Marcin Miłkowski
Institute of Philosophy and Sociology
Polish Academy of Sciences
mmilkows (at) ifispan.waw.pl

**Naturalized epistemology and artificial cognitive systems**

One of the obvious ways to start building artificial cognitive systems is to reuse existing theories of cognition. As philosophy abounds with such theories, it is not surprising that there are strong affinities between certain methodologies in AI and epistemology. For example, GOFAI approaches seem to be inspired by a view that knowledge is a formal, logical theory (as advanced by logical positivism). Connectionist approaches share a black-box behaviorist view on the cognitive agent. The current embedded and embodied AI is largely informed by phenomenological analysis of Merleau-Ponty and Heidegger, translated into engineering requirements. Evolutionary epistemology inspires attempts to evolve cognitive agents, etc.
At the same time, there can be insights in the research on artificial cognition that help develop philosophical views in epistemology. For example, the philosophical discussion of the frame problem, though often off the mark from the AI point of view, has helped many philosophers recognise how vague their ideas on knowledge were (see, for example, Dennett 1984, Fodor 1987). Generally, by building a cognitive agent that is based on a certain methodology, we can see the problems with the approach to knowledge that it is based on.

**Naturalized and artificial cognition**

Most of the epistemological conceptions just mentioned are naturalized theories of knowledge, that is they assume that the agent is a kind of a physical object in the world, and that it gains knowledge by interacting causally with the world (see Quine's 1957 naturalized epistemology manifesto). Just how those causal interactions should be accounted for and what knowledge is are the questions regarding which philosophers disagree. Most (if not all) naturalized theories of cognition should be translatable into engineering projects: if you know how cognition proceeds, you should be able to replicate the process in an artifact (or explain why it cannot be replicated at all). Dennett 2003 stresses this emphatically:

> Contemporary materialism–at least in my version of it–cheerfully endorses the assertion that we are robots of a sort–made of robots made of robots. Thinking in terms of robots is a useful exercise, since it removes the excuse that we don't yet know enough about brains to say just what is going on that might be relevant, permitting a sort of woolly romanticism about the mysterious powers of brains to cloud our judgment. If materialism is true, it should be possible ("in principle!") to build a material thing–call it a robot brain–that does what a brain does, and hence instantiates the same theory of experience that we do.

There are several benefits to researchers on artificial cognitive systems and epistemologists joining forces. The controversies around symbol grounding seem to also revolve around the question of whether knowledge might be embodied in a purely formal system; for example, is a rich enough formal system always prone to non-intended interpretations?

On the other hand, practitioners of AI, by browsing through philosophical literature that abounds with counterarguments and objections, can see whether their models of cognition are free of many known faults. So both philosophers and researchers that deal with artificial cognitive systems can benefit from sharing notes. To some extent, one can treat philosophical accounts of cognition as extremely top-down approaches, while the artificial cognition research can start with some bottom-up methodology. Of course, this is a gross simplification, as we can see general ideas in AI research

and some very detailed questions in philosophy. With naturalized philosophy, the boundaries between sciences and epistemology do become blurred.

**Android epistemology**

Of course, if you cannot build the system based on philosophical assumptions, this can rarely show decisively that the assumptions were wrong. It might just be that some of the engineering hypotheses were wrong; most epistemological conceptions mentioned are still alive and kicking, energized by attempts to get rid of the old problems. So seeing that an engineering attempt fails is not exactly most beneficial for the philosophers and engineers.

Generally, in research on cognition, one can take a biologically inspired approach or artifact-inspired approach. These approaches can intermingle, of course, as artifacts are often also biologically inspired, as are airplane wings. Still, there are artifacts that have no counterparts in the evolutionary world, such as wheels, which could not have evolved for morphological reasons but remain–in their niche–a very good technical solution. It remains an open question whether artificial systems that are today able, for example, to prove theorems using a method that is not viable for a human being (such as the four color theorem), are minimally cognitive. We may expect, however, that future cognitive systems could have capacities that are more like wheels than wings. This does not diminish the significance of computational modeling of biological cognition, of course.

Such wheel-like cognitive capacities are at least as important in naturalized epistemology as research on biological cognition. Building cognitive systems that are modeled on neither animal nor human cognition has been always a goal of epistemological theories (at least since Kant analyzed finite rational beings that are able to do natural science, but didn't assume they must be human). So joining the forces between philosophical analysis of cognition *per se* and engineering of cognitive systems can lead to *android epistemology*, as Clark Glymour dubbed it – to epistemology that can analyze non-biological cognitive systems (see Ford, Glymour & Hayes 1995; Ford, Glymour & Hayes 2006). They could be arbitrarily less or more powerful than natural cognitive systems. By juxtaposing those artificial models with natural agents, one could try to see the really cognitive aspect of their actions. At least this is the hope shared by many philosophers.

**Deflating cognition**

One of the dangers inherent in naturalized views on cognition is that cognition would be deflated and reduced to one of its properties. For example, if you think that cognition is all about countering entropy, you cannot account for some cognitive phenomena like theorem proving or representation building. Why should problems of representation ever appear if it's all about energy? Entropy is important to the functioning of all physical systems but it isn't the single most important factor in cognition.

A good way to avoid such deflationary accounts is to try to sketch a view that encompasses many aspects of cognition. But there are so many smaller and bigger theories around that it's hard to see where they are compatible. As numerous definitions of cognition on euCognition's wiki page show (http://www.eucognition.org/wiki/index.php?title=Definitions_of_Cognition), it is hard to see any common ground for an overarching theory of knowledge for cognitive systems.

Some of these definitions seem to flow from a strongly deflationary, or reductive view of cognition, such as "autonomous anti-entropy engine". It could be tempting to deflate cognition to self-awareness or learning from experience, from example. However, it would be much harder to defend the view that an autonomous agent that learns from experience but has no self-awareness really knows something. Or, that an agent that is self-aware but unable to learn can cognize. Some

definitions account for learning, but don't stress interaction with the environment or stability of the artificial system in the environment. Some account for perception but never mention action. What philosophers could attempt is to identify the various attractors in the phase space that describes possible artificial cognitive systems, at least those sketched by members of euCognition.

I can see at least eight attractors in the definition space: learning, action, knowledge, autonomy, self-awareness, adaptation, communication and information processing. This is definitely quite a chaotic list, and this analysis serves only as a departure point. So let's see what falls under those headings.

Learning is assumed to involve real-world interaction, reasoning (and reasoning includes prediction and evaluation of beliefs), search, and modifying behavior. Real-world interaction also means adapting to environment, especially when it changes.

Let us look then at adaptation. The adaptation attractor includes ideas of embodiment and embeddedness in the environment (generally, you have to be in the environment to adapt); cognition's role is to enhance fitness. Comparing artificial systems to animals seems to be a proper way to study cognition.

Adapted systems are supposed to act flexibly. This brings us to the action attractor. Action is meant to be based on learning and knowledge, flexible and embodied. Already, we can see some overlap.

What about knowledge? Knowledge is being assumed to flow from experience, and to be more exact – from learning. It involves perception; and perception involves sensing (auditory, visual etc.), and sensing is real-world interaction. Perception leads from sensory inputs to abstractions, to categories, to symbols, to concepts, to models; to some kinds of representations being stored in memory. Some say representations – especially symbolic ones – must be grounded to have any meaning. Some avoid talking about representations and try to model knowledge contents by using purely procedural algorithms. Those representations are used in planning, modeling, and in various cognitive algorithms.

The best method for checking the real meaning of symbols seems to be communication. That's why some stipulate that the cognitive system involve communication.

The system must be also relatively stable and have goals to remain an autonomous system. So it must have anti-entropic ends, and be embodied. It's also clear that a system is a system when it has a certain structure, an architecture that implements all those features mentioned.

And the last attractor, rarely mentioned explicitly: information processing. It seems obvious that real-world sensing is information processing, and that this information is used for learning, and then stored, symbolically or not, in some forms of representation or memory. The question then is whether this information processing can be purely computational or whether it must be appropriately grounded in the real world. Another question is whether there exist any information-processing methods or resources that cannot be computational in principle.

Such non-computational information might flow from self-awareness, or self-awareness might be the ability to monitor and change beliefs accordingly.

As can be seen, it is possible to put together a non-contradictory story about cognition that uses almost all of the definitions, even though they originally probably seemed incompatible. The story is quite chaotic but appears fairly convincing. What I will do next is extract some general points from this story to show that there are minimal and maximal models of cognitive systems.

**Cognition: a multiple perspective view**

Beside the thirty-nine definitions of cognition, there is one comment – Aaron Sloman's – that it makes little sense to start with definitions, and that we should try to sketch a logical geography of concepts of cognition, trying to see minimal and maximal cognitive capacities. I will sketch an introduction to such a geography; though I think that in fact, there is one, generally agreed to notion of cognition behind all those definitional attempts. The partial accounts do not necessarily contradict each other, and I will show their common ground. It will then be possible to see potential disagreements, and to construct the requirements for minimal and maximal cognitive capacities.

All proponents seem to agree that cognition leads to flexible action and general knowledge. Naturalistically, they hypothesize about the ways to achieve those goals. There must be a stable structure, i.e., an architecture that has functions contributing to flexible action and general knowledge. Processes that involve learning from experience-where by "experience" we mean both sensory experience and previously acquired knowledge of any kind-and that can realize those functions. The cognitive architecture is implemented in a stable, autonomous structure that is able to be an agent in the environment. The agent's action must be appropriate in this environment, and its knowledge adequate to it. As a result, the action may contribute to the agent's fitness.

This is a general account that everyone seems to agree with. There are disagreements about what learning is. But it is generally agreed that it may involve some kind of perception/sensing (auditory, visual, olfactory, or other), and categorization. Researchers disagree whether categories are abstract models, concepts, symbols, or abilities to act; nevertheless, everyone agrees that they are being stored in the memory and reused. All cognitive systems are able of least the forms of reasoning that are necessary to plan actions and predict the environment. The exact processes of learning or reasoning are under discussion: is it just a search? Or maybe it involves some heuristic processes? Or even formal deduction?

The flexible action, or (put slightly better) the flexible interaction with environment requires, according to some, embodiment and embedding in the environment. Others disagree, saying that in principle, all the information from the environment could be extracted and fed directly into the agent (this is another way of saying that brains in the vats could be cognitive while still in vats).

Regardless, the memory contents and categorization results, if they are to be knowledge, must be about the environment. Here the old question of intentionality arises. How is intentionality possible? Do symbols in the agents need some special grounding, or is symbol grounding a theoretical artifact, as Aaron Sloman seems to suggest, by saying that it stems from extreme concept empiricism refuted already by Kant. Either way, a coherent story about how symbols acquire meaning is something that is required to be able to speak about higher cognition.

It is controversial whether higher cognition requires any self-knowledge or self-awareness. The natural systems we know that are truly cognitive are also self-aware, but is this just a coincidence or a prerequisite? On some theories of consciousness, it is ascribed a role of a slower but discursive and rational system that is used for logical inferences and reasoning, while most practical reactions occur at a lower, unconscious information-processing level.

At least some cognitive systems are able to communicate. A recent surge in interest about social cognition seems to require accounts of the role of intersubjectivity in cognition; and communication is one of the most important features of it. Social learning is helped a lot when it is supported by effective linguistic communication. Language , however, builds upon existing semiotic abilities to use and reuse signs.

Some of these features must be present in all cognitive agents for them to be cognitive. Some others are optional, meaning that they can enhance cognitive capacities and contribute to new capacities. For example, a non-learning agent that cannot act flexibly nor acquire any general knowledge cannot be said to be cognitive. *Camera obscura* is not a cognitive agent. Minimal cognition requires at least minimal learning and action flexibility, showing that some knowledge is stored. Maybe even plants are able to cognize in this sense, and clearly some artificial agents are already are minimally cognitive in their epistemic niches (see Calvo 2007). At the same time, it is clear that no existing artificial agent is maximally cognitive, i.e. capable of full natural-language verbal reporting, introspection, efficient communication, and of reusing the knowledge intersubjectively available in the form of linguistic products such as books, novels, poems or hand-written notes. We're pretty much in the dark about how many features we should ascribe to such a maximally able cognitive agent. It remains mostly a conceptual possibility, whereas the minimal agent is likely to be implemented technically.

This, in gross simplification, is the general status quo in the field. The themes mentioned can be approached both from an engineering or AI side, and from a philosophical side – both parties know them well. The lectures that follow touch upon various aspects of this account, and try to criticize it in from a range of viewpoints.

**From function to meaning, truth and disagreement**

Ulrich Krohs will contribute to the picture by speaking about functional analysis, the point of relevance being that we need a good account of function and dysfunction to speak meaningfully about cognitive architectures. What is especially important is that most accounts cannot deal with dysfunctions. Functional characteristics of artificial cognitive systems are normative in a weak sense: you can say that an agent that is not functioning according to its architecture is dysfunctional. In this way, a functional account of cognitive capacity has implications for the question of normativity of epistemology.

Cognitive architectures must be able to acquire meanings and manipulate signs. Don Favareau will argue that in order to grasp the nature of the meaning relations we should not start with human-level signs but with biological semiotic phenomena. Abilities to use signs are more wing-like than wheel-like: they were invented by evolution, so we should take a close look at them to find out what's so distinct about them.

Maria Frappoli will deal with the question how to account for truth in a naturalized way. Her account tries to be as ecumenical as the account of cognition I outlined – we may already have lots of theories of truth but they don't necessarily contradict each other. They might describe the same elephant that we are all grasping parts of. Truth is a property that knowledge, a product of cognition, should have. It's not a property that is reducible to syntax (as has already been proved by Tarski, as we should remember while in Poland, where this workshop takes place) but is a distinctively semantic property. In other words, in order to be able to speak about truth of knowledgeobtained by artificial agents, we must account for semantics.

Of course, accounting for semantics requires that we look at the linguistic and social properties of the agents. Stevan Harnad will analyze the ways the social distribution of cognition allows agents to get categories from hearsay. By using a computer simulation, he will argue that the boost that the agents have from sharing the knowledge contributes significantly to their cognitive and survival capacities. What is important is that this lesson is learned from an artificial agent simulation.

One of the implications of social cognition – disagreement – will be focused upon in Hilary

Kornblith's talk. Reasonable disagreement is ubiquitous in environments where agents have incomplete knowledge. Negotiating between different positions is generally an NP-complete problem, which has been seen–though under a different name–already in philosophy, where incomplete knowledge is there to stay.

Without trying to summarize these and other talks, one point needs to be made clear: the aim is to focus on some specific points within the naturalized account of cognition as existing in cognitive systems, from the structures of those systems to their capacities. I hope that this year, as before, we'll see many striking affinities among different approaches.

**References**


CALVO GARZON, P., 2007, "The Quest for Cognition in Plant Neurobiology", *Plant Signaling & Behavior*, 2:4, p. 208-211.
DENNETT, D., 1984, *Cognitive wheels: The frame problem in artificial intelligence*, in C. Hookway, *Minds, Machines and Evolution*, Cambridge: Cambridge University Press, 129—151.
DENNETT, D., 2003, *What RoboMary Knows*, in: *Knowledge Argument*, Torin Alter (ed.) (draft available at http://ase.tufts.edu/cogstud/papers/RoboMaryfinal.htm).
FODOR, J., 1987, *Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres*, in: Pylyshyn, Z. (ed.), *The Robot's Dilemma. Frame Problem in Artificial Intelligence*, Ablex, Norwood, p. 139—149.
FORD, K.M., GLYMOUR, C., HAYES P., 1995, *Android Epistemology*, Cambridge, Mass.: MIT Press.
FORD, K.M., GLYMOUR, C., HAYES P., 2006, *Thinking about Android Epistemology*, Cambridge, Mass.: MIT Press.
QUINE, W.V.O, 1969, *Epistemology Naturalized*, in: Quine, W. V. *Ontological relativity and other essays*. New York: Columbia U P.