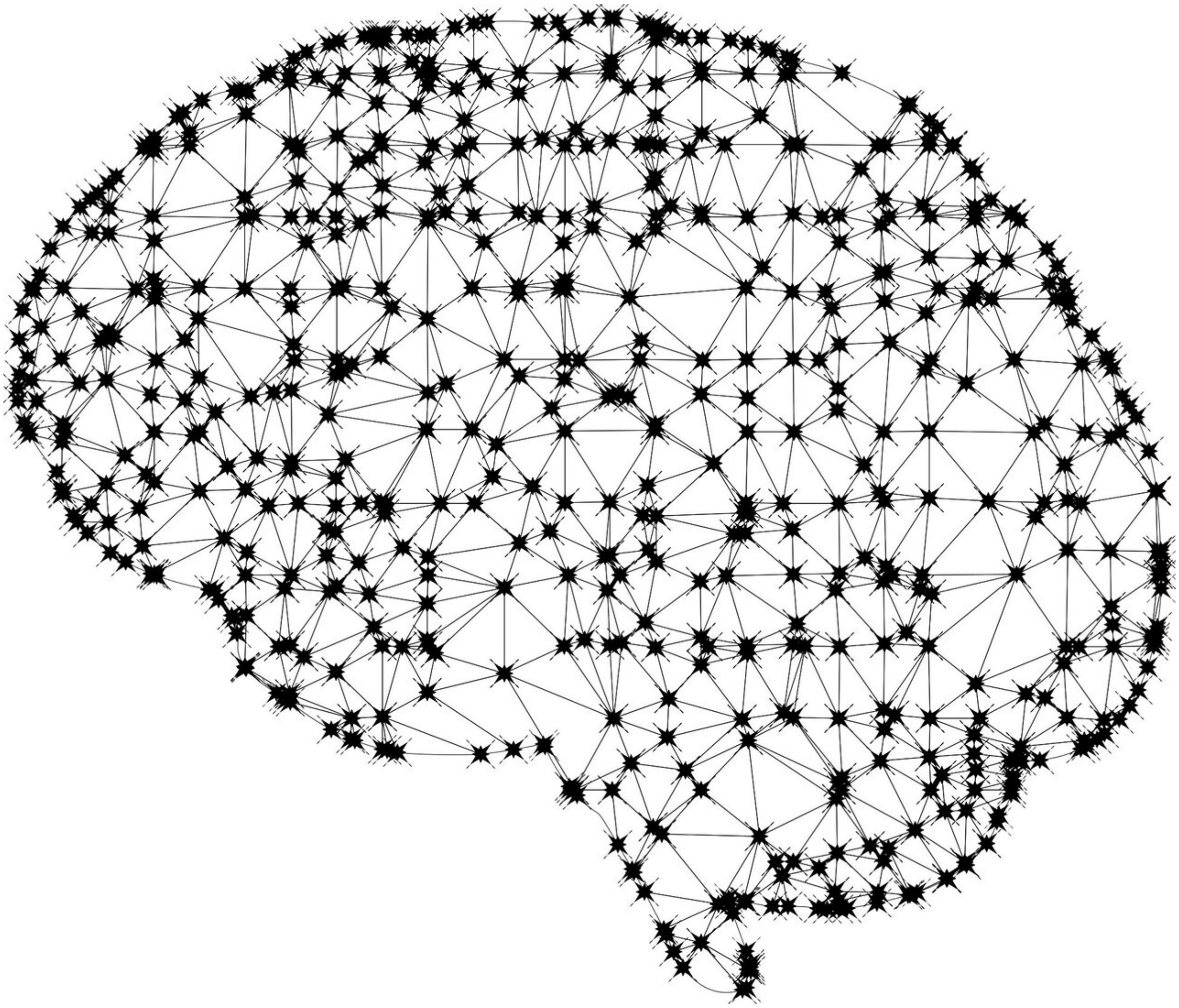
Strengthening trust in machine-learning models

March 28 2023, by Madeleine Turner



Credit: CC0 Public Domain

Probabilistic machine learning methods are becoming increasingly powerful tools in data analysis, informing a range of critical decisions across disciplines and applications, from forecasting election results to predicting the impact of microloans on addressing poverty.

This class of methods uses sophisticated concepts from probability theory to handle uncertainty in decision-making. But the math is only one piece of the puzzle in determining their accuracy and effectiveness. In a typical data analysis, researchers make many subjective choices, or potentially introduce human error, that must also be assessed in order to cultivate users' trust in the quality of decisions based on these methods.

To address this issue, MIT computer scientist Tamara Broderick, associate professor in the Department of Electrical Engineering and Computer Science (EECS) and a member of the Laboratory for Information and Decision Systems (LIDS), and a team of researchers have developed a classification system—a "taxonomy of trust"—that defines where trust might break down in a data analysis and identifies strategies to strengthen trust at each step. The other researchers on the project are Professor Anna Smith at the University of Kentucky, professors Tian Zheng and Andrew Gelman at Columbia University, and Professor Rachael Meager at the London School of Economics. The team's hope is to highlight concerns that are already well-studied and those that need more attention.

In their paper, published in February in *Science Advances*, the researchers begin by detailing the steps in the data analysis process where trust might break down: Analysts make choices about what data to collect and which models, or mathematical representations, most closely mirror the real-life problem or question they are aiming to answer. They select algorithms to fit the model and use code to run those algorithms. Each of these steps poses unique challenges around building trust. Some components can be checked for accuracy in measurable ways. For example, "Does my code have bugs?" is a question that can be tested against objective criteria. Other times, problems are more subjective, with no clear-cut answers; analysts are confronted with numerous strategies to gather data and decide whether a model reflects the real world.

"What I think is nice about making this taxonomy is that it really highlights where people are focusing. I think a lot of research naturally focuses on this level of 'Are my algorithms solving a particular mathematical problem?' in part because it's very objective, even if it's a hard problem," Broderick says.

"I think it's really hard to answer 'Is it reasonable to mathematize an important applied problem in a certain way?' because it's somehow getting into a harder space, it's not just a mathematical problem anymore."

Capturing real life in a model

The researchers' work in categorizing where trust breaks down, though it may seem abstract, is rooted in real-world application.

Meager, a co-author on the paper, analyzed whether microfinances can have a positive effect in a community. The project became a case study for where trust could break down, and ways to reduce this risk.

At first look, measuring the impact of microfinancing might seem like a straightforward endeavor. But like any analysis, researchers meet challenges at each step in the process that can affect trust in the outcome. Microfinancing—in which individuals or small businesses receive small loans and other financial services in lieu of conventional banking—can offer different services, depending on the program. For the analysis, Meager gathered datasets from microfinance programs in countries across the globe, including in Mexico, Mongolia, Bosnia, and the Philippines.

When combining conspicuously distinct datasets, in this case from multiple countries and across different cultures and geographies, researchers must evaluate whether specific case studies can reflect broader trends. It is also important to contextualize the data on hand. For example, in rural Mexico, owning goats may be counted as an investment.

"It's hard to measure the quality of life of an individual. People measure things like, 'What's the business profit of the small business?' or 'What's the consumption level of a household?' There's this potential for mismatch between what you ultimately really care about, and what you're measuring," Broderick says. "Before we get to the mathematical level, what data and what assumptions are we leaning on?"

With data on hand, analysts must define the real-world questions they seek to answer. In the case of evaluating the benefits of microfinancing, analysts must define what they consider a positive outcome. It is standard in economics, for example, to measure the average financial gain per business in communities where a microfinance program is introduced. But reporting an average might suggest a net positive effect even if only a few (or even one) person benefited, instead of the community as a whole.

"What you really wanted was that a lot of people are benefiting," Broderick says. "It sounds simple. Why didn't we measure the thing that we cared about? But I think it's really common that practitioners use standard machine learning tools, for a lot of reasons. And these tools might report a proxy that doesn't always agree with the quantity of interest."

Analysts may consciously or subconsciously favor models they are familiar with, especially after investing a great deal of time learning their ins and outs. "Someone might be hesitant to try a nonstandard method because they might be less certain they will use it correctly. Or peer review might favor certain familiar methods, even if a researcher might like to use nonstandard methods," Broderick says. "There are a lot of reasons, sociologically. But this can be a concern for trust."

Final step, checking the code

While distilling a real-life problem into a model can be a big-picture, amorphous problem, checking the code that runs an algorithm can feel "prosaic," Broderick says. But it is another potentially overlooked area where trust can be strengthened.

In some cases, checking a coding pipeline that executes an algorithm might be considered outside the purview of an analyst's job, especially when there is the option to use standard software packages.

One way to catch bugs is to test whether code is reproducible. Depending on the field, however, sharing code alongside published work is not always a requirement or the norm. As models increase in complexity over time, it becomes harder to recreate code from scratch. Reproducing a model becomes difficult or even impossible.

"Let's just start with every journal requiring you to release your code. Maybe it doesn't get totally double-checked, and everything isn't absolutely perfect, but let's start there," Broderick says, as one step toward building trust.

Paper co-author Gelman worked on an analysis that forecast the 2020 U.S. presidential election using state and national polls in real-time. The team published daily updates in *The Economist*, while also publishing their code online for anyone to download and run themselves. Throughout the season, outsiders pointed out both bugs and conceptual problems in the model, ultimately contributing to a stronger analysis.

The researchers acknowledge that while there is no single solution to create a perfect model, analysts and scientists have the opportunity to reinforce trust at nearly every turn.

"I don't think we expect any of these things to be perfect," Broderick says, "but I think we can expect them to be better or to be as good as possible."

More information: Tamara Broderick et al, Toward a taxonomy of trust for probabilistic machine learning, *Science Advances* (2023). DOI: 10.1126/sciadv.abn3999

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

Citation: Strengthening trust in machine-learning models (2023, March 28) retrieved 10 September 2024 from https://techxplore.com/news/2023-03-machine-learning.html