The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence

Marvin Minsky, Push Singh, and Aaron Sloman

■ To build a machine that has "common sense" was once a principal goal in the field of artificial intelligence. But most researchers in recent years have retreated from that ambitious aim. Instead, each developed some special technique that could deal with some class of problem well, but does poorly at almost everything else. We are convinced, however, that no one such method will ever turn out to be "best," and that instead, the powerful AI systems of the future will use a diverse array of resources that, together, will deal with a great range of problems. To build a machine that's resourceful enough to have humanlike common sense, we must develop ways to combine the advantages of multiple methods to represent knowledge, multiple ways to make inferences, and multiple ways to learn. We held a two-day symposium in St. Thomas, U.S. Virgin Islands, to discuss such a project—to develop new architectural schemes that can bridge between different strategies and representations. This article reports on the events and ideas developed at this meeting and subsequent thoughts by the authors on how to make progress.

The Need for Synthesis in Modern AI

To build a machine that has "common sense" was once a principal goal in the field of artificial intelligence. But most researchers in recent years have retreated from that ambitious aim. Instead, each developed some special technique that could deal with some class of problem well, but does poorly at almost everything else. An outsider might regard our field as a chaotic array of attempts to exploit the advantages of (for example) neural networks, formal logic, genetic programming, or statistical inference—with the proponents of each method maintaining that their chosen technique will someday replace most of the other competitors.

We do not mean to dismiss any particular technique. However, we are convinced that no one such method will ever turn out to be "best," and that instead, the powerful AI systems of the future will use a diverse array of resources that, together, will deal with a great range of problems. In other words, we should not seek a single "unified theory!" To build a machine that is resourceful enough to have humanlike common sense, we must develop ways to combine the advantages of multiple methods to represent knowledge, multiple ways to make inferences, and multiple ways to learn.

We held a two-day symposium in St. Thomas, U.S. Virgin Islands, to discuss such a project—to develop new architectural schemes that can bridge between different strategies and representations. This article reports on the events and ideas developed at this meeting and subsequent thoughts by the authors on how to make progress.[1]

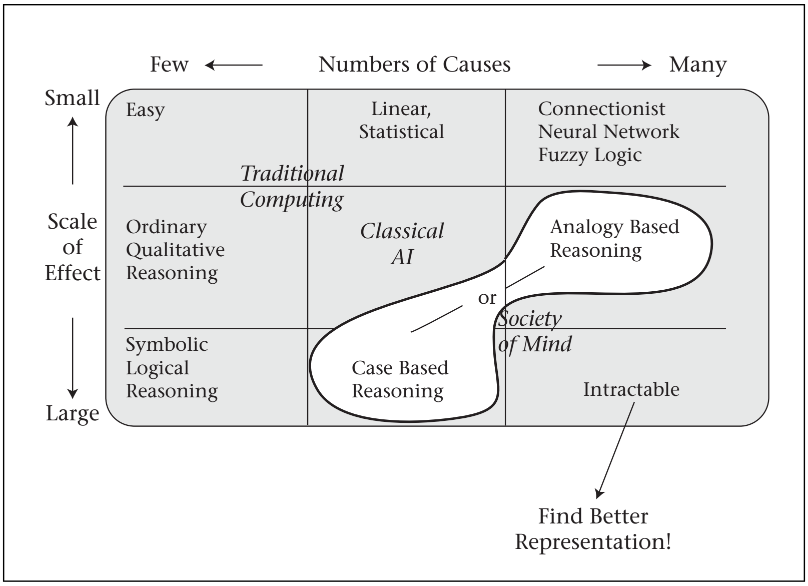


Figure 1. The Causal Diversity Matrix.

Each common AI technique is matched to problem-types with particular causal structures. Note that this is just a first draft of one idea about describing what each AI technique can do—each reader may have a different view about what their own favorite method can do, and whether these “axes” make sense for them. (Diagram adapted from Minsky’s *Future of AI Technology.*)

Organizing the Diversity of AI Methods

Marvin Minsky kicked off the meeting by discussing how we might begin to organize the many techniques that have been developed in AI so far. While AI researchers have invented many representations, methods, and architectures for solving many types of problems, they still have little understanding of the strengths and weaknesses of each these techniques. We need a theory that helps to map the types of problems we face onto the types of solutions that are available to us. When should one use a neural network? When should one use statistical learning? When should one use logical theorem proving?

To help answer these kinds of questions, Minsky suggested that we could organize different AI methods into a “causal diversity matrix” (figure 1). Here, each problem-solving method, such as analogical reasoning, logical theorem proving, and statistical inference, is assessed in terms of its competence at dealing with problem domains with different causal structures.

Statistical inference is often useful for situations that are affected by many different matched causal components, but where each contributes only slightly to the final phenomenon. A good example of such a problem-type is visual texture classification, such as determining whether a region in an image is a patch of skin or a fragment of a cloud. This can be done by summing the contributions of many small pieces of evidence such as the individual pixels of the texture. No one pixel is terribly important, but en masse they determine the classification. Formal logic, on the other hand, works well on problems where there are relatively few causal components, but which are arranged in intricate structures sensitive to the slightest disturbance or inconsistency. An example of such a problem-type is verifying the correctness of a computer program, whose behavior can be changed completely by modifying a single bit of its code. Case-based and analogical reasoning lie between these extremes, matched to problems where there are a moderate number of causal components each with a modest amount of influence. Many common sense domains, such as human social reasoning, may fall into this category. Such problems may involve knowledge too difficult to formalize as a small set of logical axioms, or too difficult to acquire enough data about to train an adequate statistical model.

It is true that many of these techniques have worked well outside of the regimes suggested by this causal diversity matrix. For example, statistical methods have found application in realms where previously rule-based methods were the norm, such as in the syntactic parsing of natural language text. However, we need a richer heuristic theory of when to apply different AI techniques, and this causal diversity matrix could be an initial step toward that. We need to further develop and extend such theories to include the entire range of AI methods that have been developed, so that we can more systematically exploit the advantages of particular techniques.

How could such a “meta-theory of AI techniques” be used by an AI architecture? Before we turned to this question, we discussed a concrete problem domain in which we could think more clearly about the goal of building a machine with common sense.

Returning to the Blocks World

Later that first morning, Push Singh presented a possible target domain for a commonsense architecture project. Consider the situation of two children playing together with blocks (figure 2).

Even in this simple situation, the children may have concerns that span many “mental realms”:

Physical: What if I pulled out that bottom block?

Bodily: Can I reach that green block from here?

Social: Should I help him with his tower or knock it down?

Psychological: I forgot where I left the blue block.

Visual: Is the blue block hidden behind that stack?

Spatial: Can I arrange those blocks into the shape of a table?

Tactile: What would it feel like to grab five blocks at once?

Self-Reflective: I'm getting bored with this—what else is there to do?

Singh argued that no present-day AI system demonstrates such a broad range of commonsense skills. Any architecture we design should aim to achieve some competence within each of these and other important mental realms. He proposed that to do this we work within the simplest possible domain requiring reasoning in each of these realms. He suggested that we develop our architectures within a physically realistic *model world* resembling the classic Blocks World, but where the world was populated by several simulated beings, and thus emphasizing social problems in addition to physical ones. These beings would manipulate simple objects like blocks, balls, and cylinders, and would participate in the kinds of scenarios depicted in figure 3, which include jointly building structures of various kinds, competing to solve puzzles, teaching each other skills through examples and through conversation, and verbally reflecting on their own successes and failures.

The apparent simplicity of this world is deceptive, for many of the kinds of problems that show up in this world have not yet been tackled in AI, for they require combining elements of the following:

Spatial reasoning about the spatial arrangements of objects in one's environment and how the parts of objects are oriented and situated in relation to one another. (Which of those blocks is closest to me?)

Physical reasoning about the dynamic behavior of physical objects with masses and colliding/supporting surfaces. (What would happen if I removed that middle block from the tower?)

Bodily reasoning about the capabilities of one's physical body. (Can I reach that block without having to get up?)

Visual reasoning about the world that underlies what can be seen. (Is that a cylinder-shaped block or part of a person's leg?)

Psychological reasoning about the goals and beliefs oneself and of others. (What is the other person trying to do?)

Social reasoning about the relationships, shared goals and histories that exist between people. (How can I accomplish my goal without the other person interfering?)

Figure 2. A Pair of Busy Youths.

Reflective reasoning about one's own recent deliberations. (What was I trying to do a moment ago?)

Conversational reasoning about how to express one's ideas to others. (How can I explain my problem to the other person?)

Educational reasoning about how to best learn about some subject, or to teach it to someone else. (How can I generalize useful rules about the world from experiences?)

Many of the meeting participants were enthusiastic about this proposal and agreed that there would be challenging visual, spatial, and robotics problems within this domain. Ken Forbus pointed out that the video game communities would soon produce programmable virtual worlds that would easily meet our needs. Several participants mentioned the success of the RoboCup competitions (Kitano et al. 1997), but some concluded that the RoboCup domain, while appropriate for those interested in the problem of coordinating multiagent teams in a competitive scenario, was very different in character from the situation of two or three people more slowly working together on a physical task, communicating in natural language, and in general operating on a more thoughtful and reflective level.

Establishing a Collection of Graded Miniscenarios

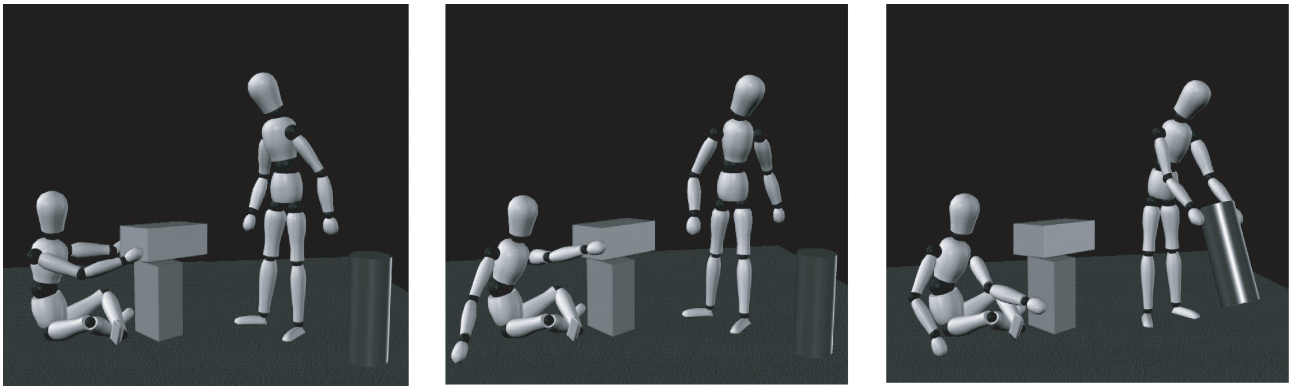
How would we guide such a project and measure its progress over time? Some participants suggested trying to emulate the abilities of human children at various ages. However, others argued that while this should inspire us, we should not use it as a plan for the project, because we don't really yet know enough about the details of early human mental development.

Aaron Sloman argued that it might be better to try to model the mind of a four- or five-year-old human child because that might lead more directly toward more substantial adult abilities. After the meeting, Sloman developed the notion of a "commonsense miniscenario," a concrete description in the form of a simple storyboard of a particular skill that a commonsense architecture should be able to demonstrate. Each miniscenario has several features: (1) It describes some forms of competence, which are robust insofar as they can cope with wide ranges of variation in the conditions; and (2) each comes with some meta-competence for thinking and speaking about what was done. For example competence can have a number of different facets, including describing the process; explaining why something was done, or why something else would not have worked; being able to answer hypothetical questions about what would happen otherwise; being able to improve performance in such ways as improving fluency, removing bugs in strategies, and expanding the variety of contexts. The system should also be able to further justify these kinds of remarks.

Sloman proposed this example of a sequence of increasingly sophisticated such miniscenarios in the proposed multi-robot problem domain:

1. Person wants to get box from high shelf. Ladder is in place. Person climbs ladder, picks up box, and climbs down.
2. As for 1, except that the person climbs ladder, finds he can't reach the box because it's too far to one side, so he climbs down, moves the ladder sideways, then as 1.
3. As for 1, except that the ladder is lying on the floor at the far end of the room. He drags it across the room lifts it against the wall, then as 1.
4. As for 1, except that if asked while climbing the ladder why he is climbing it the person answers: something like "To get the box." It should understand why "To get to the top of the ladder" or "To increase my height above the floor" would be inappropriate, albeit correct.
5. As for 2 and 3, except that when asked, "Why are you moving the ladder?" the person gives a sensible reply. This can depend in complex ways on the previous contexts, as when there is already a ladder closer to the box, but which looks unsafe or has just been painted. If asked, "would it be safe to climb if the foot of the ladder is right up against the wall?" the person can reply with an answer that shows an understanding of the physics and geometry of the situation.
6. The ladder is not long enough to reach the shelf if put against the wall at a safe angle for climbing. Another person suggests moving the bottom closer to the wall, and offers to hold the bottom of the ladder to make it safe. If asked why holding it will make it safe, gives a sensible answer about preventing rotation of ladder.
7. There is no ladder, but there are wooden rungs, and rails with holes from which a ladder can be constructed. The person makes a ladder and then acts as in previous scenarios. (This needs further unpacking, e.g. regarding sensible sequences of actions, things that can go wrong during the construction, and how to recover from them, etc.)
8. As for 7, but the rungs fit only loosely into the holes in the rails. Person assembles the ladder but refuses to climb up it, and if asked why can explain why it is unsafe.
9. Person watching another who is about to climb up the ladder with loose rungs should be able to explain that a calamity could result, that the other might be hurt, and that people don't like being hurt.

Such a system should be made to face a substantial library of such graded sequences of mini-scenarios that require it both to learn new skills, to improve its abilities to reflect on them, and (with practice) to become much more fluent and quick at achieving these tasks. These orderings should be based on such factors as the required complexity of objects, processes, and knowledge involved, the linguistic competence required, and the understanding of how others think and feel. That library could include all sorts of things children learn to do in such various contexts as dressing and undressing dolls, coloring in a picture book, taking a bath (or washing a dog), making toys out of Meccano and other construction kits, eating a meal, feeding a baby, cleaning a mess made by spilling some powder or liquid, reading a story and answering questions about it, making up stories, discussing behavior of a naughty person, and learning to think and talk about the past, the future, and about distant places, etc.



"I see you're trying to build a tower"

Recognizes the purpose behind the actions of the other agent.

Involves some spatial, physical, visual, and psychological reasoning.

"Yes but I can't reach that block"

Notices an impasse or problem that it cannot solve.

Involves knowledge about space, bodies and their abilities, problem solving, and reflection.

"I can reach it, let me get it for you"

Realizes that it can achieve that goal, and it does not conflict with its own goals.

Involves social reasoning and cooperative problem solving.

Figure 3. Reasoning in Multiple Mental Realms to Solve a Problem in the Model World.

Still, the participants had a heated debate about the adequacy of the proposed problem domain. The most common criticism was that this world does not contain enough of a variety of objects or richness of behavior. Doug Lenat suggested a solution to this, which was to embed the people within not a Blocks World, but instead somewhere like a typical house or office, as in the popular computer game *The Sims*. Doug Riecken argued that we could develop enough of the architecture within the more limited virtual world, and later add extensions to deal with a wider range of objects and phenomena.

A different response to this criticism was that in order to focus on architectural issues, it would help to simplify the problem domain, so that we could focus less on acquiring a large mass of world knowledge, and more on developing better ways for systems to use the knowledge they have. However, other participants argued that restricting the world would not entirely bypass the need for large databases of commonsense knowledge, for even this simple world would likely require hundreds of thousands or even millions of elementary pieces of commonsense knowledge about space, time, physics, bodies, social interactions, object appearances, and so forth.

Other participants disagreed with the virtual world domain. They felt that we should instead take the more practical approach of developing the architecture by starting with a useful application like a search engine or conversational agent, and extending its common sense abilities over time. But Ben Kuipers worried that choosing too specific an application would lead to what happened to most previous projects—someone discovers some set of ad hoc tricks that leads to adequate performance, without making any more general progress toward more versatile, resourceful, or "more intelligent" systems.

In the end, after long debates we achieved a substantial consensus that to solve harder problems requiring common sense, we first needed to solve the more restricted class of problems that show up in simpler domains like the proposed virtual world. Once we get the core of the architecture functioning in this rich but limited domain, we can attempt to extend it—or it extend itself—to deal with a broader

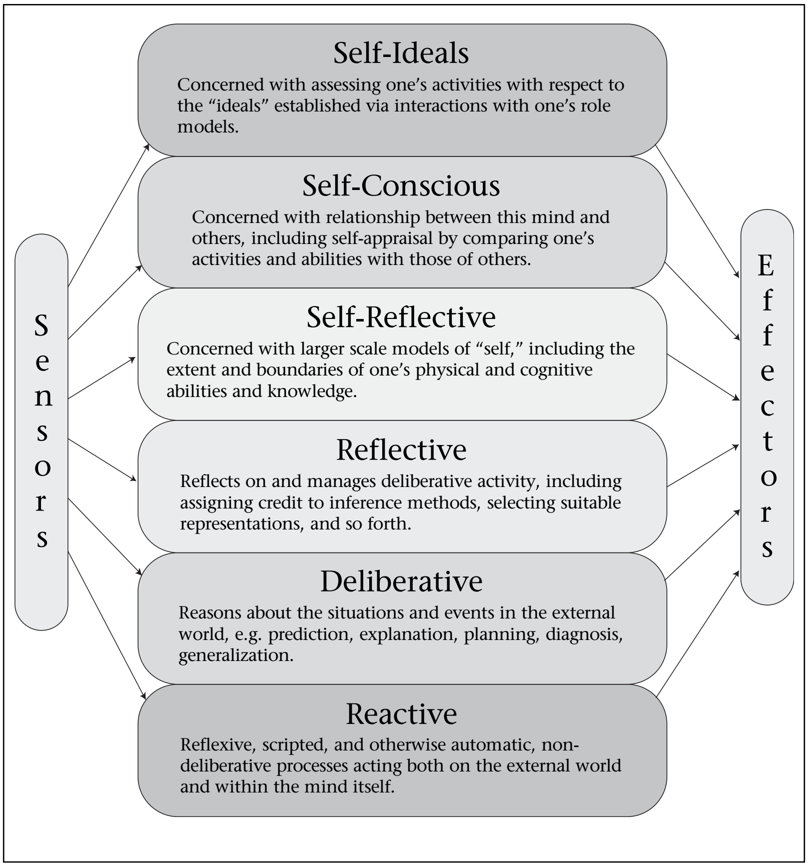

Self-Ideals
Concerned with assessing one's activities with respect to the "ideals" established via interactions with one's role models.
Self-Conscious
Concerned with relationship between this mind and others, including self-appraisal by comparing one's activities and abilities with those of others.
Self-Reflective
Concerned with larger scale models of "self," including the extent and boundaries of one's physical and cognitive abilities and knowledge.
Reflective
Reflects on and manages deliberative activity, including assigning credit to inference methods, selecting suitable representations, and so forth.
Deliberative
Reasons about the situations and events in the external world, e.g. prediction, explanation, planning, diagnosis, generalization.
Reactive
Reflexive, scripted, and otherwise automatic, non-deliberative processes acting both on the external world and within the mind itself.
Sensors
Effectors

Figure 4. Minsky's Emotion Machine Architecture.

range of problems using a much broader array of commonsense knowledge.

Large-Scale Architectures for Human-level Intelligence

In the afternoon, we discussed large-scale architectures for machines with human-level intelligence and common sense. Marvin Minsky and Aaron Sloman each presented their current architectural proposals as a starting point for the meeting participants to criticize, debug, and elaborate. These two architectures share so many features that we will refer to them together as the *Minsky-Sloman model*.

These architectures are distinguished by their emphasis on reflective thinking. Most cognitive models have focused only on ways to react or deliberate. However, to make machines more versatile, they will need better ways to recognize and repair the obstacles, bugs and deficiencies that result from their own activities. In particular, whenever one strategy fails, they'll need to have a collection of ways to switch to alternative ways to think. To provide for this, Minsky's architectural design includes several reflective levels beyond the reactive and deliberative levels. Here is one view of his model for the architecture of a person's mind, as described in his book, *The Emotion Machine,* and shown here in figure 4.

Some participants questioned the need for so many reflective layers; would not a single one be enough? Minsky responded by arguing that today, when our theories still explain too little, we should elaborate rather than simplify, and we should be building theories with more parts, not fewer. This general philosophy pervades his architectural design, with its many layers, representations, critics, reasoning methods, and other diverse types of components. Only once we have built an architecture rich enough to explain most of what people can do will it make sense to try to simplify things. But today, we are still far from an architectural design that explains even a tiny fraction of human cognition.

Aaron Sloman's *Cognition and Affect* project has explored a space of architectures proposed as models for human minds; a sketch of Sloman's H-CogAff model is shown in figure 5. This architecture appears to provide a framework for defining with greater precision than previously a host of mental concepts, including affective concepts, such as "emotion," "attitude," "mood," "pleasure," and so on. For instance, H-CogAff allows us to define at least three distinct varieties of emotions; primary, secondary and tertiary emotions, involving different layers of the architecture which evolved at different times—and the same architecture can also distinguish different forms of learning, perception, and control of behavior. (A different architecture might be better for exploring analogous states of insects, reptiles, or other mammals.) Human infants probably have a much-reduced version of the architecture that includes self-bootstrapping mechanisms that lead to the adult form.

The central idea behind the Minsky-Sloman architectures 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. There is no single underlying knowledge representation scheme or inferencing mechanism.

How do such architectures support such diversity? In the case of Minsky's Emotion Machine architecture, the top level is organized as follows. When the system encounters a problem, it first uses some knowledge about "problem-types" to select some "way-to-think" that

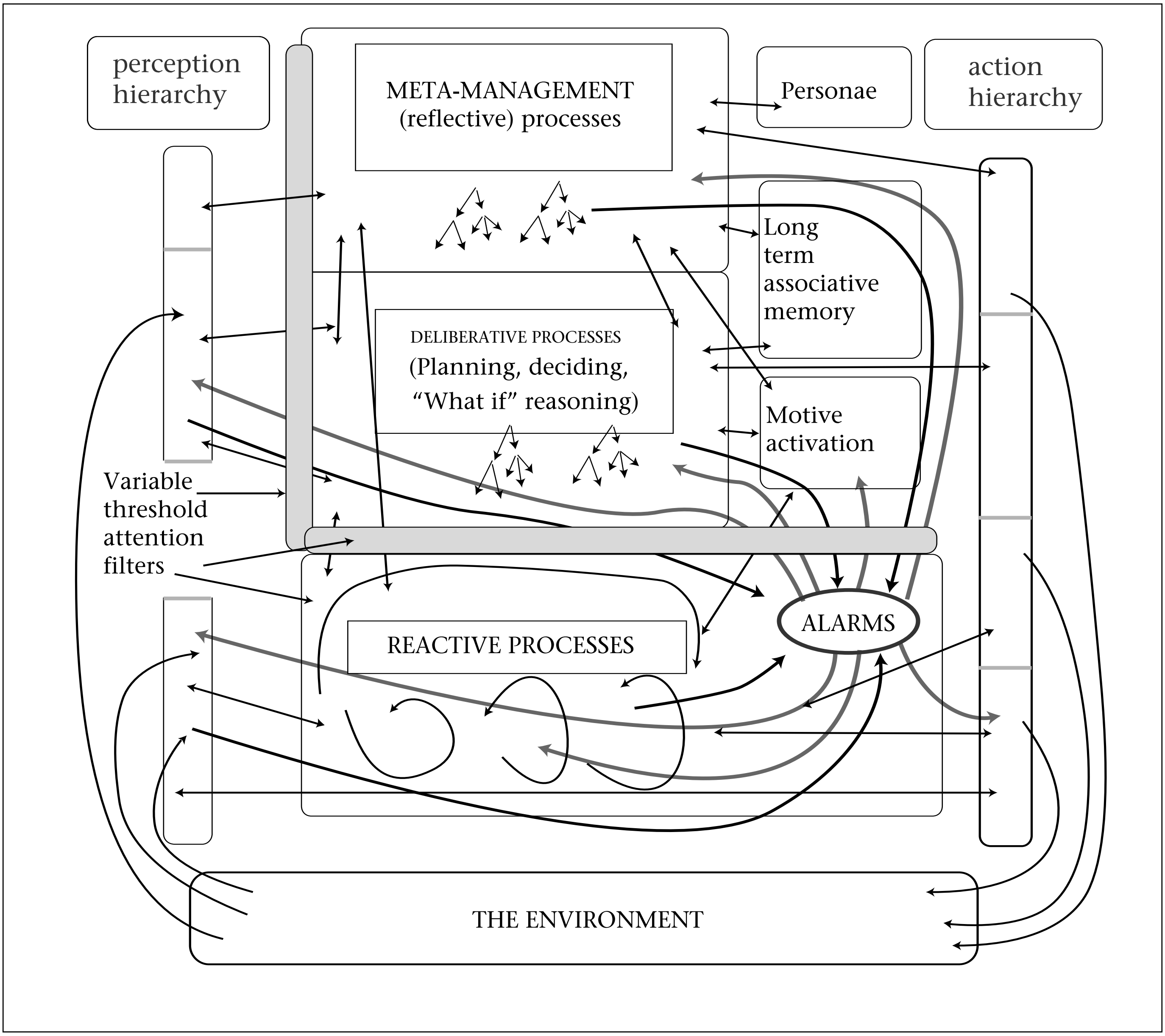

perception hierarchy
META-MANAGEMENT (reflective) processes
Personae
action hierarchy
Long term associative memory
DELIBERATIVE PROCESSES
(Planning, deciding, "What if" reasoning)
Motive activation
Variable threshold attention filters
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Figure 5. Aaron Sloman's H-CogAff Architecture.

might work. Minsky describes "ways-to-think" as configurations of agents within the mind that dispose it towards using certain styles of representation, collections of commonsense knowledge, strategies for reasoning, types of goals and preferences, memories of past experiences, manners of reflections, and all the other aspects that go into a particular "cognitive style." One source of knowledge relating problem-types to ways-to-think is the causal diversity matrix discussed at the start of the meeting—for example, if the system were presented with a social problem, it might use the causal diversity matrix to then select a case-based style of reasoning, and a particular database of social reasoning episodes to use with it.

However, any particular such approach is likely to fail in various ways. Then if certain "critic" agents notice specific ways in which that approach has failed, they either suggest strategies to adapt that approach, or suggest alternative ways-to-think, as suggested shown in figure 6. This is not done by employing any simple strategy for reflection and repair, but rather by using large arrays of higher level knowledge about where each way-to-think has advantages and disadvantages, and how to adapt them to new contexts.

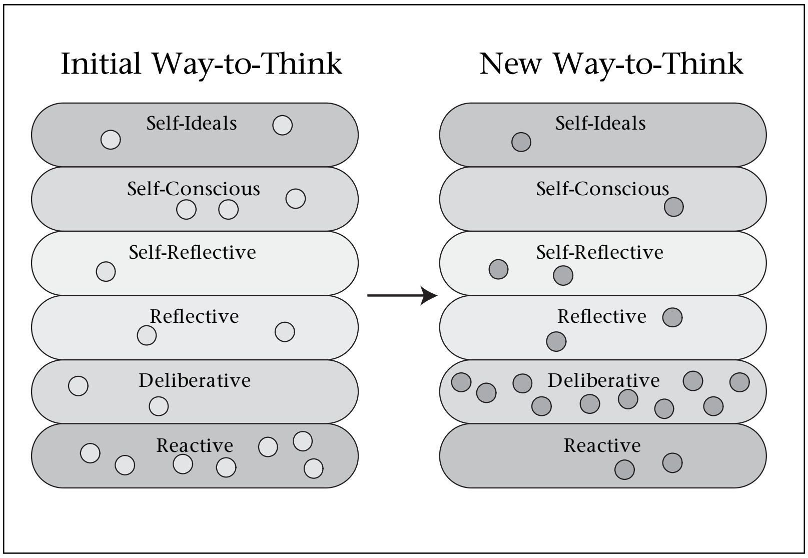


Figure 6. Switching from a Largely Reactive to a Largely Deliberative Way-to-Think.

Circles represent agents and other mental resources (fragments of knowledge, methods of reasoning, ways to learn, etc.) specific to that way-to-think, spanning the many levels of the architecture.

In Minsky's design, several ways-to-think are usually active in parallel. This enables the system to quickly and fluently switch between different ways-to-think because, instead of starting over at each transition, each newly activated way-to-think will find an already-prepared representation. The system will rarely "get stuck" because those alternative ways-to-think will be ready to take over when the present one runs into trouble, as shown in figure 7.

Here each way-to-think involves reasoning in a particular subset of mental realms. Impasses encountered while reasoning in one set of mental realms can be overcome within others. Further information about these architectures can be found in Singh and Minsky (2003), Sloman (2001), and McCarthy et al. (2002). Minsky's model will be described in detail in his new book *The Emotion Machine* (Minsky, forthcoming).

Generally, the participants were sympathetic to these proposals, and all agreed with the idea that to achieve human-level intelligence we needed to develop more effective ways to combine multiple AI techniques. Ken Forbus suggested that we needed a kind of "component marketplace," and that we should find ways to instrument these components so that the reflective layers of the architecture had useful information available to them. He contrasted the Soar project (Laird, Newell, and Rosenbloom 1987) as an effort to eliminate and unify components rather than to accumulate and diversify them, as in the Minsky-Sloman proposals. Ashwin Ram and Larry Birnbaum both pointed out that despite the agreement over the architectural proposals it was still not clear what the particular components of the architecture would be. They pointed out that we needed to think more about what the units of reasoning would be. In other words, we needed to come up with a good list of way-to-think. Some examples might include the following:

Solving problems by making analogies to past experiences

Predicting what will happen next by rule-based mental simulations

Constructing new "ways to think" by building new collections of agents

Explaining unexpected events by diagnosing causal graphs

Learning from problem-solving episodes by debugging semantic networks

Inferring the state of other minds by re-using self-models

Classifying types of situations using statistical inference

Getting unstuck by reformulating the problem situation

This list could be extended to include all available AI techniques.

Educating the Architecture

On the morning of the second day of the meeting, we addressed the problem of how to supply the architecture with a broad range of commonsense knowledge, so that it would not have to "start from scratch." We all agreed that learning was of value, but we didn't all agree on where to start. Many researchers would like to start with nothing; however, Aaron Sloman pointed out that an architecture that comes with no knowledge is like a programming language that comes with no programs or libraries.

One view that was expressed was that approaches that start out with too little initial knowledge would likely not achieve enough versatility in any practical length of time. Minsky criticized the increasing popularity of the concept of a "baby machine"—learning systems designed to achieve great competence, given very little initial structure. Some of these ideas include genetic programming, robots that learn by associating sensory-motor patterns, and online chatbots that try to learn language by generalizing from thousands of conversations. Minsky's complaint was that the problem is not that the concept of a baby machine is itself unsound, but rather that we don't know how to do it yet. Such approaches have all failed to make much progress because

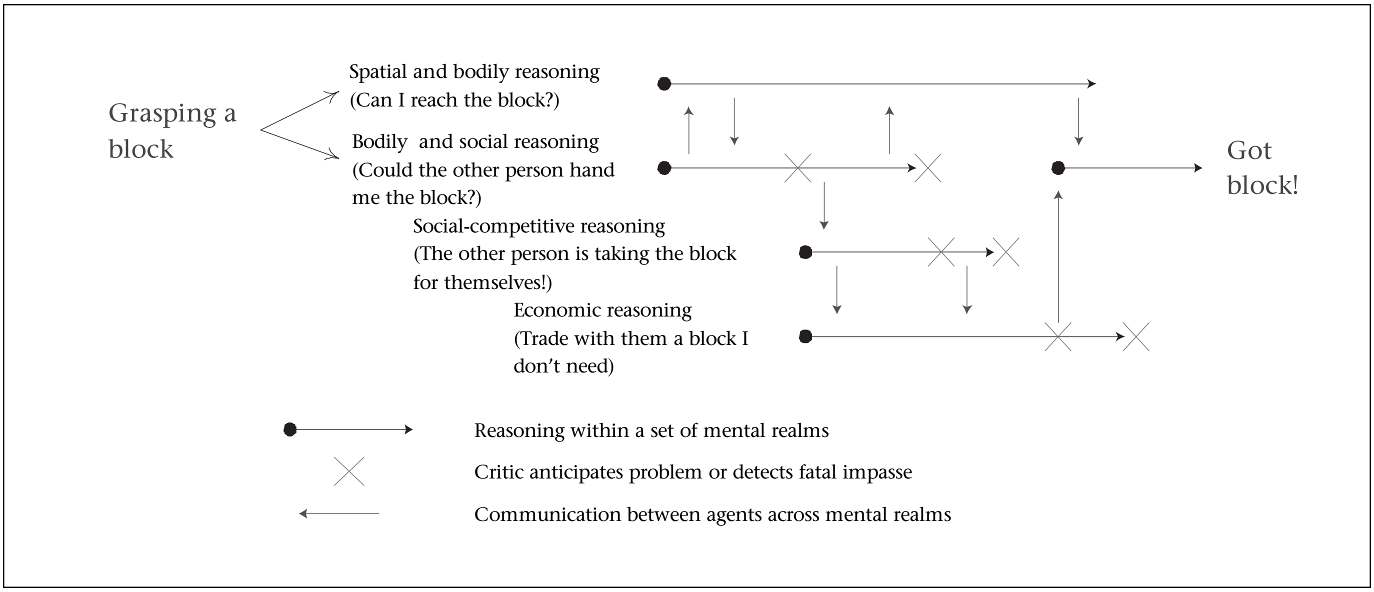

Grasping a block
Spatial and bodily reasoning
(Can I reach the block?)
Bodily and social reasoning
(Could the other person hand me the block?)
Social-competitive reasoning
(The other person is taking the block for themselves!)
Economic reasoning
(Trade with them a block I don't need)
Got block!
Reasoning within a set of mental realms
Critic anticipates problem or detects fatal impasse
Communication between agents across mental realms

Figure 7. Using Several Ways-to-Think in Parallel in a Social Blocks World.

they started out with inadequate schemes for learning new things. You cannot teach algebra to a cat; among other things, human infants are already equipped with architectural features to equip them to think about the causes of their successes and failures and then to make appropriate changes. Today we do not yet have enough ideas about how to represent, organize, and use much of commonsense knowledge, let alone build a machine that could learn all of that automatically on its own. As John McCarthy noted long ago: "in order for a program to be capable of learning something, it must first be able to represent that knowledge."

There are very few general-purpose commonsense knowledge resources in the AI community. Doug Lenat gave a wonderful presentation of the Cyc system, which is presently the project furthest along at developing a useful and reusable such resource for the AI community, so that new AI programs don't have to start with almost nothing. The Cyc project (Lenat 1995) has developed a great many ways to represent commonsense knowledge, and has built a database of over a million commonsense facts and rules. However, Lenat estimated that an adult-level commonsense system might require 100 million units of commonsense knowledge, and so one of their current directions is to move to a distributed knowledge acquisition approach, where it is hoped that eventually thousands of volunteer teachers around the world will work together teach Cyc new commonsense knowledge. Lenat spent some time describing the development of friendly interfaces to Cyc that allow nonlogicians to participate in the complicated teaching and debugging processes involved in building up the Cyc knowledge base.

Many of the participants agreed that Cyc would be useful, and some suggested we could even base our effort on top of it, but others were sharply critical. Jeffrey Siskind doubted that Cyc contained the spatial and perceptual knowledge needed to do important kinds of visual scene interpretation. Roger Schank argued that Cyc's axiomatic approach was unsuitable for making the kinds of generalizations and analogies that a more case-based and narrative-oriented approach would support. Srini Narayanan worried that the Cyc project was not adequately based on what cognitive scientists have learned about how people make commonsense inferences. Oliver Steele concluded that while we disagreed about whether Cyc was 90% of the solution or only 10%, this was really an empirical question that we would answer during the course of the project. But generally, the architectural proposal was regarded as complementary to parallel efforts to accumulate substantial commonsense knowledge bases.

Minsky predicted that if we used Cyc, we might need to augment each existing item of knowledge with additional kinds of procedural and heuristic knowledge, such as descriptions of (1) problems that this knowledge item could help solve; (2) ways of thinking that it could participate in; (3) known arguments for and against using it; and (4) ways to adapt it to new contexts.

It was stressed that knowledge about the world was not enough by itself—we also need a knowledge base about how to reason, reflect and learn, the knowledge that the reflective layers of the architecture must possess. The problem remains that the programs we have for using knowledge are not flexible enough, and neither Cyc's "adult machine" approach of supplying a great deal of world knowledge, nor the "baby machine" approach of learning common sense from raw sensory-motor experience, will likely succeed without first developing an architecture that supports multiple ways to reason, learn, and reflect upon and improve its activities.

An Important Application

Several of the participants felt that such a project would not receive substantial support unless it proposed an application that clearly would benefit much of the world. Not just an improvement to something existing, it would need to be one that could not be built without being capable of human-level commonsense reasoning.

After a good deal of argument, several participants converged upon a vision from *The Diamond Age,* a novel by Neil Stephenson. That novel envisioned an "intelligent book"—*The Young Ladies Illustrated Primer*—that, when given to a young girl, would immediately bond with her and come to understand her so well as to become a powerful personal tutor and mentor.

This suggested that we could try to build a *personalized teaching machine* that would adapt itself to someone's particular circumstances, difficulties, and needs. The system would carry out a conversation with you, to help you understand a problem or achieve some goal. You could discuss with it such subjects as how to choose a house or car, how to learn to play a game or get better at some subject, how to decide whether to go to the doctor, and so forth. It would help you by telling you what to read, stepping you through solutions, and teaching you about the subject in other ways it found to be effective for you. Textbooks then could be replaced by systems that know how to explain ideas to you in particular, because they would know your background, your skills, and how you best learn.

This kind of application could form the basis for a completely new way to interact with computers, one that bypasses the complexities and limitations of current operating systems. It would use common sense in many different ways: (1) It would understand human goals so that it could avoid the silliest mistakes. (2) It would understand human reasoning so that it could present you with the right level of detail and avoid saying things that you probably inferred. (3) It would converse in natural language so that you could easily talk to it about complex matters without having to learn a special language or complex interface.

To build such a kind of "helping machine," we would first need to give it knowledge about space, time, beliefs, plans, stories, mistakes, successes, relationships, and so forth, as well as good conversational skills. However, little of this could be realized by anything less than a system with common sense. To accomplish this we would need to pursue some sequence of more modest goals that would help one with simpler problem types—until the system achieved the sorts of competence that we expect from a typical human four- or five-year-old.

However, to get such a system to work, we would need to address many presently unsolved commonsense problems that show up in the model-world problem domain.

Final Consensus

The participants agreed that no single technique (such as statistics, logic, or neural networks) could cope with a sufficiently wide range of problem-types. To achieve human-level intelligence we must create an architecture that can support many different ways to represent, acquire, and apply many kinds of commonsense knowledge.

Most participants agreed that we should combine our efforts to develop a model world that supports simplified versions of everyday physical, social, and psychological problems. This simplified world would then be used to develop and debug the core components of the architecture. Later, we can expand it to solve more difficult and more practical problems.

The participants did not all agree on which particular larger-scale application would both attract sufficient support and also produce substantial progress toward making machines that use commonsense knowledge. Still, many agreed with the concept of a personalized teaching machine that would come to understand you so well that it could adapt to your particular circumstances, difficulties, and needs.

Ben Kuipers sketched the diagram shown in figure 8, which captures the general dependencies between the three points of consensus: Practical applications depend on developing an architecture for commonsense thinking flexible enough to integrate a wide array of processes and representations of problems that come up in the model-world problem domain.

A Collaborative Project?

At the end of the meeting, we brainstormed about how we might organize a distributed, collaborative project to build an architecture based on the ideas discussed at this meeting. It is a difficult challenge, both technically and socially, to get a community of researchers to work on a common project. However, successes in the Open Source community show that such distributed projects are feasible when the components can be reasonably disassociated.

Furthermore, this kind of architecture itself should help to make it easy for members of the project to add new types of representations and processes. However, we first would have to develop a set of protocols to support the interoperation of such a diverse array of methods. Erik Mueller suggested that such an organization could be modeled after the World Wide Web Consortium (W3C), and its job would largely be to assess, standardize and publish the protocols and underlying tools that such a distributed effort would demand.

While we did not sketch a detailed plan for how to proceed, Aaron Sloman, Erik Mueller and Push Singh listed some technical steps that such a project would need:

First, it should not be too hard to develop a suitable virtual model world, because the present-day video game and computer graphics industry has produced most of the required components. These should already include adequate libraries for computer graphics, physics simulation, collision detection, and so forth.

Second, we need to develop and order the set of miniscenarios that we will use to organize and evaluate our progress. This would be a continuous process, as new types of problems will constantly be identified.

Third, what kinds of protocols could the agents of this cognitive system use to coordinate with each other? This would include messages for updating representations, describing goals, identifying impasses, requesting knowledge, and so forth. We would consider the radical proposal to use, for this, an Interlingua based on a simplified form of English, rather than trying to develop some brand new ontology for expressing commonsense ideas. Of course, each individual agent could be free to use internally whatever ontology or representation scheme was most convenient and useful.

Fourth, we would need to create a comprehensive catalog of ways-to-think, to incorporate into the architecture. A commonsense system should be at least capable of reasoning about prediction, explanation, generalization, exemplification, planning, diagnosis, reflection, debugging, learning, and abstracting.

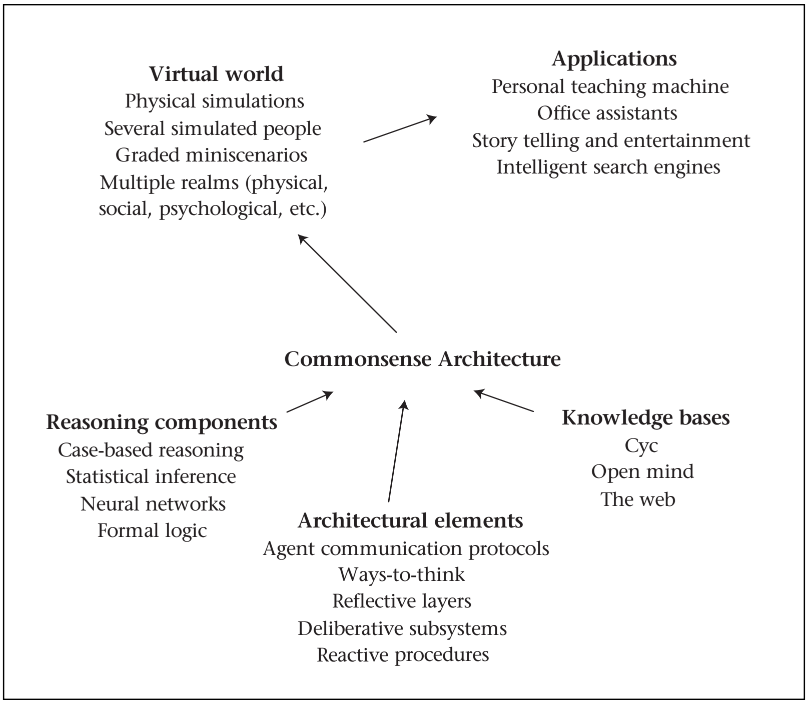


Figure 8. Dependencies Between Points of Consensus.

Fifth, what are the kinds of self-reflections that a commonsense system should be able to make of itself, and how should these invoke and modify ways-to-think as problems are encountered?

Sixth, in any case, such a system will need a substantial, general-purpose, and reusable commonsense knowledge base about the spatial, physical, bodily, social, psychological, reflective, and other important realms, enough to deal with a broad range of problems within the model world problem domain.

Finally, we might need to develop a new kind of "intention-based" programming language to support the construction of such an architecture.

Towards the Future

Since our meeting similar sentiments have been expressed at DARPA, most notably in the recent "Cognitive Systems" Information Processing Technology Office (IPTO) Broad Agency Announcement (BAA) (Brachman and Lemnios 2002), which solicits proposals for building AI systems that combine many elements of knowledge, reasoning, and learning. While we are gratified that architectural approaches are becoming more popular, we would like to see more emphasis placed on architectural designs that specifically support more common sense styles of thinking.

There was a genuine sense of excitement at this meeting. The participants felt that it was a rare opportunity to focus once more on the grand goal of building a human-level intelligence. Over the next few years, we plan to develop a concrete implementation of an architecture based on the ideas discussed at this meeting, and we invite the rest of the AI community to join us in such efforts.

Acknowledgements

We would like to thank Cecile Dejongh for taking care of the local arrangements, and extend a very special thanks to Linda Stone for making this meeting happen. This meeting was made possible by the generous support of Jeffrey Epstein.

References

Brachman, Ronald; and Lemnios, Zachary 2002. DARPA's New Cognitive Systems Vision. *Computing Research News,* 14(5):1, 8.

Kitano, Hiroaki; Asada, Minoru; Kuniyoshi, Yasuo; Noda, Itsuki; Osawa, Eiichi; and Matsubara, Hitoshi. 1997. RoboCup: A Challenge problem for AI. *AI Magazine,* 18(1):73–85.

Laird, John; Newell, Allen; and Rosenbloom, Paul 1987. SOAR: An Architecture for General Intelligence. *AI Journal,* 33(1):1-64.

Lenat, Doug. 1995. CYC: A Large-scale Investment in Knowledge Infrastructure. *Communications of the ACM,* 38(11):33-38.

McCarthy, John; Minsky, Marvin; Sloman, Aaron; Gong, Leiguang; Lau, Tessa; Morgenstern, Leora; Mueller, Erik; Riecken, Doug; Singh, Moninder; and Singh, Push 2002. An Architecture of Diversity for Commonsense Reasoning. *IBM Systems Journal,* 41(3):530–539.

Minsky, Marvin. (forthcoming). The Emotion Machine. Pantheon, New York. Several chapters are on-line at http://web.media.mit.edu/people/minsky

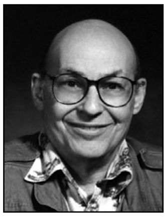
Minsky, Marvin 1992. Future of AI Technology. *Toshiba Review,* 47(7).

Singh, Push ; and Minsky, Marvin. 2003. An Architecture for Combining Ways to Think. Paper presented at the International Conference on Knowledge Intensive Multi-Agent Systems. Cambridge, Mass., September 30 – October 3.

Sloman, Aaron 2001. Beyond Shallow Models of Emotion. *Cognitive Processing,* 1(1):530-539.

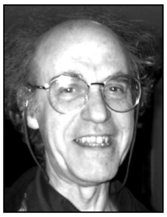
Note

[1] This meeting was held in St. Thomas, U.S. Virgin Islands, on April 14-16, 2002. The meeting included the following participants: Larry Birnbaum (Northwestern University), Ken Forbus (Northwestern University), Ben Kuipers (University of Texas at Austin), Douglas Lenat (Cycorp), Henry Lieberman (Massachusetts Institute of Technology), Henry Minsky (Laszlo Systems), Marvin Minsky (Massachusetts Institute of Technology), Erik Mueller (IBM T. J. Watson Research Center), Srini Narayanan (University of California, Berkeley), Ashwin Ram (Georgia Institute of Technology), Doug Riecken (IBM T. J. Watson Research Center), Roger Schank (Carnegie Mellon University), Mary Shepard (Cycorp), Push Singh (Massachusetts Institute of Technology), Jeffrey Mark Siskind (Purdue University), Aaron Sloman (University of Birmingham), Oliver Steele (Laszlo Systems), Linda Stone (independent consultant), Vernor Vinge (San Diego State University), and Michael Witbrock (Cycorp).



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 his B.A. and Ph.D. 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.

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 humanlike 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.



Aaron Sloman is a professor of AI and cognitive science at the University of Birmingham, UK.He received his B.Sc. in mathematics and physics (Cape Town, 1956), and a D.Phil. Philosophy, from Oxford (1962). Sloman is a Rhodes Scholar, a Fellow of AAAI, AISB, and ECCAI. He is also author of *The Computer Revolution in Philosophy* (1978) and many theoretical papers on vision, diagrammatic reasoning, forms of representation, architectures, emotions, consciousness, philosophy of AI, and tools for exploring architectures. Sloman maintains the FreePoplog open source web site and is about to embark on a large EC-funded robotics project. All papers, presentations, and software are accessible from his home page: www.cs.bham.ac.uk/~axs/