THE A.I. ENIGMA
LET'S SHINE A LIGHT INTO THE BLACK BOX

BY CLIVE THOMPSON

SAY YOU APPLY for home insurance and get turned down. You ask why, and the company explains its reasoning: Your neighborhood is at high risk for flooding, or your credit is dodgy. ¶ Fair enough. Now imagine you apply to a firm that uses a machine-learning system, instead of a human with an actuarial table, to predict insurance risk. After crunching your info—age, job, house location and value—the machine decides, nope, no policy for you. You ask the same question: "Why?" ¶ Now things get weird. Nobody can answer, because nobody understands how these systems—neural networks modeled on the human brain—produce their results. Computer scientists "train" each one by feeding it data, and it gradually learns. But once a neural net is working well, it's a black box. Ask its creator how it achieves a certain result and you'll likely get a shrug. ¶ The opacity of machine learning isn't just an academic problem. More and more places use the technology for everything from image recognition to medical diagnoses. All that decisionmaking is, by definition, unknowable—and that makes people uneasy. My friend Zeynep Tufekci, a sociologist, warns about "Moore's law plus inscrutability." Microsoft CEO Satya Nadella says we need "algorithmic accountability." ¶ All that is behind the fight to make machine learning more comprehensible. This spring, the European Union passed a regulation giving its citizens what University of Oxford researcher Bryce Goodman describes as an effective "right to an explanation" for decisions made by machine-learning systems. Starting in 2018, EU citizens will be entitled to know how an institution arrived at a conclusion—even if an AI did the concluding.

Jan Albrecht, an EU legislator from Germany, thinks explanations are crucial for public acceptance of artificial intelligence. "Otherwise people are afraid of it," he says. "There needs to be someone who has control." Explanations of what's happening inside the black box could also help ferret out bias in the systems. If a system for approving bank loans were trained on data that had relatively few black people in it, Goodman says, it might be uncertain about black applicants—and be more likely to reject them.

So sure, more clarity would be good. But is it *possible*? The box is, after all, black. Early experiments have shown promise. At the machine-learning company Clarifai, founder Matt Zeiler analyzed a neural net trained to recognize images of animals and objects. By blocking out portions of pictures and seeing how the different "layers" inside the net responded, he could begin to see which parts were responsible for recognizing, say, faces. Researchers at the University of Amsterdam have pursued a similar approach. Google, which has a large stake in AI, is doing its own probing: Its hallucinogenic "deep dreaming" pictures emerged from experiments that amplified errors in machine learning to figure out how the systems worked.

Of course, there's self-interest operating here too. The more that companies grasp what's going on inside their AIs, the more they can improve their products. The first stage of machine learning was just building these new brains. Now comes the Freudian phase: analysis. "I think we're going to get better and better," Zeiler says.

Granted, these are still early days. The people probing the black boxes might run up against some inherent limits to human comprehension. If machine learning is powerful because it processes data in ways we can't, it might seem like a waste of time to try to dissect it—and might even hamper its development. But the stakes for society are too high, and the challenge is frankly too fascinating. Human beings are creating a new breed of intelligence; it would be irresponsible not to *try* to understand it. ▥

ZOHAR LAZAR

WIRED MAGAZINE

CLIVE@CLIVETHOMPSON.NET

OCT 2016