Co-funded by the
Erasmus+ Programme
of the European Union

EUROPEAN UNIVERSITY FOR CUSTOMISED EDUCATION

STUDY GUIDE

Foundations of Trustworthy Machine Learning

Organised by

Université Polytechnique Hauts-de-France

1. IDENTIFYING DATA.

· Course Name.	*Foundations of Trustworthy Machine Learning*
· Coordinating University.	*UPHF*
· Partner Universities Involved.	*None*
· Course Field(s).	*Computer Science- AI*
· Related Study Programme.	
· ISCED Code.	*None*
· SDG.	*Goal 9: Build resilient infrastructure, promote sustainable industrialization and foster innovation* *Goal 11: Make cities inclusive, safe, resilient and sustainable*
· Study Level.	*Master ,PhD*

· Number of ECTS credits allocated.	*4 ECTS*
· Mode of Delivery.	*Online self-study*
· Language of Instruction.	*English*
· Course Dates.	*February-June*
· Schedule of the course.	*Taught hours: 36h – overall study load 140 hours*
· Key Words.	*Machine Learning, AI, Security, Privacy, Trustworthiness, Ethical AI*
· Catchy Phrase.	*Towards Trustworthy AI:* *Safely Navigating the AI landscape from a Security and Privacy perspective*

· Prerequisites and co-requisites.	- *We assume students have a foundational knowledge of AI/ML from their UG studies.* - *The study levels this course is available for Master's and PhD students* - *Required linguistic skills: English*
· Number of EUNICE students that can attend the Course.	*30 students*
· Course inscription procedure(s).	*EUNICE website*

2. CONTACT DETAILS.

· Department.	
· Name of Lecturer.	*Ihsen Alouani*
· E-mail.	*Ihsen.alouani@uphf.fr*
· Other Lecturers.	*N/A*

3. COURSE CONTENT.

This course will present an in-depth exploration of trustworthiness of AI/ML from a security and privacy perspective. The course will be research-led, incorporating recent work in the intersection between AI and Cybersecurity.

We will first introduce AI-powered cybersecurity applications like malware detection and Intrusion detection as a case study to ground the discussion of topics throughout the module.

The second part of the course will focus on the security of AI/ML models. We will explore attack types and defences specifically targeting ML models.

4. LEARNING OUTCOMES.

Successful students will be able to:

1.Understand and apply concepts and algorithms of machine learning to solve cybersecurity specific problems.

2.Implement, evaluate, and compare machine learning algorithms that are privacy-preserving and robust to attacks

3. Understand and apply concepts related to the security of AI Models, including attacks and defence methods.

5. OBJECTIVES.

By the end of the module, the students should be able to:

- *Design, train and deploy ML models for Cybersecurity purposes*
- *Design, train and deploy privacy-preserving ML models*
- *Design, train and deploy robust ML models*
- *Assess the security and privacy of a trained ML model*

6. COURSE ORGANISATION.

UNITS	
1.	*Foundations of AI* *Foundations of Cybersecurity*
2.	*AI based Cybersecurity* • *Intrusion Detection Malware Detection model (CNN opcodes / feature-based model)* • *Cyber-security specific AI concepts – Implementation pitfalls, concept-drift, bias, dataset imbalance, model evaluation*
3.	*Security of AI Models* • *Introduction to AI Security – CIA, Threat Models, Attacker Knowledge, Attacker Objectives, Training VS Inference, Types of Attacks* • *Attacks - Evasion, poisoning, backdoor-attacks,* • *Defences - Adversarial Training, Out-of-Distribution Detection*
4.	*Privacy-preserving AI Models* • *Mathematical bases of Differential Privacy* • *Attacks - model inversion, model stealing, membership inference* • *Defences – differential privacy*

LEARNING RESOURCES AND TOOLS.

- *Academic papers*
- *Practical exercices on Colab*
- *Pytorch framework*

PLANNED LEARNING ACTIVITIES AND TEACHING METHODS.

The main learning activities for this course: lectures, practical tutorials, reading

7. ASSESSMENT METHODS, CRITERIA AND PERIOD.

The assessment will be through a project.

OBSERVATIONS.

8. BIBLIOGRAPHY AND TEACHING MATERIALS.

Adversarial attacks: *https://arxiv.org/pdf/1608.04644.pdf*
Defences: *https://arxiv.org/abs/2102.01356*
Privacy attacks: *https://arxiv.org/pdf/1610.05820.pdf*
Differential Privacy: *https://arxiv.org/abs/1607.00133*
(videos, lecture notes, slides and tutorials will be provided)

Co-funded by the
Erasmus+ Programme
of the European Union