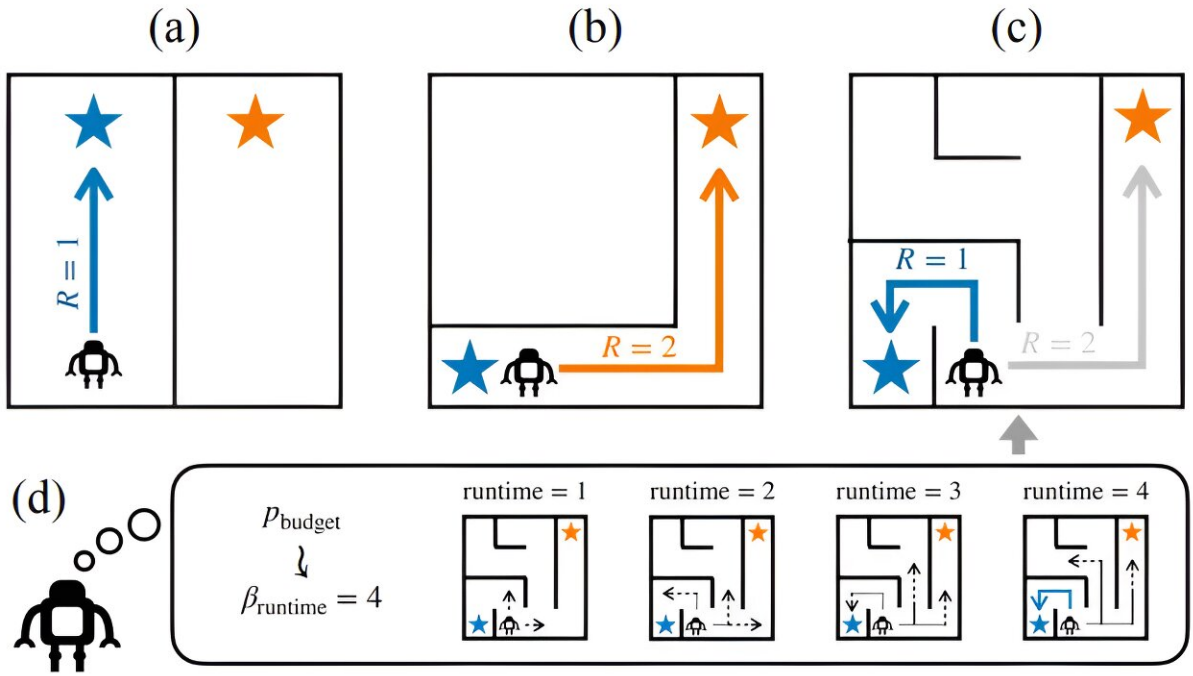
To build a better AI helper, start by modeling the irrational behavior of humans

April 19 2024, by Adam Zewe



Inferring rewards from boundedly-rational trajectories. The agent will move to the blue star (a), but prefers to move toward the orange star when both are available (b). When locating the orange star requires solving a harder search problem, however, the agent seeks the blue star instead, indicating that its search abilities are limited (c). Our proposed approach automatically infers the budget that the agent uses when planning (d). Knowing this budget, we could perhaps assist this agent by providing a targeted hint (move right) at the beginning of its trajectory. Credit: https://openreview.net/pdf?id=W3VsHuga3j

To build AI systems that can collaborate effectively with humans, it helps to have a good model of human behavior to start with. But humans tend to behave suboptimally when making decisions.

This irrationality, which is especially difficult to model, often boils down to computational constraints. A human can't spend decades thinking about the ideal solution to a single problem.

Researchers at MIT and the University of Washington developed a way to model the behavior of an agent, whether human or machine, that accounts for the unknown computational constraints that may hamper the agent's problem-solving abilities.

Their model can automatically infer an agent's computational constraints by seeing just a few traces of their previous actions. The result, an agent's so-called "inference budget," can be used to predict that agent's future behavior.

In a new paper, the researchers demonstrate how their method can be used to infer someone's navigation goals from prior routes and to predict players' subsequent moves in chess matches. Their technique matches or outperforms another popular method for modeling this type of decision-making.

Ultimately, this work could help scientists teach AI systems how humans behave, which could enable these systems to respond better to their human collaborators. Being able to understand a human's behavior, and then to infer their goals from that behavior, could make an AI assistant much more useful, says Athul Paul Jacob, an electrical engineering and computer science (EECS) graduate student and lead author of the paper on this technique.

"If we know that a human is about to make a mistake, having seen how

they have behaved before, the AI agent could step in and offer a better way to do it. Or the agent could adapt to the weaknesses that its human collaborators have. Being able to model human behavior is an important step toward building an AI agent that can actually help that human," he says.

Jacob wrote the paper with Abhishek Gupta, assistant professor at the University of Washington, and senior author Jacob Andreas, associate professor in EECS and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL). The research will be presented at the International Conference on Learning Representations (ICLR 2024), held in Vienna, Austria, May7–11.

Modeling behavior

Researchers have been building computational models of human behavior for decades. Many prior approaches try to account for suboptimal decision-making by adding noise to the model. Instead of the agent always choosing the correct option, the model might have that agent make the correct choice 95% of the time.

However, these methods can fail to capture the fact that humans do not always behave suboptimally in the same way.

Others at MIT have also studied more effective ways to plan and infer goals in the face of suboptimal decision-making.

To build their model, Jacob and his collaborators drew inspiration from prior studies of chess players. They noticed that players took less time to think before acting when making simple moves and that stronger players tended to spend more time planning than weaker ones in challenging matches.

"At the end of the day, we saw that the depth of the planning, or how long someone thinks about the problem, is a really good proxy of how humans behave," Jacob says.

They built a framework that could infer an agent's depth of planning from prior actions and use that information to model the agent's decision-making process.

The first step in their method involves running an algorithm for a set amount of time to solve the problem being studied. For instance, if they are studying a chess match, they might let the chess-playing algorithm run for a certain number of steps. At the end, the researchers can see the decisions the algorithm made at each step.

Their model compares these decisions to the behaviors of an agent solving the same problem. It will align the agent's decisions with the algorithm's decisions and identify the step where the agent stopped planning.

From this, the model can determine the agent's inference budget, or how long that agent will plan for this problem. It can use the inference budget to predict how that agent would react when solving a similar problem.

An interpretable solution

This method can be very efficient because the researchers can access the full set of decisions made by the problem-solving algorithm without doing any extra work. This framework could also be applied to any problem that can be solved with a particular class of algorithms.

"For me, the most striking thing was the fact that this inference budget is very interpretable. It is saying tougher problems require more planning or being a strong player means planning for longer. When we first set out

to do this, we didn't think that our algorithm would be able to pick up on those behaviors naturally," Jacob says.

The researchers tested their approach in three different modeling tasks: inferring navigation goals from previous routes, guessing someone's communicative intent from their verbal cues, and predicting subsequent moves in human-human chess matches.

Their method either matched or outperformed a popular alternative in each experiment. Moreover, the researchers saw that their model of human behavior matched up well with measures of player skill (in chess matches) and task difficulty.

Moving forward, the researchers want to use this approach to model the planning process in other domains, such as reinforcement learning (a trial-and-error method commonly used in robotics). In the long run, they intend to keep building on this work toward the larger goal of developing more effective AI collaborators.

More information: Modeling Boundedly Rational Agents With Latent Inference Budgets. openreview.net/pdf?id=W3VsHuga3j

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

behavior-humans.html