# Attaquer et Défendre les Modèles IA

## Atelier pratique d'IA adversariale

Adediran Abdel Farid Yessoufou, PhD

Univ Rennes 1

January 26, 2026

# Plan

1. Introduction

2. Cas pratique : la détection de spam

3. Conclusion

# Plan

# Qu'est-ce que l'IA adversariale ?

L'**Intelligence Artificielle adversariale** étudie la manière dont les systèmes d'IA peuvent être **trompés, manipulés ou contournés** par des entrées spécialement conçues.

# Qu'est-ce que l'IA adversariale ?

L'**Intelligence Artificielle adversariale** étudie la manière dont les systèmes d'IA peuvent être **trompés, manipulés ou contournés** par des entrées spécialement conçues.

Dans un contexte adversarial :

- L'attaquant ne modifie pas le modèle
- Il modifie uniquement les **données d'entrée**
- Le but est de provoquer une **mauvaise prédiction**

# Qu'est-ce que l'IA adversariale ?

L'**Intelligence Artificielle adversariale** étudie la manière dont les systèmes d'IA peuvent être **trompés, manipulés ou contournés** par des entrées spécialement conçues.

Dans un contexte adversarial :

- L'attaquant ne modifie pas le modèle
- Il modifie uniquement les **données d'entrée**
- Le but est de provoquer une **mauvaise prédiction**

Exemples :

- Tromper la reconnaissance faciale
- Contourner un filtre anti-spam
- Piéger un véhicule autonome

# Example

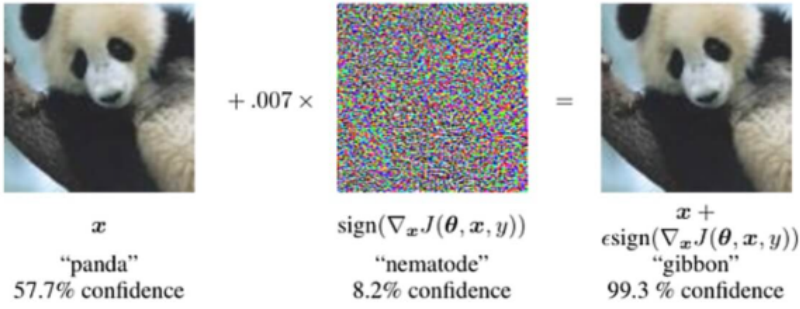


Figure: Goodfellow et al. (2014)

# Pourquoi les modèles d'IA sont vulnérables ?

Les modèles d'IA ne comprennent pas le sens réel des messages. Ils apprennent uniquement des corrélations statistiques.

Ainsi :

- De petits changements dans l'entrée peuvent produire de grands effets
- Les modèles reconnaissent des motifs, pas des intentions
- Les attaquants exploitent cet écart

# Pourquoi les modèles d'IA sont vulnérables ?

Les modèles d'IA ne comprennent pas le sens réel des messages. Ils apprennent uniquement des corrélations statistiques.

Ainsi :

- De petits changements dans l'entrée peuvent produire de grands effets
- Les modèles reconnaissent des motifs, pas des intentions
- Les attaquants exploitent cet écart

Cela crée un nouveau risque de cybersécurité : **les attaques adversariales sur l'IA**.

# Plan

# Cas pratique : la détection de spam

Nous étudions un système réel : un **classifieur de spam**.

Objectif du système :

- Classer les messages en :
  - Ham (légitime)
  - Spam (malveillant)

# Cas pratique : la détection de spam

Nous étudions un système réel : un **classifieur de spam**.

Objectif du système :

- Classer les messages en :
  - Ham (légitime)
  - Spam (malveillant)

Objectif de l'attaquant :

- Envoyer un message malveillant
- Le faire classer comme **légitime**

# Lien avec les réseaux bayésiens

Notre classifieur de spam repose sur un modèle **bayésien naïf**.

Il calcule :

$$P(\text{Spam} \mid \text{mots du message})$$

Chaque mot contribue indépendamment à la décision.

# Spam Detection

# Plan



# Conclusion

- Les modèles d'IA peuvent être trompés sans être modifiés
- Les attaques adversariales exploitent les statistiques du modèle, pas des bugs

Mais nous avons aussi vu que :

- Un bon prétraitement du texte
- L'utilisation des n-grams
- Et l'entraînement adversarial

**permettent de construire des modèles d'IA beaucoup plus robustes.**