Trends in Explainable Artificial Intelligence for Non-Experts

Elise Özalp, Katrin Hartwig, Christian Reuter

1. Introduction

Artificial intelligence (AI), cognitive systems and machine learning are already part of everyday life and have the power to transform economy and society. Although AI is widely used in data processing and domains such as the medical field or cybersecurity, the inherent pitfalls are not clear yet. According to a recent report, AI systems are increasingly biased in terms of gender, race or social background (Crawford et al. 2019). For instance, the *AI Now 2019 Report* highlights the revelations of recent audits regarding "disproportionate performance or biases within AI systems ranging from self-driving-car software that performed differently for darker- and lighter-skinned pedestrians, gender bias in online biographies, skewed representations in object recognition from lower-income environments, racial differences in algorithmic pricing, differential prioritization in healthcare, and performance disparities in facial recognition" (Crawford et al. 2019). This has led to demands from researchers and politicians for accountability and transparency of these systems. In May 2018, the European Union passed the General Data Protection Regulation (GDPR), which requires reasoning and justification of decisions based on fully automated algorithms. This shows the increasing need for *eXplainable Artificial Intelligence (XAI)* (European Union 2016).

The biases revealed and political topicality have accelerated the development of XAI, which makes use of visualizations and natural language processing to explain the reasoning behind algorithmic decisions (European Union 2016). Making algorithmic decisions more comprehensible for end users is an emerging trend in many different application domains such as cybersecurity – even when there is no artificial intelligence involved, as algo-

rithmic decisions are in any case often difficult for end users to understand (Hartwig/Reuter 2021). In general, research has shown that even when algorithmic predictions are more accurate than human predictions, both domain experts and laypeople distrust the algorithm (Narayanan et al. 2018). XAI aims at comprehensive understanding of the deployed algorithm such that user trust is built based on explanations and not only on performance. This will lead to better social acceptance of AI systems in different areas of life. It is important to note that the explanations provided by XAI systems do not only promote trust. They also help users to critically analyze the algorithms and to improve algorithms by finding mistakes.

Despite the usefulness of XAI, there are still several difficulties to address, especially regarding laypeople. The main difficulty lies in the complexity of the AI models. Often, complicated models (such as neural networks) are used to achieve the best performance. These models do not serve as an understandable explanation for non-specialists since they reason with high-dimensional numerical values. According to Miller (2019), numerical values probably do not matter for humans since we demand instead causal explanations. To handle this and other difficulties, XAI combines pedagogy, programming and domain knowledge.

In this paper we provide an overview of XAI by introducing fundamental terminology and the goals of XAI, as well as recent research findings. Whilst doing this, we pay special attention to strategies for non-expert stakeholders. This leads us to our first research question: "What are the trends in explainable AI strategies for non-experts?". In order to illustrate the current state of these trends, we further want to study an exemplary and very relevant application domain. According to Abdul et al. (2018), one of the first domains where researchers pursued XAI is the medical domain. This leads to our second research question: "What are the approaches of XAI in the medical domain for non-expert stakeholders?" These research questions will provide an overview of current topics in XAI and show possible research extensions for specific domains.

The chapter is organized as follows: Section 2 presents the foundations and related work on XAI; in Section 3, we explain the research method utilized; Section 4 identifies and surveys current trends, and in Section 5, we describe the current state of XAI in the medical domain; finally, in Section 6, we discuss how non-experts can be better integrated into XAI and which questions we should ask going forward.

2. Related Work

AI is the attempt to mimic human behaviour by constructing an algorithm for specific decision scenarios. Since the beginning of the 2010s, AI has become increasingly popular, with applications in almost every economic sector due to increased processing power. With its increasing popularity and use, AI has left the context of academic research and is now being used by non-experts of AI. However, many of the constructed algorithms are based on higher mathematics and are not transparent. XAI targets this issue by explaining AI decision-making processes and logic for end users (Gunning/Aha 2019).

We give an overview of the fundamental terminology in Table 1 and provide further literature. In general, *intelligible systems* do not have to be systems involving AI; they can be viewed as a superclass of it. Within *transparent AI*, *interpretable* and *explainable AI* can be viewed as subclasses. Even though there is a subtle difference, these terms are often used interchangeably in research. Barredo et al. (2020) present how the research interest has shifted from interpretable to explainable AI since 2012. Within the XAI research, the terms *local* and *global explanations* appear to classify explanations according to their interpretation scale (see Table 1). In many cases, *ad-hoc explainers* are used to describe local explanations (Guidotti et al. 2018). The terminology described here will build the foundation to analyze current trends in XAI for non-experts.

While the work referenced in Table 1 focuses on the foundations and the underlying idea of XAI, research in XAI generally aims at different user groups (Langer et al. 2021; Martin et al. 2021). The most targeted user group so far has been AI experts, who design the systems and aim to improve them. For research targeting machine learning experts, it is common to analyze existing use cases and create explanations around them, such as for predictive coding (Chhatwal et al. 2018) or the retrieval of video frames (Chittajallu et al. 2019). Even though XAI systems are presented, fundamental knowledge of AI methods and techniques implemented is necessary to understand these papers. The same is true for suggested techniques to create more interpretable deep neural networks (Holzinger et al. 2017; Liu et al. 2017).

Table 1: Relevant definitions for XAI with reference sources

Terminology	Description	References
Intelligible System	A system whose inner workings and inputs are exposed through transparency and explanations to the user.	(Clinciu/Hastie 2019; B.Y. Lim/Dey 2009; Mohseni et al. 2018)
Transparent AI	A system that discloses the algorithmic mechanism on the level of model, individual components and training.	(Chromik et al. 2019; Clinciu 2019; Lipton 2018; Mohseni 2018; Rader et al. 2018)
Interpretable AI	A system provided with explanations to retrace the model decision making process and predictions.	(Abdul 2018; Barredo Arrieta 2020; Clinciu 2019; Marino 2018; T. Miller 2019; Mohseni 2018)
Explainable AI	A system provided with explanations to give reasoning for algorithmic decisions.	(Kulesza et al. 2013; T. Miller 2019; Mohseni 2018; Rader 2018)
Global Explanation	Reasoning of how the overall model works, also called model explanation.	(Gedikli et al. 2014; P.L. Miller 1986; Mohseni 2018)
Local Explanation	Reasoning why a specific input leads to a certain output, also called instance explanations.	
Mental Model	The representation of how a user understands a system.	(Mohseni 2018)

More research that involves XAI in recent machine learning research can be found. Even though these research papers do not state that their target user group are AI experts, this can easily be concluded after reading them. In contrast to this research direction, research for non-experts is less prominent. While some literature also includes machine learning novices (Hohman et al. 2018; Spinner et al. 2020), there is little literature that focuses on non-experts with no technical background. In some studies (e.g. Cheng et al. 2019; Kulesza 2013; B. Lim 2011), different strategies are examined that target non-expert stakeholders, and we will look these papers at closely to observe XAI for non-experts. Kulesza et al. (2009, 2012, 2015) analyze how explanations can help non-experts to personalize and debug interactive machine learning. However, XAI mainly targets existing AI research and research only rarely specifically targets end users. Therefore, we identify this as a current research gap. In the following, we analyze the ex-

isting research on XAI for non-experts and explicate which trends are to be observed. In order to give a balanced overview of XAI for non-experts, we want to include to what extent these trends can also be found in a specific and relevant application domain where XAI is often designed for medical experts (e.g. Karim et al. 2019).

3. Methods

To analyze current trends and attain an overview of the current state of research, we decided to conduct a non-exhaustive semi-structured literature review over several platforms. Because of the high topicality, it is difficult to observe XAI for non-experts in real-life applications. Therefore, we analyze the existing literature as a foundation for later research. For our qualitative research study, we combine a constrained backtracking search strategy with a keyword search.

Mohseni et al. (2018), in which the authors present a multidisciplinary survey for the design and evaluation of explainable AI systems, serves as the first core paper for the constrained backtracking search. This paper is an in-depth survey that analyzes different aspects of XAI, including terminology, design goals, evaluation measures and frameworks. In referencing about 250 papers, it also provides a good literature overview as a starting point for further research. With the first version submitted in 2018 and the most recent in January 2020, the paper constitutes an up-to-date summary of XAI. From this work, we have selected referenced papers based on their relation to the keywords presented in Table 2, scanning titles and abstracts to decide if a paper was relevant for our context.

As the second core paper, we use Abdul et al. (2018), in which the authors conducted a literature analysis of 289 core papers on explainable systems to derive current trends and trajectories. Again, we select referenced papers based on their relation to the keywords in Table 2, scanning titles and abstracts. While the first paper targets the foundations of XAI, the second paper focuses on current trends which will allow us to combine both topics. Both papers provide a broad overview of the current state of research in XAI and will serve as core papers for our literature research. From both papers, we generally excluded papers with a focus on research on the mathematics and technical improvements of AI systems, as well as research on very specific applications that was not transferable to our research question.

Table 2: List of search keywords

Keyword	**Alternative search word**
Explainable Artificial Intelligence	Explainable AI, XAI, interpretable AI
Trends	
Interaction	Interactive
Visualization	Visual
Trust	
Bias	
Medical Domain	Clinical domain, medicine
Clinical decision support systems	CDSS
Human-computer interaction	HCI, Computer-human interaction, CHI
Non-experts	AI novices, laypeople

In order to avoid biased outcomes and to consider more recent work up to March 2021, we consider the database of IEEE, AAAI, Google Scholar, ADS, Journal of Artificial Intelligence Research and the Journal of Human-Computer Interaction. Table 2 shows search keywords and combined search terms. We use combinations such as "XAI non-experts" or "Trends Visualization XAI". However, the core search word remains "XAI" to maintain distinctions between generally AI-related research and AI-focused research.

It is crucial to point out that our database search does not make any claim to comprehensiveness, as in a first step titles and abstracts were scanned superficially and not all papers were read exhaustively. Instead, the review should be considered a semi-structured approach, giving valuable qualitative insights into trends related to XAI for non-experts. The database search led to a large quantity of publications from which we excluded publications based on their language (only English papers were included), title, abstract and application domain. We again excluded research that focused on mathematics and algorithm optimization of AI systems. Furthermore, we did not consider publications that explained specific AI application examples to AI experts. After reading several of these publications, we decided that the approaches were not transferable to laypersons due to their high complexity.

Due to the large set of literature from our core papers by Mohseni et al. (2018) and Abdul et al. (2018), the database search resulted in partial overlap

of publications. In total, we reference 42 different works including the two core papers (see below: 7. References). Many of these furnish reasoning and background for our arguments or provide examples. However, to directly derive answers for our first research question, we identified only 13 papers out of the 42 works that clearly propose trends that were transferable to general XAI for laypeople. These publications are classified into trends in Table 3.

4. Results

In the following, we will discuss our findings regarding general trends of explainable AI for non-experts and, subsequently, give more specific insights into XAI for non-experts in the medical domain..

4.1 Trends of Explainable AI for Non-Experts

When referring to AI non-experts or AI novices, we consider general AI end-users who have no previous experience in the design of machine learning algorithms and are not domain experts of the application field in which the AI operates. It is crucial to understand that XAI pursues different goals and different approaches to explanation depending on the target user. According to Mohseni et al. (2018), there are three target user groups: non-expert end-users, domain experts and AI experts. To give an example for the medical domain: the AI expert researches and designs the machine learning algorithms, the domain expert is expert in the application domain, such as a doctor in the medical domain, and the end-user non-expert is the patient with no prior medical or AI knowledge.

Mohseni et al. (2018) found four dominant goals of XAI for non-experts. The first goal of XAI is to help end-users understand how the AI system works. This is referred to as algorithmic transparency and aims to improve the users' mental model. It can also allow an end-user to improve or debug daily machine learning applications such as email categorizers (Kulesza 2009) without machine learning knowledge. Other scenarios where algorithmic transparency promotes end-user debugging in daily life applications can be found (B.Y. Lim, 2009), such as instant messenger auto-notification. Even though the end user can directly benefit from this transparency, companies that own these intelligent applications are reluctant to provide explanations

for fear of negative impact on their reputation or competitive advantage (Chromik 2019). Nevertheless, providing explanations improves user trust by letting the user evaluate the system's reliability and observe the system's accuracy for certain decisions. This user trust is influenced by the increasing amount of information on biased AI, making bias mitigation a design goal of XAI (Mohseni 2018). This can have impacts in economic scenarios, e.g. in the form of hotel recommendations (Eslami et al. 2017). But there are also societal scenarios where biased AI can have severe implications, as in the risk assessment of criminal defendants. One of the leading American tools for risk assessment found that black defendants were far more likely than white defendants to be incorrectly judged (Chouldechova 2017; Larson et al. 2016). The above mentioned GDPR now legally supports each person in accessing information about how their data is used. This goal of privacy awareness allows end-users to know which user data is influencing the algorithmic decision-making. Everyday life examples where this is also relevant can be found in personalized advertisement or personalized news feeds in social media (Eslami et al. 2015).

To achieve these goals, different explanation interfaces can be found in the literature. The main distinction is made between white-box and black-box models with interactive or static approaches. In contrast to white-box models, black-box models do not display the inner workings of the algorithm but focus on explaining the relationship between the input and output, e.g. through parameter weight influence. Since this is independent of how complicated the models are, black-box models are often used for (deep) neural networks. This also makes explaining the algorithmic concepts to AI novices unnecessary. By comparison, white-box models specifically display the inner workings of the algorithm with understandable features or transparent computations. Cheng et al. (2019) conducted a study to compare white- and black-box methods for non-expert stakeholders and evaluated them in terms of objective understanding and self-reported understanding. Users spent the same amount of time on both explanation interfaces but scored higher in objective understanding when using the white-box interface. This corresponds with the idea that more transparent explanations help user understanding. However, users did not describe an increased self-reported understanding with the white-box model and neither of the models increased the users' trust. The authors observed that greater complexity resulted in lower satisfaction. This might suggest that the black-box method could benefit from not conveying complexity if improvements to objective understanding can be

made. There are different approaches to improving explanations of black-box models (Narayanan 2018), but they are still feature-oriented or try to explain the mathematical components of the features. This loss of human interpretability is known as the "accuracy-interpretability trade-off", which states that "often the highest performing methods (e.g., deep learning) are the least explainable, and the most explainable (e.g., decision trees) are less accurate" (Gunning 2019). While high-dimensional weights and numerical features are the basis for the algorithm, social sciences argue that a person requires a causal explanation instead of a probabilistic explanation in order to be satisfied (T. Miller 2019). Aligning with this is the newest trend of "Open the Black-Box" or "Stop the Black-Box" (Rudin 2019; Rudin/Radin 2019), which suggests completely abandoning black-box models. Proponents argue that the black-box models support bias and even AI experts do not understand how predictions in complicated models are made. They also state that the improved white-box models require a significant effort to construct, especially with regard to computation and domain expertise. The idea appears promising but points more in the direction of a topic for classical AI research.

In Kulesza et al. (2015) an interactive white-box method is suggested which allows users to build an understanding of the system while exploring it. Their recommendation focuses on balancing completeness with incremental changes and reversible actions to not overwhelm the user. They state that the model should include the following types: *inputs* (features the system is aware of), the *model* (an overview of the system's decision-making process), *why* (the reasons underlying a specific decision), and *certainty* (the system's confidence in each decision). To allow further exploration and interactivity, *what if* types should be included. One of the key takeaways is that a user-focused approachthat includes a self-explanatory system and back corrections is highly beneficial to explain an AI application. According to this survey, the new watchword is interaction, with its close connections to reflection, implicit interaction and software learnability. Abdul et al. (2018) also implemented an interactive interface that let the user freely explore adjustable inputs. The interactive feature provided significantly better results in self-reported and objective understanding. However, in both cases the interactive aspect is model- and application-specific, which makes it difficult to derive generalizations for interaction in XAI.

The same problem arises regarding visualizations in XAI. Visualizations are very useful to display high dimensional data and data flow (Bach et al.

2015; Cheng 2019; Samek et al. 2019) and there are multiple approaches to using them for neural networks (Heleno et al. 2019; Montavon et al. 2018). In Liu et al. (2017), the authors present CNNVis, an interactive visual analytics system, to better understand convolutional neural networks for image processing. Hybrid visualizations are used to disclose interactions between neurons and explain the steps of neural networks. Whilst CNNVis targets machine learning experts, this is also an interesting approach to "open the black-box" for non-experts. Other literature suggests using visualization tools to explain the ML pipeline from the model input to output (El-assady et al. 2019; Spinner 2020). The model presented by Spinner et al. (2020), explAIner, is a framework for interactive and explainable machine learning. It can be used as a TensorBoard plugin and combines visual and natural language explanations, with enhancement on storytelling and justification. The framework received positive feedback in a case study but it is unclear how well AI novices can understand the explanations with no background in machine learning. Meanwhile, Google deployed an online service at the end of 2019 which provides a framework for AI explanations, including the integration of visualizations (Google Cloud 2021). This also underscores the increasingly important role that XAI is playing in the industry. However, the question of how to use visualization tools in XAI for AI novices remains open for research.

The trends identified in the most relevant papers referenced above are summarized in Table 3.

Table 3: Referenced literature with identified trends

	Black/ White Box	Explanation Types	Underlying Model	Interaction	Visualization
(Cheng 2019)	x				
(Guidotti 2018)	x				
(T. Miller 2019)	x	x			
(Kulesza 2013)	x	x			
(Kulesza 2015)	x	x		x	
(Rudin 2019)	x		x		
(Rudin/Radin 2019)	x		x		
(Abdul 2018)	x			x	
(Narayanan 2018)		x	x		
(Lim 2011)		x			
(Lim 2009)		x			x
(Heleno 2019)			x		x
(Spinner 2020)				x	x

In the next section, we seek to better illustrate the current state of XAI for laypeople on the example of an explicit application domain.

4.2 XAI for Non-Experts in the Medical Domain

Lim et al. (2018) state that the medical domain was the first application domain of XAI. More than other domains, algorithmic medical recommendations demand explanation and justification due to their high impact on human lives. This naturally acts as driving force of XAI in the medical domain. According to the research analysis (Abdul 2018), the medical and healthcare domain appears repeatedly in the context of XAI, especially in relation to fair, accountable and transparent algorithms and interpretable machine learning. As explained in Section 2, research in XAI with a focus on non-experts is very limited. Because of its importance and the availability of more literature than in other application domains, we have chosen the medical domain to analyze the current state of XAI for non-experts in an application domain.

In the medical domain, AI systems are used for critical decision-making tasks which magnifies the importance of the explanations and transparency for end-users and non-experts. The described systems are clinical decision support systems which store health knowledge and apply this to new patient observations. The resulting recommendations can help the clinicians to make choices.

Even though in general, experts and AI novices tend to distrust AI systems (Narayanan 2018), in this specific context, Goddard et al. (2012) discovered that clinicians tend to trust the system more than their own judgement. This over-trust is called automation bias and is dependent on factors such as task complexity, workload and time pressure. Bussone et al. (2015) investigate the relationship between trust, explanations and reliance of practitioners to CDSS in an exploratory between-group user study. The study involved some of Lim's and Dey's (2009) types such as confidence explanations and why explanations in natural language and also examined the explanations desired by the study participants. In contrast to Lim and Dey (2009), the participants requested more than an indicator of certainty and a significant number did not understand what this percentage even meant. From the perspective of a patient, it can also be rather unsettling if this percentage indicates a disease. Further, the system provided facts it used to make a diagnosis but the study participants requested more information for typical cases of this diagnosis. This would allow them to assess how much the suggestions fit. Additionally, the clinicians requested an explanation that allowed them to disprove other diagnoses, e.g. the second most likely diagnosis. This also supports Lim's and Dey's (2009) *why not* explanation type. Overall, Bussone et al. (2015) observe that clinicians demand explanations with the same reasoning that they use to make a diagnosis. However, the sample size of the study was very limited and the user groups studied were AI novices but experts in the medical domain.

In the study of Narayanan et al. (2018) about 600 participants were recruited to conduct a study on explanation types of recommendation systems. While one recommendation system recommended recipes, the other system diagnosed symptoms and recommended pharmaceuticals. Unlike in the study by Bussone et al. (2015), the recommendation system was less expert-oriented. Narayanan et al. (2018) found that the observations on explanations in the recipe and medical domains coincided, meaning that the application domain did not require different explanations. Another important finding is that if the participant is focused on understanding, the com-

plexity of an explanation does not result in a decrease of accuracy but rather in an increase in response time.

This leads us to the proposition that Lim's and Dey's types (Lim 2009), presented in Section 4 and supported by the observations of Bussone et al. (2015), can be transferred to explanations for patients. However, it is questionable whether or not these explanations can be used as black-box explanations when the decision support system is based on very complicated systems such as neural networks. For such systems, Holzinger et al. (2017) suggest linking vector representations of neural networks to lexical resources and knowledge bases using hybrid distributional models for the medical context. This enables step-by-step retracing of how the system developed a solution but it is questionable whether this is understandable for AI laypeople.

Overall, we observe the same problem as explained in Section 2: the literature addresses mostly AI experts or experts in the medical domain. From the literature analyzed, the findings derived are limited: Lim's and Dey's (2009) explanation types are transferable to different domains and especially the *why not* explanation type is in demand. This explanation will help to understand why the patient did not receive a different diagnosis. Further, the study by Narayanan et al. (2018) indicates that XAI can probably follow a universal explanation strategy for different domains. For this reason, we will summarize our findings and provide guidelines in the next section.

5. Discussion and Conclusion

Table 4 presents the findings of our semi-structured literature review on current XAI trends with a focus on the medical domain, summarises our most important findings and puts forward the following suggestions for the design of XAI for non-experts:

Table 4: Suggestions for the design of XAI for non-experts

Design Suggestion	Explanation for Suggestion
During the Implementation of the AI System for an Application:	
Consider explainability in the choice of the AI algorithm.	
When choosing a model, consider whether a simpler algorithm can achieve similar/better results.	According to Rudin (2019) and Rudin/Radin (2019), even AI experts do not understand how predictions in complicated models are made.
After deciding on a model, implement it such that the inner computations can be accessed and perhaps even visualized.	Liu et al. (2017) suggest hybrid visualizations to explain the steps of neural networks. We suggest implementing this directly with the system.
After Building the AI System	
Focus on explaining using a white-box approach.	
For natural language explanations, use Lim's & Dey's (2009) explanation types: *inputs*, *model*, *why*, *certainty*, *what if/why not.*	These explanation types have been requested by laypeople and proven to be reliable in different contexts (Bussone 2015; e.g., Lim 2009).
Allow the user to freely explore adjustable inputs and allow back corrections for interactive explanations.	Kulesza et al. (2015) found that interactive features lead to significantly better results in self-reported and objective understanding.
Use visual explanations to display the framework of the built AI system and to explain the computations of the data (dependent on the application domain).	Cheng (2019) suggests displaying high dimensional data and data flows using visualizations.

Identifying explanation strategies for non-experts to account for AI is an essential step in integrating AI systems into society. The task of explaining complicated systems to someone with little to no prior knowledge is generally a challenge. Which strategies can be used for AI systems? What is the current state of research? Are trends in explainable AI also observable in applications such as the medical domain? The knowledge gained from identifying these explanation strategies will be crucial for the acceptance of AI in society.

In this review, we observed current trends in XAI for non-experts. We perceived a demand for a shift from black to white-box models which ap-

pears difficult regarding complicated machine learning models. Voices in research are increasingly questioning the necessity of complicated models and suggest a simpler, well-planned architecture. At the same time, different underlying models are proposed such as hybrid models to create more self-explainable complex models. Regarding the explanations, the types suggested by Lim and Dey (2009) reappear within different independent literature, sometimes slightly modified. Therefore, we also recommend including the types: *inputs, model, why, certainty/confidence, what if*. We also propose using the *what if* type together with interactivity to support end user exploration. Generally, visualization is recommended and several frameworks are proposed. However, the problem of how to explain AI specifically to AI novices remains unsolved. This can be observed even more readily in the medical domain where trust and reliability of AI are of particular importance. The shift to white-box models, the explanation types, hybrid models and interaction can also be observed in XAI for the medical domain. Yet, all the research available is only targeted at data or AI experts.

Generally speaking, it is difficult to find research that targets non-experts. Going forward it will be important to center the research around the needs of non-experts. What information do non-experts demand from an AI system in the medical domain, such as a clinical decision support system? How do these differ from AI systems in lifestyle applications such as a spam filter? Can we directly integrate those explanations in domains where XAI is not yet established, such as cybersecurity? Further, there are open questions regarding the build of the framework. Can we build simpler AI systems that achieve similar or better results than complicated ones? Do non-experts really profit from hybrid models to understand more complicated AI systems? How important is the design of the interface in terms of interaction and visualization compared to the explanation of the system? A better understanding of these questions will guide the design of XAI for non-experts.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (SFB 1119, 236615297), by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and by the BMBF in the project CYWARN (13N15407).

References

Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y. Lim, an Mohan Kankanhalli (2018). "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda". In: *CHI '18, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 582–590. URL: https://doi.org/10.1145/3173574.3174156.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PLoS ONE* 10.7. URL: https://doi.org/10.1371/journal.pone.0130140.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera (2020). "Explainable Artificial Intelligence (XAI). Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." In: *Information Fusion*, 58.June, pp. 82–115. URL: https://doi.org/10.1016/j.inffus.2019.12.012.

Bussone, Adrian, Simone Stumpf, and Dympna O'Sullivan (2015). "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems." In: *IEEE, International Conference on Healthcare Informatics*. URL: https://doi.org/10.1109/ICHI.2015.26.

Cheng, Hao-Fei, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu (2019). "Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders." In: *CHI '19, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. URL: https://doi.org/10.1145/3290605.3300789.

Chhatwal, Rishi, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao (2018). "Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding." In: *IEEE, International Conference on Big Data (Big Data)*, pp. 1905–1911. URL: https://doi.org/10.1109/BigData.2018.8622073.

Chittajallu, Deepak Roy, Bo Dong, Paul Tunison, Roddy Collins, Katerina Wells, James Fleshman, Ganesh Sankaranarayanan, Steven Schwaitzberg, Lora Cavuoto, and Andinet Enquobahrie (2019). "XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally in-

vasive surgery videos." In: *IEEE, 16th International Symposium on Biomedical Imaging*, pp. 66–69. URL: https://doi.org/10.1109/ISBI.2019.8759428.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." In: *Big Data* 5.2, pp. 153–163.

Chromik, Michael, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek (2019). "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems." In: *IUI workshops* 2327.

Clinciu, Miruna-Adriana, and Helen Hastie (2019). "A survey of explainable AI terminology." In: *NL4XAI, Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pp. 8–13.

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas„ Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker (2019). *AI Now 2019 Report*. New York. URL: https://ainowinstitute.org/AI_Now_2019_Report.html.

El-Assady, Mennatallah, Wolfgang Jentner, Rebecca Kehlbeck, and Udo Schlegel (2019). "Towards XAI: Structuring the Processes of Explanations." In: *ACM Workshop on Human-Centered Machine Learning*, iss. May.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig (2015). "'I always assumed that I wasn't really that close to [her]' Reasoning about Invisible Algorithms in News Feeds". In: *CHI '15, Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 153–162. URL: https://doi.org/10.1145/2702123.2702556.

Eslami, Motahhare, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton (2017). "Be careful; things can be worse than they appear." In: *Eleventh International AAAI Conference on Web and Social Media* 11.1, pp. 62–70.

European Union (2016). *General Data Protection Regulation (GDPR)*, Retrieved August 5, 2021. URL: https://gdpr-info.eu.

Gedikli, Fatih, Dietmar Jannach, and Mouzhi Ge (2014). "How should I explain? A comparison of different explanation types for recommender systems." In: *International Journal of Human-Computer Studies* 72.4, pp. 367–382.

Goddard, Kate, Abdul Roudsari, And Jeremy C. Wyatt (2012). "Automation bias: a systematic review of frequency, effect mediators, and mitigators." In: *Journal of the American Medical Informatics Association* 19.1, pp. 121–127.

Google Cloud (2021). "Einführung in AI Explanations für AI Platform | AI Platform Prediction", Retrieved March 3, 2021. URL: https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (2018). "A survey of methods for explaining black box models." In: *ACM Computing Surveys (CSUR)* 51.5, pp. 1–42.

Gunning, David, and David Aha (2019). "DARPA's explainable artificial intelligence (XAI) program." In: *AI Magazine* 40.2, pp. 44–58.

Hartwig, Katrin, and Christian Reuter (2021). "Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations." In: *Behaviour and Information Technology* 41.7, pp. 1357–1380. URL: https://doi.org/10.1080/0144929X.2021.1876167.

Heleno, Marco, Nuno Correia, and Miguel Carvalhais (2019). "Explaining Machine Learning" In: *ARTECH 2019, Proceedings of the 9th International Conference on Digital and Interactive Arts*, October, Article No. 60, pp. 1–3. URL: https://doi.org/10.1145/3359852.3359918.

Hohman, Fred, Minsuk Kahng, Robert Pienta, and Duen Horng Chau (2018). "Visual analytics in deep learning: An interrogative survey for the next frontiers." In: *IEEE, Transactions on Visualization and Computer Graphics* 25.8, pp. 2674–2693. URL: https://doi.org/10.1109/TVCG.2018.2843369.

Holzinger, Andreas, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell (2017). "What do we need to build explainable AI systems for the medical domain?" iss. Ml, pp. 1–28. URL: http://arxiv.org/abs/1712.09923.

Karim, Md Rezaul, Michael Cochez, Oya Beyan, Stefan Decker, and Christoph Lange (2019). "OncoNetExplainer: explainable predictions of cancer types based on gene expression data." In: *IEEE, 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 415–422. URL: https://arxiv.org/abs/1909.04169.

Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf (2015). "Principles of explanatory debugging to personalize interactive machine learning." In: *IUI '15, Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137. URL: https://doi.org/10.1145/2678025.2701399.

Kulesza, Todd, Simone Stumpf, Margaret Burnett, and Irwin Kwan (2012). "Tell me more? The effects of mental model soundness on personalizing

an intelligent agent." In: *CHI '12, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–10. URL: https://doi.org/10.1145/2207676.2207678.

Kulesza, Todd, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong (2013). "Too much, too little, or just right? Ways explanations impact end users' mental models." In: *IEEE, Symposium on visual languages and human centric computing*, pp. 3–10. URL: https://doi.org/10.1109/VLHCC.2013.6645235.

Kulesza, Todd, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M. Burnett, Ian Oberst, and Amy j. Ko (2009). "Fixing the program my computer learned: Barriers for end users, challenges for the machine. In: *IUI '09, Proceedings of the 14th international conference on Intelligent user interfaces*, pp. 187–196. URL: https://doi.org/10.1145/1502650.1502678.

Langer, Markus, Daniel Oster, Lena Kästner, Timo Speith, Kevin Baum, Holger Hermanns, Eva Schmidt, and Andreas Sesing (2021). "What do we want from Explainable Artificial Intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research." In: *Artificial Intelligence*, pp. 103473.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin (2016). "How We Analyzed the COMPAS Recidivism Algorithm". In: *ProPublica*. Retrieved March 3, 2021. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lim, Brian (2011). *Improving Understanding, Trust, and Control with Intelligibility in Context-Aware Applications*. PHD Thesis, Human-Computer Interaction Institute.

Lim, Brian Y., and Anind K. Dey (2009). "Assessing demand for intelligibility in context-aware applications." In: *UbiComp '09, Proceedings of the 11th international conference on Ubiquitous computing*, pp. 195–204. URL: https://doi.org/10.1145/1620545.1620576.

Lipton, Zachary C. (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57. URL: https://doi.org/10.1145/3236386.3241340.

Liu, Shixia, Xiting Wang, Mengchen Liu, and Jun Zhu (2017). "Towards better analysis of machine learning models: A visual analytics perspective." In: *Visual Informatics* 1.1, pp. 48–56. URL: https://doi.org/10.1016/j.visinf.2017.01.006.

Marino, Daniel L., Chathurika S. Wickramasinghe, and Milos Manic (2018). "An adversarial approach for explainable ai in intrusion detection systems." In: *IECON, 44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3237–3243. URL: https://doi.org/10.1109/IECON.2018.8591457.

Martin, Kyle, Anne Liret, Nirmalie Wiratunga, Gilbert Owusu, and Mathias Kern (2021). "Evaluating Explainability Methods Intended for Multiple Stakeholders." In: *KI-Künstliche Intelligenz* 35, pp. 1–15. URL: https://doi.org/10.1007/s13218-020-00702-6.

Miller, Perry L. (1986). "The evaluation of artificial intelligence systems in medicine." In: *Computer Methods and Programs in Biomedicine* 22.1, pp. 3–11.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial Intelligence* 267.February, pp. 1–38. URL: https://doi.org/10.1016/j.artint.2018.07.007.

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan (2018). "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." In: *Artificial Intelligence* 1.1, pp. 1–37. URL: http://arxiv.org/abs/1811.11839.

Montavon, Grégoire, Wojciech Samek, and Klaus Robert Müller (2018). "Methods for interpreting and understanding deep neural networks." In: *Digital Signal Processing* 73, pp. 1–15. URL: https://doi.org/10.1016/j.dsp.2017.10.011.

Narayanan, Menaka, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez (2018). "How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation." In: *arXiv.org*. URL: https://arxiv.org/abs/1802.00682.

Rader, Emilee, Kelley Cotter, and Janghee Cho (2018). "Explanations as Mechanisms for Supporting Algorithmic Transparency." In: *CHI '18, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper No. 103, pp. 1–13. URL: https://doi.org/10.1145/3173574.3173677.

Rudin, Cynthia (2019). "Stop explaining black bix machine learning models for high stakes decisions and use interpretable models instead." In: *Nature Machine Learning* 1.5, pp. 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x.

Rudin, Cynthia, and Joanna Radin (2019). "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition." In: *Harvard Data Science Review* 1.2, pp. 1–9.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Cham: Springer Nature.

Spinner, Thilo, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady (2020). "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning." In: *IEEE, Transactions on Visualization and Computer Graphics* 26.1, pp. 1064–1074. URL: https://doi.org/10.1109/TVCG.2019.2934629.