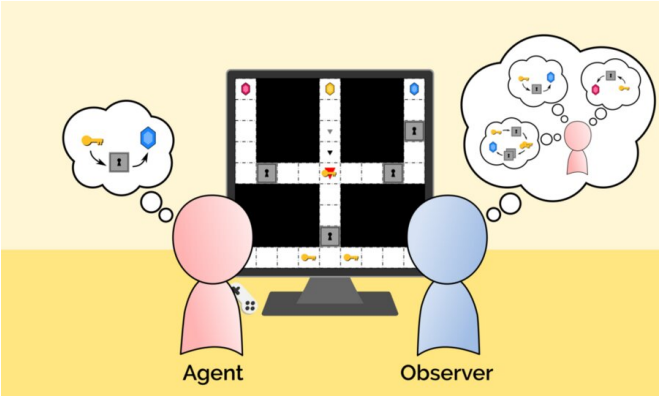
Building machines that better understand human goals

15 December 2020



An “agent” and an “observer” demonstrate how a new MIT algorithm is capable of inferring goals and plans, even when those plans might fail. Here, the agent makes a mistaken plan to reach the blue gem, which the observer infers as a possibility. Credit: Massachusetts Institute of Technology

In a classic experiment on human social intelligence by psychologists Felix Warneken and Michael Tomasello, an 18-month old toddler watches a man carry a stack of books towards an unopened cabinet. When the man reaches the cabinet, he clumsily bangs the books against the door of the cabinet several times, then makes a puzzled noise.

Something remarkable happens next: the toddler offers to help. Having inferred the man's goal, the toddler walks up to the cabinet and opens its doors, allowing the man to place his books inside. But how is the toddler, with such limited life experience, able to make this inference?

Recently, computer scientists have redirected this question toward computers: How can machines do the same?

The critical component to engineering this type of understanding is arguably what makes us most human: our mistakes. Just as the toddler could infer the man's goal merely from his failure, machines that infer our goals need to account for our mistaken actions and plans.

In the quest to capture this social intelligence in machines, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Department of Brain and Cognitive Sciences created an algorithm capable of inferring goals and plans, even when those plans might fail.

This type of research could eventually be used to improve a range of assistive technologies, collaborative or caretaking robots, and digital assistants like Siri and Alexa.

"This ability to account for mistakes could be crucial for building machines that robustly infer and act in our interests," says Tan Zhi-Xuan, Ph.D. student in MIT's Department of Electrical Engineering and Computer Science (EECS) and the lead author on a new paper about the research. "Otherwise, AI systems might wrongly infer that, since we failed to achieve our higher-order goals, those goals weren't desired after all. We've seen what happens when algorithms feed on our reflexive and unplanned usage of social media, leading us down paths of dependency and polarization. Ideally, the algorithms of the future will recognize our mistakes, bad habits, and irrationalities and help us avoid, rather than reinforce, them."

To create their model the team used Gen, a new AI programming platform recently developed at MIT, to combine symbolic AI planning with Bayesian inference. Bayesian inference provides an optimal way to combine uncertain beliefs with new data, and is widely used for financial risk evaluation, diagnostic testing, and election forecasting.

The team's model performed 20 to 150 times faster than an existing baseline method called Bayesian

Inverse Reinforcement Learning (BIRL), which learns an agent's objectives, values, or rewards by observing its behavior, and attempts to compute full policies or plans in advance. The new model was accurate 75 percent of the time in inferring goals.

"AI is in the process of abandoning the 'standard model' where a fixed, known objective is given to the machine," says Stuart Russell, the Smith-Zadeh Professor of Engineering at the University of California at Berkeley. "Instead, the machine knows that it doesn't know what we want, which means that research on how to infer goals and preferences from human behavior becomes a central topic in AI. This paper takes that goal seriously; in particular, it is a step towards modeling—and hence inverting—the actual process by which humans generate behavior from goals and preferences."

How it works

While there's been considerable work on inferring the goals and desires of agents, much of this work has assumed that agents act optimally to achieve their goals.

However, the team was particularly inspired by a common way of human planning that's largely sub-optimal: not to plan everything out in advance, but rather to form only partial plans, execute them, and then plan again from there. While this can lead to mistakes from not thinking enough "ahead of time," it also reduces the cognitive load.

For example, imagine you're watching your friend prepare food, and you would like to help by figuring out what they're cooking. You guess the next few steps your friend might take: maybe preheating the oven, then making dough for an apple pie. You then "keep" only the partial plans that remain consistent with what your friend actually does, and then you repeat the process by planning ahead just a few steps from there.

Once you've seen your friend make the dough, you can restrict the possibilities only to baked goods, and guess that they might slice apples next, or get some pecans for a pie mix. Eventually, you'll have eliminated all the plans for dishes that your friend couldn't possibly be making, keeping only the possible plans (i.e., pie recipes). Once you're sure enough which dish it is, you can offer to help.

The team's inference algorithm, called "Sequential Inverse Plan Search (SIPS)", follows this sequence to infer an agent's goals, as it only makes partial plans at each step, and cuts unlikely plans early on. Since the model only plans a few steps ahead each time, it also accounts for the possibility that the agent—your friend—might be doing the same. This includes the possibility of mistakes due to limited planning, such as not realizing you might need two hands free before opening the refrigerator. By detecting these potential failures in advance, the team hopes the model could be used by machines to better offer assistance.

"One of our early insights was that if you want to infer someone's goals, you don't need to think further ahead than they do. We realized this could be used not just to speed up goal inference, but also to infer intended goals from actions that are too shortsighted to succeed, leading us to shift from scaling up algorithms to exploring ways to resolve more fundamental limitations of current AI systems," says Vikash Mansinghka, a principal research scientist at MIT and one of Tan Zhi-Xuan's co-advisors, along with Joshua Tenenbaum, MIT professor in brain and cognitive sciences. "This is part of our larger moonshot—to reverse-engineer 18-month-old human common sense."

The work builds conceptually on earlier cognitive models from Tenenbaum's group, showing how simpler inferences that children and even 10-month-old infants make about others' goals can be modeled quantitatively as a form of Bayesian inverse planning.

While to date the researchers have explored inference only in relatively small planning problems over fixed sets of goals, through future work they plan to explore richer hierarchies of human goals and plans. By encoding or learning these hierarchies, machines might be able to infer a much wider variety of goals, as well as the deeper purposes they serve.

"Though this work represents only a small initial step, my hope is that this research will lay some of

the philosophical and conceptual groundwork necessary to build machines that truly understand human goals, plans and values," says Xuan. "This basic approach of modeling humans as imperfect reasoners feels very promising. It now allows us to infer when plans are mistaken, and perhaps it will eventually allow us to infer when people hold mistaken beliefs, assumptions, and guiding principles as well."

More information: Online Bayesian Goal Inference for Boundedly-Rational Planning Agents. arxiv.org/abs/2006.07532

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