Issue: 9 Date: March 2024

CAISS

Computation & AI for Social Science Hub

Newsletter in collaboration with The Alan Turing Institute and Lancaster University

Stop Press….Stop Press….Stop Press….Stop Press

The CAISSathon – 22nd & 23rd July 2024

New date & venue announced. Please find details on page 4, sign up now!

CAISS Editorial

CAISS spoke to James Rosie, former Senior Principal Anthropologist at Dstl and asked him for a few words on Artificial Intelligence (AI) and its ethical use.

"Given the potential for AI to shape the world around us, from managing our social media fields to the potential for self-driving vehicles, it is equally important to ask the question of what is 'right' when we use AI? However, that said, at a glance I can see how broad and complicated the questions are around AI and ethics, reflected in some of the fascinating work I have seen across the Lab in Dstl."

"Working together with Fellows from the Alan Turing Institute, the Behavioural & Social Science group, and now Dstl Newcastle, some ground breaking work asking questions around the ethics of AI and autonomous systems has been done. Although we have led research informing UK policy around weapons systems', other research has focused on how we create and implement these AI systems. One of the most important principles is that of 'explainable AI' and the idea that we should not create, let alone rely on, AI technologies if we do not understand how they produce the answer that they do. As well as the practical considerations of not 'showing the working', this would also raise ethical, and legal questions about how we attribute responsibility in life and death decisions."

"Working on the Creative Futures project, within Defence S&T Futures (more info on next page); Dstl researchers discussed with authors from the British Science Fiction Association the question of whether it would be ethical for an exceptionally powerful AI to operate behind the scenes; protecting us from harm, but at the same time perhaps depriving us of our autonomy and free will. Even more relevant to this group were the questions around large language models (LLM), such as Chat-GPT, ruining the livelihoods of professional writers. Though no one believes Chat-GPT will create the next great novel, it is a fact that already much of the commissioned work artists and authors rely on for their living, is now being done by such LLMs. To add insult to this injury, these LLMs are often being 'trained' on work produced by these authors, without their permission."

"Throughout all of these, the term 'AI' has been used to refer to algorithms such as machine learning, Large Language Models and other techniques, or their use. The questions of whether it could be possible to generate a true 'consciousness' and if it was, would it be Artificial Intelligence or simply 'Intelligence', (perhaps AI then becomes 'Another Intelligence') are even more complex" .

Continued on next page…

[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/PUB156288

Continued from previous page:

"Questions of ethics and morality are also frequently relative to who we are, so with a wider anthropological perspective, we could ask if those from cultures that did not grow up with films such as 'The Terminator' and '2001: A Space Odyssey' would have similar views on AI as we do? It is not the intent to offer any definite answers here, but rather to provide some indication of the range and breadth of the questions being asked. Not only those questions yet to be asked, but also that these are not purely theoretical and are already shaping the world around us."

Let us know what you think? We would love to hear your thoughts….

More info on Dstl Futures here:
https://www.gov.uk/government/publications/unfogging-the-future-a-dstl-biscuit-book

The EU AI Act – what does this mean for innovation?

European Union member states and the European Parliament have worked to publish "the AI act" to enable a framework of "staggered rules" based on risk These include items such as:

- Unacceptable risk systems – tech that poses a threat to people will mostly be banned
- AI systems must respect EU copyright rules and be transparent around how generative AI models have been trained
- General purpose tools like Chat GPT will be assessed on how powerful they are. Tools trained using large amounts of computing power would face more obligations and reporting restrictions.

So what: Companies will not have to implement the rules for 2 years, in which time they could be out of date before they are even implemented. Will these rules stifle innovation in this fast moving sector? Link here: Euronews

Butterflies and Chat GPT

"Prompting "is the way we talk to generative AI and large language models (LLM's). The way we construct a prompt can change a models decision on the results it provides and impact the accuracy as well. Research from the University of Southern California Information Sciences Institute shows that a minute tweak - such as a space at the beginning of a prompt can change the results. This is likened to chaos theory where a butterfly flaps its wings generating a minor ripple in the air, resulting in a tornado several weeks later in a faraway land.

The researchers, who were sponsored by the US Defense Advanced Research Projects Agency (DARPA), chose ChatGPT and applied various different prompt variations. Even slight changes led to significant changes in the results. They found many factors at play and there is more work to be done to ascertain solutions to this effect.

Why do slight changes result in such significant changes? Do the changes "confuse" the model? By running experiments across 11 classification tasks, they were able to measure how often the LLM changed its predictions and the impact on accuracy. By studying the correlation between confusion and the instances likelihood of having its answer changed (using a subset of the task with individual human annotations), they did not get a full answer.

So what: Generating LLMs which are resistant to changes and yield consistent, accurate answers is a logical next step. However, this will require a deeper understanding of why responses change under minor tweaks. Is there a way we can anticipate these resulting changes in outputs? With ChatGPT being integrated into systems at scale this work will be important for the future.

Link to original paper: https://arxiv.org/pdf/2401.03729.pdf

Continued from previous page:

Article Review

" In the Artificial Intelligence (AI) Science boom, beware: your results are only as good as your data. "

Hunter Moseley shines a light on how we can make our experimental results more trustworthy. Thoroughly vetting them before and after publication will ensure huge complex data sets are both accurate and valid. We need to question results and papers; just because it has been published does not mean it is accurate or even correct in spite of who the author may be and their credentials.

The key to ensuring the accuracy of these results is reproducibility, careful examination of the data with peers and other research groups investigating the outcomes. This is vitally important with a data set that is used in new applications. Mosely and his colleagues found something unexpected when they investigated some recent research papers. Duplicates appeared in the data sets which were used in three papers meaning they were corrupt.
In machine learning it is usual to split a data set in two and to use one subset to train a model and the other to evaluate the performance of this model. With no overlap between training and testing subsets, performance in the testing phase will reflect how well the model learns and performs. However, in their examination they found what they described as a "catastrophic data leakage" problem in that the two subsets were cross contaminated, thereby messing up the ideal separation. About one quarter of the dataset in question was represented more than once, corrupting the cross-validation steps. After cleaning up the data sets and applying the published methods again the observed performance was a lot less impressive with a drop in the accuracy score from 0.94 to 0.82. A score of 0.94 is reasonably high and "indicates that the algorithm is usable in many scientific applications", but at 0.82 it is useful but with limitations and then "only if handled appropriately".

So What: Studies that are published with flawed results obviously call research into question. If researchers do not make their code and methods fully available then this type of error can occur. If high performance is reported this may lead to other researchers not attempting to improve on results, feeling that "their algorithms are lacking in comparison." Some journals like to publish reviews of successful results so this could prevent progress in research as it is not considered valid or even worth publishing!

Encouraging reproducibility:

Moseley argues that a measured approach is needed. Where transparency is demonstrated with data, code and full results being available, a thorough evaluation and identification of the problematic dataset would allow an author to correct their work. Another of his solutions is to retract studies with highly flawed results and little or no support for reproducible research. Scientific reproducibility should not be an option.
Researchers at all levels will need to learn to treat published data with a degree of scepticism, the research community does not want to repeat others' mistakes. But data sets are complex, especially when using AI. Making these data sets and the code used to analyse them available will benefit the original authors, help validate the research and ensure rigour in the research community.
Link to full article: https://doi.org/10.1038/d41586-024-00306-2

Continued from previous page:

New Date – July 2024, The CAISSathon

Due to unforeseen circumstances we have had to reschedule the CAISSathon but are delighted to have a new date

Where: Exeter University

***When*: 22nd and 23rd July 2024**

Theme of the event: Explainability

The social responsibility of Artificial Intelligence (AI) has been under increased scrutiny as it creeps into every facet of society. Understanding the potential ramifications and harm caused by AI is of key importance, particularly as AI technology used for facial recognition, loans and mortgages and job applications (amongst others), have already been shown to be biased against ethnic minority populations and, for example, those with disabilities. Within a Defence context, understanding how AI enabled technologies can facilitate decision making is of key interest. The need for explainable and transparent AI systems is one argument to uncover bias and prevent harm. However, does explainability solve these issues? Does understanding how AI makes it decisions provide enough evidence to negate the harm or at least provide indications of potential harm? Additionally, how understandable do such explanations need to be for expert and lay users?

The Computation and AI for Social Science Hub would like to invite you to participate in our first 'CAISSathon' to explore how explainability can be conceptualised and implemented. This event will propose a series of challenges that will be collaboratively addressed throughout the two days. It will bring together individuals from government, academia and industry to brainstorm and engineer potential solutions. Researchers will have an opportunity to put knowledge into practice and solve problems which have real life implications within Defence. At the end of the two days, a portfolio of potential solutions, research questions and collaboration will be established which will inspire future investigation and lead to insightful developments in this fast moving field. Additionally, there will be a prize awarded to the team who prepare the most innovative solutions. Come and join us at this exciting event.

For more information or to request an invite please email us, at <u>any</u> of the details on the first page

Super-intelligent AI is not a thing

Panic not – says a report in Nature, LLM's will not have the ability to match or even exceed human beings on most tasks. "Scientific study to date strongly suggests most aspects of language models are indeed predictable," says computer scientist and study co-author Sanmi Koyejo. Emerging artificial "general" intelligence is no longer apparent when systems are tested in different ways. This "emergence", when AI models gain knowledge in a sharp and predictable way is nothing more than a mirage with systems' abilities building gradually.

So What: Models are making improvements but they are no where near approaching consciousness, perhaps benchmarking needs more attention paid to it – working on how tasks fit into real world activities. Link to article:
https://doi.org/10.1038/d41586-023-04094-z