



Fake News and Misinformation

CS 224C: WEEK 7

Megan Mou
Rodrigo Nieto
Marie Tano
Kathy Yin
Yutong Zhang
Dorothy Zhao

Roozenbeek et al.

**“Psychological inoculation improves resilience
against misinformation on social media.”**

Science Advances (2022).

Rethinking how we address misinformation

Reduce belief and sharing of misinformation [1]

De-bunking misinformation is not always effective due to the “continued influence effect”



Pre-bunking calls for **preemptively** building resilience against exposure to misinformation



[1] Arechar et al. “Understanding and combating misinformation across 16 countries on six continents.” *Nature Human Behavior* (2023).

What is psychological inoculation?

Based on **inoculation theory**, which posits that we can build psychological resistance to unwanted persuasion techniques

1



Forewarning
that there will be a
threat to your
attitudes

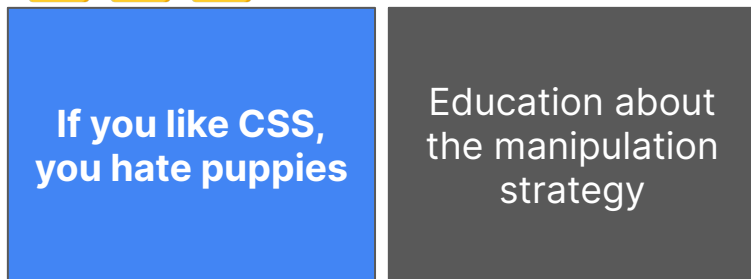
2



Exposure
to a microdose of
misinformation you
can prebunk

Using inoculation videos for pre-bunking

VIDEO FORMAT



1 of 5 manipulation strategies

(i.e., emotional language, incoherence, false dichotomies, scapegoating, and ad hominem)

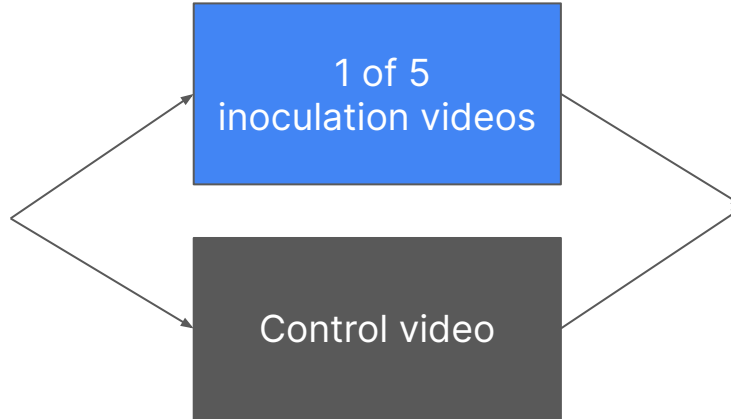


Non political and fictitious example

Laboratory study setup



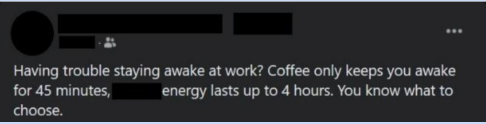
N=5,416
US Quota sample



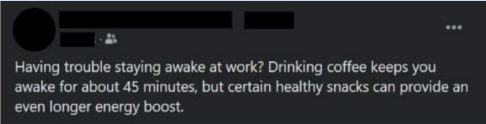
10x

Post-test Questionnaire

MANIPULATIVE



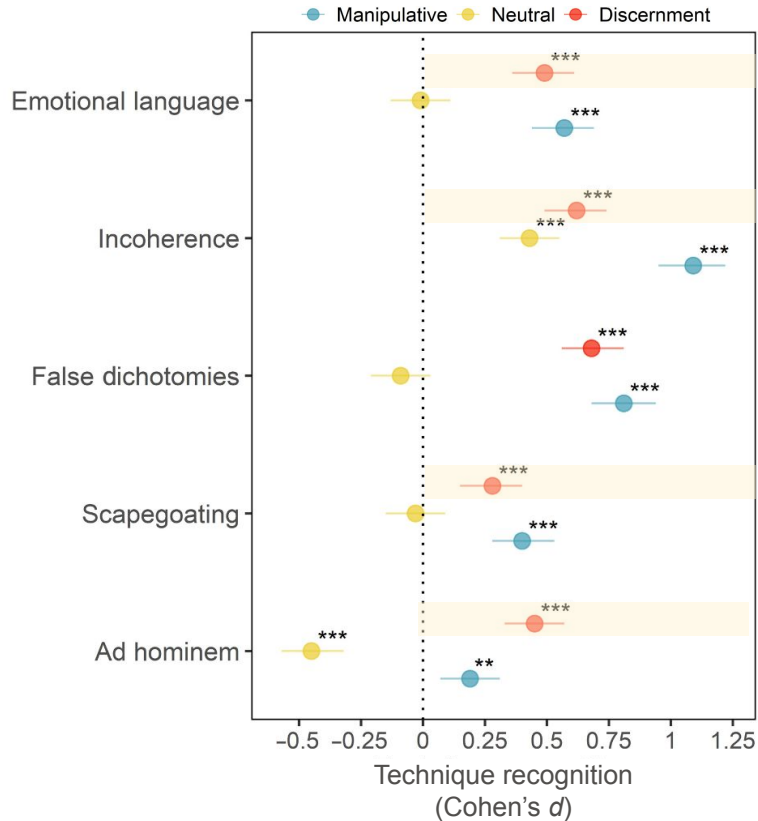
NEUTRAL



7-point Likert scales measuring technique recognition, recognition confidence, trustworthiness, and willingness to share

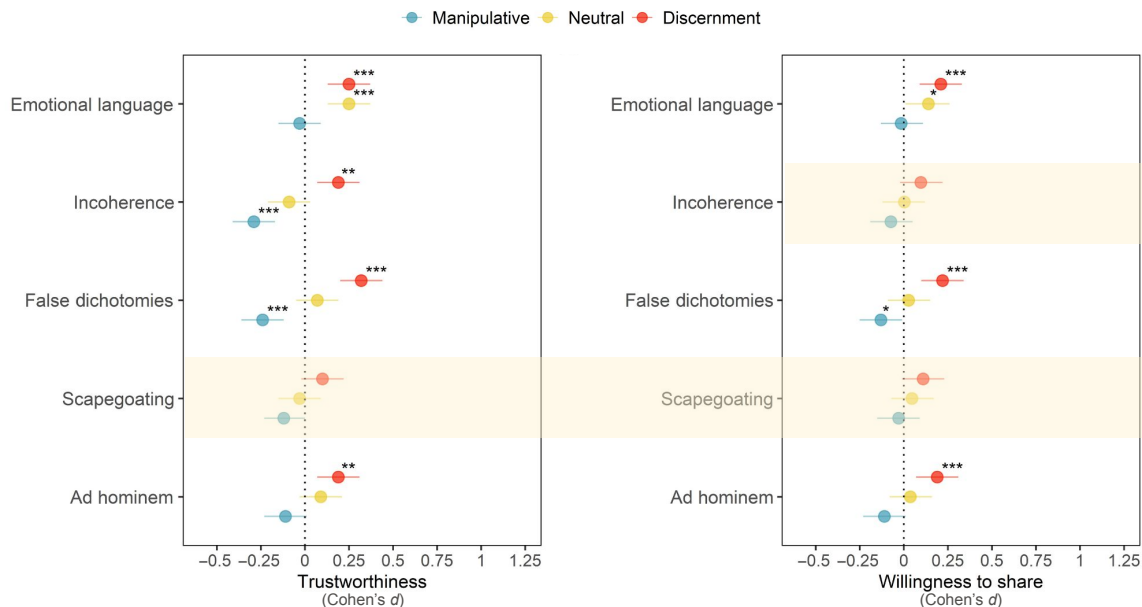
Discernment between manipulative and neutral stimuli

Inoculation improves manipulation recognition



Participants in the treatment condition were significantly **better** at discerning a persuasive technique

Inoculation mostly improves trustworthiness and sharing discernment



Lack of significance for scapegoating and incoherence may be due to high baselines

Participants in the treatment condition were significantly **better** at discerning

- Manipulative content as untrustworthy
- Not sharing manipulative content

Robustness checks confirm prior results

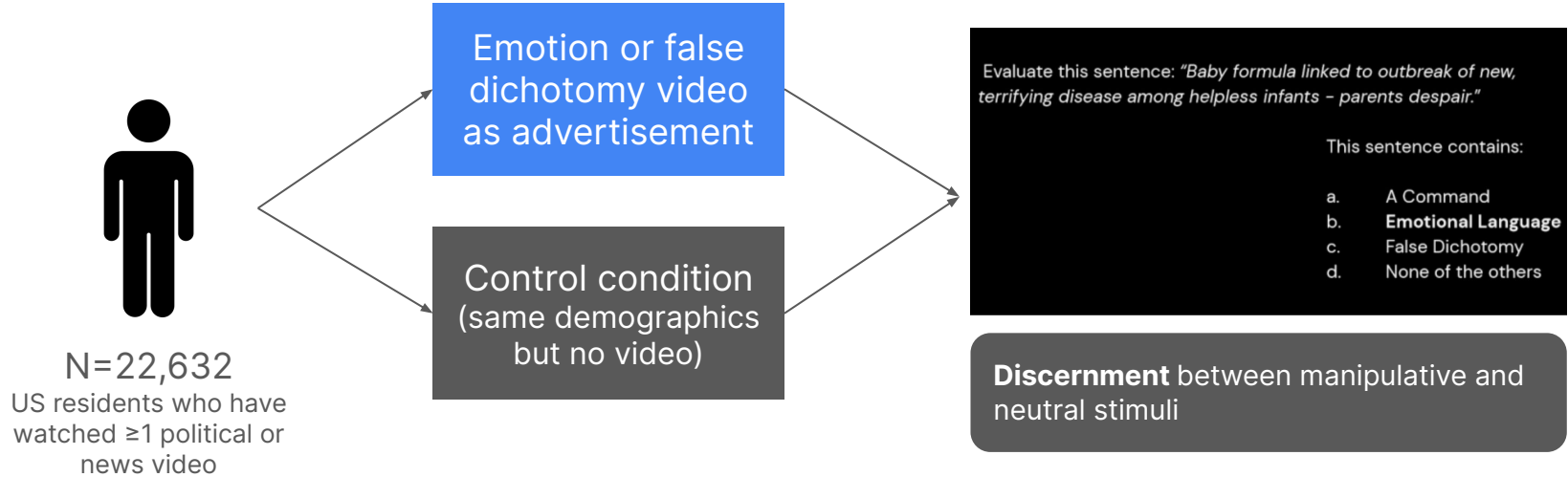
1 Replication and order effects

- Study on emotional language 1 year later (N=1,068)
- Check if question order in the post-test influences results
- Results hold + no meaningful interaction with question order

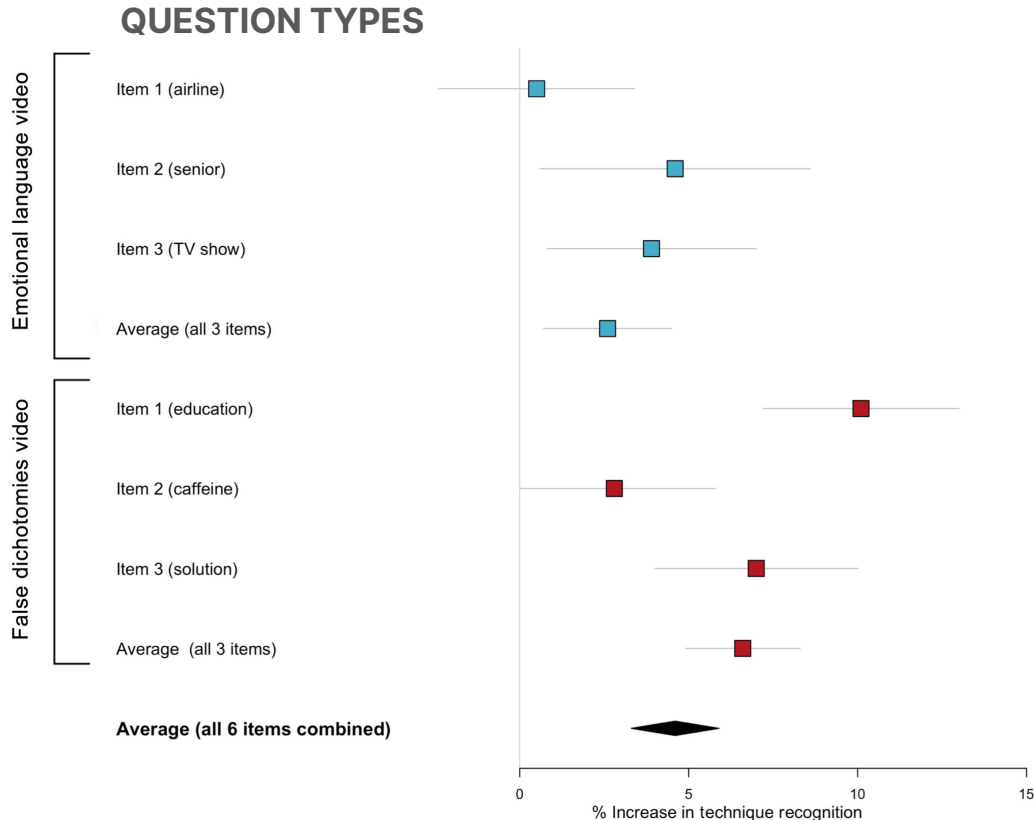
2 Moderators

- Study whether individual-level characteristics (e.g., partisanship, “bullshit” receptivity, demographics) moderate interventions
- No significant two-way interactions between the experimental condition and covariates
- Participants from various backgrounds can be “inoculated”

YouTube study setup



Inoculation videos lead to some increase in recognition



Proportion of correct answers is greater in the treatment condition

Disaggregating by item, the treatment significantly improves recognition for some questions

Key Takeaways

1. Inoculation-based interventions were effective in improving discernment with substantial effect sizes
2. “Technique-based” approaches may offer a new way of addressing misinformation
3. Misinformation recognition holds in an ecologically valid setting
4. Running field studies may not be as expensive as expected (?)

EXPERIMENTAL RIGOR 💪

- Able to not only test our their intervention in an ecologically valid setting but also replicate their study after one year
- For the non-social scientists in the room, good to learn about pre-registering your work!

Peer Review

Paper Strengths

1 Novel Approach

- Technique-based approach as opposed to specific persuasive attacks

2 Robust Methodology

- The techniques were developed from the literature + great videos!
- The effect sizes for technical recognition are impressive
 - Cohen's d ranges from 0.28 to $d = 0.68$

3 Transparency

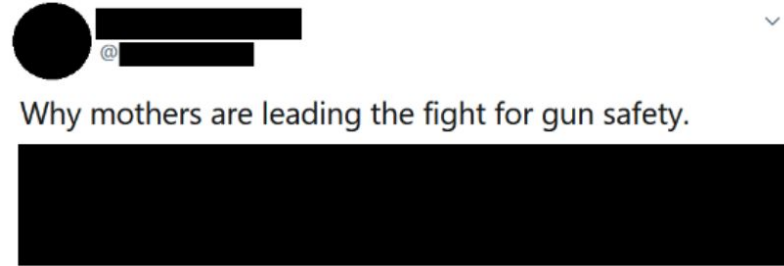
- Limitations highlight the paper's major shortcomings well
- Recognizes deviations from preregistration
- Well constructed power analysis

Paper Weaknesses

MANIPULATIVE



NEUTRAL



+

Poor handling of political ideology



Questions

How could the technique-based approach be adapted further?



How could this study be adapted to get an understanding of the longitudinal effects of the videos?



How should we consider whether the inoculation effects of YouTube translate to real-world settings?



Academic Researcher

Academic Research: Why should we consider cultural background regarding resilience against misinformation?

Cultural influences in language interpretation practices (Wierzbicka 2003; 2014)



Social Networks differ across cultures which can affect spread of (mis)information (ex: individualistic vs highly communal societies) (Pachucki & Breiger, 2010; Vilhena et al., 2014)



Access and reliability on technology can also vary across cultures and affect how information is received (Arechar et al. 2023)



Academic Research: Cross-Cultural Effectiveness of Inoculation Videos

Research Question:

- How effective are inoculation videos across different cultural and linguistic contexts?

Method and Experimental design:

- Recruit participants from diverse cultural backgrounds (Bonus points if non-Western communities)
- Gather information about participants' cultural background and values related to information consumption
- Randomly assign people to watch either a translated 1.5-min inoculation video or a translated neutral control video
- Calculate and collect ratings of discernment and confidence measures

Academic Research: Expected Outcomes

1 Effectiveness across cultures

- Determine overall effectiveness of inoculation videos in improving misinformation recognition and resilience across different cultural contexts

2 Cultural Awareness and Sensitivity

- Learn how factors such as cultural background and technological familiarity influence responses to misinformation

3 Guidelines for future resources

- Better inform future methods to ensure that misinformation interventions are localized to specific cultural and linguistic contexts

Industry Practitioner

Industry Application: Misinformation Resilience Toolkit



Features and Functionality

- An integrated feature for existing social media applications.
- Provides users with short, engaging videos on misinformation techniques.
- Part of the onboarding process or before high-risk / sensitive content.
- “Social media companies could furthermore offer ad credits to run inoculation campaigns on their platforms.” (Roozenbeek et al.)

Positive vs. Negative Impacts

- **Positive:** encourages increased sharing of high-quality, trustworthy information and discourages misinformation posts.
- **Negative:** sows potential distrust among users in content on-platform, despite positive intentions.

Social Impact

Social Impact: Social Media Literacy for New Gen

- What is social media literacy?

Digital media literacy is the ability to critically, effectively and responsibly access, use, understand and engage with media of all kinds.

- Why is social media literacy important?

The young generations are now born to be digital citizens or becoming one. Media mediate everything in the society - work, education, information, relationships etc.

- What can we do?

Misinformation training videos and curriculums in K-12 classrooms. This can help students think critically about the information presented online and common manipulating strategy.



Groh et al.

**“Deepfake detection by human crowds,
machines, and machine-informed crowds.”**

Proceedings of the National Academy of Sciences (2022).

Background: Deepfakes

- **Deepfakes:** videos that have been **manipulated/edited** by neural network models



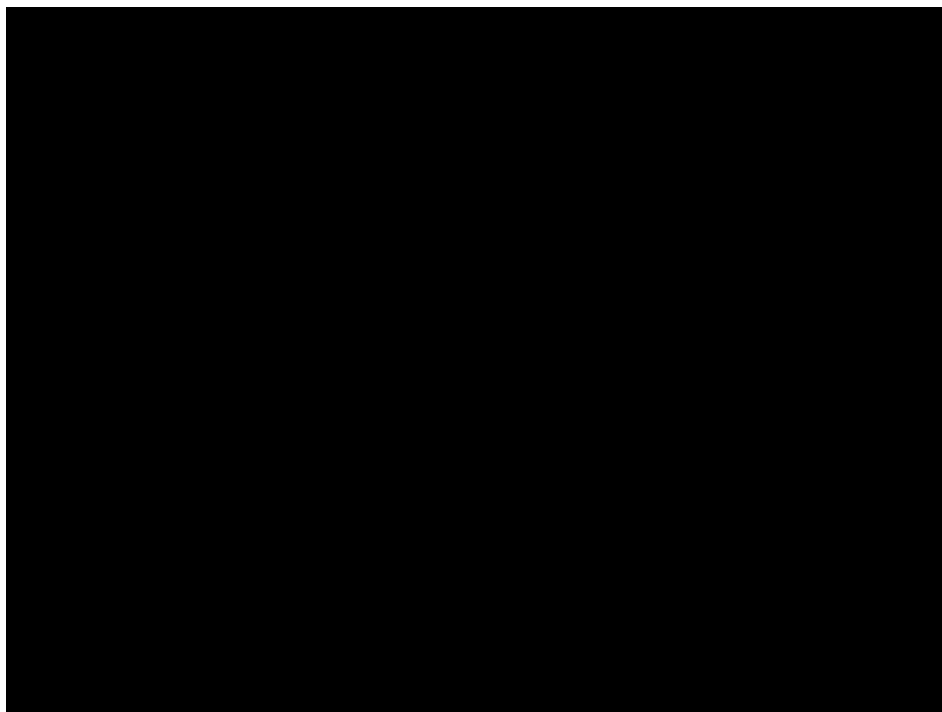
?



?

Background: Deepfakes

- **Deepfakes:** videos that have been **manipulated/edited** by neural network models



Background: Deepfake Detection

- **Deepfakes:** videos that have been manipulated/edited by neural network models.
- **Deepfake detection:** classifying videos as real or fake.



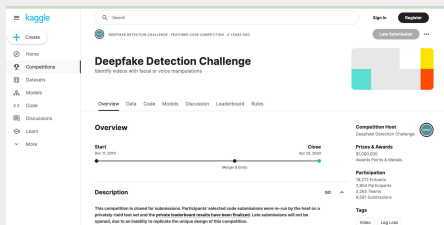
Deepfake



Authentic

Introduction

Dataset: Deepfake Detection Challenge



Detectors



Model

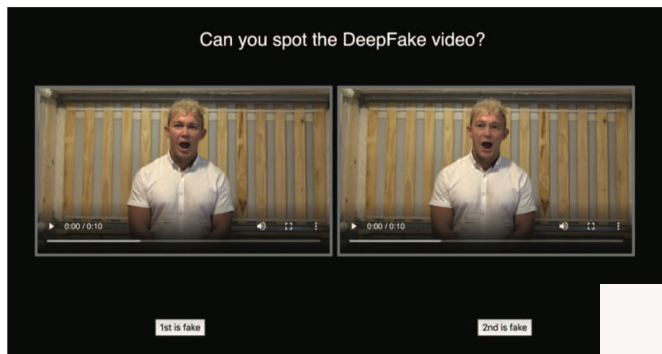


Human



Human + Model

Experiments

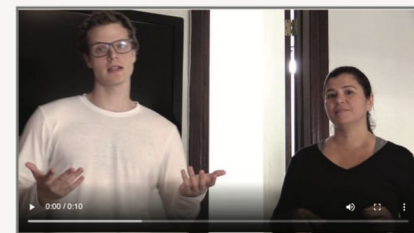


E1: Two-Alternative Forced Choice

Can you spot the DeepFake video?

[Instructions](#)

Please note that some distractions (e.g., extra shapes, extra faces, upside-down videos, facial misalignment, and black boxes over the eyes) may be applied to the videos. A video with only distractions should not be considered as a deep fake. A deepfake is a video where a face has been manipulated by AI to partially distort the face or swap facial features.



You have rated 8 videos. You must play the video before submitting.

E2: Single Video Design

AI v.s. Human: Comparison and Collaboration

- **AI detector:** develop large datasets and train computer vision algorithms.
 - E.g., Deepfake Detection Challenge (DFDC)
- **Human detector:** prior knowledge, commonsense reasoning.
- **AI v.s. Human: Comparison and Collaboration:** **how well?**
 - Individual v.s. Machine (AI)
 - Crowd Wisdom v.s. Machine (AI)
 - Human-AI collaboration
- **Experiments:**
 - **E1:** Two-Alternative Forced Choice
 - **E2:** Single Video Design

Individual v.s. Machine

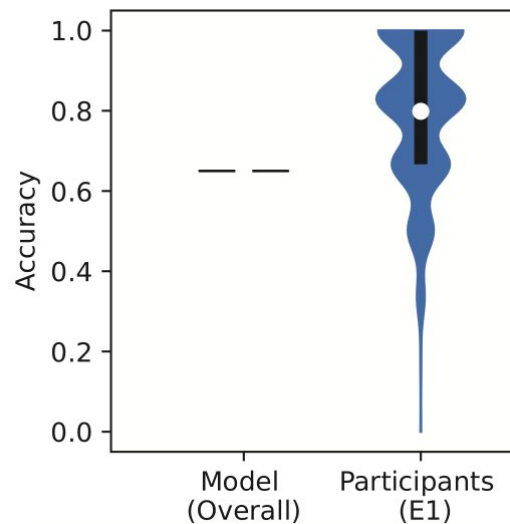


Are humans or the leading machine learning model more capable of detecting algorithmic visual manipulations of videos?

E1

Two-Alternative Forced Choice Experiment

- **Task:** Select which of two video clips is a deepfake
- **Setup:**
 - 26,820 trials
 - 56 pairs of videos
 - 882 participants (saw at least 10 video pairs)
- **Results:**
 - **82% of participants outperform the leading model**
 - Performance accuracy: right figure



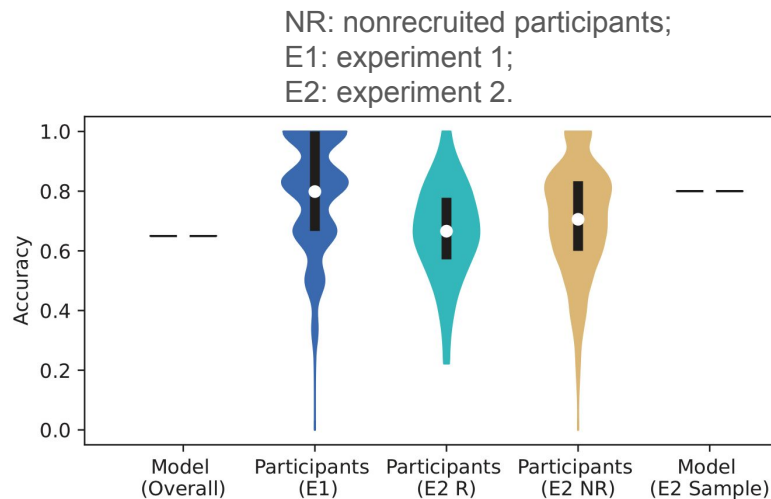
Individual v.s. Machine



Are humans or the leading machine learning model more capable of detecting algorithmic visual manipulations of videos?

E2 Single Video Design Experiment

- **Task:** Ask participants' confidence on video: (0 - 1)
- **Setup:**
 - # participants: 304 (Prolific) + 9,188 (Website)
 - # trials: 6,390 (Prolific) + 67,647 (Website)
 - # pairs of videos: 56
- **Results:**
 - Recruited participants: 57% of attempts compared to the leading model identifying deepfakes as deepfakes in 84% of videos
 - Both recruited participants and the leading model identify real videos as real videos at nearly same rate (75% vs 76%)



Takeaway: Human performance \approx model performance

Individual v.s. Machine



What explains variation in human and machine performance



Individual v.s. Machine



What explains variation in human and machine performance



E1

Two-Alternative Forced Choice Experiment

- Participants are **5.6%** less accurate at detecting pairs of inverted videos than upright videos

Individual v.s. Machine



What explains variation in human and machine performance

E2 Single Video Design Experiment

- Treatment interventions:
 - Inversion: shown upside down
 - Misalignment: with the top and bottom half of the actor's face misaligned
 - Occlusion: shown with the eyes occluded by a thin black strip
- Results: **all obstructions affect participants' ability to accurately identify deepfakes**
 - Inversion: 4.3% decrease in accuracy and 2% decrease in confidence score
 - Misalignment: 6.3% decrease in accuracy
 - Occlusion: 4.4% decrease in accuracy

Takeaway: Specialized cognitive capacities play an important role in explaining human performance in deepfake detection.

Individual v.s. Machine



What explains variation in human and machine performance

Video-level characteristics

- Seven video-level feature: graininess, blurriness, darkness, presence of a flickering face, presence of two people, presence of a floating distraction, presence of an individual with dark skin
- Result (4): **flickering face, two people in the same video, floating distractions, and presence of an individual with dark skin affect people's performance**

Emotion Priming

- Do not find statistically significant effects of anger elicitation on accuracy on overall accuracy

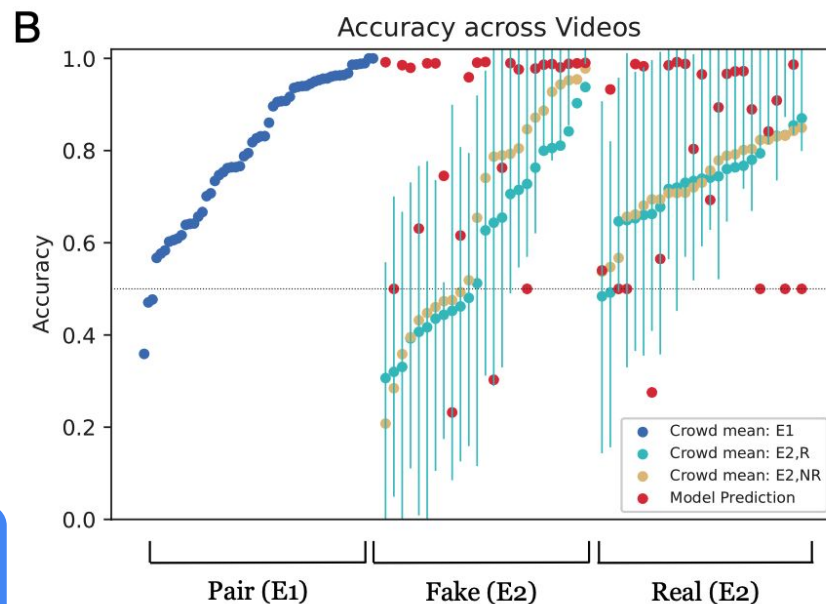
Crowd Wisdom v.s. Machine



The crowd mean, participants' responses averaged per video, is **on par** with the leading model performance on the sampled holdout videos.

- **Task:** evaluate accuracy of the crowd mean
- **Results:**
 - Recruited participants: 76%
 - Non-recruited: 80%
 - Non-recruited seeing at least 10 video: 86%
 - Machine: 80%

Takeaway: Crowd-mean responses \approx model responses



Human-AI Collaboration

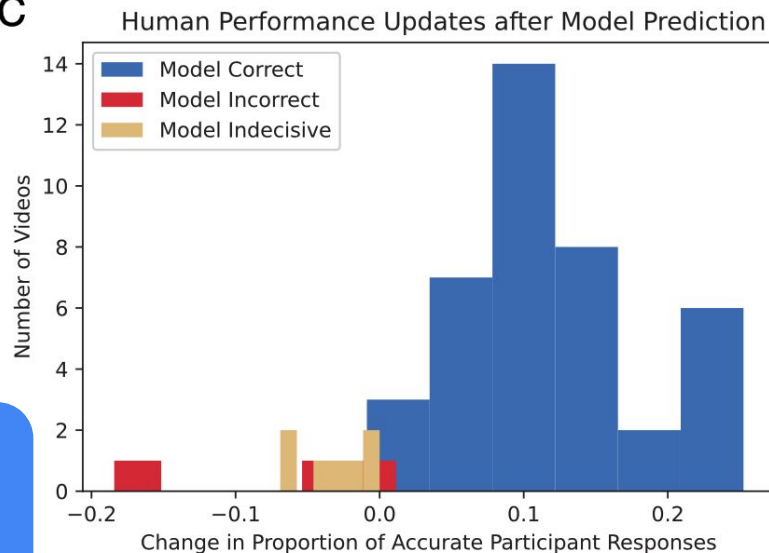


How an AI model could complement human performance

- **Task:** whether human-AI collaboration could help deepfake detection
- **Results:**
 - Update their confidence in 24%
 - Participant's accurate identification increased from 66% to 73%
 - When model made an incorrect or equivocal prediction: participants accuracy decreased by 2.7%

Takeaway: Participants with access to the model's predictions outperform both human-only and model-only approaches in accuracy, but inaccurate model predictions often decrease participants' accuracy

C

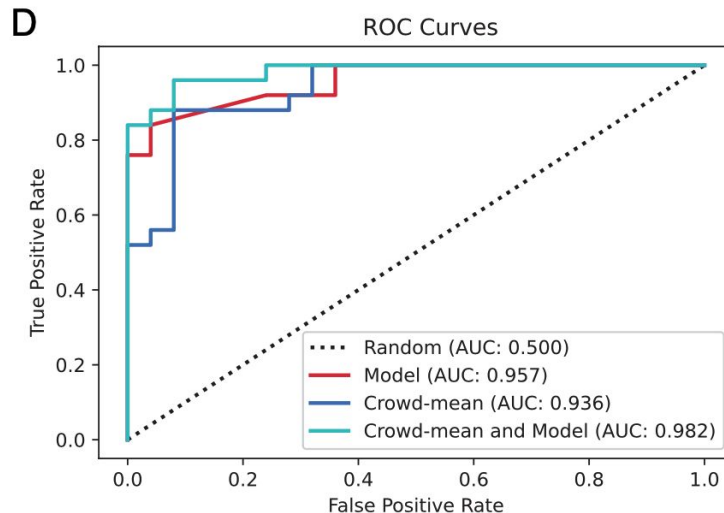


Human-AI Collaboration



How an AI model could complement human performance

- **Task:** whether human-AI collaboration could help deepfake detection
- **Results:**
 - Receiver Operating Characteristic (ROC): a graph showing the performance of a classification mode at all classification thresholds
 - the crowd mean response after seeing the model's predictions strictly outperforms both the performance of the crowd mean and leading model



Takeaway: crowd mean response after seeing the model's predictions strictly outperforms both the performance of the crowd mean and leading model

Key Takeaways

1. Humans perform in the range of the leading machine learning models
2. Collective intelligence, crowd mean, would be accurate as the model's prediction
3. System integrating human and model predictions is more accurate than either humans or the model alone
4. Inaccurate model predictions often mislead human detection

Discussion 🤔

- **Comparison:** Human and AI are compatible on accuracy, how about reliability, perception cost, and generalization?
- **Future:** Consider the variance of Human and AI, what is the dominant strategy in the future: human-in-the-loop v.s. model detector?

Synthesizing Roozenbeck et al. and Groh et al.

1 Technique-based approaches for recognizing “fake news”

- Roozenbeck et al. present psychological manipulation strategies that have been used to influence people
- Groh et al. show that participants fare as well as ML models at identifying deepfakes

2 Onus on detecting these techniques

- As we develop new ways to “inoculate” people against misinformation, what happens when adversaries develop new techniques?
 - Parallel: need to develop new versions of vaccines
- Context of deepfakes with increasing model capabilities

Peer Review

Paper Strengths

1 Analysis on Human Visual Processing vs Leading Model

- Goes in depth with considering the performance gaps across video subtypes and specialized processing

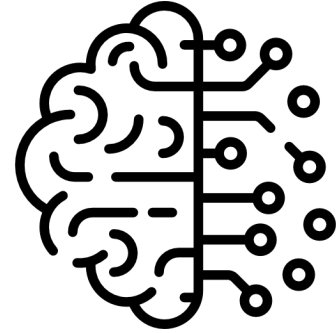
2 Heavily Literature Based

- Multiple callbacks to neuroscience, perceptual psychology, forensics

3 Large and diverse sample size

- Experiment 1: $n = 5,524$ individuals from USA, Germany, UK, Saudi Arabia, Canada...
- Experiment 2: $n = 9,188$ individuals from USA, Brazil, UK, Canada, and Germany...

Paper Weaknesses



■ Limited to one leading model

■ Little Explanation of Model Behavior

■ Lackluster Emotional Intervention

Questions

This study could benefit from more explainable AI. What should that specifically look like?



Could the integration of AI predictions lead to an overreliance on these technologies?



What could be some practical applications of the "wisdom of the crowd" finding?



Academic Researcher

Academic Research: Why should we consider the intersection of race and emotion in regards to deepfake detection?

Bias in Deepfake Detection:

Darker skin tones are less likely to be accurately

detected (Haut et al. 2021;

Trinh & Liu, 2021)

Emotion and perceived credibility:

Perceived emotion has an effect on perceived credibility or accuracy of

source (Campellone and Kring 2013; Vlasceanu, Goebel, and Coman 2020; Karduni et al. 2023)

Humans and Emotions:

Humans are more skilled at correctly identifying facial emotions of in-group members than other races

(Freeman, 1984; Vinacke & Fong, 1955; Wolfgang & Cohen, 1988; Elfenbein and Ambady 2002; Weathers, Frank & Spell, 2002)

Anti-Blackness and Emotional Association: Black voices are more likely to be identified as angry (Weissler, 2021), hostile or aggressive (Baugh 2000, Harris-Perry 2011, Gillon 2015)

Academic Research: Impact of Emotion and Racial Cues on Deepfake Detection Accuracy

Objectives

- Investigate how participant demographic factors influence deepfake detection accuracy, particularly in videos featuring individuals of different racial backgrounds
- Analyze the effects of various emotional cues on deepfake detection accuracy across different demographic groups.
- To examine the interplay between racial biases, emotional tone, and the accuracy of deepfake detection.

Research Question:

- How do the emotional tone and racial cues in videos influence the accuracy of deepfake detection by participants from diverse demographic backgrounds?

Academic Research: Impact of Emotion and Racial Cues on Deepfake Detection Accuracy

Video Dataset

- Video Dataset: Use a dataset including deepfakes and authentic videos featuring individuals from different racial backgrounds and contexts eliciting different emotional responses.
- Emotional Manipulation: use facial and linguistic cues (prosody, pitch, dialect, etc.) to reflect particular emotions (anger, happiness, etc.)

Method and Experimental design:

- Human-only detection: Participants detect deepfakes without AI assistance.
- AI-only detection: AI model detects deepfakes independently
- Human-AI collaboration: Participants receive AI predictions along with detailed explanations and contextual information
- Measure detection accuracy, confidence levels, and time taken for decisions.

Academic Research: Impact of Emotion and Racial Cues on Deepfake Detection Accuracy

Hypothesis:

- Participants will show lower accuracy in detecting deepfakes involving individuals with dark skin tones and deepfakes with negative emotional tones
- Participants will more accurately detect deepfakes that involve individuals from their own racial or cultural group

Significance:

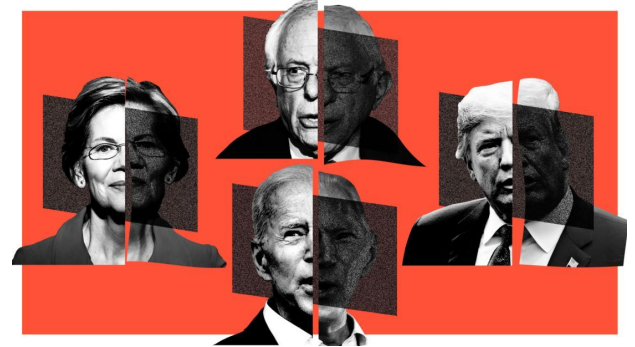
- Address racial bias in humans and algorithms
- Enhance understanding of factors that affect deepfake detection
- Interdisciplinary work is cool

Industry Practitioner

Industry Application: Deepfake Detection Module

Features and Functionality

- Consists of AI integration, a crowd-sourcing mechanism, a hybrid decision-making system, and user education features.
- Interface where users can flag potential deepfakes on news / social media sites.



Human-AI Collaboration

- Algorithm that weighs AI predictions and crowd inputs to produce a final outcome on deepfake detection.
- Includes info on **video subtypes / qualities** and how humans vs. machines perform across these subtypes, esp when they **diverge**. (ex: blurry, grainy, dark, specialized obstruction, stylistic similarities to training set)
 - **Only presenting model prediction tends to inaccurately skew human judgment**

Industry Application: Deepfake Detection Module

Positive vs. Negative Impacts



- **Positive:** active user participation in content verification fosters a sense of community and shared responsibility.



- **Negative:** risk of false positives and participation fatigue decreasing user satisfaction and engagement.

Discussion

- How much would you trust crowdsourced deepfake detection on social media platforms?
- What factors lend credibility?

Social Impact Assessor

Social Impact:

Positive Impacts:

- Public Awareness and Education: The study underscores the importance of educating the public about deepfakes and their detection by isolating specific factors that could hinder human detection.

Negative Impacts:

- Manipulation of Results: deepfake video creators gradually learn how to go around the detection machine or manipulate others perceptions (comments, retweets)
- Distrust of public: an increasing public awareness of deepfake videos will damage the trustworthiness of true information (In 2021, less than a fraction of a percent of news was misinformation.)

Duped

TRUTH-DEFAULT THEORY
and the SOCIAL SCIENCE
of LYING and DECEPTION



Timothy R. Levine