MITSloan
Management Review

The No. 1 Question to Ask When Evaluating AI Tools

SPRING 2023
ISSUE

Determining whether an AI solution is worth implementing requires looking past performance reports and finding the ground truth on which the AI has been trained and validated.

Sarah Lebovitz
Hila Lifshitz-Assaf
Natalia Levina

Vol. 64, No. 3 • Reprint #64314 • sloanreview.mit.edu

The No. 1 Question to Ask When Evaluating AI Tools

Determining whether an AI solution is worth implementing requires looking past performance reports and finding the ground truth on which the AI has been trained and validated.

By Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina

IN THE FAST-MOVING AND HIGHLY COMPETITIVE ARTIFICIAL INTELLIGENCE sector, developers' claims that their AI tools can make critical predictions with a high degree of accuracy are key to selling prospective customers on their value. Because it can be daunting for people who are not AI experts to evaluate these tools, leaders may be tempted to rely on the high-level performance metrics published in sales materials. But doing so often leads to disappointing or even risky implementations.

Over the course of an 11-month investigation, we observed managers in a leading health care organization as they conducted internal pilot studies of five AI tools. Impressive performance results had been promised for each, but several of the tools did extremely poorly in their pilots. Analyzing the evaluation process, we found that an effective way to determine an AI tool's quality is understanding and examining its *ground truth*.[1] In this article, we'll explain what that is and how managers can dig into it to better assess whether a particular AI tool may enhance or diminish decision-making in their organization.

Andy Potts

What Is the Ground Truth of the AI Tool?

The quality of an AI tool — and the value it can bring your organization — is enabled by the quality of the ground truth used to train and validate it. In general, ground truth is defined as information that is known to be true based on objective, empirical evidence. In AI, ground truth refers to the data in training data sets that teaches an algorithm how to arrive at a predicted output; ground truth is considered to be the "correct" answer to the prediction problem that the tool is learning to solve. This data set then becomes the standard against which developers measure the accuracy of the system's predictions. For instance, teaching a model to identify the best job candidates requires training data sets describing candidates' features, such as education and years of experience, where each is associated with a classification of either "good candidate" (true) or "not a good candidate" (false). Training a model to flag inappropriate content such as bullying or hate speech requires data sets full of text and images that have been classified "appropriate" (true) or "not appropriate" (false). The aim is that once the model is in production, it has learned the pattern of features that predicts the correct output for a new data point.

In recent years, there has been growing awareness of the risks of using features from the training data sets that are not representative or that contain bias.[2] There is surprisingly little discussion, however, about the quality of the labels that serve as the ground truth for model development. It is critical that managers ask, "Is the ground truth really true?"

The first step in gaining clarity into the ground truth for a tool is to investigate the metric typically used by AI companies to support performance claims, known as the AUC (short for *area under the receiver operating characteristic curve*). The AUC metric summarizes the model's accuracy in making predictions on a scale of 0 to 1, where 1 represents perfect accuracy.[3] Managers often fixate on this metric as evidence of AI quality — and take at face value the comparison with an AUC for the same prediction task done by humans.

The AUC is calculated by comparing AI outputs to ground truth categories that were used by AI designers. The AI output is considered correct if it matches the ground truth label and incorrect if it does not. The usefulness and relevance of the AUC metric is contingent upon the quality of the ground truth labels, which cannot simply be assumed to be high-quality sources of truth.

Here's the underlying problem: For many critical decisions in organizations, there is rarely an objective "truth" ready to be fed to an algorithm. Instead, AI designers construct ground truth data, and they have considerable latitude in how to accomplish this. For example, in the medical context, AI developers make significant trade-offs when choosing what ground truth will be used to train and validate a cancer diagnosis model. They could use biopsy results to serve as the ground truth, which would provide an externally validated result for whether cancer was detected. However, most patients never undergo biopsy tests (thankfully), and acquiring these results for all patients in the training data set would require enormous investment and patient cooperation.

Alternatively, developers may use the diagnosis recorded by the clinical physician overseeing a given patient at the time. This data is relatively easy to acquire from historical electronic health records. Developers could also recruit an expert physician, or a panel of experts, to produce a diagnosis for a sample of cases in the training data set, using the average or majority of their opinions as the ground truth label. Creating this type of data set may be costly and time-consuming, but it is commonly done in the medical AI community. In any case, AI developers weigh the relative costs and benefits when deciding how to assign ground truth labels — a decision that has great influence on the overall quality and potential value of the tool.

To identify an AI tool's ground truth, simply ask the vendor or developers. Verify their answers by searching for "ground truth" or "label" on technical research reports and methodology summaries. For medical tools subject to regulatory approval, this information is publicly available on the U.S. Food and Drug Administration website. We recommend deeply engaging with AI vendors and internal development teams and having open conversations about their ground truth selections, their logic behind those choices, and any trade-offs they considered. Reticence to discuss these topics transparently should be interpreted as a serious red flag.

How Objective or Externally Verifiable Is the Ground Truth?

In some contexts, what is considered the truth about a given decision outcome may be straightforward and widely agreed upon. If so, the AI ground truth may consist of more objective data sets. For example, to predict the impact of tropical storms, AI designers may rely on the volume of insurance claims and government payouts to serve as the ground truth for

labeling a weather event as highly damaging or not.

However, many AI solutions on the market focus on more subjective decision contexts, where experts often disagree about whether a decision was "true" (domains such as criminal justice, human resources, college admissions, strategic investing, and so forth). In many contexts of medical diagnosis, there is often no objective means to validate a given decision as accurate or not. In lieu of such an objective source, AI designers often use physicians' diagnostic opinions to represent the truth in their AI training data. They do this even though published medical research shows high variability and subjectivity across even the most seasoned and qualified experts, especially when it comes to making diagnoses for diseases that are very hard to differentiate.

THE RESEARCH

- The authors set out to explore how professional work is being impacted by advanced technologies.
- They conducted an in-depth qualitative field study within a health care organization actively developing, adopting, and using AI for significant decision-making tasks.
- Data collection involved 11 months of in-hospital observation, more than 40 long-form interviews, and analysis of archival documentation.

Validating experts' decisions is extremely challenging and in some cases impossible. For instance, if a patient never returns to the diagnosing clinic, one may conclude that the doctor's diagnosis was accurate and the treatment was effective, even if the patient's condition worsened and they decided to seek help elsewhere.

Such variability and subjectivity are exactly what is fueling investment in AI tools that can address these thorny decision contexts. Yet the very fact that they are subjective makes finding high-quality ground truth very challenging. Given that these decision contexts are also where we find sensitive issues involving high risk and ethical implications, it is especially important to investigate the ground truth and consider the best practices used by human experts making similar decisions unassisted by AI. How much subjectivity or variability is inherently involved in making this decision? How are decisions validated? That is, what are the established and acceptable ways to gauge the quality of experts' decisions in that particular context?

In many professions, there are accepted standards for high-quality decisions — that is, what experts agree is the best way to evaluate a given judgment, with respect to the constraints and limitations at hand. These vary significantly across contexts, organizations, and fields of expertise. Managers evaluating tools for particular decisions should ask the human experts making those same decisions what the current standards and best practices for evaluating decision quality are for that specific domain.

Examples from our study demonstrate the diversity of these standards, even within the general area of cancer diagnosis. For breast cancer diagnosis, radiologists' judgments are validated against pathology results from biopsy studies. In the case of demarcating the boundaries of brain tumors, there is no single method that experts agree upon as the clear standard for evaluating judgments. Going beyond the medical context, in the field of human resource management, is a successful job candidate the one who passes all interviews — currently the popular ground truth for AI tools in this domain — or the one who gets hired and shows superior job performance for many subsequent years?

How Does the AI Ground Truth Compare to the Ideal Standard for Experts' Critical Decisions?

Once the ideal or gold standard for assessing experts' critical decisions is clear, it is time to compare it to the AI developers' methods for determining the ground truth used to train and validate the algorithm. The following case from our study illustrates the importance of this culminating step.

Health care managers were planning to conduct a pilot study of an AI tool for breast cancer diagnosis. In the course of debating what to use as the ground truth to validate the tool's performance on their internal data, they looked at what the AI developers had used for initial performance tests. They were shocked at what they discovered.

The AI tool was designed to predict "likely cancer" or "likely benign" on the basis of one mammography image input. In this decision context, the gold standard to validate this diagnosis would involve final pathology results and long-term patient health outcomes (data that is difficult and costly to acquire). Instead, the AI designers chose to construct ground truth labels to validate the tool by asking a panel of radiologists to render a judgment after looking at a single mammogram (the same input as the AI model). When they subsequently ran human-versus-AI performance tests on the model, they claimed it was an apples-to-apples comparison, where their panel of experts performed the same decision-making task as the AI model on the basis of the same single mammogram. The results of the test were striking and made headlines: The AI tool outperformed every expert in the study.

However, in peeling back the layers of this performance report, managers in our study would discover

that the ground truth used was severely inadequate in comparison to the accepted standard in the professional field. This misalignment created dangerous misconceptions about the tool's potential value.

The managers understood that validating diagnosis decisions on the basis of a single mammogram would be ludicrous — and dangerous. If a biopsy is unavailable or ill-advised, the acceptable professional standard for reviewing such a case involves much more thorough analysis. This practice involves noting changes over multiple follow-up appointments, reviewing and comparing numerous images (such as 3D tomosynthesis images and ultrasounds), conducting physical examinations and assessing the individual's risk factors (such as age, family history, and surgical history), and even requesting additional targeted imaging. Having discovered the vast disparity between their standard and that used to establish the AI tool's ground truth, the managers in our study decided to partner with internal data scientists to design a new tool using better ground truth labels.

It is highly likely that managers will encounter AI vendors using less-than-ideal sources of ground truth, given the costs and feasibility of obtaining high-quality ground truth data and perhaps their desire to show the tool's performance in the best light. That's why it's crucial to seek tools that have been trained on ground truth data that most closely approaches the ideal standard for decision-making quality in that knowledge domain. This ground truth data should encompass experts' know-how (their real-world deliberative knowledge processes), not just their "know-what" (the decisions recorded in the labels of a data set).

If the ground truth has been constructed in a way that closely resembles the experts' gold standard, that is a green light to move to further evaluation, such as assessing fit with technical infrastructure and conducting internal pilot studies. But if the AI ground truth is inferior, we recommend caution. If it's possible to influence the development process, push to redesign the AI tool using higher-quality ground truth data. Otherwise, adopting AI tools with inadequate ground truth will pose significant risks: Decision quality will be diminished to match the lower quality dictated by the ground truth data. Moreover, as organizations and our society adopt these tools at scale, professional learning will be greatly impeded as the remnants of experts' valuable know-how are lost and replaced by the AI model and outputs. We may accept this risk if we believe that AI is learning from high-quality ground truth and is doing so faster and better than humans can, but not otherwise.

Overlooking shaky AI ground truth data for critical decisions can have severe and lasting consequences.

AI PRODUCTS ON THE MARKET ARE MEANT to dazzle, and managers may be tempted to take vendor promises and performance claims at face value, given the challenges of evaluating these tools. But overlooking shaky AI ground truth data for critical decisions can have severe and lasting consequences. We suggest that managers peel back the layers of AI performance reports to identify and assess the ground truth that these systems were built with. Only then can they effectively assess whether an AI tool will deliver sufficient value to their organization. Doing so can deliver other benefits as well: We found that managers who followed the diligent process of evaluating AI often came to reevaluate their human experts' decision processes and found ways to improve them.

Finally, policy makers and researchers should also keep in mind that ground truth decisions made by AI designers have far-reaching influence, not just in the organizations that test and adopt AI tools, but on important societal issues that will have lasting impacts. They too need to consider ground truth as part of the discussion around AI adoption. ■

Sarah Lebovitz *is an assistant professor at the McIntire School of Commerce at the University of Virginia.* **Hila Lifshitz-Assaf** *is a professor at Warwick University and a faculty affiliate at the Lab for Innovation Science at Harvard.* **Natalia Levina** *is a professor at New York University's Stern School of Business.*

REFERENCES

1. S. Lebovitz, N. Levina, and H. Lifshitz-Assaf, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," MIS Quarterly 45, no. 3 (September 2021): 1501-1526.
2. C. DeBrusk, "The Risk of Machine-Learning Bias (and How to Prevent It)," MIT Sloan Management Review, March 26, 2018, https://sloanreview.mit.edu.
3. "Classification: ROC Curve and AUC," Machine Learning Crash Course, Google, last modified July 18, 2022, https://developers.google.com.

MITSloan
Management Review

PDFs • Reprints • Permission to Copy • Back Issues

MIT Sloan Management Review articles, permissions, and back issues can be purchased on our website, **shop.sloanreview.mit.edu**, or you may order through our Business Service Center (9 a.m. - 5 p.m. ET) at the phone number listed below.