In SAB-90: Simulations of Animal Behavior, MIT Press,
S. Wilson & J. Meyer (eds), 1990, pp. 462-4.

Eight Principles for Building an Intelligent Robot*

David L. Waltz

Thinking Machines Corporation
245 First Street
Cambridge, MA 02142
USA

Brandeis University
Computer Science Department
& Center for Complex Systems
Waltham, MA 02254 USA

Abstract

We cannot expect to know the detailed "wiring diagram" of the nervous system for any intelligent creature for quite a long time. Even then, the true organization is likely to be incredibly complex and tangled. However, in order to build intelligent robots now, we need a plausible interim architecture. A functional model for robot organization is proposed, starting with a basic, first order model, which is gradually refined. In particular, it is proposed that associative memory provides a useful -- and perhaps plausible -- basis for an intelligent system.

0. Introduction

While remarkable progress is being made by neuroscientists in unraveling portions of the nervous system (see, for example, [Kosslyn 89] or [Halgren 87] for insights into the visual system and memory systems, respectively), we are still far from being able to map the wellsprings of action, intention, and decisions. Other researchers have investigated abstract models of adaptation and learning, such as genetic algorithms and classifier systems [Holland 77], or the SOAR system [Newell 87]; abstract models have been used to build explicit models of creatures (e.g. the Animat [Wilson 87]). [Drescher 89] has introduced the "schema mechanism," and his ideas have much in common with the proposals below, especially in his views on chaining, and in his key ideas on identifying and learning reliable schemas, using large amounts of statistical analysis. "Subsumption architecture" researchers in AI (e.g. [Brooks 86], [Maes 90]) hope to arrive at intelligent systems by first building a (layered) system with the abilities of, say, a cockroach, and adding yet more control layers to eventually reach greater and greater intelligence. This work is broadly within a "Society of Mind"-type theory that views intelligence as composed of a very large number of independent agents and "bureaucracies" of agents, each responsive to specific situations or patterns [Minsky 87]. While I subscribe in general to the Society of Mind view, I believe that it is both possible and valuable to divide up the model of mind somewhat differently than is done within subsumption architectures.

I propose here a model of a robot's "mind" whose components are divided up along very different lines, somewhat analogous to principal components analysis: the first component is a general associative memory model that captures general patterns and principles of behavior, while successive components add refinements, culminating in society of mind-like demons that recognize very specific situations or patterns, and then override (by priming or inhibiting) more general behaviors. Intermediate refinements include control structures that allow search and chaining of actions, as well as rote learning and generalization. Such a model fits neatly on any massively parallel computer architecture (e.g. [Hillis 85]), but can also be simulated on serial computers (though perhaps not fast enough to allow real-time performance, except in the simplest of environments).

1. Principle One

Use associative memory as the overall organizing conception.

Basic associative memory operations can capture the essence of what intelligent entities do: select relevant precedents in any situation, and act on them. "Precedents" can be actions, options, remindings, etc. This type of operation, akin to case-based reasoning (CBR) [DARPA 88, 89] and memory-based reasoning (MBR) [Stanfill & Waltz 86] is easily programmed on a massively parallel machine, and has found useful applications [Waltz 90]. A number of techniques can be used to find "relevant" items, including nearest-neighbor algorithms, and majority votes of n nearest neighbors.[1]

If only a single precedent is close to the current situation (as when the robot is operating in a familiar environment on a familiar task), then little more than an associative memory is needed in order to act intelligently. Only when two or

*This work was supported in part by the Defense Advanced Research Projects Agency, administered by the U.S. Air Force Office of Scientific Research under contract #F49620-88-C-0058.

[1]What makes a neighbor "near" is a very subtle issue, and the key open problem in CBR and MBR.

more incompatible precedents of roughly equal plausibility are found, or when nothing matches the current situation at all closely, or when all known precedents have negative outcomes, does such a system need to reason (in the ordinary AI sense), as discussed further below.

Thus, the basic action of the robot can be represented as the (behaviorist) schema:

situation --> action

Obviously, what constitutes an "action" or "situation" is also important. I am going to partially duck these issues here. The notion of situation is enlarged upon somewhat below; the notion of action depends upon the particular granularity or chunking of motor sequences built into the system, along with their parameterizations. See also [Agre & Chapman 87] and [Drescher 89].

2. Principle Two

Include innate drive and evaluation systems to provide the robot with moment-to-moment guidance for its actions.

The notion of "current situation" can be enlarged to include not only interpreted sensory inputs, but also "desires" of the robot (potentially time-varying, innate constraints, that could not be learned easily through experience; for example, a particular class of robots should "desire" to go plug themselves into recharging locations with progressively greater intensity as its battery power gets lower, and this desire should overwhelm all others as the power becomes very low). Other examples of this principle include innate wiring of a desire to avoid collisions, or a desire to minimize energy expenditure ("laziness"). Actions that are "desirable" should cause the robot to assign them a positive evaluation when they are carried out. The robot should also store the results of taking the action as part of the schema.

Revised schema:

situation + desire --> action + evaluation
evaluation =: value + results

3. Principle Three

Populate the associative memory system with sequenced rote experiences.

Actions taken, along with their evaluations should become part of the system's repertoire of precedents, to be available for guiding future actions. Storage in memory should be more likely if desires/goals are either met or thwarted, and should be less likely for actions with little or no evaluative content. Storage should also be more likely if expectations from precedents turn out to be violated. Following [Drescher 89], statistical analysis can play an important role in identifying actions that have reliable consequences from the noisy data of actual robot experience. Action sequences should be linked bidirectionally, i.e. with each schema pointing to both the next schema memory, and the previous schema memory. This linking allows control structures (see below) to chain forward from memory items that match current situations, and backward from goals (situations that match desires). Paths that include entities or objects present in the current situation should also be preferred, a kind of "middle-out" chaining.

4. Principle Four

Include mechanisms to automatically generalize across rote memories.

There are several goals here. The main ones are to combine memories of episodes that are frequently repeated, to remove irrelevant preconditions and postconditions (i.e. ones that occur unreliably), and to group actions that lead to the same outcomes (eventually paving the way for "backward chaining"). Memories that have taken part in generalizations can be garbage-collected (possibly probabilistically). Generalized memories, which consist simply of items common to all original memories, can still be matched directly to situations.

5. Principle Five

Include control structures to allow planning.

Two extensions are necessary: 1) associative retrieval of memories must be split apart from the actual performance of their actions; and 2) the retrieved situations and results of memories must be able to be used by the robot as "hypothetical situations" to trigger further retrievals, as though they were actual situations. We can assume that there is an "intention" unit, and that unless this unit is on (activated), retrieved actions will not be actually taken. The intention unit would be on whenever an emergency is detected, when in familiar situations where one precedent dominates all others, etc. Provisions also need to be made to keep track of the search, and to back up and compare values for various possible paths.

Some revised schemas:

situation' + desire(s) --> action' + evaluation'
situation' =: situation | hypothetical-situation
action' + intention --> action

Note that the associative memory can be used as a source of actions to be tried, in order of priority. This in turn can help avoid combinatorial search spaces, since branching factors can be kept much smaller than the total number of schemas.

6. Principle Six

Use specific recognizers ("demons") as censors or primers of memories and actions.

This basically allows generalizations to be very general, while still allowing the robot to cope with exceptions. Specific demons can override general memories, or can "prime" them to make their use more likely. Such demons can be formed whenever expectations are violated, for better or for worse. Here again, the work of Drescher is relevant, in particular in identifying very specific reliable schemas.

7. Principle Seven

Include control structures to encourage only controlled experiments.

If no actions with unknown consequences have been tried recently, make it more likely to try something risky; however, once one or more actions with unknown consequences have been tried, strongly prefer known, reliable schemas until a reward, punishment, or static situation results. This then allows credit to be assigned fairly, since the step/action with the unknown consequences can be assumed to be the likely cause of any surprising outcomes.

8. Principle Eight

In emergencies, or if one schema clearly dominates all others, take the best schema without further evaluation; otherwise, use search to find the best actions.

If there is no danger, and no rush, evaluate schemas by serial exploration of hypothetical actions and situations, and act on the results of this search.

9. Summary

I have argued that associative memory is a useful first-order model of intelligent behavior; in many -- perhaps most -- circumstances, such a system is able to produce appropriate actions. By adding just a few modifications and auxiliary structures one can then account for a wide range of behaviors. Whether or not these ideas have any explanatory power for real intelligence is not clear. However, the proposed model does seem to account concisely for the main phenomena of intelligent action and action selection, a problem generally ignored in AI. (It has been possible to ignore these issues in AI only by sticking with toy problems and microworlds, where the numbers of possible actions or operators that could apply at any given time is small.) The proposed model fits neatly onto the massively parallel machines that seem destined to dominate the high-end supercomputer field for the foreseeable future, and so offers one potential path that can lead to truly intelligent machines.

References

Agre, P. & D. Chapman "Pengi: an implementation of a theory of activity," *Proceedings of the 6th National Conference on Artificial Intelligence*, Los Altos, CA: Morgan Kaufmann, 1987.

Brooks, R. "A robust layered control system for a mobile robot," *IEEE Journal of Robotics and Automation, RA-2*, 1, 1986.

DARPA *Proceedings of the Case-Based Reasoning Workshops*, 1988 (Clearwater Beach, FL) and 1989 (Pensacola Beach, FL).

Drescher, G. "Made-up Minds: a Constructivist Approach to Artificial Intelligence," unpublished Ph.D. dissertation, MIT AI Lab, Cambridge, MA, September 1989.

Halgren, E. "Human hippocampal and amygdala recordings and stimulation: evidence for a neural model of recent memory," In N. Butters & L. Squires (eds.) *The Neurophysiology of Memory*. New York: Guilford, 1984, 165-181.

Hillis, W. D. *The Connection Machine*, Cambridge, MA: MIT Press, 1985.

Holland, J. "A cognitive system with powers of generalization and adaptation," Tech, Report, Dept. of Computer and Communication Sciences, University of Michigan, Ann Arbor, 1977.

Kosslyn, S., R. Flynn, J. Amsterdam, G. Wang, "Components of high-level vision: a cognitive neuroscience analysis and accounts of neurological syndromes," *Cognition 34*, 203-277, 1990.

Maes, P. "How to do the right thing," *Connection Science*, Spring 1990, to appear.

Newell, A. "Unified theories of cognition," William James Lectures, Harvard University, 1987.

Minsky, M. L. *The Society of Mind*, New York: Simon & Schuster, 1987.

Stanfill, C. & D. Waltz, "Toward memory-based reasoning," *Communications of the ACM 29*, 12, 1213-1228, December 1986.

Waltz, D. L. "Massively parallel AI," to appear in *Proceedings of AAAI-90*, Boston, August 1990.

Wilson, S. W. "Classifier systems and the Animat problem," *Machine Learning 2*, 199-228, 1987.