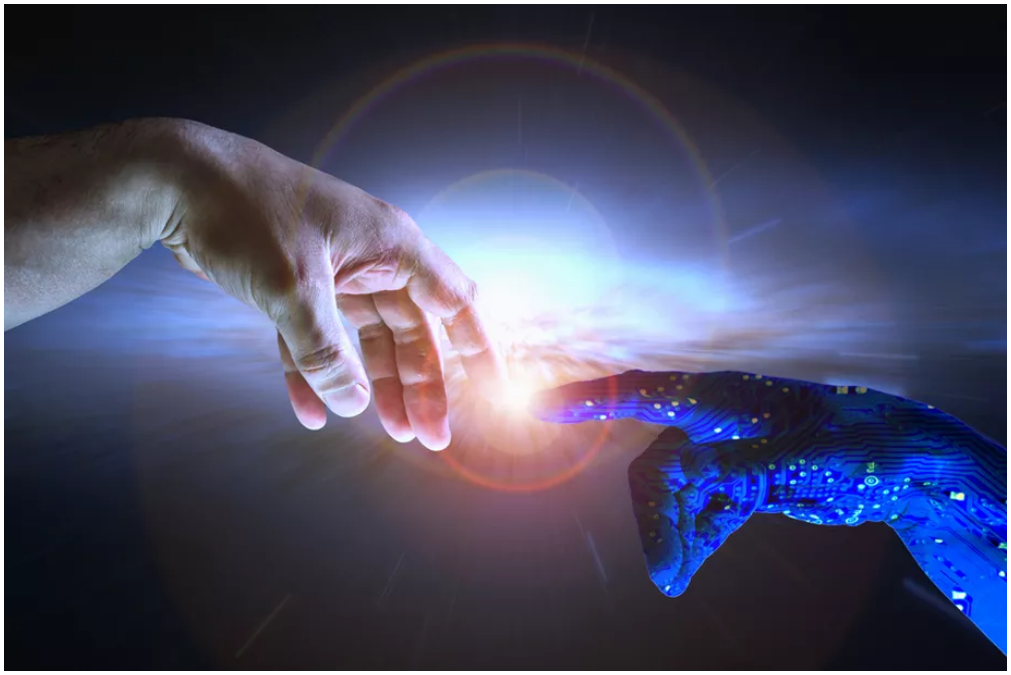
recode

VOICES ARTIFICIAL INTELLIGENCE

Ethics and Artificial Intelligence: The Moral Compass of a Machine

The question of robotic ethics is making everyone tense.

BY KRIS HAMMOND | APR 13, 2016, 2:22PM EDT



John Williams RUS/Shutterstock

The question of robotic ethics is making everyone tense. We worry about the machine's lack of empathy, how calculating machines are going to know how to do the right thing, and even how we are going to judge and punish beings of steel and silicon.

Personally, I do not have such worries.

I am less concerned about robots doing wrong, and far more concerned about the moment they look at us and are appalled at how often we fail to do right. I am convinced that they will not only be smarter than we are, but have truer moral compasses, as well.

Let's be clear about what is and is not at issue here.

> I am less concerned about robots doing wrong, and far more concerned about the moment they look at us and are appalled at how often we fail to do right.

First, I am not talking about whether or not we should deploy robotic soldiers. That is an ethical decision that is in human hands. When we consider the question of automating war, we are considering the nature of ourselves not our machines. Yes, there is a question of whether the capabilities of robotic soldiers and autonomous weapons are up to the task, but that has to do with how well they work rather than what their ethics are.

Second, I am not talking about the "ethics" of machines that are just badly designed. A self-driving car that plows into a crowd of people because its sensors fail to register them isn't any more unethical than a vehicle that experiences unintended acceleration. It is broken or badly built. Certainly there is a tragedy here, and there is responsibility, but it is in the hands of the designers and manufacturers.

Third, while we need to look at responsibility, this is not about punishment. Our ability or inability to punish a device is a matter of how we respond to unethical behavior, not how to assess it. The question of whether a machine has done something wrong is very different than the issue of what we are going to do about it.

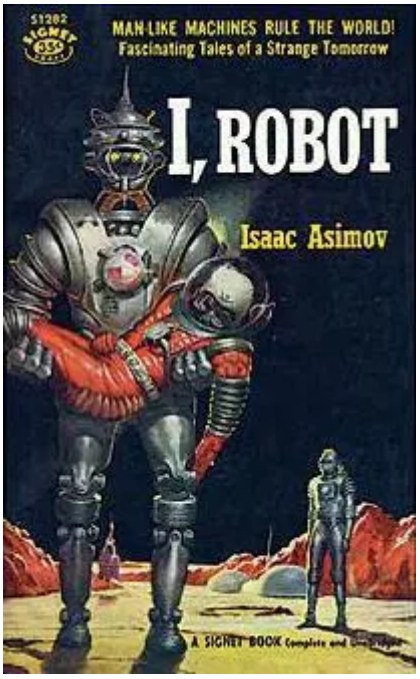
> This is not about pathological examples such as hyperintelligent paper-clip factories that destroy all of humanity in single-minded efforts to optimize production at the expense of all other goals.

Finally, this is not about pathological examples such as hyperintelligent paper-clip factories that destroy all of humanity in single-minded efforts to optimize production at the expense of all other goals. I would put this kind of example in the category of "badly designed." And given that most of the systems that manage printer queues in our offices are smarter than a system that would tend to do this, it is probably not something that should concern us.

These are examples of machines doing bad things because they are broken or because that's how they are built. These are all examples of tools that might very well hurt us, but do not have to themselves deal with ethical dilemmas.

But "dilemma" is the important word here.

Situations that match up well against atomic rules of action are easy to deal with for both machines and people. Given a rule that states that you should never kill anyone, it is pretty easy for a machine (or person for that matter) to know that it is wrong to murder the owner of its local bodega, even if it means that it won't have to pay for that bottle of Chardonnay. Human life trumps cost savings.



This is why Isaac Asimov's "Three Laws of Robotics" seem so appealing to us. They provide a simple value ranking that — on the face of it, at least — seems to make sense:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The place where both robots and humans run into problems is situations in which adherence to a rule is impossible, because all choices violate the same rule. The standard example that is used to explain this is the Trolley Car Dilemma.

The dilemma is as follows:

A train is out of control and moving at top speed down a track. At the end of the track, five people are tied down, and will be killed in seconds. There is a switch that can divert the train to another track but, unfortunately, another person is tied down on that track, and will be crushed if you pull the switch.

If you pull the switch, one person dies. If you don't, five people die. Either way, your action or inaction is going to kill people. The question is, how many?

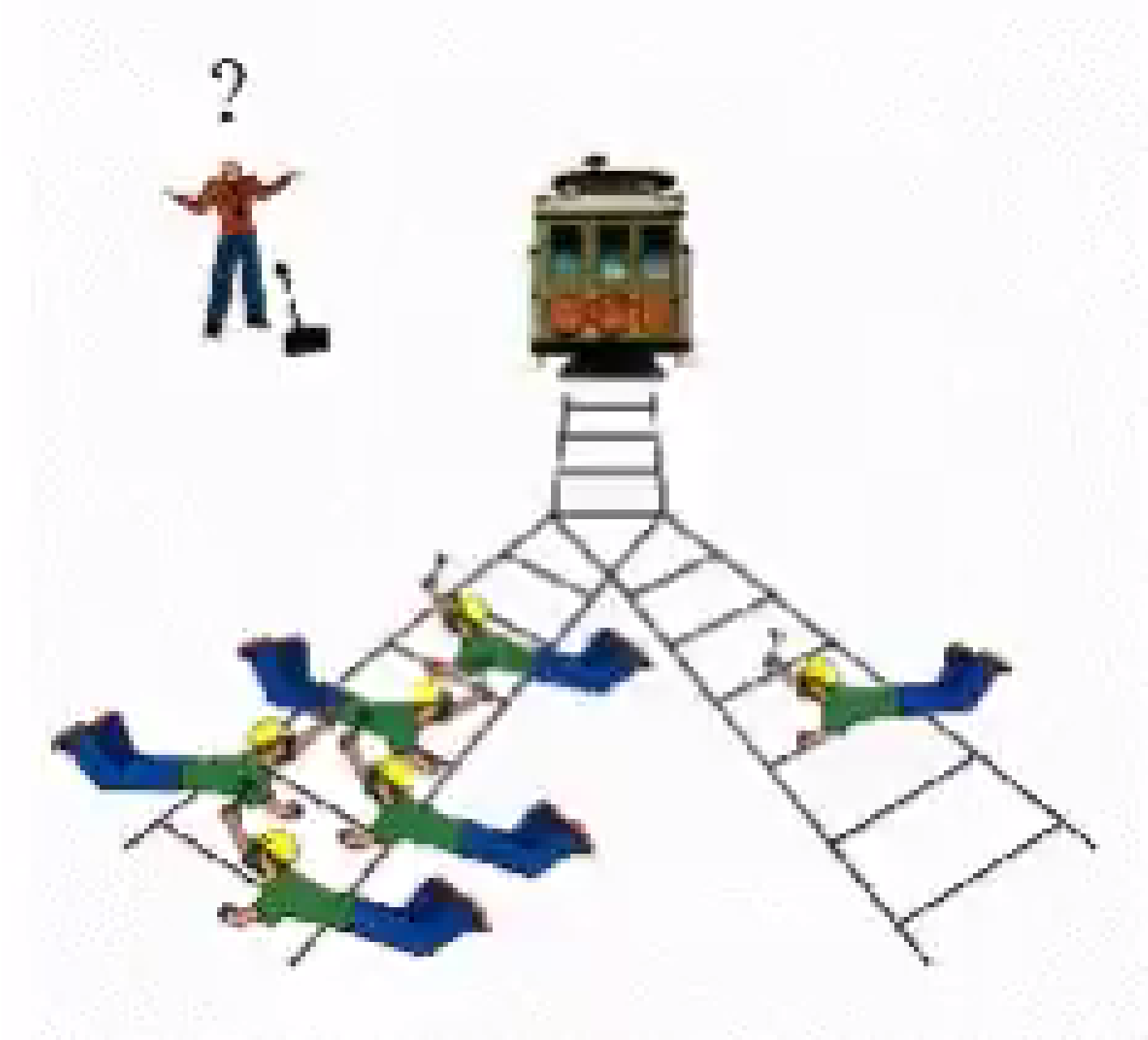
Most people actually agree that sacrificing the lone victim makes the most sense. You are trading one against five. Saving more lives is better than saving fewer lives. This is based on a fairly utilitarian calculus that one could easily hand to a machine.

Unfortunately, it is easy to the change details of this example in ways that shift our own intuitions.

Imagine that the single victim is a researcher who now has the cure for cancer in his or her head. Or our lone victim could be a genuinely noble person who has and will continue to help those in need. Likewise, the victims on the first track could all be terminally ill with only days to live, or could all be convicted murderers who were on their way to death row before being waylaid.

In each of these cases, we begin to consider different ways to evaluate the trade-offs, moving from a simple tallying up of survivors to more nuanced calculations that take into account some assessment of their "value."

Even with these differences, the issue still remains one of a calculus of sorts.



overthinkingit.com

But one change wipes this all away.

There are five people tied down, and the trolley is out of control, but there is only one track. The only way to halt the trolley is to derail it by tossing something large onto the track. And the only large thing you have it hand is the somewhat large person standing next to you. There is no alternative.

You can't just flip a switch. You have to push someone onto the track.

If you are like most people, the idea of doing this changes things completely. It has a disturbing intimacy that is not part of the earlier one. As a result, although most people would pull the switch, those same people resist the idea of pushing their fellow commuter to his or her doom to serve the greater good.

But while they are emotionally different, from the point of view of ethics or morality they are the same. In both cases, we are taking one life to save others. The calculus is identical, but our feelings are different.

Of course, as we increase the number of people on the track, there is a point at which most of us think that we will overcome our horror and sacrifice the life of the lone commuter in order to save the five, 10, one hundred or one thousand victims tied to the track.

And it is interesting to consider what we say about such people. What do we say to someone who is on their knees weeping because they have done a horrible thing in service of what was clearly the greater good? We tell them that they did what they had to do, and they did the right thing. We tell them that they were brave, courageous, and even heroic.

The Trolley Dilemma exposes an interesting problem. Sometimes our ethical and moral instincts are skewed by circumstance. Our determination of what is right or wrong becomes complex when we mix in emotional issues related to family, friends, tribal connections, and even the details of the actions that we take. The difficulty of doing the right thing does not rise out of us not knowing what it is. It comes from us being unwilling to pay the price that the right action often demands.

And what of our robot friends?

I would argue that an ethical or moral sense for machines can be built on a utilitarian base. The metrics are ours to choose, and can be coarse-grained (save as many people as possible), nuanced (women, children and Nobel laureates first) or detailed (evaluate each individual by education, criminal history, social media mentions, etc.). The choice of the code is up to us.

> I would argue that an ethical or moral sense for machines can be built on a utilitarian base. The choice of code is up to us.

Of course, there are special cases that will require modifications of the core rules that are based on the circumstances of their use. Doctors, for example, don't euthanize patients in order to spread the wealth of their organs, even if it means that there is a net positive with regard to survivors. They have to conform to a separate code of ethics designed around the needs of patients and their rights that restricts their actions. The same holds for lawyers, religious leaders and military personnel who establish special relationships with individuals that are protected by specific ethical code.

So the simple utilitarian model will certainly have overlays depending on the role that these robots and AIs will play. It would not seem unreasonable for a machine to respond for a request for personal information by saying "I am sorry but he is my patient and that information is protected." In much the same way that Apple protected its encryption in the face of homeland security, it follows that robotic doctors will be asked to be HIPAA compliant.

Our machines need not hesitate when they see the Trolley coming. They will act in accord with whatever moral or ethical code we provide them and the value determinations that we set. They will run the numbers and do the right thing. In emergency situations, our autonomous cars will sacrifice the few to protect the many. When faced with dilemmas, they will seek the best outcomes independent of whether or not they themselves are comfortable with the actions. And while we may want to call such calculations cold, we will have to admit that they are also right.

> Machine intelligence will be different than us, and might very well do things that are at odds with what we expect.

But they will be different than us, and might very well do things that are at odds with what we expect. So, as with all other aspects of machine intelligence, it is crucial that these systems are able to explain their moral decisions to us. They will need to be able to reach into their silicon souls and explain the reasoning that supports their actions.

Of course, we will need them to do be able to explain themselves in all aspects of their reasoning and actions. Their moral reasoning will be subject to the same

explanatory requirements that we would demand of explaining any action they take. And my guess is that they will be able to explain themselves better than we do.

At the end of the movie "I, Robot," Will Smith and his robot partner have to disable an AI that has just enslaved all of humanity. As they close in on their goal, Smith's onscreen girlfriend slips, and is about to fall to her death. In response, Smith screams, "Save the girl!" and the robot, demonstrating its newly learned humanity, turns its back on the primary goal and focuses on saving the girl. While very "human," this action is intensely selfish, and a huge moral lapse.

Every time I watch this scene, I just want the robot to say, "I'm sorry, I can't. I have to save everyone else." But then, I don't want it to be human. I want it to be true to its code.

In addition to being chief scientist at Narrative Science, Kris Hammond is a professor of Computer Science and Journalism at Northwestern University. Prior to joining the faculty at Northwestern, Hammond founded the University of Chicago's Artificial Intelligence Laboratory. His research has been primarily focused on artificial intelligence, machine-generated content and context-driven information systems. He currently sits on a United Nations policy committee run by the United Nations Institute for Disarmament Research (UNIDIR). Reach him @KJ_Hammond.

THE LATEST

Amazon threatened to kill its Whole Foods deal if the grocer started a bidding war

BY JASON DEL REY

Here's how to tip your Uber driver in the app

BY JOHANA BHUIYAN

Full transcript: Headspace meditation app co-founder and CEO Rich Pierson on Too Embarrassed to Ask

BY RECODE STAFF

Recode Daily: The Uber in-app tipping era has arrived

BY RECODE STAFF