A Survey of XAI Methods and Use Cases with Autonomous Vehicles

David Kaczynski
University of Michigan

***Abstract*—Artificial intelligence and, especially, machine learning have gained notoriety in recent years due to the influx in ways that they impact 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 AI/ML adopting 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 processes or about the biases that they can inadvertently learn via training. 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 automated 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 to define cohesive use cases, and we align those use cases with the various needs of engineers, consumers, and auditors of autonomous vehicle technology.**

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, cyber security, bioinformatics, 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 [1][2][3][4][5][6][7][8][9]. 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. The method of least squares linear regression has its roots in the work of mathematicians Legendre and Gauss in the early 1800s [10], and even the hot topic of training neural networks was originally published in the 1970s [11]. 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 along with powerful data centers of compute resources to perform large scale AI/ML activity. Despite "AI" being directly in the name of XAI, the field is typically 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 of 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 (DNN) excel at optimizing the relationship between input and output variables from the training data without any guidance from a human expert describing human-intuitive 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 having received any explicit 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 and accountability of these technologies. The domain of autonomous vehicles is just a one example for 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.

The burgeoning field of autonomous vehicles is outpacing the legislative efforts of the EU and of individual states in the U.S.A. to effectively enforce autonomous systems to provide trustworthy and effective explanations for their decisions. It is in automakers' best interests to establish trust with consumers to facilitate the adoption of such as an invasive and revolutionary technology. There are also incentives for the developers, scientists, and engineers who create autonomous vehicles to employ XAI methods in order to create more accurate, robust, and safer automated 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 existing literature surveys with their own perspectives.
- *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.
- *Use Cases*: Methods of XAI can be used to benefit users with a wide variety of backgrounds. We identify three general use cases for data scientists, consumers, and auditors of AI/ML systems.
- *Alignment with Autonomous Vehicles*: Existing research is identified in which XAI methods are applied in the AV domain, and future research goals are outlined.
- *Conclusion*: A summary and final thoughts are presented on the alignment between XAI and AV .

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) [12].

Sample human-interpretable domains include visual aids, natural language, or a more easily interpretable model, such as a decision tree. An explanation is the relationship between a human-interpretable concept, such as words, images, or logic, and a query of how or why a machine learning model made a decision. There is ambiguity and a lack of formality around the concept of an interpretation that is discussed further in III *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 traced back to identify the values of specific features that lead the model to its decision. On the other hand, the internals of DNNs, by default, are opaque and are not directly interpretable. 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 virtually every node in a neural network, making it challenging to isolate the contribution of the value of a single feature in the output 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 a human's perspective. Other difficult-to-explain machine learning models include support vector machines (SVM), random forests, and Gaussian belief networks (see figure 1).

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 [14] or embedding prototypes of outputs classes directly in the CNN [15]. *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 outlines, regions, or other patterns 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 [12]. 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 (see figure 2).

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[16], 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 in the output layer may not lack human interpretability. As seen in figure 3, the effectiveness of AM is largely network-specific. However, the ideal input for a neural network may grant insights into what types of features a DNN favors for an output neuron.

3) Prototypes in CNNs: Thus far in research, the concept of building a CNN with an *a priori* prototype layer has been applied in image classification tasks [15]. A prototype is essentially an image or subset of an image of one of the target labels of the classifier, either from the training data or from an image of the target class from outside of the training data. Each prototype could target a characteristic of the target class

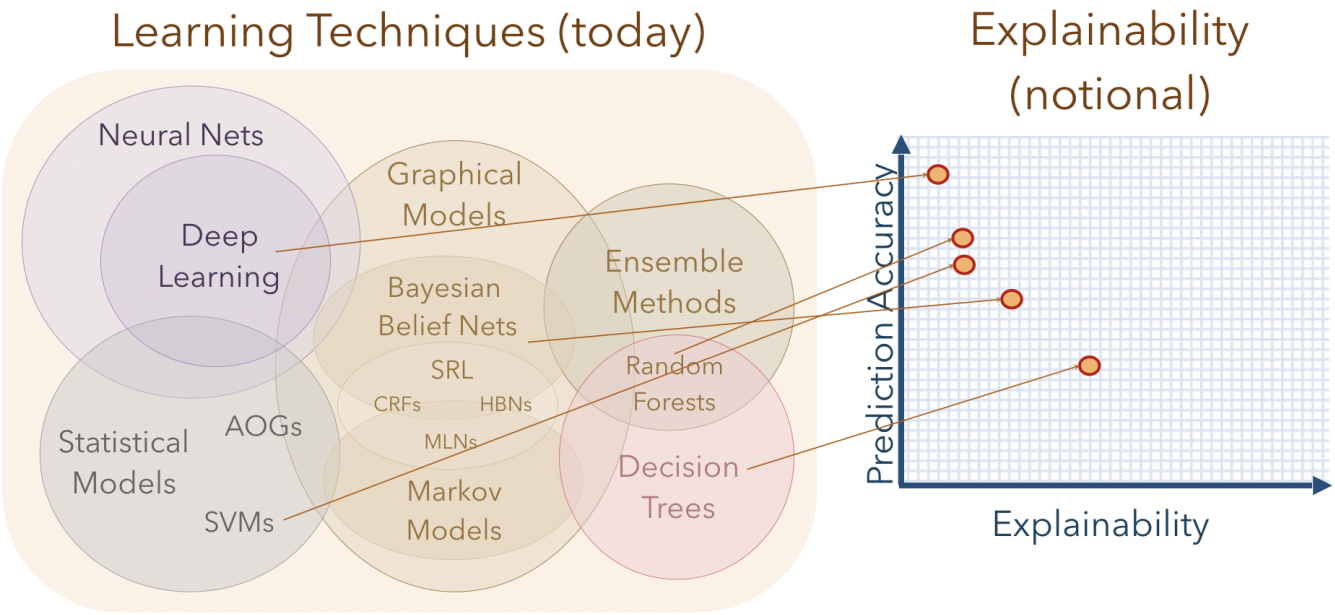

Fig. 1. A visual comparison of the explainability of machine learning methods and their relative performance [13].

.

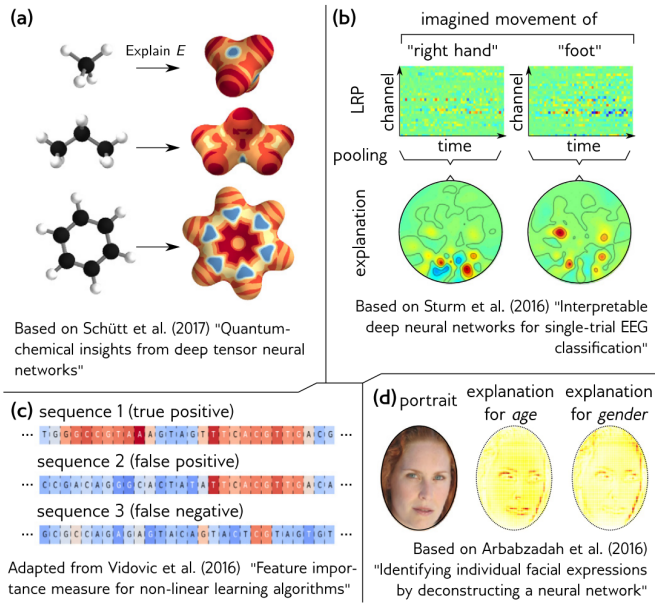

Fig. 2. Overview of several applications of machine learning explanation techniques in the sciences. (a) Molecular response maps for quantum chemistry, (b) EEG heatmaps for neuroimaging, (c) extracting relevant information from gene sequences, (d) analysis of facial appearance. [12].

that has high relevance for human experts, such as a specific color pattern on a bird or an architectural feature of a building. The class prototypes are used to construct a "prototype layer" between the convolutional layers and max pooling layer of a CNN. 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: CNNs excel at tasks of image classification, and long short-term memory (LSTM) networks excel with processing natural language, but their features can be combined *a priori* such that an LSTM can be used to generate an explanation of why a CNN made its decision [14]. Such a hybrid model is trained jointly with labeled images and textual, natural language descriptions of those images. Features of an input image are extracted from the convolutional layers of a CNN and supplied as input to an LSTM that has been trained to generate natural language responses based on the training descriptions (see figure 4)

2) Counterfactual explanations: Counterfactual explanations are an *a posteriori* method for generating bounds and rules for a classifier's decision. For example, "you were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan" [17]. Counterfactual bounds are extracted by algorithms that generate slight perturbations in an input's features until an observable change is made in the classifier's decision [17]. While the perturbation of input features may be compute intensive, the generation of counterfactual explanations requires no *a priori* knowledge of the model's internals and circumvents the need to apply model- or topology-specific methods. Counterfactual explanations work best when the input features are already human-interpretable as independent features, like income in the example above. Counterfactual explanations generated from the perturbation of individual pixels in an image may not result in human-interpretable changes in the image [17].

Fig. 3. Ideal inputs of various output classes for three different CNNs [16].

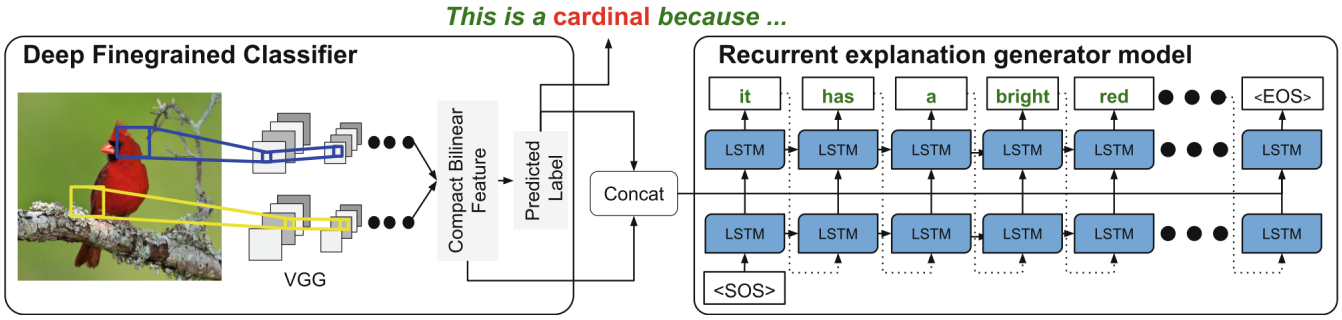

Fig. 4. Hybrid network architecture for generating classification decisions and accompanying textual descriptions of key features [14].

C. Data Provenance

While other methods of XAI mostly focus on interpreting or explaining a model's decision based on evaluating new input, data provenance can explain how a model was created or how the training data relates to a decision. Applying the principles of data provenance to the development of ML models establishes transparent data lineage throughout the life cycle of the data. Being able to trace the history of the data provides benefit particularly for data scientists and auditors of ML systems. Data scientists can develop ML models more collaboratively when they are able to log and share the various steps that have been applied in their workflows and their results, and auditors may analyze the history of the data to help identify potential issues with bias or discrimination. If an opportunity to reduce bias or discrimination is discovered, then having a healthy lineage of the data should help data scientists in reducing effort in training a new model by exercising lessons learned and observations from previous development.

The taxonomy of concepts of data provenance may be high-level or even vague at times, but the core concepts lend themselves well to practical usage, such as the modeling, capturing, and persisting metadata [18]. There exist many tools for generic data provenance activities, like managing the overhead and the scaling of metadata capture [19] or capturing data lineage by wrapping ETL pipelines [20]; however, there are also software tools that are establishing themselves as ML-centric by integrating directly with existing ML libraries or offering their own libraries for common ML activity. MLflow is a python library that provides metadata logging and collaboration features by integrating with common ML libraries, including Tensorflow, scikit-learn, Spark ML, and PyTorch [21]. H20.ai is a platform that focuses more inter-language operability by providing python, scala, R, and java APIs for their data provenance platform which includes libraries for feature extraction, dimension reduction, classification, and other common activities for training models [22]. Software tools and libraries for machine learning are becoming more accessible to wider audiences of data scientists, and there is industry demand for data provenance solutions that support heterogeneous machine learning platforms [23].

D. Model Induction

Similar to counterfactual explanations, rules can be extracted from a black-box model by submitting a very large number of inputs whose input features vary only slightly from one to the next. As bounds on the input features are discovered, these bounds result in rules which can be converted into an interpretable model, such as a decision tree of arbitrarily-constrained size. Over a large enough data set, a set of rules may be an adequate approximation of the black-box model,

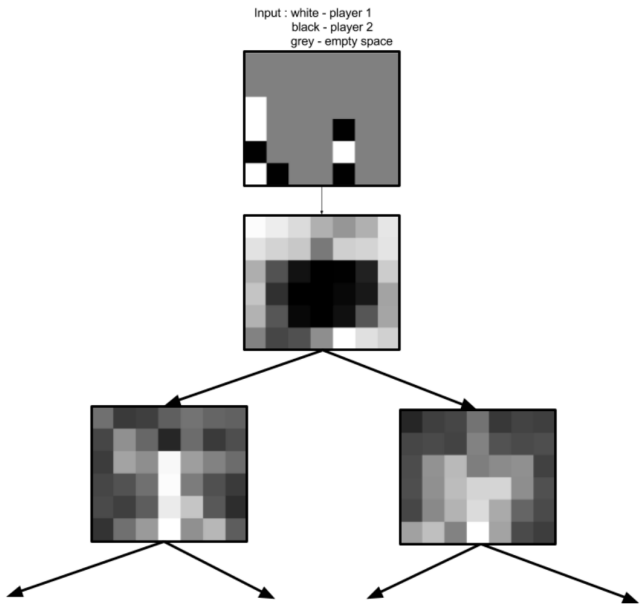

Fig. 5. The first two layers of a decision tree model extracted from a neural network that was trained to place pieces in the game of Connect Four [25].

often performing better than if a decision tree was trained directly from the data set [24]. Again, similar to counterfactual explanations, model induction works best when the input features are interpretable as independent of each other so that extracted rules may be more intuitive. However, it remains possible to apply model induction to image inputs. Frosst et al. extracted a decision tree from a neural network that was trained to play the game Connect Four [25] and observed that the first two layers of the decision tree may split the game the game into two routes: placement of pieces gravitate towards the center of the board or placement of pieces is more even distributed across all columns of the board [25]. As seen in figure 5, the interpretation of the decision tree is still highly subjective for visual inputs and relies on the observer's intuition.

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 [26]. 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 [12]. 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 [27]. 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 [28]. 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 [29]. 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, perhaps, if societal impressions of AI/ML will evolve to no longer be concerned with the inner workings of machine learning.

III. Challenges in XAI

Feature relevancy with saliency maps are a popular method of interpreting the decisions of DNNs, but methods for generating these saliency maps may be manipulated with adversarial inputs. Given a set of feature saliency maps from a DNN, it is possible to create an iterative optimizer that can take an intended input into the DNN and a target feature saliency map and generates an adversarial input that is indiscernible to the human eye [30]. The iterative optimizer perturbs the original input based on the target feature saliency map and bound by a

perturbation magnitude so that the internals of the DNN may be computed in such a way that a similar response is achieved while the explanation of the decision has been manipulated. This method of adversarial attack on feature relevance analysis on a DNN requires the adversary to have obtained previous feature saliency maps from the DNN and also the ability to inject an adversarial input into the DNN. Also, in order for the attack to have value, it must occur in a context in which the analysis of feature relevancy is occurring. These constraints limit the feasibility and value of this type of attack on real-time systems, but as seen in figure 6, the result of this type of attack can have a significant observable impact.

The interpretation of responses from machine learning models are subjective and rely on the user's intuition. The vocabulary around the field of XAI is not yet mature or homogenous, creating rifts in how researchers describe problems and approach the study of human interpretability in machine learning. Related terms in research include the study of comprehensibility and understandability, and the study of these topics can be contextualized around the concepts of interestingness, usability, acceptability, and justifiability [31]. Quantitative characteristics of models are also studied for their impact on interpretability, such as the size of a model's feature input or topology [31]. There is a lack of literature that relates the measures and concepts of interpretability to specific methods of XAI.

Stakeholders of automated decision systems cover a wide variety of backgrounds with differing domain knowledge, intentions of participation, and access to interfacing with the system. An explanation for expert users may provide little to no value to lay users, such as owners and consumers. Some ML systems, such as autonomous vehicle and facial recognition software, may be deployed in public settings in which the data subjects had no knowing participation in the ML system's decision making process. The General Data Protection Regulation (GDPR) provides a right for consumers to access the data of theirs that companies collect [32], but designing the systems and interfaces for such a wide audience of stakeholders is a complicated challenge.

Even if methods of XAI may be applied effectively, in practice, the auditing and accounting of automated decision systems in private industry is rife with tedium, complicated legacy systems, and noncooperation [33]. It is difficult for researchers to study the application of XAI in practice unless there is internal motivation for the owners of the ML process to do so because a private organization is at risk at being exposed for implementing, likely unintentionally, biased or unfair decision systems. While it is popular anecdote that machine learning systems are replacing legacy, human-driven analytics in droves, these legacy analytic processes are steeped in cultural domain knowledge in the organization which does not fit succinctly into the XAI methods that have been developed via academic experimentation [33]. The implementation of XAI methods in practice in private organizations is an uphill battle.

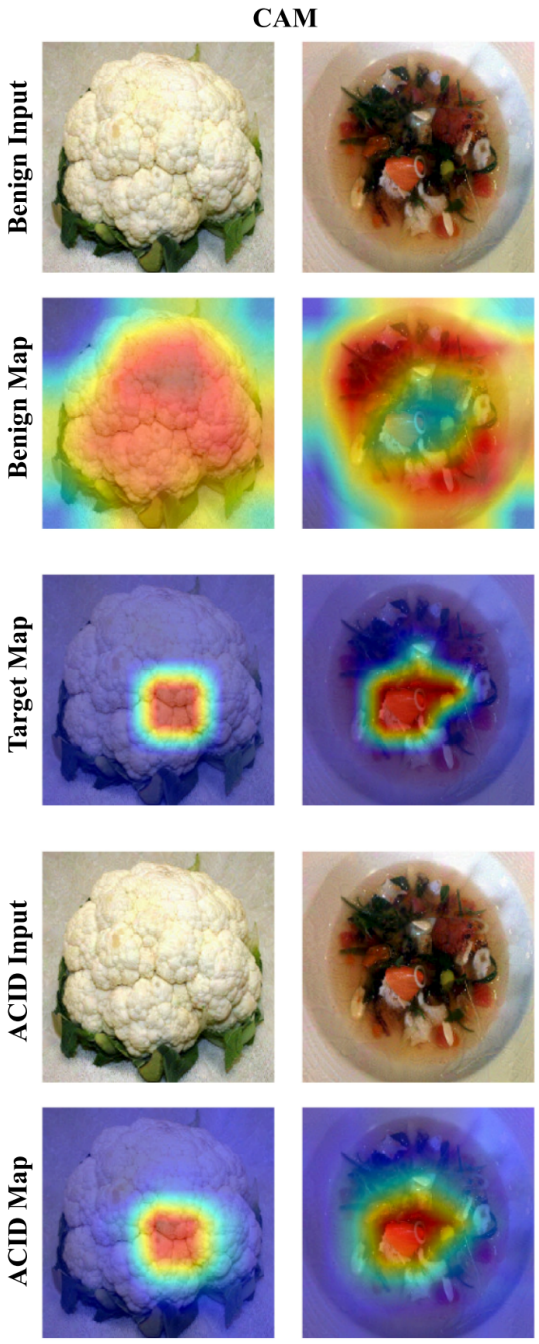

Fig. 6. Zhang et al. create adversarial inputs to deceive a DNN in attributing feature relevance based on a target map [30]

IV. Use Cases

Various methods of XAI have been layed out in the previous section, but the question still remains of how do these methods provide value to scientists, consumers, or society at large. Here, we explore the various users who interface with XAI and what value is provided.

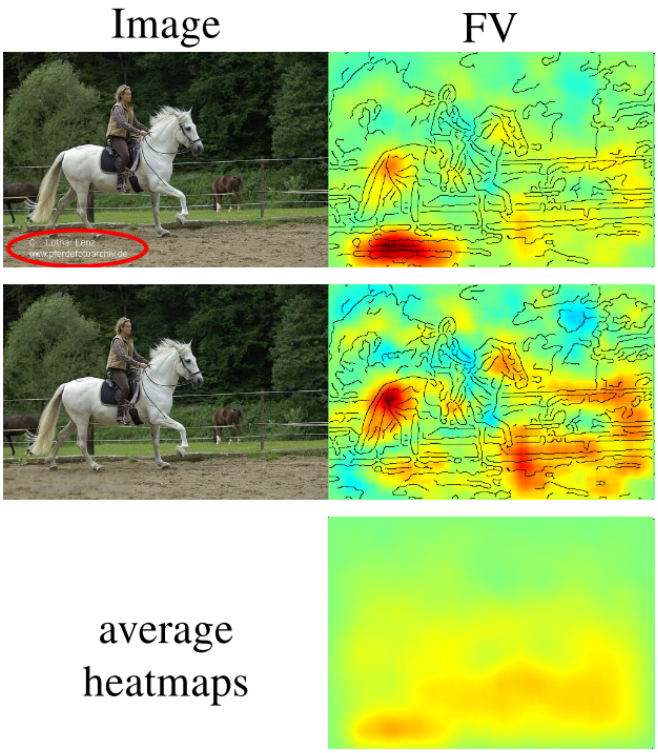

Fig. 7. Heatmaps generated via LRP for a Fisher Vector classifier, before and after the removal of a watermark from the training data. [34].

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

This use case takes an *a posteriori* approach to identifying opportunities to improve a model's performance. In current literature, the XAI methods involved are commonly visualization techniques, but other methods are still applicable. Due to the subjective nature of interpretation methods like visualization of relevant features, any action in improving the training of a model requires a human expert to analyze the explanation for potential weaknesses in the model's decision process.

In one example of improving a model through the identification of relevant features, a group of researchers used a Fisher Vector classifier on the PASCAL VOC 2007 data set [34]. After applying LRP to as a tool in explaining the classifier's decisions, it was apparent that the copyright watermark on images of horses was highly relevant in the Fisher Vector model. The researchers then removed the copyright watermark before retraining the model, and future decisions considered the horse and its surroundings as more relevant features (see figure 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

The psychological study of the trustworthiness of AI systems is as old as AI itself, but there are recent contributions to how XAI contributes to people's perceptions of ML models

(a) Husky classified as wolf (b) Explanation

Fig. 8. Heatmaps generated via LRP for a Fisher Vector classifier, before and after the removal of a watermark from the training data. [35].

and decision systems. Ribeiro et al. trained a classifier of images of wolves and huskies with intentionally biased training data and measured human subjects' trust in the model before and after receiving an explanation of a misclassified input [35]. The classifier was trained with images that were hand selected such that images of wolves had snow in the background and images of huskies did not. Human subjects observing the classifications and receiving the explanation were all students in an ML course, so each participant had a background in black-box models. Before receiving an explanation of a misclassification from the model as seen in figure 8, 10 out of 27 students trusted the model, but after the students received an explanation of the model's decision, only 3 of the students reported trusting the model. While the sample size of human subjects is small and limited to a very specific demographic, the result is intuitive that an explanation of a misclassification from a biased model directly impacts the trust in that model. In a similarly small experiment, Koo et al. establish that users of a vehicle simulator had more positive emotional responses when a driving-assistance system provided explanations for when it applied hard break events in emergency scenarios [36].

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 scrutiny, legal or otherwise

While legislation such as the EU's General Data Protection Regulation's "right to explanation" faces legitimate criticism about its scope and ability to be enforced [32], there also exists an ethical imperative for providers of AI/ML products to provide explanations for their systems' decisions. For example, in the criminal justice system, judges and parole boards may apply predictive models as tools in making their decisions, but racial discrimination has been uncovered in such tools [37] [38]. While there is a lack of enforceable legislation around regulating the accountability of automated decision systems, autonomous vehicles and vehicles with ADAS from Tesla, Google, and GM Cruise have been involved with numerous traffic incidents, ranging from trivial to fatal [39] [40] [41] [42]. It may be commercially advantageous for automakers to continue to push the capabilities of technology by designing autonomous systems that can provide effective explanations to

various stakeholders before legislation and regulations require it of autonomous vehicles.

V. XAI in Autonomous Vehicles

A. Existing Research

In one study, research subjects participated in a driving simulator in which the vehicle had a semi-autonomous feature that would initiate hard braking events in the case of an unforeseen obstacle in the path of the vehicle [36]. Researchers programmed the simulator to provide different types of verbal explanations for the hard brake events. Survey responses from the research subjects showed that drivers had not only more positive emotional responses when more descriptive explanations were provided by the semi-autonomous system, but there was also measurably safer driving habits observed by the drivers when they received rich explanations that described not only how the vehicle was behaving but also why it was choosing those behaviors.

It is not uncommon for researchers to generate saliency maps of feature importance when doing computer vision tasks with deep neural networks. These feature saliency maps can overlay neatly over the original image to help developers of these decision systems to identify if the responses from the neural network are behaving as expected. Bojarski et al. developed a method of generating saliency maps called VisualBackProp [43] and apply it to an end-to-end autonomous driving system called PilotNet. PilotNet can decide steering wheel angles given input images, bypassing the need for an intermediate perception phase that identifies segmented objects in images before making decisions. VisualBackProp confirms that PilotNet is making decisions based on human-understandable features in the images, such as the boundaries of the road, lane markers, and the boundaries of other vehicles on the road (see figure 9) [44].

Using visual explanations to explain the decisions of a trained model has multiple drawbacks which can be ameliorated with the emerging method of verbalization in XAI. While saliency maps may give insight into how a model is responding to a single input, both humans and autonomous systems make decisions based on information over time. Also, the interpretation of a visual explanation is subjective to the human audience; different observers may draw different or no insights from a single visual explanation. Kim et al. train a LSTM neural net on a dataset of images over time accompanied by human-labeled explanations of the current driving behavior [45]. The output layer of this trained model labels images over time with descriptive natural language of the driving behavior. Provided as input into this verbalization model are ”attention” maps of regions in the images that are identified as important features. These regions of high and weak attention are generated from a CNN that is used to decide acceleration and course change behavior given input images. The objective of Kim et al. of explaining a limited set of automated behavior in an autonomous vehicle via visual explanations is similar to the work of Bojarski et al [44], but the concept is taken further by generating natural language responses to the feature relevances over time.

Fig. 9. Bojarski et al. generate saliency maps of input features in their end-to-end autonomous driving system [44]

- There are a couple papers about the multimodal sensors in AV, the perception/decision architecture, and a survey of decision
- There are quite a few papers about establishing trust with automated systems, AI, and HCI. Is there a gap in establishing trust that has not been addressed in previous research?
- Might have to put the legal/auditory use case on the back burner...way out of scope for my expertise at this time.

B. Future Research in Use Case IV-A

One of the most explicit use cases for applying XAI methods to the domain of autonomous vehicles is in the support of researchers and engineers to develop more accurate and robust machine learning models, especially in the perception and decision systems of the autonomous vehicle. In the use case defined in IV-A, we identified examples of data scientists using explanations of non-intuitive, black-box ML models to improve how the models were trained and thus improve the performance of the models. While some researchers have applied XAI to CNNs that were used in an end-to-end system to decide steering angles [44], there was no published research found at the time of this writing on applying XAI towards the improvement of CNNs being applied in the perception system of autonomous vehicles for tasks such as object detection and object segmentation. In the decision systems of autonomous vehicles, deep reinforcement learning is an emerging field for the training of autonomous vehicles [46], and the resulting deep q-networks (DQNs) have not been explored for explaining their decisions onto a human-interpretable domain. Besides using explanations to discover opportunities to improve models, there is no published research on the application of data provenance as a tool to assist with the ML workflow of

developers of autonomous vehicles through collaboration and reproduction of experiments.

C. Future Research in Use Case IV-B

Trust in autonomous vehicles has been identified as having a positive, causal relationship with people's intention to use autonomous vehicle technology [47]. The influence of trust in the intention of using AV technology was broken down into various constructs (see figure /reffig:choi2015), and XAI may be applicable to two of those constructs in particular: system transparency and situation management. System transparency may be the least influential construct of trust, but it would likely be the most directly applicable area for XAI. XAI methods such as rule extraction for the decision system and verbalization techniques for the perception system may be able to map the predictions and decisions of the AV system onto the human interpretable domain of natural language. Two aspects of situation management include the system's ability to generate alternative decisions and the user's ability to control the autonomous vehicle. Rule-based and fuzzy decision systems can inherently provide a ranked ordering of other alternative decisions that were considered, and simple labeling of which decisions were made by the autonomous systems and which decisions were provided by the user can provide users with explanations on the system's situation management. It may be valuable to observe the impact of the application of XAI methods on these constructs of trust in the user experience of users of AV systems.

D. Future Research in Use Case IV-C

The legal impetus of explaining the decisions made by autonomous vehicles is often not present in regional legislation or is at risk of being unenforceable due to lack of clarity or definition. The US Department of Transportation's National Highway Traffic Safety Administration has released guidelines on the development and adoption of autonomous vehicles which defer legislation to state or municipal bodies [48], and currently, 14 states and no US territories have put legislation into law regarding autonomous vehicles [49]. California has a dedicated Department of Autonomous Vehicles which has regulation around the reporting of collisions and disengagements of autonomous vehicle systems in the state, but the requirements for describing the cause of disengagements lacks any requirements on explaining the autonomous decision system [50]. In the EU, the General Data Protection Regulation (GDPR) contains recitals that grant consumers the right to request what data of theirs companies are collecting, how companies are using it, and how decisions with their data are made. While some aspects of the GDPR are conceptually more straight-forward, such as consumers requesting the data of theirs that a company has collected, there is a lack of definition and clarity around the idea of explaining decisions that companies make with consumers' data, and it is unclear if the GDPR's so-called "right to explanation" is enforceable [32]. The future task of effectively legislating and enforcing fair, accountable, and transparent machine learning algorithms is uncertain and daunting.

First steps in approaching the modernization of the relationship between legislation and AI/ML in autonomous vehicles may include designing autonomous vehicle systems to be able to store and extract data and metadata for all relevant stakeholders, the identification of regulatory and legislative gaps on autonomous vehicles, and establishing effective language to make laws and regulations enforceable for accountable and transparent decision systems in autonomous vehicles. Stakeholders of decisions from autonomous systems can be expert users or lay users, and lay users can be anyone from the consumer of the autonomous system to individuals in the public who are acting as input to the autonomous vehicle's perception systems [51]. The challenge of how to design systems to make decision and explanation data for such a wide variety of stakeholders may be an emerging area of research in the explainability of autonomous vehicle systems. While some geographic regions explicitly lack any legislation or regulation on autonomous vehicles, even the regions that do have laws in place may find them difficult or ineffective to enforce [32]. Wachter et al. establish eight recommendations for strengthening the "right to explanation" in the GDPR, including clarification on if the consumers' "right to access" from Article 15 of the GDPR includes the explanation or just the existence of a decision, what metadata on the decision process constitutes an explanation (feature importance, decision tree model, etc.), when a decision is based solely on an automated process. The future of the legal aspect of explainable autonomous systems will be a combination of designing systems for data extraction and the refinement of laws and regulations.

VI. Conclusion

XAI is a relatively young yet increasingly popular field of research in which researchers of heterogenous backgrounds and goals struggle to find cohesion and unification. An evolving vocabulary makes it difficult for similar research to be identified and related to one another. The subjective and abstract concepts of interpretation and explanation make it difficult to quantify characteristics of models and their explanations. And the diverse array of stakeholders from a variety of backgrounds makes it difficult to design systems that support the interfaces, data collection, and explanation methods necessary to create a cohesive concept of an explainable and accountable system.

"Black box" machine learning models are pervasive in the development of autonomous vehicles. They are used in perception tasks, such as segmentation of entities from sensor data. Machine learning is also applied in the decision system of the autonomous vehicle by responding to continuous traffic scenarios to navigate to the final route destination. Where ever there is machine learning, there is an opportunity to apply methods of XAI, such as feature relevancy, explanation verbalization, extraction of explicit rules, and data provenance. While there exists literature on applying XAI methods in the context of autonomous vehicles, there is no research yet on how it may provide value to the variety of stakeholders.

The three use cases defined in section IV each can be applied in the context of autonomous vehicles, and each use

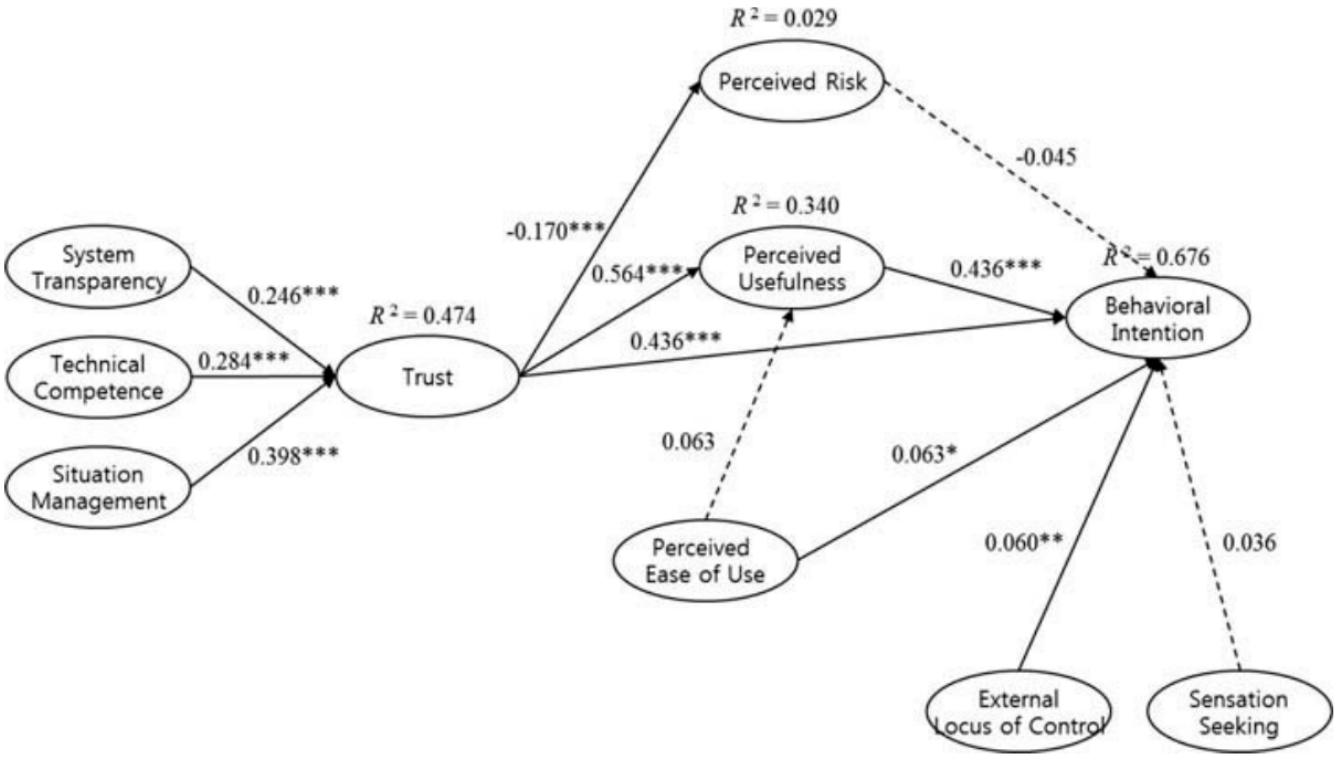

Fig. 10. Choi et al. break down the constructs of trust and influences on people's behavioral intention of using AV technology [47] *Note:* *p ¡ .05. **p ¡ .01. ***p ¡ .001.

case pertains to one of three groups of users: developers and engineers, consumers, and auditory and investigatory entities. The development process of the ML systems in the vehicle can be strengthened by data provenance techniques that improve collaboration or also by using explanations to identify opportunities to improve predictive models. Potential consumers of autonomous vehicles may feel safer when the vehicle can provide verbal explanations of the decisions that it is making. And automakers may be inclined or, potentially, required to be able to provide explanations to legal or auditory queries for the decisions made by their autonomous systems. Existing literature demonstrates how XAI methods can be applied to provide insights to developers and establish trust with end users, but there is a lack of literature on the ethical and legal obligation to provide explanations of the decisions made by autonomous vehicles.

REFERENCES

[1] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine learning in healthcare: A review," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 910–914, 2018.

[2] J. Zhang, F. Wang, K. Wang, W. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 1624–1639, Dec 2011.

[3] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *Journal of Industrial Information Integration*, vol. 6, pp. 1 – 10, 2017.

[4] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, pp. 226 – 251, 2019.

[5] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, 2018.

[6] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 1153–1176, 2016.

[7] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. P. Martínez, and V. Robles, "Machine learning in bioinformatics," *Briefings in bioinformatics*, vol. 7 1, pp. 86–112, 2006.

[8] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70 – 90, 2018.

[9] P. Lago, C. Roncancio, and C. Jiménez-Guarín, "Learning and managing context enriched behavior patterns in smart homes," *Future Generation Computer Systems*, vol. 91, pp. 191 – 205, 2019.

[10] Wikipedia contributors, "Regression analysis." [Online; accessed 2019-04-16].

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

[12] 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.

[13] D. Gunning, "Explainable Artificial Intelligence (XAI)." https://www.darpa.mil/program/explainable-artificial-intelligence, 2016.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *SIGMOD Rec.*, vol. 34, pp. 31–36, Sept. 2005.

[19] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance techniques," 2005.

[20] 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.

[21] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the Machine Learning Lifecycle with MLflow," *IEEE Technical Committee on Data Engineering*, pp. 39–45, 2018.

[22] "Overview – H20.ai 3.24.0.1 documentation." http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html.

[23] 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.

[24] G. Vandewiele, O. Janssens, F. Ongenae, F. D. Turck, and S. V. Hoecke, "Genesim: genetic extraction of a single, interpretable model," *CoRR*, vol. abs/1611.05722, 2016.

[25] N. Frosst and G. E. Hinton, "Distilling a neural network into a soft decision tree," *CoRR*, vol. abs/1711.09784, 2017.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] B. Mittelstadt, L. Floridi, and S. Wachter, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, pp. 76–99, 06 2017.

[33] 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.

[34] S. Bach, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2912–2920, 2016.

[35] 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.

[36] 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.

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

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

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

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

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

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

[43] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, U. Muller, and K. Zieba, "VisualBackProp: visualizing CNNs for autonomous driving," *CoRR*, vol. abs/1611.05418, 2016.

[44] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *CoRR*, vol. abs/1704.07911, 2017.

[45] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[46] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *CoRR*, vol. abs/1704.02532, 2017.

[47] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human Computer Interaction*, vol. 31, pp. 692–702, 2015.

[48] U.S. Department of Transportation, "Preparing for the Future of Transportation: Automated Vehicle 3.0," tech. rep., October 2018.

[49] "Autonomous vehicles — self-driving vehicles enacted legislation." http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx, March 2019. [Online; accessed 2019-04-16].

[50] "Testing of autonomous vehicles with a driver." https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing. [Online; accessed 2019-04-16].

[51] G. Ras, M. van Gerven, and W. F. G. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," *CoRR*, vol. abs/1803.07517, 2018.