As Stakes Rise, Humans Expect AI Lies

An Army Research Lab study found a gap in the practical application of artificial intelligence in that people don’t trust it.

Kevin McCaney

Wed, 02/20/2019 - 15:42

Photo credit: Jackie Niam/iStock

A key element in government plans for artificial intelligence is building trust between humans and machines so they can work together as teams. And while a lot of the focus, understandably, has been on whether AI systems can prove trustworthiness to their human counterparts, research also has found situations where humans can be a weak spot in the equation.

Trust is seen as essential to the future of human-machine teaming, whether that job is human resources, medical diagnosis, emergency response or military applications involving drones and weapons systems. As the stakes get higher, trust becomes more important, but those might be the situations where it’s most lacking, particularly when humans are already familiar with the task at hand.

Recent tests by the Army Research Laboratory, for instance, found that people were more likely to trust an AI’s recommendations when they didn’t know what the AI was talking about, while participants who were well versed in the scenario tended to disregard the AI’s advice. In that case, familiarity bred, if not outright contempt, at least a level of indifference.

The same recommendation process that works so well in the commercial sector — for choosing a restaurant or getting directions — doesn’t seem to apply in military settings, ARL scientist Dr. James Schaffer, said in an Army report. "Unfortunately, there are big gaps between the assumptions in these low-risk domains and military practice," Schaffer said.

Trusting the Force Isn’t Enough

ARL researchers, working with a team from the University of California, Santa Barbara, created a variation of the Iterated Prisoner’s Dilemma game. During the tests, players could turn on the AI system, which would appear next to the game interface, and then decide whether to take its advice. There were different versions of the AI, including one that always gave the optimal advice based on the situation, some that were inaccurate, some that required game information to be entered manually and some that gave rational arguments for their suggestions.

In the original Prisoner’s Dilemma — created at the RAND Corp. in 1950 and applied since to sociological and biological sciences — two prisoners weigh the pros and cons of sticking together or agreeing to testify against each other. The iterated game continues through a series of rounds, with players learning from past events.

But in ARL’s scenarios, the AI agent, which would assess the situation and suggest a course of action, was often left out of the loop. Players who were the most familiar with the game tended to disregard the AI, choosing instead to trust their own knowledge. The results weren’t promising. Turning off his computer and trusting his instincts may have worked out for Luke Skywalker against the Death Star, but the Force wasn’t completely with the participants in ARL’s game. When the more knowledgeable players ignored the AI’s advice — in some cases not even bothering to check with it — they performed poorly. Novice players who consulted the AI and took its advice actually did better.

"This might be a harmless outcome if these players were really doing better, but they were in fact performing significantly worse than their humbler peers, who reported knowing less about the game beforehand," Schaffer said. "When the AI attempted to justify its suggestions to players who reported high familiarity with the game, reduced awareness of gameplay elements was observed, a symptom of over-trusting and complacency."

The results also raised another question about how users see themselves in relation to AI. In a post-game questionnaire, the players who did poorly — those who were loath to heed the AI’s rational justifications — were the most likely to say they trusted AI.

"This contrasts sharply with their observed neglect of the AI's suggestions,” Schaffer said, “demonstrating that people are not always honest, or may not always be aware of their own behavior."

The research reveals another element to consider in working with complex AI systems, which, for all their impressive capabilities, often remain inscrutable. And it showed where humans can have blind spots as well. "Rational arguments have been demonstrated to be ineffective on some people, so designers may need to be more creative in designing interfaces for these systems," Schaffer said.

Solving AI’s Riddles

But the biggest gaps in building a trust factor still lie with the machines. Among the key factors is establishing that an AI's programming is unbiased, which could avoid missteps in processes such as facial recognition. Ensuring that systems are robust enough to avoid cyber exploits such as Trojans is also critical.

One problem in establishing trust is AI's current inability to explain itself. Using complex algorithms and layered processes to ingest and analyze vast amounts of data doesn't lend itself to an easy debriefing or explanation, and machines currently can't discuss, in easily understood human terms, how they reached a particular conclusion.

AI researchers in industry, academia and government have been working on getting AI systems to describe their thought processes in several ways. The Open AI research consortium, for instance, is studying how the systems think by having them debate each other. The Defense Advanced Research Projects Agency's Explainable Artificial Intelligence (XAI) project is looking to find ways an AI could use natural language in recounting its processes.

A new DARPA project is looking for machines to offer a kind of running commentary on their processes, rather than waiting until after a conclusion is reached. The Competency-Aware Machine Learning program aims to have machines continuously assess their performance and keep humans apprised.

"If the machine can say, 'I do well in these conditions, but I don't have a lot of experience in those conditions,' that will allow a better human-machine teaming," said Jiangying Zhou, a program manager in DARPA's Defense Sciences Office. "The partner then can make a more informed choice."

Zhou offered a simple analogy involving someone deciding which of two a self-driving cars would do better driving at night in the rain. The first car says it can distinguish between a pedestrian and an inanimate object 90 percent of the time, having tested itself 1,000 times. The second car might claim 99 percent accuracy, but disclose that it has tried the task less than 100 times. The rider can then make a more informed choice.

That’s a basic example (and seems like a pretty tough call), but it describes the process researchers are going for. “Under what conditions do you let the machine do its job? Under what conditions should you put supervision on it?” Zhou said. “Which assets, or combination of assets, are best for your task?”

The ultimate goal with CAML and other research projects, inside and outside government, is building trust between humans and machines. But the recent Prisoner’s Dilemma exercise proves that we still have a long way to go.

View printer friendly version

AI
machine learning
DARPA
Army Research Lab