Researchers ask AI to explain itself

19 August 2020, by Chad Boutin

NIST scientists have proposed four principles for judging how explainable an artificial intelligence's decisions are. Credit: B. Hayes/NIST

It's a question that many of us encounter in childhood: "Why did you do that?" As artificial intelligence (AI) begins making more consequential decisions that affect our lives, we also want these machines to be capable of answering that simple yet profound question. After all, why else would we trust AI's decisions?

This desire for satisfactory explanations has spurred scientists at the National Institute of Standards and Technology (NIST) to propose a set of principles by which we can judge how explainable AI's decisions are. Their draft publication, Four Principles of Explainable Artificial Intelligence (Draft NISTIR 8312), is intended to stimulate a conversation about what we should expect of our decision-making devices.

The report is part of a broader NIST effort to help develop trustworthy AI systems. NIST's foundational research aims to build trust in these systems by understanding their theoretical capabilities and limitations and by improving their accuracy, reliability, security, robustness and explainability, which is the focus of this latest publication.

The authors are requesting feedback on the draft from the public—and because the subject is a broad one, touching upon fields ranging from engineering and computer science to psychology and legal studies, they are hoping for a wide-ranging discussion.

"AI is becoming involved in high-stakes decisions, and no one wants machines to make them without an understanding of why," said NIST electronic engineer Jonathon Phillips, one of the report's authors. "But an explanation that would satisfy an engineer might not work for someone with a different background. So, we want to refine the draft with a diversity of perspective and opinions."

An understanding of the reasons behind the output of an AI system can benefit everyone the output touches. If an AI contributes to a loan approval decision, for example, this understanding might help a software designer improve the system. But the applicant might want insight into the AI's reasoning as well, either to understand why she was turned down, or, if she was approved, to help her continue acting in ways that maintain her good credit rating.

According to the authors, the four principles for explainable AI are:

- AI systems should deliver accompanying evidence or reasons for all their outputs.
- Systems should provide explanations that are meaningful or understandable to individual users.
- The explanation correctly reflects the system's process for generating the output.
- The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output. (The idea is that if a system has insufficient confidence in its decision, it should not supply a decision to the user.)

While these principles are straightforward enough on the surface, Phillips said that individual users often have varied criteria for judging an AI's success at meeting them. For instance, the second principle—how meaningful the explanation is—can

imply different things to different people, depending on their role and connection to the job the AI is doing.

"Think about Kirk and Spock and how each one talks," Phillips said, referencing the Star Trek characters. "A doctor using an AI to help diagnose disease may only need Spock's explanation of why the machine recommends a particular treatment, while the patient might be OK with less technical detail but want Kirk's background on how it relates to his life."

Phillips and his co-authors align their concepts of explainable AI to relevant previous work in artificial intelligence, but they also compare the demands for explainability we place on our machines to those we place on our fellow humans. Do we measure up to the standards we are asking of AI? After exploring how human decisions hold up in light of the report's four principles, the authors conclude that—spoiler alert—we don't.

"Human-produced explanations for our own choices and conclusions are largely unreliable," they write, citing several examples. "Without conscious awareness, people incorporate irrelevant information into a variety of decisions from personality trait judgments to jury decisions."

However, our awareness of this apparent double standard could eventually help us better understand our own decisions and create a safer, more transparent world.

"As we make advances in explainable AI, we may find that certain parts of AI systems are better able to meet societal expectations and goals than humans are," said Phillips, whose past research indicates that collaborations between humans and AI can produce greater accuracy than either one working alone. "Understanding the explainability of both the AI system and the human opens the door to pursue implementations that incorporate the strengths of each."

For the moment, Phillips said, the authors hope the comments they receive advance the conversation.

"I don't think we know yet what the right benchmarks are for explainability," he said. "At the end of the day we're not trying to answer all these questions. We're trying to flesh out the field so that discussions can be fruitful."

More information: Phillips et al., Four Principles of Explainable Artificial Intelligence. (2020). nvlpubs.nist.gov/nistpubs/ir/2 … ST.IR.8312-draft.pdf

This story is republished courtesy of NIST. Read the original story here.

Provided by National Institute of Standards and Technology

APA citation: Researchers ask AI to explain itself (2020, August 19) retrieved 12 August 2022 from https://techxplore.com/news/2020-08-ai.html