Balancing XAI with Privacy and Security Considerations

Christoforos N. Spartalis[1][0000−0001−8228−4235], Theodoros Semertzidis[1][0000−0002−6825−4328], and Petros Daras[1][0000−0003−3814−6710]

Center for Research and Technology Hellas, Information Technologies Institute,Thessaloniki, Greece
{c.spartalis,theosem,daras}@iti.gr

Abstract. The acceptability of AI decisions and the efficiency of AI-human interaction become particularly significant when AI is incorporated into Critical Infrastructures (CI). To achieve this, eXplainable AI (XAI) modules must be integrated into the AI workflow. However, by design, XAI reveals the inner workings of AI systems, posing potential risks for privacy leaks and enhanced adversarial attacks. In this literature review, we explore the complex interplay of explainability, privacy, and security within trustworthy AI, highlighting inherent trade-offs and challenges. Our research reveals that XAI leads to privacy leaks and increases susceptibility to adversarial attacks. We categorize our findings according to XAI taxonomy classes and provide a concise overview of the corresponding fundamental concepts. Furthermore, we discuss how XAI interacts with prevalent privacy defenses and addresses the unique requirements of the security domain. Our findings contribute to the growing literature on XAI in the realm of CI protection and beyond, paving the way for future research in the field of trustworthy AI.

Keywords: Trustworthy AI · Explainable AI (XAI) · Privacy · Security · Critical Infrastructures (CI)

1 Introduction

The incorporation of Artificial Intelligence (AI) into the operational procedures of Critical Infrastructures (CI) calls for enhancing both the acceptability of AI decisions and the efficiency of collaboration between AI and human operators. To this end, eXplainable Artificial Intelligence (XAI) becomes essential in creating a trustworthy and widely accepted AI workflow.

The main objective of XAI is to provide insights about the decision-making processes of AI models in a human-understandable manner. Furthermore, it can be applied at all stages of delivery process, including development, validation/verification, accountable prediction, and maintenance [32]. XAI can improve AI systems by revealing hidden facets of models and extracting new knowledge from underlying data correlations and learned strategies [2].

However, XAI methods unintentionally leak sensitive information about the training data and models at hand [48][58]. Moreover, adversaries can exploit these methods to enhance privacy and security attacks [6][25][27][40][31][32]. Undoubtedly, highly accurate AI systems with explainability, privacy and security guarantees are complex but necessary, as emphasized by regulations [13], standards [18], and EU expert groups [17].

Hence, it is important to identify overlaps, conflicts, and trade-offs to explore the ideal compromise, particularly, when implementing AI systems in safety-critical domains with significant human impact. To this end, we review recent literature, categorize the findings, and present them in a comprehensive manner. We argue that our contribution attributes researchers and practitioners to design trustworthy AI systems that meet modern requirements.

In Section 2, we provide a concise overview of the fundamental concepts necessary for a comprehensive understanding of our findings. Then, in Section 3, we delve into the main analysis of the interplay between explainability, privacy, and security in AI. Finally, in Section 4, we consolidate our general conclusions.

2 Background

In this section, we aim to establish the background for a thorough understanding of our study. We do not intend to give an exhaustive analysis of all concepts, but rather to highlight those that are absolutely essential in supporting our findings (see Section 3). To this end, we start by constructing a partial taxonomy of XAI, focusing specifically on our areas of interest. We provide a conceptual representation of the classes pertinent to our study and briefly describe them. Furthermore, we address the evaluation of explainability in AI systems by enumerating qualitative indicators of explanations and outlining different levels of evaluations. Lastly, we succinctly describe the adversarial attack tests used to assess the robustness of AI systems.

2.1 XAI taxonomy classes

Constructing an XAI taxonomy is a complex task that requires a thorough and detailed analysis. A comprehensive work on this subject can be found in [8]. In our study, we selectively emphasize only on the classes presented in our findings (see Section 3); while excluding many others. Further, XAI methods with properties relevant to multiple general groups have been uniquely categorized based on the primary focus of the reviewed article in question.

Our XAI taxonomy, as illustrated in Figure 1, offers a hierarchical conceptual representation of the different classes relevant to our study. It encompasses the scope of explanation methods, including *model-specific* and *model-agnostic* approaches, as well as the types of explained systems, such as *DL-based* and *feature-based* systems. Furthermore, it addresses the transparency of the explained models, distinguishing between *white-box* and *black-box* models. Additionally, we classify XAI methods into *intrinsic* and *post-hoc* categories, with the *post-hoc* methods further classified into *example-based*, *backpropagation-based*, *gradient-based*, and *perturbation-based* methods. For concise definitions of these taxonomy classes, please refer to Table 1.

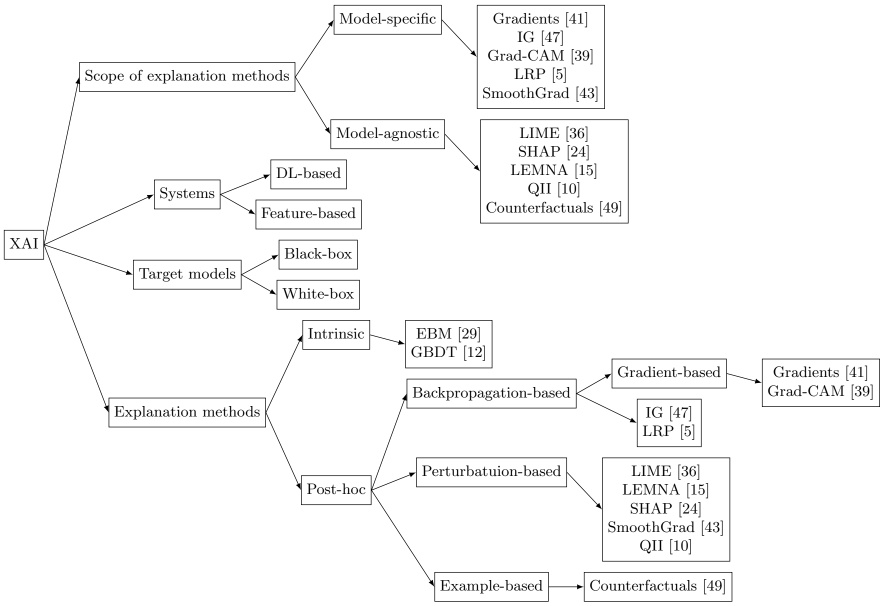

Fig. 1: A conceptual depiction of XAI taxonomy classes relevant to our findings.

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 [8].
Model-agnostic	Methods that maintain the ability to generalize across any *DL-based* system [58].
Deep Learning-based	Systems that process input data such as images, signals, or text with numerous features [32].
Feature-based	Systems that mainly process tabular data with a limited number of features, including numerical and categorical values [32].
Black-box	Models characterized by their complexity and obscurity, which pose interpretability challenges for stakeholders [52][16].
White-box	Models that are inherently interpretable and provide complete transparency, offering full access to their parameters and architecture [23].
Intrinsic	Methods that commonly impose constraints on model complexity during training to inherently increase interpretability; typically associated with model-specific methods [8].
Post-hoc	Methods applied after model training to clarify model decisions; typically associated with model-agnostic methods [8].
Backpropagation-based	Methods that leverage backpropagation to assess feature attribution in model decision-making [40].
Perturbation-based	Methods that involve querying the model with slightly modified inputs to determine feature attribution in model decision-making [6].
Example-based	Methods that use specific instances from the dataset to elucidate model behavior, without any manipulation of the features or the model itself [2].

2.2 Evaluation Criteria & Methods

In this section, we discuss the evaluation criteria and methods used to assess key aspects of trustworthy AI systems, including explainability, privacy, and security. Particularly regarding explanations, we also refer to the different levels at which evaluation can be conducted.

Explainability Despite the absence of universally accepted evaluation criteria [6][8], we have identified a set of qualitative indicators that is referred to many recent studies, including [20][4][6][8][30][52]. In Table 2, we highlight the metrics relevant to our findings and provide concise definitions.

Moreover, an intriguing aspect of this topic involves multi-level evaluation methods to assess explainability, as discussed in [8][6]. At the *functional-level*, quantitative measures are used as proxies for qualitative characteristics, eliminating the need for end-user experiments. However, solely relying on functional-level evaluation criteria can yield misleading results [6]. For a more comprehensive evaluation, it may be necessary to perform end-user experiments, involving domain experts at the *application-level* or laypersons at the *human-level* evaluation. By incorporating these different levels of evaluation, a more holistic understanding of the effectiveness and impact of explainability methods can be achieved.

Table 2: Definitions of XAI evaluation criteria.

Criterion	**Definition**
Accuracy	The extent to which the features identified as relevant in unseen data are truly so [8][6][52][44].
Completeness	The extent to which the explanations are meaningful and consistent across all possible inputs [52][37].
Comprehensibility	The degree to which end-users understand the generated explanations [8][6].
Contrastivity	The degree of difference in feature attributions assigned to different classes [6].
Efficiency	Pertains to the computational complexity and runtime of the XAI method; it measures the extent to which the typical workflow of the explainee is disrupted [6][52][28].
Faithfulness	Closely related to *accuracy*; it measures the impact on model performance when the most important features are eliminated one by one [6][4].
Fidelity	It measures the approximation quality of the surrogate interpretable model [8][6].
Robustness	It measures the resilience to both random noise and adversarial attacks [6][52].
Sparsity	The extent to which the number of features considered important is kept to a minimum [6][52][37].
Stability	To what extent the generated explanations of the same instance remain consistent across multiple runs [6][52], or similar explanations are generated for similar instances [8].
Usability	The intersection of *comprehensibility* and *efficiency* [6].

Privacy & Security Assessing privacy leakage in AI systems often involves taking the perspective of an adversary and measuring the success rate of the attack [40]. To this end, we enumerate and define popular privacy attacks discussed in the reviewed studies.

Attribute Inference refers to the disclosure of a sensitive attribute from a specific instance by utilizing the model's output and non-sensitive attributes [19][30].
Memebership Inference involves querying a particular data point to the target model to verify its presence in the training dataset [7][9][40][59].
Property Inference confirms the existence of a data point with specific properties within the training dataset [59].
Model Extraction (also known as *Model Stealing*) encompasses the construction of a surrogate model that mimics the behavior of the target model by creating a surrogate dataset and querying it to obtain the target model's decision boundaries [30][21][53][19].
Model Inversion involves reconstructing data from a private training dataset [30].

Furthermore, in Table 3, we categorize the corresponding reviewed studies based on the types of privacy attacks they employ, extending beyond the scope of explainability and security.

Table 3: Different types of privacy attacks featured in each reviewed study.

Reference	Authors	Attribute Inference	Membership Inference	Property Inference	Model Extraction	Model Inversion
[3]	Aivodji et al.				✓	
[6]	Bhusal and Rastogi		✓		✓	
[7]	Carlini et al.		✓			
[9]	Choquette-Choo et al.		✓			
[19]	Izzo et al.		✓			
[21]	Kariyappa and Qureshi				✓	
[27]	Milli et al.				✓	
[28]	Miura et al.				✓	
[30]	Oksuz et al.				✓	
[40]	Shokri et al.		✓			
[45]	Song and Shmatikov	✓				
[46]	Stadler et al.	✓	✓			
[48]	Truong et al.				✓	
[51]	Wainakh et al.					✓
[53]	Yan et al.				✓	
[55]	Yin et al.					✓
[58]	Zhao et al.					✓
[57]	Zhao et al.					✓
[59]	Zhu and Han		✓	✓		✓

Privacy and security attacks have different goals. Privacy attacks aim to leak sensitive information or violate the intellectual property of the target model, while security attacks focus on degrading the overall performance of the model

[22]. However, the robustness of an AI system to both types of attacks is evaluated using the same approach [42][56]. Thus, as we have done for privacy attacks, we define the popular security attacks found in the reviewed literature:

Poisoning refers to manipulating training data or model parameters to degrade the model performance [50].
Evasion involves manipulating input data during inference to trigger incorrect model outputs with high confidence [56].

3 Findings

Our research aims to shed light on trade-offs, challenges, and opportunities that arise from the interplay of explainability, privacy, and security in AI. In this analysis, XAI serves as a cornerstone. This signifies that the identified aspects of this analysis are primarily examined from the angle of different XAI taxonomy classes or methods. Moreover, we approach privacy from two perspectives: potential attacks and prevalent defenses. Additionally, our security analysis takes into account two facets: the security of AI systems, focusing on the protection of the AI systems per se, and security enabled by AI systems, which seeks to enhance overall security measures in various domains and applications.

3.1 Privacy Attacks

AI models are already susceptible to inference attacks [40][7]. This vulnerability can be particularly relevant to specific architecture categories, such as those used in *DL-based* systems [21] or data properties such as underrepresented population groups [19][31][40]. However, privacy risks escalate when adversaries have explanations for decision-making at hand [6][25][27][40][31][32]. For example, *model-agnostic* explanation methods can be used in *black-box* models, which are inherently more resilient to privacy attacks [6][31], to mitigate their obscurity and thus increase their vulnerability [58][53]. Another method involves *model extraction* [48]. Indeed, we argue that there is an intriguing reciprocal interaction between privacy and explainability; a privacy attack provides model interpretability and exposes it to higher risk for subsequent attacks.

In categorizing our findings into broader XAI taxonomy classes, we propose that *example-based* methods may be the most prone to privacy leakage. Comparing these with *backpropagation-based* and *perturbation-based* methods brings to light their distinctive risk profiles. The level of information leakage is so substantial that the models being explained become vulnerable not just to *membership inference* but also to *model inversion* attacks [40].

After *example-based*, *backpropagation-based* methods, particularly *gradient-based* ones, stand out as the next most significant source of privacy leak [40]. The variance of these explanations reveals statistical information on the decision boundaries of the model [6][40]. High variance indicates that a data point is close to the decision boundaries, which primarily suggests a lower probability of participating in the training set [40]. In fact, the correlation between variance and privacy leakage intensifies as the number of data features increases [40]. Based on a detailed analysis [58], *Grad-CAM* [39] is the most revealing method, followed by *Gradients* [41] and *LRP* [5]. This could be because *Class Activation Mapping (CAM)* explanations take into account a transformation of the original input, in addition to gradient information [58]. An additional justification could be that *backpropagation-based* methods, which are not *gradient-based*, such as *LRP* and *Integrated Gradients (IG)* [47], tend to lie near a low-dimensional manifold (i.e., they violate the data-manifold hypothesis) [31], and explanations with lower *fidelity* face less privacy risks [40]. Another general observation is that explanations focusing on neuron activations (like *Grad-CAM*) leak more privacy than those focusing on the model's output with respect to the input (like *Gradients*, *IG*, *LRP*, and *perturbation-based* methods) [58].

Perturbation-based methods, such as *LIME* [36] and *SmoothGrad* [43]), are found to be more resilient to *inference* and *model extraction* attacks [40][53]. This resilience may stem from their reliance on out-of-distribution (OOD) or off-manifold samples, resulting in lower *fidelity* and *stability* [31][40]. As mentioned previously, there is an underlying relation of explainability evaluation criteria and privacy leakage, as XAI methods that provide better explanations tend to leak more sensitive information [53]. Thus, a comparative analysis of *LIME*, *LEMNA* [15], and *SHAP* [24] is worth mentioning. This analysis [6] concludes that *LIME* achieves the highest *stability*, *SHAP* exhibits the lowest *faithfulness* and the highest *sparsity*, and all three methods demonstrate high scores in terms of *contrastivity* [6].

Remark 1. XAI methods "whiten" *black-box* models, increasing privacy risks.

Remark 2. Better explanations, higher exposure to privacy risks.

Remark 3. Order of XAI methods in terms of privacy leak (highest first):
Example-based > *Gradient-based* ≥ *Backpropagation-based* > *Perturbation-based*

3.2 Privacy Defences

Research in privacy-enhancing AI has converged on several prominent methods, including *Differential Privacy (DP)*, *Federated Learning (FL)*, *Homomorphic Encryption (HE)*, and anonymization techniques such as generation of *synthetic data*.

DP introduces statistical noise into the data or the model [1], which can lead to extremely convoluted decision boundaries [31], hampering the *fidelity* and *comprehensibility* of explanations [38]. However, these explanations can reveal additional information about the data or the model, increasing the privacy budget to be spent by *DP* mechanisms [31].

It has been demonstrated that *perturbation-based* methods suffer less from the adverse impacts of *DP* [38]. When employed together, these techniques can strike a balance between explainability and privacy [31][10]. Furthermore, when

DP is incorporated into inherently interpretable algorithms, such as *EMBs* [29], the objective extends beyond achieving high prediction accuracy and privacy [29]. The negative effects of noise can be mitigated post-training, and desirable constraints, like monotonicity, can be imposed. Moreover, it has been argued that the combination *DP* and *FL* could mitigate some of *DP*'s negative impacts [38].

FL is a collaborative learning process that offers a degree of privacy, as the model parameters or gradients shared by clients with the central server carry less sensitive information than raw data [26][54]. However, *model inversion* is possible using just publicly shared gradients [57][59][55], suggesting that *FL* is not foolproof in terms of privacy preservation, and *gradient-based* explanations can further worsen privacy leakage. Moreover, in this collaborative set up, clients can only make partial observations, potentially expressing doubts about model outputs or explanations [35]. We argue that resolving these conflicts might necessitate full information disclosure, which disregards privacy. On the other hand, such environments facilitate the provision of culture-based explanations that are tailored to individual clients [34].

HE allows computations on encrypted data without the need for decryption [33], and it has been effectively used in conjunction with XAI [14]. However, it comes with significant computational overhead and imposes limitations on the model architecture and types of operations, which complicates the integration with *intrinsic* constraints for interpretability [14]. Finally, the generation of *synthetic data* undermines interpretability, as it blurs the distinction between real and artificial information [46].

Remark 4. Each privacy-enhancing technique presents unique trade-offs with explainability, potentially varying across different XAI taxonomy classes.

Remark 5. Using a combination of privacy-enhancing techniques may better balance privacy and explainability.

3.3 Security Aspects

Each point discussed above is equally pertinent in the realm of security, as privacy concerns can trigger a cascading effect, thus escalating security risks. First, the data under consideration often contains sensitive information, the disclosure of which can critically compromise safety [11]. Second, if an adversary has already breached the privacy of an AI model, the process of crafting malicious samples is simpler [21].

Moreover, by revealing the inner-workings of AI systems, XAI methods can directly pose security risks [32]. The exploration of XAI in security is still not exhaustive; establishing robust defenses and defining additional prerequisites that need to be met remain open research questions [52]. The unique treatment this field necessitates arises from the participation of different stakeholders, the increased complexity of the systems, and the profound correlation between privacy and security concerns [6]. On this basis, the following evaluation criteria have been underscored as especially important: *accuracy*, *completeness*, *fidelity*, *robustness*, *stability*, and *usability* [52][6]. However, security-oriented explanations cannot yet achieve high *fidelity* and *stability* [6], perhaps due to the popularity of *DL-based* systems in the security domain [52].

Remark 6. Privacy concerns can lead to security risks.

Remark 7. The intersection of XAI and security presents unique characteristics which require further research.

4 Conclusions

Throughout our literature review, we identified numerous challenges involving explainability, privacy, and security within AI, and delved into the inherent trade-offs at their intersections. These findings were categorized according to the prevalent XAI taxonomy classes. To emphasize the contribution of our literature review, we consolidated our findings in the form of remarks in Section 3.

We underscored the significant role of XAI in bridging the vulnerability gap that exists between *black-box* and *white-box* models. Our study also emphasized the fundamental connection between explainability evaluation criteria and the success rate of adversarial attacks.

In the realm of privacy, we brought to light the potential risks induced by various XAI methods, highlighting the role of the quality of the explanations produced. *Example-based* methods pose the highest risk, followed by *Gradient-based*, other *Backpropagation-based*, and lastly, *Perturbation-based* methods. We also explored the intricate relationship between XAI and prevalent privacy defenses, highlighting the unique trade-offs associated with each privacy-enhancing technique. Moreover, we shedded light on promising results and future work directions.

Approaching the security aspects, we underlined the strong correlations of explainability and privacy vulnerabilities with security concerns. We elaborated on the additional challenges posed when incorporating XAI in security applications, emphasizing the urgent need for further research in this area.

In our future work, our aim is to implement these findings in actual use cases within CI, while exploring and taking into consideration the prioritization of requirements across different application domains.

In conclusion, we believe that our work enriches the expanding literature on XAI in the realm of CI protection and beyond. We provide a solid groundwork, supporting future research aimed at addressing challenges and balancing trade-offs inherent in trustworthy AI.

Acknowledgments

This work was partially supported by the EU funded project ATLANTIS (Grant Agreement Number 101073909).

References

1. Abadi, M., McMahan, H., Chu, A., Mironov, I., Zhang, L., Goodfellow, I., Talwar, K.: Deep learning with differential privacy. In: Proceedings of the ACM Conference on Computer and Communications Security. pp. 308–318 (2016). https://doi.org/10.1145/2976749.2978318
2. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Aïvodji, U., Bolot, A., Gambs, S.: Model extraction from counterfactual explanations. arXiv preprint arXiv:2009.01884 (2020)
4. Alvarez Melis, D., Jaakkola, T.: Towards Robust Interpretability with Self-Explaining Neural Networks. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
5. 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**(7), e0130140 (Jul 2015). https://doi.org/10.1371/journal.pone.0130140
6. Bhusal, D., Rastogi, N.: Sok: Modeling explainability in security monitoring for trust, privacy, and interpretability. arXiv preprint arXiv:2210.17376 (2022)
7. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership Inference Attacks From First Principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914 (May 2022). https://doi.org/10.1109/SP46214.2022.9833649
8. Carvalho, D., Pereira, E., Cardoso, J.: Machine learning interpretability: A survey on methods and metrics. Electronics (Switzerland) **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
9. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-Only Membership Inference Attacks. In: Proceedings of the 38th International Conference on Machine Learning. pp. 1964–1974. PMLR (Jul 2021)
10. Datta, A., Sen, S., Zick, Y.: Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In: 2016 IEEE Symposium on Security and Privacy (SP). pp. 598–617 (May 2016). https://doi.org/10.1109/SP.2016.42
11. De La Torre Parra, G., Selvera, L., Khoury, J., Irizarry, H., Bou-Harb, E., Rad, P.: Interpretable Federated Transformer Log Learning for Cloud Threat Forensics. In: Proceedings 2022 Network and Distributed System Security Symposium. Internet Society, San Diego, CA, USA (2022). https://doi.org/10.14722/ndss.2022.23102
12. Dong, T., Li, S., Qiu, H., Lu, J.: An interpretable federated learning-based network intrusion detection framework. arXiv preprint arXiv:2201.03134 (2022)
13. 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), `https://eur-lex.europa.eu/eli/reg/2016/679/oj`
14. Franco, D., Oneto, L., Navarin, N., Anguita, D.: Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition. Entropy **23**(8) (2021). https://doi.org/10.3390/e23081047
15. Guo, W., Mu, D., Xu, J., Su, P., Wang, G., Xing, X.: LEMNA: Explaining Deep Learning based Security Applications. In: Proceedings of the 2018 ACM

SIGSAC Conference on Computer and Communications Security. pp. 364–379. CCS '18, Association for Computing Machinery, New York, NY, USA (Oct 2018). https://doi.org/10.1145/3243734.3243792

16. Gürtler, M., Zöllner, M.: Tuning white box model with black box models: Transparency in credit risk modeling. Available at SSRN 4433967 (2023)
17. High-Level Expert Group on AI: Ethics guidelines for trustworthy ai. Tech. rep., European Commission, Brussels (Apr 2019), `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`
18. ISO, IEC: ISO/IEC 27001:2022(en), Information security, cybersecurity and privacy protection — Information security management systems — Requirements (2022)
19. Izzo, Z., Yoon, J., Arik, S.O., Zou, J.: Provable membership inference privacy. In: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022 (2022)
20. Jiang, H., Kim, B., Guan, M., Gupta, M.: To Trust Or Not To Trust A Classifier. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
21. Kariyappa, S., Qureshi, M.K.: Defending Against Model Stealing Attacks With Adaptive Misinformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2020)
22. Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., Vasilakos, A.: Privacy and Security Issues in Deep Learning: A Survey. IEEE Access **9**, 4566–4593 (2021). https://doi.org/10.1109/ACCESS.2020.3045078
23. Loyola-González, O.: Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access **7**, 154096–154113 (2019)
24. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
25. Malek–Podjaski, M., Deligianni, F.: Towards Explainable, Privacy-Preserved Human-Motion Affect Recognition. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 01–09 (Dec 2021). https://doi.org/10.1109/SSCI50451.2021.9660129
26. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (Apr 2017)
27. Milli, S., Schmidt, L., Dragan, A.D., Hardt, M.: Model Reconstruction from Model Explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 1–9. ACM, Atlanta GA USA (Jan 2019). https://doi.org/10.1145/3287560.3287562
28. Miura, T., Hasegawa, S., Shibahara, T.: Megex: Data-free model extraction attack against gradient-based explainable ai. arXiv preprint arXiv:2107.08909 (2021)
29. Nori, H., Caruana, R., Bu, Z., Shen, J.H., Kulkarni, J.: Accuracy, Interpretability, and Differential Privacy via Explainable Boosting. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8227–8237. PMLR (Jul 2021)
30. Oksuz, A.C., Halimi, A., Ayday, E.: Autolycus: Exploiting explainable ai (xai) for model extraction attacks against decision tree models. arXiv preprint arXiv:2302.02162 (2023)

31. Patel, N., Shokri, R., Zick, Y.: Model Explanations with Differential Privacy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1895–1904. ACM, Seoul Republic of Korea (Jun 2022). https://doi.org/10.1145/3531146.3533235
32. 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). https://doi.org/10.1109/TTS.2023.3239921
33. Phong, L., Aono, Y., Hayashi, T., Wang, L., Moriai, S.: Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Transactions on Information Forensics and Security **13**(5), 1333–1345 (2018). https://doi.org/10.1109/TIFS.2017.2787987
34. Raymond, A., Gunes, H., Prorok, A.: Culture-Based Explainable Human-Agent Deconfliction. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1107–1115. AAMAS '20, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (May 2020)
35. Raymond, A., Malencia, M., Paulino-Passos, G., Prorok, A.: Agree to Disagree: Subjective Fairness in Privacy-Restricted Decentralised Conflict Resolution. Frontiers in Robotics and AI **9** (2022)
36. Ribeiro, M., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 97–101. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-3020
37. Saeed, W., Omlin, C.: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems **263**, 110273 (Mar 2023). https://doi.org/10.1016/j.knosys.2023.110273
38. Saifullah, S., Mercier, D., Lucieri, A., Dengel, A., Ahmed, S.: Privacy meets explainability: A comprehensive impact benchmark. arXiv preprint arXiv:2211.04110 (2022)
39. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
40. 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). https://doi.org/10.1145/3461702.3462533
41. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: Proceedings of the International Conference on Learning Representations (ICLR). ICLR (2014)
42. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 180–186. ACM, New York NY USA (Feb 2020). https://doi.org/10.1145/3375627.3375830
43. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
44. Song, C., Shmatikov, V.: Overlearning reveals sensitive attributes. In: 8th International Conference on Learning Representations, ICLR 2020 (2020)
45. Song, Q., Lei, S., Sun, W., Zhang, Y.: Adaptive federated learning for digital twin driven industrial internet of things. In: IEEE Wireless Com-

munications and Networking Conference, WCNC. vol. 2021-March (2021). https://doi.org/10.1109/WCNC49053.2021.9417370

46. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic Data – Anonymisation Groundhog Day. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1451–1468 (2022)
47. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 3319–3328. PMLR (Jul 2017)
48. Truong, J.B., Maini, P., Walls, R.J., Papernot, N.: Data-Free Model Extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4771–4780 (2021)
49. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology (Harvard JOLT) **31**, 841 (2017)
50. Wahab, O.A., Mourad, A., Otrok, H., Taleb, T.: Federated Machine Learning: Survey, Multi-Level Classification, Desirable Criteria and Future Directions in Communication and Networking Systems. IEEE Communications Surveys & Tutorials **23**(2), 1342–1397 (2021). https://doi.org/10.1109/COMST.2021.3058573
51. Wainakh, A., Müßig, T., Grube, T., Mühlhäuser, M.: Label Leakage from Gradients in Distributed Machine Learning. In: 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC). pp. 1–4 (Jan 2021). https://doi.org/10.1109/CCNC49032.2021.9369498
52. 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). https://doi.org/10.1109/EuroSP48549.2020.00018
53. Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q., Dong, C.: Explanation leaks: Explanation-guided model extraction attacks. Information Sciences **632**, 269–284 (Jun 2023). https://doi.org/10.1016/j.ins.2023.03.020
54. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology **10**(2) (2019). https://doi.org/10.1145/3298981
55. Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See Through Gradients: Image Batch Recovery via GradInversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16337–16346 (2021)
56. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: 29th {USENIX} Security Symposium ({USENIX} Security 20) (2020)
57. Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)
58. 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)
59. Zhu, L., Han, S.: Deep Leakage from Gradients. In: Yang, Q., Fan, L., Yu, H. (eds.) Federated Learning: Privacy and Incentive, pp. 17–31. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-63076-8_2