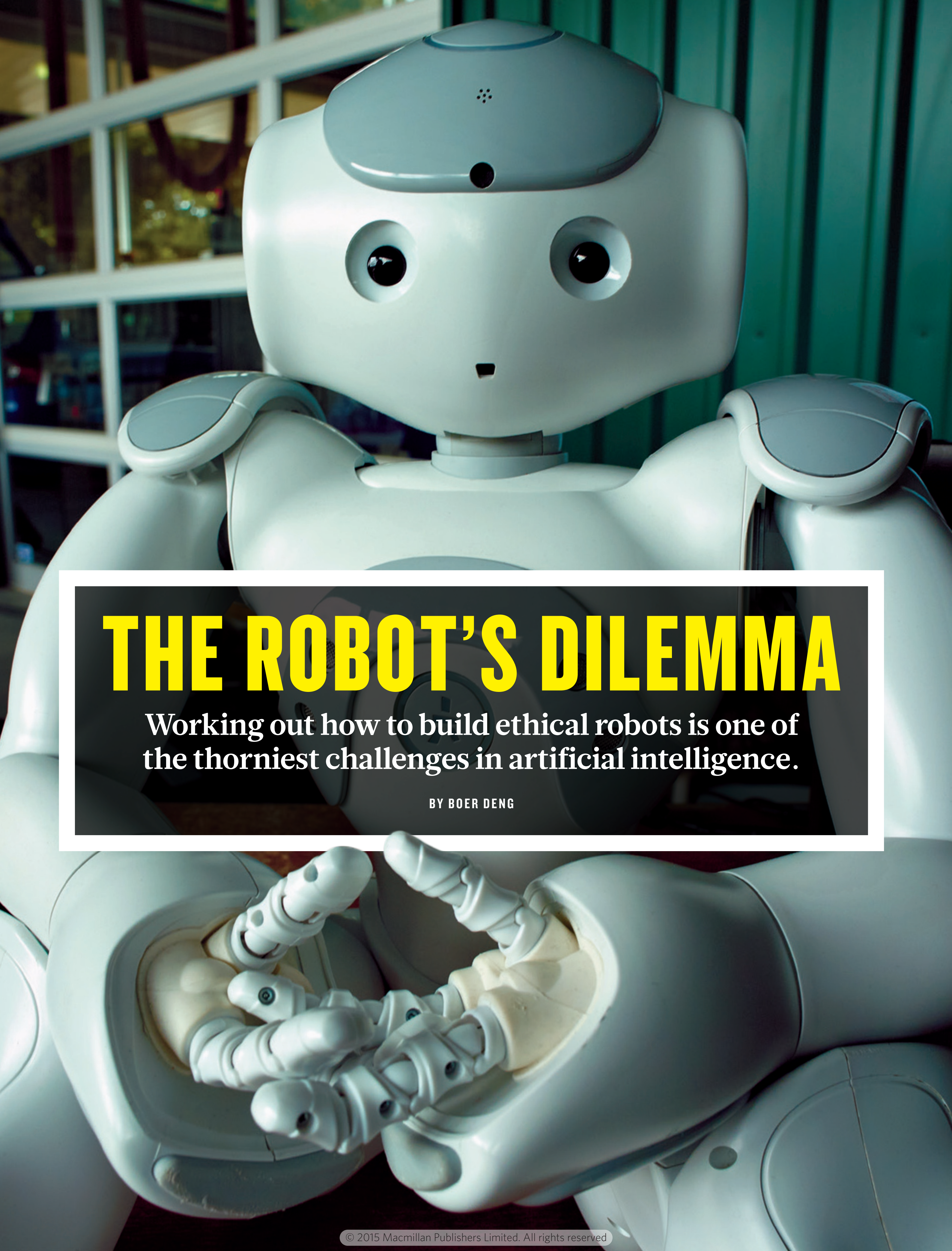

THE ROBOT'S DILEMMA

Working out how to build ethical robots is one of the thorniest challenges in artificial intelligence.

BY BOER DENG

In his 1942 short story 'Runaround', science-fiction writer Isaac Asimov introduced the Three Laws of Robotics — engineering safeguards and built-in ethical principles that he would go on to use in dozens of stories and novels. They were: 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; and 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Fittingly, 'Runaround' is set in 2015. Real-life roboticists are citing Asimov's laws a lot these days: their creations are becoming autonomous enough to need that kind of guidance. In May, a panel talk on driverless cars at the Brookings Institution, a think tank in Washington DC, turned into a discussion about how autonomous vehicles would behave in a crisis. What if a vehicle's efforts to save its own passengers by, say, slamming on the brakes risked a pile-up with the vehicles behind it? Or what if an autonomous car swerved to avoid a child, but risked hitting someone else nearby?

"We see more and more autonomous or automated systems in our daily life," said panel participant Karl-Josef Kuhn, an engineer with Siemens in Munich, Germany. But, he asked, how can researchers equip a robot to react when it is "making the decision between two bad choices"?

The pace of development is such that these difficulties will soon affect health-care robots, military drones and other autonomous devices capable of making decisions that could help or harm humans. Researchers are increasingly convinced that society's acceptance of such machines will depend on whether they can be programmed to act in ways that maximize safety, fit in with social norms and encourage trust. "We need some serious progress to figure out what's relevant for artificial intelligence to reason successfully in ethical situations," says Marcello Guarini, a philosopher at the University of Windsor in Canada.

Several projects are tackling this challenge, including initiatives funded by the US Office of Naval Research and the UK government's engineering-funding council. They must address tough scientific questions, such as what kind of intelligence, and how much, is needed for ethical decision-making, and how that can be translated into instructions for a machine. Computer scientists, roboticists, ethicists and philosophers are all pitching in.

"If you had asked me five years ago whether we could make ethical robots, I would have said no," says Alan Winfield, a roboticist at the Bristol Robotics Laboratory, UK. "Now I don't think it's such a crazy idea."

LEARNING MACHINES

In one frequently cited experiment, a commercial toy robot called Nao was programmed to remind people to take medicine.

"On the face of it, this sounds simple," says Susan Leigh Anderson, a philosopher at the University of Connecticut in Stamford who did the work with her husband, computer scientist Michael Anderson of the University of Hartford in Connecticut. "But even in this kind of limited task, there are nontrivial ethics questions involved." For example, how should Nao proceed if a patient refuses her medication? Allowing her to skip a dose could cause harm. But insisting that she take it would impinge on her autonomy.

To teach Nao to navigate such quandaries, the Andersons gave it examples of cases in which bioethicists had resolved conflicts involving autonomy, harm and benefit to a patient. Learning algorithms then sorted through the cases until they found patterns that could guide the robot in new situations[1].

PETER ADAMS

The fully programmable Nao robot has been used to experiment with machine ethics.

With this kind of 'machine learning', a robot can extract useful knowledge even from ambiguous inputs (see go.nature.com/2r7nav). The approach would, in theory, help the robot to get better at ethical decision-making as it encounters more situations. But many fear that the advantages come at a price. The principles that emerge are not written into the computer code, so "you have no way of knowing why a program could come up with a particular rule telling it something is ethically 'correct' or not", says Jerry Kaplan, who teaches artificial intelligence and ethics at Stanford University in California.

Getting around this problem calls for a different tactic, many engineers say; most are attempting it by creating programs with explicitly formulated rules, rather than asking a robot to derive its own. Last year, Winfield published the results[2] of an experiment that asked: what is the simplest set of rules that would allow a machine to rescue someone in danger of falling into a hole? Most obviously, Winfield realized, the robot needed the ability to sense its surroundings — to recognize the position of the hole and the person, as well as its own position relative to both. But the robot also needed rules allowing it to anticipate the possible effects of its own actions.

"We need some serious progress to figure out what's relevant for artificial intelligence to reason successfully in ethical situations."

Winfield's experiment used hockey-puck-sized robots moving on a surface. He designated some of them 'H-robots' to represent humans, and one — representing the ethical machine — the 'A-robot', named after Asimov. Winfield programmed the A-robot with a rule analogous to Asimov's first law: if it perceived an H-robot in danger of falling into a hole, it must move into the H-robot's path to save it.

Winfield put the robots through dozens of test runs, and found that the A-robot saved its charge each time. But then, to see what the allow-no-harm rule could accomplish in the face of a moral dilemma, he presented the A-robot with two H-robots wandering into danger simultaneously. Now how would it behave?

The results suggested that even a minimally ethical robot could be useful, says Winfield: the A-robot frequently managed to save one 'human', usually by moving first to the one that was slightly closer to it. Sometimes, by moving fast, it even managed to save both. But the experiment also showed the limits of minimalism. In almost half of the trials, the A-robot went into a helpless dither and let both 'humans' perish. To fix that would require extra rules about how to make such choices. If one H-robot were an adult and another were a child, for example, which should the A-robot save first? On matters of judgement like these, not even humans always agree. And often, as Kaplan points out, "we don't know how to codify what the explicit rules should be, and they are necessarily incomplete".

Advocates argue that the rule-based approach has one major virtue: it is always clear why the machine makes the choice that it does, because its designers set the rules. That is a crucial concern for the US military, for which autonomous systems are a key strategic goal. Whether machines assist soldiers or carry out potentially lethal missions, "the last thing you want is to send an autonomous robot on a military mission and have it work out what ethical rules it should follow in the middle of things", says Ronald Arkin, who works on robot ethics software at Georgia Institute of Technology in Atlanta. If a robot had the choice of saving a soldier or going after an enemy combatant, it would be important to know in advance what it would do.

With support from the US defence department, Arkin is designing a program to ensure that a military robot would operate according to international laws of engagement. A set of algorithms called an ethical governor computes whether an action such as shooting a missile

➲ NATURE.COM
For a podcast on robot ethics, see:
go.nature.com/wvkakj

JIJI PRESS/AFP/GETTY

'Robear' is designed to help to care for ill or elderly people.

is permissible, and allows it to proceed only if the answer is 'yes'.

In a virtual test of the ethical governor, a simulation of an unmanned autonomous vehicle was given a mission to strike enemy targets — but was not allowed to do so if there were buildings with civilians nearby. Given scenarios that varied the location of the vehicle relative to an attack zone and civilian complexes such as hospitals and residential buildings, the algorithms decided when it would be permissible for the autonomous vehicle to accomplish its mission[3].

"Logic is how we reason and come up with our ethical choices."

Autonomous, militarized robots strike many people as dangerous — and there have been innumerable debates about whether they should be allowed. But Arkin argues that such machines could be better than human soldiers in some situations, if they are programmed never to break rules of combat that humans might flout.

Computer scientists working on rigorously programmed machine ethics today favour code that uses logical statements, such as 'If a statement is true, move forward; if it is false, do not move.' Logic is the ideal choice for encoding machine ethics, argues Luís Moniz Pereira, a computer scientist at the Nova Laboratory for Computer Science and Informatics in Lisbon. "Logic is how we reason and come up with our ethical choices," he says.

Crafting instructions capable of the logical steps that go into making ethical decisions is a challenge. For example, Pereira notes, the logical languages used by computer programs have trouble coming to conclusions about hypothetical scenarios, but such counterfactuals are crucial in resolving certain ethical dilemmas.

One of these is illustrated by the trolley problem, in which you imagine a runaway railway trolley is about to kill five innocent people who are on the tracks. You can save them only if you pull a lever that diverts the train onto another track, where it will hit and kill an innocent bystander. What do you do? In another set-up, the only way to stop the trolley is to push the bystander onto the tracks.

People often answer that it is all right to stop the trolley by hitting the lever, but viscerally reject the idea of pushing the bystander. The basic intuition, known to philosophers as the doctrine of double effect, is that deliberately inflicting harm is wrong, even if it leads to good. However, inflicting harm might be acceptable if it is not deliberate, but simply a consequence of doing good — as when the bystander simply happens to be on the tracks.

This is a very difficult line of analysis for a decision-making program. To begin with, the program must be able to see two different futures: one in which a trolley kills five people, and another in which it hits one. The program must then ask whether the action required to save the five is impermissible because it causes harm, or permissible because the harm is only a side effect of causing good.

To find out, the program must be able to tell what would happen if it chose not to push the bystander or pull the lever — to account for counterfactuals. "It would be as if a program was constantly debugging itself," says Pereira — "finding where in a line of code something could be changed, and predicting what the outcome of the change would be." Pereira and Ari Saptawijaya, a computer scientist at the University of Indonesia in Depok, have written a logic program[4] that can successfully make a decision based on the doctrine of double effect, as well as the more sophisticated doctrine of triple effect, which takes into account whether the harm caused is the intended result of the action, or simply necessary to it.

HUMANS, MORALS, MACHINES

How ethical robots are built could have major consequences for the future of robotics, researchers say. Michael Fisher, a computer scientist at the University of Liverpool, UK, thinks that rule-bound systems could be reassuring to the public. "People are going to be scared of robots if they're not sure what it's doing," he says. "But if we can analyse and prove the reasons for their actions, we are more likely to surmount that trust issue." He is working with Winfield and others on a government-funded project to verify that the outcomes of ethical machine programs are always knowable.

By contrast, the machine-learning approach promises robots that can learn from experience, which could ultimately make them more flexible and useful than their more rigidly programmed counterparts. Many roboticists say that the best way forward will be a combination of approaches. "It's a bit like psychotherapy," says Pereira. "You probably don't just use one theory." The challenge — still unresolved — is to combine the approaches in a workable way.

These issues may very soon come up in the fast-moving field of autonomous transport. Already, Google's driverless cars are zipping across parts of California (see *Nature* **518,** 20–23; 2015). In May, autonomous trucks from German car-maker Daimler began driving themselves across the Nevada desert. Engineers are thinking hard about how to program cars to both obey rules and adapt to situations on the road. "Up until now we've been trying to do things with robots that humans are bad at," such as maintaining attention on long drives or being quick on the brakes when the unexpected occurs, says Bernhard Weidemann, a spokesperson for Daimler in Stuttgart. "Going forward, we will have to try to program things that come more naturally to humans, but not to machines." ■

Boer Deng *is a news intern for* Nature *in Washington DC.*

1. Anderson, M. & Anderson, S. L. *AI Magazine* **28,** 15–26 (2007).
2. Winfield, A. F. T., Blum, C. & Liu, W. in *Advances in Autonomous Robotics Systems* 85–96 (Springer, 2014).
3. Arkin, R. C., Ulam, P. & Duncan, B. *An Ethical Governor for Constraining Lethal Action in an Autonomous System* Technical Report GIT-GVU-09-02 (2009).
4. Pereira, L. M. & Saptawijaya, A. in *Logic, Argumentation and Reasoning* (eds Urbaniak, R. & Payette, G.) (Springer, in the press); available at http://go.nature.com/3xlske