

Proposal for book chapter tentatively entitled “An Architecture for Commonsense Thinking”

27 July 2003

Author: Push Singh <push@mit.edu>

Summary:

Why have AI researchers not been able to give computers human-like ‘common sense’, the ability to think about ordinary things the way people can? In our view, the source of the difficulty is that they too often seek after types of cognitive architectures, kinds of representations, and methods of inference that are based on some single simple process, theory, or principle. Despite their elegance, no single one of such techniques can capture the diversity of mechanisms needed to reason about the broad range of commonsense domains—for example, those that require reasoning about temporal, spatial, physical, psychological, social, and self-reflective matters. Ordinary commonsense thinking spans so many different types of problems and depends on so many forms of knowledge that more unified frameworks, ones that primarily make use of a single type of representation and mode of reasoning, are stretched beyond their capacity. Just as biological systems have no single, simple principle for their operation, we expect that cognitive systems will contain just as numerous and heterogeneous a variety of components.

So rather than seeking a ‘unified theory’, we seek instead to develop an architecture that can support a great diversity of cognitive processes. In this paper we briefly describe aspects of an architecture that we are developing to support the construction of AI systems resourceful enough to combine the advantages of many different ways to think about things, by making use of many types of mechanisms for reasoning, representation, and reflection. The central idea behind our architecture is that the source of human resourcefulness and robustness is the diversity of our cognitive processes: we have many ways to solve every kind of problem—both in the world and in the mind—so that when we get stuck using one method of solution, we can rapidly switch to another.

This chapter will describe an architecture under development by Push Singh and Marvin Minsky, elaborating and extending on the brief description provided in our paper “An Architecture for Combining Ways to Think”, available at

<http://www.media.mit.edu/~push/WaysToThink.pdf>

It will outline both the general structure of the architecture and catalog many of the types of agents that populate it. Some important aspects of the architecture that will be discussed include the following:

- A. The architecture supports many ‘ways-to-think’. Each such way-to-think is a type of agent society designed to solve problems by employing a certain method, for example:

*Knowing How*—We don’t see most problems as problems at all, because we already know how to solve them.

*Analogy*—Try to adapt a method you’ve used before. Few problems ever seem utterly novel because they remind us of similar ones.

*Dividing and Planning*—When we can’t solve a problem all at once, break it down into smaller parts, and regard those parts as separate goals or ‘stepping-stones’.

*Simulation*—Mental experiments in virtual worlds can help when actions are dangerous.

*Proof by Contradiction*—Try to show that your problem cannot be solved, and then look for a flaw in that argument.

*Simplifying*—First ignore the parts of the problem that seem difficult. Then restore them, one at a time. This may not solve the problem itself, but may help us to understand it.

- B. The architecture has multiple reflective layers. Our present design includes the following four layers:

*Reflective*—Thinks about the recent deliberations of deliberative level, such as whether a subgoal has gotten the system closer to a supergoal or further away.

*Self-Reflective*—Thinks about its activities with respect to large-scale models of its abilities and limits, for example, what kinds of things the system knows, how it typically behaves in similar situations, and so forth.

*Self-Conscious*—Thinks about itself in relation to others entities, for example, to compare its own skills and experiences with those of others.

*Self-Ideals*—Thinks about itself with respect to its highest level and longest term goals, perhaps by imagining what one of its ‘imprimers’ (role-models) would think of its activities.

- C. The architecture hosts a vast society of ‘mental critics’, agents whose job is to notice problems within the mind itself, for example:

*Assumed False Preconditions*—An action has failed, and the critic realizes that a precondition for that action did not hold.

*Unable to Decide*—Several methods seem to apply to the current problem, but the system has not decided on one.

*Wasted Reasoning*—While formulating a plan of action, the situation has changed and the problem no longer needs to be solved.

*Lack of Experience*—The system has had only a few experiences dealing with this type of problem.

- D. The architecture supports rapid reformulation between representations so that it is easy to switch between different ways-to-think. Some examples of types of reformulations include:

*Model panalogy*—Maintain descriptions of different models or interpretations of a situation, like seeing a cardboard box as simultaneously a folded up sheet of cardboard and as a rigid cube. Each of these interpretations may suggest different inferences or courses of actions.

*Theory panalogy*—Maintain mappings between different logical theories of the same domain. This may require translation tables that allow descriptions in one representation to be translated into the other. This is similar the notion of contexts as described by McCarthy, which uses ‘lifting rules’ that make explicit the assumptions to add and remove from assertions when transferred it from one context to another.

*Realm panalogy*—Maintain correspondences between different ‘mental realms’. For example, Lakoff and Johnson have argued that the knowledge and skills we have for reasoning about space and time are also used to help reason about social realms, and there are pervasive analogies between these seemingly very different domains.

*Structure panalogy*—Maintain connections between fragments of compositional descriptions, so as to build a larger model from multiple, incomplete partial models. For example, one might approximate a human skeleton with just a dozen bones rather the actual 206 bones of a normal adult, or as a set of sub-skeletal structures consisting of the bones of the head, neck, chest, etc. This is related to Minsky’s frame array idea where many views of an object are linked by their common parts to together form a more realistic or complete model than any individual view could form.

- E. The architecture is designed to operate within multiple reasoning domains. In particular, we have been looking at the kinds of domains that are relevant in simple scenarios involving multiple robots in a physical environment:

*Spatial*—Reasoning about the ways in which objects and the parts of objects are oriented and situated in relation to one another.

*Physical*—Reasoning about the dynamic behavior of real objects with masses and interacting surfaces.

*Bodily*—Reasoning about the capabilities of one’s physical body.

*Visual*—Reasoning about the world that underlies what can be seen.

*Psychological*—Reasoning about one’s goals and beliefs and those of others.

*Social*—Reasoning about the relationships, shared goals and histories that exist between people.

*Reflective*—Reasoning about one’s own recent deliberations.

*Conversational*—Reasoning about how to express one’s ideas to others.