Moral Machines by Wendell Wallach and Colin Allen

In the Affective Computing Laboratory at the Massachusetts Institute of Technology (MIT), scientists are designing computers that can read human emotions. Financial institutions have implemented worldwide computer networks that evaluate and approve or reject millions of transactions every minute. Roboticists in Japan, Europe, and the United States are developing service robots to care for the elderly and disabled. Japanese scientists are also working to make androids appear indistinguishable from humans. The government of South Korea has announced its goal to put a robot in every home by the year 2020. It is also developing weapons-carrying robots in conjunction with Samsung to help guard its border with North Korea. Meanwhile, human activity is being facilitated, monitored, and analyzed by computer chips in every conceivable device, from automobiles to garbage cans, and by software “bots” in every conceivable virtual environment, from web surfing to online shopping. The data collected by these (ro)bots—a term we’ll use to encompass both physical robots and software agents—is being used for commercial, governmental, and medical purposes.

All of these developments are converging on the creation of (ro)bots whose independence from direct human oversight, and whose potential impact on human well-being, are the stuff of science fiction. Isaac Asimov, more than fifty years ago, foresaw the need for ethical rules to guide the behavior of robots. His Three Laws of Robotics are what people think of first when they think of machine morality:

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 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.

Asimov, however, was writing stories. He was not confronting the challenge that faces today’s engineers: to ensure that the systems they build are beneficial to humanity and don’t cause harm to people.

Within the next few years, we predict there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight. Already, in October 2007, a semiautonomous robotic cannon deployed by the South African army malfunctioned, killing 9 soldiers and wounding 14 others—although early reports conflicted about whether it was a software or hardware malfunction. The potential for an even bigger disaster will increase as such machines become more fully autonomous. Even if the coming calamity does not kill as many people as the terrorist acts of 9/11, it will provoke a comparably broad range of political responses. These responses will range from calls for more to be spent on improving the technology, to calls for an outright ban on the technology (if not an outright “war against robots”).

A concern for safety and societal benefits has always been at the forefront of engineering. But today’s systems are approaching a level of complexity that, we argue, requires the systems themselves to make moral decisions—to be programmed with “ethical subroutines,” to borrow a
We don’t know exactly how a catastrophic incident will unfold, but the following tale may give some idea:

Monday, July 23, 2012, starts like any ordinary day. A little on the warm side in much of the United States perhaps, with peak electricity demand expected to be high, but not at a record level. Energy costs are rising in the United States, and speculators have been driving up the price of futures, as well as the spot price of oil, which stands close to $300 per barrel. Some slightly unusual automated trading activity in the energy derivatives markets over past weeks has caught the eye of the federal Securities and Exchange Commission (SEC), but the banks have assured the regulators that their programs are operating within normal parameters.

At 10:15 a.m. on the East Coast, the price of oil drops slightly in response to news of the discovery of large new reserves in the Bahamas. Software at the investment division of Orange and Nassau Bank computes that it can turn a profit by emailing a quarter of its customers with a buy recommendation for oil futures, temporarily shoring up the spot market prices, as dealers stockpile supplies to meet the future demand, and then selling futures short to the rest of its customers. This plan essentially plays one sector of the customer base off against the rest, which is completely unethical, of course. But the bank’s software has not been programmed to consider such niceties. In fact, the moneymaking scenario autonomously planned by the computer is an unintended consequence of many individually sound principles. The computer’s ability to concoct this scheme could not easily have been anticipated by the programmers.

Unfortunately, the “buy” email that the computer sends directly to the customers works too well. Investors, who are used to seeing the price of oil climb and climb, jump enthusiastically on the bandwagon, and the spot price of oil suddenly climbs well beyond $300 and shows no sign of slowing down. It’s now 11:30 a.m. on the East Coast, and temperatures are climbing more rapidly than predicted. Software controlling New Jersey’s power grid computes that it can meet the unexpected demand while keeping the cost of energy down by using its coal-fired plants in preference to its oil-fired generators. However, one of the coal-burning generators suffers an explosion while running at peak capacity, and before anyone can act, cascading blackouts take out the power supply for half the East Coast. Wall Street is affected, but not before SEC regulators notice that the rise in oil future prices was a computer-driven shell game between automatically traded accounts of Orange and Nassau Bank. As the news spreads, and investors plan to shore up their positions, it is clear that the prices will fall dramatically as soon as the markets reopen and millions of dollars will be lost. In the meantime, the blackouts have spread far enough that many people are unable to get essential medical treatment, and many more are stranded far from home.

Detecting the spreading blackouts as a possible terrorist action, security screening software at Reagan National Airport automatically sets itself to the highest security level and applies biometric matching criteria that make it more likely than usual for people to be flagged as suspicious. The software, which has no mechanism for weighing the benefits of preventing a terrorist attack against the inconvenience its actions will cause for tens of thousands of people in the airport, identifies a cluster of five passengers, all waiting for Flight 231 to London, as potential terrorists. This large concentration of “suspects” on a single flight causes the program to trigger a lockdown of the airport, and the dispatch of a Homeland Security response team to the terminal. Because passengers
are already upset and nervous, the situation at the gate for Flight 231 spins out of control, and shots are fired.

An alert sent from the Department of Homeland Security to the airlines that a terrorist attack may be under way leads many carriers to implement measures to land their fleets. In the confusion caused by large numbers of planes trying to land at Chicago’s O’Hare Airport, an executive jet collides with a Boeing 777, killing 157 passengers and crew. Seven more people die when debris lands on the Chicago suburb of Arlington Heights and starts a fire in a block of homes.

Meanwhile, robotic machine guns installed on the U.S.-Mexican border receive a signal that places them on red alert. They are programmed to act autonomously in code red conditions, enabling the detection and elimination of potentially hostile targets without direct human oversight. One of these robots fires on a Hummer returning from an off-road trip near Nogales, Arizona, destroying the vehicle and killing three U.S. citizens.

By the time power is restored to the East Coast and the markets reopen days later, hundreds of deaths and the loss of billions of dollars can be attributed to the separately programmed decisions of these multiple interacting systems. The effects continue to be felt for months. Time may prove us poor prophets of disaster. Our intent in predicting such a catastrophe is not to be sensational or to instill fear. This is not a book about the horrors of technology. Our goal is to frame discussion in a way that constructively guides the engineering task of designing AMAs. The purpose of our prediction is to draw attention to the need for work on moral machines to begin now, not twenty to a hundred years from now when technology has caught up with science fiction.

The field of machine morality extends the field of computer ethics beyond concern for what people do with their computers to questions about what the machines do by themselves. We are discussing the technological issues involved in making computers themselves into explicit moral reasoners. As artificial intelligence (AI) expands the scope of autonomous agents, the challenge of how to design these agents so that they honor the broader set of values and laws humans demand of human moral agents becomes increasingly urgent.

Does humanity really want computers making morally important decisions? Many philosophers of technology have warned about humans abdicating responsibility to machines. Movies and magazines are filled with futuristic fantasies about the dangers of advanced forms of artificial intelligence. Emerging technologies are always easier to modify before they become entrenched. However, it is not often possible to predict accurately the impact of a new technology on society until well after it has been widely adopted. Some critics think, therefore, that humans should err on the side of caution and relinquish the development of potentially dangerous technologies. We believe, however, that market and political forces will prevail and will demand the benefits that these technologies can provide. Thus, it is incumbent on anyone with a stake in this technology to address head-on the task of implementing moral decision making in computers, robots, and virtual “bots” within computer networks.

Yes, the machines are coming. Yes, their existence will have unintended effects on human lives and welfare, not all of them good. But no, we do not believe that increasing reliance on autonomous systems will undermine people’s basic humanity. Neither, in our view, will advanced robots enslave or exterminate humanity, as in the best traditions of science fiction. Humans have always adapted to
their technological products, and the benefits to people of having autonomous machines around them will most likely outweigh the costs.

However, this optimism does not come for free. It is not possible to just sit back and hope that things will turn out for the best. If humanity is to avoid the consequences of bad autonomous artificial agents, people must be prepared to think hard about what it will take to make such agents good.

In proposing to build moral decision-making machines, are we still immersed in the realm of science fiction—or, perhaps worse, in that brand of science fantasy often associated with artificial intelligence? The charge might be justified if we were making bold predictions about the dawn of AMAs or claiming that “it’s just a matter of time” before walking, talking machines will replace the human beings to whom people now turn for moral guidance. We are not futurists, however, and we do not know whether the apparent technological barriers to artificial intelligence are real or illusory. Nor are we interested in speculating about what life will be like when your counselor is a robot, or even in predicting whether this will ever come to pass. Rather, we are interested in the incremental steps arising from present technologies that suggest a need for ethical decision-making capabilities. Perhaps small steps will eventually lead to full-blown artificial intelligence—hopefully a less murderous counterpart to HAL in 2001: A Space Odyssey—but even if fully intelligent systems will remain beyond reach, we think there is a real issue facing engineers that cannot be addressed by engineers alone.

Is it too early to be broaching this topic? We don’t think so. Industrial robots engaged in repetitive mechanical tasks have caused injury and even death. The demand for home and service robots is projected to create a world-wide market double that of industrial robots by 2010, and four times bigger by 2025. With the advent of home and service robots, robots are no longer confined to controlled industrial environments where only trained workers come into contact with them. Small robot pets, for example Sony’s AIBO, are the harbinger of larger robot appliances. Millions of robot vacuum cleaners, for example iRobot’s “Roomba,” have been purchased. Rudimentary robot couriers in hospitals and robot guides in museums have already appeared. Considerable attention is being directed at the development of service robots that will perform basic household tasks and assist the elderly and the home-bound. Computer programs initiate millions of financial transactions with an efficiency that humans can’t duplicate. Software decisions to buy and then resell stocks, commodities, and currencies are made within seconds, exploiting potentials for profit that no human is capable of detecting in real time, and representing a significant percentage of the activity on world markets.

Automated financial systems, robotic pets, and robotic vacuum cleaners are still a long way short of the science fiction scenarios of fully autonomous machines making decisions that radically affect human welfare. Although 2001 has passed, Arthur C. Clarke’s HAL remains a fiction, and it is a safe bet that the doomsday scenario of The Terminator will not be realized before its sell-by date of 2029. It is perhaps not quite as safe to bet against the Matrix being realized by 2199. However, humans are already at a point where engineered systems make decisions that can affect humans’ lives and that have ethical ramifications. In the worst cases, they have profound negative effect.

Is it possible to build AMAs? Fully conscious artificial systems with complete human moral capacities may perhaps remain forever in the realm of science fiction. Nevertheless, we believe that more limited systems will soon be built. Such systems will have some capacity to evaluate the
ethical ramifications of their actions—for example, whether they have no option but to violate a property right to protect a privacy right.

The task of designing AMAs requires a serious look at ethical theory, which originates from a human-centered perspective. The values and concerns expressed in the world’s religious and philosophical traditions are not easily applied to machines. Rule-based ethical systems, for example the Ten Commandments or Asimov’s Three Laws for Robots, might appear somewhat easier to embed in a computer, but as Asimov’s many robot stories show, even three simple rules (later four) can give rise to many ethical dilemmas. Aristotle’s ethics emphasized character over rules: good actions flowed from good character, and the aim of a flourishing human being was to develop a virtuous character. It is, of course, hard enough for humans to develop their own virtues, let alone developing appropriate virtues for computers or robots. Facing the engineering challenge entailed in going from Aristotle to Asimov and beyond will require looking at the origins of human morality as viewed in the fields of evolution, learning and development, neuropsychology, and philosophy.

Machine morality is just as much about human decision making as about the philosophical and practical issues of implementing AMAs. Reflection about and experimentation in building AMAs forces one to think deeply about how humans function, which human abilities can be implemented in the machines humans design, and what characteristics truly distinguish humans from animals or from new forms of intelligence that humans create.

Just as AI has stimulated new lines of enquiry in the philosophy of mind, machine morality has the potential to stimulate new lines of enquiry in ethics. Robotics and AI laboratories could become experimental centers for testing theories of moral decision making in artificial systems.

Three questions emerge naturally from the discussion so far. Does the world need AMAs? Do people want computers making moral decisions? And if people believe that computers making moral decisions are necessary or inevitable, how should engineers and philosophers proceed to design AMAs?

Some basic moral decisions may be quite easy to implement in computers, while skill at tackling more difficult moral dilemmas is well beyond present technology. Regardless of how quickly or how far humans progress in developing AMAs, in the process of addressing this challenge, humans will make significant strides in understanding what truly remarkable creatures they are. The exercise of thinking through the way moral decisions are made with the granularity necessary to begin implementing similar faculties into (ro)bots is thus an exercise in self-understanding.

Questions:

1. How likely do you think the catastrophic incident described in the text (or something very similar) is to happen in our lifetimes? What do you think the public reaction would be?

2. After reading this text, how likely do you think it is that an Artificial Intelligence will be created that actually thinks and reasons without human help? How will this affect our human identities?

3. List a couple of examples of positives and negatives of developing artificial intelligence.

4. Have you ever had a machine or computer fail that has caused you some difficulties? Explain.