Improving Detection of Deepfakes

Study Information

Hypotheses

The study will investigate three hypotheses and two exploratory research questions.

Hypotheses:
H1: Participants provided with detection strategies will show greater detection accuracy relative to the control group.
H2: Participants provided with detection strategies will show greater detection confidence relative to the control group.
H3: There will be a bias toward categorising stimulus videos as real.

Research questions:
RQ1: Is level of interaction with stimulus videos associated with detection accuracy?
RQ2: Is detection confidence associated with detection accuracy?

Design Plan

Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding

For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Is there any additional blinding in this study?

No response

Study design

This study will employ a between-subjects experimental design. The experiment will be administered entirely online via Qualtrics.

Participants will be randomly assigned to one of two groups: a control group and a detection strategies group. Those in the detection strategies group will be provided with written strategies for detecting deepfakes. These strategies are produced by the MIT Media Lab (https://www.media.mit.edu/projects/detect-fakes/overview/). An example strategy is “Pay attention to the glasses. Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, deepfakes often fail to fully represent the natural physics of lighting.”

Participants will then be presented with a series of 20 10-second videos, half of which will be authentic and half of which will be deepfakes. Following Kobis et al. (2021), participants will be told that half of the videos are deepfakes and that they are able to watch each video as many times as they need.

The stimulus videos are sourced from the open-source Deepfake Detection Challenge Dataset (https://arxiv.org/abs/2006.07397), which contains over 100,000 clips sourced from paid actors (who all consented to have their likeness modified). Stimulus videos will be selected at random from the database until we have 20 sets of videos (an authentic and deepfake for version of each clip) which meet our inclusion criteria (e.g., consistent lighting, only one person depicted in the video, etc). Following Kobis et al. (2021), stimulus videos will be divided into two sets, each consisting of 10 deepfakes and 10 authentic videos (such that Set 1 contains an authentic video of Actor A and Set 2 contains a deepfake video of Actor A, Set 1 contains a deepfake video of Actor B and Set 2 contains an authentic video of Actor B, etc).

Before stimulus videos are presented, two validity check questions will be presented. Participants have to get this questions correct before they are able to move into viewing the stimulus videos. Validity check 1: "A deepfake is a video in which..." (a person is telling lies; a person is talking about fakes; the face and/or voice of a person has been manipulated using artificial intelligence). Validity check 2: "Which of the following statements is correct?" (Each video has a 20% chance to be a deepfake; Each video has a 50% chance to be a deepfake; All the videos are deepfakes).

Stimulus videos will be presented one at a time, with participants being asked the following after each video: “Is this video a deepfake or real?" (This video is a deepfake; This video is real) and “What is your confidence that you guessed correctly?” (measured on a continuous slide ranging from 50 = As confident as flipping a coin and 100 = 100% sure).

After the detection activity, participants’ overall detection confidence will be gauged (“Out of 20 videos, how many videos do you think you guessed correctly?”). Participants will then be asked some demographic questions (age, gender, nationality, and highest level of formal education) and whether they have completed this study previously. Participants will then be presented with feedback on their performance in the detection task.

The experiment will be anonymous; participants will not be asked to provide any identifying details. The study would likely take most participants less than 20 minutes to complete.

No files selected

Randomization

Simple randomisation will be employed to randomise participants into the control group or detection strategies group.

Randomisation will be performed using Qualtric's "randomizer" function.

Sampling Plan

Existing Data

Registration prior to creation of data

Explanation of existing data

No response

Data collection procedures

We will seek to recruit a general community sample of English-speaking adults. Any one who are 18 years or older and able to read and understand the study will be eligible to participate.

Our primary means of recruitment will be advertising on social media (primarily Facebook and Instagram), in order to engage a large and diverse potential pool of participants (Kosinski et al., 2015).

Specially, we plan to create a Facebook page for the study. We will then create a post and pay to promote that post . Facebook will allow us to select characteristics of those we want to promote the post to (e.g., adults living in English-speaking countries such as Australia, New Zealand, the UK, Canada, and the US). We hope that this will seed interest in the project and result in participants sharing the post to their networks (snowballing recruitment).

The researchers may also share this post on their private social media accounts. Text is also included in the information sheet encouraging participants to pass on the study details to others who may be interested in participating (to further promote sharing of the study).

Should these recruitment methods fail to garner the required sample size, we will then start to recruit participants from undergraduate research participation pool at the authors' institution (James Cook University). If we do collect student participants in this way, this will be noted in any publications produced.

Ethics approval to collect data has been granted for a period of 1.5 years (from April 2022).

- deepfakes power analysis.rtf

Sample size

We are aiming to recruit a sample of at least 250 participants.

Sample size rationale

A priori power analysis was conducted in G*Power to determine the total sample size needed to have adequate power to address the primary hypothesis (H1), assuming a small-to-medium effect (d = .35). This analysis indicated that an N of 204 would be sufficient (if using an α of .05, a 1-β of .80, and performing one-tailed independent samples t tests). We will aim to recruit at least 250 participants to account for non-complete responses etc.

The same analysis would require a N of 620 if assuming a small effect.

The protocols entered into G*Power for this analysis have been uploaded as a .rtf file.

Stopping rule

Ethics approval to collect data has been granted for a period of 1.5 years (from April 2022), however we are hoping to collect an adequately sized sample before this end date.

The number of participants to have completed the study will be checked intermittently (at least weekly). We will continue to collect data until a sample of at least 250 is collected. If data collection proves to be easier than anticipated, we will terminate data collected once 650 participants have completed the study; as this would adequately power the study to detect a small effect (see above).

Variables

Manipulated variables

We have manipulated whether participants were presented with strategies for detecting deepfakes. This categorical variable has two levels: detection strategies presented and no detection strategies presented (control).

The following will be presented to the detection strategies group verbatim:

Here are some strategies to help you spot the deepfake videos:
1. Pay attention to the face. High-end Deepfake manipulations are almost always facial transformations.
2. Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? Deepfakes are often incongruent on some dimensions.
3. Pay attention to the eyes and eyebrows. Do shadows appear in places that you would expect? Deepfakes often fail to fully represent the natural physics of a scene.
4. Pay attention to the glasses. Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, Deepfakes often fail to fully represent the natural physics of lighting.
5. Pay attention to the facial hair or lack thereof. Does this facial hair look real? Deepfakes might add or remove a moustache, sideburns, or beard. But, deepfakes often fail to make facial hair transformations fully natural.

6. Pay attention to facial moles. Does the mole look real?
7. Pay attention to blinking. Does the person blink enough or too much?
8. Pay attention to the size and color of the lips. Does the size and color match the rest of the person's face?

These detection strategies are produced by the MIT Media Lab (https://www.media.mit.edu/projects/detect-fakes/overview/).

No files selected

Measured variables

Outcome variables:

Video categorisation: Is this video a deepfake or real? This video is a deepfake; This video is real.

Detection confidence (per video): What is your confidence that you guessed correctly?
Slider scale anchored by 50% = "I am as confident as correctly guessing a coin flip" and 100%= "I am a 100% sure"

Detection confidence (overall): Out of the 20 videos, how many videos do you think you guessed correctly? _____

Engagement with stimulus videos: measured in time (in seconds) spent on page.

Demographic variables (for reporting on nature of sample):

What is your age? _____

What is your gender? _____

In what country do you currently reside? _____

What is your highest level of formal education?
□ Primary School
□ High School
□ TAFE/Other Vocational Studies
□ Some undergraduate study
□ Undergraduate degree
□ Some postgraduate study or a postgraduate degree

Validity checks:

A deepfake is a video in which...
...a person is telling lies
...a person is talking about fakes
...the face and/or voice of a person has been manipulated using artificial intelligence

Which of the following statements is correct?
Each video has a 20% chance to be a deepfake
Each video has a 50% chance to be a deepfake
All the videos are deepfakes

This is a serious academic study and the results will contribute to the understanding of deepfake detection. To maintain the integrity of the study's data, please indicate if you have done this study previously, so we can exclude your data.
□ This is the first time I have done this study.
□ This is NOT the first time I have done this study.
□ I am unsure.

No files selected

Indices

The video categorisation questions will be used to compute an overall score representing the percentage of videos which the participant correctly categorised ("overall detection accuracy").

Per video detection confidence scores will be used to compute "average detection confidence" scores (not to be confused with "overall detection confidence"; see above)

Time spent on each stimulus video page will be averaged to create "level of interaction with stimulus videos" scores.

"Percentage of videos categorised as real" will be calculated by adding the number of videos categorised as real and dividing by the total number of videos.

No files selected

Analysis Plan

Statistical models

H1: Participants provided with detection strategies will show greater detection accuracy relative to the control group.

An independent samples t test (or non-parametric equivalent/bootstrapping) with detection strategies group as the predictor variable and overall detection accuracy as the outcome variable.

H2: Participants provided with detection strategies will show greater detection confidence relative to the control group.

An independent samples t test (or non-parametric equivalent/bootstrapping) with detection strategies group as the predictor variable and overall detection confidence as the outcome variable.

An independent samples t test (or non-parametric equivalent/bootstrapping) with detection strategies group as the predictor variable and average detection confidence as the outcome variable.

H3: There will be a bias toward categorising stimulus videos as real.

One-sample t test comparing the percentage of videos categorised as real to 50% (baseline, given that half of all videos are real)

No files selected

Transformations

We will opt to apply non-parametric tests as opposed to transforming non-normal data.

Inference criteria

P values will be used to determine statistical significance (with an alpha of .05 being uniformly applied). Given that directional hypotheses are specified, one-tailed tests will be used. Effect size measures will be reported for all tests.

Data exclusion

Those who indicate that it is not the first time that they have done the study (or that they are unsure) will be excluded.

Those who complete the study much faster than average (under half the sample's 5% trimmed mean time to complete) will be excluded.

A combination of graphical methods and the outlier labelling rule with a 2.2 multiplier will be used to detect univariate outliers. Extreme univariate outliers will be windsorised.

Missing data

Participants who do not complete any of the detection items will be excluded.

For those with incomplete data, we will conduct a missing values analysis to determine the degree and type of missing data (e.g., missing at random). Item-level imputation methods may be applied if applicable (based on degree and type of missing data). This will be reported in any publications resulting from this study.

Exploratory analysis

We plan on conducting exploratory analyses to investigate the research questions
.
RQ1: Is level of interaction with stimulus videos associated with detection accuracy?

A Pearson's product moment correlation (or non-parametric equivalent) between overall detection accuracy and interaction with stimulus videos. Two-tailed test.

RQ2: Is detection confidence associated with detection accuracy?

A Pearson's product moment correlation (or non-parametric equivalent) between overall detection accuracy and average detection confidence. Two-tailed test.

A Pearson's product moment correlation (or non-parametric equivalent) between overall detection accuracy and overall detection confidence. Two-tailed test.

Other

Other

Literature cited

Kobis, N. C., Dolezalova, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. iScience, 24(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist, 70(6), 543-556. https://doi.org/10.1037/a0039210