that can help to consider each one's position together with that of the others, keeping in mind what moral principles are and finding in them a motivation to act, since they have been rationally justified. Justification appears to be a fundamental requirement for the condition of existence of motivation. Elsewhere, Nagel says: "To justify a choice among the alternatives to everyone it is necessary to identify both the claims of impartiality or moral equality and the claims of individual motivation, and find an arrangement which appeals to them in a feasible combination" (Nagel 1989, p. 909).

Regarding to this statement it may be seen how the results of this dissertation would help in clarifying this perspective. The pure procedural justice by John Rawls, in fact, identifies in the rational agreement the method by which moral principles are constructed: they are not imposed from the outside, neither taken for granted, but they are the result of a comparison between alternative conceptions of justice made by free and equal people – Rawls in particular proposes utilitarianism,

intuitionism and perfectionism as the three moral reasonings competing with justice as fairness. The procedure allows the parties to consider the situation from an impartial point of view, to justify the principles of justice as a product of their own reflection and, therefore, to have motivations to act.

The agreement guarantees a rational justification of the moral principles that are constructed and that become the content of the agreement. In the specific case of provided experiments, the chosen distribution principle is such that it might be justified in Nagel's terms: it applies what is right given the initial situation of unjustified inequality and, at the same time, guarantees a profit for the parties, so it holds together objective/impartial reasons and relative-agents motives. Dual justification would seem possible thanks to the achievement of an agreement behind the veil of ignorance and, since the content of the agreement is recognized as justified by the parties, it constitutes a source of motivation to act in compliance. Nagel argues that the normative cannot be replaced by the psychological, however, as another conclusion of this work, it would show how the psychological and cognitive element cannot be excluded. The fact of feeling jointly committed and of having motivations – therefore having an obligation – to act accordingly to the ex-ante decision does not end in the normative, but also spills over the psychological. According to Rawls, the psychological disposition that allows one to comply with what was previously agreed upon is the sense of justice. It would emerge ex-post, on condition that each agent recognizes the basic institutions as a result of an impartial agreement behind the veil of ignorance and with the awareness that there is public knowledge of two elements: the other members do their part, therefore act in conformity with the principles, and each agent has reason to believe that the others will also play their part, in the absence of evidence to the contrary. These conditions would make the sense of justice effective. The moral point of view on the one hand and preservation of personal points of view on the other: "The *raison d'être* of a morality is to produce reasons that overwhelm reasons for self-centred interest" (Baier

2018, p. 186, English translation by the author), and this would be the aim of moral reasons, by providing motivations to share the impartial perspective in order to solve collective action problems. Together with the reflective equilibrium, the sense of justice would give motivational force to behave in accordance with the principles of morality, questioning and evaluating the rightness of the institutions of the society in which we live. The conditions that Rawls exposes for the emergence of a sense of justice recall the conditions for the formation of a joint commitment in Gilbert's terms. What appears to be missing in the definition of Gilbertian joint commitment is the condition of validity for the joint commitment itself.

A forth conclusion of this work, therefore, would be a proposal to integrate the notion of joint commitment with a condition of validity: the deliberation procedure behind the veil of ignorance could provide an impartial and impersonal justification for obligations, stemmed from joint commitments, to act in a liberal egalitarian manner. This would be possible because of the method: the Kantian constructivism would justify that validity condition, causing a convergence of subjects' behaviour towards the liberal egalitarian criterion. As final remark on this aspect, it should be stressed again that, according to Rawls, it is the procedure that determines what is just:

"Pure procedural justice in the original position allows that in their deliberations the parties are not required to apply, nor are they bound by, any antecedently given principles of right and justice. Or, put another way, there exists no standpoint external to the parties' own perspective from which they are constrained by prior and independent principles in questions of justice that arise among them as members of one society" (Rawls 1980, pp. 523-524).

Pure procedural justice[32] is achievable among rational and reasonable agents who consider themselves and others as moral persons: it is on this assumption that, in a context of impartial and impersonal choice such as the one recreated in this setting, the procedure leads towards the liberal egalitarianism.

The last conclusion regards norm-driven behaviour, namely reasons people have for acting in a certain way. Normative reasoning proves to be relevant in the decision-making process, especially when people must decide how to behave in dilemmatic situations: the dilemma arises from the discrepancy between the positive – what one does – and the normative – what one should do. It could be said that, to explain human behaviour in the terms of (personal or collective) intentional action, a normative premise should be experimentally integrated, since it influences, *de facto*, people's behaviour. Experimental studies are, by definition, collections of data, so a positive study into people's behaviour. However, in distributive justice, reciprocal beliefs and judgments on what is meant by justice as fairness are normative arguments that play a decisive role in human conduct. Although further research is needed in explaining how, cognitively, an ex-ante agreement behind the veil of ignorance bind individuals' intentions, this project wanted to show the importance of merging ethics and moral psychology, in including normativity as a motivational factor of human practical reasoning.

[32] "We have a case of perfect PJ [Procedural Justice] if we have some independent specification of what justice requires in a certain situation, and it is possible to devise a procedure that will automatically produce the independently defined just result. We have a case of imperfect PJ if, again, we know independently what results are required by justice, but we cannot devise a procedure that will automatically produce those results […] In a case of pure PJ, however, we have no independent specification of what justice requires, but justice is a matter of importance, and there are possible procedures or rules such that, if they are followed, the outcome will automatically be just" (Nelson 1980, p.503).

References

Alvarez, M. (2016). Reasons for Action: Justification, Motivation, Explanation, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/reasons-just-vs-expl/.

Anderson, E. (1999). What is the Point of Equality?, *Ethics*, 109, pp. 287-337.

Andreoni J., and Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences, *Econometrica*, 70(2), pp. 737-753.

Aquino, K., and Reed A. (2002). The Self-Importance of Moral Identity, *Journal of Personality and Social Psychology*, 83(6), pp. 1423–1440.

Ayal, S., Gino, F., Barkan, R., and Ariely, D. (2015). Three Principles to REVISE People's Unethical Behavior, *Perspectives on Psychological Science*, 10(6), pp. 738–741.

Baier, K. (1958). *Il punto di vista razionale. Una base razionale per l'etica*, (M. Zanichelli, Trans.), Rubettino Editore.

Baldwin, T. (2006). Rawls and Moral Psychology, in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, Volume 3, pp. 248-270.

Barkan, R., Ayal, S., and Ariely, D. (2015). Ethical dissonance, justifications, and moral behaviour, *Current Opinion in Psychology*, 6, pp. 157–161.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact?, *Experimental Economics*,11(2), pp. 122–133.

Barry, B. (1995), John Rawls and the Search for Stability, *Ethics*, 105(4), pp. 874-915.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*, New York: Cambridge University Press.

Bicchieri, C. (2008). The Fragility of Fairness: An Experimental Investigation on the Conditional Status of Pro-Social Norms, *Philosophical Issues*, 18, pp. 229-248.

Bicchieri, C., and Chavez, A. (2010). Behaving as expected: Public information and fairness norms, *Journal of Behvioral Decision Making*, 23(2), pp. 161-178.

Bicchieri, C. (2016). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, Oxford University Press.

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

Bolton, G. E., Katok, E., and Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness, *Int J Game Theory*,27, pp. 269-299.

Bolton, G.E., and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition, The American Economic Review, 90(1), pp. 166-193.

Bolton, G.E., and Ockenfels, A. (2006). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment, *The American Economic Review*, 96(5), pp. 1906-1911.

Brañas-Garza, P., and Morales, A. J. (2005). Moral Framing in Dictator Games by Short Sentences, *ThE Papers 05/06*, Universidad de Granada. Departamento de Teoría e Historia Económica.

Bratman, M. (1999). *Faces of Intention: Selected Essays on Intention and Agency*, Cambridge: Cambridge University Press.

Brown, A. (2005). Luck Egalitarianism and Democratic Equality, *Ethical Perspectives: Journal of the Europena Ethics Network*, 12 (3), pp. 293-339.

Cabrales, A., Nagel R., and Rodríguez Mora, J.V. (2012). It is Hobbes, not Rousseau: an experiment on voting and redistribution, *Exp Econ*, 15, pp. 278–308.

Camerer, C.F. (2003). Behavioural studies of strategic thinking in games, *TRENDS in Cognitive Sciences*, 7(5), pp. 225-231.

Campbell, R. (2007). What is Moral Judgment?, *The Journal of Philosophy*, 104(7), pp. 321-349.

Cappelen, A.W, Hole, A.D., Sørensen, E. Ø., and Tungodden, B. (2007). The Pluralism of Fairness Ideals: An Experimental Approach, *The American Economic Review*, 97(3), pp. 818-827.

Cappelen, A. W., SørensenE. Ø., and TungoddenB. (2010). Responsibility for what? Fairness and individual responsibility, *European Economic Review*, 54(3), pp. 429-441.

Cappelen, A. W., and Tungodden, B. (2011). Distributive interdependencies in liberal egalitarianism, *Social Choice and Welfare*, 36(1), pp. 35-47.

Cappelen, A.W., Nielsen, U.H., Sørensen, E.Ø., Tungodden, B., and Tyran, J.R. (2013). Give and take in dictator games, *Economics Letters*, 118(2), pp. 280-283.

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), pp. 1398–1413.

Charness, G., and Rabin, M. (2002). Understanding Social Preferences with Simple Tests, *The Quarterly Journal of Economics*, 117(3), pp. 817-869.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments, *Journal of Economic Behavior and Organization*, 149, pp.74–87.

Cremaschi, S. (2004). L'Etica Analitica dalla Legge di Hume al Principio di Kant, in A. Campodonico (ed.), *Tra legge e virtù. La filosofia pratica angloamericana contemporanea*, Genova, Italy: Il melangolo, pp. 9-46.

Chung, A., and Rimal, R.N. (2016). Social norms: a review, *Review of Communication Research*, 4, pp. 1-28.

Cudd, A., and Eftekhari, S. (2017). Contractarianism, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/contractualism/.

Da Re, A. (2010). *Le parole dell'etica*, Bruno Mondadori.

Davis, D.D., and Holt, C.A. (1993). *Experimental Economics*, Princeton University Press, chapter I, pp. 3-66.

Davis, D.D., and Holt, C.A. (1993). Experimental Economics: Method, Problems, and Promise, *Estudios Económicos*, 8(2), pp. 179-212.

Dawes, M.R., and Messick, D.M. (2000). Social Dilemmas, *International Journal of Psychology*, 35 (2), pp. 111-116.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., and Smirnov, O. (2007). Egalitarian motives in humans, *Nature*, 446(7137), pp. 794-796.

De Vecchi, F. (2012). Ontologia sociale e intenzionalità: quattro tesi, *Rivista di estetica*, 49, pp. 183-201.

Degli Antoni, G., Faillo, M., Francés-Gómez, P., and Sacconi, L. (2016). Distributive Justice with Production and the Social Contract. An Experimental study, *EconomEtica* ,60, pp. 1-51.

Dewey, J. 1(922a). *Human Nature and Conduct*, New York: Henry Holt and Co.

Dubreuil, B., and Grégoire, J.F. (2013). Are moral norms distinct from social norms? A critical assessment of Jon Elster and Cristina Bicchieri, *Theory Dec.*,75, pp.137–152.

Eckel, C., and Grossman, P. (1996). Altruism in anonymous dictator games. Games and *Economic Behavior*, 16, pp. 181-191.

Elster, J. (1985). Rationality, Morality, and Collective Action, *Ethics*, 96(1), pp. 136-155.

Elster, J. (2006). Fairness and Norms, *Social Research*, 73, pp. 365-376.

Elster, J. (2009). Interpretation and Rational Choice, *Rationality and Society*, 21(1), pp.5-33.

Engel, C. (2011). Dictator games: a meta study, *Exp Econ*, 14, pp. 583–610.

Epstein, B. (2016). A Framework for Social Ontology, Philosophy of the Social Sciences, 46(2), pp. 147-167.

Epstein, B. (2018). Social Ontology, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/social-ontology/.

Erkal, N., Gangadharan, L., and Nikiforakis, N. (2011). Relative Earnings and Giving in a Real-Effort Experiment, *The American Economic Review*, 101(7), pp. 3330-3348.

Falk, W.D. (1947-48). "Ought" and Motivation, *Proceedings of the Aristotelian Society*, 48, pp. 111-138.

Faillo, M., and Sacconi, L.(2007). Norm Compliance: the contribution of behavioral economics models, *Discussion Paper* N. 4, University of Trento, pp. 1-36.

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), pp. 225-246.

Faillo, M., Rizzolli, M., and Tontrup, S. (2019). Thou shalt not steal (from hard-working people): An experiment on respect for property claims, *Journal of Economic Psychology*, 71, pp. 88-101.

Fehr, E., and Schmidt, K.M. (1999). A Theory of Fairness, Competition, and Cooperation, *The Quarterly Journal of Economics*, 114(3), pp. 817-868.

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism, *Nature*, 425, pp. 785-791.

Fehr, E., and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments, *The American Economic Review*, 90(4), pp. 980-994.

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans, *Nature*, 415, pp. 137-140.

Fehr, E., Naef, M., and Schmidt, K.M. (2006). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment, *The American Economic Review*, 96(5), pp. 1912-1917.

Fehr, E., and Schurtenberger, I. (2018). Normative foundations of human cooperation, *Nature Human Behaviour*,2, pp. 458-468.

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

Frankena, W.K. (1958). Obligation and motivation in recent moral philosophy, in A. I. Melden (ed.), *Essays in Moral Philosophy*. University of Washington Press.

Franzen, A., and Pointner, S. (2012). Anonymity in the dictator game revisited, *Journal of Economic Behavior& Organization*, 81, pp. 74– 81.

Freeman, S. (2019). Original Position,*The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL: <https://plato.stanford.edu/entries/original-position/>.

Frohlich, N., Oppenheimer, J.A., and Eavey, C.L. (1987). Laboratory Results on Rawls's Distributive Justice, *British Journal of Political Science*, 17(1), pp. 1-21.

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*. University of California Press.

Gauthier, D. (1986). *Morals By Agreement*, Oxford: Oxford University Press.

Giddens, A. (1971). *Capitalismo e teoria sociale. Marx, Durkheim, Weber*, (C. Cantini and M. Pogatschnig, Trans.), Il Saggiatore, chapter II, pp. 110-180.

Gilbert, M. (1989). *On Social Facts*, New Jersey: Princeton University Press.

Gilbert, M. (1990). Walking Together: A Paradigmatic Social Phenomenon, *Midwest Studies in Philosophy*, XV, pp. 1-14.

Gilbert, M. (2000). *Sociality and Responsibility: New Essays in Plural Subject Theory*, Lanham/Boulder/New York/London: Rowman and Littlefield.

Gilbert, M. (2003). The Structure of the Social Atom: Joint Commitment as the Foundation of Human Social Behavior, in F. F. Schmitt (ed.), *Socializing Metaphysics: The Nature of Social Reality*, Rowman & Littlefield Publishers, Inc.: Oxford, pp. 39-65.

Gilbert, M. (2004). Collective Epistemology, *Episteme*, 1(2), pp. 95-107.

Gilbert, M. (2004). Scanlon on Promissory Obligation: The Problem of Promisees' Rights, *The Journal of Philosophy*, 101(2), pp. 83-109.

Gilbert, M. (2006). *A Theory of Political Obligations: Membership, Commitments and the Bonds of Society*, Oxford: Oxford University Press.

Gilbert, M. (2006). Rationality in Collective Action, *Philosophy of the Social Sciences*, 36(1), pp. 3-17.

Gilbert, M. (2014). *Joint Commitment: How we make the social world*, New York: Oxford University Press.

Gino, F., Ayal, S., and Ariely, D. (2012). Self-Serving Altruism? When Unethical Actions That Benefit Others Do Not Trigger Guilt. Harvard Business School Working Paper, No. 13-028.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE, *Journal of the Economic Science Association*, 1(1), pp. 114–125.

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

Guala, F. (2005). *The Methodology of Experimental Economics*, Cambridge University Press.

Haidt, J. (2007). The New Synthesis in Moral Psychology, *Science*, 316 (5827), pp. 998-1002.

Hands, D.W. (2012). The Positive-Normative Dichotomy and Econmics, in Uskali Mäki (ed.), *Philosophy of Economics*, Elsevier, pp. 219-239.

Hart, H.L.A. (1955). Are There Any Natural Rights?, *Philosophical Review*, 64(2), pp. 175-191.

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

Hobbes, T. (2011). *Leviatano*. (G. Micheli transl.). Milano: BUR Biblioteca Univ. Rizzoli.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games, *Games and Economic Behavior*, 7(3), pp. 346–380.

Hoffman, E., McCabe, K., and Smith, V. (1996). Social Distance and Other-Regarding Behavior in Dictator Games, *The American Economic Review*, 86(3), pp. 653-660.

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

Houser, D. (2008). A note of Norms in Experimental Economics, *Eastern Economic Journal*, 34, pp. 126–128

Huanga, K., Greenea, J.D. and Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good, *Proceeding of the National Academy of Sciences of the United States of America (PNAS)*, 116(48), pp. 23989-23995.

Hume, D. (1978). *A Treatise of Human Nature* (1739). Oxford, Oxford University Press.

Kant, I. (1788), *Critique of Practical Reason*, Translated by Werner S. Pluhar (2002). Hackett Pub Co Inc.

Khadjavi, M. (2015). On the interaction of deterrence and emotions, *Journal of Law Economics and Organization*, 31(2), pp. 287–319.

Kessler, J.B., and Leider, S. (2012). Norms and contracting., *Management Science*, 58(1), pp. 62–77.

Klosko, G. (1994). Rawls's Argument from Political Stability, *Columbia Law Review*,94(6), pp. 1882-1897.

Knobe, J. and Nichols, S. (2007). *An Experimental Philosophy Manifesto* (eds.), Oxford University Press, pp. 3-14.

Kohlberg, L. (1974). The Claim to Moral Adequacy of a Highest Stage of Moral Judgment, *The Journal of Philosophy*, 70(18), pp. 630-646.

Kok-Chor, T. (2008). A Defense of Luck Egalitarianism, *The Journal of Philosophy*, 105 (11), pp. 665-690.

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

Konow, J. (2001). Fair and Square: The Four Sides of Distributive Justice, *Journal of Economic Behavior and Organization*, 46(2), pp. 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*, 41(4), pp. 1188-1239.

Konow, J. (2005). Blind spots: The effects of information and stakes on fairness bias and dispersion, *Social Justice Research*, 18(4), pp. 349–390.

Korenok, O., Millner, E.L., and Razzolini, L. (2014).Taking, giving, and impure altruism in dictator games, *Experimental Economics*, 17(3), pp. 488–500.

Korsgaard, C. M. (1996). *The sources of Normativity*, Cambridge University Press.

Krupka, E.L., and Weber, R.A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?, *Journal of the European Economic Association*, 11(3), pp. 495–524.

Landucci, S. (1993). *La critica della ragion pratica di Kant. Introduzione alla lettura*. Carocci.

Leavitt, K., Zhu, L., and Aquino, K. (2016). Good Without Knowing it: Subtle Contextual Cues can Activate Moral Identity and Reshape Moral Intuition, *Journal of Business Ethics*,137, pp. 785–800.

Levitt, S.D., and List, J.A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?, *Journal of Economic Perspectives*, 21(2), pp. 153–174.

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

List, J.A. (2007). On the interpretation of giving in dictator games, *Journal of Political Economy*,115(3), pp. 482–493.

Lindenberg, S. (1990). Homo Socio-oeconomicus: The Emergence of a General Model of Man in the Social Sciences, *Journal of Institutional and Theoretical Economics*, 146, pp. 727-748.

Lindenberg, S. (2001). Intrinsic Motivation in a New Light, *KYKLOS*, 54(2/3), pp. 317-342.

Lindenberg, S., and Steg, E. (2007). Normative, gain and hedonic goal frames guiding environmental behaviour, *Journal of Social Issues*, 63(1), pp. 117-137.

Lindenberg, S. and Steg, L. (2013). Goal-framing Theory and Norm-Guided Environmental Behavior. In H. van Trijp (ed.), *Encouraging Sustainable Behavior*, New York: Psychology Press, pp. 37-54.

Mazar, N., Amir, O., and Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance, *Journal of Marketing Research*, 45, pp. 633–644.

McClennen, E. (1989). Justice and the Problem of Stability, *Philosophy & Public Affairs*, 18(1), pp. 3-30.

Michelbach, P.A., Scott, J.T., Matland, R.E., & 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, pp. 33–40.

Mordacci, R. (1999). Agire per ragioni morali: razionalità e motivazione nelle analisi della scelta morale, *Rivista di Filosofia Neo-Scolastica*, 91(4), pp. 593-626.

Nagel, T. (1970). *The Possibility of Altruism*, New Jersey: Princeton University Press.

Nagel, T. (1973). Rawls on Justice, *The Philosophical Review*, 82(2), pp. 220-234.

Nagel T. (1986). *The View from the Nowhere*, Oxford University Press, New York; Besussi A. (it. Trad.), *Uno sguardo da nessun luogo*, Il Saggiatore, Milano 1988, chapter VIII.

Nagel, T. (1989). What Makes a Political Theory Utopian?, *Social Research*, 56(4), pp. 903-920.

Nagel, T. (2002). Rawls and Liberalism, in Samuel Freeman (ed.), *The Cambridge Companion to Rawls*, Cambridge University Press, pp. 62-85.

Nelson, W. (1980). The Very Idea of Pure Procedural Justice, *Ethics*, 90(4), pp. 502-511.

Oleson, P.E. (2001). *An Experimental Examination of alternative theories of distributive justice and economic fairness*, UMI Microform 3016508, Bell & Howell Information and Learning Company.

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms, *Journal of Economic Perspectives*, 14(3), pp.137–158.

Plato (2000). *The Republic*, ed. G.R.F. Ferrari; trans. T. Griffin, Oxford.

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

Rawls, J. (1963). The sense of justice, *Philosophical Review*, 72 (3), pp. 281-305.

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

Rawls, J. (1974). The Independence of Moral Theory, *Proceedings and Addresses of the American Philosophical Association*, 48, pp. 5-22.

Rawls, J. (1980). Kantian Constructivism in Moral Theory, *The Journal of Philosophy*, 77(9), pp. 515-572.

Rawls, J. (1999). *A theory of justice. Revised edition*, Cambridge, Massachusetts: the Belknap Press of Harvard University Press.

Rey-Biel, P., Sheremeta, R., and Uler, N. (2018). When Income Depends on Performance and Luck: The Effects of Culture and Information on Giving, *MPRA Paper* No. 83940.

Richardson, H.S. (2018). Moral Reasoning, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/reasoning-moral/.

Rigdon, M., Ishii, K., Watabe, M., and Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology* ,30, pp. 358–367.

Rodriguez-Lara, I., and Moreno-Garrido, L. (2012). Self-interest and fairness: self-serving choices of justice principles, *Exp Econ*, 15, pp. 158–175.

Rodríguez-López, B. (2013). Por qué ser justos: Son las normas de justicia sociales o morales?, *Revista Internacional de Sociología (RIS)*, 72(2), pp. 261-280.

Sacconi and Grimalda (2007). Ideals, conformism and reciprocity: A model of Individual Choice with Conformist Motivations, and an Application to the Not-for-Profit Case. In L. Bruni and P.L. Porta (eds.), *Handbook on the Economics of Happiness*, pp. 532-569.

Sacconi, L., and Faillo, M.(2008). Conformity, Reciprocity and the Sense of Justice How Social Contract-based Preferences and Beliefs Explain Norm Compliance: the Experimental Evidence, Discussion paper N. 14, University of Trento.

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, pp.171–201.

Sacconi, L., Faillo, M., and Ottone, S. (2011). Contractarian Compliance and the `Sense of Justice': A Behavioral Conformity Model and Its Experimental Support. *Analyse & Kritik*, 33(1), pp. 273-310.

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas, *Rationality and Society*, 7, pp.58–92.

Schroeder, T., Roskies, A. L., and Nichols, S. (2010). Moral Motivation, in Doris J. M. (ed.), *The Moral Psychology Handbook*, Oxford University Press, pp. 72-110.

Schurter, K., and Wilson, B. J. (2009). Justice and Fairness in the Dictator Game, *Southern Economic Journal*,76(1), pp. 130-145.

Seabright, P. (2005). The evolution of fairness norms: an essay on Ken Binmore's Natural Justice. politics, *philosophy & economics (ppe)*, 5(1), pp. 33-50.

Servátka, M. (2009). Separating reputation, social influence, and identification effects in a dictator game, *European Economic Review*, 53(2), pp. 197-209.

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

Schram, A., & Charness, G. (2012). Social and Moral Norms in the Laboratory, *eScholarship, UC Santa Barbara, Departmental Working Papers*, pp. 1-33.

Schweikard, D.P., and Schmid, H. B. (2013). Collective Intentionality, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/collective-intentionality/.

Scott, J.T., Matland, R.E., Michelbach, P.A., & Bornstein, B.H. (2001). Just Deserts: An Experimental Approach to Distributive Justice. American Journal of Political Science, 45(3), 749–67.

Shalvi, S., Gino, F., Barkan, R., and Ayal, S. (2015). Self-Serving Justifications: Doing Wrongand Feeling Moral, *Current Directions in Psychological Science*, 24(2), pp.125–130.

Simon, H.A. (1955). A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics*, 69(1), pp. 99-118.

Sinnott-Armstrong, W., Young, L., and Cushman, F. (2010). Moral intuitions. In J. M. Doris (Ed.) and Moral Psychology Research Group, *The moral psychology handbook* (p. 246–272). Oxford University Press.

Smith, M. (1994). *The Moral Problem*, Wiley-Blackwell.

Sober, E., and Wilson, D.S. (1999). *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Harvard University Press.

Stevens, D.E. (2018). *Social Norms and the Theory of the Firm: A Foundational Approach*, [Kindle DX version].

Stevenson, C.L. (1937). The Emotive Meaning of Ethical Terms, *Mind*, 46(181), pp. 14-31.

Tammi, T. (2013). Dictator game giving and norms of redistribution: Does giving in the dictator game parallel with the supporting of income redistribution in the field?, *The Journal of Socio-Economics*, 43, pp. 44-48.

Tossut, S. (2018). Margaret Gilbert, *APhEx* 18, ISSN 2036-9972.

Tuomela, R., and Miller, K. (1988). We-Intentions, *Philosophical Studies*, 53, pp. 367-89.

Tuomela, R. (2007). *The Philosophy of Sociality: The Shared Point of View*, New York: Oxford University Press.

Ullmann-Margalit, E. (1978). *The emergence of norms*. Oxford University Press.

Voigt, S. (2015). Veilonomics: On the Use and Utility of Veils in Constitutional Political Economy, in L.M. Imbeau and S. Jacob (eds.), *Behind a Veil of Ignorance?*, Springer International Publishing Switzerland, Studies in Public Choice 32, pp.9-33.

Wallace, R.J. (2020). Practical Reason, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/practical-reason/

Young, H.P. (2007). Social norms, *Discussion paper series*, 307, Department of Economics, University of Oxford.