From Radical Translation to Radical Interpretation and Back

António Zilhão
University of Lisbon

Abstract

Both Quine and Davidson put forth programs of empirical semantics satisfying the conditions that characterize the so-called "standpoint of interpretation." Quine's less ambitious program of radical translation rests upon two buttresses: causality and empathy. Davidson's more ambitious program of radical interpretation replaces causality with truth and empathy with rationality. Although the replacement of causality with intersubjective truth seems to me to be a fully justified move, I nevertheless contend that it is more realistic to develop the work of interpretation drawing upon Quine's less ambitious requirement of empathy than upon Davidson's view of human agency as rational agency. In order to substantiate this contention, I present an argument to the effect that Davidson's characterization of human agency as rational is not compatible with his other requirement that truth should provide the essential link connecting speech with environment and action.

1. The Standpoint of Interpretation

The standpoint of interpretation admits being characterized in terms of the thesis that empirical semantics has two possibility conditions. First, there has to be an essential link connecting speech with environment and action; second, there must be a behavioural core common to interpreter and interpretee. As far as I know, Wittgenstein was the first philosopher to have explicitly formulated this standpoint. The following statement of his could easily be picked up as one of its first slogans: "The common behaviour of mankind is the system of reference by means of which we interpret an unknown language" (*PI*, §206).

Both Quine and Davidson put forth programs of empirical semantics satisfying these conditions, namely, the program of radical translation and

the program of radical interpretation.

According to Quine's program of radical translation, the essential link connecting speech with environment and action is provided by causality. From his standpoint, an agent's speech has a sole "entering wedge": observation sentences. Each such sentence gets expressed, assented to or dissented from as a direct causal response to patterns of stimulation affecting the nervous extremities of the agent. These patterns constitute the stimulus-meanings of those sentences. This is how language latches on to the world.

Human speech is supposed to be a means of communication. However, patterns of stimulation of nervous extremities are not shared among different agents, even in cases in which these belong to the same linguistic community. Which is to say that stimulus-meanings are private. Thus, according to Quine's view, our possibility of achieving any understanding of the observation sentences of others and of translating them rests entirely upon our ability for projecting ourselves onto the other's situation. Quine calls such a projective ability "empathy." Quine's empathy is supposed to be perceptual, when observation sentences are the target of the interpretation, as well as inferential or grammatical, when, in order to be able to make sense of the totality of an agent's utterances, we put forth what Quine terms as 'analytical hypotheses'. Of course, a necessary condition for the success of such a practice is that interpreter and interpretee must share a common behavioural core.

Davidson's program of radical interpretation is more ambitious. According to him, the essential link connecting environment, speech and action is provided by the concept of truth. The idea that observation sentences are semantically privileged because they have stimulus-meanings and that stimulus-meanings are to be cashed out in terms of patterns of stimulation of nervous extremities has no appeal to Davidson. He rightly spots in it a survival of old-fashioned Cartesian privacy, albeit within a physicalistic framework. The causal connection Quine assumes there to exist between the world, patterns of stimulation of the speaker's nervous extremities and the content of his observation sentences still falls short of providing him with an access to a genuinely public world. According to Davidson, Quine's requirement of empathy is but a physicalistic variant of older arguments by analogy. In this sense, Davidson con-

tends that Quine's view is as vulnerable to traditional skeptic arguments regarding the possibility of our being massively mistaken about the outside world as traditional empiricist views were.

Not denying the crucial role our senses play in causally connecting us with the world and thus making our knowledge of it possible, Davidson sees the content of our sentences, observational or otherwise, to be inter-subjectively rather than sensorially determined. Content is, according to him, a matter of mutual calibration originated by the similarity of our responses to the objects around us and to each other. Thus, we create the world we talk about not by responding privately to stimuli, but by inter-acting with each other and, in so doing, observing the way others respond to the world and to ourselves. The truth conditions of any of our utter-ances do belong to this public world from the outset. That is, when we talk at all, we talk about such things as chairs, tables, rabbits or rain and not about patterns of stimulation of our nervous extremities, presumably caused by such objects. Intersubjective truth, rather than causal impact on our sensitive surfaces, is then the essential link connecting speech with environment, action and understanding.

Now, if truth is supposed to play such a crucial role in the work of interpretation, then three further conditions must also be met. First, a lot of our sentences have to be true; second, there has to be such a thing as a language user's attitude of taking a sentence to be true; third, the external manifestation of this attitude has to be intuitively recognizable by all other language users as being the manifestation of precisely this attitude, even if the content of the sentence taken to be true by the speaker is unknown to the interpreter.

Let us grant, for the sake of the argument, that these three conditions are indeed met in human linguistic intercourse. However, when a speaker speaks truly he utters not just sentences; he expresses beliefs. Now, if the meaning of the sentences through which he expresses his beliefs is to be determined by appeal to an intersubjective, public, world, then it seems that we need to know beforehand what true beliefs the speaker is sup-posed to have in each particular situation in order to be able to infer from their content the meaning of the true sentences through which they get expressed. But, in order to be able to do this, we need, first, to assume that true beliefs issue in true sentences and that true sentences express true

beliefs. Secondly, we need to know how to infer a speaker's true beliefs (and his desires) from his actions. And in order to be able to do the latter thing, we need to know something substantial about what "the common behaviour of mankind" is supposed to be, i.e., we need to know what is the particular formal structure that enables us to make sense of the purposeful behaviour of the interpretees. In other words, being centered on social interaction rather than on the epistemic subject, Davidson's approach needs to say something more specific about both the psychological background and the behavioural structure a group of agents is supposed to display in order to be classifiable as a group of language users.

Thus, rather than simply stating that the work of interpretation has to proceed having "the common behaviour of mankind as its system of reference," as Wittgenstein put it, Davidson must find a way of characterizing this behaviour in a theoretically illuminating way. That is, a way that will not get reduced either to a collection of examples of behaviours with which we feel we can empathize or to some paraphrase or other of the indexical characterization 'it is the behaviour of these creatures' together with some sort of ostension having the set of human beings as its target.

It is precisely in this respect that his work seems to have a *prima facie* great theoretical advantage over Quine's or Wittgenstein's: it aims at providing us with such a specification. It rests on two moves. Taking seriously the old Aristotelian definition of Man as the rational animal, on the one hand; and presenting an independent characterization of rational agency, on the other hand. These permit the subsequent identification of the common behaviour of mankind with the behaviour of rational agents. Therefore, what is taken to be the appropriate characterization of the latter is assumed to provide the system of reference in terms of which the work of interpretation is to proceed.

As a matter of fact, the conclusion we reach when we compare Quine's and Davidson's views on translation and interpretation is that, despite the fact that they were so close personally, they ended up putting forth two strikingly dissimilar views of what it means to be a language user. Quine's view carries out a physicalistic reconstruction of the traditional epistemological empiricist approach to the subject; Davidson's is best viewed as much more akin to David Lewis's game-theoretical approach to the nature of language than to Quine's. Be this as it may,

Davidson's program of radical interpretation soon became dominant within the field of interpretationism.

Now, it seems to me that the path Davidson follows in his attempt to specify the formal structure of human behaviour is mistaken. Therefore, although I acknowledge the theoretical appeal of his strategy, I will contend that, after having followed Davidson in leaving behind the last remnants of traditional empiricism still to be found in Quine, and thus leaving aside all talk about stimulus-meanings, one should nevertheless return to what might be viewed as a modified version of Quine's requirement of empathy, at least for the time being. In order to substantiate this contention, I will present an argument to the effect that Davidson's characterization of human agency is not compatible with his other requirement that truth should provide the essential link connecting environment, speech and action.

2. Human Behaviour and Rational Gambling

Davidson's view of rational action is basically the one to be found in Bayesian decision theory. It can be summarized thus: an action is rational if and only if it is such that it produces an outcome which is the best according to the agent's desires and the beliefs he entertains; an agent is rational if and only if his actions are, in general, rational. To be sure, such a definition does not yet yield a criterion to decide between competing courses of action which of them is the best; it is only a definition stating what it means to say of a given action that it is rational. In particular, it is a definition equating rationality with optimality. The criterion specifying how is one to assess which of a given set of possible actions is, under each circumstance, the best or one of the best is the so-called principle of "maximization of expected utility."

This principle can in turn be accounted for thus. Desires are meant to be pro-attitudes towards representations of outcomes of actions. The different outcomes taken to be possible under each particular set of circumstances are, in turn, supposed to be ranked in a scale determined by the assignment of real-numerical values to them, according to their respective degree of desirability or utility. Beliefs are meant to be truth-

evaluable attitudes towards representations of particular states of the world assumed to be relevant to the determination of the outcome of the agent's action. The likelihood the agent ascribes to the obtaining of any of these particular states of the world varies along a probability scale. To act rationally in a particular situation is then to act in a way that admits being modelled as resulting from the following succession of steps. First, the subjective probability ascribed to the obtaining of each relevant state of the world is multiplied by the subjective utility ascribed to the obtaining of each of the possible outcomes of the action. Secondly, the products so obtained are added; the sum thus obtained is then called the subjective expected utility of undertaking a given course of action. Thirdly, once steps one and two were iterated in connection with all possible courses of action, a decision to act in a certain way is reached by selecting that course of action the expected utility of which is maximal.

Human actions are part of our experience. They are events taking place in space and time before us. If a theory of rational action such as the one described above is to be taken to be a theory of human action, then its truth ought to be empirically verifiable, at least in principle. *Prima facie*, the decision-theoretical framework should admit being put to the trial of experience in the usual way. That is, the principle of maximization of expected utility should be taken to be a law relating sets of upstream con-ditions with sets of downstream consequences. The theory would then be in agreement with reality if and only if upstream conditions and down-stream consequences would relate to each other in the way prescribed by the lawlike principle. The verification of this requirement should then involve the following stages. First, the correct identification of the beliefs and desires of the agent, and of their respective strengths, on the one hand, and the correct identification of the action related to them in any particular occasion. Second, the confirmation that agents *do* act in agreement with the principle of maximization of expected utility.

However, and as a number of authors have noted, one is faced with a major problem here. It is the following. The procedures set up within the theoretical framework of Bayesian decision theory in order to allow us to arrive at a positive identification of the agent's beliefs and desires are pro-cedures that make sense only if one assumes from the outset that agents act according to the principle of maximization of expected utility. The

situation one is faced with is thus the following. In order to test the validity of a relational principle one needs, *prima facie,* to have an independent access to the *relata* such a principle relates. However, when one considers this theory carefully, one realizes that, even assuming the content of human actions to be transparent, the assumption that the relational principle holds is a necessary condition to identify the upstream *relata*, namely, beliefs and desires. In other words, there seems to exist no access to a substantial part of the relevant empirical evidence outside of the framework of the theory itself. This means that no empirical check can be made before such a framework is itself subject to a closer examination.

What such a closer examination reveals is that Bayesian decision theory is based on an analysis of what one might call 'the behaviour of a rational gambler'. That is, it is based on a couple of axioms that formalize the behaviour a gambler needs to have in a game of chance if he is to avoid losing money whatever happens. The thesis according to which this theory is empirically true boils down then to an analogical claim. Namely, the claim that human agents behave in their normal life in the same way as a rational gambler behaves in a game of chance. Now, does such a claim hold true of the facts? This is what needs to be checked.

Obviously, the fact that in a game of chance one needs to behave according to the axioms of the theory in order to avoid losing money whatever happens is by itself no proof that the theory is empirically true of human behaviour in general. It may well be the case, for all we know, that most humans are not rational gamblers anyway and that therefore in a game of chance they would easily fall in situations in which they would lose to a professional gambler no matter what happens. Or it may well be the case that despite the fact that most humans behave as rational gamblers when placed in the context of a game of chance, they do not behave in real life as if it were a game of chance, i.e., the analogical claim may be false. Therefore, if the theory is to be empirically true of human actions in general then its axioms will have to hold in most everyday life situations. This is the crucial point. As a matter of fact, if the axioms hold true of the way everyday decision problems are solved by most people, then the validity of the principle of maximization of expected utility follows by purely mathematical means. That is, such a principle is a *theorem* of the theory and not one of its axioms.

Different axiom systems and slightly different versions for what are basically the same axioms have been put forth in the relevant literature. However, two of them are, in any of the available versions, crucial for the outlining of the structure of a rational gambler's behaviour. They are the following. First, the axiom I will call "Axiom A"; it holds that a certain relation, namely, the relation "it is at least as preferred as" (or, in another version, "it is at most as preferred as") obtains between the coordinates of any pair of outcomes open to the agent's consideration and that such a relation is a weak linear ordering (or connected preorder), i.e., that it is a binary relation satisfying the following properties: transitivity, reflexivity and also connectedness. Second, the axiom I will call "Axiom B"; it states that if an option A is at least as preferred as an option B then, if options C and D result from, respectively, options A and B by the introduction of the same change in their common outcomes, then option C is at least as preferred as option D.

Axiom A can be justified thus. It is this axiom that allows utilities to be put in a one-one correlation with real numbers, that is, that allows them to be measured in such a way that both their places in the utility ranking and their intrinsic differences of value are adequately expressed by the assignment of real-numerical representatives to them. Actually, transitivity is the essential property for this purpose. As a matter of fact, the admission that failures of transitivity might exist would entail two consequences. First, the consequence that actual utility rankings are not representable by means of ordinal utility scales and that therefore the whole process of selecting that course of action the expected utility of which is highest becomes pointless. Second, the consequence that the holding of an intransitive gambling pattern is liable to make a gambler lose money for no gain whatsoever to a smart gambler, that is, the consequence that such a gambler is not rational in the decision-theoretical sense of this word.

Axiom B can be justified thus. It follows, together with other uncontroversial assumptions, from what is taken to be a very simple and intuitively appealing decision strategy. Such a strategy is the following: when trying to figure out what is the act which is more advantageous to him, the agent should consider only those elements of a set of mutually exclusive and jointly exhaustive possible futures states of the world which have

different outcomes (that is, states with the same outcomes should cancel out).

Both these axioms have already been put to test. The empirical adequacy of axiom A has been checked by experiments devised by Tversky. The empirical adequacy of axiom B has been checked and subject to extended discussions in connection with the so-called "Allais' Problem."

Tversky's experiments elaborate upon Condorcet's voting paradox. As is well known, Condorcet pointed out that social inconsistencies might arise out of consistent individual choices. In particular, he has shown that majority voting generates an intransitive pattern of preferences whenever a set of three individuals x, y and z consider a set of three options A, B and C in such a way that x ranks the options according to the scale ABC, y ranks them according to the scale BCA and z ranks them according to the scale CAB and the options are subject to a sequence of three alternative votings. Examples of this kind were later taken to provide a useful illustration to the unsolved difficulties associated with the idea of providing a social welfare function. The peculiarity of Tversky's own work consists in the fact that he has shown by means of a variety of psychological experiments that a similar pattern of intransitive preferences can arise out of the choices of a single agent who ranks his options on two dimensions in such a way that the agent's payoff matrix has the structure of a lexicographic semi-order. He has moreover shown that in a number of situations it is not uncommon for agents so to rank their options.

Allais' problem was initially introduced as a thought-experiment; however, it received later empirical confirmation. It consists in the successive consideration of two decision situations each involving two gambles. The result of the experiment is such that the set of two successive choices made by most agents violates Axiom B. These choices have the following character. In the first choice, agents typically choose a gamble in which they win a large amount of money outright in place of a gamble in which the winning of that very same amount of money is uncertain and has to be weighed against a small probability of winning 5 times as much and an even smaller probability of winning nothing. In the second choice, agents are confronted with a set of two gambles in which the dominant outcome is winning nothing. In this case, they typically choose the gamble

in which the highest prize is at stake, despite the fact that the probability of winning it is slightly inferior to the probability of winning the also large but yet 5 times smaller prize they chose in the first choice.

These and other similar experiments should have settled the question concerning the empirical validity of the theory. As a matter of fact, if the descriptive character of the theory hinges upon the descriptive character of its axioms, then, if such axioms are shown not to be empirically true of the behaviour of normal human agents in a non negligible number of decision situations, then it should follow that the theory as a whole cannot be ascribed the character of a descriptive theory of human choice behaviour. In particular, it should follow that its main theorem, namely, the theorem according to which agents maximize expected utility, cannot be ascribed the character of a lawlike principle relating the upstream conditions and the downstream consequences associated with the unfolding of human actions. And if this is so, it should also obviously follow that such a theorem cannot be simply assumed to hold in order to infer the content of an agent's beliefs and desires from such an assumption together with the knowledge of the content of some downstream effects and a bit of mathematical reasoning. Assuming that one has some independent way of identifying one's beliefs and desires, the principle codified in this theorem may or may not be a powerful normative standard for decision-making, prescribing the way one should go about making one's decisions in order to optimise one's gains; as an empirical proposition, however, it seems that it must be deemed to be false.

3. The Appeal to the Synthetic a priori

The conclusion above is far from being universally accepted though. The verdict of empirical falsity has been evaded by quite a number of authors, a leading representative of which is precisely Davidson. The main argument underlying such a response is basically the following.

In order to test the above-mentioned axioms, experiments had to be made. Now, what these experiments yielded were behavioural *data*, and these had to be interpreted. In particular, the actual choice behaviours of the agents had to be classified under particular characterizations. Other-

wise, it would not have been possible to assert that, e.g., failures of transitivity had taken place. However, the above-mentioned assumption according to which the content of our actions could be considered to be basically transparent is untrue. That is, the identification of the downstream *relata* depends as much as the identification of the upstream *relata* on the use of a theoretical framework within which particular interpretations may be considered to make sense. And there is no compelling reason why the interpretational constraints that should be at work here should not be precisely those constraints that are set up by the very same axioms the experiments in question were trying to check. Under such circumstances, instead of providing evidence in favour of the inappropriateness of the tested axioms, the results of the experiments should be reinterpreted in such a way that the content of the actions displayed in them might come out agreeing with the axioms after all. And this is something that is not particularly difficult to do. Consider for instance the following reinterpretation proposals.

Take the case of Tversky's experiments first. Someone might challenge his conclusions by making the following claims. In order to uncover the alleged existence of intransitive sets of preferences the experimenter must confront different sets of choices of the same subject with each other. Now, by necessity, the subjects go through these choice processes in time. Therefore, each choice must be separated from any other choice by a certain amount of time. Under such circumstances, why should the experimenter not simply interpret the *data* as exhibiting the plain fact that under the sort of circumstances present in Tversky's experiments subjects tend to change their minds?

Take now the case of Allais' Problem. The correctness of the verdict that Axiom B has been violated in the experiment devised by Allais depends entirely upon an unformulated assumption. This is the assumption that the consequences under consideration are to be fully characterized by their monetary values. But why should they? Why not include non-monetary considerations, such as, e.g., considerations of risk-aversion, in the way the consequences should be perceived? The inclusion of considerations of such a sort may make it possible to reinterpret the *data* in such a way that Axiom B comes out of the experiments unscathed.

In sum, the idea is that the axioms of the theory should be taken to

have the *status* of synthetic *a priori* truths about our behavioural patterns. Such a strategy can certainly not be ruled out from the set of *prima facie* admissible interpretive strategies. And given that there is no external court of appeal capable of judge once and for all the correctness or incorrectness of any particular interpretation in any particular situation, one must settle for fruitfulness in the long run. In the meantime, one should stick to the sort of reasoning by analogy first introduced by Ramsey, the outcome of which is a theory possessing two important advantages: it is conceptually simple and it lends itself to mathematical formalization.

I think that the line of reasoning described above runs the risk of leaving the doors wide open to the generation of all sorts of epicycles. As a matter of fact, I do think that, at least in the case of Tversky's experiments, the interpretation according to which subjects developed intransitive sets of preferences is not only much more intuitively plausible and charitable than the rival strategy of ascribing a flickering character to quite a number of human subjects but it also makes more cognitive sense. However, I am not going to pursue this line of argument here, as I have already done it elsewhere. What I want to stress here is the following aspect. I think that the line of escape that appeals to the synthetic *a priori,* implausible as it may be, is nonetheless a possible line of escape for decision-theorists such as Ramsey, Savage or Jeffrey. I also think that such a line of escape is not available to the theorist who made an explicit appeal to it, namely, Davidson. Below I will explain why.

4. Truth and Understanding

Both Ramsey and Savage used gambles as the means to find out what agents believed and desired and how seriously did they do it. It was a consequence of their proposal that the utility scales and the probability distributions ascribed to agents might be strongly at variance with the content of the sentences of the very same agents concerning the objects of their desires and the contents of their beliefs. Indeed, the whole idea of the enterprise was precisely to create the theoretical framework within which a scientific study of human actions would be made possible. As in other fields of study, it was only to be expected that the result of the theoretic

activity would frequently be in disagreement with people's impressionistic and intuitive assessments of the phenomena under consideration. Obviously, it is in natural discourse that such an impressionistic and intuitive assessment of psychological phenomena gets expressed. Therefore, no wide agreement was to be expected between either first-person or third-person natural discourse descriptions of intentional behaviour and the reports of the outcome of theoretically conducted behaviour experiments. As Ramsey put it, his "artificial system of psychology" was to be compared with Newtonian mechanics and not with what one might term 'Folk-Physics'. Thus, Ramsey's identifications of the belief-desire structures of agents were meant to be theoretical identifications of the causes of actions, of the same sort of, e.g., the identification of water with H2O, and not conceptual analyses of the way agents usually express their views about the unfolding of actions. Natural discourse played a role in the application of this "artificial system of psychology" only insofar as subjects had to understand the gambles they were invited to consider.

The objects to which Ramsey and Savage ascribed degrees of belief were possible states of the world and the objects to which they ascribed degrees of desire were consequences of acts. The connection between them was established through the above-mentioned gambles. Jeffrey, however, replaced propositions for both possible states of the world and consequences of acts as unified objects of belief and desire. He understood propositions as abstract objects referred to by sentences but in no way exchangeable with them. And, in particular, he considered that the sentences professed by subjects concerning the objects of their own beliefs and desires were only in an evidential relation rather than in a relation of reference with the propositions that indeed constituted the objects of the subjects' beliefs and desires. He clearly endorsed the thesis that one may be in doubt or in error about what one's beliefs and desires are and that the touchstone is actual choice. In particular, he stated that agents might be unwilling to assent to the very same sentences that are used by the theorist to correctly express the proposition that is the content of their belief or of their desire. Thus, he also thought that the outcomes of the scientific study of action might be at variance with what natural discourse has to say on such a topic. The criteria for belief and desire are, according to him, behavioural, and speech is but a particular sort of behaviour

among others.

All this is in perfect harmony with the idea according to which the axioms of Bayesian decision theory are synthetic a priori truths about the structure of human intentional behaviour. One may suspect that there is no more truth in such a contention than there is truth in the contention that the axioms of Euclidian geometry are synthetic a priori truths regulating our geometrical knowledge but that is all. Davidson's Theory of Interpretation is, however, a different matter.

This theory contains a slightly modified version of Jeffrey's version of Bayesian decision theory, namely, a version in which Davidson substitutes uninterpreted sentences for propositions as the objects of preferences. Such a substitution is needed, because Davidson uses decision theory not as the framework of a scientific theory of action but rather as the scaffold by means of the support of which the work of radical interpretation is supposed to succeed.

Now, we have seen above that, according to what seems to be our intuitive way of classifying it, the behaviour of the members of our own community is frequently at variance with the expectations decision-theory brings about. It takes an extra interpretive effort, supported by a theoretical argument carried out *in* the languages we already speak and understand, to accommodate that behaviour, under those circumstances, within the framework laid down by the theory. It may well be that the outcome of such work of accommodation is the right scientific or para-scientific account of such behaviours. But if this is true, then what normal people tend to say about such behaviours, even assuming their sincerity, is just wrong. If this is so, however, either both ordinary people's beliefs about the sources of their intentional behaviour and the sentences that express them are false or these beliefs are true but they do not issue in true sentences. Either way, the first of the above mentioned possibility conditions for Davidsonian empirical semantics, namely, the requirement that truth should be the essential link connecting speech with environment, action and understanding is not satisfied. And, if this condition is not satisfied, then, if we assume the thesis of Davidson's version of inter-pretationism to be correct, no empirical semantics of these intentional sentences is possible.

This is an interesting result. In *Word and Object,* Quine claimed

that the right way to interpret sentences of belief-ascription would be to view them as containing only one term referring the holder of the attitude and a syntactically complex but semantically simple monadic predicate, containing both the verb of the attitude and the sentence describing its content. In a paper called "Theories of Meaning and Learnable Languages," Davidson rightly claims that Quine's view entails that the segments of our natural languages containing idioms of propositional attitude must be unlearnable. From this claim, it follows that Quine's interpretation must be wrong, since no segment of any effectively spoken natural language can be unlearnable. Ironically, Davidson's charge against Quine seems to bounce back now. As a matter of fact, if both a decision-theoretical framework and truth are necessary conditions for the possibility of setting up an empirical semantics, then, if what was said above is right, lots of folk-psychological sentences must lie outside the scope of radical interpretation. According to Davidson, however, these sentences constitute the very foundation of radical interpretation. Therefore, the latter enterprise simply cannot be done.

Let me elaborate a bit more upon this. Imagine that some choice behaviours are being performed by members of foreign communities the language of which we completely ignore. If we are going to use the axioms of decision-theory as the scaffold in terms of the support of which we are going to interpret not only their actions but also what they say about them, then, if we assume, with Davidson, that the conceptual system underlying the semantics of their unknown natural language is going to be similar to ours, then their way of describing their common behaviour is going to be as wrong as ours. If this is so, having recourse to the apparatus of classical Bayesian decision theory cannot be the right way of characterizing the behavioural structure that does provide the system of reference by means of which their unknown natural language is going to be interpretable. Note that the only way of avoiding this conclusion is to postulate that the speakers of this unknown foreign language are inborn scientific psychologists. And remember that, assuming that some version or other of Ramsey's "artificial system of Psychology" is true, we *know* that we are no such thing. Thus, the introduction of such a postulate leads us outright to contradiction.

Let me stress once more that it is not my intention to present here a challenge to the intrinsic value of the decision-theoretical framework. To see that this is so, we need only remember that, within Ramsey's "artificial system of psychology," beliefs are defined as the (partial) causes of action. No assumption of transparency is made. *Qua* partial causes of his action, an agent's beliefs may well be true, and yet, still assuming sincerity, the sentences in terms of which he, or somebody else, formulates them may well be false. That is, it may well be the case that the agent (or his untrained observer) does not know or is not aware of what beliefs he in fact has. Again, that something like this should be expected to be the case was precisely the scientific motivation underlying Ramsey's development of his "artificial system of Psychology."

However, the option of accepting the decision-theoretical definition of the rational agent as basically true of us and of considering that scientific psychology should be built upon it comes with a price. This is the severing of the supposedly essential link of veridicality connecting our common ways of speaking with our common ways of acting. Severing that link has two clear consequences. The first is that ordinary ways of talking will have to be dismissed as irrelevant for the understanding of our common ways of acting; the second is that the knowledge of our common ways of acting will have to be dismissed as irrelevant for the understanding of our common ways of talking. Or, at least, they will have to be so dismissed until somebody comes along with a theory explaining in detail the convoluted ways in which our untrue ordinary way of talking about our actions is related to the true explanatory theory of action and viceversa. Such a theory does not seem to be in the offing, however.

In short, I would like to say that one cannot have it both ways: to claim, in the hope of procuring a foundation for empirical semantics, that an essential connection of veridicality should hold between speech, environment and action, on the one hand; and to severe that connection in order to make the interpretational theory conceptually simple and mathematically tractable, on the other hand. However, it seems to me that this is precisely what is happening in Davidson's Unified Theory of Interpretation.

5. Rationality, the Ways of Man and Empathy

A supporter of interpretationism will have to criticize Davidson's approach in one of the two following ways. Either he challenges the cogency of Davidson's independent characterization of rational agency, or he challenges the view according to which Man is the rational animal. Given the vagueness of the concept of rationality itself, the choice between these two criticisms gets eventually reduced to a mere choice of words.

As a matter of fact, if he enjoys the use of the term 'is rational' in association with human matters then he should want to stick to the old Aristotelian definition of Man. Under those circumstances, the interpretationist should criticize the allegedly independent definition of rational agency Davidson puts forth as wrong. But if he does not have any strong feelings associated with the use of the predicate 'is rational' to characterize the ways of Man, the interpretationist could still easily accept Davidson's definition of rational agency as correct and at the same time deny the rationality of Man. Under such circumstances, he would have to refuse the thesis that a definition of rational agency ought to be used as the system of reference by means of which any unknown language is to be interpreted and he would have to look for a less grand and more parochial way of characterizing the ways of Man. Either way, what is relevant is that Davidson's characterization of rational agency turns out not to be an appropriate way of modelling the manner in which human agency tends to be described in our natural languages. Thus, grounding the work of interpretation upon such a characterization is most certainly going to lead the interpreter into the selection, as interpretive, of non-interpretive T-theories for the natural languages of foreign communities.

Leaving undecidable controversies about the right way to use the predicate 'is rational' aside, what other means of characterizing theoretically the ways of Man do we have then? Recent research in Cognitive Psychology has consistently shown the structure underlying human agency to be of a much greater complexity than previously thought. By itself, this need not be a reason for interpretational despair. At the same time, however, evidence has also been piling up according to which humans share a largely innate set of linguistic structures, concepts and beliefs, and that these play a major role in our use and understanding of

language. This might mean that the interpretation of speech may, at least partly, be achieved independently of the acquisition of an elaborate account of the workings of other human cognitive structures, namely, of the structures underlying the triggering of intentional action. If this is true, then interpretationists do have reasons for worry.

It is at this stage that I suggest that a retreat to what might be viewed as a modified version of Quine's initial requirement of empathy can be of use. Instead of referring the analogical projection onto another of the private stimuli of the subject, the term 'empathy' would in this modified version of Quine's requirement refer the sort of inborn predisposition humans seem to have for coping with their conspecifics. (As a matter of fact, this seems to me to be the way the term is actually used by Quine in *From Stimulus to Science,* by contrast with the way the term was used by him in *Word and Object,* and other earlier essays). Be this as it may, and in spite of the fact that it is not satisfying when matters are considered from an explanatory standpoint, the appeal to such a revised concept of empathy in order to account for the possibility of linguistic understanding enables us to achieve two results. On the one hand, it enables us to stay away from a commitment to implausible accounts of linguistic understanding such as the one that views it as the result of an overt negotiation in which logically sophisticated creatures act along decision-theoretical lines and expect others to do the same. On the other hand, the lack of theoretical specificity characterizing it has the advantage of leaving room for the progressive integration into a future theory of interpretation of the description of the features of a speaker's interpersonal experience that really are salient for the triggering of the interpretational part of his linguistic abilities.

References

Allais, M. 1953. "Le Comportement de l'Homme Rationnel devant le Risque. Critique des Postulats et Axiomes de l'École Américaine." *Econometrica* **21**: 503–46.

Chomsky, N. 1986. *Knowledge of Language – Its Nature, Origins and Use.* Westport, CT.: Praeger.

Condorcet 1986. "Essay sur l'Application de l'Analyse a la Probabilité des Décisions Rendues à la Pluralité des Voix — Discours Préliminaire." *Sur les Élections et Autres Textes 1782–94.* Paris: Fayard.

Davidson, D. 1980. "Mental Events." *Essays on Actions and Events.* Oxford: Clarendon Press.

—. 1980. "Hempel on Explaining Action". *Essays on Actions and Events.* Oxford: Clarendon Press.

—. 1980. "Psychology as Philosophy." *Essays on Actions and Events.* Oxford: Clarendon Press.

—. 1984. "Theories of Meaning and Learnable Languages." *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

—. 1984. "On the Very Idea of a Conceptual Scheme." *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

—. 1984. "Radical Interpretation." *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

—. 1990. "The Structure and Content of Truth." *Journal of Philosophy* **87**: 279–328.

—. 1990. "Meaning, Truth and Evidence." In Barrett, R. and Gibson, R. (eds.), *Perspectives on Quine.* Oxford: Blackwell.

—. 1995. "Could there be a Science of Rationality?" *International Journal of Philosophical Studies* **3**: 1–16.

—. 2001. "What Thought Requires." In Branquinho, J. (ed.), *The Foundations of Cognitive Science.* Oxford: Oxford University Press.

—. 2001. "The Second Person." *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press.

—. 2001. "A Coherence Theory of Truth and Knowledge." *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press.

—. 2001. "Empirical Conten.t. *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press.

Jeffrey, R. 1983. *The Logic of Decision.* Chicago: Chicago University Press.

Kahnemann, D., and Tversky, A. 1982. "The Psychology of Preferences." *Scientific American* **246**: 160–73.

Lewis, D. 1969. *Convention — A Philosophical Study.* Cambridge, Mass.: Harvard University Press.

—. 1975. "Languages and Language." In Gunderson, K (ed.), *Language, Mind and Knowledge.* Minneapolis: University of Minnesota Press.

Machina, M. 1983. "Generalized Expected Utility Analysis and the nature of the observed violations of the Independence Axiom." In Stigum and Wenstøp (eds.), *Foundations of Utility and Risk Theory with Applications.* Dordrecht: Reidel.

Pinker, S. 1994. *The Language Instinct – The New Science of Language and Mind.* London: Penguin.

—. 1995. "Language Acquisition." In Osherson, D. N. (org.), *An Invitation to Cognitive Science – vol. 1.* Cambridge, Mass.: The MIT Press.

Quine, W. v. O. 1960. *Word and Object.* Cambridge, Mass.: The MIT Press.

—. 1975. "Mind and Verbal Dispositions." In Guttenplan, S. (ed.), *Mind and Language.* Oxford: Clarendon Press.

—. 1990. "Three Indeterminacies." In Barrett, R., and Gibson, R. (eds.), *Perspectives on Quine.* Oxford: Blackwell.

—. 1990. *Pursuit of Truth.* Cambridge, Mass.: Harvard University Press.

—. 1995. *From Stimulus to Science.* Cambridge, Mass.: Harvard University Press.

Ramsey, F. P. 1931. "Truth and Probability." *The Foundations of Probability and other Logical Essays.* London: Routledge and Kegan Paul.

Savage, L. J. 1954. *The Foundations of Statistics.* New York: Wiley and Sons.

Tversky, A. 1969. "Intransitivity of Preferences." *Psychological Review* **76**: 31–48.

—. 1975. "A Critique of Expected Utility Theory: Descriptive and Normative Considerations." *Erkenntnis* **9**: 163–74.

Tversky, A., and Kahnemann, D. 1988. "Rational Choice and the Framing of Decisions." In Bell, Raiffa and Tversky (eds.), *Decision Making — Descriptive, Normative and Prescriptive Interactions.* Cambridge: Cambridge University Press.

Wittgenstein, L. 1984. "Philosophische Untersuchungen." In *Werkausgabe – Band 1.* Frankfurt a. M.: Suhrkamp.

Zilhão, A. 1998/99. "Folk-Psychology, Rationality and Human Action." *Grazer Philosophische Studien* **56**: 1–28.

Zilhão, A. 2002. "Constrangimentos Interpretativos na Compreensão da Intencionalidade do Ponto de Vista da 3ª Pessoa." In Alonso Puelles, A., and Galán Rodríguez, C. (eds.), *Wittgenstein, 50 años después.*

Mérida, Spain: Editora Regional de Extremadura.

Department of Philosophy
University of Lisbon
Portugal
AntonioZilhao@fl.ul.pt

Resumo

Tanto Quine quanto Davidson desenvolvem programas de semântica empírica que satisfazem as condições que caracterizam o chamado "ponto de vista da interpretação." O programa menos ambicioso de Quine para a tradução radical se fundamenta em dois alicerces: a causalidade e a empatia. O programa mais ambicioso de Davidson para a interpretação radical substitui a causalidade pela verdade e a empatia pela racionalidade. Embora a substituição da causalidade pela verdade intersubjetiva nos pareça ser um movimento inteiramente justificável, contudo, argumentamos que é mais realista desenvolver o trabalho de interpretação com base ns requisito menos ambicioso de Quine, da empatia, que na concepção de Davidson, da atuação [agency] humana como atuação racional. Para fundamentar essa alegação, apresentamos um argumento segundo o qual a caracterização que Davidson faz da atuação humana como racional não é compatível com seu outro requisito de que a verdade deveria dar a ligação essencial da fala com o ambiente e a ação.