Evolving AI:
Position paper in Developmental Robotics

Pentti Kanerva
Redwood Neuroscience Institute
Menlo Park, California
pkanerva@rni.org
January 2005

Abstract

This paper addresses the difficulty inherent in our understanding of our own intelligence because we have no conscious access to the underlying mechanisms. However, we know enough about human mental processes to create conditions favorable for discovery. How to apply this knowledge toward the discovery of the underlying mechanisms and processes is discussed.

Introduction

From its inception, AI has had an ambitious goal of building machines that simulate human intelligence as manifested in our use of language, forming of abstractions and concepts, solving problems, and learning. I will call this kind of human-level AI the *high goal* of AI. In fifty years of research the task has proven extremely demanding, and the goal continues to elude us.

Achieving the high goal amounts to nothing less than understanding how the human body and mind work. This view challenges a popular tenet of AI that the actual mechanisms don't matter so long as they reproduce the sought-after behavior. However, they do matter when the target is ill-defined and the goal keeps shifting. Not that human body and mind, or behavior, are radically changing, but that our appreciation of them changes as we learn about human behavior and try to simulate it. Thus built into our quest is a dynamic that drives us toward a more and more complete understanding of the human body and mind, because it is unlikely that a very different architecture would produce the same behavior, and even if it did, how would we find it?

In this paper I will examine a major difficulty posed by our quest and will suggest a path of inquiry to overcome it or, rather, to circumvent it. It turns out that Developmental Robotics lies squarely on that path.

The difficulty comes from deep self-involvement, as suggested above. Since our quest is defined by what we humans can do, having an objective view of ourselves is crucial, but it is also very tricky. It is reminiscent of our understanding of our place in the universe when our immediate experience places us at its stationary center. Gradually we came to accept the sun-centered view of the nearby universe and the insignificance of the solar system in all of it, and only because they explained better the accumulating data. But even in this picture we are at the center, as we are the ones to whom the data are to be explained, and in terms that somehow make sense to us. A big part of the difficulty, therefore, is in assuring that our concepts are adequate for the task. How to make sure that our concepts are adequate for describing human intelligence?

Evolution

There appears to be no way for us to have an entirely objective view of ourselves and hence of the high goal of AI. The next best thing is to remove ourselves as far from the center as possible and to attack the problem indirectly. As a framework for this approach, I propose the evolution of intelligence in the animal world. Evolution recommends itself by being the only process that has actually produced the kind of intelligence we are interested in. Besides, a living record of the evolution is all around us in the form of animals at all levels of intelligence. We can look into that record and let it guide our search.

The first things to note is that we are interested in intelligence that is attributed to brains (and nervous systems), and that they process information. So the brain is a kind of a computer that runs the body and accounts for the mind, so that modeling intelligence on computers seems justified. The adaptability of brains is a major evolutionary asset that at once also becomes an engine of evolution.

What stands out in the evolutionary record? Most notably, brains become larger and nervous system more complex, and new brain structures become prominent. This matches our intuition about computers—bigger is better—including the notion that with a computer that is large and fast enough we should be able to simulate any level of human intelligence. So far we are aligned with the traditions of AI. But we can learn much more from real brains and from the bodies they inhabit. I will examine features of animals that seem highly relevant to computational modeling of intelligence.

Sensor–Motor–World Loop

Evolution takes place in a context, which we refer to as the *environment* or the *world*. The world is a complex conglomeration of resources to sustain life and threats to survival. Its complexity exceeds any individual's capacity to understand it fully, but the more an individual knows about the world, the better its changes of survival are—more is better.

Conspicuous in animals is the *prevalence of sensors*, which, together with the actuators (motors, muscles), couple the animal to the world. From an information-processing point of view, early evolution was marked by the profusion of sensors that permeate the body and respond to varying conditions of the world and of the animal itself. An animal's sensory neurons can number in the millions and they can respond to light, sound, temperature, pressure, stretch, vibration, electric and magnetic fields, chemicals, body position, body movement, and covert conditions of the body itself. It appears almost as if the body were there merely to provide a platform for a huge array of sensors! The brain's world then is this enormous amount of data constantly pouring in.

The brain evolved to deal with this massive input and to convert it into beneficial action, where also the action is defined as the activity over large numbers of neurons, ones that drive muscles and glands. Evolution worked on this design—on converting elaborate patterns of input into beneficial patterns of output in the context of the world—for several hundreds of millions of years before inventing the design that gave us symbolic thought and language, which by comparison is a very recent invention. As a rule of evolution, a new design is an elaboration of the old and relies heavily on it. We can therefore expect that our symbolic abilities rely heavily on our presymbolic abilities—not to be confused with subsymbolic—that we share with all animals.

By distinguishing between presymbolic and subsymbolic I want to draw attention to the ability of animals to comprehend the world—their being able to deal with it—without a system of abstract symbols that obey compositional syntax and semantics. Nevertheless, the brains of presymbolic animals form internal representations of the world and operate, or compute, with the representations. Missing in presymbolic animals is the ability to form arbitrary mappings between internal representations on the fly. The word subsymbolic would then refer to neural/connectionist realization of such mappings, making symbolic systems and language possible.

Handling of sensory input that is commensurate with the sensory input in animals, and converting it into beneficial action without resort to language-like symbol processing, is a major challenge to AI and robot design. To appreciate the import of this challenge, let us review what presymbolic animals can do. Assuming that symbolic language developed in primates, we can look at other mammals and birds. They learn from experience and improve by practice, they have a sense of what they can and cannot do, they remember places and things over long periods of time and act accordingly—they appear capable of imagining and planning—they communicate, they learn a social structure and their place in it, they learn by imitation, they learn rudimentary use of tools, they can be taught to obey and to perform tricks. Presymbolic animals are complete functioning autonomous systems that lack the ability to tell stories of what they have experienced or are doing.

There is a sense in which we humans understand the world and know what we are doing that is like how a dog understands the world and knows what it is doing. We should try to build that kind of foundation into our AI systems and robots first, and then base symbolic manipulation and language on it. My hunch is that this would go a long way toward solving problems in computational linguistics, for example, and could resolve ambiguity that now haunts natural-language processing. In other words, ambiguity would be resolved outside language and language-like systems such as symbolic logic, by relying on a more visceral understanding of the world.

Circuits and Representation

The brain is organized in different kinds of circuits, apparently to serve different kinds of functions. A circuit can be identified by its relative uniformity in terms of types of neurons and their connections. Remarkable about the circuits is their size. Their neurons number in the thousands to millions to billions, each with multiple—into the thousands—synapses that are like nature's transistors. Furthermore, even very simple physical and mental activity involves the activity of large numbers of neurons distributed over multiple circuits.

Such numbers are staggering, and they must be so for a reason. Among other things, they suggest that intelligence in the animal world relies on very high-dimensional, distributed representation. We already have some understanding of such representations: they allow robust systems to be built from unreliable components that are wired according to a general plan. Brains that differ in their details can nevertheless be equivalent, and the death of individual neurons does not drastically affect behavior. But these are merely superficial observations about brains. We need to know much more about brainlike representations. Some studies along these lines have been made, but the area deserves extensive in-depth research (e.g., Olshausen & Field 2004).

Neuroscience and Robotics

Neuroscience obviously studies neural circuits, and the area most closely associated with robotics is called systems neuroscience. It relates behavior to neural architecture and is a rich source of information on neuroanatomy, neural development, and the relation of neural structures to the functions they control (e.g., Swanson 2000). However, from an engineering point of view, the neuroscientists' notion of computation is incomplete. It is expressed in terms such as information flow, signal relay, facilitation, excitation, inhibition, spike train, membrane potential, synapse, and neurotransmitter. In other words, either very general or very specific, whereas key elements of computer engineering fall in between. They deal with codes and with circuits for encoding, decoding, and processing of information in terms of the codes. Computer science could make a major contribution to systems neuroscience by interpreting the neural structures in terms of circuits, codes, and information processing based on the codes, and systems neuroscience in turn could instruct AI and robotics in the organization and development of autonomous systems.

Computation Implied by the Above

Very high dimensionality of the representation—in the thousands to millions—appears to be crucial to the brain's computing. The exact nature of the dimensions seems to matter less than their number because important properties of high-dimensional spaces are evident even when the dimensions are binary. Conventional computers, by contrast, are built for low-dimensional entities: the dimensionality of the address space is usually less than 30, and the word size is usually somewhere between 8 and 64. High-dimensional entities can of course be simulated on conventional computers, although such simulations can be time-consuming.

Computing with high-dimensional distributed representations is bound to be very different from traditional numeric and symbolic computing. Quantities are inexact, patterns of input never repeat exactly, and memory works by association. Mathematically this means that the brain forms equivalence classes from its inputs and then operate on the classes, in addition to the immediate sensory data. For example, a specific person is an equivalence class over sensory inputs, and a human being is an equivalence class comprising individual persons. The brain's ability to generalize is partly due to the dimensionality of the representation.

Research into brainlike representations has exploited both the geometry and the algebra of high-dimensional spaces. Artificial neural nets and associative memory (e.g., Anderson & Rosenfeld 1988), semantic vectors such as in Latent Semantic Analysis (e.g., Landauer & Dumais 1997), and concept lattices (e.g., Widdows 2004) depend on the *geometry*; that is, they depend on the distances between, and the alignments of, points that represent meaning. Noise-tolerance and robustness of neural systems are direct results of the geometry. Holographic Reduced Representation (Plate 2003), on the other hand, relies on the *algebra*, most notably on multiplication operators that map a configuration of points from one part of the space to another and thereby potentially to new meanings. This allows symbolic composition to be realized in brainlike distributed representation and it could also provide a mechanism for analogy, which plays a key role in human intelligence.

The mathematical properties of high-dimensional spaces are rich and subtle, and much remains to be learned about them and their use for computing. They are likely to give rise to a new theory of computing that emphasizes representation and is concerned with efficiency—finding solutions fast enough that are good enough. The new theory would have a major impact on the design of computers for AI and robotics. If the reader were to take with them only one lesson from this paper, I would like it to be the need to discover the secrets of very high-dimensional distributed representations, of computing with very large patterns (Kanerva et al. 2001).

Relating to Developmental Robotics

Traditional AI strives to mimic (and exceed) human intelligence by whatever means possible. I call this *behavioral* or *functional cloning* and use the word cloning to emphasize the fact that human intelligence serves as the model, it provides the target. How we define intelligence and what we expect intelligent systems to do is drawn fundamentally from how we view ourselves.

Developmental robotics takes the human (and animal) model of intelligence closer to heart by modeling itself after human cognitive development, one reasons being that the more traditional symbolic AI has great difficulty programming general-purpose systems that deal with open-ended situations. For example, language understanding may never yield to the purely symbolic approach—nor purely nonsymbolic either.

Quoting from the symposium announcement, developmental robotics "focuses on the autonomous self-organization of general-purpose, task-nonspecific control systems." This too is close to functional cloning because it looks to duplicate overt aspects of human (and animal) intelligence—namely, how such intelligence comes to be and what it encompasses. The shift from the traditional AI is toward autonomous self-organization, because that is how human intelligence develops. There is also a trend toward *structural cloning*, as the robots incorporate more and more parts that are motivated by real nervous systems (Weng & Zhang 2002).

Developmental robotics aims at coping with unanticipated challenges of the environment. An agent—a physical body—interacts with its environment and improves its behavior by building an internal model of the interaction. The patterns received by the agent's sensors and the patterns driving the agent's effectors are the language of interaction, just as they are with animals coping in the world, or with any breed of robots. However, the makeup of the internal model makes a difference. In traditional robots it is programmed for specific tasks based on the designer's understanding of the problem. The resulting internal models are mostly logical and are naturally programmed in symbol-processing languages such as Lisp. In developmental robots one task looms over all others, namely, open-ended learning. The internal models are mostly statistical and hence prime candidates for artificial neural networks. The internal models of animals, naturally, are realized by neural networks.

The makeup of the internal model is crucial, but it is also the least understood part of human and animal behavior. For its resolution I advocate a form of structural cloning: look for ideas in structures that work and try to interpret them. Brains have a lot of organization that our neural-net models lack. Obviously very different circuits perform very different tasks, and without a proper circuit a task is unlearnable. We may be able to teach a dog a set of commands, but we cannot do it by describing the desired action in words. When the circuits that allow language to develop are missing, no amount of training will make up for it. Ultimately we want to understand the algorithms realized by the circuits and then consider different computational realizations of them.

Preparing for a Solution: Education

I will return to a topic that I began with, the difficulty of grasping human intelligence because of self-involvement—the difficulty of understanding that with which we understand. Since we have no direct access to the underlying mechanisms by introspection, we need to approach the problem indirectly. This brings forth a most intriguing issue: the better we understand the mind's workings, the better we will be able to solve difficult problems, including how the mind works in terms of its underlying mechanisms. That is to say, we can deliberately set up conditions that promote discovery. That is the ultimate in bootstrapping!

We already know enough of the psychology of human learning to take us onto this path. Discovery has a logic of its own that cannot be forced but can be facilitated. The key is in preparation. The mind needs to be prepared to *recognize* a solution if and when presented with one. It will then make its leap even if don't know how.

Preparation means becoming thoroughly familiar with—immersed in—both the problem and a broad range of topics that could contribute to the solution. The problem is manifested in human and animal behavior, and a host of disciplines bears on the solution, from biology all the way to philosophy.

Animal evolution gives us countless examples of intelligence. To be guided by it, we need a broad understanding of biology and neuroscience. Appreciating the magnitude of neural circuits is essential to discovering their algorithms.

On the behavioral and cognitive side, we need to be educated in psychology, psychophysics, and linguistics. We need to know what cognitive systems can do, how they fail, and how they can be fooled. As manifestations of the underlying mechanisms, behaviors are a prime source of challenges and checks to our exploration.

Education in computer science with its branches of AI and robotics prepare us to think in terms of algorithms. It also gives us a sense of computation as something to be quantified, something that you may need more of or less of, which translates into some tasks being difficult and others easy. Furthermore, computer simulation is an indispensable tool of exploration.

Finally, two areas of utmost importance are mathematics and engineering. The issues of representation and computation with very high-dimensional vectors are deeply mathematical, and some of the necessary math may not even have been invented yet. The field needs mathematicians who can appreciate all the educational needs as outlined above because, in the end, somebody needs to make a connection between an observed set of behaviors and a mathematical system that could explain them. That is how we put the sun at the center of the nearby universe. Without a thorough familiarity with abstract mathematical spaces, it will be impossible for anybody to see the connections and to draw the proper analogies—or for any mind to make the necessary mental leaps.

Engineering plays a special role by making things real. However, its import goes far beyond the practical, for it provides the ultimate test of our understanding. Developmental robotics is motivated to a large part by the intellectual rigor that engineering imposes on the exploration.

And let us not overlook the value of sound philosophy: we need a conceptual framework that is adequate for the task. The evolution of intelligence can help there, too. We can first develop concepts for describing the intelligence of primitive animals, in terms that are meaningful at that level, and then bootstrap to concepts for higher and higher levels of intelligence. Thus, tracking the evolution of intelligence in the animal world may well lead us to the high goal of AI. I call it the low road to the high goal.

Human and animal intelligence are properties of physical systems operating in the physical world and therefore should be capable of being built into artificial systems within the limits of existing technology. Progress in electronics manufacturing points to a future when the technology no longer is the limiting factor, whereas our understanding of the mechanisms that underly intelligence could still be. The point of this paper is that we can, in fact, organize education and devise a research strategy that lets us work out also the needed theory. Developmental robotics is a very natural component of that strategy.

References

Anderson, J.A., and Rosenfeld, E. (eds.) (1988). *Neurocomputing: Foundations of Research.* Cambridge, MA: MIT Press.

Kanerva, P., Södin, G., Kristoferson, J., Karlsson, R., Levin, B., Holst, A., Karlgren, J., and Sahlgren, M. (2001). Computing with large random patterns. In Uesaka, Y., Kanerva, P., and Asoh, H. (eds.) *Foundations of Real-World Intelligence.* Stanford, CA: CSLI Publications.

Landauer, T.K., and Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Olshausen B.A., and Field D.J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology* 14:481–487.

Plate, T. (2003). *Holographic Reduced Representation: Distributed Representation for Cognitive Structures.* Stanford, CA: CSLI Publications.

Swanson, L.W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research* 886:113–164.

Weng, J., and Zhang, Y. (2002). Developmental Robotics: A New Paradigm. *Proc. 2nd International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. EPIROB 2002, Edinburgh, Scotland.

Widdows, D. (2004) *Geometry and Meaning.* Stanford, CA: CSLI Publications.