**Ian:**

Sean to explicate what the claim on the basis of this is, what analysis would implement it, do implement it in the data from studies 1-6, and then we mutually red-team it for whether it could be used to undermine our more central claims.

NB see your comment below that reads “I don’t think these analyses are currently conducted (unless I am missing them in the analyses files.)…” for relevant content

**Sean:**

Ok. So I would like to know the following (I’ve also been referring to this in the Science and SM manuscripts):

*Question 1*: Are people aware of Deepfakes prior to the study?

To know what percentage of people were aware of Deepfakes before taking part in the study. I’m guessing this is just % for overall group

*Question 2*: Does awareness differ as a function of exposure to a Deepfake?

Same question as above but broken down for Deepfake and Genuine conditions. I’m curious to know if being exposed to a Deepfake actually alters awareness relative to non-exposure.

I’d like to know if awareness was different for the Deepfake and Genuine conditions.

I’m guessing this is a Chi Square test for differences in awareness between the two groups?

*Question 3*: The real question I’m interested in: can one be aware of Deepfakes as a concept but be poor at detecting a Deepfake when exposed to it. This is a sexy finding for me (i.e., simple awareness does not make you immune to Deepfakes and their influence).

Based on eye balling the data it seems that most people are aware of what Deepfakes are prior to the study. But then if we look at Deepfake detection in the Deepfake condition we see that most people are NOT detecting that the video was a Deepfake.

If I am right, then I would like to make the claim that most people are aware that Deepfakes are “out there” (i.e., that they exist). But they are poor at detecting when they are actually faced with one themselves.

I don’t know what stats we need to run in order to test this. In my head a simple comparison of (a) Deepfake concept awareness % scores and (b) Deepfake detection % scores in the Deepfake condition would be sufficient.

But I’m open to suggestions here.

**Ian:**

Q1

What if the percent of people aware differs greatly between our studies, will this be seen to impact results or not, even if results replicate? I.e., are our claims here correlated in any way?

Is it possible or likely we’ll be asked to include this as a moderator of our effect? And if so do they hold up to it? Are we powered to answer that question?

Q2

Hypothetically Q2 analysis is either a chi-square or a logistic regression. You’re saying that awareness(binary) ~ experient\_condition(genuine or Deepfake).

However, this experiment cannot answer this question directly. We don’t give people an exemplar of a Deepfake, some people are exposed to one and no one is (now) given any feedback about whether what they saw was really a deepfake or not. And, most people don’t detect that it was a Deepfake. The study just isn’t capable of meaningfully answering Q2 – it’s best left for a purpose built study.

**Sean**:

Q1

There are currently three studies with information on awareness rates and from eye balling that data, awareness rates appear stable across all three studies. Sure there is a risk it might be different in Study 7, but the risk is low given the stability across studies in awareness rates. If we were going in blind here then sure – I take your point. But we are not. We are going in with three studies worth of data.

Now you ask “is it a moderator of the effects”? Again there are already three studies present with data to that question. Also if the reviewer asks during review if awareness is a moderator, we can state that we cannot answer that question because we did not power the study to do so.

Q2

Thanks for highlighting the problems with the question. I’d also like to brain storm some solutions here as well. For instance, I see no reason why one cannot look at awareness rates in the genuine group, awareness rates in the Deepfake group, and compare the two. This is surely possible given the data, right? Now I agree that one cannot make *strong* conclusions on the basis of such a comparison, but one can surely state if there is a difference in awareness between these two conditions. One can also speculate about why such a difference emerged, and suggest future studies experimentally explore it. I see value in asking if people were aware of Deepfakes prior to the study, and if awareness rates differ between the genuine and Deepfaked conditions.

**Ian: NEW STUFF FROM HERE**

Q2

I’m saying that one can’t even make weak conclusions based on this data. Everything else about this study is super tight and well evidenced, and we would be introducing a wild speculation in here. It also brings the ability to answer Q1 and Q3 into doubt: if people's recollection of whether they have heard of the concept of deepfakes prior to the study is malleable based on merely being exposed to one - and exposed to it unconsciously as 80% of people dont detect it - then it wasnt a good measure of awareness in the first place. If the goal is to suggest hypotheses for future research, that can already be done using the data from the previous experiments that contained this question, without adding a very weak question to a confirmatory study. I would strongly say we drop this one.

Q3

The way that nested comments displayed in Word meant I didn’t see Q3 until now. From above:

*“I would like to make the claim that most people are aware that Deepfakes are “out there” (i.e., that they exist). But they are poor at detecting when they are actually faced with one themselves.”*

The data say just over 50% of people are aware of deepfakes, so you could say that just over half of people aware of them but it would be a stretch to say “most”. On the other hand, this question now let’s me see the utility of the awareness question beyond simply reporting a percentage awareness.

One option is that we apply the classification stats used in H3 to the subset of people who are aware of the concept prior to the study and argue that the results hold in this subset, i.e., “do people who are aware of the concept make accurate/informed judgements about content”/ “are the poor classifications in the full sample just driven by people who don’t know what a Deepfake is”. I ran them to see:

Full sample: N = 576

|  |  |  |  |
| --- | --- | --- | --- |
| **variable** | **median** | **ci\_lower** | **ci\_upper** |
| balanced\_accuracy | 0.59 | 0.56 | 0.62 |
| false\_negative\_rate | 0.73 | 0.69 | 0.78 |
| false\_positive\_rate | 0.08 | 0.04 | 0.12 |
| informedness | 0.19 | 0.13 | 0.25 |

Concept aware subset: N = 220

|  |  |  |  |
| --- | --- | --- | --- |
| **variable** | **median** | **ci\_lower** | **ci\_upper** |
| balanced\_accuracy | 0.59 | 0.54 | 0.64 |
| false\_negative\_rate | 0.72 | 0.63 | 0.8 |
| false\_positive\_rate | 0.09 | 0.04 | 0.15 |
| informedness | 0.19 | 0.09 | 0.28 |

So, descriptively at least, the results are quite similar. Because we don’t use cutoffs for inference tests in H3, I think it would be fine to simply hypothesize that that results are comparable for all four metrics when considering only the subset of concept aware people. I.e., being aware of the concept doesn’t make you obviously better at making accurate and informed judgements, and the results in the full sample aren’t merely driven by people not having a clue what a DF is.

This feels sexy but substantiated, and has implications for future studies and maybe intervention efforts (i.e., awareness raising might not be enough).

However, be careful not to phrase the question or answer as “are people who are aware of the concept **better** than those who aren’t” as this is much much more difficult to ask. The above analysis just shows results are robust in the aware subset, they don’t compare the subsets. That would be a whole other thing – sth like a Bayesian logistic model which is trying to find evidence for the null, which is very tricky for reasons I wont get into.

**Sean**: Ok. I’m trying to understand your reasoning here. As far as I can see your issue is that responding on the prior awareness of Deepfakes question *might* be influenced by exposure to a Deepfake. If so, then the awareness rate should be “uncontaminated” in the Genuine condition seeing as they are never exposed to a Deepfake (responding on the awareness question cannot be influenced by Deepfake exposure in this case). However, response to the awareness question might be “contaminated” in the Deepfake condition (because they were actually exposed to a Deepfake and this might influence their responding on the awareness question). Is this where you see an issue?

My responses is this. If exposure to a Deepfake does influence performance on the awareness question as you hypothesize, then surely a comparison between the genuine and Deepfaked groups would reveal this (i.e., the two groups are otherwise identical: thus if there is a difference between the two groups then it has to be due to the Deepfaked video). If there is no difference between the two conditions then surely this provides evidence to support the idea that exposure to the Deepfake does not artificially raise natural awareness.

For me asking and answering this question is a win-win either way. If there is a difference between the two groups in awareness rates, and if people in the Deepfaked condition who are aware still show effects, then it suggests that prior awareness is not a protective measure – something useful to know for future intervention.

If there is no difference between the two groups in awareness rates, then one can talk about how likely it is that people are aware of Deepfakes before the study, and examine if people who are (a) previously aware of Deepfakes and (b) aware they were exposed to a Deepfake in this study still show an effect. This would be the strongest argument that awareness and detection are not protective measures as often assumed.