The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence

Proceedings of the Association for Computing Machinery on Human-Computer Interaction, Volume 6, November 2022

Paper presentation

NTU IALab, December 2022

Antoni Maciąg

Overview

- Paper entirely non-technical
- Experiments conducted on the impact of AI warnings and others’ opinions on people’s ability to detect and willingness to share true and fake news

Motivation

- Previous studies exist on social impact on news sharing, and also on the impact of AI credibility indicators
- This paper analyzes how these factors interact

Research questions

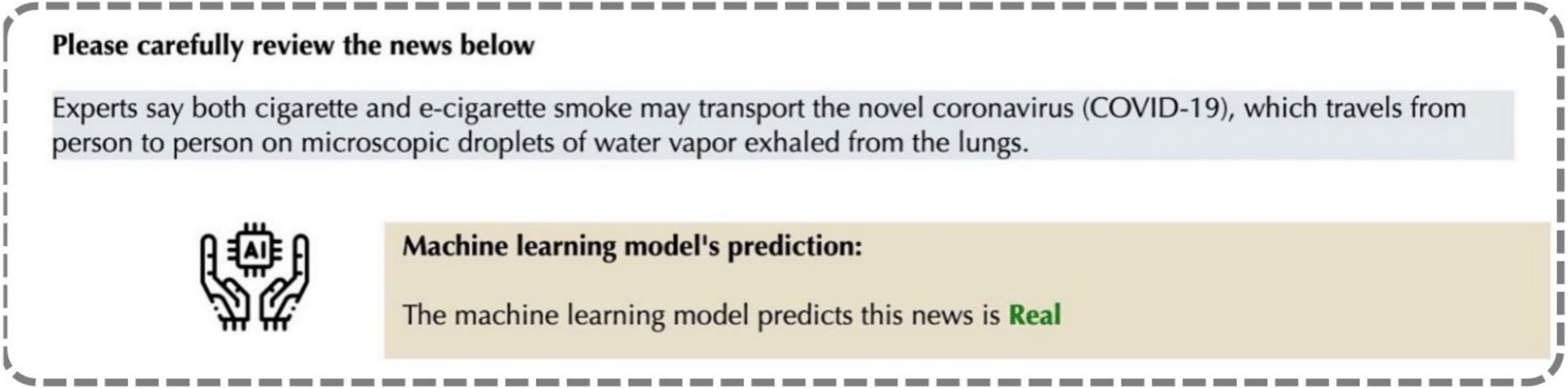
- RQ1: How does the presence and timing of AI’s indications change people’s accuracy in detecting misinformation?
- RQ2: How does the presence and the timing of AI’s indications change people’s willingness to spread the news?
- RQ3: How does the effect of AI indications change between the cases when social influence is present and not?

Experiment 1

Procedure for test subjects

- Step 1: be presented with a piece of news
- Step 2: initial judgement, report confidence 1-7
- Step 3: see chronological list of other people's judgements
- Step 4: final judgement and confidence
- Step 5: decide how likely you are to share (0% - 100%) (previous research indicates it is sensible)

Step 1



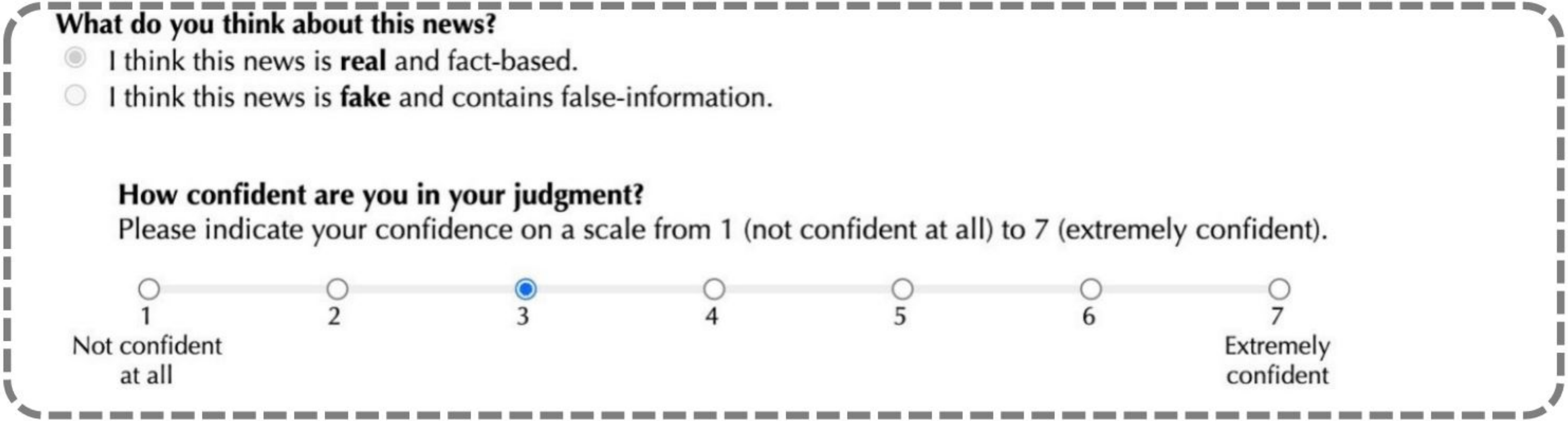
Please carefully review the news below

Experts say both cigarette and e-cigarette smoke may transport the novel coronavirus (COVID-19), which travels from person to person on microscopic droplets of water vapor exhaled from the lungs.

Machine learning model's prediction:

The machine learning model predicts this news is **Real**

Step 2



What do you think about this news?

● I think this news is **real** and fact-based.

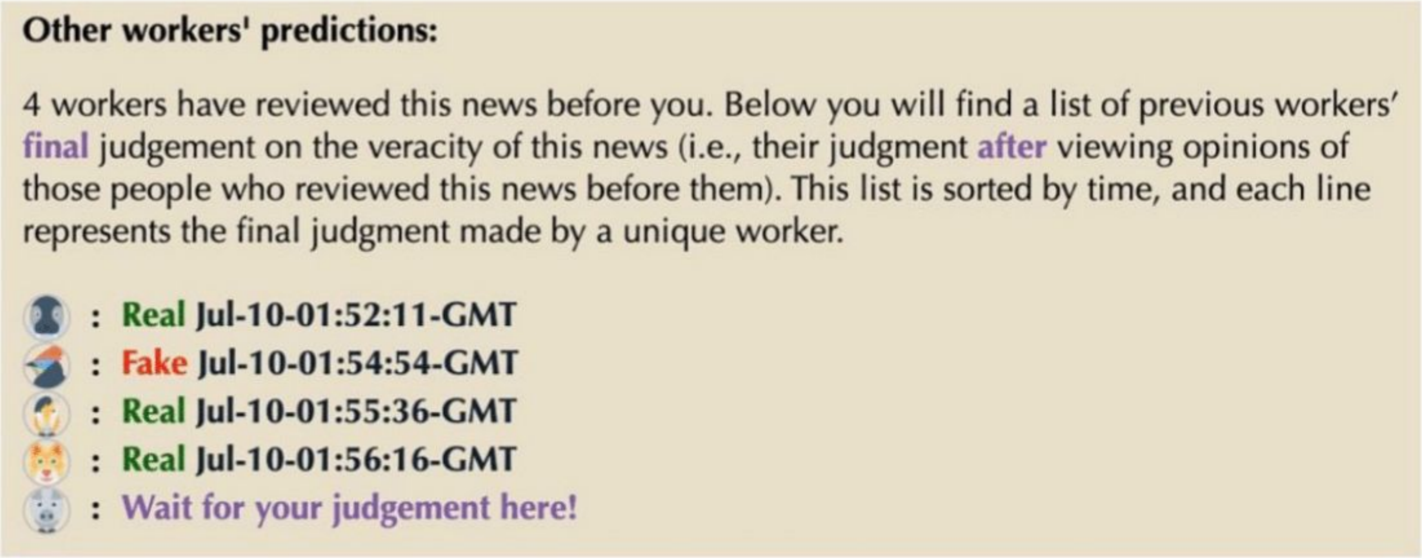
○ I think this news is **fake** and contains false-information.

How confident are you in your judgment?

Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○	○	●	○	○	○	○
1	2	3	4	5	6	7
Not confident at all						Extremely confident

Step 3

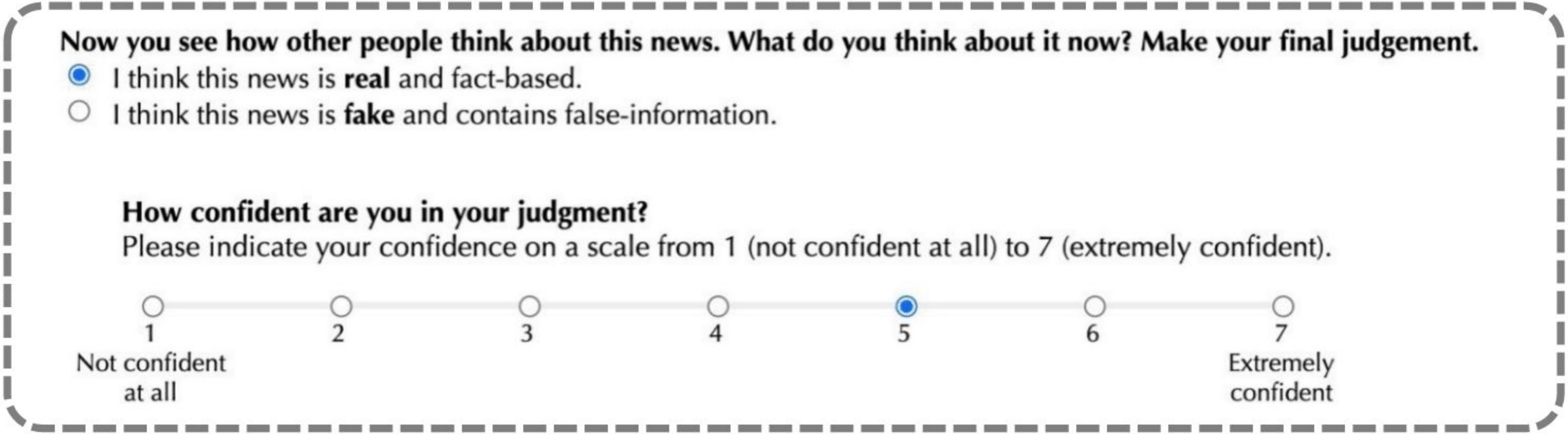


Other workers' predictions:

4 workers have reviewed this news before you. Below you will find a list of previous workers' final judgement on the veracity of this news (i.e., their judgment after viewing opinions of those people who reviewed this news before them). This list is sorted by time, and each line represents the final judgment made by a unique worker.

: **Real Jul-10-01:52:11-GMT**
: **Fake Jul-10-01:54:54-GMT**
: **Real Jul-10-01:55:36-GMT**
: **Real Jul-10-01:56:16-GMT**
: **Wait for your judgement here!**

Step 4



Now you see how other people think about this news. What do you think about it now? Make your final judgement.

● I think this news is **real** and fact-based.

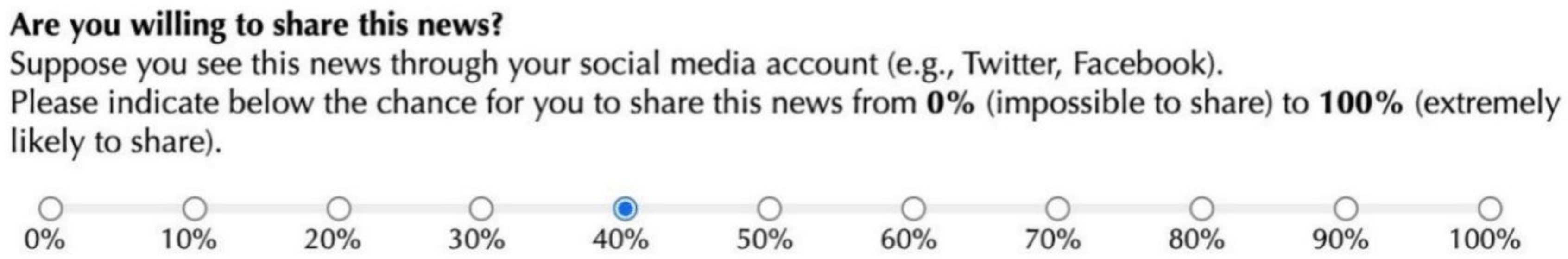
○ I think this news is **fake** and contains false-information.

How confident are you in your judgment?

Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○	○	○	○	●	○	○
1	2	3	4	5	6	7
Not confident at all						Extremely confident

Step 5



Are you willing to share this news?

Suppose you see this news through your social media account (e.g., Twitter, Facebook).

Please indicate below the chance for you to share this news from **0%** (impossible to share) to **100%** (extremely likely to share).

○ 0% ○ 10% ○ 20% ○ 30% ● 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Treatments

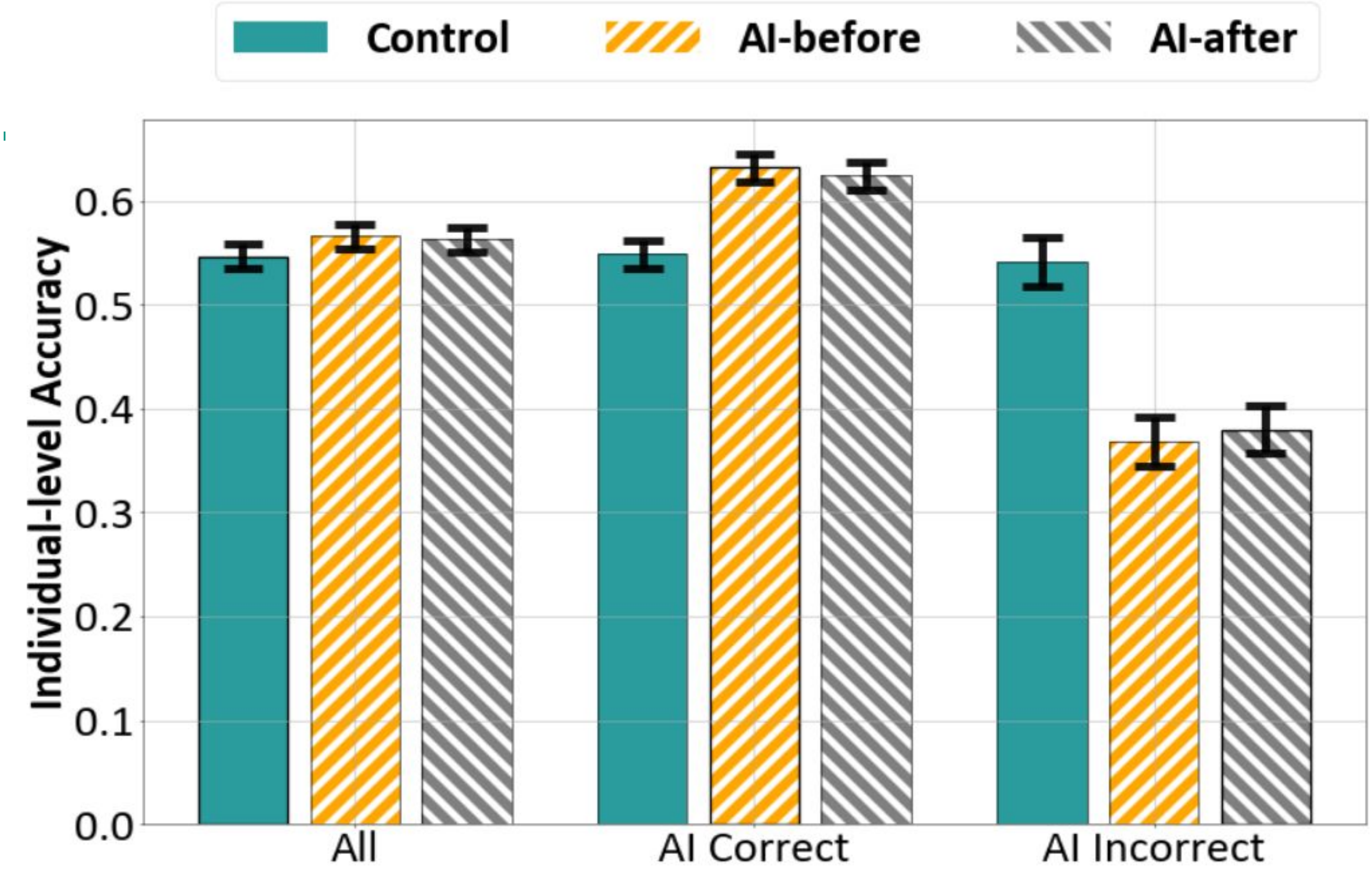
- Control: had no access to indications
- AI-before: saw indications before crowd’s opinion - Step 1
- AI-after: saw them with crowd’s opinion

More about methodology

- Based on the information cascade experiment (economics)
- Amazon Mechanical Turk, extra money for correct prediction
- Dataset: 40 news stories - 20 true and false each, COVID-19
- Model tuned to 75% accuracy, subjects knew about it

Remarks we can make on methodology

- Payment for a correct judgement was 0.05$. I guess it is still more profitable to spam random answers.
- Subjects are only shown AI prediction and know its accuracy. Why tune a model at all, instead of randomly choosing answers with the same accuracy?

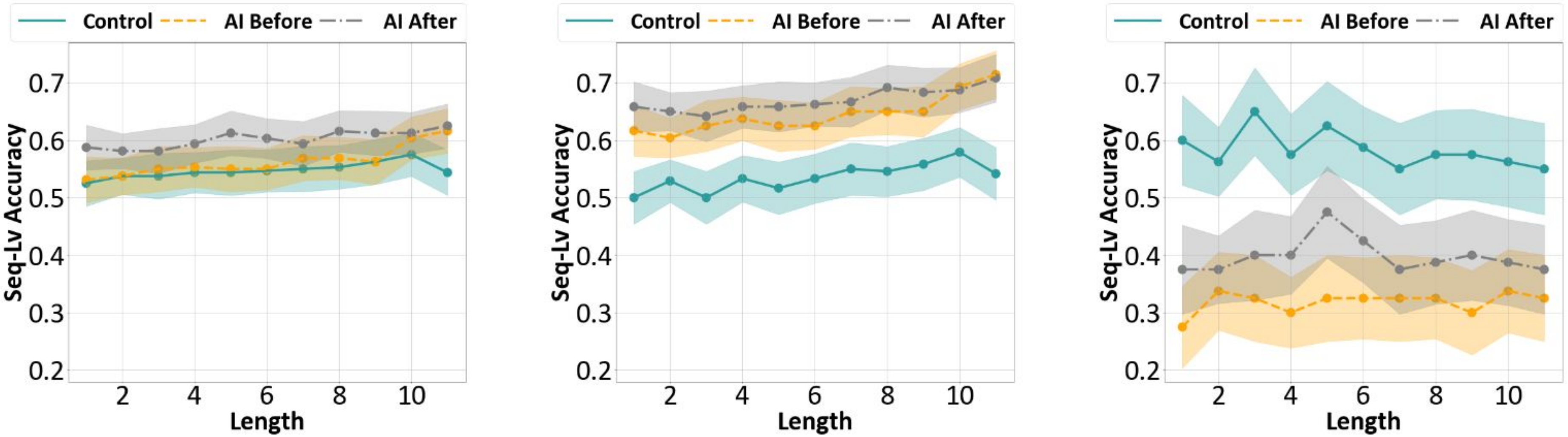

Control
AI-before
AI-after
Individual-level Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
All
AI Correct
AI Incorrect

Conclusions

- With social influence, predictions still have an effect - incorrect ones too (one-way ANOVA)
- Timing of showing them is not very relevant as far as final judgements are concerned (post-hoc Tukey, if I understand correctly)
- Even groups of people are bad at determining the correctness of AI indicators. This implies serious risks when models are wrong

Suspicious p-values

- Precise p-values rarely given, mostly just $p < 0.001$, $p > 0.05$ etc.
- Religious faith of the authors in p-value thresholds (although to be fair $p < 0.001$ is very strong evidence)
- Often $p = 0.001$ exactly, not sure if they mean $p < 0.001$ or p is often very close to 0.001 and these are approximations



(a) Seq-level accuracy (All)

(b) Seq-level accuracy (AI Correct)

(c) Seq-level accuracy (AI Incorrect)

Effect of sequence length

- They say accuracy is different between the AI-correct and incorrect case, but does not increase with sequence length
- The second one does look increasing if you ask me
- Could have investigated effects with longer length - just make up sequences of prior judgements

Impact of predictions on sharing

- AI predictions do not change people's intention to share either real news or fake news, even if we analyze the cases that the AI model makes correct and wrong predictions separately
- Minor differences but $p > 0.05$, so "not statistically significant"
- Same with spread depth

Spread depth

- Actually, for each piece of news, 4 sequences of judgements created and shown at random
- Spread lists, attaching nodes with probability as self-reported
- Defined depth as average length over 1000 spread lists
- Should be spread trees in my opinion - more realistic for Facebook friend shares etc.

Confidence

- Let x = a person's confidence in their own judgement. Then their confidence in AI's judgement is the person agrees with AI, and 8 - x otherwise.
- A person's confidence in the AI's prediction will statistically be higher if it is shown before the opinion of the crowd (ANOVA)

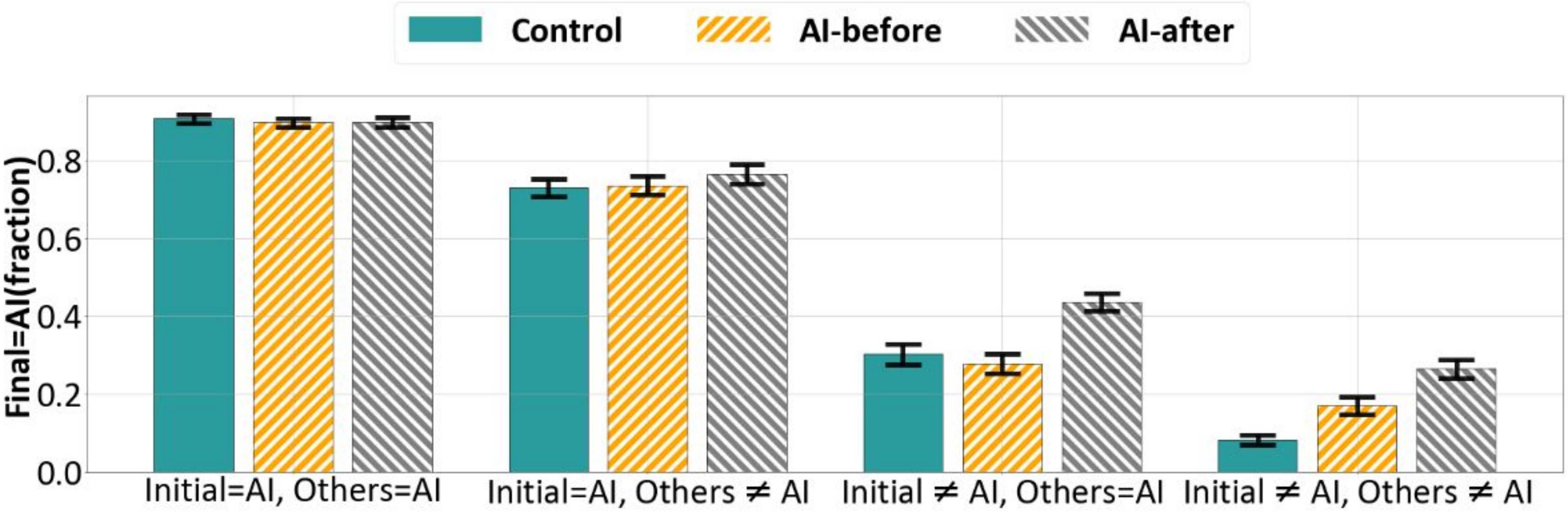


Fig. 5. The likelihood of a subject's final veracity judgment on a piece of news being the same as the AI model's prediction, when separating the data into four cases based on whether the subject's initial judgement was the same as the AI model, and whether the majority judgement of other people (i.e., subjects who reviewed this news before the current subject) was the same as the AI model. Error bars represent the standard errors of the mean.

Subjects changing their judgement

- If the subject's initial judgement agrees with AI, then whether they know it or not does not make them any more or less likely to change their decision, regardless of whether others agree
- Confidence in the AI prediction statistically the same across the three treatments, implying that seeing the model agree with a person does not change their confidence

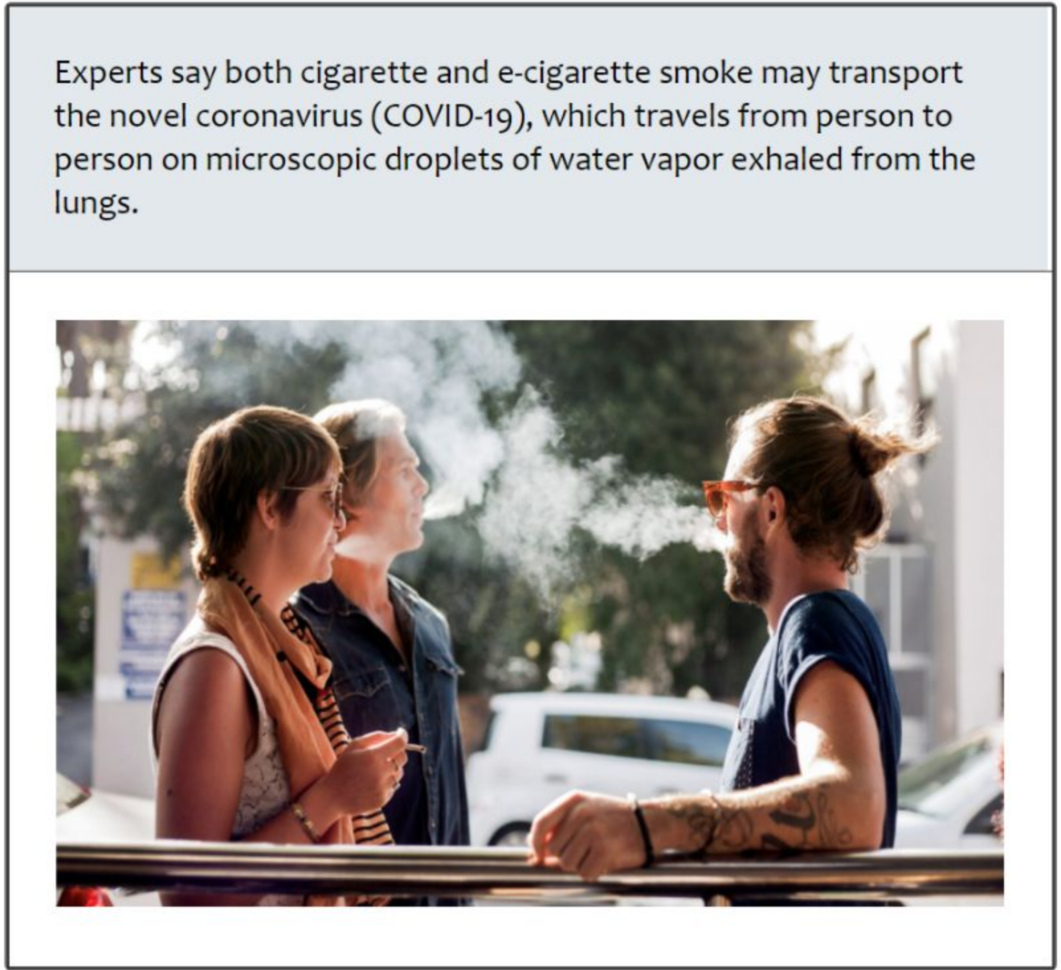
Subjects changing their judgement

- When we disagree with AI, we are most likely to change our mind for the AI-after case, and then confidence is higher too.
- This is regardless of whether others agree
- Maybe AI-before group is less likely to change to align with AI, because while formulating their initial opinions, they chose to disagree, so they have stronger opinions

Experiment 2

Differences compared to Experiment 1

- Included images
- Included no-social-influence setting
- Subjects not told about the 75% accuracy
- Used bootstrapping (1000 samples)
- New metrics



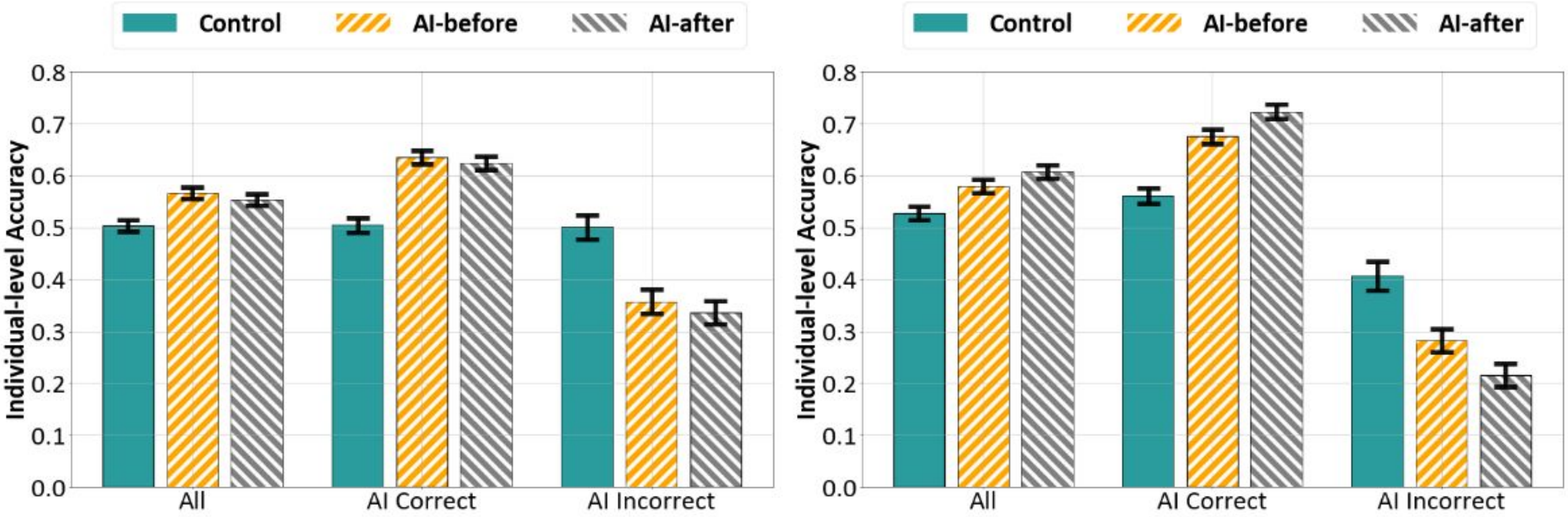
Experts say both cigarette and e-cigarette smoke may transport the novel coronavirus (COVID-19), which travels from person to person on microscopic droplets of water vapor exhaled from the lungs.

Truth and sharing discernment per subject

- If positive = true news, then truth discernment defined as (tp-fp)/(tp+fp+tn+fn)
- The same for tp=2, fn=98, fp=1, tn=99 as for tp=99, fn=1, fp=98, tn=2, so it looks flawed
- Sharing discernment - average level of willingness to share real news minus average level of willingness to share fake news

Setting with no social influence

- 3 steps: read, judge, estimate sharing probability
- The “AI-after” group will now see prediction after formulating their own judgement, and “AI-before” - along with the news



(a) With social influence

(b) No social influence

Fig. 7. The impacts of AI-based credibility indicators on subjects' individual-level accuracy in judging news veracity, for subjects who were subject to social influence (7a) and who were not subject to social influence (7b), respectively. Error bars represent the standard errors of the mean.

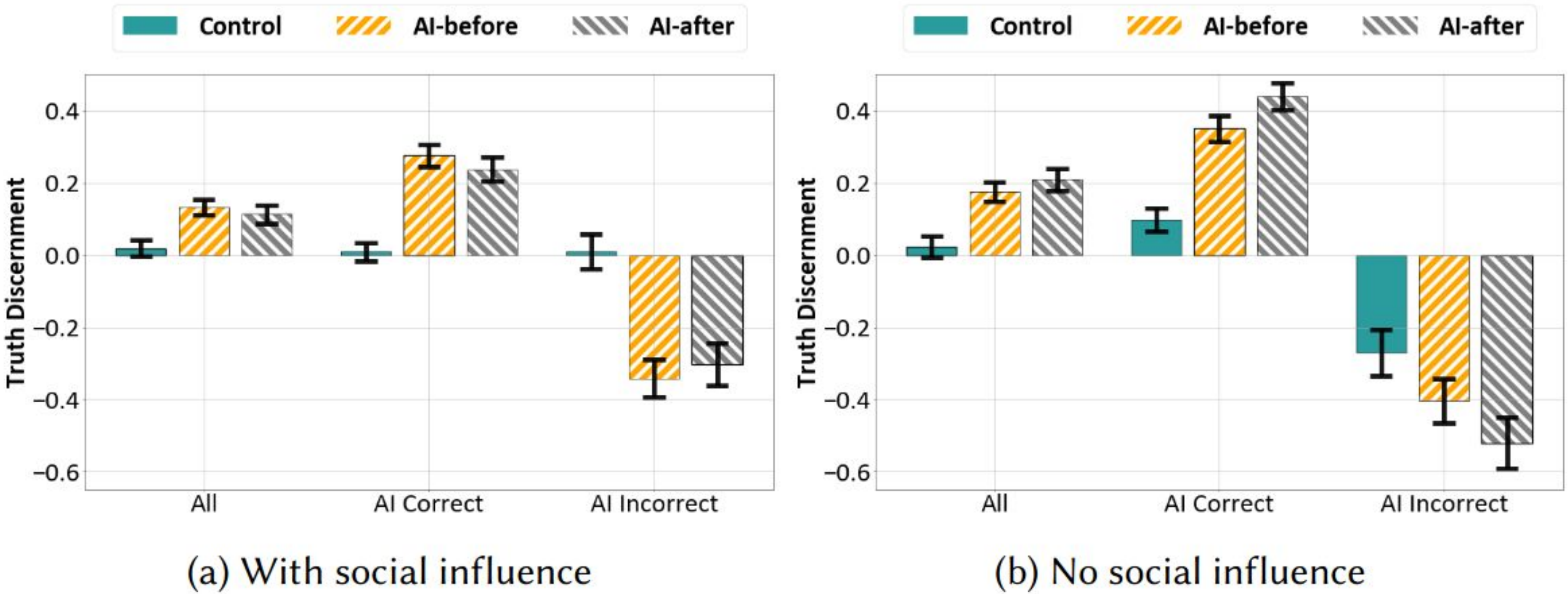


Fig. 8. The impacts of AI-based credibility indicators on subjects' truth discernment in judging news veracity, for subjects who were subject to social influence (8a) and who were not subject to social influence (8b), respectively. Error bars represent the standard errors of the mean.

Conclusions from Experiment 2

- Mostly just confirming earlier results
- Difference in sharing discernment only between control and AI-after. This implies that providing correct predictions only decreases sharing of fake news when they are shown after people have formed their own judgements

Conclusions from Experiment 2

- For AI-before, predictions' effect sizes are consistently larger (across treatments) when social influence is present. More precisely, if we estimate Cohen’s d based on a random sample from the influence-present group, and again based on another from the influence-absent group, the probability that the first one is larger is over 0.5 for all treatments.

Conclusions from Experiment 2

- This might be because providing correct AI predictions before people form their independent opinions cancels out the negative impacts of others' incorrect judgements.

Acknowledged shortcomings

- Previous research has shown that people consider repeatedly seen information to be more true. This was not tested here
- People have the tendency to befriend with like-minded individuals, which contributes to the creation of filter bubbles
- Study conducted with laypeople, and on only on COVID news. More research needed with recurring topics (e.g., climate change) to understand if the results can be generalized when people have strong prior beliefs
- The format of the news stories was fairly simple

謝謝大家~