

Review and critique of “The Society Theory of Thinking”

Depei Bao

1. Introduction

The society of mind theory is a model of human intelligence which was developed by Minsky in the early 1970s. This paper is a sequel of his 1974’s paper “A Framework for Representing Knowledge” on the frame system. In 1988, he published the book titled “the Society of Mind”, the first comprehensive description of his theory.

The main idea of the society of mind theory is that human minds work as a society of interacting agents which are mindless. The model is built layer by layer from the interactions of these simple agents. Human intelligence is that they work together.

In his pursuit of human intelligence, the core philosophy is “minds are what brains do”. So he believes that his theory provides a mind model to explain phenomena in developmental, dynamic, and cognitive psychology, and also an approach to realize artificial intelligence.

His methodology in approaching the theoretic model can be summarized in two aspects: for one thing, his theory is neither based on any psychological experiment or brain anatomy nor proves anything specific about AI or cognitive science. Instead, it is a collection of conceptual level ideas which intends to offer an alternative theory for further verification. For another thing, he constructs the model from a mind’s development perspective, i.e., how the model evolves from an early skeleton.

In this summary, I will review ideas in his paper “The Society Theory of Thinking” which mainly focuses on agent organization. The second part will inspect several critical details of his model. In the third and the fourth sections, the strengths and some suspicions are investigated.

2. Overview of the theory details

2.1 Agent organization and communication

The main element for agent organization is c-lines which connect nearby agents together. These c-lines also connect terminals—the object property slots in short term memory, so the property-value needed for the current agent (to manipulate an object) can be found in short term memory. By this arrangement, agents don’t need to send actual values of arguments to other agents. Instead, “each of an agent’s data sources is a fixed location in (short term) memory” (p426). Every c-line has a specific meaning. This means, for example, that different subagent of PUT (an agent for putting a block onto the top of a block tower) will need to know different things

about the ORIGIN (origin of the block), MOVE will need *location-of-origin* and GRASP will need *size-of-origin*. A GREEN-BLOCK object with its properties will be accessed by the meaning of ORIGIN through these c-lines.

By the connection of c-lines between close agents, on lower levels, agents become segregated into smaller and smaller subsocieties. Such gradual decentralization is called *specificity gradient*.

Agents will take inputs from several nearby levels (though the data is not directly sent from other agents). The highest of these enable groups of perhaps competitive agents. Middle levels could be seen as "context" for which of these has priority, and lowest levels as data. Agents whose outputs are above their inputs are useful in analytic recognition to provide data for upper level agents. Agents with outputs below are activating lower-level subprocesses.

2.2 A developmental explanation of the “final performance”

As mentioned before, Minsky’s theory is developed by describing how the final model is evolved from a primitive one with limited performance power. He deems that adult minds and infantile minds are different, and tries to bridge the gap between these two minds by a developmental theory.

In infancy, there is no such hierarchical organization of agents. The system is beginning life with several units and some high level common c-lines (in grossly redundant bundles) for inter-unit communication. These units work together to realize some “innate” functions, with the same computations duplicated in many, nearly parallel, layers. Later, these redundant layers slowly differentiate (function differently), as the interaction between layers is reduced under genetic control. Along with the evolution of communication paths, new agents could arise by splitting off from old ones with nearly same data connections. As development proceeds, these simple connections elaborate into the stratified, hierarchical structure.

What are the strengths

Minsky’s theory offers many innovative ideas both for AI and cognitive science. First, technically, his model is good for the representation issue in problem solving. For a general problem, it's hard to find how abstract or specific to describe problem domains. With such hierarchical organization, the problem will be decomposed from abstract to specific, taken by different agent societies. Second, the theory focuses on how the system evolves through elaborate developmental processes because Minsky believes "only a good theory of the principles of the mind's development can yield a manageable theory of how it finally comes to work"(p429). Third, the great power in viewing a mind as a society of agents, as opposed to as the consequence of some basic principle or some simple formal system, is that different agents can be based on different types of processes with different purposes, ways of representing knowledge, and methods for producing results. This idea is perhaps best summarized by the following quote:

"What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle. – Marvin Minsky, *The Society of Mind*, p. 308"

Actually, Minsky's theory doesn't try to answer what is intelligence. His theory gives models for characterizing the underlying mind mechanism. Intelligence is actually an exhibition of these multiple thinking processes.

What are the possible weaknesses

First, his model raises an interesting mind-body problem, i.e., what the activation of agents embodies? A thinking process or actually doing something? It seems the whole process is unconscious as once an agent is triggered, the muscle is actually controlled to do something. That's like once I decide to put a block onto the tower top, I don't consciously know I need to first GRASP and then move my hand. But, before I really build the tower, I may consciously plan without doing anything. Who will embody the conscious planning process?

The triggering of an agent, say, BUILDER is the emergence of the subject's goal or desire, i.e., wants to build the high tower. But what's the "want" mean? Does the "want" mean I've decided to start doing it or just want to start considering how to do it?

In his paper, he may not think about this problem. When he discussed how to realize the "context shift" of one's attention, he compared it with a recursive program to save its "context" in a push-down stack. But "doing" in a computer program is different from "doing" in human. The "thinking" of a computer program is also the "doing" of it, i.e., running programs to manipulate variables. Here context is clear—the values of variables at a certain moment. But for human, what is the difference between the context shift for thinking and the context shift for doing?

Second, how a single agent works is ambiguous. There are perhaps two alternatives: first, every agent is a planner and has his own reasoning ability. An agent knows the results of other agents, and can reason to give a plan to reach his own goal. Second, every agent is totally mindless, what he knows is just that some other agents activate it and he will activate others according to input c-lines and output c-lines. Intelligence resides in that they work together.

Third, an object is described as a property list by the frame representation. As mentioned before, an agent has a specified meaning to describe an object and its properties, e.g., location of origin. This seems flexible but we may still have various ways to describe the content. For location of origin, on higher level we have many ways to describe the location information. On lower level, like motion control, we don't explicitly describe the location of an object by symbol but we still can get that information for our control. Describing a property value is not unique. How can an agent know how the symbol he gets is represented?

Finally, even follow Minsky's approach of diversity, the intelligence problem remains but it becomes how to choose from these many choices intelligently.