SCHMIDT (Christian), « Retour sur le “voile d'ignorance” et ses implications sur les principes de justice sur la base d'une interprétation alternative de la position originelle », *Revue d'histoire de la pensée économique*, n° 5, 2018 – 1, p. 265-289

DOI : 10.15122/isbn.978-2-406-08068-8.p.0265

SCHMIDT (Christian), « Retour sur le “voile d'ignorance” et ses implications sur les principes de justice sur la base d'une interprétation alternative de la position originelle »

RÉSUMÉ – Après l'introduction des relations entre le voile de l'ignorance et l'utopie de la position originelle, la première partie analyse la position originelle comme un “monde possible” de Lewis. La deuxième partie discute la distinction entre information et connaissance au cours des séquences de l'accord des parties sur les principes de justice. La troisième partie explique le rôle du Maximin dans ces choix comme une expérience de l'esprit proche des expériences réelles et des résultats des neurosciences.

MOTS-CLÉS – Principes de justice, Maximin, mondes possibles, Rawls, voile d'ignorance

SCHMIDT (Christian), « Revisiting the “veil of ignorance” and its implication for the justice principles on the ground of an alternative interpretation of the original position »

ABSTRACT – After an introduction on the relation between the veil of ignorance and the utopian original position, the part I analyses the original position as a “possible world” in the Lewis acceptance. The part II discusses the distinction between information and knowledge at the different sequences of the parties' agreements on the principles of justice. The part III explains the role of the Maximin in the parties' choice as a thought experiment in line with real experiments and neurosciences results.

KEYWORDS – Justice Principles, Maximin, possible worlds, Rawls, veil of ignorance

# REVISITING THE "VEIL OF IGNORANCE"

## and its implication for the justice principles on the ground of an alternative interpretation of the original position

Christian Schmidt
Université Paris-Dauphine
PHARE (Paris 1)

### INTRODUCTION

The use of the metaphoric formula "the veil of ignorance" is to be understood by reference to the Rawlsian global program of a theory of justice. Indeed for finding its foundational basis, Rawls comes back to the old reference of the Social Contract, but in opening new questions about its relevance and in developing a different interpretation of its actual meaning. The first question to be solved can be formulated as follows: How humans can agree on definite Principles of justice? Rawls proposed solution is provided by the reference to a utopian "original position", where the assumption of a "veil of ignorance" takes a determinant place. So the first part is devoted to analyze its exact role as an essential component of the original position's construction. In order to deepen the perspective opened by Rawls, we suggest understanding the original position as a kind of thought experiment invented for testing the relevance of this contractual approach of a theory of justice so revised. Its logical framework can be found in the notion of possible

world proposed by Lewis. We show in a second part how this way for thinking the original position helps us to identify what kind of ignorance is required and what kind of knowledge is necessary for the success of the operation from the principles up to their conventional implementation. In the last part, the reference to the maximin criterion as a logical tool for reaching the principles of justice is re-examined in line with this understanding of the original position and the related achievement of the Two Principles. Its relevance is re-evaluate in connection with what Rawls call "the sense of justice" on the ground of the most recent behavioral and neurosciences contributions.

## I. WHAT THE "ORIGINAL POSITION" DOES REALLY MEAN?

### I.1. A WORLD WHICH IS NOT A "LITTLE WORLD" *À LA SAVAGE*, BUT REFERS TO THE LEWIS "POSSIBLE WORLDS"

At first glance, the original position looks like an *ad hoc* device for achieving a fair agreement on principles of justice by means of a bargaining (Rawls, 1972). Rawls explains that this original position is neither an historical state of affairs, nor a primitive natural state, but must be understood as a hypothetical situation. Then he quotes Kant for justifying the reference to the original position because, for him, such a hypothetical situation is the right framework for introducing the Kantian categorical imperative status of the principles of justice (Rawls, 1972, p. 252-253). But, for Rawls the aim of the original position is not only to justify the logical foundation of the principles, but also to show how such principles could be hypothetically achieved. Therefore he links up the original position to the tradition of a contractual basis for a fair political justice rather to follow a strict utilitarian perspective.[1]

1 Rawls explains why the philosophical roots of his theory of "justice as fairness" must be found neither in a pure utilitarianism, nor in a strict contractual tradition, but in an eclectic mixture which shares some features of both. He proposes to link his prospect to intuitionism through the fiction of the original position. Rawls even quotes Poincaré to

Hobbes, Locke and even more Rousseau refer to an initial convention for founding what Rousseau called the "social pact". But, in spite of his recommendation to investigate how such a convention has been emerged (Rousseau, 1762, p. 289), nothing precise can be found on the matter in Rousseau's works. Rawls however derives from Rousseau the idea of a situation where "whatever his position each is forced to choose for everyone" (Rawls, 1972, p. 140). Indeed he developed elsewhere a personal interpretation of the Social Contract where Rousseau's key notion of general will is to be understood as a view point on political justice (Rawls, 2002). So, according to this understanding, Rousseau's general will can be related to the parties' way of reasoning in the original position. Nevertheless the "*raison d'être*" of Rawls original position is to be found elsewhere.

A direction for exploring the question could be to consider the original position from the viewpoint of "possible worlds" in the Lewis' meaning, where possible worlds designate the existence of entities which differ from actual worlds but correspond to "ways things could have been" (Lewis, 1973, p. 84). Possible worlds are worlds, which are quite consistent in the logical acceptance, but according to dimensions which are not the same than those of the actual world where we live. Recent works have proposed to extend the semantic of possible world to the fictions, but sometime through different channels (Doležel, 2010; Tullman & Buckwaller, 2014). In Rawls fiction, the original position can be considered as an imaginary world where the inhabitants (the parties for Rawls), due to the "veil of ignorance", and at variance to the actual world, agree on justice principles without knowing their past, present and future own social and economic positions. Furthermore, it is because they ignore their positions that they rationally agree on principles of justice. Therefore the original position can be understood as a singular "possible world", or more precisely a category of "possible worlds", where at variance to our world the decision-makers ("the parties" in the original position) do not know anything about the consequences of their choice for themselves. Such a knowledge constraint transforms the meaning of their rational choice. In addition thanks to this translation of Rawls original position into the Lewis semantics, the "veil of

support this position in *A Theory of Justice* (Rawls, 1972, p. 22). Nevertheless, he never assumes all the logical implications of this philosophical reference.

ignorance" gives rise to a conditional proposition like "If parties had not bargain under the veil of ignorance, then they would not choose rationally the Principles of justice". On another hand, the inhabitants of such an imaginary world can generate by their agreement on Justice Principles "just modalities" into "possible worlds".

Such an understanding of the original position opens the road to a comparison between the "possible worlds" corresponding to the original position and the Savage's concepts of "worlds" and "states of the world" much more familiar to the economists. As for Savage, the "worlds" associated to the original position is for Rawls a kind of device for framing a specific decision-making problem, the fair choice of justice principles by the parties. What Savage calls "small world" is also a device for framing rational individual choice situations. For Rawls, as well as for Savage, the decision-makers are supposed to be rational. But according to Savage formulation, the individual decision makers are able to choose rationally because in a "small world", thanks to the subjective probabilities, they can derive the consequences of their choice in terms of personal utility from their knowledge about the occurrence of the possible states of the world by the means of their consequences. For that reason, Savage assumes that the decision-makers have a perfect knowledge of the impact of the possible states of the world on their own situations through their outcomes. This assumption, as Savage himself recognized, is an oversimplification of decision-making situations, which he labels "small worlds" by contrast to the "grand world", where the consequences of the identified states of the small worlds are events and not states. For sake of simplicity, Savage just assumes that the small worlds are logical partitions of the grand world (Savage, 1954, p. 82-88).

We support the idea that the "original position" must be understood at variance to what Savage calls a "small world". In the original position due to "the veil of ignorance", no one has any knowledge about the consequences of his choice on his own personal situation. Therefore the relation between the states of the world and their consequences for the decision-makers themselves can no longer support the supposed rationality of the parties. To sum up if the original position refers to possible worlds in the Lewis acceptance, it cannot be modeled as a "small world" in the Savage spirit. Therefore the questions rose to the parties in the possible world generated by the veil of ignorance cannot

be solved by means of the classical economic model of expected utility proposed by Savage and its variants imagined by his successors. Such evidence entails two major consequences. Firstly, the decision-making in the original position cannot be modeled in the Savage framework of the choice under uncertainty. Secondly, the rationality of the parties in the original position cannot be defined as the rationality of the decision-makers in the Savage "small world".

Does the "grand world" sketched out by Savage opens a way to picture the uncertainty in the Rawls possible world of the original position? As in the "grand world", there is no way in the original position to derive the consequences from the states because in the original position the states are hidden to the parties by the veil of ignorance. But the role and then the meaning of uncertainty is not the same in both cases. For Savage, the question is to find a rational treatment of uncertainty thanks to probability in subjective decision-making situations. For Rawls, uncertainty is the condition for a rational choice of accepted principles of Justice. Our prospect allows to clarify an important point and to raise a relevant question. Therefore the uncertainty derived from the veil of ignorance in the original position is different from the uncertainty considered in the expected utility models and its many variants. As for the question, the rationality of the parties under the veil of ignorance needs to be redefined.

### I.2. THE PARTIES IN THE ORIGINAL POSITION BETWEEN COMPLETE IGNORANCE AND PERFECT KNOWLEDGE

This semantic detour through the meanings of "world" in analytical philosophy helps us to see how the construction of the original position is closely dependent upon the hypotheses concerning the knowledge of the parties when they tend to agree on a social pact. The inhabitants of the original position do not know the economic and the political system where they live. We have even said that they ignore their own social position on the past (and the position of their ancestral), as well as on the future (and the position of their descendants), and so far after the conclusion of the expected social pact. They are also obviously quite ignorant of the social positions of the others who are in the same conditions. Nevertheless, in spite of those restrictions, the possible worlds generated by the representatives' choices in the original position can be

understood as "actual worlds", just different from our world according to the realistic interpretation of the possible worlds developed by Lewis (Lewis, 1973; 1986). Therefore thanks to the logical organization of this singular fiction, the relation between Rawls original position and the "possible worlds" at Lewis mode becomes understandable. This property is determinant for finding a relevant solution to the problem of rational choice of justice principles in Rawlsian terms. Indeed a social agreement deductively reached in the original position so defined is supposed to be valid and significant at any time, past, present or future in line with the Kantian prescriptions. In other words, Rawls has lately suggested in "Justice as Fairness" that the original position could be understood as a framework for a kind of thought experiment (Rawls, 1958; 2001, p. 17). In introducing the "veil of ignorance" as the crucial dimension of the possible world untitled "the original position", Rawls really opens the way to modeling political justice as the result of a rational decision-making process.

The difficulty is perhaps no more about what the parties do not know, which is carefully described by Rawls and expressed by the veil of ignorance metaphoric hypothesis, but rather about what they must know for choosing rationally the principles of justice in the possible world of the original position. A distinction must be introduced here between two kinds of knowledge: the direct knowledge of information provided by actual data and the theoretical knowledge for learning and analyzing possible information. In the original position, the parties are assumed to have a perfect knowledge of all which is necessary for the understanding the principles of justice (economics, politics, psychology,…). Rawls even adds at the frontiers between theoretical and factual knowledge of information that "parties know that their society is subject to the circumstances of justice and whatever it implies" and concludes more obviously "whatever general facts affect the choice of principles of justice" (Rawls, 1972, p. 137). In other words, if the parties are supposed to be ignorant of almost all factual data, they hold, on the contrary, a complete and perfect knowledge of all that concerns the general principles of justice. Those assumptions about the knowledge of the parties explain why they are in the same time quite ignorant by reference to its first meaning, and purely rational in their decisions according to its second meaning.

Then the question moves from the knowledge conditions to the status of the decision-makers who are the inhabitants of the possible worlds invented by Rawls. In Rawls construction, the decision-makers are not actual individual persons but parties who are the representative of individual persons. The parties so understood are at the origin of rational decisions for reaching the agreed principles of justice.

The parties must be understood as equal and free abstract supports for agreeing rationally to the options, which lead to a fair justice. So in the original condition, the decision-makers who are "the parties" are supposed to choose rationally for achieving the funding principles of justice thanks to the thick veil of ignorance previously described. This implies that decision-makers in the original position have a perfect knowledge of what the principles of justice really mean. In addition, each party knows that the others are also rational decision-makers and know the same think. To sum up, in the original position the parties know all what is necessary for choosing rationally the principles of justice on which they will agree and this knowledge is even supposed to be common Knowledge between them.

As for their knowledge, we must re-introduce explicitly the classical distinction between knowledge and information implicitly used previously for distinguishing the two kinds of Knowledge. In the original position the parties have a complete knowledge of the justice as well as a perfect common knowledge of rationality in its strong acceptance, but their information is severely bounded. The discrepancy between the knowledge, as a learning and the aptitude to understand on the one hand, and the information, as factual data on the other hand, explains the metaphor of the "veil of ignorance" chosen by Rawls to describe the singularity of the possible worlds corresponding to the original position. Indeed the veil hides to the parties a lot of factual information carefully selected for the purpose.

## II. PROCEDURAL JUSTICE AND DEGREES OF THICKNESS IN THE "VEIL OF IGNORANCE"

### II.1. FROM THE ORIGINAL POSITION TO THE FOUR FOLLOWING STAGE-SEQUENCES

If for Rawls, the agreement on justice principles is the necessary condition to build a Theory of political justice as fairness, it is not a sufficient condition, because those principles must be applied by means of fair procedures. Therefore he introduces a four stage sequences framework for the political and social applications of the principles. All of them refer to the idea of justice considered from different viewpoints: 1) Principles; 2) Constitution (basic convention); 3) Legal system (political, economic and social institutions); 4) Applications of rules to particular cases. Nevertheless Rawls defines a hierarchical order between these four stages which are supposed to succeed one another following a sequential order (Rawls, 1972, p. 195-201).

The adjunction of these four stage sequences raises some problem to the original position which is introduced as a purely hypothetical situation and so out of any temporal consideration. As we have seen, the "veil of ignorance" in its strict meaning is the major feature of the original position so understood. But if the original position is still an a-historical state it generates now a succession of sequences which are not understandable without historical references. For example, for choosing rationally a fair legal system (sequence 3), the parties must know the historical position of the society in terms of natural resources, and economic development. Moreover the reference to an historical process starts just after the first stage, because the parties are supposed to know perfectly the consequences of the chosen Principles of justice for agreeing to a constitution (sequence 2). Finally, for applying rationally the legal principles of justice to particular cases (sequence 4), the parties, who are now individual persons, must know their own position in the system. Therefore Rawls allows a progressive introduction of information from the initial stage of the original position to the last stage of his sequential framework for implementing his theory of justice as a political fairness.

Out of the "original position *stricto sensu*", the metaphoric "veil of ignorance" is now used by Rawls as an efficient device to isolate which information must be known and which information must be unknown by the decision-makers for choosing rationally the fair option at each stage of the justice procedure during the global process for achieving a just society. In the original position, the veil of ignorance hides to the parties the set of all the factual information detailed by Rawls. So it serves to guaranty their lack of factual information, while preserving their pure abstract knowledge. A complete veil of ignorance is the condition for Rawls own Kantian interpretation of the original position, from which his two Principles of justice are derived, and then become categorical imperative:

> To act from the principles of justice is to act from categorical imperative in the sense that they apply whatever in particular our aims are. This simply reflects the fact that no such contingencies appear as premises in their derivation (Rawls, 1972, p. 253).

But factual information becomes necessary for choosing rationally at the three following sequences imagined by Rawls to implement the Principles of justice. Then the thickness of the metaphoric veil allows reducing step by step the initial ignorance of all the factual information in putting at the disposal of the parties the selected data necessary for their rational choice at each stage.

For that purpose three kinds of factual information are tell out by Rawls to be associated to the three levels of the Principles applications identified by Rawls: 1- The factual consequences of the principles of justice chosen in stage 1; 2- the general facts about society where the parties are living; 3- the particular facts concerning the individual positions of the parties (Rawls, 1972, p. 200). We can summarize Rawls scheme as follows:

- In stage 1 – The Principles (the original position). The parties have no factual information, just an abstract knowledge of the fair principles (complete "veil of ignorance").
- In stage 2 – The Constitution. The parties know factual information derived from the chosen Principles of justice.
- In stage 3 – The legal system. The parties know the factual consequences of the chosen Principles + the general facts about the society.

– In stage 4 – The application of the rules (out of the original position). The parties know the factual consequences of the chosen Principles + the general facts about the society + particular facts concerning the individual positions (No "veil of ignorance").

Rawls recognizes that his scheme of the four stages derived from a well-known tradition is no more than a device for extending in a reasonable way the abstract Principles of justice to their procedural implementation. But, moving the "veil of ignorance" from the original position to the four stages changes the relation between information and rational decision-making. Of course the lack of information about parties' own positions remains the key condition for the parties to choose the fair solutions in the four stages. Nevertheless we have seen that in the original position at stage 1, the complete lack of information is a condition for choosing rationally fair Principles of justice. On the contrary, larger information becomes necessary for the rational choice of a fair decision at the stage 4. So the role of this veil is to hide to the decision–makers selected information which risk leading astray their rational choice of justice at the different stages of the procedural levels. The veil, by means of its variations, operates as a tool which allows us to treat each stage of justice as the result of the parties' rational choices.

So understood, the "veil of ignorance" contributes to connect the philosophical roots of this theory of Justice to the rational choice under uncertainty modeled by the economic theory. Nevertheless some specific features of the scheme must be underlined. First of all, due to the lack of all kinds of factual information in "the complete veil of ignorance", the original position corresponding to the stage 1 cannot be assimilated to the three following stages. Secondly, the original position is now the starting point of a dynamic and non reversible process, which engages a sequential order, from stage 2 to stage 3 and from stage 3 to stage 4. Indeed, the *raison d'être* of such a construction is to define the necessary and restricted information, to be provided to the parties for choosing rationally a fair solution in each stage of the procedural justice. But it necessitates to precise now the exact meaning of the ignorance for the parties at those different stages, and their implications for framing the rational decision-making process according to each situation all along this dynamics.

### II.2. "VEIL OF IGNORANCE" OR "VEIL OF UNCERTAINTY"?

Binmore has proposed to model Rawls approach of Justice as fairness as the result of a game of morals which opens the way to the "fair social contract" for framing the game of live (Binmore, 1994; 1998). Incidentally he pointed out an interesting distinction between the "veil of ignorance" and a "veil of uncertainty", previously proposed by Rawls for introducing his theory of justice as fairness (Rawls, 1958). According to Binmore, the "veil of uncertainty" designates a lack of knowledge about the occurrence of future events whilst the "veil of ignorance" also includes a lack of knowledge about events already fixed (Binmore, 1998, p. 214). We will follow such a Binmore distinction with a slightly different interpretation. For us, the "veil of ignorance" hides all the factual information whatever their temporal reference (ignorance = no Knowledge at all), whereas the "veil of uncertainty" only hides some kinds of information or even some features of that information (uncertainty = incomplete Knowledge). Therefore, for us, the "veil of ignorance" and the "veil of uncertainty" refer to different kinds of worlds, the a-temporal world of the original position on one hand and different temporal worlds on the other hand. The "veil of ignorance" only concerns the original position in stage 1 which corresponds to what Rawls also call the "initial position" of the original position, the other stages being related to various positions of a "veil of uncertainty".

This distinction between the "veil of ignorance" and the "veil of uncertainty" must be much more elaborated in taking into account the distinction between the information and the knowledge of the information. Ignorance designates a complete lack of knowledge whatever the correspondent information. Uncertainty, on the contrary, refers to a degree of knowledge to be associated to the information often measured along a scale between perfect certainty, quoted 1, and complete uncertainty quoted 0. To ignore something means to know nothing about it including the thing itself. To be uncertain on something means to know some information but not sufficient to be sure of the thing which gives rise to an uncertain knowledge. If the uncertainty can be estimated, its measurement takes most often the form of probabilities.[2] Furthermore, different levels of uncertainty can be distinguished: A first order, where

2 See the analysis developed by Knight (1957) as a starting point for those distinctions.

the uncertainty is estimated by a probability value (Risk); a second order, corresponding to the uncertainty about the first order estimate, and so on. Therefore the difference between a "veil of uncertainty" and the "veil of ignorance" so understood is not a question of degrees but much more a distinction which refers to logical properties. This provides an additional argument in favor of understanding the original position under the "veil of ignorance" as a "possible world" different from the other worlds only associated to a "veil of uncertainty".

In case of ignorance, there is no room for estimates, whatever the knowledge, objective as well as subjective. In case of uncertainty on the contrary, different levels of knowledge corresponding to the associated uncertainty can be introduced. Consequently, there is no way to transform gradually the "veil of ignorance" of the initial position into a "veil of uncertainty". But as soon as the world of the initial position is closed, the parties know the commitments on the Principles of justice. Then, according to the metaphor, the "veil of ignorance" disappears and the next stages of the procedural justice are elaborated under a "veil of uncertainty".

Such a difference between the "veil of ignorance" and the different thicknesses of a "veil of uncertainty" generates more complexity in the definite meaning of the original position. Indeed due to the difference of information, the agreements achieved by the parties in the original position under the "veil of ignorance" do not have the same significance than those concluded under the "veil of uncertainty". Consequently the two Principles of justice that the parties have supposed to agreeing into the original position in stage1 do not have the same epistemic status than the following other statements. In addition Rawls points out that the two principles must be themselves strictly ordered. The equal basic liberties precede the principle of fair equality of opportunities and equitable distribution.

The distinction between the ignorance and the uncertainty in the original position supports our interpretation in terms of possible world which allows introducing a significant delimitation between the original position under the "veil of ignorance" and the original position under the "veil of uncertainty". Only the first one can be really considered as a "possible world" quite different from our world. The parties in the original position under the "veil of ignorance" are not individual

decision-makers who maximize their preferences (or the utility derived from their preferences), but trustees representative of anyone individual who tend to agree on accepted Principles of justice. Due to the "veil of ignorance", they cannot assess probabilities to the possible states and they do not know the individual distribution of outcomes associated to the states. Their ignorance is not the result of a lack of certainty, but much more an intrinsic condition of their knowledge condition. Therefore the purpose of the fictive world imagined by Rawls is not to mimic a collective utilitarian process for social decision-making, but to support a thought experiment for finding freely political justice foundations.

The original position is the a-historical hypothetical world imagined by Rawls where the "veil of ignorance" is an intrinsic dimension. Just after the parties have agreed on the Principles of justice, factual necessary information is successively introduced in order to implementing the Principles during the three following stages identified by Rawls; those kinds of information can be more or less precise. They are also contingent and historically dependent. As for example, the political and economic facts about the society necessarily refer to the past, and then their knowledge is more or less certain. Furthermore, past information allows future expectations, which are by definition uncertain.

The "veil of ignorance" guarantees that the original position is a non-historical possible world, which means without historical time. But, as soon as the "veil of ignorance" is replaced by a "veil of uncertainty", the historical time becomes a necessary dimension of the world when we leave the original position. Therefore, Rawls imagines a kind of intermediate corridor where the different features of the historical time would be progressively re-introduced in the analytical domain of Justice in order to implement the Principles. But it must not hide the basic starting point where "ignorance" and "uncertainty" are two quite distinct categories without intermediate paths.

## III. THE "MAXIMIN PRINCIPLE"
## A rational reference for the parties to agree on the Principles of justice in line to the mental dimension of our sense of justice

In the Rawlsian interpretation of the Social contract, the parties in the original position will choose rationally the two Principles of justice under the "veil of ignorance". The formulation of these principles requires some preliminary clarifications. The first Principle concerns the basic equal liberties for each person; the second the economic and social fairness through the equality of opportunity and a fair rule of wealth distribution. But they are hierarchical organized. We have still noticed that from *A Theory of Justice* Rawls supports the priority of the Political basic liberties upon the economic and social equalities (Rawls, 1972, p. 61). Later he briefly argues in *Justice as Fairness* that for the same reason the equal opportunities attached to offices and positions are to be prior to the rule of the greatest benefit of the least-advantaged persons in the second Principle (Rawls, 2001, p. 43). According to Rawls view, the social and economic Principle is to be understood in a society where the basic liberties are previously defined and the difference principle must take place in a society where equal opportunities have been introduced. From a logical view point, it means that a society with equal liberties for each person is the restricted domain of interpretation for the equal opportunities so defined. Similarly a society with equal liberties and equal opportunities is the restricted domain of interpretation for the difference principle. If the consistency of the two principles is confirmed, does its necessary imply that the two principles so defined resulted from a rational choice of the parties under the "veil of ignorance" of the original position?

Due to the Knowledge conditions in the original position under the "veil of ignorance" we have seen that the rational choice of the parties cannot follow the expected utility format. Therefore Rawls proposes an alternative approach of rational choice for the parties derived from the Maximin criterion, first introduced in *A Theory of Justice* and later

argued in *Justice as Fairness.*[3] In order to justify the rationality of the Maximin criterion as a rule of choice for the parties in the original position, Rawls starts to identify three conditions to apply to the maximin rule for choosing rationally: 1) the lack of any reference to probabilistic estimates, 2) the "guaranteeable level of incomes", 3) the avoidance of worst outcomes below the "guaranteeable level" (Rawls, 2001, p. 98). Then he stated that those conditions are satisfied by the parties in the original position. Therefore the rational reference to the Maximin criterion is for Rawls the consequence of the specific features attached to the original position previously analyzed. So due to the "possible world" corresponding to the "veil of ignorance" in the stage 1 of the original position, the parties (the inhabitants of this possible world) do not follow the same kind of reasoning for choosing rationally the two Principles of justice than the inhabitant of our actual world.

### III.1. TOWARDS A RATIONAL AGREEMENT OF THE PARTIES IN A GAME THEORETICAL FORMAT

The two Principles are not only the result of individual rational choices. According to Rawls contractual approach of justice, the parties in the original position must agree to the principles of justice. Therefore game theory seems to be the appropriate analytical tool for modeling their rational choice of the principles.[4]

In a seminal paper, Kalai has pointed out a formal similarity between his proposed "proportional solution" to bargaining situations and Rawls Maximin criterion. He suggested that the proportional solutions to bargaining situations lead to apply a welfare function following the Maximin criterion. So Kalai hypotheses that starting from the maximin rule as a rational guideline for the players would also leads to the solution of the welfare game consistent with the Rawls principles of justice (Kalai, 1977). But such a direction has not been more explored.

3 Rawls has firstly tried to link directly the rule of choice associated to the parties to a principle of distributive justice by means of a fallacious similarity between the "difference principle" and the Maximin criterion (Rawls, 1972). But he recognized his error and developed a more elaborate argument in *Justice as Fairness* (Rawls, 2001).

4 In Lewis fiction of the original position the parties are supposed to choose rationally the two principles against alternative proposals. We assume, as Binmore but in a different way that the process of this founding choice by the parties is to be modeled in a specific game format.

Another approach has been developed by Howe and Roemer which takes the form of a cooperative game, where the core which is the solution of the game corresponds to an income distribution satisfying the maximin principle (Howe & Roemer, 1980). An interesting feature of Howe & Roemer treatment of Rawls justice principles in a cooperative game is to put in light the real bases of the so call "risk adverse bias". Indeed, no player in the Rawls game so modeled has an incentive to re-negotiate the Social contract agreed in stage 1. The explanation is provided by the singularity of the original position under the "veil of ignorance". In such a possible world, the guarantee level of income is a logical dimension of the principles of justice.

We have seen that Binmore from his own, has developed what he calls the "game of morals", where the bargaining takes the form of a two-person non cooperative game named Adam and Eve (Binmore, 1994; 1998). But at variance to Kalai results, the maximin does not coincide to the solution of the game so constructed, at least in its one shot bargaining version. The result can also be explained by Binmore interpretation of the "veil of ignorance". We pointed out that Binmore rejects the assumption of the thick veil of ignorance in the original position. What he calls a "thin veil of ignorance" for the Adam and Eve agreement on the Justice Principles is no more than a thick "veil of uncertainty" previously discussed (Binmore, 1994). We have shown that such a "veil of uncertainty" becomes only relevant after the agreement to the two principles in the 3 following stages of the Rawls Scheme.

More recently, Binmore has extended a different interpretation of his bargaining game of moral previously developed in his earlier version of the game of moral. The maximin strategies which secure the players in the "veil of ignorance" could lead them to reach an empathic equilibrium defined along the dynamic of an evolutionary approach, where the players acquire what Binmore call "empathic preferences" for solving a succession of coordination games (Binmore, 2014). But moving from a-temporal situation (the original position) to an evolutionary process transforms the foundations of the justice principles. As suggested by Binmore himself, the origin of those empathic preferences is to be found in human biology.[5] One must noticed that in a short paper two biolo-

5 By "empathic preferences" Binmore requires to take into account the empathic dimension into the definition of the players' preference in the initial position. The existence of

gists have also shown that the "veil of ignorance" can favor biological cooperation when the selection acts out of a self-matching mechanism (Queller & Strassmann, 2017). More generally a comparison between the role of the "veil of ignorance" in social justice and in a kind of "genomic justice" has been previously introduced by Okasha and discussed from a philosophical perspective (Okasha, 2012).

In spite of their similarity the two kinds of ignorance do not have the same meaning in both cases. Its signification in the possible world of the Rawls original position cannot be simply transposed in the genetic world. In the original position the Principles of justice are the result of a rational agreement between the parties whereas the genetic behavior is not the result of whatever rational choice. Furthermore the players of the game of justice under the "veil of ignorance" are not individuals but parties, who are representative of any kind of men or women. This implies a role of universal trustee toward all the individuals. As they are supposed to be quite rational, one can infer that they must agree on an implicit mutual assurance pact. Therefore the game plays by the parties in the original position looks like a kind of insurance game which can explain the key role of the maximin criterion. But future researches on the properties of this peculiar game would be necessary for a complete understanding of its logical properties.

### III.2. FROM THOUGHT EXPERIMENT REASONING TO ACTUAL EXPERIMENTATIONS

Let us back to our interpretation of the logical fiction call by Rawls "The original position". In terms of a possible world Rawls has found logical foundations for a contractual agreement on his two general Principles of Justice. The rational choice of the parties defined by Rawls in the original position does not coincide to its classical definition, at least in economics. We have seen that the parties for reaching this agreement do not maximize a utility function (whatever its interpretation), but follow an alternative rule, the maximin, which remains questionable from a pure rationalistic view point. Does it mean some inconsistency of the parties' strategies in the game of the original position, a criticism argued by Harsanyi from a utilitarian viewpoint on the ground of several

parties' empathic preferences is thus a necessary condition for reaching the "empathic equilibrium" in their bargaining game.

chosen counter-examples (Harsanyi, 1975). Rawls answer was to closely related the maximin principle to the very specific circumstances of the original position, where the parties must find an agreement satisfying for everyone on the two basic Principles of justice (Rawls, 2001, p. 98-104).

Rawls argument in favor of the maximin rule which explicitly relates what he often call "the device of the original position" to the "device of maximin principle" can be developed in line to our analysis of the original position as a Lewis "possible worlds". Rawls often opposes examples of decision in the everyday life to the choices of the parties in the original position. But those references also suggest the comparability of the two kinds of situations. One of the Lewis purpose with the introduction of the "possible worlds" is precisely to show that "worlds" other than ours can be understandable from our world viewpoint, thanks to his proposed modal semantic. He sometimes speaks of "ersatz worlds" to underline their accessibility from our world (the actual world) (Lewis, 1986). In order to attending this purpose, Lewis utilizes a mathematical device shared to Quine which allows for extending the dimensions of our world but he himself recognizes:

> What is interesting is not the reduction of worlds to mathematical entities, but rather the claim that the possible worlds stand in a certain one-to-one correspondence with certain mathematical entities. Call these "ersatz possible worlds". Any credible correspondence claim would give us excellent grip on the real possible worlds by their ersatz handles (Lewis, 1973, p. 90).

We suggest here a different way to utilize such "ersatz worlds". Rawls original position can be considered as a kind of "ersatz world" where the inhabitants ("the representatives") differ from the inhabitants of the actual world by the drastic reduction of their available information (complete ignorance) with its direct implication for their rational choices. One may object that all the examples of "ersatz worlds" provided by Lewis in his writings belong to physics and natural fields, whilst Rawls original position is a mental fiction. Nevertheless the idea of possible worlds can be extended to fictions and mental universes where the "things" look like different from the things look like in our actual mental world for cause of information (or rather here for lack of information) and knowledge. Lewis himself has suggested that fiction can serve for discovery modal truth, (Lewis, 1978).

Thanks to this treatment of the original position semantically linked to our actual world, we can deduce from its informational structure the logical conditions for the parties to agree on the Rawlsian two principles of justice. So, one can demonstrate that the Maximin is the rational rule for the parties to reach an agreement on the two Principles of Justice formulated by Rawls in the original position in protecting a "garanteeable level" to the worst outcomes in a kind of security game (Rawls, 2001, p. 98-99).

But the reference of the "ersatz worlds" also opens the way to other interpretations. The first one obviously consists deriving the parties agreement from a thought experiment as previously suggested. Let us recall Irvine classical definition of a thought experiment:

> A thought experiment is an instance of reasoning which attempts to draw a conclusion about how the world either is or could be by positing hypothetical and perhaps even counterfactual state of affairs (Irvine, 1991).

So, in the possible "ersatz world" characterized by the "veil of ignorance" previously defined, one can infer from a thought experiment that the parties would agree to the two Principles on the ground of the maximin choice criterion.

Following another and more ambitious interpretation the "veil of ignorance" could be understood as a framework for experimental protocols. By means of the possible correspondence between the world of the original position and our world opened by the "possible ersatz worlds" some behavioral information about the choice of Justice principles could then be directly induced from such experiments. As, for example, it could be informative, to validate (or invalidate) from a positive perspective the supposed reference of the parties to the maximin criterion. In case of positive results the maximin could be understood, not only as a normative criterion, but also as a behavioral reference for choosing the principles.

Several attempts have been implemented in that direction. Unfortunately all of them met a preliminary and decisive obstacle: How to translate accurately the strict constraint of complete ignorance in the original position into experimental protocols. Indeed, if each participant of the experiment does not know what will be his (her) own situation after their chosen wealth distribution, all of them necessarily knows a

lot of factual information about income distribution and welfare derived from our world. In addition those protocols do not design an agreement between parties as in the original position, but most often the results of individual choices between different sets of alternatives which have been previously selected by the scientists on the ground of different presumed assumptions. Those necessary discrepancies between the original position and the so call correspondent experimental protocols can explain the contrasted obtained results. As for example, whilst a first study does not find a real experimental support for the maximin criterion (Frolich & Oppenheimer, 1992), another one, on the contrary, seems to support the Rawls maximin reference (Mitchell & *al*., 1993). More recent studies tend to show that the choice of Justice principles by individuals, at least between alternative wealth distributions, is a complex operation which takes into account different rules, including of course the maximin criterion, but at variance to the parties' agreement in the original position (Scott & *al*., 2001; Michelbach & *al*., 2003). Other experimental tentative have been elaborated on the basis of a 2 players non-cooperative game, but here again the oversimplified experimental protocol does not provide a sufficient correspondence to the Rawls original position (Sarkar & Chakraborty, 2016).

### III.3. TOWARD A NEURAL APPROACH OF "THE SENSE OF JUSTICE"

Rawls has introduced another ingredient in his contractarian treatment of the Justice principles, which is formulated as a strict compliance of the parties. This compliance necessarily implies for Rawls what he calls "A sense of justice":

> There is one further assumption to guaranty strict compliance. The parties are presumed to be capable of a sense of justice and this fact is common knowledge among them (Rawls, 1972, p. 145).

The precise meanings of this compliance and of the "sense of justice" must be detailed and their relation has to be elaborated. This work has been recently beginning to be realized by a group of researchers who have previously proposed a clear distinction between the logical properties of the "*ex ante*" and the "*ex post*" perspectives during the parties' agreement in the original position (Faillo, Ottone & Sacconi, 2008). They demonstrated that under the conditions of the "veil of ignorance",

a common compliance of the parties is a necessary *ex post* constraint of their agreement. Therefore the common compliance is to be understood as an additional condition for a rational agreement on justice principles to support the maximin criterion (Sacconi & *al*., 2011).

But this compliance cannot operate without a psychological support. This psychological support is provided by what Rawls calls the sense of justice, which requires a mental capacity shared by the parties for insuring that the chosen principles will be respected. Then the sense of justice is not only a shared mental capacity, but also a capacity that the parties know that they share, and even in the form of a common knowledge between them, as supposed by Rawls himself. Furthermore, such a mental capacity does not only concern the parties when they reach their agreement in the world of the original position, but more generally all the citizens of our world for implementing the justice principles selected by the parties. Therefore Sacconi, Faillo and Ottone, who first modeled the mental requirement of compliance in a game theoretical format, were in the position to test successively its positive relevance by means of two experimental protocols (Sacconi & *al*., 2011).

The different properties of this mental capacity begin to be studied thanks to recent neurosciences studies. We have shown how the notion of empathic preferences has been introduced by Binmore in his game of moral for cause of belief coordination (Binmore, 1994, p. 336). But Binmore just contented with the interactive evolutionary mechanism repetition. The idea of an empathic foundation of the sense of justice has been deepened thanks to behavioral and neurosciences recent contributions. Mixing information from various experiments and the results of neuroimaging studies, Decety and Yoder have distinguished two kinds of empathy: the emotional empathy and the cognitive empathy. They concluded from their investigation that the sense of justice mobilizes the brain activity corresponding to the cognitive empathy, but not to the emotional empathy (Decety & Yoder, 2016).

Another study reveals a quite different facet of the sense of justice in the Rawls acceptance. When the sense of justice is applied to determinate a judgment on wealth distribution (for others), a same system of neural substrates is activated than for choosing rationally (for oneself) under uncertainty. More precisely the study reveals that a specific brain area is only activated for the experimental participants who follow the

Maximin criterion in the first case and adopt the risk adverse solution in the second case (Kameda & *al*., 2016). This result could seem, at first glance, to comfort the Harsanyi well-known criticism of the Rawls use of Maximin principle for cause of risk adverse bias. The opened question sets by those findings is much larger. As suggested by the authors, its reveal the possibility of a common mental anchor in the cognitive empathy which leads our sense of justice and in the cognitive treatment of our risk perception. Future researches are still necessary to confirm those results, but they already open the way to interesting connections between the abstract identification of justice principles and the capture of a mental sense of justice.

## REFERENCES


Ben Porath, Echanan, Gilboa, Itzak & Schmeidler, David [1997], "On the measurement of the inequality under uncertainty", *Journal of Economic Theory*, Vol. 75, N° 1, p. 194–204.

Binmore, Kenneth [1994], *Game theory and social contract*, Vol. I, *Playing fair*, Cambridge (Mass.), MIT Press.

Binmore, Kenneth [2014], "Bargaining and fairness", *Proceeding of the National Academy of the United States of America*, Vol. 111, Suppl. 3, p. 10785-10788.

Decety, Jean & Yoder, Keith [2016], "Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict, sensitivity to injustice for others", *Social Neurosciences*, Vol. 11, N° 1, p. 1–14.

Doležel, Lubomir [2010], *Possible words of fiction and history*, Baltimore, Johns Hopkins University Press.

Duhamel, David [2012], "Le programme Rawlsien apocryphe", *Oeconomia*, Vol. 2, N° 2, p. 151–177.

Faillo, Marco, Ottone, Stefania & Sacconi, Lorenzo [2008], "Compliance by believing: An Experimental exploration on social norms and impartial agreements", Working Paper N° 10, *Dipartemento di economia del 'Universita di Trento*.

Frolich, Norman & Oppenheimer, Joe A. [1992], *Choosing justice: An experimental approach to ethical theory*, Berkeley, Los Angeles, London, University of California Press.

Harsanyi, John C. [1975], "Can the maximin principle as a basis for morality: A critique of John Rawls theory", *The American Political Science Review*, Vol. 69, N° 2, p. 594–606.

Howe, Roger & Roemer, John [1981], "Rawlsian justice as the core of a Game", *The American Economic Review*, Vol. 71, N° 5, p. 880–895.

Irvine, Andrew [1991], "Thought experiments in scientific reasoning" *in* Horowitz, Tamara & Massey, Gerald J. (éd.), *Thought experiments in science and philosophy*, Lanham (MD), Rowman & Littlefied, p. 149–165.

Kalai, Ehud [1977], "Proportional solutions to Bargaining Situations: Interpersonal Utility Comparisons", *Econometrica*, Vol. 45, N° 7, p. 1623–2630.

Kameda, Tatsua, Inukai, Keigo, Higuchi, Satomi, Ogawa, Akitoshi, Kim, Hackjin, Matsuda, Tetsuya & Sakagami, Masamichi [2016], "Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decision", *Proceeding of the National Academy of the United States of America*, Vol. 113, N° 42, p. 11817–11822.

KNIGHT, Frank H. [1921], *Risk, uncertainty and profit*, Reprint, New York, Augustus M. Kelley, 1964.

LEWIS, David [1973], *Counterfactuals*, Cambridge (Mass.), Harvard University Press.

LEWIS, David [1986], *Philosophical Papers*, vol. II, New York, Oxford University Press.

LUCE, R. Duncan & RAIFFA, Howard [1957], *Games and decisions. Introduction and critical survey*, New York, John Wiley.

MICHELBACH, Philip A., SCOTT, John T., MATLAND, Richard E. & BORNSTEIN, Brian H. [2003], "Doing Rawls justice: An experimental study of income distribution norms", *American Journal of Social Sciences*, Vol. 47, N° 3, p. 523–539.

MITCHELL, Gregory, TETLOCK, Philip E., MELLERS, Barbara A., & ORDONEZ, Lisa D. [1993], "Judgements of social justice: Compromise between equality and efficiency", *Journal of Personality and Social Psychology*, Vol. 65, N° 4, p. 429-639.

OKASHA, Samir [2012], "Social justice, genomic justice and the veil of ignorance: Harsanyi's meets Mendel", *Economic and Philosophy*, Vol. 28, N° 1, p. 43–71.

PLAFF, Donald W., KAVALIERS, Martin & CHOLERIS, Elena [2008], "Mechanisms Underlying an Ability to Behave Ethically", *The American journal of bioethics*, Vol. 8, N° 5, p. 10–19.

QUELLER, David C. & STRASSMANN, Joan E. [2013], "The veil of ignorance can favour biological cooperation", *Biology letters*, Vol. 9, N° 6.

RAWLS, John [1958], "Justice as Fairness", *Philosophical Review*, N° 67, p. 164–194.

RAWLS, John [1972], *A Theory of Justice. Revised version*, Cambridge (Mass.), Harvard University Press.

RAWLS, John [2001), *Justice as Fairness. A restatement*, édité par Erin Kelly, Cambridge (Mass.), Harvard University Press.

RAWLS, John [2009], *Lectures on the History of Political Philosophy*, Cambridge (Mass.), Harvard University Press.

ROUSSEAU, Jean-Jacques [1762], « Du contrat social », in *Œuvres complètes de* …, Vol. 3 *Contrat social – Écrits politiques*, Paris, La Pléiade, Gallimard, 1964.

SACCONI, Lorenzo, FAILLO, Marco & OTTONE, Stefania [2011], "Contractarian compliance and the 'sense of justice': A behavioral conformity model and its experimental support", *Analyse & Kritik*, Vol. 33, N° 1, p. 273–310.

SARKAR, Sumit & CHAKRABORTY, Soumyakarti [2016], "Does Rawls original position induce fairness? Experimental findings on selection criteria in a discrete Nash demand game played from behind the 'Veil of ignorance'", 9-10 juin, *International Meeting de l'AFSE* (The French Association of Experimental Economics), Cergy Pontoise.

SAVAGE, Leonard. J [1972], *The Foundation of statistics*, 2e éd. (1re éd., 1954), New York, Dover publications.

SCOTT, John T., MATLAND, Richard E., MICHELBACH, Philip A. & BORNSTEIN, Brian [2001], "Just deserts: An experimental study to distributive justice norms", *American Journal of Political Sciences*, Vol. 45, No 3, p. 749–767.

TALLMAN, Katherine & BUCKWALTER, Wesley [2014], "Does the Paradox of Fiction Exist?" *Erkennthis*, No 79, No 4, p. 779–796.