Issue: 11 **Date: September 2024**

CAISS

Computation & AI for Social Science Hub

Newsletter in collaboration with The Alan Turing Institute and Lancaster University

CAISSathon, Exeter 2024

CAISS recently ran a two day workshop, "The CAISSathon" at the University of Exeter with over 40 participants from Industry, Academia, Dstl and other government agencies. The aim was to develop ideas for research projects around the transparency and explainability of AI systems, whilst touching on other areas such as fairness. The event opened with two fascinating keynote talks from Dr Ali Alameer from the University of Salford on Communicating Fairness in AI and Major Jordan Blee on Trust and Transparency – Utilising Artificial Intelligence within a military context. Both talks received considerable interest and sparked interesting lively discussions and debates.

Delegates were then divided into 7 teams and decided on a pre-prepared challenge in the area of transparency and explainability to work on over the next day or so. They could also have come up with their own idea, although some of the pre-prepared challenges were modified by some groups. An example challenge was:
"Is it sensible and/or achievable, to try to apply explainability to all aspects of an AI based system? If not what aspect should it apply to and how can the developer be guided to identify those aspects?" By the end of the first day each team gave a very brief summary stating the challenge they had chosen and the work they had done.
Day two started with a brief overview given by Dr Sophie Nightingale, who leads the Lancaster CAISS team, with a chance for people to ask any questions they may have had. This was then followed by a Red Teaming exercise where the teams paired up with one team, the Blue team, explaining what they had done so far and the other team, the Red Team, constructively challenged the approach, enabling the Blue team to reflect and respond to the challenges. Once each team had been Red teamed, they finalised their work and prepared short five minute presentations which were given at the end of the event. Some of these presentations were very professionally produced! The "Skewed Squad" team were voted for as the best idea by everyone with their idea around the ethnomethodology of decision making.
What next: The CAISS team will be summarising the outcomes and hopefully some of these outcomes from the event will feed into future research. CAISSathon 2 will go into the planning stage for next year, taking into account the feedback received from delegates, which was overwhelmingly positive.

[dstl]
The Science Inside

The Alan Turing Institute

Lancaster University

[ ] 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 the QR code

DSTL/PUB161604

Conference Reports CAISS out & about

AI vs Human Guidance

Dr Joe Pearson from the Lancaster CAISS Team presented at the BASS24 conference held this year in St.Andrews, Scotland, on some of the work that the CAISS hub has been conducting: He presented on how Artificial Intelligence (AI) is increasingly used to assist decision-making in public and private sector organisations. However, there is a lack of research and understanding about whether humans use algorithms effectively when making decisions. The current work examined whether humans successfully utilise guidance purportedly provided by humans or an algorithm in a decision-making task. Participants (N = 295, M_{age} = 33.79) provided judgements on the authenticity of 80 faces (40 real, 40 artificial) presented alongside guidance ostensibly from human experts or an algorithm that was correct only 50% of the time. Analyses revealed that individuals were equally likely to make judgements consistent with guidance from humans or an algorithm but, encouragingly, were more likely to respond consistently with guidance when it was correct. Signal detection analyses identified a tendency to classify faces as real, alongside a reduced ability to discriminate between real and synthetic faces with greater self-reported trust in AI, when participants received algorithmic guidance.

So what: These results are promising as they suggest humans can follow AI guidance when it is correct and disregard it when it is not. However, further work is required to understand human-algorithm interactions when guidance accuracy is higher and, therefore, is considered more trustworthy. One takeaway from the conference was that there are calls for research at the intersection of human and AI relations, and calls for work examining instances of this relationship where algorithms designed to support decision-making are improperly deployed as decision-makers themselves.

IC2SC

Whilst Joe was in Scotland, Dr Matt Asher was in the USA presenting at the IC2S2 conference on using AI for Literature reviews. IC2S2 is the annual conference of the International Society for Computational Social Science. Matt gave a presentation titled, “Literature Analysis with Algorithms, AI and Human Assistance” which was in the “Methods & Innovations” session, alongside other presentations on using LLMs for processing social data. He presented an overview of CAISS’s work on automating literature reviews, using AI, computational techniques and manual literature review techniques. While there was little time to go into much detail of any of the work streams individually, the point was the overall comparison between the three approaches, the conclusion of which is that while there are only small areas of overlap between the three results, they each have a component to contribute to automatically working through increasing amounts of existing literature. This introduced the idea of the Textual Information Measurement & Mark-up Interface (TIMMI) tool developed by the CAISS team, it is a visual literature analyses tool concept, which, despite being far less technically deep than many of the surrounding presentations, seemed to be well received as a concept.

So what: Humans cannot be replaced yet in terms of automating literature reviews, but augmenting human work in this area could lead to faster results and be more efficient. Where time and budgets are tight this approach could be an ideal solution but it should not simply be used as a “short cut” to high quality research.

"AI now beats humans at basic tasks – new benchmarks are needed, says major report".

Artificial Intelligence (AI) is now so sophisticated that it can at times nearly match or exceed humans in tasks such as reading comprehension, image classification and competition level maths. As the development is so rapid the benchmarks and tests for evaluating these are becoming obsolete. The Artificial Intelligence Index Report 2024 which was published by Stanford University makes fascinating reading as it documents the meteoric progress in machine learning systems over the past decade. https://aiindex.stanford.edu/report/

The report is hefty at over 400 pages long and interestingly was copy edited with the aid of AI tools. It notes that there is a lack of standardised assessments for the "responsible use of AI" and this creates problems when assessing systems in terms of risks.

Industry is responsible for some of the rapid rise in "notable machine learning systems last year" (2023) with 51 systems, with academia contributing less at around 15; this could be due to academia spending more time on evaluating and testing systems. This "tough testing" is necessary to assess the large language models (LLMs) which are behind chatbots. One such test "the Graduate Level Google Proof Q&A benchmark (GPQA) has been developed by David Rein at New York University. It is made up of around 400 multiple choice questions, PhD level scholars can obtain around 65% correct answers in their field, chance would be 25% and when scholars had access to the internet they could only score 34% outside of their respective fields. AI systems were scoring 30-40% correct last year but Claude 3 from AI company Anthropic scored around 60%, this illustrates the rapid advancements and how hard it is to have a benchmark test.

So What: Assessing AI using benchmarks needs to keep pace with the exponential development of systems. Historically a benchmark could be useful for up to ten years, now it is for just a few. This rapid rise in development calls for ethical regulation which will promote responsible AI use as well environmental consideration. Costs are also an issue, energy use is high and the amount of water used to cool the data centres is enormous. As performance of AI systems rises so will the costs and potentially also the risks. Less is more could really be an ideal future position in this regard.
Link to full article: https://www.nature.com/articles/d41586-024-01087-4

AI Systems could run out of training data

A report by the Associated Press hypothesises that within 2-8 years the text required to train LLMs will either run out or have created a bottleneck. Computing researcher Tamay Besiroglu says it will be hard to "really scale up models efficiently anymore" in this scenario. However, computer engineer Nicolas Papernot counters with "we don't necessarily need to train larger and larger models".

So what: Scaling up models historically has been an important way of expanding capabilities and improving the quality of a models outputs. More skilled AI models for specialised tasks will mitigate against model collapse where the AI systems are trained using the same outputs they are producing. This can also result in the bias, any unfairness and mistakes becoming baked into the model data. High quality data is needed whether this is from humans or synthetic, but currently there is not a "gold standard" solution to this evolving problem.

Link: https://apnews.com/article/ai-artificial-intelligence-training-data-running-out-9676145bac0d30ecce1513c20561b87d

CAISS Autumn/Winter Talks 24

CAISS Talk Series – These will be hosted on MS Teams

Tuesday 1st October 2024	Dr Suzanne McClure, Exeter University
Tuesday 12th November 2024	Dr Victoria Nockles, The Alan Turing Institute
Tuesday 3rd December 2024	Dr Alex Hardy, Liverpool University
Tuesday 14th January 2025	Professor Steve Meers, Dstl

Invitations will be sent out by email nearer the time – everyone welcome

Are chatbots corrupting peer review?

Dalmeet Singh Chawla writes in Nature magazine about the report that has identified dozens of adjectives that could signal a particular text has been written with the aid of chatbots. These "buzzwords" could be the hallmarks of Artificial Intelligence (AI)-written text in peer review reports. Are researchers turning to ChatGPT and other AI tools to evaluate the work of others? Up to 17% of peer review reports analysed since the release of ChatGPT that were submitted to four major computer science conferences could have been modified. What is unclear is whether AI was used to write the whole report or used as an editing tool. Debora Weber-Wulff, a computer scientist at the HTW Berlin – University of Applied Sciences in Germany says we should be concerned considering how often AI tools "hallucinate"; this is where an AI tool invents facts or data to fill in missing information and is often wrong. The problem with this hallucinating is that we do not know when it is happening.

A survey in Nature in 2023 *Nature* survey, had 1,600 scientists respond who said they had used generative AI to write papers, with 15% saying they had used it for literature reviews and to write grant applications. In the asXiv study undertaken by a team led by Weixin Liang who is a computer scientist at Stanford University, California, the team developed a technique to search AI text for adjectives used more frequently by AI than humans. Some of the adjectives that had increased significantly since the chatbots use had become mainstream are: commendable, innovative, meticulous, intricate, notable and versatile. When people were short of time and submitted their papers nearer to deadlines the use of these adjectives increased. However, a study looked at 25,000 peer reviews with 10,000 manuscripts accepted for publication across 15 Nature Portfolio journals between 2019 and 2023 and did not find the same "spike" in usage of the same adjectives since the release of ChatGPT. Some publishers e.g. Springer, asks peer reviewers not to upload manuscripts into generative AI tools.

The work of Liang inspired Andrew Gray a bibliometric support officer at University College London to look at peer reviewed studies between 2015 and 2023 to ascertain if the same adjectives appear as well as certain adverbs. He found that "commendable", "meticulous" and "intricate" usage had increased since ChatGPT. Estimates in this study suggested that 1% of all scholarly studies published in 2023 had to some extent used chatbots.

So What: Is it bad or good to use AI tools to help with reviewing papers? Perhaps it is more important that transparency and accountability are the focus with an estimate of how much of the review text has been produced by using generative AI and acknowledge the fact it has? Weber-Wulf feels that "peer review has been corrupted by the use of AI systems". There are also copyright implications as tools will be given access to "confidential, unpublished material". However, using certain adjectives to identify if AI tools have been used could simply be due to using automatic language translation systems. Greater transparency is needed to address this.

Link: https://www.nature.com/articles/d41586-024-01051-2

CAISS Notes podcast: What is CSS? Here: https://wp.lancs.ac.uk/caiss/podcasts/