Harvard
Business
Review

Risk Management

The AI Transparency Paradox

by Andrew Burt

December 13, 2019

Jorg Greuel/Getty Images

Summary. In recent years, academics and practitioners alike have called for greater transparency into the inner workings of artificial intelligence models, and for many good reasons. Transparency can help mitigate issues of fairness, discrimination, and trust — all of... **more**

In recent years, academics and practitioners alike have called for greater transparency into the inner workings of artificial intelligence models, and for many good reasons. Transparency can help mitigate issues of fairness, discrimination, and trust — all of which have received increased attention. Apple's new credit

card business has been accused of sexist lending models, for example, while Amazon scrapped an AI tool for hiring after discovering it discriminated against women.

At the same time, however, it is becoming clear that disclosures about AI pose their own risks: Explanations can be hacked, releasing additional information may make AI more vulnerable to attacks, and disclosures can make companies more susceptible to lawsuits or regulatory action.

Call it AI's "transparency paradox" — while generating more information about AI might create real benefits, it may also create new risks. To navigate this paradox, organizations will need to think carefully about how they're managing the risks of AI, the information they're generating about these risks, and how that information is shared and protected.

Some recent studies illustrate these trends. Let's start with a research paper by scholars at Harvard and the University of California, Irvine published last month. The paper focused on how variants of LIME and SHAP, two popular techniques used to explain so-called black box algorithms, could be hacked.

To illustrate the power of LIME, the 2016 paper announcing the tool explained how an otherwise incomprehensible image classifier recognized different objects in an image: an acoustic guitar was identified by the bridge and parts of the fretboard, while a Labrador Retriever was identified by specific facial features on the right side of the dog's face.

LIME, and the explainable AI movement more broadly, have been praised as breakthroughs able to make opaque algorithms more transparent. Indeed, the benefit of explaining AI has been a widely accepted precept, touted by both scholars and technologists, including me.

But the potential for new attacks on LIME and SHAP highlights an overlooked downside. As the study illustrates, explanations can be intentionally manipulated, leading to a loss of trust not just in the model but in its explanations too.

And it's not just this research that demonstrates the potential dangers of transparency in AI. Earlier this year, Reza Shokri and his colleagues illustrated how exposing information about machine-learning algorithms can make them more vulnerable to attacks. Meanwhile, researchers at the University of California, Berkeley, have demonstrated that entire algorithms can be stolen based simply on their explanations alone.

As security and privacy researchers focus more energy on AI, these studies, along with a host of others, all suggest the same conclusion: the more a model's creators reveal about the algorithm, the more harm a malicious actor can cause. This means that releasing information about a model's inner workings may actually decrease its security or expose a company to more liability. All data, in short, carries risks.

The good news? Organizations have long confronted the transparency paradox in the realms of privacy, security, and elsewhere. They just need to update their methods for AI.

To start, companies attempting to utilize artificial intelligence need to recognize that there are costs associated with transparency. This is not, of course, to suggest that transparency isn't worth achieving, simply that it also poses downsides that need to be fully understood. These costs should be incorporated into a broader risk model that governs how to engage with explainable models and the extent to which information about the model is available to others.

Second, organizations must also recognize that security is becoming an increasing concern in the world of AI. As AI is adopted more widely, more security vulnerabilities and bugs will surely be discovered, as my colleagues and I at the Future of Privacy Forum recently argued. Indeed, security may be one of the biggest long-term barriers to the adoption of AI.

Last is the importance of engaging with lawyers as early and as often as possible when creating and deploying AI. Involving legal departments can facilitate an open and legally privileged environment, allowing companies to thoroughly probe their models for every vulnerability imaginable without creating additional liabilities.

Indeed, this is exactly why lawyers operate under legal privilege, which gives the information they gather a protected status, incentivizing clients to fully understand their risks rather than to hide any potential wrongdoings. In cybersecurity, for example, lawyers have become so involved that it's common for legal departments to manage risk assessments and even incident-response activities after a breach. The same approach should apply to AI.

In the world of data analytics, it's frequently assumed that more data is better. But in risk management, data itself is often a source of liability. That's beginning to hold true for artificial intelligence as well.

Andrew Burt is the managing partner of Luminos.Law, a boutique law firm focused on AI and analytics, and a visiting fellow at Yale Law School's Information Society Project.