New models sense human trust in smart machines

December 11 2018, by Emil Venere

How should intelligent machines be designed so as to “earn” the trust of humans? New models are informing these designs. Credit: Purdue University photo/Marshall Farthing

New "classification models" sense how well humans trust intelligent

machines they collaborate with, a step toward improving the quality of interactions and teamwork.

The long-term goal of the overall field of research is to design intelligent machines capable of changing their behavior to enhance human trust in them. The new models were developed in research led by assistant professor Neera Jain and associate professor Tahira Reid, in Purdue University's School of Mechanical Engineering.

"Intelligent machines, and more broadly, intelligent systems are becoming increasingly common in the everyday lives of humans," Jain said. "As humans are increasingly required to interact with intelligent systems, trust becomes an important factor for synergistic interactions."

For example, aircraft pilots and industrial workers routinely interact with automated systems. Humans will sometimes override these intelligent machines unnecessarily if they think the system is faltering.

"It is well established that human trust is central to successful interactions between humans and machines," Reid said.

The researchers have developed two types of "classifier-based empirical trust sensor models," a step toward improving trust between humans and intelligent machines.

The work aligns with Purdue's Giant Leaps celebration, acknowledging the university's global advancements made in AI, algorithms and automation as part of Purdue's 150th anniversary. This is one of the four themes of the yearlong celebration's Ideas Festival, designed to showcase Purdue as an intellectual center solving real-world issues.

The models use two techniques that provide data to gauge trust: electroencephalography and galvanic skin response. The first records

brainwave patterns, and the second monitors changes in the electrical characteristics of the skin, providing psychophysiological "feature sets" correlated with trust.

Forty-five human subjects donned wireless EEG headsets and wore a device on one hand to measure galvanic skin response.

One of the new models, a "general trust sensor model," uses the same set of psychophysiological features for all 45 participants. The other model is customized for each human subject, resulting in improved mean accuracy but at the expense of an increase in training time. The two models had a mean accuracy of 71.22 percent, and 78.55 percent, respectively.

It is the first time EEG measurements have been used to gauge trust in real time, or without delay.

"We are using these data in a very new way," Jain said. "We are looking at it in sort of a continuous stream as opposed to looking at brain waves after a specific trigger or event."

Findings are detailed in a research paper appearing in a special issue of the Association for Computing Machinery's Transactions on Interactive Intelligent Systems. The journal's special issue is titled "Trust and Influence in Intelligent Human-Machine Interaction." The paper was authored by mechanical engineering graduate student Kumar Akash; former graduate student Wan-Lin Hu, who is now a postdoctoral research associate at Stanford University; Jain and Reid.

"We are interested in using feedback-control principles to design machines that are capable of responding to changes in human trust level in real time to build and manage trust in the human-machine relationship," Jain said. "In order to do this, we require a sensor for

estimating human trust level, again in real-time. The results presented in this paper show that psychophysiological measurements could be used to do this."

The issue of human trust in machines is important for the efficient operation of "human-agent collectives."

"The future will be built around human-agent collectives that will require efficient and successful coordination and collaboration between humans and machines," Jain said. "Say there is a swarm of robots assisting a rescue team during a natural disaster. In our work we are dealing with just one human and one machine, but ultimately we hope to scale up to teams of humans and machines."

Algorithms have been introduced to automate various processes.

"But we still have humans there who monitor what's going on," Jain said. "There is usually an override feature, where if they think something isn't right they can take back control."

Sometimes this action isn't warranted.

"You have situations in which humans may not understand what is happening so they don't trust the system to do the right thing," Reid said. "So they take back control even when they really shouldn't."

In some cases, for example in the case of pilots overriding the autopilot, taking back control might actually hinder safe operation of the aircraft, causing accidents.

"A first step toward designing intelligent machines that are capable of building and maintaining trust with humans is the design of a sensor that will enable machines to estimate human trust level in real time," Jain

said.

To validate their method, 581 online participants were asked to operate a driving simulation in which a computer identified road obstacles. In some scenarios, the computer correctly identified obstacles 100 percent of the time, whereas in other scenarios the computer incorrectly identified the obstacles 50 percent of the time.

"So, in some cases it would tell you there is an obstacle, so you hit the brakes and avoid an accident, but in other cases it would incorrectly tell you an obstacle exists when there was none, so you hit the breaks for no reason," Reid said.

The testing allowed the researchers to identify psychophysiological features that are correlated to human trust in intelligent systems, and to build a trust sensor model accordingly. "We hypothesized that the trust level would be high in reliable trials and be low in faulty trials, and we validated this hypothesis using responses collected from 581 online participants," she said.

The results validated that the method effectively induced trust and distrust in the intelligent machine.

"In order to estimate trust in real time, we require the ability to continuously extract and evaluate key psychophysiological measurements," Jain said. "This work represents the first use of real-time psychophysiological measurements for the development of a human trust sensor."

The EEG headset records signals over nine channels, each channel picking up different parts of the brain.

"Everyone's brainwaves are different, so you need to make sure you are

building a classifier that works for all humans."

For autonomous systems, human trust can be classified into three categories: dispositional, situational, and learned.

Dispositional trust refers to the component of trust that is dependent on demographics such as gender and culture, which carry potential biases.

"We know there are probably nuanced differences that should be taken into consideration," Reid said. "Women trust differently than men, for example, and trust also may be affected by differences in age and nationality."

Situational trust may be affected by a task's level of risk or difficulty, while learned is based on the human's past experience with autonomous systems.

The models they developed are called classification algorithms.

"The idea is to be able to use these models to classify when someone is likely feeling trusting versus likely feeling distrusting," she said.

Jain and Reid have also investigated dispositional trust to account for gender and cultural differences, as well as dynamic models able to predict how trust will change in the future based on the data.

Provided by Purdue University

Citation: New models sense human trust in smart machines (2018, December 11) retrieved 9 April 2024 from https://techxplore.com/news/2018-12-human-smart-machines.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.