The Detect Fakes Experiment: DeepFakes, Can You Spot Them (#40298)

Created: 05/01/2020 02:07 PM (PT)
Public: 12/08/2021 05:09 PM (PT)

Author(s)
Matt Groh (MIT) - groh@mit.edu
Ziv Epstein (MIT) - zive@mit.edu
Rosalind Picard (MIT) - picard@media.mit.edu

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
The main question is how do emotion and tacit knowledge play a role in a visual perception task to differentiate DeepFake videos with AI generated facial manipulations from real videos.

A secondary question is how do human participants' performance compare to a machine learning model's performance?

3) Describe the key dependent variable(s) specifying how they will be measured.
The key dependent variable is a binary variable defined by whether the participant guessed correctly or not. The experiment follows a two-alternative forced choice (2AFC) method whereby participants see two videos presented and are instructed to guess which of two videos is fake. We will evaluate both how people discern and learn to discern fake from real videos.

A secondary key variable is response time. Response time is measured by the number of seconds it takes to respond to each 2AFC task.

4) How many and which conditions will participants be assigned to?
We will randomize interventions both across individuals and within individuals.

Across Individuals

There are four randomized interventions across individuals: (i) three emotion priming interventions: (a) control (b) anger (c) anxiety (d) happy, (ii) one reflection prime (a) control (b) delayed load time on wrong answers, (iii) one visual perception manipulation (a) upside down video on 3rd and 8th video (b) upside down video on 4th video and 8th video, (iv) one crowd-sourced hint prime (a) control (b) crowd-sourced hint. Participants will be cross-randomized to the four interventions.

The crowd-sourced hint will be available later in the course of the experiment once enough data has been collected to offer crowd-sourced hints. If we have over 20,000 users participate, we will also include two more interventions: (a) varying whether we give participants feedback and (b) varying whether we give participants general hints.

Within Individuals

The order in which videos are seen will be randomly assigned to each individual. This randomization will include three sets of video difficulty as determined by machine learning model performance.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will examine treatment effects on the accuracy of participant, j, on manipulated video, i. We plan to focus our analysis on the first ten pairs of videos seen by a user.

The first model hypothesizes that the best fit linear model will involve a logarithmic transformation and the second model estimates treatment effects separately for each image position. All models use Huber-White (robust) standard errors clustered at the participant level with video fixed effects.

We will analyze how emotion priming affects accuracy and improvement in accuracy (and response times).

Previous research on emotion and fake news suggests that emotion impedes discernment (Martel, Pennycook, and Rand 2019). While the recent research on emotions and fake news suggests a null hypothesis that emotions negatively affect discernment, we form our hypothesize based on the Appraisal Tendency Framework (Lerner and Keltner 2001). Specifically, we hypothesize that inducing anger increases discernment relative to the control group by increasing attention and depth of processing (Lerner & Tiedens 2007). We hypothesize that inducing anxiety decreases discernment relative to the control group by adding cognitive load due to uncertain existential threats (Han, Lerner, Keltner 2007). We hypothesize that inducing happiness increases discernment relative to the control group because it is associated with an elevated sense of certainty (Han, Lerner, Keltner 2007).

We will analyze how the machine learning model's confidence predicts the human participant's accuracy (and response times).
We will analyze how perceptual manipulations (rotating the video) affects accuracy and improvement in accuracy (and response times).

We hypothesize decreased discernment in upside down videos.

We will analyze how crowd-sourced attention priming affects accuracy and improvement in accuracy (and response times).

We hypothesize increased discernment in attention priming interventions.

We will analyze how the forced delay upon wrong answers affects accuracy and improvement in accuracy (and response times).

There is an option to ask for a hint. We will also look at the propensity of clicking the hint button based on image order and how clicking on the hint button predicts accuracy on the current video pair and future video pairs.

We will also look at effects of interventions regardless of image order.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We want to exclude participants who do not participate in good faith (e.g. the participants who click haphazardly without paying attention to the videos). We will exclude observations where individuals answered too fast to have paid attention to the task. We expect that discernment takes on average 5-45 seconds per pair of videos because some videos can be discerned on their first frame and some videos need to be replayed multiple times. Based on the distribution of response times, we will identify a cut-off somewhere between 2-5 seconds that fairly identifies arbitrary clicking patterns.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
The sample size will be determined by the popularity of our experiment. We will export and analyze data 3 months after the thousandth person participates in the experiment, or after the 100,000th person participates, whichever comes first. We will not engage in optional stopping. All observations collected within this experimental window will be analyzed.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
For the analysis, the matrix of covariates will include video placement on the website (left/right on desktop or top/bottom on mobile), time, location, mobile vs. desktop, type of browser, high-level video features (lighting conditions, type of manipulation (e.g. adding mustache vs. brightening eyes vs. adding wrinkles vs. adding glare to glasses, etc.), perceived race/gender/age of the people in the videos), and reported demographics of users. We will look at heterogeneous effects of these covariates on discernment and learning rates. For example, do race and gender (or does mobile phone vs. desktop) play a role in the difficulty for people to discern DeepFakes?