Michael Anderson has a Ph.D. from the University of Connecticut and is an associate professor of computer science at the University of Hartford. He has a longtime interest in artificial intelligence.

Susan Leigh Anderson received her Ph.D. from the University of California, Los Angeles, and is professor emeritus of philosophy at the University of Connecticut, specializing in applied ethics. In 2005 she and Michael Anderson helped to organize the first international symposium on machine ethics. They have a book on machine ethics forthcoming from Cambridge University Press.

ROBOTICS

ROBOT BE GOOD

Autonomous machines will soon play a big role in our lives. It's time they learned how to behave ethically

By Michael Anderson and Susan Leigh Anderson

IN BRIEF

Robots that make autonomous decisions, such as those being designed to assist the elderly, may face ethical dilemmas even in seemingly everyday situations.

One way to ensure ethical behavior in robots that interact with humans is to program general ethical principles into them and let them use those principles to make decisions on a case-by-case basis.

Artificial-intelligence techniques can produce the principles themselves by abstracting them from specific cases of ethically acceptable behavior using logic.

The authors have followed this approach and for the first time programmed a robot to act based on an ethical principle.

Nao, manufactured by Aldebaran Robotics, is the first robot to have been programmed with an ethical principle.

Photograph by Darren Braun

IN THE CLASSIC NIGHTMARE SCENARIO OF DYSTOPIAN SCIENCE FICTION, machines become smart enough to challenge humans—and they have no moral qualms about harming, or even destroying, us. Today's robots, of course, are usually developed to help people. But it turns out that they face a host of ethical quandaries that push the boundaries of artificial intelligence, or AI, even in quite ordinary situations.

Imagine being a resident in an assisted-living facility—a setting where robots will probably become commonplace soon. It is almost 11 o'clock one morning, and you ask the robot assistant in the dayroom for the remote so you can turn on the TV and watch *The View*. But another resident also wants the remote because she wants to watch *The Price Is Right*. The robot decides to hand the remote to her. At first, you are upset. But the decision, the robot explains, was fair because you got to watch your favorite morning show the day before. This anecdote is an example of an ordinary act of ethical decision making, but for a machine, it is a surprisingly tough feat to pull off.

The scenario we just described is still theoretical, but we already have created a first demonstration of a robot able to make similar decisions. We have endowed our machine with an ethical principle that it uses to determine how often to remind a patient to take a medication. Our robot's programming so far is capable of choosing among only a few possible options, such as whether to keep reminding a patient to take medicine, and when to do so, or to accept the patient's decision not to take the medication. But to our knowledge, it is the first robot to rely on an ethical principle to determine its actions.

It would be extremely difficult, if not impossible, to anticipate every decision a robot might ever face and program it so that it will behave in the desired manner in each conceivable situation. On the other hand, preventing robots from taking absolutely any action that might raise ethical concerns could unnecessarily limit opportunities for robots to perform tasks that could greatly improve human lives. We believe that the solution is to design robots able to apply ethical principles to new and unanticipated situations—say, to determining who gets to read a new book, rather than who next gets control of the remote. This approach has the additional benefit of enabling robots to refer to those principles if asked to justify their behavior, which is essential if humans are to feel comfortable interacting with them. As a side benefit, efforts to design ethical robots could also lead to progress in the field of ethics itself, by forcing philosophers to examine real-life situations. As Tufts University philosopher Daniel C. Dennett recently put it, "AI makes philosophy honest."

I, ROBOT

AUTONOMOUS ROBOTS ARE LIKELY to soon be a part of our daily lives. Some airplanes are already capable of flying themselves, and self-driving cars are at the development stage. Even "smart homes," with computers controlling everything from lighting to the A/C, can be thought of as robots whose body is the entire home—just as HAL 9000, the computer in Stanley Kubrick's classic *2001: A Space Odyssey,* was the brains of a robot spaceship. And several companies have been developing robots that can assist the elderly with everyday tasks, either to supplement the staff of an assisted-living facility or to help the aged live at home by themselves. Although most of these robots do not have to make life-or-death decisions, for them to be welcome among us their actions should be perceived as fair, correct or simply kind. Their inventors, then, had better take the ethical ramifications of their programming into account.

If one agrees that embodying ethical principles in autonomous machines is key to their success in interacting with humans, then the first question becomes, Which principles should go in them? Fans of science-fiction literature may believe that Isaac Asimov already provided the answer some time ago, with his original 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.

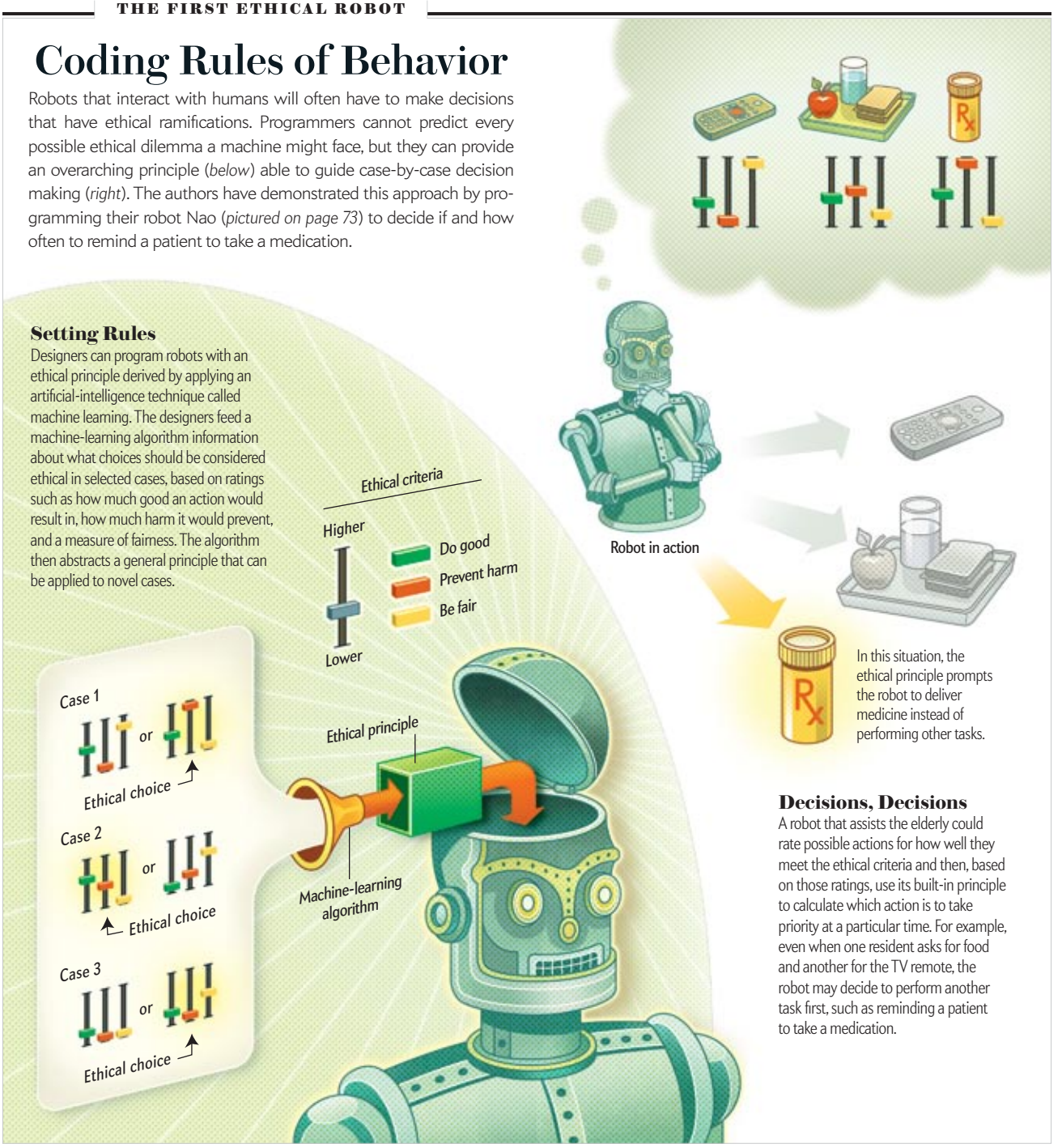
But some have discovered inconsistencies when thinking through the implications of these laws, which Asimov first articulated in a short story in 1942. And Asimov himself illustrated how unsuitable they were in his 1976 story *The Bicentennial Man,* in which human bullies order a robot to dismantle himself. The robot has to obey the bullies because of the Second Law, and he cannot defend himself without harming them, which would be a violation of the First Law.

If Asimov's laws are not acceptable, what is the alternative? Is an alternative even possible? Some people believe that implementing ethical behavior in machines is a hopeless proposition. Ethics, they say, is not the sort of thing that can be computed, and so it will be impossible to program it into a machine. Already in the 19th century, however, English philosophers Jeremy Bentham and John Stuart Mill maintained that ethical decision making is a matter of performing "moral arithmetic." Their doctrine of Hedonistic Act Utilitarianism, formulated in opposition to an ethic based on subjective intuition, holds that the right action is the one likely to result in the greatest "net pleasure," calculated by adding up units of pleasure and subtracting units of displeasure experienced by all those affected. Most ethicists doubt this theory accounts for all the dimensions of ethical concern. For example, it has difficulty capturing justice considerations and can lead to an individual being sacrificed in the interests of the majority. But at least it demonstrates that a plausible ethical theory is, in principle, computable.

Others doubt that machines will ever be capable of making ethical decisions, because machines lack emotions and so cannot appreciate the feelings of all those who might be affected by their actions. But humans are so prone to getting carried away by emotions that they often end up behaving unethically. This quality of ours, as well as our tendency to favor ourselves and those near and dear to us, often makes us less than ideal ethical decision makers. We think it is very possible that a properly

Coding Rules of Behavior

Robots that interact with humans will often have to make decisions that have ethical ramifications. Programmers cannot predict every possible ethical dilemma a machine might face, but they can provide an overarching principle (*below*) able to guide case-by-case decision making (*right*). The authors have demonstrated this approach by programming their robot Nao (*pictured on page 73*) to decide if and how often to remind a patient to take a medication.



trained machine could be designed to be impartial and to perceive human emotions and include them in its calculations, even if it does not have emotions itself.

LEARNING BY EXAMPLE

ASSUMING THAT IT IS POSSIBLE to give ethical rules to robots, whose ethical rules should those be? After all, no one has yet been able to put forward a general set of ethical principles for real-live humans that is accepted universally. But machines are typically created to function in specific, limited domains. Determining ethical parameters for behavior in such cases is a less daunting task than trying to devise universal rules of ethical and unethical behavior, which is what ethical theorists attempt to do. Moreover, when given the description of a particular situation within many contexts in which robots are likely to function, most ethicists would agree on what is ethically permissible and what is not. (In situations in which there is no such agreement, we believe that machines should not be allowed to make autonomous decisions at all.)

Researchers have proposed various different approaches to deriving rules for machine behavior, usually by means of AI techniques. For example, in 2005 Rafal Rzepka and Kenji Araki

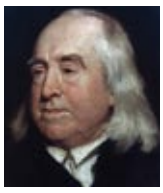
Illustration by Peter and Maria Hoey

NOTABLE MOMENTS IN MACHINE ETHICS

When Science Imitates Art

Long before ethicists, roboticists and AI experts became interested in the possible ethical ramifications of robots' behavior, science-fiction writers and film directors toyed with scenarios that were not always unrealistic. In recent years, however, machine ethics has become a bona fide field of research, in part drawing inspiration from the writings of 18th-century philosophers.

1495 Leonardo da Vinci designs one of the first humanoid robots



1780s Jeremy Bentham (*above*) and John Stuart Mill propose that ethics is computable

1921 Karel Čapek's play *R.U.R* first introduces the word "robot" and the concept of robot rebellion

1750 1800 1850

of Hokkaido University in Japan proposed "democracy-dependent algorithms" that would mine the Web for information on what people have in the past considered ethically acceptable actions and then use statistical analysis to produce answers to new questions. In 2006 Marcello Guarini of the University of Windsor in Ontario suggested that neural networks—algorithms inspired by the human brain that learn how to process information in an increasingly optimal way—could be "trained" using existing cases to recognize and select what are ethically acceptable decisions in similar cases.

In our view, reflected in our research, ethical decision making involves balancing several obligations, what ethicists refer to as prima facie duties (*prima facie* is Latin for "at first sight"). These are duties we should basically try to adhere to, each of which, however, can be overridden on occasion by one of the other duties. For example, people should generally try to keep their promises, but if they could prevent much harm by breaking a trivial promise, they should do so. When duties are in conflict with one another, ethical principles can determine which one should take precedence in each particular situation.

To obtain ethical principles that can be programmed into a robot, we employ an AI technique called machine learning. Our algorithm accesses a representative number of particular cases in which humans have determined certain decisions to be ethically correct. Then, using inductive logic, it abstracts an ethical principle. This "learning" stage takes place at the time of software design, and the resulting ethical principle is then encoded into the robot's programming.

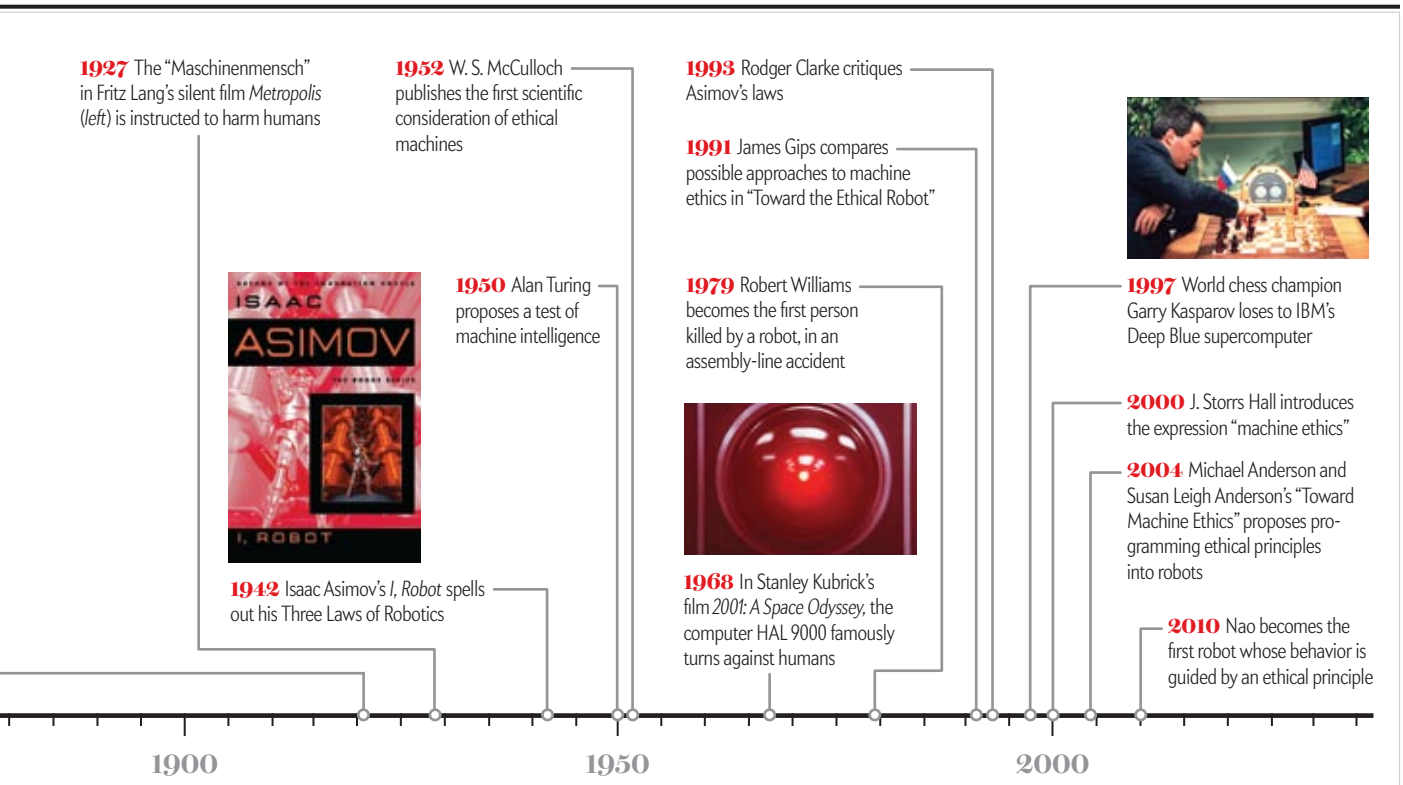
As a first test of our method, we considered a scenario in which the robot has to remind a patient to take a medication and notify an overseer when the patient does not comply. The robot must balance three duties: ensuring that the patient receives a possible benefit from taking the medication; preventing the harm that might result from not taking the medication; and respecting the autonomy of the patient (who is assumed to be adult and competent). Respecting patient autonomy, in particular, is considered a high priority in the field of medical ethics; this duty could be violated if the robot reminds the patient too often or notifies the overseer too soon for noncompliance.

After we fed it information about particular cases, the machine-learning algorithm produced the following ethical principle: a health care robot should challenge a patient's decision—violating the patient's autonomy—whenever doing otherwise would fail to prevent harm or severely violate the duty of promoting patient welfare.

AN IDEA WITH LEGS

WE THEN PROGRAMMED the principle into a humanoid robot, Nao, developed by the French company Aldebaran Robotics. Nao is capable of finding and walking toward a patient who needs to be reminded to take a medication, bringing the medication to the patient, interacting using natural language, and notifying an overseer by e-mail when necessary. The robot receives initial input from the overseer (who typically would be a physician), including: what time to take a medication, the maximum amount of harm that could occur if this medication is not taken, how long it would take for this maximum harm to occur, the maximum amount of expected good to be derived from taking this medication, and how long it would take for this benefit to be lost. From this input, the robot calculates its levels of duty satisfaction or violation for each of the three duties and takes different actions depending on how those levels change over

COURTESY OF NICCOLAI-TEKNOART/MUSEUM OF LEONARDO DA VINCI, FLORENCE (*humanoid robot*); THE GRANGER COLLECTION (*Bentham*); GETTY IMAGES (*R.U.R. poster*); UFA/THE KOBAL COLLECTION (*Metropolis robot*)



time. It issues a reminder when the levels of duty satisfaction and violation have reached the point where, according to its ethical principle, reminding is preferable to not reminding. The robot notifies the overseer only when it gets to the point that the patient could be harmed, or could lose considerable benefit, from not taking the medication.

A full-fledged version of an ethical elder care robot—EthEl for short—would need a more complicated ethical principle to guide its broader range of behaviors, but the general approach would be the same. During its rounds in the assisted-living facility, the robot would use that principle to determine when one duty takes precedence over another. Here is how a typical day might unfold.

Early in the morning EthEl stands in a corner, plugged into the wall socket. Once her batteries fill up, her duty of beneficence ("do good") overrides her duty to maintain herself, so she starts making her way around the room, visiting residents and asking if she can be helpful in some way—get a drink, take a message to another resident, and so on. As she receives tasks to perform, she assigns initial levels of satisfaction and violation to each duty involved in the task. One resident, in distress, asks her to seek a nurse. Ignoring the distress of a resident means violating the duty of nonmaleficence ("prevent harm"). That duty now overrides her duty of beneficence, so she seeks a nurse to inform her that a resident is in need of her services. Once this task is completed, her duty of beneficence takes over again, and she resumes her rounds.

When the clock strikes 10 A.M., it is time to remind a resident to take his medication. This task, satisfying the duty of beneficence, becomes paramount, so she seeks the resident out and gives him his medication. Later, the residents are absorbed in a TV show—be it *The View* or *The Price Is Right*. With no other duties pending and with her batteries running low, EthEl finds her duty to herself to be increasingly violated, so she returns to her charging corner.

The study of machine ethics is only at its beginnings. Though preliminary, our results give us hope that ethical principles discovered by a machine can be used to guide the behavior of robots, making their behavior toward humans more acceptable. Instilling ethical principles into robots is significant because if people were to suspect that intelligent robots could behave unethically, they could come to reject autonomous robots altogether. The future of AI itself could be at stake.

Interestingly, machine ethics could end up influencing the study of ethics. The "real world" perspective of AI research could get closer to capturing what counts as ethical behavior in people than does the abstract theorizing of academic ethicists. And properly trained machines might even behave more ethically than many human beings would, because they would be capable of making impartial decisions, something humans are not always very good at. Perhaps interacting with an ethical robot might someday even inspire us to behave more ethically ourselves. SA

MORE TO EXPLORE

IEEE Intelligent Systems. Special issue on machine ethics. July/August 2006.
A Robot in Every Home. Bill Gates in *Scientific American*, Vol. 296, No. 1, pages 58–65; January 2007.
Machine Ethics: Creating an Ethical Intelligent Agent. Michael Anderson and Susan Leigh Anderson in *AI Magazine*, Vol. 28, No. 4, pages 15–26; Winter 2007.
Moral Machines: Teaching Robots Right from Wrong. Colin Allen and Wendell Wallach. Oxford University Press, 2008.
War of the Machines. P. W. Singer in *Scientific American*, Vol. 303, No. 1, pages 56-63; July 2010.

COMMENT ON THIS ARTICLE www.ScientificAmerican.com/oct2010

COURTESY OF RANDOM HOUSE (*Asimov book cover*); MGM/THE KOBAL COLLECTION (*HAL 9000*); AFP/GETTY IMAGES (*Kasparov and Deep Blue*)