# TrustFake

## Trustworthy AI for Social Media Deepfake Detection

Project n°2

# 01

## Context

### Deepfakes go viral

AI-generated faces, voices and videos now spread on social media at the speed of news — eroding public trust in what we see.

### Detectors break in the wild

Compression, resizing, re-encoding and adversarial perturbations degrade detectors that look perfect in the lab.

### Confidence can be hacked

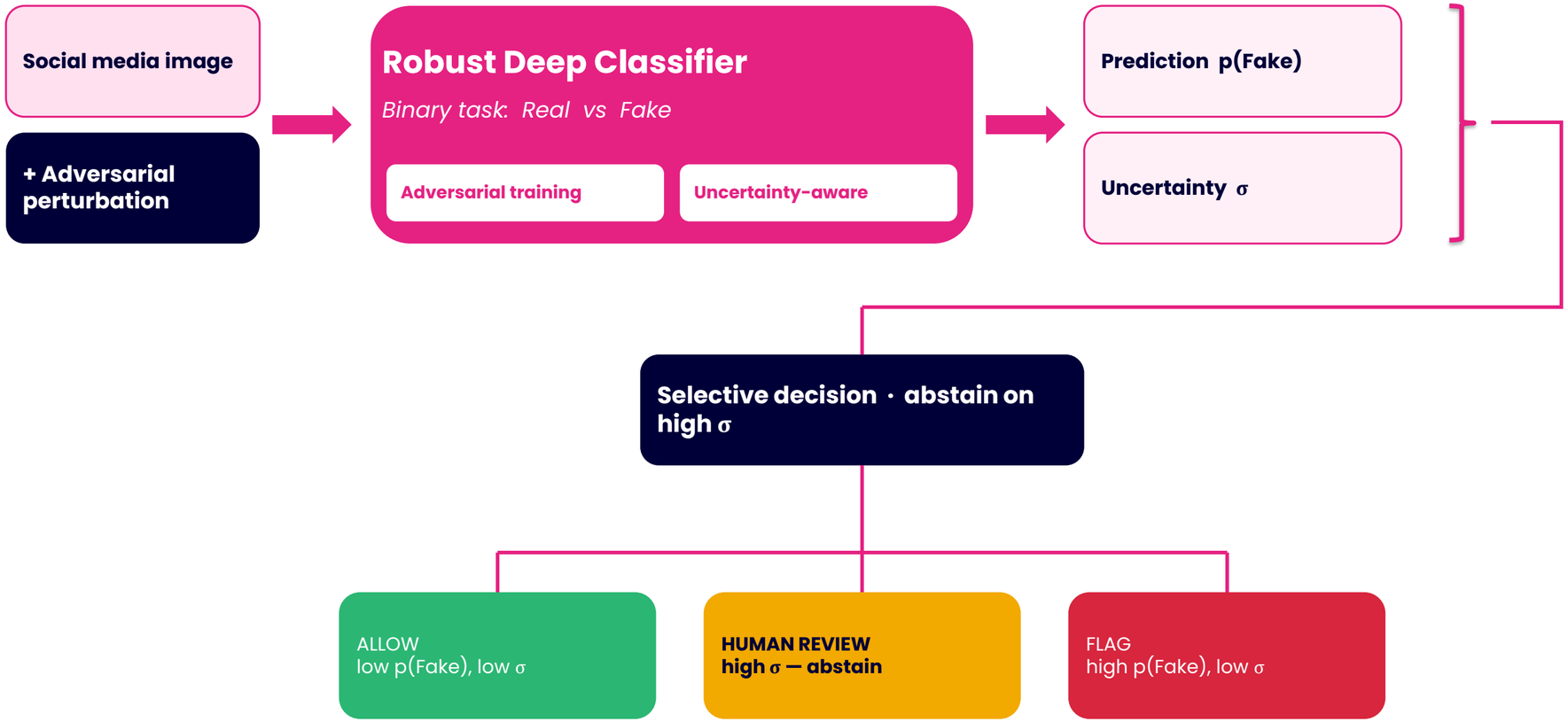
Moderation pipelines trust model confidence to flag content. Attackers can inflate it — slipping fakes past human review.

**TrustFake builds detectors that stay honest, robust under perturbation, calibrated under attack, safe for human-in-the-loop moderation.**

TRAIL
TRUSTED AI LABS

# 02

## Project objective




TRAIL
TRUSTED AI LABS

# 03

# Methodology

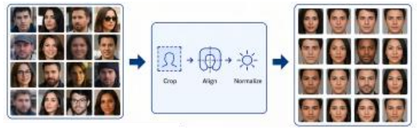
TRAIL
TRUSTED AI LABS

# TrustFake: Trustworthy AI for Social Media Image Deepfake Detection

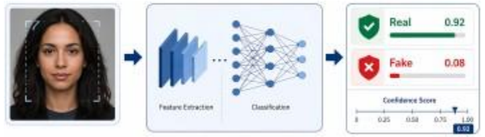
*From data to trustworthy decisions: detect, stress-test, harden, and act.*

## WP1 Deepfake Detection Framework

### 1. Data Pipeline



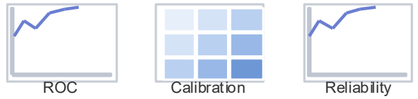

### 2. Task Pipeline




Baseline Detection Model

### 3. Evaluation Pipeline

Evaluate with robust metrics and calibration analysis




**D1. Baseline framework and evaluation protocol**

## WP2 Adversarial Attack for Predictive and Confidence Manipulation

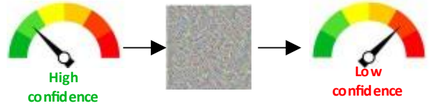
### 1. Prediction-targeted attacks

Generate perturbations to flip the model prediction (real ↔ fake)

Real | Fake

### 2. Confidence/uncertainty-targeted attacks

Manipulate model confidence and uncertainty without necessarily changing the prediction



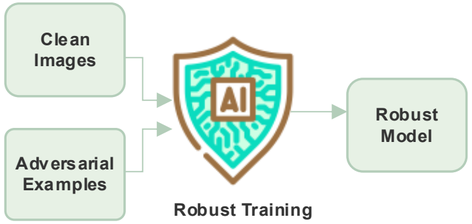

Analyze vulnerabilities across models, attacks and metrics

**D2. Attack package**

## WP3 Robust Adversarial Training for Reliable Deepfake Detection

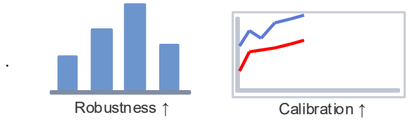
### 1. Defense Pipeline

Implement robust training strategies centered on adversarial training and TRADES




### 2. Comparative Robustness Evaluation

Compare robustness and reliability against attacks

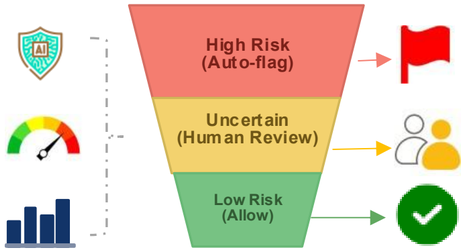



**D3. Defense package**

## WP4 Selective Moderation and Demonstrator

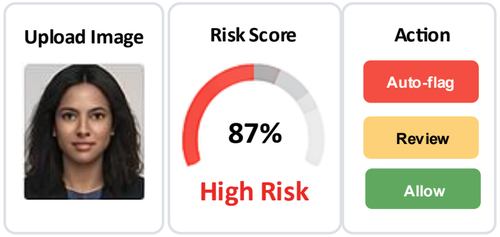
### 1. Selective Moderation Layer

Risk-aware decision making




### 2. End-to-end Demonstrator

Interactive web interface for analysis and moderation




**D4. Selective moderation prototype**

## PROJECT OBJECTIVES & IMPACT

**Detect**
Build a strong baseline for deepfake detection.

**Stress-test**
Identify and analyze vulnerabilities via adversarial attacks.

**Harden**
Improve robustness and reliability under attacks and uncertainty.

**Act**
Enable risk-aware decisions with human oversight.

**Impact**
Safer social media, greater trust, and responsible AI.

# 04

# Concrete implementation

01

DATASET

## SID-Set

300k images · real, fully synthetic, tampered. Realistic social-media content (OpenImages, FLUX, latent-diffusion edits).

Binary: Real vs Fake | Balanced subset | HuggingFace

02

BASELINE

## ResNet detector

Fixed backbone for fair, controlled comparison across all training regimes.

ResNet-18 | ResNet-50

03

UNCERTAINTY

## Confidence beyond softmax

Estimate per-image uncertainty to drive selective abstention in moderation.

MC-Dropout | Deep Ensembles

04

ROBUSTNESS

## Attack & defend

Stress-test both predictions and confidence, then harden with robust training.

PGD · AutoAttack | Uncertainty attacks | Adv. training · TRADES

TRAIL
TRUSTED AI LABS

# 05

# Founding team and expertise sought

## Founding team

**Nicolas Sournac**

**Team leader - Multitel**

Multitel
INNOVATION CENTRE

Trustworthy AI · robustness · reliability and safety of deep neural networks for computer vision

**Ahmed Baha Ben Jmaa**

**Team leader - Multitel**

Multitel
INNOVATION CENTRE

2D/3D computer vision · multimodal ML · robust and reliable vision systems

## Expertise sought

| | |
|---|---|
| **Deep learning & CV** | PyTorch, backbone training, image preprocessing |
| **Adversarial ML** | FGSM/PGD/AutoAttack, attack evaluation |
| **Uncertainty** | calibration, MC-Dropout, ensembles, abstention |
| **Prototype** | moderation UI, metrics dashboard, integration |

TRAIL
TRUSTED AI LABS

Unil.

civis

# TRAIL

TRUSTED AI LABS

## THANK YOU FOR YOUR ATTENTION !

WWW.TRAIL.AC
CONTACT@TRAIL.AC

ULB Université Libre de Bruxelles · UMONS University of Mons · UCLouvain · Liège université · Université de Namur · LIEU Liaison Entreprises-Universités · Multitel Innovation Centre · cetic · Cenaero · sirris · S3 · AI4Belgium · digital wallonia .be

Cofinancé par l'Union européenne