Collaborative Sciences Center for
ROAD SAFETY

Coffee and Conversation Speaker Series

Proceedings from "Beyond the Trolley Problem: Why Vehicles Must Be Ethical"
Monday, December 3, 2018

Kurt Gray, Associate Professor
Department of Psychology and Neuroscience
University of North Carolina at Chapel Hill

In the final "Coffee and Conversations" talk of 2018, the theme was centered on ethics in autonomous vehicles.

Ethical decisions are hard enough for humans. The "Trolley Problem" is a popular thought experiment in which one is operating a trolley on track to kill a group of people. Pulling a lever would divert the trolley into killing one person. A decision to do nothing yields a certain death toll, while a decision to divert means killing one person.

"People have a deep-seeded aversion to machines making important decisions, even when those decisions can save lives," said Kurt Gray, Associate Professor in the Department of Psychology and Neuroscience at the University of North Carolina at Chapel Hill. Gray says his work suggests "we don't want machines making these decisions."

Delegates from nine African countries were also in attendance as part of U.S. Department of State's Bureau of Educational and Cultural Affairs Exchange Programs.

Gray began by discussing his research in social and moral psychology. The field grew out of philosophy, from normative ethics. His work looks at how people make moral decisions, by asking subjects to judge the "right" answer for moral dilemmas, such as the Trolley Problem.

The nature of morality is such that people care a lot about it, according to Gray. It is also subjective. For example, people feel very strongly about abortion whether they are for or against it. This is the case with vegetarians versus meat eaters, or those opposed to removing the controversial monument to Confederate soldiers, also known as Silent Sam, and those opposed to keeping it.

This concept extends to daily value judgments. "A dog might be cute," said Gray. "Or it might be a machine that turns kibble into [excrement]."

Experience, permissibility, and agency play a big role in these judgment decisions about ethical actions. People do not think of robots as learning from personal experiences. In fact, there is the theory of the uncanny valley: As objects become more human-like, we are at first fascinated and then repulsed. [1] The actions of robots and machines is not ethically permissible and the ability for machines to make decisions is not clear. All of this has implications for autonomous vehicles in the future, said Gray.

"If robots are making moral judgments and seem to lack those experiences of mind, then maybe we don't want them doing so," observed Gray. For example: Drones are fully controlled by pilots to launch missiles. Do we want drones making those decisions? Gray examined these issues in his studies.

In the first study, Gray's team focused on assessing his subjects' sense of permissibility about robots and humans. Humans are considered more forgivable than cars. An example of this is the Autonomous Vehicle (AV) that killed a woman in Tempe, Arizona, in March 2018. [1]

Another study had subjects evaluate the parole decisions of robots versus humans. Humans were considered better. "People ascribe a lot less 'mind' to machines than to humans, which is not surprising," said Gray. Machines are seen as having less agency and less permissibility. When drones make mistakes that result in collateral damage or "friendly fire," it is less permissible, or acceptable, than when a human does this.

Another study took inventory of people's reactions to machines assisting in medical procedures, since there are more robotics in medicine these days, said Gray. The perception is that machines make worse outcomes, even when the outcome is the same.

"Machines won't feel bad if they make the wrong decision," said Gray. However, there are cases where machines are better, but what is the tipping point whereby there is acceptance of machines? Gray demonstrated through the results of another study that increasing the benefits of robotic medical procedures made a difference in subjects' opinions about it. Expert computer decision systems may be better than doctors. Even when learning that 1,000 hospital beds were saved, people were at 50% on whether a doctor or the robots should treat patients. Some of the reactions included the following:

- "I trust computers more than humans."
- "Emotion should not be avoided no matter how much money it could save."
- "I had a hard time considering what was more important, Ultimately, I felt the number I came to, 10, was a good number because eventually it outweighed the doctor."

Other studies attempted to assess what changes need to be made for more acceptance of machine learning and artificial intelligence for decision-making. Gray found the following concepts:

- Increasing Expertise: “Give the robot some experience.”
- Increasing Emotional Capacity: “Teach the robot to apologize.”
- Restricting to Advisory Role: “You're not actually letting the machine make the decision.”

Gray said he believes the aversion will surely fade. As more functions become automated by machines, the uncanny valley is no longer as steep. [2] This change will happen gradually. Currently we are proximal agents, which means we are closer to our actions. Distal is the opposite: Our ability to take action, or our agency, is further away when robots or machines are taking the actions. [3]

In ethics this comes to play when we ask whom is responsible when a child kills someone? Do we blame the child, or do we blame parents, neighborhood, or teachers?

Gray then overlaid this subject with how machines make decisions. It is a three-step process.

- Sense: Appreciate the meaning of stimuli. When we think of humans sensing the world, we mean humans making sense of their experience. For robots, this is running a regression model.
- Plan: Think ahead to achieve a goal. Machines have a better understanding when it comes to making decisions on steps to take. Machines playing games is an example of this.
- Act: Freely implement plans into the world. We measure a machine’s ability to execute decision by the degrees of freedom to action. Highly flexible machines are held more accountable for decisions they make.

The big question, according to Gray, is how do we line this up with moral decision making?

References:

[1] Mori, M., MacDorman, K., Kageki, N. 2012. "The Uncanny Valley." *Institute of Electrical and Electronics Engineers Spectrum*. pp. 98-100. June 12, 2012. Accessed December 7, 2018. https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley

[2] Wakabayashi, D. (2018). Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. *The New York Times*. Accessed December 11, 2018: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[3] Osman, Magda. *Future-Minded: The Psychology of Agency and Control.* 2014. St. Martin's Press. 240 pages. Pages 160-162.