Abstracta 1:2 pp. 125 – 144, 2005

ARE OUR BRAINS SUBCUTANEOUS MACHINES OF TRUTH-OPTIMIZATION?

António Zilhão

Abstract
Strategies purporting to determine the meaning of inner states of belief-content in terms of their inferential role usually assume the inner structure of the human inferential competence to be that of first order logic plus identity. Considerations of computational complexity and cumbersomeness of representation tend to undermine the plausibility of combining such strategies with this assumption. In this paper I contend that appealing to rules of default reasoning won't make things turn out any better for the inferential role functionalist.

1. Meaning Constraints on the Individuation of States of Belief-Content

It is nowadays widely acknowledged that material thinking, understood as the activity of biologically or otherwise physically realized systems of rules for truth-optimization, cannot be defined in terms of perfect or near perfect deductive ability. On the other hand, no hard core of rules for truth-optimization can be determined *a priori* as being that core of rules in terms of the following of which the necessary conditions for material thinking admit being established[1]. Therefore, a strategy purporting to determine the *meaning* of inner states of belief-content in terms of the inferential role they are supposed to play in inner-state-webs reproducing networks of propositional inference has to be species-relative or, at least, cognitive-architecture-relative. As a consequence, the initial enterprise of Inferential Role Functionalism has to be somewhat watered down. That is, instead of trying to account for intentionality in general, the supporters of this perspective can hope to go no further than striving to provide an account of *human* intentionality only.

A necessary condition for the achievement of this goal is the clarification of the inner structure of the inferential competence of humans. Now, given the parallelism inferential role functionalists assume there to obtain between networks of inner content-states and networks of inferentially connected propositions, this inferential competence has to be such that it may account for the structure underlying the language within which the above-mentioned connections between propositions are to be represented. Following in the wake of Davidson, inferential role functionalists tend to claim this structure to be that of a language of first-order quantification theory plus identity.

However, if we are going to imagine inner systems of states of belief-content displaying the inferential complexity of the human system of propositionally individuated belief-contents,

[1] I take it this was convincingly shown by Cherniak (Cherniak 1986: Ch. 2).

strictly logical relations cannot be the only relations by means of which inner states of belief-content are to be individuated. Such relations will have to be of, at least, another type. That is, besides the logical relations obtaining between any particular belief-content C and all those belief-contents which are logically connected with C by means of logical rules, there will also have to be *conceptual* relations obtaining between C and sets of other belief-contents, the terms featuring in which are semantically related to, at least, some of the terms featuring in C.

In effect, accounts of the structure of human languages in terms of languages of first-order quantification theory plus identity do try to incorporate this second aspect within their theoretical framework. Let us then see how do they do it and how is it possible to imagine that this latter type of relation may be implemented within a materially realized inner system of truth-optimization.

The picture laid down by the view of the intentional structure of the mind espoused by the inferential role functionalist is basically the picture of a hierarchy. Thus, at the bottom level, the level at which the stimuli from the outside world are supposed to exert their impact upon the mind, simple contents are supposed to be originated; such contents are thus wholly determined by their input conditions. Above in the hierarchy, one finds those other complex contents whose only constituents are simple contents linked by means of logical terms. These contents are supposed to be determined both by the input conditions determining their simple constituents and by the web of logical relations associated with the logical terms connecting them. At a higher level in the hierarchy, there are supposed to be found contents the constituents of which are, at least partly, not directly determined by any specifiable input conditions. Now, if such contents are to be functionally/inferentially identifiable, it has to be possible to put forth clearly defined structural constraints on the states embodying them other than the purely logical ones. That is, the latter constraints will not suffice, if higher-level belief-contents sharing the same logical form but having different meanings are to have a unique place in the structure of the general inferential network.

Patterns of structural connections depicting the *conceptual* or *semantic* relations obtaining between such contents and other contents will then have to be formalized in order to make these further constraints available. The structural constraints in question (such as relations of, e.g., entailment) are, in turn, supposed to be determined by those relations which obtain between the *meanings* of the non directly observable terms and expressions featuring in the propositions indexing such content-states and other terms, observable or not (such as relations of, e.g., inclusion). The web of connections determined by these relations is precisely what is sometimes called the *conceptual role* of a concept. The semantic rules that regulate conceptual roles are in turn characterized by inferential role functionalists in terms similar to those Carnap once used to characterize what he called the *meaning postulates* of a linguistic system (Carnap 1956: Appendix D)[2].

The idea of a meaning postulate can be clarified in the following way. Consider the case of an artificial first-order language L to be used to depict a very simple reality and correspondingly endowed with only a few predicates. Consider also that in such a language L all vagueness has been eliminated from all the concepts referred to by those predicates. Finally, imagine that such a language is physically implemented in some sort of computational system. Within such a language, it would be possible to lay down, for each predicate not supposed to be wholly determined by any putative input conditions, an exhaustive list of precise rules by means of which such a predicate would be wholly defined. Thus, to fall under a certain concept X, for instance, would be exhaustively defined in such a list as falling under, e.g., the conjunction of the concepts Y, W and Z and to fall under a certain other concept Q would be exhaustively defined as falling under the conjunction of the concepts X, R and S. Some or all of the concepts Y, W, Z, R and S might or might not be definable in terms of input conditions. If not, then the concepts by means of which they would be in turn defined might or might not be definable in terms of such conditions and so on. Accompanying such a top down hierarchy of concepts, there would also be a logical hierarchy of sentences. Thus, under these circumstances, a sentence S1 ascribing to the object *a* the property X would immediately entail in L the following sentences: a sentence S2 ascribing to *a* the property Y, a sentence S3 ascribing to *a* the property W, a sentence S4 ascribing to *a* the property Z, a sentence S5 ascribing to *a* the properties Y and W, a sentence S6 ascribing to *a* the properties Y and Z, a sentence S7 ascribing to *a* the properties W and Z, and a sentence S8 ascribing to *a* the properties Y, W and Z. Similarly, a sentence P1 ascribing to the object *b* the property Q would immediately entail in L a set of sentences derived from P1 in the same way as S2-S8 were derived from S1, and so on.

If we now follow in the wake of Inferential Role Functionalism and try to imagine a first-order language L' intended to depict a much more complex reality, such as one of the realities humans are supposed to have beliefs about, we will immediately realize that lots and lots of predicates are going to be needed in order to account for all the properties such a language needs to represent. If, for each term representing one such property, an explicit definition of the above-mentioned kind is going to be produced, such definitional rules will eventually end up constituting a gigantic list. If we now imagine that such a language is going to be physically implemented in some computational system or, in agreement with Inferential Role

[2] On the side of Inferential Role Functionalism see Loar (1981: 81ff) and Block (1986: 628).

Functionalism, in our brains, the size of the rule-list ceases to be just an inconvenience and threatens to become a difficult practical problem. Both the silicon and the cerebral computational system will necessarily have limited resources and an enormous amount of those resources will be mobilized by the storage and handling of such a list. Obviously, that would be a highly undesirable state of affairs.

The idea of introducing meaning postulates in the system is then put forth in order to respond to such a problem. Thus, instead of laying down explicitly precise rules for the use and understanding of each of the predicative terms of the language, one would only need to lay down explicitly the meaning relations obtaining between such terms. The sentences S1-S8 above are an example of what such relations might be like. Thus, in the example above, rather than producing explicit definitions of what it is to be an X or an Y or a Q or an S in terms of, e.g., conjunctions of other properties, one would only produce the meaning postulates according to which being, e.g., a Q entails being an X, being an X entails being a W, etc. Such terms would then be wholly characterized in terms of these sentential concatenations. The latter would then stipulate the way the terms in question would contribute to the inferential role of the sentences they would be part of.

From Carnap's standpoint, it is thanks to meaning postulates that one is able to account for the relations of analyticity existing within a linguistic system. Thus, given the combined facts that a formal system of first-order logic plus identity does not tell us anything about the relations obtaining between, e.g., the predicates 'x is a bachelor' and 'x is unmarried' and that, at the same time, we want to say that 'John is unmarried' is *entailed* by 'John is a bachelor', we introduce in the characterization of the language within which we want to talk about bachelors and unmarried people the following meaning postulate P1:

$$P1 = (\forall x)\{B(x)\rightarrow [H(x)\&{\sim}M(x)]\},$$

in which B(x) represents the property of being a bachelor, H(x) represents the property of being a man (i.e., a human male) and M(x) represents the property of being married. After this introduction, it will indeed be true in the enlarged system that having the property of being a bachelor entails having the property of being unmarried. And this entailment obtains irrespective of the presence in the system of any explicit definitions of 'Bachelor' or 'Married' or 'Man'. Thus, the sentence 'If John is a bachelor, then John is unmarried' can be classified as an analytic sentence because it follows logically from P1 within the system of first-order logic with identity enlarged with at least P1 as a meaning postulate.

There is a disagreement between inferential role functionalists and Carnap as to the status of meaning postulates. Predictably, the former claim that by laying down structural constraints reproducing meaning postulates one is not thereby committed to any notion of analyticity (Loar 1981: 81-2; Block 1986: 629). Indeed, their idea is precisely that questions of meaning should be grounded in questions of mind computational fact rather than in questions of conventional choice of a language. Thus, the identification of a particular set of meaning postulates is to be seen, according to the supporters of this idea, as an empirical hypothesis about the inner organization of that part of our central control organs within which inner content-states are supposed to be manipulated. Regardless of such a disagreement, both Carnap and inferential role functionalists agree on two things. First, that the meaning relations that are supposed to be codified in the meaning postulates for a language intended to account for the contents of our usual beliefs and desires define an enormous network of inferential connections obtaining between the sentences of the languages containing such predicates. Second, that the meaning relations in question have to be those which *grosso modo* underlie our linguistic habits.

In order to assess the success of this strategy, one has to investigate whether or not it is indeed able to meet this latter *desideratum.* If it is not, then we will be left completely in the dark about how on earth inferential role might possibly account for the “narrow” meaning of our putative inner states of propositional attitude-content.

2. Negative Meaning Constraints

The first question to ask in order to assess this strategy is the following. How many types of meaning-postulate are there to be found? That is, are all meaning connections obtaining between predicates of a natural language as easily characterizable as the above-mentioned connection obtaining between 'Bachelors' and 'Unmarried Men'? If not, what sort of meaning postulates regulates the semantic connections by means of which the meaning of most terms and expressions of our natural languages is supposed to be represented? Let us begin answering this question by first considering Carnap's own suggestions on the subject (Carnap 1966: 257-64). According to Carnap, if one were to launch an empirical study in order to find out what were the meaning postulates of a natural language, one would presumably find among them universal affirmative sentences such as 'All red-headed woodpeckers are birds', 'All birds are animals', 'All men are rational', 'All men are animals', etc[3]. The relations established by means of these sentences would thus regulate the conceptual space within which we talk about such things and understand other people's talk about them. And, indeed,

[3] Loar's clarification of the idea of a M-constraint (his term for meaning postulate) revolves largely around the same kind of examples, which he characterizes as examples of relations of sub-kind to kind and of determinable-determinate. See Loar (Loar 1981: 82-3); and also Block (Block 1986: 628-9).

from the above-mentioned meaning postulates, one can derive by simple logical means other sentences such as 'All red-headed woodpeckers are animals', 'Some animals are men', 'Some birds are red-headed woodpeckers', etc., the assent to which is arguably also to be considered to be equally necessary to ascribe to someone the understanding of natural language concepts such as the concepts 'woodpecker', 'bird', 'man', etc.

However, our actual knowledge of the meaning of terms such as 'woodpecker' or 'bird' is not wholly reflected by their representation within a first-order logical language by means of relations of inclusion only. As a matter of fact, our understanding of the meaning of 'woodpecker' reflects itself not only in the fact that we are supposed to be able to infer that to be a woodpecker entails being a bird, and therefore being an animal, but also in the fact that we are supposed to be able to infer also that to be a woodpecker entails *not* being a mammal or *not* being a plant. If, however, we would assume that the kind of sentences present in Carnap's examples of meaning postulates for declarative terms were somehow prototypical, we would quickly realize that, under such circumstances, it would not be possible to derive within such languages any relations of exclusion between concepts. Therefore, questions such as 'Is a red-headed woodpecker a mammal?' would have no possible answer within these linguistic systems either. That is, although the existing meaning postulates would not allow an organism or machine endowed with these linguistic systems to answer affirmatively to questions of this kind, they would not allow it to answer negatively to them either. As a matter of fact, in order to be able to derive a sentence such as 'No woodpecker is a mammal' within a network of first-order logic and associated meaning postulates, one needs to have available explicit *negative* meaning postulates such as, e.g., 'No bird is a mammal'.

Thus, if we are to imagine an intentional system programmed with a first-order logical language the behavior of which is supposed to be similar to our own linguistic behavior, we will need to add to its stock of affirmative meaning postulates another kind of meaning postulates, namely, negative meaning postulates. The latter regulate the relations of exclusion obtaining among the concepts referred to in that language. As a matter of fact, it seems to be as intuitive to claim that the understanding of the concept 'woodpecker' entails the assent to the sentence 'All woodpeckers are birds' as it is to claim that such an understanding entails either the assent to the sentence 'No woodpecker is a mammal' or the dissent from the sentence 'Some woodpeckers are mammals'.

We have thus realized that a linguistic system of first-order logic intended to account for our use and understanding of ordinary language and therefore for our use and understanding of sentences of intentional-state-ascription will need to include among its meaning postulates sentences expressing relations of exclusion between concepts. So far, such a realization does not seem to be in contradiction with anything Carnap suggested. However, although Carnap did not contend, and probably did not think, that the web of meaning postulates regulating our use of the descriptive terms of our natural languages consists only in universal affirmative sentences establishing relations of inclusion between concepts, neither he nor Loar or Block seem to have noticed an important problem associated with the necessary inclusion within such a web of negative meaning postulates.

This problem is a problem of practical implementation and underlies any attempt to implement a stock of negative meaning postulates in a computational system. It can be stated thus: what and how many negative meaning relations are there associated with each affirmative meaning relation? That is, when we know that, e.g., all woodpeckers are birds, how many things do we know that woodpeckers are not? Simply raising the question makes us already see the enormity of the problem. The number of possible negative meaning connections vastly exceeds the number of possible affirmative meaning connections. Moreover, even if we consider that the actual affirmative meaning connections obtaining between the terms of our language can be restricted to a bare minimum, still, *each* such meaning postulate we might want to consider to be indispensable to the characterization of any of the terms in question will have to be accompanied by a huge number of negative meaning postulates. As a consequence, it might easily turn out to be completely unfeasible to represent explicitly all the negative meaning connections that might be implicit in the use of a complex language system.

Such an unfeasibility problem leaves then the inferential role functionalist having to face the following dilemma: on the one hand, in order to avoid the unfeasibility in question, negative meaning postulates seem to have to be implicitly rather than explicitly represented in the inferential network accounting for the inner computations actually performed by the believer's central control organ upon its stock of stored content-states; on the other hand, however, in order to be able to generate those relations of exclusion between concepts we consider to be associated with the understanding of the meaning of the concepts in question, a first-order logic linguistic system needs the negative meaning postulates in question to be explicitly represented by such a network.

Prima facie, the inferential role functionalist will have thus to choose between two equally uncomfortable options: either he preserves first-order logical representation and, in order to preserve feasibility, he abandons the idea that he might be able, by means of such a representation, to account for all the meaning connections we tend to consider to be necessarily associated with the use and understanding of the descriptive terms of our language, that is, he abandons the idea that he might be able to account for conceptual roles in general; or he tries to preserve the integrity of such connections and abandons first-order logic representation. Needless to say, each of these options contradicts a fundamental assumption of his standpoint.

3. Exceptions

In the previous section, I reviewed the implementation problem involved with the representation of universal negative meaning postulates in a computational system endowed with a language of first-order logic. In this section, I will focus on a wider problem. It is the following. Is it legitimate to represent by means of universally quantified sentences in general, be them affirmative or negative, the meaning connections in terms of which our use and understanding of ordinary language descriptive terms is supposed to be regulated? This new question does not deal with implementation problems associated with a particular kind of sentences, as the former did, but with a different sort of problem, namely, the problem of determining whether or not, independently of questions of implementation, a certain type of representation is or is not adequate to represent the cognitive competence it purports to represent.

The problem associated with representing meaning postulates as universally quantified sentences is that, given its rigidity, such a form of representation does not seem to be able to account for the existence of exceptions. And exceptions are to be found falling under almost every concept referred to by ordinary language predicates.

In order to illustrate the aforementioned problem, let us consider the term 'bird', for instance[4]. Can the normal understanding of this term in ordinary language contexts be adequately represented in terms of the web of relations of inclusion regulated by sentences such as 'All birds are warm-blooded', 'All birds are vertebrates', 'All birds are animals', etc. and of relations of exclusion regulated by sentences such as 'No bird is a mammal', 'No bird is a reptile', etc.? The crucial question associated with the introduction of these putative meaning postulates is the following. If we decide to select as essential to determine meaning those semantic connections which are regulated by postulates which apply with no exception to every single bird, we will obviously get no exceptions; we will, however, get too much generality (mammals are also warm-blooded vertebrates, for instance). But if we decide to select as essential to determine meaning those characteristics which, among warm-blooded animals, apply only to birds, neither will the semantic characterization of 'bird' be given by the web of

[4] This is the standard example used in the literature to illustrate this problem. See, e.g., Ginsberg (1993: ch.11); or Oaksford and Chater (1993: 38).

connections established by that set of universally quantified sentences the members of which are 'All birds are feathered', 'All birds lay eggs', 'All birds fly', etc. The reason justifying the latter statement is that, although some of the characteristics ascribed to all birds in such sentences apply only to birds and to no other type of warm-blooded creature, there are too many birds to which some of such characteristics do not apply. Thus, the conjunction of sentences ascribing such characteristics to all birds will be vulnerable to too many counterexamples (male birds do not lay eggs, penguins, ostriches, rheas, emus and kiwis do not fly, and the chicks of different bird species are born featherless; all of them are, however, birds).

Attempts to provide a solution within a first-order logic representation to the problem of, simultaneously, accounting for exceptions and avoiding both the excess of generality and the vulnerability to counter-examples can be tried, but with no success. Three possibilities to provide such a solution seem to be *prima facie* available.

The first possibility is the one that is simultaneously the most immediate and the one that is most obviously unacceptable. It is to include in the proposed set of meaning postulates sets of sentences such as, e.g., the following:

P2 = 'All birds fly', i.e., $(\forall x)[B(x)\rightarrow F(x)]$;
P3 = 'All kiwis are birds', i.e., $(\forall x)[K(x)\rightarrow B(x)]$;
P4 = 'No kiwi flies', i.e., $(\forall x)[K(x)\rightarrow \sim F(x)]$.

Such a set of sentences expresses naturally some of the intuitively essential meaning connections obtaining between the terms 'bird' and 'kiwi'. It is therefore natural to want to have them codified in our meaning postulates. This representation of these meaning connections has an obvious drawback, however. It is the fact that, from such a set of sentences, the contradictory sentence P5 is obviously derivable:

P5 = ‘All kiwis fly and do not fly’, i.e., $(\forall x)\{K(x)\rightarrow [F(x)\&\sim F(x)]\}$.

That is, the attempt thus to interpret the meaning postulates underlying our use of descriptive words such as 'kiwi' or 'bird' gives rise to an inconsistent representation. And an organism endowed with such a combination of meaning postulates associated with a language L of first-order logic would be an organism completely unable to establish any connections between the information it might have that a particular bird was a kiwi and the flying capacities of that bird. But that would be unacceptable on two accounts. First, it would violate the only constraint on systems of truth-optimization left by the abandonment of the view that there would have to be a hard-core of logical rules, namely, the constraint on consistency; secondly, it would contradict the assumption that it is precisely the holding of such conceptual connections that is supposed to regulate the semantics of our representational system and is thus supposed to account for the possibility of individuating a substantial part of our belief contents.

A second attempt to try to reach the above-mentioned goal would be to replace those universally quantified sentences featuring in meaning postulates that originate counterexamples with the corresponding existentially quantified sentences. Under those circumstances, however, the inferential potential of these postulates would be totally lost. As a matter of fact, if we know beforehand that some birds fly and if we are informed that a particular set of creatures is a set of birds, then it will not be possible to infer anything about the flying capacities of such creatures. But, if these postulates do not establish relations of inclusion among the relevant concepts and therefore do not draw clearly defined inferential paths associated with the concept one is trying to clarify, then no distinction will be established between those characteristics which are allegedly essential to the characterization of the concept, such as, in the present case, flying, egg-laying or being feathered, and those characteristics which are obviously inessential, such as color, size, color of the eyes, etc. That is, if we consider the representation of the following two sentences P6 and P7:

P6 = 'Some birds fly'
P7 = 'Some birds are yellow',

in first-order quantification theory, namely, $(\exists x)[B(x)\&F(x)]$ and $(\exists x)[B(x)\&Y(x)]$, with B(x) standing for ´x is a bird´, F(x) standing for ‘x flies’ and Y(x) standing for ‘x is yellow’, we will immediately realize that they have exactly the same logical force. Why the former rather than the latter should be selected as a meaning postulate appropriate to characterize the meaning of 'bird', would then have to remain an unexplained choice.

Finally, a third possibility would be to preserve the idea that meaning postulates should be represented by means of universally quantified sentences but to try to account for the existence of exceptions by means of the complexification of the antecedents of such sentences. For instance, the meaning postulate P2 = 'All birds fly', i.e., $(\forall x)[B(x)\rightarrow F(x)]$, would be replaced by the following meaning postulate P8:

P8 = 'All birds, with the exception of penguins, ostriches, rheas, emus, kiwis and chicks, fly', i.e., $(\forall x)\{[B(x)\&{\sim}P(x)\&{\sim}O(x)\&{\sim}R(x)\&{\sim}E(x)\&{\sim}K(x)\&{\sim}C(x)]\rightarrow F(x)\}$, with P(x) standing for ´x is a penguin´, O(x) standing for ‘x is an ostrich’, R(x) standing for ‘x is a rhea’, E(x) standing for ‘x is an emu’, K(x) standing for ‘x is a kiwi’, and C(x) standing for ‘x is a chick’.

The problem engendered by this suggestion is that the property of flying would not be represented as associated with birds in general, but only with those birds which do not belong to any of the non-flying species. Thus, it could not really be said that the general concept of bird would be being clarified by such a meaning postulate. As a matter of fact, if enough exceptions would be listed in the antecedent of one such expression, any property of any bird could be selected to feature in its consequent, and hence too many meaning postulates for 'bird' would be available. Thus, quantitative criteria would have to be laid down in order to distinguish between those antecedents exhibiting a small enough number of conjuncts, and thus acceptable to be selected as antecedents of genuine meaning postulates, and those antecedents exhibiting too large a number of conjuncts, and thus unacceptable to be selected as antecedents of genuine meaning postulates. However, it would not at all be guaranteed that such a quantitative criterion of choice for antecedents of meaning postulates would match our intuitions of what it is and what it is not essential for something to have in order to be intuitively considered to be a bird. And if the introduction of some other criteria of essentiality were needed, we would have returned to our starting point.

A consequence of the state of affairs characterized above would be that an organism endowed with meaning postulates or, in Loar’s vocabulary, M-constraints, like these, and placed in a situation in which it would have to act based upon incomplete information would be as unable to deduce anything about most of the relevant characteristics of birds from the knowledge that a particular set of creatures were a set of birds as would be an organism the cognitive and linguistic apparatus of which would be endowed with existentially quantified meaning postulates. As a matter of fact, unless the means were given to it to check whether or not the creatures in question were penguins, ostriches, rheas, emus, kiwis or chicks, no inference could be made as to the flying capacities of such creatures. But organisms tend to have to act and to react quickly in situations in which the amount of relevant information available is typically scarce; under such circumstances, it would probably be to the advantage of an organism to be able to infer that a particular unknown member of a set has those properties which are deemed to be prototypical for that set and to act accordingly, even if the prototype does not apply uniformly to all members of the set. Thus, to have a mind endowed with such meaning postulates would not only be to have a mind presumably different from our own, but it would also be to have a mind the inner organization of which, rather than being to the evolutionary advantage of the organism possessing it, would probably constitute an evolutionary handicap for such an organism.

The analysis of the three above portrayed ways of trying to integrate the existence of exceptional cases within a first-order logic representation of the semantic connections we intuitively consider to be associated with our use and understanding of the descriptive terms of our natural languages seems to point to the apparently inescapable dilemma already encountered in the previous section. Such a dilemma, which is the one the inferential role functionalist has to face, is the following: either he preserves first-order logical representation and he abandons the idea that by means of such a representation he is accounting for the way we use and understand the descriptive terms of our natural languages; or he stays faithful to the project of accounting for such an use and understanding and he has to abandon the idea that such uses and understanding might be adequately rendered in terms of their putative first-order logical representation.

If he chooses the first horn of the dilemma, the inferential connections he will be codifying will not be those that are supposed to define the "narrow" meaning of our inner states of propositional attitude-content and the whole effort will be psychologically useless. If he chooses the second horn of the dilemma, two possibilities are still open before him. The first is the presentation of some other system of rules of truth-optimization in terms of which such use and understanding might be accounted for. The second is to abandon the idea that our inferential competence is a competence for truth-optimization and therefore to abandon using the knowledge we have of abstract systems of truth-optimization as a psychological tool to account for the meaning of our inner states of propositional attitude-content independently of their external effects. Obviously, the only path he can choose without abandoning his basic theoretical commitment is the path defined by the first of the two possibilities opened up by the choice of the second horn of the dilemma. It will therefore be the target of my analysis in the next section.

4. Exceptions, Negative Meaning Constraints and Default Reasoning

What other system of rules for truth-optimization could there be that might be successfully used to account for the problem of how to establish negative meaning constraints and the problem of the existence of exceptions? Some claims have been made recently that these problems could be accounted for by means of mechanisms implementing a non-classical but equally logical form of representation, namely, *default reasoning*. According to the relevant literature, default reasoning has been tried in many natural language understanding systems studied in the field of Artificial Intelligence[5]. Given this state of affairs, the question naturally arises whether or not the facts about the way we do seem to establish meaning relations

[5] See, e.g., Reiter (Reiter 1985: 402).

between concepts in everyday language use are a consequence of a putative underlying subcutaneous fact, namely, the fact that our minds are also endowed with default rules of reasoning. If this were to be the case, then it might still be possible to rescue Inferential Role Functionalism by reconstructing along the lines laid down by a suitable theory of default reasoning the image of the structural constraints organizing the inferential network within which inner states of belief-content would be given meaning.

Default reasoning is a form of inference of the following kind: 'if such and such cannot be derived from a given belief set, then infer this and that'. The basic idea underlying such a reasoning strategy is that, instead of having to have explicitly stored beliefs about whether certain relations between concepts obtain or not in order to be able to derive further beliefs about the obtaining or non obtaining of other specific relations between concepts or about the obtaining or non obtaining of the consequences assumed to be derivable from them, a cognizer implementing such a strategy will be allowed to infer that such further relations between concepts obtain or not by means of the simple application of the form of inference presented above. Given the fact that this form of inference is supposed to be implemented in a procedural rule the cognizer's brain is endowed with, the cumbersome explicit representation within his belief set, needed by classical logical procedures, of what may, in many cases, be a huge set of beliefs about conceptual connections will be avoided. This fact notwithstanding, the same deductive results will be achieved. The cognizer will thus be able to "jump" to the intended conclusion and, therefore, a substantial reduction in the complexity of both the representational and the computational apparatus will be achieved.

In particular, the following default rule R1 might be capable of *prima facie* solving the problems raised in section 2:

R1 = ´For any predicate and any object, if it is not derivable that a particular predicate is ascribable to a particular object, then it is derivable that such a predicate is not ascribable to such an object.'

The rule R1 admits being formalized thus:

$$R1 = (\forall x)(\forall \Pi)\nvdash \Pi(x)\vdash {\sim}\Pi(x) \text{ (Reiter 1985, pp. 404-5).}$$

A cognizer whose brain were endowed with an implementation of this rule would not need to have an explicit representation of any negative meaning postulates in its belief set; the simple fact that the explicitly represented affirmative meaning postulates would not allow the cognizer to derive from its belief set a sentence S ascribing a certain predicate Q to a particular object *a* known to satisfy some other specific predicate P would be enough to allow the cognizer in question to infer by means of a single application of the rule above the negative sentence S' denying that *a* falls under Q. Using classical deductive reasoning procedures such a sentence would only be derivable from a negative meaning postulate establishing a relation of exclusion between the concepts P and Q.

As an example of this kind of default reasoning consider the following case. Imagine that a cognitive apparatus is implemented with a first-order language L enriched with a set of affirmative meaning postulates interpretable as regulating the use of terms such as 'red-headed woodpecker', 'bird', 'mammal', 'animal', etc. Imagine further that a pair of sentences, T and U, ascribes to Woody the properties of being, respectively, a red-headed woodpecker and a mammal. Now, consider that T belongs to the set of belief-contents stored in the knowledge-base of the cognizer, whereas U neither belongs to that set nor is derivable from it. If we now imagine that the default rule mentioned above were also part of this cognitive apparatus, the simple activation of this rule would engender the immediate inference of the sentence U' = 'Woody is not a mammal'. This inference would thus have been performed despite the fact that neither U' nor a universal meaning postulate such as, e.g., 'No red-headed woodpecker is a mammal' or 'No bird is a mammal', from the conjunction of which with the appropriate affirmative meaning postulates U' might have been directly deduced according to the traditional rules of first-order logic, would be explicitly represented and stored in the cognizer's belief set. Thus, no questions of feasibility associated with the explicit representation in the belief set of all the negative meaning postulates connected with the use of any of the concepts of the language would need to be raised. Such a belief set would contain only a moderate amount of affirmative meaning postulates and any of the potentially unlimited number of relations of exclusion obtaining among the concepts of L would be inferred by default.

Similarly, another default rule could also be put in place in order to *prima facie* solve the problems dealt with in section 3, namely, the problems involved in accounting for the existence of exceptions. A default rule of reasoning accounting for the obvious looseness of our everyday use of concepts, let us call it R2, might be the following. Assuming that within a first-order language L enriched with a set M of meaning postulates a concept P would, in general, be characterizable in terms of a set N of affirmative meaning postulates establishing relations of inclusion between the concept P and certain other concepts Q, R and S; and assuming also that there would be certain exceptional cases of objects falling under P but not falling under Q, and that such objects would typically fall under the sub-kinds K, L and M; then, for any object x falling under P, if it could not be explicitly derived in L that such an object would fall under the sub-kind K or under the sub-kind L or under the sub-kind M, then the inference would be allowed that such an object would fall under Q[6]. Thus, instead of depending on the capacity to check in each case whether or not a particular object *a* falling under P also falls under sub-kinds K or L or M, the possibility of deducing that a particular object falling under P also falls under Q is automatically guaranteed by the default rule, provided no explicit knowledge of the falling of *a* under sub-kinds K, L or M is available.

A particular example of an application of the aforementioned default rule is the following. Suppose we know that Woodstock is a bird but that we do not know what kind of a bird Woodstock is. Using default rule R2 we are allowed to conclude that Woodstock can fly, provided that no explicit information is available in our belief set classifying Woodstock under any one of the non-flying kinds of bird. Thus, in order to ascribe to Woodstock the property of flying, there is no need to establish explicitly the negative information that Woodstock is not a penguin or an ostrich or a rhea or an emu or a chick, etc. Such an explicit verification would probably be a very costly process in terms both of time and resources. The fact that no explicit affirmative information about Woodstock being an ostrich, or an emu, or a penguin, etc., would be available in the belief set, would actually be sufficient to allow the cognizer to infer that Woodstock would be a flyer. However, if the acquisition of further information would eventually end up making the cognizer generate the new belief that Woodstock was after all a penguin, or an emu, or an ostrich, etc., then the previous conclusion would be defeated, and the default belief about the flying capacities of Woodstock would be replaced by an explicit belief ascribing to Woodstock a lack of flying capacities.

As the use of the term 'defeated' indicates, this latter rule of default reasoning is actually a rule formalizing a type of reasoning usually called *defeasible reasoning.* A piece of reasoning is considered to be defeasible if and only if it is a piece of reasoning in which the conclusions one draws by means of the use of, e.g., typical meaning postulates can be passed over by the obtaining of new information contradicting one or some of the assumptions codified in the meaning postulates themselves. Thus, a particularity of defeasible reasoning is, more generally, the fact that, contrary to classical processes of inference, a regular conclusion obtained from a set M of premises is not necessarily a regular conclusion obtainable from a set N of premises which is such that N is a superset of M. In other words, the growth of the extension of the set of regular conclusions is not monotonically connected with the growth of the extension of the set of sentences present in the belief set of the cognizer and thus usable as premises. This is why this kind of reasoning is also called *nonmonotonic reasoning* (Ginsberg 1993: 217).

[6] Reiter presents a particular version of this rule too (Reiter 1985: 407).

5. Closed World Assumption and Feasibility

In the present section, I will discuss whether or not the appeal to rules of default reasoning is a possible way of rescuing the view of Inferential Role Functionalism on how inferential relations are supposed to account for the individuation of the “narrow” meaning of our putative inner states of propositional attitude-content. I will present two objections to this move.

The first objection is the following. As used in computational practice, a default rule such as R1 can be implemented in an inferential system only under the assumption that the set of beliefs feeding such an inferential system embodies a perfect knowledge of the domain that it purports to represent. This assumption is usually called the "closed world assumption" (Reiter 1985: 405). It entails that it has to be the case that all possible positive facts relevant to the full representation of the domain are actually represented in the system's belief set. Otherwise, there would be no guarantee whatsoever that a default rule such as R1 would not be systematically generating falsities. In order to illustrate this idea let us consider the following example. Let us suppose that a first-order language L has been defined together with a set of meaning postulates regulating the way concepts such as 'dolphin', 'maritime mammal', etc., are to be understood. If it would not be possible to generate within a given belief set expressible in L the belief that the sensorial apparatus of dolphins is endowed with a module for eco-location, the computational system of which such a belief set would be a part would automatically generate by default the belief that the sensorial apparatus of dolphins is *not* endowed with a module for eco-location. However, as it is widely known, that would be a falsity. Nevertheless, it is hardly disputable that the meaning postulates regulating the use of concepts such as 'dolphin', 'maritime mammal', and so on, are not supposed to contain any sort of meaning connections relating these concepts with concepts such as, e.g., the concept 'bearer of a sensorial apparatus endowed with a module for eco-location'.

In short, given the combined facts that Loar's or Block's meaning constraints are basically a subcutaneous computational realization of Carnap's meaning postulates, and that Carnap's meaning postulates are supposed to be the formalized counterpart to the analytical sentences of a given observational language L, it follows that a computational system combining a relevant set of meaning postulates for a given observational language L with a default rule such as R1 would, when properly queried, immediately infer the negation of any true synthetic statement not independently stored in the system's belief set. As a matter of fact, the synthetic statements of a language L are, by definition, precisely those statements that cannot be derived from the logical rules and the meaning constraints of the language alone. Thus, one such statement not explicitly stored in the system's belief set would not be derivable and, therefore, its negation would be taken to be true by default. As a consequence, ignorance would be an impossible state for such a system to be in. In spite of the fact that sometimes people do display the tendency to explicitly deny the holding of facts they are ignorant of, it seems to me to be too excessive to maintain that the computational representation of the structure of our minds is such that it implies the impossibility of our generating states of ignorance about any aspect of the world.

The second objection is the following. According to Church's Theorem, there can be no effective procedure allowing us to determine whether or not an arbitrary formula of the theory of quantification is a theorem of first-order logic. But if that is the case, it may well happen that no procedure will be available in an arbitrary case to decide that a particular first-order formula Φ is not derivable from a particular belief set. Under such circumstances, however, it would not be possible to set the putative corresponding default rule in motion. At this stage, we have to remember ourselves that a default rule such as R2 is a rule the aim of the implementation of which is to simplify the computational work of the system. A system using R2 is a system which is dispensed of having to check explicitly that a particular individual *i* falling under a general predicate P does not fall under those exceptional sub-kinds of P the members of which do not fall under a predicate Q, usually associated with P. If the system is endowed with a rule such as R2, it is enough, in order to ascribe Q to *i,* that no belief be derivable according to which *i* falls under the exceptional sub-kinds in question. But, in order to ascertain that no such belief is derivable, the whole of the creature's belief set must be exhaustively searched. However, according to Church's Theorem, there is no way of guaranteeing that a computational procedure aimed at checking whether or not a belief of the kind in question is derivable from a given belief set will terminate. This means that the subroutine upon which the application of the default rule is based may never terminate and therefore the default rule itself may be inapplicable. Reiter himself refers this problem as a "peculiar and intuitively unacceptable behavior" of a default theory (Reiter 1985: 408). Thus, the adoption of this reasoning strategy as a means to reduce the complexity of the representational apparatus runs the risk of creating an even worse problem, namely, the problem of defining an approach to the computational problem it purports to solve that not only is too complex and cumbersome but actually unfeasible.

6. Conclusion

The two objections mentioned above seem to me to indicate that the appeal to rules of default reasoning in order to characterize the pattern of inferential procedures in terms of which meaning postulates might fit appropriately in our psycho-semantic network will not be successful. As a matter of fact, after this discussion, we ended up having to face the following two discouraging conclusions. First, an original problem of computational explosion was evaded by the introduction of an implausible form of representation. Second, an original problem brought about by a cumbersome and unsatisfactory form of representation was evaded by the introduction of a procedure which is bound to suffer from a problem of computational intractability actually similar to the problems one is confronted with when one tries to understand a thinker's inferential competence in terms of perfect deductive ability. Thus, we are no better off now than we were at the outset.

Therefore, unless inferential role functionalists are able to come up with a new, physically as well as cognitively realistic, account of the relevant human inferential competences, we are left completely in the dark about the effective means in terms of which the meaning of our putative inner states of propositional attitude-content might be given by their inferential role. And if we are left in the dark about that, then Inferential Role Functionalism becomes a somewhat void philosophical doctrine.

António Zilhão
Universidade de Lisboa

Bibliography

Block, N. 1980: "What is Functionalism?" in his *Readings in Philosophy of Psychology.* London: Methuen, vol. I, pp. 171-83.

Block, N. 1986: "Advertisement for a Semantics for Psychology" in *Midwest Studies in Philosophy*, 10, pp. 615-78.

Block, N. 1990: "Can the Mind change the World?" in Boolos, G. (ed.) *Meaning and Method – Essays in Honor of Hilary Putnam.* Cambridge: Cambridge University Press, pp. 137-70.

Block, N. and Campbell, J. 1987: "Functional Role and Truth Conditions" in *Proceedings of the Aristotelian Society,* Supplementary Volume 61, pp. 157-81.

Carnap, R. 1956: "Meaning Postulates" in his *Meaning and Necessity – A Study in Semantics and Modal Logic.* Chicago: University of Chicago Press, 2nd edition, Appendix D, pp. 222-29.

Carnap, R. 1966: "Analyticity in an Observation Language" in his *Philosophical Foundations of Physics.* New York: Basic Books, pp. 247-56.

Cherniak, C. 1986: *Minimal Rationality.* Cambridge (MA): The MIT Press.

Davidson, D. 1984: "Theories of Meaning and Learnable Languages" in his *Inquiries into Truth & Interpretation.* Oxford: Clarendon Press, 1984, pp. 3-15.

Davidson, D. 1984: "Truth and Meaning" in his *Inquiries into Truth & Interpretation.* Oxford: Clarendon Press, 1984, pp. 17-36.

Davidson, D. 1984: "Radical Interpretation" in his *Inquiries into Truth & Interpretation.* Oxford: Clarendon Press, 1984, pp. 125-139.

Ginsberg, M. 1993: *Essentials of Artificial Intelligence.* San Francisco (CA): Morgan Kaufmann.

Kleene, S.C. 1967: *Mathematical Logic.* New York: Wiley & Sons.

Levesque, H.J. 1988: "Logic and the Complexity of Reasoning" in *Journal of Philosophical Logic,* 17, pp. 355-89.

Levesque, H.J. & Brachman, R.J. 1985: "A Fundamental Tradeoff in Knowledge Representation and Reasoning" in Brachman & Levesque (eds) *Readings in Knowledge Representation.* Los Altos (CA): Morgan Kaufman, pp. 41-70.

Loar, B. 1981: *Mind and Meaning.* Cambridge: Cambridge University Press.

Oaksford, M. & Chater, N. 1993: "Reasoning Theories and Bounded Rationality" in Manktelow and Over (eds) *Rationality – Psychological and Philosophical Perspectives.* London: Routledge, pp. 31-60.

Reiter, R. 1985: "On Reasoning by Default" in Brachman & Levesque (eds) *Readings in Knowledge Representation.* Los Altos (CA): Morgan Kaufman, pp. 402-10.