Explainable AI

A Taxonomy of the Different Methods on the Field

Jan Rodríguex Miret

June 20, 2021

Contents

1 Introduction

Nowadays, most artificial intelligence (AI) state-of-the-art systems consist of artificial neural networks with hundreds of neuron layers that account for millions or even billions of neurons and parameters. Furthermore, the nature of these networks makes it nearly impossible to know what were the actual reasons for a predicted output of the model.

This lack of interpretability is a problem because we are losing control and understanding of these systems while relying on them for more and more tasks and decisions. These systems, though, are not exempted from bias and unexpected behaviors, and they can be easily fooled with conscious or unconscious tricks.

If we just keep building more complex networks and blindly trusting their outcomes in an increasingly autonomous world, the consequences could be catastrophic at some point. This is why we need the proper tools to ensure that the system is behaving in a comprehensible manner, understanding what is going on at each time step and why it is returning this solution and not another. This is the goal of Explainable AI (XAI), to obtain AI models that satisfy this set of properties.

In this document, a historical overview and explanation of why this problem has gained attention recently are introduced first (Section 2), along with a clearer definition of what XAI is and its differences from other similar terms 3. Then, some of the most common families of methods and the specific techniques for different types of models are presented in Section 4. After that, the international regulations, programs, and interests are explained in Section 5. Finally, the challenges still remaining in the field and the conclusions of this work are discussed in Section 6.

2 Contextualization of the problem

Artificial Intelligence has suffered many highs and lows since its birth in the mid-twentieth century, where a period of incredible growth and investment precedes a loss of faith in the field and a withdrawal of resources. There have been 3 major cycles throughout history.

The first golden age of AI is usually considered to last from 1956 to 1974, with systems that were based on manually specified rules. These rules had to be defined by an expert through traditional conditional programming which sometimes required to extract meaningful features from the raw data. This approach was sufficient to create some sort of recognizable intelligent behavior in some rather simple tasks in a variety of domains like natural language or machine vision, among others. Many AI systems made use of heuristics to solve computationally unfeasible problems, especially in reasoning and planning. Also, the first and most simple neural network was introduced by Roosenblatt in 1958: the perceptron [12], although its applications were very limited.

These systems were very primitive and rather easy to follow. Engineers and designers had to specify exactly what to look for, meaning that they had complete control over what was going on and the model was predictable.

Then, after the first AI winter, came another flourishing period of success and development in the field, from 1980 to 1987. The backpropagation algorithm for efficiently training neural networks was proposed in 1986 by Hinton and Rumelhart [13], though connectionism was still lacking sufficient computational resources to create complex neural networks. Researchers started to gain more interest in using statistical methods to create these AI models, in what is called machine learning.

Machine learning models are built by applying statistics to a data set of samples and trying to distinguish the existent patterns in them. This kind of model, thus, relies on the data that is being

used and the specific task that it is trying to solve. Again, features must be extracted in most cases to obtain meaningful results. The complexity of machine learning-based models can be much higher than a simple rule-based one, but still can be quite interpretable in some cases.

Finally, after another AI winter that lasted until 1993, resources invested in AI research just kept growing and growing. Computational power had been increasing at a tremendous pace (Moore's law), enabling new algorithms and techniques to be feasibly performed. In 2012, AlexNet made use of the backpropagation algorithm to train a deep convolutional neural network (CNN) using graphics processing units (GPUs). The model won the ImageNet Large Scale Visual Recognition Challenge and entailed a major improvement over previous techniques. This and many other breakthroughs further arouse the attention of countries and enterprises to invest in the field.

Since then, neural networks have become larger and larger, in what is called deep learning (artificial neural networks with several layers). The natural language model GPT-3 reached an astounding 175 billion parameters in 2020 [4]. Nonetheless, with this increase in model complexity, systems are treated as "black boxes" and are being evaluated end-to-end: giving some input and assessing the quality of the corresponding output but knowing very little about the steps.

This problem of not being able to understand its internal "thought process" holds true disregarding the type of learning used to train the network: supervised, unsupervised, reinforcement, etc [10]. In fact, the problem is not something uniquely attributable to neural networks, but also to much simpler machine learning techniques if not performed with certain guarantees (for instance if we do not know what the data is representing).

However, in the case of deep learning the "black box" problem is more exaggerated compared to traditional simpler techniques. Moreover, it is intrinsic to the nature of artificial neural networks and how they work. The augmenting use of these AI-based systems in many fields and our daily lives is posing an also augmenting threat to society and mistrust towards AI.

This is why we have seen efforts towards Explainable AI in recent years, aiming at solving this opacity problem in current systems. Figure 1 shows a clear growth in interest in this topic since 2016.

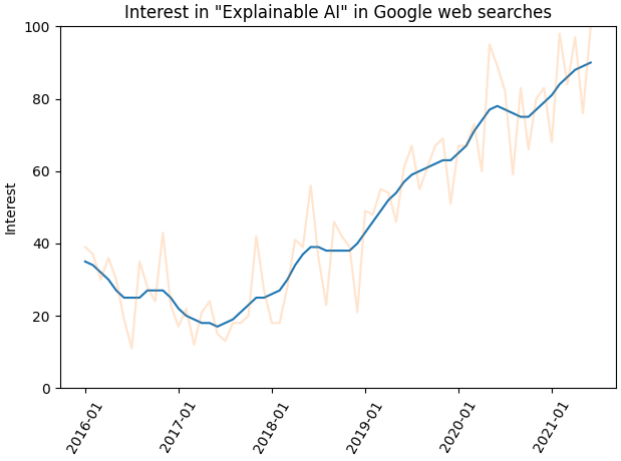


Figure 1: Monthly interest in the topic "Explainable artificial intelligence" in Google web searches from January 2016 to June 2021. Results were smoothed using a Gaussian filter with $sigma = 3$ to see clearer patterns (in blue). Data from 2004 to 2016 was not used because in 2016 they improved the collection of data, but the topic reached a similar interest as now in 2004 and 2010. Source: author, with original data (in faded orange) extracted from Google Trends [https://trends.google.com/trends/explore?date=2016-01-01%202021-06-18&q=%2Fg%2F11fxvbm8wd].

3 What defines an Explainable AI?

For an AI model to be explainable, it needs to satisfy a set of constraints that are defined in this section. There are several definitions of which are these properties to fulfill, but all of them have the common goal of trying to make the results of an AI model understandable for a human.

According to the work of [9], from the U.S. National Institute of Standards and Technology (NIST), every explainable AI system should be built considering four important principles, which are the following:

- **Explanation**
- **Meaningful**
- **Explanation Accuracy**
- **Knowledge Limits**

3.1 Explanation

The AI system must give evidence, support, or reasoning for its outputs. This principle by itself does not oblige the AI to provide correct evidence, but just to be able to explain it, without any quality assessment on them.

3.2 Meaningful

The $Meaningful$ principle is fulfilled if the system's explanations are understood by the recipient. Significantly, the explanations may have to adapt to various groups of users that require different explanations. Not only that, but the same AI's explanation could result in different interpretations for the recipients, due to their prior knowledge, and even for a single individual can change over time the more used it gets to the task.

Satisfying this principle is thus not easy at all and means that developing an AI system requires to account also for human factors, apart from computational ones.

3.3 Explanation accuracy

With the *Explanation* and $Meaningful$ principles, we ensure that a system is able to give explanations that are understandable for the user, but they do not have to be true, which is achieved by imposing the *ExplanationAccuracy* principle. By fulfilling this principle, an AI system must give an explanation that actually reflects its process of generating the output. For that, we need an explanation accuracy, which differs from the typical decision accuracy that assesses the performance of a model's output compared to its target. Recently, many works are trying to develop new metrics for explanation accuracy, like [6] and [16].

Again, we may use different explanation accuracy metrics for different recipients. Note that a very descriptive explanation will usually be more accurate but at the cost of being far less meaningful.

3.4 Knowledge Limits

Until now, we are assuming that the system is operating within its knowledge limits, meaning that the domain and the task to perform are the same all the time. Nonetheless, we have to ensure that the system is able to detect that it is dealing with some input that lies outside of the scope that it was trained on, and thus, inform that the outcome is not reliable and should not be trusted.

This could be the case when an AI trained to distinguish dog breeds is presented with the image of a cat. The system should output that it could not find any dog in the image. Similarly, if the image provided is too blurry, the system could say that a dog is found but it is not sure about its breed, instead of just giving the breed with higher probability.

Note that satisfying all principles at the same time is complicated and a balance among them is necessary.

3.5 Differences between XAI and similar terms

Once we have defined what an Explainable AI system must have, we can compare it to other similar terms that are related but not the same.

3.5.1 Explainable versus Responsible AI

Explainable and Responsible AI are often confused as they both aim to improve the trust and robustness of these intelligent systems with transparency. However, they differ in how and when are they performed.

Explainable AI tries to detail the reasons for its outcome. Therefore, this explanation can only be performed once it has computed the solution.

On the other hand, Responsible AI aims at preventing the misuse or accidents derived from these systems. It takes place before the incident. One of the ways to achieve that is to have explainable models and run them in simulations such that we can be much more sure of their outcomes and correct them if needed.

3.5.2 Explainable versus Interpretable AI

Explainability and interpretability are two qualities that can refer to the same thing in some contexts and which are used interchangeably in the literature sometimes. However, there are some differences between them worth mentioning.

Explainable AI tries to answer why the model gives this output and not another, focusing on providing understandable explanations of the system process. Meanwhile, interpretability is more related to how intuitive and predictable is one model by nature. For instance, a decision tree is much more interpretable than a deep neural network.

Of course, the two terms are related in the sense that usually the more interpretable a model is, the easier it will be to make it explainable. However, we could have a rather interpretable AI model that does not fulfill the requirements of what we expect for an explainable model.

4 Families of methods

There are several ways to obtain explainable AI, which can be divided into main families of methods. Some of them assess the explainability as a quantitative measure (numeric) while others need a qualitative judgment by a human. Notably, some methods depend on the specific type of model that is being used, while others are completely agnostic about that.[3]

4.1 Transparent models

First of all, there are some machine learning models like Decision Trees, K-Nearest Neighbors, Rule-base Learners, Bayesian Models, and Logistic/Linear Regression that can be considered transparent in some cases. This means that they are understandable by themselves. A model is *transparent* if three conditions are satisfied: a human can think of it as a whole and manually compute the results, a human can understand each part without any additional tool, and the algorithm must be understandable and easy to predict.

Note that a single perceptron can also be considered *transparent* if the number of inputs is low, while a decision tree with hundreds of rules cannot.

These transparent models have huge interpretability and they are self-explanatory, so there is no need to apply any XAI technique. However, most of the ML models need to apply some post hoc explainability methods, which are described next.

4.2 Visualization

This group of techniques is aimed at providing the user with visual explanations of diverse nature that are useful to have a clearer view of what is going on at each time step. These visualization techniques are also useful to display the results of applying other techniques such as feature importance and Sensitivity Analysis (SA), described in Sections 4.4 and 4.6 respectively. Therefore, they are usually used when dealing with images and in combination with other families of methods. Some proposed visualization techniques that are model-agnostic can be found in [5].

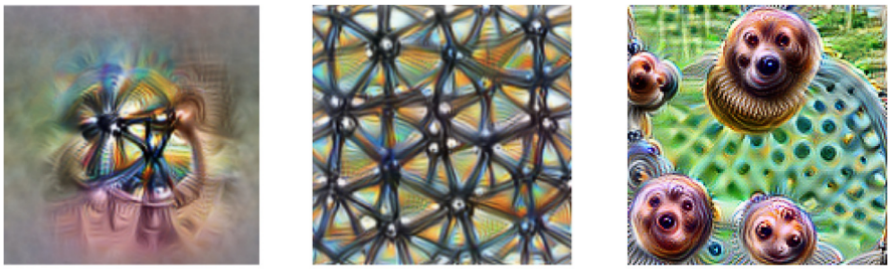
This family of techniques is also very useful for understanding where a convolutional neural network is focusing its attention, as in the case of GradCAM [14] (see Figure 2).

(a) Original image (b) Grad-CAM using ResNet (Dog)

Figure 2: Using Grad-CAM to visualize what are the input pixels that the model had to pay attention to detect the dog in the image. Image extracted from [14]

.

Another visualization technique very common is to visualize the feature relevance as it is later explained in Section 4.4. In the case of a CNN, it could be very useful to have an insight into the actual features that the network is searching for (see Figure 3).



(a) Neuron (b) Channel (c) Layer

Figure 3: Different visualizations of features that correspond to a neuron (a), a channel (b), and the complete layer (c) in a deep CNN network. Image extracted from [8]

4.3 Text explanations

Similarly to visualization techniques, the system can provide the user with text explanations describing the model justification or a generated caption of the processed image to understand better what is happening. For example, in the work of [17], a CNN is combined with a Recurrent Neural Network (RNN) to automatically give an explanation of the content of the image. This allows us to have a more detailed explanation than just using visual attention (see Figure 4), like in the case of Grad-CAM already mentioned.

(a) Correct example

(b) Incorrect example

Figure 4: A method with visual and textual explanations (white indicates attended regions for the underlined word). Image extracted from [17]

.

4.4 Feature importance

With this family of explanation techniques, the user can know the features that were most important for computing the final prediction of the model. It measures the influence of each feature on the output so that the recipient can understand better what has the model been based on.

One of the most used methods in the literature is SHAP (SHapley Additive exPlanations), proposed in [7], which uses an additive importance for each feature using their defined SHAP values, that are computed by satisfying a set of properties. Figure 5 shows an example of SHAP.

One advantage of SHAP is that it can be applied to any machine learning model, including deep neural networks.

Other simpler statistical approaches to explain feature importance or relevance exist in the literature. One way is to compute the variance of a data set that each feature explains, which can be performed directly or after applying a dimensionality reduction technique such as Principal Component Analysis (PCA). These techniques are not too precise and are not currently very used in the

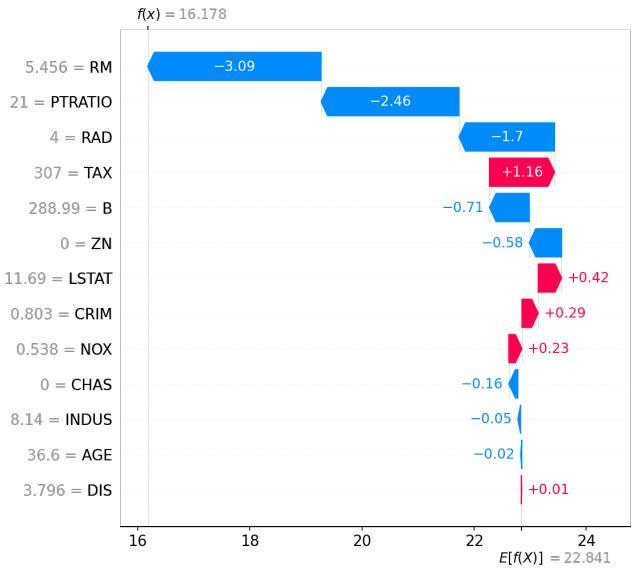


Figure 5: A waterfall plot showing the additive nature of SHAP values on a house pricing problem. Starting from our prior expectation $E[f(x)]$, we add the SHAP value of each feature one at a time until we reach the final prediction $f(x)$. Image extracted from SHAP Python package documentation [https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html]

literature, but can serve as an approximation.

4.5 Model simplification

One typical technique to improve interpretability, and thus make it easy for a model to be explainable, is to reduce the complexity of the model. This usually depends on the specific kind of model that has to be simplified.

In the case of decision trees, we can prune branches without sufficient elements, limit the maximum depth or stop the expansion of the tree if the elements of the subset that has to be partitioned are already very similar (using entropy or Gini index).

On the other hand, neural network models are often overkill and can be much reduced once trained (by training another smaller network that has to output the same as the larger one). For simple applications, we could gain interpretability by decreasing the number of layers or neurons to be used.

Significantly, an analysis of the compromise between the simplification and accuracy of the model must be addressed.

4.6 Sensitivity analysis

This family of methods is used to compare how sensitive is the output to variations of each of the inputs. This is called *sensitivityanalysis* and is one of the most common techniques for explainable AI since it can be applied pretty much to all types of models.

For it, we usually want to find the minimum change that is needed for an input to significantly change the output. A combined approach with this technique and exemplification (refer to Section 4.8) is used in IBM Watson OpenScale to give the user information about the minimum changes to

the input to change the outcome, as well as the maximum changes to the input to keep the same outcome (see Figure 6).

Figure 6: An example of sensitivity analysis applied for explainable AI. Image extracted from OpenScale video in IBM Watson web page [https://www.ibm.com/watson/explainable-ai]

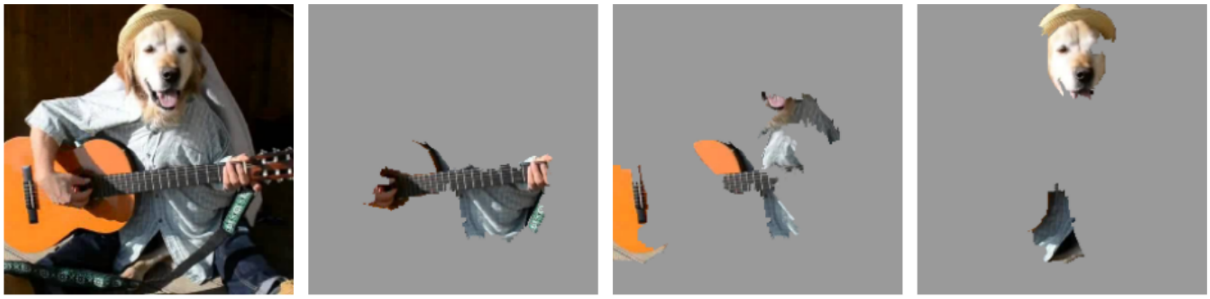
4.7 Local explanations

Note that in some cases, we are interested in global explanations (e.g. what are the most relevant features to discriminate our data?), but in other situations, we are more interested in specific explanations of a concrete output (e.g. what are the most relevant features that the model used to discriminate this sample?).

Also, sometimes *localexplanations* refers to the strategy of diving the computation of the solution into different parts and providing the user with these partial explanations.

These strategies are especially used in Support Vector Machines (SVMs) and deep neural networks [3].

An explanation technique that lies in this category and is used extensively used is LIME (Local Interpretable Model-agnostic Explanations) [11], which assumes that all models behave linearly on a local scale. An example of LIME can be seen in Figure 7



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 7: LIME explanation of an image classification prediction made by Google's Inception neural network for the top 3 classes. Image extracted from the original paper

4.8 Exemplification

The last family of explainable AI techniques is the exemplification. With that, the model provides an example of an instance that fulfills a desired property.

For instance, we can give the user the most similar sample to the one that is being handled. Similarly, during the development of the model, engineers could use the samples with the highest loss or error to try to guess what has failed and why, exploring the limitations of the system.

As we can see, there are several techniques that can be used to explain the model's outcomes. Most of them are typically used in a combination of some sort (especially with visualization).

5 Regulations, programs and international interest

As a hot topic it is, the path to Explainable AI has increasingly been regulated and lots of institutions and programs are encouraging researchers to help its development. Some of them are mentioned in this section, with the purpose of getting a broader sense of what is actually going on in the field and who are the ones that are actually pushing forward.

In 2016, the Defense Advanced Research Projects Agency (DARPA) from the United States created the Explainable AI (XAI) program, with the goal of producing more explainable machine learning models while maintaining a high level of prediction accuracy. With it, they hope to improve the human understanding, trust, and management of these systems [15].

The program gained investment during these years but it is now starting to lose it, by nearly halving it from 17.200 million dollars in 2021 to 9.324 in 2022. The reason for this decrease is due to "the shift from development and integration of explainable machine learning techniques and systems to testing, performance assessment, and transition", meaning that many techniques have been developed and now they just need to be adopted and evaluated, which does not need as many resources, they think [2].

Contrary, the overall budget for "Artificial Intelligence and Human-Machine Symbiosis" project will increase from 178.162 million USD in 2021 to 193.274 in 2022, meaning that the XAI program investment within the whole project is reduced from a 9.6% to 4.8%.

Among the three superpower governments (the United States, China, and European Union), only the latter has developed stricter measures to ensure that Explainable AI becomes a reality. The European Commission reached an agreement to regulate AI systems in an effort to achieve more trustworthy AI systems [1]. This regulation also states the bans to attend for the misuse of these technologies. The document is about AI in general but mentions the need for explainable and interpretable models.

On the other hand, the U.S. and China are not so cautious about this responsibility and trust of AI because they do not want to put obstacles in the race to AI supremacy. Nonetheless, China is also gaining more and more weight in many fields of AI and although it is not as regulated as in Europe, both the U.S. and China are still the most prolific countries in XAI research.

Apart from the big three, XAI is usually a hot topic in the most developed countries, especially in Asia. Figure 8 shows a map with the proportional interest in XAI within the country. It is an approximation based on Google web searches, and the fact that it is something relative from within the country do not reflect the real absolute interest. The country with the most interest is South Korea, while the others' interest is proportional to that (e.g. China has 91% of South Korea's interest).

Following many Asian countries and Israel, we find Norway, which stands out from other European countries like Austria, Netherlands, Greece, Spain, and United Kingdom. Some other countries not appearing in the image are Germany (16th with 14), the United States (20th with 12%), India (21th with 12), and France (28th with 7).

6 Conclusions

This document has presented what is Explainable AI, why we need it, and how can be obtained using different techniques for different kinds of models. Furthermore, several advantages of XAI systems are discussed, mainly the path to more trustworthy, robust, and accurate models. The beneficiaries of XAI include normal citizens, businesses, and public services, with a safer and clearer explanation

Figure 8: The interest by region in the topic "Explainable artificial intelligence" in Google web searches from January 2016 to June 2021. Source: adapted from Google Trends [https://trends.google.com/trends/explore?date=2016-01-01%202021-06-18&q=%2Fg%2F11fxvbm8wd].

.

of these automated systems.

Explainable AI is one of the needed solutions to not lose control over our own creations, since we cannot follow their decisions blindly. It is true that AI systems can outperform humans in many tasks, but it is not an excuse to not supervise their decisions or actions.

The impact of XAI is assured in many fields like healthcare, financial services, and criminal justice, among others. It is especially desirable in those applications where the decision could be high-risk, biased, or fooled easily.

Despite all the development that has been done in recent years, XAI is still a pretty new field regarding the new state-of-the-art deep learning approaches that are being proposed. Therefore, many challenges are remaining to be solved and ensure that even the most complex models are explainable. It is easy to imagine what would be the benefits of being able to interpret these models: an incredible source of knowledge never seen before. That would help us improve our AI models and even understand how some functions in natural life work (e.g. learning, visual cognition).

References

[1] On artificial intelligence: A european approach to excellence and trust. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, February 2020.

[2] Department of defense fiscal year (fy) 2022 budget estimates. https://www.darpa.mil/attachments/DARPA_PB_2022_19MAY2021_FINAL.pdf, May 2021.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[5] P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348, 2011.

[6] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017.

[7] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[8] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[9] P. Phillips, A. Hahn, P. Fontana, D. Broniatowski, and M. Przybocki. Four principles of explainable artificial intelligence (draft), 2020-08-18 2020.

[10] E. Puiutta and E. M. Veith. Explainable reinforcement learning: A survey, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[12] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[15] M. Turek. Explainable artificial intelligence (xai). https://www.darpa.mil/program/explainable-artificial-intelligence, 2016.

[16] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.