Explainable AI: A Review

Dipta Chatterjee

13 April, 2022

Abstract

This paper provides a brief analytical review of the current state-of-the-art in relation to the explainability of artificial intelligence in the context of recent advances in machine learning and deep learning. The paper starts with a brief historical introduction and a taxonomy, and formulates the main challenges in terms of explainability building on the recently formulated National Institute of Standards four principles of explainability. Recently published methods related to the topic are then critically reviewed and analyzed. Finally, future directions for research are suggested.

Keywords : black-box models, deep learning, explainable AI, machine learning, prototype-based models, surrogate models

1 Introduction

Artificial intelligence (AI) and machine learning (ML) have demonstrated their potential to revolutionize industries, public services, and society, achieving or even surpassing human levels of performance in terms of accuracy for a range of problems, such as image and speech recognition (Mnih et al., 2015) and language translation (Young et al., 2018). However, their most successful offering in terms of accuracy—deep learning (DL) (LeCun et al., 2015)—is often characterized as being "black box" and opaque (Pasquale, 2015; Rudin, 2019). Indeed, such models have a huge number (many millions or even a billion) of weights (parameters) which are supposed to contain the information learned from training data. Not only is the number of these weights very large, but their link to the physical environment of the problem is extremely hard to isolate. This makes explaining such forms of AI to users highly problematic. Using opaque, "black box" models is especially problematic in highly sensitive areas such as healthcare and other applications related to human life, rights, finances, and privacy. Since, the applications of advanced AI and ML, including DL, are now growing rapidly, encompassing the digital health, legal, transport, finance, and defense sectors, the issues of transparency and explainability are being recognized increasingly as critically important. For example, a search in Google Trends (https://trends.google.com/trends/) reveals that in the last decade publications using the terms "DL" and "explainable AI" (XAI) both grew significantly, but while the curve for DL is now in a saturation stage over the last 3 years or so, the curve for XAI is growing exponentially starting precisely 3 years ago when the saturation in regards to

2 Brief Historical Perspective

AI was closely linked to both ML and to logic and symbolic forms of reasoning from its inception in the middle of the 20th century (Samuel, 1959; Smolensky, 1987). ML and data-driven statistical techniques gained momentum in recent years due to an unprecedented increase in the number and complexity of data available (now the majority of data are unstructured, with many more images/videos as well as text/speech in comparison to the 20th century) (Bishop, 2006; Goodfellow et al., 2014). Historically, the first methods of AI, such as decision trees (Quinlan, 1990), symbolic AI (Smolensky, 1987), expert systems, fuzzy logic, and automated reasoning (Robinson & Voronkov, 2001), as well as some forms of artificial neural networks (ANNs), for example, radial-basis function (RBF) architectures and linguistic, prototype-based, representations were significantly more interpretable and self-explainable than the more recent and more efficient forms such as support vector machines (SVMs) (Hearst et al., 1998) and most other forms of ANNs. In the last few years, explainability has become an important issue not only for scientists, but also for the wider public including, regulators, and politicians. As AI and ML (and, especially, DL) become more wide spread and intertwined with human-centric applications, and algorithmic decisions become more consequential to individuals and society, attention has shifted back from accuracy to explainability (Angelov & Soares, 2020; Core et al., 2006; Pedreschi et al., 2019). Complex and "black box" (Pasquale, 2015; Rudin, 2019) types of models can easily fool users (Nguyen et al., 2015) and, in turn, this can lead to dangerous or even fatal consequences (Stilgoe, 2020). Opening the "black box" is critically important not only for acceptability within society, but also for regulatory purposes. (In 2019 the US Congress passed

the Algorithmic Accountability Act (MacCarthy, 2019) and the EU enshrined the right for an explanation to the consumer (Core et al., 2006; Goodman & Flaxman, 2017; Pedreschi et al., 2019).) The current data-rich environment brought the temptation to take shortcuts from raw data to solutions using a very large number of abstract, purely numerical parameters (Angelov & Soares, 2020; Rudin, 2019; Stock & Cisse, 2018), without providing a deep insight into, and understanding of, the underlying dependencies, causalities, and internal model structures. The issue of explainability is an open research question for some of the most successful (in terms of accuracy) forms of ML such as SVMs, DL, and many of the ANNs (Bishop, 2006), as Figure 2 illustrates. In the above context, the main question is not so much: Can we get an XAI solution?, but Can we get a highly accurate XAI solution comparable to the accuracy that DL would provide? Table 1 illustrates some results for the Caltech-101

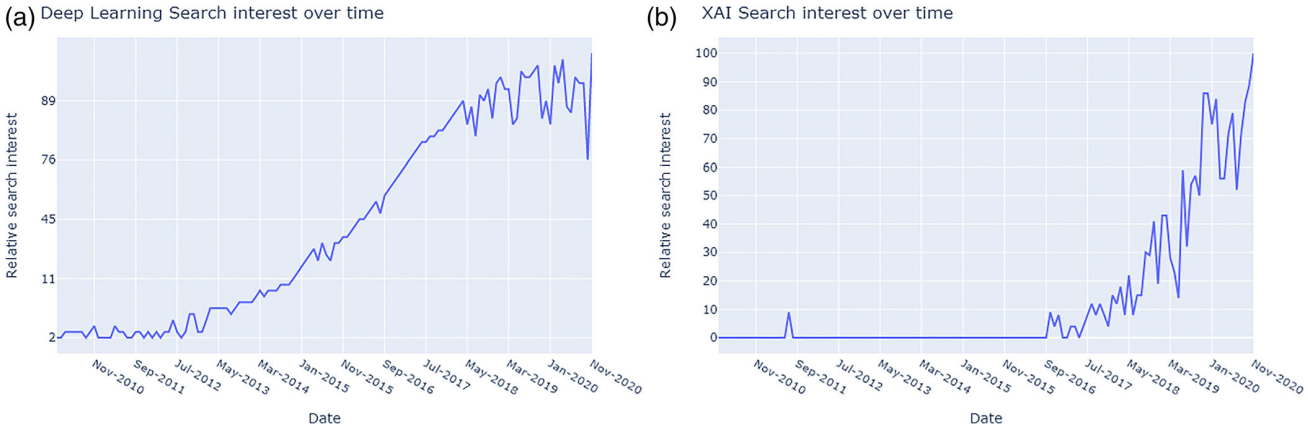

Figure 1: Illustrates the interest evolution towards two terms according to Google Trends: (a) deep learning (DL), (b) explainable artificial intelligence (XAI)

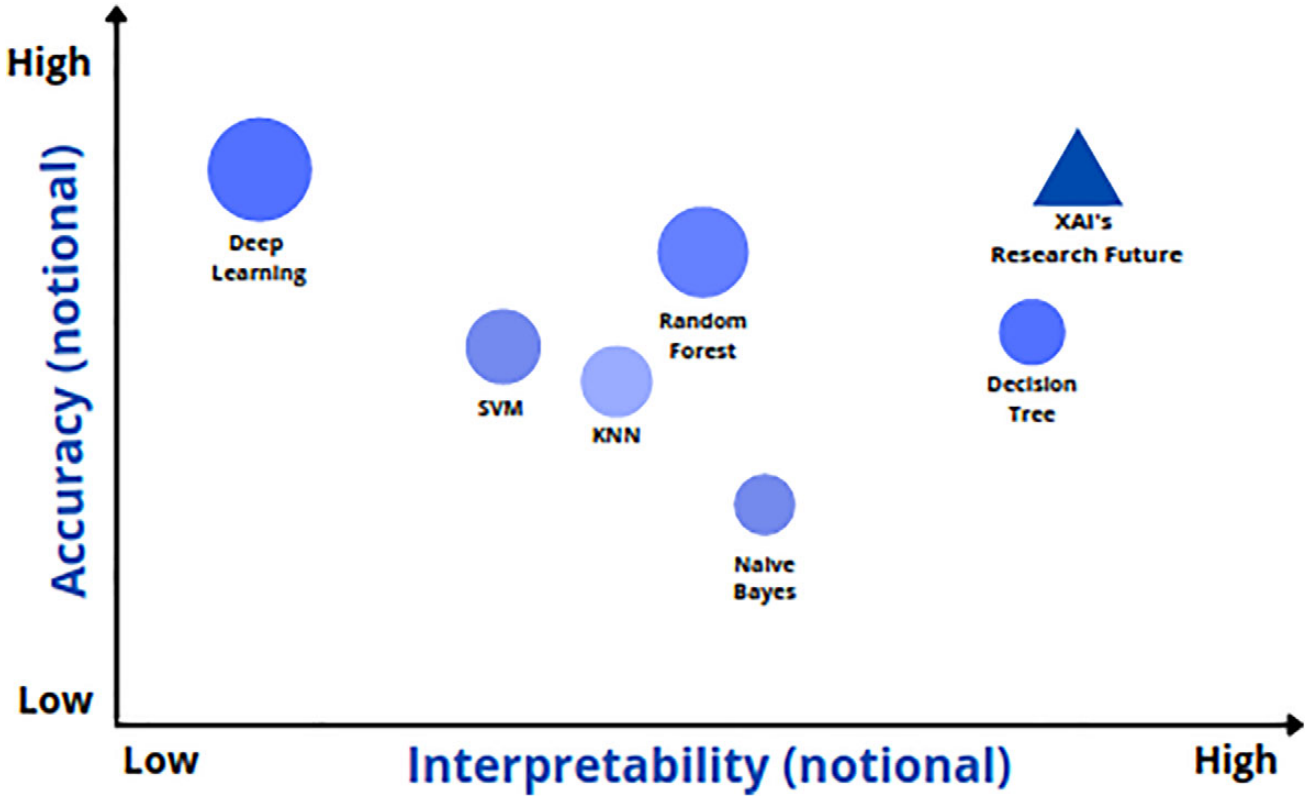

Figure 2: Accuracy vs. interpretability for different machine learning models

Table 1: Performance comparison for the Caltech-101 dataset

Method & **Accuracy**	**Time (s)**	**Parameters**	**Interpretability**	
xDNN (Angelov High	Soares, 2020)	94.31%	362	4 per class
VGG-16 (Simonyan Very low	Zisserman, 2014)	90.32%	18,332	138.000.000
ResNet-50 (He et al., 2016)	90.39%	12,540	23.000.000	Very low
Random forest (Breiman, 2001)	87.12%	412	20,000	Medium
SVM (Hearst et al., 1998)	86.64%	783	15,000	Low
kNN (Peterson, 2009)	85.65%	221	300 and all data	Low
Decision tree (Quinlan, 1996)	86.42%	236	5 rules per class	High
Naive Bayes (Rish, 2001)	54.84%	323 & 409,700	Medium	

3 XAI Taxonomy

In the literature, a variety of terms exist to indicate the opposite of the "black box" nature of some of the AI and ML, and especially DL, models. We distinguish the following terms:

- Transparency: a model is considered to be transparent if, by itself, it has the potential to be understandable. In other words, transparency is the opposite of "black-box" (Adadi & Berrada, 2018).
- Interpretability: is defined as the capacity to provide interpretations in terms that are understandable to a human (Gilpin et al., 2018).
- Explainability: is related with the notion of explanation as an interface between humans and an AI system. It comprises AI systems that are accurate and comprehensible to humans (Gilpin et al., 2018).

  Although these terms are similar in their semantic meanings, they confer different levels of AI to be accepted by humans. For more details, the ontology and taxonomy of XAI at a high level can be detailed as below:

- Transparent model: Typical transparent models (Adadi Berrada, 2018) include k-nearest neighbors (kNN), decision trees, rule-based learning, Bayesian network, and so on. The decisions from these models are often transparent, although transparency, as a property, is not sufficient to guarantee that a model will be readily explainable
- Opaque model: Typical opaque models (Pasquale, 2015; Rudin, 2019) include random forest, neural networks, SVMs, and so on. Although these models often achieve high accuracy, they are not transparent.
- Model agnostic: Model-agnostic XAI approaches (Dieber & Kirrane, 2020) are designed with the purpose of being generally applicable. As a result, they have to be flexible enough, so that they do not depend on the intrinsic architecture of the model, thus, operating solely on the basis of relating the input of a model to its outputs.
- Model-specific: Model-specific XAI approaches often take advantage of knowing a specific model and aim to bring transparency to a particular type of one or several models (Bach et al., 2015).
- Explanation by simplification: By simplifying a model via approximation (Tritscher et al., 2020), we can find alternativesto the original models to explain the prediction we are interested in. For example, we can build a linear modelor a decision tree around the predictions of a model, using the resulting model as a surrogate to explain the more complex one. Explanation by feature relevance: This idea is similar to simplification. Roughly, this type of XAI approaches attempts to evaluate a feature based on its average expected marginal contribution to the model's decision, after all possible combinations have been considered (Chen et al., 2019; Pedreschi et al., 2019).
- Visual explanation: This type of XAI approach is based on visualization (Chattopadhay et al., 2018). As such, the family of data visualization approaches can be exploited to interpret the prediction or decision over the input data.
- Local explanation: Local explanations (Selvaraju et al., 2017) approximate the model in a narrow area, around a specific instance of interest, and offer information about how the model operates when encountering inputs that are similar to the one we are interested in explaining.

The ML literature predominantly uses the term "interpretability" as opposed to "explainability," but according to Burkart and Huber (2020), interpretability itself is insufficient as it does not cover all possible problems associated with understanding "black-box" models. To gain the trust of users, and acquire meaningful insights about the causes, reasons, and decisions of "black-box" approaches, explainability is required rather than simple interpretability. Although, explainable models are interpretable by default, the opposite is not always true. The existing literature (Adadi Berrada, 2018) divides XAI taxonomy by:

- Scope (local (Bach et al., 2015; Selvaraju et al., 2017) and global (Chen et al., 2019; Pedreschi et al., 2019)).
- Usage (post hoc, e.g., surrogate models (Dieber & Kirrane, 2020; Pedreschi et al., 2019; Tritscher et al., 2020) and intrinsic to the model architecture, e.g., explainable-by-design (Soares, Angelov, Biaso, et al., 2020; Soares, Angelov, Costa, et al., 2020)).

  Methodology (focused on the features (Chen et al., 2019; Selvaraju et al., 2017) or on the model parameters (Dieber Kirrane, 2020)). In recognition of the growing importance of this topic, NIST published in August 2020 Four principles of XAI (Phillips et al., 2020), which define the following fundamental principles which an AI must honor to be considered an XAI as follows:

- Explanation: this principle states that an AI system must supply evidence, support; or reasoning for each decision made by the system.
- Meaningful: this principle states that the explanation provided by the AI systemmust be understandable by, and meaningful to, its users. As different groups of users may have different necessities and experiences, the explanation provided by the AI system must be fine-tuned to meet the various characteristics and needs of each group.
- Accuracy: this principle states that the explanation provided by the AI system must reflect accurately the system's processes.
- Knowledge limits: this principle states that AI systems must identify cases that they were not designed to operate in and, therefore, their answers may not be reliable.

Figure 3 depicts the ontology of the XAI taxonomy. Transparent models can easily achieve explainability, while opaque models require post hoc approaches to make them explainable. The categories of post hoc approaches are illustrated accordingly.

4 Review of the state of the art

Current research on XAI is still mostly limited to sensitivity analysis (Arrieta et al., 2020), layer-wise feature relevance propagation and attribution (Tritscher et al., 2020), local pseudo explanations by LIME (Dieber & Kirrane, 2020), gametheoretic Shapley additive explanations (Chen et al., 2019), gradient-based localization, and Grad-CAM (Selvaraju et al., 2017) or surrogate models. In this section, some of the more widely used methods are outlined.

4.1 Features-oriented methods

SHapley Additive exPlanation (SHAP) (Lundberg Lee, 2017) is a game-theoretic approach to explain ML predictions. SHAP seeks to deduce the amount each feature contributed to a decision by representing the features as players in acoalition game. The payoff of the game is an additive measure of importance, the so called Shapley value, which represents the weighted average contribution of a particular feature within every possible combination of features. As such, local and global interpretations of a model are consistent and the average prediction is fairly distributed across all Shapley values, meaning that contrasting comparisons between explanations are possible. However, if the model is not additive then interpretation of the Shapley values is not always transparent, as predictive models may have non independent pay-off splits. Furthermore, while SHAP can be considered model agnostic, optimized implementations of the SHAP algorithm to all model types is not immediately straight forward or efficient. Class activation maps (CAMs) are specific to CNNs. CAMs represent the per-class weighted linear sum of visual patterns present at various spatial locations in an image (Zhou et al., 2016). More formally, global average pooling is applied to the final convolutional feature map in a network, before the output layer. These pooled feature maps are then used as the input features to a fully connected layer and output through a loss function. By projecting the weights of the output back to the previous convolutional layer, the areas in the input image with greater influence over the CNNs' decision are highlighted per-class and visible through a heatmap representation. CAMs cannot be applied to pre-trained networks and networks that

do not adhere to the specified fully convolutional network architecture. Additionally, spatial information can be lost by the fully connected layer and map scaling. Two generalizations of the base CAM model, Grad-CAM (Selvaraju et al., 2017) and Grad-CAM++ (Chattopadhay et al., 2018), try to further increase the explainability of CNNs. Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) generalizes CAM to any arbitrary CNN architecture and without retraining. The gradients for any target class are fed into the final convolutional layer and an importance score computed in respect to the gradients. As with other methods, a heatmap representation of the Grad-CAM indicates which regions of the input image were most important in the CNN's decisions. However, Grad- CAM produces only coarse-grained visualizations and cannot explain multiple instances of the same object in an image. Grad-Cam++ (Chattopadhay et al.) considers the weighted average of the gradients to overcome these drawbacks. Feature oriented methods provide insights into where a decision is taking place in terms of the input, but fall short of a human level explanation of how and why the model came to those decisions. Consequently, a human could not exactly reproduce the explanations rendered by the model.

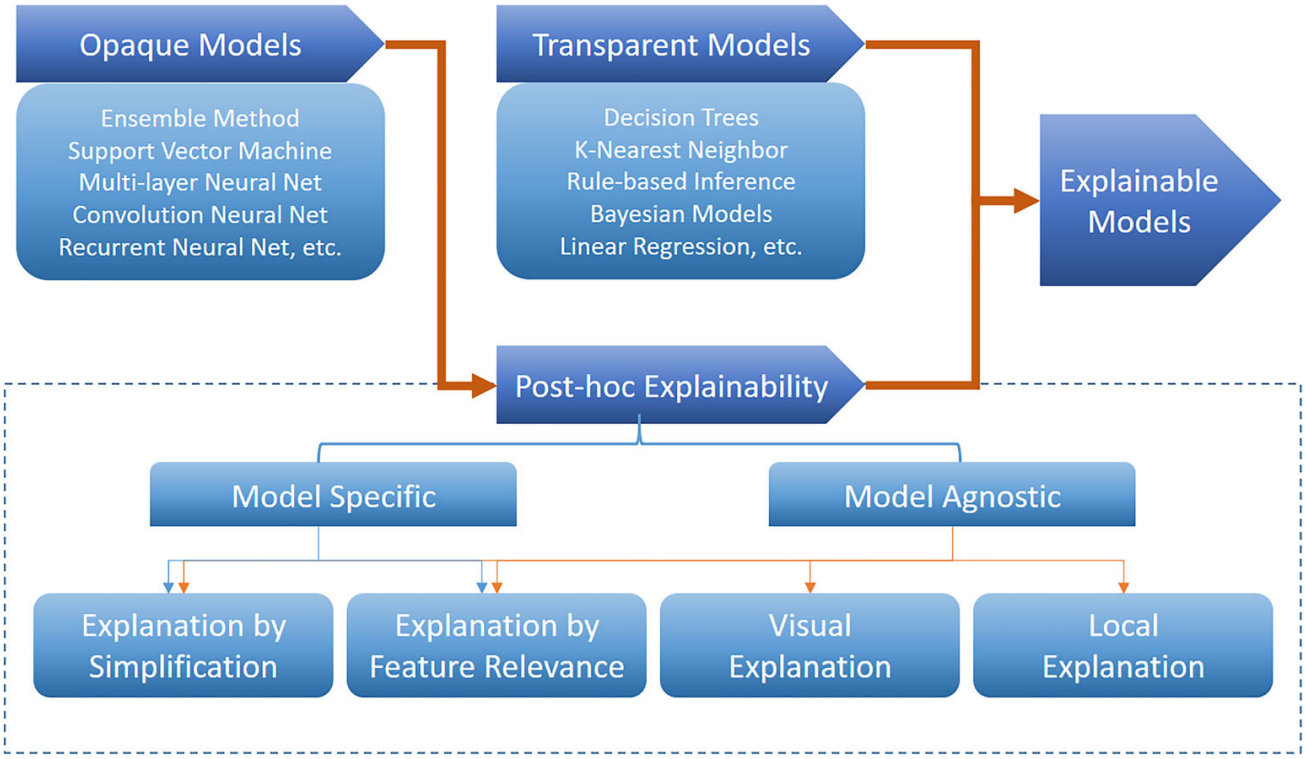

Figure 3: The high-level ontology of explainable artificial intelligence approaches

4.2 Global methods

For features with precise semantic definitions, global attribution mappings (GAMs) (Ibrahim et al., 2019) can explain a neural network's predictions on a global level, across subpopulations, by formulating attributions as weighted conjoined rankings. The advantages are that different subpopulations can be captured through a tuneable granularity parameterGAMs find a pair-wise rank distance matrix between features and a K-medoids clustering algorithm used to group similar local feature importances into clusters. The medoid of each cluster then summarizes the pattern detected in each cluster as a global attribution. This approach is therefore relevant to feature exploration among different subpopulations of samples. Gradient-based saliency maps (Simonyan et al., 2013) are a visualization technique which render the absolute value of the gradient (in respect to the input features) of the majority predicted class as a normalized heatmap. The pixels with a high activation are highlighted and correspond to areas that are most influential (i.e., salient). The method's explanation lies in the ability for a user to look at what features in the image are being used in the classification decision. However, the absolute value means that gradients of neurons with negative input are suppressed when propagating nonlinear layers. As with feature-oriented methods, gradient-based saliency maps do little to communicate decisions beyond model diagnostics. In Ancona et al. (2018), deep attribute maps are presented as a technique for rendering the explainability of gradient-based methods. Importantly, the proposed framework illustrates evaluations between different saliencey-based explanation models. Simply, the gradient of the output is multiplied by the respective input to generate an explanation of a model's prediction in the form of a heatmap. Red and blue colors indicate positive and negative contributions, respectively, to the output decision. Explanations are sensitive to noisy

gradients and variations in the input. Deep attribute maps alone cannot explain why two models produce similar or different results

4.3 Concepts models

Concept activation vectors (CAVs) were introduced by Kim et al. (2021), a technique to explain globally the internal states of a neural network by mapping human understandable features to the high-level latent features extracted by the neural network. As such, CAVs represent the degree to which these abstract features point towards a set of human understandable concepts chosen by a user. Of course, a certain amount of human bias is imposed, but by explaining the associated concept it becomes possible to determine any defects in the decision-making process the model has learned; for instance, if certain characteristics are mistakenly seen as important. Subsequently, automatic concept based explanations (Ghorbani et al., 2019) extract CAVs automatically without human supervision, thereby removing human bias. Instead of being chosen, the human understandable concepts are segmented at various spatial resolutions from in-class images. Nevertheless, concept-based methods are reliant on the concepts being uniquely meaningful to the class, and the effectiveness of explanation is adversely affected if a chosen concept is commonly present in multiple classes.

4.4 Surrogate Models

Local interpretable model-agnostic explanations (LIME) (Dieber & Kirrane, 2020) is a model-agnostic technique to create locally optimized explanations of ML models. LIME trains an interpretable surrogate model to learn the local behavior of a global "black box" model's predictions. For image classification, an input image is divided into patches of contiguous superpixels (i.e., an image object) and a weighted local model is then trained on a new set of permuted instances of the original image (i.e., some superpixels are turned to gray). The intuition is then that by changing aspects of the input data that are human understandable (spatial objects) and learning the differences between those perturbations and the original observations, one can learn what about the input contributed to each class score. However, these explanations are not always informative or reliable at a human level if the parameters that control the perturbations are chosen based solely on heuristics.

4.5 Local, pixel-based methods

Layer-wise relevance propagation (LRP) (Bach et al., 2015) uses predefined propagation rules to provide an explanation of a multilayered neural network's output in respect to the input. The method renders a heatmap, thereby providing insight into which pixels contributed to the model's prediction and the extent to which they did. Accordingly, LRP highlights positive contributions to a network's decision. While LRP can be applied to an already trained network, this process is post hoc and therefore provides only a simplified distillation of the features' role in the decision and is only applicable if the network implements backpropagation. DeconvNet (Noh et al., 2015) uses a semantic segmentation algorithm which learns a deconvolution network and, therefore, provides insights about pixel contribution during the classification process. Similarly, a deep belief network (Hinton et al., 2006) was proposed to improve the interpretability of traditional neural networks.

4.6 Human-centric methods

The above methods, despite their advantages, do not provide clear explanations understandable to humans. They rather "barely scratch the surface" of the "black box" aiming for "damage limitation" with post hoc hints about the features (attribute allocation) or localities within an image. This is radically different from the way people reason and make decisions, make associations, evaluate similarities, and draw an analogy that can be articulated in court or to another expert (e.g., in medicine, finance, law or other area). The aforementioned methods do not answer the fundamental questions of model structure and parameters relating to the nature of the problem and completely ignore reasoning. Recently, in Angelov and Soares (2020) a cardinally different approach to explainability was proposed which treats it as a human-centric (anthropomorphic) phenomena rather than reducing it to statistics. Indeed, humans compare items (e.g., images, songs, and movies) in their entirety and not per feature or pixel. People use similarity to associate new data with previously learned and aggregated prototypes (Bien Tibshirani, 2011) while statistics is based on averages (Bishop, 2006).

5 Explainability-Critical Application

The frequency and importance of algorithms in applications have lead regulators and official bodies to develop policies that provide clearer accountability for algorithmic decision-making. One such example is the European Union's General Data Protection Right, which some have interpreted as a "Right to Explanation" (Goodman

& Flaxman, 2017). Although the extent of this right is in dispute, the discourse around such topics has reinforced that automated systems must avoid inequality and bias in decisions. Furthermore, they must fulfill the requirements for safety and security in safety-critical tasks. Consequently, there has been a recent explosion of interest in XAI models in different areas. Recently, it has been reported that XAI has been applied in several critical domain applications such as medicine (Holzinger et al., 2017), the criminal justice system (Dressel Farid, 2018), and autonomous driving (Cysneiros et al., 2018). In the medical domain there is a growing demand for AI approaches, most notably during the COVID-19 pandemic. However, AI applications must not only perform well in terms of classification metrics, but need also to be trustworthy, transparent, interpretable, and explainable, especially for clinical decision-making (Holzinger et al., 2017). Soares, Angelov, Biaso, et al. (2020), for example, offered an explainable DL approach for COVID-19 identification via computed tomography (CT) scans. The proposed approach was reported to surpass mainstream DL approaches such as ResNet (He et al., 2016), GoogleNet (Szegedy et al., 2015), and VGG-16 (Simonyan & Zisserman, 2014) in terms of accuracy, F1 score and other statistical metrics of performance, but critically, this approach is based on prototypes which, in this case, represent a CT scan that a radiologist can clearly understand. The prototypes are examples of CT scans of patients with or without COVID. This approach can be expanded readily to include more classes, such as "mild" or "severe" COVID, and so on, or go to the level of superpixels as in Tetila et al. (2020). Furthermore, the proposed deep neural network has a clear and explainable architecture (with each layer having a very clear meaning and using visual images of CT scans so the decision can easily be visualized). Couteaux et al. (2019) proposed an explainable DeepDream approach where the activation of a neuron is maximized by performing gradient ascent of a given image. The method has output curves that show the evolution of the features during the maximization. This favors the visualization and interpretability of the neural network and was applied for tumor segmentation from liver CT scans (Couteaux et al., 2019). Another example application of XAI is the criminal justice system. In some countries such as the United States automated algorithms are being used to predict where crimes will most likely occur, who is most likely to commit a violent crime, who is likely to fail to appear at their court hearing, and who is likely to re-offend at some point in the future (Dressel Farid, 2018). One such widely used criminal risk assessment tool is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). Although the data used by COMPAS do not include an individual's race,other aspects of the data may be correlated to race that can lead to racial biases in the predictions. Therefore, explanations of such critical decisions are necessary to favor fairness and reduce racism during the decisions (Dressel & Farid, 2018). As discussed by Soares and Angelov (2019), prototype-based algorithms can be a solution to reduce bias and favor fairness as one can check and balance the prototypes generated to guarantee a fairer decision. Moreover, the approach proposed in Soares and Angelov (2019) also provides human explainable rules to assist specialists during decision-making. Applications based on NLP also benefit from XAI. Mathews (2019) presented an explainable approach for tweet data classification based on LIME. XAI techniques for anomaly and fraud detection are also explored by different authors as a means of enhancing users' trust (Smith-Renner et al., 2019; Xie & Philip, 2018). Another application domain in which there is a growing number of applications and interest towards XAI is defined as autonomous systems (these may be airborne, maritime or land-based individual vehicles with a control system or swarms). Self-driving vehicles, for example, are automated systems that are expected to be used in possibly an unknown environment (Das Rad, 2020). In this context, the trust and acceptance of such systems require transparency, in contrast to "black-box" solutions. For example, a recent crash (on 18 March 2018) by an autonomous car owned by Uber led to the operator being charged with negligent homicide (Stilgoe, 2020) two and a half years later. It is, therefore, critically important (not only from the point of view of public perception and trust which can make or break market perspectives, but also from a purely regulatory and legal perspective) to have transparent, interpretable, and explainable, non-"black-box" models in use. This can lead to more reliable systems which are necessary to guarantee safety and meet regulations (Das & Rad, 2020). Recently, examples of prototype-based approaches were published in which XAI was used for understanding the visual scene (Soares et al., 2019) and the situation awareness of a self-driving car on a highway/motorway/autobahn through the so-called vector of affordance indicators (relative velocities and distances to the neighboring vehicles) (Soares, Angelov, Costa, et al., 2020). Not only were the accuracy, F1 score and other statistical measures reported to be comparable with, or surpass conventional DL methods, but the model was clearly explainable to a human in the form of linguistic rules and visual means. Moreover, for cases when the situation on the road is deemed to be generated from a class that was never used in training (a completely new type of scene) it was reported that conventional DL methods can make an incorrect prediction with a high confidence, which may have very damaging consequences for autonomous vehicles, passengers, legal outcomes, and trust. Instead, Soares et al. (2019) proposed a self-evolving approach, which can pro-actively learn from new situations due to its prototype nature, and also provide explainable rules. These safety mechanisms are very important for critical applications such as autonomous driving. Table 2 summarizes the applications mentioned in this section:

Table 2: XAI critical applications—summary

Method	Application
Holzinger et al. (2017)	Medicine
Dressel and Farid (2018), Soares and Angelov (2019)	Criminal justice system
Soares, Angelov, Biaso, et al. (2020)	COVID-19 identification
Couteaux et al. (2019)	Tumor segmentation
Mathews (2019)	NLP
Smith-Renner et al. (2019), Xie and Philip (2018)	Anomaly and fraud detection
Soares et al. (2019)	Novelty detection
Soares, Angelov, Costa, et al. (2020)	Soares, Angelov, Costa, et al. (2020)

6 Further Discussion

XAI aims to help humans to understand why a machine decision has been reached and whether or not it is trustworthy. Consequently, XAI is inevitably a paradigm on how to bridge machine intelligence and human intelligence, with the goal being to enable and widen the acceptance of AI systems by human subjects. In this sense, XAI can be interpreted as "AI for people."

6.1 Critical importance of XAI

Even though intelligent systems offer great possibilities, the research initiative of XAI raises concerns of giving such intelligent systems too much power without the ability to explain the decision-making process lying underneath such complex systems to domain experts (e.g., medics, lawyers, financial experts, etc.) in terms, and in a form, understandable to them. This not only helps understand specific decisions made by such systems, but also encourages researchers to create more human-like (anthropomorphic) solutions as well as inspiring the study and increased understanding of the brain as a natural information processing phenomenon. Moreover, since machines are taking over the decision process in many daily situations, user rights have to be protected. Intelligent machines still mostly cannot process abstract information or real-world knowledge unless it is converted to a form understandable by the algorithm (features, outputs, and labels). The above critical issue has become extremely important in many AI application areas. For example, the decision from an automated diagnosis system may influence the treatment plan of a patient, and doctors need to understand why such a decision was made and evaluate the underlying risks. If we consider farming-assisting autonomous drones, the farmers need to know why, when, and where drones decide to perform automated spraying of water or pesticides. Thus, a trustworthy XAI system becomes a critical prerequisite for AI to be applied to practically any real-world problem. Much research is now being focused on how to handle such kinds of problems

6.2 Bridge the gap between DL and neuroscience via XAI

DL as the state-of-the-art AI technique has its roots in the emulation of the human brain. To make deep neural networks explainable, an ultimate goal is to find a way to match human intelligence and find a way to build a humanmade "brain" that can interpret the neuronal activities in the human brain or at least, at a functionally higher level, map the deep architectures to the layered information processing units in the brain. There are two important differences between the features of current mainstream DL and the human brain. First, the human brain is more like an analogue circuit without the ability to store high precision parameters. Second, neurons in the human brain are highly interconnected instead of the carefully "handcrafted" architectures of the current mainstream DL. It is curious, therefore, that the mainstream DL literature is very critical of so-called "handcrafted" features (Goodfellow et al., 2014), but is slow to accept that the architectures it is pushing forward are "handcrafted", highly problem-specific and with multiple meta-parameters such as stride, kernel sizes, number of layers, and so on. With the above concerns, XAI can help bridge the gap between DL and neuroscience in a mutually beneficial way. On one side, neuroscience and psychology can help build rationalized XAI models that are more easily understood by humankind (Byrne, 2019; Taylor Taylor, 2020). On the other side, XAI models derived from deep neural networks can also help in understanding the mechanisms of intelligence in the human brain (Fellous et al., 2020; VU et al., 2018). The ultimate goal of XAI could be redefined as the pursuit of fully understanding how human intelligence originates from neurons.

6.3 Future directions

One promising direction for future research is to focus on prototype-based models (Angelov & Gu, 2018; Angelov Soares, 2020) rather than on abstract and highly embedded architectures. Prototype-based models are not new as such (Bien & Tibshirani, 2011)—starting with the simplest (and highly efficient example of kNN), through RBF types of ANNs and IF... THEN rules. The power of prototype-based models was noted by Tibshirani in (Bien Tibshirani, 2011), but so far these were not developed in the context of DL where they can combine a deeper architecture with a clearly explainable form of representation. Despite its efficiency, the kNN method is, strictly speaking, not a learning method, because it requires all the data to be available and stored. Some sparsity is needed which can result from simple unsupervised forms of learning such as clustering or more complex end-to-end auto-encoders. There is an established misconception that the only form of learning is parametric learning through optimization (minimization) of a cost (or loss) function. In fact, people learn by acquiring prototypes from data samples using similarity. Following this logic, the learning in prototype-based models revolves around the position and properties of the prototypes in the feature/data space as opposed to the parameters/weights-centered approach that dominates the mainstream. In addition, there is a principle difference between similarity and statistical learning (i.e., the two alternative approaches to evaluate the difference and divergence between two data items). Similarity can be defined over a pair of data items/samples while statistical measures require a large (theoretically infinite) number of independent data observations. Another promising direction is to build Turing's type-B random machines (or unorganized machines) (Jiang & Crookes, 2019; Webster, 2012), also random Boltzmann machines, which can possibly lead to a generalized AI. The inclusion of new neuro-scientific findings into XAI models will make research on XAI more rationalized, and vice versa: such a cross-disciplinary exploitation will make XAI not only meaningful for AI researchers but also help solve century-old challenges on how to understand human intelligence, ultimately. Open research questions in this area include: (i) how best to determine the network/model architecture?; (ii) how best to extract and represent features?; (iii) what are the best distance metrics and what are the implications?; (iv) which is the best optimization method?; and (v) how to determine the best set of prototypes that represent the data best (if a prototype-based method is being used)?

7 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. http://arxiv.org/abs/1711.06104

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). Neural Networks, 130, 185–194.

Angelov, P. P., Gu, X. (2018). Toward anthropomorphic machine learning. Computer, 51, 18–27.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 10, e0130140.

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. The Annals of Applied Statistics, 5, 2403–2424.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Burkart, N., & Huber, M. F. (2020). A survey on the explainability of supervised machine learning. arXiv preprint arXiv:2011.07876.

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In Proceedings of the twenty-eighth international joint conference on artificial intelligence (IJCAI-19) (Vol. 1, pp. 6276–6282).

Campbell, M., Hoane, A. J., Jr., & Hsu, F.-H. (2002). Deep blue. Artificial Intelligence, 134, 57–83. Chang, H. S., Fu, M. C., Hu, J., & Marcus, S. I. (2016). Google deep mind's alphago. OR/MS Today, 43, 24–29.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter conference on applications of computer vision (WACV) (pp. 839–847).

Chen, H., Lundberg, S., Lee, S.-I. (2019). Explaining models by propagating Shapley values of local components. arXiv preprint arXiv: 1911.11888.

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S., & Rosenberg, M. (2006). Building explainable artificial intelligence systems. In AAAI (pp. 1766–1773).

Couteaux, V., Nempont, O., Pizaine, G., & Bloch, I. (2019). Towards interpretability of segmentation networks by analyzing DeepDreams. In Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support (pp. 56–63). Springer.

Cysneiros, L. M., Raffi, M., do Prado Leite, J. C. S. (2018). Software transparency as a key requirement for self-driving cars. In 2018 IEEE 26th international requirements engineering conference (RE). IEEE (pp. 382–387).

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint arXiv:2006.11371.

Dieber, J., Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of lime. arXiv preprint arXiv:2012.00093.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4, eaao5580.

Fellous, J. M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2020). Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. Frontiers in Neuroscience, 13, 1346.

Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019). Towards automatic concept-based explanations. http://arxiv.org/abs/1902.03129

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE (pp. 80–89).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 3, 2672–2680.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38, 50–57.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their Applications, 13, 18–28.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527–1554.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI

systems for the medical domain? arXiv preprint arXiv:1712.09923.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks, 3, 551–560.

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Global explanations of neural networks: Mapping the landscape of predictions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19. Association for Computing Machinery (pp. 279–287). https://doi.org/10.1145/3306618.3314230.

Jiang, R., & Crookes, D. (2019). Shallow unorganized neural networks using smart neuron model for visual perception. IEEE Access, 7, 152701–152714.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2021). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). http://arxiv.org/abs/1711.11279

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774 https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43d
Abstract.html

MacCarthy, M. (2019). An examination of the algorithmic accountability act of 2019. Available at SSRN 3615731.

Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. In Intelligent computing-proceedings of the computing conference (pp. 1269–1292). Springer.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518, 529–533.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 427–436).

Nilsson, N. J. (2014). Principles of artificial intelligence. Morgan Kaufmann.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520–1528).

Pasquale, F. (2015). The black box society. Harvard University Press.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, pp. 9780–9784).

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4, 1883.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020) Four principles of explainable artificial intelligence.

Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20, 339–346.

Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28, 71–72.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, pp. 41–46).

Robinson, A. J., Voronkov, A. (2001). Handbook of automated reasoning (Vol. 1). Gulf Professional Publishing.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3, 210–229.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618–626).

Simonyan, K., Vedaldi, A., Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 [cs].

Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smith-Renner, A., Rua, R., Colony, M. (2019). Towards an explainable threat detection tool. In IUI workshops.

Smolensky, P. (1987). Connectionist AI, symbolic AI, and the brain. Artificial Intelligence Review, 1, 95–109.

Soares, E., Angelov, P. (2019). Fair-by-design explainable models for prediction of recidivism. arXiv preprint arXiv:1910.02043.

Soares, E., Angelov, P., Biaso, S., Froes, M. H., Abe, D. K. (2020). SARS-Cov-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-Cov-2 identification. medRxiv.

Soares, E., Angelov, P., Costa, B., Castro, M. (2019). Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In 2019 international joint conference on neural networks (IJCNN). IEEE (pp. 1–8).

Soares, E. A., Angelov, P. P., Costa, B., Castro, M., Nageshrao, S., Filev, D. (2020). Explaining deep learning models through rule-based approximation and visualization. IEEE Transactions on Fuzzy Systems, 1, 1–10.

Stilgoe, J. (2020). Who killed Elaine Herzberg? In Who's driving innovation? (pp. 1–6). Springer.

Stock, P., Cisse, M. (2018). ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In Proceedings of the European conference on computer vision (ECCV) (pp. 498–512).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1–9).

Taylor, J. E. T., Taylor, G. W. (2020). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. Psychonomic Bulletin and Review, 28, 6276–6282.

Tetila, E., Bressem, K., Astolfi, G., Sant'Ana, D. A., Pache, M. C., Pistori, H. (2020). System for quan-

titative diagnosis of COVID-19-associated pneumonia based on superpixels with deep learning and chest CT. ResearchSquare, 1, 1-13. https://doi.org/10. 21203/rs.3.rs-123158/v1

Tritscher, J., Ring, M., Schlr, D., Hettinger, L., Hotho, A. (2020). Evaluation of post-hoc XAI approaches through synthetic tabular data. In International symposium on methodologies for intelligent systems (pp. 422–430). Springer.

VU, M., Adalı, T., Ba, D., Buzsaki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V., Widge, A., Mayberg, H., Sapiro, G., Dzirasa, K. A. (2018). A shared vision for machine learning in neuroscience. Journal Neuroscience, 18, 1601–1607.

Webster, C. S. (2012). Alan turing's unorganized machines and artificial neural networks: His remarkable early work and future possibilities. Evolutionary Intelligence, 5, 35–43.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. Proceedings of the IEEE, 78, 1550–1560.

Xie, S., Philip, S. Y. (2018). Next generation trustworthy fraud detection. In 2018 IEEE 4th international conference on collaboration and internet computing (CIC). IEEE (pp. 279–282).

Young, T., Hazarika, D., Poria, S., Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13, 55–75.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. (2016). Learning deep features for discriminative localization.

Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems Richard Tomsett 1 Dave Braines 1 2 Dan Harborne 2 Alun Preece 2 Supriyo Chakraborty 3

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList Marco Tulio Ribeiro Microsoft Research Sameer Singh Univ. of California, Irvine

Semantics of the Black-Box: Can knowledge graphs help make deep learning systems more interpretable and explainable? Manas Gaur, Keyur Faldu, Amit Sheth

https://www.researchgate.net/publication/350312766$_{T}owards_{V}erifying_{R}esults_{f}rom_{B}iomedical_{N}LP_{M}achine_{L}earning_{M}$

Towards Explainable Artificial Intelligence Wojciech Samek1 and Klaus-Robert M¨uller2 ,Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany wojciech.samek@hhi.fraunhofer.de,Technische Universit¨at Berlin, 10587 Berlin, Germany,Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea,Max Planck Institute for Informatics, Saarbr¨ucken 66123, Germany klaus-robert.mueller@tu-berlin.de

Explaining Explanations: An Overview of Interpretability of Machine Learning Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology Cambridge, MA 02139 lgilpin, davidbau, bzy, abajwa, specter, lkagal@ mit.edu

Alber, M., Lapuschkin, S., Seegerer, P., H¨agele, M., Sch¨utt, K.T., Montavon, G., Samek, W., M¨uller, K.R., D¨ahne, S., Kindermans, P.J.: iNNvestigate neural networks!. Journal of Machine Learning Research 20(93), 1–8 (2019)

Ancona, M., Ceolini, E., ¨Oztireli, C., Gross, M.: Gradient-based attribution methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science 11700, Springer (2019)

Arras, L., Arjona-Medina, J., Gillhofer, M., Widrich, M., Montavon, G., M¨uller, K.R., Hochreiter, S., Samek,W.: Explaining and interpreting LSTMs with LRP. In: 14 W. Samek and K.-R. M¨uller Explainable AI:

Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science 11700, pp. 211238. Springer (2019)

Arras, L., Montavon, G., M¨uller, K.R., Samek, W.: Explaining recurrent neural network predictions in sentiment analysis. In: EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA). pp. 159–168 (2017)

Arras, L., Osman, A., M¨uller, K.R., Samek, W.: Evaluating recurrent neural network explanations. In: ACL'19 Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 113-126 (2019)

Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6541–6549 (2017)

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Wood, A.: Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134 (2017)

European Commission's High-Level Expert Group: Draft ethics guidelines for trustworthy AI. European Commission (2019)

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision 111(1), 98–136 (2015) 17Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 (2017)

Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: IEEE International Conference on Computer Vision (CVPR). pp. 3429–3437 (2017)

Fong, R., Vedaldi, A.: Explanations for attributing deep neural network predictions. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science 11700, pp. 149167. Springer (2019)

Goodman, B., Flaxman, S.: European union regulations on algorithmic decisionmaking and a "right to explanation". AI Magazine 38(3), 50–57 (2017)

Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2125–2126 (2016)

The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective Satyapriya Krishna$_1$, $TessaHar$

https://www.linkedin.com/pulse/collection-useful-slides-quotes-ai-ethics-xai-murat-durmus/

https://www.linkedin.com/feed/update/urn:li:activity:6707954976506822656/