What is a black box? A computer scientist explains what it means when the inner workings of AIs are hidden

May 23 2023, by Saurabh Bagchi

Credit: Unsplash/CC0 Public Domain

For some people, the term "black box" brings to mind the recording devices in airplanes that are valuable for postmortem analyzes if the unthinkable happens. For others it evokes small, minimally outfitted theaters. But black box is also an important term in the world of artificial

intelligence.

AI black boxes refer to AI systems with internal workings that are invisible to the user. You can feed them input and get output, but you cannot examine the system's code or the logic that produced the output.

Machine learning is the dominant subset of artificial intelligence. It underlies generative AI systems like ChatGPT and DALL-E 2. There are three components to machine learning: an algorithm or a set of algorithms, training data and a model. An algorithm is a set of procedures. In machine learning, an algorithm learns to identify patterns after being trained on a large set of examples—the training data. Once a machine-learning algorithm has been trained, the result is a machine-learning model. The model is what people use.

For example, a machine-learning algorithm could be designed to identify patterns in images, and training data could be images of dogs. The resulting machine-learning model would be a dog spotter. You would feed it an image as input and get as output whether and where in the image a set of pixels represents a dog.

Any of the three components of a machine-learning system can be hidden, or in a black box. As is often the case, the algorithm is publicly known, which makes putting it in a black box less effective. So to protect their intellectual property, AI developers often put the model in a black box. Another approach software developers take is to obscure the data used to train the model—in other words, put the training data in a black box.

The opposite of a black box is sometimes referred to as a glass box. An AI glass box is a system whose algorithms, training data and model are all available for anyone to see. But researchers sometimes characterize aspects of even these as black box.

That's because researchers don't fully understand how machine-learning algorithms, particularly deep-learning algorithms, operate. The field of explainable AI is working to develop algorithms that, while not necessarily glass box, can be better understood by humans.

Why AI black boxes matter

In many cases, there is good reason to be wary of black box machine-learning algorithms and models. Suppose a machine-learning model has made a diagnosis about your health. Would you want the model to be black box or glass box? What about the physician prescribing your course of treatment? Perhaps she would like to know how the model arrived at its decision.

What if a machine-learning model that determines whether you qualify for a business loan from a bank turns you down? Wouldn't you like to know why? If you did, you could more effectively appeal the decision, or change your situation to increase your chances of getting a loan the next time.

Black boxes also have important implications for software system security. For years, many people in the computing field thought that keeping software in a black box would prevent hackers from examining it and therefore it would be secure. This assumption has largely been proved wrong because hackers can reverse-engineer software—that is, build a facsimile by closely observing how a piece of software works—and discover vulnerabilities to exploit.

If software is in a glass box, then software testers and well-intentioned hackers can examine it and inform the creators of weaknesses, thereby minimizing cyberattacks.

This article is republished from The Conversation under a Creative

Commons license. Read the original article.

Provided by The Conversation