622.755 – Introduction to Cybersecurity Winter Term, 2020/21

Project Write Up

Roberto Tonino 12040367

1 Introduction

In this document I will describe why a cybersecurity framework named DeepSecure can be used to help the development of Trustworthy AIs. In Section 2 I will discuss who has given a set of guidelines to help addressing an Artificial Intelligence and why it needs to be considered trustworthy. In Section 3 I will explain how DeepSecure, a framework used to build secure and performant Deep Learning (DL) distributed algorithms, helps to achieve the result of a Trustworthy AI by providing a security-proven framework.

2 Trustworthy AI

In 2018, the European Union promoted its strategy to foster the development and the use of Artificial Intelligence. The strategy aims to increase collaboration between the State Members, to make more data available and to promote made in Europe Artificial Intelligence that is cutting-edge, ethical and secure [1]. In its document, the European Union listed the main activities that are needed to achieve this goal, which are summarized in [2, p. 4], and are:

- increasing public and private investments in AI to boost its uptake;
- preparing for socio-economic changes;
- ensuring an appropriate ethical and legal framework to strengthen European values.

The last of this three activities has led to the creation of a group composed of 52 members with different backgrounds, that are academia, civil society and industry, named the High-Level Expert Group on Artificial Intelligence (AI HLEG) [3].

One of the AI HLEG tasks is to write a document called *Ethics Guidelines for Trustworthy AI*, which consists in a set of guidelines that can be used as a support when building AI products and that ensures their reliability, security and social acceptance.

The creation of a document of this level of importance is guided by the modern necessity of data privacy and protection and by the need of protection of the people from malicious activities that can be performed through a continuously evolving technology. A commercial product must be reliable and secure to its users, not harming them in any way.

3 How DeepSecure is Trustworthy

A Trustworthy AI holds a large number of technical, social, legal and ethical features. DeepSecure, in this document, is seen as one of the many components that help to achieve this result. DeepSecure alone cannot achieve all the requirements listed in [2].

Some of the features that a Trustworthy AI is required to have are [2, p. 14]:

- being respectful of privacy;
- being robust, both from a technical and a social perspective.

One of the major problems of our times is the data privacy and protection. It is widely known that big tech companies track their users to perform some not well defined activities, and often users are unaware of this [4, 5]. With the arising of Artificial Intelligence systems and algorithms, the necessity of the data protection remains a very important aspect to achieve. The cited big tech companies are pioneers in this field, mainly because of their large investments.

One of the tasks that this companies are performing is Deep Learning. A characteristic of Deep Learning systems is the need of input data that is used to produce a result. Depending on the application, the input can be some kind of private and sensible data: for example, the task of analyzing medical photos to detect a particular disease requires as input private photos of the patients.This is certainly a problem in matter of the patients' (and, more broadly, users') data privacy and protection.

Remaining in the example situation above, the company that develops the Deep Learning medical algorithm has the need to maintain the algorithm parameters secret, because they are often the results of large research efforts and a large amount of time spent for the training of the DL algorithm.

In this situation, there is a problem on both sides: the data of the two parties needs to remain secret and not be disclosed to the public. The setting of the problem is therefore the following: there is the need of the computation of a function between two parties, but both of the parties do not want to reveal their input. In cybersecurity this is a problem addressed as Multi Party Computation (MPC). The literature provides different techniques to use in MPC; one of the oldest and most valuable techniques are Yao's Garbled Circuits.

Yao's Garbled Circuits operate on functions transformed in boolean circuits, with a maximum of two inputs. The advantage of this technique primarly consist in the fact that the total evaluation cost does not depend of the depth of the circuit but it is instead constant. DeepSecure enhances Yao's Garbled Circuits with a handful of optimizations in order to maximize performance [6, p. 3], but most importantly it creates a setting in which both inputs are maintained private and oblivious to the other party.

In terms of robustness, DeepSecure can be considered secure from a technical perspective. Given that for an MPC setting there are no IND-CPA-like security definitions, it is still possible to prove the security of an algorithm of this kind by setting a “Honest but curious” security model, in which all the parties involved in the communication follow the protocol, but some malicious ones could potentially learn more information than what it is permitted. DeepSecure achieves security in this model by using two secure cryptographic protocols: Oblivious Transfer (OT) and Yao’s Garbled Circuits [6, p. 3].

In the *Ethics Guidelines for Trustworthy AI* document we read how AI systems should “be protected against vulnerabilities that can allow them to be exploited by adversaries” [2, p. 16]. This is a crucial point because systems of this kind are very powerful and can cause a lot of harm to all the involved stakeholders. An example of this situation is the leakage of private data used by a DL system, which can be used to threaten all the people that used the system without knowing its flaws and trusting it. Frameworks like DeepSecure help to achieve the robustness requirement for a Trustworthy AI by using secure-proven cryptographic protocols and by placing the security of users and their data as a first class citizen of the solution.

4 Conclusions

In this document I described how a framework like DeepSecure can help the new arising technologies to maintain data privacy and protection, following the important guidelines made available by the European Union in the *Ethics Guidelines for Trustworthy AI* document.

Acronyms

AI Artificial Intelligence. 5-1, 5-2, 5-3

AI HLEG High-Level Expert Group on Artificial Intelligence. 5-1

DL Deep Learning. 5-1, 5-2, 5-3

EU European Union. 5-1, 5-3

IND-CPA Indistinguishability under Chosen Plaintext Attacks. 5-2

MPC Multi Party Computation. 5-2

OT Oblivious Transfer. 5-3

References

[1] European Commission. *Member States and Commission to work together to boost artificial intelligence "made in Europe"*. European Commission, 2018. URL: `https://ec.europa.eu/commission/presscorner/detail/en/IP_18_6689`.

[2] High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI.* Apr. 2019. URL: `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

[3] European Commission. *High-Level Expert Group on Artificial Intelligence.* European Commission, June 2018. URL: `https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence`.

[4] Alex Kantrowitz. *Here's How Facebook Tracks You When You're Not On Facebook.* BuzzFeed News, Apr. 2018. URL: `https://www.buzzfeednews.com/article/alexkantrowitz/heres-how-facebook-tracks-you-when-youre-not-on-facebook`.

[5] Anthony Cuthbertson. *Google's private browsing incognito mode leaks way more personal data than you might think.* The Independent, Aug. 2018. URL: `https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-chrome-incognito-mode-personal-data-private-browser-a8502386.html` (visited on 02/09/2021).

[6] Bita Darvish Rouhani, M. Sadegh Riazi, and Farinaz Koushanfar. *DeepSecure: Scalable provably-secure deep learning.* 2017. URL: `https://arxiv.org/abs/1705.08963`.