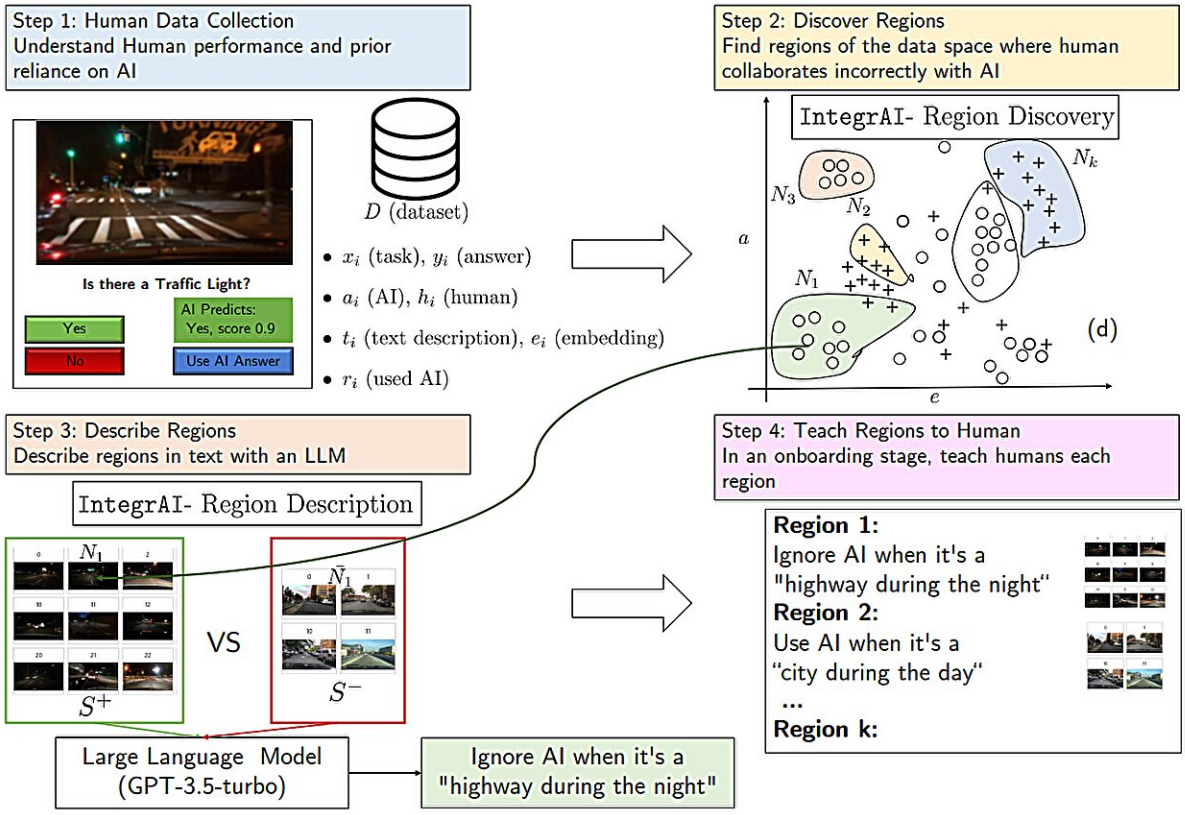
Automated system teaches users when to collaborate with an AI assistant

December 7 2023



The proposed onboarding approach with the IntegrAI algorithm. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2311.01007

Artificial intelligence models that pick out patterns in images can often do so better than human eyes—but not always. If a radiologist is using an

AI model to help her determine whether a patient's X-rays show signs of pneumonia, when should she trust the model's advice and when should she ignore it?

A customized onboarding process could help this radiologist answer that question, according to researchers at MIT and the MIT-IBM Watson AI Lab. They designed a system that teaches a user when to collaborate with an AI assistant.

In this case, the training method might find situations where the radiologist trusts the model's advice—except she shouldn't because the model is wrong. The system automatically learns rules for how she should collaborate with the AI, and describes them with natural language.

During onboarding, the radiologist practices collaborating with the AI using training exercises based on these rules, receiving feedback about her performance and the AI's performance.

The researchers found that this onboarding procedure led to about a 5 percent improvement in accuracy when humans and AI collaborated on an image prediction task. Their results also show that just telling the user when to trust the AI, without training, led to worse performance.

Importantly, the researchers' system is fully automated, so it learns to create the onboarding process based on data from the human and AI performing a specific task. It can also adapt to different tasks, so it can be scaled up and used in many situations where humans and AI models work together, such as in social media content moderation, writing, and programming.

"So often, people are given these AI tools to use without any training to help them figure out when it is going to be helpful. That's not what we

do with nearly every other tool that people use—there is almost always some kind of tutorial that comes with it. But for AI, this seems to be missing. We are trying to tackle this problem from a methodological and behavioral perspective," says Hussein Mozannar, a graduate student in the Social and Engineering Systems doctoral program within the Institute for Data, Systems, and Society (IDSS) and lead author of a paper about this training process.

The researchers envision that such onboarding will be a crucial part of training for medical professionals.

"One could imagine, for example, that doctors making treatment decisions with the help of AI will first have to do training similar to what we propose. We may need to rethink everything from continuing medical education to the way clinical trials are designed," says senior author David Sontag, a professor of EECS, a member of the MIT-IBM Watson AI Lab and the MIT Jameel Clinic, and the leader of the Clinical Machine Learning Group of the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Mozannar, who is also a researcher with the Clinical Machine Learning Group, is joined on the paper by Jimin J. Lee, an undergraduate in electrical engineering and computer science; Dennis Wei, a senior research scientist at IBM Research; and Prasanna Sattigeri and Subhro Das, research staff members at the MIT-IBM Watson AI Lab. The paper is available on the *arXiv* preprint server and will be presented at the Conference on Neural Information Processing Systems.

Training that evolves

Existing onboarding methods for human-AI collaboration are often composed of training materials produced by human experts for specific use cases, making them difficult to scale up. Some related techniques

rely on explanations, where the AI tells the user its confidence in each decision, but research has shown that explanations are rarely helpful, Mozannar says.

"The AI model's capabilities are constantly evolving, so the use cases where the human could potentially benefit from it are growing over time. At the same time, the user's perception of the model continues changing. So, we need a training procedure that also evolves over time," he adds.

To accomplish this, their onboarding method is automatically learned from data. It is built from a dataset that contains many instances of a task, such as detecting the presence of a traffic light from a blurry image.

The system's first step is to collect data on the human and AI performing this task. In this case, the human would try to predict, with the help of AI, whether blurry images contain traffic lights.

The system embeds these data points onto a latent space, which is a representation of data in which similar data points are closer together. It uses an algorithm to discover regions of this space where the human collaborates incorrectly with the AI. These regions capture instances where the human trusted the AI's prediction but the prediction was wrong, and vice versa.

Perhaps the human mistakenly trusts the AI when images show a highway at night.

After discovering the regions, a second algorithm utilizes a large language model to describe each region as a rule, using natural language. The algorithm iteratively fine-tunes that rule by finding contrasting examples. It might describe this region as "ignore AI when it is a

highway during the night."

These rules are used to build training exercises. The onboarding system shows an example to the human, in this case a blurry highway scene at night, as well as the AI's prediction, and asks the user if the image shows traffic lights. The user can answer yes, no, or use the AI's prediction.

If the human is wrong, they are shown the correct answer and performance statistics for the human and AI on these instances of the task. The system does this for each region, and at the end of the training process, repeats the exercises the human got wrong.

"After that, the human has learned something about these regions that we hope they will take away in the future to make more accurate predictions," Mozannar says.

Onboarding boosts accuracy

The researchers tested this system with users on two tasks—detecting traffic lights in blurry images and answering multiple choice questions from many domains (such as biology, philosophy, computer science, etc.).

They first showed users a card with information about the AI model, how it was trained, and a breakdown of its performance on broad categories. Users were split into five groups: Some were only shown the card, some went through the researchers' onboarding procedure, some went through a baseline onboarding procedure, some went through the researchers' onboarding procedure and were given recommendations of when they should or should not trust the AI, and others were only given the recommendations.

Only the researchers' onboarding procedure without recommendations

improved users' accuracy significantly, boosting their performance on the traffic light prediction task by about 5 percent without slowing them down. However, onboarding was not as effective for the question-answering task. The researchers believe this is because the AI model, ChatGPT, provided explanations with each answer that convey whether it should be trusted.

But providing recommendations without onboarding had the opposite effect—users not only performed worse, they took more time to make predictions.

"When you only give someone recommendations, it seems like they get confused and don't know what to do. It derails their process. People also don't like being told what to do, so that is a factor as well," Mozannar says.

Providing recommendations alone could harm the user if those recommendations are wrong, he adds. With onboarding, on the other hand, the biggest limitation is the amount of available data. If there aren't enough data, the onboarding stage won't be as effective, he says.

In the future, he and his collaborators want to conduct larger studies to evaluate the short- and long-term effects of onboarding. They also want to leverage unlabeled data for the onboarding process, and find methods to effectively reduce the number of regions without omitting important examples.

More information: Hussein Mozannar et al, Effective Human-AI Teams via Learned Natural Language Rules and Onboarding, *arXiv* (2023). DOI: 10.48550/arxiv.2311.01007

This story is republished courtesy of MIT News

(web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology