Should Machine Learning Models Report to Us When They Are Clueless?

Roozbeh Yousefzadeh,[1*] Xuenan Cao[2]

[1]Yale Center for Medical Informatics and VA CT Healthcare System
[2]Yale MacMillan Center for International and Area Studies
New Haven, CT 06510, USA

[*]E-mail: roozbeh.yousefzadeh@yale.edu.

The right to AI explainability has consolidated as a consensus in the research community and policy-making. However, a key component of explainability has been missing: extrapolation, which describes the extent to which AI models can be clueless when they encounter unfamiliar samples (i.e., samples outside a "convex hull" of their training sets, as we will explain down below). We report that AI models extrapolate outside their range of familiar data, frequently and without notifying the users and stakeholders. Knowing whether a model has extrapolated or not is a fundamental insight that should be included in explaining AI models in favor of transparency and accountability. Instead of dwelling on the negatives, we offer ways to clear the roadblocks in promoting AI transparency. Our analysis commentary accompanying practical clauses useful to include in AI regulations such as the National AI Initiative Act in the US and the AI Act by the European Commission.

A consensus has consolidated in the research community and policy-making about the right to reasonable explanations for people affected by decisions made by Machine Learning and Artificial Intelligence models (*1, 2*). In 2020, the National Artificial Intelligence Initiative Act in the United States recognized the need to improve the reliability of artificial intelligence methods. In 2021, the AI Act by the European Commission drafted a highly sophisticated product safety framework to rank and regulate the risks of AI-driven systems. Both acts hover above the key concern of the right to explanation without landing precisely on it. One fundamental element about the right to explanation has been neglected, that is, extrapolation, which AI and ML models frequently perform. We propose that regulations incorporate articles requiring AI and ML models to report "overconfidence:" the output of an automated system should clarify whether it has extrapolated or not, and in which directions.

AI and ML, broadly defined, is a set of mathematical methods automating the learning process. Using certain algorithms, a model learns from a training set (data on which the model is trained), then uses the learned phenomenon to make decisions and predictions in the world at large. In a medical setting, a model may learn from the clinical outcomes of a cohort of patients, and possibly predict with some accuracy for new patients that walk through the door of a hospital. It would be commonsensical for any health provider to inquire how a new patient compares with the cohorts of patients in the training set and whether the new patient's information falls within the range of information in the training set. Extrapolation is a mathematical concept describing just that. In an extreme case of extrapolation, a new patient could have some rare and complicated form of liver disease that the model has never seen before, and therefore, the model's output for this patient may not be reliable. If a nurse encounters a patient with features that he has never seen before, he may elevate the situation to an expert physician. A physician may also need to elevate certain cases to a committee of experts. One would expect, quite reasonably, a nurse or a physician to elevate such cases and seek further expertise. However, this

procedure of escalating and reporting has so far eluded the attention of the research community and has been overlooked in regulations guiding the use of AI.

In math, there are well-defined algorithms for verifying whether a model is extrapolating, and if so, in which directions and dimensions. A training set, however small or large, forms a convex hull. Think of it as a dome. Any new sample, e.g., information about a new patient, will either fall within that convex hull or outside it. When a new data point is outside the convex hull of its corresponding training set, a model will need to extrapolate to process it. Conversely, when a new data point is within the convex hull of its training set, the model would interpolate. The concept of convex hull dates back to at least Isaac Newton (*3*). Extrapolation also has a rich literature in pure and applied mathematics (*4*) and cognitive science and psychology (*5*).

Whether a model has extrapolated is a piece of information lying at the heart of the right to explanation. In automated decision-making, if a model is making vital decisions or predictions about a patient with features not similar to any sample it has seen before, the model should be mandated to report and elevate the case to human experts.

Consider, in a case of loan applications being decided by an automated system, extrapolation might happen for an applicant because she is an immigrant, relatively young, and very well educated and the model has not seen any profile of an immigrant as young and educated in the training set. In such a case, the model might not make a sound decision and it would be reasonable to have a loan officer look over the model's decision. If an automated process rejects a loan application while extrapolating, the model should report the direction and extent of extrapolation.

In the research community, there have been discussions about whether machine learning models interpolate or extrapolate. Some researchers assume that models are predominantly interpolating between their training samples (*6, 7*) (think of it as under the dome or within the convex hull) and do not often extrapolate. All the datasets we have investigated prove to be ex-

trapolating frequently enough to be taken seriously. On the other hand, a group of researchers recently reported that in datasets with more than a 100 features, learning always amounts to extrapolation (*8*). This notion is realistic, but two issues arise. First, it leaves out many applications where datasets have less than 100 features. Second and more importantly, this position can be used to trivialize extrapolation. Scholars have argued that since extrapolation happens frequently, it must be trivial. Our results show the opposite. If we continue to believe that extrapolation is trivial, people affected by it may not be entitled to know about this fundamental issue.

Many applications of AI and ML are based on datasets with 10 to 50 features. Extrapolation in such applications is not trivial, nor negligible. When we studied, for example, the adult income dataset, a benchmark case for studying social applications of machine learning, about half of its testing samples required some extrapolation. Some of these extrapolations may be considered negligible, but for a considerable portion of testing samples, the extent of extrapolation is far from negligible. We see the case of a woman in the US workforce originally from Thailand, with high education in a managerial position, but in the lower-income bracket. The training set of this dataset did not have any sample close to her, so significant extrapolation in the dimensions of age, native-country, race, education level, and weekly work hour has to happen by any model trained on this dataset. We projected this woman's information to the convex hull of training set and saw that in these dimensions, both collectively and individually, the projections significantly differ from hers. What we see is not just an outlier here – such levels of extrapolation are neither rare nor predominant. Consider another case in the healthcare domain. We investigated a dataset from the Veterans Affairs Healthcare System (*9*) with more than one million patient records. Performing a 5-fold cross-validation, about 15% of patient records in the testing set required extrapolation. For many of these, extrapolation was too extensive from the medical perspective to be considered proper. These trends persist in all the other datasets we

studied. Extrapolation cannot be dismissed as trivial. In any respect, the affected person should have a right to know that model extrapolated when it made that decision for her.

Extrapolation can lead to good decisions. We do not suggest the prohibition of extrapolation. Models extrapolate, inevitably. Yet, before experts claim the benefits of extrapolation, information about extrapolation can be made available. The issue we raise here is promoting transparency by making information about extrapolation readily available. When a nurse encounters a patient with unfamiliar features, we expect this to be noted and reported to the physician. Analogically, when a model makes a decision, whether it has extrapolated or not, the information about extrapolation should be explicit, not hidden.

Determining whether a model has extrapolated adds a negligible computational burden. So the proposal does not add roadblocks but helps resolve issues of distrust. Transparency about extrapolation will increase the trust in using these automated systems. For example, in the medical setting, increased transparency can help gain the confidence of expert physicians. Understandably, physicians may not be willing to use automated systems unless the model provides adequate explanations about its recommendations.

The community recognizes that AI and ML models may have shortcomings and unacceptable biases (*10, 11*). In the past two decades, data collection from various realms of life, together with growing computational power, has allowed the practice of learning from data to spread widely, leading to the emergence of a field known as data science, ML, and AI. This widespread practice can be viewed as a democratization of mathematical modeling and data analysis as researchers from one discipline often contribute to other disciplines by way of deploying AI and ML tools. Yet this democratization has sometimes happened at the expense of domain expertise and interpretability. Models that fall under the umbrella of AI and ML are usually complex mathematical functions that are difficult to interpret (*12*), hence the proper name "black-box models." Requiring explanations about the rationale behind the model's deci-

sions has entered the public policy domain and regulations, but the knowledge about whether a model has extrapolated or not has been neglected. Regulations of AI explainability could benefit by adding practical clauses that AI and ML models should report when they extrapolate, and potentially also the direction and extent of their extrapolation.

References

1. S. Wachter, B. Mittelstadt, *Columbia Business Law Review* pp. 494–620 (2019).

2. D. Coyle, A. Weller, *Science* **368**, 1433 (2020).

3. I. Newton, D. T. Whiteside, *et al.*, *The Mathematical Papers of Isaac Newton* (2008).

4. C. Brezinski, M. R. Zaglia, *Extrapolation methods: theory and practice* (Elsevier, 2013).

5. R. Yousefzadeh, J. A. Mollick, *Workshop on Human and Machine Decisions at NeurIPS* (2021).

6. M. Belkin, D. Hsu, S. Ma, S. Mandal, *Proceedings of the National Academy of Sciences* **116**, 15849 (2019).

7. T. Webb, *et al.*, *International Conference on Machine Learning* (PMLR, 2020), pp. 10136–10146.

8. R. Balestriero, J. Pesenti, Y. LeCun, *arXiv preprint arXiv:2110.09485* (2021).

9. A. C. Justice, *et al.*, *Medical Care* **44**, S13 (2006).

10. C. Rudin, *Nature Machine Intelligence* **1**, 206 (2019).

11. B. Eshete, *Science* **373**, 743 (2021).

12. R. Yousefzadeh, D. P. O'Leary, *La Matematica* (2021).