Aleksander Madry on building trustworthy artificial intelligence

December 16 2018, by Kim Martineau

Aleksander Madry is a leader in the emerging field of building guarantees into artificial intelligence, which has nearly become a branch of machine learning in its own right. Credit: CSAIL

Machine learning algorithms now underlie much of the software we use, helping to personalize our news feeds and finish our thoughts before

we're done typing. But as artificial intelligence becomes further embedded in daily life, expectations have risen. Before autonomous systems fully gain our confidence, we need to know they are reliable in most situations and can withstand outside interference; in engineering terms, that they are robust. We also need to understand the reasoning behind their decisions; that they are interpretable.

Aleksander Madry, an associate professor of computer science at MIT and a lead faculty member of the Computer Science and Artificial Intelligence Lab (CSAIL)'s Trustworthy AI initiative, compares AI to a sharp knife, a useful but potentially-hazardous tool that society must learn to weild properly. Madry recently spoke at MIT's Symposium on Robust, Interpretable AI, an event co-sponsored by the MIT Quest for Intelligence and CSAIL, and held Nov. 20 in Singleton Auditorium. The symposium was designed to showcase new MIT work in the area of building guarantees into AI, which has almost become a branch of machine learning in its own right. Six faculty members spoke about their research, 40 students presented posters, and Madry opened the symposium with a talk the aptly titled, "Robustness and Interpretability." We spoke with Madry, a leader in this emerging field, about some of the key ideas raised during the event.

Q: AI owes much of its recent progress to deep learning, a branch of machine learning that has significantly improved the ability of algorithms to pick out patterns in text, images and sounds, giving us automated assistants like Siri and Alexa, among other things. But deep learning systems remain vulnerable in surprising ways: stumbling when they encounter slightly unfamiliar examples in the real world or when a malicious attacker feeds it subtly-altered images. How are you and others trying to make AI more robust?

A: Until recently, AI researchers focused simply on getting machine-learning algorithms to accomplish basic tasks. Achieving even average-

case performance was a major challenge. Now that performance has improved, attention has shifted to the next hurdle: improving the worst-case performance. Most of my research is focused on meeting this challenge. Specifically, I work on developing next-generation machine-learning systems that will be reliable and secure enough for mission-critical applications like self-driving cars and software that filters malicious content. We're currently building tools to train object-recognition systems to identify what's happening in a scene or picture, even if the images fed to the model have been manipulated. We are also studying the limits of systems that offer security and reliability guarantees. How much reliability and security can we build into machine-learning models, and what other features might we need to sacrifice to get there?

My colleague Luca Daniel, who also spoke, is working on an important aspect of this problem: developing a way to measure the resilience of a deep learning system in key situations. Decisions made by deep learning systems have major consequences, and thus it's essential that end-users be able to measure the reliability of each of the model's outputs. Another way to make a system more robust is during the training process. In her talk, "Robustness in GANs and in Black-box Optimization," Stefanie Jegelka showed how the learner in a generative adversarial network, or GAN, can be made to withstand manipulations to its input, leading to much better performance.

Q: The neural networks that power deep learning seem to learn almost effortlessly: Feed them enough data and they can outperform humans at many tasks. And yet, we've also seen how easily they can fail, with at least three widely publicized cases of self-driving cars crashing and killing someone. AI applications in health care are not yet under the same level of scrutiny but the stakes are just as high. David Sontag focused his talk on the often life-or-death consequences when an AI system lacks robustness. What are some of the red flags when

training an AI on patient medical records and other observational data?

A: This goes back to the nature of guarantees and the underlying assumptions that we build into our models. We often assume that our training datasets are representative of the real-world data we test our models on—an assumption that tends to be too optimistic. Sontag gave two examples of flawed assumptions baked into the training process that could lead an AI to give the wrong diagnosis or recommend a harmful treatment. The first focused on a massive database of patient X-rays released last year by the National Institutes of Health. The dataset was expected to bring big improvements to the automated diagnosis of lung disease until a skeptical radiologist took a closer look and found widespread errors in the scans' diagnostic labels. An AI trained on chest scans with a lot of incorrect labels is going to have a hard time generating accurate diagnoses.

A second problem Sontag cited is the failure to correct for gaps and irregularities in the data due to system glitches or changes in how hospitals and health care providers report patient data. For example, a major disaster could limit the amount of data available for emergency room patients. If a machine-learning model failed to take that shift into account its predictions would not be very reliable.

Q: You've covered some of the techniques for making AI more reliable and secure. What about interpretability? What makes neural networks so hard to interpret, and how are engineers developing ways to peer beneath the hood?

A: Understanding neural-network predictions is notoriously difficult. Each prediction arises from a web of decisions made by hundreds to thousands of individual nodes. We are trying to develop new methods to make this process more transparent. In the field of computer vision one of the pioneers is Antonio Torralba, director of The Quest. In his talk, he

demonstrated a new tool developed in his lab that highlights the features that a neural network is focusing on as it interprets a scene. The tool lets you identify the nodes in the network responsible for recognizing, say, a door, from a set of windows or a stand of trees. Visualizing the object-recognition process allows software developers to get a more fine-grained understanding of how the network learns.

Another way to achieve interpretability is to precisely define the properties that make the model understandable, and then train the model to find that type of solution. Tommi Jaakkola showed in his talk, "Interpretability and Functional Transparency," that models can be trained to be linear or have other desired qualities locally while maintaining the network's overall flexibility. Explanations are needed at different levels of resolution much as they are in interpreting physical phenomena. Of course, there's a cost to building guarantees into machine-learning systems—this is a theme that carried through all the talks. But those guarantees are necessary and not insurmountable. The beauty of human intelligence is that while we can't perform most tasks perfectly, as a machine might, we have the ability and flexibility to learn in a remarkable range of environments.

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