When should someone trust an AI assistant's predictions?

January 19 2022, by Adam Zewe

Researchers have created a method to help workers collaborate with artificial intelligence systems. Credit: Christine Daniloff, MIT

In a busy hospital, a radiologist is using an artificial intelligence system to help her diagnose medical conditions based on patients' X-ray images.

Using the AI system can help her make faster diagnoses, but how does she know when to trust the AI's predictions?

She doesn't. Instead, she may rely on her expertise, a confidence level provided by the system itself, or an explanation of how the algorithm made its prediction—which may look convincing but still be wrong—to make an estimation.

To help people better understand when to trust an AI "teammate," MIT researchers created an onboarding technique that guides humans to develop a more accurate understanding of those situations in which a machine makes correct predictions and those in which it makes incorrect predictions.

By showing people how the AI complements their abilities, the training technique could help humans make better decisions or come to conclusions faster when working with AI agents.

"We propose a teaching phase where we gradually introduce the human to this AI model so they can, for themselves, see its weaknesses and strengths," says Hussein Mozannar, a graduate student in the Clinical Machine Learning Group of the Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Institute for Medical Engineering and Science. "We do this by mimicking the way the human will interact with the AI in practice, but we intervene to give them feedback to help them understand each interaction they are making with the AI."

Mozannar wrote the paper with Arvind Satyanarayan, an assistant professor of computer science who leads the Visualization Group in CSAIL; and senior author David Sontag, an associate professor of electrical engineering and computer science at MIT and leader of the Clinical Machine Learning Group. The research will be presented at the

Association for the Advancement of Artificial Intelligence in February.

Mental models

This work focuses on the mental models humans build about others. If the radiologist is not sure about a case, she may ask a colleague who is an expert in a certain area. From past experience and her knowledge of this colleague, she has a mental model of his strengths and weaknesses that she uses to assess his advice.

Humans build the same kinds of mental models when they interact with AI agents, so it is important those models are accurate, Mozannar says. Cognitive science suggests that humans make decisions for complex tasks by remembering past interactions and experiences. So, the researchers designed an onboarding process that provides representative examples of the human and AI working together, which serve as reference points the human can draw on in the future. They began by creating an algorithm that can identify examples that will best teach the human about the AI.

"We first learn a human expert's biases and strengths, using observations of their past decisions unguided by AI," Mozannar says. "We combine our knowledge about the human with what we know about the AI to see where it will be helpful for the human to rely on the AI. Then we obtain cases where we know the human should rely on the AI and similar cases where the human should not rely on the AI."

The researchers tested their onboarding technique on a passage-based question answering task: The user receives a written passage and a question whose answer is contained in the passage. The user then has to answer the question and can click a button to "let the AI answer." The user can't see the AI answer in advance, however, requiring them to rely on their mental model of the AI. The onboarding process they developed

begins by showing these examples to the user, who tries to make a prediction with the help of the AI system. The human may be right or wrong, and the AI may be right or wrong, but in either case, after solving the example, the user sees the correct answer and an explanation for why the AI chose its prediction. To help the user generalize from the example, two contrasting examples are shown that explain why the AI got it right or wrong.

For instance, perhaps the training question asks which of two plants is native to more continents, based on a convoluted paragraph from a botany textbook. The human can answer on her own or let the AI system answer. Then, she sees two follow-up examples that help her get a better sense of the AI's abilities. Perhaps the AI is wrong on a follow-up question about fruits but right on a question about geology. In each example, the words the system used to make its prediction are highlighted. Seeing the highlighted words helps the human understand the limits of the AI agent, explains Mozannar.

To help the user retain what they have learned, the user then writes down the rule she infers from this teaching example, such as "This AI is not good at predicting flowers." She can then refer to these rules later when working with the agent in practice. These rules also constitute a formalization of the user's mental model of the AI.

The impact of teaching

The researchers tested this teaching technique with three groups of participants. One group went through the entire onboarding technique, another group did not receive the follow-up comparison examples, and the baseline group didn't receive any teaching but could see the AI's answer in advance.

"The participants who received teaching did just as well as the

participants who didn't receive teaching but could see the AI's answer. So, the conclusion there is they are able to simulate the AI's answer as well as if they had seen it," Mozannar says.

The researchers dug deeper into the data to see the rules individual participants wrote. They found that almost 50 percent of the people who received training wrote accurate lessons of the AI's abilities. Those who had accurate lessons were right on 63 percent of the examples, whereas those who didn't have accurate lessons were right on 54 percent. And those who didn't receive teaching but could see the AI answers were right on 57 percent of the questions.

"When teaching is successful, it has a significant impact. That is the takeaway here. When we are able to teach participants effectively, they are able to do better than if you actually gave them the answer," he says.

But the results also show there is still a gap. Only 50 percent of those who were trained built accurate mental models of the AI, and even those who did were only right 63 percent of the time. Even though they learned accurate lessons, they didn't always follow their own rules, Mozannar says.

That is one question that leaves the researchers scratching their heads—even if people know the AI should be right, why won't they listen to their own mental model? They want to explore this question in the future, as well as refine the onboarding process to reduce the amount of time it takes. They are also interested in running user studies with more complex AI models, particularly in health care settings.

More information: Hussein Mozannar, Arvind Satyanarayan, David Sontag, Teaching Humans When To Defer to a Classifier via Exemplars. arXiv:2111.11297v2 [cs.LG], arxiv.org/abs/2111.11297