Detecting Deepfakes – An experiment on the role of motivation (#46610)

Created: 08/24/2020 02:24 AM (PT)
Public: 10/07/2021 02:18 AM (PT)

Author(s)
Ivan Soraperra (University of Amsterdam) - i.soraperra@uva.nl
Nils Köbis (University of Amsterdam) - n.c.kobis@gmail.com
Barbora Dolezalova (University of Amsterdam) - Dolezalova.Bara@seznam.cz

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?
This study looks at the effect of motivation on the ability to detect deepfake videos. Overall, we have 3 treatments: Control Treatment where motivation is not manipulated; Awareness Treatment where motivation is manipulated by highlighting the negative consequences that deepfakes can have; Financial Incentive Treatment where motivation is manipulated by rewarding the correct classification of videos. We have the following hypotheses:
Hypothesis 1: Compared to the Control Treatment, participants in the Awareness Treatment perform significantly better at detecting deepfake videos.
Hypothesis 2: Compared to the Control Treatment, participants in the Financial Incentive Treatment perform significantly better at detecting deepfake videos.

3) Describe the key dependent variable(s) specifying how they will be measured.
The dependent variable is the number of correctly identified videos (either deepfake or not).

4) How many and which conditions will participants be assigned to?
There will be 6 conditions in a 3 (between subjects: Control vs. Awareness vs. Financial Incentive Treatments) x 2 (within subjects: Fake videos vs. Authentic videos) design. The treatments consist of
- Control : No incentivization for accuracy
- Awareness: Motivate participants to answer accurately by raising awareness about harmful consequences of deepfakes via an article
- Financial Incentives: Motivate participants to answer accurately by providing financial incentives for accuracy
In each treatment participants watch 16 videos. The probability of a video to be a deepfake is 50%.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
To analyze the overall effect of motivation, we will compare the individual rate of accuracy across treatments using a one-way ANOVA.
To test the heterogeneous treatment effects for fake and authentic videos we will use a linear probability model with robust standard errors clustered at individual level. As explanatory variables we will include the treatment dummies, the indicator variable for the authenticity of the video (dummy: fake vs. real), and their interaction. Logit and Probit models will be used to assess the robustness of the results.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We do not plan to exclude participants from the analysis, but we will check the robustness of the results to the inclusion of control variables assessing whether the subject clicked on the video and/or is able to recall the content of the video, which is assessed by several control questions after watching the video.

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.
In total we will collect 3360 decisions from 210 participants (16 decisions per participant and 70 participant per treatment).

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
As an exploratory question, we will collect the beliefs of subjects about their ability to detect deepfakes. This is done in two ways: (i) asking the confidence level of a correct detection after each video; (ii) asking to guess the overall number of videos they correctly identified at the end of the experiment (incentivized with a small bonus for a correct guess). We expect to observe overconfidence and a positive interaction between overconfidence and the motivation treatments.
Finally, as a second exploratory question, we will look at the differential effect of Motivation on the ability to detect the fake and the authentic videos. Here we expect that motivation treatments can have a positive effect on the ability to detect fake video and a negative effect on the ability to detect authentic videos due to the stronger focus on deepfakes.