To appear in Visions of Mind, Darryl Davis ed.

An Architecture for Cognitive Diversity

Push Singh and Marvin Minsky

23 April 2004

{push, minsky}@media.mit.edu

Media Lab
Massachusetts Institute of Technology
20 Ames St.
Cambridge, MA 02139
United States

Abstract

To build systems as resourceful and adaptive as people, we must develop cognitive architectures that support great procedural and representational diversity. No single technique is by itself powerful enough to deal with the broad range of domains every ordinary person can understand—even as children, we can effortlessly think about complex problems involving temporal, spatial, physical, bodily, psychological, and social dimensions. In this chapter we describe a multiagent cognitive architecture that aims for such flexibility. Rather than seeking a best way to organize agents, our architecture supports multiple 'ways to think', each a different architectural configuration of agents. Each agent may use a different way to represent and reason with knowledge, and there are special 'panalogy' mechanisms that link agents that represent similar ideas in different ways. At the highest level, the architecture is arranged as a matrix of agents: Vertically the architecture divides into a tower of reflection including the reactive, deliberative, reflective, self-reflective, and self-conscious levels; Horizontally the architecture divides along 'mental realms' including the temporal, spatial, physical, bodily, social, and psychological realms. Our goal is to build an AI system 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.

Keywords

cognitive architecture, commonsense reasoning, human-level intelligence, multiagent systems, multiple representations

1 Introduction

How can we build a machine with the intelligence of a person? There is no shortage of ideas for how to implement in machines aspects of human intelligence, for example, methods for recognizing faces, parsing the syntactic structure of sentences, or planning paths through cluttered spaces. Yet all such techniques fail miserably in comparison to people when it comes to 'common sense' domains—such as recognizing arbitrary objects in arbitrary scenes, answering questions about the simplest children's story, or stuffing a pillow into a pillow case. The problem, as we see it, is that the field of AI has focused on solutions to problems that can be captured in the form of single, simple methods, algorithms, and representations, when in fact the human world is so

varied and complicated that any single such solution fails when presented with problems even slightly different from those they were programmed to handle.

How can we build AI systems that are not so fragile? We believe that to build systems as resourceful and adaptive as people, we must develop cognitive architectures that support great procedural and representational diversity. No single technique is by itself powerful enough to deal with the broad range of domains every ordinary person can understand—even as children, we can effortlessly think about complex problems involving temporal, spatial, physical, bodily, psychological, and social dimensions. Ordinary thinking spans so many different types of problems and depends on so many forms of knowledge that unified frameworks, ones that primarily make use of a single type of representation and technique for inferencing and learning, 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.

The *Society of Mind* theory (Minsky, 1986) presents one possible framework for engineering great cognitive diversity. In this theory the mind is seen as an immense collection of 'agents' that perform a wide range of functions, such as expecting, predicting, repairing, remembering, revising, debugging, acting, comparing, generalizing, exemplifying, analogizing, simplifying, and many other cognitive tasks. Agents are not based on any one principle, but instead employ a great variety of different methods for learning, representation, and reasoning. In fact, the emphasis in the Society of Mind theory is less on the techniques used by any particular type of agent, but instead on how groups of these agents can be organized into communities with more capabilities than any individual agent could possibly have. However, the impact of the Society of Mind theory was mixed—while today there is a thriving field concerned with building complex multiagent systems, few such systems aspire to human-level intelligence.

In this chapter we describe a possible architecture for organizing agents into a flexible, human-like Society of Mind. Rather than seeking a best way to organize agents, our architecture supports multiple 'ways to think', each a different architectural configuration of agents. Each agent may use a different way to represent and reason with knowledge, and there are special 'panalogy' mechanisms that link agents that represent similar ideas in different ways. At the highest level, the architecture is arranged as a matrix of agents: Vertically the architecture divides into a tower of reflection including the reactive, deliberative, reflective, self-reflective, and self-conscious levels; Horizontally the architecture divides along 'mental realms' including the temporal, spatial, physical, bodily, social, and psychological realms. Our goal is to build an AI system 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.

2   Ways of thinking

Our architecture is designed to support a vast diversity of agents, numbering at least in the millions, each roughly on the scale of a small unit of knowledge or subroutine of a computer program. How can we organize a system this large? In our architecture, at any time only a subset of these agents are active—and each of these states produces a specific 'way to think'. This is illustrated in Figure 1 below.
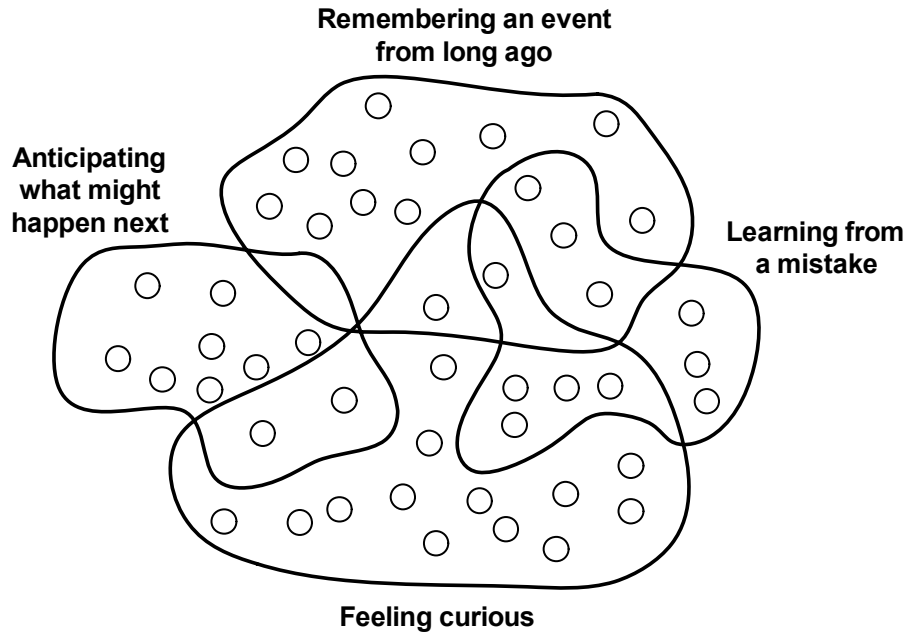
Figure 1. Each way to think results from the activity of a particular subset of mental agents.

In other words, the architecture is not a single kind of 'machine', based on a single type of algorithm or method of reasoning. Instead, in different contexts it transforms into a different machine by switching on different subsets of agents, where the activity of each subset results in a different way of thinking about things. Some examples of these ways to think include:

- Solving problems by making analogies to past experiences (e.g. Carbonell (1986))
- Predicting what will happen next by rule-based mental simulations (e.g. Kuipers (1986))
- Constructing new 'ways to think' by building new collections of agents (e.g. Minsky (1980))
- Explaining unexpected events by diagnosis using causal structures (e.g. Davis (1984))
- Learning from problem-solving episodes by debugging semantic networks (e.g. Winston (1970), Sussman (1973))
- Classifying types of situations using statistical inference (e.g. Pearl (1988))
- Getting unstuck by reformulating the problem situation (e.g. Amarel (1968))

These ways to think are intended to span the full range of AI methods. At the same time, because each of these ways to think is the result of the activity of a set of agents, new ways to think can be formed by assembling together new collections of agents. 'Ways to think' are an evolution of the *K-lines* idea from Minsky's Society of Mind theory (Minsky, 1980). K-lines are special agents whose primary job is to switch on other sets of agents. This provides a simple but effective mechanism for disposing a mind towards engaging relevant kinds of problem solving strategies, retrieving particular fragments of knowledge, selecting or prioritizing sets of goals, invoking memories of particular experiences, and bringing to bear other mental resources that might help in coping with a problem. Each way to think is more or less self-contained, and the mind can be seen as a distributed collection of such ways of thinking with no 'central control'—the decentralized vision of cognition presented in the *Society of Mind*.

What controls which ways to think are active at any moment, and when to switch to new ways to think? In our architecture, there are special 'critic' agents concerned primarily with selecting ways to think. At the highest level these critic agents can be regarded as chronic or persistent questions and concerns, for example:

- What will happen next following this event?
- What would explain why this event occurred?
- What is the best thing for me to do now?
- What can I learn from this failure?
- What might go wrong while performing this action?
- What could be the negative consequences of taking this action?
- Why is that person taking that action?

Each of these mental questions leads to other questions and ways of thinking that can attempt to address them. If we wish to predict what might happen next in a situation, we may try to remember what happened next in a similar situation in the past. If we wish to learn from a failure, we may initiate a credit assignment process that traces back along the causal dependencies among recent events. And so forth.

When a way of thinking begins to fail, the architecture can switch to another more appropriate way to think. This happens through the operation of critic agents that recognize not problems in the outside world, but rather classes of failures and impasses within the mind itself. When such an impasse is detected, these critics can select alternative ways to think, as shown in Figure 2 below.
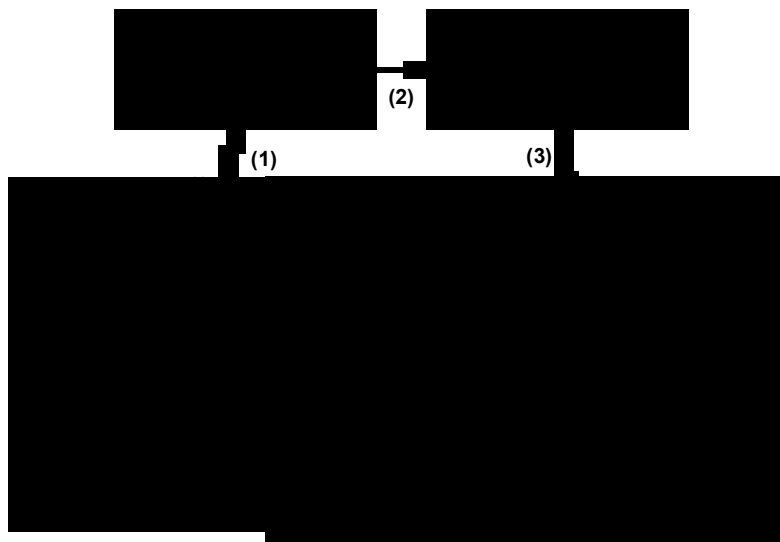
Figure 2. When one way of thinking is beginning to fail, mental critics recognize the failure and select alternative ways to think.

3   Multiple representations

When one way to think becomes ineffective the architecture tries to switch to another. Normally, this would require a certain 'start up time' where the agents of the new way to think gather the information they need to do their jobs. However, our architecture performs transitions more efficiently by having special support for multiple representations, to allow agents that represent similar information to easily synchronize what they know. When an agent writes to a

representation, it updates the representations of related agents in parallel, including the ones used by agents that are at the moment quiescent. Thus, when the architecture selects a new way to think, instead of having to start from scratch, it finds many of its agents to be already prepared for the situation.

We do not use any single technique for coordinating representations across multiple agents, and instead make use of a family of processes for synchronizing and sharing information. We refer to these together as *panalogy* (a term that derives from 'parallel analogy'). Here are some of the methods of panalogy we use:

Event panalogy. Maintain the correspondences between the elements of action and event descriptions across multiple representations. For example, when we imagine the consequences of buying a fancy new car, we can rapidly switch between considering the effects of that purchase on our social status (which it may improve) and on our financial situation (which it may hurt.) This form of panalogy let us assess the consequences of an action or event from a great many different perspectives at once—for in the ordinary, common sense world, actions and events usually have a wide range of important physical, social, psychological, economic, and other types of consequences.

Model panalogy. Maintain descriptions of different models or interpretations of a situation, like seeing a window simultaneously as both an obstacle and as a portal. Each of these interpretations may suggest different inferences or courses of actions, and if we discover that in fact the window is not locked, inferences based on the 'portal' interpretation are already available for use. This form of panalogy is valuable because it takes advantage of the notion that a problem often becomes trivial when we look at it from just the right perspective. A planning problem represented one way might require an immense amount of search, but when represented in another way might be solved by simple hill climbing.

Theory panalogy. Maintain mappings between different theories of the same domain. For example, we may choose to use one theory of time where events are treated as atomic points on a timeline, or use another theory of time where events are treated as occurring over intervals on a timeline. When the first theory is unable to answer a question about, for example, the total duration of some set of actions or the order in which they occurred, we might switch to the second theory. This form of panalogy is useful because it is difficult to find the 'best' way to represent fundamental commonsense subjects such as space, time, causality, goals, and so forth. We argue instead that there is no best 'upper level ontology' for describing such entities, and that we should instead employ multiple theories about foundational matters.

Realm panalogy. Maintain analogies between different 'mental realms', large-scale commonsense domains such as the spatial, temporal, and social realms. Lakoff and Johnson (1990) have argued for example that the knowledge and skills we use for reasoning about space and time are also used to help reason about social realms, for in language there are pervasive metaphors that exist between these seemingly very different domains. This form of panalogy is important because it is clear from language that it is possible to exploit such metaphors to simplify communication about abstract matters, and we suspect that such metaphors may serve similar roles within the mind as well (see Boroditsky (2000) for some recent evidence that temporal ideas have their roots in spatial notions.)
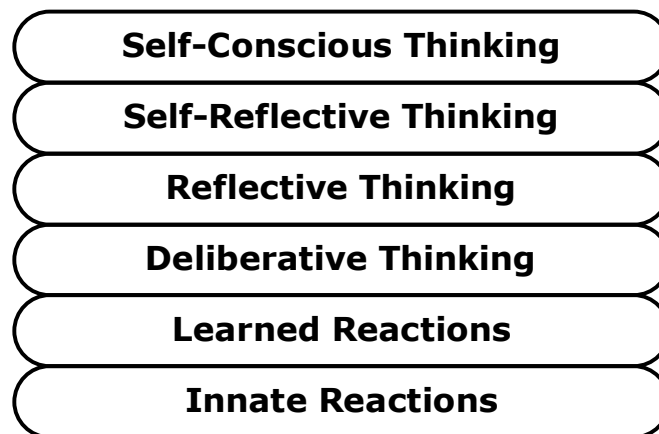
Abstraction panalogy. Maintain connections between different abstract descriptions. For example, one might approximate a human skeleton with just a dozen limbs rather than the actual 206 bones of a normal adult, or focusing on particular sub-skeletal structures such as the bones of

the right leg. Each of these different abstractions can linked by their common parts to together form a more realistic or complete model than any individual abstraction could form. This form of panalogy is powerful because it lets us link together a variety of 'simplifications' of a situation, each useful for a different type of problem. If we are trying to grasp a pair of scissors it may be useful to think about each of our fingers separately, but if we are trying to push closed a heavy door we may instead think of the palm of our hand and its five fingers as a single unit that applies pressure to the door.

Ambiguity panalogy. Maintain links between ambiguous senses of predicates. For example, the preposition 'in' can refer to a wide range of relations far more specific than any division provided by ordinary dictionary senses. Rather than selecting any particular such relation when describing a situation, we can instead maintain the ambiguity between those relations, which then lets us draw on our understanding of all those related senses to answer questions about how one thing could be 'in' another. This form of panalogy lets us bypass one of the basic difficulties in building symbolic systems—namely, that it is incredibly challenging and perhaps impossible to define any given symbol precisely enough that we and others will use it only as intended in the future. Just as the meanings of words evolve with their use, and quickly come to acquire multiple new senses in different contexts, so should the meanings of symbols.

4   Multiple layers of reflection

In any multiagent system that regularly faces new situations; the existing community of agents will sooner or later run into problems. An important feature of our architecture is that it designed to be highly self-reflective and self-aware, so that it can recognize and understand its own capabilities and limitations, and debug and improve its abilities over time. In contrast, most architectural designs in recent years have focused mainly on ways to react or deliberate—with no special ability to reflect upon their own behavior or to improve the way they think about things. In our architecture, agents are organized into a tower of reflection consisting of six layers, as shown in Figure 3 below.

Self-Conscious Thinking
Self-Reflective Thinking
Reflective Thinking
Deliberative Thinking
Learned Reactions
Innate Reactions

Figure 3. The agents of our architecture are divided into layers, each managing and reflecting upon the layers beneath.

Each of these layers is responsible for recognizing and responding to problems within the agents in the layers beneath. To do this, we again make use of special 'critic' agents that bridge these layers. The job of critics is to notice problems in the agents in the layers beneath, and select ways

to deal with those problems. The function of each layer and examples of the critics that populate each layer are described below:

Innate and learned reactions. Instinctive reflexes and learned responses to opportunities and emergencies in the world. These consist of critics that detect specific types of problems in the world and switch on ways to react to those problems. Much of the behavior of animals can be described by networks of such critics. For example:

- *I hear a loud noise → Move to a quieter place*
- *I feel hungry → Follow the smell of food*
- *I am far from something I want → Walk towards it*
- *I feel scared → Run quickly to a safe place*

Deliberative thinking. When faced with a difficult problem, it is useful to build a model of the situation in our minds, for example, as a network of goals, actions, and their effects, in which we can search for a solution. The agents of the deliberative layer reason about the situation by engaging in various types of mental deliberation, for example, prediction, explanation, planning, diagnosis, generalization, and so on. Even the simplest problems may result in large search spaces, and deliberative critics help us search those mazes more effectively:

- *Action A did not quite achieve my goal → Try harder, or try to find out why*
- *Action A worked but had bad side effects → Try some variant of that action*
- *Achieving goal X made goal Y harder → Try them in the opposite order*
- *These events do not chain → Change one of their end points to match*

Reflective thinking. When faced with a hard problem that we are not making much progress on, we may need to reflect on the techniques that we are using to solve that problem. This may involve activities such as assigning credit for success or failure to particular inference methods or types of knowledge, selecting or modifying the knowledge representation structures we have been using, and so forth. Reflective critics assess the performance of recent deliberations in this way, and suggest high level changes to the way we are approaching the current situation:

- *The search has become too extensive → Find methods that yield fewer alternatives*
- *You have tried the same thing several times → Some manager agent is incompetent*
- *You overlooked some critical feature → Revise the way you described the problem*
- *You cannot decide which strategy to use → Formulate this as a new problem*

Self-reflective thinking: When reflecting on the methods we use fails to help, we may criticize ourselves. The self-reflective layer is concerned with large-scale models of the "self", including the extent and boundaries of one's physical and cognitive abilities and knowledge. Self-reflective critics look for highly entrenched long-standing deficiencies and weaknesses in our knowledge and methods, and suggest significant courses of action to deal with such problems:

- *I missed an opportunity by not acting quickly enough → Set up a mental alarm that warns me whenever I am about to do that*
- *I can never get this exactly right → Spend more time practicing that skill*
- *I let my other interests take control → Tell one of my friends to scold me when I get distracted*
- *I do not seem to have the knowledge I need → Quit this and go to graduate school*

Self-conscious thinking. It is occasionally useful to imagine what others might think of our activities, and how others might approach these same problems. This layer is concerned with the relationship between one's mind and those of others, and performs self-appraisals by comparing one's abilities and goals with those of others. Self-conscious critics resemble self-reflective critics, but operate at a more social level by imagining what others, and especially people whom we respect, might think of us:

- *I think I am good at this task → Can I do it as well as the best people I know?*
- *My mentor would not have made this mistake → What would he have done in this situation?*
- *Others will think less of me if I keep failing at this → Maybe I should give up doing this sort of thing*
- *How is it that other people can solve this problem? → Find someone good at this problem and spend time with them*

By employing multiple layers of mental critics, we need not build architectures under the impossible constraint that agents always produce the correct inference or perfect suggestion for a course of action. Instead, when the architecture fails, it can examine its own recent activity and self-models to attempt to diagnose and deal with the problem, so that next time it does not make the same type of mistake.

5 The many realms of common sense thought

It is not enough for our architecture to possess many mechanisms for representation, reasoning, and reflection. In addition, it must actually *know* things to cope with the great complexity of the human world. In our view, an architecture that comes with no knowledge is much like a programming language that comes with no libraries or example programs—it is very difficult to get started with it or put it to practical use. What sorts of knowledge should a commonsense architecture possess? If one stops to think about the range of types of things that people know about and the kinds of problems people can solve, it is clear that the list of enormous, and at first glance may seem to consists of an entirely haphazard collection of knowledge and skills.

However, while the range of things that an adult human knows about is vast, there is a much more limited class of things that we can expect all people to be able to think about, and especially, the average young child to be able to think about. If we limit the scope of our study in this way, we can approach more systematically the problem of determining what our architecture should know about, how to represent that knowledge, and how to teach it that knowledge. We have been enumerating a list of *mental realms*, the general commonsense domains that all people including children have at least some expertise in. We regard these mental realms as so fundamental that it would be reasonable to regard the inability to reason in terms of one these mental realms as a serious cognitive deficiency.

What are some of the important mental realms? We do not yet have a well-defined, definite list of such realms, but the following are good examples of what we mean by a realm:

- **Spatial**: The spatial realm is concerned with representing the shapes, relative positions and orientations of places, objects, and their parts. It is also concerned with the motion of objects and the paths that they take through space, as well as the relative spatial relationships between objects as they move above. It is the knowledge and processes of the spatial realm we use when solving problems like determining whether objects are close enough to reach,

whether it is possible for us to squeeze through a narrow passageway, or how to fit several pieces of wood together to build a table.

- **Physical**: The physical realm is concerned with representing the dynamic behavior of real objects, such as how different objects respond to various forces, perturbations, and other physical interactions. We all know, for example, that you can push things with a stick, but cannot pull them with a stick—unless the end of the stick is somehow 'attached' to the object we are pulling, or unless the end of the stick is curled into a hook, which lets us convert a 'pull' into a 'push' on the other side.

- **Bodily**: The bodily realm is largely concerned with representing the abilities of your body, such as how far you can reach in different directions from different initial postures, what procedure you should follow to grasp an object of a given shape, or whether you are strong enough to pick up a particularly large object. The bodily realm, combined with the spatial and physical realm, constitutes much of the knowledge a humanoid robot would need to get around in the world and physically manipulate the objects it encounters, such as putting a pillow in a pillow case, hanging a set of curtains, or throwing a tennis ball to another robot.

- **Social**: The social realm is concerned with representing the relationships, mutual dependencies, and interactions that exist and occur between social entities such as people and animals. This includes matters such as whether your goals are compatible with the goals of others, keeping track of the people you know and the experiences you have shared with them, predicting how someone you know might behave in different situations, and so on. It is the social realm that lets us infer, for example, that someone who is laughing and smiling while talking to someone else is probably enjoying spending time with them.

- **Psychological**: The psychological realm is concerned with representing matters of our own psychology, such as how long it takes us to learn some new subject, whether we are capable of arguing some point or whether we must admit ignorance, the goals that we presently have and their relative priorities, and so on. The psychological realm is about the many types of problems that occur within our own minds. In our view very little is known about this realm in comparison to many of the other realms we have discussed, simply because to understand this realm well requires in itself a detailed architectural model of the mind.

These are just a few of the realms we have been exploring. We have found it useful to further subdivide these realms into more specialized subrealms concerned with more specific matters. For example, in the bodily realm, we may choose to separate knowledge about how to manipulate objects dexterously with our hands from knowledge about how to use our legs to walk over complex terrains.

Each of these realms involves a substantial amount of knowledge. However, realms are distinguished not only by the knowledge they include, but also by the methods of reasoning they support. Problems within certain realms may for efficiency use specialized representations and reasoning methods. For example, in the spatial realm, a specialized planner that is designed for a three dimensional Euclidean space may be more suitable to solving spatial path planning problems than some more general technique that can search some arbitrary search space.

The notion of mental realms has helped to organize the types of commonsense knowledge that our architecture will need. It has also been very useful for educational purposes. To the beginner, the idea of 'common sense' can seem vague and undifferentiated, yet after we introduce to them

the concept that common sense can be divided and organized into large-scale specialties, they often seem to better understand what we mean by common sense. It should be noted, however, that we have not been using the notion of mental realms in any definite technical sense—that is, a realm does not refer to any particular computational object, except very loosely as that set of representations and processes that are used to cope with a certain wide class of problems.

6   The realm-layer matrix

We have found it to be useful to merge the previous two ideas of reflective layers and mental realms into the matrix shown in Figure 4 below. Each cell of the matrix consists of populations of agents that think about a certain mental realm at a certain level of reflection. While there are problems with this diagram—for example, what does it mean for there to be agents in the instinctive-psychological cell?—we have found that more often than not that there seem to be interesting processes at play in each of these cells.
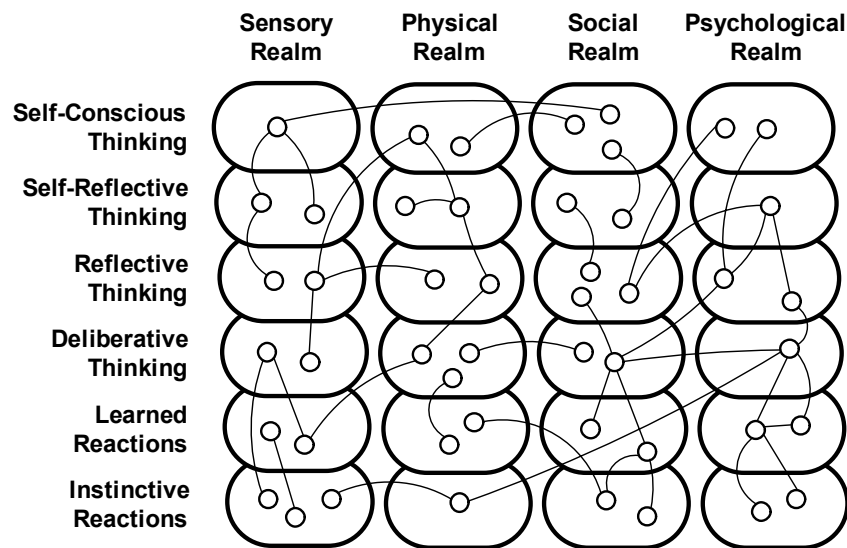

Figure 4. The architecture can be divided into a matrix of cells, for example, the physical-deliberative cell or the social-reflective cell.

Let us examine a single vertical slice of the architecture, for example, the social realm. At the lowest reactive levels there are processes for recognizing that someone is smiling at you, for smiling back at them, and so on. At the deliberative level there may be models of how people react to different sorts of social actions, which includes knowledge such as someone who smiles kindly at you probably has no malicious intent, or perhaps recognizes you. At the reflective level there may be processes for understanding why we made a social mistake about classifying our relationship with someone else—for example, it is not always the case that someone who smiles at you knows you, but had you jumped to that conclusion by mistake, and in fact they were trying to introduce themselves to you. At the self-reflective level you may decide that you are no good at remembering people's faces and need to do something about that problem. At the self-conscious level you may decide that the other person thinks less of you for making that mistake, resulting in a feeling of mild embarrassment.

7   An Example Scenario

We are developing a concrete implementation of this architecture in the context of an 'artificial life' scenario where two simulated people in a virtual world work together to build complex structures from simple objects like sticks, balls, and blocks, as in the simulator screenshot shown in Figure 5 below.
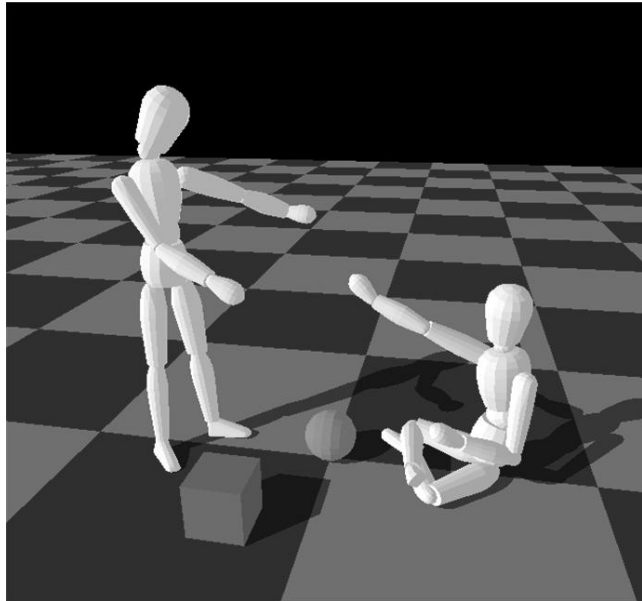
Figure 5. A simulated world

While this domain may seem sparse, its simplicity hides a great depth of issues. In particular, the mental realms we have discussed so far all show up in some form in this domain. Because the world is physically realistic, the people must reason about the effects of gravity on objects and the forces that must be applied to move them. Because the people have synthetic vision systems, they must reason about whether objects that seem to have disappeared behind bigger ones are in fact really still there. Because there are two people, they must reason about the social challenges that arise between them, such as conflicts between their goals and possible opportunities for cooperation. To solve problems in this world requires reasoning simultaneously about the physical, social, psychological, and several other mental realms.

Consider the simple scenario shown in Figure 6, depicting two people named Alpha and Beta working together to build a tower. Let us examine Alpha's thoughts during the first two frames of this situation, where it reaches for a block, fails, then realizes that Beta may be able to help. Even this seemingly simple problem requires commonsense reasoning across multiple realms and multiple levels: reactive, deliberative, and reflective processes across the physical, bodily, spatial, perceptual, and social realms:
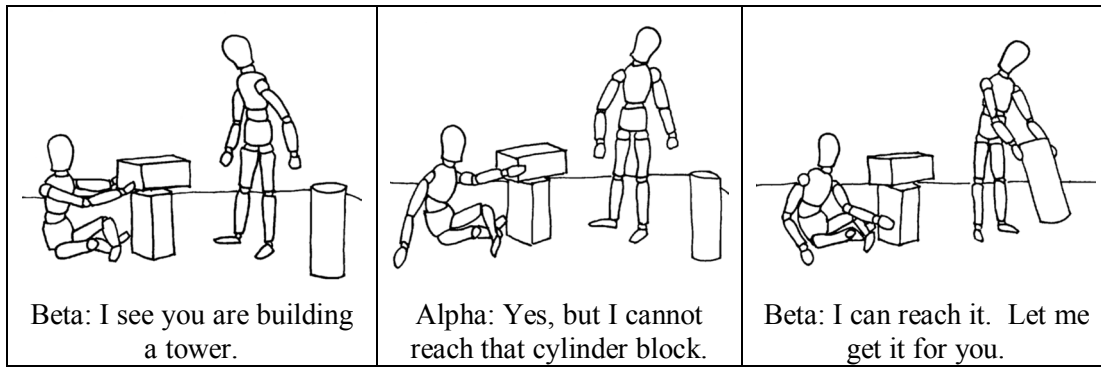
Figure 6. Alpha (left) builds a tower with Beta (right)

Frame 1: Alpha wants to build a tower three blocks high. It tries and fails to reach the cylinder block.

- *Deliberative Physical (Goals)*: I want to build a tower of blocks three high.
- *Deliberative Physical (Situation)*: I have built a tower two blocks high.
- *Self-Reflective Physical (Known Methods)*: What do I know about assembling structures?
- *Deliberative Ownership (Goals)*. I need to obtain a new block to build a tower three high.
- *Reflective Priorities (Expectation)*: If I am building a tower then I will not be able to achieve my other goals.
- *Deliberative Physical (Situation)*: There are three blocks in this room.
- *Deliberative Resources (Situation)*: Those three blocks are ingredients for the desired tower.
- *Reflective Visual & Spatial (Question)*: Perhaps there is another hidden block?
- *Reflective Procedural (Memory)*: I have never built a tower exactly three blocks high.
- *Reflective Debugger (Method)*: Abstracting my goal may result in finding a suitable method.
- *Reflective Debugger (Method)*: Replace 'three' by 'several'.
- *Deliberative Spatial (Expectation)*: Placing a block on top of a tower will make the tower higher.
- *Deliberative Bodily (Situation)*: There is a cylinder block nearby that I can possibly reach.
- *Deliberative Spatial (Expectation)*: I might not be able to reach that cylinder block.
- *Reactive Bodily (Action)*: Produce appropriate muscle actuations to produce desired hand trajectory.
- *Deliberative Bodily (Situation)*: My arm is at full length and I do not have the cylinder block in hand.
- *Reflective Bodily (Critic)*: The current method has completely failed.
- *Self-Reflective Bodily (Method)*: Find another method.

Frame 2: Alpha sees Beta and asks for help.

- *Self-Reflective Bodily (Critic)*: No method is available for easily reaching that block from here.
- *Self-Reflective Social (Selector)*: Switch to social way of thinking for Obtaining Help.

- *Deliberative Spatial (Situation)*: Beta is nearby.
- *Deliberative Social (Situation)*: Beta may be able to help me.
- *Deliberative Spatial (Inference)*: Beta seems to be near enough to the block to reach it.
- *Deliberative Communication (Method)*: Ask Beta for help.
- *Deliberative Psychological (Goal)*: Beta may have other things to do.
- *Deliberative Social (Situation)*: Beta may want that block for itself.
- *Deliberative Linguistic (Situation)*: Beta says it notices I am building a tower.
- *Deliberative Social (Inference)*: Beta understands my larger goal.
- *Deliberative Social (Inference)*: It would cost Beta very little to help me.
- *Reactive Linguistic (Action)*: Say "I cannot reach that cylinder block".
- *Reactive Gestural (Action)*: Point at cylinder block.

So we see that even this simple seemingly trivial exchange exercises many of the cells our matrix of commonsense agents. The details here are greatly simplified—every step involves many more agents than are listed, and many iterations of thought must be involved in producing and refining the solutions to the subproblems encountered by those agents.

8   Conclusions

This chapter elaborates on a similar discussion in Singh and Minsky (2003). More details about our architectural design are available in Minsky's book *The Emotion Machine* (forthcoming), in McCarthy et al. (2002), and in Minsky *et al.* (forthcoming).

One might question the need for an architecture with its many lists, catalogs, and other accumulations of kinds of components and features. There will surely be those who find such approaches inelegant, and instead would prefer something simpler, perhaps based on some new mathematical principle or universal method of learning or reasoning. But our view is that any approach that seeks to build something as complex as a human mind will need to consist of a great accumulations of representations and methods, for many of the same reasons that a typical modern computer requires many thousands of small files and programs to operate. Can we realistically expect something comparable to the human mind to be reduced to some simple algorithm or principle given the range of things it must be able do? We hope that our architectural design will change the way AI researchers picture what an AI system should look like, and convince people to value systems less on some ethereal notion of elegance and more based on their speed, flexibility, and all-around resourcefulness.

Acknowledgements

References

Saul Amarel (1968). *On representations of problems of reasoning about actions*. In Michie, D. (Ed.) Machine Intelligence, 3(3):131—171. Elsevier.

Lera Boroditsky (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition,* 75(1):1—28.

Jaime Carbonell (1986). Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In Michalski, R., Carbonell, J., & Mitchell, T. (Eds.) *Machine Learning: An Artificial Intelligence Approach.* Morgan Kaufman Publishers: San Mateo, CA.

Randall Davis (1984). Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, 24:347—410.

Benjamin Kuipers (1986). Qualitative simulation. *Artificial Intelligence*, 29:289—338.

George Lakoff and Mark Johnson (1990). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

John McCarthy (1993). Notes on formalizing context. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI 1993)*. Avignon, France.

John McCarthy, Marvin Minsky, Aaron Sloman, Leiguang Gong, Tessa Lau, Leora Morgenstern, Erik Mueller, Doug Riecken, Moninder Singh, and Push Singh (2002). An architecture of diversity for commonsense reasoning. *IBM Systems Journal,* 41(3):530—539.

Marvin Minsky (1980). K-lines, a theory of memory. *Cognitive Science*, 4: 117—133.

Marvin Minsky (1986). *The society of mind.* New York: Simon and Schuster.

Marvin Minsky (forthcoming). *The emotion machine*.

Marvin Minsky, Push Singh, and Aaron Sloman (forthcoming). The St. Thomas commonsense symposium. *AI Magazine.*

Judea Pearl (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.

Push Singh (2003). A preliminary collection of reflective critics for layered agent architectures. *Proceedings of the Safe Agents Workshop (AAMAS 2003*). Melbourne, Australia.

Push Singh and Marvin Minsky (2003). An architecture for combining ways to think. *Proceedings of the International Conference on Knowledge Intensive Multi-Agent Systems*. Cambridge, MA.

Gerald J. Sussman (1973). *A computational model of skill acquisition* (Phd thesis). Department of Mathematics, MIT.

Patrick H. Winston (1970). *Learning structural descriptions from examples* (Phd thesis). Department of Electrical Engineering, MIT.

Author Information

Push Singh

Push Singh is a doctoral candidate in MIT's department of Electrical Engineering and Computer Science. His research is focused on finding ways to give computers human-like common sense, and he is presently collaborating with Marvin Minsky to develop an architecture for commonsense thinking that makes use of many types of mechanisms for reasoning, representation, and reflection. He started the Open Mind Common Sense project at MIT, an effort to build large-scale commonsense knowledge bases by turning to the general public, and has worked on incorporating commonsense reasoning into a variety of real-world applications. Singh received his B.S. and M.Eng. in Electrical Engineering and Computer Science from MIT.

Marvin Minsky

Marvin Minsky has made many contributions to AI, cognitive psychology, mathematics, computational linguistics, robotics, and optics. In recent years he has worked chiefly on imparting to machines the human capacity for commonsense reasoning. His conception of human intellectual structure and function is presented in *The Society of Mind* which is also the title of the course he teaches at MIT. He received the BA and PhD in mathematics at Harvard and Princeton. In 1951 he built the SNARC, the first neural network simulator. His other inventions include mechanical hands and other robotic devices, the confocal scanning microscope, the "Muse" synthesizer for musical variations (with E. Fredkin), and the first LOGO "turtle" (with S. Papert). A member of the NAS, NAE and Argentine NAS, he has received the ACM Turing Award, the MIT Killian Award, the Japan Prize, the IJCAI Research Excellence Award, the Rank Prize and the Robert Wood Prize for Optoelectronics, and the Benjamin Franklin Medal.