Can robots have a conscience?

Moral Machines: Teaching Robots Right from Wrong
by Wendell Wallach and Colin Allen
Oxford University Press: 2008.
288 pp. $29.95

Artificial moral agents do not exist but are easily imagined: driverless trains that choose to turn away from a track on which five engineers are working, or autonomous armed drone aircraft that can distinguish between legitimate and unsanctioned targets. In *Moral Machines*, ethicist Wendell Wallach and philosopher Colin Allen pose three questions: "Does the world need artificial moral agents? Do people want computers making moral decisions? And how should engineers and philosophers proceed to design such agents?"

In contrast to Hollywood's fantasies of intelligent but malignant doom machines and researchers' speculations about machine-based transcendence, *Moral Machines* is modest, accurate and informative. The authors provide clear accounts of the basic ethical and philosophical issues, presupposing no technical background. To ask whether non-conscious machines can be real moral agents, they focus on 'functional morality': "Moral agents monitor and regulate their behaviour in light of the harms their actions may cause or the duties they may neglect." The book covers a wide range of approaches, organizing current research into top-down application of traditional ethical theories, bottom-up evolutionary or learning strategies, and work on implementing emotions in computers.

As no robot is close to realizing functional morality, the book's discussion may seem premature. The authors argue for an early start. But there is a risk that such early framing of issues can become powerfully salient; witness the influence of Isaac Asimov's Three Laws of Robotics, an example of a hierarchical, rule-based morality for robots. The authors' stated goal is to frame discussion in a way that guides the engineering task of designing artificial moral agents. But there are three main problems about the way they frame the topic.

First, the authors stretch their case, both in terms of the need for moral evaluations and the systems they analyse. In answer to the question of whether the world needs these artificial moral agents, they use the example of the power blackout in the northeastern United States in 2003, in which "software agents and control systems at … power plants activated shutdown procedures, leaving almost no time for human intervention". Consequently, they argue that "there is a need for autonomous systems to weigh risks against values". The example is surprising: large-scale networked electricity infrastructure is a long way from the robot vacuum cleaners and nurses that might need a functional morality. Although the book focuses mainly on physical robots and the software simulations used to design them, the authors deliberately increase the scope of their topic to include software agents, or bots. But they do not discuss the related ethical issues, such as privacy, raised by programs such as "data-mining bots that roam the Web". Asking people whether they would want computers making moral decisions may yield different answers than if you asked them the same about physical robots, in which issues of control and responsibility are simpler because robots are local.

Second, the focus on autonomy in the authors' definition of robot, by stressing "independence from direct human oversight", forecloses the important alternative option of remotely controlled robots. An example of this type of technology is robotic surgery, which promises great benefits but raises far fewer philosophical and ethical issues than other applications. Similarly, remotely piloted Predator drone aircraft raise no new ethical issues. In these cases, one could simply insist on and develop better technologies for remote human oversight and control. For example, following the commuter-train wreck in Los Angeles, California, in September 2008, it was proposed that surveillance cameras in train cabs — instead of extra moral education for the drivers — could alleviate the failures of human autonomy; in this case, the driver reportedly being distracted by texting. Thus, autonomy should not be stressed in the definition of the target technologies: "Should a good autonomous agent alert a human overseer if it cannot take action without causing some harm to humans? (If so, it is sufficiently autonomous?)"

Third, the book advocates implementing human morality, as it is the only one we know about. This choice is not so obvious. For the foreseeable future, robots will be inferior to humans in their moral decision-making capacity. So Asimov's hierarchical morality has appeal, compared with a human-based morality that stresses the equality of all moral agents. The morality of dogs would be a good alternative. Similar to dogs, robots will vary in their ability to make morally appropriate judgements. My family has had well-trained dogs that we trusted off-leash, others that could be trusted only when temptations such as squirrels or cats were absent, and some that needed leashes and even muzzles. All were pack animals, focused on a leader and unlike all our cats in this respect. A worthy goal for near-term robot ethics would be machines that we could classify in this way, so that we could give each the level of trust and control that lets them serve us well.

Moral Machines looks well in advance at robot ethics, but the jury is out on whether this book will set the agenda or if it is too premature to be influential. ■

A. WYANT/GETTY IMAGES

Robots could be given varying levels of morality depending on their role.

Peter Danielson is Mary and Maurice Young Professor of Applied Ethics at the Centre for Applied Ethics, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada. He is author of *Artificial Morality* and editor of *Modeling Rationality, Morality, and Evolution*.
e-mail: pad@ethics.ubc.ca