On the Concept of Concept in the Context of Autonomous Agents

Paul Davidsson

Department of Computer Science, Lund University
Box 118, S–221 00 Lund, Sweden
fax: +46 46 131021
e-mail: Paul.Davidsson@dna.lth.se

Abstract

This paper deals with some fundamental questions regarding the concept of concept in the context of autonomous agents. The most basic of these is defining what it actually means for someone to have a concept. Rather than trying to state a number of conditions that should be satisfied in order to have the concept, it is concluded that having a concept is a matter of degree, which can be defined in terms of the functions the concept can serve. The more functions it can serve and the better it can serve these functions, the higher is the degree to which one has the concept. Moreover, the distinction between entity and dispositional theories of concepts is discussed, and it is concluded that they are complementary in that both perspectives are necessary to get a full picture of the concept of concepts. A conceptualistic entity theory and a dispositional theory based on which functions the concept should be able to serve are then put forward and discussed in a representational framework that supports these functions. Furthermore, we discuss the meaning of concepts, i.e., the problem of interpreting the symbols used to designate concepts, and give some arguments of why an autonomous agent should have the ability to interpret (some of) its own descriptions. We examine the work carried out within the field of logical semantics, and conclude that since traditional truth conditional semantics requires a human who grounds the meaning of elementary symbols, i.e., one who assigns objects and sets of objects to constants and predicates, this approach is not appropriate. Instead, a subjective intensionalistic approach based on the grounding of symbols is suggested, which is more in line with the verificationist and procedural approaches to semantics. Finally, we show that theories of meaning are closely linked with views on universals.

Keywords: concept, autonomous agent, epistemology, knowledge representation, semantics

1 Introduction

Although concepts "... are assumed to be the basic constituents of thought and belief" [Smi89], it is a fact that: "Evidence that the notion of concept is understudied in AI is easy to find" [Kir91]. While this oversight is serious for AI in general, it is even more so when it comes to the development of autonomous agents.[1]

[1] By autonomous agent we will here mean a system capable of interacting independently and effectively with its environment via its own sensors and effectors in order to accomplish some given or self-generated task(s). Thus, humans and most animals can in this sense also be regarded as autonomous agents. In the following, however, we will by autonomous agents refer to artificial ones.

To begin with, let us make explicit the relation between traditional AI systems, or even computer systems in general, and autonomous agents in terms of how they interact with the environment. Traditional AI-systems typically need a human operator who observes the environment (i.e., the problem) and describes the relevant parts of it to the computer. The results of the computer's computations are interpreted by the operator who then performs the required actions. (See Figure 1.) Thus,

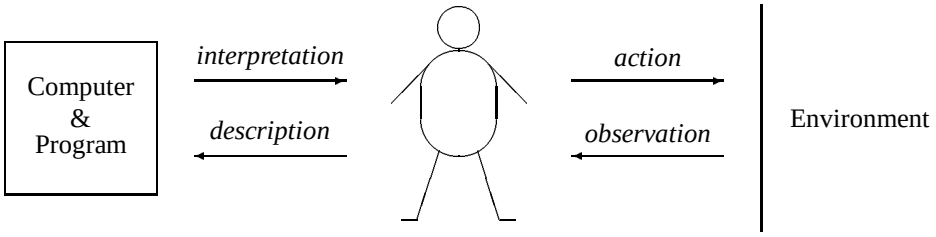

Figure 1: Traditional AI-system.

the system can be seen as a tool, extending some of the (mental) capabilities of the operator, but is useless on its own as it is dependent on the operator's basic (mental and physical) capabilities.

An autonomous agent, on the other hand, must observe the environment by itself and turn these observations into descriptions for further computations. Moreover, it must, also by itself, interpret the results of its computations and then perform the appropriate actions. (See Figure 2.)

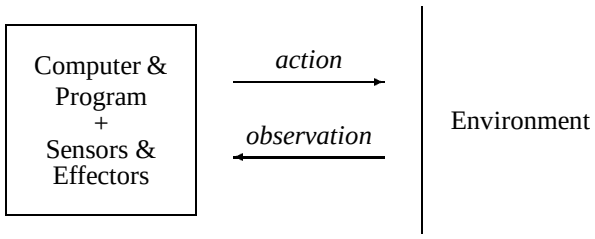

Figure 2: Autonomous agent.

Thus, the interpretation of the symbols used to represent knowledge about the environment in traditional systems is done solely by humans. The systems themselves do not need to know what the symbols stand for. Specifically, this has led to that concepts often are represented only by a name (designator), and relying instead on the understanding of the concept in the mind of the operator. However, this simple notion of concepts is not sufficient for an autonomous agent that must know what the symbols stand for, in order to be able to perform the appropriate actions. Consequently, it needs a richer notion of concepts, probably one that is more like that of humans.

2 What does it mean to have a concept?

Although Kirsh [Kir91] has pointed out that in AI "...the worry about what it is to have a concept is seldom articulated", there have been made some attempts in related fields to state explicitly what it means to have a concept. For instance, in philosophy the following suggestions (among others) has been put forward (cf. Heath [Hea67]). To have a concept "x" is:

- to know the meaning of the word 'x'
- to be able to pick out or recognize a presented x, or to be able to think of x (or x's) when they are not present
- to know the nature of x, to have grasped or apprehended the properties which characterize x's and make them what they are.

As we can notice, these conditions are rather vague, especially if we try to apply them to artificial agents. Several questions remain open: What is it for a computer system to know the meaning of (this will be discussed in detail below), to think of, and to apprehend something? What is the nature of a concept? It is only the first part of the second condition, to be able to recognize instances of the class "x", that seems reasonably straightforward.

A proposal that is easier to comprehend, comes from Smith [Smi88], a cognitive psychologist, who suggests that: "To have a concept of *X* is to know something about the properties of *X's* instances." However, it seems that this condition is too weak and underspecified; it seems not to capture the full meaning of "having a concept". Kirsh [Kir91], on the other hand, gives the following view (in AI terms) on the problem:

> We cannot just assume that a machine which has a structure in memory that corresponds in name to a structure in the designer's conceptualization is sufficient for grasping the concept. The structure must play a role in a network of abilities; it must confer on the agent certain causal powers [Bir91]. Some of these powers involve reasoning: being able to use the structure *appropriately* in deduction, induction and perhaps abduction. But other powers involve perception and action—hooking up the structure via causal mechanisms to the outside world. (p.10)

From this, and the above discussion, it seems appropriate to draw two conclusions.

1. It is perhaps not adequate to treat "having a concept" as a two-valued predicate (i.e., either you do have the concept or you do not). Instead, we should think in terms of a continuum of degrees of having the concept.
2. Rather than trying to state a number of conditions that should be satisfied for having the concept, it seems that a more fruitful approach would be to ask which *functions*, or purposes, (cf. causal powers) a concept should have.

Thus, the more functions a particular agent's representation of a particular concept can serve and the better it can serve these functions, the higher is the degree to which the agent has the concept. The functions of concepts will be discussed in Section 4.

We will in the following concentrate on concepts associated with classes of concrete objects (e.g., "chair" and "dog"). Thus, we will not discuss concepts of singular objects or more abstract concepts (e.g., concepts associated with events and situations, and Kant's a priori categories). Nor will we discuss concepts referring to *properties* of objects (e.g., red and large). In my opinion, a considerable source of confusion is the fact that all these kinds of concepts are discussed simultaneously (in contrast to treating them separately) in the philosophical literature.

2.1 Entity Theories versus Dispositional Theories

An examination of the philosophical literature on concepts reveals that it is possible to distinguish two perspectives on concepts: (1) that they should be seen as entities and (2) that they are just dispositions or capacities. Entity theories identify concepts with individual entities of one kind or another, for instance, "subsistent" word meanings, abstract ideas in the mind, and external unitary forms (cf. Plato). Dispositional theories, on the other hand, suggest that concepts are essentially habits or capacities for: the right use of words, the production of suitable conditioned responses, recognition, or image formation (cf. Hume, Kant).

It is the author's opinion that these views are complementary in the sense that both are necessary to get a complete picture of concepts in the context of autonomous agents. We need an entity theory since a concept, in some way or another, has to be represented in the memory of an agent, i.e., of a computer.[2] A dispositional theory, on the other hand, is needed because in order to decide how to implement concepts in an autonomous agent, we must know what they should be used for.

Closely associated with entity theories is the old philosophical problem of *universals*, which will be discussed in the next section. In the section after that, we will sketch a dispositional theory of concepts in terms of which functions they should serve.

3 The Problem of Universals

When discussing entity theories of concepts it is important to clarify which fundamental view one has on universals. There are basically two camps: realists and non-realists. Realists, such as Plato, believe that universals are non-mental, mind-independent entities that exist in themselves. Non-realists, on the other hand, argue that universals are mental, mind-dependent entities that would not exist if there were no minds. The most common non-realist theory is *conceptualism*, which was suggested by the classical British empiricists (e.g., Locke and Berkeley). According to conceptualism, there is some general mental representation (i.e., concept) that mediates between a word and what that word stands for. A more radical non-realistic approach is *nominalism*, suggested by Hobbes among others, which argues that not even concepts (general concepts) are necessary, i.e., only words are general.

As may have been understood by earlier statements, we are here adopting a non-realist stance. Thus, rather than being a priori entities, it is supposed that categories are the inventions, or constructions, of an agent or a collective of agents, used to structure the environment in order to facilitate cognition. Examples of collectively formed categories are "chair" and "ostrich". More personal categories invented by a particular individual are, for example, "the-things-that-are-mine" and "articles-relevant-for-my-thesis".[3] Moreover, since we also assume that the agent has a structure (i.e., a concept) in its mind that represents a category, we can classify our view as conceptualistic.

In short, we suggest (as an entity theory) that a concept is an internal representation of a class, or category, of external objects. This category is not an objective, a priori entity, but a construction of some agent(s) in the domain.

---

[2] We are here assuming a symbolic approach, whether this representation must be regarded as an entity also in a connectionist system is, however, unclear.

[3] In the field of Machine Learning, both the learning of concepts representing categories invented by humans (i.e., learning from examples) and categories formed by the learning system itself (i.e., learning by observation, or concept formation) have been studied.

4 The Functions of Concepts

To begin with, we can point out that concepts seem to be the very stuff on which reasoning and other cognitive processes are based. Actually, it is difficult to think of a mental activity that does not make use of concepts in one way or another. However, it is possible to distinguish several functions of human concepts, some of them are:

- stability functions
- cognitive economical functions
- linguistic functions
- communicative functions
- metaphysical functions
- epistemological functions
- inferential functions.

This list is inspired by different philosophers and cognitive scientists, in particular by Rey [Rey83] and Smith [Smi88]. However, we will not always use the terms in exactly the same ways as they do.

Concepts give our world *stability* in the sense that we can compare the present situation with similar past experiences. For instance, when confronted with a chair, we can compare this situation with other situations where we have encountered chairs. Actually, there are two types of stability functions, intrapersonal and interpersonal. Intrapersonal stability is the basis for comparisons of cognitive states within an agent, whereas interpersonal stability is the basis for comparisons of cognitive states between agents.

By partitioning the set of objects in the world into classes, in contrast to always treating each individual entity separately, we decrease the amount of information we must perceive, learn, remember, communicate and reason about. In this sense we can say that classes, and thus concepts, promote *cognitive economy*. For instance, by having one representation of the class "chair" instead of having a representation for every chair we have ever experienced, we do not have to remember that the chair we saw in the furniture-shop yesterday can be used to rest on.

The *linguistic function* is mainly providing semantics for linguistic entities (words), so that they can be translated and synonymy relations be revealed. For instance, the fact that the English word "chair" and the Swedish word "stol" have the same meaning enables us to translate "chair" into "stol" and vice versa. Furthermore, it seems that it is the linguistic function together with the interpersonal stability function that makes it possible for us to communicate (by using a language). Thus, the *communicative function* is reducible to the linguistic and the interpersonal stability functions.

In philosophy, metaphysics deals with issues concerning how the world is, while epistemology deals with issues concerning *how we know* (believe, infer) how the world is. Thus, we might say that the *metaphysical functions* of a concept are those that determine what makes an entity an instance of a particular class. For example, we can say that something actually is a chair if it has been made with the purpose of seating one person (or something like that).[4] The *epistemological functions* then, are those that determine how we decide whether the entity is an instance of a particular class. For instance, we recognize a chair by size, material, form, and so on. A better example for illustrating

[4] We use the word "metaphysic" in a more pragmatic way than is common in philosophy. In our notion that which makes an entity an instance of a particular class is decided by some kind of consensus amongst the agents in the domain.

this distinction is the class "gold". Something is actually a piece of gold if it has a particular atomic structure. However, when we recognize a piece of gold, we use other features such as: color, weight, and so on.[5] We should note that both these functions are related to classification: the metaphysical considers what actually makes an entity an instance of a particular class, whereas the epistemological considers how an agent decides whether the entity is of a particular class.

Finally, concepts allow us to *infer* non-perceptual information from the perceptual information we get from perceiving an entity, and to make predictions concerning it. In this sense, we can say that concepts enable us to go beyond the information given. For instance, by perceptually recognizing a chair we can infer that it can be used to rest on, or by recognizing a scorpion we can infer that it is able to hurt us. This is maybe the most powerful function of concepts; it emphasizes the role of concepts as the central element of cognition. As Smith [Smi88] writes: "Concepts are our means of linking perceptual and non-perceptual information ... they serve as entry points into our knowledge stores and provide us with expectations that we can use to guide our actions." In addition to prediction, concepts allow us to explain relationships, situations, and events.

4.1 The Function of Concepts in AI

The most advanced notion of concepts in AI is perhaps the one used in the field of Machine Learning (ML). A typical definition of concepts in this field is provided by Rendell [Ren86]: "The term concept is usually meant as a description of a class (its intension)." However, a description of this kind is often limited to describing how to recognize instances of the class. Thus, such representations are learned for a certain purpose; they are meant to serve a certain function. Namely, to discriminate between members and non-members of the class,[6] which corresponds to the classification tasks mentioned in the last section.

It should be stressed that this narrow view of concepts does not cause any problems in applications where the learning system is used as a tool to improve human performance. Because in this case, discriminating between members and non-members of a class is often exactly what we want the system to do — it might be precisely the function we want the learned concept to serve. The other functions that concepts normally serve are taken care of by the human operator. In an autonomous agent setting, on the other hand, there is no such operator available. Thus, a more powerful way of representing classes, able to support all the desired functions is needed. In fact, a conceptual framework for doing this has already been presented (cf. Davidsson [Dav93a]). In this framework a concept is represented by a composite description consisting of several components. The main idea is that since a single description is not likely to be able to provide all the functions desired, one should have different kinds of representations for different purposes. For example, one should normally not use the same concept description for both perceptual categorization and high-level reasoning.

The most important parts of the composite description are the following: the *internal designator*, which is the name (symbol) used by the agent to refer to the category, the *epistemological representation*, which is used to recognize instances of the category[7], and the *inferential representation*, which is a collection of "encyclopedic" knowledge about the category and its members, one that can be used to infer non-perceptual information and to make predictions. Thus, there is an obvious mapping between these parts and the functions they are supposed to serve: the internal designator supports the

[5] For human concepts this distinction is maybe not as clean-cut and unproblematic as described here (cf. Lakoff [Lak87]) but nevertheless it suits our purposes very well.

[6] Or, to discriminate between members of the class and members of other classes (cf. discriminative versus characteristic descriptions).

[7] Representations corresponding to the epistemological representation are often referred to as object models in the field of computer vision and concept descriptions in the field of machine learning.

intrapersonal stability, the epistemological representation supports the epistemological function, and the inferential representation supports the inferential function.

In addition, there are optional parts of the composite description such as: the *external designator*, necessary for communicating agents in multi-agent systems, which supports both interpersonal stability and the linguistic functions, and the *metaphysical representation*, perhaps not necessary for any agent, which supports the metaphysical function.

The category "chair" can be used to illustrate the idea of the composite description. The concept's English name "chair" could serve as the internal designator (but also as the external designator) and some sort of 3-D object model of a typical chair as the epistemological representation. The encyclopedic knowledge in the inferential representation can include such facts as: that chairs can be used to sit on, that they are often made of wood, and the like. Note that this is only a general framework in the sense that it does not specify how the different components should be represented.

We also should note how the idea of the composite description stresses the possibility of having a concept in different degrees. Depending on how many parts of the description the agent has, it could be said to have the concept to a lesser or larger degree. For instance, it could only have the external designator (i.e., the agent only knows the word commonly used to refer to the category), or the epistemological representation (i.e., the agent is able recognize instances of the category, but does not know anything about the category explicitly). Another factor that determines the degree to which the agent has the concept is how well developed the representations are (e.g., the amount of encyclopedic knowledge contained in the inferential representation).

5 The Meaning of (Symbols Designating) Concepts

In Section 2 we saw that one suggested definition of "having a concept" was to know the meaning of the word (i.e., symbol) that designates the concept. Although this ability to interpret symbols was not explicitly discussed in our survey of the functions of concepts, it seems as a fundamental ability that deserves some discussion. Moreover, it has been argued that the ability to interpret and reason *about* its own descriptions is desirable, perhaps necessary, if we want to create a more powerful kind of agents than the presently available (cf. Astor, Davidsson, and Ekdahl [ADEG90, DAE94, EAD95]).

Since the interpretation of descriptions in formal languages has been studied within the field of logical semantics, it seems as a good idea to closer examine the work in this field (especially if we believe that the representations of an agent are best expressed in a language similar to predicate logic). In particular, we will investigate whether it is possible to program an agent to autonomously interpret symbols according to the principles of logical semantics (a task for which these theories were not originally intended).

5.1 Logical Semantics

For logicians, semantics is *truth conditional*, i.e., to know the meaning of a logical description, or formula, is to know what the world have to be like for it to be true. Thus, to give the meaning of a description is to specify its truth-conditions in terms of necessary and sufficient conditions. For instance,

> '$\forall x\, P x$' is true if and only if every object in the domain has the property denoted by P.

As Tarski [Tar56] has pointed out, the truth-conditions of a description must be specified in a *meta-language* to the language in which the description is formulated, i.e., the *object language*. In the example above, the object language is first-order predicate logic whereas the meta-language is ordinary English.

However, this seems not to get us very far. As Baker and Hacker [BH84] have pointed out, we may ask ourselves: "...the question whether metalinguistic equivalences on Tarski's model actually connect expressions with reality or whether, on the contrary, they merely constitute a translation manual for rendering expressions in the object language into other expressions in the metalanguage." (p.126) In other words, traditional truth conditional semantics attempts to *describe* the meaning of the symbols. However, this only leads to another set of symbols, which would likewise need to be interpreted, and in the end to an infinite regression.

Above we have mainly discussed interpretation of logical *sentences*. But, since the meaning of a sentence is typically defined as a function of the meanings of its constituents, the problem of interpreting constants and predicates has to be solved. To begin with, the semantic value of a constant is an individual, i.e., the *entity* it designates, not another symbol. Consequently, the interpretation of a constant is its assignment to some member in the domain of interpretation. However, as the semantic value of a predicate is a set of individuals, it is predicates that are most interesting to us (since a concept represents a class of objects, i.e., a set of individuals). For instance, it is possible to denote the concept "chair" by the predicate $Chair$.[8] The interpretation of a predicate is then its assignment to a set of members in the domain. Another way to express the difference between the interpretation of sentences and primitive symbols (such as constants and predicates) is that, whereas the meanings of sentences are systematically determined by their composition, the meanings of primitive symbols are arbitrary (cf. Haugeland [Hau85]).

How then, are these assignments actually carried out? This is often explained in terms of an abstract mathematical *model* of those entities in the world making up the semantic values of symbols in the description language. Formally, a model is an ordered pair: $<A, F>$ where A is a set of individuals, and F is a function which assigns semantic values of the appropriate kind to basic expressions. However, it seems that there is a tacit assumption that it is the logician himself who actually makes these assignments, much in the same way as the operator interprets the symbols (e.g., the output) of a traditional AI-system. Wittgenstein [Wit22] suggested that the process of analysis of sentences must ultimately terminate in truth-functions of elementary sentences, each of which is composed of symbols incapable of further analysis. Meaning is assigned to these "indefinables" by their direct association with simple objects through "elucidations". He concluded that these methods of projection cannot be described in language. Thus, the interpretation of symbols designating concepts seems to go beyond the scope of traditional truth conditional semantics. However, if we want to realize an autonomous agent capable of interpreting (some of) its own descriptions we have to implement these methods (at least partly) in some language.

As described earlier, to know the meaning of a logical formula according to truth-conditional semantics is to know the conditions under which the formula is true. But exactly what does it mean to know these conditions? We argued above that just the ability to explicitly *state* the conditions does not get us very far. Instead, we argued that one, in some way or another, must be able to *recognize* the situations when the conditions are true. In a similar vein, Dummett [Dum75, Dum76] has suggested a *verificationist semantics* where the meaning of a formula consists of those conditions that would *verify* the formula, i.e., specifications of the procedures that would verify the formula. In this case, the meaning of a formula becomes an epistemological concept rather than a metaphysical since this approach requires a set of procedures that when applied recognize whether the formula is true or not.[9] Related to this approach is the notion of *procedural semantics* put forward by Johnson-Laird

[8] In the framework of composite concepts outlined above, $Chair$ would correspond to the internal designator of the concept "chair".

[9] Related to this is ones conception of truth. Dummett [Dum78] makes a distinction between *realism*, which treats truth as an objective property of descriptions independent of human cognition, and *anti-realism*, which suggests that, at least, some fundamental propositions are made true by conditions associated to human recognition mechanisms. Whereas

[JL77, JL83] which suggests that the meaning of a symbol consists of a set of procedures that operate on it, for instance, computing its truth value.

5.2 Symbol Grounding

Harnad [Har90] has suggested an alternative view to the problem of giving meaning to symbols, which, in fact, is rather similar to Wittgenstein's (and our) position. According to him, the meaning of the system's symbols should be grounded in its ability to identify and manipulate the objects that they are interpretable as standing for. He proposes a hybrid system with both symbolic and connectionist components, stating: "In a pure symbolic model the crucial connection between the symbols and their referents is missing ..." (p.344)

He argues that three kinds of representations are necessary: *iconic representations*, which are the sensor projections of the perceived entities, *categorical representations*, which are "learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections", and *symbolic representations*, which consist of symbol strings describing category membership. Both the iconic and the categorical representations are assumed to be non-symbolic.

He concludes that a connectionist network is the most suitable for learning the invariant features underlying categorical representations[10] and thus for connecting names to the icons of the entities they stand for. The function of the network then is to pick out the objects to which the symbols refer. However, although it seems clear that a pure symbolic system does not suffice (since sensors do not provide symbolic representations), regarding connectionist networks alone as being capable of serving this function appears too limited.

As has been described in [Dav93b], the problem of symbol grounding becomes easier to resolve if it is viewed in terms of the general concept representation framework presented above. In this approach, it is essentially the vision system, through its use of epistemological representations that are parts of the same structure as the corresponding symbols, which permits grounding, or the connection between symbols (internal designators) and their referents (objects in the world), to be carried out. However, the main point to be made is that the epistemological representation does not have to be a connectionist network. Rather, it can be virtually any representation the perceptual system can successfully use to identify (i.e., categorize) objects.

Note that we are here discussing the meaning of symbols referring to (classes of) concrete objects, which are directly perceivable by the perceptual system. Symbols referring to more abstract entities, on the other hand, probably get their meaning more from internal structure than from external grounding (cf. Sloman [Slo86]).

5.3 The Meaning of Symbols in Present and Future Autonomous Agents

In present autonomous agents the programmer has typically tried to foresee all possible situations that the agent might find itself in, and to program his own grounding based on his own experiences (which probably are meaningless for the agent as it does not have the same kinds of experiences as humans, e.g., artificial agents do not have the same kind of sensors as humans). However, in order to cope with situations not anticipated by the programmer, the agent must by itself be able to make sense of the symbols used to represent its knowledge about its environment and of its problem solving capabilities (cf. [ADEG90, DAE94, EAD95]). Thus, as pointed out earlier, it must be able to

traditional truth conditional semantics assumes the former, we are here suggesting the latter as being more appropriate.

[10] Harnad seems here to assume that categorization always is based on invariant features, a position often referred to as the classical view. There are, however, strong empirical evidence suggesting that this position is not accurate (cf. Smith and Medin [SM81]).

interpret and reason about these symbols. In the last section we stressed that the agent must ground the primitive symbols by itself, most notably by constructing its own concepts.[11]

As a consequence, the agent will develop subjective and individual concepts.[12] Moreover, and in line with the view that meaning should be regarded as an epistemological rather than metaphysical concept, Dorffner and Prem [DP93] have pointed out that: "No objective world has to be assumed, except some causal dependencies between sensory signals and internal states and except for what is expressed in the meta-level representations that have to be part of the system." How much pre-programmed knowledge of this kind (cf. meta-level representations) an agent must be equipped with is, however, an open question. Our guess is that, at least, information about the structure of a concept, i.e., its division into the different components of the composite concept framework, must be "hard-wired" into the agent. (Moreover, some basic, or primitive, concepts seem necessary.)

Dorffner and Prem also note an important insight that follows from this, namely, that one problem with many of the proposed theories of semantics[13] is that they try to explain meaning as not being tied to individual agents. From the above discussion it seems clear that the agent itself cannot be ignored in a complete theory of meaning.

5.4 Theories of Meaning and Theories of Universals

We can now see that there is a deeper connection between theories of meaning and theories of universals. However, let us first of all make a distinction between the *extension* and the *intension* of a symbol. Whereas the extension of a symbol is those objects which it designates, the intension is often defined as those properties possessed by all and only the objects that the symbol designates.[14]

Traditional truth-conditional semantics is basically an extensional theory of meaning. Intensional theories of meaning, on the other hand, can be divided into objective intensionalism, which holds that meanings are objective intensions, and subjective intensionalism, which holds that meanings are subjective intensions. Thus, the position that has been sketched in this paper can be characterized as subjective intensionalism. We regard concepts as subjective intensions of predicates, i.e., individually formed descriptions of classes of objects. Moreover, we agree with Barwise and Etchemendy [BE89] (cf. their theory of situation semantics) in that predicates should be treated as primitive symbols, and not defined in terms of the set of objects they designate.

To sum up, we can express the connections between theories of meaning and theories of universals as follows: extensional theories of meaning are consistent with a nominalistic view on universals, objective intensionalism with realism, and subjective intensionalism with conceptualism.

6 Summary

We have dealt with some fundamental questions regarding the concept of concepts in the context of autonomous agents. The most basic one was that of defining what it actually means for someone to have a concept. As a result of our efforts we realized that having a concept is a matter of degree.

[11] For a comprehensive treatment of different aspects of concept acquisition by autonomous agents, see Davidsson [Dav94].

[12] However, it might be able to associate them with the external designator that other agents use to refer to the category of objects that the concepts represent (e.g., by supervised learning), which in the end will make the agent able to communicate with other agents.

[13] There are, however, newer approaches (other than those presented above) that may not be affected by this criticism. Devlin [Dev91], for instance, has suggested that the root of most of the problems with classical logic is that it is based on the concept of truth. Instead, he proposes a logic based on the concept of information.

[14] Also this definition seems to assume a classical view of concepts. However, we will here assume a more general notion of intension, i.e., one that is not tied to a particular view of concept.

Moreover, it was concluded that trying to state a number of conditions that should be satisfied for having the concept is very difficult. It seems that a more fruitful approach would be to ask which functions a concept should have — the more functions it can serve and the better it can serve these functions, the higher is the degree to which one has the concept.

Moreover, we identified a distinction between entity and dispositional theories of concepts, and suggested that they are complementary in that both perspectives are necessary to get a full picture of the concept of concepts. We put forward a conceptualistic entity theory and a dispositional theory based on which functions the concept should be able to serve. A representational framework that supports these functions was then described.

Finally, we discussed the meaning of concepts, or rather, the problem of interpreting the symbols used to designate concepts, and gave some arguments of why an autonomous agent should have this ability. We examined the work carried out within the field of logical semantics, and concluded that traditional truth conditional semantics requires a human (preferably a logician) who grounds the meaning of elementary symbols, i.e., one who assigns objects and sets of objects to constants and predicates. Instead, we suggested a subjective intensionalistic approach based on the grounding of symbols, which is more in line with verificationist and procedural semantics. We also showed that theories of meaning are closely linked with views on universals.

Acknowledgements

I wish to thank Eric Astor, Bertil Ekdahl, Peter Gärdenfors, Bengt Hansson and Olof Tilly for helpful comments and suggestions.

References

[ADEG90] E. Astor, P. Davidsson, B. Ekdahl, and R. Gustavsson. Anticipatory planning. Technical Report LU–CS–TR: 90–69, Dept. of Computer Science, Lund University, Lund, Sweden, 1990. Also in Advance Proceedings of the European Workshop on Planning, 1991.

[BE89] J. Barwise and J. Etchemendy. Model-theoretic semantics. In M.L. Posner, editor, *Foundations of Cognitive Science*, pages 207–243. MIT Press, 1989.

[BH84] G.P. Baker and P.M.S. Hacker. *Language, Sense & Nonsense*. Basil Blackwell, 1984.

[Bir91] L. Birnbaum. Rigor mortis: A response to Nilsson's "Logic and artificial intelligence". *Artificial Intelligence*, 47(1):57–77, 1991.

[DAE94] P. Davidsson, E. Astor, and B. Ekdahl. A framework for autonomous agents based on the concept of anticipatory systems. In R. Trappl, editor, *Cybernetics and Systems '94, Volume 2*, pages 1427–1434. World Scientific, 1994.

[Dav93a] P. Davidsson. A framework for organization and representation of concept knowledge in autonomous agents. In *Scandinavian Conference of Artificial Intelligence – 93*, pages 183–192. IOS Press, 1993.

[Dav93b] P. Davidsson. Toward a general solution to the symbol grounding problem: Combining machine learning and computer vision. In *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?, (FS–93–04)*, pages 157–161. AAAI Press, 1993.

[Dav94] P. Davidsson. Concepts and autonomous agents. LU–CS–TR: 94–124, Dept. of Computer Science, Lund University, Sweden, 1994.

[Dev91] K. Devlin. *Logic and Information*. Cambridge University Press, 1991.

[DP93] G. Dorffner and E. Prem. Connectionism, symbol grounding and autonomous agents. Technical Report No. 17, Austrian Research Institute for Artificial Intelligence, Vienna, Austria, 1993.

[Dum75] M.A.E. Dummett. What is a theory of meaning? In S. Guttenplan, editor, *Mind and Language*, pages 97–138. Clarendon Press, 1975.

[Dum76] M.A.E. Dummett. What is a theory of meaning? II. In Gareth Evans and John McDowell, editors, *Truth and Meaning: Essays in Semantics*, pages 67–137. Clarendon Press, 1976.

[Dum78] M.A.E. Dummett. *Truth and Other Enigmas*. Duckworth, 1978.

[EAD95] B. Ekdahl, E. Astor, and P. Davidsson. Towards anticipatory agents. In M. Wooldridge and N.R. Jennings, editors, *Intelligent Agents — Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence 890*, pages 191–202. Springer Verlag, 1995.

[Har90] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[Hau85] J. Haugeland. *Artificial Intelligence – The Very Idea*. MIT Press, 1985.

[Hea67] P.L. Heath. Concept. In P. Edwards, editor, *Encyclopedia of Philosophy*. Macmillan, 1967.

[JL77] P. Johnson-Laird. Procedural semantics. *Cognition*, 5:189–214, 1977.

[JL83] P. Johnson-Laird. *Mental Models*. Harvard University Press, 1983.

[Kir91] D. Kirsh. Foundations of AI: The big issues. *Artificial Intelligence*, 47(1):3–30, 1991.

[Lak87] G. Lakoff. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. The University of Chicago Press, 1987.

[Ren86] L. Rendell. A general framework for induction and a study of selective induction. *Machine Learning*, 1(2):177–226, 1986.

[Rey83] G. Rey. Concepts and stereotypes. *Cognition*, 15:237–262, 1983.

[Slo86] A. Sloman. Reference without causal links. In J.B.H. du Boulay, D. Hogg, and L. Steels, editors, *Advances in Artificial Intelligence – II*, pages 369–381. North Holland, 1986.

[SM81] E.E. Smith and D.L. Medin. *Categories and Concepts*. Harvard University Press, 1981.

[Smi88] E.E. Smith. Concepts and thought. In R.J. Sternberg and E.E. Smith, editors, *The Psychology of Human Thought*. Cambridge University Press, 1988.

[Smi89] E.E. Smith. Concepts and induction. In M.L. Posner, editor, *Foundations of Cognitive Science*, pages 501–526. MIT Press, 1989.

[Tar56] A. Tarski. *Logic, Semantics, Metamathematics*. Oxford, 1956. second edition, J. Corcoran (ed.), Hacklett, 1983.

[Wit22] L. Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, 1922.