



HAL
open science

PICTURE: Physical and Intrinsic Security of Embedded Neural Networks

Pierre-Alain Moellic

► **To cite this version:**

Pierre-Alain Moellic. PICTURE: Physical and Intrinsic Security of Embedded Neural Networks. WISG 2022 - 14ème Workshop interdisciplinaire sur la sécurité globale, Jan 2022, Paris (Virtual event), France. . ⟨hal-04628373⟩

HAL Id: hal-04628373

<https://hal.science/hal-04628373v1>

Submitted on 28 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

PICTURE

Physical and Intrinsic Security of Embedded Neural Networks

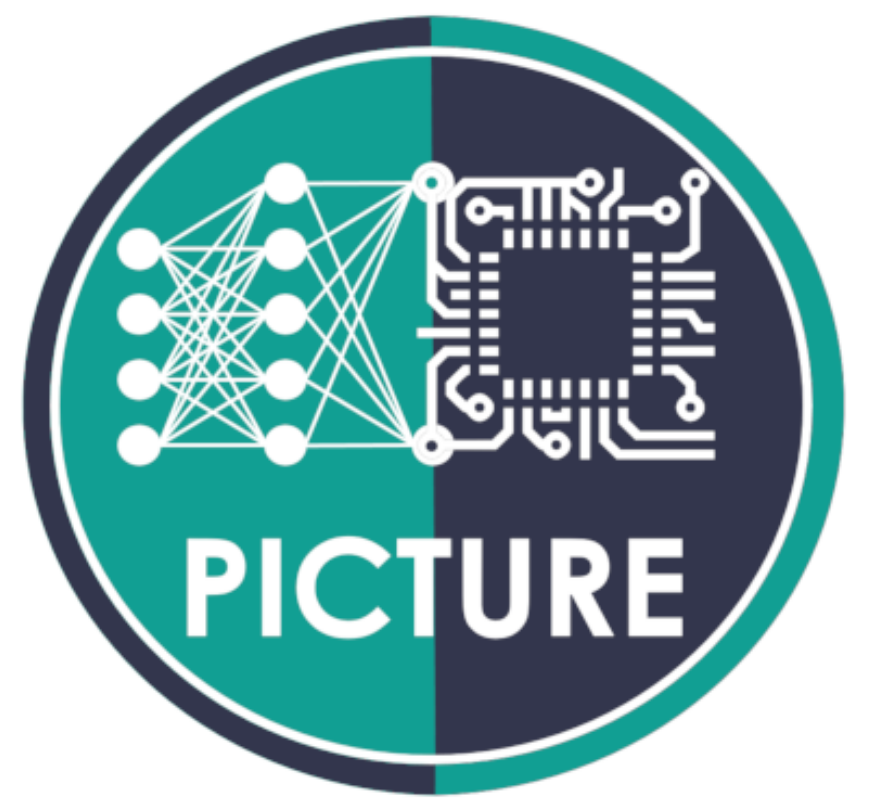
anr[©] agence nationale de la recherche

Appel : AAPG (CE39)

Année : 2020

Instrument : PRCE

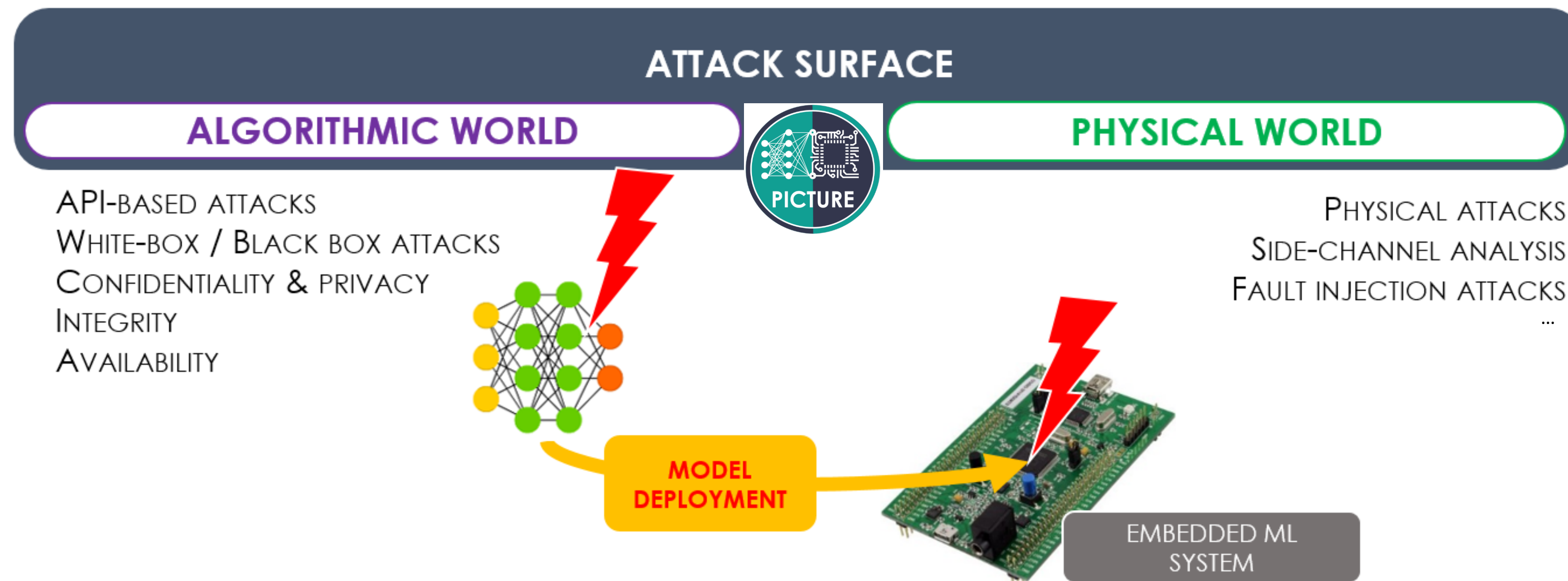
Contact : Pierre-Alain MOELLIC



COORDINATEUR : Pierre-Alain MOELLIC (CEA LETI)

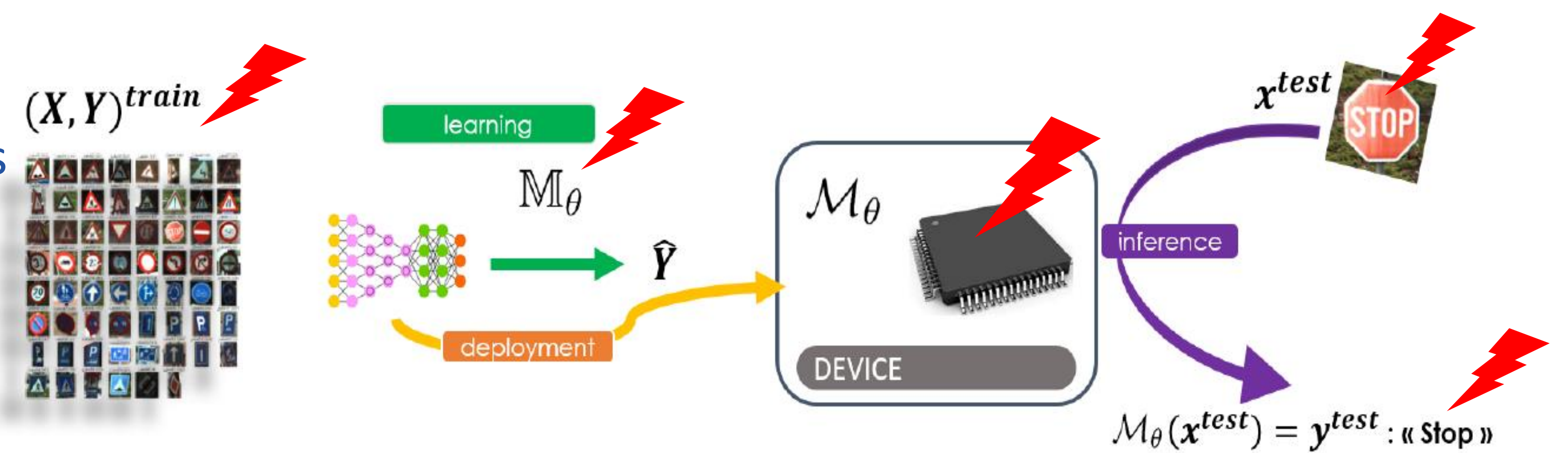
PARTENAIRES :
CEA-LETI, Mines Saint-Etienne, IDEMIA,
STMICROELECTRONICS

PICTURE s'intéresse à la sécurité des réseaux de neurones embarqués (logiciels) en considérant une surface d'attaque regroupant les menaces algorithmiques (*adversarial examples, model extraction...*) et physiques (*side-channel, fault injection*) contre l'intégrité, la confidentialité et la disponibilité des modèles.



CONTEXTE ET OBJECTIFS

- Déploiement massif des modèles de Machine Learning
- Données / Tâches / Plateformes HW critiques
- Le ML Pipeline traditionnel est menacé à tous les étages
- Etat de l'art conséquent des attaques algorithmiques: *Adversarial Examples, Poisoning Attacks, Model Extraction...*



Nos objectifs

1. Démontrer la criticité d'attaques combinant failles théoriques et physiques
2. Evaluer des contre-mesures physiques combinées à l'état de l'art des défenses algorithmiques et proposer des nouvelles défenses.
3. Disséminer des « bonnes pratiques » et suivre les actions de régulation/standardisation/certification.

MÉTHODOLOGIE ET RÉSULTATS

Méthodologie

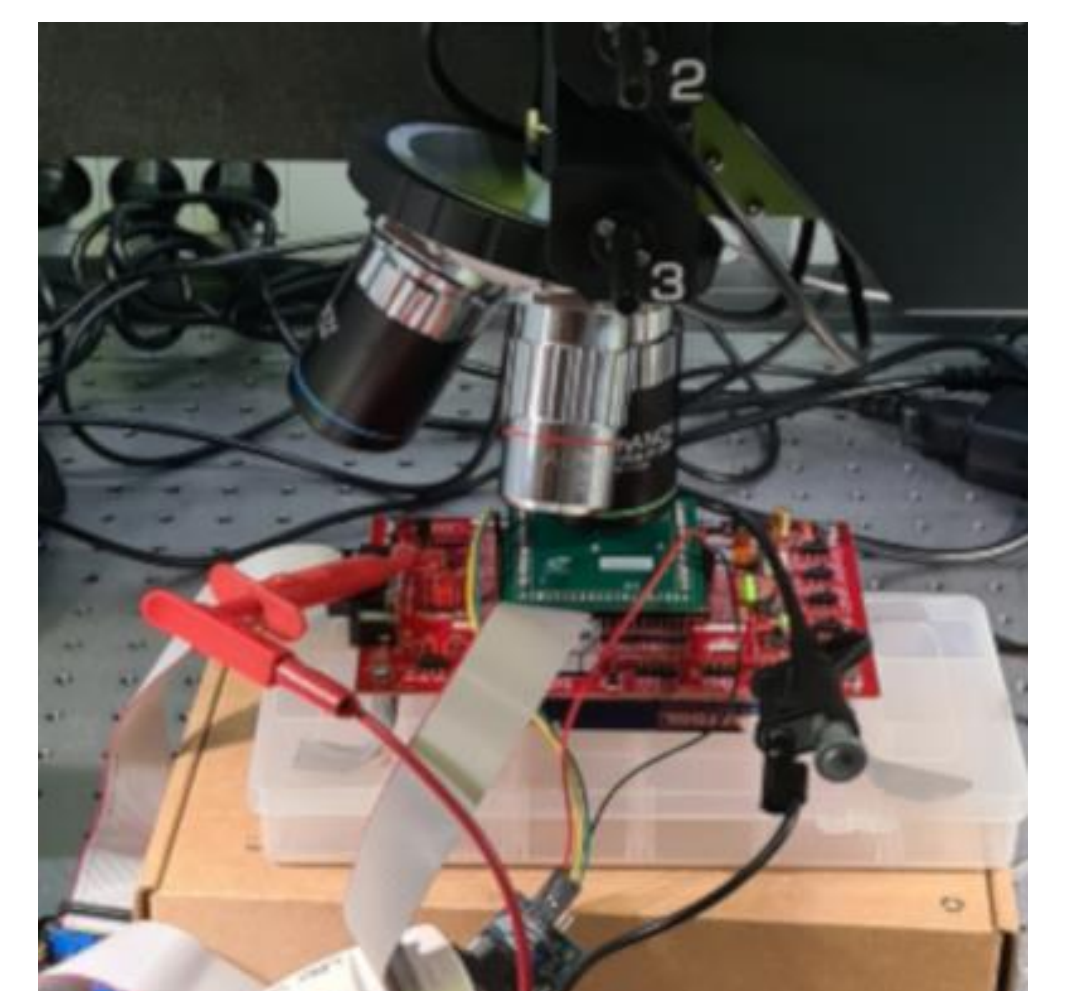
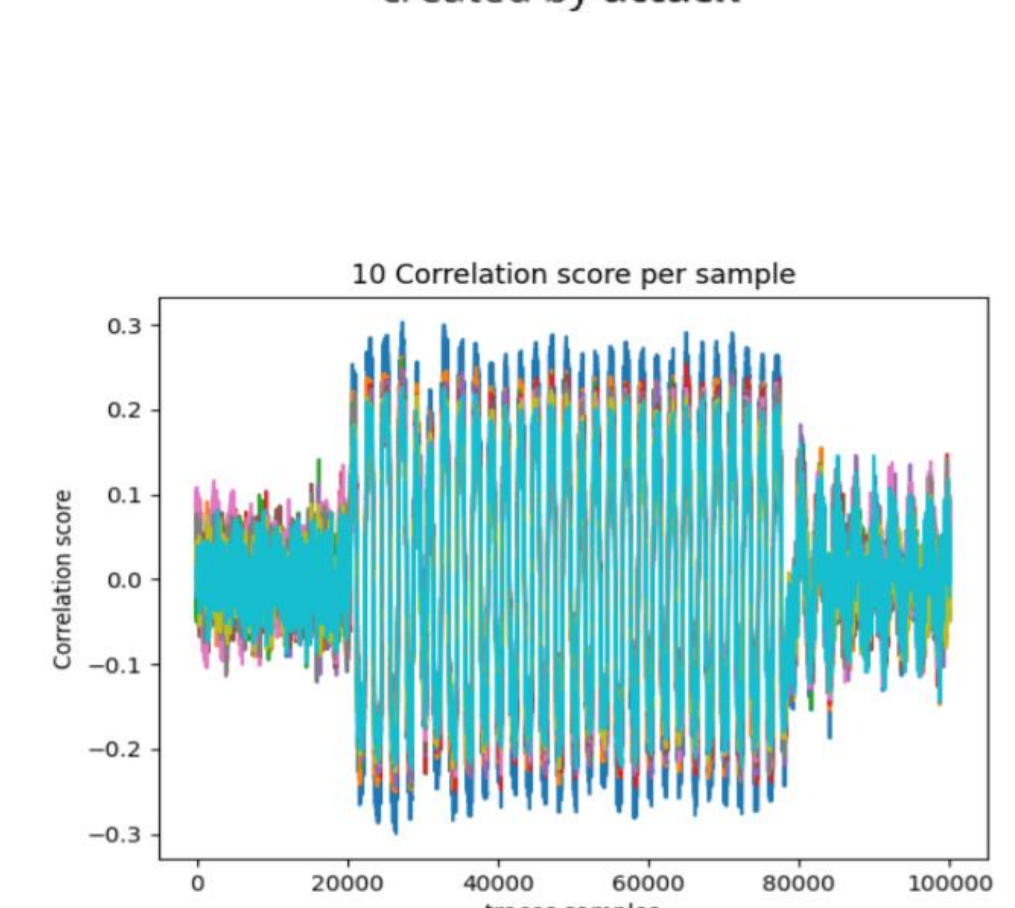
- Considérer une surface d'attaque globale et analyser conjointement failles physiques et algorithmiques:
 - ✓ *Fault injection analysis + adversarial perturbation*
 - ✓ *Side-channel analysis + model extraction*
- Analyser des modèles et plateformes réelles sur des cas d'usage critiques:
 - ✓ *Face Recognition systems*
 - ✓ *IoT*
- Méthodologie incrémentale WP Attaque ↔ WP Défense + Evaluation

Dissémination

- PUBLIQUE "State-of-the-art: Attacks & Threat Models"
- 3 Publications 2021 (IEEE IJCNN, IEEE World Forum IoT)



(small) adversarial perturbation created by attack



<https://picture-anr.cea.fr>

@AnrPicture

25 et 26
JANVIER

2022

wisg²²
WORKSHOP INTERDISCIPLINAIRE SUR LA SÉCURITÉ GLOBALE

UNIVERSITÉ DE BORDEAUX