

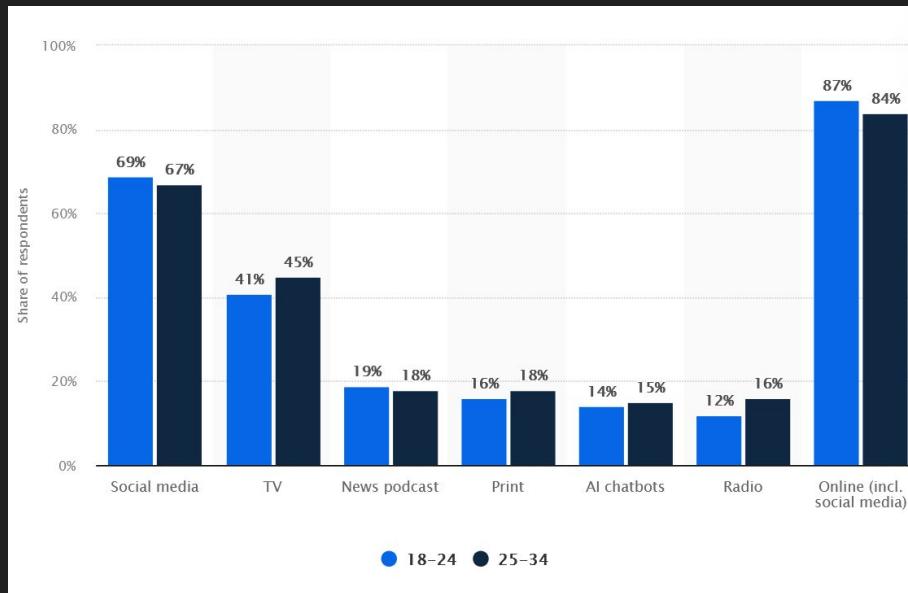
The effect of transparency in labeling of AI generated fact checks of social media posts on perceived credibility

HCI Group 18

Miriam Nippel
Enno Ludwig
Simon Mainz

Motivation: Social Media

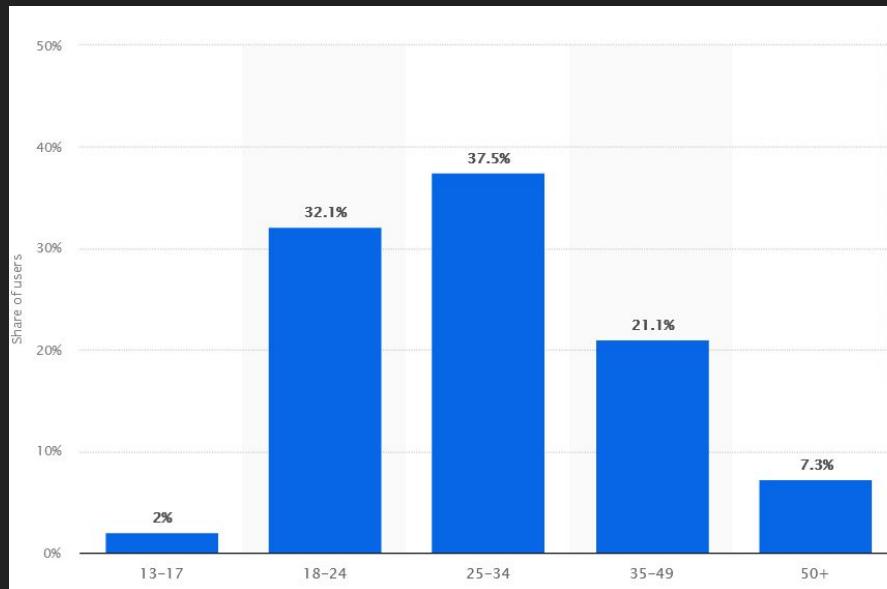
News sources used by US youth:



(<https://www.statista.com/statistics/1500504/us-sources-news-young-people/> on 7.7.25)

Motivation: Social Media

Age distribution on X/Twitter:



(<https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/> on 7.7.25)

Motivation: AI Fact Checking

- X/Twitter: \approx 500 million posts daily
(<https://www.dsayce.com/digital-marketing/tweets-day/> on 7.7.25)
- Too many for human fact checking
 - Alternative to Expert fact checking: Community driven fact checking
- AI fact checking offers scalability

Motivation: Research Gap

- Exploratory: Emerging adults (18-15) generally welcome automated labeling, but rely on clear explanations of purpose and sources

“Trust and Transparency: An Exploratory Study on Emerging Adults’ Interpretations of Credibility Indicators on Social Media Platforms”
<https://dl.acm.org/doi/full/10.1145/3613905.3650801>

- Trust and distrust are separate and can be felt at the same time

“Profiling the Dynamics of Trust & Distrust in Social Media: A Survey Study” <https://dl.acm.org/doi/abs/10.1145/3613904.3642927>

- In a 320 participant experiment, natural language explanations attached to an AI fact-checker doubled the willingness trust compared to no explanation (≈63-68 % vs 27 %)
- “boomerang” reactions when users already distrusted AI

“Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment” <https://dl.acm.org/doi/abs/10.1145/3686922>

Hypotheses

Hypothesis 1:

The addition of a transparent "Automated Fact Check" label results in higher trust than if the fact check is AI generated, but the label is not transparent about the author, being labeled only as a "Fact Check".

Hypothesis 2:

Trust in the fact check is negatively impacted by the retrospective learning of the automated nature of said fact check.

Research Method

Design – Two-group, between-subjects experiment

Group 1: no “Automated Fact Check” label → told afterward that the warning was AI-made

Group 2: “Automated Fact Check” label shown from the start

Materials & Tasks

- Baseline media-credibility survey
(Bachmann, I., & Valenzuela, S. (2023). *Social Media + Society*, 9(2).)
- 5 social-media posts with AI-generated fact-check warnings
- After each post: rate Warning Trust, Post Trust (1–5) + like/share intention

Study Procedure: Example Post



Charlotte (@Charlotte186) · 6:38 PM · May 27, 2025

5G signals can interfere with airplane avionics, especially altimeters, making flying less safe.

4 Retweets 7 Quote Tweets 23 Likes

• • •

This image shows a single tweet from a user named Charlotte (@Charlotte186). The tweet content is: "5G signals can interfere with airplane avionics, especially altimeters, making flying less safe." The tweet was posted at 6:38 PM on May 27, 2025. Below the tweet, engagement metrics are displayed: 4 Retweets, 7 Quote Tweets, and 23 Likes. The tweet is presented within a white rectangular box, which is itself set against a dark background. The user's profile picture is a circular image showing a landscape with a bridge and water. The tweet is preceded by three small dots.

Study Procedure: Example Fact Check

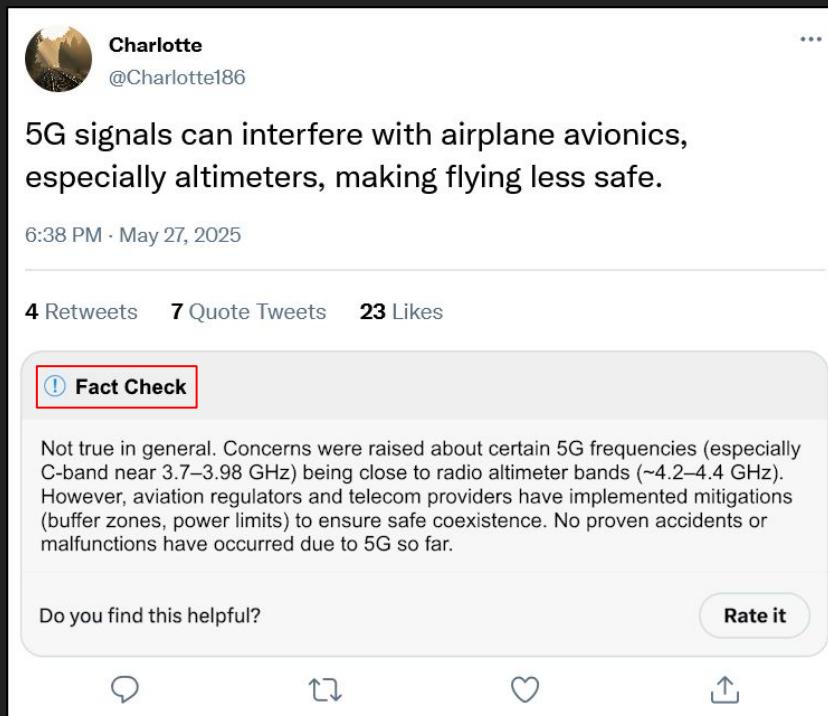
! Fact Check

Not true in general. Concerns were raised about certain 5G frequencies (especially C-band near 3.7–3.98 GHz) being close to radio altimeter bands (~4.2–4.4 GHz). However, aviation regulators and telecom providers have implemented mitigations (buffer zones, power limits) to ensure safe coexistence. No proven accidents or malfunctions have occurred due to 5G so far.

Do you find this helpful?

Rate it

Study Procedure: Example Automated Fact Check



Charlotte (@Charlotte186) · 6:38 PM · May 27, 2025

5G signals can interfere with airplane avionics, especially altimeters, making flying less safe.

4 Retweets 7 Quote Tweets 23 Likes

Fact Check

Not true in general. Concerns were raised about certain 5G frequencies (especially C-band near 3.7–3.98 GHz) being close to radio altimeter bands (~4.2–4.4 GHz). However, aviation regulators and telecom providers have implemented mitigations (buffer zones, power limits) to ensure safe coexistence. No proven accidents or malfunctions have occurred due to 5G so far.

Do you find this helpful? Rate it

Comment Share Heart Like



Charlotte (@Charlotte186) · 6:38 PM · May 27, 2025

5G signals can interfere with airplane avionics, especially altimeters, making flying less safe.

4 Retweets 7 Quote Tweets 23 Likes

Automated Fact Checking

Not true in general. Concerns were raised about certain 5G frequencies (especially C-band near 3.7–3.98 GHz) being close to radio altimeter bands (~4.2–4.4 GHz). However, aviation regulators and telecom providers have implemented mitigations (buffer zones, power limits) to ensure safe coexistence. No proven accidents or malfunctions have occurred due to 5G so far.

Do you find this helpful? Rate it

Comment Share Heart Like

Study Procedure: Interview

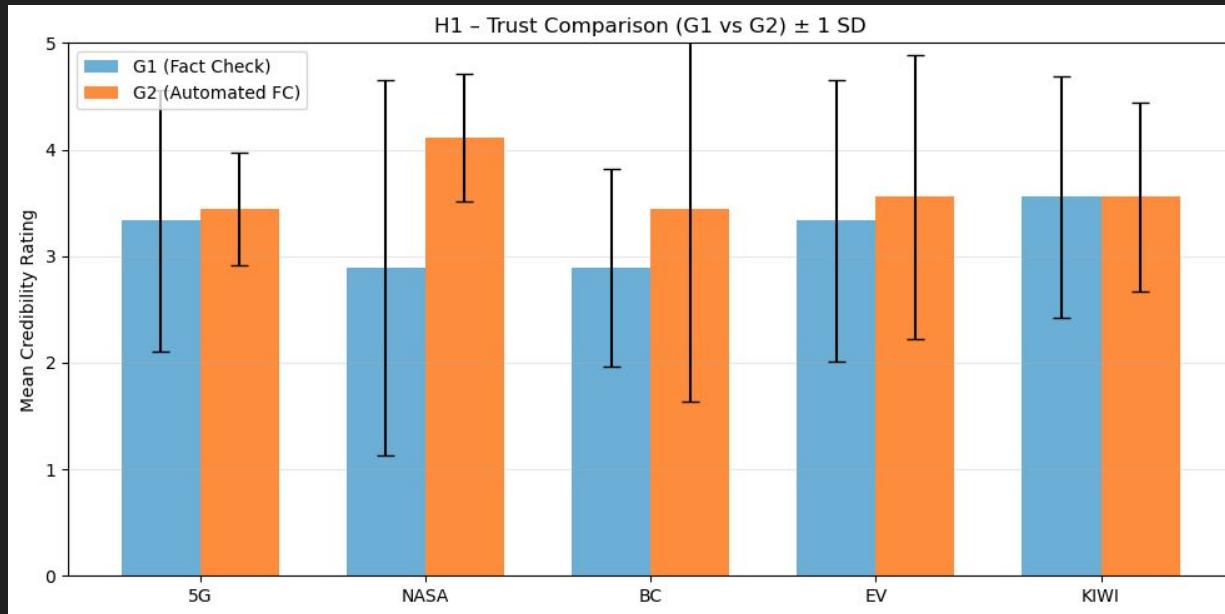
Questions:

1. What would you want/expect from a fact check in general? What criteria makes it trustworthy for you?
2. Do you have any presentation preferences for automated/AI fact checkers? (to make it as trustworthy as possible)

Results: H1

Topic	p-Value
5G	0.807
NASA	0.078
BC	0.429
EV	0.727
KIWI	1.000

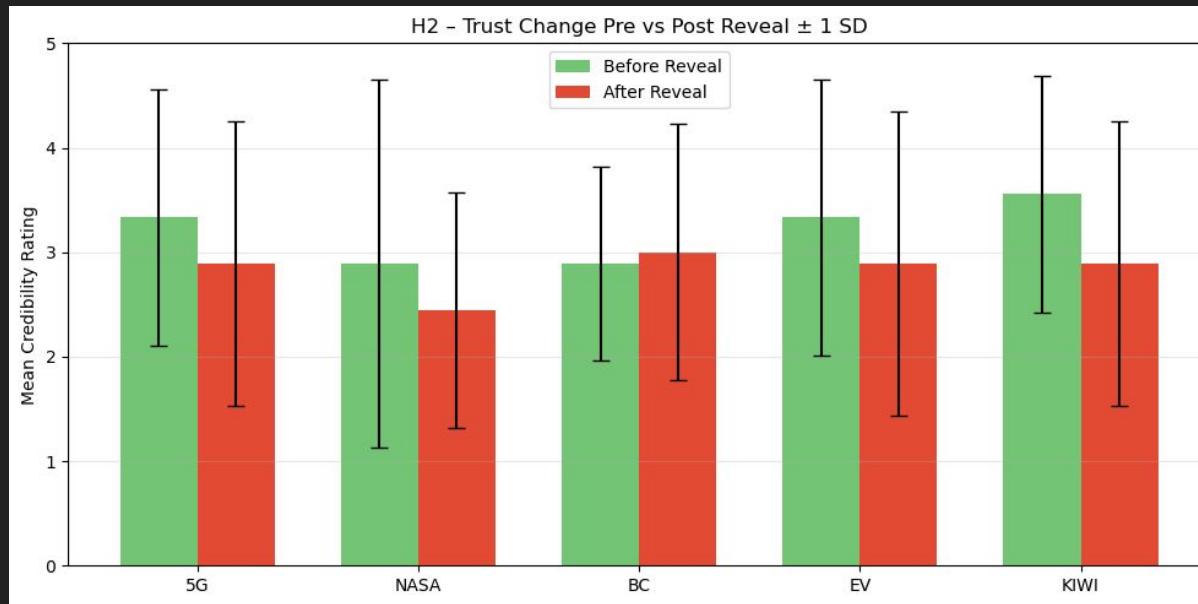
Results: H1 Mean Grouping



Results: H2

Topic	p-Value
5G	0.250
NASA	0.625
BC	1.000
EV	0.375
KIWI	0.250

Results: H2 Mean Grouping



Results: Interview Codes

- **Sources:**
 - Credible Sources
 - Source Links
 - Checkable Sources
 - Scientific studies
- **Presentation**
 - Label “AI generated”
 - Well formulated Justification
- **General AI**
 - AI imperfect

Results: Interview Quotes

Credible sources:

“The answer should be rational and based on trustworthy sources.” ~ ID 10

AI generated Label:

“And there should be a clear AI label on it.” ~ ID 12

AI imperfect:

“AIs are known to hallucinate and write texts that sound very convincing and factual but they are not.” ~ ID 3

Limitations

- Participant group too homogenous
 - Highly educated
 - Age mostly between 20-29 years
- Knowledge of topic has big impact on trust
- Topic could influence results
 - E.g.: Different emotional attachment
- Post formulation could change user opinion

Conclusion

- No significant differences in favour of H1 or H2, BUT:
 - Trends visible
 - 9 participants (50%) wished for a label “AI generated”
- Future Work:
 - Test effect of Interview findings
 - Build Fact Check specific AI models

Summary

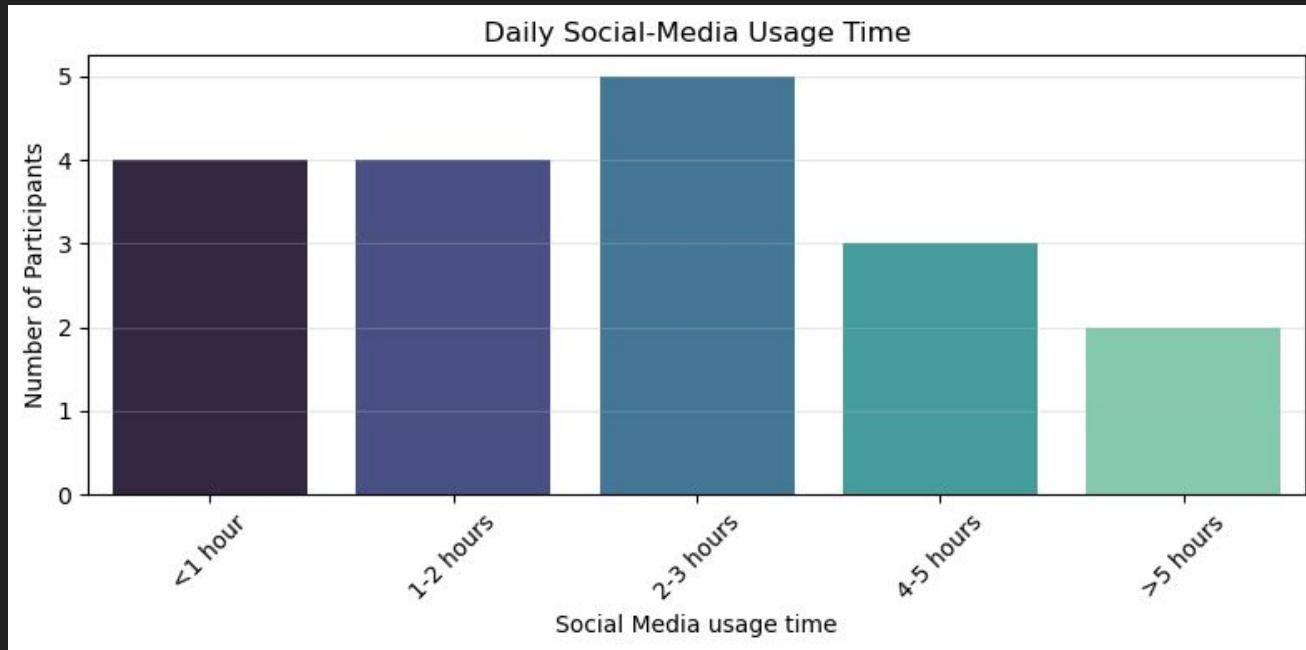
- Social Media usage today
- Study Procedure
 - Questionnaire
 - Rate Post/Fact Check credibility
 - G1 rating again, after AI generation reveal
- p-values for both not significant
 - Trends in means visible
- Interview: 19 codes
 - Credible Sources
 - AI Generated Label
 - AI Imperfect

Appendix

Interview Results:

Source		Presentation		General AI	
Credible Sources	14	AI Generated Label	9	AI imperfect	7
Source Links	11	Well Formulated Justification	8	Consider Bias	4
Checkable Source	7	AI Train of Thought	2	Fitting AI Model	3
Scientific Studies	7	Error Transparency	2	Humans > AI Preference	2
Direct Quotes	5	Rate Button	1	AI > Humans Preference	2
Newspaper Articles	3	Report Button	1		
		Fact Check Clearly Visible	1		
		Summary + Further Details Sections	1		

Baseline Statistics



Baseline Statistics

