Securing XAI through Trusted Computing

Maximilian Becker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
maximilian.becker@kit.edu

Abstract

The escalating use of Artificial Intelligence (AI) and Machine Learning (ML) systems underscores the need for transparency and data security. This paper explores the fusion of Explainable AI (XAI) with trusted computing technologies such as Trusted Platform Modules (TPMs) and Trusted Execution Environments (TEEs). Highlighting the synergy between XAI, aimed at elucidating ML decision-making, and trusted computing, which fortifies system integrity, this study introduces novel approaches. Specifically, it proposes leveraging TEEs to protect user privacy during XAI computation and TPMs to verify system trustworthiness. This integration seeks to augment trust in AI systems by securing personal data processing and ensuring system integrity, thereby potentially reshaping the landscape of trust in AI technologies.

1 Introduction

AI and ML have become ever more utilized in recent years. This also increases the demand for XAI to increase the transparency of these systems. Because such systems are used more frequently they also process more and more personal data. The European Union is planning on putting forward certain transparency requirements for high-risk applications of AI [4]. This would increase the

demand for XAI even more. It order to convince users to share their personal data with AI systems they need to trust the systems. Trust is often mentioned in the XAI literature [9]. It is hard to trust a system you do not understand; XAI can help in this regard by making the AI system more understandable. However, this trust is not warranted if the system processing the personal data can be tempered with. To avoid this trusted computing technologies can be used. This combination could ensure trust in the system through trusted computing and trust in the AI through XAI.

We present some background on XAI and trusted computing in Section 2. Afterwards in Section 3 we show some concepts on how XAI and trusted computing can be combined. Section 4 ends with a short summary.

2 Background

Here we present some background, first on XAI and later about trusted computing technologies.

2.1 XAI

The goal of XAI is to make ML systems more transparent by generating explanations [2]. These explanations can explain individual decisions of ML models or the models as a whole. There are many different techniques utilized in XAI and the explanations themselves can also take different forms. The explanations can show different features and how important they are for the model, they can show alternatives that can change a prediction or show correlations between different features. Different explanation methods can be used for different use cases and user groups. Some explanation methods are very technical and produce complicated graphs. These methods can help developers to gain deep insight into the models they train. There are also methods that are oriented more towards end users. Such methods can for example deliver explanations in natural language understandable without technical knowledge.

The European Union is working on the Artificial Intelligence Act (AI Act) [4], a legislation that puts forward transparency requirements for high-risk applications

of AI. This regulation would require XAI to fulfill these requirements. The AI Act specifically mentions that ”Users should be able to interpret the system output and use it appropriately”. This focus on users of XAI systems would increase the demand for user-centered explanations.

2.2 Trusted Computing

The goal of trusted computing is to create trust in the hardware of a system as well as the software running on that system. The Trusted Computing Group (TCG)[1] creates standards around trusted computing. Most notably they created the specifications for the Trusted Platform Module (TPM)[2].

A TPM is a trusted hardware component build into many modern systems [5]. TPMs can add IT-security functionalities to a system. The main purpose of TPMs revolves around encryption. They can generate cryptographic keys and store generated keys as well as keys provided to the TPM. These keys can encrypt and decrypt data provided to the TPM. Additionally, the TPM can encrypt keys with a Storage Root Key to store them outside the TPM.

A TPM can be used to verify that a system is in a trusted software state [11]. To achieve this it has several Platform Configuration Registers (PCRs) which can be used to store a hash of the system state. Because the number of registers is limited new states are concatenated to the current hash, hashed again and the register gets overwritten. TPMs also have a functionality called enhanced authorization. Keys can be bound to policies and can only be used if e.g. expected values are stored in the PCRs. Alternatively the PCR values can be provided to users or applications to verify the system state from outside.

The process of verifying the state of a system from outside is called remote attestation. According to Coker et. al. ”Remote attestation is the activity of making a claim about properties of a target by supplying evidence to an appraiser over a network.” [3]. This process involves two parties: The attester provides its current state and the verifier wants to verify this state [11]. There are two

[1] `https://trustedcomputinggroup.org/`
[2] `https://trustedcomputinggroup.org/resource/tpm-library-specification/`

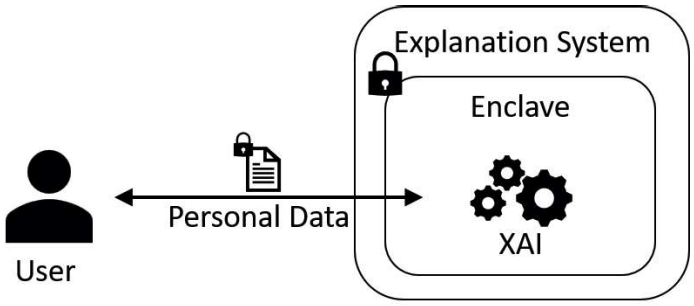
approaches to remote attestation: explicit attestation and implicit attestation. For explicit attestation with TPMs the attester provides the verifier with different values signed by the TPM including the PCR values from the TPM. The verifier checks these values by comparing them to know, trusted values. The attestation also involves a so called nonce that the verifier sends the attester at the beginning of the process to avoid replay attacks. Implicit attestation uses the enhanced authorization features of the TPM. The attester has to communicate with the verifier with a key that can only be used if the attester is in a known, trusted state. The key is bound to a policy an can therefore only be used if the system is in a trusted state. Through the key the verifier can verify that the attester can be trusted. At the beginning the verifier also sends a nonce that is then signed by the TPM and send back to verify that the system state is current.

Another trusted computing technology are Trusted Execution Environments (TEEs) [10]. A TEE is an isolated part of a processor with encrypted memory. It can be used to create enclaves in which data can be processed confidentially without other processes, even with higher privileges, being able to access the data. Examples for this technology are ARM TrustZone [1] and Intel Software Guard Extensions (Intel SGX) [6]. A similar technology is Intel Trust Domain Extensions (Intel TDX) [7] which builds on SGX. The idea there is to secure a complete virtual machine from its host. This means that for example a cloud provider can not access the VM but everyone that has access to it has to be included into the trust model.

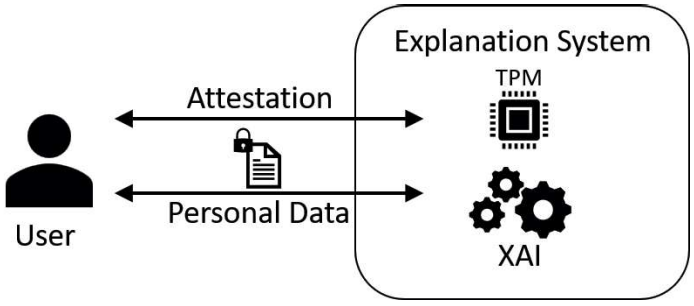
3 Explainable AI and Trusted Computing

XAI literature often mentions that XAI can be used to increase trust in AI systems [9]. However XAI is not sufficient to trust the system from an IT-security standpoint. To achieve this trusted computing can be used. There are different trusted computing technologies that are used in different contexts and give different guarantees. Trusted Execution Environments (TEEs) can create secured enclaves that are protected from the rest of the system. Trusted Platform Modules (TPMs) can be used to verify that a system is in a known, trustworthy state. Both technologies could be used to protect XAI applications.

To our knowledge there are no existing approaches that combine XAI and trusted computing. Section 3.1 presents a concept for TEEs and Section 3.2 a concept for TPMs. Figure 5.1 shows the two concepts for combining XAI and trusted computing.



(a) Calculating explanations in an enclave.



(b) Attesting the system state with a TPM.

Figure 3.1: Two concepts for combining XAI and trusted computing.

3.1 Securing XAI with TEEs

There are already some works that combine machine learning and trusted computing. One method that could be applied to XAI is origami inference proposed by Narra et. al. [8]. The proposed method is used to do privacy-preserving inference of deep neural networks using Intel SGX. First the encrypted user data that should be protected is received in the SGX enclave. The inference of the first layers of the deep neural network are calculated in a privacy-preserving way using SGX. Because the input is almost impossible to reconstruct after a

few layers, later layers can be executed as usually. This is done because TEEs have limited storage and processing capabilities. The input data is kept safe but execution time is greatly reduced compared to calculating everything in the enclave. XAI approaches also need to process personal data if the explanations are designated to end users. The methods need to evaluate the ML model, often multiple times to generate explanations. Origami inference could be used here to preserve the users privacy while maintaining an acceptable processing time. Figure 3.1(a) shows the concept of calculating explanations in enclaves without origami inference.

3.2 Securing XAI with TPMs

Another way to secure user data when generating XAI explanations is the use of TPMs. Remote attestation explained in section 2.2 can be used to verify that the system calculating the explanations is in a trustworthy state. Explicit as well as implicit attestation can be used to achieve this. This would require an attestation step to be executed before the use of the XAI application which could however be completely transparent to the user. An application such as an XAI dashboard could execute implicit attestation and the user would only be able to connect to the dashboard when the attestation succeeds. Figure 3.1(b) shows the concept of attesting the state of an explanation system with a TPM.

4 Summary

We presented two concepts for creating more trust in XAI systems by using trusted computing technologies. The first concept utilizes TEEs while the second one uses TPMs. These technologies could be especially interesting when the scope of the system involves processing personal data. In such scenarios the utilization of trusted computing could increase the users' trust in the system and make them more comfortable sharing personal data with the system.

References

[1] *ARM Security Technology Building a Secure System using TrustZone Technology*. 2008. URL: https://documentation-service.arm.com/static/5f212796500e883ab8e74531?token=.

[2] Nadia Burkart and Marco F Huber. "A survey on the explainability of supervised machine learning". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[3] George Coker et al. "Principles of remote attestation". In: *International Journal of Information Security* 10 (2011), pp. 63–81.

[4] European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[5] *Enterprise Security: Putting the TPM to Work*. 2008. URL: https://trustedcomputinggroup.org/wp-content/uploads/TPM-Applications-Whitepaper.pdf.

[6] *Intel Software Guard Extensions*. URL: https://www.intel.com/content/www/us/en/architecture-and-technology/software-guard-extensions.html.

[7] *Intel Trust Domain Extensions*. 2021. URL: https://cdrdv2-public.intel.com/690419/TDX-Whitepaper-February2022.pdf.

[8] Krishna Giri Narra et al. "Origami inference: private inference using hardware enclaves". In: *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*. IEEE. 2021, pp. 78–84.

[9] Atul Rawal et al. "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives". In: *IEEE Transactions on Artificial Intelligence* 3.6 (2021), pp. 852–866.

[10] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. “Trusted execution environment: what it is, and what it is not”. In: *2015 IEEE Trustcom/BigDataSE/Ispa*. Vol. 1. IEEE. 2015, pp. 57–64.

[11] *TCG Trusted Attestation Protocol (TAP) Information Model for TPM Families 1.2 and 2.0 and DICE Family 1.0*. 2019. URL: https://trustedcomputinggroup.org/wp-content/uploads/TNC_TAP_Information_Model_v1.00_r0.36-FINAL.pdf.