A Survey of XAI Methods and Alignment with Autonmous Vehicles

David Kaczynski
University of Michigan

***Abstract*—Artificial intelligence and, especially, machine learning have gained notoriety in recent years do to the influx in ways that they impact the people's lives. AI/ML can replace and supplement human experts and provide automation in ways like never before, such as medical diagnosis, smart home technology, and autonomous vehicles. With new roles like these so close to the personal lives of the general public, there has been growing concern from political and technological leaders about putting too much trust in the automated decision-making process of AI/ML systems or about the biases that they can learn through the training process. The internals of some AI/ML methods are inherently difficult to interpret, e.g. deep neural networks. The combination of the societal interest in making AI/ML systems more accountable combined with the difficult-to-interpret nature of some decision systems has brought about a new field of research frequently referred to as explainable artificial intelligence, or XAI. There is currently sparse literature on the overlap between XAI and autonomous vehicles. In this survey, the relationship between the demand for and value provided by XAI is explored, and we align those use cases with the various backgrounds of engineers, consumers, and auditors of autonomous vehicles.**

I. Introduction

A. History and Concepts

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 [add citations here]. 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?".

Artificial intelligence is a broad term with an amorphous definition that changes over time, but machine learning can be more explicitly defined: 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 and by creating a model that can be applied to new data. While the term "machine learning" has been added to the English lexicon only recently, the foundation of machine learning was paved as far back as the 19th century. Linear regression has its roots in the work of mathematicians Legendre and Gausse in the early 1800s [add citation], and even the hot topic of training neural networks was published in the 1970s [1]. The catalyst for bringing these methods to the forefront of modern methods of AI and automation is the combination of the broad availability of data and powerful data centers of compute resources 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.

Deep learning is a relatively new branch of machine learning methods based off the older concept neural networks. Thanks to advancement in GPU technology, the training of deeper, more complex neural networks has become affordable and available to common individuals, such as researchers, hobbyists, and data science professionals. These deep neural networks (DNNs) excel at optimizing the relationship between input and output variables from the training data without any input from a human expert describing rules or conditions. This process of training a DNN inevitably creates a model whose internals are not able to be understood via human inspection. While the internals of the model are opaque, the optimized relationship that the model learned from the training data is able to quickly make decisions without ever having received any direct rules or instructions, making it ideal in situations like computer vision and natural language processing where explicit rules are difficult to define.

Through progress and controversy, artificial intelligence and machine learning are interfacing with more people in more ways every year, and there is a general awareness being raised about the legitimate concerns of the trustworthiness of these technologies. Medical diagnosis, autonomous vehicles, and national security are just a few examples of ways in which machine learning can directly impact the health and well being of individuals' lives. XAI is a new tool in the frontier of establishing traceable, trustworthy, and accountable decision systems.

B. Organization of Paper

This paper is organized into the following sections.

- *Background*: Specifically, what are existing methods of XAI? Also, there are some other pretty darn thorough surveys on XAI.
- *Use Cases*: Methods of XAI can be used to the benefit users with a wide variety of backgrounds. We identify four general use cases for data scientists and consumers.

- *Alignment with Autonomous Vehicles*: The perspectives of both engineers and consumers of autonomous vehicles (AV) are considered as the relationship between XAI and AV are explored.
- *Challenges*: XAI is a relatively young field, lacking formality, and there are several legitimate criticisms against it. It is important to recognize the shortcomings of XAI and assess the current feasibility and progress in overcoming its obstacles.
- *Conclusion and Future Research*: A summary of the alignment between XAI and AV is presented along with an outline of how identified obstacles may be approached.

II. Background

XAI is the interpretation and explanation of machine learning models. In order to go further, the concepts of "interpretation" and "explanation" warrant a more formal definition:

- an **interpretation** is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of, and
- an **explanation** is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression) [2].

Interpretations are often presented in the form of a heat map layer on an image, natural language generated to either describe a decision or to describe a boundary of what a model can or cannot do, or an easier-to-explain model, such as a decision tree. An explanation is the relationship between a human-interpretable concept, such as words or shapes, some aspect of a machine learning model, such as how or why a decision was made. There is ambiguity and a lack of formality around the concept of an explanation that is discussed further in V Challenges in XAI.

Some trained ML models may have easier-to-explain internals than others. Decision trees are a common example of an explainable ML model. The output of the decision tree can be directly back to identify the values of specific features that lead the model to its decision. On the other hand, the internals of deep neural networks (DNN), by default, are opaque and are not directly interpretable. The any node in the output layer of a neural network has many inputs with many weights, each of which may have many more inputs with many weights. The value of a single input feature can propagate through potentially every node in a neural network, making it challenging to isolate the contribution of the value of a single feature in the decision of the model. In addition to the challenge in tracing a single feature through the a DNN, a single feature, such as the pixel in an image, may not have an isolated interpretation for humans to understand. In that sense, the relevancy of an isolated input feature may not have meaning or value from the human's perspective. Other difficult-to-explain machine learning models include support vector machines (SVM), random forests, and Gaussian belief networks.

There are various methods of XAI that can be applied in machine learning, both during or after the training of the ML model. A priori methods are those in which a traditionally black box model can be constructed in such a way that it either is easier to explain with other methods or can generate an explanation alongside its traditional output. Examples of generated explanations include using a LSTM DNN alongside a CNN to generate natural language explanations [3] or embedding prototypes of outputs classes directly in the CNN [4]. A posteriori methods of XAI include visualizations techniques such as Deep Taylor decomposition and Layer-wise Relevance Propagation to generate heat maps that can be over-layed on the original input to help identify patterns in the relevancy of input features, such as highlighting shapes or regions that either supported or detracted from the network's decision.

In this section, methods of XAI or organized into the types of artifacts that are generated: visualization, verbalization, data provenance, and model induction.

A. Visualization Techniques

1) LRP: Layer-wise relevance propagation (LRP) is an a posteriori method of generating heat maps, or saliency maps, to highlight the positive and, sometimes, negative relevancy of input features in the output layer, or decision, of a DNN. LRP is functionally a backward pass of the output layer back through the neural network to the input layer. At its core, LRP is not a mathematical function but a set of constraints that defines properties of the relationship between layers in a DNN. Data scientists can substitute existing or new activation functions to apply different highlight with different levels of sensitivity the relationship between the input and output layers of the DNN.

The heat maps generated by LRP are not limited to image inputs. Researchers have applied heat maps generate dy LRP to natural language, genetic sequences, and 3D models of molecules.

2) Activation maximization: Activation maximization is an a posteriori method for generating an input for a model that maximizes the activation of a specific output neuron[5], such as a class label. In theory, the input that maximizes the output would be some sort of ideal or target input, but in practice, the input that maximizes the activation of a neuron may not resemble other similar training samples. Further literature searching needs to be done in this area to identify cases in which valuable relationships were discovered via AM.

3) Prototypes in CNNs: Thus far in research, the concept of building a CNN with a prototype layer has been applied in image classification tasks. A prototype is essentially an image of one of the target labels of the classifier. The prototype image can either be a subset of an image from the training data or it can be images of the subject class from outside of the training data. When the prototype layer is constructed from subsets of training images, the theory is that each prototype represents some interpretable, defining characteristic of the class, such as a color pattern on a bird or the shape of ears on a bear.

Once a CNN has been constructed using a prototype layer, a data scientist can inspect the activation of various prototypes from that layer to gain insights into the importance of those characteristics in the CNN's decision.

B. Verbalization of Explanations

1) Generating explanation in parallel: Verbalization of CNN decisions [3]

2) Counterfactual explanations: An introduction of counter factual explanations as a method of explaining decisions without interpreting "black box" internals [6]

3) Rule-based decision systems: While rule-based decision systems may not be the result of a machine learning method of training, these rule-based systems excel at verbal explanations. Soar is a good example of this.

C. Data Provenance

Data provenance is the attribution of the origins of data and the transformation that it undergoes in its journey. Data provenance can be used to trace decisions and analysis to the raw input data. Also, data provenance can be used to analyze how specific training samples influence an ML model.

1) LAMP: Calculating the contribution of individual training samples to a trained model's decision [7]

2) Metadata persistence: Heterogenous tools are available for modern data scientists, but much of the workflow of a data scientist remains the same: clean, extract, train, and evaluate (oversimplified). Data provenance can be used to create relationships between the raw sources of data through all of the transformations and training that generated an ML model, along with relationships to various evaluation metrics. In this context, data provenance can be considered a method of XAI that explains how a model was created and how its performance compares to the performance of other models, all the way from raw training data to statistical analysis of test results. Architectures, data models, and wrappers have been developed to aid data scientists in capturing valuable metadata on the workflow of developing ML models.

D. Model Induction

Model induction is the generating a more explainable model, such as a decision tree, from a black box model, such as a DNN. Model induction has limit

E. Existing Surveys in XAI

Although XAI is a relatively young field of research, appearing in publications circa 2016, there are at least hundreds of publications on the topic along with accompanying literature surveys that draw from the breadth of research. Surveys vary in their perspective of XAI, the methods used to categorize literature, and the scope of methods that they cover.

Hohman et al. conduct a thorough literature search of visualization techniques of XAI [8]. The authors' perspective begins with an analysis of the taxonomy of questions that can be asked of machine learning models. The visualization techniques are labeled with various helpful metadata, such as what type of visualization is produced, if the XAI method is a priori or a posteriori, and more. This metadata on the visualization techniques allows the reader to easily map the various visualization techniques to the questions that they can answer.

Montavon et al. present a far more focused literature search on a specific method of visualization in XAI and how it has been applied by other researchers, which domains, and to what ends [2]. For a large portion of the paper, the method of Layer-wise Relevance Propagation (LRP) is broken down into its core mathematical principles so and trace those principles through a conceptual CNN to demonstrate how LRP may be applied to generate heat maps of input features to understand feature relevancy. The authors to provide examples of LRP being applied in domains such as computer vision, genetic engineering, and chemistry.

The approach of Abdul et al. is to plot the relationship of topics around the interpretation and explanation of machine learning through the use of network graph visualizations [9]. This contribution to the surveying of XAI literature provides a much needed compass since the terminology in XAI, and even the name of the field itself, is still evolving rapidly. We can also observe the relationship of topics in XAI as they weave through such domains as psychology, human-computer interaction, and the sociological aspects of accountability and fairness.

Guidotti et al. take a more broad approach to XAI methods by analyzing to what type of ML models have researchers been applying explanation methods, what were the interpretable domains of their inputs (e.g. images, natural language), and what methods of explanation were applied [10]. In this survey, a unique amount of attention was paid to methods of rule or model extraction, in which data scientists can extract a decision tree model or natural language rules as approximations of the original ML model.

Adadi et al. also take look at the methods of XAI as a whole and describe considerable detail on the non-visualization methods of XAI [11]. In addition to their discussion on the taxonomy of XAI methods, this survey tracks the history of the terminology around XAI, exposing the burgeoning growth of the field and providing insight on the demands being placed on the field. The authors classify the value provided by XAI into four different areas: explain to justify, explain to control, explain to improve, and explain to discover.

In some sense, the abundance of literature in such a short period of time speaks to the demand for further research in the field and to the wide-held interest across people of many backgrounds. Time will tell if XAI can deliver the value that people seek or, maybe, if societal impressions of AI/ML will evolve to no longer be concerned with the inner workings of machine learning.

III. Use Cases

We've layed 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 data set and make the model more robust

- Montavan et al provide multiple real examples of researchers using a popular XAI method called Layer-wise

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 data set.
- The image of a dog was misclassified as a wolf [12]. 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 datavset.
- Liu et al focus on interpreting the decisions of a model used to identify cyber security threats [13]. 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. [7]

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 [12].
- 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 [14]

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 [15]. 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 [16] [17] [18] [19]
- Researchers discuss non-technical challenges and high-level technical challenges in investigating transparent model design in private industry [20]
- 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 [21] [22].
- A magazine article describes the GPDR's "right to explanation" and includes a survey of research in identifying and rectifying discrimination [23]

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 [24].
- A team at from Amazon developed a platform and storage schema for effectively persisting and querying the activity of heterogenous ML tools [25], 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 Use Cases span users with a wide variety of backgrounds.

V. Challenges in XAI

- Feature relevancy can be compromised by using a mask on the input that is virtually undetectable to the human eye [26].
- 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." [27]
- 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 [20].

VI. Conclusion and Future Research

Apply XAI to IV *Alignment with Autonomous Vehicles* and also identify V *Challenges in XAI* that may be worked towards.

References

[1] P. Werbos and P. J. (Paul John, "Beyond regression : new tools for prediction and analysis in the behavioral sciences /," 01 1974.

[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] 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.

[4] C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: deep learning for interpretable image recognition," *CoRR*, vol. abs/1806.10574, 2018.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Ẅhy 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.

[13] 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.

[14] 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.

[15] 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.

[16] R. Read, "For the first time, google admits its autonomous car is at fault in fender-bender," *Washington Post*, 2016.

[17] The Tesla Team, "An update on last week's accident." https://www.tesla.com/blog/update-last-week

[18] E. Ackerman, "Fatal tesla self-driving car crash reminds us that robots aren't perfect," *IEEE Spectrum*, 2016.

[19] P. Bhavsar, K. Dey, M. Chowdhury, and P. Das, "Risk Analysis of Autonomous Vehicles in Mixed Traffic Streams," tech. rep., 2017.

[20] 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.

[21] R. Wexler, "When a Computer Program Keeps You in Jail," *The New York Times*, June 2017.

[22] J. Angwin, L. Jeff, S. Matta, and L. Kirchner, "Machine bias," *Pro Publica*, 2016.

[23] 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.

[24] 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.

[25] 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.

[26] X. Zhang, N. Wang, S. Ji, H. Shen, and T. Wang, "Interpretable deep learning under fire," *CoRR*, vol. abs/1812.00891, 2018.

[27] 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.