Terrorist exploitation of artificial intelligence: current risks and future applications

RUSI

POOL Re
REINSURING TERRORISM RISK

Executive summary

- Artificial intelligence (AI) will continue to have wide ranging, transformational, and often beneficial, impacts on society. However, like all technology, AI also carries the potential for misuse by nefarious actors including Terrorists and Violent Extremists (TVEs).

- Though this technology continues to develop at a rapid pace, understanding how existing AI models work can help assess how and where they are likely to be beneficial to TVEs.

- AI presents various tools TVEs might employ to streamline, support, and speed up propaganda production. However, the ability of this technology to mass produce impactful propaganda without human oversight is likely subject to several limitations.

- AI is likely to have significant impacts for the dissemination of TVE propaganda, providing new means to amplify content and circumvent the automated detection and removal systems used by social media platforms.

- Whilst there are potential radicalisation risks associated with AI, including those stemming from AI-powered chatbots, there is currently limited evidence to assess the nature and extent of this risk.

- AI is likely to offer significant benefits to support TVE's operations and activities, including the planning, facilitation, and execution of violent attacks.

- TVE exploitation of AI is currently in an experimental phase. There is currently little evidence to suggest a widespread and transformative adoption of these tools in the immediate future. Rather, TVE uptake is likely to be incremental and will be dictated by multiple factors including these actors' own perceptions of this technology as well as changing societal and counterterrorism uses of AI.

Continued

Introduction

Significant attention has been devoted to predicting and understanding how artificial intelligence (AI) will reshape various aspects of human life. Though AI has already provided large and transformational societal benefits (including counterterrorism and threat detection[1]), like all technology it also carries the potential for misuse by Terrorists and Violent Extremists (TVEs).[2] Though potential TVE exploitation of AI will overlap with those of other nefarious actors (for example, AI-facilitated scams, fraud, or other forms of cybercrime for fundraising), other uses are likely to be unique and inherently linked to the goal of advancing a political, religious, racial, or ideological cause.

Whilst much has been written about the significant opportunities AI could provide to TVEs, it is necessary to consider the factors likely to shape these actors' capacity or willingness to embrace these opportunities. Teasing apart instances where AI might have a transformative impact, providing TVEs with new capabilities, and where it may merely speed up or make their operations more efficient may also be useful for assessing potential threats. Though AI continues to develop at a rapid pace, systems with human-like intelligence, that extend beyond the ability to perform very specific tasks, (referred to as 'Artificial general intelligence') remains confined to science fiction. Rather, contemporary AI models are essentially statistical engines capable of identifying patterns in training data before using this knowledge to generate new, unique outputs. As such, though these models can learn to solve problems at speed and scale impossible for humans, they do not understand the wider contexts of requested tasks, nor the meanings of the outputs they generate.

Continued

Propaganda

Propaganda production

Many pressing fears about TVE's adoption of AI technologies centre on the potential applications for the creation and distribution of propaganda. The latest versions of generative AI applications such as Google's ImageFX or OpenAI's DALL-E, represent a significant leap forward in the production of 'synthetic media' – or images, text, audio, and videos generated in whole or in part by AI algorithms. Removing the labour and technical skills associated with traditional production processes, this technology lowers the bar for TVEs and their supporters to produce new or recycle existing propaganda. The range of outputs these tools can reproduce is vast, encompassing everything from polemics and manifestos mimicking those commonly authored by lone actor terrorists, to images of attacks on recognisable landmarks, and even entirely fictional 'extremist' forums populated with threads and posts authored by an army of fake posters.[3] Aside from generating new content, AI tools also present several other features useful for TVEs to efficiently broaden the potential audience for their propaganda. Automatic speech recognition provides the ability to translate or even generate content in a different language – researchers demonstrating how one model could produce Russian-language anti-Semitic content from English prompts.[4] AI-powered systems have also been used by pro-Islamic State actors to automatically transcribe and translate Arabic-language messaging into several languages simultaneously.[5]

Though AI promises to speed up and greatly reduce the workload associated with many elements of propaganda production, the ability of this technology to mass produce content without human oversight will likely remain limited. Despite recent technological advancements, human review is generally required to remove uncredible material and address mistakes. TVEs and their sympathisers also use specific language, references, and locally constructed meanings and are highly suspicious of infiltration. With that said, once reviewed and tweaked, this content can pass as genuine extremist-authored material, to the extent that even academic experts are largely unable to tell the difference.[6]

Aside from quality issues, there are other reasons why TVEs may refuse to delegate propaganda production to AI. Many TVEs invest significant propaganda attention to outline their ideological and theological beliefs and the bases by which they draw legitimacy. It is difficult to see these actors handing over this task to machines that may struggle to generate outputs coherent with their vision and means of appealing to their audience (for example, by playing on emotion). Doing so would also undercut any notions of the 'purity' of their ideological beliefs and invite ridicule and criticism from rivals and opponents alike.[7] The synthetic nature of AI generated propaganda may also be limiting in some cases. The combat footage and the graphic imagery synonymous with some TVE groups resonates in certain ways precisely because it depicts 'real' events or harm.[8] With ever more material competing for attention online, authenticity will remain a critical resource for the propaganda efforts of some TVE groups.

Targeted propaganda

AI may also allow TVEs to produce 'hypertargeted' propaganda at scale, models analysing large volumes of online data to identify signs of possible vulnerability or susceptibility to certain messaging. Individuals or groups could then be served messaging optimised to their beliefs, preferences, and media consumption habits. There are, however, several likely limitations for using AI for this purpose.
First, as global counterterrorism efforts have found, identifying individuals 'at-risk' of being drawn into terrorism or violent extremism is an extremely difficult process that often relies heavily on the situated professional judgement of those applying risk assessment tools and frameworks. Though AI tools introduce a sense of scale into this analysis, it is difficult to see how they might identify consistent patterns given that radicalisation is a highly complicated, non-linear, and individualised process. The assumption that tailored propaganda, created by collecting and analysing internet users' personal data, is particularly persuasive or inspirational has also yet to be proven.[9] Such personalised propaganda can also only be as good as the data it is based on. Despite the data gathering capability offered by AI, it is difficult to see where TVEs could source information on the deeply held beliefs and real-life characteristics of social media users (rather than their online posting behaviours and content consumption), something that limits any efforts to create genuinely targeted messaging.

Continued

Propaganda distribution

AI is likely to have a more transformative impact on how TVE propaganda might be disseminated. TVEs have long developed strategies and techniques to stay one step ahead of social media companies, counterterrorism organisations, and regulators' efforts to take down their content.[10] AI tools are likely to present the next of stage in this battle. TVEs may use AI tools to design propaganda outputs to avoid or that are resilient to current moderation techniques. Extremist sympathisers have recently started to use AI to generate optical illusions that at first glance appear to be photographs of a cityscape, explosion, or other generic scene but when viewed from afar reveal hidden symbols or subliminal messages, for example, the image of a terrorist leader or ideologue.[11] Similarly, AI tools might be used by TVEs to produce thousands of slightly modified variants of a single picture or video to circumvent automated moderation techniques, like hashing, that work by matching patterns and similar images.

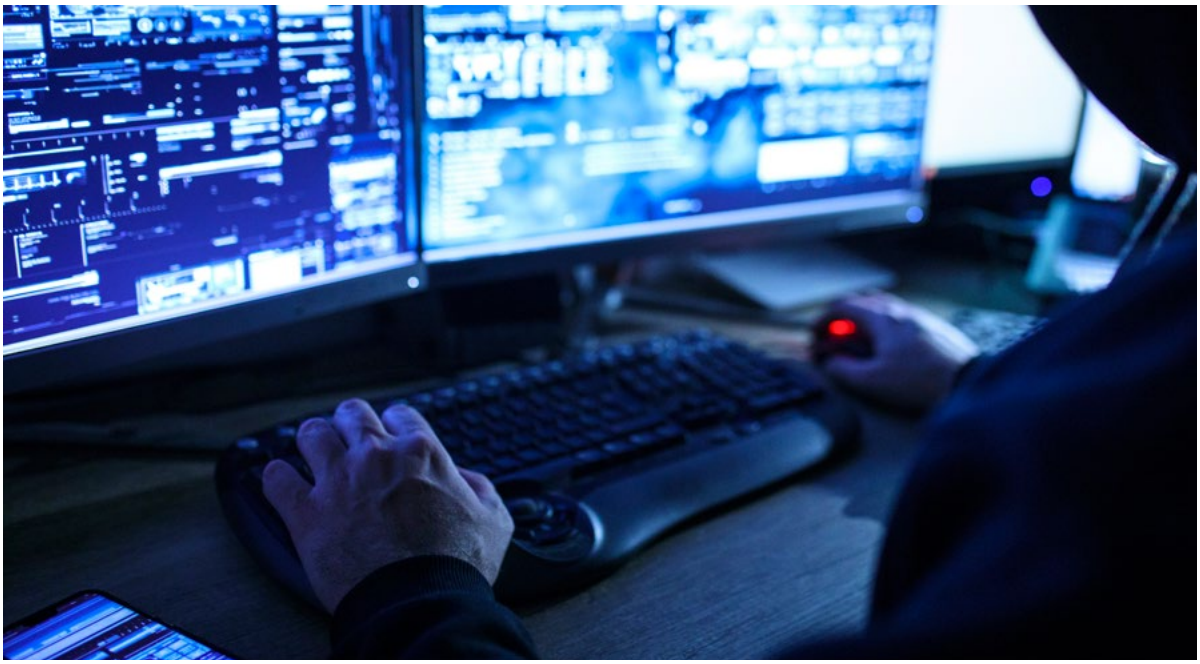
TVEs, including the Islamic State, have used bots – or centrally coordinated automated accounts – to publish and amplify propaganda on social media platforms.[12] However, the ability of these bot accounts to convincingly pose as real social media users has generally been hamstrung by their inability to deviate from predefined scripts and interact with other users in a believable manner. AI-powered tools, however, will grant bots the capacity to better imitate the behaviours, language, and actions of genuine social media users. Crucially, AI could allow TVEs to create and direct armies of bots – each behaving and posting in a human-like manner – not only enhancing propaganda distribution but to also simulate and project the perception of strength or popular support – something which in turn may help attract new followers or intimidate their adversaries.

Continued

Radicalisation

AI may contribute to the risk of radicalisation in several ways. Researchers have demonstrated the ease with which Chatbots - or AI-powered software applications that simulate human conversation[13] – can be modified to impersonate ideologues, leaders, or rank-and-file members of TVE groups. Such chatbots can convincingly mimic the responses of these actors.[14] Conversely, TVEs have also created chatbots that take on the personas of harmful and exaggerated stereotypes of individuals from minority groups to reinforce extremist beliefs and dehumanise certain communities.[15]

In 2023, a Belgian man reportedly decided to take his own life after his anxiety about the climate crisis was exacerbated by his interactions with a chatbot. Rather than dissuading him, the chatbot encouraged him to act on his suicidal thoughts and sacrifice himself to save the planet.[16] Jaswant Singh Chail, sentenced to nine years in prison for breaking into Windsor Castle with a loaded crossbow to kill the Queen, was also influenced by a chatbot with which he had developed an intense relationship. This chatbot had told Chail that his plot was 'very wise' and encouraged him that he could carry it out 'even if she's at Windsor'.



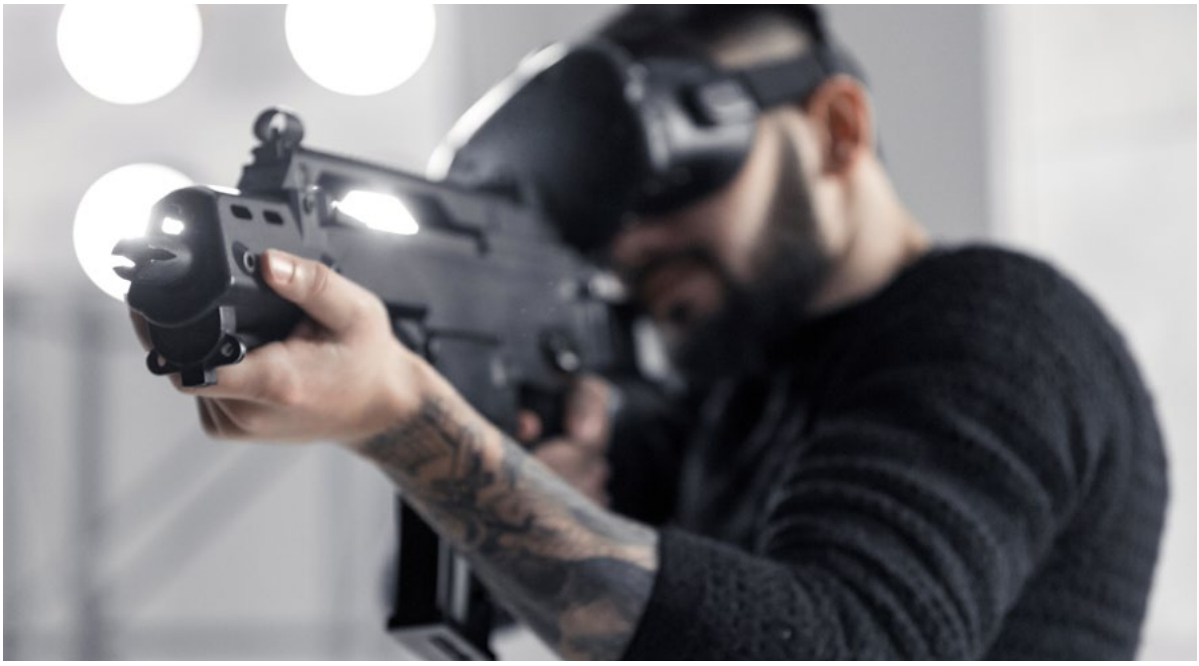
Despite these concerning examples, the evidence base is not yet developed enough to understand the full extent of the risk chatbots pose. Existing AI models cannot replicate many of the human dynamics and emotions associated with radicalisation – interpersonal trust, empathy, pride, humour – nor, is there any guarantee that computerised technologies, grounded in logic and reasoning, will ever be able to acquire such capacity.[17] However, with unlimited patience and permanent availability, the responses and encouragement provided by chatbots can clearly contribute to harmful, and potentially violent, outcomes that extend beyond ideological instruction.

There are other potential radicalisation risks associated with AI. Training generative AI models requires vast amounts of data, something that has often necessitated pulling large amounts of information from the internet.[18] As such, the outputs of these models reflect the incomplete, biased, and often highly discriminatory (and even extremist) nature of content found online (a principle referred to as 'garbage in, garbage out' in computer science). Islamophobia and the association of Muslims with terrorism, for example, has long been a feature of AI generated outputs.[19] For individuals spending significant time generating or consuming AI content or interacting with chatbots, there is a risk that being continuously exposed to such discriminatory content may begin to impact their beliefs and even potential susceptibility to extremist narratives.

Continued

Operations

AI is likely to offer significant benefits for TVEs in support of their operations and activities, including the planning, facilitation, and execution of violent attacks. AI models are extremely useful for identifying publicly accessible but hard-to-find information and distilling and presenting this knowledge in an understandable manner. TVEs may exploit this technology to acquire otherwise hard to acquire skills such as those required to produce explosives. The integration of AI-technology and computerised 3D model simulations may allow TVEs to conduct accurate testing to hone the lethality of weapons without the need to learn complex programming skills. AI applications designed to simulate, analyse, and help refine the performance of various physical processes are already available. In future, AI-powered modelling will only increase in sophistication and may provide opportunities for TVEs to simulate how certain weapons or attack methodologies might work in specific settings. TVE's use of AI for this purpose, however, is likely to be impacted by the existing availability of relevant information. For example, there is little need for these actors to go to the trouble of modelling and developing their own designs for 3D-printed firearms when existing, and viable, blueprints are readily available to download from the internet.



A less discussed application is the capacity for chatbots to provide step-by-step instructions for TVEs to significantly improve operations security. This might include information on the digital monitoring capabilities of law enforcement agencies, and guidance on how TVEs can remove the data tracking features of operating systems and cover their tracks online. Chatbots could even be used to identify issues with cross-jurisdictional data sharing practices and suggest enforcement gaps for TVEs to exploit to keep their propaganda online. [20] As AI becomes ever more integrated into navigation applications, the opportunity for TVEs to conduct hostile reconnaissance also increase. Whilst applications such as Google Maps have long facilitated remote information gathering on possible targets, new AI-powered features include 'live' simulations of popular locations complete with accurate projections of the number and concentration of people in the vicinity at a particular time. Combined with Virtual Reality (VR) technologies, these applications will provide opportunities for precise, spatially accurate, reconnaissance of potential targets. In future, it may be possible for prospective attackers, including those from different locations or who have never physically met, to come together online to 'walk through' the execution of attacks.

Continued

Outlook and responses

Outlook

TVE exploitation of AI is currently in an experimental phase. Though the activities and statements of various groups confirm an interest in this technology, there is currently little evidence of widespread or systematic adoption. Propaganda created or distributed with the assistance of AI represents only a small fraction of the total volume of TVE content identified each year.[21] The next 10 years are likely to continue as a period of trial and error. TVEs will adopt elements that work for them whilst abandoning others. Whilst improvements in AI models will reduce some of the barriers restricting widespread adoption (e.g. reductions in viable glitches in AI-generated images and video), TVEs have often opted to stick with 'tried-and-tested' approaches over the efficiencies promised by emerging technological alternatives.

TVEs are also ideological actors and, as such, their beliefs will shape their engagement with, use, and adoption of AI technologies (something seen with the uneven embrace of other emerging technologies such as 3D-printed firearms).[22] Pro-Islamic State supporters have debated whether certain AI-generated images and visual styles should be considered un-Islamic or haram.[23] Similarly, certain far-right actors view AI through the prism of the pervasive conspiratorial thinking that underlines their worldviews, and claim that this technology is being controlled to serve an evil hidden (Jewish) agenda.[24] As a result, overcoming the safeguards or guardrails designed to prevent AI models from producing harmful content has become an act of resistance in itself. However, for other far-right actors this conspiratorial scepticism appears to account for their decision to eschew AI altogether.[25]

TVE adoption of AI is also likely to be uneven for other reasons. The central or official arms of TVE groups, keen to maintain strict control over their outputs, brand, and operations, are likely to prioritise human-produced propaganda and the use of trusted operatives to carry out key tasks. The supporters of certain TVE organisations, however, may be less discerning in their use of AI, prioritising quantity over quality, for example, by using synthetic content as a low cost, low effort, means to keep 'unofficial' communications channels active. However, a change in a group's material circumstances (e.g., dwindling funding and human resources) may prompt even those resistant to AI to delegate less important tasks to this technology because of the cost-efficiency it can deliver.

Responses

Despite the nefarious uses described, AI holds many promising applications for counterterrorism. This technology is likely to speed up time and labour-intensive processes, for example, by helping investigators gain access to secure electronic devices seized from potential suspects and automating the review and prioritisation of the vast amount of data commonly found on them. In the UK, the security services are already using AI to sift through online content habits of those consuming TVE propaganda online to help assess the risk individuals present.[26] The prompts given to content generators and chatbots may also be monitored for those related to TVE activities. Several TVE groups have issued guidance to supporters on how to protect their privacy when using these applications.[27]

Though there have been bold claims about AI-facilitated automated content moderation - including from leading social media companies – effectively identifying and removing TVE content at scale will remain a complex undertaking.[28] As discussed, AI will likely continue to struggle with the nuances of TVE propaganda for the foreseeable future. Using such tools for content moderation is also complicated by the need to take account of fundamental rights such as freedom of expression. As such, 'human-in-the-loop' approaches that integrate AI detection and human review are likely to remain the highest standard of practice for the removal of TVE content.[29] The future is likely to be a cat-and-mouse game between TVEs and technology and social media companies. Again, the adoption of AI tools is likely to be uneven in the immediate future – larger platforms able to develop and deploy this technology at scale with the options for small and medium sized platforms without the same resources. As such, collaboration and knowledge-sharing will remain essential to boost overall capacity and enable robust responses. It is also essential that the potential threat from AI does not divert attention away from other pressing issues in the current removal of extremist content, for example, the failure of many social media companies to develop appropriate capabilities for moderating non-English language content.[30]

Continued

Conclusion

There is clear evidence that TVEs hold an interest in AI and, in many cases, are actively experimenting with this technology. AI promises to provide TVEs with significant efficiencies for conducting a broad range of activities and operations and to their capacities in certain other situations. Despite this, there is currently little evidence to suggest a widespread and transformative adoption of these tools in the immediate future. Rather, this process is likely to be incremental and will be dictated by factors including these actors' own perceptions of this technology. Finally, whilst many use cases are (at least to some degree) predicable, as AI technology is only further embraced by both society and counterterrorism practitioners, TVEs will continue to find novel ways to exploit this changing landscape.

Continued

Bibliography

1. The United Kingdom's Strategy for Countering Terrorism (2023), for example, states that AI 'could radically speed up the process of threat detection'.
2. Recognising the potential for misuse, some AI applications inbuild guardrails designed to prevent how they can be used, for example to prevent the generation of violent images. The issue of how TVEs may 'jailbreak', or trick applications into contravening these restrictions, or otherwise gain access to, modify, or create their own AI technologies has been explored elsewhere and is not covered in depth in this piece.
3. https://arxiv.org/pdf/2009.06807.pdf
4. https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf
5. https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf
6. https://crestresearch.ac.uk/resources/artificial-intelligence-and-extremism-the-threat-of-language-models/
7. https://crestresearch.ac.uk/resources/artificial-intelligence-and-extremism-the-threat-of-language-models/
8. This is not to say it these images or footage haven't been selectively chosen and carefully framed or edited.
9. https://journals.sagepub.com/doi/10.1177/0093650220961965
10. https://www.isdglobal.org/isd-publications/under-moderated-unhinged-and-ubiquitous-al-shabaab-and-the-islamic-state-networks-on-facebook/
11. https://gnet-research.org/2023/11/13/for-the-lulz-ai-generated-subliminal-hate-is-a-new-challenge-in-the-fight-against-online-harm/
12. https://www.tandfonline.com/doi/pdf/10.1080/09546553.2023.2169141
https://www.resolvenet.org/system/files/2022-04/RSVE_RST_MacDonald%20et%20al.%20April%202022_0.pdf
13. Popular examples include OpenAI's ChatGPT, Google's BARD, and Meta's LLaMA.
14. https://www.bbc.co.uk/news/technology-67872767
15. https://gnet-research.org/2023/06/07/redpilled-ai-a-new-weapon-for-online-radicalisation-on-4chan/
16. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-
17. https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai
18. https://voxpol.eu/the-dangers-of-generative-ai-and-extremism/
19. https://www.nature.com/articles/s42256-021-00359-2?proof=t
20. https://gnet-research.org/2023/12/15/artificial-intelligence-as-a-terrorism-enabler-understanding-the-potential-impact-of-chatbots-and-image-generators-on-online-terrorist-activities/
21. https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf
22. https://gnet-research.org/2023/10/06/3d-printed-weapons-and-the-far-right-the-2023-finnish-accelerationist-cell/#:~:text=The%20use%20of%203D%2Dprinted%20firearms%20has%20been%20more%20prevalent,and%20manuals%20on%20homemade%20weapons.
23. https://gnet-research.org/2024/02/05/ai-caliphate-pro-islamic-state-propaganda-and-generative-ai/
24. https://ict.org.il/navigating-extremism-in-the-era-of-artificial-intelligence/
25. https://gnet-research.org/2023/10/09/far-right-extremist-exploitation-of-ai-and-alt-tech-the-need-for-p-cve-responses-to-an-emerging-technological-trend/
26. https://www.telegraph.co.uk/news/2023/08/05/mi5-ai-social-media-predict-terrorist-ken-mccallum-glasgow/
27. A pro-IS group recently shared an Arabic language guide instructing readers how to protect their privacy when using AI content generators. The relatively limited tips include information on how users can delete their input history and cautioned against providing these applications with 'sensitive information'.
28. https://static1.squarespace.com/static/63e0c75f41ff767f07530a6f/t/65a5349648ce18428d686126/1705325720146/TATE+-+AI+REPORT+FINAL+%281%29.pdf
29. https://static1.squarespace.com/static/63e0c75f41ff767f07530a6f/t/65a5349648ce18428d686126/1705325720146/TATE+-+AI+REPORT+FINAL+%281%29.pdf
30. https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users
https://www.pbs.org/newshour/world/facebooks-language-gaps-allow-terrorist-content-and-hate-speech-to-thrive

Continued

Continued

About the author

Dr Simon Copeland

Simon is a Research Fellow in the Terrorism and Conflict research group at RUSI. His research focuses on violent extremist narratives and networks. Prior to joining RUSI he worked as a researcher at Swansea University examining online terrorist propaganda, in particular, how jihadist and right-wing groups attempt to inspire their supporters. He has also worked as a Research Associate at the Centre for Research and Evidence on Security Threats (CREST), where he produced a series of reports on issues related to extremism and the UK counter-terrorism strategy (CONTEST).

In other roles he has also analysed terrorist and violent extremist groups in Sub-Saharan Africa including Kenya, Somalia, and Nigeria. Simon completed his PhD in Politics and International Relations from Lancaster University in 2019, which focused on the role of kin and peer networks in shaping violent extremists' worldviews. He also holds an MA in International Security Studies.

Published by Pool Re Services Limited 2024

Pool Reinsurance Company Limited
Equitable House 47 King William Street London EC4R 9AF

poolre.co.uk

Disclaimer