Artificial intelligence can offset human frailties, leading to better decisions

12 May 2022

Credit: Anton Grabolle / Better Images of AI / Human-AI collaboration / CC-BY 4.0

Modern life can be full of baffling encounters with artificial intelligence—think misunderstandings with customer service chatbots or algorithmically misplaced hair metal in your Spotify playlist. These AI systems can't effectively work with people because they have no idea that humans can behave in seemingly irrational ways, says Mustafa Mert Çelikok. He's a Ph.D. student studying human-AI interaction, with the idea of taking the strengths and weaknesses of both sides and blending them into a superior decision-maker.

In the AI world, one example of such a hybrid is a "centaur." It's not a mythological horse-human, but a human-AI team. Centaurs appeared in chess in the late 1990s, when artificial intelligence systems became advanced enough to beat human champions. In place of a "human versus machine" matchup, centaur or cyborg chess involves one or more computer chess programs and human players on each side.

"This is the Formula 1 of chess," says Çelikok. "Grandmasters have been defeated. Super AIs have been defeated. And grandmasters playing with powerful AIs have also lost." As it turns out, novice players paired with AIs are the most successful. "Novices don't have strong opinions" and can form effective decision-making partnerships with their AI teammates, while "grandmasters think they know better than AIs and override them when they disagree—that's their downfall," observes Çelikok.

In a game like chess, there are defined rules and a clear goal that humans and AIs share. But in the world of online shopping, playlists or any other service where a human encounters an algorithm, there may be no shared goal, or the goal might be poorly defined, at least from the AI perspective. Çelikok is trying to fix this by including actual information about human behavior so that multi-agent systems—centaur-like partnerships of people and AIs—can understand each other and make better decisions.

"The 'human' in human-AI interaction hasn't been explored much," says Çelikok. "Researchers don't use any models of human behavior, but what we're doing is explicitly using human cognitive science. We're not trying to replace humans or teach AIs to do a task. Instead, we want AIs to help people make better decisions." In the case of Çelikok's latest study, this means helping people eat healthier.

In the experimental simulation, a person is browsing food trucks, trying to decide where to eat, with the help of their trusty AI-powered autonomous vehicle. The car knows the passenger prefers healthy vegetarian food over unhealthy donuts. With this criterion in mind, the AI car would choose to take the shortest path to the vegetarian food

truck. This simple solution can backfire, though. If the shortest path goes by the donut shop, the passenger may take the wheel, overriding the AI. This apparent human irrationality conflicts with the most logical solution.

Çelikok's model uniquely avoids this problem by helping the AI figure out that humans are time-inconsistent. "If you ask people, do you want 10 dollars right now or 20 tomorrow, and they choose 10 now, but then you ask again, do you want 10 dollars in 100 days or 20 in 101 days, and they choose 20, that is inconsistent," he explains. "The gap is not treated the same. That is what we mean by time-inconsistent, and a typical AI does not take into account non-rationality or time-inconsistent preferences, for example procrastination, changing preferences on the fly or the temptation of donuts." In Çelikok's research, the AI car will figure out that taking a slightly longer route will bypass the donut shop, leading to a healthier outcome for the passenger.

"AI has unique strengths and weaknesses, and people do also," says Çelikok. "The human weakness is irrational behaviors and time-inconsistency, which AI can fix and complement." On the other hand, if there is a situation where the AI is wrong and the human right, the AI will learn to behave according to the human preference when overridden. This is another side result of Çelikok's mathematical modeling.

Combining models of human cognition with statistics allows AI systems to figure out how people behave faster, says Çelikok. It's also more efficient. Compared to training an AI system with thousands of images to learn visual recognition, interacting with people is slow and expensive, because learning just one person's preferences can take a long time. Çelikok again makes a comparison to chess: a human novice or an AI system can both understand the rules and physical moves, but they may both struggle to understand the complex intentions of a grandmaster. Çelikok's research is finding the balance between the optimal moves and the intuitive ones, building a real-life centaur with math.

More information: Mustafa Mert Çelikok, Frans A. Oliehoek, Samuel Kaski, Best-Response Bayesian Reinforcement Learning with Bayes-adaptive POMDPs for Centaurs. arXiv:2204.01160v1 [cs.AI], arxiv.org/abs/2204.01160

Provided by Aalto University