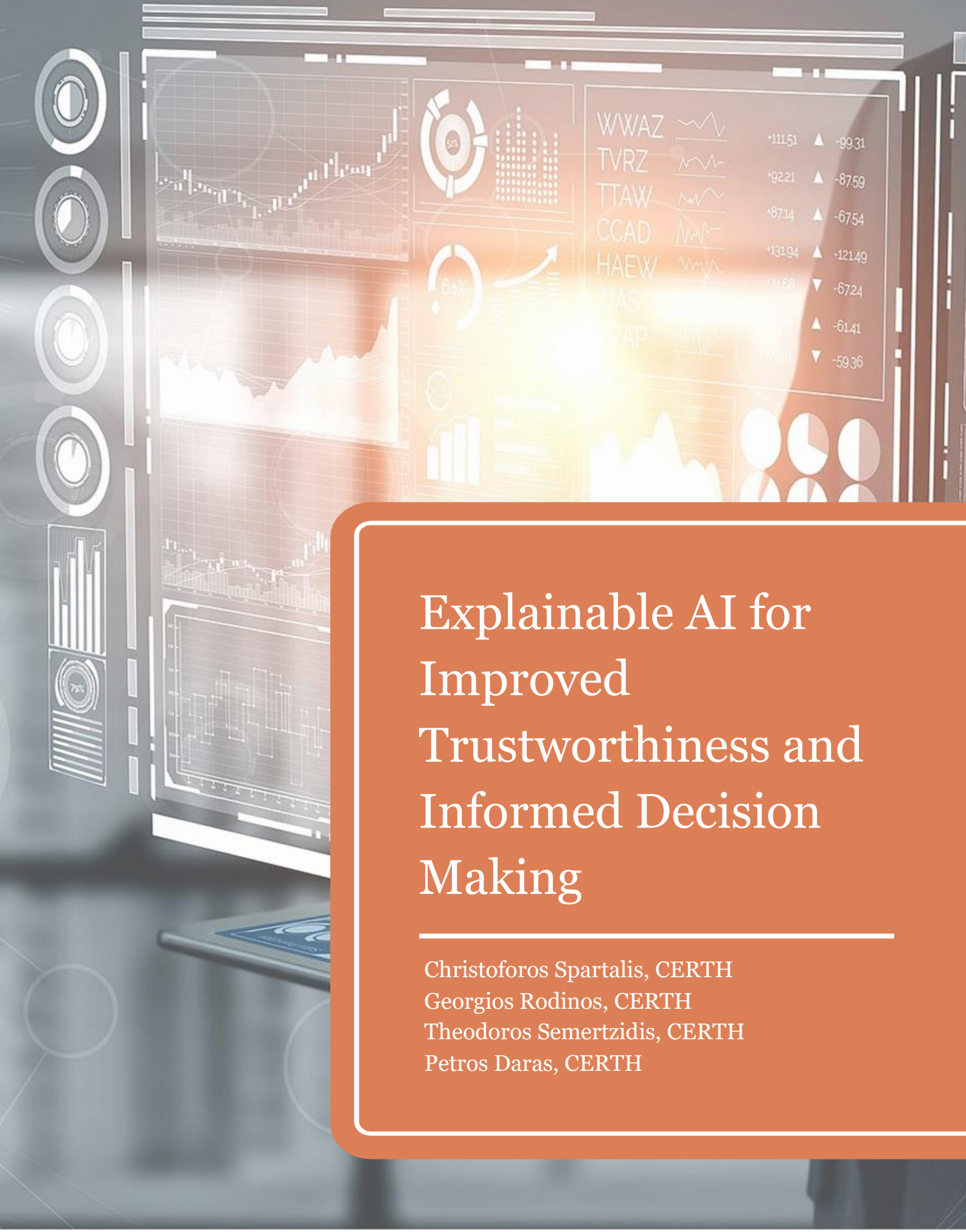

Explainable AI for Improved Trustworthiness and Informed Decision Making

Christoforos Spartalis, CERTH
Georgios Rodinos, CERTH
Theodoros Semertzidis, CERTH
Petros Daras, CERTH

Co-funded by the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Grant Agreement No. 101073909.

Explainable Artificial Intelligence for Improved Trustworthiness and Informed Decision Making

Christoforos Spartalis, Georgios Rodinos, Theodoros Semertzidis, and Petros Daras (CERTH)

This short paper presents the work performed in Task 4.2 in providing explainable AI (XAI) in the ATLANTIS federated learning environment. This set of tools will provide XAI approaches for summarizing incidents, providing recommendations, highlighting emerging risks and rewriting them in a simplified way, that lowers the barriers for cross-CI operators to accept and adopt them. An umbrella of tools will be delivered to address different types of data (e.g. images, videos, tabular data, text, social network posts etc.) and different scenarios in the defined ATLANTIS Large Scale Pilots. The short paper also discusses security issues that XAI may introduce and barriers in applying the different techniques in real world CIs.

1. Introduction

A crucial part in building trust and engaging users with AI systems is to enable deeper understanding of the inner-workings and parameters that affect the final decisions of the AI systems. Following this requirement, the main objective of XAI is to provide insights about the decision-making process of AI models in a human-understandable manner. However, XAI methods may unintentionally expose sensitive information about the training data and the models at hand. Thus, adversaries can exploit them to enhance security attacks. As such, the use of XAI in security critical applications should be carefully studied and evaluated.

Moreover, the explanations provided should be designed to be easily understood by cross-CI operator who may lack technological background. The explanations should also present high fidelity and stability since they are offered in a safety-critical context. Finally, these tools should unveil different aspects of explainability, allowing the end-users to choose the explanations that are more suitable per case.

2. The Current State of Affairs in XAI

The current situation in Explainable Artificial Intelligence (XAI) within the research field and on the market has been driven by the concept of Trustable AI [1]. This concept encompasses various requirements, including explainability, privacy, security, accountability, and continuous monitoring through end-user feedback. These demands, as have been articulated in regulations [2], standards, and guidelines [3] by EU expert groups [4], may present trade-offs at some point.

To shed light on these matters, in Figure 1, we present a hierarchical conceptual representation of the XAI taxonomy classes (i.e., different types of XAI) that we have highlighted based on the considerations we introduced. In Table 1, we provide a brief description of these classes.

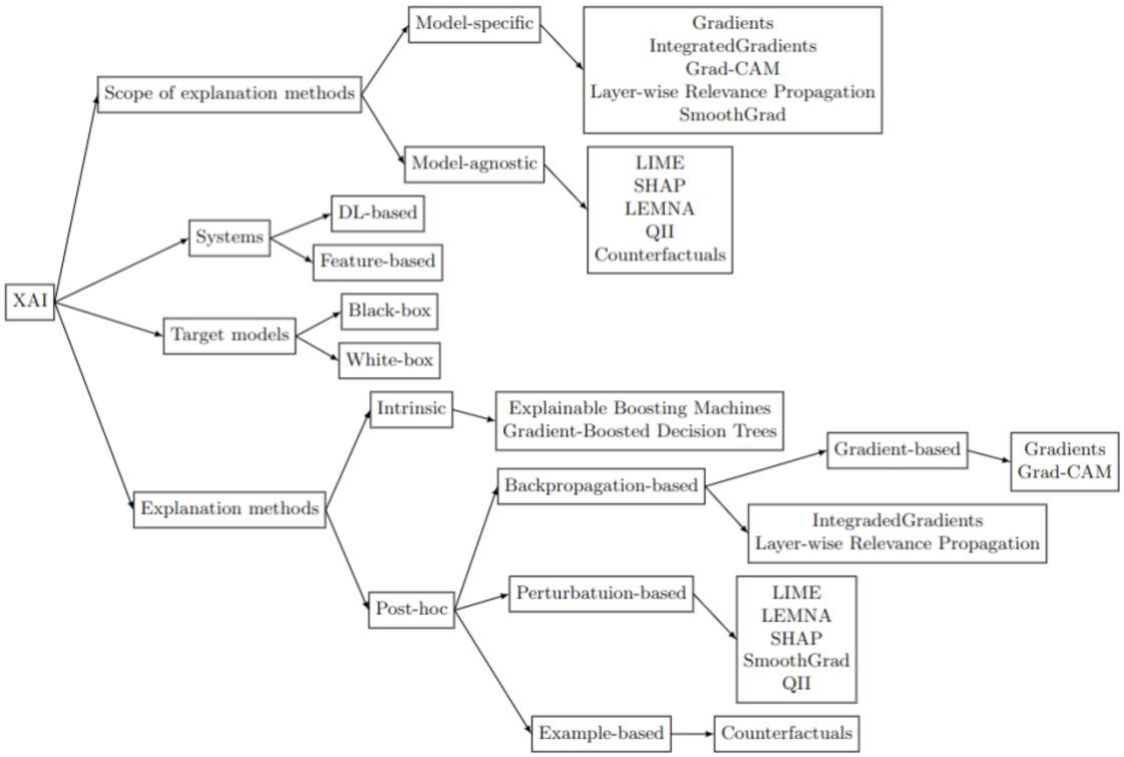

Figure 1. A conceptual depiction of XAI taxonomy classes relevant to our findings.

Table 1. Definitions of XAI taxonomy classes.

Taxonomy Class	**Definition**
Model-specific	Methods exclusive to certain model classes that are highly relying on their internal parameters and mechanisms, such as weights and gradients [5].
Model-agnostic	Methods that maintain the ability to generalize across any *DL-based* system [6].
DL-based	Systems that process input data such as images, signals, or text with numerous features [7].
Feature-based	Systems that mainly process tabular data with a limited number of features, including numerical and categorical values [7].
Black-box	Models characterized by their complexity and obscurity, which pose interpretability challenges for stakeholders [8][9].
White-box	Models that are inherently interpretable and provide complete transparency, offering full access to their parameters and architecture [10].
Intrinsic	Methods that commonly impose constraints on model complexity during training to inherently increase interpretability; typically associated with model-specific methods [5].
Post-hoc	Methods applied after model training to clarify model decisions; typically associated with model-agnostic methods [5].
Backpropagation-based	Methods that leverage backpropagation to assess feature attribution in model decision-making [11].
Perturbation-based	Methods that involve querying the model with slightly modified inputs to determine feature attribution in model decision-making [12].
Example-based	Methods that use specific instances from the dataset to elucidate model behaviour, without any manipulation of the features or the model itself [13].

Some useful remarks that we have highlighted in [14] are that white-box models are more vulnerable to privacy and security attacks than black-box ones, XAI methods "whiten" black-box models increasing privacy and security risks, and that the order of XAI methods in terms of privacy leak (from highest to lowest) is as follows: *Example-based, Gradient-based, Backpropagation-based,* and *Perturbation-based.*

3. The Role of WhiteBoxXAI-FL and BlackBoxXAI-noFL

We attempted to integrate insights from the growing XAI literature into the design process of our components. Notably, we developed an XAI component named WhiteBoxXAI-FL. It distinguishes itself by targeting white-box models, recognized for their interpretability yet susceptible to privacy and security threats. In augmenting the explanatory capacity of this module, we employ a Gradient-based XAI method named Grad-CAM [15], known for delivering explanations with higher fidelity and stability than other Backprobation-based and Perturbation-based methods. However, it also carries the higher potential risk of privacy leakage exploitable by adversaries. To address this concern, we explicitly specify that this component is tailored for deployment in a Federated Learning environment—a prevalent privacy-enhancement technique—thus mitigating the adverse impacts of our methodological choice.

Moreover, we offer another XAI component named BlackBoxXAI-noFL. It is designed for inherently obscure black-box models. This obscurity prevents, to some extent, adversaries from deducing insights about the model and the data. To augment the resilience of this module against privacy leaks that could possibly pose significant security issues, we employ a Perturbation-based XAI method named SHAP [16]. While these methods are recognized for leaking the minimum possible privacy, there is a trade-off with explanations of lower quality. Consequently, we ensure, to a certain extent, that this module can be seamlessly integrated into a centralized environment within a safety-critical context, such as ATLANTIS. The key characteristics of the implemented components and their main contributions are detailed in Table 2.

Table 2. Main contribution of methodological choices to the design of the XAI components.

Component	**Characteristics**	**Explainability/ Interpretability**	**Privacy/ Security**
WhiteBox-FL	White-box target model	X	
	Gradient-based XAI method	X	
	Federated Environment		X
BlackBox-noFL	Black-box target model		X
	Perturbation-based XAI method		X

Another aspect of explainability that we strive to incorporate into the design process of these modules is the ease of understanding for end-users. Enhancing this characteristic would facilitate the acceptance and adaptation of AI-driven decisions by end-users.

4. The Research and Development Path in ATLANTIS

ATLANTIS aims to strengthen the Cyber-Physical-Human (CPH) security of vital Critical Infrastructures (CI). Working on a project related to CI is challenging and it frequently comes with its own set of problems and obligations. The objective is to analyse potential risks in real time by using FL and XAI techniques. These methodologies will provide summaries of events as well as recommendations to improve cross-CI operator acceptability and adoption.

To meet the aforementioned objectives, two methods were implemented, as is illustrated in the *XAI Roadmap* in Figure 2.

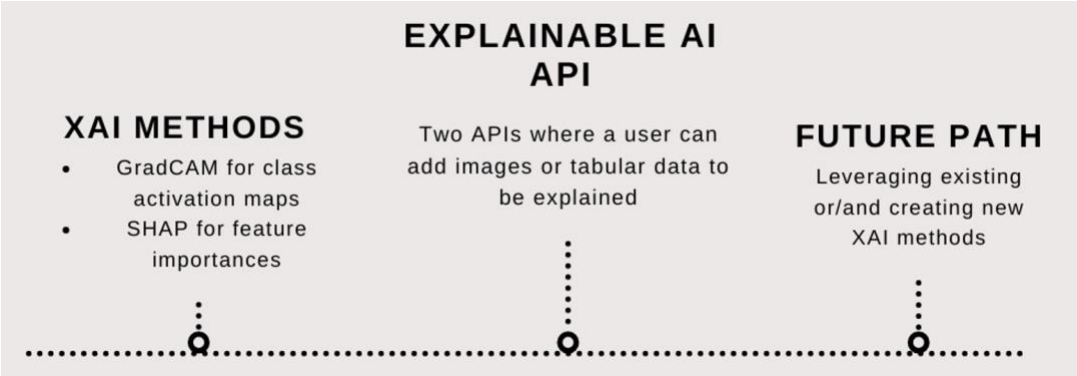

Figure 2. XAI roadmap.

The first one is Grad-CAM [15], a technique used for visualizing the regions of an input image that have the most influence on the model's decision. It overlays a heatmap over the input image, highlighting the areas where the model focused its attention throughout the decision-making process. This visualization can be particularly useful in image-based tasks, allowing you to understand which parts of an image were crucial for the model's decision. An illustration can be seen in the following image in Figure 3.

Figure 3. Class Aviation Map example from the Explainable AI API.

The second one SHAP [16] , which allows you to breakdown the prediction into contributions from individuals features, giving you insights into the relevance of each feature and the influence it has on the model's choice. For instance, in Figure 4 we can see the result regarding network intrusion, however it will not be described in detail since it is specialized for domain experts.

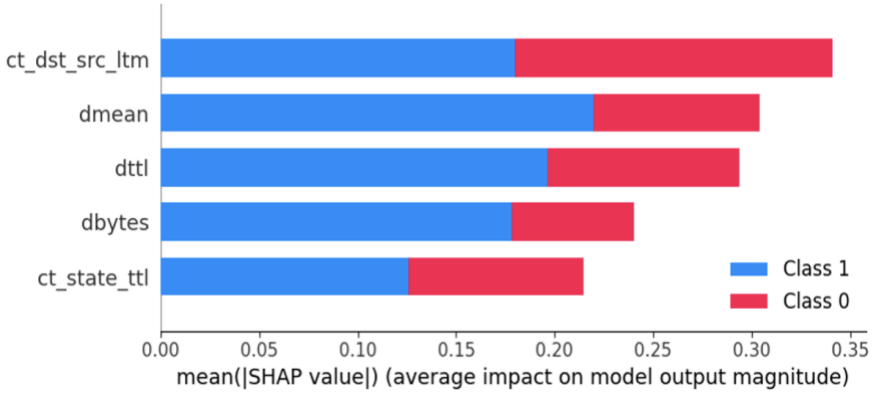

Figure 4. Feature Importance example from the Explainable AI API.

We concluded on these two methods after a thorough review of recent literature in XAI, particularly in the safety-critical domain. We have provided valuable insights into the complex interplay of explainability, privacy, and security, incorporating them into the design process as discussed in the previous section.

Our future work primarily focuses on the continued enhancement of the WhiteBoxXAI-FL and BlackBoxXAI-noFL components. This involves experimenting XAI techniques beyond Grad-CAM [15] and SHAP [16] while adhering to the requirements outlined in Section 3 and also integrating concepts and methodologies from related research domains to ensure the provision of high-quality explanations without exposing the AI systems to privacy and security threats [11][17][18]. The goal is to create explanations that are easily understood by end-users while uncovering new aspects of explainability.

The inclusion of a human-in-the-loop criterion in the design process is crucial. Incorporating human judgment and feedback into the AI decision-making process can improve comprehensibility. One possible approach is to create interactive interfaces leveraging the power of recent advances in Large Language Models (e.g., GPT3), that allow users to engage with the model and obtain real-time explanations for specific inputs. Based on unique requests (e.g., prompts), this user-driven method allows for a more tailored and precise explanation. In addition, allowing humans to provide feedback on the model's outputs can help the model perform better over time, since users can fix mistakes or provide suggestions to help the model learn and adapt.

New research approaches in the domain of Explainable AI are to create interpretable models by-design [19][20][21] and convert the model's decisions into human readable concepts rather than try to explain complex models [22].

In general, human-in-the-loop methods recognize the value of human intuition and domain expertise. User engagement can be leveraged to improve the capabilities of AI-based systems, making them comprehensible and trustworthy in real-world applications.

5. The Challenges and Barriers

While Explainable Artificial Intelligence (XAI) tools offer valuable insights into the decision-making processes, their implementation comes with various challenges and barriers [22].

For instance, complex models, such as deep neural networks, can be challenging to interpret. The intricate relationships between features in these models make it difficult to generate clear and understandable explanations. In addition, ensuring that XAI tools are applicable across different types of models is a challenge. Model-agnostic methods may not capture certain nuances specific to particular architectures. Finally, XAI tools may expose sensitive information in the training data, raising privacy concerns. This is particularly relevant when dealing with personal or confidential data. Thus, there might be limitations regarding the data availability.

The aforementioned challenges lead to some possible barriers. These may vary from the understanding the internal workings of highly complex models that may require advanced expertise in both the model architecture and the domain to the design of model-specific interpretability methods which may require additional effort, limiting the generalizability of the solution. Furthermore, implementing privacy-preserving XAI methods or developing mechanisms to handle sensitive information securely can be challenging mainly because access in model or/and data may be limited [23]. Finally, legal expertise may be required to navigate and comply with data protection and transparency regulations [24].

Another important aspect to take into consideration is to identify the needs of different stakeholders. End-users, who may not have a technical background, might struggle to understand the complex explanations generated by XAI tools. This could lead to misinterpretations or a lack of trust in the explanations [23]. Bridging the gap between technical and non-technical stakeholders is essential to ensure effective communication and understanding of the interpretability or explainability results. Explainability focuses more on technical individuals and tries to manifest the inner workings of a model whereas interpretability is more oriented to the description of the output of a model and targets non-experts.

6. The Benefits and Impact

In our design process, we place a strong emphasis on privacy and security considerations. Our goal is to strike a desirable balance in the development of each XAI component, ensuring explainability, privacy, and security. Simultaneously, we offer stakeholders two options, each with distinct features, allowing them to select the most suitable solution for their specific needs.

Moreover, as part of our design strategy, we incorporate Federated Learning to safeguard against the potential leakage of private information in explanations, thereby maintaining the confidentiality of sensitive data. This is particularly critical as breaches in privacy within CIs can lead to security concerns with far-reaching implications for individuals' well-being.

An interesting research field is the Large Language Models (LLMs). Even though it can be challenging in terms of explainability due to their complexity and the lack of explicit rules governing their behaviour, it can be utilized in different ways. For instance, incorporating

user feedback and allowing users to query the model for specific explanations can improve the overall interpretability.

7. Future Outlook

Our future work aims at digesting to a greater extent the security and privacy considerations in the development of the XAI toolset. Moreover, we will elaborate on human-in-the-loop scenarios to provide explanations that are easier to interpret from users with no technological background. Additionally, we aspire to investigate new aspects of explainability, approaching this subject from different angles to unveil new facets of the data and models at hand. To this end, we will attempt to transfer knowledge from growing relevant research fields.

For example, we aim to investigate how specific datapoints trigger specific nodes of a deep neural network. From this point of view, we could get insights about the decision-making of the AI system, and the data point per se. In the changing scene of cybersecurity, this implementation will help not only to identify how different data points affect the decision process, but will open the road for implementations that they will try to mitigate its influence in the DNN.

8. Conclusions

In this paper, we have presented our components designed for explainability, aiming to empower end-users and CI operators to accept the decisions of AI systems and adjust their actions accordingly. We have emphasized the crucial factors that have influenced recent research in the field, shaping the design process of the components that seek to strike a balance between explainability and privacy and security considerations. Finally, we have proposed paths for future work to enhance the comprehensibility of the provided explanations, particularly in relation to human-in-the-loop scenarios and LLMs. Additionally, we have suggested future directions that could enable the components to adapt to the evolving landscape of the cybersecurity domain.

References

[1] Yang, Guang, Qinghao Ye, and Jun Xia. "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond." *Information Fusion* 77 (2022): 29-52.

[2] European Commission: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) (2016).

[3] ISO, IEC: ISO/IEC 27001:2022(en), Information security, cybersecurity and privacy protection — Information security management systems — Requirements (2022).

[4] High-Level Expert Group on AI: Ethics guidelines for trustworthy ai. Tech. rep., European Commission, Brussels (Apr 2019).

[5] Carvalho, D., Pereira, E., Cardoso, J.: Machine learning interpretability: A survey on methods and metrics. Electronics (Switzerland) 8(8), 832 (2019).

[6] Zhao, X., Zhang, W., Xiao, X., Lim, B.: Exploiting Explanations for Model Inversion Attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 682–692 (2021).

[7] Petkovic, D.: It is Not "Accuracy vs. Explainability"—We Need Both for Trustworthy AI Systems. IEEE Transactions on Technology and Society 4(1), 46–53 (Mar 2023).

[8] Gurtler, M., Zollner, M.: Tuning white box model with black box models: Transparency in credit risk modeling. Available at SSRN 4433967 (2023).

[9] Warnecke, A., Arp, D., Wressnegger, C., Rieck, K.: Evaluating Explanation Methods for Deep Learning in Security. In: 2020 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 158–174 (Sep 2020).

[10] Loyola-Gonzalez, O.: Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access 7, 154096–154113 (2019).

[11] Shokri, R., Strobel, M., Zick, Y.: On the Privacy Risks of Model Explanations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 231–241. ACM, Virtual Event USA (Jul 2021).

[12] Bhusal, D., Rastogi, N.: Sok: Modeling explainability in security monitoring for trust, privacy, and interpretability. arXiv preprint arXiv:2210.17376 (2022).

[13] Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160 (2018).

[14] C. Spartalis, T. Semertzidis, P. Daras: Balancing XAI with Privacy and Security Considerations, Cyber-Physical Security for Critical Infrastructures Protection, CPS4CIP 2023.

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

[16] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

[17] Saifullah, S., Mercier, D., Lucieri, A., Dengel, A., & Ahmed, S. (2022). Privacy meets explainability: A comprehensive impact benchmark. *arXiv preprint arXiv:2211.04110*.

[18] Kuppa, A., & Le-Khac, N. A. (2020, July). Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In *2020 International Joint Conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

[19] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, *32*.

[20] Nauta, M., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2744-2753).

[21] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19187-19197).

[22] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

[23] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

[24] Ebers, M. (2020). Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s). *An Overview of the Current Legal Framework (s)(August 9, 2021). Liane Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics*.

Front cover image by Tung Nguyen via Pixabay.
https://pixabay.com/users/tungnguyen0905-17946924