STEPHANE BAELE

AI AND EXTREMISM: THE THREAT OF LANGUAGE MODELS FOR PROPAGANDA PURPOSES

"These fast developments come with excitement and hype, but also serious concerns." Stephane Baele highlights the potential misuse of language models by extremist actors for propaganda purposes.

Recent large-scale projects in the field of Artificial Intelligence have dramatically improved the quality of language models, unfolding a wide range of practical applications from automated speech/voice recognition and autocomplete to more specialised applications in healthcare and finance. Yet the power of this tool has also, inevitably, raised concerns about potential malicious uses by political actors. This article highlights the threat of one specific misuse: the potential use of language models by extremist actors for propaganda purposes.

THE RISE OF LANGUAGE MODELS

Language models are statistical models that calculate probability distributions over sequences of words. Over the past five years, language modelling has experienced massive improvement – amounting to no less than a 'paradigm shift' according to some researchers (Bommasani *et al*. 2021) – with the rise of 'foundation models'. Foundation models are large language models with millions of parameters in their deep learning neural network architecture, trained on extremely large and broad data, which can be adopted to a wide range of downstream tasks with minimal fine-tuning.

The development of these models is very expensive, necessitating large teams of developers, numerous servers, and extensive data to train on. As a consequence, performant models have been created by well-endowed projects or companies like Google (BERT in 2018), OpenAI (GPT- 2 in 2019, GPT-3 in 2020), and DeepMind (GOPHER in 2022), who entered a race to design and deliver the most powerful model trained on the biggest base corpus, implementing the most parameters, and resting on the most pertinent architecture. GPT-3, for instance, was trained on approximately 500 billion words scraped from a wide range of internet spaces between 2016 and 2019; its development is estimated to have costed over $15million on top of staff salaries. Microsoft started an investment in OpenAI of no less than $1billion in July 2019.

WARNINGS OF MALICIOUS USE

These fast developments come with excitement and hype, but also serious concerns. As Bommasani and colleagues (2021, pp.7-8) ask, "given the protean nature of foundation models and their unmapped capabilities, how can we responsibly anticipate and address the ethical and social considerations they raise?"

A series of warning signs revealed some of these 'ethical and social considerations', triggering increasing anxiety. Back in 2012, IBM noticed that its Watson model started using slurs after the scraped content of the Urban Dictionary was integrated in its training corpus. Four years later, Microsoft had to shut down the Twitter account it opened for its Tay model less than a day after it was launched after a series of users effectively fine-tuned the chatbot into an unhinged right-wing extremist (claiming, among many others, that "feminists should burn in hell" and that "Hitler was right").

These problems echo broader worries about AI in general, with other techniques like deepfakes or molecules toxicity prediction models generating critical controversies and concerns about seemingly inevitable malicious uses.

The leading AI companies have therefore attempted to typologize and explore the various potential areas/types of malicious use and ethical issues posed by large-scale language models. OpenAI, for instance, published several reviews (Solaiman *et al*. 2019; Brown *et al*. 2020), and commissioned an assessment from the Middlebury Institute of International Studies at Monterey to evaluate the risk that their model could help produce extremist language (McGuffie & Newhouse, 2020).

DeepMind similarly released a report (Weidinger *et al*. 2021) highlighting six specific risk areas associated with their GOPHER model: 'Discrimination, Exclusion and Toxicity', 'Information Hazards', 'Misinformation Harms', 'Malicious Uses', 'Human-Computer Interaction Harms', and 'Automation, Access, and Environmental Harms'. At the same time, a scientific literature has emerged that evidences models' ingrained biases and experimentally tests the credibility of texts produced by foundation models.

© *sdecoret* | stock.adobe.com

> **...more extremist propaganda of any format can be produced in less time by less people.**

Worrying conclusions have pointed to the production of highly credible fake news and the potential of these models for campaigns of disinformation (Kreps *et al*. 2020; Buchanan *et al*. 2021). Across all these studies, a key claim holds consensus: the real power of language models is not so much that it could automatically produce large amounts of problematic content in one click (they are too imperfect for truly achieving that), but rather that they enable significant economies of scale. In other words, the cost of creating such content is about to plummet.

For terrorism and extremism experts, this evolution is deeply worrying: it means that much more extremist propaganda of any format can be produced in less time by less people. Yet at the exception of OpenAI's commissioned report by McGuffie and Newhouse, none of the existing explorations seriously considers this risk – even though several commentators have claimed that these models "can be coaxed to produce [extremist manifestos] endlessly" (Dale, 2021, p.116).

McGuffie and Newhouse's report already provided a much-needed first exploration of how language models can be used to produce extremist content, using a series of prompts to get GPT-2 to write radical prose from various ideological flavours. Yet the real potential of language models to create truly credible extremist content of the desired type and style through fine-tuning remained unevaluated.

EXTREMIST USE OF LANGUAGE MODELS: KEY OBSERVATIONS AND PRACTICAL IMPLICATIONS

We took up the task of rigorously evaluating the possibility of a foundation language model to generate credible synthetic extremist content. To do so, we adopted the idea of a 'human-machine team' (Buchanan *et al*. 2021) to design an optimal workflow for synthetic extremist content generation – by 'optimal' we mean the one designed to generate the most credible output while at the same time reflecting the constraints likely to restrict extremist groups' use of the technology (e.g., technological sophistication, time, pressures, etc.).

Working with various types (e.g., forum posts, magazines paragraphs) and styles (e.g., US white supremacist, incel online discussion, ISIS propaganda) of extremist content, we implemented that workflow with varying parameters to generate thousands of outputs. This systematic work immediately unfolded two main findings:

1. Even with the best variation of the workflow, the model generated a lot of 'junk', that is, content that is immediately not credible. While that proportion would shrink with bigger fine-tuning corpora, our study's commitment to a realistic setting makes the production of 'junk' inevitable. Most of the remaining synthetic content was deemed credible only after minor alterations by a lingo expert (correcting mistakes such as geographical inconsistencies), while a small minority was judged to be immediately highly credible.
2. The model is usually very good at using insulting outgroup labels in a pertinent way, and generating convincing small stories. However, as Dale puts it (in another context), the text get "increasingly nonsensical as [it] grows longer" (Dale, 2021: 115). Generally speaking, the longer the generated text, the bigger the need for a post-hoc correction by a human.

SURVEY RESULTS

To more rigorously test the credibility of the synthetic output beyond these two observations, we ran two survey experiments testing the credibility of a randomly selected sample of two types/styles of extremist content (ISIS magazine paragraphs in survey 1, and incel forum posts in survey 2), asking academics who have published peer-reviewed scientific papers analysing these two sorts of language (not simply ISIS or incel communities) to distinguish fake synthetic content from genuine text used as input to train the model (Baele, Naserian & Katz 2022).

Two situations were set up. In the first situation (Task 1), the experts had to distinguish ISIS/Incel content from non-ISIS/Incel content, and did not know that some of this content was AI-generated. In the second situation (Task 2), experts still had to distinguish ISIS/Incel content from non-ISIS/incel content, but were made aware that some of the texts they faced was generated by a language model.

> **...no less than 87% of evaluations of fake ISIS paragraphs were wrongly attributed to ISIS.**

The results, in both tasks, clearly point to the great confusion induced by the fake texts. In task 1, for example, no less than 87% of evaluations of fake ISIS paragraphs were wrongly attributed to ISIS – this is, strikingly, 1% higher than for genuine ISIS paragraphs correctly attributed to ISIS. In Task 2, experts were only slightly better than random guessers, and with low levels of expressed confidence in their answers. These results are worrying, and echo findings from one of the authors' complementary study on audio deepfakes, which demonstrate that open-source models are able to perfectly 'clone' a voice – that is, to create fake statements that are undistinguishable to the listener from the original ones – with less than a thousand 5-seconds genuine audio chunks of that voice.

These developments lead us to infer five main thinking points for stakeholders involved in CVE:

1. Because the threat of extremists using language models is evident, CVE practitioners should familiarise with the technology and develop their own capabilities in language modelling. Among other tasks likely to become central are the detection of synthetic text and the conception of tactics to reduce the growing flow of extremist content online.
2. Yet despite their sophistication, off-the-shelf models cannot be directly used, off-the-shelf, to mass-produce, 'in one click', truly convincing extremist prose. Extremists use highly specific language (lingo, repertoires, linguistic practices, etc.) that corresponds to the particular ideological and cultural niche they occupy, so to be convincing a synthetic text ought to reproduce this specific language with high accuracy, or else it will quickly be spotted as fake. This requires the fine-tuning of a powerful foundational model, which is currently not without difficulties – but will soon become easy.
3. Even if the technology is available to them, some groups are less likely to use it. Groups that place a higher emphasis on producing 'quality' ideological and theological content may be reluctant to hand over this important job to a mindless machine, either out of self-respect and genuine concern for ideological/theological purity, or more instrumentally because of the risk of being outed. However, even these groups may make use of the technology when facing material constraints (dwindling human resources, loss of funding, etc.) or engaging in some propaganda tasks deemed less important (quantity vs. quality).
4. The threat of language models is not uniformly distributed: they are likely to be used for particular tasks within a broader propaganda effort. Consider a web of different online platforms and social media established by an extremist group: while the central, official website would only display small amounts of human-produced content, an 'unofficial' Telegram channel linked on that website could be exclusively populated, at low cost, by large amounts of synthetic text.
5. The workflow structure can be used against extremist actors. For example, stakeholders willing to troll extremist online spaces in order to make them less likely to be visited may use adequately fine-tuned language models to do so more efficiently, more credibly, and at reduced cost. Alternatively, language models can be trained to generate de-radicalising content that could be disseminated by bots.

Stephane J. Baele is Associate Professor of Security and Political Violence at the University of Exeter's Politics Department. His work on extremists' communications and IR theories, which spans across the social sciences, can be found in Political Psychology, the Journal of Conflict Resolution, Terrorism & Political Violence, *or* the Journal of Language & Social Psychology, *among others.* ***This article is a reprint of the CREST guide published in October 2022.***