“Meet the Future of AI: Countering Sophisticated and Advanced Disinformation”

Summary of Conference Discussions and Outcomes

K. Bontcheva (University of Sheffield)

The event "Meet the Future of AI: Countering Sophisticated & Advanced Disinformation" took place in Brussels on 29 June 2023. It was co-organised by the Horizon Europe-funded research projects AI4Media, AI4Trust, TITAN, and vera.ai – together with the European Commission. We welcomed close to 100 participants on-site at VRT in Brussels, as well as an online audience of over 150 participants online.

Challenges and Opportunities

First we discussed the dangers of generative AI being harnessed by disinformation agents to launch highly credible large-scale disinformation campaigns across different platforms and media. Unfortunately the latest Large Language Models, including ChatGPT, are not designed to “speak the truth”, and getting them to generate highly convincing disinformation is very easy and extremely cheap. Even when the ChatGPT model itself admits that it is not supposed to generate disinformation, it still provides false text as well.

Consequently, as the fluency and affordability of LLMs increase from one month to the next, so does the threat of their wide-ranging misuse for the creation of affordable, large-scale disinformation campaigns.

So we need to act against this and do so urgently! One of the ways that we are foreseeing to act against this is to, for example, have AI-generated images and videos watermarked for provenance. Watermarking, however, is not a panacea, while at the same time it is getting harder and harder for humans (and AI tools) to tell apart authentic from AI-generated content.

Policy Responses

In terms of policy responses, we discussed at length the potential of the AI Act, the DSA (Digital Services Act) and the 2022 Code of Practice on Disinformation. At present, the Code of Practice is the key self-regulatory tool aimed at helping researchers fight disinformation (pending the incoming DSA and the AI Act).

However, a key limitation of the Code is in its self-regulatory nature, as there are no consequences for the Very Large Online Platforms (VLOPs) when they drop out from the Code of Practice, or when they are sending insufficiently informative reports. At the same time, the Code puts the onus on AI researchers to prove that VLOPs are not doing enough with respect to tackling disinformation, without the necessary mechanisms in place to ensure that the vital data which researchers, journalists, and fact-checkers need for this is made freely available by the VLOPs. Therefore, in order to make the Code a success in terms of accountability and transparency, an urgent coordinated action is needed between all stakeholders, led by policy makers.

There's also a big ethical dimension to all of this, and the AI research on disinformation going forward. There are studies, for example, by media organisations using AI tools that detect hate speech on Facebook and thus prove that Meta’s current moderation practices fall remarkably short. To do this, however, journalists subscribe and follow private Facebook groups, which unfortunately many AI and social science researchers cannot do due the strongly enforced restrictions of their ethics boards.

Barriers

At the same time, companies are scraping millions of images off the internet and using them to train face recognition software. Again wouldn't all researchers love to have that data set?

Research institute's ethics boards will clearly reject such activities, too, and they would be right to do so. However, there are companies out there that do that, and yet, they are not facing ethical or legal repercussions. We are thus facing a power asymmetry problem that is actually due to commercial organisations that are pouring billions into LLMs using scraped data off the internet. The key question then becomes: can we enlist policy makers, legal and ethics experts, and other stakeholders to set out new ground rules for what is acceptable for companies and researchers to do in terms of data access, scraping, and model training?

This leads us to the biggest major stumbling block that researchers face: lack of data access. Without a doubt, all major societal debates around elections, wars, pandemics, and other major events are also taking place on social media platforms, and yet, researchers are increasingly facing bigger barriers to studying these. Most recently Twitter withdrew free data access for researchers, but so are Meta, Reddit, TikTok, etc.

So on one hand, we need to have effective responses against AI-generated disinformation using the models these very large platforms create and train on the social media data they have, but at the same time, researchers are denied access to that same data for the purpose of research in the public good, especially with respect to training state-of-the-art disinformation and hate speech detection AI models. So, again, there's a massive imbalance here.

There's also a massive imbalance in terms of funding received. Again, on one hand companies invest billions into LLMs and their NLP (Natural Language Processing) and speech processing labs with hundreds of very highly paid researchers, while at the same time EU and national funders can barely afford tens of millions across a handful of research projects, and researchers need to bid competitively for this every few years. This unfortunately leads to a large overhead in terms of simply sustaining even a tiny advantage that researchers may have managed to create.

Of course, these project-based funding models mean that the effort of each research project and lab are pretty much "siloed", time-bounded and inevitably there are certain overlaps between them, which further diminishes the scale that researchers can achieve.

At the same time, policy responses take years to develop, whereas generative AI models barely take months. So how do we reconcile these matters? Researchers are already pretty much behind the curve in terms of their ability to train LLMs, so this issue needs to be addressed very urgently, by all relevant stakeholders.

Potential Research Avenues

Researchers from the four Horizon Europe-funded research projects that co-organised the event (AI4Media, AI4Trust, TITAN, and vera.ai) are in the process of developing state-of-the-art AI models for the detection and analysis of online disinformation, including coordinated campaigns, AI generated images and videos, ChatGPT-generated disinformation, etc. These are all areas of much needed research and it is giving some hope going forward.

But at the same time, given the barriers we face, it is clear that we as researchers need to join forces in order to succeed. To best exploit the limited data and funds available, we as researchers may need to go a little bit against our current research practices, which means that different research groups often compete with each other to produce the best models and the most cited publications. However, for the benefit of society, it seems advisable to seriously think about changing the way we work, making the best use of our limited resources, as compared to those available to companies.

One lever that we have going forwards is to try and lobby with funders, the Commission, and other stakeholders. Good leverage for that are the forthcoming EU and UK elections. Given the threat of AI-generated disinformation, the stakes have never been higher.

Next Steps

1. **Urgent call for mediation and data access:** researchers from these four projects have undertaken to work together with the EDMO research community and produce a common letter emphasising our difficulties with data access from the VLOPs, as well as ethical and legal dimensions, in an attempt to lobby for urgent mediation between the research community and the VLOPs. Data access within the next 6 months is vital because without that our projects and efforts will not be as successful as they otherwise would be. Further co-signatories from other European and nationally funded projects will be sought too.
2. **Working group of European AI researchers on countering disinformation:** working in close collaboration with EDMO, other stakeholders and users of AI disinformation detection tools (e.g. fact-checkers), we will aim to establish a working group and thus create a common voice and hopefully help us reach the required critical mass needed to elicit urgent action from policy makers, VLOPs, etc.
3. **Further joint events and forums** will be organised from the autumn 2023 onwards, starting with panels on AI's impacts, data dependence, and transparency at, e.g., the 2023 Conference on Disinformation in Krakow and a second edition of this joint research conference, planned for mid-2024.
4. **Creating training datasets together:** data from the platforms is badly needed, but isn't enough. Researchers badly need human-labelled data to fine-tune AI algorithms. Researchers from the four projects have thus committed to identifying common types of labelled data that they need, and to join forces in creating such annotated data for training and evaluation much faster, in more languages.

The Big Ask: CERN-like European Infrastructure for AI Research and Open Source Tools

Other than Internet-scale datasets, very large compute facilities (including hundreds of powerful GPUs) are the second key enabler of LLM development. This is yet another uneven playing field where companies have a huge advantage over publicly-funded AI researchers, especially those from smaller EU countries such as, e.g., Bulgaria, Romania, and Slovakia. This begs the question of how the European Commission and national funders can work together to create a very large, shared hardware infrastructure and facility, which can then transform AI research much in the same way in which CERN transformed physics research.

The challenge that we are facing now with AI and the damages it can do to society is very, very significant and we need to not only act fast, but also act together, especially as Europe has many languages.

Such a joint facility would not be sufficient without it being complemented by open source tools for data access, transformation, and processing. They are badly needed not only for replicability and transparency reasons, but also to avoid duplication of the already scarce effort that publicly funded researchers have at their disposal.

Even if we take just these four research projects, each of them has spent some research effort on data collection from platforms such as Telegram, TikTok and YouTube, as well as data cleaning, storage, harmonisation, and access.

Therefore, such open-source tools would enable the research community to solve these basic data access and storage issues together, and to really focus our scarce resources on the AI research itself, which is what can really make a difference.

Contact: K.Bontcheva@sheffield.ac.uk

More on EDMO: www.edmo.eu