The Methods and Approaches of Explainable Artificial Intelligence

Mateusz Szczepański[1,2], Michał Choraś[1,2], Marek Pawlicki[1,2], and Aleksandra Pawlicka[1]

[1] ITTI Sp. z o.o. Poznań, Poland
[2] UTP University of Science and Technology, Bydgoszcz, Poland

Abstract. Artificial Intelligence has found innumerable applications, becoming ubiquitous in the contemporary society. From making unnoticeable, minor choices to determining people's fates (the case of predictive policing). This fact raises serious concerns about the lack of explainability of those systems. Finding ways to enable humans to comprehend the results provided by AI is a blooming area of research right now. This paper explores the current findings in the field of Explainable Artificial Intelligence (xAI), along with xAI methods and solutions that realise them. The paper provides an umbrella perspective on available xAI options, sorting them into a range of levels of abstraction, starting from community-developed code snippets implementing facets of xAI research all the way up to comprehensive solutions utilising state-of-the-art achievements in the domain.

Keywords: xAI · AI · Intelligent Systems · Explainability

1 Introduction

Since **Artificial Intelligence (AI) models** have become sophisticated enough to outclass many competing approaches in their respective fields, their popularity has been on the rise [1]. With initiatives such as **autonomous vehicles**, **various recommendation systems** (e.g., used by Netflix or Google Sybil), **personal assistants** and many more, intelligent systems are being instilled in everyone's lives.

This increasing ubiquity, along with the black-box nature of the best performing solutions, has led to some serious concerns [1][2][3], such as the questions of finding whether the model is unbiased [3], guaranteeing the security of the AI models [4], ensuring the model's decisions are right [5], or deciding whether to trust a system, the decisions of which cannot be understood [1].

The need to answer those questions has initiated the concept of **Explainable Artificial Intelligence (xAI)** [1]. Its main concern is to deliver the tools and methods that allow human operators to understand the driving forces behind the decisions made by AI [6]. The field also relies on the achievements of other disciplines, such as psychology or sociology [2].

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_1

Following the expansion of deep learning solutions, the search for rational explanations to the decisions taken by Artificial Intelligence has gained wider recognition [6]. This very year, a number of papers in the field have been published. Some of them present a general overview of the concept [7][8], while others focus on specific, particular features of the Explainable AI [9][10]. Finally, scientific papers which recommend using xAI in a particular field, or prove how beneficial this kind of application would be, have been published, e.g., [11][12][13], etc.

At present, the discipline is expanding in a dynamic way, enjoying its renaissance [3] and attracting the attention of the biggest corporations, such as Google [14] and IBM [15].

In other words, the accuracy obtained by AI is not the only factor that must be considered at this moment. The ability to understand the decision processes driving AI seems to be of crucial importance, too [5]. This subject has recently started to attract a wider audience [2]. Therefore, the following paper aims to become a starting point for exploring Explainable Artificial Intelligence, the main approaches and available solutions . It is structured as follows: firstly, the notion of Explainable Artificial Intelligence is introduced, with the criteria for explanations and some practical issues. Then, an overview of xAI taxonomies solutions is performed, and lastly, an umbrella perspective of the solutions that utilise xAI is given. The above approach is summarised in the conclusion section that follows.

2 Explainable Artificial Intelligence

The following subsection goes into the details of explainable artificial intelligence, and its advantages over the classical, black box approach to AI are illustrated.

2.1 The issue about the black-box Artificial Intelligence

In psychology, there is the term of the ”*Clever Hans effect*” [16]. The name comes from a horse which was famous for its ability to answer questions and solve arithmetic equations, communicating the results by tapping its hoof. However, it later turned out that instead of being a genius, the animal could simply read the cues from the body language of the person asking questions, and stopped tapping accordingly [1]. Today, the “Clever Hans effect” refers to a situation when, in the course of a flawed experiment, the questioner cues the desired behaviour in an unintentional way.

As scientists have learned, this effect is not limited to animals and humans, but also applies to artificial intelligence models as well. There have been observed the cases of models that were successful in performing their tasks only when very specific conditions were met (e.g., a model recognised boats provided that there was water in the picture, too) [1]. This issue may carry adverse implications.

One of the main concerns of today is related to the application of AI in predictive policing. For example, it has been brought to the public’s attention that some discriminatory practices generated ”dirty data”. The data, having been directly ingested by the predictive policing system, posed the risk of reinforcing and amplifying deeply ingrained biases [17]. This in turn might easily have led to disrespecting individual rights, human dignity and undermining justice [18]. In fact, it has indeed been observed that the intelligent criminal justice system had been deciding whether a person deserved parole or not based on their ethnicity [19]. This particular incident has since become a valid argument illustrating the need for artificial intelligence to be transparent, especially in high stake decision processes. An unexplainable system is unverifiable, and therefore untrustworthy. Probably no end user would wish to trust such a system with their lives. Actually, the matter caused so much controversy that a few jurisdictions in the US have ceased their use of predictive policing, whilst in Europe it is being argued that it would be better to pause the use of it until the systems become explainable and transparent enough [17].

2.2 Exploring Explainable artificial intelligence

As stated before, the AI-based systems need to be transparent. So much so, in some cases the transparency has been required by law [3]. Therefore, new solutions needed to be found. Thus, the essence of Explainable Artificial intelligence has become that, **given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand** [20]. In order to start discussing explainability, one should then first define the term. In the literature, there exist several terms which, in the context of AI are often used interchangeably to describe a very similar concept, i.e. ”explainability”, ”interpretability”, ”understandability”, ”comprehensibility”, ”intelligibility” and ”transparency”. However, there are slight differences between them, or rather, the terms have somewhat different undertones, and there is still an ongoing discussion concerning what they actually mean and what they differ in [1][2][3][20].

In order to clarify this issue, Table 1 presents the meaning of the synonyms in detail. For the sake of this paper, the term ”explainability” was selected, due to its broadest scope, active nature and its already established position in the subject literature.

2.3 The criteria for explanations

Generally, all of those considerations lead to the objective of determining **what constitutes a good explanation**. To begin with, as Carvalho et al. highlight, an important distinction must be made between the aim of achieving a *correct* explanation and the *best* explanation. Generally speaking, there are **non-pragmatic and pragmatic theories of explanation**. The former group is concentrated on achieving **correct** explanations, while the latter searches for **good** explanations [3].

The non-pragmatic theories usually assume that there is only one, true reason behind the actions of an intelligent system. Their aim is to unveil this reason,

term	definition
explainability	Refers to the extent to which human users are able to comprehend and literally explain the mechanisms that drive the learning of an AI/ML system [21]. It is an active feature of a model; the term refers to the actions taken by the model to clarify its inner working [22]
interpretability	Related to the aspects concerning observing the outputs of an AI system. The more predictable the changes of the system outputs when having switched algorithmic parameters, the higher the system's interpretability. Otherwise stated, it concerns the extent to which humans are able to forecast the results produced by an AI system, relying on various inputs [21]. It is a passive feature of the model [22]
understandability	Used to describe the situation where the user is able to comprehend and generate explanations of how the model works (its way of functioning), without being offered any description of the processes within the learning model [22][23].
intelligibility	In the context of AI, it is understood in a very similar way to understandability [22][23].
comprehensibility	Is used to describe the capability of the learning model to outline the knowledge it has learnt in a manner that the user can understand [22].
transparency	A transparent model is one which does not need any other interface or process to be understood, i.e. it is understandable by itself [22].

Table 1. The terms used when discussing explainability in the context of AI

but whether or not it is understandable for an audience, is beyond their concern. On the other hand, the pragmatic theories include the listener as an important part of the whole process. Explanation must be formulated in the manner that the audience can understand and use.

The pragmatic theory adds a powerful tool to the theoretical arsenal of xAI researchers and designers: the **Rashomon effect** [3]. It states that an event can have multiple explanations; i.e., more than one explanation can actually be found, and a person can select the one that fits their goals best, while still keeping some level of ”*truthfulness*”. However, though certainly useful, it still leaves the matter of selecting the”*best*” explanation from all of the ”*good*” ones.

There have been a number of attempts to solve this issue [1]. General guidelines, as well as more objective measures of quality have been suggested. For example, Hansen and Rieger present the ”*xAI Desiderata*”, proposed by Swartout and Moore in 1993:

1. **Fidelity**: the explanation must be a reasonable representation of what the system actually does.
2. **Understandability**: Involves multiple usability factors including terminology, user competencies, levels of abstraction and interactivity.
3. **Sufficiency**: Should be able to explain function and terminology, and be detailed enough to justify decision.
4. **Low Construction Overhead**: The explanation should not dominate the cost of designing AI.
5. **Efficiency**: The explanation system should not slow down the AI significantly[1].

Though developed in the context of expert systems, it still remains true for modern AI systems. It also presents a challenge for the community, because designing a solution that adheres to all the principles is not an easy task. As regards the **Quantitative Interpretability Indicators** [24][25], that is the indicators that can be measured and compared, there have been the attempts to formulate those, preferably in a universal manner. The **Axiomatic Explanation Consistency Framework**[25][26] is one of such endeavours. It measures to what degree an explanation method achieves the objective of attaining explanation consistency, and is based upon three axioms [3]:

- **Identity** - Identical objects must have identical explanations.
- **Separability** - Nonidentical objects cannot have identical explanations.
- **Stability** - Similar objects must have similar explanations.

2.4 A range of practical issues

Besides the above-mentioned theoretical aspects, there exist a number of other practical issues. At present, most top performing models are Artificial Neural Networks (ANN). These work by utilising layers of connected computation units called neurons [27]. Though each one on its own is only able to solve simple mathematical problems, together they form complex equations capable of diagnosing cancer, for instance [28]. This ability to generate more abstract concepts based on the simpler ones [29] is what gives Neural Networks their power, but is also the main reason for why achieving their explainability is a non-trivial task. There can be thousands or millions of neurons that interact with one another. Somehow, they are able to form some sort of representations that allow performing advanced tasks. How can those concepts be grasped, though? And even if one is able to frame the concept, the question remains of how to present it to people in an understandable way. Finally, there are also the issues of accuracy loss and a drastic increase of additional overhead.

3 An overview of xAI taxonomies

In the recent years, many approaches to explainability have been developed. Many attempts at taxonomising the domain have also been undertaken. One of those attempts can be found in [30]. A comprehensive and in-depth survey on xAI can be found in [31], where authors place considerable effort to handle the formalisms and multidisciplinarity of the field. A brief attempt at a user-centered taxonomy was placed in [32]. A preliminary taxonomy of human subject evaluation can be found in [33]. There is also a comprehensive taxonomy of xAI presented in [20], which includes the methods for both shallow and **Deep Learning (DL)**.

To begin with, the main division present within xAI should be pointed out, i.e., the distinction between the models that inherently have some level of explainability and the ones that need to utilise external means to achieve it. Arrieta et al. present further decomposition of the first category based on the domain, within which the model is transparent [20]. They highlight three main classes:

1. **Simulatable models** - the models that can be fully comprehended and simulated by humans,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_1

2. **Decomposable models** - the models that every part of which, i.e., input, parameter and calculation, can be explained,
3. **Algorithmically transparent models** - the process that generates the output can be understood by a man [20].

Generally, linear models, decision trees rule-base systems etc. are inherently transparent, with the degree varying across the mentioned domains. Nevertheless, with the increasing complexity, these explainable properties can be lost. For example, in case of decision trees, when they get too deep and wide, it becomes quite difficult to follow the paths that a system uses to generate predictions [34].

Unfortunately, most models do not possess this natural transparency; therefore, external methods are needed. Those techniques fall into the wide category of the **post-hoc explanations**. They **”aim at communicating understandable information about how an already developed model produces its predictions for any given input”** [20]. In other words, they make opaque system explainable to some degree.

The post-hoc methods are further split into the **model agnostic** and **model specific** ones. The former means that a method can be used by different Machine Learning models, while the latter marks those designed to explain specific algorithms. Of course, those can be divided even further. The authors of [20] propose to organise the agnostic methods as follows:

- **Feature relevance explanation** - the techniques based on measuring the importance that each feature has for the model’s prediction,
- **Explanation by simplification** - the methods where a new, simpler model is built. It resembles the original and keeps a similar performance score, but the level of its complexity has been lowered,
- **Visual Explanation** - as the name suggests, the algorithms belonging to this category employ some form of graphical representation to explain an opaque model.

A good example of an agnostic method is the **Local Interpretable Model-Agnostic Explanation (LIME)** [35], which trains an interpretable linear model around the prediction. It falls into the category of ”*explanation by simplification*” and has achieved a significant popularity [1]. Another popular agnostic method is **Shapley Additive exPlanations (SHAP)**[36]. It is a game-theory based framework that calculates an additive feature importance score for each prediction using the Shapley values [20].

As already mentioned, the model specific approaches are designed for particular algorithms. Although they lose the flexibility offered by the agnostic approaches, they may allow for a higher level of fidelity and accuracy. All in all, they were made to leverage the traits of the model they explain. Though the tools are being searched for which can explain shallow models, such as **Support Vector Machines (SVM)**, the main focus is on something else. Since the top performing artificial intelligence systems are usually based on deep learning, it should be no surprise that the methods designed to explain them attract the most attention [20]. There is a variety of approaches dedicated to them. It

DOI: 10.1007/978-3-030-77970-2_1

should be mentioned though that many agnostic methods prove useful for explaining various aspects of deep networks, e.g., the SHAP [36] or LIME [20]. Nonetheless, there are the methods that make sense only with ANN. **Layer-wise relevance propagation (LRP)**[37] is an example of such a method [1]. Founded theoretically on Deep Taylor decomposition, it propagates the output backwards through the network in order to calculate the impact of the input. Like in the case of image recognition, it is expected that the pixels representing the object one wants to detect have a higher score than the others. Of course, it is not the only one. In addition, there are the attribution methods, such as Grad-CAM, hybrid approaches, the systems which combine other deep learning algorithms to automatically generate textual explanation, and many other ways to achieve explainability of DL systems [20]. The final section of this paper will present several of them.

4 An overview of xAI Solutions

4.1 xAI Methods

Developers and scientists have been looking for practical solutions that will fulfil the pressing need for xAI in modern intelligent systems [1]. This search has ultimately led to the creation of many new algorithms, together with the ways to use them in practice.

To begin with, there are standalone methods developed that are available to the community. Those usually take the form of a source code which the developer can download from the portals like GitHub. In some cases, standard copy-paste procedures are enough to use them as part of the program. This is a rather "*low-level*" approach. When there is the need for more of them, it can quickly become cumbersome and unpractical; even more so if each one of them has its own set of dependencies.

4.2 xAI Libraries and Frameworks

One level of abstraction above the code fragments there are modules, libraries and frameworks. Those provide the practitioners with whole collections of methods in a single package. iNNvestigate [38] is a good example. This library can be simply imported using Python's package manager pip. It allows a developer to quickly use algorithms such as PatternNet, PatternAttribution [39], and different variants of LRP [40]. Another representative for this category is Skater [41]. It provides completely different methods from iNNvestigate, like bfPartial Dependence or LIME.

The last example for this category is the **AI Explainability 360 Open Source Toolkit** from IBM. It presents itself as one of the best frameworks currently available for the practitioners. It offers a diverse selection of algorithms like **ProtoDash** [42] or **Contrastive Explanation Method (CEM)**[43], even improving some of them [15]. Additionally, in contrast to the libraries mentioned earlier, it also provides some metrics to evaluate the quality of the explanation, though it is still quite limited. All of this is backed up by an extensive amount of materials, tutorials and guidelines, which makes it easier to start working with xAI.

4.3 xAI as Part of the System

A popular alternative for frameworks is designing and implementing solutions integrated into a specific system. The main benefit of this approach comes with full customisation, allowing to cater for the specific needs of stakeholders and their product. Explainability is therefore a natural part of the whole and should seamlessly integrate with the rest of the solution. On the other hand, the main disadvantage is the need for additional resources necessary to develop an xAI module from scratch. Additionally, this solution requires the personnel to have expert knowledge about the subject. Therefore, it is suggested to follow this path only if there is a viable reason to do so.

An example from the financial technology market is Flowcast [44]. The solution offers machine learning products for money lending companies. Smartcredit is one of them and is supposed to help in making decisions about financing thin-file **small and medium-size enterprises (SMEs)**, i.e., companies with small amount of traditional financial data used by banks in classic loan application process. This often leads to the rejection of such applicants, although some of them are potential good clients. The creators of the solution claim that this market offers 540 billion dollars' worth of financeable opportunities. Therefore, their system was designed to collect information from non-traditional sources like transaction data, to help the lender get a better picture of an SME company and assess the risks more accurately. Their platform supports a selection of ML algorithms, one of them being a variant of the boosted trees algorithm [45]. As explainability is crucial in the finance sector and the mentioned algorithm is naturally opaque, they had to find a way to clearly explain system assessments. Thus, they use SHAP along with **Natural Language Processing (NLP)** to generate plain-text sentences explaining the output in layman's terms. This is supposed to provide the description of why the system made such a decision, what must be done to change it and the level of confidence in it. They highlight it that the risk professionals employing their platform can access up to top ten reasons why each decision was made. The quality of those explanations is tested by focus groups comprised of risk management professionals and consumers.

The concluding examples of system with an integrated xAI module come from the area of cybersecurity. There is work in the domain of xAI geared towards explaining the decisions of Artificial Neural Networks used as an intrusion detection system. The solution leverages aggregations of decision trees to find the closest explanations for a classified sample [46] To protect network environments from unwanted, malevolent activity, **intrusion detection systems (IDS)** are deployed. As mentioned earlier, the systems that employ some form of ML have become very popular. The ones with best performance usually utilise some form of deep learning. As it was explained in [3], this opaqueness raises concerns and

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_1

fosters lack of trust. This is a serious issue in the field of cybersecurity, where a wrong decision can lead to dramatic consequences. An expert needs clear understanding, in order to be able to make the right decisions. The authors of [6] present a way to help with that. On a sample dataset, they have trained two deep neural networks to act as IDS. Then, they attached an explainability module that uses the earlier-mentioned SHAP algorithm. The explanation is provided using simple charts that clearly show the features and their contributions. Additionally, the paper introduces a new way to show global relations between feature values and classes. It still needs extensive testing to prove both feasibility and resistance to sophisticated types of attacks. As a final note, it should be clarified that the authors of [6] present their solution as a framework. In this paper, the framework is treated as a collection of ready-made algorithms and tools that support some way of developing a piece of software. Therefore, because this solution would still have to be implemented and integrated into an IDS by a developer, it was placed in this subsection.

As all of the examples above illustrate, *”xAI as part of the system”* is, even with its shortcomings, a valid and fairly popular approach. However, it is not a proper solution if one does not have the knowledge and resources necessary to use xAI this way. Similarly, this is not the best solution for those who only want to validate a model or gain some additional insights into the data without a ’deep dive’ into the domain. The last subsection proposes solutions to this issue.

4.4 xAI as a Service

In this section, a promising way of delivering xAI to the companies, developers and scientists is discussed. It is called **”xAI as a Service” (xAI-S)**. As mentioned earlier, implementing the explainable part manually has its unique benefits. However, in most cases it would need excessive resources and would not prove to be as worthy in the long run. Following, there are several examples of *xAI lending* services.

One company offering such service is called **DarwinAI**. On their website [47], they present **The Gensynth Platform**. It is designed to help developers build deep learning models faster, by automatic generation of high performance neural networks that can be deployed in many environments. The fact that it also offers explainability is even more important from the point of view of this work. Their materials show that this is achieved by **Generative Synthesis**. The crux of it is to use another AI model, which will learn how the observed ANN works and generate a compact version of it. Thanks to this, a mathematical model explaining the decision process can be constructed. So far, it has been applied by companies such as Intel, Nvidia and Audi.

Fiddler Labs also offer their own system that helps to achieve explainability [48]. However, while the DarwinAI tool seems to focus more on supporting quick development of deep learning models, Fiddler is all about xAI. While it offers a way to understand AI predictions using methods such as SHAP or **Integrated Gradients**[49], it does not limit itself to them. The official materials highlight other capabilities of the platform, like continuous monitoring of the deployed

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_1

models. It can be utilised to detect abnormalities in deployed models or catch data anomalies by rising settable alerts. The system also investigates feature relationships, for example by comparing distributions across dataset splits or explaining performance within a specific subset. Last but not least, it allows to test *"What-if scenarios"* i.e. check how different values of input features impact model's decisions [48]. All of that is complemented by the inclusion of human feedback in the workflow, and modern user interface.

The final example illustrates that even the biggest corporations are developing an interest in xAI and the possibilities it offers. *Google Explainable AI* is a part of the Google Cloud platform and has been released in beta version. It is a collection of ready-made tools and frameworks, rather than a streamlined solution, providing a supplement for other products offered by the Google AI Platform. Nevertheless, some of the solutions, like the **What-If Tool**, can be used within a range of environments. Owing to its diverse nature, it is hard to unambiguously classify this whole collection into one category. Nonetheless, these tools are developed to support an existing development service, so the platform roughly falls into the same category as Fiddler and Darwin AI. The mentioned frameworks and tools offer a range of advantages. The mentioned What-If Tool, for example, allows checking feature attribution, test different scenarios to see their impact on the model, examine it for fairness, compare it with others and more; all of that delivered in the form of an interactive dashboard. The official website presents the full list of features and tools available, along with in-depth descriptions, guidelines and tutorials [14]. The platform integrates xAI implementations of Axiomatic Attribution for Deep Networks [49], Sampled Shapley [50], eXplanation with Ranked Area Integrals (XRAI) [51] and others.

There are of course many more startups, products and frameworks that either offer or utilise explainability, which are not included in this this section; Rulex [52], Kyndi [53], H20.ai [54], to name just a few [55].

4.5 Current Initiatives and Research Projects

Apart from business and development solutions mentioned in previous sections, there are also several research initiatives and projects looking into the future of AI and xAI.

Obviously, most research projects worked on explainability for image recognition and image retrieval tasks. However, there are projects that touch upon many other domains, like physics etc.

Explainability is one of the challenges recognized by the SPARTA, a Horizon 2020 cybersecurity pilot project, funded by the European Commission. In particular, SAFAIR Programme (Secure and Fair AI Systems for Citizens) of the H2020 SPARTA project focuses on security, explainability, and fairness of AI/ML systems, especially in the cybersecurity domain [56]. Explainability is also one of the factors closely interlinked with ELSA (ethical, legal, societal) activities of SPARTA. Both the SPARTA project and the SAFAIR Programme have started in 2019, and the results are expected by 2022.

DOI: 10.1007/978-3-030-77970-2_1

Explainability in cybersecurity domain is very challenging (not as visually comprehensible as heatmaps of images), but such aspects are also in the agenda of SIMARGL (Secure Intelligent Methods for Advanced Recognition of Malware and Stegomalware) project working on malware and stegomalware detection mechanisms.

Another H2020 project dealing with the explainability of AI solutions is the Transparent, Reliable and Unbiased Smart Tool for AI (TRUST). It aims at creating an AI platform which is going to be trustworthy and collaborative, and employ explainable by design models and learning models. All the while, the learning process that is going to be adopted is said to be ”human-centric” and integrate cognition [57].

In her plenary talk, [58] explored the research field of xAI, used to ”overcome the shortcomings of pure statistical learning” and provide the results in the form that could be comprehended by human users [58].

5 Conclusions

This paper discusses the concept of xAI and describes some of its noteworthy solutions.

Explainability is worth being brought to AI models for a variety of reasons:

- **Explainability helps to root out the ”*Clever Hans*” models** [1]. An opaque model is by its nature difficult to debug or verify. In that case, only the results are visible, not the process. It forces a developer to follow tedious and inefficient approaches in order to find possible inconsistencies. This slows down the whole development and makes it unstable, which in turns increases the risk of obtaining faulty models. However, if the decision process is clear to the designer, many potential problems immediately become apparent.
- **Explainability is a cornerstone of reliability**. This statement results directly from the previous one. Deployed models face challenges such as **Concept Drift** (changes in the hidden context that can induce more or less radical changes in the concept of interest [59]) and **Data Decay**[60]. These are caused by the change of data relevance and its dynamics over time. To alleviate those, both the model and data have to be regularly verified to stay relevant, and, consequently, reliable.
- **Explainability brings trust in the system's decisions** [1]. People will not use the tools they do not trust. It is especially true when the stakes are high. A physician deciding about the treatment needs to know the reasons for the system reaching such a diagnosis in order to verify it and decide whether they should agree with it or not. Either way, a clear picture is necessary to make a decision based on AI system's output.
- **Explainability reduces bias and supports fairness** [3]. By understanding the principles behind system's decision, it is possible to identify unwanted biases that are present in a dataset. This helps to build models that support our modern ethics, instead of deepening unfair treatment based on race, gender or orientation.

DOI: 10.1007/978-3-030-77970-2_1

- **Explainability allows to gain additional insights into the domain** [1][31]. An explainable system can detect and unveil unknown relations present in the data. This may lead to new discoveries and studies, making the transparent AI models valuable for the scientific community.

Raising awareness about AI explainability and implementing it across various sectors is still an ongoing process. Even though the questions of explaining AI models to people without losing their accuracy are not easy, the scientific community keeps searching for answers. Since xAI enjoys its renaissance, many new approaches were developed in the recent years [20]. Though a perfect one does not exist, a lot of them show promise and have already proved to be useful.

We hope, this work may serve as a reference point to understand the various tools and xAI solutions/problems better.

Acknowledgment

This work is funded under the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

References

1. Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR., "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning," *Lecture Notes in Computer Science*, vol. 11700, 2019.
2. Miller T., "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, 2019.
3. Carvalho D. V., Pereira E. M., Cardoso J. S., "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, 2019.
4. Pawlicki M., Choraś M., Kozik R., "Defending network intrusion detection systems against adversarial evasion attacks," *FGCS*, vol. 110, 2020.
5. Choraś M., Pawlicki M., Puchalski D., Kozik R., "Machine Learning - the results are not the only thing that matters! What about security, explainability and fairness?," *International Conference on Computer Recognition Systems, LNCS*, vol. 12140, 2020.
6. Wang M., Zheng K., Yang Y., Wang X., "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, 2020.
7. Vilone G., Longo L., "Explainable Artificial Intelligence: a Systematic Review," 2020.
8. Xie, N., Ras, G., van Gerven, M., Doran, D. , "Explainable Deep Learning: A Field Guide for the Uninitiated," 2020.
9. Stoyanovich, J. and Van Bavel, Jay J., West, Tessa V., "The imperative of interpretable machines," *Nature Machine Intelligence*, 2020.
10. Roscher R., Bohn B., Duarte M.F., Garcke J. , "Explainable Machine Learning for Scientific Insights and Discoveries," *CoRR*, 2019.
11. Tjoa E., Guan E., " A Survey on Explainable Artificial Intelligence: Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*.

12. Ghosh A., Kandasamy D., “Interpretable Artificial Intelligence: Why and When,” *American Journal of Roentgenology.*
13. Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F-M., von Tengg-Kobligk, H., Summers, R. M., Wiest, R., “On the Interpretability of Artificial Intelligence in Radiology,” *Radiology: Artificial Intelligence.*
14. https://cloud.google.com/explainable-ai.
15. Arya V., Bellamy R.K.E., Chen P.Y., Dhurandhar A., Hind M., Hoffman S.C., Houde S., Liao Q.V., Luss R., Mojsilovíc A., Mourad S., Pedemonte P., Raghavendra R., Richards J., Sattigeri P., Shanmugam K., Singh M., Varshney K.R., Wei D., Zhang Y., “One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques,” 2019.
16. Samhita L., Gross H., “The “Clever Hans Phenomenon” revisited,” *Communicative integrative biology*, 2013.
17. Greene T., “AI Now: Predictive policing systems are flawed because they replicate and amplify racism,” *TNW*, 2020.
18. Asaro, P.M., “AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care,” *IEEE TSM*, 2019.
19. Wexler R., “When a computer program keeps you in jail: How computers are harming criminal justice,” *New York Times*, 13.06.2017.
20. Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, 2020.
21. Choraś M., Pawlicki M., Puchalski D., Kozik R., “Machine Learning - The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness?,” *ICCS*, vol. 4, 2020.
22. Gandhi, M., “What exactly is meant by explainability and interpretability of AI?,” *Analytics Vidhya*, 2020.
23. Taylor, M. E., “Intelligibility is a key component to trust in machine learning,” *Borealis AI*, 2019.
24. Doshi-Velez, F., Been K., “Considerations for evaluation and generalization in interpretable machine learning.,” *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 2018.
25. Doshi-Velez, F., Been K., “Towards a rigorous science of interpretable machine learning.,” *arXiv preprint:1702.08608*, 2017.
26. Honegger, M., “Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions.,” *arXiv preprint:1808.05054*, 2018.
27. Russel S., Norvig P. *Artificial Intelligence: A Modern Approach*, 2010.
28. Liu S., Zheng H., Feng Y., Li W., “Prostate cancer diagnosis using deep learning with 3D multiparametricMRI,” *MedicalImaging2017: Computer-AidedDiagnosis*, 2017.
29. Goodfellow I., Bengio Y., Courville A. *Deep Learning*, 2016.
30. Lipton, Z. C., “The mythos of model interpretability,” *Int. Conf. ”In Machine Learning: Workshop on Human Interpretability in Machine Learning”*, 2016.
31. Adadi A., Berrada M., “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *ICCS*, vol. 6, 2018.
32. Weina J., Carpendale S., Hamarneh G., Gromala D., “Bridging AI Developers and End Users: an End-User-Centred Explainable AI Taxonomy and Visual Vocabularies,” *IEEE Vis*, 2019.

33. Chromik, M., Schuessler, M., "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI," *ExSS-ATEC@ IUI*, 2020.
34. Blanco-Justicia, A., Domingo-Ferrer, J., "Machine learning explainability through comprehensible decision trees," *Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science*, vol. 11713, 2019.
35. Ribeiro, M., Singh, S., Guestrin, C., "Why should I trust you?: explaining the predictions of any classifier," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA.*
36. Lundberg, S, M., Su-In L., "A unified approach to interpreting model predictions.," *Advances in neural information processing systems*, 2017.
37. Bach, S., Binder A., Montavon G., Klauschen F., Müller K.R., Samek W., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.," *PloS one 10*, vol. 7, 2015.
38. Alber M., Lapuschkin S., Seegerer P., Hagele M., Schutt K., Montavon G., Samek W., MullerK., Dahne S., KindermansP., "iNNvestigate neural networks!," *arXiv*, 2018.
39. P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," 2017.
40. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, *Layer-Wise Relevance Propagation: An Overview*, pp. 193–209. Cham: Springer International Publishing, 2019.
41. https://github.com/oracle/Skater, Accessed: 30.12.2020.
42. Gurumoorthy, K.S., Dhurandhar A., Cecchi G, Aggarwal C., "Efficient Data Representation by Selecting Prototypes with Importance Weights," *ICD, IEEE*, 2019.
43. Dhurandhar, A., Chen,P., R. Luss, Tu,C., Ting, P., Shanmugam, K., Das, P., "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in Neural Information Processing Systems*, 2018.
44. https://flowcast.ai, Accessed: 30.12.2020.
45. https://resources.flowcast.ai/resources/big-data-smart-credit-white-paper/, Accessed: 18.03.2021.
46. Szczepański, M., Choraś M., Pawlicki M., Kozik R., "Achieving Explainability of Intrusion Detection System by Hybrid Oracle-Explainer Approach," *IJCNN*, 2020.
47. https://darwinai.com, Accessed: 30.12.2020.
48. https://www.fiddler.ai, Accessed: 30.12.2020.
49. Sundararajan, M., Taly, A., Yan, Q., "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.
50. Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., Rogers, A., "Bounding the estimation error of sampling-based Shapley value approximation," *arXiv:1306.4265*.
51. Kapishnikov, A., Bolukbasi, T., Viégas, F., Terry, M., "Xrai: Better attributions through regions," *IEEE International Conference on Computer Vision*, 2019.
52. https://www.rulex.ai, Accessed: 30.12.2020.
53. https://kyndi.com, Accessed: 30.12.2020.
54. https://www.h2o.ai, Accessed: 30.12.2020.
55. https://www.ventureradar.com, Accessed: 30.12.2020.
56. https://www.sparta.eu/programs/safair/, Accessed: 18.03.2021.
57. https://cordis.europa.eu/project/id/952060, Accessed: 30.12.2020.
58. Zanni-Merk, C., "On the Need of an Explainable Artificial Intelligence," 2020.
59. Widmer, G., Kubat M., "Learning in the presence of concept drift and hidden contexts.," *Machine learning 23*, vol. 1, 1996.
60. https://peterasaro.org/writing/Asaro_PredictivePolicing.pdf, 30.12.2020.