Symbols and subsymbols for representing knowledge: a catalogue raisonné*

Marcello Frixione
Department of Philosophy
and Department of Communications, Computer and Systems Science (DIST),
University of Genoa, Italy

Giuseppe Spinelli
Department of Communications, Computer and Systems Science (DIST), University of Genoa, Via Opera Pia 11A, 16145 Genoa
Italy (spinelli@dist.unige.it)

Salvatore Gaglio
Department of Electrical Engineering,
University of Palermo, Italy

Abstract

Traditional artificial intelligence studies generally approach the problem of representing knowledge following the so-called knowledge representation hypothesis, as formulated by Brian Smith.
More recently the development of the connectionist paradigm has questioned the symbolic approach to the study of the mind bringing about a more articulated view of the problem. This article singles out five possible approaches to the problem of knowledge representation in cognitive science: compositional symbolic approaches, local non-compositional approaches, distributed non compositional approaches, cognitive subsymbolic approaches and "neural" subsymbolic approaches. In particular, in the subsymbolic cognitive approach the elements that make up the representation system are not symbols with an ascribed meaning nor do they correspond to anatomic entities at a neurological level; rather they are to be considered as "theoretical constructs" of a theory of the cognitive level which permit the deduction (in the sense of "computation") of cognitive behaviours which cannot be otherwise modelled. We consider the development of models of this kind to be essential to a computational approach to the problem of reference without hypothesizing "magical qualities" of the mind (in the sense of assuming a necessary connection between mental symbols and their referents), while remaining within a functionalist vision, in the wider sense, which does not make reference to the specific physical properties of the neural hardware.

1 A geography of knowledge representation

"Traditional" approaches to artificial intelligence generally adopt the so-called *knowledge representation hypothesis* as formulated by Smith [1982] to tackle the problem of knowledge representation.

Smith's formulation states "any process capable of reasoning intelligently about the world must consist in part of a field of structures, of a roughly linguistic sort, which in some fashion represent whatever knowledge and beliefs the process may be said to possess" [Brachman and Levesque, 1985, p. 33]. He also states that there is "an internal process that 'runs over' or 'computes with' these representational structures ... this ingredient process is required to react only to the 'form' or 'shape' of these mental representations, without regard to what they mean or represent - this is the substance of the claim that computation involves formal symbol manipulation" (ibid.). This, substantially, is the approach behind the formalisms used in A.I. (logic formalisms, frames, production rule systems). This approach follows the hypothesis that the mind is a formal symbol system [Newell, 1980].

The development of the connectionist paradigm, by questioning the symbolic approach to the study of the mind, has further articulated the situation. The connectionists disdain the hypothesis that mental structures are, in a wide sense, of a linguistic nature and that cognitive activities can be reduced to formal symbol manipulation. In addition connectionist theories have brought about a renewed interest in the lower levels of mental organisation and neural hardware. This is not however a head-on collision between two monolithic positions, the "classical" stance on the one side and the connectionist stance on the other. Recent debate between traditional and connectionist models has become more articulated. Following mainly McClelland *et al.* [1986], Fodor and Pylyshyn [1988], Smolensky [1988] we have drawn up the following "geography" of the approaches to the knowledge representation problem in cognitive science (see Figure 1):

- 1. compositional symbolic approaches
- 2. local non compositional symbolic approaches

* This work was partially supported by a grant of the Italian Ministry of Education (MPI)

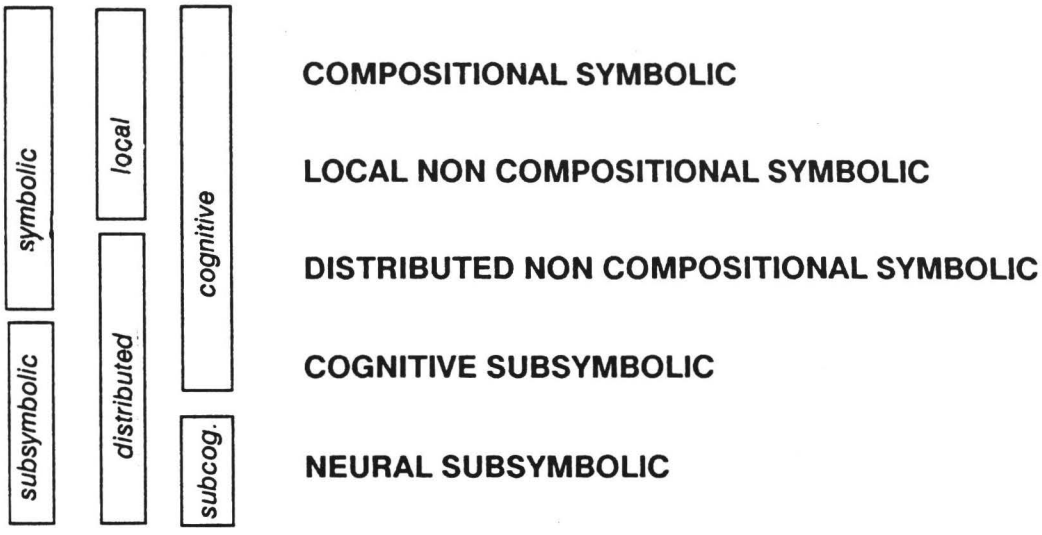

Figure 1

- 3. distributed non compositional symbolic approaches
- 4. cognitive subsymbolic approaches
- 5. neural subsymbolic approaches.

Point 1. comprises the approaches in line with Smith's knowledge representation hypothesis. Following Fodor and Pylyshyn [1988] we can say they are characterized by their adoption of the principle of *compositionality of meaning* . In brief, the syntax of the representation system distinguishes between syntactically atomic expressions and complex expressions. Complex expressions can be generated from atomic expressions using recursive syntactic rules, and the meaning of a complex expression is a function of its syntactic structure and the meaning of the atomic expressions in it. It can be hypothesized that each representation system of this type can be given a model theoretic semantics. If, in fact, the representation formalism has a syntactic structure which can be characterized in terms of recursive rules, and if the meaning of each construct is a function of the meaning of its components, then each construct can be interpreted in a set theoretic model by singling out semantic rules parallel to the syntactic rules, which interpret the complex expressions starting from the interpretation of primitive expressions. Model theoretic semantics has given good results for many formalisms with a compositional structure (examples are the development of Kripke models for intensional logics, Montague semantics, denotational semantics for programming languages). So it may be considered extensible to any well defined representation language of this type.

The approaches covered by point 2. are symbolic, in the sense that every element in the representation system is a symbol, that is an entity with a meaning ascribed, and local (i.e. non-distributed) in the sense that each representation is "localized" in a specific element of the system. However, the composition principle does not apply in this case: there are no syntactic rules which can generate complex representations, each symbol is, in a certain sense, atomic. These approaches cannot be given a model theoretic semantics except in a fairly banal manner. Without compositional syntactic rules it is impossible to define parallel semantic rules and every symbol in the representation system should be interpreted separately in the model. "Local connectionist" models, where each conceptual entity is represented by a distinct unit in the network, are of this type. It is this lack of compositionality that Fodor and Pylyshyn identify as the characteristic that distinguishes connectionist models from classical models [Fodor and Pylyshyn, 1988, pp. 15-19].

In a distributed model "each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" [Hinton *et al*., 1986]. Point 3. comprises connectionist models distributed on micro-features, where each node of the network is a symbolic entity, in that it represents a micro-feature and it possesses a distinct semantic interpretation. The "high level" concepts however are not represented by units, rather they are distributed on the network as activation patterns. These models are subconceptual (in the sense that the representation of a concept may not be localized in a distinct element of the representation system) but not subsymbolic (the units representing micro-features are given a meaning).

We shall discuss point 4. later. Point 5. comprises the "neural" connectionist models. In these the representations are distributed on non-symbolic units that are assumed to have exact anatomic and neurophysiological correspondences. Fodor and Pylyshyn [1988] consider this type of approach not relevant at the cognitive level as it involves lower levels of mental organisation. Moreover, although this type of model can be theorized, it is difficult to realize in view of the empirical evidence available.

Finally, as regards point 4., in this approach the subsymbolic units are not considered as having meaning (for

example in the sense of representing micro-features), nor as corresponding to physical entities identifiable at the neurological level. We maintain that they should rather be placed among the "theoretical constructs" of a theory of the cognitive level, which allow the deduction (in the sense of computation) of cognitive behaviours that cannot be otherwise modelled. This type of model is not considered by Fodor and Pylyshyn [1988], who see the only possible alternatives for a distributed representation in points 3. and 5. However, 4. is the position held by Smolensky [1988] (although his examples are often more suitably identified in point 3. - see for example the distributed representation of different kinds of room in Rumelhart et al. [1986]). An analogous position is that of Hofstadter [1985], from whom Smolensky says he drew some of his main ideas. Although the type of computation used in the subsymbolic paradigm is "inspired" by neural computation, Smolensky states that: "the fundamental level of the subsymbolic paradigm, the subconceptual level, lies between the neural and the conceptual levels" [Smolensky, 1988]. Further, "the subconceptual level seems at present rather close to the conceptual level, while we have little grounds for believing to be close to the neural level" (ibid.). Again: "subsymbolic models should not be viewed as neural models" (ibid.).

To sum up, the approaches at points 1., 2. and 3. can be characterized as *symbolic* in a wide sense, as they contrast with the *subsymbolic* approaches at points 4. and 5. (see Figure 1). Approaches 3., 4. and 5. are to some extent *distributed* in contrast with the *local* approaches of points 1. and 2. Lastly, while the approaches of 1. to 4. are at a *cognitive* level, approach 5. involves *subcognitive* levels (anatomic).

2 The need for a cognitive subsymbolic level

While recognizing the importance of compositional symbolic models of cognitive activities (if only because non-compositional models have not yet succeeded in offering a plausible alternative in many cases), there are problems which seem to suggest that there are valid theoretical reasons for hypothesizing a "cognitive subsymbolic" level.

Reference is the one case in which a subsymbolic computation seems necessary also within a compositional symbolic paradigm. We do not subscribe to Fodor's solipsistic assumptions [Fodor, 1980], that functional relations relevant to a computational theory of the mind are exclusively those between symbols, and that all semantic concepts, including reference, are not relevant to such a theory. We would tend rather to agree with Harman's statement that "of primary importance are functional relations to the external world in connection with perception, on the one hand, and action, on the other" [Harman, 1987, p. 67]. Harman does not specify such relations. Sloman [Sloman and Cohen, 1986] also expresses the need for links in order to anchor reference to symbols. In this paper we suggest that such links or relations could be envisaged in a subsymbolic framework.

We have seen how the semantics of first type representation systems can be characterized using the tools of model theory. Model theoretic semantics does not claim to be an empirically adequate characterization of reference as a cognitive process. However, apart from its empirical inadequacy, it contains certain *a priori* limits. The *empirical* inadequacy of model theoretic semantics is related to the fact that a human does not establish the reference of complex symbolic expressions (consider the truth value of modal statements) following the rules of model theoretic semantics (in the case of natural language and presumably in the case of expressions of the language of thought). By *a priori* limits we mean a type of limit that would also be valid for a hypothetical "non-anthropomorphic" cognitive entity that uses the rules of model theoretic semantics.

The analysis of such a priori limits could clarify the problems that arise in a symbolic representation paradigm in elaborating a model of reference from a cognitive point of view. It is well known that no formal symbol system can univocally determine its own model. In a logical system, sets of meaning postulates can be used to limit the set of admissible models, but they are not sufficient to fix a particular model. In model theoretic semantics the link between a system of symbols and the objects whereon the symbols are interpreted is established via an *interpretation function*. In a logical language, for example, such a function interprets the individual constants on objects (of the appropriate type) in the domain, the one argument predicate letters are interpreted on subsets of the domain and so on. The interpretation function calculates the value of complex expressions starting from the value of primitive symbols. However in this last case the interpretation function is taken as given and, in any case, characterizable only in a purely extensional manner, like a table which associates, for example, to each predicate symbol, a subset in the domain. This, from a cognitive viewpoint, is not sufficient since an adequate representation of conceptual entities must take into account the manner in which the reference is established, that is, how the interpretation function is calculated. This function (and its inverse) must be (partially) computable, and it must be (in a general sense) possible to know "how they are made", that is, it must be possible to identify (in the wide sense, and perhaps in the connectionist sense) an algorithm which calculates them. As we are dealing with functions which map symbols on objects in the world and viceversa, they cannot be defined exclusively in terms of other symbols as this would imply an infinite regression. The interpretation function must therefore be calculated, at least for some symbols in the system (the "primitives") by some subsymbolic device.

Figure 2 shows the drawing of an over-simplified hypothetical model of how the reference of symbols directly connected with sensorial inputs could be computed. The input units are connected in a causal, non-symbolic, man-

ner with the world. None of these units is a representation of anything in particular. The internal units are, in turn, non-symbolic entities and particular activation patterns on them correspond to expressions of the representation formalism at the symbolic level. The lack of a subsymbolic layer of this type would mean postulating the existence of symbols which directly "grasp" their referents, thus falling into the category of "magical" reference theories which are discussed below.

Fodor and Pylyshyn state "there are, no doubt, cases where special empirical considerations suggest detailed structure/function correspondences between different levels of system's organization. For example, the input to the most peripheral stages of vision ... at these stages it is reasonable to expect an anatomically distributed structure to be reflected by a distributed functional architecture" [Fodor and Pylyshyn, 1988, p. 63]. From our point of view however the problem does not lie in the fact that for lower level cognitive functions a greater structure/function correspondence is necessary, and therefore a functional structure corresponding more closely to anatomic structure. Strictly speaking it is not even necessary for subsymbolic computation to be in any sense "neural" or inspired to the anatomic structure of the nervous system. The problem is a priori and not empirical (and so, much less does it originate from considerations of an anatomic nature): an understanding of the relation of reference of a system of mental symbols cannot exclusively make recourse to other symbols as this would be simply shifting the problem. On the other hand, no symbol in the sense of syntactic object has reference as a result of some intrinsic quality, since there would be a necessary connection between symbols and referents, and mental symbols do not enjoy any "special status" in this regard. According to Putnam, "what is important to realize is that what goes for physical pictures also goes for mental images, and for mental representations in general; mental representations no more have a necessary connection with what they represent than physical representations do. The contrary supposition is a survival of magical thinking" [Putnam, 1981, p. 3]. To explain reference the symbolic level needs to be abandoned and a subsymbolic level as theorized by Smolensky could be adopted to model that non-symbolic activity of the mind which arranges that symbols refer to something. (Note that Putnam's statement, which is inspired by Wittgenstein's criticism of the concept of mental image [Wittgenstein, 1958], likens mental images to other types of mental representation: from the point of view of reference, symbol systems and mental images both fail for the same reasons.)

It is obvious that, in the real world, reference, in symbol systems such as natural language, is an extremely complex phenomenon: symbols use many "hooks" by which they grasp reality. Analogously, we can hypothesize that also the reference of mental symbols comes about by means of different modalities (sensorial, motorial, etc.) variously

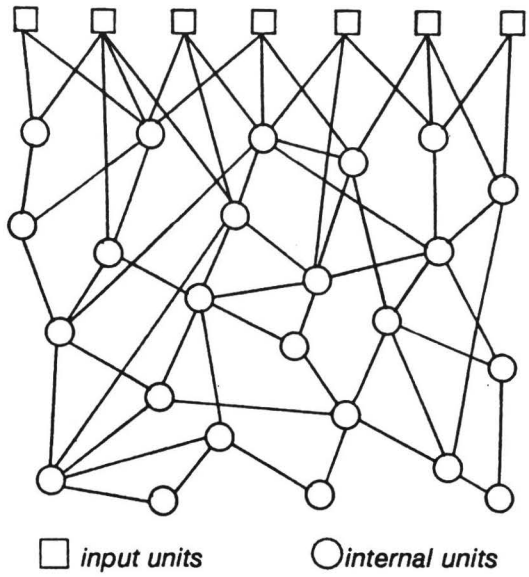

Figure 2

interrelating, and depends to a large extent on components of a cultural and therefore of a partly symbolic nature. However, we consider that recourse to a subsymbolic cognitive activity is inescapable.

We do not intend to adopt a "referentialist" position, which reduces the problem of meaning to that of establishing the reference of terms denoting material things. However, there is a real problem of reference and it is difficult to imagine meaning without any causal relation with the world.

The situation is even more complex if the assumptions of "ingenuous realism" are not accepted, and a position, which holds that the world is not intrinsically organized according to a fixed ontology and completely independent of the cognitive activities of the observer, is assumed. (A similar position is that of Lakoff [1986], which he calls *experientialist* in contrast with *objectivist* positions.) In this case the problem is not only to recognize and label ontological ly given entities, but also to impose a categorization on reality. Such categorization must necessarily include a non symbolic component: "before" the categorization is carried out there is no sense in using symbols as there is nothing they can refer to.

Reference is linked to sensorial and motorial activities not only in the obvious sense that many referents are perceived as objects of the senses, but also because a symbolic and *non symbolic* interaction with other members of the community of speakers is fundamental both in learning how to use a system of symbols and in the reference of "abstract" terms. Sensorial and motorial aspects are the basis of such interaction. The idea that an intelligent system can have symbols which *refer* to numbers, laws or feelings (whatever *refer* may mean in such a case) is unconceivable without tools which also allow a non-sym-

bolic interaction with the world and with other intelligent systems. A system closed to the world can be only a formal manipulator of tokens, and the fact that in such systems certain tokens or sets of tokens are readable as symbols depend exclusively on the interpretation of an observer outside that system.

Denying that a subsymbolic activity of the mind is relevant at the cognitive level means, either supposing that mental symbols have a "magical" talent for reference or that the problem of reference has no cognitive relevance and relegating it to a lower level of organisation. To assume a "subsymbolic cognitive level", even if at present completely hypothetical, could be the way to tackle the problem of reference without introducing "magical virtues" of the mind, maintaining a "functionalist" vision, in the wider sense, which does not make specific reference to the physical properties of neural hardware.

Acknowledgements

We thank Diego Marconi, Luisa Montecucco and Carlo Penco for their helpful and insightful comments, and Edward Farrants for preparing the English version.

References

[Brachman and Levesque, 1985] Ronald J. Brachman and Hector J. Levesque (eds.). *Readings in knowledge representation.* Morgan Kaufmann, Los Altos, Ca., 1985

[Fodor, 1980] Jerry A. Fodor. Methodological solipsism as a research strategy in psychology. *Behav. Brain Sciences*, 3 : 63-73, 1980

[Fodor and Pylyshyn, 1988] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28 : 3-71, 1988

[Harman, 1987] Gilbert Harman. (Nonsolipsistic) conceptual role semantics. In Ernest Le Pore (ed.), *New directions in semantics*. Academic Press, New York, 1987

[Hinton *et al.*, 1986] Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. *Distributed representations*. In [McClelland *et al.*, 1986]

[Hofstadter, 1985] Douglas R. Hofstadter. Waking up from the Boolean dream, or, subcognition as computation. In Douglas R. Hofstadter, *Metamagical themas*, pp. 631-665. Basic Books

[Lakoff, 1986] George Lakoff. Cognitive semantics. *Versus*, 44-45 : 119-154, 1986

[McClelland *at al.*, 1986] James L. McClelland, David E. Rumelhart and the PDP Research Group. *Parallel distributed processing.* MIT Press, Cambridge, Mass., 1986

[Newell, 1980] Allen Newell. Physical symbol systems. *Cognitive science*, 4 : 135-183, 1980

[Putnam, 1981] Hilary Putnam. *Reason, truth and history.* Cambridge University Press, Cambridge, 1981

[Rumelhart *et al.*, 1986] David E. Rumelhart, Paul Smolensky, James L. McClelland, and Geoffrey E. Hinton. *Schemata and sequential thought processes in PDP models.* In [McClelland *et al.*, 1986]

[Sloman and Cohen, 1986] Aaron Sloman and Jonathan Cohen. What sorts of machines can understand the symbols they use? *Proc. Arist. Soc.*, Suppl., 60 : 61-95, 1986

[Smith, 1982] Brian C. Smith. *Prologue to "Reflection and semantics in a procedural language"*. In [Brachman and Levesque, 1985]

[Smolensky, 1988] Paul Smolensky. On the proper treatment of connectionism. *Behav. Brain Sciences*, 1988

[Wittgenstein, 1958] Ludwig Wittgenstein. *The blue and brown books*. Basil Blackwell