Moral Decision Making with Autonomous Systems

Discussion Leads: Rahul Mangharam and Johannes Betz

We focus on the question:
Who or what is morally responsible for the decisions made by an autonomous system?

Read the article and come with your written arguments. This note frames Thursday's discussion.
"Killer Robots" by Robert Sparrow, in Journal of Applied Philosophy, Vol. 24, No. 1, 2007

As engineers, we design systems to have a certain function: move parts in a warehouse, interpret and implement a surgeon's hand movements, or drive passengers around. Aside from functionality, we also design safety, security and reliability into our systems. With autonomous systems, we will be asked to design *ethics* into our systems, since they will share the world with us and make decisions autonomously, and some of these decisions are bound to involve ethical questions. How can we design and program an autonomous system that is ethical? Indeed, what does it mean for an autonomous system to be ethical? Or even, just *responsible* for its actions? Are there certain autonomous systems that simply cannot be ethical? Moreover, when speaking about designing ethical systems, *whose ethics are we talking about*?

The engineers who will be called upon to build and program autonomous systems must have an awareness of these issues and a solid foundation upon which to base their thinking about them. One can safely say that the engineer's own ethical obligation and professional code require them to think carefully about these questions and present their employers and the general public with a principled analysis of the ethical dilemmas facing the autonomous system they design - and in some cases, choose not to design.

This assignment asks us to discuss some of the ethical considerations inherent in the design of autonomous systems generally, not just self-driving cars.

Where does responsibility lie?

Since the early days when it became apparent that self-driving cars are a real possibility, engineers, regulators and jurists started asking who would be liable when the car injures or kills someone though a mistake of its own? For instance, take the case of the Uber incident in Arizona, where an SUV in Uber autonomous mode struck and killed 49-year-old Elaine Herzberg on Sunday March 18, 2018[1]. Who is *responsible* for her death? Possibilities include:

- The car's passenger, though they were not in physical control of the car at the time
- The car's autonomy manufacturer (in this case, Uber[2])
- The designer of the computer vision algorithm that failed to identify the woman with sufficient

[1] The Guardian article here https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

[2] We are assuming the mechanical vehicle itself, manufactured by Volvo, is not at fault: the brakes didn't fail, the tires didn't slip, etc.

confidence
- The company that provided the dataset on which the computer vision algorithm (a deep neural net) was trained
- The city and state regulators who allowed Uber to test its cars on public roads at such an early stage

Note that the entity or entities responsible are not necessarily the ones that get sued in court, since the choice of who to sue depends on many other factors, including whether they're capable of paying damages, whether they're affected by negative publicity, etc. In particular, *moral responsibility*, which is what we're concerned with here, is different from *legal responsibility*, although a determination of moral responsibility is likely to affect the determination of legal responsibility.

We will explore this issue in our discussion, but in a different context, which brings out the urgency and salient traits of this issue in sharp relief: namely, in the context of autonomous weapons, that make the kill decision and implement it independently of humans. This is not because we think autonomous weapons are fine and dandy - we don't. Rather, it is because thinking about autonomous weapons makes it very clear what is at stake in terms of responsibility assignment. Who is morally responsible when an autonomous robot takes a decision to kill a human being?
