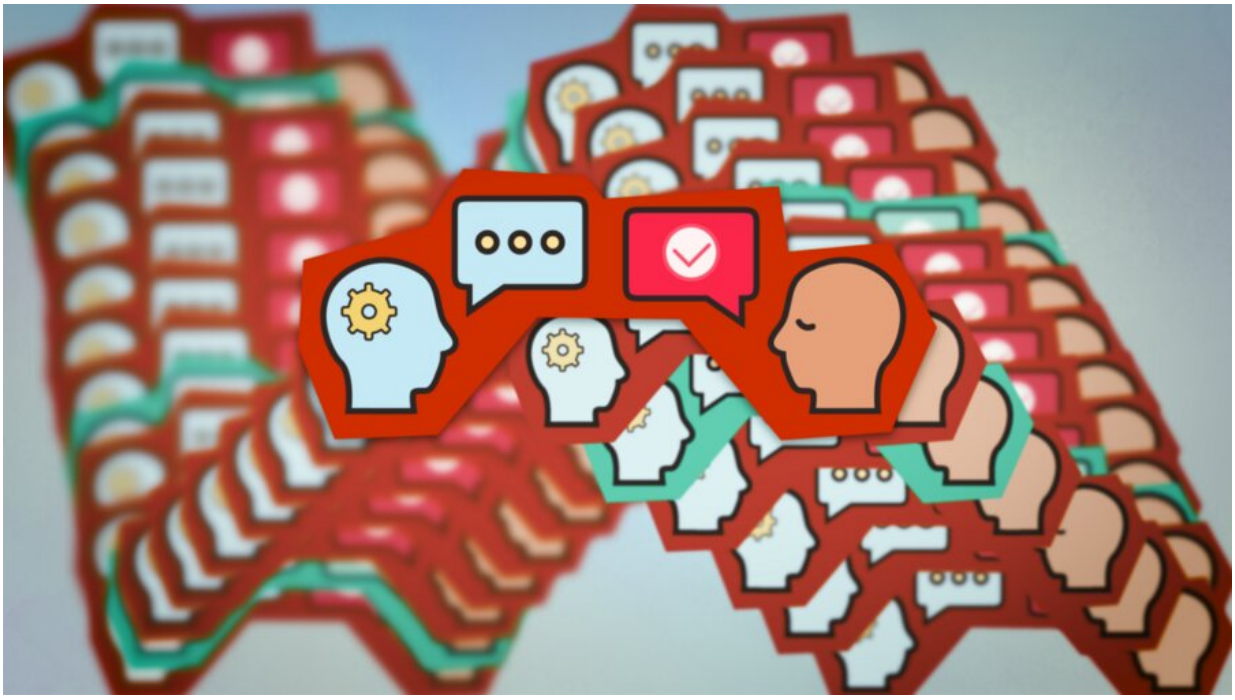
A method to interpret AI might not be so interpretable after all

October 16 2023, by Kylie Foy



A study finds humans struggle to understand the outputs of formal specifications, a method that some researchers claim can be used to make AI decision-making interpretable to humans. Credit: Bryan Mastergeorge, Massachusetts Institute of Technology

As autonomous systems and artificial intelligence become increasingly common in daily life, new methods are emerging to help humans check that these systems are behaving as expected. One method, called formal

specifications, uses mathematical formulas that can be translated into natural-language expressions. Some researchers claim that this method can be used to spell out decisions an AI will make in a way that is interpretable to humans.

MIT Lincoln Laboratory researchers wanted to check such claims of interpretability. Their findings point to the opposite: Formal specifications do not seem to be interpretable by humans. In the team's study, participants were asked to check whether an AI agent's plan would succeed in a virtual game. Presented with the formal specification of the plan, the participants were correct less than half of the time.

"The results are bad news for researchers who have been claiming that formal methods lent interpretability to systems. It might be true in some restricted and abstract sense, but not for anything close to practical system validation," says Hosea Siu, a researcher in the laboratory's AI Technology Group. The group's paper, currently available on the *arXiv* preprint server, was accepted to the 2023 International Conference on Intelligent Robots and Systems held earlier this month.

Interpretability is important because it allows humans to place trust in a machine when used in the real world. If a robot or AI can explain its actions, then humans can decide whether it needs adjustments or can be trusted to make fair decisions. An interpretable system also enables the users of technology—not just the developers—to understand and trust its capabilities. However, interpretability has long been a challenge in the field of AI and autonomy. The machine learning process happens in a "black box," so model developers often can't explain why or how a system came to a certain decision.

"When researchers say 'our machine learning system is accurate,' we ask 'how accurate?" and 'using what data?' and if that information isn't provided, we reject the claim. We haven't been doing that much when

researchers say 'our machine learning system is interpretable,' and we need to start holding those claims up to more scrutiny," Siu says.

Lost in translation

For their experiment, the researchers sought to determine whether formal specifications made the behavior of a system more interpretable. They focused on people's ability to use such specifications to validate a system—that is, to understand whether the system always met the user's goals.

Applying formal specifications for this purpose is essentially a by-product of its original use. Formal specifications are part of a broader set of formal methods that use logical expressions as a mathematical framework to describe the behavior of a model. Because the model is built on a logical flow, engineers can use "model checkers" to mathematically prove facts about the system, including when it is or isn't possible for the system to complete a task. Now, researchers are trying to use this same framework as a translational tool for humans.

"Researchers confuse the fact that formal specifications have precise semantics with them being interpretable to humans. These are not the same thing," Siu says. "We realized that next-to-nobody was checking to see if people actually understood the outputs."

In the team's experiment, participants were asked to validate a fairly simple set of behaviors with a robot playing a game of capture the flag, basically answering the question "If the robot follows these rules exactly, does it always win?"

Participants included both experts and nonexperts in formal methods. They received the formal specifications in three ways—a "raw" logical formula, the formula translated into words closer to natural language,

and a decision-tree format. Decision trees in particular are often considered in the AI world to be a human-interpretable way to show AI or robot decision-making.

The results: "Validation performance on the whole was quite terrible, with around 45 percent accuracy, regardless of the presentation type," Siu says.

Confidently wrong

Those previously trained in formal specifications only did slightly better than novices. However, the experts reported far more confidence in their answers, regardless of whether they were correct or not. Across the board, people tended to over-trust the correctness of specifications put in front of them, meaning that they ignored rule sets allowing for game losses. This confirmation bias is particularly concerning for system validation, the researchers say, because people are more likely to overlook failure modes.

"We don't think that this result means we should abandon formal specifications as a way to explain system behaviors to people. But we do think that a lot more work needs to go into the design of how they are presented to people and into the workflow in which people use them," Siu adds.

When considering why the results were so poor, Siu recognizes that even people who work on formal methods aren't quite trained to check specifications as the experiment asked them to. And, thinking through all the possible outcomes of a set of rules is difficult. Even so, the rule sets shown to participants were short, equivalent to no more than a paragraph of text, "much shorter than anything you'd encounter in any real system," Siu says.

The team isn't attempting to tie their results directly to the performance of humans in real-world robot validation. Instead, they aim to use the results as a starting point to consider what the formal logic community may be missing when claiming interpretability, and how such claims may play out in the real world.

This research was conducted as part of a larger project Siu and teammates are working on to improve the relationship between robots and human operators, especially those in the military. The process of programming robotics can often leave operators out of the loop. With a similar goal of improving interpretability and trust, the project is trying to allow operators to teach tasks to robots directly, in ways that are similar to training humans. Such a process could improve both the operator's confidence in the robot and the robot's adaptability.

Ultimately, they hope the results of this study and their ongoing research can better the application of autonomy, as it becomes more embedded in human life and decision-making.

"Our results push for the need to do human evaluations of certain systems and concepts of autonomy and AI before too many claims are made about their utility with humans," Siu adds.

More information: Ho Chit Siu et al, STL: Surprisingly Tricky Logic (for System Validation), *arXiv* (2023). DOI: 10.48550/arxiv.2305.17258