# Interpretable Machine Learning

Andreea Teodora Patra

As stated by Brunton and Kuts (1): ‘Data science is not replacing mathematical physics and engineering, but is instead augmenting it for the twenty-first century, resulting in more of a renaissance than a revolution’. In classical artificial intelligence (AI), researchers would write programs that encode the knowledge needed to carry out a particular task. Over the years, this knowledge focused approach has been replaced by machine learning (ML) which tries to learn from data by statistical means (2). Using the information obtained, it would be programmed to forecast or perform various tasks such as speech recognition. However, the core problem throughout time remained trust. The European regulations imposed in 2018 called for safe and secure AI in safety critical tasks, therefore requiring data scientists to provide explanations for their algorithms. This paper will explore situations in which machine learning models can be trusted through an analysis of academia papers.

Modern machine learning depends greatly on the exact details of large training sets and if it encounters something too different from what it has experienced before, it would produce unacceptable results. For instance, in 2016, IBM claimed that Watson, the AI system that won at the game "Jeopardy!" would be able to read medical literature and make recommendations that human doctors would miss. However, in less than 2 years, the project was cancelled as the diagnosis was considered unsafe and incorrect (2). Stanford computer scientist Judy Hoffman has also shown that a self-driving car whose vision system was trained in one city can do significantly worse in another (2). Despite this history of missed milestones, machine learning has been developing every day with examples such as deep learning that is used in a wide range of applications from predicting earthquake aftershocks to object recognition. This success has been driven by two major factors: hardware and big data, gigabytes or terabytes of data. Advances, which allowed for memory efficiency and parallel computing, helped process the vast amount of data and they can be seen today in the form of Hadoop, MapReduce, Elasticsearch, GPU or neuromorphic hardware.

A surge in interpretable machine learning has been stimulated by the appearance of trojaning attacks. Targeted errors can be achieved as any valid input can be stamped using a trojan trigger, which is a mutation of the initial model (3). This could result in a security camera that relies on face or voice recognition to be compromised. Furthermore, scientists proved that it is possible to make use of evolutionary algorithms of gradient ascent to create images unrecognizable to human eyes than Deep Neural Networks classify them with 99% probability (4). Despite its state-of-art results, deep learning is considered a ‘black box’ model and its main approach is learning. It follows a process of trial and error and adjustment that strengthens the weights for a particular configuration of inputs to a particular output (2). Thus, experts struggle to understand why particular neural networks make the decisions they do. In rare cases we have insights into the behaviour of nodes due to millions of numerical parameters. Unfortunately, these increase in attacking methods is not the only challenge of machine learning. Gary Marcus and Ernest Davis (2) put also an emphasis on change in the domain over time, racially biased training set and the programmer's attempt to precisely anticipate any perturbations.

At the moment, models are evaluated using accuracy metrics, but the real world is significantly more complex, and those specific metrics may not be indicative of the product's goal. Before diving into the various explanation techniques, we need to have a clear understanding of the terms that define a machine learning algorithm as reliable. Interpretability is defined as the explanation of internals of a system in a way that is understandable to humans ((5),(6)). It should efficiently answer the question (7) : "Can we understand on what the ML algorithm bases its decision?". Transparency is the understanding of the mechanism by which the model works ((5)). For example, SVM can become non-transparent due to the decision to use RBF kernel on Euclidean distances. LIME algorithm (8) gives an interpretable representation of a 'black box' model by approximating it locally with an interpretable model (decision trees, linear models, Naïve Bayes). The approximation is done by constructing a linear model on the weighted sample in the neighbourhood of an instance. For example, if our data would contain symptoms of patients and the model's goal is to try and predict whether the patient has a particular disease, the interpretable representation would be a vector of symptoms. Not only we would be able to assess the faithfulness of our model, but we would also be able to decide which features help generalize. Even if one can understand the benefits of explainable machine learning, Herman (9) notes that human evaluations imply a strong bias towards simpler descriptions. As a consequence, questions like "When would it be unethical to simplify an explanation to better persuade users?" will arise. Despite many successful cases of simple, transparent models ((12),(13)), they cannot be applied to more complex problems such as object recognition. In these cases, interpretability comes at the cost of efficiency and accuracy.

In conclusion, in order for machine learning models to achieve acceptance, they need to provide satisfactory explanations of their decisions and underlying mechanisms. Caution will always be required when working towards scaling our models and deploying them into the real world. Data scientists would need to account for shifts from the training data as well as for biases and a greater diversity of evaluation metrics should be used. From my point of view, a change in scientist belief's will be needed. As many experiments proved (8), a high accuracy does not always result in a model that generalizes. Developments in the methods used to develop explanations of models are impressive, but we still need a way to assess that explanation and provide a quantification of the uncertainty.

# References

[1] Steven L. Brunton and J. Nathan Kutz, *Data-Driven Science and Engineering*, Cambridge University Press, 2019. doi: 10.1017/9781108380690

[2] Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence we can trust*, Pantheon Books, 2019

[3] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, *Trojaning attack on neural networks*, 2017.

[4] Nguyen A, Yosinski J, Clune J, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.* In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015

[5] Ribana Roscher, Bastian Bohn, Marco F. Duarte and Jochen Garcke, *Explainable Machine Learning for Scientific Insights and Discoveries*, IEEE Access 8 (2020): 42200–42216.

[6] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, Lalana Kagal, *Explaining Explanations: An Overview of Interpretability of Machine Learning*

[7] C. Casert, T. Vieijra, J. Nys, and J. Ryckebusch, *Interpretable machine learning for inferring the phase boundaries in a nonequilibrium system* Physical Review E, 99(2):023304, feb 2019. doi: 10.1103/PhysRevE.99.023304.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, *Why should i trust you?: Explaining the predictions of any classifier*, arXiv:1602.04938v3

[9] B. Herman, *The promise and peril of human evaluation for model interpretability* arXiv:1711.07414, 2017.

[10] Finale Doshi-Velez and Been Kim, *Towards a Rigurous Science of Interpretable Machine Learning*, arXiv:1702.08608v2

[11] Forough Poursabzi-Sangdeh, Daniel G.Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan and Hanna Wallach, *Manipulating and Measuring Model Interpretability*, arXiv:1802.07810v3

[12] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Knowledge Discovery and Data Mining (KDD)*, 2015

[13] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, *Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model*, Annals of Applied Statistics, 2015.