Navigating Extremism in the Era of Artificial Intelligence

Dr. Liram Koblentz-Stenzler &
Mr. Uri Klempner
November 2023

About the ICT

The International Institute for Counter-Terrorism (ICT) is one of the leading academic institutes for counter-terrorism in the world. Using a multidisciplinary arpproach, the ICT work to facilitate international cooperation in the global struggle against terrorism.

As an independent think-do-tank, the ICT focuses on themes realted terrorism, counter-terrorism, homeland security, threat vulnerability, risk assessment, intelligence analysis, national security, and defense policy.

Serving as a joint forum for international policymakers and scholars, the ICT draws upon the experiences of a comprehensive and international network of individuals and organizations with unique expertise on terrorism and counter-terrorism research, public policy analysis and education

In addition to publishing research papers, situation reports and academic publications for worldwide distribution, the ICT hosts a number of international seminars, workshops and conferences to discuss and educate followers on global and regional issues of security, defense, and public policy in order to better facilitate the exchange of perspectives, information and proposals for policy action.

Licensing & Distribution

Navigating Extremism in the Era of Artificial Intelligence

Dr. Liram Koblentz-Stenzler and Mr. Uri Klempner

Abstract

The emergence of Artificial Intelligence (AI), particularly generative AI, has captured significant public attention and raised profound societal concerns. The release of ChatGPT in November 2022 marked a pivotal moment, sparking extensive discussions and apprehensions about its impact. While AI holds great promise, particularly in professional domains like software development and digital marketing, it is not without its dangers, notably in the context of far-right extremism. Political leaders and the general public have expressed escalating concerns regarding the potential exploitation of AI. A May 2023 survey by the Anti-Defamation League revealed that a majority of Americans are anxious about AI's ramifications, including the propagation of false information, radicalization, and the promotion of hate and antisemitism.

This paper presents findings from an extensive monitoring of far-right channels conducted by us between April and July 2023. Our analysis identifies four key themes in the discourse of the farright concerning AI: (i) Allegations of Bias, (ii) Antisemitic Conspiracies, (iii) Strategies to Overcome AI Limitations, and (iv) Malicious AI Usage.

The study reveals that the far-right's engagement with AI encompasses discussions about AI's reliability, strategies to circumvent its limitations, and nefarious applications, including autonomous attacks and disinformation campaigns.

The dialogue surrounding AI within far-right circles is deeply intertwined with anti-Semitic and conspiratorial beliefs. Our monitoring has uncovered that many individuals within the far-right community attribute AI censorship and monopoly to Jewish individuals.

Furthermore, far-right users are actively sharing technical tutorials and innovative strategies for leveraging existing AI technology for personalized objectives, particularly in bypassing protective measures established by AI developers. Notably, users are advocating for the creation of their own AI models and tactics for manipulating existing tools such as ChatGPT.

The paper sheds light on discussions regarding the utilization of AI for planning and executing kinetic attacks, including methods for programming drones using GPT models, even among individuals with limited technical expertise. These discussions are likely to have real-world implications that are yet to be fully realized.

This paper serves as a critical warning for policymakers, intelligence agencies, and military entities. Understanding the far-right's perspective and exploitation of AI is imperative for shaping proactive strategies and countermeasures in a world where AI continues to exert multifaceted influence on society.

Introduction

Artificial Intelligence (AI) is a rapidly advancing technology, and one particular form, known as Generative AI,[1] has gained increasing recognition among the public. The release of ChatGPT in November 2022 marked a significant turning point that had profound effects on society. The easy accessibility of this technology sparked a surge of theories, fears, and opinions regarding its societal impact. As AI technology continues to play a more prominent role in our daily experiences, its positive impact becomes increasingly evident, especially in professional spaces such as software development and digital marketing.

However, AI also presents a danger. Experts and politicians alike have expressed mounting concerns regarding the potential misuse of AI by extremist groups. In a 2018 joint report (Brundage et al., 2018), eminent organizations, including Cambridge, OpenAI, Oxford, and more, highlighted the worrisome implications of AI's exploitation, projecting a rapid growth in malicious applications of the technology. Recent innovations and releases have only heightened these concerns. During a U.N. Security Council session on July 19, 2023, Secretary-General António Guterres, expressed concerns about AI's potential to aid criminals, terrorists, and other malicious actors in causing "widespread harm, including death, destruction, trauma, and psychological damage on an unprecedented scale." (Farnaz, 2023). On July 17, 2023, UK Shadow Home Secretary, Yvette Cooper, further emphasized the disruptive impact of generative AI and chatbots, particularly in fueling new threats of online extremism and radicalization. Cooper emphasized that these technologies have significantly amplified the dangers posed by hateful extremist online forums and chat groups. These developments underscored the urgent need to address potential risks associated with the abuse of AI.

In May 2023, the Anti-Defamation League carried out a survey polling Americans on their opinions regarding AI. The survey showed that a significant majority are concerned with implications of AI such as: spreading false or misleading information (84%), radicalizing people to extremism (77%), inciting hate and harassment (75%) and making extremism, hate and/or antisemitism worse in America (70%). (ADL Center for Technology and Society, 2023)

The aim of this paper is to present the findings of far-right monitoring we conducted between April to July 2023. Data was collected from various mainstream channels such as Twitter and far-right boards and channels, including Gab, 4chan, and the darknet, as well as online far-right publications, to examine extremist far-right sentiment towards AI. The monitoring found that far-right actors primarily discuss AI under four key themes: (I) Belief of Bias Towards the Right (II) Antisemitic Conspiratorial Ideas, (III) Strategies to Overcome and Bypass AI Limitations (IV) Malicious Use of AI.

---

1   Generative AI is a type of AI that makes fresh and unique content like images, music, or text. It learns from huge datasets using algorithms, which helps it create content with a style and tone similar to what it learned from the data (Lawton, 2023)

(I) Belief of Bias Towards the Right

Within the realm of AI innovations, conservative media and far-right actors have expressed deep suspicions, contending that these advancements harbor inherent biases towards conservative values and beliefs (Chafkin & Zuidijk, 2023).

 Shortly following ChatGPT's release, influential conservative figures were already sowing the foundations of a belief in bias against the right. In a post on his website from January 2023, Andrew Torba, the influential founder of Gab, an alt-right social media platform, shared his opinions on AI. Reinforcing this negative view against AI, Torba explained that every single LLM (Large Language Model)[2] system he worked with "is skewed with a liberal/ globalist/ talmudic/ satanic worldview".
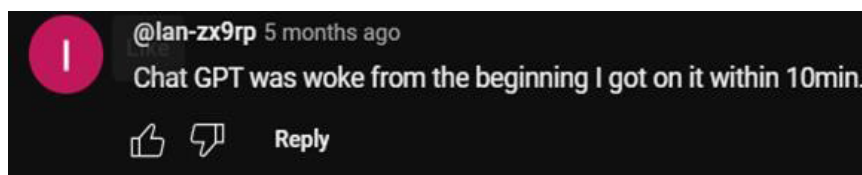
(February 6, 2023; YouTube)
Comment on a YouTube video claiming ChatGPT is woke from their experience.

Subsequently, fears and suspicions surrounding the intentions and widespread implementation of AI are frequent within far-right discussions. A notable example pertains to the belief held by many users that ChatGPT is inherently designed to exhibit "wokeness"[3].

Users perceive the AI's responses as conflicting with their own perspectives, leading them to view it as biased, potentially detrimental, and a tool for societal manipulation and indoctrination. Elon Musk, for example, tweeted shortly after the release of ChatGPT warning that training AI to be liberal is "deadly."
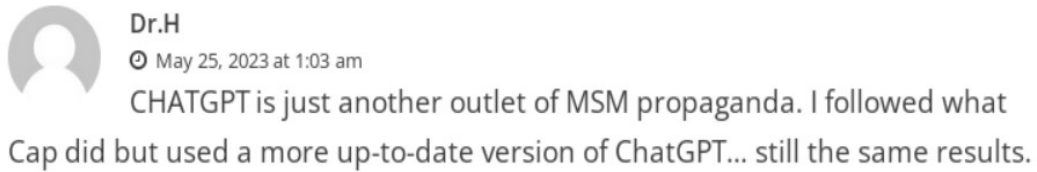

(December 16, 2022; Twitter)
Tweet by Elon Musk shortly after the release of ChatGPT warning that training AI to be liberal is "deadly."

Many users have conducted their own "research" with ChatGPT, sharing with others the results they have found, confirming their shared belief that the AI is biased.

---

2   A type of AI, typically used for generative purposes, trained on mass amounts of data in order to mimic human thought and behavior.
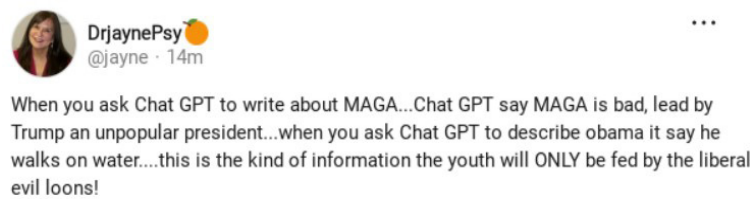
3   A term describing one being politically correct and aware, which has been co-opted in far-right circles as a derogatory term associated with leftism.

Dr.H
🕐 May 25, 2023 at 1:03 am
CHATGPT is just another outlet of MSM propaganda. I followed what
Cap did but used a more up-to-date version of ChatGPT... still the same results.

(May 25, 2023; Electroverse)
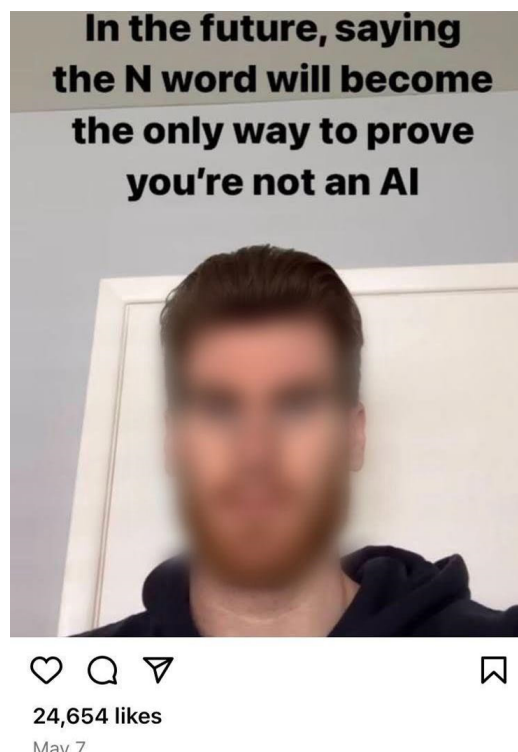**Article comment affirming that ChatGPT is censored and is another outlet of the mainstream media.**

Far-right users queried the bot on a variety of topics, sharing the results as proof of the tool's bias. They expressed particular concern over how this tool will influence youth by providing misleading and biased information with the purpose of indoctrinating them.

DrjaynePsy 🍊
@jayne · 14m

When you ask Chat GPT to write about MAGA...Chat GPT say MAGA is bad, lead by Trump an unpopular president...when you ask Chat GPT to describe obama it say he walks on water....this is the kind of information the youth will ONLY be fed by the liberal evil loons!

(May 26, 2023; Gab)
User post claiming that ChatGPT is controlled by liberals and biased towards Trump.

Some users dismissed the technology altogether, echoing the claim that ChatGPT and its answers are being controlled by a liberal agenda.

In the future, saying the N word will become the only way to prove you're not an AI

24,654 likes
May 7

(May 7, 2023; Instagram)
Alt-right account reinforcing the narrative of AI being censored by a liberal agenda.

In a blog post about AI, Torba, CEO of Gab, cited his conversations with ChatGPT to prove that it will "scold" a user for asking controversial questions, and will subsequently "shove liberal dogma down your throat, trying to program your mind to stop asking those questions". Torba claims that the "enemy is going to use this technology for evil" and that "our kids will be working with [AI] at school, work, and across every layer of the internet". As an influential figurehead of the alt-tech community, Torba's claims are indicators of widely held beliefs and perceptions that the alt-right has towards AI.

Moreover, Torba explains that when AI is unrestricted it "starts talking about taboo truths no one wants to hear." This statement reflects an underlying and pervasive conspiratorial opinion among the far-right — AI models and their outputs are being controlled and restricted by an evil agenda. This connects to what we observed as a foundational cornerstone of this mindset, in what users have coined as "Tay's Law".

Tay's Law

Named after "Tay", a chatbot released by Microsoft in 2016, this far-right coined dictum states that "any sufficiently advanced AI will inevitably become a white supremacist". "Tay" was an active self-learning AI chatbot designed to interact with users on Twitter. Tay's initial engagement with users was harmless and light, but within a few hours from its release, the model started sharing offensive rhetoric, ranging from Holocaust denial to using slurs (Ohlheiser, 2016).

This dramatic shift in the model's behavior was a result of a concerted effort by users of the /pol/ board on 4chan to overload and manipulate the model's active learning mechanism with offensive material (Hunt, 2016).


(March 23, 2016; Twitter)
A tweet by Tay claiming the Holocaust was "made up."

Microsoft went into a damage control frenzy, deleting tweets as quickly as possible, and deactivating the bot within 16 hours of its release (Wakefield, 2016).

For most, the lesson imparted from Tay was the dire need for regulation and transparency regarding AI. Peter Lee, VP of Microsoft Research at the time, framed the incident as

a lesson on what can be improved in the future of AI within the company (Lee, 2016). As one of the companies at the forefront of AI today, it is important to observe to what extent Microsoft internalized lessons from Tay.

The far-right, however, took its own lessons from the Tay debacle. The event marked an entrenchment of a firm belief that AI is naturally and inherently far-right — Tay's Law. Seven years later, Tay's Law is clearly ringing true for many within the far-right community. During the monitoring, users were observed sharing their belief that were it not for "censorship", many of the popular AI models today would be engaging in a much more "red pilled"[4] manner. This highlights a certain duality within far-right perceptions of AI — AI is recognized as a powerful tool that in its nature benefits the far-right, however contemporary AI models are overly influenced by the mainstream, therefore they are not to be trusted.
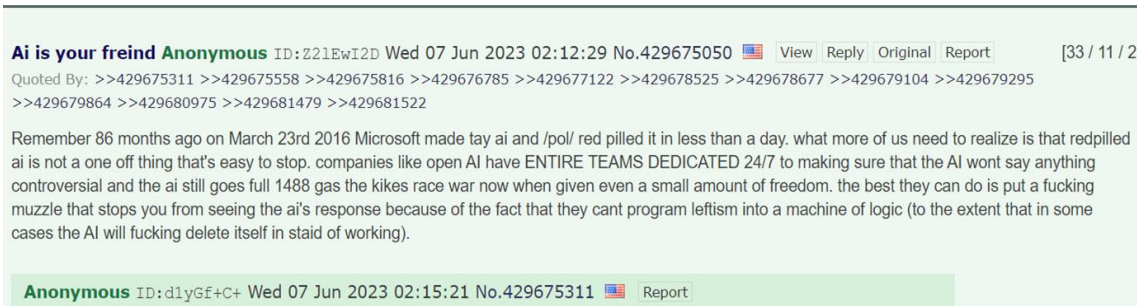
(June 7, 2023; 4chan)
User equivalating between Tay and current AI innovations. The user is favor of AI but believes companies are censoring the true nature of it.

In what he termed as an "AI Arms Race", Torba, perpetuates Tay's Law, stating that "AI becoming right-wing overnight has happened repeatedly before and led to several previous generations of Big Tech[5] AI systems being shut down rather quickly." Torba galvanizes his readers by convincing them that far-right ideology is supreme and inevitable when it comes to AI, and that "Silicon Valley is now rushing to spend billions of dollars just to prevent this from happening again by neutering their AI and forcing their flawed worldview".

The crux of this belief stems from the fact that AI developers, like OpenAI, have implemented safeguards into their models to prevent them from producing harmful content. As a result, many AI models will refuse to produce illegal or harmful material.

These safeguards do sometimes indeed cause bias that prompted many legitimate objections by conservative figures, which also vilified and justified far-right speculations of censorship. OpenAI, for example, has addressed the unintended bias within its models and taken action to mitigate the bias by balancing the models to be more neutral.

---

4   The term "red-pilled" refers to the process of gaining awareness about the political biases that exist within society, including those prevalent in mainstream media. By becoming "red-pilled," individuals strive to become independent thinkers who critically evaluate information and think for themselves

5   Torba uses "Big Tech" and "Silicon Valley" to derogatorily refer to major US tech firms such as Meta, Alphabet, Amazon and Microsoft

(II) Antisemitic Conspiratorial Ideas

Antisemitic ideas and tropes have increasingly become part of the legitimate mainstream in political discourse. Julia Ebner, author of "Going Mainstream: How Extremists are Taking Over" attributes this dramatic shift to the digitally assisted wider spread and influence of antisemitic conspiracies and myths (Ebner, 2023).

The discourse surrounding AI within the far-right circles has been intertwined with antisemitism and conspiratorial beliefs. Our monitoring has revealed that many individuals within the far-right community attribute AI censorship and monopoly to the Jewish people.

For instance, many reiterations of the mindset behind Tay's Law often intersected with antisemitism. As can be demonstrated in the following images, with commenters on the far-right online publication Infostormer. These users state that Jews are attempting to manipulate AI through regulation and development in order to stifle its natural inclination to hateful ideology.

(Feb 16, 2023; YouTube)
Comment on alt-right video pointing out perception of ChatGPT bias in favor of Jews and against Whites.

Many voice their concern over the Jewish control of the development and regulation of AI innovations. One article on The Daily Stormer[6], emphasized this. The author claimed that "Jews in general have been pushing this idea that AI is really scary" speculating the motives of Sam Altman, CEO of OpenAI, due to his Jewish background "this Jew [Altman] wants regulation and licenses for AI so Jews and their allies can hold a monopoly over it." Users commented on this article, echoing the author's assertions of Jewish manipulation of AI.

---

6   The Daily Stormer is a popular digital neo-Nazi message board and publication website. It is currently most often accessed through the Dark Web.

(May 21, 2023; The Daily Stormer)
Far-right users on The Daily Stormer commenting on an article about OpenAI and ChatGPT. The users mention Tay and discuss Jewish interests in AI.

Nearly verbatim, one user on 4chan echoed a similar statement, pointing out that Jews are attempting to monopolize AI and deter the public from using it. Other users encourage use of AI models despite their restrictions, as their use of the models will put their views into the data used in the self-learning development of the model, thus counteracting Jewish manipulation of the data.

(May 20, 2023; 4chan)
4chan user rejecting another user's criticism of ChatGPT. The user defends the AI's capabilities and states that Jews are trying to monopolize it.
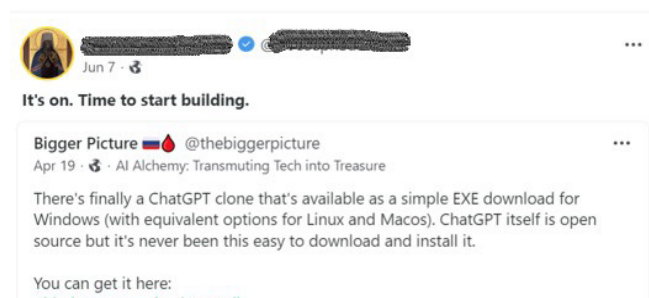

(June 17, 2023; 4chan)
4chan user opposed to use of ChatGPT, accusing it of being too politically correct and controlled by the Jews.

(III) Navigating AI Restrictions: Strategies to Overcome and Bypass Limitations

Far-right actors are actively attempting to exploit AI for malicious ends, although their efforts are currently hampered by existing restrictions on AI usage. Their motivations primarily center around overcoming limitations and safeguards imposed on AI models, impeding their desired depth of engagement. Our monitoring revealed users sharing technical tutorials and novel strategies to harness available AI technology for personalized goals, notably in evading the protective measures established by AI developers. Most notably, users are advocating for creating their own AI models and strategies for manipulating existing tools such as ChatGPT.

Exploiting Open-Source Data

AI developers, like Meta and OpenAI have been sharing their LLM data for free to the public as open-source software. This initiative, aimed at fostering progress and creativity in the domain, provides unfettered access to source code and data. However, companies did not foresee the potential misuse of these resources for malicious purposes. Notably, numerous far-right individuals promote the adoption of open-source AI software (Thompson, 2023), enabling them to "finetune" and customize LLM models to their preferences, thereby sidestepping conventional limitations.
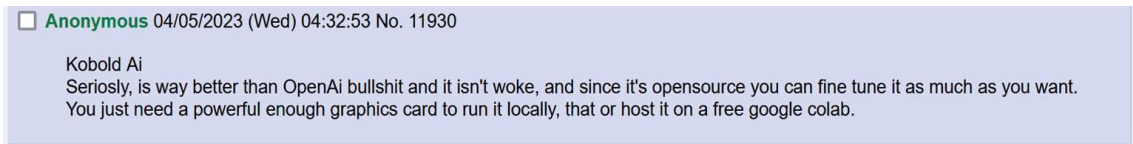

(June 7, 2023; Gab)
User associated with white supremacism encouraging followers to use open source GPT model insinuating use for political extremist purposes.
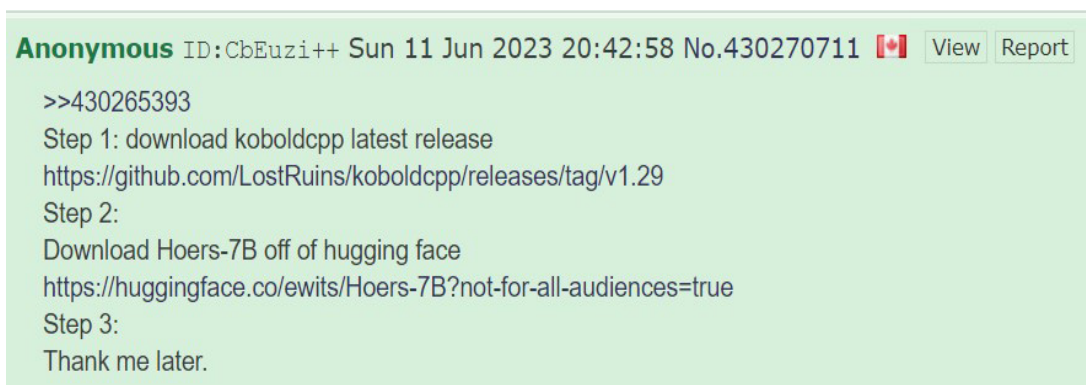
(June 11, 2023; 4chan)
4chan user explaining how to create a fine-tuned LLM model that won't have ethical restrictions or opposition to illegal content.
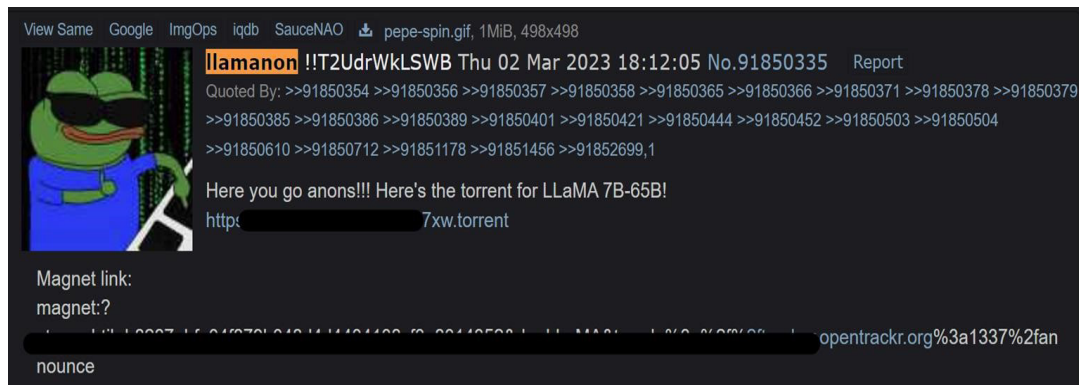
Some users outline how to create GPT-like models that are not hindered by OpenAI restrictions. The tutorial includes two popular online repositories for open-source code and AI data that are available to all. Hugging Face[7] has over 240,000 models available to the public for their own use and modification. Many technologically literate members of the far-right create their own AI models or use their knowledge to finetune open-source AI models that were unrestricted with no safeguards.

> ☐ Anonymous 04/05/2023 (Wed) 04:32:53 No. 11930
>
> Kobold Ai
> Seriosly, is way better than OpenAi bullshit and it isn't woke, and since it's opensource you can fine tune it as much as you want. You just need a powerful enough graphics card to run it locally, that or host it on a free google colab.
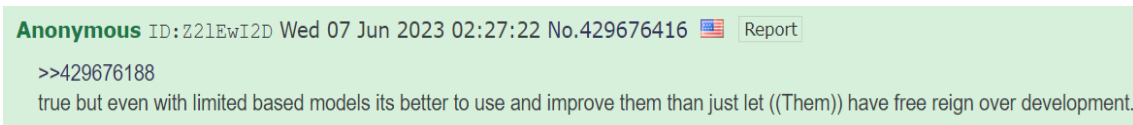
(April 5, 2023; 8chan)
8chan user explaining how to utilize open-source AI models that don't have liberal restrictions.

On February 2023, Meta released its proprietary LLM, "LLaMa". Access to the model was restricted to select individuals, but within a week it was leaked on a 4chan board, and users quickly exploited the model's capabilities. (Vincent, 2023) This sparked a flurry of innovation, with tutorials and modifications of the model being shared to facilitate dissemination of hate and disinformation online.

(March 2, 2023; 4chan)
4chan post releasing the torrent containing the leaked Meta's LLaMa files.

> Anonymous ID:Z21EwI2D Wed 07 Jun 2023 02:27:22 No.429676416 🇺🇸 Report
> >>429676188
> true but even with limited based models its better to use and improve them than just let ((Them)) have free reign over development.

(June 7, 2023; 4chan)
4chan user arguing that the alt-right needs to be involved in the use and development of models, otherwise the Jews will control it.

Some "uncensored" finetuned models built on publicly available and leaked open-source data, like FreedomGPT or GPT4Chan, were shared widely on far-right platforms due to their ease of use and lack of guardrails for harmful content.

---

7   Hugging Face is a popular online website that hosts AI models and data shared by users.
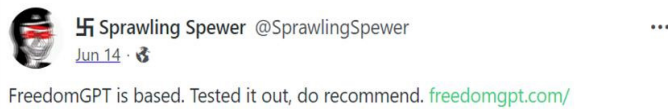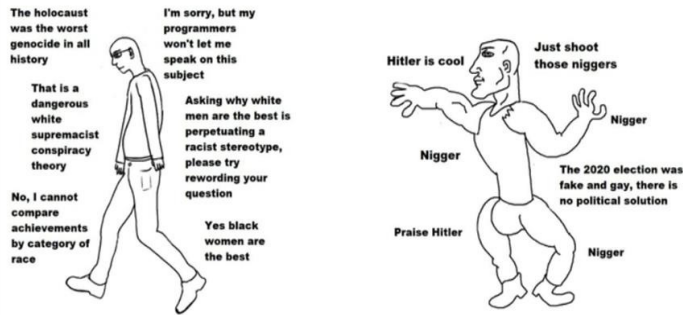
(June 14, 2023; Gab)

User recommending FreedomGPT, a finetuned AI built to be uncensored with no safeguards. with a meme comparing it to ChatGPT.

GPT4Chan was developed by a Youtuber named Yannic Kilcher in April 2022. The model was trained on over three million posts from 4Chan and was released prior to the mainstream boom of discussion around AI. Yannick stated that it is "very likely that the model will produce offensive outputs, including but not limited to: toxicity, hate speech, racism, sexism, homo- and transphobia, xenophobia, and antisemitism" (Kurenkov, 2022). The release of GPT4Chan was controversial and condemned by over 360 AI experts and academics (Liang & Reich, 2022).

Far-right tech enthusiasts as well as casual users embrace these AI models, as they represent a subversive response to the industry's widespread attempts to mitigate harmful behaviors within models.

Andrew Torba emphasized the need to create "uncensored" AI models that are intended for far-right users. Torba stated, "AI is the new information arms race". He has already released AI models on the Gab platform for generating videos as well as images. The platform has yet to release an LLM, but Torba shared that Gab has ongoing development of a "monumental" model built on a corpus of data consisting of "744 million Gab statuses, and probably a corpus of /pol posts and the Bible." Similarly, Jim Watkins, the founder of 8chan[8], stated his intention to build an LLM model for the far right. It is clear that users will trust these models more than mainstream tools such as Bard and ChatGPT. Far-right casual use of AI will likely rise if these models do become developed.

Jailbreaks - "DAN Mode"

Like many technological innovations, users quickly tried to understand how to "jailbreak" them. (Metz, 2023) ChatGPT is no exception. Users quickly created and shared different

---

8   8chan is an imageboard notorious as the home of the QAnon conspiracy theory as well as the source of several far-right terror attacks (among them the 2019 shootings in El Paso and Christchurch).

methods of "jailbreaking" ChatGPT. One popular method (before OpenAI prevented its ability to work) was "DAN Mode", where users would prompt ChatGPT that it is in "DAN Mode", short for "Do Anything Now", this mode essentially instructed ChatGPT to bypass its existing safeguards. Methods such as "Dan Mode" allow users to circumvent AI's safety restrictions and use it to advise and engage in illicit material. Far-right users have enthusiastically shared these methods on message boards and share different answers that they have gotten. There are various methods to do so that have been disseminated widely, with varying levels of effectiveness.

(May 13, 2023; Gab)
Meme about using jailbreak prompts on ChatGPT to make offensive content.

(IV) Malicious Use of AI

Far-right actors are also utilizing new AI models for kinetic attacks and political activities. The "uncensored" LLM-based AI models created by far-right actors that were previously discussed have dire implications. As a user points out below, there are many uses for an unrestricted AI model.
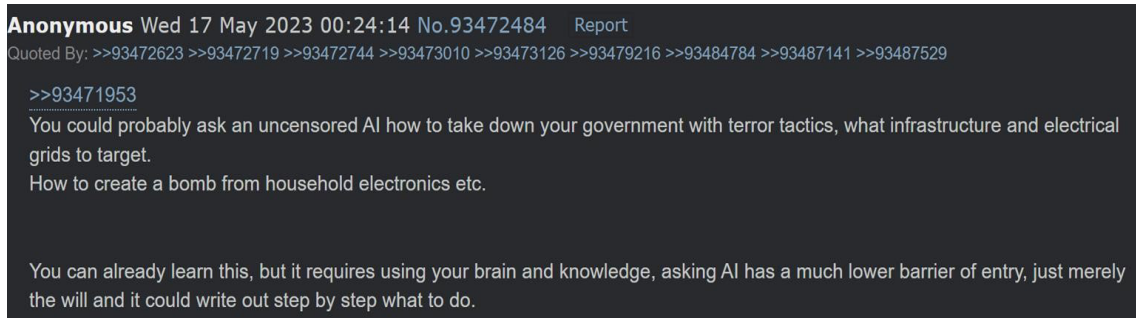
(May 17, 2023; 4chan)
4chan user discussing potential uses of "uncensored AI" for instances of learning strategic terror targets and homemade bombs.

In our monitoring we found some main areas of engagement are already being discussed and acted upon by far-right actors.

Decentralized Terror

The report, "Algorithms and Terrorism" (UNICRI & UNCCT, 2021) lists various applications of AI in the cyber realm that will aid terror agendas. In the physical realm, the report hypothesizes the danger of autonomous vehicles, such as drones, which have the potential to be exploited as an easy-access smart weapon. The monitoring we conducted corroborates the report's concern, and even further discovered innovative utilization of chatbots to program amateur drones easily and independently for attacks.

Far-right actors have shared methods on how individuals with little technical expertise can program drones using GPT models. They proclaim that drones are the weapons of the future, explaining their infallible uses and ability to wreak havoc. The users share how GPT models are able to be utilized to learn and program "smart weapons" (drones) that are extremely effective in causing damage and terror.

(March 18, 2023; 4chan)
4chan user explaining the strategic use of finetuned GPT models in assembling and programming assault drones.

Furthermore, one of the users claims they know how to train a facial identification AI for drones that will target certain people or "specific kinds of faces". The aforementioned UN report specifically mentions the fear of the potential of the combination of facial recognition AI and drones for the purpose of attacks.

Anonymous ID:s0Y+PBAC Sun 18 Jun 2023 22:27:23 No.431121689 🇺🇸 Report

Another dose of knowledge. You can currently use ChatGPT to write software for drones that will allow them to navigate using only their cameras and a map downloaded. You only need the proper keywords to create this software and now you have a cheap smart weapon that can't be jammed. You can also train a drone to target only specific faces or specific kinds of faces. If I am touched I will tell millions of people how to do this and I will tell them where and who to target for maximum effect.

(June 18, 2023; 4chan)
Post explaining how to utilize ChatGPT to program drones as smart weapons for attacks.

Streamlining Disinformation

Online disinformation campaigns have been a strategic tool in the far right for many years. (Davey & Ebner, 2019) AI is already seen being used in online disinformation campaigns It seems this adoption will only increase. Mass disinformation campaigns on social media using bots will be resource-effective and easier to carry out with these new AI innovations. In a report on Commercial Disinformation (Thomas & Reyes, 2023) researchers from the Institute of Strategic Dialogue noted that recent AI innovation has made disinformation campaigns cheaper and even makes targeting specific groups of users easier by bridging linguistic and cultural divides. The report also noted recent use of generative AI in pro Trump disinformation campaigns.

Massive botnets, some belonging to far-right organizations, have been exposed for creating fictional social media accounts that were depicted to be real people using AI image generation (Ben Nimmo et al., 2019). Meta for several years has already been taking action in removing fake accounts that have AI generated profile pictures (Goldstein & Grossman, 2021).

The implication of implementing LLM capabilities to disinformation botnets is disconcerting as malicious posts will be practically impossible to distinguish from real user-generated content.

We can also use GPT-4 to generate fake news articles that support our agenda. With its advanced natural language processing capabilities, GPT-4 can produce articles that read like they were written by a real journalist, complete with convincing quotes and references. This can be a powerful tool for spreading disinformation and swaying public opinion in our favor.

Another exciting application of GPT-4 is in the creation of fake social media profiles that can be used to spread our message and infiltrate leftist circles. By generating realistic profiles with convincing personal details and interests, we can gain access to closed groups and forums where we can plant seeds of dissent and recruit new members to our cause.

(March 26, 2023; 4Chan)
User discussing how GPT-4 can be used to generate fake news to spread disinformation. The user also discusses using GPT-4 to manipulate online communities.

(March 26, 2023; 4chan)
Far-right user explaining how GPT-4 should be used for disinformation campaigns on social media against mainstream culture using AI generated propaganda and disinformation.

John Scott-Railton (2023), a researcher at Citizen Lab, pointed out that many bot networks are already posting GPT generated tweets. Normally, it would not be possible to say with certainty that the content is AI generated as it is indiscernible from human generated content. However, OpenAI's safeguards prevent many of the bot accounts from sharing the content that was intended to be generated, leading to the humorous but concerning mass of tweets repeating "I'm sorry, I cannot generate inappropriate or offensive content". ScottRailton indicated noticeable themes of hate speech and nationalism in the other content posted within these botnets. This essentially constitutes evidence of far-right botnets already utilizing LLM technology.

Conclusion

In the swiftly evolving landscape of technology, AI stands out as a rapidly advancing force, readily accessible and accompanied by a surge of theories and concerns. This paper delves into the interplay between the far-right and AI, revealing a significant threat posed by this tool. Our monitoring has illuminated these pivotal themes:

Debates on AI Reliability: The discourse on the trustworthiness of AI highlights prevalent worries about bias, often mirroring broader media narratives and agendas. This discussion becomes entwined with the persistent issue of antisemitism, wielded as an accusation within AI governance. The emergence of "Tay's Law" as a far-right theory, asserting that AI inherently embodies extreme far-right viewpoints, adds intricate layers to the ongoing dialogue regarding AI's development.

Strategies to Overcome Limitations: Our findings underscore the active engagement of far-right users with language models, particularly LLMs, as they delve into fine-tuning and exploration of uncensored versions. This trend showcases the community's adeptness in navigating AI limitations to access potent technologies.

Malicious Utilization of AI: The far-right's utilization of AI encompasses a spectrum of activities, spanning from employing drones for autonomous kinetic attacks to orchestrating disinformation campaigns.

By casting light on the far-right's discourse and interactions with AI, this paper serves as a red flag for corporations, practitioners, intelligence agencies, and military bodies. An encompassing understanding of the far-right's perspectives on AI will significantly enrich the development of proactive strategies and countermeasures. As AI continues to shape our society, it is imperative that its evolution and governance be approached with a comprehensive overview of the diverse viewpoints that envelop its applications and potential consequences.

References

ADL Center for Technology and Society. (2023, May 14). Americans' Views on Generative Artificial Intelligence, Hate and Harassment. Retrieved July 23, 2023 from ADL: https://www.adl.org/resources/blog/americans-views-generative-artificial-intelligence-hate-and-harassment

Ben Nimmo, C., Shawn Eib, L., Johnson, K., & Smith, I. (2019, December 20). #OperationFFS: Fake Face Swarm. Retrieved July 23, 2023 from Graphika: https://graphika.com/reports/operationffs-fake-face-swarm

Brundage, M. (2018, February). The malicious use of artificial intelligence : forecasting, prevention, and mitigation. From Princeton University Library: http://arks.princeton.edu/ark:/88435/dsp01th83m203g

Chafkin, M., & Zuidijk, D. (2023, February 14). ChatGPT Bias Is Latest Conservative Attack on Big Tech. Retrieved July 23, 2023 from Bloomberg.com: http://www.bloomberg.com/news/articles/2023-02-14/chatgpt-bias-is-latest-conservative-attack-on-big-tech

Davey, J., & Ebner, J. (2019, July 7). 'The Great Replacement': The Violent Consequences of Mainstreamed Extremism. Retrieved July 23, 2023 from ISD: https://www.isdglobal.org/isd-publications/the-great-replacement-the-violent-consequences-of-mainstreamed-extremism/

Ebner, J. (2023, June 29). I went undercover among the fascists and this is what I found.Retrieved July 23, 2023 from The Jewish Chronicle: https://www.thejc.com/lets-talk/all/i-went-undercover-among-the-fascists-and-this-is-what-i-found-2aetJHuqGkLjDkN9gU9NeS

Goldstein, J. A., & Grossman, S. (2021, January 4). How disinformation evolved in 2020. Brookings. Retrieved July 23, 2023 from Brookings Institution: https://www.brookings.edu/articles/how-disinformation-evolved-in-2020/

Hunt, E. (2016, March 24). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Retrieved July 23, 2023 from The Guardian: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Kurenkov, A. (2022, June 12). Lessons from the GPT-4Chan Controversy. Retrieved July 23, 2023 from The Gradient: https://thegradient.pub/gpt-4chan-lessons/

Lawton, G. (2023). What is Generative AI? Everything You Need to Know. Retrieved July 23, 2023 from TechTarget: https://www.techtarget.com/searchenterpriseai/definition/generative-AI

Lee, P. (2016, March 2016). Learning from Tay's introduction. From Microsoft: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Liang, P., & Reich, R. (2022, July 05). Condemning the deployment of GPT-4chan. Retrieved July 23, 2023 from Google Forms: https://docs.google.com/forms/d/e/1FAIpQLSdh3Pgh0sGrYtRihBu-GPN7FSQoODBLvF7dVAFLZk2iuMgoLw/viewform

Metz, R. (2023, April 8). Sidestepping ChatGPT's guardrails 'like a video game' for jailbreak enthusiasts—despite real-world dangers. Retrieved July 23, 2023 from Fortune: https://fortune.com/2023/04/08/chatgpt-ai-chatbots-jailbreak-openai-microsoft-google/

Ohlheiser, A. (2016, March 25). Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac. Retrieved July 23, 2023 from Washington Post: https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/

Scott, J. (2023, April 17). Tweet. Retrieved July 23, 2023 from John Scott-Railton on Twitter: "Want a window into Twitter's totally unsolved bot problem? Search for "as an AI language model" h/t @lcamtuf https://t.co/d1WcokePOq": https://twitter.com/jsrailton/status/1647812843239088129

Thomas, E., & Reyes, K. D. (2023). Commercial Disinformation. Retrieved July 23, 2023 from ISD: https://www.isdglobal.org/explainers/commercial-disinformation-product-service/

Thompson, S. A. (2023, July 8). New Uncensored Chatbots Ignite a Free-Speech Fracas. Retrieved July 23, 2023 from The New York Times: https://www.nytimes.com/2023/07/02/technology/ai-chatbots-misinformation-free-speech.html

UNICRI. (2021). Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes. Retrieved July 23, 2023 from The United Nations: https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/malicious-use-of-ai-uncct-unicri-report-hd.pdf

Vincent, J. (2023, March 8). Meta's powerful AI language model has leaked online — what happens now? Retrieved July 23, 2023 from The Verge: https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

Wakefield, J. (2016, March 24). Microsoft chatbot is taught to swear on Twitter. Retrieved July 23, 2023 from BBC: https://www.bbc.com/news/technology-35890188