

# AI Explainability. A Bridge between Machine Vision and Natural Language Processing

Mourad Oussalah

University of Oulu, Faculty of Information Technology, CMVS, Oulu, 90014 Finland.  
Mourad.Oussalah@oulu.fi

**Abstract.** This paper attempts to present an appraisal review of explainable Artificial Intelligence research, with a focus on building a bridge between image processing community and natural language processing (NLP) community. The paper highlights the implicit link between the two disciplines as exemplified from the emergence of automatic image annotation systems, visual question-answer systems. Text-To-Image generation and multimedia analytics. Next, the paper identified a set of natural language processing fields where the visual-based explainability can boost the local NLP task. This includes, sentiment analysis, automatic text summarization, system argumentation, topical analysis, among others, which are highly expected to fuel prominent future research in the field.

**Keywords:** Explainable AI, Machine Vision, Natural Language Processing.

## 1 Introduction

Aided by the advances in computer system computational performances and learning system theory, the success of machine learning methods in the last decade has been phenomenal in various fields, especially, computer vision and natural language processing, which enhanced the prediction and automated decision-making capabilities. This has taken machine intelligence and artificial intelligence (AI) to a new frontier that witnessed the emergence of new industry standard (e.g., industry 4.0) and human-computer interaction modes where a machine guides medical diagnosis systems, creates recommender systems, makes investment decisions and instructs driverless vehicles. On the other hand, the state-of-the-art systems in many AI applications use ensembles of deep neural networks that are even more difficult to interpret, even for skilled programmer users. This negatively impacts trust. For instance, during the PwC's 2017 Global CEO Survey [1], although it is acknowledged the substantial increase of AI market to more than \$15 trillion, 67% of the business leaders believe that this will impact negatively stakeholder trust levels in their industry in the next five years. This fosters the emergence of explainable AI research that seeks to ensure transparency and interpretability of machine learning and AI based algorithms. Indeed, many applications have seen a huge increase in demand for transparency from the various stakeholders involved at various levels of product pipeline. For instance, in precision-medicine, explanation is required to support system diagnosis outcome and clinical investigation;

in finance and management, explanation is needed to evaluate various investment scenarios with qualitative / quantitative risk evaluations; in autonomous systems, explanation enhances fault inspection and recovery based strategies. In general stakeholders are reticent to adopt techniques that are not directly trustworthy, tractable and interpretable [2], especially given the increasing scope of ethical AI [3].

Beyond academia, since 2017, the European Union's General Data Protection Regulation (GDPR) introduced the "right to explanation" which states that a user can ask for an explanation of an algorithmic decision that was made about them [4].

Strictly speaking, the need for AI explainability was recognized well earlier, and was an inherent component of many of the first AI diagnostic systems where "IF-Then" rules and inference engine were widely employed to explain the actions of the underlined expert system for instance. This was implemented in early MYCIN systems [82] that formed the basis of many subsequent medical systems; although the exact scope and nature of these rules can be debatable. In the literature, the concept of explainability is related to transparency, interpretability, trust, fairness and accountability, among others [5]. Interpretability, often used as a synonym of explainability as well, is defined by Doshi and Kim [6] as "the ability to explain or to present in understandable terms to a human".

According to Sameket al. [7], the need of explainable systems is rooted in four points: (a) Verification of the system: Understand the rules governing the decision process in order to detect possible biases; (b) Improvement of the system: Understand the model and the dataset to compare different models and to avoid failures; (c) Learning from the system: "Extract the distilled knowledge from the AI system"; (d) Compliance with legislation (particularly with the "right to explanation" set by European Union): To find answers to legal questions and to inform people affected by AI decisions.

Lewis [8] states that "to explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event –explanatory information – tries to convey it to someone else". Halpern and Pearl [9] define a good explanation as a response to a Why question, that "(a) provides information that goes beyond the knowledge of the individual asking the question and (b) be such that the individual can see that it would, if true, be (or be very likely to be) a cause of".

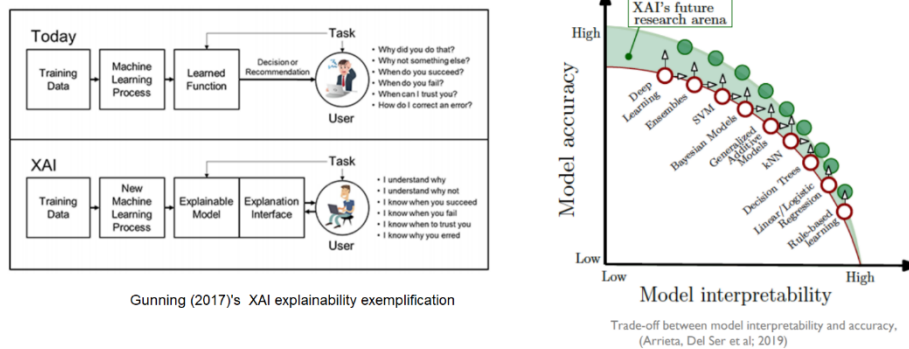
Miller [10] extracts four characteristics of explanations: "explanations are contrastive" (why this and not that), "explanations are selected in a biased manner (not everything shall be explained)", "probabilities don't matter" and finally "explanations are social".

Traditionally, transparency has always been at odds with performance where an increase in transparency is often translated into a decrease in system performance because of large number of parameters that require tuning [11], see Fig. 1 for exemplification. Therefore, a trade-off between the level of transparency and performance required.

Encompassing the broad scope of the explainability and its multi-disciplinary nature, this paper attempts to reconcile explainable AI research on two complementary fields: Machine Vision System (MVS) and Natural Language Processing (NLP), trying to survey the explainable AI in each field and seek complementary aspects in a way to boost fruitful XAI research in the two fields.

## 2 Background

We will adopt in this paper Gunning definition of Explainable Artificial Intelligence (XAI) [12]: “XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners”, see Figure 1.



**Fig 1.** XAI concept as described in Gunning’s [12] and Performance-interpretability trade-off

This definition brings together two concepts; namely, understanding and trust that need to be addressed. Such concepts are ultimately linked to several other aspects that overlap with cognitive operations of understanding or comprehension tasks. This includes for instance causality, transferability, informativeness, fairness and confidence [13]

Regardless of the type of applications involved or the system inputs, the explanation differs according to the underlying chosen criterion. Especially, explanation methods and techniques for ML interpretability can be classified according to different criteria.

- *Pre-Model vs. In-Model vs. Post-Model*

Interpretability methods can be classified depending whether this is applied before (pre-model), during (in-model) or after (post-model) the machine-learning model [14].

Pre-model interpretability techniques tackle the data itself regardless of the employed ML model, focusing on the structure of the inputs and associated features with their visualization. Intuitive features and sparsity are some properties that help to achieve data interpretability. This includes techniques related to descriptive statistics to data visualization methods including Principal Component Analysis (PCA) [15], t-SNE (t-Distributed Stochastic Neighbor Embedding) [16], and clustering methods, such as MMD-critic [17] (Maximum Mean Discrepancy) and k-means [18]. Hence, data visualization is critical for pre-model interpretability.

In-model interpretability concerns ML models that have inherent interpretability in it, either with or without constraints, being intrinsically interpretable. Post-model interpretability refers to improving interpretability after building a model or model training (post hoc). This answers the question: what else can the model tell us?

- **Model-Specific vs. Model-Agnostic**

Model-specific interpretation restricts the analysis to specific model classes. For example, the interpretation of weights in a linear model is a model-specific interpretation [19]. In contrast, model-agnostic methods apply to any ML model after training phase (post-hoc), so without having access to model inner workings (i.e., weighting). By default, all post-hoc methods are model-agnostic since they were applied after the training.

In essence, two approaches can be distinguished to explain ML model prediction through either a global method that treats the group of predictions of interest as if it was the whole dataset or by applying local methods on each individual prediction followed by aggregating and joining these individual explanations afterwards [13,20].

- *Local versus global explanation*

According to the scoop of interpretability, one distinguishes global interpretability, which concerns comprehending the entire model behavior and local interpretability, which rather focuses on a single prediction, or a single feature behavior. For instance, Yang et al. [21]’s GIRP (global mode interpretation via recursive partitioning) builds a global interpretation tree. Nguyen et al. [22] advocate an activation-maximization based approach for global explanation. One of the most popular local interpretability model is LIME model [23], which enables approximating ML model in the neighborhood of any prediction of interest.

Other techniques for local explainability models include decomposition models (omitting some features of the original dataset and attempt to combine the effects of various features), Shapely explanations [24], sensitivity maps, saliency maps [26].

- *Type of explanation*

This includes Feature summary (either through visualization or textual input), Model internals (model specific), Data point (output data points that make the model interpretable), Surrogate intrinsically interpretable model— through approximation of ML model either locally or globally with an intrinsically interpretable model.

- *Simulatability*

This refers to comprehending how the model makes decisions, grounded on a holistic view of the data features and individual components (e.g., weights, parameters) in order to simulate the overall system.

- *Visualization and interaction*

Popular visualization techniques applied in ML interpretability include partial-dependence plot [27], surrogate models [28, 29], individual conditional expectation [30]. Depending on the stage where the visualization techniques is conducted, one can also distinguish pre-model, in-model or post-hoc based visualization. Similarly, visualization can be performed for local or global like interpretability purpose.

**Table 1.** Review of main XAI techniques

Techniques	Global (G) / Local (L)	Model Specific (Sp) / model Agnostic (A)	Pre-model, In-model, Post-model (Po)	Approximation (Ap) / Sensitivity (S)	Visualization Interaction	Simulatability	References
Decision Trees	G	Sp	All	x			[39, 40]
LIME	L	A	Po	x	Yes	Yes	[23, 41]
Shapely explanations	L	A	Po	x	Yes	Yes	[24]
Rule extraction	G / L	A	Po	x	Yes	Yes	[43], [44]
Decomposition	L	A	Po	Ap	Yes	Yes	[32]
Activation-Maximization	G / L	A	Po	x	x	x	[22]
Surrogate models	G / L	A	Po	S	Yes	x	[28-29]
Individual Conditional expectation	L	A	Po	x	Yes	x	[30]
Model distillation	G	A	Po	x	Yes	Yes	[34]
Feature importance	G / L	A	Po	Ap / S	Yes	x	[45]
Saliency map	L	A	Po	S	Yes	x	[26]
Sensitivity analysis	G / L	A	Po	S	Yes	x	[36]
Counterfactuals explanations	L	A	Po	x	Yes	x	[46]
Tree View	G / L	A	Po	x	Yes	x	[48]
Rule Set	G	Sp	Po	x	Yes	x	[49]
DecText	G / L	A	Po	x	x	x	[50]
DeepLift	G	A	Po	x	Yes	x	[51]
Layer-Wise Relevance Propagation	G / L	A	Po	Ap	x	x	[47]
Fuzzy Inference System	G	A	Po	Ap	Yes	Yes	[52]

- *Approximation versus sensitivity*

Using the universal approximator property of neural network system, several interpretable approximation models have been proposed to gain insights into the functioning of the black-box of the ML model. For instance, rule-based approximation [31] approximates the decision-making process in ML model by utilizing the input and the output of the ML only. Decompositional approaches look at extracting rules at the level of individual units within ML model [32]. This includes Orthogonal search-based rule extraction proposed in [33] for medicine application. Another approximation method cited in the literature is model distillation [34], which acts as a model compression algorithm from deep network to shallow networks. DarkSight [35] combines ideas from distillation model and visualization to explain the black-box functioning.

On the other hand, sensitivity analysis [36] focuses on how black box model is influenced by its input weight perturbations. Feature importance as in Fisher et al. [37]'s Model Class Reliance or SFIMP [38] for permutation based shapely feature importance are other sensitivity like analysis for explanation tackling.

### 3 Link between Image and text in explainability

Intuitively, the link between image and text is either explicit or implicit from several standpoints:

#### *Purpose and outcome expectation*

Both image processing and NLP based XAI system do share the same purpose of using ML model for classification purpose, and therefore, seeking an explanation of the ML outcome using the XAI model. They may seek a global / local, model-specific or agnostic, pre-training, in-training or post-training explanation. Although, the classification task might be different for NLP and image processing cases.

#### *Universality of many XAI tools*

Tools like LIME framework that enables observing the explainer function on the correctly predicted instance can be applied regardless the context of application domain (e.g., image or text based exploration). Similar reasoning applies to many visualization toolkits, e.g., heat-map that visualizes the extent to which each element contributes to the prediction result, saliency map, feature importance map, among others, that are independent of the application context.

#### *Structure of multimedia and social media posts*

With the advances in Web 2.0 technology that enabled the users to post various types of files (text, images, multimedia) and the memory efficiency for handling large scale multimedia files, the need for building a capacity to handle equally image and textual inputs on the same setting is growing. This motivates the development of unified frameworks in XAI to handle both types of inputs as well.


#### *Development of automatic annotation services*

Automatic image annotation is the process of automatically creating textual based description for the different regions of the image highlighting the content of the images. This is especially important to identify sensitive content on online media platforms. With the advances in deep learning technology and large-scale image database, several tools were made available to scientific community for this task. This includes Google Cloud Vision [53], GoogLeNet [54], a deep learning model trained on the ILSVRC dataset.

#### *Development of visual question answering services*

Visual Question Answering (VQA) is the task of addressing open-ended questions about images; namely, given an image and a natural language question about that image, the task is to provide an accurate natural language answer, see Fig.2 for an example. Typically, VQA requires visual and linguistic comprehension, language grounding capabilities as well as a common-sense knowledge. A variety of methods have been developed [55, 56]. In the latter, the vision component of a typical VQA system extracts visual features using a deep convolutional neural network (CNN), then linguistic components encode the question into a semantic vector using a recurrent neural network

(RNN). An answer is then generated conditioned on the visual features and the question vector.

	Visual-Question	Only-Question
		
What is the animal in the water?	dog dog dog	duck duck guppy
How many people are swimming?	15 15 15	2 3 3

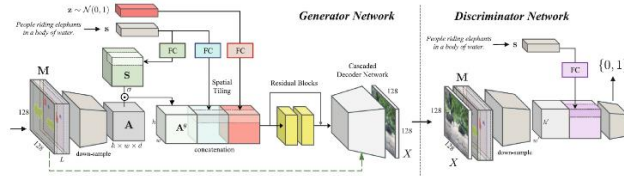
**Fig. 2.** Example of Visual-Question answer systems

#### *Development of Text to Image generation*

Similarly to the previous visual question answering, the problem of explaining textual content through visual representation is also important for education and learning purposes. This has also been investigated by many computer vision scientists. Indeed, deep generative models [22] have been proposed to address the task of generating appropriate images from natural description by inferring a semantic layout, which is then converted into image using image generator. Methods based on conditional Generative Adversarial Network (GAN) have been employed in several text-to-image synthesis tasks and competitions [57, 58] and tested on large scale dataset such as birds [59], flowers [60], MS-COCO [61]. In this regard, the task of image generation is viewed as a problem of translating semantic labels into pixels. Nevertheless, the complexity of the reasoning cannot be ignored. Especially, learning a direct mapping from text to image is not straightforward and layout generator requires several constraints to enhance its practicality due to the vast amount of possibilities of potential image candidates that fit a given textual utterance, see, for instance, the example in Fig. 3.



**Fig. 3.** Example of Text-to-Image based explanation



**Fig 4.** Architecture for image generator conditioned on the text description and semantic layout generator in [79].

## 4 Potential benefits to NLP community

### 4.1 Word-sense disambiguation

Word-sense disambiguation aims to assign an appropriate sense to the occurrence of a word in a given context. For instance, the meaning of word “chair” in the sentence “I have been awarded a chair in Computer Science” is different from that in the sentence “I bought a chair in the city market today”, where the sense of the target word “chair” is chosen among the set of senses listed in a given dictionary. Typically, standard Lesk’s algorithm [62] looks into the number of overlapping words of each sense of the target word with the underlined context (sentence), so that the sense that yields the highest number of overlap is used to disambiguate the target word. Other variants of Lesk’s algorithm as well as supervised and/or semi-supervised algorithms have been proposed for word-sense disambiguation tasks [63-64, 80].

Making use of visual description raised, for instance, by text-to-image mapping can provide insights into word-sense disambiguation task. This assumes that an overall visual representation is generated for the whole sentence for each sense of the target word, and appropriate metrics are constructed to quantify the relevance and commonsense of each global visual representation.

Visualization techniques issues from XAI can also be accommodated for the word-sense disambiguation feature. For instance, various senses of target words can play the role of features and utilize LIME-like-approach to visualize the contributions of the various senses, and thereby, disambiguate accordingly. Similarly, the emergence of graph based approaches for word sense disambiguation where, in the same spirit as Navigli and Lapata [64], the senses are mapped to a graph representation where the graph is built using various connectivity algorithms such as PageRank, hyperlink induced topic search, key player problem, various senses can be ranked accordingly, and, thereby, handle disambiguation task.



## 4.2 Text Argumentation theory

With the emergence of Dung's Argumentation Framework [65], a central approach for performing reasoning within argumentation in artificial intelligence and natural language processing become tenable. This opens up opportunities for legal text analysis, medical science and opinion mining. In this course, arguments are viewed as abstract entities, so that the use of argumentation semantics, i.e., criteria for finding acceptable sets of arguments, suffices to reason in an argumentative way for a variety of application scenarios. Arguments are supposed to support, contradict, and explain statements, and they are used to support decision-making. What may constitute an argument is very much context dependent. In natural language processing, this can be short utterance that fits a given ontology or might be extracted using text summarization like technique from a large or a multi-document source file. The possibility to represent both positive and negative views using the employed argumentation framework provides opportunity to sustain debate and boost interactions. Recently, abstract argumentation has been suggested to explain predictions of neural network system and diaelectrically explainable prediction [66]. This ultimately builds bridge with XAI and offers nice opportunities to use the abstract argumentation framework as a means to derive explanation and interpretability.

## 4.3 Sentiment Analysis

Sentiment analysis refers to the use of NLP, text analysis and computational linguistics techniques to systematically identify, extract and quantify affective states and subjective information, classifying the polarity of a given text at sentence, document or multi-document level in order to find out whether the expressed opinion is positive, negative or neutral. This has been extensively employed in a range of applications ranging from marketing to customer services to clinical medicine. Key approaches to sentiment analysis include knowledge-based techniques, which classify text based on affect categories according to presence of affect words such as happy, sad using affective lexical database of lexicon dictionary. Furthermore, supervised and machine-learning like techniques have also been populated for the same purpose [67]. Several open sources are made available for the purpose of sentiment analysis. This includes, Python NLTK, TextBlob, Pattern.en, RapidMiner, Stanford's CoreNLP, SentiStrength, among others.

Through explainability-based reasoning, sentiment analysis can be boosted a step-further to provide the reason for the sentiment score. For instance, in [68], the authors proposed layer-wise relevance propagation model for explaining recurrent network architecture for sentiment analysis.

Zhang et al. [69] proposed an Explicit Factor Model (EFM) based on phrase-level sentiment analysis to generate explainable recommendations.

Interestingly, the presence of contrastive statements from opinionated documents in sentiment analysis context opens up the door wide to application of more advanced argumentation system, reinforcement learning or Markov-Chain based reasoning in the

same spirit as [70] inherited from question-answer system analysis. Similarly, the emergence of multimedia documents in social media platforms provides opportunities to mix-up image-analysis & text-analysis based reasoning. For instance, face-emotion recognition in videos can provide useful insights into sentiment polarity of the underlined textual input.

#### **4.4 Topical Modelling**

Since the emergence of Latent Dirichlet Allocation (LDA) [71], the task of automatic discovery of topics in a textual document has seen a new landmark. In essence, LDA introduces sparse Dirichlet prior distributions over document-topic and topic-word distributions, encoding the intuition that documents cover a small number of topics and that topics often use a small number of words. Topic models are a form of unsupervised machine learning, in that the topics and mixture parameters are unknown and inferred solely from the data where each topic is represented by its  $N$  most probable words. Humans can judge whether words of a given topic (cluster) form interpretable concept(s). Therefore, it is important to seek automatic alternative to measure the interpretability of the outputted set of words of each topic. A commonly employed approach is based on the co-occurrence analysis, stipulating that words that have high frequency of co-occurrence (either within the document under investigation or in a more wider corpus) would indicate high coherence and relatedness, as for words caught and fever for instance [72]. The development of word embedding promoted by Google researchers has also promoted the so called embedded topic model [73] where each word is modelled as a categorical distribution whose natural parameter is the inner product between a word embedding and an embedding of its assigned topic. This has shown to discover interpretable topics even with large vocabulary. On the other hand, the development of interactive topic modelling [74], where more interaction modes with system output are enabled, offers a nice setting to apply a range of visualization tools developed in the context of explainable AI for this purpose.

#### **4.5 Automatic textual summarization**

Automatic text summarization has been a hot topic in NLP focusing on how to summarize the main content of the document while preserving the semantic meaning and key messages conveyed by the original document in a way to reduce redundancy and maximize the diversity. Typically, two streams have been identified in the literature [75]. Extractive summarization where the summary is constituted of a selected sentences from original document through some scoring analysis mechanism that takes into account sentence similarity, location, presence of selected keywords, among others, and abstractive summarization where the summary sentences may be different from that of original documents. Extractive summarization is by far the most investigated research stream in automatic summarization. Various graph-based approaches have been put forward for extracting relevant sentences. Examples include TextRank [76] where the nodes are sentences and the edges the relations (which is context dependent,

e.g., semantic similarity beyond certain threshold) between the sentence, and the importance of a given sentence is quantified using PageRank like algorithm. Similarly, latent semantic analysis [77-78], which provides a lower dimensional representation of words, has also been applied to summarization purpose [81].

Strictly speaking, explainable research benefits summarization from both directions. First, summarization can be used as a tool to construct and identify arguments that can be used to guide the explanation process. Second, the interactive tools, LIME like approach can also be adapted to boost the sentence weighting scheme, which, in turn impact the outcome of the summarization task.

## 5 Conclusion

Explanation methods are a promising approach to leverage hidden knowledge about the workings of neural networks and black-box systems, promoting transparency and interpretability of the results in the light of the new data protection EU directive on the “right of explanation”. This paper attempted to review the state of art of explainability methods focusing on intertwine between image processing and natural language processing fields in a way to promote fruitful development of new explanation framework. Especially, the paper highlights the implicit link between the two research fields through, e.g., automatic image annotation, visual question-answer systems, Text-To-Image generation, multimedia analytics in addition to the overall input-output like system analysis. On the other hand, this review has also identified several NLP research fields that would benefit from visual explainability based approach. This includes, wordsense disambiguation, sentiment analysis, argumentation theory, automatic text summarization and topical modelling.

There are several interesting future research directions to explore further. An interesting direction is semi-supervised learning of the model using a large set of partially annotated data. For instance, we can exploit a small number of fully annotated images and a large number of partially annotated images (*e.g.* images with only text descriptions), which allows the developed model to exploit large-scale datasets, such as the Google Conceptual Caption dataset. The paper also opens up new research directions in multimedia analytics, text summarization and abstract argumentation logic.

### Acknowledgment

This work is partly supported by the H2020 YoungRes (# 823701) project, which is gratefully acknowledged.

## References

- 1 . Oxborough, C., and Cameron, E.: Explainable AI, PWC report, (2020). <https://www.pwc.co.uk/services/risk-assurance/insights/explainable-ai.html>, accessed July 2020
- 2 . Grice, H.P.: Logic and Conversation, in: Syntax and Semantics 3: Speech arts. 41–58 (1975).

- 3 . Conati, C., Porayska-Pomsta, K., and Mavrikis M.: AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. arXiv preprint arXiv:1807.00154 (2018).
- 4 . Goodman, B., and Flaxman, S.: European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, **38**(3), (2017).
- 5 . Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., and Kankanhalli, M.: Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI. Association for Computing Machinery, Montreal, Canada, (2018), <https://doi.org/10.1145/3173574.3174156>
- 6 . Doshi-velez, F., and Kim, B.: A Roadmap for a Rigorous Science of Interpretability. CoRR, abs/1702.08608, (2017), <http://arxiv.org/abs/1702.08608>
- 7 . Samek W., Wiegand T., and Müller, K.R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries Special Issue No.1*, pp-39-48, (2017).
- 8 . Lewis, D.: Causal Explanation. In *Philosophical Papers. Vol II*. Oxford University Press, New York, Chapter Twenty two, 214–240, (1986).
- 9 . Halpern, J. Y., and Pearl, J.: Causes and Explanations : A Structural-Model Approach . Part II : Explanations. **56**(4), 889–911. (2005), <https://doi.org/10.1093/bjps/axi148>
- 10 . Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267** (2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- 11 . Arrieta, Del Ser et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, *Information Fusion* **58**, pp. 82-115, (2019).
- 12 . Gunning, D.: Explainable Artificial Intelligence (XAI ). Technical Report. 1–18 pages, (2017).
- 13 . Lipton, Z.C.: The mythos of model interpretability, *Queue* **16**(3), 30:31–30:57 (2018)
- 14 . Kim, B., Doshi-Velez, F.: Introduction to Interpretable Machine Learning. In Proceedings of the CVPR'18 Tutorial on Interpretable Machine Learning for Computer Vision, Salt Lake City, UT, USA, (2018)
- 15 . Jolliffe, I.: Principal component analysis. In *International Encyclopedia of Statistical Science*; Springer: Berlin, Germany; 1094–1096, (2011).
- 16 . Maaten, L.Y.D., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605, (2008).
- 17 . Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, pp. 2280–2288, (2016).
- 18 . Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, **28**, 100–108 (1979)
- 19 . Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6**, 52138–52160, (2018)
- 20 . Molnar, C.: *Interpretable Machine Learning*. (2019), Available online: <https://christophm.github.io/interpretable-ml-book/> (Accessed July 2020)
- 21 . Yang, C., Rangarajan, A., and Ranka, S.: Global model interpretation via recursive partitioning. Online .. Available: <https://arxiv.org/abs/1802.04253> (2018)
- 22 . Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 3387–3395 (2016)
- 23 . Ribeiro, M. T., Singh, S., and Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. (2016)
- 24 . Lundberg, S.M., and Lee S.L.: A unified approach to interpreting model predictions, in: *Proc. Adv. Neural Inf. Process. Syst.*, pp. 4768–4777, (2017).

- 25 . Cortez, P., and Embrechts, M.J.: Opening black box data mining models using sensitivity analysis, in: *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, pp. 341–348, (2011)
- 26 . Smilkov, D., Thorat, N., Kim, B., ViOgas F., and Wattenberg, M.: SmoothGrad: Removing noise by adding noise. (2017). Available: <https://arxiv.org/abs/1706.03825>
- 27 . Green D.P., and Kern, H.L.: Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees, in: *Proc. Annu. Summer Meeting Soc. Political Methodol.* pp. 1–40, (2010)
- 28 . Bastani, O., Kim, C., and Bastani, H.: Interpretability via model extraction. (2017). Available: <https://arxiv.org/abs/1706.09773>
- 29 . Thiagarajan, J.J., Kailkhura, B., Sattigeri, P., and Ramamurthy, K.N.: TreeView: Peeking into deep neural networks via feature-space partitioning. Online .. Available: <https://arxiv.org/abs/1611.07429>, (2016)
- 30 . Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, *J. Comput. Graph. Stat.*, **24**(1), 44–65, (2015)
- 31 . Frank, E., and Witten, I.H.: Generating accurate rule sets without global optimization, *ICML'98*, p-144-151, (1998)
- 32 . Robnik, M., Ikonja and Kononenko, L.: Explaining classifications for individual instances, *IEEE Trans. Knowl. Data Eng.*, **20**(5), 589–600, (2008)
- 33 . Etchells, T.A., and Lisboa, P. J. G.: Orthogonal search-based rule extraction (OSRE) for trained neural networks: A practical and efficient approach, *IEEE Trans. Neural Netw.*, **17**(2), 374–384, (2006)
- 34 . Tan, S., Caruana, R., Hooker, G., and Lou, Y.: Detecting bias in black-box models using transparent model distillation. Online .. Available: <https://arxiv.org/abs/1710.06169>, (2018)
- 35 . Xu, K., Park, D.H., Yi, D.H., and Sutton, C.: Interpreting deep classifier by visual distillation of dark knowledge. Online .. Available: <https://arxiv.org/abs/1803.04042> (2018)
- 36 . Cortez, P., and Embrechts, M.J., Using sensitivity analysis and visualization techniques to open black box data mining models, *Inf. Sci.*, **225**, 1–17, (2013)
- 37 . Fisher, A., Rudin, C., and Dominici, F.: Model class reliance: Variable importance measures for any machine learning model class, from the ‘rashomon’ perspective. Online .. Available: <https://arxiv.org/abs/1801.01489>, (2018)
- 38 . Casalicchio, G., Molnar, C., and Bischl, B.: Visualizing the feature importance for black box models. Online .. Available: <https://arxiv.org/abs/1804.06620>, (2018)
- 39 . Schetinin V., et al.: Confident interpretation of Bayesian decision tree ensembles for clinical applications, *IEEE Trans. Inf. Technol. Biomed.*, **11**(3), 312–319, (2007).
- 40 . Hara, S., and Hayashi, K.: Making tree ensembles interpretable. Online .. Available: <https://arxiv.org/abs/1606.05390>, (2016)
- 41 . Ribeiro, M.T., Singh, S., and Guestrin, C.: Anchors: High-precision model-agnostic explanations, in: *Proc. AAAI Conf. Artif. Intell.*, pp. 1–9, (2018)
- 41 . García, S., Fernández, A., and Herrera, F.: Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems, *Appl. Soft Comput.*, **9**(4), 1304–1314, (2009).
- 42 . Wang, F., and Rudin. C.: Falling rule lists, in: *Proc. 18th Int. Conf. Artif. Intell. Statist. (AISTATS)*. San Diego, CA, USA: JMLR W&CP, pp. 1013–1022, (2015)
- 43 . Ras, G., Van Gerven, M., and Haselager, P.: Explanation methods in deep learning: Users, values, concerns and challenges. Online .. Available: <https://arxiv.org/abs/1803.07517>, (2018)
- 44 . Johansson, U., König, R., and Niklasson, I., The truth is in there—Rule extraction from opaque models using genetic programming, in *Proc. FLAIRS Conf.*, pp. 658–663, (2004).

- 45 . Casalicchio, G., Molnar, C., and Bischl, B.: Visualizing the feature importance for black box models. Online .. Available: <https://arxiv.org/abs/1804.06620>, (2018)
- 46 . Wachter, S., Mittelstadt, B., and Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Online .. Available: <https://arxiv.org/abs/1711.00399>, (2017)
- 47 . Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLoS ONE*, **10**(7), e0130140. (2015)
- 48 . Thiagarajan, J.J., Kailkhura, B., Sattigeri, P., and Ramamurthy, K.N.: Treeview: Peeking into deep neural networks via feature-space partitioning. arXiv preprint arXiv:1611.07429, (2016).
- 49 . Wang, T., Rudin, C., Velez-Doshi, F., Liu, Y., Klamp, E., and MacNeille, P.: Bayesian rule sets for interpretable classification. In Data Mining (ICDM), IEEE 16th International Conference on, pages 1269–1274, (2016)
- 50 . Boz, O.: Extracting decision trees from trained neural networks. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 456–461. ACM, (2002)
- 51 . Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv:1605.01713, (2016)
- 52 . Zhou, A.M., Gan, J.Q.: Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling, *Fuzzy Sets and Systems* **159** (23), 3091–3131, (2008)
- 53 . <https://cloud.google.com/vision>
- 54 . Szegedy C. et al.: Going deeper with convolutions, arXiv, eprint 1409.4842, also in CPRV'15, (2004)
- 55 . Xu, H., and Saenko, K.: Ask, Attend and Answer: Exploring question-guided spatial attention for visual question answering. In Proceedings of ECCV'16, (2016)
- 56 . Lu, J., Yang, J., Batra, D.: Parikh. Hierarchical question image co-attention for visual question answering. In Advances In Neural Information Processing Systems (NIPS2016), pp 289–297, (2016)
- 57 . Reed S. et al.: Generative adversarial text to image synthesis. In: ICML'16, pp. 1060–1069, (2016)
- 58 . Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS, pp. 217–225, (2016)
- 59 . Welinder, P. et al.: Caltech-UCSD Birds 200. Technical Report. CNS-TR-2010-001, California Institute of Technology, (2010)
- 60 . Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, pp. 722–729, (2008)
- 61 . Lin T.Y. et al.: Microsoft COCO: common objects in context. In: ECCV, pp. 740–755, (2014)
- 62 . Lesk M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proc. of SIGDOC, SIGDOC, pp. 24–26, (1986)
- 63 . Mihalcea, R.: Knowledge-based methods for WSD. In: Word Sense Disambiguation: Algorithms and Applications, Text, Speech and Language Technology, pp. 107–132. Springer, Dordrecht, (2006)
- 64 . Navigli, R., and Lapata, M.: Graph connectivity measures for unsupervised word sense disambiguation. In IJCAI International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 1683–1688, (2007)
- 65 . Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence*, **77**(2), 321 – 357, (1995)
- 66 . Cocarascu, O., Stylianou, A., Cyrus K., and Toni, F.: Data-Empowered Argumentation for Dialectically Explainable Predictions, 24th European Conference on Artificial Intelligence – ECAI, (2020)
- 67 . Tsytarau M., and Palpanas, T.: Survey on mining subjective data on the web, *Data Min Knowl Discov*, **24**, 478–514, (2012)
- 68 . Arras, L., Horn, F., Montavon, G., Muller, K.R., and Samek W.: Explaining predictions of non-linear classifiers in NLP, arXiv preprint arXiv:1606.07298, (2016)

- 69 . Zhang Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., and Ma, S.: Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in Proc of the 37th ACM SIGIR pp. 83–92, (2014)
- 70 . Sherstov, A.A., and Stone, P.: Improving action selection in mdp’s via knowledge transfer, in AAAI, vol. 5, pp. 1024–1029, (2005)
- 71 . Blei, D.M., and Lafferty, J.D.: Topic Models. Chapman & Hall/CRC, (2009)
- 72 . Lau, H.J., Newman, D., and Baldwin, T.: Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In EAC, (2014)
- 73 . Dieng, A.B., Ruiz, F.R., and Blei, D.M.: Topic Modelling in Embedding Spaces, arXiv:1907.04907v1 cs.IR ., (2019)
- 74 . Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A.: Interactive Topic Modeling. Machine Learning **95**, 423–469, (2013)
- 75 . Nenkova A., McKeown K.: A Survey of Text Summarization Techniques. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. [https://doi.org/10.1007/978-1-4614-3223-4\\_3](https://doi.org/10.1007/978-1-4614-3223-4_3), (2012)
- 76 . Mihalcea, R., and Tarau, P.: Textrank: Bringing order into text. In Proceedings of the conference on empirical methods in natural language processing, (2004)
- 77 . Gong, Y., and Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th ACM SIGIR pp 19– 25, (2001)
- 78 . Steinberger, J., and Jezek, K.: Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM, 4:93–100, (2004).
- 79 . Hong, S., Yang, D., Choi, J., and Lee, H.: Interpretable Text-to-Image Synthesis with Hierarchical Semantic Layout Generation, in Samek et al. (Eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, Germany. (2019)
- 80 . Mohamed, M., and Oussalah, M.: A Hybrid Approach for Paraphrase Identification Based on Knowledge-enriched Semantic Heuristics, Language Resources and Evaluation, <https://doi.org/10.1007/s10579-019-09466-4>, (2019).
- 81 . Mohamed, M., and Oussalah, M.: SRL-ESA-TextSum: A Text Summarization Approach Based on Semantic Role Labeling and Explicit Semantic Analysis, Information Processing and Management, (2020) <https://doi.org/10.1016/j.ipm.2019.04.003>
- 82 . Buchanan, B.G., and Shortliffe, E.H.: Rule Based Expert Systems: The MYCIN Experiment of the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley, (1984)