Elements of an Epistemology of Embodied AI

Erich Prem*
The Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Wien, Austria
erich@ai.univie.ac.at

Abstract

This paper discusses epistemological aspects of embodied AI as an engineering and as a cognitive science. The paper is abaout a warning and a proposal. The warning concerns potentially misleading assumptions about how spatial characteristics can be used to generate an alternative approach to grounding cognition. The proposal is to concentrate on biological transformations of epistemological questions that have lead to the development of modern ethology. These have been proven to be useful in the design of control systems for behavior-based robots. This also leads to a reacknowledgement of finality in the description and design of autonomous systems.

Key words: embodied AI, epistemology, robotics, Kant, theoretical biology, finality, teleology.

Introduction

In its short history, embodied Artificial Intelligence has challenged a sizeable number of foes. Among the list of opponents we find classical robotics and Artificial Intelligence (AI) in computer science, cognitivism in psychology, and objectivism in philosophy. The provocation lies in embodied AI's attack on a fundamental assumption of modern Western science. Dating back to medieval philosophy (or to Descartes, if you prefer) this assumption has been the primacy of the mental in the study of human cognition. 'Mental' here does not only refer to the opposite of 'physical' but also means 'rational' which is often considered opposed to emotion and intuition.

To extent that embodied AI tries to replace this predominance of the mental and rational by an emphasized acknowledgement of the bodily basis of cognition (Brooks 91) it threatens the disciplines mentioned above, which have a long tradition in disregarding the human body.

*In Proc. of the AAAI Fall Symposium on Embodied Cognition and Action, AAAI Press, Menlo Park, CA, 1996.

Rationality as language

Throughout AI, psychology, and also in philosophy (especially within the Anglo-American "analytic" tradition) the notions of human rationality and reasoning are centered around a strong emphasis of language. This is not only true for our human capability of producing speech and text reports but also for ideas, imaginations, beliefs, and other mental phenomena which are considered to be of inherently language-like nature.

As the outstanding example for this claim consider the interpration of logic in AI and the study of cognition as a 'science of reasoning'. A careful analysis reveals that logic is but a set of simplified abstractions over natural language which have originally been described so as to unmask the arguments of sophists (Prem 95). In this light, the claim that logic uniquely relates to the laws of thinking is hard to maintain. Nevertheless, human linguistic capabilities–their scientific explanation–and the efforts of their technological simulation have been taken as core characteristics of intelligence with respect to humans or programs.

It would be premature, however, to grant that embodied AI breaks with this tradition of language-centered accounts of reason. Quite to the contrary, main proponents of embodied AI make extensive reference to authors such as (Johnson 87) or (Lakoff 87) who argue that the basis of meaning and understanding in humans is the body and its environmental interaction. A closer look at their understanding of meaning and metaphor reveals another exclusive concentration on human linguistic capabilities.

This position is unfavourable because it misses the chance to abandon misguided consequences of this focus of interest in AI. This chance would be a fundamental reconsideration of questions concerning the meaning of words or thoughts, the nature of concepts, approaches to problem solving as mere artefacts of a linguistic view of cognition, rather than simply study these questions in the wrong way.

An alternative to this view of human cognition must include other important phenomena. For instance, musical understanding is often considered as genuine human, related to culture and individual interaction.

The bodily basis and emotional component of muscial understanding (Jackendorff 87) make it an interesting case for studying the priors of human cognition apart from language and inferential reasoning.

Criticizing pure reason: mind from the body

Another potential epistemological problem of embodied intelligence research could arise from an insufficient deliberation of how bodily interaction can form the basic source of the development of intelligence. Mark Johnson's proposals of image schemata as a set of tools of reason appeal to a predominantly spatial understanding in the sense that spatial relations between objects are used as metaphors for relations between abstract concepts. Some of the schemata can also be transformed to a time domain, or rather, the spatial transformation of a time-based image schema is used as the depiction of a metaphor. This concept of an image schema is Johnson's version of how the mind operates on raw sensory data, i.e. it is a description of a tool of reason. In Kantian terminology this is the pure concept of the mind or an a priori of human cognition.

In this version of embodiment, however, the danger arises to confuse the condition and the conditioned. Kant's statement that space and time are only formal products of our human experience should be taken more seriously when it comes to the question as to how spatial properties can be used to generate forms of abstract reasoning. Throughout embodied AI it is often taken as granted that image schemata arise from interaction with the properties of 3-dimensional space. Let us assume for a moment that Kant was right in that space and time are only a priori conditions of experience. Let us furthermore assume that spatial characteristics should be used to generate reason and therefore also condition experience. Then the question arises which of the two processes is the more fundamental one: which conditions which? In other terms, the problem is that spatial *experience* cannot be used to generate *spatial* experience. Another formulation would be that Johnson's characteristic relations depicted as image schemata are already based on what they try to explain: conditions of human experience.

The solution to these confusing subtleties of epistemology lie in a concentration on a view of system epistemics that is more oriented towards biology and has a sensory-motor perspective rather than a timely-spatial one. In such a perspective the fundamental building-blocks of cognitive abilities are control schemata for motor patterns that arise from perceptual interaction with the system environment. The drives for the system arise from within the system as needs or goals.

Theoretical biology and functional circuits

As soon as 1930 Jakob von Uexküll described a view of biology which bases the study of animals on the animal's view of the world rather than on a scientist's "objective" view of the animal and its environment. This is basically a Kantian turn in producing better predictions of how an animal will behave in a given context.

As an example consider the difference between the two following descriptions of the tick's feeding behavior:

1. The tick attacks warm-blooded animals like humans or deer when they make contact with the trees or grass inhabited by the tick.
2. The tick bites when making contact with anything which has a superficial temperature of 37 C and emits butyric acid.

While the first description is immediately easy to understand, the second certainly has a higher predictive value. The analysis which is necessary to come up with the second way of describing the tick behavior consists in a careful study of a tick's sensory organs and reflexes. In fact, the second version is more a description of *how the tick sees the world* in human terms. For the tick there are no humans, deer, trees, grass, etc. All that governs the tick behavior in the feeding context are specific features of two environmental qualities: temperature and chemical concentration.

> The sensations of the mind become properties of things during the construction of the world, or, one could also say, the subjective qualities construct the objective world. [J.v.Uexküll]

However, at the point where Kant's considerations lead to a discussion of categories as the final set of tools of reason to bring the "manifoldness of experience into the unity of concepts", von Uexküll develops descriptions of sensor (and actuator) spaces. His intention is to describe, how the

> marking signs of our attention turn into marks of the world. [Ibd.]

The basis of this process is formed by goal-driven interaction with the environment. The basic construct for explaining this interaction space is the description of *functional circuits.* Figure 1 depicts Uexküll's view of such a circuit. A "thing" in the animal's world is only "effector-" and "receptor-bearer". It can be thought of as a generator for signals to receptor organs and as a receptor of manipulations through effectors.

The formation of sensory experience is not only based on inter-*action.* Even more importantly, the interaction has a specific purpose. Such a purpose turns the object from a collection of merely causally operating parts of physical entities into a meaningful assembly of things which are integrated in a purposeful whole. The essential point is *to understand how the*

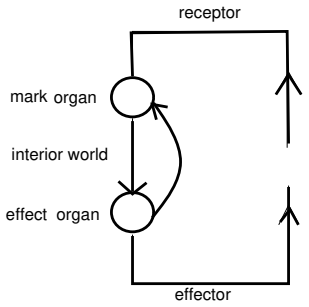

Figure 1: Action circuit as described by Jakob von Uexküll (1923).

thing is embedded in an action and how this action is embedded in a purposeful interaction with the world. In order to fully understand the system's world, our task consists in the dissection of the functional world (i.e. the whole of the subject's functional circuits).

Such a point of view is surprisingly close to the credo of behavior-based robotics where the descriptive strategy outlined above is turned into a design method. Starting from functional interaction circuits, the engineer tries to develop a minimalist architecture that fulfils the system requirements. An example for this strategy can be found in (Connell 90).

Summarizing Uexküll's position, there can be no understanding of animals without clarifying how they see the world, or better, what makes up the animal's world. Most notably, no such understanding seems possible without having gained insight into the animal's meaningful whole of functional circuits. To the modern, enlightened scientist such a view is dangerously close to the teleology which has been systematically eliminated from biology in the last century. However, there is a perfectly scientific version of finality that can help in the explanation and construction of embodied AI systems. Such a turn in describing representational elements in embodied AI systems even seems necessary, as will be argued in the next section.

Teleology

Consider an adaptive autonomous system that exhibits physical interaction with its environment. A part of such a behavioral system (Brooks 85) is schematically depicted in figure 2.

In this system, the behavior generated by the transfer function is learned based on a training signal. Let us assume that the training signal serves to optimize some criterion that is of importance to the system. It might, for example, assist in the provision of food. Following a description by (Rosen 85), we realize that the system's input is I, while the adaptation is determined by the optimization criterion r. Of course, it is reasonable to believe that there is a linkage between the "predicate" to be learned and the observables of the

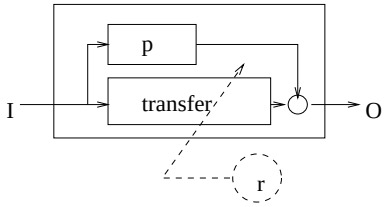

Figure 2: Behavioral module (after (Connell 90)) that transfers an input to an output if the applicability predicate p is true. The transfer function is learned based on some training signal r.

system environment I. Two things happen in this picture.

Firstly, the learning mechanism that selects the proper parameters generates a picture of the external linkage (between I and the predicate to be learned) *within* the system. Secondly, the adaptation must on principle be in a certain sense slower than the system-environment interaction. Thus, the system implicitly generates a *model* of the linkage, and also, of the system-environment interaction. (For an extensive treatment of these system-theoretic properties cf. (Rosen 85).)

The result of such a learning or selection mechanism is a transfer function that "predicts" external reinforcement, i.e. it drives the system in a way that fitness is optimized before it is evaluated. This is why Rosen calls such systems *"anticipatory."* Representations (data structures, models) in the transfer functions become shaped based on their predictive value with respect to maximizing fitness. As a further consequence, these representations must be properly explained with reference to the future outcome of the system's interaction with its environment. This results in a finalistic or teleological terminology.

Note that the selection mechanism itself works perfectly causally. It operates based on inputs and "rewards". But the generation of some kind of internal model (be it symbolic, connectionist, or statistical) makes it necessary to change the merely causal description of the system and the system's "representational" framework to eithe a finalistic or, probably, intentional one. The physical embodiment of our system is an important fact in this context. Bodily interaction is hard to describe and measure perfectly. Therefore, the subjective component becomes so strong that objectivity lies in the system's "objectives" rather than in purely physical properties. (Without this being unphysical, of course.)

As innocent as this descriptional framework may look, it has a rather strong influence on the system's representational framework. It is now likely that sensory impressions of the system are categorized in classes that form items of the same usefulness to the system. In the same sense that "chairs" are properly

described by their function "for sitting" for humans, objects in an embodied system's environment will now be classified due to their functional properties. It is clear that such a representational frame can be conceptually opaque in relation to human concepts.[1]

Additionally, the system will appear to behave depending on future events. This happens, because the actions are chosen so as to maximize reward or fitness that is evaluated later, based on an internalized goal-oriented model of system-environment interaction. This model, however, is based on the system's past experience.

There is evidence at the neurophysiological level that exactly this kind of finalistic indicative representation plays a major role in sensory-motor body-environment interaction (Tanji et al. 94). The subject centered viewpoint of Uexküll has also been well supported by neurophysiological evidence. Experiments by (Graziano et al. 94) show that premotor neurons play a major role in the coding of visual space. The evidence suggests that the encoding of the spatial location of an object happens in arm-centered coordinates rather than using a retinocentric represenation. This again points to the way how system-environment interaction of an embodied system creates models that are heavily oriented by the system's functional needs.

Ontology

This finalistic view brings with it the development of a rather distinct system ontology. The ontological position described here is so surprisingly similar to the existential-ontological philosophy of Martin Heidegger (Heidegger 27) that it is worth describing a few points of contact between both ontologies. Our notion of *things* in the animal's world can be best compared to what Heidegger calls *equipment*. In the human Being's everyday practices things in our world make sense because we can use them.

> We shall call those entities which we encounter in concern *"equipment"*. In our dealings we come across equipment for writing, sewing, working, transportation, measurement. The kind of being which equipment possesses must be exhibited. (Heidegger 27, p.68, taken from (Dreyfus 91))

The entities that will be encountered this way are not objects in the above sense. We do not simply add a functional predicate to them. *Dealing* with them is our primordial way of having them, not some bare perceptual cognition. To paraphrase Heidegger, "hammering" does not know about this property of being a tool. Instead, the more we are immediately engaged in coping with the problem of fixing something, the less the hammer is taken as an object which can be used in-order-to hammer (Heidegger 27, p.69). Strictly speaking, for Heidegger nothing like one equipment in this sense exists. This is because anything which we are using is embedded in a whole of multiple references to other tools and purposes. The hammer thereby refers to nails, tables, wood, etc.–i.e. a whole world of equipment and also of meaningful coping with the world. As long as we are engaged in "hammering"–in a purposeful dealing with equipment–and this equipment simply is "available", we do not even think about it. In such a situation the tools are simply "ready-at-hand".

> The world presents itself in the equipmental nexus, in the reference to a previously seen whole. (Heidegger 27, p.75, my translation)

The world does not consist of things which are "ready-at-hand", because it is only in situations of breakdown that the equipment can be recognized as one thing primarily identified by its sensory or physical properties. In these situations the things are deprived of being "ready-at-hand", creating mere *occurrentness*.

For Heidegger then, the fact that the world usually does not present itself as a world (in the usual scientific sense of the word) is the

> condition of the possibility of the non-entering of the available from the inconspicuous phenomenal structure of this being-in-itself. (Heidegger 27, p.75, my translation)

This view opposes any tradition which believes that things can be identified with reference to their sensory properties. Basically, this belief is based on the Cartesian assumption that extension must be essential characteristic of substance.

> [...] *Descartes* is not merely giving an ontological misconception of the world, but that his interpretation and its basis have lead to *skipping* the phenomenon of the world as well as the being of the [...] innerworldy being. (Heidegger 27, p.95, my translation)

In the end, this is one of the main sources of the problems of traditional robotics. From the idea that sensory and physical properties would be primordial it follows that a physical theory must be used to decide upon (detect, describe, deal with) objects encountered in the world. Moreover, such a theoretical approach must be used to find out whether a table could also be used as a chair. Any usage of tools and any way of dealing with the world therefore have to be explained with respect to those senory qualities. In a (remotely) existenial-ontological view, however, this problem simply does not arise in this way, because dealing with things for a specific purpose is the prevailing mode of encountering them, or rather: to create them. The argument therefore, is not that theoretical objects cannot exist, but that their functional properties must remain inaccessible if functions are not taken as the primordial source of creating things.

Of course, in this view a new difficulty arises, which we have not mentioned so far: How do all different views of one object become integrated? How can the hammer lying over there and the one in my hand become recognized as one hammer-thing? It is basically

[1] "And if a lion could speak, we would not understand it." (Wittgenstein 53)

at this point where many researchers would argue that human linguistic capabilities play a major role. But it is only at this point, and not before the questions concerning system ontology and epistemology have even been considered.

Conclusion

In this paper, I have tried to show that embodied Artificial Intelligence is a field of research that will have to address scientific problems in a way that is very different from comparable approaches in traditional AI or cognitive science. It will have to avoid the programmatic pitfalls of traditional AI which have lead the research in this field into a direction that was too much concentrated on a linguistic and mathematic view of intelligence. This view implicitly considered physics as of central importance to engineering solutions and as a general metaphor of how research in AI had to be pursued. Instead, embodied AI will have to ensure a reacknowledgement of natural elements, be they evolution, biology, ethology, or physiology.

Contrary to what people in the field of traditional AI have proposed (perhaps most prominently (Minsky 85)), "functions" may not be some additional property attached to an object, but at the very heart of what things actually are, i.e. of what there is in the world. The conditions of the possibility of object constitution are, of course, constrained by the sensory system. Knowledge about the nature of objects can only be gained by understanding the different actions of the system. The actions, and the related behaviors, must be based on understanding functional circuits. For the system engineer this means that the primary task consists in the design of a functional world of the autonomous system. Such a system, hence, will never be *auto*-nomous, but only *hetero*-nomous.

Acknowledgments

The Austrian Research Institute for Artificial Intelligence is sponsored by the Austrian Federal Ministry of Science, Transport and the Arts. A part of this research was completed during a visit to the MIT AI Lab. The author wishes to thank Rodney Brooks for his support.

References

Brooks, R.A. 1985. A Robust Layered Control System for a Mobile Robot. AI-Memo 864. Cambridge, MA.: AI-Laboratory, Massachusetts Institute of Technology.

Brooks, R.A. 1991. Intelligence without Representation. In Special Volume: Foundations of Artificial Intelligence, *Artificial Intelligence,* 47(1–3).

Connell, J.H. 1990. *Minimalist Mobile Robotics.* San Diego, Calif.: Academic Press.

Dreyfus H.L.: 1991. *Being-in-the-world.* Cambridge, MA.: MIT Press.

Heidegger, M. 1927. *Sein und Zeit.* (Being and Time.) Tübingen: Niemayer.

Jackendorff R. 1987. *Consciousness and the Computational Mind.* Cambridge, Mass.: MIT Press.

Johnson, M. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* Chicago, Ill.: University of Chicago Press.

Graziano M.S.A., Yap G.S., and Gross C.G. 1994. Coding of Visual Space by Premotor Neurons. *Science,* 11 November 1994, 266, pp. 1054–1057.

Kant, I. 1781. *Kritik der reinen Vernunft.* (A critique of pure reason.)

Lakoff, G. 1987. *Women, Fire and Dangerous Things; What Categories Reveal about the Mind.* Chicago, Ill.: University of Chicago Press.

Minsky, M. 1985. *The Society of Mind.* New York, NY.: Simon & Schuster.

Prem E. 1995. Symbol Grounding and Transcendental Logic. In Niklasson L. & Boden M.(eds.): *Current Trends in Connectionism.* 271–282. Hillsdale, NJ.: Lawrence Erlbaum.

Rosen, R. 1985. *Anticipatory Systems.* Oxford, UK: Pergamon.

Tanji J., and Shima K. 1994. Role for supplementary motor area cells in planning several movements ahead. *Nature,* 371 (6496), pp. 413–416.

Uexküll, J. von. 1928. *Theoretische Biologie.* (Theoretical Biology.) Frankfurt/Main: Suhrkamp.

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