TReC 2026 Project Proposal
TrustFake
Trustworthy AI for Social Media Image Deepfake Detection

Domain of Application	Media, Culture & Digital Humanities
Scientific Theme	Trustworthy AI
Team Leaders	Nicolas Sournac; Ahmed Baha Ben Jmaa

1. Project Title

TrustFake: Trustworthy AI for Social Media Image Deepfake Detection

2. Profile of the Team Leader(s) & Expected Team Composition

Team Leaders.

- **Nicolas Sournac** is a researcher in the Fundamental AI & Optimization Group of Multitel's Artificial Intelligence Department. His research focuses on trustworthy AI, with particular expertise in the robustness, reliability, and safety of deep neural networks, especially for computer vision.
- **Ahmed Baha Ben Jmaa** is a researcher in the Computer Vision Group of Multitel's Artificial Intelligence Department. His research focuses on 2D/3D computer vision and multimodal machine learning for real-world applications. More recently, his work has expanded toward trustworthy AI, with a particular interest in robust and reliable computer vision systems.

Expected Team Composition. We are seeking participants with strong skills in deep learning, computer vision, and Python/PyTorch. Prior experience in adversarial machine learning, uncertainty estimation, or generative AI would be an asset, but is not required. We particularly value motivated candidates who are interested in trustworthy AI and socially relevant media applications, and who can contribute to data and evaluation pipelines, adversarial attack modeling, robust training, uncertainty-aware decision strategies, or prototype development.

3. Abstract

Recent advances in generative AI have made synthetic and manipulated images increasingly realistic, raising major challenges for media reliability and deepfake detection on social media platforms. While existing detectors can perform well in controlled settings, their reliability often degrades under realistic perturbations such as compression, resizing, re-encoding, and adversarial manipulation. This limitation is particularly critical in moderation pipelines, where decisions depend not only on predicted labels but also on model confidence to determine whether content should be automatically accepted, flagged, or escalated to human review.

In this context, **TrustFake** proposes a trustworthy AI framework for social media image deepfake detection, centered on robustness, confidence reliability, and selective decision-making. The project will develop a modular pipeline covering deepfake detection, adversarial attack modeling on both predictions and confidence estimates, and robust adversarial training strategies for reliable detection. It will further implement a selective moderation layer based on uncertainty-aware abstention, using a relevant large-scale dataset, to support the development of more reliable and practically deployable deepfake detection tools for online media moderation.

4. Background Information & Problem Statement

Deepfake detection systems are increasingly integrated into social media moderation pipelines, where decisions are not based on predictions alone but also on model confidence [1, 2]. In practice, confidence thresholds are used to determine whether content is automatically removed, flagged for human review, or left untouched. This makes uncertainty estimation a critical component of system reliability. However, this reliance introduces a subtle but serious vulnerability: an attacker can manipulate a model to become over-confident on incorrect or suspicious content [3–7]. By artificially inflating confidence scores, malicious deepfakes that should trigger human review can instead bypass scrutiny and be automatically accepted or insufficiently checked. This undermines the safeguard role of uncertainty, effectively turning a risk-aware system into an over-trusting one.

This challenge is particularly important in social media image deepfake detection, where images are subject to realistic transformations such as compression, resizing, re-encoding, re-digitization, and transmission artifacts, and where detectors may also face previously unseen manipulations or deliberate adversarial perturbations. As a result, methods that perform well in controlled settings may become unreliable in deployment conditions that are much closer to real moderation scenarios.

TrustFake addresses this problem at the intersection of deepfake detection, adversarial robustness, and uncertainty-aware trustworthy AI. It builds on foundational work in robustness evaluation, including white-box attacks such as FGSM, PGD, and AutoAttack [8–10], and on robust learning methods such as adversarial training and TRADES [11, 12]. It also draws on uncertainty estimation methods such as Monte Carlo Dropout and deep ensembles [13, 14], which provide confidence estimates beyond standard softmax scores.

Despite these advances, an important research gap remains: current approaches rarely study, within a unified setting, how realistic perturbations, adversarial attacks, and confidence-based moderation decisions interact in social media image detection. In particular, the relationship between robustness at the prediction level and reliability at the confidence level remains insufficiently understood. TrustFake is therefore motivated by the need to move beyond closed-world accuracy toward a more realistic notion of trustworthy detection, where a model must remain effective under perturbations while preserving meaningful confidence estimates for selective decision-making.

5. Project Objectives & Concrete Implementation

The objective of TrustFake is to develop a trustworthy AI framework for social media image deepfake detection, with a focus on robustness, confidence reliability, and selective decision-making. Rather than only optimizing detection accuracy, the project will study whether detectors can remain reliable under realistic perturbations and adversarial manipulation while preserving meaningful confidence estimates for human-in-the-loop moderation.

Figure 1 provides a visual overview of the proposed TrustFake workflow, illustrating the progression from baseline deepfake detection to adversarial stress-testing, robust training, and uncertainty-aware selective moderation.

The project is structured around four scientific challenges, each mapped to a concrete objective, work package, and deliverable. This structure ensures that the project remains scientifically grounded and feasible within the two-week camp. The resulting mapping is summarized in Table 1, which provides an overview of the logical progression from scientific challenges to concrete project outputs.

WP1 - Deepfake Detection Framework

WP1 establishes the baseline detection and evaluation framework. It includes a data pipeline for loading, preprocessing, and splitting SID-Set (see Section 6); a task pipeline defining the core detection setting; and an evaluation pipeline for measuring predictive performance, uncertainty quality,

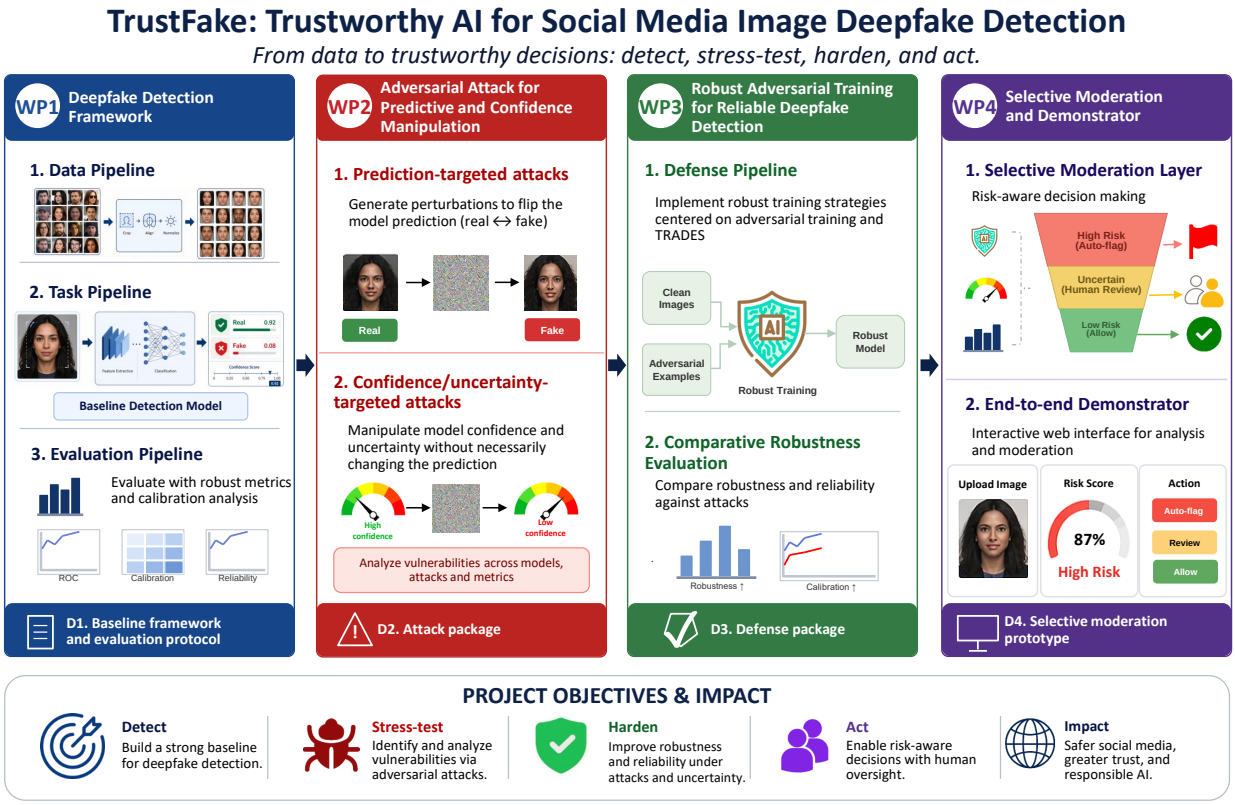

Figure 1: Overview of the TrustFake project workflow.

and selective decision behavior. The core task will focus on binary real-vs-fake detection to ensure feasibility, while the three-class setting distinguishing real, synthetic, and tampered images will be considered as an optional extension. The evaluation pipeline will include accuracy, F1-score, AUROC, risk-coverage curves, AURC, and AUGRC. **Deliverable:** D1 - a baseline detection framework with data, task, and evaluation pipelines.

WP2 - Adversarial Attacks for Predictive and Confidence Manipulation

WP2 develops the adversarial attack component of the project. It will cover prediction-targeted attacks, which aim to alter the detector output, and confidence/uncertainty-targeted attacks, which aim to manipulate the detector confidence estimates. This distinction is central to the project: in moderation pipelines, a model can fail not only by predicting the wrong label, but also by being over-confident when wrong. The attack module will support white-box attacks such as FGSM and PGD, with stronger evaluation such as AutoAttack considered where feasible. It will also include one focused family of confidence-targeted attacks to assess whether uncertainty estimates remain informative under adversarial manipulation. **Deliverable:** D2 - an attack package including prediction-targeted attacks, confidence-targeted attacks, and a vulnerability assessment.

WP3 - Robust Adversarial Training for Reliable Deepfake Detection

WP3 focuses on robust learning and defense. It will implement a defense pipeline centered on robust adversarial training strategies, primarily standard adversarial training and TRADES. The goal is to assess whether robust training can improve adversarial resilience without degrading confidence reliability or selective decision behavior. To ensure a controlled comparison, experiments will use a fixed backbone, fixed data protocol, fixed evaluation pipeline, and comparable training budgets. The resulting models will be evaluated under clean, realistic, and adversarial conditions, with particular

Table 1: Mapping between scientific challenges, research objectives, work packages, and deliverables.

Scientific challenge	Objective	Work package	Deliverable
C1. Reliable deepfake detection under realistic social-media conditions.	O1. Build a baseline deepfake detection framework for realistic evaluation.	**WP1 - Deepfake Detection Framework:** data pipeline, task pipeline, evaluation pipeline.	**D1.** Baseline detection framework.
C2. Manipulation of predictions and confidence estimates.	O2. Model adversarial attacks targeting both detector outputs and confidence estimates.	**WP2 - Adversarial Attacks for Predictive and Confidence Manipulation:** prediction-targeted and confidence-targeted attacks.	**D2.** Attack package with vulnerability assessment.
C3. Robustness without loss of confidence reliability.	O3. Study robust adversarial training for reliable detection.	**WP3 - Robust Adversarial Training for Reliable Deepfake Detection:** adversarial training, TRADES, comparative evaluation.	**D3.** Defense package and comparative results.
C4. Translation into moderation decisions.	O4. Implement an uncertainty-aware selective moderation workflow.	**WP4 - Selective Moderation Layer and Demonstrator:** abstention policy and end-to-end prototype.	**D4.** Selective moderation prototype.

attention to the trade-off between detection performance, adversarial robustness, uncertainty quality, and abstention reliability. **Deliverable:** D3 - a defense package with robust training implementations and comparative results.

WP4 - Selective Moderation Layer and Demonstrator

WP4 translates model outputs into a practical moderation workflow. It will implement a selective moderation layer based on uncertainty-aware abstention, where predictions can be automatically accepted, flagged, or escalated to human review depending on confidence and uncertainty thresholds. This layer will be evaluated using deployment-oriented indicators such as coverage, residual risk, and review rate. The project will integrate the main components into a lightweight demonstrator showing the input image, predicted label, confidence or uncertainty score, and resulting moderation decision. **Deliverable:** D4 - a selective moderation prototype demonstrating uncertainty-aware human-in-the-loop decision support.

6. Project Dataset

The project will primarily rely on **SID-Set**, introduced in *SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model* [1]. SID-Set is a large-scale benchmark designed for deepfake detection in realistic social media settings, addressing limitations of earlier datasets in terms of diversity, realism, and annotation quality.

SID-Set contains approximately 300,000 images, evenly split between real, fully synthetic, and tampered images. This structure supports both binary detection and finer-grained multi-class classification. For the core camp implementation, TrustFake will focus on binary real-vs-fake detection to maximize feasibility, while the real/synthetic/tampered setting will be treated as an optional extension.

The dataset is well suited for this project because it covers realistic and diverse content. Real images are sourced from OpenImagesV7 [15], synthetic images are generated using FLUX based on prompts and image-caption resources such as Flickr30k and COCO [16–18], and tampered images are produced using latent diffusion models to modify specific objects in COCO images [19]. The dataset includes image data, labels, and segmentation masks for manipulated regions. Although localization will not be the core objective of the two-week project, the availability of masks provides a natural direction for future extensions.

Dataset access: `https://huggingface.co/datasets/saberzl/SID_Set`. The expected data format includes image files, class labels (real, synthetic, tampered), and manipulation masks when available. The project will initially use a curated, balanced subset for rapid iteration and will scale up experiments as time and compute resources allow.

The dataset is publicly accessible, which reduces data-access risk and supports immediate experimentation during the camp. The project does not require the collection of new personal data, and no NDA-dependent dataset access is planned. A curated balanced subset will be prepared before the camp to ensure rapid prototyping, while larger-scale experiments will be run only if time and compute resources allow.

7. Detailed Work Plan

The project is organized as a focused two-week implementation plan, with a short pre-camp preparation phase to ensure that workshop time is dedicated to scientific development, experimentation, and integration. The detailed sequence of activities, milestones, and expected outputs is presented in Table 2. To complement this detailed schedule, Table 3 provides a Gantt-style overview of the temporal allocation of work packages and deliverables across the main project phases.

Although the work packages are presented as logically ordered components, they will be executed through parallel subgroups. WP1 provides the common experimental foundation, while WP2, WP3, and WP4 progress in parallel once the baseline interfaces are available. The final day is deliberately reserved for integration, demonstration polishing, reporting, and presentation preparation rather than major new implementation.

Task distribution. The work will be organized into four complementary streams. One subgroup will focus on data, task, and evaluation pipelines. A second subgroup will implement prediction-targeted and confidence-targeted attacks. A third subgroup will work on robust adversarial training and comparative evaluation. A fourth subgroup will integrate the selective moderation layer and demonstrator. The team leaders will coordinate the scientific protocol, ensure consistency between work packages, supervise experimental design, and support integration of final outputs.

This work plan is deliberately scoped to remain feasible within the two-week camp. Optional extensions, such as three-class detection, additional backbones, deep ensembles, out-of-distribution detection, or localization, will only be pursued if time and compute resources allow.

Risk mitigation. The project has been deliberately scoped to reduce implementation risk. Data-access risk is limited by relying on SID-Set, a publicly accessible dataset, and by preparing a curated balanced subset before the camp. Compute risk will be handled by using one primary backbone and scaling experiments only if resources allow. Scope risk is controlled by focusing on binary real-vs-fake detection, with multi-class detection, localization, additional backbones, deep ensembles, and out-of-distribution detection treated as optional extensions. Ethical and legal risks are limited because the project uses an existing research dataset and does not collect new personal data during the camp.

Table 2: Detailed implementation timeline.

Period	Main activities	Milestones and outputs
Pre-camp	Environment setup, SID-Set access, data loading utilities, baseline code structure, and initial metric implementations.	Experimental scaffold ready before the camp.
Day 1	Project kick-off, final task definition, team organization, validation of the core scope, shared repository setup, and allocation of subgroups.	Final experimental protocol; task allocation; shared implementation interfaces.
Days 2-3	**WP1 - Deepfake Detection Framework:** implementation of the data, task, and evaluation pipelines; baseline detector setup; clean evaluation; and initial realistic perturbation evaluation. In parallel, WP2 starts from the first available baseline interfaces by preparing and implementing prediction-targeted attack components.	First working baseline; data/task/evaluation pipelines; initial prediction-targeted attack components.
Days 4-5	**Parallel work phase I:** WP2 continues the adversarial attack package with prediction-targeted attacks and confidence/uncertainty-targeted attacks. WP3 starts robust adversarial training experiments from the training pipeline and baseline model.	**D1.** Baseline detection framework; first attack results; first robust training runs.
Days 6-8	**Parallel work phase II:** WP2 finalizes the attack package and vulnerability assessment. WP3 continues robust adversarial training and comparative evaluation. WP4 starts the selective moderation layer using baseline and intermediate model outputs, and integrates uncertainty scores, risk-coverage outputs, and abstention rules into the moderation workflow.	**D2.** Attack package with vulnerability assessment; robustly trained detector variants; first integrated selective moderation layer.
Day 9	Cross-WP integration and comparative analysis across standard and robust models under clean, realistic, and adversarial conditions. WP4 finalizes the selective moderation prototype and consolidates the decision workflow.	**D3.** Defense package + comparative results; **D4.** demo-ready selective moderation prototype; final analysis figures.
Day 10	Final integration, demonstrator polishing, report consolidation, and presentation preparation.	Final selective moderation prototype; final figures, summary report, and presentation material.

Table 3: Gantt-style overview of work packages and associated deliverables across the TrustFake project timeline.

Work package / output	Pre-camp	1	2	3	4	5	6	7	8	9	10
Setup and experimental scaffold	•										
WP1 - Deepfake Detection Framework / D1		•	•	•							
WP2 - Adversarial Attacks / D2			•	•	•	•					
WP3 - Robust Adversarial Training / D3					•	•	•	•	•		
WP4 - Selective Moderation Prototype / D4							•	•	•	•	
Final integration and presentation									•	•	•

8. Multidisciplinarity and Reusable TRAIL Factory Brick

Multidisciplinarity between STEM and SSH: Yes. TrustFake is technically grounded in computer vision, adversarial machine learning, uncertainty quantification, and robust training. At the same time, its application domain concerns media reliability, misinformation, content moderation, and human-in-the-loop decision-making. The project therefore naturally connects STEM methods with SSH-relevant questions around trust in digital media, decision support, and the governance of AI-assisted moderation.

Reusable TRAIL Factory brick: Yes. The project is expected to deliver reusable components that can later support applied AI processes: a deepfake detection framework, an adversarial attack package, a robust training package, and a selective moderation prototype. These components could be adapted into a reusable evaluation and decision-support brick for trustworthy multimedia AI, helping organizations test whether visual AI systems remain reliable under realistic perturbations and adversarial manipulation.

Bibliographic References

[1] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025.

[2] Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. pages 20379–20389, 2025.

[3] Ido Galil and Ran El-Yaniv. Disrupting deep uncertainty estimation without harming accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, pages 3–14, 2017. doi: 10.1145/3128572.3140444.

[5] Kathrin Grosse, David Pfaff, Michael Thomas Smith, and Michael Backes. The limitations of model uncertainty in adversarial settings. *arXiv preprint arXiv:1812.02606*, 2018.

[6] Huimin Zeng, Zhenrui Yue, Yang Zhang, Ziyi Kou, Lanyu Shang, and Dong Wang. On attacking out-domain uncertainty estimation in deep neural networks. In *Proceedings of the Thirty-First*

International Joint Conference on Artificial Intelligence (IJCAI), pages 4893–4899, 2022. doi: 10.24963/ijcai.2022/678.

[7] Emanuele Ledda, Giovanni Scodeller, Daniele Angioni, Giorgio Piras, Antonio Emanuele Cin, Giorgio Fumera, Battista Biggio, and Fabio Roli. On the robustness of adversarial training against uncertainty attacks. *Pattern Recognition*, page 112519, 2025.

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[11] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[12] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019.

[13] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[16] Black Forest Labs. FLUX.1-dev. Hugging Face model card, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev.

[17] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.