

A Survey of XAI Methods and Alignment with Autonomous Vehicles

David Kaczynski
University of Michigan

Abstract—This is my abstract. So, I’ve been looking a lot at the interpretation and explanation of black box machine learning methods, namely, neural networks, because they’re mysterious, fascinating, and a little bit scary. There are lots of options available for tools for developing neural networks, but the workflow of a data scientist is pretty well defined: data preparation, feature extraction, dimensionality reduction, training, testing, and analysis of results. We are specifically going to be focusing on a specific set of methods for analyzing results known as XAI, eXplainable Artificial intelligence. In this paper, we explore the history and background that has popularized the use of black box machine learning methods, their relationships with society, the problems that XAI hopes to solve, and the challenges it faces in doing so. Then, we take our exploration of XAI a step further by aligning its use cases with the development and use of autonomous vehicle technology.

I. INTRODUCTION

A. Soft introduction. ML/AI is hot stuff, smart cars, autonomous homes, internet of stuff, etc etc

Machine learning and artificial intelligence (AI/ML) have gained mainstream attention as consumer products featuring these technologies permeate into everyday life. Healthcare, transportation, supply chains, stock markets, social media, national security, genetic engineering, political science, and smart homes are all applying AI/ML to make faster, more accurate decisions and to automate tasks that previously required a human expert [citations needed]. As a wider audience is exposed to AI/ML, new questions are on the tips of people’s tongues, like “what is the difference between artificial intelligence and machine learning?” and “do I need to be worried about computers making decisions that directly impact my health, security, and privacy?”.

B. What is XAI? What is machine learning? What can it do? What is a black box model?

While artificial intelligence is a more familiar term and has a broader, amorphous definition, machine learning is a branch of AI that can be defined relatively explicitly. Machine learning is the algorithmic processing of training data to create a computer program that can be used for repeatable tasks, such as making decisions or extracting insights from data. In this sense, the computer can said to be “learning” by looking at existing data to create a model that can be applied to new data. Some methods, such linear regression, are old news scooby doo. Even the hot topic of neural networks was introduced while you were in grade school. What **is** new is the availability of data, the raging locomotive research interest,

and powerful GPUs and distributed commodity hardware to actually perform large scale AI/ML activity. Despite AI being directly in the name, the field of XAI typically is more focused on so-called “black box” machine learning models, such as deep neural networks.

C. What is an explainable model? Why are block box models growing in popularity? What are they used for?

Deep learning is a relatively new set of methods of machine learning based off an older concept called neural networks. Thanks to advancement in GPU technology, cheaper and faster processing of linear equations has become affordable and available to a wider audience of consumers, including researchers, hobbyists, and professionals. They excel and these things but not at those things, but hey, they’re even getting better at those things like creeping us out with people faces.

D. Why do we need XAI? Why should we care about “cracking the black box”? What can it be used for?

Love it or hate it, AI/ML are like the Patriots: they’re here, they’re winning, and there’s nothing you can do about it. Our lives are on the line. Seriously. Medical diagnosis, autonomous vehicles, personal and private details from the IoT...for safety, trust, and progress, XAI is srs, srsly.

E. Organization of paper

1) *Background:* Specifically, what are existing methods of XAI? Also, there are some other pretty darn thorough surveys on XAI.

2) *Use Cases:*

3) *Challenges:*

4) *Alignment with Autonomous Vehicles:*

5) *Conclusion and Future Research:*

II. BACKGROUND

Currently, there is no formal definition of what is or is not an explanation of the output of a machine learning model.

Let’s summarize these surveys

- Comprehensive survey of visualizations as XAI techniques [1]
- Methods of LRP & heatmaps and their various use cases in research[2]
- A comprehensive survey of a variety of XAI techniques, includes sensitivity analysis (like heatmapping), rule and

model extraction, activation maximization, and more [3], but it doesn't mention verbalization as a technique

- Network graphs plotting relationships between topics and sub-topics related to XAI [4]
- Another broad survey of techniques in the field of XAI along with a discussion on its history and fundamental concepts, but again, no mention of verbalization as a technique[5]

A. Visualization Techniques

Introductory paper of LRP [6]

Activation Maximization [7]

B. Verbalization of Explanations

Verbalization of CNN decisions [8]

An introduction of counterfactual explanations as a method of explaining decisions without interpreting "black box" internals [9]

C. Data Provenance

Data provenance to trace decisions and analysis to the raw input data

Calculating the contribution of individual training samples to a trained model's decision [10]

D. Model Induction

I'll need to put something here

III. USE CASES

We've laid out various methods of XAI, but the question still remains of how do these methods provide value to anyone. Here, we explore the various users who interface with XAI and what value is provided.

A. As a data scientist, I can use the explanation of model decisions to identify opportunities to improve the training dataset and make the model more robust

- Montavan et al provide multiple real examples of researchers using a popular XAI method called Layerwise Relevance Propagation (LRP) for convolutional neural networks [2]. In one example, it was discovered that a computer vision classifier was classifying horses based on a watermark that was present in the training dataset.
- The image of a dog was misclassified as a wolf [11]. Feature relevancy revealed that the snowy ground around the dog was deemed as the most relevant input pixels. The model can be improved by adding images of dogs with snowy backdrops to the training dataset.
- Liu et al focus on interpreting the decisions of a model used to identify cybersecurity threats [12]. The researchers interpret the model's decisions that failed to successfully label threats. The interpretations help the researchers identify how to generate specific perturbations in the training data that improve the training of the model

against previously misidentified threats. The accuracy of the model improves.

- LAMP is a tool that can trace the decision of a trained model to the importance that individual samples had on that decision. [10]

B. As a potential consumer of an AI/ML product, such as an autonomous vehicle or virtual assistant, an explanation of how the product is making its decisions will improve my trust in the product

- Human subjects are asked about how much they trust a classifier before and after its explanations are made available [11].
- Decisions from autonomous vehicles and advanced driver assistance systems (ADAS) may be explained using XAI methods to answer the questions "How?" and "Why?". Koo et al measure both the emotional impact and impact in making safer decisions when drivers are provided answers to these questions [13]

C. As a provider of an AI/ML product, I am responsible for providing an explanation for how my product makes its decisions in the case of being the subject of an investigation or defendant in a lawsuit

- Human subjects conduct fictitious banking scenarios, both with and without discrimination-aware data mining (DADM) tools [14]. The accuracy and presence of discrimination in decisions was compared across tools.
- Autonomous vehicles and vehicles with ADAS from Tesla, Google, and GM Cruise have been involved with numerous traffic incidents, ranging from trivial to fatal [15] [16] [17] [18]
- Researchers discuss non-technical challenges and high-level technical challenges in investigating transparent model design in private industry [19]
- In the criminal justice system, judges and parole boards may apply predictive models as tools in making decisions. Racial discrimination has been uncovered in such tools [20] [21].
- A magazine article describes the GDPR's "right to explanation" and includes a survey of research in identifying and rectifying discrimination [22]

D. As a member of a team of data scientists, I want to improve collaboration and reduce the duplication of effort by being able to effectively store, query, and trace how and on what data our AI/ML models were trained, and, potentially, be able to trace a decision to the relevancy of each training item

- Researchers develop a tool on the distributed computation framework Spark that traces individual records through transformations and their relationship to aggregated or derived values [23].
- A team at Amazon developed a platform and storage schema for effectively persisting and querying the activity of heterogeneous ML tools [24], providing the ability to replicate training of models and a platform for conducting analytics across development of ML models.

IV. ALIGNMENT WITH AUTONOMOUS VEHICLES

The use cases defined in III span users with a wide

V. CHALLENGES IN XAI

- Feature relevancy can be compromised by using a mask on the input that is virtually undetectable to the human eye [25].
- The concept of an explanation is subjective; there is no quantitative way of saying "yes, this decision is explained" or "no, this decision is not explained." [26]
- Due to the diverse audience of XAI, an explanation may be too technical or not technical enough to people of different backgrounds.
- Private industry is rarely incentivized to expose algorithms and ML models to researchers [19].

REFERENCES

- [1] F. M. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 1–20, 2018.
- [2] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1 – 15, 2018.
- [3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, pp. 93:1–93:42, Aug. 2018.
- [4] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), pp. 582:1–582:18, ACM, 2018.
- [5] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [6] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, pp. 1–46, 07 2015.
- [7] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *NIPS*, 2016.
- [8] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 3–19, Springer International Publishing, 2016.
- [9] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *CoRR*, vol. abs/1711.00399, 2017.
- [10] S. Ma, Y. Aafer, Z. Xu, W.-C. Lee, J. Zhai, Y. Liu, and X. Zhang, "LAMP: Data provenance for graph based machine learning algorithms through derivative computation," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 786–797, 2017.
- [11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 1135–1144, ACM, 2016.
- [12] N. Liu, H. Yang, and X. Hu, "Adversarial Detection with Model Interpretation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, (New York, NY, USA), pp. 1803–1811, ACM, 2018.
- [13] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 9, pp. 269–275, Nov 2015.
- [14] B. Berendt and S. Preibusch, "Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence," *Artificial Intelligence and Law*, vol. 22, pp. 175–209, Jun 2014.
- [15] R. Read, "For the first time, google admits its autonomous car is at fault in fender-bender," *Washington Post*, 2016.
- [16] The Tesla Team, "An update on last week's accident," <https://www.tesla.com/blog/update-last-week>
- [17] E. Ackerman, "Fatal tesla self-driving car crash reminds us that robots aren't perfect," *IEEE Spectrum*, 2016.
- [18] P. Bhavsar, K. Dey, M. Chowdhury, and P. Das, "Risk Analysis of Autonomous Vehicles in Mixed Traffic Streams," tech. rep., 2017.
- [19] M. Veale, M. Van Kleek, and R. Binns, "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), pp. 440:1–440:14, ACM, 2018.
- [20] R. Wexler, "When a Computer Program Keeps You in Jail," *The New York Times*, June 2017.
- [21] J. Angwin, L. Jeff, S. Matta, and L. Kirchner, "Machine bias," *Pro Publica*, 2016.
- [22] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, pp. 50–57, 2017.
- [23] M. Interlandi, A. Ekmekji, K. Shah, M. A. Gulzar, S. D. Tetali, M. Kim, T. Millstein, and T. Condie, "Adding data provenance support to Apache Spark," *VLDB Journal*, vol. 27, no. 5, pp. 1–21, 2017.
- [24] S. Schelter, J.-H. Böse, T. Klein, and S. Seufert, "Automatically Tracking Metadata and Provenance of Machine Learning Experiments," in *Machine Learning Systems Workshop*, 2017.
- [25] X. Zhang, N. Wang, S. Ji, H. Shen, and T. Wang, "Interpretable deep learning under fire," *CoRR*, vol. abs/1812.00891, 2018.
- [26] A. Bibal and B. Frénay, "Interpretability of machine learning models and representations: an introduction," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 04 2016.
- [27] V. Curcin, E. Fairweather, R. Danger, and D. Corrigan, "Templates as a method for implementing data provenance in decision support systems," *Journal of Biomedical Informatics*, vol. 65, pp. 1–21, 2017.
- [28] S. Lapuschkin, A. Binder, K.-R. Müller, and W. Samek, "The LRP Toolbox for Artificial Neural Networks," *Journal of Machine Learning Research*, 2016.