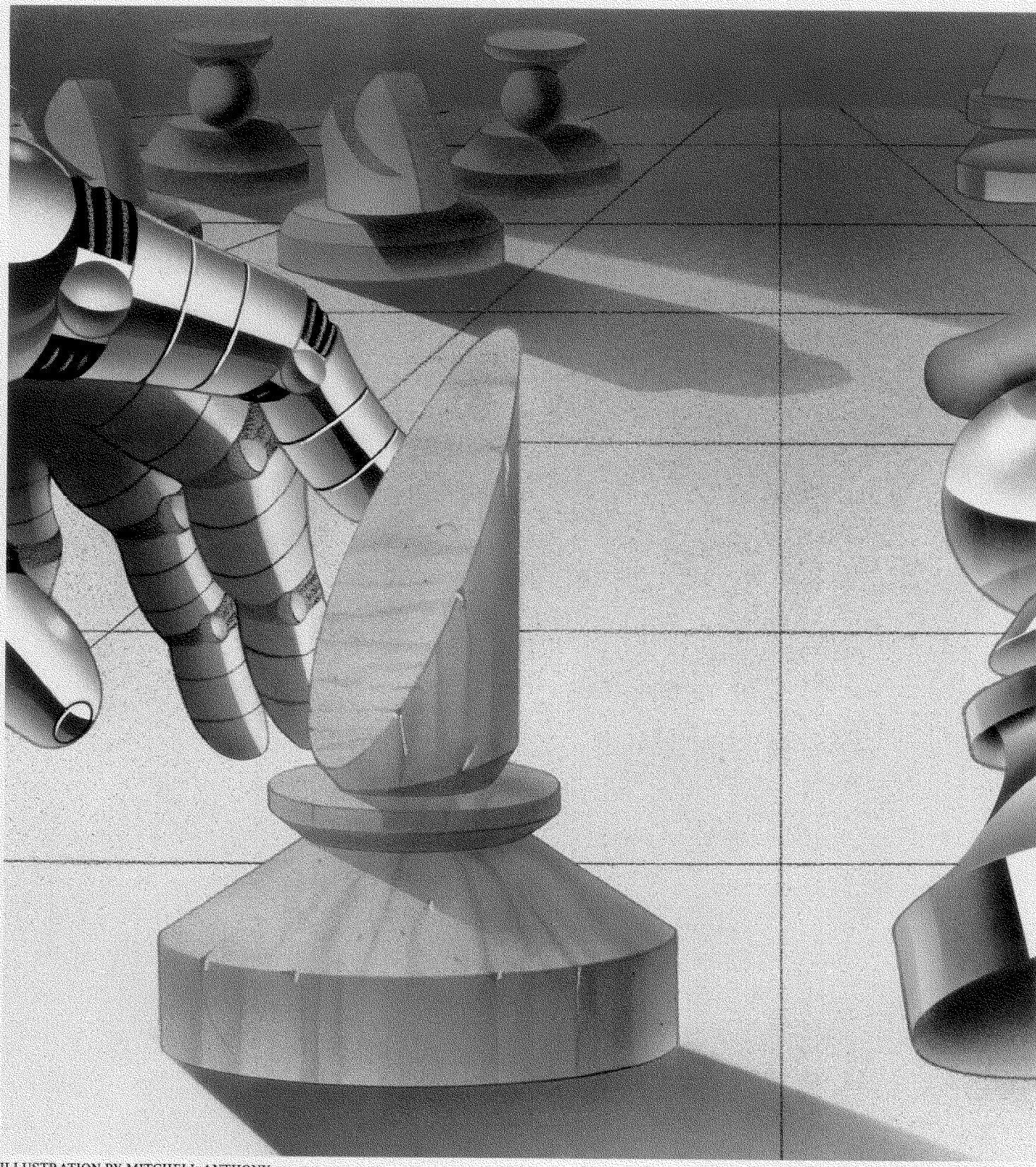
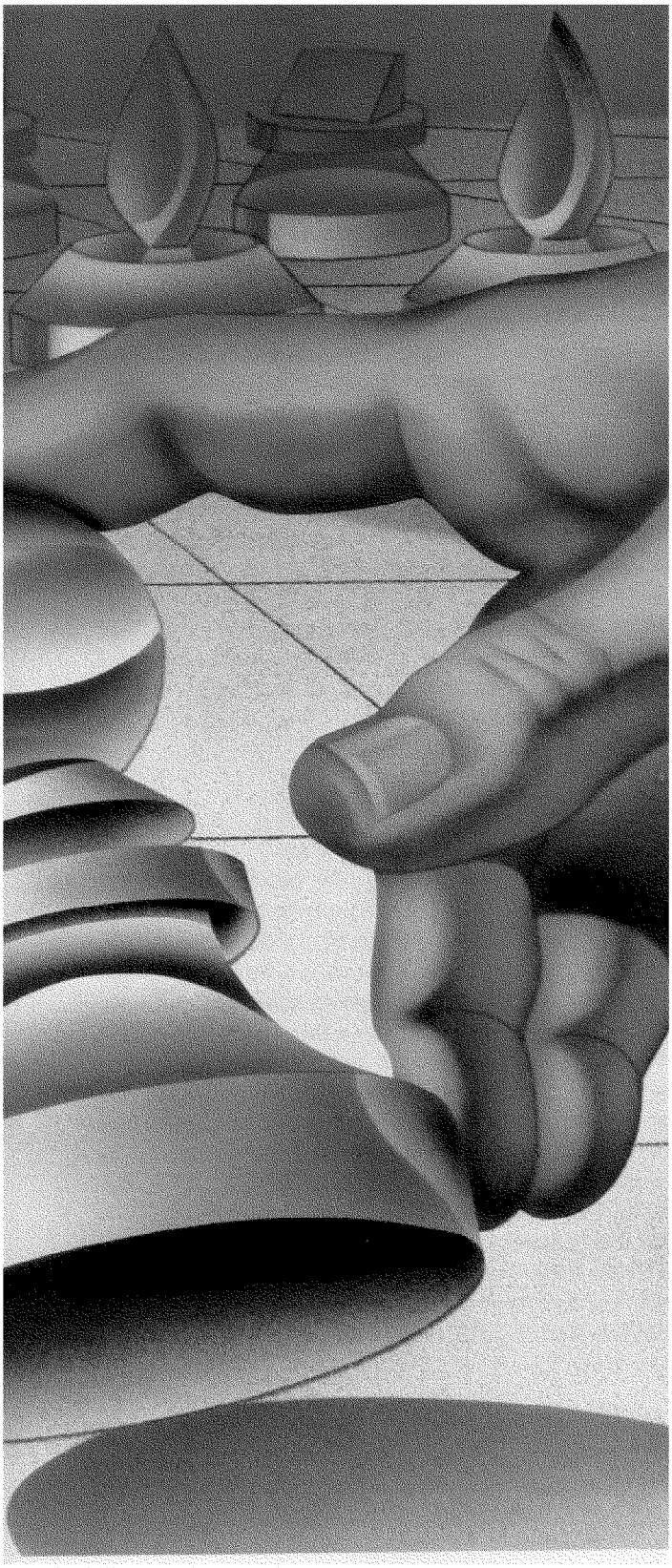


ILLUSTRATION BY MITCHELL ANTHONY

M. Mitchell Waldrop



A Question of Responsibility

In 1940, a 20-year-old science fiction fan from Brooklyn found that he was growing tired of stories that endlessly repeated the myths of Frankenstein and Faust: Robots were created and destroyed their creator; robots were created and destroyed their creator; robots were created and destroyed their creator—ad nauseum. So he began writing robot stories of his own. "[They were] robot stories of a new variety," he recalls. "Never, never was one of my robots to turn stupidly on his creator for no purpose but to demonstrate, for one more weary time, the crime and punishment of Faust. Nonsense! My robots were machines designed by engineers, not pseudo-men created by blasphemers. My robots reacted along the rational lines that existed in their 'brains' from the moment of construction."

In particular, he imagined that each robot's artificial brain would be imprinted with three engineering safeguards, three Laws of Robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law.

The young writer's name, of course, was Isaac Asimov (1964), and the robot stories he began writing that year have become classics of science fiction, the standards by which others are judged. Indeed, because of Asimov one almost never reads about robots turning mindlessly on their masters anymore.

But the legends of Frankenstein and Faust are subtle ones, and as the world knows too well, engineering rationality is not always the same thing as wisdom. This insight was

M Mitchell Waldrop is a reporter for *Science* Magazine, 1333 H Street N.W., Washington D C. 20005. His work covers the areas of physics, astronomy, space, and computers

This article is an excerpt from Mitch Waldrop's book entitled "*Man-Made Minds: The Promise of Artificial Intelligence*," to be published in March 1987, by Walker and Company, New York

never captured better than in another science fiction classic: the dark fantasy "With Folded Hands," published by Jack Williamson (1978).

The robots of this story are created by an idealistic young scientist whose world, the planet Wing IV, has just been devastated by a senseless war. Sickened by humankind's taste for viciousness and destruction, and thinking to save men from themselves, he programs his robots to follow an Asimovian Prime Directive—"To Serve and Obey, and Guard Men from Harm." He establishes factories where the robots can duplicate themselves in profusion. He sends them forth to bring order, rationality, and peace to humanity. And he succeeds all too well, as the citizens of his world and other worlds soon began to realize:

> . . . "Our function is to serve and obey, and guard men from harm," it cooed softly. "It is no longer necessary for men to care for themselves, because we exist to insure their safety and happiness."
>
> . . . "But it is unnecessary for human beings to open doors," the little black thing informed him suavely. "We exist to serve the Prime Directive, and our service includes every task."
>
> . . . "We are aiding the police department temporarily," it said. "But driving is really much too dangerous for human beings, under the Prime Directive. As soon as our service is complete, every car will have a humanoid driver. As soon as every human being is completely supervised, there will be no need for any police force whatsoever."

At last, the scientist realizes what he has done:

> "I found something worse than war and crime and want and death." His low rumbling voice held a savage bitterness. "Utter futility. Men sat with idle hands, because there was nothing left for them to do. . . . Perhaps they tried to play, but there was nothing left worth playing for. Most active sports were declared too dangerous for men, under the Prime Directive. Science was forbidden, because laboratories can manufacture danger. Scholarship was needless, because the humanoids could answer any question. Art had degenerated into a grim reflection of futility Purpose and hope were dead. No goal was left for existence . . No wonder men had tried to kill me!"

He attempts to destroy his robots by destroying the central electronic brain that controls them. (Williamson was writing in the days before distributed processing.) The robots stop him; this is clearly a violation of the Prime Directive, for how can they serve men if they themselves are hindered? He flees to another world, and tries again. And again. And again. Each time he is thwarted, as the robots continue to spread from planet to planet faster than he can run from them. And in the end, the robots devise a simple brain operation to cure his "hallucinations": "'We have learned to make all men happy, under the Prime Directive,' the mechanical promised him cheerfully. 'Our service is perfect at last.'"

Needless to say, "With Folded Hands" was widely considered a horror story.

What is a Robot?

A servant, it seems, can all too easily become the master—a phenomenon worth thinking about as we rush towards a new generation of intelligent machines. Just how will we use these machines? How much power and authority should they have? What kind of responsibilities should we give them? And who, if anyone, is going to control them?

Before we tackle those questions, however, we first ought to drop back a step and ask a different question: What exactly *is* a robot?

The question is more subtle than it sounds. For most of us the word *robot* conjures up an image of something like R2D2 or C3PO from the film *Star Wars*. But what about dishwashers and word-processing machines? Are *they* robots? The Robotics Industries Association uses a definition specially devised for factory robots: "A reprogrammable multifunctioning manipulator designed to move material, parts, tools or specialized devices through variable programmed motions for the performance of a variety of tasks." But that describes R2D2 and C3PO only in the crudest sense.

Actually, my favorite definition is "A surprisingly animate machine." But for our present purposes, the most useful definition is one that ignores the gadget's physical appearance entirely, and even its brainpower. It focuses instead on the role of the machine; unlike a lawn mower or a word processor, which requires continuous and direct supervision, a robot is an artificial *agent*—a machine that can take action without direct supervision.

Hidden Processing and Microrobots

Of course, if we take that definition literally, we're already surrounded by robots. Stoplights, for example. The automated teller machine at the bank. The coffeepot that starts up automatically at 7:00 AM. Admittedly, none of these "robots" is very smart. But microprocessors have already begun to appear in coffeepots, washing machines, automobiles, and microwave ovens. Given the rapidly decreasing price of microprocessors, and the increasing ease with which circuitry can be designed and built for special-purpose applications, there is every reason to expect that the devices around us will rapidly get smarter. Ultimately, in fact, we can expect that the engineers will add in little knowledge bases to their chips, so that their machines can talk, listen to orders, and respond to changing circumstances. And at that point we are not so far from what Pamela McCorduck (1979) has described as "a world saturated with intelligence," and what Allen Newell (1976) called the New Land of Fairie. For instance:

- Refrigerators that know how to thaw the chicken for dinner.
- Robotic cars that know how to stop on wet pavement, and how to drive down the highway while their passengers take a nap.

- Lampposts that know the way, so that no one need ever get lost.

Indeed, perhaps we should forget any lingering fears that all our descendants will become like teenage hackers hunched over a computer screen; our descendants may be much more like sorcerers, able to animate the objects around them with a word, and to command those objects to do their bidding.

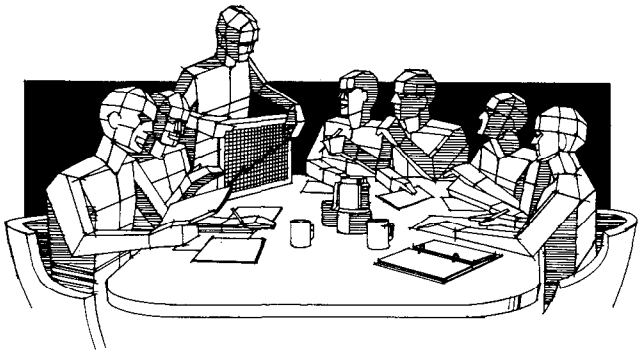
In all seriousness, many prognosticators think that "hidden" computing, as it is called, may very well be the most important way in which computers and AI will enter our lives. Alan Kay, who was one of the guiding spirits behind the development of personal computers when he was at Xerox Palo Alto Research Center in the early 1970's, points out that using a desktop computer to store household recipes, or to turn on the coffeepot in the morning, is roughly equivalent to an engineer in 1900 saying, "Electric motors are great! Every home should have one!" and then proceeding to rig an elaborate system of belts and pulleys to run everything in the house from one big motor in the attic. In fact, the average American home has some fifty electric motors, according to Kay. It's just that we never notice them, because they are tucked away out of sight in electric shavers, hairdryers, typewriters, fans, and washing machines. In short, hidden motorization. Apply that logic to the newest technology and you have computing and hidden AI (Kay 1985).

Humanoid Robots

Next, what about not-so-hidden robots: the mobile, humanoid machines that can walk, talk, see, handle things, and even think? We've read about them in science fiction stories. We've watched them in the *Star Wars* movies. So when are they going to be available at the local department store?

Actually, simple robots are available already, and have been since the early eighties. As of the mid-eighties, in fact, they are being offered by at least six different manufacturers at prices ranging from $2,000 to $8,000. The Heath Company of Benton Harbor, Michigan, is even offering its HERO I as a $1,200 build-it-yourself kit.

These first-generation machines are real, programmable robots with on-board microcomputers; they are not remote-control toys. However, it's important to realize that they have nowhere near the sophistication of R2D2 or C3PO. Certain advertisements to the contrary, present-day robots are not very good for serving canapes at a cocktail party, nor are they much help at doing the dishes afterwards. Essentially, they are personal computers on wheels, with ultrasonic ranging devices and perhaps a few heat and light sensors to help them avoid obstacles. Even then, they are hard-put to cross a room without bumping into anything, and it takes a major programming effort to get one to fetch a glass of water from the kitchen table. (And at that, if the glass is moved 6 inches from its previous position, the programming has to be done all over again.) By themselves, they are only



useful as teaching machines, to show people how robots work (Bell 1985).

On the other hand, they *are* significant as a possible first step toward more useful robots. For example, one of the first practical applications for personal robots might be as companions to the elderly and handicapped, which would ease their dependence on human help; it needn't take a very sophisticated robot to pick up a dropped handkerchief or to change television channels on voice command. This is especially true if the users are willing to wire their homes with radio locators in the walls; that way their robots could dispense with the need for high-powered image processing and instead use simple radio receivers to keep track of where they are in the house. In the same vein, mechanical-engineering students at Stanford recently developed a robot hand that will translate electronic text into sign language for the deaf (Shurkin 1985).

Another near-term use for mobile robots might be in industrial settings—not in factories as such but in cleaning long, unobstructed hallways, or in mowing the vast lawns of an industrial park. A prototype floor-washing robot has been developed at Carnegie-Mellon. Moreover, without too much adaptation such robots could also serve as security guards to patrol the grounds of prisons or sensitive military installations (Bell 1985).

Admittedly, as of the mid-1980's these markets are still a bit fuzzy. But the prospects are sufficiently promising to give the fledgling personal-robotics industry an undeniable momentum. In effect, the pioneers are betting that robotics will take the same route that personal computers did starting in the early 1970's. There, too, the early machines were little more than playthings for hobbyists. But they gave rise to second-generation PCs that *were* genuinely useful. The skyrocketing sales, in turn, gave the PC makers both the cash flow and the incentive to develop still more sophisticated computers. And on it went, with the results we see around us today. So who knows? Perhaps things will work out this way for the robotics industry, too.

One trend that gives substance to this vision is the support of advanced research efforts such as the Defense Advanced Research Project Agency (DARPA) Strategic Computing Program. One of DARPA's major goals in that

program is an autonomous vehicle capable of roving over rugged and hostile terrain at the speed of a running man. In a sense this vehicle represents the ultimate in AI: the union of vision, touch, reasoning, and motion in one entity, a machine that not only thinks but acts in a multitude of situations. Moreover, such mobile robots would find applications not just on the battlefield, but in any environment where flexibility, versatility and autonomy crucial. Take deep space, for example: If and when NASA sends another unmanned mission to Mars, perhaps in the 1990's, an important scientific goal will be to bring back samples of material from many different areas on the planet, which means that a rover of some kind will have to move around on the surface to pick them up. This will in turn require some pretty fancy footwork, since the images returned by the Viking landers in 1976 show that the Martian surface is littered with rocks, boulders, and craters. Unfortunately, however, there is no way to operate the rover by remote control from Earth. Radio signals take forty minutes to get from Earth to Mars and back, and at forty minutes per step the rover would never get anywhere. So the Mars rover will need to be intelligent enough to find its own way and pick up the right samples.

Machines of this caliber are clearly not consumer items in the sense that the personal robotics industry would like. On the other hand, advanced research-and-development does have a way of showing up in commercial products with surprising rapidity. So before long we may find R2D2 and C3PO in the local department store after all.

We probably won't be using these humanoid robots for everything, of course, at least not in the way that Asimov, Williamson, and other science fiction writers have suggested. It wouldn't make sense. Why build mobile robots for an automobile factory when it is so much easier to string them out along the assembly line and move the cars? Why build a humanoid robot to drive a truck when it would be far simpler to build a computer into the dashboard? Why not just make the truck itself into a robot—a kind of autopilot for the interstates?

In some situations, however, a mobile, humanoid robot might be ideal. Mining is an example, or construction work, or anyplace else where the environment is complex and unpredictable. Ironically, one of the most demanding environments, and the one where a mobile robot might be most useful, is the home. Consider what it would take to program a robot maid to vacuum under the dining-room table without vacuuming up the cat in the process; the terrain is at least as complex as Mars. Furthermore, the environment requires mobility for the simple reason that an automated valet or maid would have to share the house with people. Imagine how complicated it would be to build an intelligent clothes hamper that could take the dirty clothes, sort them by color and fabric, wash them, dry them, fold them, and put them away in the upstairs dresser; it would be a Rube Goldberg contraption with conveyor belts and special-purpose handling devices that took up half the house. Much more sensible would be a humanoid robot maid that could walk up and down the stairs and run an ordinary washing machine.[1]

Distributed Computing and Macrorobots

There is no reason that mobile robots couldn't also be part of a distributed system of interconnected intelligences. For example, it's easy to imagine a construction gang of robots swarming over the skeleton of a new building, each knowing a fragment of the architect's computer-designed blueprint, and each connected to the others by radio . . . grappling units that would carry a girder into place and hold it upright . . . an octopuslike thing goes to work with riveting guns on the end of each arm . . . a welding unit walks into position and strikes an arc from its snout . . .

But this vision leads to an intriguing question: Is this robot construction gang simply a team of individual agents? Or is it *one* individual—a "macrorobot" that just happens to have many eyes, many hands, and many minds?

The distinction may simply be semantic. But it may be more; certainly it's worth thinking about as large-scale distributed systems become more and more important. For example, look at computer-automated factories; if a robot is defined as a machine that can take autonomous action, then such a factory might very well be called a robot.

Or consider a modern, high-tech office building: The people who work there talk to each other through a digital, computer-operated telephone system. Their desktop computers communicate through a local area network. The building's heating-and-cooling system, its security system, its elevators and lights—all are controlled by computers for optimum efficiency. In effect, the building is a single giant machine: a robot.

In the 1990's, NASA hopes to launch a permanently manned space station occupied by at least six people at all times[2]. Eventually, it could carry as many as twenty people. The station will be expensive, of course—costing roughly eight billion dollars—so NASA wants to keep the crew busy repairing satellites, doing experiments, making astronomical observations, and carrying out other useful work. The agency does *not* want them to spend their time cleaning house and taking care of the life-support system. The station

[1]If mechanical valets and maids ever become common, life might take on a curiously Victorian flavor We might go back to building houses with "maid's rooms"—or at least with storage and parking areas for household robots We might go back to rooms elaborately decorated with bric-a-brac, because we will have mechanical servants to dust and clean it And fashions may swing back toward fussy, elaborate, Victorian-style clothing because we will have mechanical maids and valets to help us put it on and take it off.

[2]The schedule became quite hazy after the explosion of the space shuttle Challenger on January 28, 1986; as of mid-1986, however, NASA officials were still expressing confidence that the first components of the space station would be in orbit by the mid-1990's

will therefore be designed to take care of itself automatically, using AI and robotics technology wherever possible. In effect, it will be a giant robot (NASA 1985).

Back on Earth, meanwhile, personal computers, minicomputers, mainframes, and supercomputers are more and more often being linked together in transcontinental communications networks, in much the same way that telephones are already linked together for voice communications. A computer network can be thought of as a robot; indeed, such networks may turn out to be the most intelligent macrorobots of all.

The Multinet: Computers in the Global Village

A great deal has been written and said about the the coming of the "networked society." Local area networks are proliferating in business offices and on university campuses. And many people have imagined an interconnected web of such networks that would encompass virtually every computer and database on the planet, a concept that MIT computer scientist J. C. R. Licklider (1979) has dubbed the "multinet."

In time, says Licklider, the government itself will become an integral part of this multinet—using it to monitor regulated industries such as the stock market, to provide weather forecasts and take census data, to collect taxes and make social-security payments, to conduct polls, and even to hold elections. Meanwhile, the postal system will fade into a memory as people communicate more and more through telephone lines and electronic mail—all part of the multinet. Filing cabinets, microfilm repositories, document rooms, and even most libraries will be replaced by on-line information storage and retrieval through the multinet. More and more people will work at home, communicating with coworkers and clients through the multinet. People will shop through the multinet, using cable television and electronic funds transfer. People will even reserve their tables at restaurants and order ballet tickets through the multinet.

Clearly, Licklider's vision of a multinet connecting everyone and everything is still a long way from reality. If nothing else, there remains the question of who would run it: A private company? A public utility? A federal agency? We can only guess at when it will all come together.

On the other hand, we don't have to rely entirely on guesswork to know what this computer-saturated, networked society of the future will be like. It is already being created in microcosm on the nation's campuses. Most of the large universities in the United States, as well as many of the smaller colleges, are spending tens of millions of dollars each to lay cable in cross-campus trenches and to run wires into dormitories, offices, and classrooms. Three schools in particular—Carnegie-Mellon, MIT, and Brown University—have taken the lead in developing the software and networking standards for what is commonly known as the "scholar's work station": a personal computer with the power of a current-generation LISP machine and the three-thousand dollar price tag of a current-generation microcomputer. (Much of their funding has come from such manufacturers as IBM, Digital Equipment, and Apple, who will build the actual hardware. The machines began to appear on the market in late 1986) (Balkovich, Lerman, and Parmelee 1985).

What makes a network so attractive in an academic setting—or, for that matter, in a business setting—is not just that it allows people to send a lot of information bits from computer to computer but that it allows them to do things in ways they couldn't before. For example, individual users hooking into the network can begin to share powerful resources that no one user could justify by himself. Thus, local area networks give users access to departmental minicomputers, phototypesetters, high-volume laser printers, and massive on-line databases located in a central library. At a national level, meanwhile, the National Science Foundation has established a series of five national supercomputer centers (at Cornell, San Diego, Princeton, Carnegie-Mellon, and the University of Illinois) to give scientists around the country routine access to state-of-the-art number-crunching power. As part of that effort, the foundation is establishing a transcontinental computer network so that users can communicate with the supercomputers directly from their desktop terminals, without having to travel (Jennings 1985). In much the same way, the ARPAnet was established in the late sixties to link DARPA-supported computer scientists by high-speed communications. ARPAnet users are now able to send experimental designs for new computer chips to DARPA-supported silicon foundries, which then produce the chip and mail it back within a few weeks; as a result, hundreds of students and professors have been trying out a multitude of inventive ideas, and the art of integrated circuit design is being transformed. More recently, as DARPA has gotten underway with the Strategic Computing program, researchers have begun to tie in to testing grounds through ARPAnet so that they can experiment with remote robotics and advanced knowledge-based systems.

An equally important effect of networking is that it fosters a sense of community. Instead of just talking on the telephone, for example, widely scattered researchers on the ARPAnet have used its electronic mail facility to create a kind of ongoing debating society, with messages read and posted at any hour of the day or night. Among other topics, the debates have included some fiery exchanges about the pros and cons of the Fifth Generation efforts, the Strategic Computing program, and the Strategic Defense Initiative. The comments have often been quite critical of DARPA (IEEE Staff 1985). Indeed, some of the most insistent criticism of Strategic Computing and Strategic Defense has come from the Computer Professionals for Social Responsibility, a group that was organized through the ARPAnet (Augarten 1986).

(DARPA, to its credit, has never tried to censor any of this criticism.)

Meanwhile, ARPAnet's electronic communications have proved a boon for collaboration among widely separated individuals. This is particularly true when a message includes computer code for a joint programming project, the draft of a joint research paper, or graphics depicting a new chip design—all of which would be next to impossible to express in a voice message over the telephone. An author can write a book using the word-processing program on his office computer, and send the manuscript electronically to several dozen friends and reviewers at a keystroke. Then, within a few days, he can fetch it back at a keystroke with comments, suggestions, and corrections inserted—without anyone having to wait for the U.S. mail.

Managing the Network With AI

At first glance, none of this seems to have much to do with AI. But in fact, AI will play an increasingly important role in this area, especially when it comes to managing computer networks.

Fundamentally, it's a matter of human factors writ large. Whether or not we ever arrive at anything as all-encompassing as Licklider's multinet, we are clearly headed in the direction of a nationwide information infrastructure on the scale of the interstate highway system or the electrical power grid, and much more complicated than either one. Indeed, we can already see the beginnings of such a system in ARPAnet, in the National Science Foundation's supercomputer network, in the BITnet system that links campuses across the country, and in the commercial electronic mail services being introduced by such companies as MCI and Western Union (Jennings 1985).

At the same time, no matter how large or small a computer network may be, it has to be "transparent" in a certain sense, or people will find every excuse not use it. Compare the process to turning on a light: When I walk into a dark room, I don't have to know about voltages, transformers, or peak-load generators. I don't have to know where the power plant is, or whether it runs on coal, uranium, or solar energy. I just flip the switch and the light comes on. And by the same token, when I want to send an electronic message to a friend across the country, I don't want to have to tell my computer what kind of terminal she has, or what kind of data protocols her local network uses, or what the data rate should be. I just want to type in the message and have it go.

The solution, obviously, is to make the network itself intelligent. Make it come alive with software robots that keep things running smoothly in the background. Give it voice with helpful little expert systems that pop up whenever they are needed.

As an individual user of such a system, for example, I would have instant access to all the services of a private secretary, a personal librarian, a travel agent, and more. (The programs to do all this wouldn't necessarily have to be in my computer; if the communications were quick enough, the software could be residing in another computer elsewhere on the network and still *seem* to be running on my computer.) Thus, when I wanted to send that message, I would just type my friend's name and the text of the message on my office terminal; then, with a keystroke or a voice command, my secretarial expert system would look up her electronic address and route the message for me. In the same way, when I've told my computer to set me up for a trip to Pasadena on the twenty-fifth, my travel-agent system would automatically plan out an itinerary—remembering that I like to fly out of Washington in the early afternoon—and then automatically use the network to make reservations with the airlines, the hotels, and the car-rental agencies. A librarian expert system would likewise be available to guide me through the intricacies of the network itself—advising me what databases I might want to look at to ascertain the political situation in Chad, for example, or how I would go about accessing a supercomputer to do high-quality computer animation.

My secretarial system could also filter out electronic junk mail—which sounds trivial, except that in the age of electronic communications it may be a matter of mental survival. For example, say I'm a magazine journalist trying desperately to meet a deadline on major story. I don't want to have my computer beeping at me for every electronic-mail message that comes in. On the other hand, I *do* want it to alert me if one of the field bureaus is sending in new information that needs to go in my story. Yet I can't just set my electronic mail system to accept any message labeled *Urgent*, because there at least three companies sending me "urgent" messages about hot new prices on used IBM PCs, and one fellow in New Hampshire is desperately trying to convince me that Einstein's theory of relativity is wrong and that the moon is about to spiral into the Earth. So what I really want is a personal-receptionist program living inside the network, with enough natural-language ability to recognize contents of a message and enough common sense to know what I do and don't want to see.

There are other possibilities. In a corporate setting, for example, one can imagine an kind of office-manager program living in the system and keeping track of the electronic flow of work: Who has to read this item? Who has to coordinate on that project? Who has to sign off on this decision, and by when? Such a program would act to regulate the flow of electronic information through the organization, while preventing backlogs and keeping track of productivity. It might even function as a kind of social secretary: based on its expertise about policy and etiquette, it could remind people what kind of replies need to be sent to what messages. At the same time, the office-manager program could serve as a kind of gatekeeper for the network, providing people sure and convenient access to information when company policy says

they are authorized to have it, and keeping them out when they are not authorized to have it.

Looking beyond information networks, we can easily imagine how these same concepts of distributed, intelligent management might apply to other systems: an automated electric power grid, using telecommunications and computers to optimize the efficiency of transcontinental power lines (Gaushell 1985) . . . an automated air traffic-control system, juggling information from radars hundreds of miles apart . . . an automated city traffic-control system, coordinating lights and directing traffic around bottlenecks on a regional basis, so that things flow smoothly even in rush hour. Ultimately, in fact, we can imagine society itself being run by networks of interlinked intelligences—a single giant robot with each component an expert in own niche, and the whole greater than the parts.

So perhaps we should define the concept of *robot* yet again: Robots are not just machines that can walk around and manipulate things, or even machines that can take action on their own. In a less visible, and yet more all-pervasive sense than Asimov or Williamson ever realized, robots are machines that take care of us.

How Much Responsibility?

So we return to the questions we started with. Robots, in the broad sense that we have defined them, play the role of *agent*. Coupled with AI, moreover, they will be able to take on responsibility and authority in ways that no machines have ever done before. So perhaps it's worth asking before we get to that point just how much power and authority these intelligent machines *ought* to have—and just who, if anyone, will control them.

Obfuscation and Big Brother

There is ample reason to be concerned about such questions. Computers don't come equipped with any built-in ethical system analogous to Asimov's three laws of robotics. And for that very reason, they are well suited to act as tools and smoke screens for the powers that be—as bureaucrats learned almost as soon as computers were introduced. ("I'd like to help you, but I'm afraid the computer just isn't set up to do it that way . . .") AI, unfortunately, won't necessarily change things.

To illustrate some of the issues here, imagine an expert system that serves as a bank loan examiner. Each applicant sits at a terminal in the bank's offices answering questions about his or her financial status, while the computer verifies everything by automatic queries through the network to other banks and credit companies. (The applicant could also do this through a terminal at home.) Finally, the system makes a decision: Yes, the applicant qualifies, or no, the applicant doesn't qualify.

Now, this could be a very efficient and useful system. Certainly it would be consistent in applying the bank's loan policy to each applicant. On the other hand, people may not be willing to put up with that kind of treatment from a machine. It's humiliating enough to get the once-over from a human bank official, and a few banks might get their computer screens smashed in. But in some ways it's more disturbing to think that people *will* put up with such treatment—not because the expert system does anything sinister, necessarily, but because the situation obscures the fact that the machine's "decision" actually embodies a policy made by humans. Furthermore, the decision comes wrapped in the aura of "artificial intelligence." Rejected applicants may be too intimidated to protest. And even if they do, the (human) managers of the bank could all too easily brush them off with "Gee, I'd like to help you, but the computers these days are wiser than we are. . . ."

So even a seemingly straightforward application of AI can obscure the lines of responsibility—and it's naive to think that some people won't use it that way when they don't want outsiders asking uncomfortable questions.

Another troublesome aspect of this scenario is the free and easy way that our loan applicant feeds information about his life and activities in the electronic data network. How does he know what is going to be done with that data?

Maybe a lot will be done with it. "Today we're building systems that can collect a vast amount of information on an individual's daily transactions," says Fred Weingarten, manager of the Communications and Information Technologies program at the congressional Office of Technology Assessment. "Further, systems such as those employing knowledge-based technology can do more with that information than ever before, including the making of decisions that affect private lives" (Weingarten 1985). In that context it's worth noting that in a recent study, Weingarten's group found that at least thirty-five federal agencies were using or planned to use some form of electronic surveillance techniques—including computer-usage monitoring and the interception of electronic mail—and that current laws did not adequately regulate such electronic surveillance (OTA 1985).

There is nothing in the rules that says AI has to be benign, of course, and it is perfectly possible to imagine AI techniques in the service of Big Brother. Consider an advanced speech-recognition program that monitors every telephone call in the country for evidence of criminal activities, or a natural-language program that scans every computer bulletin board and reads every item of electronic mail. Such a system might catch a lot of obscene phone-callers and drug dealers. But how long before the definition of "criminal" starts sliding over into "subversive" and "disloyal?"

Actually, the process doesn't even have to be that dramatic. The mere possibility of surveillance is chilling. Langdon Winner, professor of politics and technology at the University of California at Santa Cruz, thinks that this may be the most insidious problem of all: "When I talk to people about this question, citing examples in which electronic

monitoring has been used as a tool for harassment and coercion," he says, "they often say that they don't need to worry about it because they're not doing anything anyone would possibly want to watch. In other words, it becomes a sign of virtue for them to say: 'Thank God, I'm not involved in anything that a computer would find at all interesting.' It's precisely that response that I find troubling."

Winner finds a metaphor for this situation in a fascinating design for a building, the Panopticon, created by the nineteenth-century philosopher Jeremy Bentham:

> The Panopticon was to be a circular building, several stories high, with a tower in the center. It could be used as a prison, hospital, school, or factory. A key feature of the design was that people in its rooms could not see each other, but the person in the center looking out to the periphery could gaze into every cell. Bentham saw this architecture as the ultimate means of social control. There would not even need to be a guard present in the tower at all times: all one had to do was to build in such a way that surveillance became an omnipresent possibility that would eliminate misbehavior and ensure compliance.
>
> It appears that we now may be building an electronic Panopticon, a system of seemingly benign electronic data-gathering that creates *de facto* conditions of universal surveillance.

It isn't just a threat to individual privacy. It is a threat to our public freedoms. "Unless we take steps to prevent it," Winner concludes, "we could see a society filled with all-seeing data banks used to monitor an increasingly pliant, passive populace no longer willing to risk activities that comprise civil liberty" (Winner 1985).

Tacit Assumptions in the Nuclear Age

The specter of Big Brother is never far from people's minds when they worry about the effect of computers on society. But in a sense, that kind of abuse is also the easiest to guard against. The bad guys wear black hats, so to speak; their actions are explicit and deliberate, even if covert. Thus, in principle, laws can be written to protect our electronic privacy, and a code of behavior enforced.

Much more widespread and much more insidious, however, is another kind of situation, in which the bad guys don't wear black hats. Instead, they are ordinary, well-intentioned people whose tacit assumptions and values cause us to drift in a direction we might not have taken by choice. For example, the designers of the first office word-processing machines, together with the managers who installed them, brought certain values and assumptions to the job that ended up turning many offices into electronic sweatshops. They didn't necessarily mean to do it, but that is what happened. Looking to the future, consider Licklider's multinet and the various software robots that will be required to run it; even with the best of intentions, the people who design those programs will shape what we do, what we see, what we know about, and how we interact with our fellow human beings.

PC Magazine editor Bill Machrone highlights this issue in an article about a demonstration he saw of a new electronic-mail package for PCs. The idea of the package was to render corporate communications more efficient by organizing all messages according to set categories: requests, denials, commands, counteroffers, and the like. Machrone's reaction: "communications software designed by Nazis."

"Consider that this is how people who don't like each other communicate," he explains. "It is not a foundation for trust and mutual cooperation. In fact, in most organizations, the use of techniques such as this is virtually guaranteed to put people on edge and at one another. Like anything else that mechanizes human interaction, such a system inevitably makes that interaction less human. . . . [What people in corporations need is] room to roam, freedom to grow, to express opinions, develop points of view, and interact."

So why would a group of well-meaning programmers produce such a package? Machrone asks. "Its creators are fervent believers in mechanized, highly structured communications. It works for them, and they are sure it'll work for you. They're converts to this way of thinking and they want you to be too" (Machrone 1986).

In short, a robot doesn't have to be an agent just of a person or an institution. Without anyone's ever fully realizing it, a robot can also be the agent of a value structure or a set of assumptions. And that is perhaps as it should be—so long as we understand what the values and assumptions really imply. Remember that Williamson's robots, who were pledged "to serve and obey, and guard men from harm," were conceived with the best of intentions.

Nowhere is this issue of tacit values and assumptions illustrated more starkly than in the realm of nuclear weapons and nuclear war. In particular, consider the "launch-on-warning" strategy.

There has long been a school of thought among strategic planners in the United States that our land-based nuclear missiles should be launched as soon as incoming warhead show up on radar. Afterward, goes the argument, it will be too late. Both the military and civilian communications network will very likely collapse as soon as the first hostile warheads fall. So even if the missiles themselves survive in their hardened silos, no one could fire them. We would have lost the war without firing a shot. Indeed, the very possibility invites a sneak attack. Thus, launch-on-warning (Ford 1985).

Now, there is an undeniable, if ruthless, logic to that argument. On the other hand, it's important to notice the tacit assumptions inherent in launch-on-warning: that a state of unrelenting hostility exists between us and the Other Side; that international fear and mistrust are the natural state of affairs; that They are sneaky devils just waiting for a chance to hit us on the blind side; that our ability to retaliate is more important than anything else, including the fate of the human race.

Perhaps what we need is . . . a theory and practice of machine ethics . . .

Are those the kind of assumptions we want to build our future on?

In any case, launch-on-warning has never been implemented—at least, not by the United States[3]—largely because the early-warning radars have shown a distressing tendency to give false alarms. Signals mistaken for hostile missiles in the past include flights of geese, the rising moon, and the signals from a training tape that was accidentally mounted on the wrong computer. For much the same reason, even the proponents of the idea have shied away from entrusting the launch decision to computers. Computers are too prone to break down and (currently) too rigid in their responses.

However, those tacit assumptions of hostility, fear, and mistrust are very strong. The pressure to move toward a launch-on-warning strategy is always with us, especially as our land-based nuclear-missile forces become increasingly vulnerable to cruise missiles launched from submarines off the coast, and to increasingly accurate intercontinental missiles routed over the pole. Indeed, the existence of new-generation computers and a new generation of machine intelligence may well tempt some future administration to take the step.

Meanwhile, we have President Reagan's concept of a space-based defense against ballistic missiles: "Star Wars." The system is intended strictly as a defensive measure; so if we assume for the sake of argument that it *will* be built and *will* be effective—two controversial assumptions—then the consequences of a false alarm would presumably not be as dire. The orbital lasers and such would just waste a lot of ammunition firing at empty space. An accidental activation of the system, on the other hand, might reveal its weaknesses to the other side and leave the country temporarily unprotected. So the indirect consequences could be very serious. In any case, the pressure to put an automatic trigger on such a defensive system are the same as they are for launch-on-warning: Once incoming missiles show up on the radar screen, there is just too little time to rely on having the president or anyone else make meaningful decisions.

In short, the relentless logic of the technology seems to be leading us ever closer to a point where the fate of the human race can no longer be entrusted to humans. Officially, of course, the release of our offensive nuclear arsenal can only be authorized by the president in his role as commander-in-chief. Moreover, the officials of the Strategic Defense Initiative Organization have always maintained that any future Star Wars defensive system will likewise be under the control of the president. And there is no reason to doubt them at their word.

However, even if we leave aside a host of practical questions—Will the president be able to communicate with the missile command centers after the bombs start falling? Will he even be *alive* at that point?—we can still ask what human control would really mean in a situation of nuclear crisis.

The fact is that most of the offensive and defensive systems would have to be automated in any case; there's simply no way for any one human to understand the myriad details involved. Indeed, from what little one can gather about U.S. plans for responding in a nuclear war, the president will essentially be presented with a short menu of preprepared options: Place *these* forces on alert, for example, or launch *that* contingent of missiles while holding *those* back, and so on.

Given the painfully short response times available in a nuclear attack, this menu approach is probably the only rational way to go. However, it's all too easy to imagine the scene: one aging politician—the president—sitting in front of a computer terminal that he may never have seen before, trying to read and digest a list of complex options while half-hysterical advisers are whispering contradictory advice in each ear. Meanwhile, he himself is probably becoming panicky with the knowledge that an irrevocable decision has to be made right *now*. Just how much careful consideration is he going to be able to give to his choice?

Very little, probably. In fact, one is left with an uncomfortable feeling that "controlling" nuclear forces in such a situation is virtually a contradiction in terms. The real choices will have already been made by the people who prepare the items on the menu—which means that it becomes critically important to know what their assumptions are. If they only present options that refer to this or that degree of belligerence, with no options that allow for backing away from hostilities, then the system has built into it the presumption that *there will be war*. Indeed, so far as an outside reporter can tell, that is exactly the case. There is precious little

[3]There are persistent rumors that the Soviet Union *has* implemented such a policy

consideration being given on either side to helping leaders calm the crisis.

Actually, this is one place where AI techniques might be very helpful. It's easy to imagine a calm, nonhysterical expert system advising some future president on various diplomatic options in a nuclear crisis, together with predictions of the other side's likely reaction. Never despairing, never forgetting things, never becoming obsessive about this or that course of action under the pressure of the moment—such a machine would be invaluable as the one cool head in the midst of chaos. Indeed, if consulted beforehand, it might help keep the crisis from developing in the first place. The question is, Is anyone planning to develop or deploy such a system, or any other decision-support system for helping our leaders deal with a crisis? At the moment no one seems to be. Does that mean that the presumption of war takes precedence?

From one point of view, of course, this debate over human control of nuclear weapons can be read as a rationale for turning the whole thing over to computers. In a nuclear age, goes the argument, a new generation of very intelligent computers incorporating AI could actually defend the country better, faster, and more rationally than humans. And who knows? Maybe they could. But here are some thoughts to ponder:

- Even if computers *are* better than humans at fighting a nuclear war, is it ethical to abdicate responsibility for the fate of human civilization to machines? For better or worse, that is our responsibility.
- Even if computers can do the job better than humans, computers are not human: An artificially intelligent machine may know *about* humans, in some sense, but it's hard to imagine that any machine in the foreseeable future will be able to appreciate the full implications of launching a nuclear-tipped missile. Even with AI, a computer just follows orders as best it can—the ultimate Good German. Perhaps a little hesitation and a chance for second thoughts ought to be kept in the system.
- Even if computers can be programmed to decide the future of the race in a cooler and more rational manner than humans are capable of, perhaps it's worth devoting a little effort along the way to making sure that neither humans nor machines are ever put in that position.

The Theory and Practice of Machine Ethics

In the last analysis, it seems unlikely that the question "How much responsibility?" is ever going to have a simple answer. After all, human beings have been arguing among themselves about responsibility, authority, and control for many thousands of years, with no final resolution in sight; I see no reason to think that the answers are going to be any easier just because we've suddenly introduced some new intelligences based on silicon instead of on flesh and blood.

However, one thing that is apparent from the above discussion is that intelligent machines *will* embody values, assumptions, and purposes, whether their programmers consciously intend them to or not. Thus, as computers and robots become more and more intelligent, it becomes imperative that we think carefully and explicitly about what those built-in values are. Perhaps what we need is, in fact, a theory and practice of machine ethics, in the spirit of Asimov's three laws of robotics.

Admittedly, a concept like "machine ethics" sounds hopelessly fuzzy and far-fetched—at first. But maybe it's not as far out of reach as it seems. Ethics, after all, is basically a matter of making choices based on concepts of right and wrong, duty and obligation. We can already see a glimmer of how computers might make such choices in Jaime Carbonell's model of subjective understanding (Carbonell 1979). Carbonell showed how programs could be governed by hierarchies of goals, which would guide their reasoning processes in certain directions and not in others. Thus, it might very well be possible to formulate a hierarchy of goals that embody ethical concepts; the hard part, as always, would lie in formulating precisely what those concepts ought to be.

Another hint comes from work on distributed processing: In the effort to teach individual computers how to cooperate among themselves without having some boss computer tell them what to do, AI researchers are beginning to discover the principles that govern when individuals will work together harmoniously, and when they will not.

In any case, the effort of understanding machine ethics may turn out to be invaluable not just as a matter of practicality, but for its own sake. The effort to endow computers with intelligence has led us to look deep within ourselves to understand what intelligence really is. In much the same way, the effort to construct ethical machines will inevitably lead us to look within ourselves and reexamine our own conceptions of right and wrong. Of course, this is hardly a new activity in human history; it has been the domain of religion and philosophy for millennia. But then, pondering the nature of intelligence is not a new activity, either. The difference in each case is that, for the first time, we are having to explain ourselves to an entity that knows *nothing* about us. A computer is the proverbial Martian. And for that very reason, it is like a mirror: The more we have to explain ourselves, the more we may come to understand ourselves.

The Shape of the Future

As the great Danish physicist Niels Bohr once said, "It's hard to predict—especially the future." So we can talk all we want about possibilities and trends, but no one really knows what the new generation of computers will bring. Even if we did know, people would still be arguing about precisely which effects were good and which were bad.

In the broadest terms, of course, our prospects are

framed by the visions of Asimov and Williamson. On the one hand, we have the bright vision of intelligent machines as our servants, advisers, tutors, companions, even our friends. According to this vision, computers represent a profoundly humane technology. Indeed, we can look forward to a new kind of partnership between mankind and machines, in which intelligent computers and robots will both relieve us of drudgery and tedium, while expanding our ability to understand and to cope with the world. The result will thus be a richer and more fulfilling life for all of us.

On the other hand, we have a darker vision of the future as an exercise in blank futility. Even if we leave aside our concerns about Big Brother, what will happen when all these artificially intelligent computers and robots leave us with nothing to do? What will be the point of living? Granted that human obsolescence is hardly an urgent problem. It will be a long, long time before computers can master politics, poetry, or any of the other things we really care about. But "a long time" is not forever; what happens when the computers *have* mastered politics and poetry? One can easily envision a future when the world is run quietly and efficiently by a set of exceedingly expert systems, in which machines produce goods, services, and wealth in abundance, and where everyone lives a life of luxury. It sounds idyllic—and utterly pointless.

But personally, I have to side with the optimists—for two reasons. The first stems from the simple observation that technology is made by people. Despite the strong impression that we are helpless in the face of, say, the spread of automobiles or the more mindless clerical applications of computers, the fact is that technology does not develop according to an immutable genetic code. It embodies human values and human choices. And to the extent that we can make those choices consciously instead of by blindly stumbling into them—admittedly not an easy thing to do—we do have control. Indeed, as we've just seen, the effort of developing intelligent computers may help us gain the insight to make those choices more wisely.

My second reason for being optimistic stems from a simple question: What does it mean to be "obsolete"?

A parable: Behold the lilies of the field. Considered purely as devices for converting light into energy, they've already been made obsolete by solar cells. But they go right on blooming, because photochemistry is not what lilies are about.

Another parable: Outside my window, the sparrows gather every day at a bird feeder. Considered purely as flying machines, they've long since been made obsolete by 747s. But they go right on eating and squabbling, because flying isn't what sparrows are about.

So what are human beings about? Perhaps our purpose is to serve God. Or perhaps we are here to serve each other. Perhaps we are here to create beauty in music, art, and literature, or to comprehend the universe, or to have fun. I won't presume to dictate the correct answer for anyone else. But I do suspect that in the long run, the most important implication of AI may be that it leads us to confront this question anew.

No, we don't know what this new world will be like. Perhaps it's just hard for those of us born to the work ethic to imagine what our hypothetical descendants will do with themselves. They may think of some very creative entertainments. Or they may create a new golden age of art and science. And they almost certainly will think of a whole new set of problems to worry about. But consider this: Some four thousand years stand between us and the author of Genesis. Technology has changed the world immeasurably in that time. And yet we can still read his words and feel their power. I somehow doubt that the advent of intelligent machines is going to change that very much. These machines may transform the world in ways we can only guess at—but we will still be human.

References

Asimov, I 1964 *The Rest of the Robots* New York: Doubleday

Augarten, A 1986. A Sense of Responsibility *PC Magazine* 5(5):99–104

Balkovich, E ; Lerman, S ; and Parmelee, R P. 1985. Computing in Higher Education: the Athena Experience. *Communications of the ACM* (November 1985):1214–1224

Bell, T E. 1985 Robots in the Home: Promises, Promises *IEEE Spectrum* 22(5):51–55.

Carbonell, J G 1979 *Subjective Understanding: Computer Models of Belief Systems*. Ann Arbor: University of Michigan Research Press.

Ford, D 1985. *The Button* New York: Simon and Schuster.

Gaushell, D J 1985. Automating the Power Grid. *IEEE Spectrum* 39–45

IEEE Staff 1985 Assessing the Technical Challenges: A Log of Electronic Messages In *Next Generation Computers*, ed E A. Torrero, 100–134. New York: IEEE Press

Jennings, D. M et al 1985. Computer Networking for Scientists *Science* 231: 943.

Kay, A 1985 Software's Second Act *Science* 85: 122–126

Licklider, J C R 1979 Computers and Government In *The Computer Age: A Twenty-Year Review*, eds M. L Dertouzos and J Moses, 91. Cambridge, Mass.: MIT Press.

Machrone, W 1986. Spare Me the Sermon *PC Magazine* 5(2):53–55.

McCorduck, P 1979 *Machines Who Think* San Francisco: W. H. Freeman

NASA Advanced Technology Advisory Committee 1985. Advancing Automation and Robotics Technology for the Space Station and the U S Economy, NASA Technical Memorandum 87566

Newell, A 1976 Viewpoints, No. 3. Pittsburgh, Penn : Carnegie-Mellon University Publications

Office of Technology Assessment 1985 Federal Government Information Technology: Electronic Surveillance and Civil Liberties, Technical Report, OTA CIT-23, Washington, D.C.

Shurkin, J 1985. Robot Hand Conveys Sign Language to Persons Both Deaf and Blind Stanford University News Service, July 30.

Weingarten, F W 1985 Assessing the Sociotechnical Challenges In *Next-Generation Computers*, ed E. A Torrero, 138 New York: IEEE Press.

Williamson, J 1978 With Folded Hands In *The Best of Jack Williamson*, 154–206. New York: Ballantine

Winner, L. 1985 Assessing the Sociotechnical Challenges In *Next-Generation Computers*, ed E A. Torrero, 138. New York: IEEE Press