Safeguarding STEM Education and Scientific Knowledge in the age of Hyper realistic AI-Generated Data

Drs. Christopher Doss & Jared Mondschein, RAND Corporation
Dr. Conrad Tucker, Carnegie Mellon University
Dr. Lance Bush, Challenger Center

Challenge:

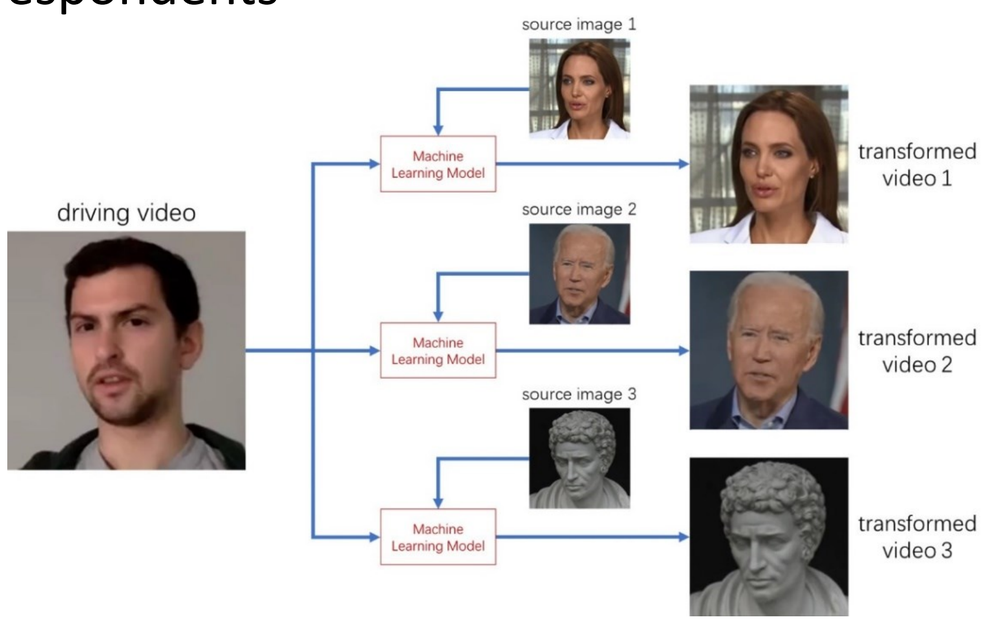
- AI lower barriers to mass creation and dissemination of manipulated digital content (deepfakes)
- The risk of exposure among education stakeholders has increased as learners and educators rely on the Internet for information
- To date we do not know the level of vulnerability or what video and personal characteristics moderate vulnerability

Solution:

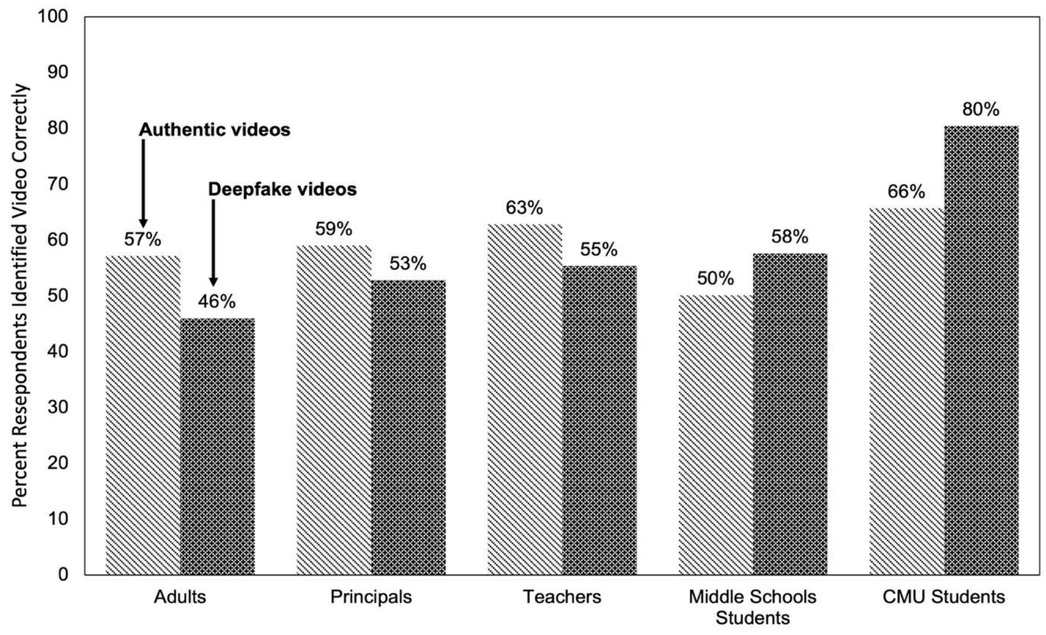
- Fielded randomized controlled trials in surveys to understand the vulnerabilities of U.S. adults, educators, K-12 students and college students to climate change deepfakes
- Asked respondents which aspects of videos helped them determine the videos' authenticity
- Asked about climate change knowledge, political orientation, learning habits, social media use, and perceived threat of deepfakes

Scientific Impact:

- Deepfakes are of sufficient quality to induce confusion that leaves all stakeholders vulnerable
- Adult populations were more vulnerable than undergraduate and K-12 students
- Those who reported more trust in information sources exhibited more vulnerability
- More exposure to potential deepfake videos increased vulnerability
- Social context of videos may help to better identify deepfakes, but this was used less often by respondents



Driving video can animate still images to create realistic deepfake videos



Percent of Responses Correctly Identifying the Authenticity of Videos, by Video Authenticity and Population

Notes: Each bar represents the percentage of responses that correctly identified the authenticity of videos by population and deepfake video status. Tabulations in the adult, principal, and teacher populations are weighted to be nationally representative.

Broader Impacts – Society

- Characteristics of more vulnerable stakeholders were identified
- Potential threat of unmitigated deepfakes was quantified and the urgency for mitigation revealed
- Importance of social context of deepfakes was quantified

Broader Impacts – Society

- Future mitigation strategies can target the most vulnerable stakeholders
- Results can guide content of mitigation strategies (e.g., understanding social aspects of deepfakes, instilling healthy distrust of information, etc.)

Broader Impacts – Participation

- Findings can inform content of mitigation strategies that can be fielded in:
  - K-12 schools
  - Universities
  - Informal learning centers
  - Online with partnerships with tech companies

Award ID#: 2039612