**Ciaran’s comments on the data analysis/R code**

Comment 2 (analyses\_exp\_1-6\_preregistered.Rmd - Differences from zero by source\_valence condition):

* All the IAT\_D2 values for negative condition have much different impact/significance than for the positive condition. I can imagine the effect being moderated by a lot of confounders. Will we need to comment on a reason why?
* mean\_self\_reported\_evaluation has similar effects for both positive and negative. Contradiction to above observation?
* mean\_intentions (only in experiment 6) negative mean+std much different (lower) than positive effect. Both positive and negative valence means are negative in sign (different from self evaluations). E.g., even though people evaluate Chris similarly, they dont intend to share videos.

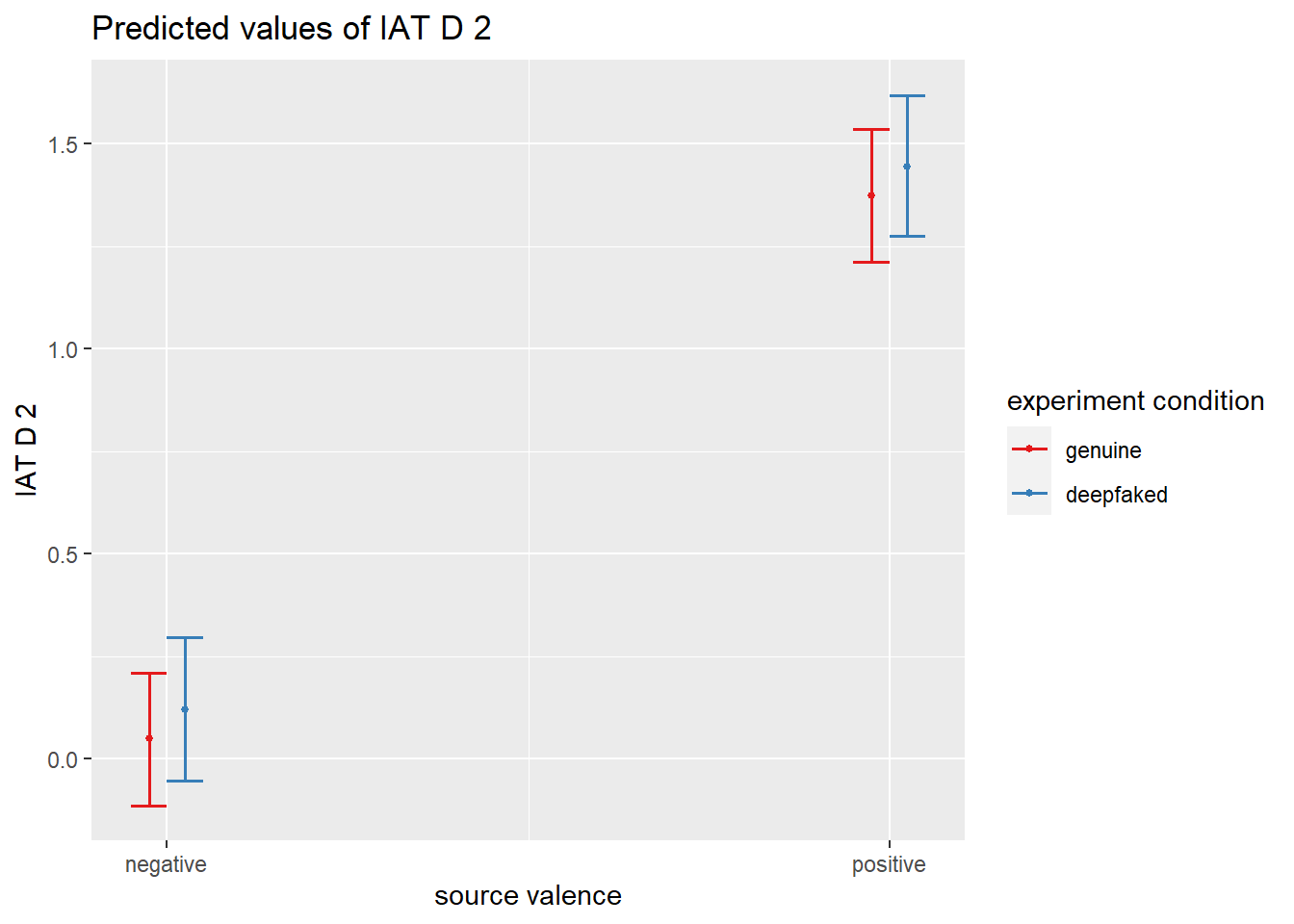
Comment 3 ( analyses\_exp\_1-6\_preregistered.Rmd - Differences between genuine and deepfaked content (after rescoring for source\_valence)

* For Experiment 6 ONLY (and not 1-3): mean\_self\_reported\_evaluation\_recoded\_for\_source\_valence is different between deepfake and genuine. Is this consistent with statistical power value, e.g., expect some false negatives?

Comment 4 (analyses\_exp\_1-6\_preregistered.Rmd - Correlations):

* Has some comments that might need removed before final submission to openscience.

Comment 5 (analyses\_exp\_1-6\_exploratory.html – IAT D2 plot of postive/negative vs genuine/deepfake):

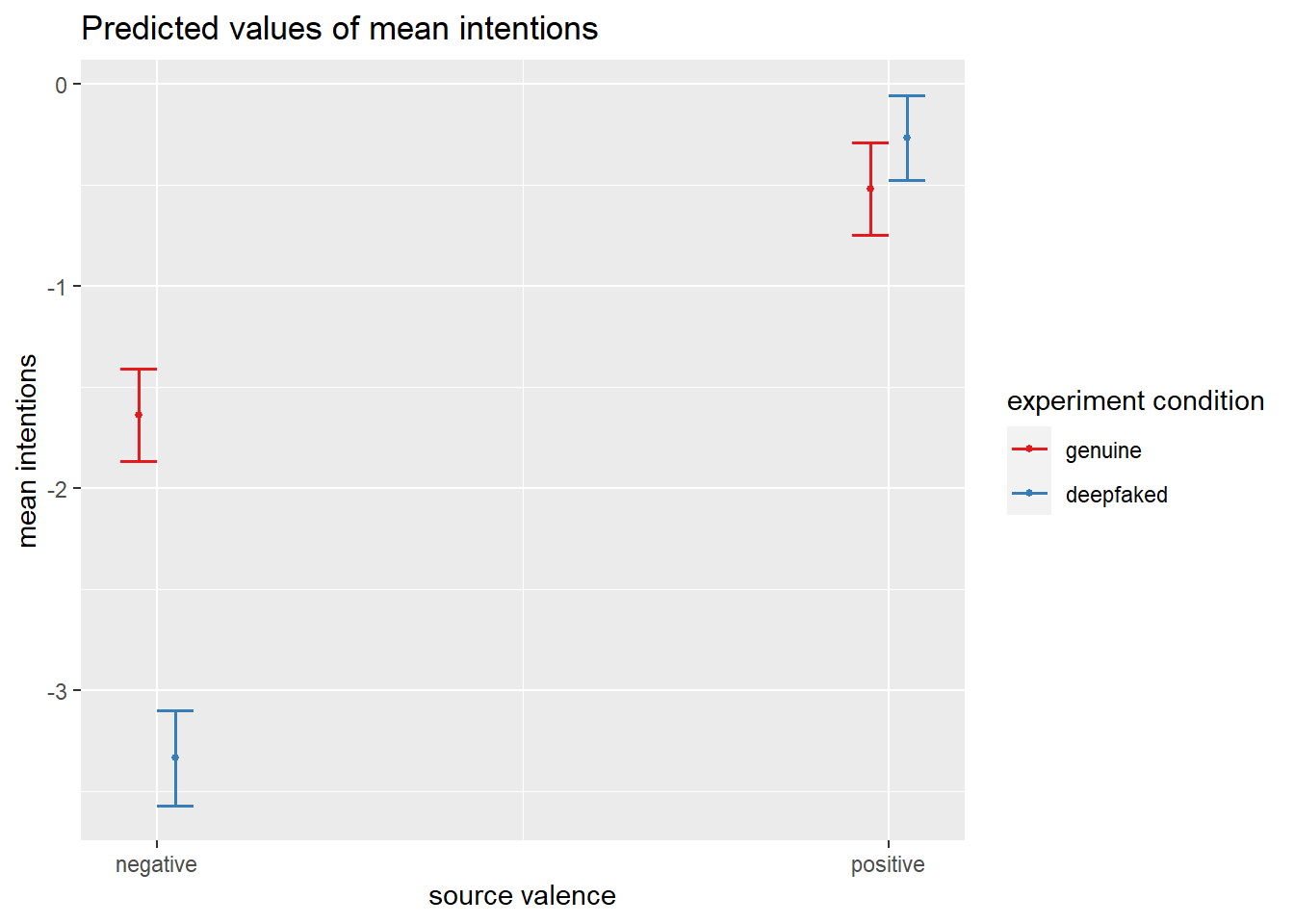


For the IAT D2 plot, the negative condition does not have an impact (both genuine and deepfake overlap with zero). The positive condition does have a non-zero impact. As such, it is only for the positive condition that it is useful to claim that the deepfake has the same impact as the as the genuine content (e.g., non-inferior is passed). For the negative condition, non-inferiority looks like it will still pass, but the impact is zero, so not very useful to say its non-inferior but the content had no effect. I understand that we’re taking the difference between the negative and positive as the measure of effect, though something to think about.

Comment 6 (analyses\_exp\_1-6\_exploratory.html)

* Is the 90% lower bound on the CL in the non-inferiority tests,instead of 95%, because the difference is the non-inferiority margin? If so, should we say this?

Comment 7 (analyses\_exp\_1-6\_exploratory.html)



This particular data looks odd compared to the rest. Mostly that the negative has a difference between genuine and deepfake. I wouldnt expect people to share the video more if the (negative) content is the same. I can imagine people sharing more if they detected it was a deepfake (which they didn’t on large, and didn’t for any other metric = IAT D2 or self evals). If I was a reviewer, I would comment on this. I agree that this can happen, and it is possible to get a good fit with the interaction effects, this just looks weird taking everything else into account.

Comment 8 (analyses\_exp\_1-6\_exploratory.html)

| awareness | detection | detection\_probability\_MAP | CI\_95\_lower | CI\_95\_upper |
| --- | --- | --- | --- | --- |
| TRUE | TRUE | 0.144 | 0.067 | 0.250 |
| TRUE | FALSE | 0.379 | 0.260 | 0.567 |
| FALSE | TRUE | 0.058 | 0.010 | 0.115 |
| FALSE | FALSE | 0.391 | 0.231 | 0.538 |

* Probability of detecting deepfake if unaware: 0.058
* Probability of detecting deepfake if aware: 0.144

The other two columns are also worth mentioning: you’re 2.6 more likely to detect a deepfake if you’re aware of the concept of deepfakes, but you’re still a lot more likely to not detect them regardless of whether you’re aware or not (0.391 and 0.379).

Comment 9 ( analyses\_exp\_1-6\_exploratory.html - H5a)

* Can I verify that the experimental group does not contain people who are aware AND detected deepfakes. I currently think the experimental group looks like just the people who were aware of deepfakes (which would be correct for RQ5, but the text around H5a, H5b, H5c makes it sound like the experimental group also detected).

Comment 10 (analyses\_exp\_1-6\_exploratory.html – H5b)

* “(model 4) to estimate a 95% Confidence Interval on standardized effect size change in IAT D2 scores between Source Valence conditions in the **Deepfaked** video condition subgroup”

Can I very the deepfake control group is correct, as in H5a the control group says its the genuine condition?

Comment 11 (analyses\_exp\_1-6\_exploratory.html –H7a,b,c)

* Would it be useful to ask both questions: if you detect a deepfake, are you just as influenced as compared to I) genuine content, or ii) people who were given a deepfake and did not detect it? E.g., the effect\_deepfake and effect\_deepfake\_detected look like they’d be similar if they had comparable stats.

Comment 12 (power\_analysis\_via\_simulation.Rmd)

* H2a,b,c all have a power of approx 0.80%. Is this usually sufficient?