Balancing XAI with Privacy and Security Considerations

C. Spartalis[1] T. Semertzidis[1] P. Daras[1]

[1]Center for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece

Co-funded by the
European Union

CPS4CIP, September 2023

eXplainable AI (XAI)

XAI objectives:

Justify	the decision-making of AI models
Control	all stages of delivery process for accountable prediction and maintenance
Improve	AI systems by revealing hidden facets of models
Extract	new knowledge from underlying data correlations and learned strategies

Why XAI for CI?

- Acceptability of AI decisions
- Efficiency of collaboration between AI and human operators

Motivation / Contribution

- XAI leaks information about the training data and the AI model per se
- Malicious actors can use explanations to enhance privacy and security attacks
- Regulations, standards, and guidelines from EU expert groups require explainability, privacy, and security guarantees

- Review the recent literature
- Identify overlaps, conflicts, and trade-offs
- Organize the findings on some major XAI taxonomy classes
- Present findings in a comprehensive manner by providing essential background knowledge of the fundamental concepts
- Contribute to the growing literature on XAI in the realm of CI protection
- Attribute researchers and practitioners to build AI systems that meet modern requirements

Partial Taxonomy of XAI

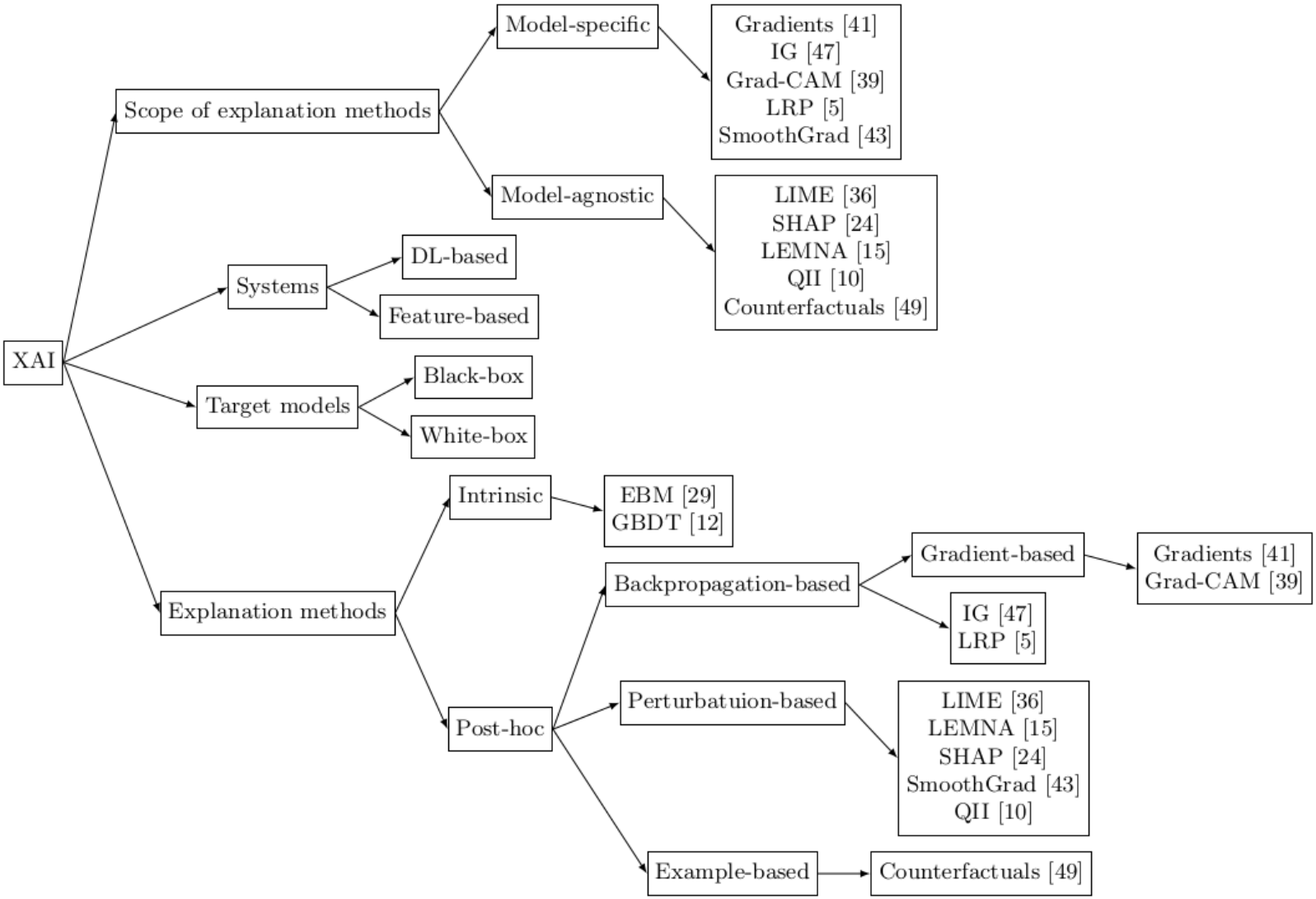

Figure: A conceptual depiction of XAI taxonomy classes relevant to our findings.

Evaluation Criteria of Explainability

Challenge

Absence of universally accepted evaluation criteria of explainability

Table: Definitions of XAI evaluation criteria.

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

Evaluation Criteria of Privacy & Security

Assessing privacy and security breaches

Taking the perspective of an adversary and measuring the success rate of attacks.

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

Authors	Attribute Inference	Membership Inference	Property Inference	Model Extraction	Model Inversion
Aivodji et al.				✓	
Bhusal and Rastogi		✓		✓	
Carlini et al.		✓			
Choquette-Choo et al.		✓			
Izzo et al.		✓			
Kariyappa and Qureshi				✓	
Milli et al.				✓	
Miura et al.				✓	
Oksuz et al.				✓	
Shokri et al.		✓			
Song and Shmatikov	✓				
Stadler et al.	✓	✓			
Truong et al.				✓	
Wainakh et al.					✓
Yan et al.				✓	
Yin et al.					✓
Zhao et al.					✓
Zhao et al.					✓
Zhu and Han		✓	✓		✓

Methodology

- Interplay of explainability, privacy, and security $\rightarrow$
  Trade-offs, challenges, and opportunities
- XAI serves as the cornerstone
- Two perspectives of privacy
  - Potential attacks (inference, model extraction and model inversion attacks)
  - Privacy-enhancement techniques (DP, FL, HE, synthetic data generation)
- Two facets of security
  - Security of AI systems (defenses against poisoning and evasion attacks)
  - Security enabled by AI systems (e.g., the scenario of a security operation center)

Findings

Privacy Attacks

- Black-box models $+$ Model-agnostic XAI methods $=$ Privacy leakage
- Example-based XAI methods $\rightarrow$ Higher fidelity and stability $\rightarrow$ More privacy leakage
- Backpropagation-based methods reveal statistical information about the decision boundaries
  - Higher variance $\rightarrow$ Data point close to the decision boundary $\rightarrow$ Less probable to participate in the training set

Findings

Privacy Attacks

- Gradient-based methods leak the most among Backpropagation-based methods
  - CAM methods (e.g., Grad-CAM) take into account additional information
  - Non-Gradient-based methods (e.g., LRP, IG) violate data-manifold hypothesis $\rightarrow$ Lower fidelity
  - Explanations focusing on neuron activation (e.g., Grad-CAM) leak more privacy
- Perturbation-based methods (e.g., LIME, SmoothGrad, SHAP, LEMNA) are more resilient to privacy attacks
  - OOD or off-manifold perturbed inputs $\rightarrow$ Lower fidelity and stability
  - LIME $\rightarrow$ High stability
  - SHAP $\rightarrow$ High sparsity $\rightarrow$ Security applications (alert handling)

Findings

Privacy-enhancement Techniques

- DP $\rightarrow$ statistical noise into the data or the model
  - It hampers fidelity and comprehensibility
  - Explanations increase the privacy budget to be spent by DP mechanisms
  - Perturbation-based methods suffer less from DP
  - Mitigation of negative effects of DP (DP $+$ EBMs and DP $+$ FL)
- FL $\rightarrow$ a collaborative training process that offers a degree of privacy
  - Model inversion attacks using gradients $\rightarrow$
    XAI methods revealing gradient information pose more risks
  - Clients make partial observations $\rightarrow$
    Doubts about the generated explanations
  - Provision of culture-based explanations; tailored to individual clients
- HE $\rightarrow$ computation on encrypted data without the need for decryption
  - Significant computational overhead
  - Limitation on the type of operations
- Synthetic data $\rightarrow$ Interpretability undermining

Findings

Security Aspects

Challenge

The exploration of XAI in the security domain is not yet exhaustive

- Privacy - Security interdependence
  - Privacy concerns can escalate to security risks
  - Breached privacy $\rightarrow$ Crafting malicious samples is simpler
  - Disclosure of sensitive information $\rightarrow$ Safety risks
- Privacy-security correlation, increased system complexity, different stakeholders $\rightarrow$ Unique treatment
- Prominent evaluation criteria of explanability in the security domain
  - Accuracy, Completeness, Fidelity, Robustness, Stability, and Usability
  - Popularity of DL-based systems hampers fidelity and stability

Conclusions

Remark 1: Privacy Attacks

XAI methods "whiten" black-box models, increasing privacy risks

Remark 2: Privacy Attacks

Better explanations $\rightarrow$ higher exposure to privacy risks

Remark 3: Privacy Attacks

Order of XAI methods in terms of privacy risks (higher first)
Example-based $>$ *Gradient-based* $\geq$ *Backpropagation-based* $>$ *Perturbation-based*

Remark 4: Privacy-enhancement Techniques

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

Conclusions

Remark 5: Privacy-enhancement Techniques

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

Remark 6: Security Aspects

Privacy concerns can lead to security risks

Remark 7: Security Aspects

The intersection of XAI and security presents unique characteristics that require further research

Thank you for your attention!