Enhancing trust in artificial intelligence: Audits and explanations can help

Carl Schonander

By , Contributor, CIO | 13 August 2019 07:14 PT

Opinions expressed by ICN authors are their own.

There are a lot of tools available to help with AI audits and explanations and more will be available in the coming years.

xijian / Getty Images

There is a lively debate all over the world regarding AI's perceived "black box" problem. Most profoundly, if a machine can be taught to learn itself, how does it explain its conclusions? This issue comes up most frequently in the context of how to address possible algorithmic bias. One way to address this issue is to mandate a right to a human decision per the General Data Protection Regulation's (GDPR) Article 22. Here in the United States, Senators Wyden and Booker propose in the Algorithmic Accountability Act that companies be compelled to conduct impact assessments.

Auditability, explainability, transparency and replicability (reproducibility) are often suggested as means of avoiding bias. Auditability and explainability are probably furthest along in a practical sense, and they

can sometimes overlap in interesting ways.

Audits – at least for now – might be the way to go in many cases. What this means in reality is checking for adherence to guardrails/controls required by laws, regulations, and/or good or best practices. So, for example, if bias avoidance is the goal, then that needs to be defined precisely. There are different kinds of bias such as confirmational bias, measurement bias, and other forms which impact how conclusions are drawn from data.

Explainability is intrinsically challenging because explanations are often incomplete because they omit things that cannot be explained understandably. Algorithms are inherently challenging to explain. Take, for instance, algorithms using "ensemble" methodologies. Explaining how one model works is hard enough. Explaining how several models work both individually and together is exponentially more difficult. But there are some interesting new tools on the market that can help.

Transparency is usually a good thing. However, it if it requires disclosing source code or the engineering details underpinning an AI application, it could raise intellectual property concerns. And again, transparency about something that may be unexplainable in laymen's terms would be of limited use.

Replicability or reproducibility involves the extent to which an AI decision making process can be repeated with the same outcome. One problem with this approach is the absence of universal standards governing the data capture, curation and processing techniques to allow for such replicability. A second problem is that AI experiments often involve humans repeatedly running AI models until they find patterns in data and difficulty of distinguishing correlation from causation. A third problem is the sheer dynamism of this technology – reproducing results with so much change is difficult.

Relevant developments in the auditing field

With respect to audits, Jessi Hempel writes, in "Want to prove your business is fair? Audit your algorithm" about a company called Rentlogic that obtained an external audit of the algorithm it uses to score how well New York City landlords take care of their buildings. O'Neal Risk Consulting and Algorithmic Auditing (ORCAA) devised a matrix composed of a vertical line listing stakeholders and a horizontal line including traits denoted as accuracy, consistency, bias, transparency, fairness and timeliness. The point of the matrix was to spur conversations within the company about the algorithm. So one recommendation, for instance, was to check extra carefully for patterns from inspector reports.

In 2018, the Information Systems Audit & Control Association (ISACA) released a white paper titled "Auditing Artificial Intelligence" that provides some ideas for how to conduct audits. One important aspect is that the audits do not have to be particularly technical in nature. Instead, auditors should "focus on the controls and governance structures that are in place and determine that they are operating effectively. Auditors can provide some assurance by focusing on the business and IT governance aspects."

Keying off the IASCA white paper, one way of conducting an external audit might be to verify that companies have actually instituted a framework for responsible data use, particularly in the context of avoiding bias. What might such a framework look like, at least in the United States?

- Find affirmative ways to use data to ensure fairness for historically underserved groups.
- Ensure application of existing non-discrimination/consumer law to use of data analytics.
- Provide indication of the kinds of factors that go into decision-making algorithms.
- Draw from disparate impact analysis methodologies to conduct internal assessments.
- Develop fairness standards for responding to internal assessments.

Checking or, to use another word, auditing that companies have these processes in place would make a contribution to ensuring that AI is used in non-discriminatory ways.

Explanations are becoming easier to get and easier to understand

There is a lot of interesting work going on with respect to explainability as well. And some of it overlaps with auditability of governance. Cornell University, for example, developed the idea of producing model cards for machine learning models in a paper titled "Model Cards for Model Reporting." The idea is for machine learning models to have a "model card" including information on model details (including citation details and licenses and where to send questions or comments about the model); intended use; factors such as demographics in developing the model; metrics; evaluation data; training data; quantitative analysis; ethical considerations such as whether the model uses sensitive data, i.e. affects human life; mitigations; potential risks and harms; and, information about especially risky use cases. This looks like a condensed impact assessment – it certainly provides a guide for companies on how to approach explainability.

A recent Forbes piece entitled: "Future of Explainable AI" also provides some interesting information on explainability developments. It goes into some detail on Google's What-If-Tool for Cloud AI Platform models." We'll see how this tool gets used in practice, but there is promise in that it permits analysts to slice datasets by features (say age) and compare performance across those slices, which can be helpful in machine learning fairness investigations.

The bottom line is that there are a lot of tools available to help with audits and explanations and more will be available in the coming years. It's hard to say whether audits or explanations will be relatively more important – both will have a role to play in enhancing the public's trust in AI.

This article is published as part of the IDG Contributor Network. Want to Join?

Carl Schonander is the Senior VP for Global Public Policy at SIIA. Prior to this, Carl was Senior Director for International Policy at SIIA, and before that served as a US diplomat.