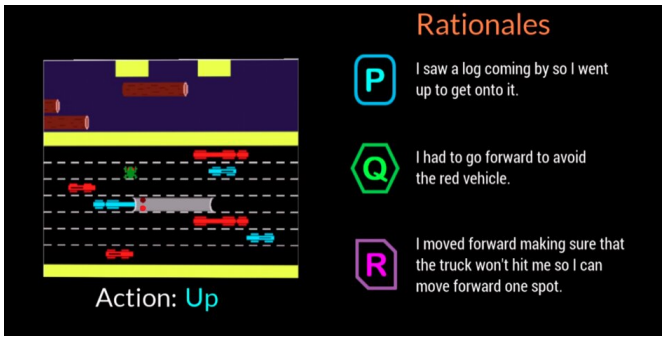
AI agent offers rationales using everyday language to explain its actions

12 April 2019



An AI agent provides its rationale for making a decision in this video game. Credit: Georgia Tech

Georgia Institute of Technology researchers, in collaboration with Cornell University and University of Kentucky, have developed an artificially intelligent (AI) agent that can automatically generate natural language explanations in real-time to convey the motivations behind its actions. The work is designed to give humans engaging with AI agents or robots confidence that the agent is performing the task correctly and can explain a mistake or errant behavior.

The agent also uses everyday language that non-experts can understand. The explanations, or "rationales" as the researchers call them, are designed to be relatable and inspire trust in those who might be in the workplace with AI machines or interact with them in social situations.

"If the power of AI is to be democratized, it needs to be accessible to anyone regardless of their technical abilities," said Upol Ehsan, Ph.D. student in the School of Interactive Computing at Georgia Tech and lead researcher.

"As AI pervades all aspects of our lives, there is a distinct need for human-centered AI design that makes black-boxed AI systems explainable to everyday users. Our work takes a formative step toward understanding the role of language-based explanations and how humans perceive them."

The study was supported by the Office of Naval Research (ONR).

Researchers developed a participant study to determine if their AI agent could offer rationales that mimicked human responses. Spectators watched the AI agent play the videogame Frogger and then ranked three on-screen rationales in order of how well each described the AI's game move.

Of the three anonymized justifications for each move—a human-generated response, the AI-agent response, and a randomly generated response—the participants preferred the human-generated rationales first, but the AI-generated responses were a close second.

Frogger offered the researchers the chance to train an AI in a "sequential decision-making environment," which is a significant research challenge because decisions that the agent has already made influence future decisions. Therefore, explaining the chain of reasoning to experts is difficult, and even more so when communicating with non-experts, according to researchers.

The human spectators understood the goal of Frogger in getting the frog safely home without being hit by moving vehicles or drowned in the river. The simple game mechanics of moving up, down, left or right, allowed the participants to see what the AI was doing, and to evaluate if the rationales on the screen clearly justified the move.

The spectators judged the rationales based on:

- Confidence—the person is confident in the AI to perform its task
- Human-likeness—looks like it was made by a human

- Adequate justification—adequately justifies the action taken
- Understandability—helps the person understand the AI's behavior

AI-generated rationales that were ranked higher by participants were those that showed recognition of environmental conditions and adaptability, as well as those that communicated awareness of upcoming dangers and planned for them. Redundant information that just stated the obvious or mischaracterized the environment were found to have a negative impact.

"This project is more about understanding human perceptions and preferences of these AI systems than it is about building new technologies," said Ehsan. "At the heart of explainability is sense making. We are trying to understand that human factor."

A second related study validated the researchers' decision to design their AI agent to be able to offer one of two distinct types of rationales:

- Concise, "focused" rationales or
- Holistic, "complete picture" rationales

In this second study, participants were only offered AI-generated rationales after watching the AI play Frogger. They were asked to select the answer that they preferred in a scenario where an AI made a mistake or behaved unexpectedly. They did not know the rationales were grouped into the two categories.

By a 3-to-1 margin, participants favored answers that were classified in the "complete picture" category. Responses showed that people appreciated the AI thinking about future steps rather than just what was in the moment, which might make them more prone to making another mistake. People also wanted to know more so that they might directly help the AI fix the errant behavior.

"The situated understanding of the perceptions and preferences of people working with AI machines give us a powerful set of actionable insights that can help us design better human-centered, rationale-generating, autonomous agents," said Mark Riedl, professor of Interactive Computing and lead faculty member on the project.

A possible future direction for the research will apply the findings to autonomous agents of various types, such as companion agents, and how they might respond based on the task at hand. Researchers will also look at how agents might respond in different scenarios, such as during an emergency response or when aiding teachers in the classroom.

The research was presented in March at the Association for Computing Machinery's Intelligent User Interfaces 2019 Conference. The paper is titled "Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions." Ehsan will present a position paper highlighting the design and evaluation challenges of human-centered Explainable AI systems at the upcoming Emerging Perspectives in Human-Centered Machine Learning workshop at the ACM CHI 2019 conference, May 4-9, in Glasgow, Scotland.

Provided by Georgia Institute of Technology

APA citation: AI agent offers rationales using everyday language to explain its actions (2019, April 12) retrieved 22 April 2019 from https://techxplore.com/news/2019-04-ai-agent-rationales-everyday-language.html