Issue: 7 Date: November 2023

CAISS

Computation & AI for Social Science Hub

Newsletter in collaboration with The Alan Turing Institute and Lancaster University

CAISS Talk Series Reports

The first CAISS talk was held in September with a fabulous session from Dr Sharon Glaas on "Mitigating Researcher Bias in Linguistic Studies".

Sharon started by defining bias as "who gets to talk and who is listened to". For example do stay at home Mums have a voice? Sharon studies linguistics – the systematic study of language and communication – functional and descriptive not prescriptive e.g. linguistic sources of persuasion. She reminded us how the social world is studied based on how it is constructed. Some highlights:

- Linguistics frequently work in an interdisciplinary, multi disciplinary way – working with other disciplines highlights the issue of bias in ways of thinking.
- How you talk about something affects how you view it. E.g. Pro-life versus anti abortion.
- Constructivist versus positivist perspectives, the social world versus the real or natural world – what is the truth and how is meaning perceived?
- Bias is part of the world that we live in. We cannot remove it but we need to be aware of it and try to mitigate it.
- Media literacy is most important.
- One of the biggest red flags is the use of Large Language models (LLM's) and how they are being framed. AI does not know things it just repeats them.
- People "pull down" on large chunks of language and a LLM will just predict what the next word is.
- Language is an issue as LLM's do not learn.

Sharon also elaborated on her interesting work in a corpus assisted study of political and media discourses around the EU in the lead up to Brexit. She found that the pro –EU stance of the Guardian was systematically undermined by three key themes:

- Discourses of **Conflict** between UK / EU and EU / Member states
- Discourses of **Disparity** of citizen's experience (EU not working)
- Discourses of **Threat** to the UK and an existential risk to the EU.

We all have linguistic biases – ways of conceiving and talking about things that are grounded in our world view. Sharon does not believe it is possible to entirely eliminate bias from our work – but awareness and transparency help mitigate the issue. She stressed in her conclusion that it is important to understands the impact of those biases as use of LLM's and AI tools become more prevalent.

We had excellent feedback from Sharon's talk, one delegate said it was "the best one hour briefing they had heard in a very long time".

[ ] Work with us

Please email us if you wish to be added to the mailing list.

[ ] Contribute

Have something to share, please let us know and we can feature it in a future issue.

[ ] Spread the Word

Please feel free to pass this onto your colleagues.

[ ] Contact Details

caiss@lancaster.ac.uk

caiss@dstl.gov.uk

To email us at Dstl, scan QR code

DSTL/PUB153332

Continued from previous page:

CAISS were privileged to have Professor Wendy Moncur from the University of Strathclyde deliver our second talk in October.

Wendy leads the Cybersecurity Group and her research focuses on online identity, reputation, trust and cybersecurity and crosses many disciplinary boundaries. Her current research – the 3.6million AP4L project – develops privacy enhancing technologies (PET's) to support people going through sensitive life transitions. The research is looking at four transition groups: (i) living with cancer, (ii) leaving the armed forces, (iii) LGBT+ and (iv) relationship breakdowns.

Wendy talked to us about "Navigating bias in online privacy research". She stressed that it is important that we ask the right questions and whilst doing this we also ask the right people. As researchers what do we ourselves "bring" to the research as we use our own "interpretive lens" and it is important that we communicate our findings clearly so that others can understand the results.

Wendy then went on to discuss our individual online identities, this is co-constructed, made up of data about an individual posted by themselves and by other people and organisations. The internet in general is swimming in personal data, the minute we share anything we have lost control – once it is "out there" this information persists. Her explanation of how threads of personal data can be used to construct information regarding an individual was very thought provoking; e.g. if you share your Strava run data then someone can easily ascertain your home address or where you work if you run in your lunch break!

To mitigate against bias in the research Wendy advocated the following:

- Draw out information on digital privacy in sensitive contexts
- Allow for self reflection
- Foster participants' ability for self-expression
- Facilitate richer, more comprehensive stories and descriptions
- Enable non-experts to be heard
- Avoid assumptions and bias.

To further reduce the researcher bias and ensure that the vocabulary was robust the research team worked hard to increase the list of descriptive terms they used, checked out further terms with the University Librarian and also with the advisory board of people living with the transitions under investigation. This led to a very big list! For the workshops that ensued participants were asked to map their life transition on line with questions as prompts. Then empathy mapping was used to help further remove bias and deliver a shared understanding of the user across the research team. Next metaphor cards were used with the groups asked to consider potential technological solutions as opposed to just challenges, needs and practices. Finally participants ideas were prioritised using the MoSCoW tool (Must have, Should have, Could have, Will not have).

Sociodemographic groups were discussed in that older people 70 plus tend to read but don't comment on line, 30 to 60 year olds have a lot to say and younger people are happy to share information but in general have a more robust awareness of online security.

Results have indicated that the "Transition Continuum" is not a straight line and this is being explored further. Useful design insight for developing Privacy Enhancing Tools is that people's experiences are not necessarily linear or instantaneous and can extend over a long period. For the future privacy settings ideally need to be more like a dial than a switch.

Continued from previous page:

Article Review

Social Media Algorithms warp how people learn from each other, research shows.

William Brady, Assistant Professor of Management and Organisations at Northwestern University.

Interactions especially on social media are influenced by the flow of information controlled by algorithms. These algorithms are amplifying the information that sustains engagement – and could be described as "click bait". Brady suggests that a side effect of this clicking and returning to the platforms is that "algorithms amplify information that people are strongly biased to learn from". He has called this "PRIME" – prestigious, in-group, moral and emotional information. This type of learning is not new and would have served a purpose from an evolutionary perspective – learning from prestigious individuals is efficient as we can copy the successful behaviour. Also from a moral point of view, those who violate moral norms can be sanctioned as it would help the community maintain cooperation.

With social media this PRIME information is giving a poor signal as prestige can be faked and our feeds can be full of negative and moral information which will lead to conflict rather than cooperation. This can foster dysfunction as social learning should support cooperation and problem solving, but the algorithms are designed to increase engagement only. Brady calls this "mismatch functional misalignment".

So what, why does this matter?

People can start to form incorrect perceptions of their social world, this can lead to a polarisation of their political views, seeing the "in group" and "out-group" as being more sharply divided than they really are. The author also found that the more a post is shared the more outrage it generates. So when these algorithms amplify moral and emotional information the misinformation is included in this and is itself then amplified.

What next?

Research in this area is new and there is some controversy around whether this type of online polarisation being amplified spills over into the offline world is debateable. More research is needed to understand the outcomes that occur "when humans and algorithms interact in feedback loops of social learning". For research to continue ethical concerns such as privacy need to be considered. Brady would like to see "what can be done to make algorithms foster accurate human social learning rather than exploit social learning biases". He suggests we need an algorithm that "increases engagement while also penalising PRIME information".

Link: https://www.scientificamerican.com/article/social-media-algorithms-warp-how-people-learn-from-each-other/

****CAISS Autumn/Winter Speaker Series****

Each month from November – December 2023 we will be hosting a talk from an expert in their field addressing an aspect of bias in their work.

Dr Lewys Brace, **7th November 2023 -** Senior Lecturer in Computational Social Science, University of Exeter

Xiao Hui Tai 5th December 2023 - Assistant Professor in Statistics from the University of California Davies

Look out for the Teams meeting links which will be dropping into your mailbox before each talk. Unfortunately we cannot record these talks. Each one will take an hour in total with a chance for some questions and answers at the end. They promise to be fascinating!

Don't miss these brilliant speakers

Continued from previous page:

CAISS Bytes

Anirban Ghosal, senior writer for Computerworld, August 2023 discusses how OpenAI are planning to use GPT-4 LLM for content moderation, and how this could help to eliminate bias. By automating the process of content moderation on digital platforms, especially social media, GPT-4 could interpret rules and nuances in long content policy documentation, as well as adapting instantly to policy updates. The company believe AI can help to moderate online traffic and relive the mental burden on a large number of human moderators. The company posit that custom content policies could be created in hours, and they could use data sets containing real-life examples of policy violations in order to label the data. Traditionally people label the data and this is time consuming and expensive.

People will then be used to read the policy and assign labels to the same dataset without seeing the answers. Using these discrepancies the experts can ask GPT-4 to explain the reasoning behind its labels, look into policy definitions, discuss the ambiguity and resolve any confusion. This iterative process will have many steps with data scientists and engineers before the LLM can generate good useful results.

So What: Using this approach should lead to a decrease in inconsistent labelling and a faster feedback loop. Results should be more consistent. Undesired biases can creep into content moderation during training, although results and output will need to be carefully looked at and further refined by maintaining humans in the loop, therefore, bias could be reduced. Industry experts suggest that this approach has potential and could lead to a massive multi-million dollar market for Open AI.

Link: https://www.computerworld.com/article/3704618/openai-to-use-gpt-4-llm-for-content-moderation-warns-against-bias.html

An article in Nature magazine discusses:
How can we "stop deepfakes from sinking society"

It is easy for AI to generate convincing images and videos, but we need to ensure we can guard against the harm such deepfakes could cause. The US Defence Advanced Research Projects Agency lead, Wil Corvey says we should question not "how much of this is synthetic" but instead "why was this made?" One problem identified is that "people are not used to generative technology" says Cynthia Rudin from Duke University, North Carolina. There is no degree of scepticism as this technology has exploded onto the scene rather than developed slowly. Some deepfake images are fun and for entertainment, but others can be used to deceive and carry out fraudulent activities. One way to help detect these synthetic images is to watermark them by altering pixels in a method that is imperceptible to the naked eye but is picked up on analysis; or to tag a file's metadata to authenticate an image.

So What: There is a losing battle occurring to detect deepfakes, we need greater technological literacy and tools at our disposal to counteract the harmful deepfakes. As our own Dr Sophie Nightingale says "People's ability to really know where they should place their trust is falling away. And that's a real problem for democracy". With major elections due in many countries there is possibly a big threat to contend with.

Link: https://www.nature.com/articles/d41586-023-02990-y