DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection

RIT

Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Raymond Ptucha, Matthew Wright (Rochester Institute of Technology) and Andrea Hickerson (University of South Carolina)
saniat.s@mail.rit.edu

The Ethics and Governance of Artificial Intelligence Initiative

Deep-learning-based face-swap video and audio manipulations (Deepfakes) are getting easier to make, reducing trust in media.

This project, DeFake, aims to develop a tool to help journalists in the efficient detection of these manipulations before they reach the general population. We performed user studies with 11 journalists and media verification workers to develop a prototype and identify the requirements of a deepfake detection tool.

Current Process

Media verification triggered by:
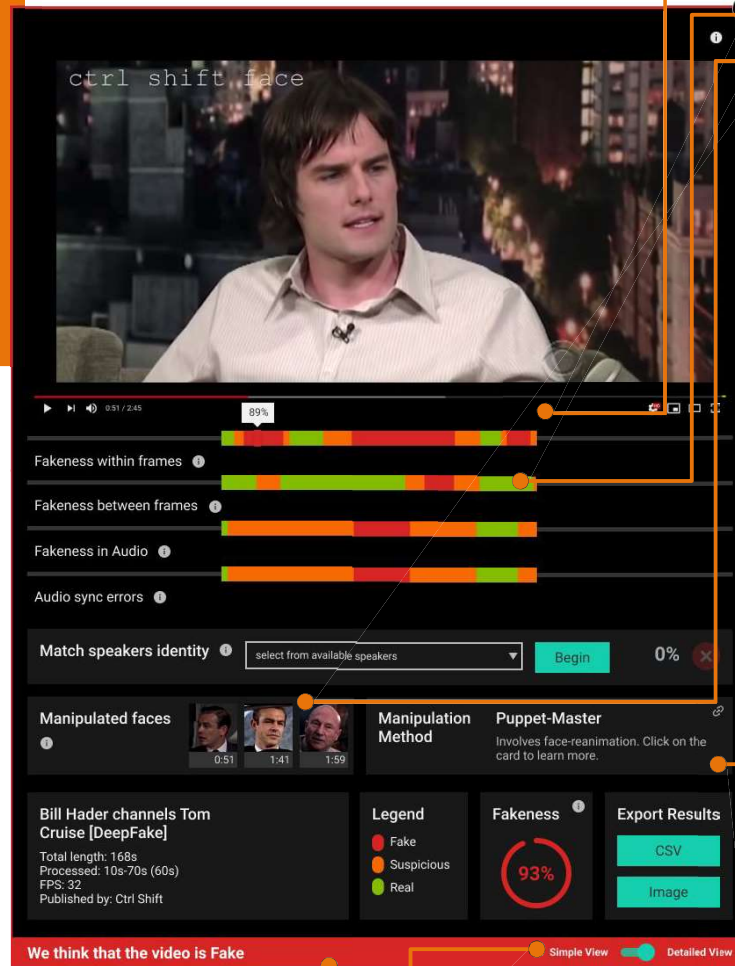**bipartisan sources** and **polarizing content.**

- No deepfake detection tools exist.
- Low awareness.
- Underestimating quality of manipulations.

	Local Orgs.	National / Int'l Orgs.
Verification Resources	**Low.** Smaller operations.	**Medium.** Dedicated teams.
Verification Procedures	**Minimal.** Using self recorded or pre-verified content.	**High.** Manual context verification. Google Reverse Image Search, TinEye, InVid Project

Requirements for Journalists
Providing explainability.

Fakeness within frames
Fakeness between frames
Fakeness in Audio
Audio sync errors

Match speakers identity

Manipulated faces

Manipulation Method — Puppet-Master — Involves face-reanimation. Click on the card to learn more.

Bill Hader channels Tom Cruise [DeepFake]
Total length: 168s
Processed: 10s-70s (60s)
FPS: 32
Published by: Ctrl Shift

Legend
● Fake
● Suspicious
● Real

Fakeness 93%

Export Results
CSV
Image

We think that the video is Fake   Simple View — Detailed View

Fakeness indicators to get verdict at a glance.

Simpler view (only bottom row)

Descriptions of detected methods.

Variety of distinct detection methods to allow journalists make informed judgements.

Time-distributed fakeness scores.

Visualization of manipulated areas/faces.

Technical Requirements
- Accuracy over speed
- Less false negatives
- Open Source
- Logins and rate limiting

"Accuracy is the most important thing in journalism. Anyone who tells you otherwise is not a good journalist.
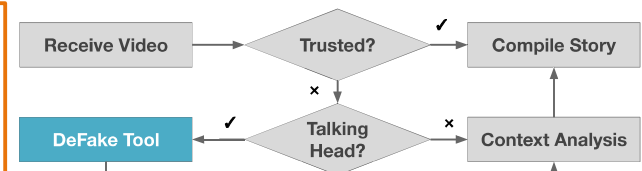— study participant

What Videos to Focus on?

- Talking head videos.
- Public addresses
- Varying compressions

Deepfake example: Bill Hader to Tom Cruise and Seth Rogen

Expected Workflow
The tool to be used for supplementary information.

Moving Forward
Beta testing a deployed tool on journalists for iterative improvements to the interface & workflow.