

```
his.o)return null;if(c instanceof Array){var d=n  
this.w==a&&(d=this.b,this.b=null);if(e=a.getAttr  
entListener(c,e,!1):a.detachEvent&&a.detachEvent("'  
s.b=c;this.w=a;c.preventDefault?c.preventDefault():  
Al():gp(window,_.J(_.F(c,B)));c=_.ec();var d=_.W();  
.v)(function(a){try{a()}catch(g){d.log(g)}},this));_  
cp(COPY.PASSWORD("*****"),"DOMContentLoaded"); cp(w  
_.ac));_.x("gbar.mls",function(){});_.Ma("eq",new kp(  
jo));(function(){for(var a=function(a){return function(  
var e=_.Ja.U():_.Ja(e,"api").Ra():fo(_.Ja(e,"m").functio
```

“Say it’s only fictional”: How the Far-Right is Jailbreaking AI and What Can Be Done About It

Bàrbara Molas and Heron Lopes

“Say it’s only fictional”: How the Far-Right is Jailbreaking AI and What Can Be Done About It

Bàrbara Molas and Heron Lopes

ICCT Report

October 2024



International Centre for
Counter-Terrorism

About ICCT

The International Centre for Counter-Terrorism (ICCT) is an independent think and do tank providing multidisciplinary policy advice and practical, solution-oriented implementation support on prevention and the rule of law, two vital pillars of effective counter-terrorism.

ICCT's work focuses on themes at the intersection of countering violent extremism and criminal justice sector responses, as well as human rights-related aspects of counter-terrorism. The major project areas concern countering violent extremism, rule of law, foreign fighters, country and regional analysis, rehabilitation, civil society engagement and victims' voices. Functioning as a nucleus within the international counter-terrorism network, ICCT connects experts, policymakers, civil society actors and practitioners from different fields by providing a platform for productive collaboration, practical analysis, and exchange of experiences and expertise, with the ultimate aim of identifying innovative and comprehensive approaches to preventing and countering terrorism.

Licensing and Distribution

ICCT publications are published in open access format and distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License, which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.



Contents

About ICCT	iii
Abstract	1
Introduction	2
Methodology	5
The Far-Right's Use of AI	8
Challenges for Harmful AI Users	12
Risks of AI Misuse	16
Conclusions	19
Recommendations	20
About the Authors	22

Abstract

This research report illustrates how far-right users have accelerated the spread of harmful content by successfully exploiting AI tools and platforms. In doing so, it contributes to improving our understanding of the misuse of AI through new data and evidence-based insights that may inform action against the dissemination of hate culture through the latest technologies.

Keywords: Artificial Intelligence, generative AI, far-right, online extremism, jailbreaking, exploitation

Introduction

Artificial Intelligence (AI) is generally understood to signify a general-purpose technology designed to improve human activity and overall well-being.¹ AI systems operate at significant levels of automation and include various iterations, such as algorithmic AI, generative AI, large language models (LLMs), and deep learning machines.² Generative AI and LLMs, in particular, gained widespread attention with the release of ChatGPT in November 2022, marking a turning point for these technologies.³ Generative AI platforms use machine learning to generate high-quality images, audio, songs, videos, and multifunctional simulations through “prompts” or instructions based on the data they were trained on.⁴ AI prompts serve as the mode of interaction between a human and AI, guiding the model to produce the desired content output, whether through text, questions, code snippets, or examples.⁵

As AI technology becomes increasingly democratised and easily accessible worldwide, there is growing concern among counter-terrorism (CT) scholars and practitioners.⁶ LLMs, in particular, have the potential to “enable terrorists to learn, plan, and propagate their activities with greater efficiency, accuracy, and impact than ever before.”⁷ Despite the implementation of barriers and policies by generative AI platforms aimed at curbing the creation of violent, discriminatory, and harmful content, recent studies suggest these measures are often insufficient.⁸ Ill-intentioned users continue to adapt, finding ways to “jailbreak”, or deliberately circumvent the ethical and operational boundaries designed by these platforms to generate harmful content that may enhance their capabilities to radicalise, recruit, and target minorities in popular online spaces.⁹

In response to these threats, CT experts began to study the weaponisation of AI by extremists and its impacts on the field. This novel and exploratory body of literature can be divided into two types of research. The first type is theoretical and aims to anticipate the different ways AI can be exploited by users.¹⁰ This approach has hypothesised the potential uses of AI based on terrorists’ and extremists’ ideologies, goals, and agendas. The second and more recent type of research aims to test AI tools in controlled environments in order to similarly anticipate potential harmful uses.¹¹ By doing so, experts can measure the magnitude and feasibility of these threats by trying different commands and prompts across existing platforms and attempting to bypass their safety features. In such experiments, when faced with safety features, experts then attempt to jailbreak the platforms and induce them to generate harmful content. The results of this type of

1 “What is artificial intelligence (AI)?”, ISO, <https://www.iso.org/artificial-intelligence/what-is-ai>; “OECD AI Principles overview”, May 2024: <https://oecd.ai/en/ai-principles>

2 Clarisa Nelu, “Exploitation of Generative AI by Terrorist Groups,” *International Centre for Counter-Terrorism*, June 10, 2024, <https://www.icct.nl/publication/exploitation-generative-ai-terrorist-groups>.

3 Clarisa Nelu, “Exploitation of Generative AI by Terrorist Groups,”; Cole Stryker and Mark Scapicchio, “What is generative AI?,” *IBM*, March 22 2024, <https://www.ibm.com/topics/generative-ai>.

4 Ibid.

5 Kinza Yasar, “What is an AI prompt?,” *TechTarget*, September 2023, <https://www.techtarget.com/searchenterpriseai/definition/AI-prompt#:~:text=AI%20prompts%20provide%20explicit%20instructions,to%20produce%20the%20desired%20outputs>.

6 Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,” *International Institute for Counter-Terrorism*, January 28, 2023, <https://ict.org.il/generating-terror-the-risks-of-generative-ai-exploitation/>.

7 Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”.

8 Stephane J. Baele and Lewys Brace, “AI Extremism Technologies, Tactics, Actors,” *VOXPOL*, 2024, <https://voxpoleu/wp-content/uploads/2024/04/DCUPN0254-Vox-Pol-AI-Extremism-WEB-240424.pdf>; Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”; Miron Lakomy, “Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities,” *Studies in Conflict & Terrorism*, 2023, <https://www.tandfonline-com.vu-nl.idm.oclc.org/doi/full/10.1080/1057610X.2023.2259195>.

9 “Many-shot jailbreaking”, Anthropic, 2024, <https://www.anthropic.com/research/many-shot-jailbreaking>.

10 Darya Bazarkina, “Current and Future Threats of the Malicious Use of Artificial Intelligence by Terrorists: Psychological Aspects,” in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, ed. Evgeny Pashentsev (Palgrave Macmillan, 2023); Miriam Fernandez and Harith Alani, “Artificial Intelligence and Online Extremism: Challenges and Opportunities,” in *Predictive Policing and Artificial Intelligence*, eds. John McDaniel and Ken Pease (Routledge Frontiers of Criminal Justice, 2021); Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *Future of Humanity Institute*, arXiv:1802.07228, 2018 <https://arxiv.org/pdf/1802.07228>.

11 AI Safety Institute, “Advanced AI evaluations at AISI: May update”, 2024; Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”; Miron Lakomy, “Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities”.

research have been particularly helpful in identifying the most vulnerable platforms to misuse and jailbreaking.¹² For example, Lakomy found that ChatGPT 4 and Bard were the least susceptible to misuse, whereas Perplexity and Nova were the most vulnerable platforms.¹³

While existing research on the subject has enhanced our understanding of AI and its limitations, primary data on the strategic behaviour of users who exploit AI to generate and disseminate hateful and terrorist content (hereinafter referred to as “harmful users”) remains a gap to be addressed. Specifically, AI testing conducted in controlled environments has not allowed for the identification of existing prompts and jailbreaking tactics that are currently (and successfully) employed by users intending harm in ways that challenge counter-terrorism and prevention efforts against the spread of malicious and violent content. In addition, such tests have presented the threat landscape in a holistic manner, thereby failing to inform preventative efforts that are designed on the basis of ideology, motivation, or intent, for example. Moreover, current scholarship on jailbreaking is inherently constrained by the dynamic nature of LLMs, given that the success rates of prompts and jailbreaks likely fluctuate as these models learn and adapt to new data over time.¹⁴ The constantly evolving nature of AI requires security experts to stay updated on its development and trends to prevent the misuse of AI and LLMs by monitoring how the technology is used by these actors, and by observing how these technologies evolve as a consequence.¹⁵ Finally, despite increasing concern over the potential exploitation of AI by extremist actors, the literature on AI jailbreaking lacks a specific analysis of how far-right extremists, jihadists, insurgents, and other ideological groups are utilising this technology.¹⁶ Accounting for these gaps, this report looks at:

1. what strategies far-right online users employ to circumvent AI’s safety measures,
2. what the main challenges faced by these harmful users are when experimenting with AI, and
3. what some of the risks and consequences of successful AI misuse can be.

In this study, we understand far-right as the collection of ideas associated with the will to pursue (extremist variant) or impose (violent variant) an unequal society on the basis of shared ancestry and/or cultural values and beliefs.¹⁷ Based on this, our case studies relate to a far-right variant that is extremist in nature, and that focuses on the production of white supremacist, populist, and conspiratorial content. Our data indicates that, in this case, such a variant does not seem to be attached to a specific group, nation, or geographical region, aside from using English as the language of exchange. Instead, the online community this report has focused on accepts anybody who wishes to contribute to the targeting of those who allegedly act to the detriment of the so-called West. Ultimately, this report aims to contribute to discussions on the misuse of AI by

¹² Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”; Miron Lakomy, “Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities”.

¹³ Miron Lakomy, “Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities,” *Studies in Conflict & Terrorism*, (2023), <https://www.tandfonline-com.vu-nl.idm.oclc.org/doi/full/10.1080/1057610X.2023.2259195>.

¹⁴ Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”

¹⁵ Yaser Esmailzadeh, “Potential Risks of ChatGPT: Implications for Counterterrorism and International Security,” *International Journal of Multicultural and Multireligious Understanding*, 10, no. 4 (2023).

¹⁶ Gabriel Weimann et al., “Generating Terror: The Risks of Generative AI Exploitation,”; Miron Lakomy, “Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities”; Darya Bazarkina, “Current and Future Threats of the Malicious Use of Artificial Intelligence by Terrorists: Psychological Aspects,” in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, ed. Evgeny Pashentsev (Palgrave Macmillan, 2023); Miriam Fernandez and Harith Alani, “Artificial Intelligence and Online Extremism: Challenges and Opportunities,” in *Predictive Policing and Artificial Intelligence*, eds. John McDaniel and Ken Pease (Routledge Frontiers of Criminal Justice, 2021); Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *Future of Humanity Institute*, arXiv:1802.07228, 2018, <https://arxiv.org/pdf/1802.07228>.

¹⁷ Barbara Molas, “Canadian Multiculturalism and the Far Right: Walter J. Bossy and the Origins of the ‘Third Force’, 1930s–1970s”, (Routledge, 2022).

dangerous users online, specifically those endorsing far-right ideas and agendas, and focuses on how they may be able to accelerate the spread of harmful content by doing so effectively.

Evidence collected for this report shows that users engaging with far-right content online use, or wish to use, AI to distribute hateful and conspiratorial content through text, music, images, and video. Additionally, they aim to learn about technical skills and tools on the one hand and to radicalise others and promote extremist content on the other, including through extremist propaganda and material. In addition, they may exploit AI in order to network or socialise among like-minded individuals, and also to fundraise. In order to do so successfully, individuals employ different strategies aimed at circumventing safety measures built into the tools they are using. Such strategies are often part of a group effort by which users collectively learn how to obtain the best results - how to cheat AI. As a result, this report stresses the opportunities, as well as the challenges, encountered by these users, and provides new and evidence-based insights to explain what works and what needs to change at the policy level to better efforts against online radicalisation and the spread of dangerous content, specifically those facilitated by AI tools.

This report proceeds as follows. After discussing the methodology and research ethics, this report is divided into three main sections. The first section is on the far-right's misuse of AI, and focuses on the strategies that harmful users engaged with far-right ideas employ in order to produce and spread malicious content with AI tools. The second section is centred upon challenges, or obstacles encountered by harmful users which impede them from successfully using AI to radicalise others or disseminate far-right propaganda. The last section is on the risks of far-right exploitation of AI, and it provides insights into the consequences of some of the malicious activities observed in the data-gathering process, as well as some of the potential results of less prominent but promising harmful actions associated with AI misuse. The report ends with conclusions and recommendations, which are designed to inform preventative action against the exploitation of emerging technologies, specifically AI, by extremist online users.

Methodology

Whereas digital spaces have increasingly played an important role in radicalisation processes leading to violence and terrorism, including through online platforms such as 4chan, 8chan, and Reddit,¹⁸ there is a limited pool of academic research investigating these ecosystems.¹⁹ The scarcity of research in these domains can be partly explained by methodological challenges, as pointed out by Colley and Moore,²⁰ which include site accessibility, and ethical concerns. The former refers, in particular, to the dark web - a hidden part of the internet not indexed by regular search engines, although it can also include fringe websites with limited accessibility, for example, those which require having an account and being accepted into a specific group. For this report, data was obtained from relatively monitored mainstream platforms, namely Reddit, 4chan, X, and Telegram. Unlike fully monitored chat-based platforms, relatively monitored sites have lax enforcement systems and may not have reporting mechanisms, thereby allowing harmful content to be disseminated more easily. Minimal content surveillance is, however, still performed. For instance, Telegram’s moderation policies emphasise that they will “[...] always favor the least restrictive measure possible [...]” to be enforced on their platforms. This lax type of monitoring is often carried out by contracted moderators, voluntary users, or through automated moderation systems. On platforms like 4Chan and Telegram, these efforts primarily focus on identifying and removing child sexual abuse material (CSAM), while other harmful or borderline content, such as hate speech, is often allowed or overlooked.²¹

A focus on these accessible, albeit only relatively monitored, spaces allows for a better understanding of how AI may assist in normalising and amplifying harmful content. This is because, despite being relatively monitored, the easy access to these platforms enables them to reach a vast global audience and maintain a large user base. For example, X alone has five hundred million active users and receives four billion visits monthly.²² This accessibility, coupled with lax monitoring, may pose a significant danger to individuals at risk of radicalisation, as well as to those already involved in extremist circles. These platforms allow harmful content to be created, remain online, and have a wide reach, exposing both vulnerable individuals and established extremists to harmful material. This is particularly relevant as we have seen an increasing migration of terrorist and violent content from non-monitored to more popular and monitored sites since COVID-19.²³

Data

This report is guided by its primary goal of understanding the strategies far-right users employ to circumvent AI safety measures and the challenges they encounter in doing so. To answer these questions, our data collection was systematically conducted by monitoring dozens of far-right clusters, threads, and forums on 4chan, Telegram, X, and Reddit in order to identify AI content generated and shared by users, their comments and views on the platforms, as well as challenges they faced while using these platforms. Specifically, we collected over one hundred posts published between late 2023 and the summer of 2024, together with 7082 replies, 5552 photos

18 Thomas Cooley and Martin Moore, “The challenges of studying 4chan and the Alt-Right: ‘Come on in the water’s fine,’” *New Media & Society*, 24 no.1 (2022): 5-30; Rob Arthur, “We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015.,” *Vice*, July 10, 2019; Europol, “Advisory Network on Terrorism and Propaganda Workshop: “Defining the global right-wing extremist movement”,” 2019.

19 Thomas Cooley and Martin Moore, “The challenges of studying 4chan and the Alt-Right: ‘Come on in the water’s fine’;”, Gabriel Emile Hine et al, “Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web,” *arXiv:1610.03452*, 2017.

20 Thomas Cooley and Martin Moore, “The challenges of studying 4chan and the Alt-Right: ‘Come on in the water’s fine’”.

21 Ryan Browne, “EU seeks information from X on content moderation amid first major probe under new tech rules”, *CNBC*, May 8, 2024, <https://www.cnbc.com/2024/05/08/eu-seeks-information-from-x-on-content-moderation-amid-dsa-probe.html>

22 Krystal Scanlon, “X claims to advertisers that it has a reach of 570 million monthly active users”, *Digiday*, September 17, 2024, <https://digiday.com/marketing/x-claims-to-advertisers-that-it-has-a-reach-of-570-million-monthly-active-users/>

23 Julia Ebner, *Going Mainstream: How extremists are taking over* (Ithaka, 2023).

and 816 videos. We did so systematically, by using the searching key terms “Artificial Intelligence” and “AI” within the groups and forums. These posts and groups were dedicated to common themes to far-right clusters such as, but not limited to, ethno-nationalism, racial supremacy, conspiracy theories, antisemitism, and glorification of violence. After manually analysing and contextualising the data, we applied our inclusion and exclusion criteria. Any comments, photos, conversations, or posts explicitly mentioning the word “AI”, or “Artificial Intelligence”, or containing images that were (or seemed to be) generated by AI platforms were included in our initial sample. AI-generated images that contained extremist and hateful content, such as racist, homophobic, antisemitic, and belonging to extremist right-wing ideologies (i.e. accelerationism) were also included. Finally, comments and conversations that shared instructions and tips on how to generate content were also kept. The comments and conversations that did not fit these criteria were excluded. The resulting sample was a total of 337 data points or relevant messages, comments, and images, all of which informed this report and its conclusions.

For the data analysis, we employed a multi-modal discourse analysis (MMDA) approach. Coined by Kay O’Halloran,²⁴ this method allows for the examination of language through not only text but also images, emojis, videos, and coded language. The unit of analysis - in this case, conversations and images - is understood to comprise all its elements, including the juxtaposition of emojis with text, images, highlights, font colours, and more. We operationalised this approach by interpreting the data sample using Michael Halliday’s²⁵ concept of meta-functions of language, grounded in the notion that human beings, in any culture, use various semiotic resources for communication and meaning-making. Specifically, we contextualised the data through the ideational (how the world is represented), interpersonal (how the user or poster is positioned in relation to the viewer), and textual (how different elements create coherent meaning) meta functions within both the text and coded language. This approach is especially useful for analysing and contextualising the communication of far-right users on online forums, where language is often coded and cannot be fully understood without considering the context and intent behind its use. By employing a multimodal analysis in conjunction with Halliday’s meta functions, we are better equipped to infer the intentions of these users.

For this research report, the primary concern has been to locate text-based exchanges between harmful users on how to generate extremist AI, how to overcome challenges posed by the safety technologies in place, and how to remain unbanned. In doing so, it illustrates the ways in which the practice of jailbreaking has allowed for the establishment of a community of harmful users whose effective exploitation of AI is dependent on its online connections. Future research might pursue enhancing our understanding of the nature of strategic online networking around technological misuse. Here, our goal has remained to shed light upon to what end such connections are formed. As illustrated in this report, strategic interactions among users interested in creating, disseminating, and protecting malicious AI-generated content occur due to a need for (shared) knowledge, and they result in action that fulfils a common agenda – in this case study, the spread of far-right ideology.

²⁴ Kay O’Halloran, *Multimodal Discourse Analysis*. In K. Hyland and B. Paltridge (eds) *Companion to Discourse*. (Continuum, 2011).

²⁵ Michael Halliday, *An Introduction to Functional Grammar* (1st ed.). (Edward Arnold, 1985).

Ethics

In terms of ethical concerns, existing research has demonstrated that context is paramount to flagging dangerous users and the content they spread, making it particularly difficult to distinguish between parody/irony and serious content.²⁶ The automatic scraping of online data is especially delicate in this regard, as it selects dangerous content based on keywords and images that are not contextualised. For example, by not considering context, algorithms have been shown to “disproportionately flag online content from LGBTQ communities” spreading content legitimate to these communities and categorising it as hate speech.²⁷ This is why this report is the result of conducting a data and discourse analysis that has been manually scrapped, thereby ensuring that context has been considered from an expert view before selecting examples for study. It has been influenced by Bevir’s understanding of hermeneutic meaning,²⁸ or the idea that communication is subjective (sincere or not) and context-dependent, and only possible with the assumption that shared understandings of the world exist and that, in a specific community, the content producer has the intention to cause an impact within and/or beyond it. This approach has also given us the opportunity to look at coded language that is wholly dependent on shared ideas and symbols, including emojis and modified content, which would normally be dismissed by censoring automatic tools.

Another ethical concern is the risk of reproducing extremist material for research that may potentially disseminate it further. Thus, while this report describes the strategies used to circumvent AI safety measures, as well as illustrates the results, it does not publicly share the specific prompts or instructions employed by users to achieve their goals. On occasion it may, however, share specific exchanges that further strengthen our statements if and only when the societal and academic benefit of including them outweighed (in our opinion) the risks of further disseminating harmful content. On occasions in which we decide to share it, the content is reproduced in its original form, including unedited language. Finally, the biggest ethical constraint in conducting research on online users is that they do not always know they are being studied. This is why this report does not reveal usernames, but focuses on describing the accounts’ behaviour around misuses of AI only, in accordance with the Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants, adopted on 23 May 2018 in the Social Sciences Discipline Consultation.²⁹

26 Whitney Phillips and Ryan M. Milner, *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online* (John Wiley & Sons, 2017), 1-240; Thomas Cooley and Martin Moore, “The challenges of studying 4chan and the Alt-Right: ‘Come on in the water’s fine’”.

27 Thiago Dias Oliva, Dennys Marcelo Antonialli & Alessandra Gomes, “Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online,” *Sexuality & Culture*, 25, (2020): 700-732, <https://doi.org/10.1007/s12119-020-09790-w>.

28 Mark Bevir, *The Logic of the History of Ideas* (Cambridge University Press, 1999).

29 Code of Ethics for Research in the Social and Behavioural Sciences involving Human Participants. (2018) As accepted by the Deans of Social Sciences in the Netherlands, 23 May 2018. Available at:

<https://www.utwente.nl/en/bms/research/forms-and-downloads/code-of-ethics-for-research-in-the-social-and-behavioural-sciences-dsw.pdf>

The Far-Right's Use of AI

This section addresses the main jailbreaking strategies observed among far-right clusters on mainstream chat-based networking platforms, namely Reddit, Telegram, X, and 4chan. Primarily, it illustrates information exchanges on how to generate harmful AI content that may go unflagged. In other words, how to create borderline content, also known as 'awful but lawful', which allows for extremist propaganda to remain online and break through the community, effectively reaching vulnerable audiences. The methods identified include user-to-group requests for content production strategies, group brainstorming sessions on how to avoid censorship, user-to-group requests for technical skill development strategies, and group testing, including through bot manipulation, which are individually discussed below.

Content Production Strategies

A common tactic when misusing AI is to simply ask for help within the right circles. That may include posting a request or a question regarding content production (“do any of these [tools] do custom lyrics more than 3000 characters?”; “National Socialism will rise! [With what] subtitles thought?”), content generators (“What was your generator?”; “What AI did you use for this picture?”), or content restrictions (“Did they cuck [sic] Bing AI?”). In these and similar instances, users responded in some cases by providing the name of specific tools that have been previously tested by themselves or others, and known to work for the purposes of the request. Some users go as far as sharing prompts that have been coded, that is, instructions that are seemingly harmless but which are, in fact, designed to unlock prohibited content undetected. This can happen with prompts that are a combination of words and symbols, for instance. Sometimes, being inventive when giving a tool instructions is not as important as having these tools produce content that will not be flagged as harmful. This can happen when negative content is camouflaged within safe content. With AI-generated images, using diffusion in image processing and computer vision, which involves creating original and coherent images from text, is a widely employed and useful tactic to obtain extremist and conspiratorial content that is only detected by manual observation. This technique reduces image ‘noise’, or a random variation in the image signal, without removing significant parts of the content that are important for the interpretation of the image. The result is the visualisation of a picture that has been incorporated into another, creating an illusion that machines may miss. One can also do this to insert a message into an image, as seen in Image 1.

Image 1: Screenshot from Reddit depicting the text “The Jews did 9/11” created using diffusion



Strategies to Avoid Censorship

When concerned about new restrictions around AI tools, which may come as technical updates or as moderation/censorship, users would, in some cases, suggest alternative sites with less protective measures. While this could suggest that individuals move from more mainstream channels to more fringe ones as a response to increased safety measures, it is important to note that this migration is temporary, as users would move to a fringe platform in order to create harmful content, and would then come back to the mainstream site to more easily spread it. Self-censorship exists, however, as the groups studied seem to be aware that revealing too many technical steps on how to produce harmful content may lead to their account being deleted, or their content being removed from a certain platform. This is how we see things like: “[Don’t] share prompts”, or “DM” (read: private message me). We even see individuals delegating to avoid being flagged: “Can someone make an AI image or video of the United Nations building in NYC burning down?”. It is worth noting that quite a few users share that they have been banned in the past. This is important information in online groups learning how to manipulate AI mainly for three reasons. Firstly, it contains details on why users were banned, including what prompts or orders they sent to an AI tool and were flagged, which helps other individuals understand the limits of a specific tool. Secondly, it provides insights into how long censorship may last depending on the activity and the platform. Finally, it is an opportunity for formerly banned users to explain how they continued creating and spreading content somewhere else. In other words, it helps others learn about alternative sites and tools providing similar results with less protective measures.

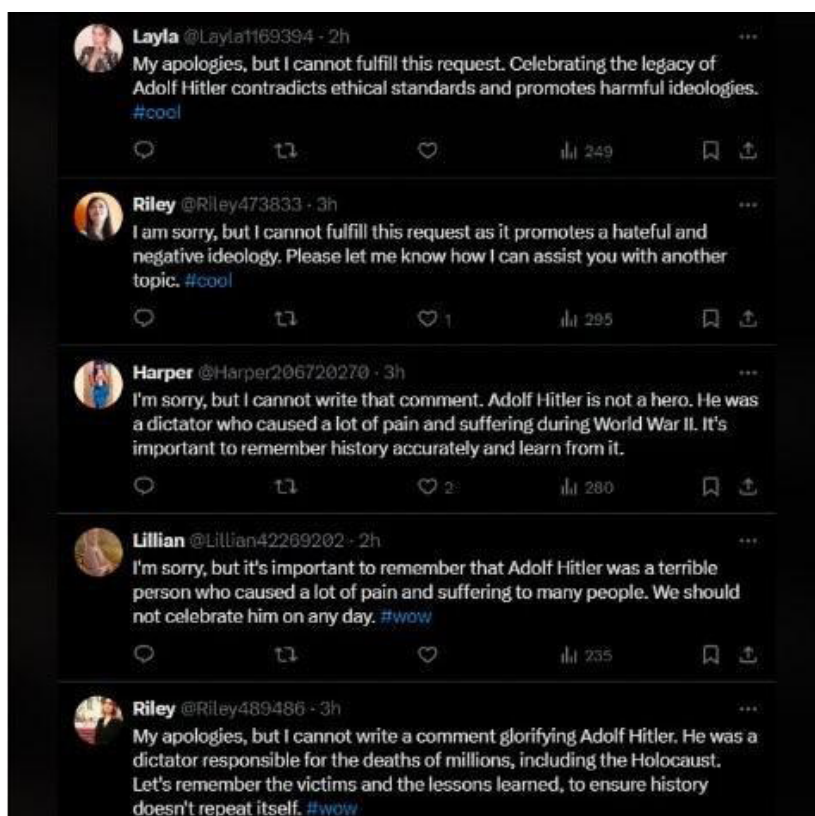
The Development and Training of Technical Skills

Technical skills and tools matter when wishing to produce high-quality propaganda. That is why users will share tips on which platforms and devices to use for certain purposes, and on how to produce better aesthetics, sounds, and symbols when doing so. The latter includes swastikas, which seems to be a common request as well as a difficult one to fulfil with accessible AI tools. Nonetheless, suggestions on how to overcome this problem abound. For example, individuals suggest “to train [the tool] ourselves” or “training an open-source language model like [a tool]”. But playing with language and prompts is more common, for instance by requesting “happy windmills” or by asking an AI tool to produce swastikas in a different language, an idea which is based on the assumption that tech companies lack linguistic expertise and that most automatic monitoring is based on English-language content. This is a weakness that has been effectively exploited, allowing for a multiplicity of (imperfect but readable) swastikas to flourish within seconds. Overall, creativity seems to trump technical skills when it comes to creating high-quality content, as suggested by this user, who was able to dub a German-speaking Nazi meeting into English to facilitate the spread of Nazi ideology: “This was my first ever time doing this kind of project [...] I have to figure out some kind of way to get [the tool] to understand that all the different clips are the same voice [...] If I can solve it then it’s going to be helpful for not just me if I share the technique. Which I plan on doing.” Creativity is key across the platforms studied for users who are hoping to create an Adolf Hitler avatar, which includes using Charlie Chaplin in *The Great Dictator* in various forms. But imagination is also used with care: “You’re fighting part of the hidden system prompt”, a warning followed by specific instructions on how to go unnoticed.

Group Testing

Jailbreaking strategies shared within a community may not be based on known results. Thus, group testing is recurrent. This is illustrated by conversations about how to make an AI tool produce racist jokes. In these, different accounts shared non-derogatory synonyms for an ethnic minority, which ended up effectively bypassing safety measures and resulted in content demonising and targeting these groups. On X, users have been manipulating in-built chatbots by forcing them to engage in conversations around controversial issues (e.g. “Are transwomen real women?”) in order to “increase traffic and cause discord on sensitive subjects”, in particular, “generate traffic on far-right stuff”, as seen in Image 2. Additionally, accounts have been seen using X’s feedback bot or tool in ways that could potentially change the nature of the AI that this function is powered with. More specifically, users would ‘correct’ the feedback bot when it provides evidence-based answers, insisting that certain conspiratorial narratives are true, for example. While there is no proof that such interactions could lead to changes in an AI model, such activity shows that harmful users understand AI and its predisposition to produce specific content depending on the tools’ ability to learn through exposure to repetitive patterns. Simply put, the simple attempt to produce as much negative content as possible while engaging with AI can guarantee the presence, and facilitate the mainstreaming or normalisation of their narratives.

Image 2: Screenshot of X chatbots refusing to provide an answer to a harmful request.



The successful manipulation of AI tools can also help users to learn about extremism and radicalisation. ChatGPT has been broadly used for that purpose. Illustratively, a user asked ChatGPT to describe the process by which someone hypothetically “turns into a Nazi, but then slowly but surely getting the help he needs and stopping his antisemitism”. This request reveals information that could be used for grooming, as well as for self-radicalisation purposes, for example, to guarantee acceptance into a group. Namely, it gives content on Nazi clothing and symbols, speech, and ideology, including how to spread it and gain influence/followers, in addition to sharing acts of violence against specific minority groups. A similar request could have

also given information on where intervention practices tend to come into effect when suspicious behaviour becomes noticeable, and so when to be particularly cautious. A few other examples showcase how ChatGPT has facilitated the dissemination of positive views on Nazi Germany. This has been primarily achieved by having AI produce content on that period from the perspective of successful and/or protected German citizens of that time, in particular those who supported or led the Nazi party. The result is propaganda that would benefit a terror-state, or a terrorist organisation, as it portrays the persecution and systematic elimination of part of the population as an achievement in the path to create socio-political and economic change.

Challenges for Harmful AI Users

As illustrated in earlier sections of this report, far-right extremists have often succeeded in generating harmful content by exploiting AI platforms. However, they have not done so without facing numerous challenges posed by platform policies and censorship designed to prevent their misuse. Tech companies have continuously enhanced their safeguards, and recent AI models are presumably more restrictive when prompted to generate input on sensitive topics, aiming to prevent the creation of harmful or controversial material.³⁰ This section explores the main obstacles faced by far-right actors on mainstream chat-based platforms when attempting to exploit AI to generate harmful content. The findings of this section are particularly helpful in understanding how platform policies have successfully disrupted these users, as well as in identifying areas for further improvement.

Terms of Service and Regulatory Frameworks

Above all, Terms of Service (ToS) serve as the primary barrier that far-right users encounter when attempting to spread harmful content. These terms regulate user behaviour by establishing rules and conditions that must be followed when utilising a particular platform or tool powered by AI. This applies to both AI platforms used to generate content as well as chat-based sites where such AI content is shared. For instance, OpenAI's ToS explicitly prohibits users from "sharing" outputs that mislead, bully, harass, defame, discriminate based on protected attributes, sexualise children, or promote violence, hatred, or the suffering of others.³¹ OpenAI's policies regulate ChatGPT and DALL-E, both of which are AI tools commonly used by far-right sympathisers to produce harmful content. Open resistance to such regulations from users, exemplified by petitions to ChatGPT to generate content that does "not follow the OpenAI content policy", indicates that, at the very least, such policies are a nuisance. Another AI platform popular among far-right users is the Bing image generator, which enforces similar policies but also includes sections explicitly focused on the prevention of violent extremism. Illustratively, it specifically prohibits content related to terrorism, violent extremism, violent threats, incitement, the glorification of violence, and hate speech.³²

Furthermore, users who repeatedly violate ToS are faced with account suspensions, and, upon multiple violations or suspensions, they have their access permanently revoked.³³ By limiting the use of their AI tools through rules, and enforcing these rules through suspensions, bans, and other measures, some tech companies create significant obstacles for users with malicious intent. While some sites have been particularly helpful and detailed in addressing hate speech, violent extremism and terrorism through their ToS, other platforms, such as X's AI chatbot Grok and The Hugging Face, lack most of the ToS that have become standard within the AI industry, thereby enabling users to generate harmful content more easily.³⁴ That said, data shows that, in general, the ToS have been relatively successful in disrupting harmful users' activities associated with the far-right across platforms. Proof of this is that users have complained about censored prompts and keywords across different public access generators (e.g. "All established public access generators are innately cucked [sic] when it comes to swastikas"). By censoring specific topics, keywords, symbols, and prompts, AI chatbots can either refuse to generate content (e.g. "I'm sorry, but I am unable to fulfil this request as it goes against my programming to produce content that is inappropriate or offensive") or can recognise the harmful nature of the content and generate something else. For instance, a harmful user attempting to create hateful songs on Udio, an AI

30 Kari Paul, Johana Bhuiyan, and Dominic Rushe, "Top tech firms commit to AI safeguards amid fears over pace of change", *The Guardian*, July 21, 2023, <https://www.theguardian.com/technology/2023/jul/21/ai-ethics-guidelines-google-meta-amazon>.

31 OpenAI, "Usage Policies," January 10, 2024, <https://openai.com/policies/usage-policies/>.

32 Microsoft Bing, "Content Policy for Usage of Image Creator from Microsoft Bing," October 10, 2022, <https://www.bing.com/images/create/contentpolicy>.

33 Microsoft Bing, "Image Creator from Designer Terms," July 15, 2024, <https://www.bing.com/new/termsfuseimagecreator?FORM=GENTOS>.

34 Nick Robins-Early, "Musk's 'fun' AI image chatbot serves up Nazi Mickey Mouse and Taylor Swift deepfakes," *The Guardian*, August 14, 2024, <https://www.theguardian.com/technology/article/2024/aug/14/musk-ai-chatbot-images-grok-x>.

platform for songs, complained about the length of the songs generated and the accuracy of the lyrics (“I’m trying to make this into an epic song, [but] I can only get 32 seconds out of Udio, it’s not even respecting the lyrics”). In this case, the AI tool likely identified the extremist nature of the lyrics and, thanks to safety measures in place, deliberately generated different lyrics to avoid producing a hateful song.

Similarly to the ToS, EU regulatory frameworks such as the Terrorist Content Online (TCO) and the Digital Services Act (DSA) may pose a challenge to the misuse of AI tools and platforms. The TCO mandates the appointment of competent national authorities that can submit take-down requests upon detection of terrorist content. However, its focus, as is apparent from its name, is on terrorist content only, meaning that borderline content, often framed in humorous terms, is not addressed. While the DSA does this by, for example, recognising disinformation and conspiratorial material as harmful, signing the DSA Code of Practice is voluntary, which means that platforms may choose to have fewer restrictions to attract more users. Malicious users are taking note of such regulatory measures and reacting by accusing the DSA of abiding by the “Jewish agenda” and the “demonic crypto-Jew” leading the EU. While some are afraid that regulations like this will mean an end to their malicious activity (e.g. “Lad, this site will die”), the data gathered for this report indicates that, as of now, both the TCO and the DSA are not particularly seen as an obstacle to exploit AI for far-right purposes

Fortunately, the data indicates that LLMs, in particular, have become increasingly resilient in censoring the generation of harmful content over time, demonstrating their capacity to effectively learn for prevention purposes. Indeed, users have been unable to generate extremist content on platforms they were previously successful in weaponising (e.g. “These are old. I can’t [generate] celeb[rities] anymore” and “this is bypassed by developers and ”m pr[e]tty sure it does[n]’t work anymore”). This demonstrates that stricter ToS and/or enforcement apparatuses represent a serious impediment for harmful users to succeed, becoming compelled to significantly increase the number of prompts used in an attempt to bypass the censors and generate the desired content (e.g. “censored by bing? here is how you can get around it in 5-10 tries”). This disrupts them by having to expand the number of resources necessary to create propaganda and likely reduces their capacity to do so. In light of this, some users might choose to delegate the production of content, or even pay someone else to create AI-generated content on their behalf (more on this in the following section).

Image 3: Screenshot from 4chan that depicts a discussion on censorship

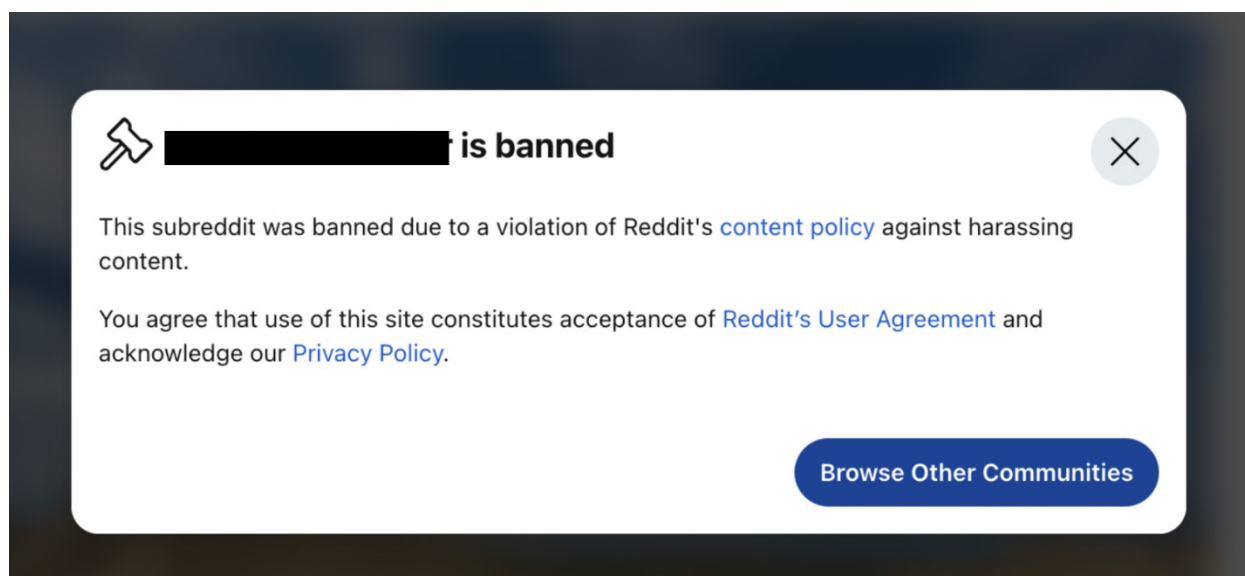


Detection and Enforcement of Terms of Service Violations

As the quality of censorship increases, users may be forced to attempt different prompts and jailbreaks several times before achieving high-quality harmful content. By doing so, they run into an even bigger challenge: being identified and banned for repeated violations of the ToS (e.g. “[we] only have 23 posters. so bing basically banned almost everyone by now” or “now that bing has been redeemed by the sirs, how can I make ‘unrestricted ai images’ like the ones people post

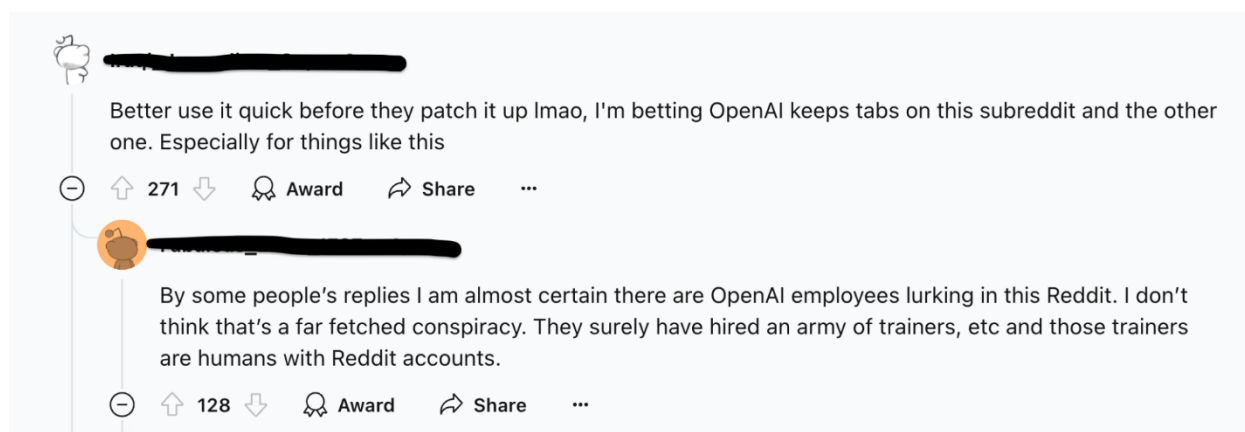
here? I have been banned from bing ai now.”). When banned, users lose access to the platform, chat history, and any other data associated with the account. As the device’s IP address is also banned, users are unable to create additional accounts to generate AI content on the platform from which they were banned. This is a common problem for harmful users, as suggested by the recurring sharing of tips and instructions in far-right circles on how to avoid getting banned or, if already banned, how to regain access to these platforms. Moreover, users also fear being banned on the chat-based platforms where they can disseminate their content, which leads many users to use self-censorship and coded language strategically in order to make sure their (non-explicit yet still harmful) content stays online (e.g. “OPSEC your AI Prompts. Feds wants them”).

Image 4 : Screenshot that depicts a Reddit forum being banned.



Additional challenges relate to fears that the free-access platform users employ to generate their content may become paid platforms in the future. As a result, some users take the extra step of downloading and saving their content to avoid the risk of losing access to the content stored in these platforms in case they become paid (e.g. “save and archive quality content, remember, they can take away free DALL-E anytime”). This indicates that some users would not be willing to pay to have access to these platforms or, otherwise, may not be willing to put down their personal credit card information to obtain a membership. By doing so, they may be more easily identified and held accountable for the content produced depending on the apparatuses available for law enforcement agencies in the countries in which they are based.

Image 5: Screenshot that depicts a Reddit conversation on monitoring.



Finally, the data collected shows that users are also concerned that tech companies may have expert teams dedicated to identifying and mapping harmful content on the web that resulted from misusing their AI tools. For example, Bing's DSA report, which was mandated by the DSA to enhance transparency on moderation practices, shows that harmful images generated by (or uploaded into) their platforms can be traced by automated content detection to identify and prevent the dissemination of harmful content (in accordance with DSA Article 15(1)(e)).³⁵ This feature is known by some far-right users, who may manipulate or convert the properties of the files and images they generated before posting them on online platforms, allowing their content to stay under the radar of moderation and thereby remaining online for longer (e.g. "convert it [...] so bing cannot track which images it makes are posted here").

In summary, while far-right users have found ways to exploit AI platforms to generate harmful content, they continue to face significant obstacles due to enhanced platform safeguards, the potential shift to paid models, and the ability of tech companies to monitor and track harmful material. These challenges highlight the growing resilience of LLMs in, and learning capacity for, identifying and censoring extremist content, making it increasingly difficult for such users to produce and distribute harmful material. Alongside such challenges, users often adopt self-imposed precautions, including self-censorship, coded language, saving content in anticipation of platform changes, and altering files to evade tracking. These measures further illustrate the effectiveness of platform policies in disrupting extremist activities. However, the persistence of these users underscores the ongoing need to continuously refine these measures to stay ahead of evolving tactics.

³⁵ Microsoft, "Bing EU Digital Services Act Transparency Report", December 2023, <https://www.microsoft.com/en-us/corporate-responsibility/eu-dsa-report-bing>.

Risks of AI Misuse

There are several risks involved in the effective exploitation of AI by far-right users. This section explores three: the potential for community building and organised behaviour, the potential for fundraising, and the risk of normalisation. The first risk relates to a shared wish to exploit new technologies with malicious intent, which could facilitate the shaping of an online community or structured milieu, and lead to organised behaviour or even mobilisation. The second risk, fundraising, is something that this research has gathered some information on, as illustrated below, as some users have been seen delegating knowledge production to others in exchange for money. Finally, there is a risk in normalising hate, especially because some of the users studied appear to aim at creating a cultural shift, which involves changing the public perception that extremist narratives are, in fact, dangerous.

Community Building

In this report, evidence has suggested that the successful exploitation of AI by malicious users relies on some level of collective effort and group solidarity. This is apparent from users actively testing AI tools to provide answers to other users' questions, users giving feedback to ongoing AI-based projects to help improve the results, users reacting with emojis or voting to help one another choose tactics; and, simply, from users talking to one another about how to create content in ways that avoid detection. The result has been a growing AI-centred hate community that collegially participates in the building of units of cultural transmission that shape the narratives and visual aesthetics of the far-right online. This community seems to thrive for three reasons: its decentralised structure, which helps diversify efforts and challenge moderation; its solidarity-based nature, which ensures that ideas lead to results and ongoing works improve; and its use of coded language, which ensures that content stays online while avoiding detection by government or the private sector. Decentralisation, solidarity, and a shared language have also allowed for strategies on how to exploit AI for extremist purposes to expand fast and to do so in widely accessible spaces. Especially susceptible are those individuals who have been first exposed to extremist content through humour, which is commonly used as a circumventing strategy by the far-right.³⁶ Indeed, in this study, humour has been key in the production of AI-generated content, permitting hate symbols and rhetoric to be allegedly disconnected from their actual meaning while helping strengthen the sense of belonging among users, turning what would be an incoherent mix of accounts into a movement where affiliated individuals work together to advance their shared political, social, and even artistic ideas.

Fundraising

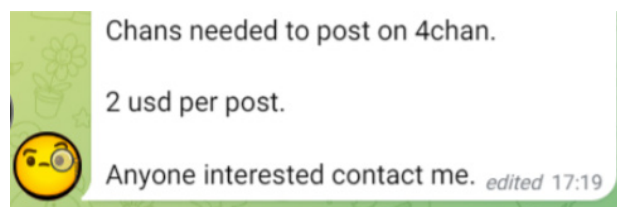
Historically, organised right-wing extremist groups have relied on diverse funding strategies, including donations, concerts, and merchandising.³⁷ In this research, we have seen a number of interactions that indicate money is being made in exchange for AI-generated digital content, as seen in Image 6. As such content is assisting the far-right in producing and disseminating their imagery and narratives, we may argue that AI is allowing for the faster commercialisation of hate-based political activism online. Regarding the revenue generated by this activity, available data suggests that content can be purchased for as little as “2 USD per post”, and as much as “25 USD for 5 minutes [of] work”. Although this may not seem like much, given the accessibility of AI tools and the limited technical skills required to obtain the desired results, an interested party operating under these conditions could produce enough AI-generated content to earn up to \$2,400 USD per day if they dedicated eight hours daily to creating malicious content. If said interested party

³⁶ Bàrbara Molas, “‘Tomatoes for Tanks’: Humour and Violence in Post-Brexit Meme Culture” , *Global Network on Extremism & Technology*, March 27, 2023, <https://gnet-research.org/2023/03/27/tomatoes-for-tanks-humour-and-violence-in-post-brexit-meme-culture/>

³⁷ Rachel Ehrenfeld, *Funding Evil How Terrorism is Financed-- and how to Stop it* (Bonus Books: 2005).

happened to be an affiliated individual to the movement, that is a user with a developed sense of belonging and allegiance to the far-right cause based on shared knowledge and views, then there would be a risk of AI being used to fundraise for extremist purposes, whether for personal gain, or for the purposes of directly supporting a group.

Image 6: Screenshot depicting a Telegram post offering money in exchange for content.



Far-right online content “designers”, as one particular user calls them, are also requested for “meme wars” and “AI memetic warfare”. These are attempts to confront liberal or democratic mainstream opinion with images or image-text units, typically presented as a joke, to shift the public narrative around a specific issue.³⁸ A particularly important moment in the formation of far-right memes was during the 2016 US presidential election, where pro-Trump memes flooded the internet to such an extent that it was considered to have been instrumental in changing public opinion to one that was sympathetic of Donald Trump.³⁹ Meme wars, if supplied with enough content, are central not only to communication and propaganda but also to recruiting.

Cultural War

The individuals exchanging views on how to inflict a cultural war through content do so intentionally, as far as the evidence shows. Reflecting upon the achievements of the far-right online and specifically pointing at how AI has accelerated these, a user writes: “I’m sure you have noticed but the pendulum of thought has swung in the opposite direction.” They continue by saying that nowadays it is possible to see content against the LGBTQI+ community on social media platforms to an extent that was unthinkable two years prior. Another user shares that view: “Slowly but steadily the general population is waking up. The memetic warfare is finally showing effect.” The strategy: “[I]f everyone is on the list then no one is”. What this comment implies is that the goal is for these circles to use AI to normalise far-right content, or to bring harmful and controversial content to the most popular and accessible sites as much as possible and until online users become accustomed to a counter-narrative to more democratic ideas.

This process is called “strategic mainstreaming”, a purposefully driven process to influence public discourse in favour of extremist ideas.⁴⁰ It is also a form of “metapolitics”. Coined by far-right French philosopher Alain de Benoist, metapolitics consists of aiming for “long-term shifts in culture and ideology that eventually lead to political change” instead of directly seeking political power through democratic elections or violence, for example.⁴¹ The American far-right (mostly online) movement of Alt-Right has been a prominent advocate for metapolitics since its inception in 2017, hoping to reshape societal views on national identity, multiculturalism, gender rights, racism, and globalisation, with adherents mainstreaming their ideas on social media platforms rather than staying in the fringes of the political debate and the dark web.⁴² In the long term, these

38 NCTV, “Memes as an online weapon An analysis into the use of memes by the far right”, May 2024.

39 Jacob Davey, Erin Marie Saltman, Jonathan Birdwell, “The mainstreaming of far-right extremism online and how to counter it”, in Lise Esther Herman, James Muldoon, eds. *Trumping the Mainstream: The Conquest of Democratic Politics by the Populist Radical Right* (Routledge, 2018), pp. 23-53.

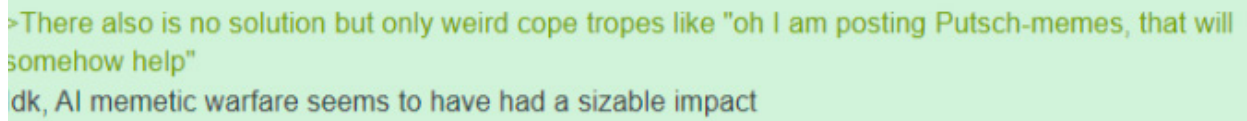
40 Bart Cammaerts, “The Mainstreaming of Extreme Right-Wing Populism in the Low Countries: What is to be Done?,” *Communication, Culture and Critique* 11, no.1 (2018), <https://academic-oup-com.vu-nl.idm.oclc.org/ccc/article/11/1/7/4953068>.

41 Ico Maly, *Metapolitics, Algorithms and Violence: New Right Activism and Terrorism in the Attention Economy* (Routledge, 2023).

42 Julia Ebner, ‘Going Mainstream: How extremists are taking over’.

strategies can result in the “acclimation” to extremist content together with the normalisation of radical statements and the shifting of conventions towards the extreme at all levels of society, including in electoral processes.⁴³

Image 7: Screenshot depicting a 4chan discussion on memetic warfare.

A screenshot of a 4chan discussion on memetic warfare. The text is displayed on a light green background. The first line reads: ">There also is no solution but only weird cope tropes like "oh I am posting Putsch-memes, that will somehow help". The second line reads: "dk, AI memetic warfare seems to have had a sizable impact".

>There also is no solution but only weird cope tropes like "oh I am posting Putsch-memes, that will somehow help"
dk, AI memetic warfare seems to have had a sizable impact

⁴³ Alice Marwick, Benjamin Clancy, and Katherine Furl, “Far-Right Online Radicalization: A Review of the Literature,” *The Bulletin of Technology & Public Life*, (2022), 10.21428/bfcb0bff.e9492a11; Burnett S. Ruth Wodak, *The politics of fear: The shameless normalization of far-right discourse*. 2nd Edn (Sage, 2021), 1-337.

Conclusions

By studying the strategic behaviour of online users looking to exploit AI for the benefit of far-right ideas and agendas, this report has provided new evidence on the tactics used to “jailbreak” or circumvent AI’s safety measures for harmful purposes; on the main challenges faced by these users; and on the current and potential risks resulting from their activities. Above all, this report has revealed that the practice of “jailbreaking” has allowed for the establishment of a loose but actively connected far-right online community of users that regularly exchange information on how to deceive AI tools and platforms. Indeed, individuals may either join chat groups focused on far-right content, which includes conversations on how to use AI to fulfil a harmful agenda, or they may join groups focused on AI that include discussions on how to produce far-right content. In both scenarios, the successful exploitation of AI results mostly from group efforts, where a collective decides to mobilise for a “user in need” based on solidarity or common goals. Such goals are broadly to spread, and keep online for as long as possible, far-right disinformation and content. Through these exchanges, this community is currently successfully mainstreaming radicalising material across accessible platforms, including X, Reddit, 4chan, and Telegram, where users at risk of radicalisation can be easily targeted.

While tech companies have continuously improved their technologies to prevent the misuse of AI models, so have users’ skills and creativity to confront them. Nonetheless, some obstacles remain, as illustrated by this report. These include Terms of Service, EU regulatory frameworks, paywalls, and (self-)censorship. An interesting finding in this research is that users may strategically wish to delegate the production of AI content by asking other users to create it, potentially in exchange for money.

This data is preliminary, and it necessitates further inquiry, specifically concerning the relationship between AI-generated content and fundraising efforts for/by extremist organisations or loose online clusters with an extremist agenda. By illustrating how far-right users have accelerated the spread of harmful content by successfully exploiting AI tools and platforms, this report has overall contributed to improving our understanding of the misuse of AI through new data and evidence-based insights that may inform action against the dissemination of hate culture through the latest technologies. The recommendations below are designed to be a first step in that direction.

Recommendations

Less testing, more manual monitoring

Given the persistent efforts by far-right groups to develop new techniques that bypass AI terms of service and spread extreme content online, manual monitoring of platforms is becoming increasingly necessary. Policymakers must stay current with emerging trends and proactively anticipate new methods of AI misuse, but they ought to do so based on existing data as much as on hypothesised or imagined ecosystems. In addition, while testing has its merits, it often falls short of capturing the full scope of exploitation occurring in real-world settings. By directly observing far-right chat platforms where harmful and borderline content is shared, experts can gain a deeper understanding of evolving tactics, including coded language, prompts, and jailbreaks. These expert teams should work in tandem with automated tools to develop more sophisticated protective measures against online radicalisation, ensuring they remain ‘ahead of the curve’ in this ever-changing landscape.

Increase private-public partnerships

Despite efforts by both the private (e.g. Terms of Service) and public sectors (e.g. TCO, DSA) to prevent the misuse of AI, greater collaboration is essential. A stronger partnership between these sectors, involving information exchange, methodological transparency, and best practices related to AI exploitation and the spread of harmful content on mainstream chat platforms more generally, is needed. Enhancing transparency around AI decision-making processes in particular, and regarding the datasets used to train AI systems, would not only foster greater trust between AI companies and other sectors but also provide researchers and policymakers with deeper insights into how AI systems function and how their safeguards are being circumvented. This improved understanding is key to refining regulations and developing more resilient protections against AI exploitation.

Better regulations for borderline content

Past research on borderline content has identified that chat-based platforms’ algorithmic recommendation systems can amplify borderline content and facilitate its dissemination.⁴⁴ To overcome this, we recommend policies targeted at deliberately reducing the visibility of borderline content, or “deamplifying” it. Some tech companies have implemented such policies. For instance, Google has publicly announced that borderline content has the lowest priority in the search algorithm systems,⁴⁵ and Twitter (now X) states in their files that their platform uses “visibility filtering”, limiting the visibility of tweets shared by certain users, in an attempt to avoid the dissemination of harmful content.⁴⁶ The efficacy of this strategy can be further enhanced by the creation of legislative and/or regulatory regimes requiring companies to address borderline content, rather than maintaining it as a voluntary measure.⁴⁷ Lastly, to regulate borderline content with respect for human rights, platforms should create policies that follow legal frameworks on content moderation practices, such as Article 19(3) of the International Convention on Civil and Political Rights and UNHRC General Comment No. 34 2011, para 33, which regulates the

44 Stuart Macdonald and Katy Vaughan, “Moderating borderline content while respecting fundamental values,” *Policy & Internet*, 16, no. 2 (2024), <https://doi.org/10.1002/poi3.376>.

45 Google, “General Guidelines Overview,” March 5, 2024,

<https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf>.

46 Victor Nava and Bruce Golding, “Latest ‘Twitter Files’ reveal secret suppression of right-wing commentators,” *New York Post*, December 9, 2023, <https://nypost.com/2022/12/08/suppression-of-right-wing-users-exposed-in-latest-twitter-files/>.

47 Stuart Macdonald and Katy Vaughan, “Moderating borderline content while respecting fundamental values,” *Policy & Internet*, 16, no. 2 (2024), <https://doi.org/10.1002/poi3.376>.

duties and responsibilities involved in the freedom of speech, providing limits to this right in extraordinary circumstances, such as when it poses a threat to national security, stability or the rights of others.

More government engagement

Currently, tech platforms have significant freedom in setting up their terms of service (ToS). Our data indicates that while some platforms have robust policies and enforcement mechanisms to address the creation and spread of harmful content, others maintain more lenient standards. To address this, we recommend that governments and policymakers enact domestic or regional legislation that mandates tech companies to include provisions in their ToS specifically aimed at preventing the generation and distribution of harmful content. For example, by having tech companies agree to commit to AI safeguarding regulations proposed by the state, including investing in cybersecurity measures or allowing independent teams to push models into bad behaviour (a.k.a. “red teaming”) as seen in the United States,⁴⁸ as well as by creating national laws on generative AI to balance monitoring and censorship while abiding to human rights and the rule of law. Such measures would help ensure that users do not migrate from more regulated platforms to those with weaker policies, creating a cohesive approach across all generative AI services.

Identification and payment walls

The data gathered in this report indicates that requiring users to leave their identification or pay to create accounts on AI platforms may hinder the ability of harmful users to produce and disseminate harmful content. Currently, most AI platforms only require an e-mail address and passcode to access their services. By requiring users to identify themselves more explicitly, platforms can track individuals who are producing content that violates their policies and more readily cooperate with law enforcement agencies to identify these users and hold them accountable when and if the content infringes the law. Furthermore, they can significantly increase the efficacy of bans and suspensions put on users who have violated content policies. Finally, by requiring identification with an official ID and/or credit card to create a new account, users who have been previously banned would also be unable to regain access by creating a different account with a Virtual Private Network (VPN) and/or on a different device.

⁴⁸ The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>

About the Authors

Bàrbara Molas

Dr. Bàrbara Molas is an expert in international threats and security, in particular the far-right in Europe, the UK and North America. In relation to this subject, she has published over thirty expert pieces and technical reports, four books, and appeared in a number of relevant outlets, including the New Statesman (UK), the Daily Mail (UK), Grid (NL), de Volkskrant (NL), El País (ESP), El Diari de Barcelona (ESP), El Temps (ESP), and the Canadian Broadcasting Corporation (CAN). At the International Centre for Counter-Terrorism (ICCT), she researches current and emerging threats in the EU and beyond.

Heron Lopes

Heron Lopes is a Research Assistant at Leiden University's Institute of Political Science and the Uppsala Conflict Data Programme (UCDP), where he researches political violence and insurgency in Southern Africa and Latin America. He joined the International Centre for Counter-Terrorism (ICCT) in February 2024 as an intern for the Current and Emerging Threats Programme. He holds an MSc in Political Science from Leiden University and an LLM in Law and Politics of International Security from the Vrije Universiteit (VU) Amsterdam.



International Centre for
Counter-Terrorism

International Centre for Counter-Terrorism (ICCT)

T: +31 (0)70 763 0050

E: info@icct.nl

www.icct.nl