BETA
Bureau
d'économie
théorique
et appliquée

Documents de travail

« Between economics and philosophy: a reappraisal of the Rawls–Harsanyi debate »

Auteurs

Juan CARVAJALINO, Herrade IGERSHEIM

Document de Travail n° 2026 – 07

Mars 2026

Bureau d'Économie
Théorique et Appliquée
BETA

https://www.beta-economics.fr/

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Université de Strasbourg · cnrs · Université de Lorraine · INRAE · AgroParisTech · Université Haute-Alsace

Between economics and philosophy: a reappraisal of the Rawls–Harsanyi debate

Juan CARVAJALINO, University of Paris 8
Herrade IGERSHEIM, CNRS and University of Strasbourg[1]

Abstract. This article revisits the debate between John Rawls and John Harsanyi by drawing on newly explored archival materials. Traditionally viewed as a short-lived, technical disagreement of the 1970s over the rational criterion for choice under uncertainty—the maximin versus average utility rules—their exchange in fact spanned nearly four decades, from their first encounter in 1964 to the late 1990s. The paper reconstructs this dialogue to reveal its ethical and philosophical depth, showing that what began as a technical dispute gradually evolved into a confrontation over the moral foundations of justice. The paper traces four stages of this evolving relationship, emphasizing Harsanyi's later overlooked "philosophical turn" and his continuing attempts to defend utilitarianism against Rawls's egalitarianism. By revealing all the facets of their exchange, the study enriches our understanding of the modern dialogue between economics and philosophy and of the enduring opposition between utilitarian and egalitarian conceptions of social justice.

Keywords. John Rawls, John Harsanyi, Maximin, Utilitarianism, Social justice.

JEL Codes. B21, B31, D60.

[1] Corresponding author: BETA, 61 avenue de la Forêt Noire, F-67085 Strasbourg Cedex. igersheim@unistra.fr
We are particularly grateful to John Weymark for helpful discussions, as well as to Yuji Ando, Kotaro Yonemura, and seminar participants at the 20th conference Charles Gide in Bordeaux, the 89th Japanese Society for the History of Economic Thought in Hirosaki, the 17th INEM Conference in Bayreuth, and Rikkyo University, for valuable comments. The archivists of the University of Harvard and Berkeley are also warmly thanked.

‘I might just say that I don’t think we are
as far apart as it may seem at first sight’
(JHP, Rawls to Harsanyi, July 18, 1973).

1. Introduction

In the opening to the collective volume *Justice, Political Liberalism and Utilitarianism: Themes from Harsanyi and Rawls*, the three co-editors—Marc Fleurbaey, Maurice Salles, and John Weymark—emphasise that ‘Harsanyi and Rawls deserve our most respectful tribute for their fundamental contributions to utilitarianism and liberal egalitarianism, respectively, and, more generally, for helping to bring questions of social justice to the fore after many years of relative neglect’ (Fleurbaey *et al.* 2008, 2). On the one hand, John Harsanyi (1920–2000), both economist and philosopher, was born in Budapest in 1920, obtained a PhD in philosophy in 1947 and moved to Australia in 1950. In 1956, Harsanyi switched to economics and obtained a PhD at Stanford under the supervision of Kenneth Arrow, continuing his academic career at Berkeley from 1964. In 1994, he was awarded the Nobel Prize in economics, together with Reinhard Selten and John Nash, for their pioneering work on equilibrium in non-cooperative game theory, while a large part of his work is also devoted to welfare economics, more precisely to a rehabilitation of utilitarianism based on modern decision theory.

On the other hand, John Rawls (1921–2002) was born in Baltimore in 1921 and studied philosophy at Princeton, where he obtained his PhD in 1950. Rawls went on to teach at prestigious institutions such as Oxford, Cornell and MIT, before moving to Harvard in 1962. In 1971, John Rawls published his major work *A Theory of Justice*, which marked a turning point both for political and moral philosophy and normative economics by making a real attempt to reconcile economic and philosophical concepts in order to define social justice. Rawls’ contribution legitimised the rejection of utilitarian objectives in favour of equality as the primary aim of modern economic theories of justice.

Although Harsanyi’s and Rawls’s theories share similar aims—namely, to ground a theory of justice on the basis of rational choice theory—the debate of the mid-1970s that saw them opposed over the appropriate decision criterion under uncertainty was nonetheless highly contentious. This debate is well known and has been much discussed

in the literature (see, notably, on this very issue: Binmore 1989; Audard 2002; Weymark 2005; Fleurbaey *et al.* 2008; Duhamel 2012; Kandil 2014; Moehler 2018; Binmore 2021). For most scholars, the dispute between the two thinkers begins with the publication of *A Theory of Justice* and ends with Harsanyi's critiques of Rawls in 1975. Yet this framing reduces the exchange in two ways: first, the debate is typically viewed as a brief exchange of papers over a short period; second, it is usually interpreted as a purely technical disagreement about the rational criterion of choice under conditions of uncertainty, between Harsanyi's average utility principle and Rawls's maximin rule as embodied in his principles of justice.[2]

This article reexamines the relationship between Rawls and Harsanyi, seeking to go beyond the technicalities by drawing on their archives. Based on these new sources, we are able to bring to light a far more nuanced and much longer story between our two protagonists. First, the duration of their exchanges extends well beyond the 1970s, and in fact spans nearly forty years—from their first meeting in 1964 to the late 1990s. Second, beyond the familiar technical question of the appropriate decision criterion under conditions of uncertainty, the nature of their dialogue evolved considerably over time, ultimately centring on the ethical and normative foundations of their respective theories and revealing the profound divergence of their moral ideals. This story unfolds in four main stages. We show that their disagreement actually began several years before the publication of *A Theory of Justice*, during their first encounter in October 1964 (1). This initial exchange was followed by the well-known technical debate in the mid-1970s (2). Then, as that confrontation began to be brought to a close, Harsanyi's work took an increasingly philosophical turn (3), eventually leading him to formulate a broader critique of Rawls's theory (4).

Several recent works have addressed the relationship between Rawls and the economists in a more contextually oriented way, notably making use of Rawls's archival material (Amadae 2003; Peart and Levy 2008; Hawi 2016; Galisanka 2017; Coker 2021; Jackson and Stemplowska 2021; Igersheim 2022; Guizzo and Paré-Ogg 2023; Igersheim 2023). Yet, to the best of our knowledge, no previous contribution has attempted to reconstruct the whole dispute between Rawls and Harsanyi from a historical perspective. Doing so provides a deeper understanding of their respective positions and, above all, brings to light the disagreement between their respective conceptions of ethics, something which underlies the technicalities of the mid-1970s debate.[3] Moreover, in the light of our

[2] Note that the term 'technical' was used by Harsanyi in 1977 (635).

[3] Apart from a few exceptions, like Moelher (2018, 95), the ethical dimensions of their confrontation

analyses some elements of their works can be better interpreted. In addition to Rawls's evolution in the early 1980s, which can partly be explained by his difficult interactions with economists such as Arrow, Musgrave, and Samuelson, culminating with Sen's attack in 1980, we argue that Harsanyi's shift toward philosophy can also be understood as a response to his confrontation with Rawls's theory and as part of his effort to defend utilitarianism not only on the formal grounds of decision theory and welfare economics, but also on a philosophical level.

As such, this article may be relevant to scholars interested in the crossdisciplinarity interrelations between philosophy and history of economics. By revealing the duration and the conceptual depth of the Harsanyi–Rawls dispute, we bring to light some of the lesser-known facets of the modern debate between utilitarianism and liberal egalitarianism.

2. A first programmatic confrontation under the auspices of Buchanan and Tullock

At the beginning of the 60s, both Rawls's and Harsanyi's works were still rather nascent, and their similarities and differences had not yet been brought out. Yet, from their early writings, one can already draw out their respective concerns for utilitarianism strongly attached to cardinal utility on the one hand, and for equality and fairness on the other.

Harsanyi, for his part, returned from Australia to the United States at the beginning of the 60s, securing a position as Professor of Economics at Wayne State University in Detroit, before moving to Berkeley in 1964 (Weymark 2008). At that time, Harsanyi had already gained a certain degree of recognition in welfare economics thanks to two influential papers published in the *Journal of Political Economy* in 1953 and 1955. In the 1953 paper, he famously argues that the concept of cardinal utility, reintroduced by von Neumann and Morgenstern (1944), is also relevant for welfare economics and can be used to compute social welfare. According to Harsanyi, each individual possesses, on the one hand, personal preferences represented by a utility function, and, on the other hand, moral preferences represented by a social welfare function. The latter corresponds to impartial judgments made by an external observer who considers himself as having an equal probability $1/n$ of being any one of the n members of society. In this way, Harsanyi supports a rational-choice–theoretic defence of utilitarianism that incorporates cardinal utility and interpersonal comparisons of utility. This result is referred to by Weymark

have rarely been examined.

(1991) as Harsanyi's impartial observer theorem. Continuing along this line of research, Harsanyi demonstrated in his 1955 paper that if both personal and moral preferences satisfy the von Neumann–Morgenstern axioms for choices under uncertainty, then the cardinal social welfare function can be expressed as a weighted sum of the cardinal individual utility functions. This second result once again supports a utilitarian framework and is known as Harsanyi's social aggregation theorem (Weymark 1991).

On the other hand, at Harvard from 1962 onward, Rawls devoted most of his time to completing *A Theory of Justice* (Pogge 2007). After having attended a seminar by William Baumol in the fall of 1950 (Rawls 1991), his determination to combine economics and philosophy in order to define social justice grew stronger, eventually leading him to the conviction that economics 'alone' could help clarify certain aspects of justice (JRP, Rawls to Mueller, October 16, 1972).[4] However, in the early 1960s, Rawls seems to have been only indirectly acquainted with John Harsanyi's contributions. In 1963, he noted in a footnote—citing Rothenberg (1961)—that 'an earlier proposal of Harsanyi's, similar to that of the natural objection, is equivalent to the principle of utility', before referring readers to Harsanyi's seminal papers from 1953 and 1955. In this context, Rawls advocated what he termed an 'open society', namely, one in which rational agents would affirm the two famous principles of justice. He described the 'natural objection' as the concern that rational individuals might, under certain conditions, prefer to enter a caste system 'if they each found their expectations of well-being in the caste system sufficiently attractive', knowing that a person's expectation is 'given by the expression $\sum p_i w_i$ where p_i is the percentage of persons in the i^{th} position and w_i is an index of the average level of well-being of those in the i^{th} position' (1963, 81 and 85).[5] Therefore, a caste system is based on the presupposition that there exists a prior agreement on a real and equal sharing of risks in the case of the basic structure of the society; but, according to Rawls, this presupposition does not hold.[6] In other words, Rawls claims that the principle of utility 'is not the correct principle of justice to apply to the constitution of the social system itself' (*ibid.*, 86). A second objection to the credibility of an individual choice in favor of a caste system is that persons have obligations especially regarding their children. In such a case, the persons do not want to take chances and would prefer the two principles of justice. The sharp ethical contrast between the Rawlsian open society—which upholds equal liberty through the ideal of equal citizenship—and the

[4] JHP refers to John Harsanyi's papers kept at Berkeley, while JRP refers to John Rawls's papers kept at Harvard.

[5] Before 1971, he will again briefly refer to Harsanyi's 1953 and 1955 papers in 1968.

[6] On the weakness of this argument, see Binmore (1989).

utilitarian caste system is highly representative of the moral opposition that would later emerge between Rawls and Harsanyi concerning the very foundations of social organisation.

Given that both Rawls and Harsanyi worked at the intersection of economics, philosophy, and mathematics, and pursued broadly similar aims, their eventual meeting on American soil in the course of the 60s was largely a matter of time and circumstance. The occasion arose under the auspices of James Buchanan and Gordon Tullock who, having just published *The Calculus of Consent* (1962), launched a series of interdisciplinary meetings bringing together scholars interested in the processes of political decision-making.

The inaugural meeting of the Committee for Non-Market Decision Making took place in October 1963 at the University of Virginia in Charlottesville. Rawls attended the second meeting in October 1964, with Anthony Downs, John Harsanyi, Gordon Tullock and William Riker as chairmen. In the paper he presented, Rawls argued that the principles of justice are 'those principles which rational persons would agree upon or consent to unanimously from an original position of equality' and that 'the difference principle goes beyond the notion of (Pareto) efficiency to a principle of justice' (Amadae 2003, 150). While some minor disagreements began to surface between Rawls and Buchanan and Tullock during this period (see, for instance, Rawls 1963; Jackson and Stemplowska 2021), a far more pointed and enduring dispute emerged between Rawls and Harsanyi, as reflected in their correspondence.

In a letter sent by Rawls to Harsanyi shortly after their encounter, he wrote: 'Whatever the ethical merits of my principle [of difference] I feel that it must be rational (consistent) in that it may be understood as ignoring all individuals in society except one (the worst off) and maximising with respect to him. If his preferences are consistent, so should the social welfare function be consistent. … He is, if you like, Arrow's dictator. In terms of your 1955 JPE article where you show the function is V = $\sum_i a_i U_i$, all the $a_i = 0$ except a_n where *n* is the worst off man. All this may be odd, but it seems consistent with basic rationality assumptions' (JRP, Rawls to Harsanyi, October 1964). Harsanyi's reply came swiftly, marking the beginning of a lifelong dispute between the two thinkers. In a highly didactic and tightly argued eight-page letter, Harsanyi sought to demonstrate that Rawls's social welfare function—namely, $W(X) = \min[U_1(X), U_2(X)]$ with X being a risky prospect—'very definitely does violate' Harsanyi's 1955 *Journal of Political Economy* theorem. According to Harsanyi, his theorem requires the coefficients a_i to be constants;[7]

[7] Harsanyi used an implicit assumption in his 1955 proof which enables one to guarantee the uniqueness of the individual weights (see Fleurbaey *et al.* 2008, 16).

therefore, it cannot accommodate the maximin principle, which would allow the coefficients to shift from zero to non-zero (or vice versa) depending on whether an individual *i* is the poorest member of society.

Although the tone of the exchange remained courteous, it is clear that Harsanyi regarded Rawls's borrowings from theoretical economics as inappropriate, given what he saw as Rawls's insufficient mastery of the underlying mathematics. Notably, from this very first exchange, Harsanyi restricted his critique to the mathematical and economic aspects of Rawls's ideas, refraining from engaging their philosophical dimensions, even though, holding a PhD in philosophy, he could have done so with natural authority.

This initial correspondence already contained the roots of the much-debated dispute that would later intensify following the publication of *A Theory of Justice*. Even more interesting is how this first exchange invites a fresh reading of relevant passages of *A Theory of Justice* that can be understood as direct and pointed responses to Harsanyi's criticisms.

3. *A Theory of Justice* and the 'technical' debate between Rawls and Harsanyi

In 1971, Rawls published his major work *A Theory of Justice*. In light of his earlier private confrontation with Harsanyi, it is hardly surprising that Rawls devoted extensive passages to addressing Harsanyi's attacks. In line with his criticism of the caste system, he also took aim at the very idea of employing rough concepts of risk and probability to resolve central ethical questions—such as determining the basic institutions that will structure society over the long term and ensuring social justice within them. Rawls refuted both Harsanyi's utilitarianism and his use of expected utility theory, arguing that it would lead the impartial observer to gamble on the structural principles of society's most basic institutions.

Rawls devoted two sections of his magnum opus to these subtle issues. In §27, 'The Reasoning Leading to the Principle of Average Utility', he begins by explaining under what circumstances the principle of average utility might be valid. When the veil of ignorance is complete—that is, when the parties know nothing about their own or others' particular preferences, abilities, or the societies they represent—one can imagine a 'hypothetical newcomer' assuming 'that there is an equal likelihood of his turning out to be anyone'. In this case, 'his prospect is highest for that society with the greatest average utility'. Rawls concludes: 'If we waive the problem of interpersonal comparisons of utility, and if the parties are viewed as rational individuals who have no aversion to risk

and who follow the principle of insufficient reason in computing likelihoods (the principle that underlies the preceding probabilistic calculations), then the idea of the initial situation leads naturally to the average principle' (Rawls 1971, 143). Rawls further notes that the same reasoning can be applied not only with the 'traditional' sense of utility—tied to the satisfaction of desire and cardinal interpersonal comparisons—but also with the modern sense of utility adopted by contemporary economic theory: a measure of cardinal utility representing the choices of economic agents, derived 'from the Neumann–Morgenstern construction, which is based on choices between prospects involving risks' (*ibid.*, 143); and adding in a footnote regarding the latter that 'how this might be done was shown by J. C. Harsanyi' (*ibid.*, 144 fn. 25).

In §28, 'Difficulties with the Principle of Average Utility', Rawls explains why his own theory ultimately departs from such reasoning and from the principle of average utility itself—and so, implicitly, from Harsanyi's approach. The reason relies on the specificity of the decision the parties have to make in the original position, i.e., to agree on the principles of justice that will shape their societies for decades, and on the specific features circumscribing the original position via the veil of ignorance: no knowledge of their desires and ends; no knowledge of their social circumstances in their respective society, nor of its array of techniques; finally, even if they had clues about the last two points, they have no grounds to rely on one probability distribution over them rather than another. In this very precise sense, the Rawlsian veil of ignorance is complete and 'leads directly to the problem of choice under complete uncertainty' (*ibid*., 149). Then, if the parties' decision in the original position is to be genuinely rational, it cannot be based on probability judgments that lack an objective factual basis. The principle of insufficient reason, which implies that when there is no objective basis for probabilities, individuals assume subjective probabilities, is ruled out—given the fundamental importance of the decision for the parties and for their descendants. Further, Rawls remarks: 'It may be surprising that the meaning of probability should arise as a problem in moral philosophy, especially in the theory of justice. It is, however, the inevitable consequence of the contract doctrine which conceives of moral philosophy as part of the theory of rational choice' (*ibid.*, 149). To address this difficulty, Rawls adds that the principles of justice should not depend on specific attitudes towards risk. Thanks to the veil of ignorance, the parties thus ignore if they have an unusual aversion to (or predilection for) taking chances.[8]

[8] On this point, Rawls's reasoning appears inconsistent: according to his own account, the parties may have von Neumann–Morgenstern utility functions, which incorporate risk aversion. In that case, under the

Thus, considering the detailed discussion in these two sections, it is perhaps unsurprising that most economists came to view *A Theory of Justice* as 'a classic decision problem, that of the rational choice of an isolated individual in a situation of probabilisable (or non-probabilisable) uncertainty' (Dupuy 2002, 120, our translation).

The gap between Rawls's conception of justice and the interpretation of his work by economists has progressively come to be widely recognised, notably since Sen's *Resources, Values, and Development* (1984), where he coined the term 'Rawlsian economics' to describe the tendency in welfare economics to judge social states by the utility level of the worst-off individual. The technical debate that set Rawls and Harsanyi against each other in two famous articles (Rawls 1974b; Harsanyi 1975) stands out as one of the clearest illustrations of the emergence of the so-called 'Rawlsian economics. Harsanyi's review of *A Theory of Justice*, entitled 'Can the Maximin Principle Serve as a Basis for Morality?' and that was in circulation as a working paper by May 1973 (JHP, Ctn. 2), was intended for publication in the *American Political Science Review* in 1974 as part of a symposium on Rawls. Around the same time, Rawls himself was preparing an article for the *American Economic Review*, in particular to respond to Arrow's 1973 book review. Since Harsanyi had sent his manuscript to Rawls as early as July 1973 (JRP, Harsanyi to Rawls, July 7, 1973), the quirks of publication schedules gave Rawls the chance to respond before Harsanyi's article appeared in print, in his 1974 *AER* piece—prompting Harsanyi to append a postscript to his own paper.

In his article, Harsanyi set out to argue that 'Rawls's attempt to suggest a viable alternative to utilitarianism does not succeed' (1975, 594). He framed the debate in terms of two competing approaches to decision-making under uncertainty: the maximin principle, which, he noted, had been shown since the mid-1950s to produce serious paradoxes, and the prevailing Bayesian approach, which endorsed expected-utility maximisation. Harsanyi sought to demonstrate that the maximin rule produced 'highly irrational conclusions' and carried 'unacceptable moral implications', except in those cases where 'the maximin principle is essentially equivalent to the expected-utility maximization principle' (*ibid.*, 595). Although Harsanyi was fully familiar with Rawls's broader theory, the well-known examples he used to illustrate his attacks are equally revealing of the misunderstandings into which many economists fell when attempting to

modern conception of utility, the principle of average utility is not neutral with respect to risk. This problem was noted by Arrow in his 1973 review of *A Theory of Justice* (Arrow 1973, 256). Rawls was aware of the confusion and made several attempts to defuse the criticisms on this issue (JRP, Rawls to Arrow, June 26, 1972, Rawls to Alexander, April 12, 1973).

engage with it: the issue of selecting principles of justice is reframed as a problem of choosing among lotteries in everyday decision-making.

Answering Rawls's implicit criticism in *A Theory of Justice*,[9] Harsanyi sought also to establish the ethical legitimacy of both probability and von Neumann–Morgenstern utility functions in his own approach. He first noted that the equiprobability assumption may be viewed not only as an instance of the principle of insufficient reason but also as a moral postulate requiring that 'we must give the same *a priori* weight to the interests of all members of the society', while the maximin principle amounts to giving 'unity or near-unity probability to the possibility of the worst-off individual in society', which can be seen as very unfair by the descendants of the parties (*ibid.*, 599). According to Harsanyi, impartiality towards members of the society is thus a moral requirement, which means that everyone should receive an equal consideration. Moreover, he argued that 'vNM utility functions have a completely legitimate place in ethics because they express the subjective importance people attach to their various needs and interests' (*ibid.*, 600).

In his brief *American Economic Review* article (1974b), in a simple footnote after having emphasised that the maximin is 'a macro not a micro principle', Rawls added that this qualification 'affects the force of Harsanyi's counterexamples' (1974b, 142 and fn. 4), without even attempting to engage with other aspects of Harsanyi's critique. Harsanyi underscored this point in his sharp postscript, writing: 'I am really astonished that a distinguished philosopher like Rawls should have overlooked the simple fact that the counterexamples I have adduced have nothing whatever to do with scale at all … . I cannot see how anybody can propose the strange doctrine that scale is a fundamental variable in moral philosophy, without giving credible answers to these questions at the same time' (Harsanyi 1975, 605). Drawing also on Arrow's criticisms of Rawls as compared with utilitarianism, Harsanyi concluded that both their papers undermined 'Rawls's theory as a serious competitor to utilitarian theory … . For this reason, I find it rather unfortunate that Rawls's paper does not even try to answer this criticism at all' (*ibid.*, 606).

According to most scholars, the Rawls–Harsanyi debate does not extend much longer than these two papers (Rawls 1974b, Harsanyi 1975). Rawls did, however, return to the controversy in 1977, later incorporated as Lecture VII of *Political Liberalism*. After acknowledging that 'the first principles of justice as fairness are plainly not suitable for a

[9] While §27 and 28 of *A Theory* do not explicitly address Harsanyi's views, an unpublished 1985 manuscript entitled 'Reply to Harsanyi on Maximin' (JRP) makes this clear. Harsanyi himself was not mistaken on this point, as he wrote: 'Rawls discusses my model primarily in Chapters 27 and 28 of his book' (598).

general theory', he added: 'I cannot reply adequately to Harsanyi's forceful objections here, but I note the following: the maximin was never proposed as a basis for morality; in the form of the difference principle it is one principle constrained by others that applies to the basic structure'.

Beyond the published record that has long framed the debate, archival sources shed new light on its actual unfolding. In September 1985, Rawls drafted a four-page unpublished manuscript entitled 'Reply to Harsanyi on Maximin' (JRP). Rawls began his reply by identifying two areas of agreement with Harsanyi. First, he concurred that the maximin rule is not satisfactory 'as a rule for guiding choices in ordinary life': the maximin 'is not only foolish, but simply crazy' (JRP, 'Reply to Harsanyi on Maximin', unpublished, September 1985). Second, he agreed with Harsanyi that 'the most appropriate general principle of rational choice for individuals is that of maximizing expected utility' (*ibid.*). Yet Rawls insisted that the 'highly unusual circumstances' of the original position required the parties 'to make basic departures from how we reason in ordinary life' (*ibid.*). Here lies the core disagreement between the two scholars. Rawls emphasised once again that, in the original position, 'the idea of probability does not apply', and that the parties must therefore adopt a different form of rational deliberation. By contrast, Harsanyi, according to Rawls, persists in the mistaken belief that 'the idea of probability always applies', even if only in the form of 'as-if probabilities'. In the original position, however, such a trick is unavailable: only probabilities grounded in good evidence would be admissible. Lacking such evidence, and faced with radical uncertainty about the future, the parties' only way to secure the protection of basic liberties is to reject the principle of average utility and instead reason according to the maximin rule.

From this unpublished reply (and thus almost certainly unknown to Harsanyi), the reasons for the divide between the two thinkers become even clearer. Beyond their technical disagreement over the appropriate criterion of choice under uncertainty, what is at stake is the very nature of Rawlsian contractualism: while Harsanyi challenges its foundations, Rawls questions Harsanyi's ability to construct a genuine theory of justice grounded in philosophy rather than relying solely on the technical tools of economic theory. In a letter to Sidney Alexander, Rawls remarked: 'Harsanyi doesn't discuss in enough detail what the effects of risk aversion from the moral standpoint are likely to be' (JRP, Rawls to Alexander, April 12, 1973). Quite interestingly, it is precisely from this moment of direct confrontation with Rawls that Harsanyi's intellectual trajectory began to shift: toward a deeper exploration of his own theory of morality as well as the connections between

morality and rational choice theory. In the process, his critique of Rawls's work would become increasingly explicitly philosophical in character.

4. Harsanyi's 'philosophical turn'

It is obviously no accident that Harsanyi's more pronounced 'philosophical turn' takes place as early as the mid-1970s, in the immediate aftermath of his confrontation with Rawls. In 1976, Harsanyi published a book, with a preface by Arrow, entitled *Essays on Ethics, Social Behavior, and Scientific Explanation*, which brings together thirteen papers written between 1953 and 1975. The first part of the book, which contains five articles—including the two from 1953 and 1955 as well as the critique of Rawls from 1975—is particularly interesting for our purposes, insofar as it seeks to provide 'a modern decision-theoretical foundation of a utilitarian ethical theory' (Harsanyi 1976, ix). Yet it is in an article published the following year, 'Morality and the Theory of Rational Behavior', that the economist continues his effort to present his ethical theory in a synthetic way, and show the manner in which it draws on the modern Bayesian theory of rational behaviour under risk and uncertainty. In doing so, Harsanyi responds more systematically both to Rawlsian critiques concerning the use of probabilities and the use of von Neumann–Morgenstern utility functions—both of which are deeply embedded in Harsanyi's ethics—while also reusing certain passages and examples from his 1975 article. After this paper, Harsanyi's criticism of Rawls can no longer be understood as a mere technical objection, or as one that misses its target in light of the breadth of Rawls's philosophical construction. Harsanyi's opposition to Rawls takes the form of an ethical theory that closely combines philosophical analysis and mathematical reasoning, much like Rawls's own, with nothing less as its objective than to offer 'a unique rational answer to the philosophical question, "What is morality?"' (Harsanyi 1977, 654), moral behaviour itself being a special form of rational behaviour.

According to Harsanyi, the entire moral content of his ethical theory derives from three intellectual traditions: Adam Smith, Kant, and the utilitarian school, which in his view is the most important. He further adds that the combination of these elements was made possible only thanks to the Bayesian concept of rationality, 'a very crucial ingredient of [his] theory' (*ibid.*, 627). His profound disagreement with Rawls's theory is made clear from the introduction: 'all nonutilitarian theories of morality, including John Rawls's very influential theory and several others, at one point or another involve some highly irrational moral choices, representing major departures from a rational pursuit of common

human and social interests which, in my view, is the very essence of morality' (*ibid.*, 626). Harsanyi's thesis is especially strong in that he defends the idea that 'the emergence of modern decision theory has made ethics into an organic part of the general theory of rational behavior' (*ibid.*, 627).

Building on his earlier 1953 work, Harsanyi then develops the equiprobability model of moral value judgments, understood as a particular kind of preference judgment. For a moral value judgment to be authentic, and not merely a matter of personal preference, it must stem from an 'equiprobability postulate', that is, 'the fictitious assumption of having the same probability of occupying any possible social position' (1977, 632). This postulate leads directly to the conclusion that a rational individual, whether or not they belong to the society in question, will always choose the social system that maximises the average of individual utilities. If the individual belongs to that society, their choice can be interpreted as guided by moral preferences—which, according to a similarity postulate, tend to be the same for every member of society. Conversely, if the individual does not belong to the society, their choice can be interpreted as that of an impartially sympathetic observer, impartiality corresponding to the equiprobability postulate, and sympathy referring to the possibility of making interpersonal utility comparisons grounded in empathy.[10] Alternatively, following his 1955 article, Harsanyi also indicates that it is possible to obtain an axiomatic justification of the utilitarian theory of rationality based solely on the Bayesian postulates of rationality and the Pareto principle. This further reinforces his thesis, since the ethical assumptions regarding the form of personal and moral preferences—both conceived as von Neumann–Morgenstern utility functions—are weaker than those required in the equiprobability model of moral value judgments.[11]

Revisiting Rawls's theory of justice, Harsanyi eventually observes that he developed his equiprobability model at the beginning of the 50s, while Rawls later independently proposed a 'very similar model'. He adds: 'the difference does not lie in the nature of the two models, which is based on almost identical qualitative assumptions. Rather, the difference lies in the decision rule, namely, the maximin principle, which was fairly different n that Rawls avoids any use of numerical probabilities' (Harsanyi 1977, 634–635).[12]

---

[10] More precisely, 'all interpersonal utility comparisons are transformed into intrapersonal utility comparisons for the observer' (Fleurbaey *et al*., 2008, 15).

[11] Sen (1976) has opposed a well-known 'standard objection' to Harsanyi's use of von Neumann-Morgenstern expected utility theory, claiming that the latter is an ordinal one and thus cannot justify Harsanyi's utilitarian conclusions (for a recent survey, see Weymark 2005).

[12] That Harsanyi was actually a precursor to Rawls (after Vickrey) in developing the idea of the original position is also already highlighted by Arrow (1973, 250; foreword in Harsanyi 1976) and by Harsanyi

Although Rawls was aware of this 1977 article (later republished in *Utilitarianism and Beyond* in 1982, an edited volume by Amartya Sen and Bernard Williams, to which Rawls also contributed)—as evidenced by his reference to it in *Justice as Fairness: A Restatement* (2001, 100 fn. 22)—he does not discuss it in his unpublished 1985 manuscript, which is somewhat regrettable. In any case, one can stress that this period undeniably marks a genuine 'philosophical turn' in Harsanyi's work, as he fully embraces and makes explicit his ambition to bring philosophy and mathematics into dialogue through Bayesian decision theory. This turn is also clearly reflected in his scholarly activities and professional correspondence. Beginning in the late 1970s, Harsanyi engaged in increasingly intensive exchanges with a growing number of philosophers, among them John Watkins, John Smart, John Broome, Richard Hare, Robert Nozick, David Gauthier, and Richard Brandt (JHP, Ctn. 2 and 3). The topics discussed in these exchanges largely revolved around the central pillars of his moral theory: rule versus act utilitarianism and their respective relation to the concept of justice, his disagreement with Rawls, the types of preferences permitted within utilitarianism, the contractarian tradition, the distinction between risk and uncertainty, and related issues.

His involvement with philosophy and philosophers extended further, as he also participated in conferences devoted to fostering dialogue between philosophy and economics. One such event was the May 1983 conference 'Philosophy, Justice and Economics', jointly organised by the Departments of Philosophy and Economics at the University of Waterloo (Canada), where philosophers such as David Gauthier and Allan Gibbard were present (JHP, Ctn. 3). But his strong reluctance toward Rawls's work does not weaken over time. In a 1987 interview with Lorenzo Sacconi of the University of Milan, Harsanyi declared: 'Among past thinkers, I have great respect for Adam Smith (both as an economist and as moral philosopher), Alexis de Tocqueville, Max Weber, and Joseph Schumpeter. Among our contemporaries, I have great respect for Quine, Richard Brandt, and Richard Hare' (JHP, Harsanyi to Sacconi, March 27, 1987). Notably, John Rawls does not appear on Harsanyi's shortlist of contemporary philosophers. This tension also emerges in his correspondence with Norman Jacobson, a political science professor at Berkeley, concerning the possible arrival of Jürgen Habermas at Berkeley. Jacobson wrote: 'I agree with you … that the field of Social Choice Theory is not exactly a sparkling one just now, and that John Rawls is probably the leader in the field, not an especially encouraging sign' (JHP, Jacobson to Harsanyi, June 22, 1980).

himself in his 1975 paper against Rawls: the notion of the original position 'played an essential role in [his] own analysis of moral value judgements, prior to its first use by Rawls in 1957' (1975, 595).

5. The 'second' Harsanyi against Rawls

Rawls and Harsanyi articulate contrasting views of the link between ethics and rationality. Rawls grounds ethics in contractualism: The principles of fairness chosen in the original position are justified independently of outcomes or behaviour, with justice lying in strict adherence to the letter of the rules. By contrast, Harsanyi embeds ethics in utilitarian decision theory, treating moral principles as axiomatised constraints on rational choice whose authority derives from consequences and behavioural guidance, privileging the rule's spirit. As Harsanyi's work became more philosophical over time, his critique of Rawls deepened, adding to his rejection of the maximin rule the emphasis of the tensions between rules and behaviour or letter and spirit.

First, Harsanyi developed an increasingly marked critique of Rawlsian contractualism, already present in 1975 in embryonic form (598). In an exchange with philosopher of science Roger Rosenkrantz regarding his recent book *Inference, Method and Decision: Towards a Bayesian Philosophy of Science* (Rosenkrantz 1977), Harsanyi rejected Rosenkrantz's use of the term 'contractarian' to describe a utilitarian approach to ethics, adding that 'Rawls has revived the term "contractarian" specifically to contrast his own view with utilitarian ideas' (JHP, Harsanyi to Rosenkrantz, May 7, 1980). According to Harsanyi, any contractarian theory of morality ends up in a 'vicious circle', since 'all contracts derive their moral binding force from the fact that most of us subscribe to a moral code that makes contracts morally binding. Consequently, our moral code is logically prior to any binding contract we may make, so that our moral code itself cannot be the result of a social contract' (JHP, Harsanyi to Binmore, May 15, 1988). This opposition to grounding a theory of morality in a social contract led him not only to reject Rawls's framework, but also to reject David Gauthier's conception as articulated in *Morals by Agreement* (Gauthier 1986; see Harsanyi 1987 and JHP, Harsanyi to Sacconi, May 27, 1987).

Second, Harsanyi also challenged the lexical ordering of the principles of justice advocated by Rawls. He did so at the conference 'Justice, Political Liberalism, and Utilitarianism in Honour of John Harsanyi and John Rawls' held at the University of Caen in June 1996, in a talk entitled 'John Rawls's Theory of Justice: Some Critical Comments'.[13] The list of participants was impressive: beyond Maurice Salles, John

[13] Due to the condition of his health, Rawls was not able to attend it.

Weymark and Marc Fleurbaey (who co-edited the collective volume resulting from the conference, only published in 2008, long after the deaths of Rawls and Harsanyi in 2001 and 2000, respectively), attendees included philosophers such as Richard Arneson, John Broome, Robert Sugden, and Philip Pettit, as well as social choice scholars and game theorists including John Roemer, Claude d'Aspremont, Philippe Mongin, François Maniquet, and Kenneth Binmore. In his talk, Harsanyi criticised the 'rather complicated hierarchy' of principles proposed by Rawls and his 'rather vague and unconvincing' supporting arguments (Harsanyi 2008, 73). In particular, he argued that the absolute priority of the principle of basic liberties was mistaken, as it implied that liberty is 'infinitely more important' than other social values, such as economic efficiency or reductions in economic inequalities. Harsanyi emphasised that, in real societies, people are often willing to sacrifice some individual freedom to enhance other social values, since trade-offs between values are inevitable. In his view, the only absolute priority in moral reasoning is that of 'our moral duties over personal interests and over all other nonmoral considerations' (*ibid.*, 74).

Third, Harsanyi is also critical of Rawls's rejection of meritocracy, accusing him of 'denying moral credit' (Harsanyi 2008, 74). This critique already appears in a 1991 letter Harsanyi sent to Rawls commenting on a manuscript of *Justice as Fairness: A Restatement* (JHP, Harsanyi to Rawls, October 28, 1991). In the letter, Harsanyi's critique focuses entirely on this point, which had never been mentioned before. He specifically challenges the idea that individuals should receive no moral credit for their efforts in cultivating good character and morally praiseworthy behaviour, regardless of their circumstances. He compares Rawls's position to 'hard determinism', which, in his view, denies that people have free will and should bear no moral responsibility for their behaviour, character, and choices. In his 1996 lecture at Caen, Harsanyi further distinguishes between philosophers who are incompatibilists, those who 'take the view that determinism and free will are incompatible' (Harsanyi 2008, 77), and himself, whom he describes as a compatibilist. On the one hand, he argues that talented individuals are deserving if they effectively develop their abilities for the benefit of society; they merit some form of reward, which he considers 'a matter of justice'. On the other hand, he raises the complex issue of the 'working poor', who should be entitled to receive more, even if this conflicts with the difference principle (unlike a group of 'undeserving poor'), and such redistribution would not constitute an injustice. Finally, Harsanyi maintains that great scientists and creators (Einstein, Gödel, etc.) deserve social and even material recognition, even if their work does not serve the most disadvantaged in society: 'society

needs its most educated and most able members. To keep them happy, it should support science, philosophy, and other cultural activities' (*ibid.*).

6. Concluding remarks

Revisiting the Rawls–Harsanyi debate in the light of archival material allows us to better grasp the nuances and philosophical implications of each thinker's position. Beyond the literature's focus on the mid-70s technical dispute, an examination of their archives reveals two different aspects: first, that Harsanyi's criticism of Rawls spanned almost forty years; second, that it touched on several fundamental aspects of their work, thus revealing all facets of the dispute between our two authors. While the central point of the technical debate between Rawls and Harsanyi lay in the opposition between the maximin principle and the average utility principle, a broader philosophical debate emerges from Harsanyi's writings, raising profound ethical and moral questions as his theory of morality progressively takes shape. From the 1980s onward, Harsanyi's philosophical critique developed on three main levels. First, his challenge to the social contract questioned the very foundations of the social structure proposed by Rawls. Second, Harsanyi criticised Rawls's priority of the first principle of justice, arguing that societies do require trade-offs between freedoms and other social values. Third, he strongly opposed Rawls's non-compatibilist stance, which, in his view, disregards the effort individuals can make to cultivate their talents for the benefit of society. Note that these latter elements may raise questions, given their relatively underdeveloped nature and their potential incompatibility with Harsanyi's moral theory.[14] Nevertheless, they demonstrate the extent to which Harsanyi is in disagreement with Rawls, far beyond a mere technical dispute.

This led Harsanyi to state the following in a 1996 interview conducted by d'Aspremont and Hammond during the 1996 Caen conference:

> *Cl. d'Aspremont*: If Rawls were to abandon [his difference principle], and propound a theory based instead on expected utility, would most of your disagreements be resolved?

[14] On this point, elements of an answer might be found in the final book that Harsanyi intended to publish toward the end of his life. Entitled *Morality, Equality and Individual Excellence: A Somewhat Unorthodox Utilitarian Theory*, it is listed as a 'Book Ms in Progress' in a bibliography of Harsanyi dating from December 1997 and was published in a special issue of *Games and Economic Behavior* following his death. This unfinished manuscript does not appear in his archives.

> *J. Harsanyi*: No, I don't think so… And I much more, much more object to his basic moral views than to decision-theoretical views. (2001, 395)

Without the reconstruction of the entire debate between the Harvard philosopher and the Berkeley economist that we have undertaken in this article, this statement would remain difficult to fully appreciate.

References

Amadae, S.M. (2003). *Rationalizing Capitalist Democracy. The Cold War Origins of Rational Choice Liberalism*. The University of Chicago Press.

Arrow, K.J. (1973). Some ordinalist-utilitarian notes on Rawls' *Theory of Justice*. *The Journal of Philosophy*, 70, 245–263.

Audard, C., 2002, Utilitarisme et éthique publique : le débat avec Rawls, *Cités*, 10, 49–62.

Binmore, K., 1989, Social Contract I: Harsanyi and Rawls, *The Economic Journal*, 99(395), 84–102.

Binmore, K., 2021, John Rawls Versus John Harsanyi, in *Imaginary Philosophical Dialogues*, Springer, Cham.

Buchanan, J.M., 1976, A Hobbesian interpretation of the Rawlsian difference principle, *Kyklos*, 29, 5–25.

Buchanan, J.M. & Tullock, G., 1962, *The Calculus of Consent*, Ann Arbor, MI: University of Michigan Press.

Coker, D.C., 2021, Rawls and Knight: Connections and Influence in A Theory of Justice, *Research in the History of Economic Thought and Methodology*, 39C, 77–98.

D'Aspremont, C. & Hammond, P.J., 2001, An Interview with John C. Harsanyi, *Social Choice and Welfare*, 18, 389–401.

Duhamel, D., 2012, Le programme rawlsien apocryphe, *Oeconomia*, 2, 151–177.

Dupuy, J.-P., 2002, *Avions-nous oublié le mal?* Bayard.

Fleurbaey, M., Salles, M. & Weymark, J.A. (eds), 2008, *Justice, Political Liberalism and Utilitarianism, Themes from Harsanyi and Rawls*, Cambridge University Press.

Forrester, K., 2019, *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy*, Princeton and Oxford: Princeton University Press.

Galisanka, A., 2017, Just Society as a Fair Game, *Journal of the History of Ideas*, 18(2), 299–308.

Galisanka, A., 2019, *John Rawls. The Path to a Theory of Justice*, Cambridge (Mass.): Harvard University Press.

Gauthier, D., 1986, *Morals by Agreement*, Oxford: Clarendon Press.

Guizzo, D. & Paré-Ogg, C., 2023, Economics with(out) ethics? An interdisciplinary encounter between public economists and John Rawls in the 1970s, *The European Journal of the History of Economic Thought*, 30(5), 906–933.

Harsanyi, J.C., 1953, Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking, *Journal of Political Economy*, 61, 434–435.

Harsanyi, J.C., 1955, Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility, *Journal of Political Economy*, 63, 309–321.

Harsanyi, J.C., 1975, Can the Maximin Principle Serve as a Basis for Morality? A critique of John Rawls's Theory, *American Political Science Review*, 69, 594–606.

Harsanyi, J., 1976, *Essays on Ethics, Social Behavior and Scientific Explanation*, Reidel Publishing Company.

Harsanyi, J.C., 1977, Morality and the Theory of Rational Behavior, *Social Research: An International Quarterly*, 44(4), 623–656.

Harsanyi, J.C., 1987, Morals by Agreement, David Gauthier, Oxford: Clarendon Press, 1986, 297 pages, *Economics and Philosophy*, 3(2), 339–351.

Harsanyi, J.C., 2008, John Rawls's Theory of Justice: Some Critical Comments, in M. Fleurbaey, M. Salles & J.A. Weymark (eds), *Justice, Political Liberalism and Utilitarianism, Themes from Harsanyi and Rawls*, Cambridge University Press, 71–79.

Hawi, R., 2016, *John Rawls. Itinéraire d'un libéral américain vers l'égalité sociale*, Classiques Garnier.

Igersheim, H., 2022, Rawls and the economists: the (im)possible dialogue, *Revue Économique*, 73(6), 1013–1037.

Igersheim, H., 2023, Samuelson against Rawls's gratuitism: some lessons on the misunderstandings between Rawls and the economists, *The European Journal of the History of Economic Thought*, 30(5), 883–905.

Jackson, B. & Stemplowska, Z., 2021, A Quite Similar Enterprise… Interpreted Quite Differently? James Buchanan, John Rawls and the Politics of the Social Contract, *Modern Intellectual History*, 18(4), 1010–1033.

Kandil, F., 2014, La justice est aveugle: Rawls, Harsanyi et le voile d'ignorance, *Revue Économique*, 65(1), 97–124.

Moehler, M., 2018, The Rawls–Harsanyi Dispute: A Moral Point of View, *Pacific Philosophical Quarterly*, 99, 82–99.

Peart, S.J. & Levy, D.M. (eds), 2008, *The Street Porter and the Philosopher*, Ann Arbor: University of Michigan Press.

Pogge, T., 2007, *John Rawls. His Life and Theory of Justice*, Oxford University Press.

Rawls, J., 1963, Constitutional Liberty and the Concept of Justice. In Rawls, J., 1999, *Collected Papers*, Harvard University Press.

Rawls, J., 1967, Distributive Justice. In Rawls, J., 1999, *Collected Papers*, Harvard University Press.

Rawls, J., 1968, Distributive Justice: Some Addenda. In Rawls, J., 1999, *Collected Papers*, Harvard University Press.

Rawls, J., 1971 [1999], *A Theory of Justice*, Revised Edition, Oxford University Press.

Rawls, J., 1974a, Reply to Alexander and Musgrave, *The Quarterly Journal of Economics*, 88, 633–655.

Rawls, J., 1974b, Some Reasons for the Maximin Criterion, *American Economic Review*, 64, 141–146.

Rawls, J., 1977, The Basic Structure as Subject, *American Philosophical Quarterly*, 14(2), 159–165.

Rawls, J., 1991, Questions on Reflection, *Harvard Review of Philosophy*, 1, 44–54.

Rawls, J., 1993, *Political Liberalism*, Columbia University Press.

Rawls, J., 1999, *Collected Papers*, Harvard University Press.

Rawls, J., 2001, *Justice as Fairness: A Restatement*, ed. E. Kelly, Harvard University Press.

Rosenkrantz, R., 1977, *Inference, Method and Decision*, Dordrecht-Holland: D. Reidel.

Rothenberg, J., 1961, *The Measurement of Social Welfare*, Englewood Cliffs, N.J.: Prentice-Hall.

Sen, A., 1976, Welfare inequalities and Rawlsian axiomatics, *Theory and Decision*, 7, 243–262.

Sen, A.K., 1984, *Resources, Values and Development*, Oxford: Basil Blackwell.

Sen, A.K. & Williams, B., 1982, *Utilitarianism and Beyond*, Cambridge University Press.

Von Neumann, J. & Morgenstern, O., 1944, *Theory of Games and Economic Behavior*, Princeton University Press.

Weymark, J., 1991, A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism, in J. Elster & J.E. Roemer (eds), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 255–320.

Weymark, J., 2005, Measurement Theory and the Foundations of Utilitarianism, *Social Choice and Welfare*, 25(2–3), 527–555.

Weymark, J., 2008, John Charles Harsanyi, in N. Koertge (ed.), *New Dictionary of Scientific Biography*, vol. *3*, Detroit: Charles Scribner's Sons, 247– 253.