Don't trust "explainable AI": Proper validation is what matters

Momin M. Malik, PhD
Center for Digital Health, Mayo Clinic, Rochester, MN, United States

DEFINING ARTIFICIAL INTELLIGENCE

BACKGROUND

Artificial intelligence is typically defined in terms of what it *aspires* to do ("th[e] activity devoted to making machines intelligent"), or what it *seems* to do ("a set of technologies... inspired by—but typically operate quite differently from—the ways people use their nervous systems and bodies to sense, learn, reason, and take action"), rather than *how* it does so.

The "how" is **by using statistical correlations**. Rather than, like in statistical modeling, trying to reason about correlations (e.g., by controlling for confounders), AI models find optimal sets of correlations that match to an output (in statistical terms, they are often "high-dimensional" and "non-parametric" statistical models, but statistical models nonetheless).

Thus, we can **define AI as the instrumental use of statistical correlations to match to previously observed output**, rather than to use correlations to try and model how the world works.

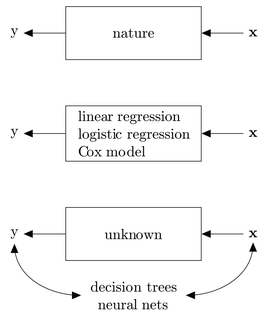

Fig 1. Reproduction of figure 1 from Leo Breiman's 2001 landmark article[3] that theorized machine learning (represented by the bottom diagram) as a new style of statistical modeling.

Because of the way that we fit AI models to retrospective data (the only data we have), and also test performance using retrospective data, **if AI models are made from observational data, then they cannot "predict" the results of an intervention.** In this sense, the use of the term "prediction" can be misleading, because it is "passive" prediction.

For example, multiple sets of models that look very different to one another, suggesting vastly different underlying causal processes, can "predict" equally well in retrospective data,[12] and so are not reliable guides to how to *intervene*. Breiman called this the "Rashomon problem" and saw it as a point against traditional statistical modeling, which seldom considers the full range of plausible models (but he did not engage with the question of what type of modeling better supports deciding how to intervene).

"EXPLAINABILITY"

AI models are highly flexible, often by finding correlations in ways that are no longer interpretable. E.g., they may find correlations after taking multiple nonlinear transformations of combinations of multiple sets of variables.

Rudin[15] defines "interpretability" as when models that can be understood by looking at them, for example by looking at the coefficients of a linear regression model.

"Explainability" is, for models that are not inherently interpretable, sets of descriptions of how particular model inputs are linked to particular model outputs.

The problem is when we take either our interpretations, or explanations, to be statements about the world rather than about the model, and base our trust of AI models on how well explanations fit with our causal intuitions.

Again in his 2001 article, Breiman[3] hails statistical "decision and regression trees" as an example of a highly-interpretable model, as exemplified by a presentation he gave to a group of Colorado judges for a model of sources of delay in criminal courts:

> "A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, 'I knew those guys in District N were dragging their feet'."

Was District N really dragging its feet? Or was it due to, say, a lack of resources? Decision and regression trees only pick up what correlates most highly; they do not consider confounding, which is what we should look at when considering whether or not to blame District N.

This is an excellent and early example of the dangers of interpretability (let alone explainability): if end users do not understand that a model is built only on correlations, then they can be misled by "interpreting" causality.

There are some explicit acknowledgments of this tension.

> "one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern."[6]

> "Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation."[5]

> "Another problem is that such an interpretation might explain the behavior of the model but not give deep insight into the causal associations in the underlying data... The real goal may be to discover potentially causal associations that can guide interventions."[9]

POSITIVE EXAMPLE OF VALIDATION: CARDOSO ET AL. (2016)[4]

"OUTCOME REASONING"

Many, perhaps most problems, in healthcare involve interventions, and intervention requires knowledge of causality, such that AI may actually not be appropriate. To take the example of hospital readmission, we don't actually want to predict hospital readmissions, but rather, identify *preventable* hospital readmissions.[11] Even if we had a model that perfectly predicted hospital readmissions, if it could not guide us towards how to intervene to *decrease* readmissions, then it would not necessarily be of any use.

Problems that can be addressed only using correlations have been called "prediction policy problems,"[8] "y-hat problems,"[12] and more recently, "outcome reasoning."[2] Despite multiple arguments that such problems are widespread, are they really? Probably not; but they do at least exist.

An exemplary case is that of a "gene signature" for metastatic breast cancer. In 2002, van't Veer et al[16] published a study that used a custom statistical decision tree to find 70 genes that correlate with metastatic breast cancer in 117 patients. In the retrospective data, this model was optimal (better than clinician decision-making at identifying metastatic breast cancer); so should we use it? Certainly, it is worth running a clinical trial. It took 14 years to get to running such a trial and publishing the results,[4] but the setup and result are instructive.

SETUP

If the ML model and the clinical judgment both agreed on high risk for metastatic breast cancer, patients were treated with chemotherapy. If both agreed on low risk, patients were not. For those patients where the model and clinical judgment disagreed, patients were randomized into chemotherapy.

RESULT

They found that chemotherapy was *worse* for patients with high ML risk, and low clinical risk! But for those with high clinical risk and low ML risk, chemo was similar (so, avoid unhelpful pain). Conclusion: the ML would be a poor *replacement* for clinical judgment, but can help *catch clinical false positives*.

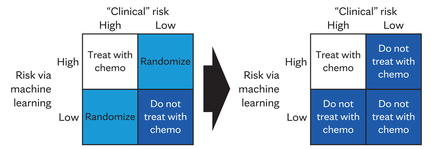

Fig 2. Schematic diagram of experimental setup and result.

LESSONS

Without this RCT, we would never know how to properly use the correlations! There are issues, like how the true target should be "gene signature for responding to chemotherapy", and how the initial model was made from very little data, but in general this is a good exemplar.

CAVEATS

MY DEFINITION OF ARTIFICIAL INTELLIGENCE

My proposed definition is entirely about *supervised statistical machine learning*. It excludes unsupervised machine learning (synonymous with clustering in statistics), and it also excludes "Good Old-Fashioned AI" (GOFAI) like expert systems and symbolic approaches.

However, the AI that is the focus of hope and anxiety is almost exclusively supervised statistical machine learning. Even embeddings, which are sometimes classed as "unsupervised", are calculated by models to impute held-out words (i.e., the held-out words are an outcome). GOFAI did not prove scalable and/or empirically effective.

Could a different type of AI, not based on supervised statistical machine learning, arise in the future? Perhaps, but then we should address it based on how *it* works. Technology-agnostic approaches make it difficult to address the actual issues, which makes irrelevant any potential "future-proofing" of more open definitions.

There is also causal learning, which seeks to infer causal structure. But this is based on very strong and unrealistic assumptions[7] and should be used with great caution.

IS STATISTICS CAUSAL?

Traditional statistical modeling does not necessarily succeed in modeling causality.[14] Observational causal inference techniques from econometrics, like instrumental variables, propensity score matching, and regression discontinuity design, are reasonable but can dramatically fail without us knowing.[1]

Still, despite the seeming impossibility of modeling causality, if our goals really are causal, then we should be doing this rather than thinking we can find an adjacent problem that solvable with correlations and apply that solution.

Lastly, machine learning can be used as a part of causal inference techniques, including propensity score matching and related techniques of g-estimation, inverse probability of treatment weighting (IPTW), Targeted Maximum Likelihood Estimation (TMLE), and "Double Machine Learning."

I do not include these as part of AI, as these require the same sort of careful causal setup as does interpreting causality from statistical models. Still, these techniques, especially when used with Structural Causal Models (SCMs), are a good way to do "cautious causal inference".[13]

WHAT TO DO?

1. Recognize "Artificial Intelligence" as *an instrumental use of statistical correlations*. In automating tasks, AI works—and can fail—in all and exactly the ways that correlations work and can fail. We can reason about potential failure points by thinking of where and how correlations can break down.
2. *Do not take explanations of how AI models arrive at outputs as explanations of how the world works*—AI models are based on correlations, but understandings of the world are in terms of causality. We can (and inherently do) easily concoct plausible causal stories to explain these correlations and see that they "make sense", but we should avoid doing so (correlations is not causation!).
3. If a user, e.g., a clinician, would only accept ("trust") a model based on causality, *then AI is unacceptable*, whether or not it is explainable/explained. There might be significant work to do in convincing clinicians that there are times and places where we do not need causality, and finding how to determine this; but *this* is what we need, rather than addressing technological reluctance or skepticism.
4. Having explanations of model outputs is neither a necessary not a sufficient basis for relying on AI. **We ought to accept or reject AI based on rigorous, out-of-sample (preferably experimental) tests of performance, not on whether explanations are present and "make sense".**

REFERENCES

1. Arceneaux K, Gerber AS, Green DP. A cautionary note on the use of matching to estimate causal effects: an empirical example comparing matching estimates to an experimental benchmark. Sociol Methods Res. 2010;39(2):256–82. doi: 10.1177/0049124110378098.
2. Baiocchi M, Rodu J. Reasoning using data: two old ways and one new. Obs Stud. 2021;7(1):3-12. doi: 10.1353/obs.2021.0016.
3. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001 Aug;16(3):199-231. doi: 10.1214/ss/1009213726.
4. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med. 2016 Aug 25;375(8):717-29. doi: 10.1056/NEJMoa1602253.
5. Caruana R, Lou Y, Gehrke K, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15); 2015 Aug 10-13. New York: ACM Press; 2021. p. 1721–30. doi: 10.1145/2783258.2788613.
6. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608 [Preprint]. 2017 Mar 2: [13 p.]. Available from: https://arxiv.org/abs/1702.08608.
7. Freedman DA. Graphical models for causation, and the identification problem. Eval Rev. 2004 Aug;28(4):267-93. doi: 10.1177/0193841X04266432.
8. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. Prediction policy problems. Am Econ Rev. 2015;105(5):491-5. doi: 10.1257/aer.p20151023.
9. Lipton ZC. The myth of model interpretability. KDnuggets. 2015 Apr;15(13). Available at: https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html.
10. Malik MM. A hierarchy of limitations in machine learning. arXiv: 2002.05193 [Preprint]. 2020 Feb 29: [68 p.]. Available from: https://arxiv.org/abs/2002.05193.
11. Marafino BJ, Schuler A, Liu VX, Escobar GJ, Baiocchi M. Predicting preventable hospital readmissions with causal machine learning. Health Serv Res. 2020 Dec;55(6):993-1002. doi: 10.1111/1475-6773.13586.
12. Mullainathan S, Spiess J. Machine learning: An applied econometric approach. J Econ Perspect. 2017;31(2):87-106. doi: 10.1257/jep.31.2.87.
13. Ogburn EL, Shpitser I. Causal modelling: the two cultures. Obs Stud. 2021;7(1):179-83. doi: 10.1353/obs.2021.0006.
14. Pearl J. Causality: models, reasoning, and inference. 2nd ed. Cambridge (UK): Cambridge University Press; 2009. Epilogue, The art and science of cause and effect; p. 401-28.
15. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi: 10.1038/s42256-019-0048-x.
16. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-6. doi: 10.1038/415530a.