Inquiring Minds topic – 13 October 2017

Roger Palms, Moderator

HOW DO WE TEACH ROBOTS RIGHT FROM WRONG? SOON THE PROBLEM WON'T BE HYPOTHETICAL

*By **Mark Wyner***

Digital Trends contributor Mark Wyner wonders how we go about teaching artificial intelligence right from wrong.

To consider:

1. As robots are built to perform certain acts, how can we teach them when not to perform those acts?
2. If robots can be programed to create other robots and the original program contains an ethical or moral flaw, the "progeny" will act the same way as the "parent." Will those fault-bearing robots out-produce us who are morally-conscious humans?
3. Will enacting laws stop malevolent robot creators? Can we even know the potential to do harm before that harm is done?
4. At what point could the human creator lose control of what he has created? Will that crossing point be discovered only after the robots are loose on their own?
5. If it can be done, it will be done. If malevolent people can create evil Twitter bots, what will sophisticate malevolent people create with their artificial intelligence-bearing robots?

Twitter has admitted that as many as 23 million (8.5 percent) of its user accounts are autonomous Twitterbots. Many are there to increase productivity, conduct research, or even have some fun. Yet many have been created with harmful intentions. In both cases, the bots have been known to behave with questionable ethics – maybe because they're merely minor specimens of artificial intelligence (AI).

Humans are currently building far-more sophisticated machines which will face questions of ethics on monumental scales. So how do we make sure they make the right choices when the time comes?

Build it in or teach it

The key factor in successfully building autonomous machines that coincide symbiotically with human beings is ethics. And there are basically two ways to program ethics into machines:

First, you can hard code them into their operating systems. The concern here is that ethics are subjective. The ethics in the machine are contingent upon the ethics in its creator. But we humans do not always align in our morality. We fight wars over ethical differences. So as we build autonomous machines to be ethical, we're building within the confines of our existing disparities.

Second, you can provide some guidelines, then allow the machine to learn its own ethics based on its own experiences. This passive approach leaves plenty of room for misinterpretations of morality, contingent upon which behaviors are observed. Consider the recent meltdown of Microsoft's Twitter AI, Tay, who was tricked into tweeting racist slurs and promotions of genocide based on a false inference of accepted ethical normalcy.

A team at Georgia Tech is working on the latter, teaching cognitive-systems to learn how to behave in socially acceptable ways by reading stories. A reward system called Quixote is supposed to help cognitive systems identify protagonists in stories, which the machines use to align their own values with those of human beings. It's unclear what methods Microsoft used with Tay. But if their techniques were preemptive, as with Georgia Tech's learning system, we're a long way from solidifying ethics in Artificial Intelligence.

Ethical paralysis

Now, all of this is based on the idea that a computer can even comprehend ethics. As Alan Winfield shows in his study *Towards an Ethical Robot*, when a computer encounters an ethical paradox, the result is unpredictable, often paralyzing. In his study, a cognitive robot (A-robot) was asked to save a "human" robot (H-robot) from peril. When the A-robot could save only one of two H-robots, it dithered and conceded in its own confusion, saving neither.

There is an ancient philosophical debate about whether ethics is a matter of reason or emotion. Among modern psychologists, there is a consensus that ethical decision making requires both rational and emotional judgments. As Professor Paul Thagard notes, "ethical judgments are often highly emotional, when people express their strong approval or disapproval of various acts. Whether they are also rational depends on whether the cognitive appraisal that is part of emotion is done well or badly."

Decisions with consequences

So, if cognitive machines don't have the capacity for ethics, who is responsible when they break the law? Currently, no one seems to know. Ryan Calo of the University of Washington School of Law notes, "robotics combines, for the first time, the promiscuity of data with the capacity to do physical harm; robotic systems accomplish tasks in ways that cannot be anticipated in advance; and robots increasingly blur the line between person and instrument."

The crimes can be quite serious, too. Netherlands developer Jeffry van der Goot had to defend himself — and his Twitterbot — when police knocked on his door, inquiring about a death threat sent from his Twitter account. Then there's Random Darknet Shopper, a shopping bot with a weekly allowance of $100 in Bitcoin to make purchases on Darknet for an art exhibition. Swedish officials weren't amused when it purchased ecstasy, which the artist put on display. (*Though, in support for artistic expression they didn't confiscate the drugs until the exhibition ended.*)

In both of these cases, authorities did what they could within the law, but ultimately pardoned the human proprietors because they hadn't explicitly or directly committed crimes. But how does that translate when a human being unleashes an AI with the intention of malice?

The ironic reality is that we are exploring ways we can govern our autonomous machines. And beyond our questionable ability to instill ethics into them, we are often bemused by their general behavior alone. When discussing the methods behind their neural networks, Google software engineer Alexander Mordvintsev revealed, "… even though these are very useful tools based on well-known mathematical methods, we actually understand surprisingly little of why certain models work and others don't."

Can we keep up?

All things considered, the process for legislation is arduously slow, while technology, on the other hand, makes exponential haste. As Vivek Wadhwa of Singularity University explains, "the laws can't keep up because … laws are essentially codified ethics. That we develop a consensus as a society about what's good and what's bad and then it becomes what's right and what's wrong, and then it becomes what's legal and what's illegal. That's the way the progression goes. On most of those technologies we haven't decided what's good or bad."

If the law does catch up, we may be writing our own doom. All of that talk about robots taking over the world? Maybe they just jaywalk en masse until we imprison so much of our race that we become the minority in autonomous beings. Checkmate.

For additional information see Alan Winfield's Web Log: "Towards an Ethical Robot."