# Technical forum


## Machine intelligence and the Turing Test

Alan Mathison Turing, the British mathematician, philosopher, and logician, proposed in 1950 that if a computer could successfully mimic a human during an informal exchange of text messages, then, for most practical purposes, the computer might be considered intelligent.[1] This soon became known as the Turing Test (TT), most typically conducted as anonymous exchanges of English-language text between computers. A panel of judges poses questions to the contestants in order to determine which are human and which are programs. There have been many such experiments since Turing proposed the challenge, but there is considerable disagreement as to what passing the test means, and whether passing it tells us much at all.[2]

We use the TT here as a means of identifying artificial intelligence (AI) technologies that will have a pivotal role in creating more intuitive machine-human interactions. We have chosen six technologies supporting certain computer behaviors that could significantly increase the practical value of computers. In what follows we summarize and editorialize on where each of these technologies stands today, relying heavily on the findings of the conference and workshop "Machine Intelligence and the Turing Test"[3] held last year at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York.

The first technology we address is *natural language understanding* (NLU). We confess to a particular passion for this part of AI, so fundamental for communication and yet still full of nuances, poorly understood, and hard to symbolize.[4] Even if computers could understand plain English, this would just be the beginning.

Our second technology is *machine reasoning* (MR). The TT judges ask the contestants questions intended to flush out the mere mechanical responses of a computer. To fool the human judges, a computer will need to provide reasonable answers, answers that are relevant within the context set by earlier exchanges.




0018-8670/02/$5.00 © 2002 IBM

Since TT questions can be about any subject, contestants need a very significant knowledge base covering a wide range of human activities such as sports, politics, health, and food. In addition to the rational, objective knowledge about ourselves and our society, the computer's knowledge base needs to also include "rules of thumb," myths, "old wives tales" and urban legends, as well as the complex relationships between facts, theories, conjectures, and judgments. We need *knowledge representation* (KR) technology to represent this information in all its complexity.

Creating the on-line body of knowledge would be itself a daunting task, and *knowledge acquisition* (KA) is our fourth AI technology with direct application for any would-be TT winner. The manual effort required to capture this knowledge would be enormous, so computers need to be programmed to listen and learn automatically.

Two less prominent technologies complete our set. There is growing interest in the dynamics of dialog and the role of identity in computer-human exchanges. This has led to theories and experiments in *dialog management*,[5] and to experiments in making computers react to human emotions.[6]

## Natural language understanding

Natural language understanding (NLU) is the technology enabling computers to extract meaning from text—easy and natural for humans, but notoriously difficult as computation. NLU is a key component of software that can accept commands and queries from humans in their own language and produce answers whose meanings are automatically extracted from electronic texts. Even limited comprehension has significant business benefit. For example, today's imperfect approximations to NLU are widely used in: abstracting of trends and important events from news sources; summarizing vast repositories of text; and supporting natural language queries for accessing online help.

Expectations for NLU are well beyond the limited capabilities of the first-generation technologies that are built into current search engines, classification engines, and information extraction systems.[7] Current techniques do not handle semantics very well because of the many-to-many mappings between syntactic forms and semantic meanings. One form can have many meanings (e.g., "time flies like an arrow"); whereas the same meaning can be paraphrased in different forms ("announce/unveil/come out with a new product"). This is usually no problem for humans because we interpret meaning relative to context—prior knowledge shared by author and reader (or speaker and hearer).

The grand challenge of NLU is to simulate the human capability to:

- Create a store of prior knowledge (see the section on knowledge representation, later)
- Create a representation of the meaning of the current text
- Integrate this representation into the knowledge store (see the section on knowledge acquisition, later)

A representation of the meaning of text is created "bottom up" (small semantic units are fitted together into larger representations) and "top down," by homing in on the specific topic the document discusses. Significant progress has been made with the semantics of the smallest units, such as people, places, organizations, and temporal expressions (successfully embedded in IBM text analysis products[8]) and, increasingly, relations in the sentence that hold among these are also recognized (e.g., announce [IBM, hard disk drive]). Both symbolic and statistically based machine learning techniques have been successful and continue to develop. Advances continue on disambiguation of common words[9] and in translation between pairs of languages.

Progress has been slower beyond the word and sentence level. One well-known requirement, for example, is pronoun resolution—determining what a word such as "it" or "they" refers to. Algorithms were developed over ten years ago and for a while they improved, but they seem to have reached a plateau. The next major advance in NLU will probably come from the use of inferences based on knowledge of the world. For example, in the text "IBM has unveiled the Ultrastar* 36Z15 . . . It is the fastest in the industry," we need to know that IBM is a computer manufacturer, that computer manufacturers produce computers and parts, and that machines have properties such as speed to determine that "it" refers to the disk drive rather than to IBM.

The complete semantic task includes understanding the relations among entities discussed in the text, the actions and events they are engaged in, and temporal and causal sequences. There has been some success in understanding the main events (*who* did *what*

to *whom*) in specific and narrow domains,[10] but expanding to larger domains remains a challenge.

Statistical "top down" techniques for topic identification have been developed mainly within the information retrieval community, with notable successes by IBM.[11] Current classifiers group large collections of documents into more specific topics, automatically producing a taxonomy like that of the Yahoo!** service.[12] These techniques continue to be crucial for identifying the "aboutness" of a document.

The progress of NLU to date has been encouraging in the areas of syntactic parsing, language-pair translation, semantic analysis in narrow domains, and statistically based information retrieval. Now is the time to concentrate on a deeper semantic understanding of text in larger domains. The domain-independent and complete NLU required for TT-like tasks will remain elusive for many years, but incremental progress can be made, and measured, within broadly defined domains and with respect to specific tasks.

## Machine reasoning

The human capability for reasoning is another important aspect of our intelligence that machines have not fully captured. What we call "reasoning ability" is in fact a bundle of different abilities, such as:

- Simple inference within the KR system. For example, if we are told that Clyde is an elephant, we can deduce that he probably has four legs and that he needs food, water, and oxygen to survive.
- Search within the KR system. For example, find an African animal that is large, gray, and four-legged, with big-floppy ears and a long prehensile nose.
- More difficult inference. Some kinds of inference, such as theorem-proving, are much more difficult for humans than the kind of KR-based inference described above. There seems to be a qualitative difference in the effort required, and it may be that different mechanisms come into play.
- Planning and problem solving. In domains such as chess and factory scheduling, computers already exhibit problem-solving performance that is better than human performance. However, humans still excel at tasks requiring broad, diverse knowledge, flexibility, and the ability to learn, generalize, and transfer skills from one domain to another.
- Plan recognition and the ability to reason explicitly about plans. Humans can create plans; they can also recognize and explain what another person is trying to do. They can examine their own plans and explain costs, risks, and alternatives to clients or coworkers.
- Creativity. Some say that creativity is just competent problem solving that happens to lead to a surprising result, and in that sense, machines can sometimes be creative, but we believe fundamental elements of creative problem solving are still missing.
- Applying recipes. It appears that we humans store our knowledge of procedures in the form of recipes or scenarios, rather than as rigid programs. This same knowledge can be used to produce new plans and to recognize the structure and components of other plans that we encounter. This is another kind of knowledge that we must represent and store effectively.

The expert-system tools of the 1980s and early 1990s developed much of the basic machinery for machine reasoning. The Soar[13] system, developed at Carnegie Mellon University by the late Allen Newell and his students, explored ways of combining rule-based problem solving with powerful learning and chunking mechanisms, so that the system's performance would improve over time. However, much still remains to be done to make this problem-solving more flexible and use knowledge of all kinds to guide the problem-solving process.

## Knowledge representation

For problems that require breadth of understanding—what we sometimes call "common sense"—current computing systems fall far short of human ability. The most critical missing piece is the ability to deal with large amounts of knowledge of many kinds, and to make that knowledge effective in perception and problem solving.[14]

Many kinds of knowledge are required for human-like capability. Predicate calculus can in principle be used to represent all the types listed below, but efficiency concerns push the system to more specialized representations for some of these types:

- Declarative statements
- Linguistic knowledge
- Procedural knowledge
- Naive physics
- Recognition knowledge
- Tactile/kinesthetic knowledge
- Visual knowledge
- Social knowledge

It seems unlikely that any single approach to knowledge representation (KR) will adequately cover all of these areas. Merely *representing* and *storing* each kind of knowledge is not sufficient; we must also make the knowledge *effective*. Each kind of knowledge requires appropriate representation, machinery for efficient search and inference, and some way to acquire and digest knowledge. Then all the knowledge types must fit into an architecture that allows them to work together effectively.

Declarative knowledge figures most prominently in the TT, which is focused on natural language input and output. There are several approaches both inside and outside IBM (most notably the Cyc** knowledge base[15]) for building a declarative KR system with powerful search and inference capabilities. These systems include a large base layer of knowledge that spans most domains—physical objects, materials, people, organizations, common actions and behaviors, and so on. They are then enriched with more specialized knowledge for each domain of interest. Although specialists may be required to build the most fundamental "roots of the universe" knowledge, it needs to be easy for nonspecialists to extend the knowledge base.

## Knowledge acquisition

A system needs hundred of thousands to several million knowledge elements to approximate the knowledge of a human being. The challenge is to automate the KR process using a variety of techniques, including "learning" (as in statistical modeling and machine learning), in both supervised form (where the answer is provided to the learning algorithm) and unsupervised form (where the system observes the data without knowing the answer and has to infer it). To illustrate, a team at the University of Pennsylvania manually parsed a million words and provided parse trees for about 40000 sentences. Their *Treebank*[16] has been used by many researchers worldwide to create and improve parsers for broad domain English for various applications. Similarly, the creation of annotated databanks for other purposes will be a key ingredient in improving the state-of-the-art of the component technologies. At the other extreme is the knowledge in the Cyc knowledge base—a million facts manually entered over 15 years with 450 person-years of effort. One Cycorp researcher, when pressed, estimated that Cyc contains perhaps 2 percent of the required knowledge. But Cyc is at an inflection point and can start exploring methods to automate the KA and extract knowledge from the tens of terabytes of on-line text available on intranets and the Internet.

Another crucial aspect of KA is that new knowledge is needed on a daily basis, so the process of acquiring it has to be intrinsic to any system to keep up with the demands of deployed applications. The efficient creation of a whole cycle of knowledge update, from statistical learning to manual acquisition, is key to managing the total cost of operating these "intelligent" systems and a fundamental activity in creating the technologies.

## Dialog management

A dialog is a sequence of interactions between participants with a shared context and a shared set of goals. Dialog management refers to the analysis of user utterances in the context of the current discourse, figuring out an appropriate response, and conveying it to the user.

Existing dialog systems differ in the degree to which users can take the initiative and steer the conversation. Directed dialog systems force users to constrain their input and stay on predefined dialog paths. Mixed initiative systems constrain the user input only when it is imperative to have a clear understanding of user intentions (e.g., to get confirmation before executing a stock purchase). Over the last 20 years much progress has been made in replacing rigid hierarchical directed dialog systems with mixed initiative systems offering a more open mode of conversation.[17] Of course, we are still a long way away from completely open user initiative dialogs as characterized by the TT.

Narrow domain dialog systems are finding widespread use inside and outside IBM for a whole gamut of applications ranging from buying stocks to finding information to directory assistance. The big challenge confronting dialog researchers is to build systems that can converse with humans about topics not limited by a few predefined forms or templates. Another issue is one of meta-knowledge: Does the system know how much it knows or does not know? Can its behavior degrade gracefully when it encounters the limits of its knowledge? Promising new approaches combining statistical information retrieval, information extraction, and dialog systems may help answer the above questions.

## Emotion

An intelligent system may benefit from having access to information about the intentions or the emotional states of humans. Systems can already begin to recognize and use affective information in a variety of forms. Ultimately, systems might do this by observing facial expressions and body language, recognizing patterns in physiological measures, analyzing the affective content of text and speech, and inferring emotional states from interactive behavior.

Understanding what people are feeling can guide the computer interaction with its users, from help messages to the use of true-to-life computer-generated speech. Human conversational partners who do not give affective cues—tone of voice, choice of words, gestures—are perceived as flat and unresponsive partners. Without affective cues, misunderstandings abound. However, today it is difficult to find reliable indicators of emotion that are not obscured by individual differences in the way humans experience and express emotion.

From the user's perspective, giving our systems the ability to understand and appropriately respond to affective content may raise computers from their current socially inept role to a role more consistent with human conversational expectations.

## Epilogue

Most of the AI technology used in products today is based on linguistic models of knowledge and linguistic processing techniques. An important next step will be to combine linguistics with a large database of assertions representing commonsense facts about the world, in the hopes of producing much more human-like reading and conversational systems. Systems like Cyc and NETL[18] aim to give machines common sense by amassing a large collection of commonsense assertions, then reasoning about them with the help of a logic engine.

Although there is no doubt as to the business value of this direction, it is interesting to note that memorized linguistic assertions play only a small part in intelligent human behavior, and formal first-order logic has no significant role in human cognitive processing. This opens the way for approaches that seek to incorporate aspects of humans-as-systems beyond the purely linguistic, both to enhance the linguistic skills of our systems and to develop skills that may not be accessible with purely linguistic approaches.

One such approach creates learning machines that discover the facts by themselves much as humans do. Common sense is acquired by the machine sensing its environment directly and learning from that experience. In order to learn about catching a baseball, the machine might interact with the physical world through sensors and effectors designed for vision and motion. Common sense involved in a particular linguistic domain might be acquired by reading texts and conversing with humans.

Other emerging areas include the study of machines that are based on models of brain behavior, and although these approaches are relatively immature and more speculative in nature than the technologies based on linguistic models, they may allow us to make new, significant, and perhaps revolutionary progress in AI. On the other hand, although we doubt that AI technology based on purely linguistic models will ever be mistaken for a human, or ever pass an unconstrained TT, we believe that it will make the largest AI contribution to business and society for years to come.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of Yahoo! Inc. or Cycorp, Inc.

## Cited references and notes

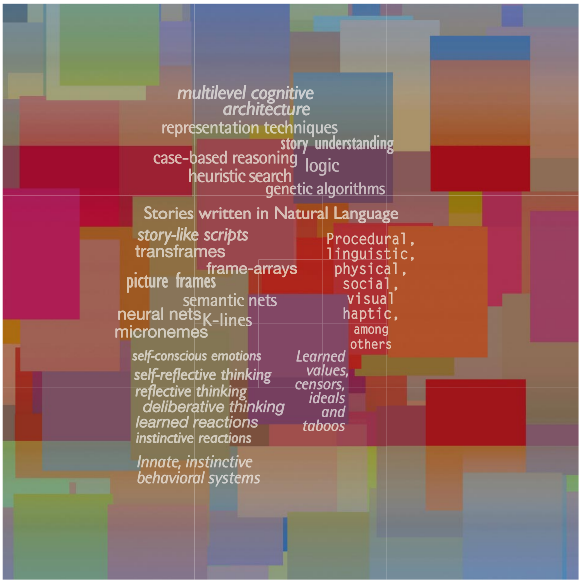

1. A. M. Turing, "Computing Machinery and Intelligence," *Mind* **59**, 433–460 (1950).
2. A formal TT yearly contest, sponsored by Hugh Loebner and The Cambridge Center for Behavioral Studies, accords a $2000 prize and medal to the most human-like computer contestant. Among the most well-known critics of the contest is Marvin Minsky, a professor of computer science at the Massachusetts Institute of Technology (MIT) who is considered by many to be "the father of AI." Minsky has wittily sponsored a "Minsky Loebner Prize Revocation Prize."
3. In May 2001 the IBM Academy of Technology and IBM Research held a conference and workshop on "Machine Intelligence and the Turing Test." The conference speakers were: Jaime Carbonell (Carnegie Mellon University), Barbara Grosz (Harvard University), Jerry Hobbs (SR International), John Laird (University of Michigan), Doug Lenat (Cycorp, Inc.), Michael Mauldin (Virtual Personalities & Carnegie Mellon University) and Rosalind Picard (MIT). The IBM organizing team comprised: Joe Bigus, Ian Brackenbury (chair), Scott Fahlman, Joe Londa, Clifford Pickover, Yael Ravin, and Alan Webb. The authors gratefully acknowledge the contributions to the workshop report by: Nancy Alverado, Scott Fahlman, Charles Peck, and Steve R. White, fragments of which are given here in condensed form.
4. Y. Bar-Hillel, "Automatic Translation of Languages," *Advances in Computers*, D. Booth and R. E. Meagher, Editors, Academic Press, New York (1960). This classic article on the NLU challenge is still often cited today.

5. See, for example, the *Proceedings of the ANLP/NAACL 2000 Workshop on Conversational Systems*, Seattle, WA, May 2000.
6. R. W. Picard, "Toward Computers That Recognize and Respond to User Emotion," *IBM Systems Journal* **39**, Nos. 3&4, 705–719 (2000).
7. "Enterprise Portals: Web Interfaces for Employees, Partners, and Customer Communities," META Group (September 1999).
8. Intelligent Miner for Text, IBM Corporation, http://www.ibm.com/software/data/iminer/fortext.
9. See http://www.itri.brighton.ac.uk/events/senseval.
10. See http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.
11. This technique has been successfully used by customers and partners to route customer e-mail to the right expert, based on the e-mail content. See http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/email198.html.
12. Yahoo! is a large Web site featuring vast, manually maintained, taxonomies covering all manner of general-interest topics such as DIY (Do-It-Yourself), medicine, cooking recipes, arts, and sciences. Find out more at http://www.yahoo.com.
13. J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An Architecture for General Intelligence," *Artificial Intelligence* **33**, 1–64 (1987).
14. We are indebted to S. E. Fahlman of IBM and Carnegie Mellon University (http://www-2.cs.cmu.edu/~sef/) for contributions to the sections on KR/KA and machine reasoning.
15. The Cyc knowledge base is an extensive knowledge base and inference engine system, with a core of over 1 000 000 hand-entered assertions (or "rules") designed to capture a large portion of what we consider knowledge about the world. The effort was pioneered by Doug Lenat in 1984. For more information, see the Cycorp, Inc. Web site at http://www.cyc.com/.
16. See the Linguistic Data Consortium page at the University of Pennsylvania, http://www.ldc.upenn.edu/.
17. See http://www.darpa.mil/ipto/research/com/index.html.
18. S. E. Fahlman, *NETL: A System for Representing Real World Knowledge*, MIT Press, Cambridge, MA (1979).

*Accepted for publication May 16, 2002.*

I. Brackenbury
IBM Software Group
Somers, New York

Y. Ravin
IBM Corporate Technology
Somers, New York




## An architecture of diversity for commonsense reasoning

Although computers excel at certain bounded tasks that are difficult for humans, such as solving integrals, they have difficulty performing commonsense tasks that are easy for humans, such as understanding stories. In this Technical Forum contribution, we discuss commonsense reasoning and what makes it difficult for computers. We contend that commonsense reasoning is too hard a problem to solve using any single artificial intelligence technique. We propose a multilevel architecture consisting of diverse reasoning and representation techniques that collaborate and reflect in order to allow the best techniques to be used for the many situations that arise in commonsense reasoning. We present story understanding—specifically, understanding and answering questions about progressively harder children's texts—as a task for evaluating and scaling up a commonsense reasoning system.

In the fall of 2001, a proposal was developed by Marvin Minsky, Erik Mueller, Doug Riecken, Push Singh, Aaron Sloman, and Oliver Steele for a project to develop a human-level commonsense reasoning system. The basic proposal was (1) to develop certain ideas of Minsky and Sloman about a multilevel cognitive architecture, and (2) to develop the system in a way that would exploit many existing artificial intelligence techniques for commonsense reasoning and knowledge representation, such as case-based reasoning, logic, neural nets, genetic algorithms, and heuristic search.

We proposed to organize a meeting at which we would bring together many of the major established researchers in the area of commonsense knowledge and reasoning. Riecken organized a preliminary meeting at the IBM Thomas J. Watson Research Center in March 2002, at which many IBM researchers were invited to discuss and react to this general subject as well as to present their own ideas. Afterwards, the specific proposal was discussed in more detail by specialists in commonsense knowledge and reasoning at a meeting held on St. Thomas, Virgin Islands, in April 2002, and hosted by Jeffrey Epstein. This Technical Forum contribution focuses on the preliminary meeting, but also contains some material presented at the April meeting, including some material from Minsky's forthcoming book *The Emotion Machine*.[1]

At the IBM meeting, a broad consensus was reached on three main points. First, there was agreement that the community should strive toward solving a nontrivial problem that would require a level of knowledge, and a capability of reasoning with that knowledge, beyond what is demonstrated by current systems. The problem put forward was that of story understanding. An important advantage of the story understanding task is that standardized tests are available to evaluate students on their reading comprehension skills. Moreover, these tests require the use of commonsense reasoning skills. It is thus possible to evaluate the performance of any story understanding system against that of students at different reading levels.[2]

Second, there was consensus that the story understanding task provides a strong testbed for evaluating a commonsense reasoning system. Not only does such a system need several different forms of reasoning, representation, and learning, but it also needs them to work in conjunction with each other. In addition, the task highlights the importance of using and reasoning with common sense. This is illustrated by a sentence from a story about a child and her grandfather: "He gently takes my elbow as we walk so that I can help show him the path." Knowledge of the fact that the grandfather is blind, and the commonsense facts that people ordinarily use their sight to find paths and that blind people are unable to see, enable the inference that the child is guiding the grandfather and not merely pointing out the path, another frequent sense of the word "show." Absence

of this commonsense knowledge could lead to the incorrect interpretation of the word "show."

Third, there was agreement on the need to develop a testbed architecture for representation and reasoning that allows different systems and representations to work with each other. Researchers often try to solve a problem using just one form of representation and reasoning. But such an approach does not work well for sufficiently complex problems such as story understanding. In contrast, enabling various techniques to collaborate will allow the best techniques to be used for a given situation. Any such architecture must provide metalevel control and knowledge that will enable different techniques to determine whether or not they are suited for a given task, to decide what other techniques may be better for the task, and to communicate information and share partial results with each other.

## What makes commonsense reasoning difficult

Commonsense reasoning—the sort of reasoning we would expect a child to do easily—is difficult for computers to do. Certainly, the relative paucity of results in this field does not reflect the considerable effort that has been expended, starting with McCarthy's paper "Programs with Common Sense."[3] Nevertheless, the problem remains unsolved. What is it about commonsense reasoning that makes it difficult to automate? Various explanations have been suggested, some of which we discuss in this section.

**McCarthy's commonsense informatic situation.** The knowledge needed to solve a commonsense reasoning problem is typically much more extensive and general than the knowledge needed to solve difficult problems. McCarthy points out that the knowledge needed to solve well-formulated problems in fields such as physics or mathematics is bounded.[4] In contrast, there are no *a priori* limitations to the facts that are needed to solve commonsense problems: the given knowledge may be incomplete; one may have to use approximate concepts and approximate theories; one will generally have to use nonmonotonic reasoning to reach conclusions; and one will need some ability to reflect upon one's own reasoning processes. Morgenstern provides an example of the commonsense informatic situation in the problem of two friends arranging to meet for dinner at a restaurant.[5]

**Explicit vs implicit knowledge.** Commonsense knowledge is often implicit, whereas the knowledge needed to solve well-formulated difficult problems is often explicit. For example, the knowledge needed to solve integrals can be found in explicit form in a standard calculus textbook. However, the knowledge needed to arrange a dinner meeting exists in vague, implicit form. Implicit knowledge must first be made explicit, which is a time-consuming task requiring a serious knowledge engineering effort.

**Domain knowledge.** A huge amount of knowledge is needed to do even simple forms of commonsense reasoning. For example, to figure out what sorts of objects will work as stakes in a garden—a reasoning task that seemingly demands no effort—requires knowledge of plant materials, how plants grow, flexibility and hardness, shapes of plants, soil texture, properties of wind, spatial reasoning, and temporal reasoning.[6] Although there have been a number of efforts to capture large amounts of world knowledge, most notably the Cyc** project,[7] we are not at this point aware of any knowledge base that contains the information necessary to reason about stakes in a garden or about fumbling for an object in one's pocket.

This Technical Forum piece does not present a solution to these difficulties. Rather, we are attempting to see how far we can progress on an important commonsense reasoning problem even in the presence of such difficulties.

## Story understanding as a vehicle for studying commonsense reasoning

Story understanding requires addressing the commonsense informatic situation. A story understanding system should be able to read and understand a story, and demonstrate its understanding by (1) answering questions about the story, (2) producing paraphrases and summaries of the story, and (3) integrating the information the story contains into a database. Further, useful results from this work will have a direct impact on many business products and services.

**A brief history of story understanding systems.** Starting in the 1960s,[8] researchers have studied story understanding and have built systems that can read and answer questions about simple stories. An early system built by Charniak[9] used a single mechanism, test-action demons, for making inferences in understanding. In the 1970s, Schank and Abelson[10] pro-

posed scripts, plans, and goals as knowledge structures for understanding. These knowledge structures were incorporated into the SAM[11] and PAM[12] story understanding systems.

In the 1980s, knowledge structures for emotions, story themes, and spatial/temporal maps were incorporated into BORIS.[13] AQUA[14] used case-based reasoning to retrieve and apply explanation patterns in order to answer questions raised by anomalies encountered while reading a story. CRAM[15] used a connectionist approach to story understanding.

Recent story understanding systems have adopted the approach of understanding a story by building and maintaining a simulation that models the mental and physical states and events described in the story, as demonstrated in ThoughtTreasure.[16] The advantage of this approach is that it is easy to answer questions about the story simply by examining the contents of the simulation.

Critical problems for story understanding systems. The story understanding systems built so far work only on the particular stories they are designed to handle. For example, SAM[11] handles five stories, BORIS[13] three, AQUA[14] five, and ThoughtTreasure[16] three. What prevents story understanding systems from scaling up to hundreds of previously unseen stories?

We contend that story understanding research is blocked on three critical problems: (1) complexity of the structure of natural language, (2) necessity for large commonsense knowledge bases, and (3) combinatorial explosion in the understanding process.

*Complexity of the structure of natural language.* Rare is the simple subject-verb-object sentence that maps into a simple proposition. More typically, text contains numerous language phenomena such as adverbials, compound nouns, direct and indirect speech, ellipsis, genitive constructions, and relative clauses.[17] Present-day syntactic and semantic parsers have trouble producing accurate parses of typical story sentences.

*Necessity for large commonsense knowledge bases.* Understanding even simple stories requires knowing a huge number of facts. For example, understanding the first paragraph of *The Cat in the Hat* requires knowing about children's play, how children can be affected by winter weather, their relationship to their parents, and notions of discipline, boredom, surprise, and risk. Similarly, as Davis[18] points out, the first paragraph of *The Tale of Benjamin Bunny* assumes familiarity with concepts of quantity, space, time, physics, goals, plans, needs, and communication.

*Combinatorial explosion in the understanding process.* Multiple possible interpretations arise at all levels of language. Words are ambiguous as to part of speech and word sense. Sentences are syntactically ambiguous. There are several possible explanations for any action of a story character, several possible explanations for those explanations, and so on. We get a combinatorial explosion: the understanding process must search an extremely large space of possibilities.

Approaches to critical problems in story understanding. What can be done? We propose a three-pronged approach. First, to deal with the complexity of the structure of natural language, we make a major cut in complexity by going back to books for early readers. Second, to deal with the necessity for large commonsense knowledge bases, we propose to identify the domains most frequently used in a restricted set of stories and to address these first. Last, to deal with the combinatorial explosion in the understanding process, we propose a new paradigm for commonsense reasoning: an architecture of diversity.

*Early readers.* Early reader texts are designed for preschool and kindergarten students. These texts employ a small or controlled vocabulary, short sentences, and limited language constructions. Working with early reader texts will enable us to effectively solve the language front-end problem using existing research techniques.

*Text annotation for domain identification.* We cannot hope to deal with the commonsense informatic situation head-on. The point of McCarthy's 1996 paper[4] is that any domain can be relevant to a particular problem: when reading a story, any area of knowledge may be necessary for comprehension. This is less true for stories designed for very young readers; although, as our examples above show, a great many concepts and domains are still needed for full comprehension even of early reader texts. Nevertheless, we believe we can make progress by choosing to address those domains that most frequently turn up in children's stories. Such an approach would, we hope, make the problem tractable.

We thus propose the following corpus-based approach. We start with a corpus of stories at the preschool and kindergarten levels and divide the corpus into a development set and a test set. We manually annotate each story in the development set with an informal inventory of what domains of commonsense knowledge and reasoning must be addressed in order to understand the story. We sort the domains by their frequency and attempt to develop methods to understand the domains that occur most frequently. We start with the most frequent domain, proceeding to the next most frequent domain, and so forth. Development proceeds on the development set, and a final evaluation of the generality of the system is conducted on the previously unseen test set. We iterate this process on successively higher reading levels, progressing to stories designed for Grades 1, 2, and 3. This approach, based on an incremental series of experiments, will enable a significant research focus at each step on an architecture of diversity.

To demonstrate how this approach would work, we formed a corpus of 15 early reader stories and annotated them as to the domains of common sense necessary for understanding them. The vocabulary size was 561 words. The top 10 domains of common sense are shown in Table 1. This provides us with a path for research in understanding the story corpus: focus on handling the most frequently appearing domains of common sense.

Dealing with these concepts is by no means trivial. We plan to leverage the extensive work that has been done in these areas. Such work includes: ThoughtTreasure,[16] NETL2,[19] Cyc,[7] Shanahan's formalization of time,[20] the RCC formalization of space,[21] and Kuipers's Spatial Semantic Hierarchy.[22] We will also employ rapid knowledge formation techniques such as Open Mind.[23]

## An architecture of diversity

Many attempts to build intelligent computers have hunted for a single mechanism (such as universal subgoaling, propagation rules, logical inference, probabilistic reasoning) or representation (such as production rules, connectionist networks, logical formulas, causal networks) that would serve as a basis for general intelligence. Why have these approaches so far failed to achieve human-level common sense?

Table 1 Early reader corpus: top 10 domains of common sense

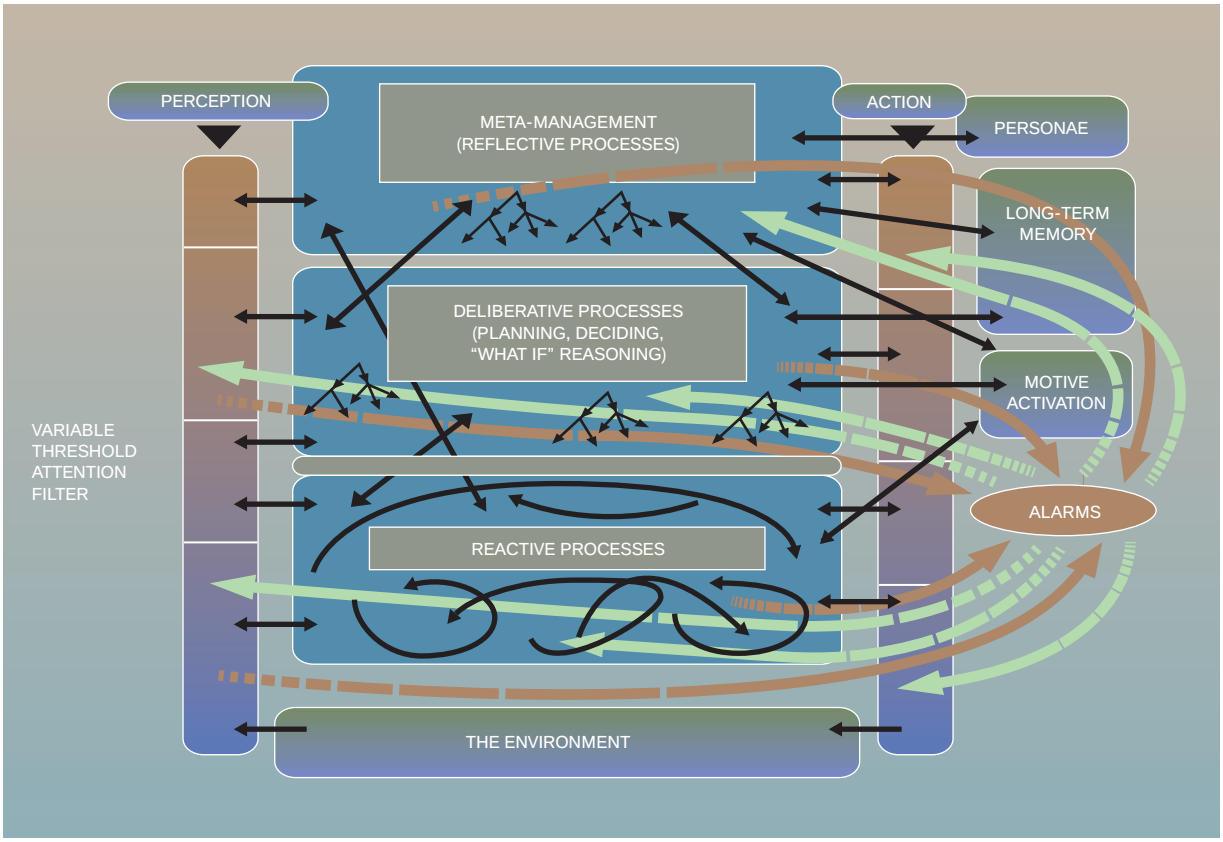
| Domain | Number of Stories | Percentage of Stories |
|---|---|---|
| space—location | 14 | 93.3 |
| space—motion | 11 | 73.3 |
| math—counting | 10 | 66.6 |
| attitude—positive | 9 | 60.0 |
| speech act | 9 | 60.0 |
| space—size | 8 | 53.3 |
| space—grasping | 7 | 46.6 |
| sound—speech | 7 | 46.6 |
| logic—universal quantification | 7 | 46.6 |
| space—housing | 6 | 40.0 |

We believe that the problem is too large to solve using any single approach. Human versatility must emerge from a large-scale architecture of diversity in which each of several different reasoning mechanisms and representations can help overcome the deficiencies of the other ones.[24,1] Our hypothesis is that such an architecture can overcome the combinatorial explosion problem in story understanding.

**Multilevel cognitive architecture.** We conjecture that the information processing architecture of a human is something like the three-level architecture developed by Sloman in the Cognition and Affect project[25] (H-Cogaff), shown in Figure 1. This conjecture is based on evidence of many kinds from several disciplines, and constraints on evolvability, implementability in neural mechanisms, and functionality.[26]

Reactive processes are those in which internal or external states detected by sensors immediately trigger internal or external responses. Deliberative processes are those in which alternative possibilities for action can be considered, categorized, evaluated, and selected or rejected. More generally a deliberative mechanism may be capable of counterfactual reasoning about the past and present and hypothetical reasoning about the future. The depth, precision, and validity of such reasoning can vary. Meta-management processes add the ability to monitor, evaluate, and to some extent control processes occurring within the system in much the same way as the whole system observes and acts on the environment. The three layers operate concurrently and do not form a simple dominance hierarchy. Arrows represent flow of information and control, and boundaries need not be sharp in all implementations.

Figure 1 The H-Cogaff three-level architecture




A. Sloman, "Beyond Shallow Models of Emotion," *Cognitive Processing,* Vol. 1, No. 1 (2001).

The reactive and deliberative layers differ in that the deliberative layer evolved much later and requires a far more sophisticated long-term memory, as well as symbolic reasoning capabilities using a short-term reusable memory. The meta-management layer may have evolved at a still later time and requires explicit use of concepts referring to states of an information processing architecture. The earliest organisms, such as most existing organisms, were totally reactive. Deliberative and meta-management layers evolved later. Adult humans appear to have all three types of processing, which is probably rare among other animals.

One of the key features that gives H-Cogaff its generality is the fact that different components, instead of forming parts of simple pipelines, can concurrently send information of various kinds to arbitrarily many other components, allowing a wide variety of feedback mechanisms and triggering mechanisms.

In story understanding, the meta-management level may control the deliberative level in a number of ways.

- If the deliberative level is spending too much time considering certain details and those details are not crucial to the story, the meta-management level will make the deliberative level stop.
- If the deliberative level is spending too much time on a task that does not relate to the goal of reading the story, the meta-management level will make the deliberative level stop.
- If the deliberative level becomes confused, the meta-management level will tell it to go back and reread. The deliberative level may have ruled out a possibility earlier that needs to be reconsidered in light of new information.

Minsky further elaborates the H-Cogaff architecture into the six-level architecture called "Model Six"

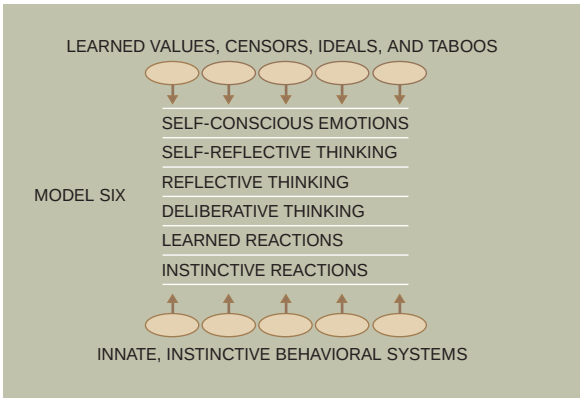
shown in Figure 2.[1] At its bottom lies a "zoo of instinctive subanimals" built upon ancient, ancestral systems that still maintain our bodies and brains. These include systems for feeding, breathing, heating, sleeping, and other systems that keep us alive. The deliberative and reflective levels are engaged to solve more difficult kinds of problems. The self-reflective level is engaged when the problems involve our relationships with our past and future selves. At the top lies machinery that we acquire from our societies, such as suppressors and censors, imprimers and values, and our various kinds of self-ideals.

Multiple reasoning and representation schemes and levels. An architecture of diversity would embed representations from natural language to micronemes[27,1] as depicted in Figure 3. The representations depicted include frames, transframes, frame-arrays, K-lines, and micronemes. A frame is a representation based on a set of slots to which other structures can be attached.[28] Each slot is connected to a default assumption that is easily displaced by more specific information. A transframe is a particular type of frame representing the causal trajectory between the initial and resulting states representing a situation that a legal action was performed on. A frame-array is a collection of frames that share the same slots, making it easy to change perspective with respect to physical viewpoint or other mental realms. A knowledge-line or K-line is a wirelike structure that attaches itself to whichever resources are active in solving a problem. The K-line simplifies activation of those same resources when solving a similar problem in the future. Micronemes are low-level features for representing the many cognitive shades and hues of a context. In Figure 3, new evolved structures are made from older lower-level ones, and the tower shown might be a plausible Darwinian brain-development scheme.

Table 2 shows just a few of the diverse representation and reasoning schemes useful for domains of story understanding.

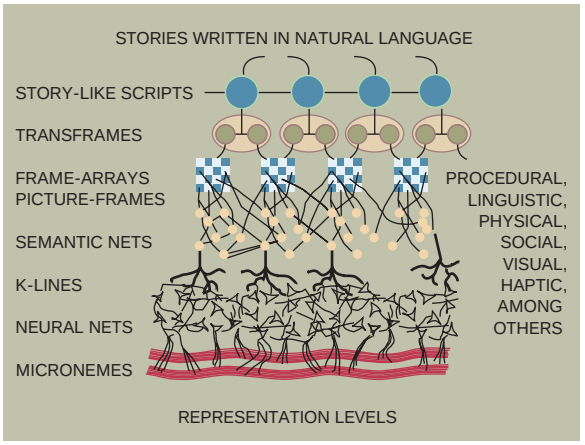
We propose to address the commonsense reasoning problem starting with stories for very young readers. However, to demonstrate all of the different ways we think when understanding a story, and what we would eventually expect a commonsense story understanding system to be able to handle, consider the following adult story (the discussion here is condensed from Reference 1).

Figure 2 The Model Six six-level architecture




M. Minsky, *The Emotion Machine*, Pantheon, New York (forthcoming).

Figure 3 Multiple reasoning and representation schemes and levels




From M. Minsky, "Common Sense-Based Interfaces," *Communications of the ACM,* Vol. 43, No. 8, 67-73 (2001). Copyright 2001 ACM. Reprinted by permission.

*Joan heard a ring and picked up the phone. Charles was answering her question about how to use a certain technique. He suggested she read a certain book, which he would soon bring to her since he had planned to be in her neighborhood. Joan thanked him and ended the call. Soon Charles arrived and gave her the book.*

Following are a few of the understandings an adult reader would have after hearing the story.

Table 2 Diverse schemes for story understanding domains

| Domain | Representation/Reasoning Schemes |
|---|---|
| space | frame, generalized cylinder model, interval logic, occupancy grid |
| time, action effects | causal model, event calculus, situation calculus, transframe |
| reactivity | neural net, production system, subsumption architecture |
| schemas, scripts | finite automaton, frame, frame-array, generalized Petri net |
| subgoaling | first-order logic, K-line, marker passing, semantic net |
| emotions, attitudes | microneme, neural net, temporal modal logic |

- *Joan heard a ring*. She recognizes it as a telephone bell and feels the need to respond quickly. She knows how to use the telephone.
- *She picked up the phone*. She is subsequently holding the phone to her ear.
- *Charles was answering her question*. Charles and Joan are not in the same room. Charles also knows how to use the telephone.
- *He suggested she read a certain book*. Joan probably now feels some relief, since she knows where to find the knowledge she needs.
- *He had planned to be in her neighborhood*. Joan will not be surprised when he arrives, because she will remember that he said he would come.
- *He gave her the book*. Will she have to give it back? The story does not tell us that.

These conclusions are based on reasoning and representations in many realms, as follow.

***The physical realm.*** In this realm, *give* might mean the motion of the book through space. This could be represented as a transframe that starts with Charles's hand holding the book and ends with Joan's hand carrying it. One must know a lot about physical things and how they behave in space and time.

***The social realm.*** In this realm, *give* may signify social acts that can alter the relationships of the actors. What were Charles's motives or his attitudes? Clearly, he was not returning a loan. Was he hoping to ingratiate himself? Or was he just being generous? How will Joan feel about Charles after he gives her the book? One must know a lot about what people are, and a certain amount about how people work.

***The dominion realm.*** Given *Charles gave Joan the book*, one infers not only that Joan is holding the book, but also that, at least for a time, she possesses the right to use it.

***The conversational realm.*** How do conversations work? Consider how many elaborate skills are involved in a typical verbal exchange. One has to keep track of what is being discussed, what one has previously told the listener, and what the listener knows. Thus conversations are partly based on knowledge of how human memories work and what is commonly known in one's culture. One has to make sure the listener has understood what was said and why it was said. One certainly needs to know how to speak and to understand some of what one may hear.

***The procedural realm.*** How does one make a telephone call? One must first find a phone and dial a number. Then once the connection has been established, one says hello, talks a bit, and eventually leads into why one called. At the end, one says goodbye and hangs up the phone. Generally, such scripts have certain steps that are specified, while other steps provide for more room to improvise.

***The sensory and motor realms.*** Each of the above steps raises questions. For example, it takes only one second or so for one's arm to reach out in order to pick up the phone. How can one do that so quickly?

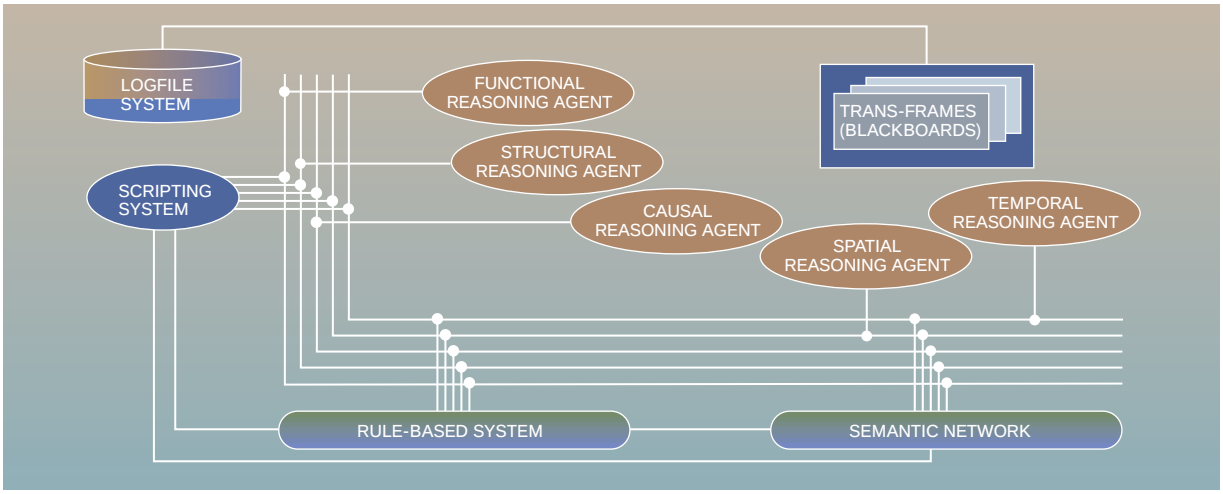
***The kinesthetic, tactile, and haptic realms.*** Using a telephone or any other physical object engages a great base of body-related knowledge and skills. One anticipates how the phone will feel against one's ear or sandwiched between shoulder and cheek. One expects certain haptic sensations such as the feel of the phone's weight. One strengthens one's grip when the phone starts to slip.

***The temporal realms.*** People have elaborate models of time where events are located in futures and pasts that are represented in relation to other times and events or in anecdotal stories.

***The economic realm.*** People know and reason about the costs incurred by each action or transaction in terms of money, energy, space, or time.

***The reflective realm.*** People know about themselves. One knows to some degree what one can or cannot do, what kinds of problems one can solve, how one's thinking and memory works, and what sorts of things one is able to learn.

Figure 4 The M system




From D. Riecken, "An Architecture of Integrated Agents," *Communications of the ACM,* Vol. 37, No. 7, 107-116 (1994). Copyright 1994 ACM. Reprinted by permission.

Along with these positive kinds of knowledge, one also has negative knowledge about what might go wrong when using a phone. One must know what to do if one gets a wrong number, if there is no answer, or if a modem or intercept recording is reached.

Example system with architecture of diversity. Thus far, the Sloman and Minsky architectures are theoretical constructs and have not yet been implemented. However, there are examples of working systems that capture the spirit of such architectures. One such example is the M system depicted in Figure 4.[29] M integrates multiple reasoning processes and representations to serve as an assistant to a user collaborating with other workers within a virtual meeting room that hosts multimedia desktop conferencing. M serves to recognize and classify the actions performed by the participants as well as the objects upon which the actions are applied; example actions and objects are brainstorming on a whiteboard, coauthoring a document, and creating and working with other artifacts.

## Next steps

The two recent meetings held in March 2002 at the IBM Thomas J. Watson Research Center and in April 2002 on St. Thomas indicate that there is a dedicated group of recognized researchers interested in working together on a project to develop a solution to commonsense reasoning. We are now planning to undertake some of the next steps in a plan for such a project. The inspiration for this work comes from Minsky's past and forthcoming work. We close with his thoughts on how such a project might be realized, as follows.

Our goal is to aim toward a critical "change of phase" that will come when we cross a threshold at which our systems know how to improve themselves. This is something that all young children can do, but we do not know enough about how they do it; so one goal of the project must be to develop better models of how normal people think.

We will start by trying to implement some of the architectures proposed over the past decade. There already exist many useful schemes for representing and using knowledge mostly of a factual nature for use on what we call the deliberative level. However, there has not been enough work on the higher reflective and self-reflective levels that humans use, as they learn to improve their thinking itself. Any such system, we claim, will need additional kinds of meta-resources, which will include systems that manage, criticize, and modify the already operating parts of the structure.

In the field of AI we already have many resources related to this, for example, neural networks, for-

mal logic, relational databases, genetic programs, statistical methods, and of course the heuristic search, planning, and case-based reasoning schemes of earlier years. However, our goal is not to discuss which method is best. Instead we will try to develop a plan of how to incorporate into one system the virtues of many different approaches. Of course, each such scheme has deficiencies and our hope is that our system can escape from these by using higher-level, more reflective schemes that understand what each of those other schemes can do and in what context they are most effective.

**Trademark or registered trademark of Cycorp, Inc.

## Cited references and notes


1. M. Minsky, *The Emotion Machine*, Pantheon, New York (forthcoming). Several chapters are on line at http://web.media.mit.edu/~minsky.
2. The use of reading comprehension tests as a metric for evaluating story understanding systems was previously proposed in L. Hirschman, M. Light, E. Breck, and J. Burger, "Deep Read: A Reading Comprehension System," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, June 1999, Association for Computational Linguistics (1999).
3. J. McCarthy, "Programs with Common Sense," *Proceedings of the Symposium on Mechanisation of Thought Processes*, Her Majesty's Stationery Office, London (1958), pp. 77–84.
4. J. McCarthy, "From Here to Human-Level Intelligence," *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning* (KR'96), Cambridge, MA, November 1996, Morgan Kaufmann, San Mateo, CA (1996), pp. 640–646.
5. L. Morgenstern, "A Formal Theory of Multiple Agent Nonmonotonic Reasoning," *Proceedings of the Eighth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA (1990), pp. 538–544.
6. E. Davis, "The Naive Physics Perplex," *AI Magazine* **19**, No. 4, 51–79 (1998).
7. D. Lenat, "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM* **38**, No. 11, 32–38 (1995).
8. More details can be found in E. T. Mueller, "Story Understanding," to appear in *Encyclopedia of Cognitive Science*, Nature Publishing Group, London (2002).
9. E. Charniak, *Toward a Model of Children's Story Comprehension*, Technical Report AITR-266, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA (1972).
10. R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding*, L. Erlbaum Associates, Hillsdale, NJ (1977).
11. R. E. Cullingford, *Script Application: Computer Understanding of Newspaper Stories*, Technical Report YALE/DCS/tr116, Computer Science Department, Yale University, New Haven, CT (1978).
12. R. Wilensky, *Understanding Goal-Based Stories*, Technical Report YALE/DCS/tr140, Computer Science Department, Yale University, New Haven, CT (1978).
13. M. G. Dyer, *In-Depth Understanding*, MIT Press, Cambridge, MA (1983).
14. A. Ram, *Question-Driven Understanding: An Integrated Theory of Story Understanding, Memory, and Learning*, Technical Report YALE/DCS/tr710, Computer Science Department, Yale University, New Haven, CT (1989).
15. C. Dolan, *Tensor Manipulation Networks: Connectionist and Symbolic Approaches to Comprehension, Learning, and Planning*, Technical Report 890030, Computer Science Department, University of California, Los Angeles, CA (1989).
16. E. T. Mueller, *Natural Language Processing with ThoughtTreasure*, Signiform, New York (1998), full text of book available on line at http://www.signiform.com/tt/book/.
17. L. G. Alexander, *Longman English Grammar*, Longman, London (1988).
18. E. Davis, *Representations of Commonsense Knowledge*, Morgan Kauffman, San Mateo, CA (1990).
19. S. E. Fahlman, *NETL: A System for Representing and Using Real-World Knowledge*, MIT Press, Cambridge, MA (1979).
20. M. Shanahan, *Solving the Frame Problem*, MIT Press, Cambridge, MA (1997).
21. D. A. Randell, Z. Cui, and A. G. Cohn, "A Spatial Logic Based on Regions and Connection," *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, CA (1992), pp. 165–176.
22. B. Kuipers, "The Spatial Semantic Hierarchy," *Artificial Intelligence* **119**, 191–233 (2000).
23. P. Singh, "The Public Acquisition of Commonsense Knowledge," *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA, March 2002, American Association for Artificial Intelligence (2002).
24. M. Minsky, *The Society of Mind*, Simon & Schuster, New York (1985).
25. A. Sloman, "Beyond Shallow Models of Emotion," *Cognitive Processing* **1**, No. 1 (2001).
26. A. Sloman, "Architectural Requirements for Human-Like Agents both Natural and Artificial," K. Dautenhahn, Editor, *Human Cognition and Social Agent Technology*, John Benjamins, Amsterdam (2000), pp. 163–195.
27. M. Minsky, "Common Sense-Based Interfaces," *Communications of the ACM* **43**, No. 8, 67–73 (2001).
28. M. Minsky, "A Framework for Representing Knowledge," AI Laboratory Memo 306, Artificial Intelligence Laboratory, Massachusetts Institute of Technology (1974), reprinted in *The Psychology of Computer Vision*, Patrick Winston, Editor, McGraw-Hill, New York (1975).
29. D. Riecken, "An Architecture of Integrated Agents," *Communications of the ACM* **37**, No. 7, 107–116 (1994).


*Accepted for publication May 17, 2002.*

J. McCarthy
Stanford University
Stanford, California

M. Minsky
Massachusetts Institute of Technology
Cambridge, Massachusetts

A. Sloman
University of Birmingham
Birmingham, UK

L. Gong
IBM Research Division
Hawthorne, New York

T. Lau
IBM Research Division
Hawthorne, New York

L. Morgenstern
IBM Research Division
Hawthorne, New York

E. T. Mueller
IBM Research Division
Hawthorne, New York

D. Riecken
IBM Research Division
Hawthorne, New York

M. Singh
IBM Research Division
Hawthorne, New York

P. Singh
Massachusetts Institute of Technology
Cambridge, Massachusetts