3IF1135 - Sécurité de l'IA

Responsables : **Pierre-François GIMENEZ**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE RENNES**

Nombre d'heures d'études élèves (HEE) : **35**

Nombre d'heures présentielles d'enseignement (HPE) : **15**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Ce module est un cours de découverte de l'utilisation de l'IA pour la Cybersécurité ainsi que les attaques que l'on peut mener contre les méthodes basées sur de l'IA. Le cours ne cherche pas à être exaustif sur l'ensemble des méthodes d'IA: au contraire, il se focalise sur les attaques contre ces systèmes, notamment celles qui ciblent les réseaux de neurones et dégradent les performances de classification, ce qui pourrait impacter les systèmes de détection d'intrusion ou d'analyse de malware. De plus, le cours ambitionne de décrire les méthodes de l'état de l'art qui implémentent des mécanismes de défense dans les réseaux de neurones. Cette partie du cours sera abordée au travers de ces aspects recherche par des chercheurs du domaine. Une partie pratique permettra aux étudiants d'implémenter des attaques et d'évaluer l'efficacité des défenses.

Prérequis :

- Programmation en Python (cours SIP)

Plan détaillé du cours (contenu) :

- Les réseaux de neurones et de leur apprentissage (CM, 3h)
- Classification d'image par réseau de neurones convolutifs (TP, 3h)
- Attaques contre les réseaux de neurones (CM, 3h)
- Défenses contre ces attaques (CM, 3h)
- Attaque et défense d'un classifieur (TP, 3h)

Déroulement, organisation du cours :

Cours magistraux (9h)
TP (6h)

Organisation de l'évaluation :

Évaluation du second TP

Moyens :

Les TP utiliseront la bibliothèque pytorch (open-source).

Acquis d'apprentissage visés dans le cours :

- Développer un système de classification basé sur un réseau de neurones
- Evaluer l'impact d'attaques contre un classifieur
- Défendre un réseau de neurones contre ces attaques

Description des compétences acquises à l'issue du cours :

C2: develop a classification system based on a neural network
C3: apply state-of-the-art attack and defense methods

Bibliographie :

- Deep Learning, Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press, 2016
- Adversarial Machine Learning: Attack Surfaces, Defence Mechanisms, Learning Theories in Artificial Intelligence, Aneesh Sreevallabh Chivukula, Xinghao Yang, Bo Liu, Wei Liu, Wanlei Zhou, 2023