Deloitte.

Webinar

Society in the loop artificial intelligence: Autonomous vehicles and beyond

Audience questions answered

Deloitte recently hosted a webinar about the societal implications of artificial intelligence (AI), featuring Iyad Rahwan, associate professor of media arts and sciences at the MIT Media Lab. Jim Guszcza, Deloitte Consulting LLP's US chief data scientist, served as host. The discussion focused on the need to reflect societal values and norms in the design of AI algorithms, as well as ideas for crowdsourcing societal expectations. Given the rapid pace of AI technological development, gaining societal consensus on the ethics of AI is a very timely topic.

The audience posed some interesting questions during the discussion. Here's a glimpse into them.

What is human-machine collective intelligence?

Put simply, human-machine collective intelligence combines the "intelligence" of humans and algorithms in ways that can enable better predictions, judgments, and decisions than are possible by relying on just one method or the other. Each have their own strengths, which counterbalance the weaknesses of the other. Therefore they can think better together. Freestyle Chess, famously discussed by Garry Kasparov, is a classic illustration. Two amateur players using three ordinary computers loaded with chess-playing software were able to beat grandmaster-level players and more powerful computers. "Kasparov's Law" states that: Weak human + machine + better process is superior to a strong human + machine + inferior process.

This notion of humans and machines counterbalancing one another's strengths and weaknesses is relevant when considering autonomous vehicles and ethical decisions on the road. On the one hand, computer algorithms make consistent decisions, are unaffected by emotions, and do not get tired. On the other hand, humans have the ability to make decisions under conditions that change rapidly, without warning. According to Thomas Malone of MIT, who has done extensive work in the area, the combined human-computer approach may be especially useful in situations where patterns are difficult to discern, where data are difficult to codify, or where sudden changes occur unexpectedly. For example, we still have pilots overseeing flights despite the widespread use of auto-pilot technology.

Who should decide AI ethics? Government, corporations, consumers?

All of the above need to be part of the dialogue. The spirit of Rahwan's research at the MIT Media Lab is that societal values must be researched using empirical methods, and reflected in the design of AI applications. For example, at moralmachine.mit.edu, anyone can express his or her opinion on how autonomous vehicles should make decisions in a variety of morally fraught scenarios. This yields a crowdsourced view of values that are important in societies around the globe. This type of research can help articulate what Rahwan calls an "Algorithmic Social Contract" that can help ensure that the behavior of autonomous systems reflects societal values. Given the rapid pace at which AI technologies are reshaping societal landscapes, the need to articulate algorithmic social contracts is both practical and pressing.

Will AI have different ethical rules in different countries, analogous to varying tax laws today? Or will we need a global standard in place?

This notion presents an interesting question as there are some global standards that work somewhat uniformly on an international level — such as human rights and environmental agreements. However, there are other sets of rules that vary from country to country, such as privacy rights. In the case of AI, it is likely that there will be both general societal contracts that are global in nature and others that will vary from country to country. Rahwan's research has already started to identify how judgments pertaining to specific ethical scenarios vary from country to country.

Can AI be trained to model itself off real human decisions in like situations?

Not only can AI model itself off real human decisions, this is largely how AI works today. For example, radiologists learn how to identify cancer cells by reviewing x-rays over and over again. A machine learns much the same way. Once a human has labeled the x-rays with cancerous cells, the machine uses these original human inputs to identify occurrences of these patterns on future x-rays. The algorithm's accuracy improves with the more x-rays it reviews. Machines can analyze massive amounts of data more rapidly and accurately than humans. This can help us make similar future judgments and decisions with greater accuracy and less bias.

Could AI produce a better ethical decision making process than humans?

AI can help humans make better ethical decisions, but not replace human judgment altogether. Machines and human minds each have their relative strengths and weaknesses. For example, machines don't have common sense, creativity, imagination, or intuition. However, machines are far better at logical reasoning, speed, and consistency. Thanks to these strengths, AI can help humans make better decisions. But AI cannot evaluate context, make original ethical judgments, or weigh important pieces of information not represented by the data originally used to train the algorithms being used. In such applications as hiring, medical, or judicial decisions, humans will need to remain in the loop. With that said, using machines to augment — rather than automate — human judgments and decisions can enable them to be made more ethically, consistently, and accurately.

How will errors be identified for human review?

The design of new AI systems should include ways to identify and correct erroneous outputs in order for the systems to "learn" and improve. Society also needs to be disciplined about how to use AI: in order to correctly interpret and use the output of an algorithm, people must be given a general understanding of the algorithm's assumptions, premises, and methods for arriving at its conclusions. This understanding can allow humans to understand why certain results are produced, and ultimately know when to override its decisions. Not all overrides are due to mistakes. Humans have access to important information that the algorithm is not privy to. For example, a machine learning algorithm may see bad credit indicators, but a human adjuster might understand that a person's identity was stolen and that they actually have exemplary credit once the illegal activity has been filtered out. People should be equipped with the knowledge and ability to recognize and resolve algorithmic biases and errors in order for them to be able to collaborate with machines in ways that reflect societal values.

You can view the **full webinar** and a summary of its contents on our **Vitamin D blog** as well as **subscribe** to receive invitations to future webinars in this series.