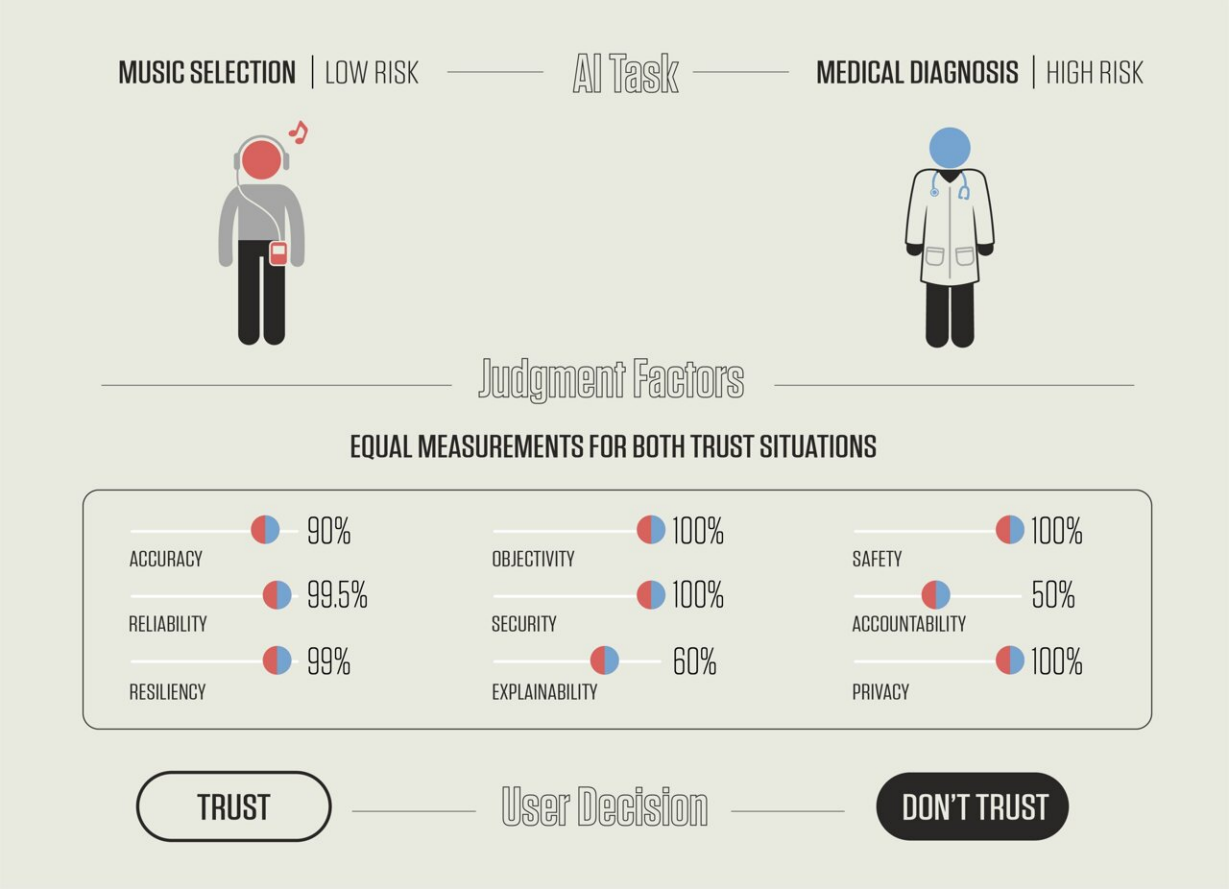
Proposed method for evaluating user trust in artificial intelligence systems

May 20 2021, by Chad Boutin



NIST's new publication proposes a list of nine factors that contribute to a human's potential trust in an AI system. A person may weigh the nine factors differently depending on both the task itself and the risk involved in trusting the AI's decision. As an example, two different AI programs — a music selection algorithm and an AI that assists with cancer diagnosis — may score the same on all nine criteria. Users, however, might be inclined to trust the music selection algorithm but not the medical assistant, which is performing a far riskier task.

Credit: N. Hanacek/NIST

Every time you speak to a virtual assistant on your smartphone, you are talking to an artificial intelligence—an AI that can, for example, learn your taste in music and make song recommendations that improve based on your interactions. However, AI also assists us with more risk-fraught activities, such as helping doctors diagnose cancer. These are two very different scenarios, but the same issue permeates both: How do we humans decide whether or not to trust a machine's recommendations?

This is the question that a new draft publication from the National Institute of Standards and Technology (NIST) poses, with the goal of stimulating a discussion about how humans trust AI systems. The document, Artificial Intelligence and User Trust(NISTIR 8332), is open for public comment until July 30, 2021.

The report contributes to the broader NIST effort to help advance trustworthy AI systems. The focus of this latest publication is to understand how humans experience trust as they use or are affected by AI systems.

According to NIST's Brian Stanton, the issue is whether human trust in AI systems is measurable—and if so, how to measure it accurately and appropriately.

"Many factors get incorporated into our decisions about trust," said Stanton, one of the publication's authors. "It's how the user thinks and feels about the system and perceives the risks involved in using it."

Stanton, a psychologist, co-authored the publication with NIST computer scientist Ted Jensen. They largely base the document on past research

into trust, beginning with the integral role of trust in human history and how it has shaped our cognitive processes. They gradually turn to the unique trust challenges associated with AI, which is rapidly taking on tasks that go beyond human capacity.

"AI systems can be trained to 'discover' patterns in large amounts of data that are difficult for the human brain to comprehend. A system might continuously monitor a very large number of video feeds and, for example, spot a child falling into a harbor in one of them," Stanton said. "No longer are we asking automation to do our work. We are asking it to do work that humans can't do alone."

The NIST publication proposes a list of nine factors that contribute to a person's potential trust in an AI system. These factors are different than the technical requirements of trustworthy AI that NIST is establishing in collaboration with the broader community of AI developers and practitioners. The paper shows how a person may weigh the factors described differently depending on both the task itself and the risk involved in trusting the AI's decision.

One factor, for example, is accuracy. A music selection algorithm may not need to be overly accurate, especially if a person is curious to step outside their tastes at times to experience novelty—and in any case, skipping to the next song is easy. It would be a far different matter to trust an AI that was only 90% accurate in making a cancer diagnosis, which is a far riskier task.

Stanton stressed that the ideas in the publication are based on background research, and that they would benefit from public scrutiny.

"We are proposing a model for AI user trust," he said. "It is all based on others' research and the fundamental principles of cognition. For that reason, we would like feedback about work the scientific community

might pursue to provide experimental validation of these ideas."

More information: Artificial Intelligence and User Trust (NISTIR 8332): nvlpubs.nist.gov/nistpubs/ir/2 … ST.IR.8332-draft.pdf

Provided by National Institute of Standards and Technology