The Ethics of AI:
What does AI do when Humans cannot agree what is "Right"?

Stuart Madnick

Working Paper CISL# 2019-20

October 2019

Cybersecurity Interdisciplinary Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

The Ethics of AI:
What does AI do when Humans cannot agree what is "Right"?

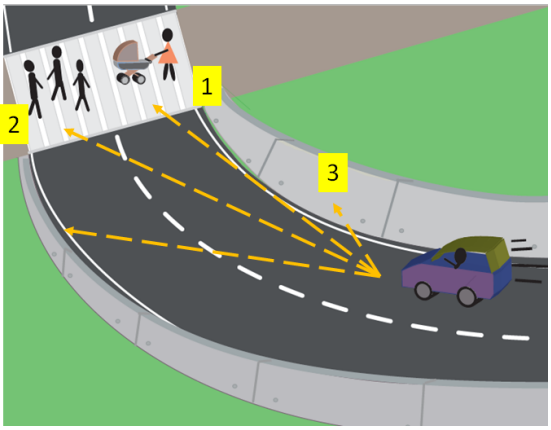
Stuart Madnick
MIT Sloan School of Management

Stuart Madnick (@smadnick2) is the John Norris Maguire Professor of Information Technologies at the MIT Sloan School of Management and the founding director of the Cybersecurity at MIT Sloan (CAMS) research consortium.

One of the appealing features of artificial intelligence is the ability to come up with the "right" answer automatically, faster and more reliably than humans. In many cases, the right answer is singular and unambiguous, such as 2 + 2 = 4 (3.9 or 4.1 are close, but not right.). But, as is often the case in matters of ethics, what is AI to do if humans cannot agree on the right answer?

This challenge has two important consequences. First, it can delay the introduction or acceptance of new AI applications, such as autonomous vehicles. Second, it will require that management be prepared to explain and justify the rationale for how their AI will make these decisions.

In a class I teach, I use the following example to illustrate the point. Imagine an autonomous vehicle, under the control of AI, is driving down a steep mountain in Switzerland. It makes a sharp turn and the sensors and object-recognition software realize that: (1) There is a woman pushing a baby carriage in the cross walk to the right, (2) Three gentleman entered the intersection on the left, and (3) There are concrete barriers on the left and right.



A classic thought exercise reveals some of the challenges of artificial intelligence in no-win ethics scenarios. PHOTO: CYBERSECURITY AT MIT SLOAN

The software also realizes that given the speed of the vehicle, the condition of the pavement and the distances involved, it would not be able to stop before coming across the intersection.

So, what should the vehicle do? When I ask my students, including senior executives, for a definitive, instant answer, there is a strong reluctance to respond. But situations like this will soon be reality and the consequences cannot be ignored.

To back up and reduce the stress a notch, I have my students explain what is wrong about each choice. The discussion goes somewhat as follows:

1. Hitting the woman with the baby sounds terrible. Can you imagine the press saying that "the car deliberately killed them?" The baby might have grown up to be an important inventor or even president.

2. Killing the three gentleman violates the ethics principle known as Utilitarianism. That is, given a choice, it is better to save three people than two people.
3. Causing the car to go into either of the concrete barriers would definitely kill the "driver" (though he or she was not actually controlling the car; the AI was). But imagine trying to sell a car with a sticker on it: "Notice: If necessary, this car will kill the driver." Furthermore, from an ethics perspective, shouldn't protecting the driver be of highest importance?

For managers, this simple example makes clear the challenges of addressing ethical questions for AI. There will likely be many such situations as we apply AI to an increasing number of applications beyond autonomous vehicles. Of course, there is no decision that everyone would agree with. But much like the medical profession had to determine how to triage patients in the aftermath of a disaster, managers will need to develop a set of thoughtful principles, and not leave such decisions to some programmer of the AI. And they will need to be prepared to defend those principles.

At the same time, those principles have to be balanced at against other considerations. For example, in the case of autonomous vehicles, computers do not get drowsy or distracted, and they can respond with split-second speed. There are estimates that even with current-quality AI, autonomous vehicles could save a tremendous number of lives (an estimated 3,287 lost in car crashes each day, globally).

If a driver nods off or doesn't notice a person running onto the road in time and hits a pedestrian, it is very tragic, but we would usually forgive that as a human accident. But, what if an autonomous vehicle makes one mistake that we would accept from a human—can we forgive AI?

Under the basic principles of Utilitarian ethics, saving thousands of lives even at the cost of a few is obviously the right thing to do—but could *you* forgive the computer, especially if it was your relative who was killed?

Ethical decisions are indeed hard, and AI will increasingly raise these dilemmas. The dialogue on these matters must be started now, by creators of the science, by business leaders responsible for its uses, and by society which will have to live with the consequences