Connectionism, Symbol Grounding, and Autonomous Agents

Georg Dorffner, Erich Prem
Austrian Research Institue for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
and
Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna
email: {georg, erich}@ai.univie.ac.at

Abstract

In this position paper we would like to lay out our view on the importance of grounding and situatedness for cognitive science. Furthermore we would like to suggest that both aspects become relevant almost automatically if one consequently pursues the original ideas from connectionism. Finally we discuss the relevance of grounding for theories of meaning and the possible contribution of symbol grounding for autonomous agents.

What is grounding and why is it useful?

Grounding of concepts and symbols is an important perspective for cognitive science. This perspective concentrates on two things: First it points out that in any cognitive agent there are important pathways between the sensors and effectors on one side, and the agent's concepts and symbols on the other side. Neglecting those pathways leads to neglecting the very essence of the concepts that make them part of the agent's mental processes. Secondly, it explains how concepts can be developed through interactive behavior in an environment and thus puts emphasis on the history and subjective context behind any conceptual scheme. Grounding in this sense is not limited to linking concepts to sensors and effectors, but also to any other internal mental state thus constituting the context for concept formation.

What the former aspect is concerned it is often argued that any classical symbol system in AI can be envisioned as being linked to sensors and effectors, and therefore any AI system is or can be grounded. What the second aspect is concerned, it is sometimes argued that if, for instance, someone copied all mental states from one agent into another one would get a functioning systems while being able to forget about any history or subjective experiences (Minsky, personal communication).

From a certain perspective, both points are true. Of course, when copying someone's brain into a functionally equivalent computer program or the like (let's suppose for the moment we really could) one could no longer really say that this new agent has had experiences which helped it to ground its symbols. This is not the point. What the symbol grounding problem suggests is that one cannot *merely* copy that agent's concepts and its use of symbols [1] into another agent without losing some or all of their functionality. In other words, in order to get the new agent to make use of the symbols in *exactly* the same way as the original one, one needs to also copy *all* of the original sensors plus the pathways between them and the concepts (with all their imprinted weights, to use connectionist terms). Experiences or interactions with the environment do not explain us how at any specific moment in time sensors are connected to symbols, but they explain us how the pathways were or could be established (other than through copying).

So what does that tell us for cognitive science or AI? It does not tell us that we could never ever achieve a fully artificially intelligent system without caring about an agent's experiences and the resulting grounding. However, at the moment we would not know of any method – other than resorting in large parts to trial and error – of how to build in detail those said pathways between the sensors and the concepts. What that story does tell us is that viewing concepts and symbols as "to be grounded" through mechanisms that are a result from the agent's ex-

[1] For the further discussion we want to make clear that we distinguish between concepts as mental states resulting from categorization and symbols as signs (or labels) that are made to refer to concepts by the agent.

periences will get us much farther in designing (and understanding) such artificial agents. Thus the symbol grounding problem does not point to a theoretical impossibility of building AI machines without an explicit account of grounding (implicitly, however, every successful machine will give such an account). This has already been pointed out by (Harnad, 1990). Instead it gives us directions toward plausible and efficient design of artificial cognitive systems which can use symbols. Those directions tell us that we should emphasize on agents that – like us – can see, hear and feel the world and act in it, and by this way acquire concepts and meaningful symbols. And it tells us that we should not approach symbols and conceptual knowledge divorced from any perceiving and acting agent, because if we do we will have the same problems as when trying to copy (only) the symbols from one agent into another (as suggested above).

Grounding as a logical consequence of radical connectionism

In this section we would like to discuss grounding in the framework of connectionist models.

Recently, we have argued for a "radical" version of connectionism to be embraced by those who are aiming for a truly alternative approach to the classical paradigm in cognitive science and artificial intelligence (Dorffner, 1991). By "radical" we understand the consequent emphasis of those aspects which are novel as compared to the symbolic, or cognitivist, approach. At the center of such aspects is *self-organization* in a neural network through learning rules, replacing explicit design or knowledge engineering. The basic idea is that a connectionist model can acquire "knowledge" through adaptive reactions to a stream of inputs, thus freeing the model's creator from the need for explicit design of that knowledge.

Many (tiny) models with at least one so-called hidden layer have given prove to this hypothesis that indeed a connectionist model can acquire some behavior that was not explicitly programmed. However, connectionist modelers have often been criticized for claiming just that but at the same time using explicitly designed representations on the model's inputs and outputs (such as in (Sejnowski and Rosenberg, 1987) or (Rumelhart, D.E. and McClelland, J.L., 1986)). Thus, a consequent extension of self-organizing models would be to place inputs and outputs where no such representation is needed. With respect to cognitive models this means the use of pure sensory input and motor output, since their representations are somewhat self-evident (and also more accessible in nature). We have called such representations "immediately grounded." By designing a cognitive model such that the only interfaces with its environment are of sensory or motor signal nature, in theory the need for explicit design of world knowledge can be erased. What remains are prewired architectural schemes so as to facilitate certain self-organizations over others – in a sense "innate" structures which in nature are created by evolution[2].

In this framework, any conceptual knowledge of a cognitive system develops through self-organization based on adaptive interaction with the environment. In other words, besides from what is given through the "meta-level" representations (the innate structures), all aspects about such purely connectionist concepts are acquired through the very system's own behavior in and interaction with its environment. Concepts are thus the "systems' own" and their meaning is no longer parasitic on the concepts of others (the system designer). Thus they are automatically grounded in Harnad's sense (Harnad, 1990).

In conclusion, we can say that if connectionists want to maximally distinguish their models from classical AI models through maximally emphasizing self-organization, as a logical consequence they will end up with grounded concepts (provided they do not deny the nature of concepts), and sub-sequently symbols.

Symbol grounding and meaning

We would like to suggest that when considering symbols in cognitive science we should look at "cognitively relevant" symbols which in our view are linguistic entities, that is, signs in the semiotic sense actively used by an agent to refer (see (Dorffner, 1992)). Thus a model of symbol grounding would contribute to a theory of linguistic meaning, at least on the lexical level. To look at what this contribution could be, we come back to a modeling framework like radical connectionism.

If designers wanted to explicitly program an artificial agent and its grounded concepts and symbols, they would tend to program their own grounding based on their own experiences. The limitations of such a design becomes clear when we are dealing with agents that cannot share all the same types of experiences, that is, viusal, acoustic, tactile, and other sensations, plus emotions, feelings, and motivations we humans have. We are very far from building ma-

[2]We are not denying that it might possibly turn out that finding such structures might even be a bigger practical problem than explicit design of agents.

chines that inlcude all of that. Artificial agents have their own, so far rather primitive ways of experiencing the world. There is no principal reason, however, why they should not be capable of acquiring conceptual schemes and the ability to use language. This means that we can learn to understand the principles of language by building artificial language users (Allen 1990) but we cannot expect such users to have exactly our kind of language. What we can expect is that their language shares enough common components with ours in order to enable us to communicate with them. This is underlined by the observation that we ususally design artificial agents with some subset of our ways of experiencing the world – i.e. simple visual or acoustic input, simple motor skills, etc. – and with a tendency to pursue a subset of our own goals.

The morale of this crucial observation is that we cannot expect artificial concept schemes to exactly mirror any of the human schemes we have access to. This makes self-organization – for instance, radical connectionism – as a basis for cognitive models very important. In this sense, radical connectionism can lead to models of agents that develop their subjective and individual concepts and subsequently learn their use of words we use in front of them (see also Brooks, this volume). 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. This is a view of language that strongly coincides with the ideas of radical constructivism (Maturana and Varela, 1980; Winograd and Flores, 1986). This approach, at first made necessary since explicit design appeared impossible, now at the same time obviously leads to a consistent and coherent theory of semantics. It is coherent as it puts the individual agent in the foreground and defines meaning with respect to its capabilities of responding to the world. Such agents learn the meaning of words by using them in certain situations (cf. late Wittgenstein). The meaning of an utterance depends on how each individual maps it onto its internal states and can thus differ for each individual. Universal (logical) principles can only enter the scene in two ways. Either they are the result from "meta-level" representations (the innate, globally pre-wired architecture) common to all individuals. Or they are an epiphenomenon resulting from the tendency of each individual to maximize some kind of success of its own linguistic behavior (e.g. the rate of "being understood"). In the latter case, the global principles are projected into a group of individuals' behavior by an observer through their own means of conceptualizing.

As a result, grounding and a framework like radical connectionism do not only add to existent theories of semantics but seem to that make a truly coherent one possible. Virtually all theories of semantics which wanted to explain meaning as not being tied to individual agents (see (Lyons, 1977) for an overview) have failed in one respect or another of achieving this.

Situatedness and Autonomous Agents

We shall now revisit our main arguments in the light of designing autonomous agents. The history of building robots which move around in an unknown and unstructured environment is known today as having produced a series of optimistic predictions many of which had to be withdrawn shortly after their announcement. Classical robotics has (metaphorically spoken) suffered from the lack of an adequate measurement device which – when sensing the physical world – would directly deliver clear symbolic descriptions of the objects encountered by the robot as well as of the situation in which the agent finds itself. Although research on the construction of such devices has only proceeded very slowly, research programs were undertaken as if such a meter would soon become available and the problem was left to the people in the physics or the AI (computer vision) department. It is only recently that robot researchers like Rod Brooks (Brooks, 1991) have abandoned the idea of waiting for such a meter to be developed. Instead, they have restricted themselves to simple sensors, which are continually used by the agent in order to meet the agent's need to act. In this way the artificial agents achieve high interaction dynamics with their environment.

Several reasons account for difficulties when trying to develop the above mentioned meter, among which we find unpredictability of basic measuring devices, large amounts of data, unknown features of the environment and difficulties in clearly describing necessary and sufficient constraints on the observed data. Whenever programmers tried to exactly specify what objects there are in the world and which set of physical characteristics they possess, they had to realize that they could only account for a very limited number of aspects of a small set of objects. This endeavor of an exact specification of the measurement data corresponds to the program of eliminative materialism, where all things in the world (objects and their features) are reduced to descriptions of their physical properties. Such an approach totally neglects the inherently functional aspect of all categorizations and ontologies of autonomous systems. Intelligent

autonomous agents – through their way of dealing with the world and dealing with it on the basis of adaptive control structures – strongly emphasize this functional perspective with respect to their own goals and own interaction space.

The history of robotics provides us with no evidence whatsoever that this sort of situatedness as a basis for grounding could be replaced by a careful designer. Again, the point here is not that there is an in-principle argument against overcoming this deficiency, the point *is* instead that the burden of proof has shifted to those who keep building their arguments on meters, storage, computation power, and design methods which do not exist today and may never be available in the required quality. On the contrary, there is evidence that situatedness and active interaction with the "objects" in the world constitutes the only way to bootstrap the construction of such a meter. This evidence consists in the fact that now for the first time, we are able to construct robots which interact with their unstructured and previously unknown environment in a highly dynamical and at the same time robust way. Both, interaction and robustness give rise to the phenomenon called *autonomy*. The possibility of "bootstrapping" the construction of a system with grounded symbols is based upon this autonomy. Designed with a set of simple needs and drives and equipped with several simple dynamically evaluable sensors an agent can be constructed which first builds a "conceptual" framework for objects, situations, and features important to the agent. Only based on such a conceptual schema the next step of communicating these context-dependent subjective concepts becomes possible. It is in this second step that we can try to let the symbols in again, that we can use some sort of a physically embodied sign for being a means of communicating between two agents (e.g. a written or spoken word or a blinking light).

The possible contribution of behavior-based robotics and symbol grounding research to AI would be to eventually come up with the above mentioned meter which delivers some symbolic descriptions of aspects of the world. But these symbols would then no longer be arbitrary in the strong sense as put forward by traditional AI. Not only would their meaning depend on the context of the utterance and on the whole situation including internal states of the agent, the social aspect of communicating to somebody else would also have to play a constitutive role in the selection of certain symbols for communication and in deciding what and when to communicate to whom. Some sort of symbolic communication is badly needed, especially if the internal structures of an agent are subsymbolic or continually trained and thus changing. How else should we tell an agent what to do? Consider the case of a simple find & fetch agent for use in my appartment[3]. Besides of the usual requirements like moving around, not getting stuck, etc., this system would need to know (i) what to fetch and (ii) where to bring it to. It would be extremly valuable to train the agent on the notion of my glasses and of my desk drawer. These symbols should be grounded in the agent to make it use a physical embodiment chosen by myself (e.g. the typed symbol "gls" or the spoken word "glasses") for an object experienced through its interactions with it (e.g. failure in fetching it, breaking it, distinguishing it from my sun glasses). To bring it to a desired location, the robot also needs to know labels of several places in the appartment. Work on this last aspect has already been successful within the field of behavior-based robotics (Mataric, 1992).

Grounded Symbols and Explanations

One reason for the many objections to the necessity of grounding is the fear of inherently relativistic structures connecting elements which are not simply representations of objects or features of objects "out there in the world". The programmed (designed) elements in the sort of radically connectionist networks which we have proposed here cannot easily be interpreted. Moreover, they are not in any sense stable. Consequently, when it comes to grounded symbols, it is feared that any such notion as a symbol's "Sinn" (Frege's use of the term) is abandoned and thus the secure basis of designing and understanding technical systems would be corrupted. This criticism, however, is based on the argument that any – in our sense – alternatively generated behavior is indeterministic and therefore impossible to explain and to maintain. An argument like this confuses the level on which both processes – explanation and maintanance – must proceed. In the conventional view it is believed that debugging means to directly edit the rules which generate the agent's behavior. As opposed to this, we suggest that it only means to give to the agent the possibility of adopting an adequate behavior, thus disregarding rules within the agent at all. It is the history of structural couplings to the world which must lead to the construction of the necessary concepts. If the appropriate set of concepts does not arise, then the construction process must be changed. Consequently, an explanation of an autonomous agent has to take the interaction space between two dynamical

[3]This example has been used by Tom Mitchell.

systems (the agent and its environment) into account. This can perhaps be done in a mathematically exact way using "complex systems talk" (Smithers, 1991) or from the standpoint of an external observer using "the intentional stance" (Dennett, 1987). However, we cannot expect such theories to have an inherently simple and neat structure of atomic elements which correspond to the atoms of our own descriptive ontology. What we *can* expect from grounded symbols is that they actually support finding explanations for the behavior of an adaptive autonomous intelligent agent. These symbols have the power of relating internal states of the agent to the environment and to the actions and the goals of the autonomous system itself. The point which we would like to make here is that such a system can be explained by using *its own* descriptions of the situations. We can construct predictive rules based on the system's own situation descriptions.

For a simple example consider our fetch & find robot again. It has now learned a set of symbols for objects and locations. But for some reason it cannot find its way out of the kitchen, instead it always bumps into the table. Asking the agent about its location, i.e. having the agent use one of its grounded location symbols, it could e.g. answer with the "bathroom"-symbol. In this case, we could easily discover that the system wrongly "believes" that it currently is in the bathroom, maybe because of similar sensor readings. (Consequently, it bumps into the table instead of leaving the bathroom through the door.)

We are assuming that we have already trained the system to use a set of symbols in the same way as we (the system observers) would use them. Of course, there is no guarantee that the system will always use our words correctly. This problem has been intensively discussed within a philosophical debate around the later works of Ludwig Wittgenstein (Wittgenstein, 1953). In Kripke's interpretation (Kripke, 1982), the main achievement of Wittgenstein was to have shown that there is no other guarantee for what somebody *means* with a word but some outer criterion to be found in his or her behavior. According to this position it is useless to search for facts within the speaker which could guarantee what is meant with a word. And it is also impossible to ensure that the word will always be used correctly in the future. With respect to our problem here this position holds that as long as the system correctly uses my *words* we have to say that the system means the same as we do. It is in this sense that we *understand* the situations in which the system is and that we can use grounded symbols for explaining the agent's actions.

Acknowledgments

The Austrian Research Institute for AI is sponsored by the Austrian Federal Ministery for Science and Research. We would like to thank Robert Trappl for having made this research possible.

References

Brooks, R. 1991. Intelligence without Reason. In *Int. Joint Conf. on AI (IJCAI 91)*, 1–1. Morgan Kaufman.

Dennett, D. 1987. *The Intentional Stance.* Cambridge, MA: MIT Press.

Dorffner, G. 1991. "Radical" Connectionism for Natural Language Processing. Working Notes of the AAAI Symposium on Connectionist Natural Language Processing.

Dorffner, G. 1992. On Redefining Symbols and Reuniting Connectionism with Cognitively Plausible Symbol Manipulations. Technical Report TR-92-13, Austrian Research Institute for AI.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42:335–346.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language.* Cambridge, MA: Harvard University Press.

Lyons, J. 1977. *Semantics.* Cambridge, UK: Cambridge University Press.

Mataric, M. 1992. Integration of Representation Into Goal-Driven Behavior-Based Robots. *IEEE Transactions on Robotics and Automation* 8(3):304–312.

Maturana, H. and Varela, F. 1980. *Autopoiesis and Cognition.* Reidel, Dordrecht.

Rumelhart, D.E. and McClelland, J.L., . 1986. On Learning the Past Tenses of English Verbs. In McClelland, J. L.*et al.*(Eds.), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Vol.II.* Cambridge, MA: MIT Press.

Sejnowski, T. and Rosenberg, C. 1987. Parallel Networks that Learn to Pronounce English Text. *Complex Systems* 1:145–168.

Smithers, T. 1991. Taking Eliminative Materialism Seriously: A Methodology for Autonomous Systems Research. In Varela, F. J.*et al.*(Eds.), *Towards a Practice of Autonomous Systems*, 31–40.

Winograd, T. and Flores, F. 1986. *Understanding Computers and Cognition.* Norwood, NJ: Ablex.

Wittgenstein, L. 1953. *Philosophische Untersuchungen (Philosophical Investigations).* Frankfurt/Main: Suhrkamp. (1986).