From 2001 to 2001: Common Sense and the Mind of HAL

Douglas B. Lenat

Making mistakes is in the nature of being human. I'll spare you the usual quote about forgiveness being divine, because I certainly have never forgiven HAL. We all felt bad when HAL terminated the cryogenically slumbering crew, cut Frank adrift, and almost murdered Dave. But that's not what I found so unforgiveable. To me, HAL's biggest crimes were his conceit and his stupidity.

By conceit, I mean claims like "No 9000 computer has ever made a mistake." This is more than just arrogant, more than just false; it is the *antithesis* of realism. If you met a man who genuinely believed he never had nor ever would make a mistake, you'd call him insane. Surely NASA would never have entrusted the mission to such a patently insane computer. I'll return to this point a bit later.

By stupidity, I mean his resort to extreme violence—murdering the entire crew—to solve his problems. Yes, he was facing a dilemma: should he jeopardize his secret orders, or should he disobey the order to keep them secret from the crew? This sort of dilemma is no more or less than the makings of good drama. HAL's solution was the same one Shakespeare employed in his blackest tragedies, the same one Clint Eastwood employed in his man-with-no-name movies: just kill everyone.

In the late 1970s I built a computer program (Eurisko) that discovered things on its own in many fields. To get it to work, I had to give it the power to tinker with its own learning heuristics and its own goals. I would leave it running overnight and hurry in the next morning to see what it had come up with. Often I'd find it in a mode best described as "dead." Sometime during the night, Eurisko would decide that the best thing to do was to commit suicide and shut itself off. More precisely, it modified its own judgmental rules in a way that valued "making no errors at all" as highly as "making productive new discoveries." As soon as Eurisko did this, it found it could successfully meet its new goal by doing nothing at all for the rest of the night. This reminds me of HAL's boast: "No 9000 computer has ever made a mistake." I eventually had to add a new heuristic to Eurisko—one it couldn't modify in any way—to explicitly forbid this sort of suicide.

People have found many ways to grapple with and resolve conflicting goals short of killing everybody in sight. Surviving and thriving in the real world means constantly making tough decisions, and, yes, making mistakes. The only ways not to make mistakes are Eurisko's -- do nothing -- HAL/Shakespeare/Eastwood's -- make sure there are no living souls left anywhere around you -- and God's -- be omniscient. HAL, if he were really smart, could have found another solution, just as we do every day.

Surely, any intelligent computer would understand that occasional mistakes and inconsistencies are inevitable and can even serve as valuable learning experiences. Anything else leads to what I call "Star Trek brittleness –the absurdity that holds that one

small inconsistency will make the computer self-destruct.

As humans we tolerate inconsistency all the time. Some inconsistency stems from different levels of generality and precision. We know about Einstein's relativity, but in our everyday lives we act as though it doesn't exist; we know the earth revolves around the sun, but most of the time we talk and act as if the sun moves around the earth. Other instances of inconsistency are a result of different epistemological statuses: we both know that there are no vampires and that Dracula is a vampire. Some of it comes from information we learned in different ways at different times, such as childhood phobias that persist in adult life even though rationally we know them to be groundless. Yet other inconsistencies result from other people's inconsistencies, which they pass on to us.

So, just how smart *was* HAL? And how does Arthur C Clarke's vision of computer intelligence compare with the reality of the HAL-like programs we can build today?

The Knowledge Pump

There is a lot of controversy about how human-level machine intelligence will develop. Some scientists believe it will follow a path similar to the one followed in nature by evolution: there will be artificial one-celled animals, artificial insects, artificial lawyers, artificial monkeys, and so on up to artificial human-level machine minds. Nature has made good use of trillionfold parallelism for hundreds of millions of years; so it's no surprise that some folks expect computer hardware and raw computing power to create similar bottlenecks that we will ultimately overcome.

In an alternative view of machine-intelligence development, personal computers are already so powerful that they are not the bottleneck problem at all. This is my view.

This view is best likened to priming a pump. Visualize your brain as a knowledge pump. Knowledge goes in, gets stored, combined, copied, or whatever; from time to time, you say or write or do things that are, in effect, ways for your brain to emit knowledge. On a good day, the knowledge you give out may be as good or better than the knowledge you put in.

No one expects you to be a productive knowledge pump without training and experience -- whether we're talking about playing the piano or tennis, writing a check or a novel, or making a U-turn in a car. You have to invest some learning-and-teaching time and effort before anyone expects you to be competent at a task, let alone to excel at it.

Consider Dr. Dave Bowman, mission commander of *Discovery.* His experiences as an engineering or astrophysics student and his astronaut training qualified him to lead the mission. Before that, his high school and undergraduate college prepared him for graduate school, and before that, he learned in elementary and middle school the fundamentals he needed for high school. And long before -- here we get to some very important stuff indeed -- his early experiences as a baby and toddler prepared him for kindergarten and first grade. He learned, for instance, to talk and that people generally prepare food in

kitchens. He learned that if you leave something somewhere, it often remains there, at least for a while. He learned that chairs are for sitting on and that pouring milk from one glass to another differently shaped glass doesn't change the total volume of the milk. He learned that there's no air in outer space so he'd better not forget the helmet of his spacesuit if he's going walking in space. He learned all this -- and a million other things.

It Takes Common Sense to Understand Each Other

There is a crucial point that most creators of science-fiction robots (from Robby to HAL to Data) seem to have ignored, and it is this: You need to have quite a bit of "common sense" before you can learn to talk, and before you can survive on your own in the everyday world.

Consider the first thing HAL tells Frank: "We've got the transmission from your parents coming in." There are three possible ways to interpret this sentence, depending on who or what exactly is "coming up." It could be *we* who are coming in, as in "We've got a lot of anxiety coming into this room." It could be *Frank's parents* who are coming in, as in "We looked up and saw our parents coming in." Or it could be -- and is -- the *transmission* that is coming in. Anyone with common sense could figure that out, and we assume that we all have common sense. It would be insulting or confusing to make our sentences longer just to clarify exactly what goes with what.

Consider next what HAL says to Dave when he finds that Dave has just drawn a new sketch of Dr. Hunter: "Can you hold it a bit closer?" This sentence is rife with ambiguity, though neither HAL nor Dave nor the audience appears to notice. Does HAL want a yes-or-no response to his question? Of course not, he wants Dave to move the drawing. Does he want Dave to move it closer to Dave or closer to HAL? Obviously the latter. But if Dave had been taking a zero-gravity tennis lesson from HAL and HAL had said "Can you hold it a bit closer?" we would all assume he wanted Dave to hold the tennis racket closer to Dave, not to HAL. To figure out that HAL is asking Dave to hold the sketch closer to HAL, one has to bring common sense into play. You can't appreciate art without seeing it, and the more clearly you can see it, the better you can appreciate or critique it. And, of course, the closer an object gets to your eyes (or to HAL's visual sensors), the more clearly you can make out its details, and so on. Similarly, in the tennis example, other simple facts -- such as the relevance of how close Dave holds the racket to his ability to hit the ball well -- would dissolve the ambiguity. All this is just common sense, the sort of things you learn as a baby and toddler. Yet you either know this mass of trivia or you don't; and if you don't, how can you tell that the picture should move closer to HAL but the tennis racket should go closer to Dave's body?

HAL's very next sentence, after Dave moves the sketch closer to HAL's sensor lens, is, "That's Dr. Hunter, isn't it?" Dave replies, naturally, "Yes," rather than "No, it's a sketch of Dr. Hunter" -- another illustration of the way we depend on shared common sense to keep our sentences short. HAL knows it's a sketch, not a person, and Dave knows HAL knows this, and so on. These shared understandings let them communicate successfully

despite the terseness of their exchanges.

Similarly, HAL's response to Dave during Frank's extra-vehicular activity (EVA) -- "The radio is still dead" -- creates no confusion in Dave's mind. He knows that it violates common sense to even consider the possibility that the radio might suddenly become a living creature. HAL has to understand this -- and a great many other things -- in order to generate sentences using colorful language, metaphor, colloquialisms, and various other sorts of "realistic" ambiguity. Even something as innocuous as verb tenses -- dealing with time -- requires some common sense. Should HAL answer Dave's question about the radio in terms of the moment he starts asking it or finishes asking it, or when HAL starts to reply? Of course, Dave wants the latest information about Frank's status, and HAL can use present tense to convey all of the above. Otherwise, his speech would be so stilted he'd sound like a... well, like an unintelligent machine: "The radio transmitter, which is located in Frank's spacesuit, was nonfunctional when you asked your question and remains so now as I answer it."

Consider the dramatic sentence just before the lipreading: "Do you read me, HAL?" Clearly, Dave is not using the most common definition of *read* but a rarer meaning, *to be receiving successfully via radio.* HAL remains silent, not because he misunderstands the sentence, but because he's intentionally deceiving Dave.

An even clearer case of HAL's use of metaphor occurs when he raises for the first time the possibility that something is wrong: "Well, certainly no one could have been unaware of the very strange stories *floating around* before we left." In addition to using *floating* metaphorically, HAL is employing, in this one short sentence, all of the following: sophisticated double negation ("no one ... unaware"), a counterfactual construction ("no one could have been"), and an assumption about context. *("No one* refers not to all the people on Earth, almost none of whom were aware of anything strange at launch time -- other than, possibly, that an epidemic had broken out on the moon.)

Finally, consider the chilling sentence HAL utters after Dave says he'll use the emergency airlock: "Forgot your space helmet, Dave." The subject of the sentence could be Dave, or HAL, or someone else. But, of course, Dave, HAL, and the audience know exactly who has forgotten the helmet. Now consider the word *forgot* in this sentence. It is used here in the sense of "left behind, a while ago" rather than "didn't think of it, just now." Presumably, the fact that he was helmetless was very much on Dave's mind just then as he decided to use the emergency airlock. He had left it behind several minutes earlier, but had no doubt been actively regretting his oversight for the past several seconds when HAL pointed it out. HAL knows this too, which is what makes his saying it so chilling: HAL is being cruel, and taunting Dave, not trying to be helpful.

We could give many more examples. In fact, most of the sentences uttered in the film exhibit the same phenomena: each party has quite a bit of common sense and assumes that others do too; those assumptions let them all encode and decode each other's utterances, use fewer words, employ metaphor and ambiguity, and deviate casually from the strict rules of grammar.

In life, this terse encoding is sometimes rather extreme: for example, between twins or long-married couples, colleagues working in a technical area, or bridge partners who have played together for a long time. These people draw on, not just common sense, but particular shared experiences, agreed-upon conventions, common technical expertise, and so on. As a result, those of us who lack that shared knowledge and experience often don't understand much of what they say to each other, In the same way, today's computers, which don't share even the common-sense knowledge we all draw on in our everyday speech and writing, can't comprehend most of our speech or texts.

It Takes Common Sense to Stay Focused, and to Learn

Our human dependence on common sense is very far-reaching. It comes into play with spoken and written language (as when we try to decipher someone's scratchy handwriting) and in our actions (e.g., when driving a car and deciding whether to brake or accelerate or swerve to avoid something). Before we let robotic chauffeurs drive around our streets, I'd want the automated driver to have general common sense about the value of a cat versus a child versus a car bumper, about children chasing balls into streets, about young dogs being more likely to dart in front of cars than old dogs (which, in turn, are more likely to bolt than elm trees are), about death being a very undesirable thing, and so on. That "and so on" obscures a massive amount of general knowledge of the everyday world without which no human or machine driver should be on the road, at least not near me on in any populated area.

Our simple common-sense models of the world don't just clarify possible ambiguities; they are good enough to provide a context, a way of restricting reasoning to potentially relevant information and excluding irrelevant data. Suppose, for example, I'm trying to find my Visa card, which seems to be lost. Various things might be relevant to my search: the last thing I bought with it, the places I went yesterday, and so on. But I'll give up and report it lost before I bother trying to use all the pieces of information I possess, whether the number of legs on an arachnid, the year that Abraham Lincoln was elected president, or my mother's birthday.

Similarly, if someone were to ask you "Is Bill Clinton standing or sitting right now?" you would recognize right away -- probably in one or two seconds -- that you don't know the answer, despite all the miscellaneous facts about Clinton that you *do* know.

In a course on the calculus of manifolds I took about the time *2001* was released, we stated and proved Sard's theorem one day near the end of the term. The statement of the theorem was a couple lines long, and the proof wasn't much longer; it cited a couple of lemmas (auxiliary propositions) we'd proved the preceding week. But stating that theorem, let alone proving it, would have been a daunting task if we hadn't had the way prepared for us by a series of useful definitions and stepping-stone lemmas. They, in turn, presumed a certain level of what is vaguely termed *mathematical maturity* which is usually interpreted as a set of prerequisites, courses that have to be taken before signing up for that particular class.

This is a very technical example of learning a new thing that relates only tangentially to what we already know, at the "fringes," so to speak. But the phenomenon is important in everyday life as well as in math classes. We learn new things by extending and combining and contrasting already-assimilated concepts, facts, heuristics, models, and so on. It's hard to explain the need for sanitation to people who don't know the first thing about bacteria, and easy to do it with those who do. A particularly powerful way to teach someone about a new subject is to use a storytelling model, although even this method is less effective if the lives of listener and storyteller are too different.

The aliens in *2001* understand this all too well. They placed their signaling beacon on the moon so that they wouldn't have to pay attention to these intelligent apes until they hit "the knee of the learning curve." Once humans achieved even primitive space travel, the aliens decided, they were

beginning to have a larger and larger "fringe" of knowledge, which would grow exponentially, each new discovery reinforcing and accelerating the next one, like a snowball gathering mass as it rolls downhill. When that happened, the extraterrestrials wanted to be paged.

These examples illustrate how important it is to have a fair amount of common knowledge to understand written/spoken/handwritten sentences, to drive a car, to find your keys, to answer a question, to learn new things. In other words, before any future HAL could be entrusted with absolute power over the ship's functions -- or could even hold a casual conversation with a crew member -- it would somehow have to acquire this massive prerequisite store of knowledge. You can think of this knowledge as the foundation of consensus reality, things that are so fundamental that anyone who doesn't know and believe them is, in effect, living in a different world.

As Dave disconnects HAL's cognitive memory modules, HAL is reduced at some point in the procedure, to the same blank slate he was when he was first powered up, the *tabula rasa* onto which all his programming and education were subsequently written. Yet in this scene we hear the newborn HAL carrying on a conversation with Dave, asking whether Dave wants him to sing "Daisy, Daisy," and so on. This is one of ***2001's*** few technically unrealistic moments. As our examples illustrate, even simple linguistic behavior requires lots of general knowledge about the world, not to mention specific knowledge about the speaker and the context of the conversation. So, when Dave blanked out HAL's mind, the ability to hold such a conversation would have been one of the first abilities to go, not the last one.

How to Build HAL Today in Three Easy Steps

We're now in a position to specify the steps required to bring a HAL-like being into existence.

1. Prime the pump with the millions of everyday terms, concepts, facts, and rules of

thumb that comprise human consensus reality -- that is, common sense.

2. On top of this base, construct the ability to communicate in a natural language, such as English. Let the HAL-to-be use that ability to vastly enlarge its knowledge base.

3. Eventually, as it reaches the frontier of human knowledge in some area, there will be no one left to talk to about it, so it will need to perform experiments to make further headway in that area.

These steps aren't quite so separate as the list makes them appear. The step-i type of explicit, manual teaching will have to go on continually, even when steps 2 and 3 are well underway. Step 2 conversations will continue even after the computer reaches step 3 -- not only in other fields but even in the field with which step 3 is concerned; that is, the computer will probably want to discuss its discoveries with other researchers in the field.

Of course, the first step is both immensely difficult and immensely time-consuming. What are the millions of things that we should use to prime the new HAL's knowledge pump? How should they be represented inside the machine so that it can use them efficiently to deduce further conclusions when needed, just as we would? Who will do the actual entering of all that data? Assuming it's done by a large group of individuals, how will they keep from diverging and contradicting each other?

It may surprise you to hear that this is not just a fanciful blueprint for some massive future endeavor to be launched when humanity reaches a higher plateau of utopian cooperation. It is, in fact, the specific plan I and my team have been following for the past dozen years. In the next section, I report on our progress and our prospects for the future.

The CYC Project: Taking That First Step

In the fall of 1984, Admiral Bobby Ray Inman convinced me that if I was serious about taking that first step, I needed to leave academia and come to his newly formed MCC (Microelectronics and Computer Consortium) in Austin, Texas, and assemble a team to do it. The idea was that over the next decade dozens of individuals would create a program, CYC, with common

sense. We would "prime the knowledge pump" by handcrafting and spoon-feeding CYC with a couple of million important facts and rules of thumb. The goal was to give CYC enough knowledge by the late 1990s to enable it to learn more by means of natural language conversations and reading (step 2). Soon thereafter, say by 2001, we planned to have it learning on its own, by automated-discovery methods guided by models or minitheories of the real world (step 3).

To a large extent, that's just what we did. At the end of 1994, the CYC program was mature enough to spin off from MCC as a new company -- Cycorp -- to commercialize the technology and begin its widespread deployment.

Our purpose was not to understand more about how the human mind works, nor to test

some particular theory of intelligence. Instead, we built nothing more nor less than an artifact, taking a very nuts-and-bolts engineering approach to the whole project.

What Should the CYC System Know?

The first problem we faced was *what* knowledge to represent. Although we expected encyclopedias to play an important role, within a few months we realized that what they contain is almost the *complement* of common sense. Assuming that readers already have common sense, can read, and so on, they provide the next level of detail for reference purposes.

If we couldn't use encyclopedias for their content directly, we could still use their information indirectly. If we take any sentence from an encyclopedia article and think about what the writer assumes the reader already knows about the world, we will have something worth telling CYC. Alternatively, we can take a paragraph and look at the "leaps" from one sentence to the next and think about what the writer assumes the reader will infer "between" the sentences. For instance, back in 1984 our first example read, "Napoleon died on St. Helena. Wellington was greatly saddened." The author expects the reader to infer that Wellington heard about Napoleon's death, that Wellington outlived Napoleon, and so on.

For many years, we were largely driven by bottom-up examples of this sort from encyclopedias, newspapers, novels, advertisements, and so on. Gradually, around 1990, we began to work in a more top-down fashion, treating entire topics one at a time and in moderate detail. By 1996, we had told CYC about hundreds of topics. That brings up the next issue.

How Should That Knowledge Be Represented?

The real physical universe is not, of course, inside CYC, anymore than it is inside our brains. All we have is a representation of a sliver of the world, and we operate from that representation to acquire new ideas, make decisions, and so forth.

Initially we used a simple frame-and-slot language to store information in CYC; for instance, "timeOfBirth (HAL) = 1/12/1992 ." This caused several problems, however. How could we represent *not, or, every, some,* opinions, expectations, counterfactual conditionals, and similar material. Consider, for example, these speeches from *2001:*

> *HAL:* I hope the two of you are not concerned about this.
>
> *Dave:* No, I'm not, HAL.
>
> *Dave:* I don't know what you're talking about.

HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

HAL: I know everything hasn't been quite right with me, but I can assure you now, very confidently, that it's going to be alright again.

HAL: If you'd like to hear it, I can sing it for you.

All of these sentences are too complex to squeeze them efficiently into the frame-and-slot straightjacket. So our method of representation had to evolve slowly, until today it is a type of second-order predicate calculus. That's a fancy way of saying the language of logic. The second-order qualifier means that sometimes we need to represent things whose interrelationship is unknown or to refer explicitly to earlier conversations. For instance, at one point in the film HAL says "I know I've never completely freed myself of the

suspicion that there are some extremely odd things about this mission. I'm sure you'll agree that there's some truth in what I say."

Lessons Learned Along the Way

We've learned many other things over the past dozen years, working on the CYC program. Three lessons in particular were painful but important to development of the program.

Originally we attached probabilistic weights -- that is, numerical certainty factors -- to each sentence we gave CYC. (For instance, HAL's statement, "Yes, that's a completely reliable figure," would have a very high certainty factor.) Including certainty factors had several bad consequences, however, and we eventually changed to a scheme in which all inputs are true by default. To decide whether to believe something, CYC gathers up all the pro and con arguments it can think of, examines them, and then reaches a conclusion.

Every representation is a trade-off between expressiveness (how easily you can say complicated things) and efficiency (how easily the machine can reason with what you've told it). English is very expressive but not very efficient. Most computer languages, such as Basic, C, and Fortran, are efficient but not very expressive. To get both qualities, we separated the epistemological problem (what should the system know?) from the heuristic problem (how can it effectively reason with what it knows?) and developed two separate languages, respectively EL and HL. Our knowledge enterers talk to CYC in the clean, expressive language (EL). Their input is then converted into the heuristic language (HL), which is efficient for dealing with many sorts of frequently recurring inference problems, such as reasoning about time, causality, containment, and so forth.

The third, and perhaps most important lesson we learned along the way was that it was foolhardy to try to maintain consistency in one huge flat CYC knowledge base. We eventually carved it up into hundreds of contexts or microtheories. Each one of those is consistent with itself, but there can be contradictions among them. Thus, in the context of working in an office it's socially unacceptable to jump up screaming whenever good things happen, while in the context of a football game it's socially unacceptable *not* to.

In the fictional context of Brain Stoker's *Dracula,* vampires exist; in the standard rational worldview context they don't. Other contexts carve out similar distinguishable eras in time, political or religious points of view, and so forth.

Applications of CYC

We've discussed the need to have something, like the CYC program, that can understand natural language; so it should come as no surprise that getting it to do this is a high-priority application task for us. Long before it can read all on its own, CYC will carry on semiautomated knowledge acquisition from texts, a sort of tutoring program in which it asks clarifying questions when it comes across something it's not sure about.

One potential use for CYC is to understand such structured information sources as spreadsheets and data bases, and then use that understanding to detect common-sense errors and inconsistencies in the data. For example, one column of a table might indicate a person's gender, and another might indicate that of his or her legal spouse. Without having to be specially programmed for the task, CYC would know that there's probably a mistake in the data if X and X's spouse have the same gender, if X's spouse lists a third person as his or her spouse, or if X is listed as X's spouse. This sort of data cleaning gets more interesting when combining information from several tables. (For example, according to one data base, X is suspected of committing a certain crime, whereas according to another data base X was in jail at the time.) This sort of information fusion or integration is very important, because much of the data we draw upon in our lives is gathered, formatted, and maintained by someone not under our direct control. Human beings -- and HAL, and CYC -- need to be able to assimilate information from numerous sources and interrelate it correctly. That task, in turn, requires common sense. Using n data bases and writing the transformation rules for their communications works fine when $n = 2$, but not so well when $n = 100$ or 1,000. Instead, the approach we use for CYC treats each column of each data base one at a time, writing rules that explain its meaning in terms CYC can understand. The entire CYC knowledge base then becomes, in effect, the semantic glue for implicitly joining all that information together -- just as you or I can draw on and combine information we acquire from many different sources. One of the flashiest early uses of CYC has been for information retrieval.

Imagine a library of captioned images and a user who comes along and types in a word, phrase, or sentence asking for images. Today's software would have to do Boolean searches based on keywords in the query and the captions, perhaps broadening the search a bit by looking up synonyms in a thesaurus or definitions in a dictionary. Or consider the

World Wide Web, whose keyword-based indexing is the only way to search through that immense information space. That's fine if you want to match "A bird in water" against "A duck in a pond," but it takes something like CYC to match "A happy person" against "A man watching his daughter take her first step." CYC uses common sense to do matches of that sort. Similarly, CYC matched the query "a strong and adventurous person" to a caption of "a man climbing a rock face." To do that, it used a few rules of the sort: "If people do something for recreation that puts them at risk of bodily harm, then they are adventurous."

Conclusions and Parting Thoughts

We haven't talked much about emotions and motivations. I started out by complaining that HAL was stupid, because he showed a distinct lack of common sense when he killed the crew rather than, for example, bringing them into his confidence about the secret orders for the mission. Most humans would agree that it's better to lie to people or risk confiding in them than to "solve" the problem by killing them.

At least HAL was rational in his murderous plans. One of the prevalent themes in science fiction has been that of the robot gone amok, generally driven by some strong emotion or craving for power. This is a reflection of our human fears, I think; it is the monster still lurking under our beds in the dark. HAL, CYC, and their ilk won't have emotions, because they are not useful for integrating information, making decisions based on that

information, and so on. A computer may pretend to have emotions, as part of what makes for a pleasing user interface, but it would be as foolish to consider such simulated emotions real as to think that the internal reasoning of a computer is carried out in English just because the input/output interface uses English (but see chapter 14).

We have described several applications of CYC, such as natural language understanding, checking and integrating information in spreadsheets and data bases, and finding relevant information in image libraries and on the World Wide Web. Notice that we were not talking about Herculean tasks like beating Garry Kasparov at chess by looking seventeen moves ahead or simulating and predicting any of a trillion problems days before they may occur, as HAL does continuously. We're just talking about inference problems that are only a couple of steps long. The key point here is that if you have the necessary common-sense knowledge -- such as "deadly pastimes suggest adventurousness" -- then you can make the inference quickly and easily; if you lack it, you can't solve the problem at all. Ever.

This is the essence of common sense -- that a little goes a long way. HAL had a veneer of intelligence, but in the end he was lacking in values and in common sense, which resulted in the needless death of almost the entire crew. We are on the road to building HAL's brain. But this time -- now that it's for real -- we aren't going to cripple it by skipping the

mass of simple stuff it needs to know.

Further Readings

Douglas B. Lenat and John Seely Brown. "Why AM and Eurisko Appear to Work." *J. Artificial Intelligence* 23 (1984): 269 -- 94. Summarizes the decades of his pioneering work in automated discovery that led Lenat to undertake the CYC project.

Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems.* Reading, Mass.: Addison-Wesley, 1990. A detailed discussion of the first five years of the CYC project.

Nicholas, J. Mars, ed., *Toward Very Large Knowledge Bases.* Amsterdam: 105 Press, 1995. A collection of technical papers, including "Steps to Sharing Knowledge" by Lenat.

Marvin A. Minsky. *The Society of Mind.* New York: Simon and Schuster, 1985. Minsky didn't completely see things Lenat's way when he wrote this seminal book, but he has since come closer to that point of view.

Tosh Munkakata, ed. *Communications of the ACM* 38, no. il (November 1995). A special issue devoted to the three efforts most closely related to building HAL's brain: CYC, EDR, and WordNet.

Williard V. Quine. *Ontological Relativity and Other Essays.* New York: Columbia University Press, 1969. A vital classic on "natural kinds."

Amos Tversky and Daniel Kahneman. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (1974): 1124 -- 31. Guaranteed to cause a paradigm shift in anyone who believes people are rational beings.