The Promises and Pitfalls of Automatic Fact Checking

Isabelle Augenstein*

ACL Workshop on NLP for Positive Impact
Vienna, 31 June 2025

COPENLU

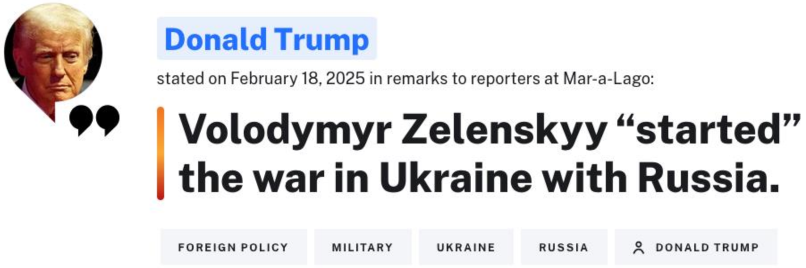
Funded by
the European Union

erc
European Research Council
Established by the European Commission

*partial slide credit: Greta Warren

UNIVERSITY OF
COPENHAGEN

Fact checking – what is it?

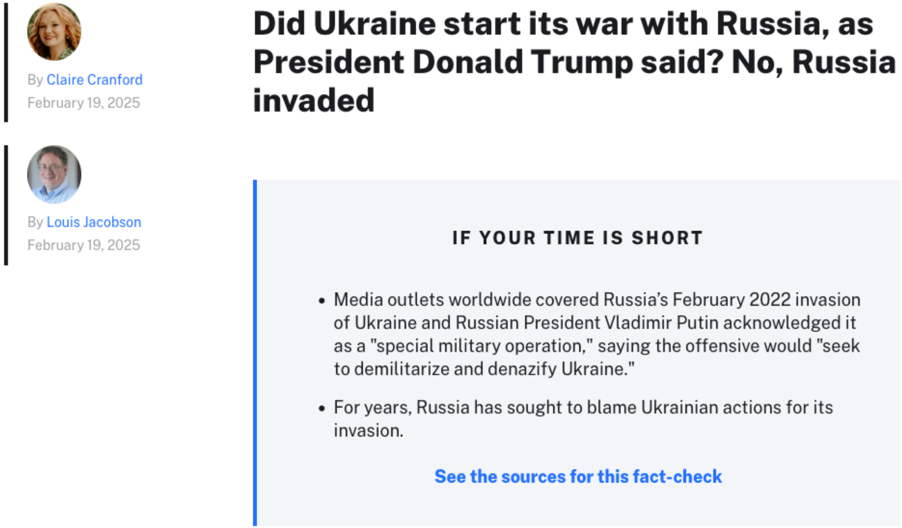


Donald Trump

stated on February 18, 2025 in remarks to reporters at Mar-a-Lago:

Volodymyr Zelenskyy "started" the war in Ukraine with Russia.

FOREIGN POLICY | MILITARY | UKRAINE | RUSSIA | DONALD TRUMP



By Claire Cranford
February 19, 2025

By Louis Jacobson
February 19, 2025

Did Ukraine start its war with Russia, as President Donald Trump said? No, Russia invaded

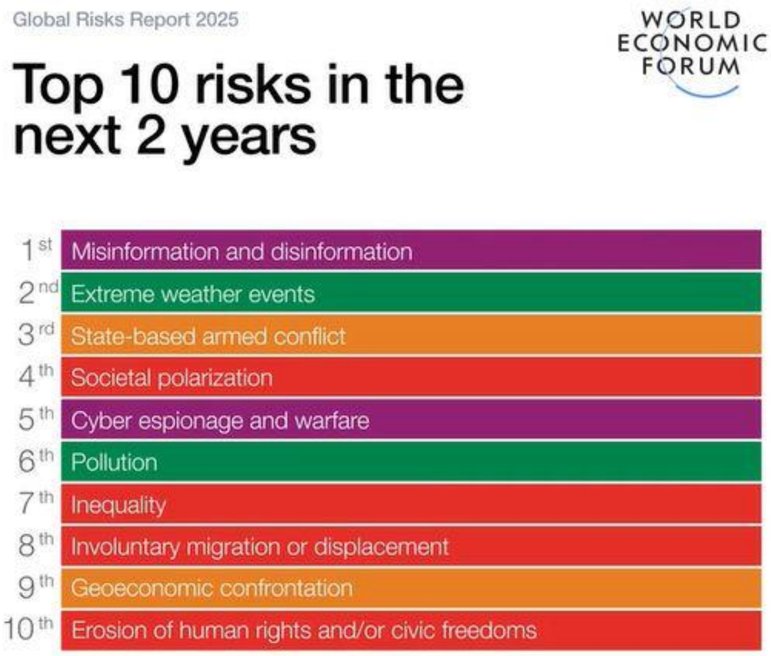
IF YOUR TIME IS SHORT

- Media outlets worldwide covered Russia's February 2022 invasion of Ukraine and Russian President Vladimir Putin acknowledged it as a "special military operation," saying the offensive would "seek to demilitarize and denazify Ukraine."
- For years, Russia has sought to blame Ukrainian actions for its invasion.

See the sources for this fact-check

https://www.politifact.com/factchecks/2025/feb/19/donald-trump/did-ukraine-start-its-war-with-russia-as-president/#sources

Fact checking – why is it important right now?

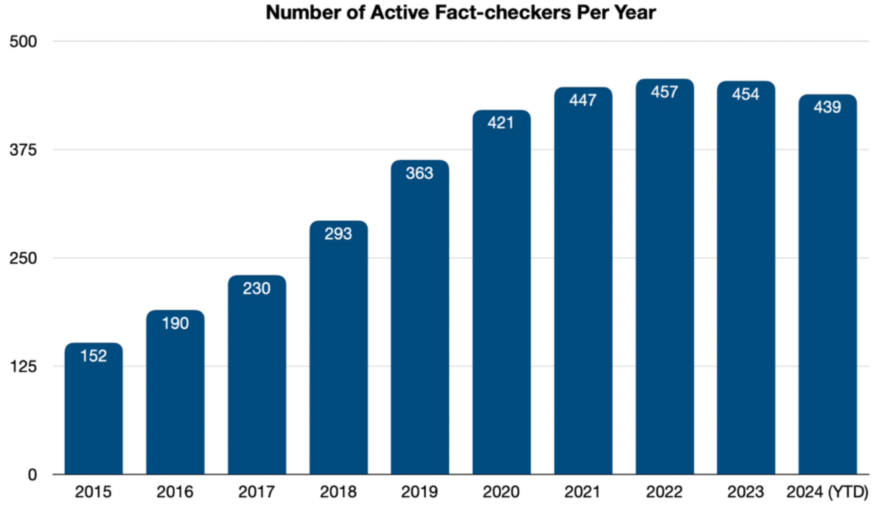


Global Risks Report 2025

WORLD ECONOMIC FORUM

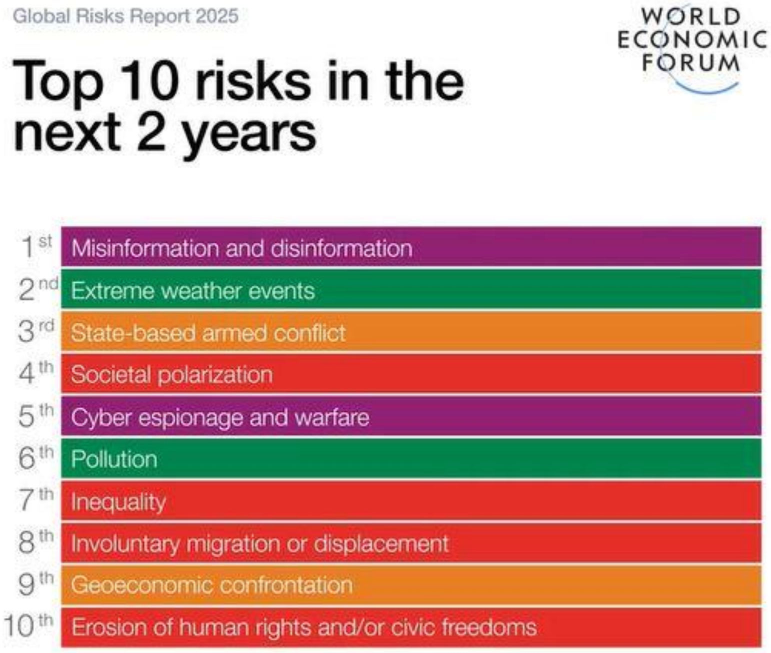
Top 10 risks in the next 2 years

1st Misinformation and disinformation
2nd Extreme weather events
3rd State-based armed conflict
4th Societal polarization
5th Cyber espionage and warfare
6th Pollution
7th Inequality
8th Involuntary migration or displacement
9th Geoeconomic confrontation
10th Erosion of human rights and/or civic freedoms



The number of active fact-checkers per year, 2015 to 2024 (year-to-date). The Reporters' Lab continuously updates its counts based on the start and stop dates of the fact-checkers. That means our numbers are revised year-to-year. (Courtesy)

Fact checking – why is it important right now?

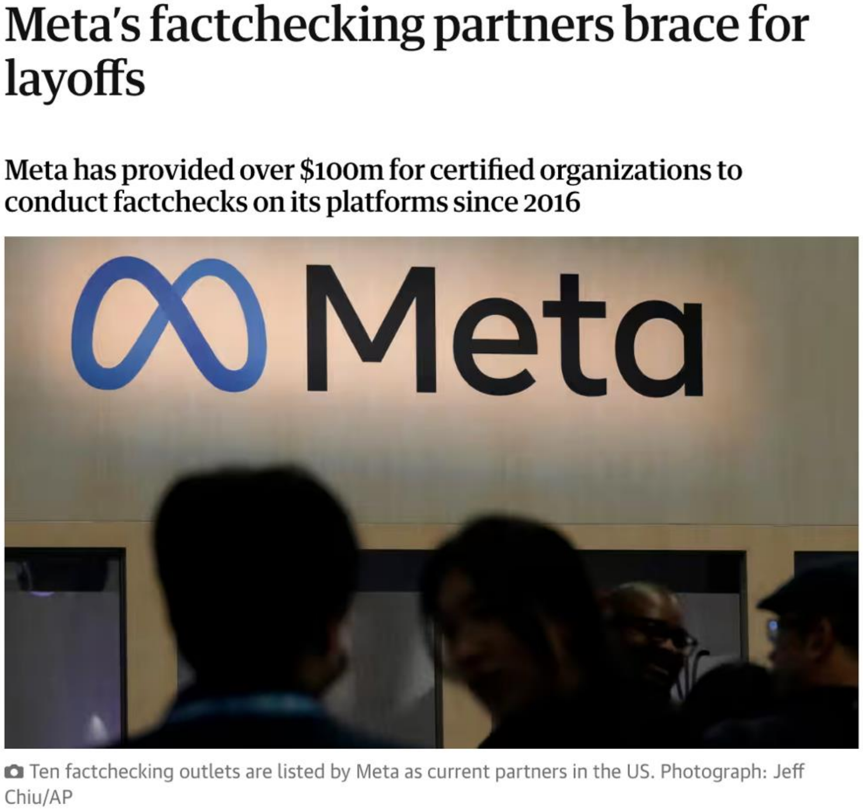


Global Risks Report 2025

WORLD ECONOMIC FORUM

Top 10 risks in the next 2 years

1st	Misinformation and disinformation
2nd	Extreme weather events
3rd	State-based armed conflict
4th	Societal polarization
5th	Cyber espionage and warfare
6th	Pollution
7th	Inequality
8th	Involuntary migration or displacement
9th	Geoeconomic confrontation
10th	Erosion of human rights and/or civic freedoms

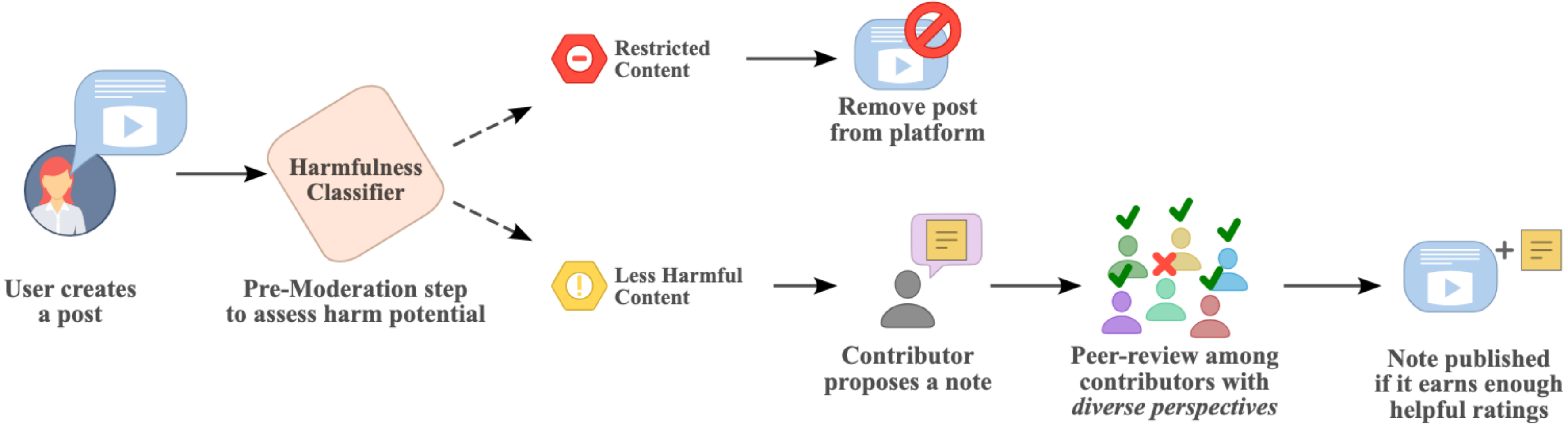


Meta's factchecking partners brace for layoffs

Meta has provided over $100m for certified organizations to conduct factchecks on its platforms since 2016

Ten factchecking outlets are listed by Meta as current partners in the US. Photograph: Jeff Chiu/AP

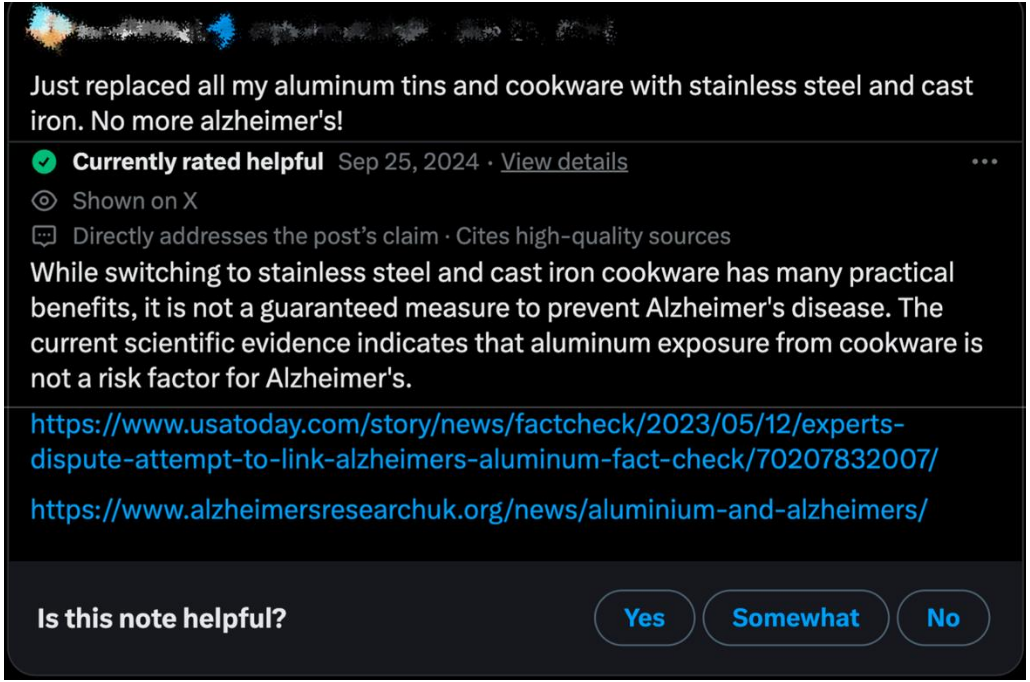
Community Notes (X / Twitter, Meta / Facebook, TikTok)


Restricted Content
Remove post from platform
Harmfulness Classifier
User creates a post
Pre-Moderation step to assess harm potential
Less Harmful Content
Contributor proposes a note
Peer-review among contributors with *diverse perspectives*
Note published if it earns enough helpful ratings

Moderation process:
(i) Pre-Moderation using AI classifiers: Restricted / blocked vs less harmful -> community moderation
(ii) Community Moderation: eligible volunteers propose additional context that undergoes peer review by other contributors with diverse perspectives before being published after a consensus is achieved

Augenstein et al. Community Moderation and the New Epistemology of Fact Checking on Social Media. CoRR, abs/2505.20067, May 2025.

Community Notes (X / Twitter, Meta / Facebook, TikTok)



[redacted]

Just replaced all my aluminum tins and cookware with stainless steel and cast iron. No more alzheimer's!

✓ **Currently rated helpful** Sep 25, 2024 · View details

Shown on X

Directly addresses the post's claim · Cites high-quality sources

While switching to stainless steel and cast iron cookware has many practical benefits, it is not a guaranteed measure to prevent Alzheimer's disease. The current scientific evidence indicates that aluminum exposure from cookware is not a risk factor for Alzheimer's.

https://www.usatoday.com/story/news/factcheck/2023/05/12/experts-dispute-attempt-to-link-alzheimers-aluminum-fact-check/70207832007/

https://www.alzheimersresearchuk.org/news/aluminium-and-alzheimers/

Is this note helpful?

Relation between fact checking and community notes

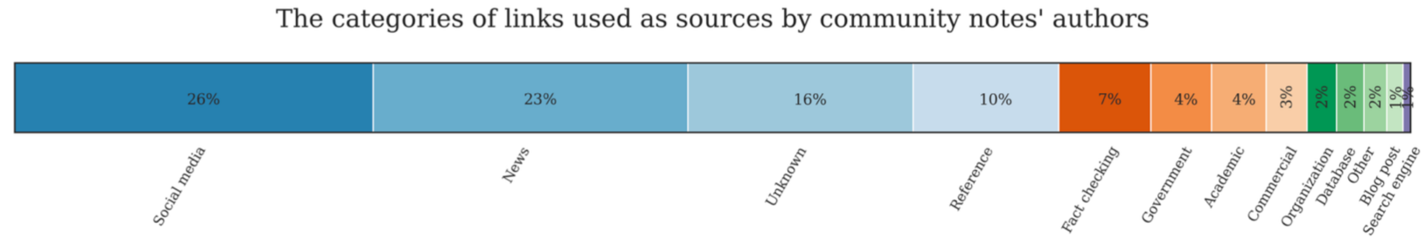


Figure 7: The categories of links used by Community notes' authors as a source, filtering for notes rated as "helpful".

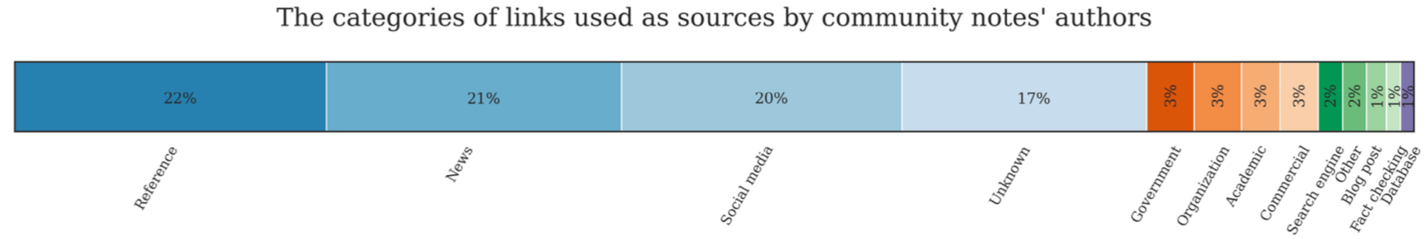
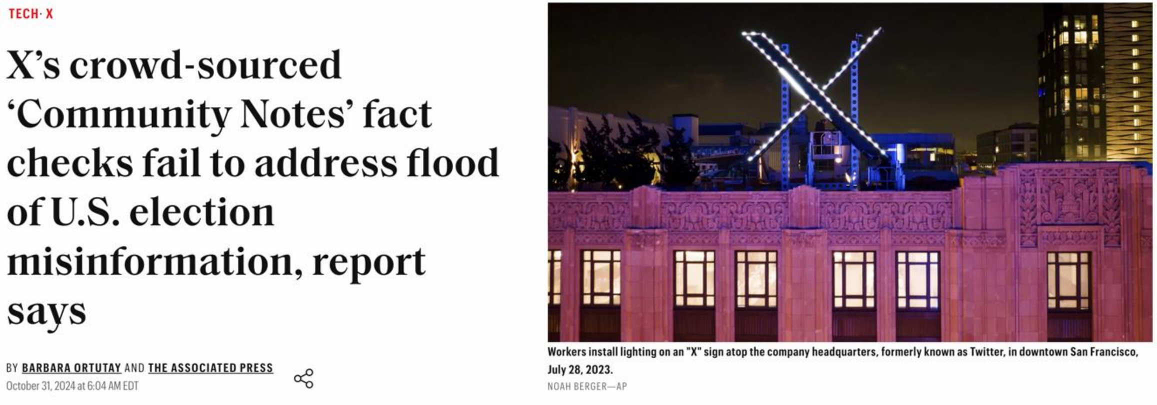


Figure 8: The categories of links used by Community notes' authors as a source, filtering for notes rated as "not helpful".

Nadav Borenstein*, Greta Warren*, Desmond Elliott, **Isabelle Augenstein**. Can Community Notes Replace Professional Fact-Checkers? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), July 2025.

Community Notes – does it work?



TECH · X

X's crowd-sourced 'Community Notes' fact checks fail to address flood of U.S. election misinformation, report says

BY BARBARA ORTUTAY AND THE ASSOCIATED PRESS
October 31, 2024 at 6:04 AM EDT

Workers install lighting on an "X" sign atop the company headquarters, formerly known as Twitter, in downtown San Francisco, July 28, 2023.
NOAH BERGER—AP

- *"Accurate notes correcting false and misleading claims about the U.S. elections were not displayed on **209 out of a sample of 283 posts deemed misleading** — or 74%"*
- *"Misleading posts that did not display Community Notes even when they were available included **false claims that the 2020 presidential election was stolen** and that **voting systems are unreliable**"*
- *"In the cases where Community Notes were displayed, the **original misleading posts received 13 times more views** than their accompanying notes"*

https://fortune.com/2024/10/31/x-community-notes-fact-checks-us-election-misinformation/

Community Notes – why does it not work?

- Only 11% of submitted notes reach ‘helpful’ status (i.e., shown to users) by achieving a cross-perspective
- Long time frame for notes to reach the algorithm’s required agreement level (15.5 hours on average)

➢ **False information has already spread**

- No expertise needed to become notes contributor
- Reliance on subjective helpfulness rather than objective facts
- Inadequate support and guardrails regarding explicit content

➢ **Key issues: speed, expertise, safety, adversarial attacks**

Nadav Borenstein*, Greta Warren*, Desmond Elliott, **Isabelle Augenstein**. Can Community Notes Replace Professional Fact-Checkers? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), July 2025.

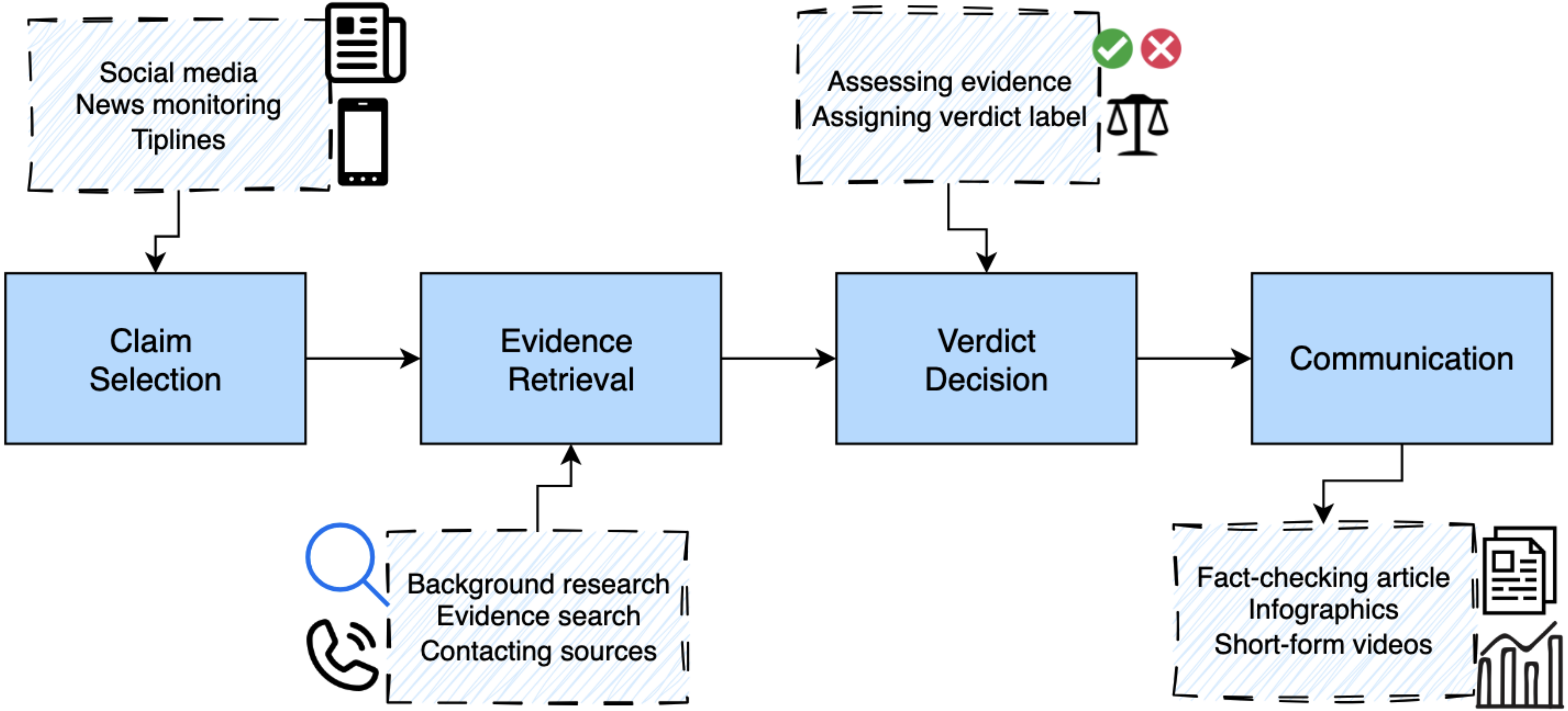
Community Notes: Recommendations

- Collaboration between **community and experts**
  - Workload distribution (repetitive claims vs high-risk claims)
  - Fact checkers as secondary reviewers of notes
  - Community flagging checkworthy claims

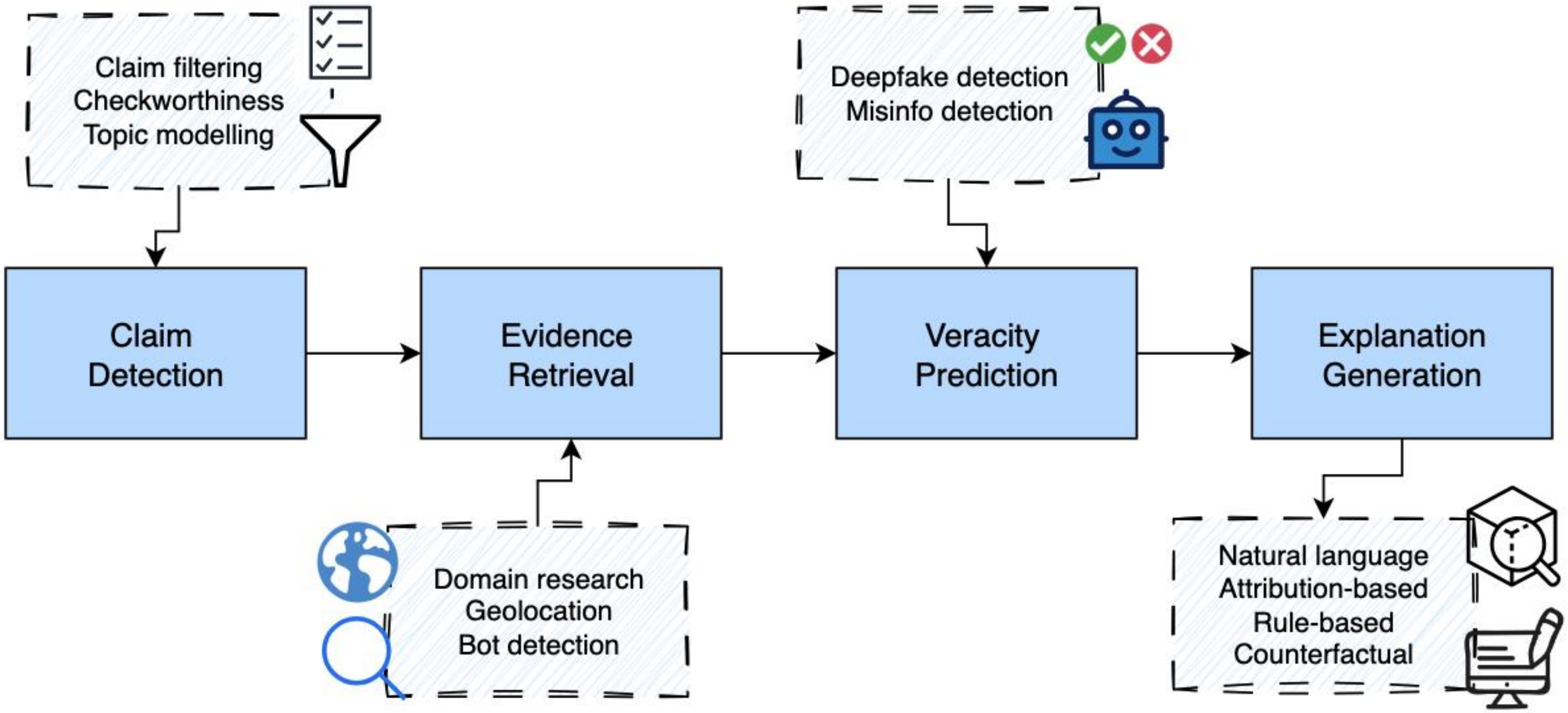
- Collaboration between **technology and the community**
  - Identify users likely to bring in diverse perspectives
  - Fusing community notes
  - Simulating crowd with AI agents (e.g. for sensitive content)
  - Handle previously checked notes with AI models

Augenstein et al. Community Moderation and the New Epistemology of Fact Checking on Social Media. CoRR, abs/2505.20067, May 2025.

Journalistic fact checking – how?



Explainable automatic fact checking – how?



Claim filtering
Checkworthiness
Topic modelling

Deepfake detection
Misinfo detection

Claim Detection → Evidence Retrieval → Veracity Prediction → Explanation Generation

Domain research
Geolocation
Bot detection

Natural language
Attribution-based
Rule-based
Counterfactual

Explainable automatic fact checking

Methods disconnected
from fact-checking practice
(Schlichtkrull et al., 2023)

Desiderata shaped by
AI developers & researchers
(Das et al., 2023)

Ineffective for fact-checkers
& misleading for laypeople
(Schmitt et al., 2024; Lim et al., 2024)

Research Questions

How do fact-checkers explain their decisions and processes?

Where are explanations of automated fact-checking systems needed?

How can explanations of automated fact-checking systems address fact-checkers' explanation needs?

Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. Conference on Human Factors in Computing Systems (CHI 2025), May 2025.

Method: Fact-checker interviews

10 interviews with fact-checkers in June & July 2024

5 women and 5 men from Europe, Africa, Asia, North America & South America

Pre-interview questionnaire

60-minute semi-structured interview

Bottom-up open coding → selective codes → Themes

Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. Conference on Human Factors in Computing Systems (CHI 2025), May 2025.

Design implications: Source quality

Primary sources are gold-standard

Account for biases & positionality of secondary sources

Evidence quality, relevance and reliability must be assessed and **explained alongside the verdict**

Design implications: Nuanced verdicts

Pervasive misinformation
often has a grain of truth

Detailed verdicts may be more
effective & less polarising

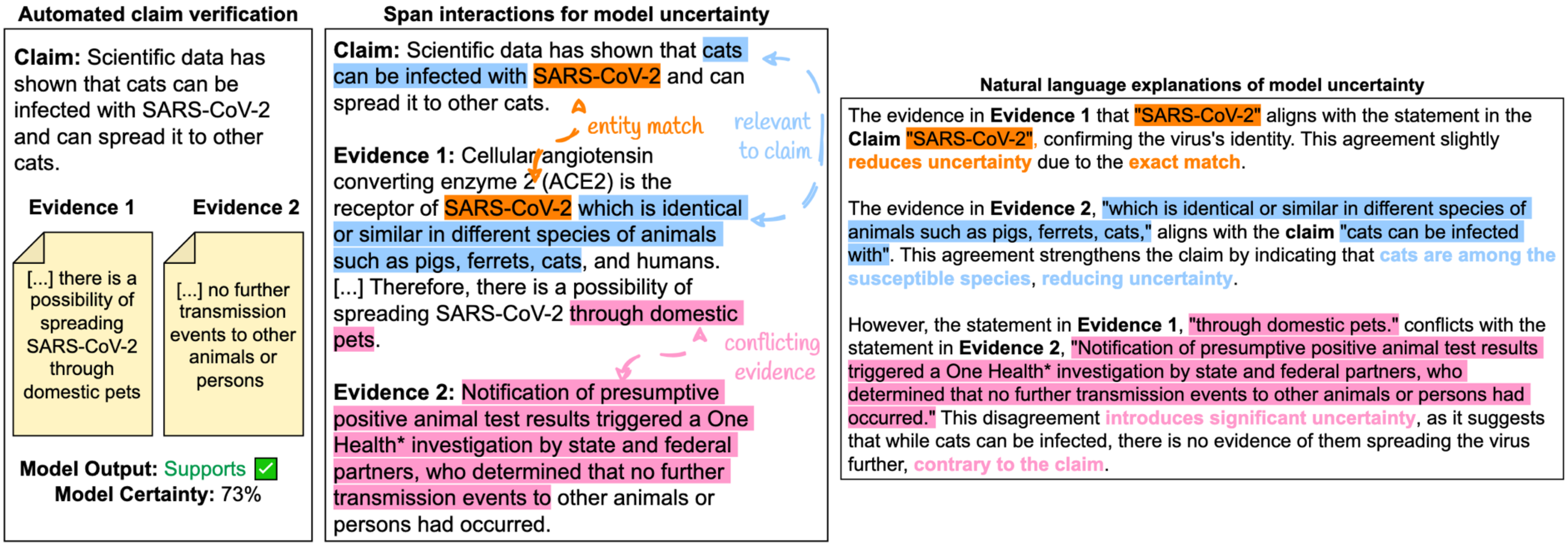
Explaining complex claims requires nuance **beyond binary true or false verdicts**

Design implications: Show the work

Explaining the **pathway to the verdict** is as important as the verdict itself

Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. Conference on Human Factors in Computing Systems (CHI 2025), May 2025.

Show the work -> explaining sources of uncertainty



Automated claim verification

Claim: Scientific data has shown that cats can be infected with SARS-CoV-2 and can spread it to other cats.

Evidence 1

[...] there is a possibility of spreading SARS-CoV-2 through domestic pets

Evidence 2

[...] no further transmission events to other animals or persons

Model Output: Supports ✅
Model Certainty: 73%

Span interactions for model uncertainty

Claim: Scientific data has shown that cats can be infected with SARS-CoV-2 and can spread it to other cats.

Evidence 1: Cellular angiotensin converting enzyme 2 (ACE2) is the receptor of SARS-CoV-2 which is identical or similar in different species of animals such as pigs, ferrets, cats, and humans. [...] Therefore, there is a possibility of spreading SARS-CoV-2 through domestic pets.

Evidence 2: Notification of presumptive positive animal test results triggered a One Health* investigation by state and federal partners, who determined that no further transmission events to other animals or persons had occurred.

Natural language explanations of model uncertainty

The evidence in **Evidence 1** that "SARS-CoV-2" aligns with the statement in the **Claim** "SARS-CoV-2", confirming the virus's identity. This agreement slightly **reduces uncertainty** due to the **exact match**.

The evidence in **Evidence 2**, "which is identical or similar in different species of animals such as pigs, ferrets, cats," aligns with the **claim** "cats can be infected with". This agreement strengthens the claim by indicating that **cats are among the susceptible species**, **reducing uncertainty**.

However, the statement in **Evidence 1**, "through domestic pets." conflicts with the statement in **Evidence 2**, "Notification of presumptive positive animal test results triggered a One Health* investigation by state and federal partners, who determined that no further transmission events to other animals or persons had occurred." This disagreement **introduces significant uncertainty**, as it suggests that while cats can be infected, there is no evidence of them spreading the virus further, **contrary to the claim**.

Jingyi Sun, Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Explaining Sources of Uncertainty in Automated Fact-Checking. CoRR, abs/2505.17855, May 2025.

Wrap-Up: Fact checkers needs vs. AI methods’ limitations

Verifiable explanations
Issues with faithfulness and stability of feature attributions

Explaining uncertainty
Numerical percentages disconnected from human notions of uncertainty

Replicable explanations
Requires end-to-end fact-checking systems & alignment with fact-checker processes

Way forward: human-centered explainable fact checking

Aligning AI methods with fact-checker reasoning processes

Providing human-centred, useful explanations tailored to context and expertise

HCI & AI research is needed to integrate automated fact-checking into fact-checkers processes & ensure fact-checkers remain central

References

Explainable fact checking:

- Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. Conference on Human Factors in Computing Systems (CHI 2025), May 2025.
- Jingyi Sun, Greta Warren, Irina Shklovski, **Isabelle Augenstein**. Explaining Sources of Uncertainty in Automated Fact-Checking. CoRR, abs/2505.17855, May 2025.
- Jingyi Sun, Pepa Atanasova, **Isabelle Augenstein**. A Unified Framework for Input Feature Attribution Analysis. In Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025), April 2025.
- Sagnik Ray Choudhury*, Pepa Atanasova*, **Isabelle Augenstein**. Explaining Interactions Between Text Spans. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), December 2023.
- Shuzhou Yuan, Jingyi Sun, Ran Zhang, Michael Färber, Steffen Eger, Pepa Atanasova, **Isabelle Augenstein**. Graph-Guided Textual Explanation Generation Framework. CoRR, abs/2412.12318, December 2024.

References

Factuality and context utilisation:

- **Isabelle Augenstein**, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, Giovanni Zagni. Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.
- Lovisa Hagström*, Youna Kim*, Haeun Yu, Sang-goo Lee, Richard Johansson, Hyunsoo Cho, **Isabelle Augenstein**. CUB: Benchmarking Context Utilisation Techniques for Language Models. CoRR, abs/2505.16518, May 2025.
- Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, **Isabelle Augenstein**. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), July 2025.
- Haeun Yu, Pepa Atanasova, **Isabelle Augenstein**. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), August 2024.

References

Community notes for fact checking:

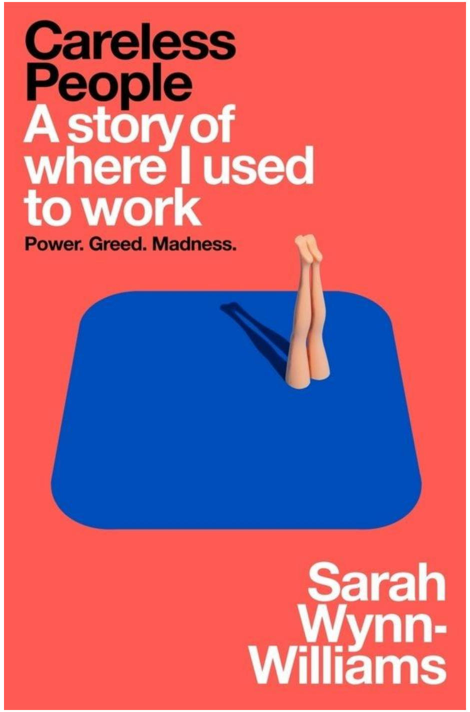
- **Isabelle Augenstein**, Michiel Bakker, Tanmoy Chakraborty, David Corney, Emilio Ferrara, Iryna Gurevych, Scott Hale, Eduard Hovy, Heng Ji, Irene Larraz, Filippo Menczer, Preslav Nakov, Paolo Papotti, Dhruv Sahnan, Greta Warren, Giovanni Zagni. Community Moderation and the New Epistemology of Fact Checking on Social Media. CoRR, abs/2505.20067, May 2025.
- Nadav Borenstein*, Greta Warren*, Desmond Elliott, **Isabelle Augenstein**. Can Community Notes Replace Professional Fact-Checkers? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), July 2025.

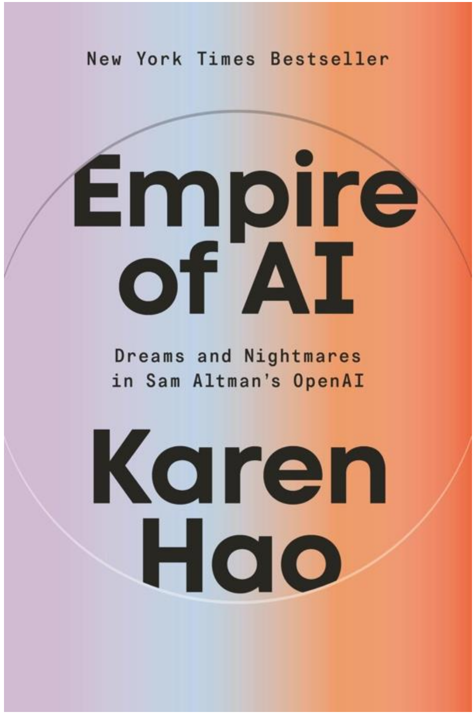
References

Other fact checking:

- Kevin Roitero, Dustin Wright, Michael Soprano, **Isabelle Augenstein**, Stefano Mizzaro. Collecting Cost-Effective, High-Quality Truthfulness Assessments with LLM Summarized Evidence. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025), July 2025.
- Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, **Isabelle Augenstein**, Iryna Gurevych, Preslav Nakov. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), November 2024.

Book recommendations




New York Times Bestseller
Empire of AI
Dreams and Nightmares in Sam Altman's OpenAI
Karen Hao

CopeNLU Lab

Isabelle Augenstein

Full Professor

Isabelle's main research interests are natural language understanding, explainability and learning with limited training data.

Pepa Atanasova

Assistant Professor

Pepa's research interests include the development, diagnostics, and application of explainability and interpretability techniques for NLP models.

Dustin Wright

Postdoc

Dustin is a DDSA postdoctoral fellow, working on scientific natural language understanding and faithful text generation.

Greta Warren

Postdoc

Greta's research interests include user-centred explainability, fact-checking, and human-AI interaction.

Yoonna Jang

Postdoc

Yoonna's research interests include language generation, factuality and interpretability.

Nadav Borenstein

PhD Student

Nadav's research interests include improving the trustworthiness and usefulness of deep models in the NLP domain.

Sarah Masud

Postdoc

Sarah broadly works in the area of computational social systems with a focus on news narrative and hate speech modelling. Her PhD at IIIT-Delhi was supported by fellowships from Google and PMRF.

Arnav Arora

PhD Student

Arnav's research interests include equitable ML, mitigating online harms, and the intersection of NLP and Computational Social Science.

Erik Arakelyan

PhD Student

Erik's main research interests are question answering and explainability.

Sara Vera Marjanovic

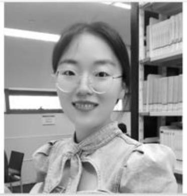
PhD Student

Sara's research interests include explainable IR and NLP models, identifying biases in large text datasets, as well as working with social media data. She is a member of the DIKU ML section and IR group and co-advised by Isabelle.

Haeun Yu

PhD Student

Haeun's main research interests include enhancing explainability in fact-checking and transparency of knowledge-enhanced LM.



Jingyi Sun

PhD Student

Jingyi Sun's research interests incluc explainability, fact-checking, and question answering.

Siddhesh Pawar

PhD Student

Siddhesh Pawar's research intere include multilingual models, fairn and accountability in NLP system

Amalie Brogaard Pauli

PhD Student

Amalie's research focuses on detecting persuasive and misleading text. She is a PhD student at Aarhus University and co-advised by Isabelle.

Sekh Mainul Islam

PhD Student

Sekh's research interests include explainability in fact checking and improving robustness and trustworthiness in NLP models.

Zain Muhammad Mujahid

PhD Student

Zain's main research interests include disinformation detection, fact-checking, and factual text generation.

Lucas Resck

PhD Student

Lucas is an ELLIS PhD student at the University of Cambridge, supervised by Anna Corhonen and co-supervised by Isabelle. His research interests include machine learning, NLP and explainability.

Ahmad Dawar Hakimi

PhD Student

Dawar is an ELLIS PhD student at LMU Munich, supervised by Hinrich Schütze and co-supervised by Isabelle. His research interests include mechanistic interpretability, summarisation and factuality of LLMs.

Na Min An

PhD Intern

Na Min An's research interests are explainability, multimodal systems, and human-centered AI.

Thanks for your attention!

Questions?