Cahiers de Recherche
Série « Décision, Rationalité, Interaction »

Cahier DRI-2014-02

Representation Theorems and the Semantics of Decision-theoretic concepts

Mikaël Cozic et Brian Hill

INSTITUT D'HISTOIRE ET DE PHILOSOPHIE DES SCIENCES ET DES TECHNIQUES

Cahiers de recherche de l'IHPST

Série « Décision, Rationalité, Interaction »

Sous la responsabilité scientifique de Mikaël Cozic et Philippe Mongin

Representation Theorems and the Semantics of Decision-theoretic Concepts[1]

Mikaël Cozic[2] & Brian Hill[3]

Résumé : La théorie de la décision contemporaine accorde une importance considérable à une famille de résultats mathématiques qu'on appelle les théorèmes de représentations. Ces théorèmes relient des critères pour évaluer les options qui s'offrent au décideur (comme le critère de l'espérance d'utilité) à des axiomes qui portent sur ses préférences (comme l'axiome de transitivité). Plusieurs raisons ont été avancées pour expliquer ou défendre l'importance de ces résultats. L'objectif de cet article est d'évaluer leur rôle *sémantique* : dans cette perspective, les théorèmes de représentation ont pour fonction de fournir des définitions des concepts décisionnels mobilisés dans les critères d'évaluation (comme ceux d'utilité ou de probabilité subjective, qui sont mobilisés par le critère de l'espérance subjective d'utilité). Nous examinerons cette fonction en comparant les théorèmes de représentation aux théories philosophiques de la signification des termes dits théoriques.

Mots-Clés : théorie de la décision, axiomatisation, termes théoriques, utilité, probabilité

Abstract: Contemporary decision theory places crucial emphasis on a family of mathematical results called representation theorems, which relate criteria for evaluating the available options (such as the expected utility criterion) to axioms pertaining to the decision maker's preferences (for example, the transitivity axiom). Various claims have been made concerning the reasons for the importance of these results. The goal of this article is to assess their *semantic* role: representation theorems are purported to provide definitions of the decision-theoretic concepts involved in the evaluation criteria (such as those of utility or subjective probability that feature in the subjective expected utility criterion). In particular, this claim shall be examined from the perspective of philosophical theories of the meaning of theoretical terms.

Keywords : decision theory, axiomatization, theoretical terms, utility, probability

Classification JEL: B41, D81.

[1] This work was supported by the ANR "Institute of Cognitive Studies" under grant ANR-10-LABX-0087 IEC. Mikaël Cozic's contribution was also supported by a Junior Fellowship of the Institut Universitaire de France. The authors wish to thank the participants of seminars or conferences in Paris (SPS 2009, SOPHA 2012), Nancy (LMPS 2011), Munich (Center for Mathematical Philosophy), Rotterdam (INEM 2013) and Helsinki (EPSA 2013). We are particularly grateful to Itzhak Gilboa, Hannes Leitgeb and Philippe Mongin for their comments.

[2] Université Paris-Est Créteil, Institut Universitaire de France & IHPST. E-mail : mikael.cozic@ens.fr

[3] Centre National de la Recherche Scientifique & HEC Paris. Email address: brianhdehill@gmail.com.

1. **Introduction**

One of the particularities of contemporary decision theory, as practiced since the Second World War, is its *axiomatic* style. The axiomatic study culminates in mathematical results called *representation theorems*, results that are of critical importance in the discipline. Von Neumann & Morgenstern in the second edition of the *Theory of Games and Economic Behavior* (1944/1947) and Savage in his *Foundations of Statistics* (1954/1972) paved the way by proposing two representation theorems for the expected utility criterion[1]. Sixty years after these pioneering contributions, much of the theoretical research in decision sciences is still structured by the formulation of these representation theorems.[2] Why are these results held to be so important? Three sorts of roles may be assigned to representation theorems, which can explain, individually or collectively, their importance[3]:

> **Role (1):** Representation theorems provide '*foundations*' for the concepts introduced in the evaluation criteria[4] which are at the heart of decision models and according to which options are assessed. (Typical examples include the concepts of utility or subjective probability that are part of the subjective expected utility criterion.)[5] These concepts will be referred to hereafter as '*decision-theoretic concepts*'.
>
> **Role (2):** Representation theorems furnish an understanding of the *content* of an evaluation criterion, whether it is considered descriptively or normatively. This understanding allows us to *assess* the criterion from one or other of these points of view.[6]
>
> **Role (3):** Representation theorems play an *architectonic* role in the development of the different evaluation criteria. They allow comparison of different criteria, understanding of the logical relations between them, and may suggest new ones.

It is difficult to overestimate the importance of representation theorems in the traditional paradigm of decision theory. P. Wakker certainly reflects a widespread opinion among contemporary theorists when he claims that representation theorems 'change the status' of decision-theoretic concepts, turning them from '*ad hoc*' concepts into '*scientifically well-founded*' ones.[7] The current

[1] In fact, their work was preceded by that of F. Ramsey (1926). But the latter, which remained relatively unknown for a considerable period, certainly did not play the same structuring role in the history of the discipline. See Bradley (2004) for an excellent treatment of Ramsey's approach.

[2] From this point of view, the importance of representation theorems in decision theory is similar to that of completeness theorems in mathematical logic. The relationship with mathematical logic, and in particular with its notion of an axiomatic approach, is analyzed in detail by Mongin (2003).

[3] Gilboa (2009, p. 48) highlights the roles (1) and (2), and recognizes the two versions of role (1), semantic and operational, that we shall identify below.

[4] We use 'evaluation criterion' to refer to the rule for forming preferences over options; a typical example is the subjective expected utility criterion. See Section 2 for further discussion.

[5] Gilboa (2009) refers here to 'meta-theoretical interpretation'.

[6] Gilboa (2009) refers here to the descriptive and normative interpretations of the representation theorems, respectively. Others, to designate what we call role (2), refer to the axiomatic 'justification' of an evaluation criterion.

[7] Wakker (2010), p. 34.

article is part of a broader methodological project, which aims to examine each of these roles in order to provide an overall evaluation of the importance of representation theorems. This project has several motivations. The first lies in the fact that, despite the importance that they attribute to these results, decision theorists are generally not very explicit about their reasons for holding them in such esteem.[8] Secondly, the decision sciences are currently undergoing major changes, in particular in their relationship with cognitive science. In these fields of research and especially in so-called *behavioral economics*, representation theorems do not seem to be as important as they are in traditional decision theory. Finally, the few explanations that have been offered of the importance of representation sometimes reflect doctrines that have been largely abandoned in philosophy of science and in philosophy of language (notably operationalism and behaviorism). This leaves one wondering whether the importance attributed to these results is simply a relic of these doctrines, or whether contemporary philosophy of science has more valid reasons for assigning them a key role.

The aforementioned roles for representation theorems are generally presented as non-exclusive and largely independent; it is thus possible to examine them in isolation from each other. In this paper, we do just this for role (1), which states that they provide foundations for the concepts introduced in the evaluation criteria. Ever since the pioneering work of F. Ramsey (1926), the project of giving foundations for decision-theoretic concepts has involved two goals (at times without clearly distinguishing them). On the one hand, there is the *semantic* goal of elucidating or determining their *meaning* and, on the other, the *operational* goal of providing a method of *measuring* them. The importance of the operational goal is not to be underestimated: it was (and still is) one of the main reasons for elaborating representation theorems. It also explains why results originating in decision theory have become one of the paradigms of the *representational theory of measurement* (see Krantz et al. (1971)). Nevertheless, the modest aim of this article is to examine the first of these goals, the semantic one: is it true that representation theorems provide or determine the meaning of the concepts involved in the evaluation criteria?

Recognising that the concepts involved in the evaluation criteria have strong analogies to theoretical terms in the theories of the natural sciences, our strategy will be to analyze representation theorems through the lens of accounts of the meaning of theoretical terms developed in the philosophy of science. Despite the strong affinities between the general form common to representation theorems and an influential account of theoretical terms (due to Ramsey, Carnap and Lewis), the confrontation provides a challenge rather than a comfort to those wishing to defend a semantic role for representation theorems. Meeting this challenge requires either adopting a radical eliminativist position concerning theoretical terms or an anti-holist attitude towards meaning. In the latter, more interesting case, the defense nuances the semantic contribution of representation theorems: it is not the theorems as such, but their proofs (and only certain types of proofs) that provide the meaning of the decision-theoretic concepts. A similar space of, at times specific, positions can be drawn out under the main alternative account of theoretical terms. In a word, as far as the purported semantic contribution of representation theorems is concerned, we find that the situation is more subtle than generally realized: a semantic contribution may be defended, but it relies on particular philosophical positions about meaning, and comes from some unexpected quarters.

In Section 2, we review the pioneering representation theorems and provide a general formulation covering them. In Section 3, we link this general formulation to the specific features of

[8] Gilboa (2009), Wakker (2010) and Dekel & Lipman (2010) are recent exceptions.

decision-theoretic concepts. In Section 4, we introduce the main accounts of the semantics of theoretical terms, and in particular the Ramsey-Carnap-Lewis (RCL) account. In Section 5, we apply it to evaluation criteria. Sections 6 and 7 contain our examination of the semantic goal of representation theorems: first of all, we draw general morals from the viewpoint of the RCL theory, before turning to the specific, and prima facie unclear, issue of the purported attraction of representation theorems in the eyes of an anti-holist. It is also briefly noted that a similar space of positions is in principle available under the main alternative account of the semantics of theoretical terms. We present our conclusions in Section 8.

2. The General Form of Representation Theorems

Representation theorems involve three basic ingredients:

(i1) A framework: this includes, first of all, a set of options O among which the decision makers must choose. The options usually have a particular structure. It also includes the decision maker's own preference relation regarding the options in O.

(i2) An evaluation criterion **EC:**[9] this criterion determines a preference relation over O based a set of parameters x_1, ,..., x_n, that are assumed to be relevant - the decision-theoretic concepts. Formally, $\mathbf{EC}(O, x_1, ,\ldots, x_n) \subseteq O\times O$.

(i3) A set of axioms **Ax** on the preference relation on O×O. For example, the transitivity axiom is common to most axiomatizations.

A representation theorem *for an evaluation criterion **EC*** asserts a logical relationship between a preference relation $\preccurlyeq$ satisfying the set of axioms **Ax** and its being represented by the **EC**. More precisely, a *weak* representation theorem establishes that *if* a preference relation $\preccurlyeq$ satisfies the axioms **Ax**, *then* there are values (say x_1^*, ,..., x_n^*) for the parameters of the **EC** evaluation criterion such that the preference relation is $\preccurlyeq$. In other words: if $\preccurlyeq$ satisfies **Ax**, then there exist x_1^*, ,..., x_n^* such that $\mathbf{EC}(O, x_1^*, ,\ldots, x_n^*) = \preccurlyeq$. A *strong* representation theorem establishes complete logical equivalence. The two types of results that we have just defined are representation theorems *stricto sensu*. Usually, decision theorists establish representation theorems *lato sensu,* which include a representation theorem *stricto sensu* and a *uniqueness theorem*. This states that the values of the x_1^*, ,..., x_n^* parameters are unique, in the appropriate sense.

For example, in so-called situations of *risk* – where the probabilities of the outcomes that will be generated by the different options available are given (one often speaks of 'objective' probabilities) – the options are represented by lotteries, i.e., probability distributions $p \in \Delta(C)$ on the

[9] Wakker (2010) speaks of 'decision models' whereas we use the term 'evaluation criterion'. We prefer the word 'evaluation' over 'decision' because the immediate result of the criterion, as we see it, is a preference relation and not a selection of one or several options. A decision criterion selects a certain number of options from among O. The evaluation criteria that generate complete orderings normally correspond to decision criteria under which the agent chooses one of the best actions according to the generated preference relation.

set of outcomes C.[10] The most basic representation theorem involves a set **VNM** of axioms on preferences (that generally contains: transitivity and completeness, independence, continuity), and establishes the following:

Theorem (von Neuman and Morgenstern, 1944/1947)

The VNM axioms are satisfied if and only if there exists a utility function such that for all $p,\ q \in \Delta(C)$

$$p \preccurlyeq q \Leftrightarrow \sum\nolimits_{c\, \in C} p(c).u(c) \ \leq \sum\nolimits_{c\, \in C} q(c).u(c)$$

Moreover, u is unique up to positive affine transformation.[11]

This theorem concerns the expected utility criterion, which states that the preferences of agent *i* over the lotteries are determined by the expected utility calculated with i's utility function on outcomes. In this theorem, the utility function is the only 'parameter' involved in the theorem. Note, as is clear from this example, that several individuals can use (or be represented by) the same criterion, even if they have different preferences: any difference stems from the fact that they do not have the same utility functions.

To take another well-known example, in situations of *uncertainty* – where the probabilities that certain acts yield particular consequences are no longer exogenously given – the options of choice are *acts* $f \in$ A, i.e. functions from the set S of possible states of nature to the set C of outcomes. The main representation theorem in this domain relates a set **Sav** of axioms on preferences with the expected utility criterion, as follows:

Theorem (Savage, 1954/1972)

If **Sav** is satisfied, then there exists a (finitely additive[12]) probability distribution p on S and a utility function such that for all $f,\ g \in$ A

$$f \preccurlyeq g \Leftrightarrow \int_S u\big(f(s)\big)dp \ \leq \int_S u\big(g(s)\big)dp$$

Moreover, p is unique and u is unique up to positive affine transformation.

Note that, unlike the case of risk, the expected utility criterion for decision under uncertainty involves two 'parameters': beyond the decision maker's utility, there is a probability distribution, generally interpreted as a representation of the decision maker's beliefs. Both of the examples just given are representation theorems *lato sensu*, in the terminology introduced above.

A remarkable feature of representation theorems is their particularly rigorous conceptual discipline: whilst the evaluation criterion links the preferences to the decision-theoretic concepts in it, the axioms involve 'pure' properties of preferences.[13] We will refer to this pureness of axioms as

[10] This set of options has additional structure, relating to the operation of mixing lotteries for example.

[11] In other words, if the equation (1) is valid for the two utility functions u and u^0 , then there exists real numbers a and b, the first of which is positive, such that u^0 = au + b.

[12] The technical details of these results are of no particular interest to this study, and shall be ignored in what follows.

[13] This conceptual discipline is not unanimously respected in decision theory however. Thus, in the framework of Jeffrey's decision theory (1965/1983), J. Joyce (1998, chap. 4) formulates a representation theorem

preferentialism. Preferentialism is not self-evident. Even in theoretical economics, axiomatizations do not necessarily conform to this principle. An interesting example, which relates directly to the theory of individual decision making, is that of the axiomatic literature on choice in situations of 'total' uncertainty which embraces the axiomatic approach to evaluation or decision-making criteria (Milnor, 1954; Luce and Raiffa, 1957/1985; Maskin, 1979). In this framework, the options are Savage-type acts, and a utility function for the outcomes is generally given. The axioms can then involve this utility function; for example*,* the axiom of linearity requires that the application of the evaluation or decision-making criteria be invariant under positive affine transformations of the utility function. This is the kind of axiom that is excluded by preferentialism.[14]

3. Decision-theoretic concepts and their 'definitions'

As can be inferred from the previous remarks, preferentialism is related to the status of the decision-theoretic concepts. In general, these are accepted to be *subjective*, because of three different properties:

> (s1). They are the characteristics of the decision maker, and can *vary* from one decision maker to another. *A contrario*, the expected gain of a monetary lottery (i.e., a lottery whose outcomes are sums of money) is an 'objective' property of the lottery, which does not vary from one decision maker to another.
>
> (s2). One generally assumes that decision-theoretic concepts are not directly observable. In the abstract framework of decision theory, it is usually considered that it is the decision maker's *choices* or *behavior* that are observable. Often, but not always, decision theorists write as if the preferences are also observable; tacitly it is being assumed that there is a sufficiently strong relationship between preferences and choice. For some, such a relationship holds by definition: according to the *semantics of the revealed preference*, to prefer o_1 to o_2 is defined in terms of choosing (or intending to choose) o_1 over o_2.[15] In this case, the observability of preferences is claimed quite literally. For others, preferences are interpreted as states of mind preceding the choice, but it can be assumed firstly that if a decision maker knows that she is choosing between o_1 and o_2 and prefers o_1 to o_2, then she will choose o_1 rather than o_2, and secondly that consideration is limited to only those situations where decision maker is aware of the options available.[16] This assumption provides a sufficiently strong link between preferences and choice to allow one to infer the latter from the former. According to this second view, the observability of preferences is rather a simplification.

whose axioms are divided into two sets. The first is made up of 'pure' conditions regarding the preferences, but the second involves a comparative subjective probability relation.

[14] In fact, it is interesting to note that Milnor (1954) interprets the utility function in his axioms as a von Neumann-Morgenstern utility function. This implicitly implies that the von Neumann and Morgenstern axiomatics is assumed by (and thus plays a foundational role in relation to) this type of axiomatic study for choice under 'total' uncertainty. Maskin (1979, p. 320), for example, explicitly makes this assumption.

[15] This view has been recently endorsed by Gul and Pesendorfer: "To say that a decision maker prefers *x* to *y* is to say that he never chooses *y* when *x* is also available."(Gul & Pesendorfer 2008, p. 20).

[16] Fishburn (1970) takes this point of view: 'For a connection between decision and preferences, we shall assume that preferences, to a greater or lesser extent, govern decision and that, generally speaking, a decision maker would rather implement a more preferred alternative than one that is less preferred.' (p. 1)

(s3). Decision-theoretic concepts are often considered as (or as closely tied to) mental attitudes. The probability distribution p and the utility function u are thus frequently interpreted as quantifying the decision maker's *degrees of belief* and *degrees of desire,* respectively.[17]

The subjectivity of decision-theoretic concepts, and particularly characteristic (s2), explains in part why decision theorists feel the need for an axiomatization, and why canonical axiomatizations are preferentialist in form: contrary to preferences, which enjoy an *epistemic privilege*, decision-theoretic concepts are not considered to be 'given' to the modeler. Moreover, this justifies the existential form in which decision-theoretic concepts appear in representation theorems. If there was a way of observing, or otherwise determining, the value of the decision-theoretic concepts of an agent i, if for example there was only one decision-theoretic concept x and that its value was proven to be x^*, then the agent would choose according to the evaluation criterion **EC** only if $\mathbf{EC}(0, x^*) = \preccurlyeq_i$. In this case, a representation theorem that links the criterion **EC** with the set of axioms **Ax** is far less informative regarding the question of knowing whether agent i actually satisfies the criterion **EC**. By simply knowing that the agent satisfies **Ax**, one cannot conclude that $\mathbf{EC}(0, x^*) = \preccurlyeq_i$: the axioms imply only the existence of a x such that the preferences are represented by x according to **EC**, but do not guarantee that $x = x^*$.[18]

Conventional decision theory not only gives preferences an epistemic privilege over decision-theoretic concepts: it readily assigns representation theorems the task of 'giving foundations for' these concepts. We will focus on the semantic version of this foundational role. The semantic goal of representation theorems emerges in decision theorists' claims that these results '*give meaning*' to decision-theoretic concepts or provide them with a *definition*; for example it is common to present representation theorems for choice under uncertainty as 'definitions' of the concept of subjective probability.[19] Sometimes the claim is spelt out in terms of '*behavioral* definitions'[20] or 'definitions *in terms of preference*'[21] – the variations are understandable in the light of the different interpretations of the concept of preference discussed above. However it is not easy to determine if, and in what way, representation theorems do provide or elucidate the meaning of the decision-theoretic concepts involved in an evaluation criterion: each representation theorem does not have its own *definiens,* strictly speaking.

[17] See in particular Jeffrey (1983, chap. 4).

[18] Similar remarks could be made about revealed preference theory. This theory is usually applied to choice under certainty, where one basically assumes that an agent chooses what she prefers and that her preferences are transitive and complete. Moreover, one then assumes that it is the choices and no longer the preferences that are 'given'. The basic concept of rationalizability presents preferences in an existential form: a decision maker's behavior is rationalizable if there exists a preference relation which can generate it. One often considers that a decision maker with rationalizable behavior chooses according to the theory of choice under certainty. Strictly speaking, this holds only under the hypothesis that the decision maker's preferences are not given independently of her choice behavior. If one assumes that the preferences of decision maker i are given, and that her choice behavior can be rationalized by a relation other than then one can no longer say that the i chooses according to the theory.

[19] Aumann and Anscombe (1963), Machina and Schmeidler (1992), Karni (1993), Gilboa (2009, pp. 128-129).

[20] Gilboa (2009).

[21] Jeffrey (1983, p. 74).

4. The Semantics of Theoretical Terms

We propose to deal with the question of the semantic contribution of representation theorems by drawing on the literature in the general philosophy of science on *the meaning of the 'theoretical' terms*. Given the purported difference in observability between the decision-theoretic concepts and the preferences involved in representation theorems (point (s2)), there is a natural analogy between the former and theoretical terms of the natural sciences, such as 'electron' or 'gene'. Although this relationship has not, to our knowledge, been explicitly discussed in the philosophy of economics, it provides a promising avenue for assessing the semantic claims made for representation theorems. After briefly recalling the main theories of the semantics of theoretical terms in this section, we turn, in subsequent sections, to the application of these theories to the case of decision-theoretic concepts and representation theorems.

The problem of the meaning of theoretical concepts is usually presented as follows: one assumes that theory *T* is formulated in a certain language, as a set of propositions, and that one can distinguish, in one's conceptual repertoire, between two categories of terms. In the neo-positivist tradition, theoretical terms are contrasted with observational terms, where a term is considered observational when you can determine through observation whether or not it applies to an entity in its domain of application. D. Lewis (1970, 1972) liberalizes the distinction: the 'theoretical' terms are terms that are *introduced* by a theory *T*, and they are contrasted with terms whose meaning was determined *prior* to the theory *T*. For the discussion here, there is no need to decide between these distinctions. As standard, we will use the word *t*-terms to designate theoretical terms, and *o*-terms to designate observational terms *or* those introduced prior to the theory *T*.

Broadly speaking, there are currently two families of approaches to the semantics of theoretical terms in the literature: a descriptivist approach, and a causal-historical-based approach. For reasons that shall become clear below, much of the focus here will be on the descriptivist approach, which is naturally formulated in terms of definitions. One way of determining the meaning of a theoretical concept is to provide it with an *explicit definition*, that is, to formulate, using the *o*-terms, a complex property equivalent to the theoretical concept. Contemporary notions of the meaning of theoretical concepts are largely derived from the long-accepted observation[22] that, in general, one cannot find or expect to find these types of explicit definitions in scientific theories. A natural proposal is thus that theories *implicitly define* the theoretical concepts they contain. This is the idea behind one of today's most influential theories, namely David Lewis's (1970, 1972), which expands on the logico-semantic studies initiated by F. Ramsey (1929) and continued by R. Carnap (1959, 1966). For the sake of simplicity, we refer to it as the Ramsey-Carnap-Lewis (RCL) approach. In short, the main idea is that theoretical terms derive their meaning from the context of the theory in which they appear, a theory that defines them *implicitly* and *simultaneously*. Carnap and Lewis each propose a method to *make explicit* the way in which these theories implicitly confer meaning on their theoretical terms.[23] In the

[22] See Hempel (1950, pp.56-7) and Carnap (1956): 'It is true that empiricists today generally agree that certain criteria previously proposed were too narrow; for example, the requirement that all theoretical terms should be definable on the basis of those of the observation language...' (p. 39).

[23] In fact, there are a variety of possible applications of the Ramsey-Carnap-Lewis approach. Carnap (1966) refers to a first-order language that contains the symbols of theoretical predicates, and the Ramsey sentence is generated by existentially quantifying on *second*-order variables. Carnap also considers that when one is interested in mathematizing theories as is the case in contemporary physics, the Ramsey sentence will quantify in rich mathematical domains (natural numbers, the classes of natural numbers, the classes of classes of numbers,

sequel, we shall mainly follow the Lewis formulation. According to Lewis, it is possible to provide *explicit definitions* by the following method:

1. Assume that we start with a T theory comprising the t-terms $t_1,...,t_n$ as well as observational and logico-mathematical terms:

$$T[t_1, ,\ldots, t_n] \ (\mathbf{T})$$

2. Replace the t-terms by variables, stipulating that each variable not only be realized, but that it be so *uniquely*:

There exist unique $x_1, ,\ldots, x_n$ such that $T[x_1, ,\ldots, x_n]$ (**Qu-T** or **T's Lewis sentence**)[24]

3. Define the theoretical terms as the (unique) things that satisfy $T[x_1, ,\ldots, x_n]$. More exactly, each term t_i is defined as the i-th component *T's* unique realization, or

t_i is the unique x_i such that there exist unique $x_1, ,\ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ such that $T[x_1, ,\ldots, x_n]$ (**Def-Lew-*t_i*** or the **Lewis definition of *t_i***)[25]

In Lewis's theory, if there exists a unique realization of T, then t_i denotes the i-th component of this realization, whereas if there is no unique realization (if there are several realizations, or if there are none at all), then t_i denotes nothing, like a term of fiction.[26]

The RCL semantics for theoretical terms has natural affinities with the descriptivist theory of common nouns and proper names in the philosophy of language; in this domain, the main challenger to this theory is the direct-reference or causal-historical theory. Rather than associating a description

etc.). He refers to this language as the '*extended observational language*'. On the other hand, in Lewis (1970), the domain of individuals is assumed to be enriched, the theoretical terms are constants, and the Ramsey sentence (or the Lewis sentence) is generated by existentially quantifying on *first-order* variables. To take yet another example, Ketland (2004) uses the resources of model theory and formulates the original theory T in a multi-sorted first-order logic: the language is interpreted in two domains: a domain of observable individuals and a domain of theoretical individuals. The predicate symbols are interpreted either as relations in the domain of individuals that can be observed (observational predicates), or as relations in the domain of theoretical individuals (theoretical predicate), or as relationships between individuals in the two domains (mixed predicates).

[24] This is usually denoted $\exists\,!\; x_1, ,\ldots, x_n \;\; T[x_1, ,\ldots, x_n]$The Lewis sentence differs from the *Ramsey sentence*, logically weaker, which affirms the existence of a realization of $T[x_1, ,\ldots, x_n]$ without requiring that the latter be unique.

[25] What we denote $t_i = \iota x_i \exists ! x_1 ... x_{i-1} x_{i+1} ... x_n T[x_1 ... x_n]$ using the notation ι for the defined description.

[26] In two texts that have remained relatively unknown until recently, Carnap (1959, 1961) develops definitions of the sort proposed later by Lewis, but without requiring that the uniqueness of the realization. He uses the Hilbert operator that can be seen as an *indefinite description* operator; in the case of a language with a theoretical term t, the latter is defined as designating any of the entities that realizes the Ramsey sentence of the theory. For more detail, see the introduction to Carnap's text (Psillos, 2000).

to a term – understood as its meaning – and fixing the reference of the term to be whatever satisfies that description, the direct-reference theory provides an account of how the reference of a term is fixed in terms of the causal chain of uses of the term, culminating in an original reference-fixing act: the reference of the term is the object referred to by the speaker in that act (Kripke 1972/1980, Putnam 1975). Whilst this account, as developed for common nouns and proper names, has evident shortcomings when applied to theoretical terms (Enç, 1976, Psillos, 1999),[27] a version can be developed for them, which, to the extent that it draws on the descriptivist as well as the causal-historical approach, can be called, following Psillos (1999), the causal-descriptivist theory.[28] To the best of our knowledge, this is currently the main alternative to RCL semantics as an account of the meaning of theoretical terms. Psillos formulates his causal-descriptivist position as follows: 'a theoretical term *t* typically refers by means of a core causal description of a set of kind-constitutive properties, by virtue of which its referent *x* is supposed to play a given causal role in respect of a certain set of phenomena.'[29]

Note that the framework provided by the causal-descriptivist account of theoretical terms does not *prima facie* provide a fruitful context for understanding the claims made for the semantic importance of representation theorems. In particular, the view does not fit easily with the language of decision theorists. Firstly, like the causal-historical approach, the central question under this account is not the meaning of a theoretical term, but its reference (or how it is fixed). As such, it is at odds with the standard language used by decision theorists who, as noted in Section 3, speak unabashedly of *definitions*. Secondly, the language and philosophy of causal-descriptivist theories is not that which is used by most decision theorists, nor that with which they would be most comfortable. Picking out the reference suggests that subjective probabilities or utilities exist in some strong sense of the word, and the insistence on a *causal* description is at odds with the purported 'as if' character of representations.[30] Whilst these ontological or metaphysical concerns may not be inexorable,[31] they seem sufficient reason to focus our discussion on the evaluation of representation theorems in the light of the RCL semantics.

5. Decision-Theoretic Concepts as Theoretical Terms

Now let us consider how Lewis's method for defining theoretical terms could be applied to an evaluation criterion (**EC**), such as the subjective expected utility criterion. Firstly, one must separate the vocabulary of the evaluation criterion into *t*-terms and *o*-terms. Many of the points made in

[27] For instance, on a purely causal account of the reference of theoretical terms, it is very difficult to explain how one could become convinced that an entity or property posited by a scientific theory (e.g., *phlogiston*) does not exist. Such cases of referential failure are hard to explain without involving some descriptions.

[28] To the best of our knowledge, the main two existing developed versions of this theory of the meaning of theoretical terms in the literature are due to Enç (1976) and Psillos (1999). Whilst there are differences between their two theories, they shall not be central for the purposes of the current discussion.

[29] Psillos (1999, pp. 295-296).

[30] A view of this kind is put forward, for example, by F. Gul and W. Pesendorfer in their methodological criticism of neuroeconomics. After having claimed that '…standard economics does not study the causes of preferences', they say the following about a representation theorem they give as example: 'While the formula is suggestive of a mental process this suggestiveness is an expositional device not meant to be taken literally' (Gul & Pesendorfer, 2008, p. 15).

[31] For example, the emphasis on the causal aspect is not shared by all versions of the causal-descriptive theory; Enç for instance does not require the causal element.

Section 3, and in particular characteristic (s2), justify the assumption that (i) the preferences belong to the o-terms, while (ii) the decision-theoretic concepts used by the evaluation criteria are t-terms. Under this assumption, there is a distinction between the **(EC)** language, which contains both t-terms and o-terms, and what could be called the '*simple language of preference*', in which properties of preferences such as transitivity can be expressed. In particular, the axioms in representation theorems are formulated in the simple language of preference. Note that although, as is evident from the remarks cited in Section 3, decision theorists' own positions are fully compatible with this way of drawing the line between t- and o-terms, it is not beyond debate. In particular, whilst characteristic (s2) justifies the distinction, (s3) might, at first glance, seem to undermine it: given the affinities between decision-theoretic and folk-psychological concepts (for example, between subjective probability and belief or between utility and desire), it may be asked to what extent the concepts used by decision theorists are truly theoretical terms, rather than (pre-existing) ordinary-language terms. However, there are reasons to suspect that an objection along these lines is flawed. As Enç (1976) has pointed out in his discussion of similar examples from the natural sciences, there is a difference between terms such as 'heat' and 'magnet' on the one hand and 'caloric' or 'magnetic field' on the other, and theories of theoretical terms can quite uncontroversially be understood as pertaining specifically to the latter sorts of terms. It seems that the analogy between, for instance, the pair belief-subjective probability and heat-caloric is sufficiently strong to justify treating decision-theoretic concepts as t-terms for the purposes of this discussion.[32]

Focusing on the concept of utility, the steps in the definition of this concept are as follows:

1. For all $o_1, o_2 \in$ O, $o_1 \preccurlyeq o_2 \Leftrightarrow E_p u(o_1) \leq E_p u(o_2)$ (**EC**)

2. There exists (appropriately) unique utility and probability functions u and p such that, for all $o_1, o_2 \in$ O, $o_1 \preccurlyeq o_2 \Leftrightarrow E_p u(o_1) \leq E_p u(o_2)$ (**Qu-EC**)

3. the utility of the decision maker 'is' the (appropriately) unique function u such that there exists a unique p resulting in $, o_1 \preccurlyeq o_2 \Leftrightarrow E_p u(o_1) \leq E_p u(o_2)$ for all o_1, o_2 (**Def-Lew-*u***)

One could obviously proceed in the same manner for the concept of subjective probability. Several remarks are in order. First, this way of defining subjective utility and probability is very similar to the way in which, in the philosophy of mind, functionalists (such as Lewis himself) characterize ordinary beliefs and desires. More exactly the definitions are similar to *forward-looking* features in the characterization of mental states, i.e. features that refer to their *effects*, in contrast with *backward-looking* features, which refer to their *causes*.[33] Secondly, (**EC**) and (**Qu-EC**) have

[32] Note that whilst this analogy with the case of the natural sciences works well on the descriptive interpretation of representation theorems (see Section 1), there is an important disanalogy with the normative interpretation, insofar as the normative dimension is absent in the natural sciences. This disanalogy does not endanger the point made in the text: if there is sufficient difference between belief and subjective probability to treat the latter as a theoretical term on the descriptive interpretation of representation theorems and evaluation criteria, there appears to be no reason why such a difference would not subsist under the normative interpretation.

[33] Indeed, in his pioneering article, Ramsey (1926) is very close to (**Def-Lew-*P***) when he states:

consequences that are expressible in the simple language of preference, some of which, several idealizations notwithstanding, can be tested. Transitivity of the decision maker's preferences, for example, is a consequence of (**EC**) as well as of (**Qu-EC**). Indeed, the starting point of the Ramsey-Carnap-Lewis approach is the result that *T* and its Ramsey sentence logically imply the same sentences expressible in the 'observational fragment of the language' – that is, in the current case, in the simple language of preference.

What does this confrontation with the RCL approach to defining theoretical terms have to say about representation theorems? It is immediately evident that the product of the second step of Lewis's method, (**Qu-EC**), is *exactly* one of the two propositions involved in the representation theorem *lato sensu*.[34] However, according to Lewis's procedure, (**Qu-EC**) is *sufficient* to provide, in an explicit way, the meaning of subjective concepts. This is what was demonstrated with (**Def-Lew-*u***). Hence, even on the basis of the RCL semantics for theoretical terms, which is the most amenable to the terms of the discussion had by decision theorists, the semantic contribution of representation theorems is far from evident.

6. RCL Semantics and Representation Theorems

Rather than giving a straightforward justification of the semantic claims for representation theorems, the confrontation with RCL semantics has raised two central questions: in what sense (if any) does a representation theorem provide a definition of the subjective concepts in the corresponding evaluation criterion? And why should such a definition be considered superior or more satisfying to a definition à la Lewis, which can be formulated simply on the basis of the evaluation criterion? The aim of this Section and the next one is to provide answers to these questions.

Since, as is evident from the preceding discussions, representation theorems provide something over and above the Lewis definition of theoretical terms, any purported semantic role for representation theorems would have to rest on some reservation about or criticism of the Lewis definition. Three such reservations have been expressed in the literature, each of which may in principle provide a semantic role for representation theorems.

(r1). The first concerns the purported *elimination* of theoretical terms in the Lewis definitions. The theoretical *terms* are strictly speaking *eliminated* in Lewis (and Ramsey) sentences, but they are replaced by existentially quantified variables of the appropriate type. In this sense, there is no fundamental difference between the interpretations of the original *T* theory and its Lewis (or Ramsey) sentence: the two make certain assertions about the domain of the original theory's theoretical terms. This point, already emphasized by Hempel (1958)[35] and by Lewis himself,[36] makes the purported

'I suggest that we introduce as a law of psychology that [a man's] behaviour is governed by what is called the mathematical expectation, that is to say if *p* is a proposition about which he is doubtful, any goods or bads for whose realization *p* is in his view a necessary and sufficient condition enter into his calculations multiplied by the same fraction, which is called the 'degree of his belief in *p*'. We thus define degree of belief in a way which presupposes the use of the mathematical expectation.'

[34] The representation theorem *stricto sensu*, without a uniqueness result, links the axioms with T's Ramsey sentence.

[35] '...the Ramsey-sentence associated with an interpreted theory T' avoids reference to hypothetical entities only in letter - replacing Latin constants by Greek variables - rather than in spirit. For it still asserts the existence of certain entities of the kind postulated by T', without guaranteeing any more than does T' that those entities are observable or at least fully characterizable in terms of observables' (p. 81)

elimination quite *trivial*. Similar points hold for the Lewis definitions: each theoretical term is defined *without using* the other theoretical terms. But it is defined in *the extended language of preference*, to use Carnap's (1959) term modified for the case at hand, rather than the *simple* language of preference, and involves moreover the existentially quantified versions of the other theoretical terms. The Lewis definition of utility (**Def-Lew-*u***), for example, involves the quantified version of the subjective probability concept.

(r2). The second worry concerns the *holistic* character of the Lewis (or Carnap) definitions. These definitions allow each theoretical term to be defined individually, but they only do so by using the whole theory, or more precisely its Lewis (or Ramsey) sentence, and therefore the quantified versions of the other theoretical terms. Given this is interdependence, this account determines the meaning of theoretical terms in a way that can be characterized as *holist.*[37] The holism of the definitions can be set out as follows:

(h1). the definition of each theoretical term involves the whole (Ramsey or Lewis sentence of the) theory; it follows in particular that

(h2). the definition of each theoretical term takes into account the other theoretical concepts in their quantified version, so that the semantic dependency relationship forms a network rather than a chain starting from the preference concept (the *o*-term of the theory); and that

(h3). the definition of a theoretical term takes into account all of the individual's preferences, and not a limited set of them, insofar as it specifies the overall contribution of what the term indicates for the determination of preferences;

[36] 'My proposal could be called an elimination of theoretical terms; for to define them is to show how to do without them. But it is better called a vindication of theoretical terms; for to define them is to show that there is no good reason to want to do without them.' (p. 427)

[37] On semantic holism in general, see Pagin (2006). The holism we are concerned with is closer to what Peacocke (1997, pp. 243-4) refers to as 'local holism', by contrast with generalized holism, relating to the language as a whole, which is widely discussed in the philosophy of language. The characterization of the Lewis defintions as `holist' may be objected to. Psillos (2008), for example, deals with the question of whether or not the late Carnap's account of theories and theoretical terms come under semantic holism. He discusses, in particular, Carnap–type explicit definitions that are based on the ε–operator. According to Psillos, Carnap saw in these definitions a way to 'restore a total semantic atomism to theoretical terms'. This sentence seems to conflict with (r2). It seems to us that the resulting 'semantic atomism' is too superficial to clearly avoid the charge of 'holism'. Once again, although the explicit definition of a theoretical term does not involve other theoretical terms, it does use their existentially quantified expressions. Indeed, just after having argued that Carnap definitions should be considered as atomic, Psillos (2000, p. 157) qualifies his claim as follows: 'To be sure, each and every theoretical term is explicitly defined relative to the *n*-tuple *t* of the theoretical terms of the theory. Still, however, relative to this *n*-tuple, the meaning of each and every theoretical term of the theory can be fully disentangled from the meanings of the rest' (our emphasis). This seems to be in accord with the qualification of such definitions as locally holist, in the sense of Peacocke. It is nonetheless certain that Carnap saw his definition as a valuable step forward compared to his prior use of Ramsey sentences to disentangle the analytical content of a theory: 'I thought very briefly about that question [whether there is a way of giving explicit definitions for all the theoretical terms in the observation language] years ago and I just dismissed it from my mind, because it seems so obvious that is is impossible...Now , it is possible. I found that only a few weeks ago and I hope I have not made a mistake...So, in the hope that there is something in it, I will now present the way of doing this by explicit definitions, which is really so surprising that I still can hardly believe it myself.' (Carnap, 1959, p. 168)

(h4). the theoretical terms are defined globally rather than 'argument-by-argument': the definition does not give the conditions under which it is true, for example, that the subjective probability associated with an event E is α.

(r3). A final noteworthy characteristic of the Lewis definitions, often discussed in the literature on the semantics of theoretical terms, is that if theory T is false (more exactly, if its Lewis sentence is false), then the theoretical terms it contains do not denote anything.

As suggested above, each of these points could in principle be used as reasons in favor of a semantic contribution for representation theories. It is straightforward to see how the last point, characteristic (r3), might be used as in a semantic argument in favor of representation theorems: to the extent that the axioms facilitate the evaluation of the truth of the evaluation criterion, they help determine whether the decision-theoretic concepts refer at all. Such an argument would make the semantic contribution of representation theorems tributary to their evaluative role (role (2) in Section 1), and any proper appraisal of the argument would require a full analysis of the extent to which representation theorems are essential to assessing the descriptive adequacy of an evaluation criterion.[38] Since this brings us beyond the semantic question that is the topic of this article, we shall leave this argument aside and focus the discussion on worries (r1) and (r2).

For each of these two worries, one can *a priori* conceive of a 'weak' version and a 'strong' version, themselves providing reasons of differing strengths in favor of axiomatization. The 'weak' reason pertains to meaning indirectly, via the (psychological) concept of understanding: the idea is not that the Lewis sentence (**Qu-EC**) or the Lewis–type definitions (**Def-Lew-*t_i***) do not give the meaning of decision-theoretic concepts, but rather that the axiomatization furnished by a representation theorem, formulated as it is in the simple language of preference, yields a *better understanding of them*. The, more radical, 'strong' reason brings into doubt the acceptability of (**Qu-EC**) or (**Def-Lew-*t_i***) definitions. Factoring in the two strengths, one obtains four reasons for axiomatizing an evaluation criterion (**EC**):

(r1-w) The RCL approach characterizes the theoretical concepts involved in evaluation criteria in an *extended* language of preference rather than a simple language, and that these concepts are *harder to grasp* in the former than in the latter. By contrast, the axioms that are related to the evaluation criterion by the representation theorem are formulated in the simple language of preference.

(r1-s) The strong version of this worry originates in an eliminativist attitude towards theoretical terms; the Lewis (or Ramsey) sentences realize only a trivial form of elimination, whilst a genuine elimination is desired. The suggestion would be that an axiomatisation provided by a representation theorem is a genuine elimination of the decision-theoretic concepts.

(r2-w) Another motivation, related to the holistic nature of the Lewis definitions, does not to reject them definitions because they are holistic, but rather notes that, because of their holism, understanding the theoretical concepts essentially requires understanding the theory itself. Axiomatization aids understand of the concepts, because by providing an alternative formulation of the original theory, it facilitates understanding of the theory.

[38] For example, it would have to be explained what a set of axioms that are equivalent to the use of an evaluation criterion brings to the assessment of the descriptive adequacy of the criterion over and above the testing of salient consequences of the criterion.

(r2-s) The strong version of this point is based, in contrast, on an anti–holistic attitude: it rejects the way in which the subjective concepts are defined by the Lewis method because it is holist.

The two 'weak' reasons (r1-w)-(r2-w) for axiomatizing an evaluation criterion are not insignificant, but it is doubtful that they account for the importance traditionally attributed to representation theorems. At best, they highlight the advantages of axiomatic formulations for the *understanding* of the decision-theoretic concepts, and more generally of the evaluation criterion: the formulation in the simple language of preference, and the involvement of a limited number of properties that can be immediately and independently grasped makes comprehension easier. One could call these advantages *heuristic* or *psychological*, but one could also speak of a contribution to the meaning of the concepts, as the decision theorists do, as long as one takes a broad view of 'meaning', which is no doubt far less precise than those dominant in philosophy.[39]

Among the strong potential reasons in favor of axiomatisation, the first, (r1-s), makes a clear difference between the RCL semantics of theoretical terms and the axiomatisation offered by representation theorems. However, as an argument in favor of the latter, it relies on an eliminativist atttitude towards the decision-theoretic concepts that appear in evaluation criteria, and such an attitude is not generally considered as particularly attractive nowadays. Furthermore, on the descriptive question of understanding the importance of representation theorems in the discipline, it is doubtful that decision theorists have the same ontological scruples as eliminativists.

As concerns the remaining reason (r2-s), stemming from an anti–holistic attitude, it is unclear how the anti-holist criticism can be used as an argument in favor of representation theorems. Indeed, on examination of many representation theorems – for instance Savage's theorem for decision under uncertainty – it is evident that the decision-theoretic concepts are derived *collectively* from the axioms, and concomitantly with the evaluation criterion. What advantage could such theorems offer, as far as holism goes, over the Lewis definition of decision-theoretic concepts? It is to this question that we now turn.

7. Holism and Preferential Definitions

At first glance, it may seem that, for an anti-holist, representation theorems have no advantage over Lewis definitions of theoretical terms. Before jumping to such a conclusion, more careful consideration is required of the *proofs* of representation theorems. For many important representation theorems, there exist proofs which, for reasons that will become obvious, can be qualified as *constructive*. For example, in the algebraic proof of the von Neumann Morgenstern[40] representation theorem, a utility function u on the C set of outcomes is defined as follows:

$$u(c) = \alpha \text{ if and only if } c \sim [c^*,\alpha\ ;\ c_*,(1-\alpha)] \quad (\textbf{Def-pref}^{\textbf{Risk}}\textbf{-}\boldsymbol{u})$$

[39] This type of consideration is emphasized by Von Neumann & Morgenstern (1944/1947, p. 25) in the following excerpt, cited in Mongin (2003), who analyses it in detail: 'The axioms should not be too numerous, their system should be as simple and transparent as possible, and each axiom should have an intuitive meaning by which its appropriateness may be judged directly. In a situation like ours this last requirement is particularly vital in spite of its vagueness: we want to make an intuitive concept amenable to mathematical treatment and to see as clearly as possible what hypotheses this requires.'

[40] See for example Kreps (1988). Other proofs, including some that are not constructive in the sense used here, are given in Gilboa (2009).

where c^* is one of the best outcomes (from the decision maker's point of view), c_* one of the worst and $[c^*,\alpha ; c_*,(1-\alpha)]$ the lottery that yields c^* with probability α and c_* with probability $(1-\alpha)$. This type of definition is relevant for the matter we are dealing with because it does not directly involve the evaluation criterion (in this case, expected utility) and does not define utility by its role in the evaluation criterion. Moreover the utility is determined 'argument-by-argument', for each outcome $c \in C$. Indeed, the definition is essentially a *measurement method*: it proposes a way of measuring the value of the utility function for each of its possible arguments. Moreover, this definition assumes that the decision maker's preferences satisfy a certain number of conditions (first and foremost, that for each outcome c, there is a unique α such that $c \sim [c^*,\alpha ; c_*,(1-\alpha)]$. In fact, the proof proceeds by showing that the axioms guarantee that these assumptions hold. (**Def-pref$^{\text{Risk}}$-*u***), the definition of utility in terms of preference, is thus both atomic and operational. To this extent, it provides an aspect on which von Neumann Morgenstern's representation theorem fairs better than the Lewis definition in the eyes of an anti-holist. The reason (r2-s) can be used to formulate an argument in favor of the representation theorem in this case. A definition of the theoretical term that does not suffer from the charge of holism affecting the Lewis definition is provided: by an appropriate constructive proof of the representation theorem, rather than by the theorem itself .

The case of subjective expected utility is obviously more interesting, since it involves 'defining' not one but two subjective concepts. There is no reason that the points made above for choice under risk should continue to hold. Yet Savage's proof of his representation theorem (1954/1972) shows how to define the two concepts of utility and subjective probability *sequentially* from preferences. One starts by constructing a definition of qualitative (or relational) probability $>_P$, then of quantitative probability P and finally of utility u. The definition of the qualitative probability relation, from which the construction starts, is as follows:

$E_1 >_P E_2$ if and only if $\exists x, y \in C$ such that $x \succ y, xE_1y \succ xE_2y$ (**Def-pref$^{\text{Unc}}$->$_P$**)

where $E_1, E_2 \subseteq S$ are events, C is the set of outcomes and xE_iy is the act whose outcome is x if E_i is the case and y if not. According to (**Def-pref$^{\text{Unc}}$->$_P$**), an agent considers E_1 more probable than E_2 precisely when, if she prefers one outcome to another, she prefers the bet that delivers the preferred outcome if E_1 holds over than the bet that delivers the preferred outcome if E_2 holds. Savage's construction then goes on to define quantitative probability (**Def-pref$^{\text{Unc}}$-*P***), before ending with a definition (**Def-pref$^{\text{Unc}}$-*u***) of the utility function, which is derived essentially using the von Neumann-Morgenstern method mentioned above.[41] It is important to highlight the difference between the sequential organization of the definitions (**Def-pref$^{\text{Unc}}$-*u***) and (**Def-pref$^{\text{Unc}}$-*P***), and the Lewis definitions (**Def-Lew-*u***) and (**Def-Lew-*P***), which are given simultaneously. In particular the definitions of subjective probability and utility drawn from the proof of the representation theorem are, unlike the Lewis definitions, naturally ordered by the construction: (**Def-pref$^{\text{Unc}}$-*P***) does not assume (**Def-pref$^{\text{Unc}}$-*u***), whereas (**Def-pref$^{\text{Unc}}$-*u***) builds upon (**Def-pref$^{\text{Unc}}$-*P***).[42]

If one assumes that the axioms are satisfied, then the definitions in terms of preferences (**Def-pref**) seem to satisfy the strictest empiricist and operationalist criteria. Indeed, they correspond to

[41] Savage (1954/72), pp. 75-6.

[42] See Carnap (1936, 1956). Peacocke (1997, p. 229) puts forward order as a requisite for anti–holism.

what Carnap (1936/37) refers to as '*explicit definitions*'.[43] The benchmark representation theorems for the expected utility criterion thus provide instances where there is a strong semantic argument in favor of representation theorems. These theorems, and more specifically the proofs discussed above, contain explicit definitions of decision-theoretic concepts that, in the eyes of an anti-holist (ie. someone moved by (r2-s)), are preferable to the Lewis definitions.

Note that, since this defense pertains not so much as the representation theorem *result* as to its *proof*, it has perhaps unexpected consequences for which representation theorems, and proofs, can be thought of as giving the meaning of decision-theoretic concepts. For example, there are several different proofs of the von Neumann Morgenstern result (Gilboa, 2009, Ch . 9), and whilst, as we saw above, some involve explicit definitions of utilities, others do not. An example is the proof (given in the previous reference) using a separation argument: the proof applies a separating hyperplane theorem to deduce the existence of a utility function, without explicitly defining the function. The preceding discussion would suggest that only the former, and not the latter proof provides the meaning of the decision-theoretic concepts, and hence accomplishes the semantic role attributed to the representation theorem.

Moreover, the dependence on the particular forms of the proofs implies that whilst some representation theorems may give the meaning of the decision-theoretic concepts in a satisfactory way, others may not. For example, the standard proof of the benchmark representation theorem for so-called maximin expected utility model, provided by Gilboa & Schmeidler (1989), relies heavily on a separating hyperplane theorem, and, like the proof of the von Neumann Morgenstern theorem mentioned above, does not provide an explicit definition of one of the decision-theoretic concepts involved (the set of probability measures). We are not aware of other proofs of this theorem; if no other proofs have been proposed, then the preceding considerations imply that this representation theorem *does not* provide the meaning of the decision-theoretic concepts involved, or at least, that it contains no semantic contribution beyond the Lewis definition of the appropriate decision-theoretic concepts. A new, appropriately constructive proof would be required, in which the decision-theoretic concepts are explicitly defined.

In summary, a strong semantic contribution for representation theorems can be identified, under the RCL account of theoretical terms, if one adheres either to some form of reductionist empiricism[44] (r1-s) or to a particularly robust form of semantic anti–holism (r2-s). In the latter case, the contribution comes not from the result itself, but from its proof, and more specifically from the definitions of decision-theoretic concepts involved in it. This may have interesting consequences in of itself: in particular, it implies that only certain proofs of representation theorems have a semantic contribution – those involving definitions of the sort described – and that, in principle, a representation theorem established by a proof that is not of this sort has no semantic interest until an appropriate alternative proof is required.

As a final aside, note that the space of possible defenses of representation theorems, whilst drawn out under the RCL semantics of theoretical terms, can be mapped into the causal-descriptivist account discussed in Section 4. Recall that the central reference-giving characteristic under this

[43] If one retains explicitly the dependence on the axioms – for example, in the case of the definition of utility, taking the statement 'if the **vNM** axioms are satisfied, then any lottery P has a utility of α iff U(P) = α iff $P \sim [c^*,\alpha ; c_*,(1-\alpha)]$' – one obtains what Carnap refers to as a *bilateral reduction sentence*.

[44] The term 'reductionist' refers to here to the bilateral reduction sentences: a reductionist empiricist is an empiricist who requires bilateral reduction sentences for theoretical terms.

theory is the 'kind-constitutive properties that are a core part of a causal description'. Different positions as to what counts as the kind-constitutive properties of the decision-theoretic concepts open up different possible statuses for representation theorems. If none of the definitions discussed in the previous sections could be considered among the kind-constitutive properties of the decision-theoretic concepts, then obviously representation theorems would have no semantic contribution. If, however, the evaluation criterion (**EC**) – or rather the property associated with it, which is given by (**Def-Lew$^{\mathbf{EC}}$**) – were to be considered as a kind-constitutive property, then it would contribute to 'giving the meaning' of these terms. As we have seen above with respect to the RCL semantics, the question then boils down to *what, if anything, is added* by the set of axioms, which the representation theorem states to be equivalent to (**Def-Lew$^{\mathbf{EC}}$**). The arguments proposed above could in principle be made, though their force will depend on the specificities of the causal-descriptivist account. On the one hand, the criticism of the purported triviality of (**Def-Lew$^{\mathbf{EC}}$**) on eliminativist grounds sits uncomfortably in the context of a causal-descriptivist approach, which, as emphasized by Psillos, is strongly anchored in a realist philosophy of science. On the other hand, the fact the proof of representation theorems may involve different definitions of decision-theoretic concepts may be relevant in a causal-desciptivist perspective. Although, the difference in holism between (**Def-Lew-*u***) and (**Def-pref$^{\mathbf{Risk}}$-*u***), say, may be irrelevant to a causal-descriptivist, the latter definition may nevertheless be considered as semantically important, if it identifies a kind-constitutive property by virtue of which the referent of the term 'utility' plays a particular causal role. In this case, representation theorems – or more precisely their proofs – could be defended on the grounds that they reveal kind-constitutive properties crucial to reference-fixing that go beyond those involved in the evaluation criterion. Naturally, this defense rests on a position concerning the kind-constitutive properties of decision-theoretic concepts, and not on a rejection of holism.

8. Conclusions

What should one conclude from the preceding analyses? First of all, it is clear that the plausibility of the claims of semantic relevance for representation theorems are strongly dependent on one's view on the semantics of theoretical terms, and on issues such as holism and eliminativism. Unsurprisingly, causal-historical accounts, or the causal-descriptivist accounts of the semantics of theoretical terms inspired by them, don't seem to be appropriate for the understanding of the purported semantic relevance of representation theorems. More surprising, perhaps, is how difficult it is, even on a descriptivist account of theoretical terms that is conceptually more coherent with the framework implicitly evoked by decision theorists, to identify a role for representation theorems. Invoking an account of the semantics of theoretical terms – the RCL approach – that is *prima facie* more amenable to the sorts of claims made by decision theorists does not in and of itself vindicate these claims. Other philosophical positions are required, and even then, the contribution comes in unexpected forms.

Basically, under the RCL account of theoretical terms and given the assumption that concept of preference is an acceptable basic concept,[45] someone who adheres to a form of reductionist empiricism or to a particularly robust form of semantic anti–holism will consider 'constructive'

[45] For the empiricist, that would mean that the preferences are observable or can be reduced to observational concepts; for the anti–holist, that would mean that the meaning of the preference concept is determined according to anti–holistic criteria.

representation theorems to be necessary to give the meaning of the decision-theoretic concepts involved in an evaluation criterion. However, it is not the result of the theorem, but rather its proof, and more specifically the explicit definitions that feature in it, that constitutes the central semantic contribution. These positions can account for the importance traditionally attributed to representation theorems, albeit via a somewhat unexpected route.

However, it leaves many alleys open to someone who wishes to deny their importance. Even if you subscribe to the RCL account of the meaning of theoretical terms, denying these positions suffices to shed doubt on the semantic importance of representation theorems. For example, if, like Weirich (2001), one considers that 'we do not have to be stricter in decision theory concepts than in physics' while taking into consideration, as one often does, that theories in physics do not satisfy the strictest criteria of either the reductionist empiricists or the anti–holists, then representation theorems are not as indispensable as they are sometimes claimed to be. This point has an interesting corollary for the development of economics and the decision sciences: under such a liberal view of the meaning of theoretical terms, there is no *semantical* flaw in attaching less importance to representation theorems, as seems to be common today in behavioral economics. That said, even if representation theorems are not considered indispensable to give the meaning or reference of decision-theoretic concepts, this does not imply that they are without value, including from the semantic perspective: we can certainly accept that they may allow us to better understand an evaluation criterion and the concepts involved in it. Perhaps it is a sense of 'meaning' that is loose enough to cover these aspects that some decision theorists have in mind when they claim that representation theorems give the meaning of decision-theoretic concepts. Whether or not this is the case, the semantic importance of representation theorems is far less simple and obvious than it would appear: in its strongest form, according to which these results are essential for attributing a meaning to decision-theoretic concepts, it is dependent on very specific, and controversial philosophical notions.

9. References

Anscombe, F.J. and Aumann, R.J. (1963) "A Definition of Subjective Probability", *The Annals of Mathematical Statistics*, 34(1), pp. 199-205

Bradley, R. (2004) "Ramsey's Representation Theorem", *Dialectica*, 58(4), pp. 483-497.

Bridgman, P.W. (1927) *The Logic of Modern Physics*, New York: MacMillan

Carnap, R. (1936/37) "Testability and Meaning", *Philosophy of Science*, 3, 1936 and 4, 1937

Carnap, R. (1956) "The Methodological Character of Theoretical Concepts", in H. Feigl & M. Scriven (eds) *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Minnesota Studies in the Philosophy of Science, 1, Minneapolis: University of Minnesota Press, pp.38-76

Carnap, R. (1959) "Theoretical Concepts in Science", *Studies in History and Philosophy of Science*, 31(3), pp. 158-72

Carnap, R. (1966) *An Introduction to the Philosophy of Science*, New York: Basic Books

Dekel, E. & Lipman, B. (2010), "How (Not) to Do Decision Theory", *Annual Review of Economics*, 2, pp. 257-282

Enç, B. (1976) "Reference of Theoretical Terms", *Noûs*, , 10(3), pp. 261-282

Gilboa, I. (2009) *Theory of Decision under Uncertainty*, Cambridge: CUP

Gilboa, I. & Schmeidler, D. (1989), "Maxmin Expected Utility with a Non-Unique Prior', *Journal of Mathematical Economics*, 18, pp. 141-153

Gul, F. & Pesendorfer, W. (2008), "The Case for Mindless Economics" in Caplin, A. & Schotter, A. (eds) *The Foundations of Positive and Normative Economics*, pp. 3-41, Oxford: Oxford University Press

Hempel, C. (1950) "Problems and Changes in the Empiricist Criterion of Meaning", *Revue Internationale de Philosophie*, 41, pp. 41-63

Hempel, C. (1958) "The Theoretician's Dilemma: A Study in the Logic of Theory Construction", in *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: University of Minnesota Press

Jeffrey, R.C. (1965/1983) *The Logic of Decision*, 2nd. ed., Chicago: University of Chicago Press

Joyce, J.M. (1999) *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press

Karni, E. (1993) "A Definition of Subjective Probabilities with State-Dependent Preferences", *Econometrica*, 61(1), pp. 187-198

Krantz, D.H., Luce, D.R., Tversky, A. and Suppes, P. (1971), *Foundations of Measurement*, vol. 1, Dover

Kreps, D. (1988) *Notes on the Theory of Choice*, Boulder: Westview Press

Kripke, S. (1972/1980) *Naming and Necessity*, Cambridge, Mass.: Harvard University Press

Lewis, D. (1970) "How to Define Theoretical Terms", *Journal of Philosophy*, 67, 427-46. Republished in *Philosophical Papers* vol. I (OUP, 1983), 78-95

Lewis, D. (1972) "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy*, 50, 249-58. Republished in *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 248-61

Machina, M.J. and Schmeidler, D. (1992) "A More Robust Definition of Subjective Probability", *Econometrica*, 60(4), pp. 745-780

Maskin, E. (1979) "Decision-Making under Ignorance with Implications for Social Choice", *Theory and Decision*, 11, pp. 319-337

Milnor, J. (1954) "Games against Nature", in Thrall, Coombs & Davis (eds.) *Decision Processes*, New York: Wiley

Mongin, Ph. (2003) "L'axiomatisation et les théories économiques", *Revue économique*, 54, 2003, p. 99-138. von Neumann, J. & Morgenstern, O. (1944/47) *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press

Pagin, P. "Meaning holism", in Lepore, E. and Smith, B. (eds) *Handbook of Philosophy of Language*, Oxford UP, pp. 213-232

Papineau, D. (1996) "Theory-Dependent Terms", *Philosophy of Science*, 63(1), pp. 1-20

Peacocke, C. (1997) "Holism", in Hale, B. and Wright, C. (eds) *A Companion to the Philosophy of Language*, London: Blackwell, pp. 227-247

Psillos, S. (1999) *Scientific Realism. How Science Tracks Truth*, London: Routledge

Psillos, S. (2000) "Rudolf Carnap's 'Theoretical Concepts in Science' "*Studies in the History and Philosophy of Science*, 31(3), pp. 151-8

Psillos, S. (2008) "Carnap on Incommensurability", *Philosophical Inquiry*, 30(1), pp. 135-156 Putnam, H. (1975) "The Meaning of 'Meaning'", in *Mind, Language and Reality*, *Philosophical Papers*, vol. 2, Cambridge: Cambridge University Press

Ramsey, F.P. (1926) “Truth and Probability”, in *The Foundations of Mathematics and other Logical Essays*, Ch. VII, p.156-198, London: Kegan, 1931

Ramsey, F.P. (1929) “Theories” in Mellor, D.H. (ed.) *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, London: RKP, 1978, pp. 101-125

Savage, L. (1954/72) *The Foundations of Statistics*, New York: Dover

Weirich, P. (2001) *Decision Space. Multidimensional Utility Analysis*, Cambridge: Cambridge University Press

Wakker, P. (2010) *Prospect Theory*, Cambridge: Cambridge University Press