ISN
ETH Zurich

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

The Moral Code: How to Teach Robots Right and Wrong

By Nayef Al-Rodhan, September 11, 2015

At the most recent International Joint Conference on Artificial Intelligence, over 1,000 experts and researchers presented an open letter calling for a ban on offensive autonomous weapons. The letter, signed by Tesla's Elon Musk, Apple co-founder Steve Wozniak, Google DeepMind CEO Demis Hassabis, and Professor Stephen Hawking, among others, warned of a "military artificial intelligence arms race." Regardless of whether these campaigns to ban offensive autonomous weapons are successful, though, robotic technology will be increasingly widespread in many areas of military and economic life.

Over the years, robots have become smarter and more autonomous, but so far they still lack an essential feature: the capacity for moral reasoning. This limits their ability to make good decisions in complex situations. For example, a robot is not currently able to distinguish between combatants and noncombatants or to understand that enemies sometimes disguise themselves as civilians.

To address this failing, in 2014, the U.S. Office of Naval Research offered a $7.5 million grant to an interdisciplinary research team from Brown, Georgetown, Rensselaer Polytechnic Institute, Tufts, and Yale to build robots endowed with moral competence. They intend to capture human moral reasoning as a set of algorithms, which will allow robots to distinguish between right and wrong and to override rigid instructions when confronted with new situations.

The idea of formalizing ethical guidelines is not new. More than seven decades ago, science-fiction writer Isaac Asimov described the "three laws of robotics"—a moral compass for artificial intelligence. The laws required robots to protect humans, obey instructions, and preserve themselves, in that order. The fundamental premise behind Asimov's laws was to minimize conflicts between humans and robots. In Asimov's stories, however, even these simple moral guidelines lead to often disastrous unintended consequences. Either by receiving conflicting instructions or by exploiting loopholes and ambiguities in these laws, Asimov's robots ultimately tend to cause harm or lethal injuries to humans.

Today, robotics requires a much more nuanced moral code than Asimov's "three laws." Robots will be deployed in more complex situations that require spontaneous choices. The inevitable next step, therefore, would seem to be the design of "artificial moral agents," a term for intelligent systems endowed with moral reasoning that are able to interact with humans as partners. In contrast with software programs, which function as tools, artificial agents have various degrees of autonomy.

However, robot morality is not simply a binary variable. In their seminal work *Moral Machines,* Yale's Wendell Wallach and Indiana University's Colin Allen analyze different gradations of the ethical sensitivity of robots. They distinguish between operational morality and functional morality. Operational morality refers to situations and possible responses that have been entirely anticipated and precoded by the designer of the robot system. This could include the profiling of an enemy combatant by age or physical appearance.

The most critical of these dilemmas is the question of whose morality robots will inherit.

Functional morality involves robot responses to scenarios unanticipated by the programmer, where the robot will need some ability to make ethical decisions alone. Here, they write, robots are endowed with the capacity to assess and respond to "morally significant aspects of their own actions." This is a much greater challenge.

The attempt to develop moral robots faces a host of technical obstacles, but, more important, it also opens a Pandora's box of ethical dilemmas.

Whose values?

The most critical of these dilemmas is the question of whose morality robots will inherit. Moral values differ greatly from individual to individual, across national, religious, and ideological boundaries, and are highly dependent on context. For example, ideas of duty or sacrifice vary across cultures. During World War II, Japanese banzai attacks were supported by a cultural expectation that saw death as a soldier's duty and surrender as an unforgivably shameful act. Similarly, notions of freedom and respect for life have very different connotations in peacetime or war. Even within any single category, these values develop and evolve over time.

Human morality is already tested in countless ways, and so too will be the morality of autonomous robots and artificial intelligence. Uncertainty over which moral framework to choose underlies the difficulty and limitations of ascribing moral values to artificial systems. The Kantian deontological (duty-based) imperative calls for rigid ethical constraints on one's actions. It requires acting in a way that reflects universal values and sees humanity as an end, not as a means. In contrast, utilitarianism stresses that one should calculate only the consequences of one's action—even if that action is not initially recognizably moral—and choose the most beneficial course. However, do we trust a robot to anticipate and weigh the numerous possible consequences of its actions? To implement either of these frameworks effectively, a robot would need to be equipped with an almost impossible amount of information. Even beyond the issue of a robot's decision-making process, the specific issue of cultural relativism remains difficult to resolve: no one set of standards and guidelines for a robot's choices exists.

For the time being, most questions of relativism are being set aside for two reasons. First, the U.S. military remains the chief patron of artificial intelligence for military applications and Silicon Valley for other applications. As such, American interpretations of morality, with its emphasis on freedom and responsibility, will remain the default. Second, for the foreseeable future, artificial moral agents will not have to confront situations outside of the battlefield, and the settings in which they will be given autonomy will be highly constrained.

Learning by doing

Even if the ethical questions are eventually answered, major technical challenges would still remain in coding something as abstract as morality into transistors.

There are two mainstream approaches. First is the top-down approach, which requires encoding specific moral values into an algorithm. These moral values are determined by the robot's developers and can be based on frameworks such as religion, philosophical doctrines, or legal codes. To many neuroscientists and psychologists, this approach holds severe limitations. It devalues the fundamental role that experience, learning, and intuition play in shaping our understanding of the world and thus our moral code.

The second approach is bottom-up and is based on letting robots acquire moral competence through their own learning, trial and error, growth, and evolution. In computational terms, this system is extremely challenging, but the advent of neuromorphic computing could make it a reality. Neuromorphic ("brainlike") chips aim to replicate the morphology of human neurons and emulate the neural architecture of the brain in real time. Neuromorphic chips would enable robots to process data similarly to humans—nonlinearly and with millions of interconnected artificial neurons. This would be a far cry from conventional computing technology, which relies on linear sequences of calculations. This may sound like science fiction, but IBM has already developed the TrueNorth chip, which is able to mimic over one million human neurons. Robots with neuromorphic chips would possess humanlike intelligence and be able to grasp the world in unique (humanlike) ways.

The ability to learn and experience offers no guarantee that a robot would consistently adhere to a "high" moral code. A robot equipped with a neuromorphic chip may appear ideal, but it does not promise "moral" outcomes in all situations, simply because human morality itself is often suboptimal and flawed.

In fact, the dissimilarity between robots and humans is sometimes touted as their greatest advantage. Proponents of moral robots argue that a robot, unlike a human, could not be affected by the stress of combat or succumb emotionally under pressure. While humans are inconsistent and get bored or tired, robots could apply codes of conduct more systematically. For instance, they would not act erratically or shoot indiscriminately at a crowd in a moment of panic.

Robots with neuromorphic chips would possess humanlike intelligence and be able to grasp the world in unique (humanlike) ways.

Neuromorphic chips, and the humanlike behavior they may bring, would therefore not necessarily be a net gain in terms of moral benefits. Robots could develop humanlike weaknesses: hesitation, selfishness, or misunderstandings that could hinder their ability to accomplish their duties.

An existential risk?

If humans successfully develop neuromorphic chips that enable robots to grasp the world in humanlike ways, what would constitute robots' moral framework? There are several possible answers, but I prefer to look to neuroscience.

Neuroscience and brain imaging today suggest that humans are inherently neither moral nor immoral, but amoral. We function as a "predisposed tabula rasa." That is, our moral compass is shaped by our upbringing and environment, but our propensity to be moral varies according to our perceived emotional self-interest. Humans are also fundamentally egoistic: our actions will, in most cases, be guided by our desire to maximize our chances of survival. The fundamental human instinct for survival and dominance is coded in our genetics and is a powerful motivator throughout our existence.

The very concept of making moral robots implies that they cannot be originally amoral. Even with neuromorphic technology, they cannot learn moral values from absolute scratch; they would still be programmed with basic preferences or biases established by their programmers. Eventually, a more sophisticated robot capable of writing its own source code could start off by being amoral and develop its own moral compass through learning and experience. Such robots, like humans, might ultimately be driven by self-interest and an intrinsic desire to ensure their own survival.

If this comes to pass, the implications are daunting. Robots might compete with humans for survival and dominance. Alternatively, robotics could be used to enhance human cognition. The future is uncertain. In the best-case scenario, robots will be successfully programmed with benign moral values and will constitute no threat. However, a more likely scenario is the development of autonomous robots that may be amoral or even immoral—a serious challenge to the future of humanity.

Nayef Al-Rodhan (@SustainHistory) is a Philosopher, Neuroscientist and Geostrategist. He is an Honorary Fellow at St. Antony's College, University of Oxford, UK, and Senior Fellow and Head of the Geopolitics and Global Futures Programme at the Geneva Centre for Security Policy, Geneva, Switzerland. Author of Neo-statecraft and Meta-Geopolitics. Reconciliation of Power, Interests and Justice in the 21st Century (LIT: Zurich, 2009)

This article was originally published on the *Foreign Affairs* Website on August 12, 2015, and published on the *ISN ETH Zurich* Website on September 11, 2015. http://isnblog.ethz.ch/technology/the-moral-code-how-to-teach-robots-right-and-wrong