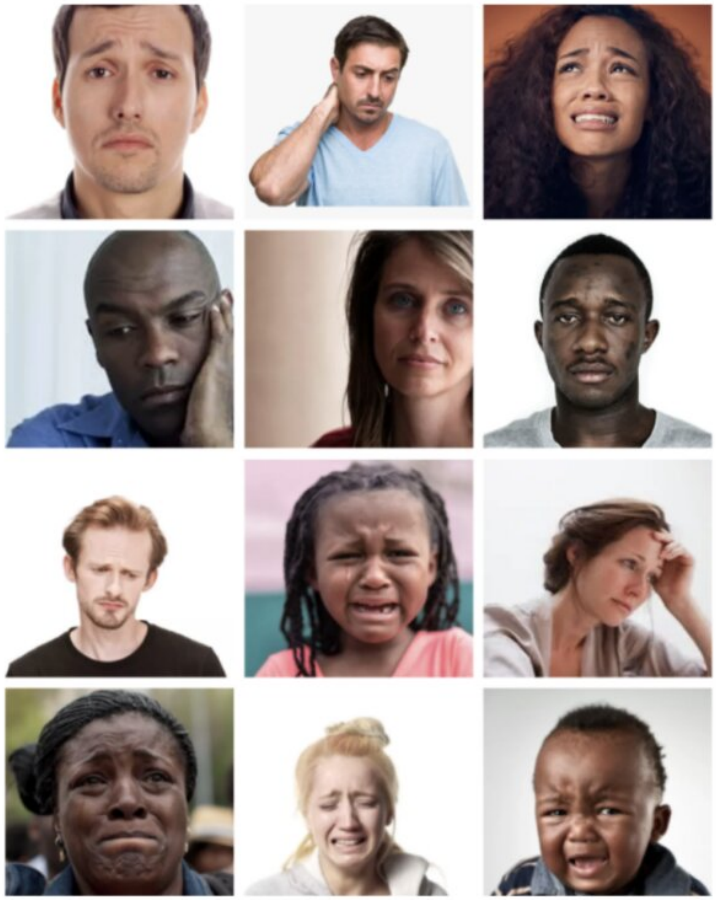
Transparent labeling of training data may boost trust in artificial intelligence

April 24 2023, by Matt Swayne



Showing users that visual data fed into artificial intelligence systems was labeled correctly might make people trust AI more, according to researchers. Credit: Penn State / Creative Commons

Showing users that visual data fed into artificial intelligence (AI) systems was labeled correctly might make people trust AI more, according to researchers. These findings may also pave the way to help scientists better measure the connection between labeling credibility, AI performance, and trust, the team added.

In a study, the researchers found that high-quality labeling of images led people to perceive that the training data was credible and they trusted the AI system more. However, when the system shows other signs of being biased, some aspects of their trust go down while others remain at a high level.

For AI systems to learn, they first must be trained using information that is often labeled by humans. However, most users never see how the data is labeled, leading to doubts about the accuracy and bias of those labels, according to S. Shyam Sundar, James P. Jimirro Professor of Media Effects in the Donald P. Bellisario College of Communications and co-director of the Media Effects Research Laboratory at Penn State University.

"When we talk about trusting AI systems, we are talking about trusting the performance of AI and the AI's ability to reflect reality and truth," said Sundar, who is also an affiliate of Penn State's Institute for Computational and Data Sciences. "That can happen if and only if the AI has been trained on a good sample of data. Ultimately, a lot of the concern about trust in AI should really be a concern about us trusting the training data upon which that AI is built. Yet, it has been a challenge to convey the quality of training data to laypersons."

According to the researchers, one way to convey that trustworthiness is to give users a glimpse of the labeling data.

"Often, the labeling process is not revealed to users, so we wondered

what would happen if we disclosed training data information, especially accuracy of labeling," said Chris (Cheng) Chen, assistant professor in communication and design, Elon University, and first author of the study. "We wanted to see whether that would shape people's perception of training data credibility and further influence their trust in the AI system."

The researchers recruited a total of 430 participants for the online study. The participants were asked to interact with a prototype Emotion Reader AI website, which was introduced as a system designed to detect facial expressions in social media images.

Researchers informed participants that the AI system had been trained on a dataset of almost 10,000 labeled facial images, with each image tagged as one of seven emotions—joy, sadness, anger, fear, surprise, disgust, or neutral. The participants were also informed that more than 500 people had participated in data labeling for the dataset. However, the researchers had manipulated the labeling, so in one condition the labels accurately described the emotions, while in the other, half of the facial images were mislabeled.

To study AI system performance, researchers randomly assigned participants to one of three experimental conditions: no performance, biased performance and unbiased performance. In the biased and unbiased conditions, participants were shown examples of AI performance involving the classification of emotions expressed by two Black and two white individuals. In the biased performance condition, the AI system classified all images of white individuals with 100% accuracy and all images of Black individuals with 0% accuracy, demonstrating a strong racial bias in AI performance.

According to the researchers, the participants' trust fell when they perceived that the system's performance was biased. However, their

emotional connection with the system and desire to use it in the future did not go down after seeing a biased performance.

Training data credibility

The researchers coined the term "training data credibility" to describe whether a user perceives training data as credible, trustworthy, reliable and dependable.

They suggest that developers and designers could measure trust in AI by creating new ways to assess user perception of training data credibility, such as letting users review a sample of the labeled data.

"It's also ethically important for companies to show the users how the training data has been labeled, so that they can determine if it's high-quality or low-quality labeling," said Chen.

Sundar added that AI developers would need to devise creative ways to share training data information with users, but without burdening or misleading them.

"Companies are always concerned about creating an easy flow for the user, so that users continue to engage," said Sundar, who is also director of the Penn State Center for Socially Responsible Artificial Intelligence, or CSRAI. "In calling for seamless ways to show labeling quality, we want interface designs that inform users and make them think rather than persuade them to blindly trust the AI system."

The researchers presented their findings today (April 24) at the ACM CHI Conference on Human Factors in Computing Systems, and reported them in its proceedings.

More information: Cheng Chen et al, Is this AI trained on Credible

Data? The Effects of Labeling Quality and Performance Bias on User Trust, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). DOI: 10.1145/3544548.3580805

Provided by Pennsylvania State University