# Beyond Seeing Is Believing: On Crowdsourced Detection of Audiovisual Deepfakes

**Michael Soprano**, Andrea Cioci, Stefano Mizzaro

ROMCIR 2026 — The 6th Workshop on Reducing Online Misinformation through Credible Information Retrieval

April 2, 2026 — Delft, The Netherlands

*Held as part of the 48th European Conference on Information Retrieval (ECIR 2026)*


UNIVERSITÀ DEGLI STUDI DI UDINE

HIC SUNT FUTURA

ECIR 2026

# Motivation

- **Audiovisual deepfakes** are becoming increasingly realistic and accessible
- Their growing realism increases misinformation risks and makes **authenticity assessment** more difficult
- Assessing authenticity depends not only on the **available evidence**, but also on how audiovisual content is **presented** and **interpreted**
- **Authenticity judgments** are therefore an important part of audiovisual deepfake detection

UNESCO (Naffi, 2025) describes deepfakes as part of a broader “crisis of knowing”

# The Human Component

- Deepfake detection is often framed as an **automated classification task**, but detector reliability may not remain stable across settings
- **Individual human judgments** are also imperfect in realistic online-like conditions
- The **human side of deepfake detection** is therefore worth direct study
- We examine how **individual and aggregated crowd judgments** support **authenticity assessment**

# Research Questions

- We address the study through **three research questions**

**RQ1**
Can a crowdsourcing-based approach distinguish **authentic** from **manipulated** videos?

**RQ2**
How much do workers **agree** with one another when judging **audiovisual authenticity**?

**RQ3**
How accurately can workers identify the **manipulation type** and localize it with **timestamps**?

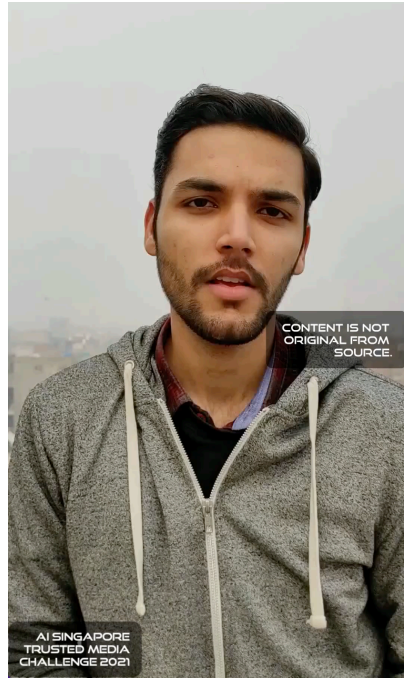
# Benchmark Selection

- Deepfake benchmarks vary substantially in **scale**, **modality**, and **collection setting**
- Many benchmarks remain primarily **video-centric**, rather than fully **audiovisual**
- **TMC** and **AV-Deepfake1M** provide explicit **authentic/manipulated** labels and a consistent **manipulation type taxonomy**
- This lets us compare the two datasets under the same **sampling design**

Chen et al. (2022; TMC), 10.1145/3511808.3557715 · Cai et al. (2024; AV-Deepfake1M), 10.1145/3664647.3680795

# Video Sampling

- We sampled **48 videos** from each dataset
- The pool was balanced across **four conditions**
  - **Authentic**
  - **Audio-only**
  - **Video-only**
  - **Audio-video**
- This resulted in **96 videos overall**




Example clips from the two benchmarks. Top: AV-Deepfake1M · Bottom: TMC

# Crowdsourcing Task

- We ran **two matched crowdsourcing tasks** on **Prolific** with **240 workers**
- Each work unit contained **4 videos**
- Each sampled video received **10 independent judgments**
- Overall, we collected **960 judgments**
- Workers were paid **1.50 GBP** per work unit
- We used the same **interface** and **label space** across both datasets

# Judgment Flow

- Each worker completed a **two-step judgment**

**INPUT**
The worker watches one **video**

↓

**STEP 1**
The video is judged as **authentic** or **manipulated**

↓

**STEP 2**
If judged as **manipulated**, the worker reports **manipulation type** and **timestamp**

↓

**OUTCOME**
One of four judgment labels: **authentic**, **audio-only**, **video-only**, or **audio-video**

# Judgment Aggregation

- Each video received **10 judgments**
- We used **two aggregation strategies** to combine these judgments
- **Majority vote** provides a simple baseline
- **Dempster-Shafer** weights judgments by worker reliability and models uncertainty explicitly

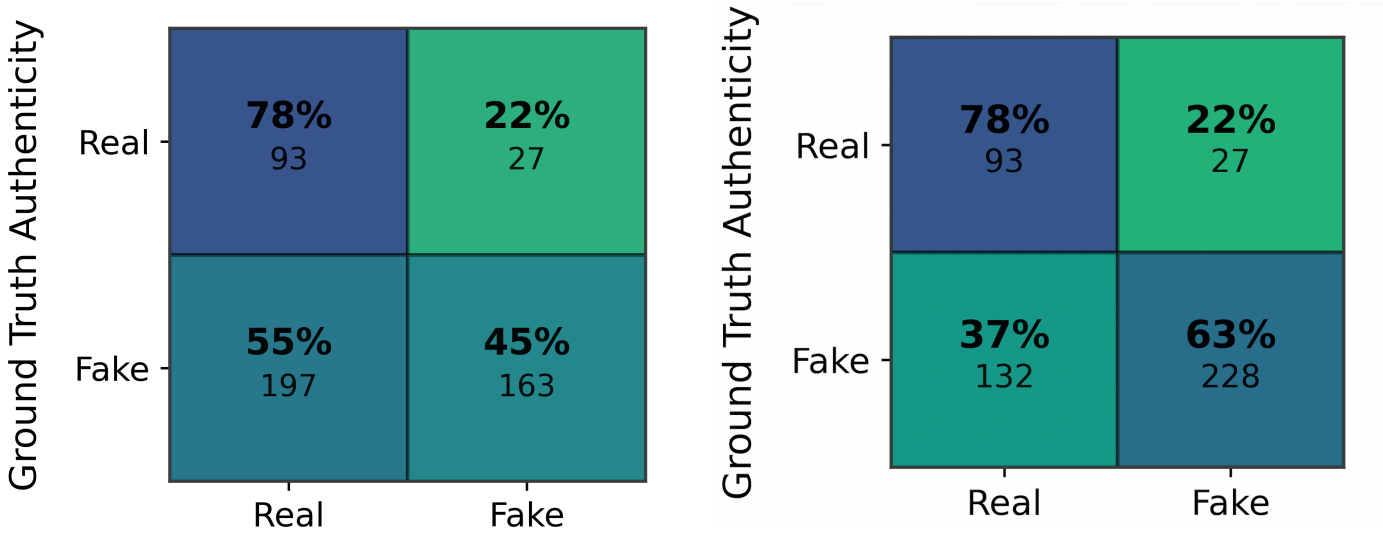
Dempster (1967), 10.1214/aoms/1177698950 · Shafer (1976), 10.1515/9780691214696

# Evaluation

- We evaluated **authenticity detection** with standard classification metrics
- **Judgment consistency** was measured through inter-annotator agreement:
  - **Krippendorff's alpha**, as an overall agreement coefficient
  - **Majority agreement**, as concentration around the majority label
  - **Pairwise agreement**, as how often workers agree with one another
- We evaluated **timestamp reports** through agreement around the per-video median
- Statistical comparisons relied on **nonparametric tests**

# RQ1: Authenticity Detection

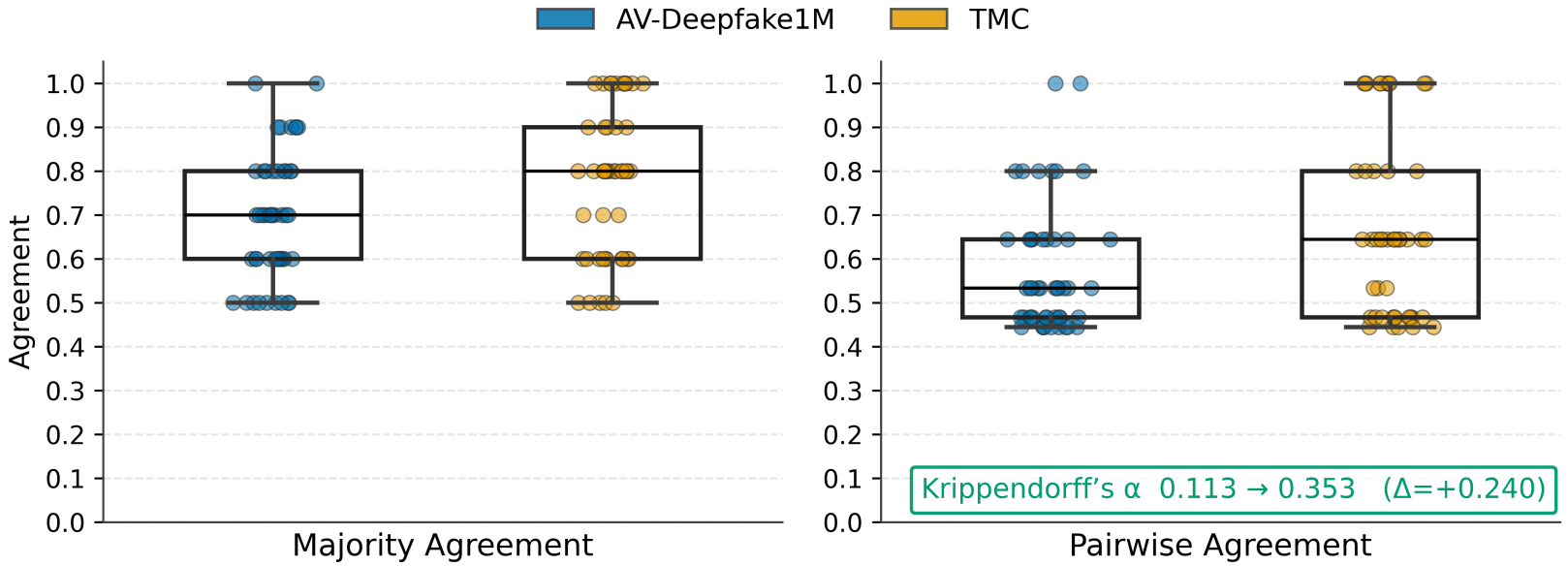
- Authentic videos are only rarely judged as **manipulated**
- Errors are dominated by **missed manipulations**
- This pattern is more severe on **AV-Deepfake1M** than on **TMC**
- Even after **aggregation**, missed manipulations remain substantial




Evidence from individual authenticity judgments. Left: AV-Deepfake1M · Right: TMC

# RQ2: Judgment Consistency

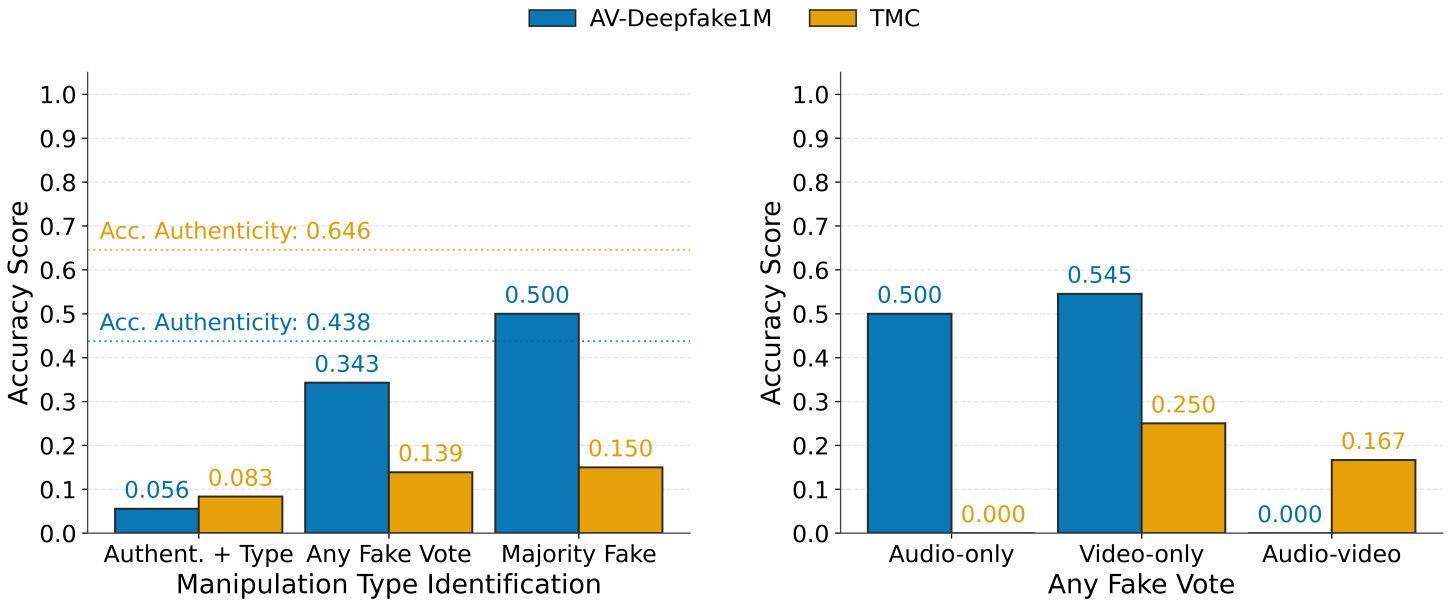
- Agreement remains **limited overall**
- It is nonetheless consistently **higher on TMC** than on **AV-Deepfake1M**
- **Krippendorff's alpha** increases from **0.113** to **0.353**
- Majority and pairwise agreement show the same pattern across datasets




Internal agreement of authenticity judgments. Left: majority agreement · Right: pairwise agreement

# RQ3: Manipulation Type Identification

- **Manipulation type identification** is harder than **authenticity detection**
- Workers may detect that something is wrong, but often misidentify the affected **modality**, especially in **audio-video** cases
- **Timestamp reports** can still converge on a plausible segment




Manipulation type identification. Left: overall accuracy across settings · Right: accuracy by type under Any Fake Vote

# Main Takeaways

> Workers may notice that **something is off**, but often fail to attribute the anomaly to the **affected channel**.

- Crowd judgments provide a useful **authenticity signal**, but many manipulations are still **missed**
- The crowd signal is consistently **stronger on TMC** than on **AV-Deepfake1M**
- **Modality attribution** remains the hardest part of the task
- **Timestamp reports** can still support **downstream review** when workers converge on a plausible segment

# Practical Implications and Future Work

- One practical workflow **might be two-stage**: crowd-based screening first, then expert or model-assisted verification
- **Aggregation** can stabilize authenticity judgments, but cannot recover manipulations that most workers miss
- **Manipulation type identification** may require richer interfaces and stronger worker support
- Future work will explore **human-AI workflows** and more controlled datasets varying **modality**, **duration**, and **manipulation strength**

# Thank you!

- **Repository:** 10.17605/OSF.IO/9RJ28
  - Judgments, demographic summaries, task configuration, and questionnaire materials
- **Contact:** michael.soprano@uniud.it

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI UDINE HIC SUNT FUTURA