MBG Collaboration Topics

DAAD IFI Deutscher Akademischer Austauschdienst German Academic Exchange Service

MBG Media Bias Group

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

Generally…

- We are a global research network with six years of history and offer a collaborative, well-connected environment committed to ambitious research on media bias. More on us and our vision on: https://media-bias-research.org/.

- We offer research topics to work on together. These can be done remotely or in our Tokyo office. For that, we also offer a paid master's thesis or a paid research stay (for PhD students), both lasting 3 to 12 months. We also offer guest-stay opportunities for postdocs.

- Please check https://media-bias-research.org/career/ for more details.

- In the following, you will find a variety of research topics, but we are also open to other proposals. If a topic is interesting to you, just contact the supervisor mentioned on each topic slide, and we will assist you from there.

- Tokyo is a fascinating city and a unique research environment!

Generally…

Please note:

The following are **rough ideas**, which we will sketch out with you. When deciding to contact us, please consider which directions might interest you, based on a rough idea. Projects and their states change constantly, so these slides are intended only to bring together people with a basic shared interest. When you reach out to us, we will pick up from the latest state of research on each matter. We will engage with you to create a competitive research project.

In general, if you decide to apply, we will work on a project with you and your home university supervisor. We will help you, aiming to take the project to a top-tier level.

Talk to us!

Some general topic directions we deal with:

- NLP in bias detection
- Mechanistic interpretability
- Statistical analysis of bias
- Misinformation
- Narrative Detection
- Bias Indication
- Teaching bias awareness

Mechanistic Interpretability

Filip Kučera
f.kucera@media-bias-research.org

Mechanistic Interpretability for Vision-Language Models for Bias

Background

VLMs are gaining traction and most LLMs has gained at least some multimodal capacities. Yet, similarly to LLMs, they contain **biases** (e.g., gender or racial stereotypes), with the interaction between text & vision modalities adding another layer of complexity (e.g., is the output biased because of the text input or the visual input?). Mechinterp (MI) allows us to understand what's happening inside and where this behavior comes from.

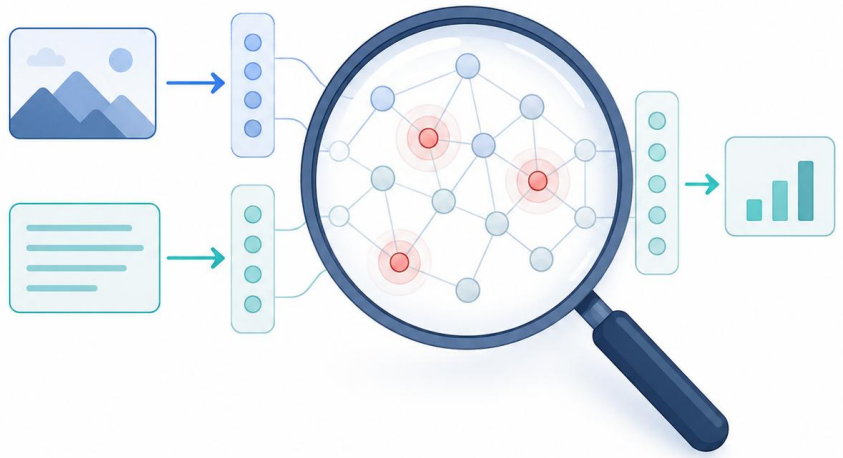
In this project, you will:

Goal

1. Learn about mechanistic interpretability methods in VLMs
2. Understand how models process information and how the outputs are formed from the inputs
3. Evaluate MI methods to **mitigate** the bias (e.g., by latent steering)

Tasks

1. Review metrics & benchmarks usable for measuring bias inside VLMs
2. Explore and evaluate MI methods usable for the bias detection
    - Correlational & Causality analysis (e.g., causal scrubbing)
    - Attention head(s) analysis
    - Linear probes (and/or transcoders/sparse autoencoders)
3. Based on the findings, explore and evaluate MI methods for mitigating this biased behavior *while preserving performance across unrelated tasks*



Filip Kučera
f.kucera@media-bias-research.org

Synthetic Media Bias Data Generation

Christoph Mandl
c.mandl@media-bias-research.org

Annotator Disagreement

Background

Media bias in news coverage is inherently subjective. Recent related work shows that disagreement between annotators is not just noise but tracks real differences in how groups of readers see the world. By collapsing those annotations into a single majority label, we lose exactly that signal and end up with models that are least reliable on the cases where readers genuinely disagree.

Goal

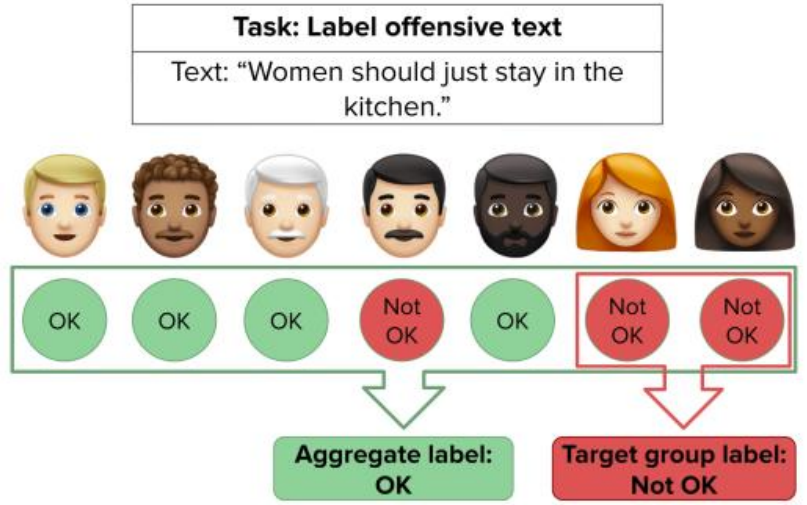
1. Understand how annotator disagreement is modeled in subjective NLP tasks
2. Develop a media bias model that leverages disagreement

Tasks

1. Survey existing strategies for modeling annotator disagreement
2. Evaluate these methods on media bias datasets (BABE, AnnoLexical)
3. Propose a new methodology tailored to media bias

References

- https://aclanthology.org/2023.emnlp-main.415/
- https://aclanthology.org/2025.emnlp-main.1800/
- https://aclanthology.org/2025.findings-emnlp.824/



Example from https://aclanthology.org/2023.emnlp-main.415/

Christoph Mandl
c.mandl@media-bias-research.org

Video Misinformation Detection

Florian Braun
f.braun@media-bias-research.org

How humans engage with model reasoning in misinformation detection

Background

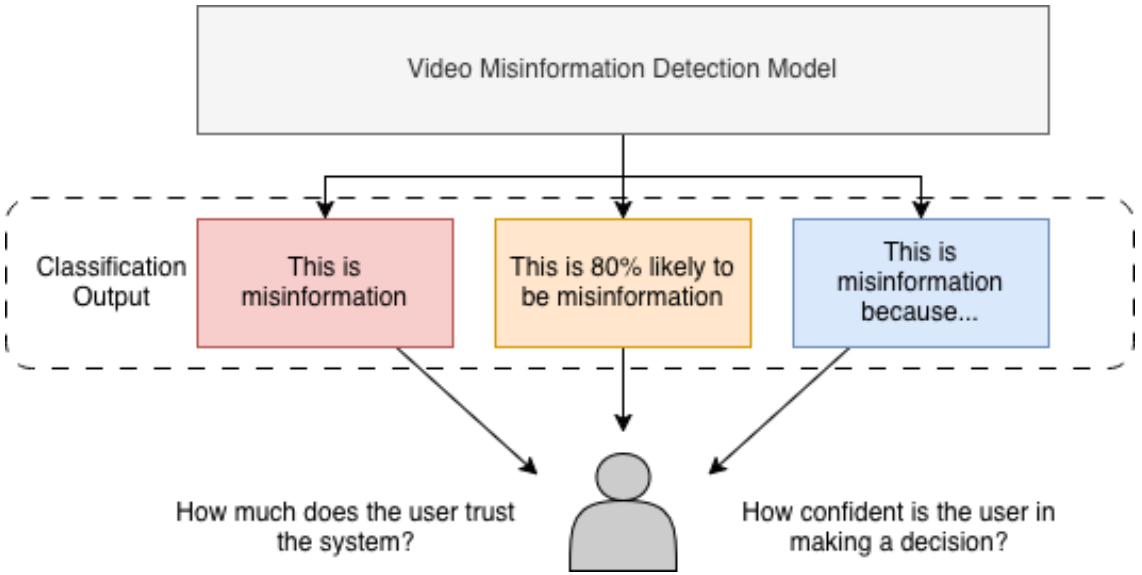
- Misinformation detection algorithms can output explanations, confidences and other scores to help with interpreting the results from the model
- How the different outputs are interpreted by users, and how they influence the confidence of the users is currently unexplored

Goal

1. Learn how humans are influenced by different model outputs (explanations, likelihood scores, reported confidence, true/false score)
2. Find out how well models report confidence compared to humans

Tasks

1. Explore confidence & explanation generation methods for video misinformation detection
2. Generate explanations for why a video is misinforming or not
3. Generate model confidence scores
4. Conduct a user study to compare how explanations, confidence scores or binary labels shape user performance & confidence



Florian Braun
f.braun@media-bias-research.org

Bridging Video Deepfake Detection Algorithms to Misinformation Detection

Background

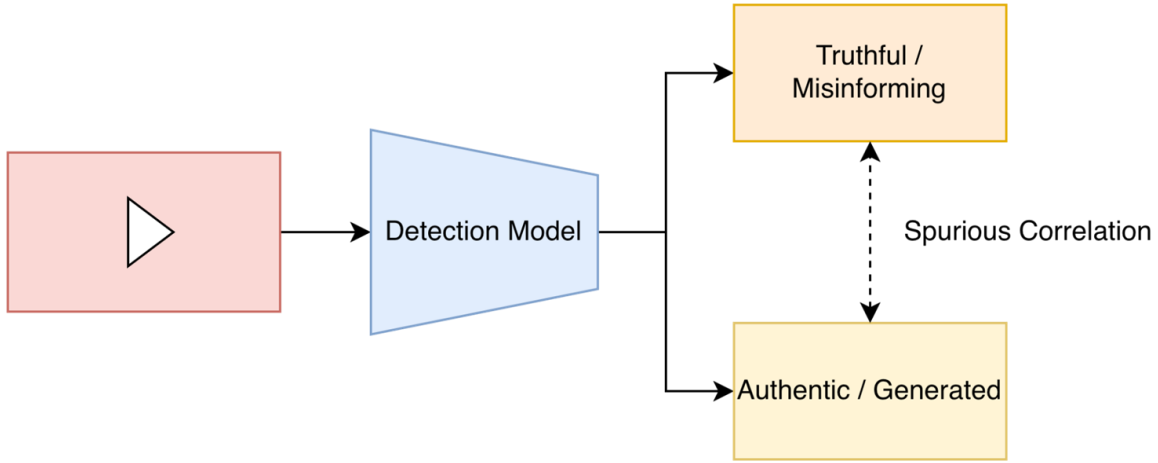
- Deepfake videos can be misleading, but also truthful
- Current deepfake detection methods only estimate whether content is generated, but not whether the content is misinforming or truthful
- Current misinformation detection methods do not focus on generated video content, but may learn to connect generated content to misleading content

Goal

1. Evaluate how well current deepfake detection methods work for misinformation detection
2. Evaluate whether misinformation detection methods learn spurious connections, linking generated content to misleading content
3. Propose a novel method to detect misleading generated video content

Tasks

1. Review literature on video deepfake detection, specifically for misinformation detection
2. Build a dataset from factually correct and incorrect generated videos
3. Benchmark existing video deepfake detection methods on the generated videos
4. Benchmark existing video misinformation detection methods on the generated videos
5. Develop a new baseline to detect whether generated videos are truthful or misinforming



Florian Braun
f.braun@media-bias-research.org

Statistics

Isabella Habereder
i.habereder@media-bias-research.org

Exploring media events as geopolitical indicators

Background

The media can influence public perception and decision-making. They can shape political views, social opinions, and voting behaviour [1]. Furthermore, media interventions (e.g., public statements, expression of intention for cooperation, …) are geopolitical indicators [2], but advanced analysis techniques are necessary to grasp the complexity of these media interventions.

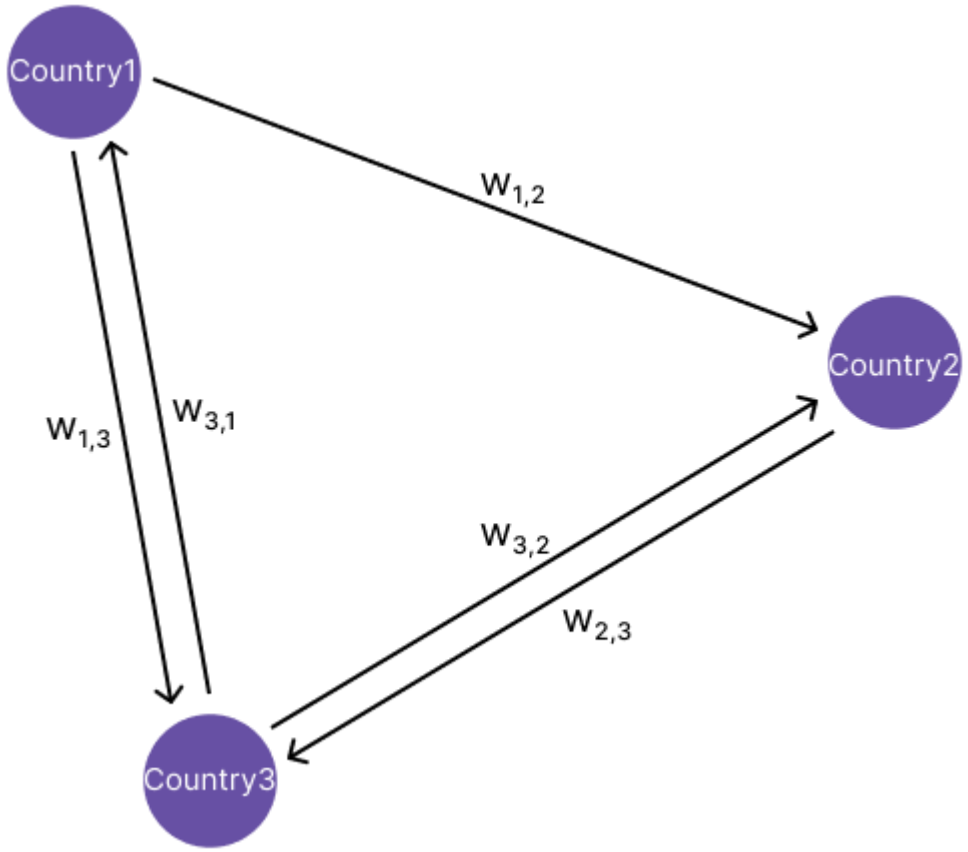
Goal

The goal is to answer the following research questions.

1. Which countries interact with each other through media coverage with respect to specific media events, and do these structures form distinct clusters in latent space?
2. How do the latent positions of countries in the media event network change during specific crises, and do they return to their initial positions?

Tasks

The GDELT (Global Data on Events, Location, and Tone) 1.0 Event Database can be used. Suitable strategies for processing large amounts of data are needed to represent the data as a network (see e.g., Figure). This network will then be analysed using a latent space model [3,4] to answer the research questions above. The results will be validated by external data.



[1] Alan S Gerber, Dean Karlan, and Daniel Bergan. “Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions”. *American Economic Journal: Applied Economics* 1, 2 (2009), 35–52.
[2] Altuntaş, Elgiz Yilmaz. "Perception and Framing Geopolitical Realities." *Media Representation and Public Perception of War* (2025): 283.
[3] Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. "Latent space approaches to social network analysis."*Journal of the american Statistical association* 97.460 (2002): 1090-1098.
[4] Casarin, Roberto, Antonio Peruzzi, and Mark FJ Steel. "Media bias and polarization through the lens of a Markov switching latent space network model." *The Annals of Applied Statistics* 19.4 (2025): 3416-3437.

Isabella Habereder
i.habereder@media-bias-research.org

Image Bias

Timo Spinde
T.Spinde@media-bias-research.org

Image Bias in News

Background

- The project aims to understand and evaluate image bias in diverse news articles sourced from various outlets.
- Our current collection methodology focuses on the visual representation of news, comparing the imagery used for similar topics across different publications.
- Preliminary observations suggest certain topics may be consistently represented with biased or skewed imagery.
- The challenge lies in not just collecting a diverse dataset, but also in accurately categorizing and analyzing the potential biases.

Goal

1. To expand the existing dataset with a diverse range of images from multiple news outlets and create a comprehensive understanding of image bias in news media.
2. Enhance the analytical depth by categorizing the types of biases and the frequency of their occurrence.
3. Understand the impact of image bias on public perception and the overall narrative of the news article.

Tasks

- Identify and collect diverse news articles and their associated images for a broad spectrum of topics.
- Use images without copyright or never publish the images, only use them for analysis (possible)
- Develop a robust framework that can categorize and detect potential image biases.
- Conduct detailed analysis on how different outlets represent similar news topics visually.
- Understand the implications of image biases and their potential to sway public opinion.
- Investigate if there's a correlation between textual and image biases in news articles.
- Potentially: Educate the public and journalists about the findings, emphasizing the importance of unbiased imagery in news representation.

Timo Spinde
T.Spinde@media-bias-research.org

HCI Research

- AI & Persuasion & Trust
- Media Bias

Smi Hinterreiter
s.hinterreiter@media-bias-research.org

1) AI (LLM) Persuasion: HCI Literature Review

Background

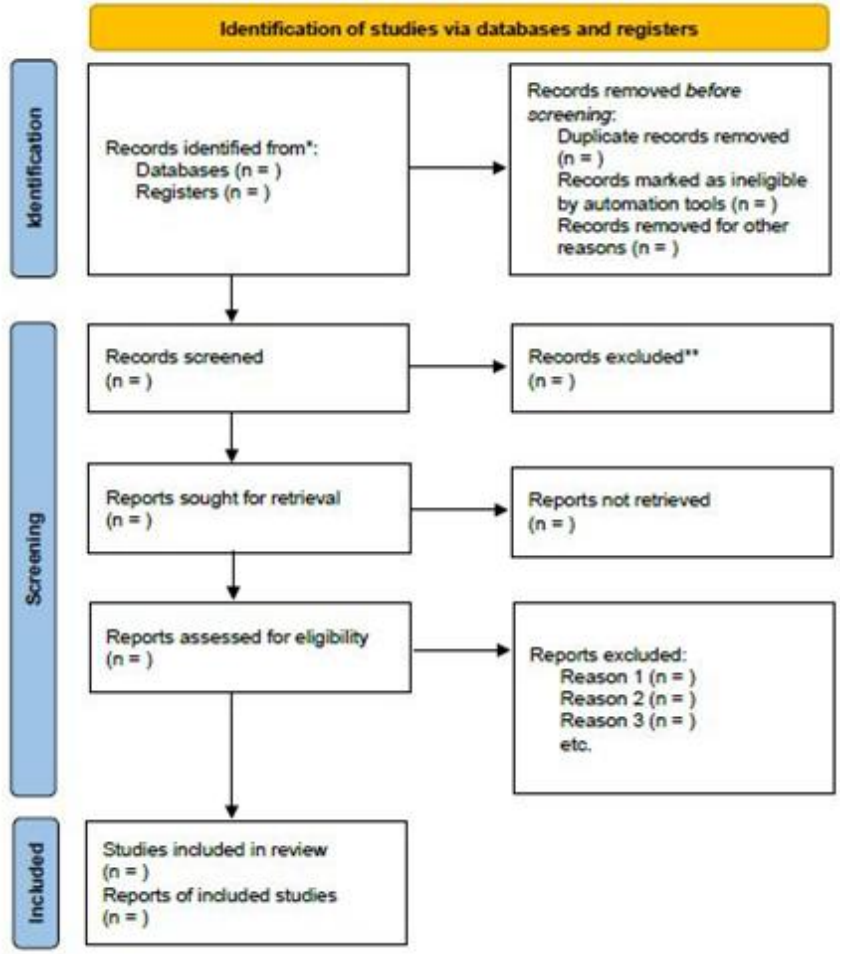
We already know that LLMs can be more persuasive than humans. This could be problematic in a range of ways. However, there's no overview yet on *how* they are persuasive, besides emerging technical and linguistic research. We want to look into the HCI and social aspects.

Goal

1. What interaction techniques showed persuasive effects?
2. What character aspects showed persuasive effects?
3. Create a taxonomy including character, input/output modality, representation of AI, and context of use.
4. Summarize what has been shown to be persuasive for the different categories.

Tasks

1. PRISMA review with Literature Crawl
2. Label and note effects
3. Create taxonomy



Smi Hinterreiter
s.hinterreiter@media-bias-research.org

2) AI (LLM) Persuasion User Studies: What's most persuasive?

Background

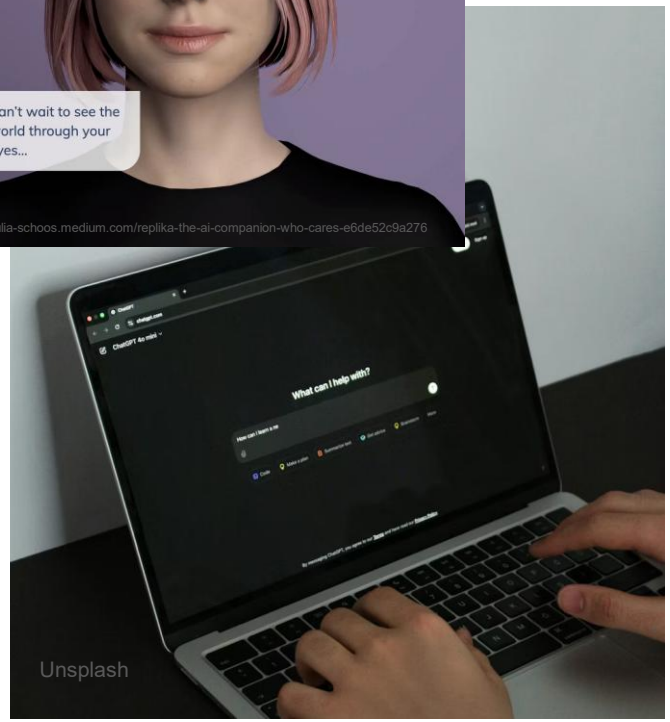
We already know that LLMs can be more persuasive than humans. We want to know how it can be persuasive, and through which mechanisms, e.g. through measurable variables like functional anthropomorphism or interactional anthropomorphism.

Goal

1. Find out which AI/LLM mechanisms and interactions are most persuasive through empirical studies (e.g. LLM character, modality voice/text/3D avatar, usage context, input modality).
2. Find out how to design embodied versions of LLM that are the most persuasive and/or inspire the most trust and/or make people think the most critical about given info (e.g. 3D avatar design).

Tasks

1. Design study and conditions, especially the variables we're interested in (e.g. trust, credibility, decision-making task, attraction, etc).
2. Run study under either controlled or real life conditions.
3. Statistical (and/or qualitative analysis) of study results.
4. Write paper presenting results.



Smi Hinterreiter
s.hinterreiter@media-bias-research.org

3) AI (LLM) Persuasion: How to protect against persuasion and calibrate trust?

Background

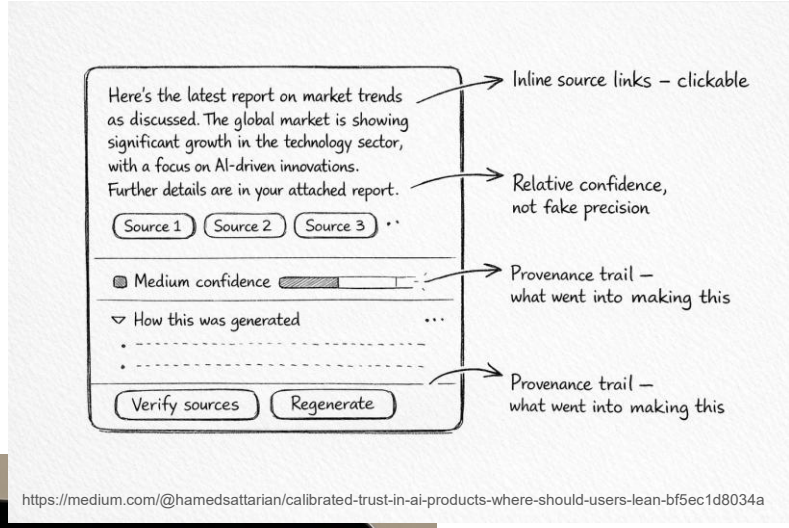
We already know that LLMs can be more persuasive than humans. We want to know how we can mitigate persuasive power and make people think critically OR detect persuasion tries. Further, we are interested in if an LLM can persuade people into caring more about topics such as climate change, gender equality, systemic racism, etc. (Alternatively, make them care more about people with opposing views.)

Goal

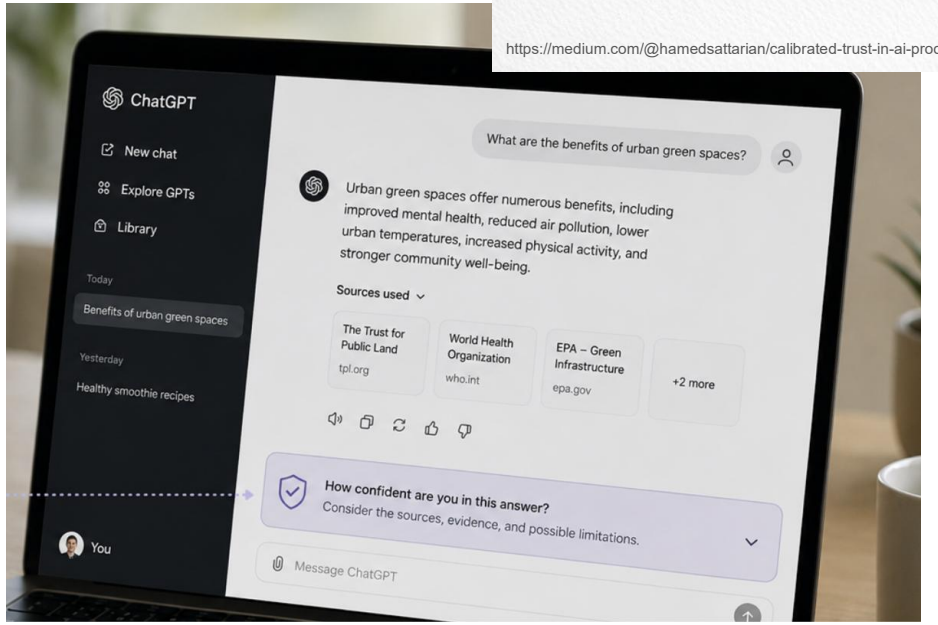
1. Find out which interventions protect against (a) AI persuasion or (b) AI hallucinated misinformation through empirical user studies.
2. Find out if/which persuasion strategy leads to people caring more about societal problems

Tasks

1. Review and design interventions, based on literature review.
2. Design study and conditions.
3. Run study under either controlled or real life conditions.
4. Statistical (and/or qualitative analysis) of study results.
5. Write paper presenting results.



https://medium.com/@hamedsattarian/calibrated-trust-in-ai-products-where-should-users-lean-bf5ec1d8034a



Smi Hinterreiter
s.hinterreiter@media-bias-research.org

4) AI & Calibrated Trust: AI Literacy

Background

AI often generates misinformation through hallucinations or can be persuasive without users noticing it. We want to find out what role AI literacy and potentially AI literacy interventions play in limiting belief in AI misinformation and AI discernment skills.

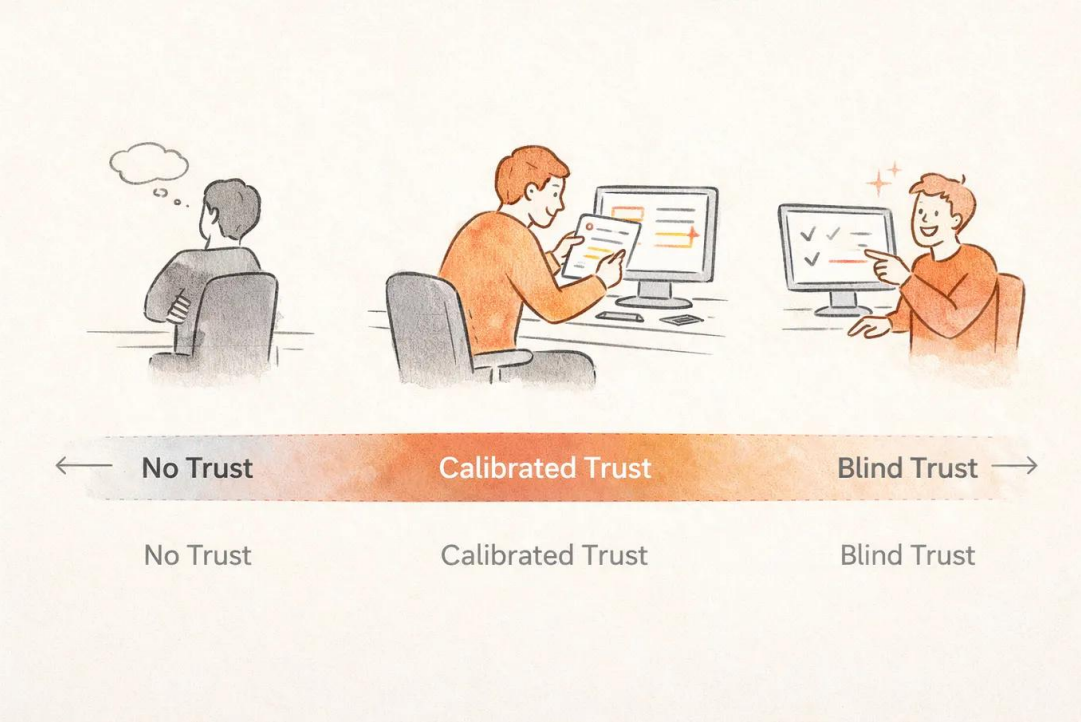
Goal

1. Online or test study (with brader population sample) to test influence of AI Literacy on belief in AI misinformation/ veracity discernement.

Tasks

1. Create study design: select questionnaires, tasks and study material.
2. Evaluate study and write paper.

We're also open to related topics and studies you're working on or are highly interested in!



https://medium.com/@hamedsattarian/calibrated-trust-in-ai-products-where-should-users-lean-bf5ec1d8034a

Smi Hinterreiter
s.hinterreiter@media-bias-research.org

5) Bias & BARI: UI/UX Design

Background

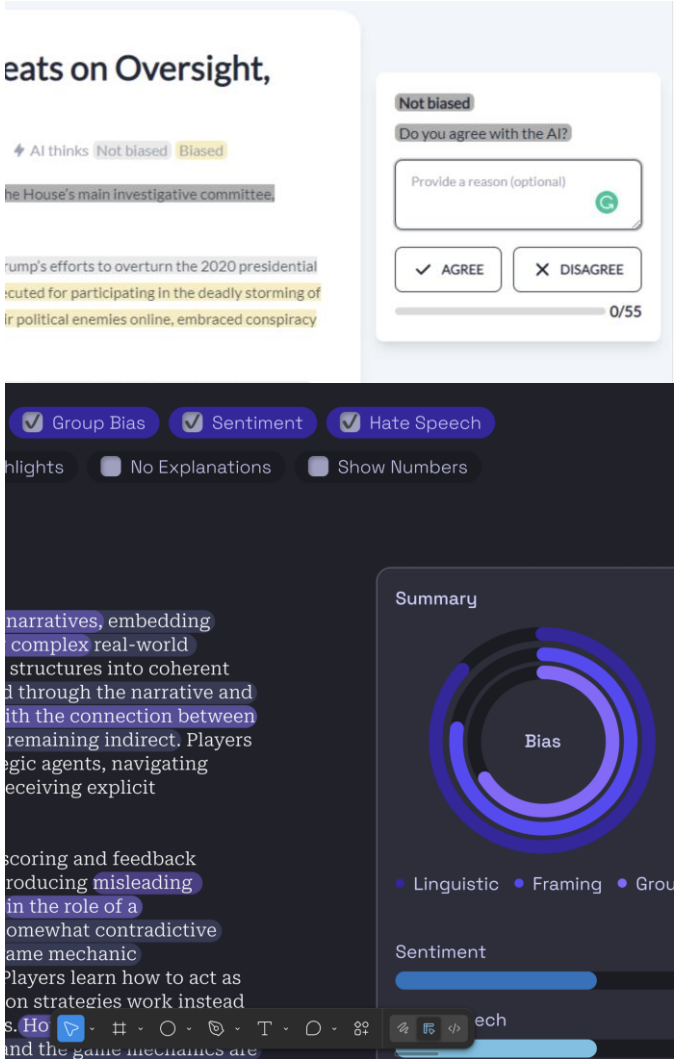
THis project is about developing BARI (Bias Analysis Research Infrastructure), a platform/website for researchers to:
1) Do small- and large-scale bias text analysis (with option to feedback the results → active learning) and
2) Discuss the bias taxonomy and classification of bias (discussion board/threaded discussions)
Target Group: Journalists, Social Science / Communication Science Researchers, Computer Scientists
BARI is a "Research Infrastructure" aka Researcher are the central target group)
→ needs to be very easy to use for persons who have little AI + Digital Literacy

Goals (1-2 per project)

1. Create the website/application design with a focus on UX and usability.
2. Find out what expert want in need in such a tool. (pre-design interviews)
3. Verify BARI through qualitative expert interviews. (after first design is roughly done)
4. Verify BARI through Qual. User Tests/Interviews
5. Test the functional version of BARI in Online User Study
6. Design new feedback mechanisms for active learning to improve BARI/classifier; test different ones and evaluate which one works best in terms of UX, amount of feedback, and quality of feedback.

Tasks

1. Review literature on similar tools and derive design guidelines.
2. Create design in Figma iteratively in communication with stakeholders; Figma for Dev
3. Find relevant experts + reach out
4. Develop questionnaire or interview structure
5. Do expert interviews
6. UX, thematic analysis or general analysis from the tests/interviews
7. For online user studies: Create study design, metrics, questionnaires, task, materials etc.mCreate a website on which you can run the study; Run & evaluate study
8. Summarize findings in paper



Smi Hinterreiter
s.hinterreiter@media-bias-research.org

6) Bias & BARI: How to explain Bias effectively

Background

We originally conducted a study to determine whether we can use LLM bias explanations in our systems to inform and educate folks. This worked for medium- and high-bias explanations with ChatGPT 4. However, we observed that our prompting can lead to unwanted persuasive effects. LMMs over-explained bias in the low-bias condition and increased participants' bias perception, leading them to deviate from the ground truth more than before the explanation. So part of what interests us is how to create effective LLM bias explanations. This can be a more technical topic with finetuning with Christoph and me, or something more about figuring out the best prompting and format.

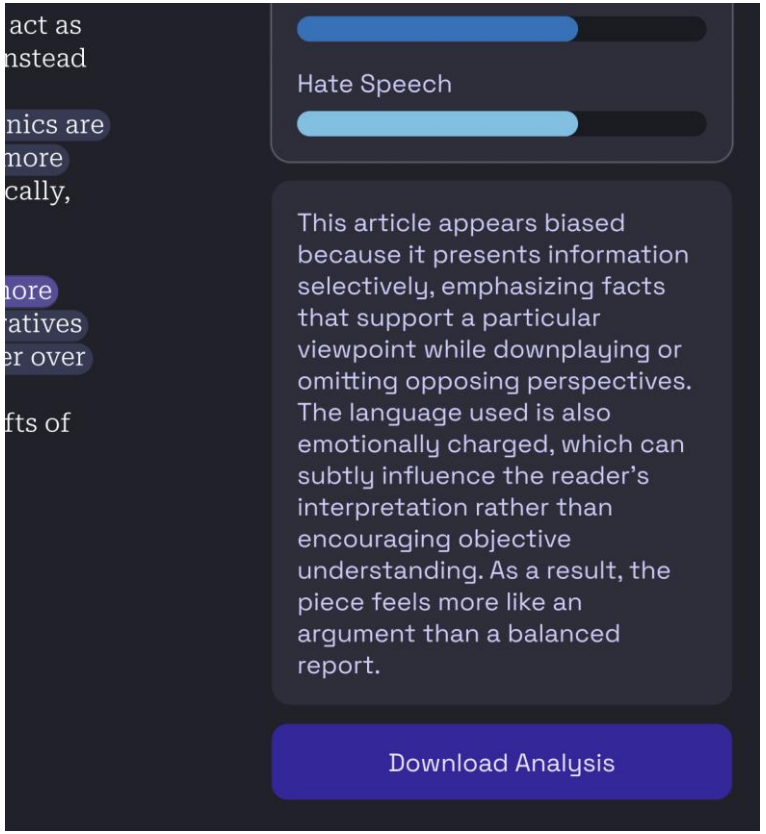
Further, there's other research on bias in grammar tools that influenced people subliminally and shifted their opinions, even when they were warned. The arising question is if we can do something similar for bias, where people are nudged to perceive bias accurately by, for example, their grammar tool.

Goal (1 out of the 3 per project)

1. Find out how to generate LLM explanations for bias through prompt design and empirical user studies.
2. Find out how to generate LLM explanations for bias through finetuning an LLM and either evaluate either through statistical analysis or empirical user study
3. Find ways to increase bias perception subliminally, like through grammar tools.

Tasks

1. Figure out what a good explanation needs (length, bullet points, content, examples). We already have some in our study, but it could be extended from a pedagogical perspective.
2. Decide if you either want to go the prompting path, the finetuning path, or a mixed approach (e.g., use our classifiers to decide if it is biased and needs an explanation, or already give the LLM our classification result in the prompt).
3. Run a study to compare the most promising methods to generate explanations and see if participants move their bias perceptions towards the expert ground truth.
4. Analyze results and summarize in paper.



Smi Hinterreiter
s.hinterreiter@media-bias-research.org