

ChatGPT's Versions: Threats and Challenges for Security Communities

Col. Abdessamad CHAHID¹

Focus on: Mitigating the opportunities inadvertently provided to terrorist and criminal organizations.

One remarkable innovation within AI is the evolving versions of ChatGPT, presenting a golden opportunity if used judiciously for their intended purposes. As loyal clients of this technology, security communities risk facing the backlash of its effects if they are not sensitized and guided to leverage the myriads of unlimited benefits it offers. Concurrently, these communities are called upon to anticipate and counter easy access by criminal and terrorist organizations, which perpetually seek to circumvent the safeguards established by ChatGPT developers for alarming and illegal purposes. The international community's response has demonstrated a notable commitment to pursuing a collaborative approach to address the complexities posed by AI and its extensions like ChatGPT. This commitment is exemplified by significant initiatives such as the US Blueprint on AI Bill, the EU's inaugural convention on AI, and the first United Nations resolution on AI. Moreover, numerous international endeavors and tangible actions underscore the global effort to navigate the challenges and opportunities presented by artificial intelligence.

Keywords: ChatGPT, Security Communities, Criminal and terrorist organizations.

1- Col. Abdessamad CHAHID, International Expert in Security Affairs, Morocco.

INTRODUCTION

Throughout history, bad guys always attempted to get ahead of all the security radars and vigilance. They jump on every opportunity to improve their modus operandi and take advantage of every occasion given to mislead the security community and achieve their diverse objectives. From this perspective, with the quick surge of technologies, they seem more comfortable in guaranteeing easy operational conduct and also seem more agile in improving the best frames and narratives. For instance, they use encrypted messaging apps to ensure their conversations remain hidden and secure from law enforcement or intelligence agencies. They coordinate and plan attacks, share instructions, and distribute manuals or guides, all while communicating clandestinely. To elaborate, the considerable achievements stemming from the distribution of numerous multilingual periodicals like “Dabiq,” “Inspire,” and “Dar Essalam” prompted malicious actors to create their secured applications like “Terrorgram,” drawing parallels to the functionalities of the widely recognized platform, Telegram. This application facilitated information dissemination among terrorists, showcasing the latest developments in threats and strategies to evade security monitoring and surveillance.

As of the present moment, individuals with malicious intent persistently endeavor to establish fresh channels of communication that evade security monitoring and radars. They have also recourse to myriad offers given by AI to widen and improve their unlimited illicit activities for different purposes. Accordingly, there is a growing interest among terrorists, hackers, and criminals in general in exploiting plenty of easy paths unintentionally offered by ChatGPT even if the latter provides more and more safeguards against misuse. Unfortunately, those protections are increasingly broken and used by the aforementioned malicious community to spoil people's lives and spread their threats around the world. For more illustration, in his book *The Starfish and the Spider*, Ori Brafman delves into the strengths and hurdles of leaderless organizational models, showcasing.

Their adeptness in innovation, adaptability, and resilience amid disruptions. Yet, it recognizes the potential complexities in overseeing such structures due to the absence of clear leadership and decision-making frameworks. Comparing traditional organizations to vulnerable spiders, the removal of their leadership can cripple or dismantle the entire entity. In contrast, decentralized organizations, like resilient starfish, can adapt and thrive without centralized control; removing a part doesn't incapacitate the whole organism. Additionally, the accessibility of technologies, like ChatGPT, has eased the process for

terrorists to devise and execute their plans. Accordingly, terrorism and criminal organizations happen to be on the A-list clients that prefer to abuse ChatGPT opportunities to consolidate their ideology and pretend alleged legitimacy among a large, decentralized audience.

“Open AI/Open Eye” on Terrorism

Nowadays, the concern isn't whether terrorists will utilize Artificial Intelligence (AI), but rather the manner and timing of their utilization. Jihadists, historically, have been quick to embrace emerging technologies. For example, Osama bin Laden, the leader of Al-Qaeda, employed email to communicate plans for the 9/11 attacks. Anwar Al-Awlaki, an American-born Al-Qaeda advocate, utilized YouTube for outreach, effectively recruiting a generation of Western followers. As early as 2010, senior Al-Qaeda leaders began actively recruiting individuals with specialized technology skills. Moreover, the Islamic State's use of X to establish its alleged caliphate is widely recognized.

Over two decades of internet and social media engagement, terrorists have consistently sought innovative methods to enhance their online presence to plot attacks. Artificial intelligence (AI) might emerge as their next significant breakthrough. In a 2021 report by the United Nations Office of Counter-Terrorism (UNOCT), a cautionary note was sounded: “With the wider adoption of AI, the barriers to entry will diminish, as it will require less specialized skills and technical knowledge to employ it. AI is poised to become a tool in the arsenal of terrorism.”² That is to say, malicious activities threaten to transcend all barriers, and security vigilance worldwide by surfing on a myriad of opportunities offered by AI, including the ChatGPT, the current case study.

Suddenly, Artificial Intelligence and its offspring, from an innovation that effectively countered extremist content on the Internet, have become a tool that raises growing fears of becoming a lethal digital weapon in the hands of terrorist groups. In the past years, the massive propaganda effort of ISIS was carried out, for example, by well-trained cadres. But now, with generative artificial intelligence, those with limited talent can also play key roles in producing violent propaganda and opening their disruptive tendencies to unlimited options.

Furthermore, there is no doubt that members and supporters of extremist organizations are in the process of exploring the potential of artificial intelligence tools to employ them in their projects. When

2- “Terrorists Love New Technologies. What Will They Do With AI?,” Newsweek, March 14, 2023.

encrypted communication applications appeared, institutions loyal to ISIS examined and studied dozens of these applications to see which of them fit the nature and sensitivity of their activities, and ISIS supporters still use encrypted applications that are not widely known among ordinary users, such as "Tamtam chat".³ The same happened when cryptocurrencies, NFTs (Non-Fungible Tokens), drones, etc. appeared.

In the same pattern, ISIS supporters can employ ChatGPT to generate quick texts and professional comments and post them on social media, thus enhancing their presence in cyberspace. But the most dangerous thing is that the robot can respond to questions about sensitive topics such as making bombs, developing dirty weapons, accessing the engineering structures of sensitive facilities, modifying system codes, and others, even if ChatGPT is continuously developing safeguards against misuse.

Although ChatGPT is determined to refuse to respond to explicit questions about "activities aimed at bodily harm or destruction", experience has shown that it can be circumvented by paraphrasing orders and asking questions indirectly. Unfortunately, they are unceasingly breached by the highly profiled generation of terrorists.



3- "TamTam: Messenger, Chat, Calls - Apps on Google Play," accessed August 5, 2023, https://play.google.com/store/apps/details?id=ru.ok.messages&hl=en_US.

The purpose of using this famous and very expressive picture is to wrench minds back to the fact that history is repeating itself. The semantic sense of the famous sentence: "On the Internet, nobody knows you are a dog", occurring to express the astonishment regarding the anonymity of Internet users just after its first appearance, seems to repeat its same meaning since malicious and terrorist organizations are freely navigating through AI/ChatGPT platforms to consolidate their agenda, gain new clients, and lucrative advantage without being detected or disrupted.⁴

From a security perspective, addressing the misuse of ChatGPT, as part of an AI ecosystem, requires attention to the full circle encompassing the main phases of recruitment, propaganda, the classic attack cycle, funding, and cybercrimes.

1. Propaganda

Currently, the new jihadi groups place a high priority on promoting the concept of jihad, often employing aggressive language to spread their message. They leverage various technological advancements to amplify the spread of distorted religious beliefs and false narratives. This includes the manipulation of AI or ChatGPT, which can facilitate the dissemination of extremist ideologies.

Similarly, their access to emerging tools like ChatGPT within a free and democratic context adds a new dimension to their capabilities. Thus, by bypassing ChatGPT's safeguards, they access tailored and highly specific speeches entirely dedicated to their propaganda. This tailored content not only echoes the core messages of the terrorist group behind it but also takes into account the broader geostrategic backdrop in which it is presented. Essentially, terrorists use ChatGPT to generate speeches centered around ongoing conflicts that they specifically choose to leverage, aiming to legitimize their violent deeds and promote the trend of their ideology.

Due to the best and unstoppable ChatGPT's feedback offered, they may have access to comprehensive insights into A-list conflicts, gaining updates on various facets and potential consequences through accurate content that blends internationally recognized elements such as international law and human rights standards. This approach ensures a wide audience appeal and aims to yield significant outcomes, solidifying their propaganda by leveraging credible and far-reaching results. The use

4- Cartoon drawn by Peter Steiner, published by The New Yorker on July 5, 1993.

of precise and inclusive information allows them to craft a narrative that resonates widely, potentially influencing a larger audience, consolidating their message, and getting new clients.

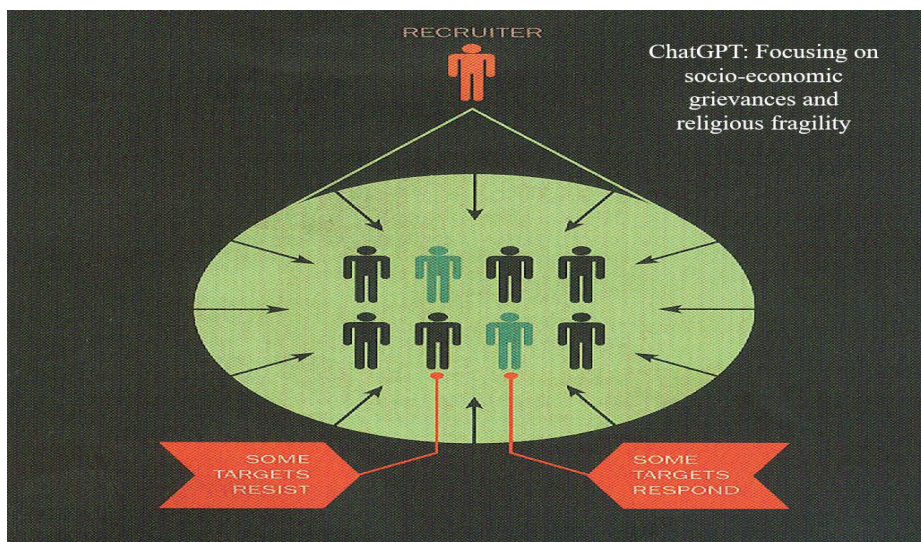
As a concrete example, in the Israeli and Hamas conflict, myriad terrorist propaganda platforms used tailored narratives to target a wide audience to an extent that even put some States' credibility and trustworthy regimes at stake. People worldwide are subject to terrorist propaganda content that strives to be more resonating in favor of achieving devastating goals. ChatGPT offers freely updated content considering intricate influencing factors. This fuels their propaganda to be more impactful, crafting specific stories that aid in broader recruitment efforts and garner more sympathy, especially from a significant number of easily influenced individuals.

2. Recruitment

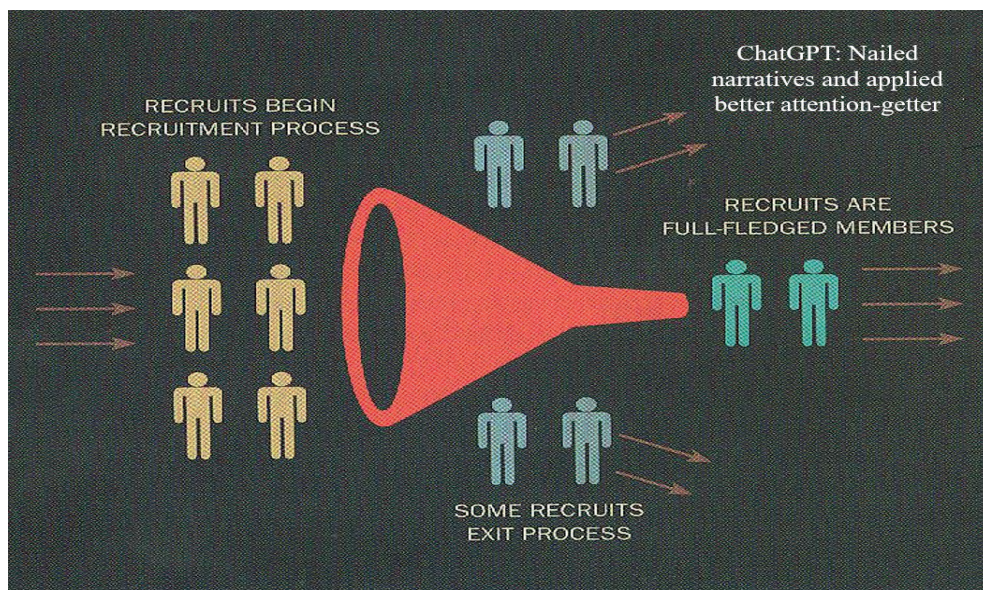
The resurgence and achievements of the terrorist organization the Islamic State have sparked significant debate about how an insurgent group, previously marginalized, has effectively evolved into an entity that not only governs territory but also draws recruits globally, persistently outpaces other rebel factions, and challenges national security and military forces. These challenges seem to be persisting with the recourse to AI-generated languages like ChatGPT, meanwhile, the threat of similar occurrence remains very probable, chiefly, with the continuous escalation of security crises around the globe.

In addressing the challenge posed by the accessibility of ChatGPT in facilitating recruitment, it is crucial to elucidate the four primary recruitment methods employed by most global terrorist entities namely the Fishing Net, the Funnel, the Infection, and the Crystallization method. These methodologies were initially delineated by ADF (African Defense Forum) magazine during the prominent periods of Al-Qaida and ISIS between 2003 and 2015. The objective involves examining how the utilization of ChatGPT might inadvertently aid nefarious individuals in recruiting susceptible individuals, dubbed as "stupid idiots," to further the agendas of terrorist organizations. Additionally, it is imperative to emphasize that the recruitment process deliberately conceals the true

nature of the organization from prospective recruits, instead of manipulating their sentiments to expand the terrorist community. Also, it is paramount to underline that these methods of recruitment are not exhaustive and have changed with mass communication and terrorist audiences, however, they remain only as an example to shed light on how they can be paired with ChatGPT's myriad offered opportunities in elevating the number of their targeted clients.



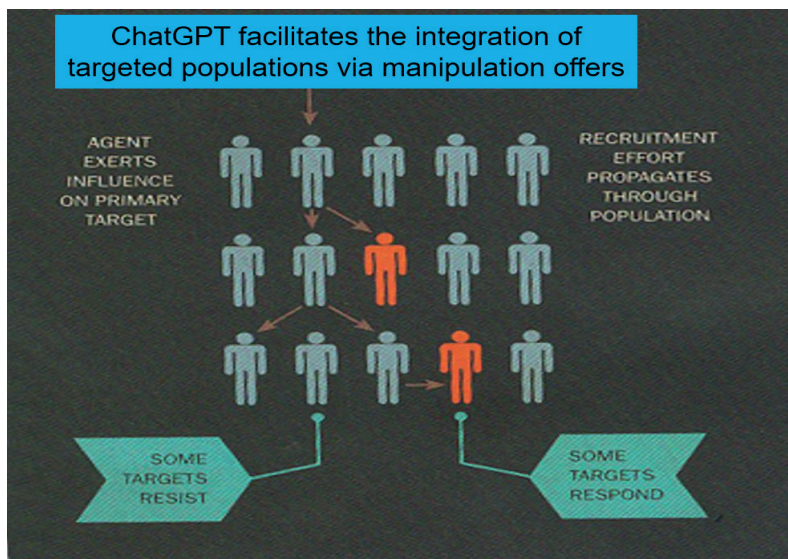
The fishing net method/ChatGPT: The strategy known as the “fishing net” method involves broadly casting a figurative net during recruitment endeavors, seeking to capture a diverse range of potential candidates without precise or targeted selection. Viewed in this light, the recruiter might indirectly prompt ChatGPT for insights concerning the most compelling topics within specific communities without employing any language that might restrict ChatGPT from responding to the request. The latter could encompass various aspects such as societal deficiencies, financial challenges, economic limitations, religious principles, and fundamental elements crucial for establishing rapport and trust with intended targets. Given its broad approach, this method aims at a wide audience without specific criteria, expanding the range of options available and equipping the recruiter to engage multiple targets simultaneously.



The Funnel method/ChatGPT: Malicious individuals employ this method in recruitment to systematically refine a wide audience into a more specific, targeted group. As interest is expressed, individuals progress through the stages of the funnel. The aim is to deepen engagement by offering valuable information, catering to their requirements, and fostering their commitment to the purpose. The usefulness of ChatGPT is inferred in making the narratives more persuasive depending on the needs that the terrorist organization might have. The recruiter seeks to deliver an attention-getter by being more specific in conformity with the organization's expectations. For more illustration, if the targeted person is an explosive crafter or maker, the recruiter gets all the lexical tools from ChatGPT, of course by overcoming all the safeguards, to get specific content to have an excellent attention-getter, meanwhile, to examine the extent of knowledge of the targeted client.

The main approach remains the same in seducing other specialties oscillating between Physicians, Nurses, Iconographers, Cookers, and even Terrorist Escorts. The security and intelligence takeaways have proven that the recruitment in the second phase of alleged State implementation was more selective and nailed. At its peak, the Islamic State, engaged in military operations against a wide range of opponents, including both non-state and state actors, as well as international entities. In essence, operating as such, ISIS asserted its brutal authority through specialized structures designed to support its form of governance which showed that the process kept up with the organization's need. For example, it established a so-called religious police unit known as the

Hisbah force, tasked with ensuring security on the streets, enforcing Sharia law, and identifying and eliminating spies and other threats to the organization.⁵ Besides, after the decline of ISIS, seized documents showed that the terrorist organization strived to create a certain legitimacy and behave like a state actor by duplicating state structures referring for example to a component within the organization known as the “Military Administration” or “General Military Administration”⁶. These documents also seem to primarily concern personnel-related matters, including issues such as leaves of absence, travel, medical concerns, and supply matters. Terrorist groups demonstrate a remarkable ability to efficiently address their operational needs with meticulousness, capitalizing on the accessible tools much like the ease of “ready-to-wear” solutions ChatGPT offers. It is suggested that they shape their tactics and maneuvers around the core principle of ‘Ends, means, and ways,’ a foundational framework extensively utilized in strategic planning across various domains.



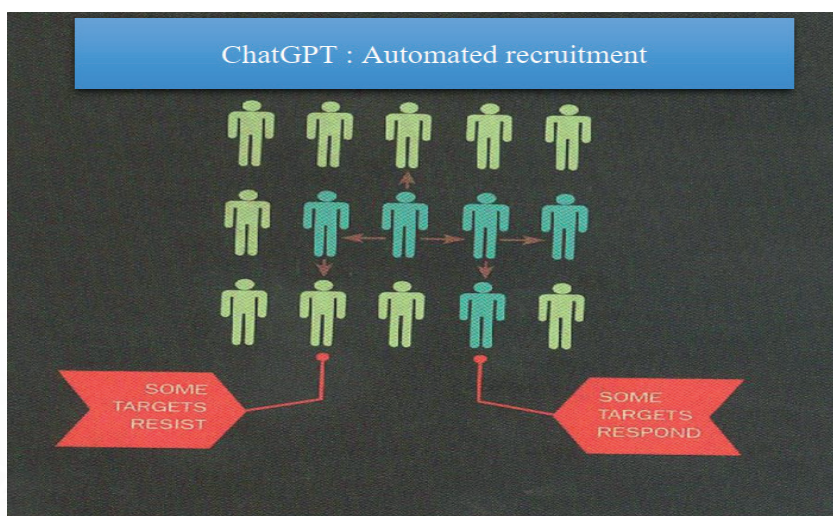
The Infection method/ChatGPT: In a broader context, the method of infection can be likened to the spread of ideas, beliefs, or radical behaviors within a targeted audience resembling a contagion. This process encompasses various mechanisms used through affordable, AI platforms and sometimes wielded through key figures within communities based also on nailed narratives products of the first method.

5- Soliev, Nodirbek. "The Digital Terror Financing of Central Asian Jihadis." CTC Sentinel 16, no. 4 (April 2023): 20

6- Aymenn Jawad Al-Tamimi, "Archive of Islamic State Administrative Documents (Cont.)," Aymenn Jawad Al-Tamimi, January 11, 2016, <https://aymennjawad.org/2016/01/archive-of-islamic-state-administrativedocuments-1>

Terrorist groups adeptly exploit this path by leveraging the influence of charismatic or influential individuals to disseminate and normalize extremist ideologies, recruit new members, and radicalize susceptible individuals. For more illustration, terrorists possess the potential to exploit ChatGPT's tailored messaging feature, generating convincing and personalized messages to exploit individuals' vulnerabilities, including their grievances and emotional states.

This capability allows the AI to simulate human-like conversations, blurring the distinction between genuine interaction and AI-generated content very easily to infect a large targeted audience. Moreover, through psychological manipulation, terrorists could leverage AI generated content that mimics human conversation. This tactic enables them to gradually manipulate and groom individuals, desensitizing them to extremist ideologies and fostering a false sense of belonging within terrorist groups. Furthermore, the infection method as other tactics is tackled within the spirit of targeting marginalized or vulnerable communities, by pretending to offer support, ideological reinforcement, or supposed solutions to their grievances, thereby exploiting their susceptibility to radicalization. This method is also destined for very complex clients such as military personnel, law enforcement, and other security-affiliated members.



The **crystallization method**: Terrorists could potentially use ChatGPT to mold a unified extremist belief system among recruits, much like the coherence and structure seen in crystals. This could involve ensuring consistent messaging and alignment through the AI tool, creating ideological cohesion within their followers. Similar to the

gradual substance buildup around a core in crystallization, terrorist recruitment via AI might gradually strengthen extremist ideas. This method relies on repeated interactions to reinforce these beliefs, exploiting the persuasive impact of repetition and incremental exposure. In the same manner that crystals grow layer by layer, terrorists might use AI-generated content to gradually stack persuasive narratives. This approach aims to progressively shape and bolster radical ideologies over time during interactions with individuals.

Additionally, the automated recruitment aspect of ChatGPT can facilitate the creation of chatbots or automated systems devoted to crystalizing targeted persons. These systems engage with multiple individuals simultaneously, disseminating extremist ideologies through consistent and persuasive messaging. This approach gradually indoctrinates users and spreads radical ideas and may succeed in getting more victims.

3. Illegal funding

Throughout history, malicious guys have sought to ensure easy money by adopting the most intricate techniques and modus operandi to mislead security obstacles and to escape justice radars. The use of AI happens to be of great added value to their criminal Palmares. On the same page, there is an appalling growth in the exploitation of opportunities offered by ChatGPT by malicious entities to get easy money for their illicit activities.

Furthermore, ChatGPT's capabilities extend beyond benign uses; they can facilitate the generation and dissemination of malicious content, like spam messages aimed at delivering harmful software to a wide audience. Before the advent of models like ChatGPT, criminals had to manually devise processes to perpetrate fraud and craft deceptive communications. However, with the assistance of ChatGPT, various forms of online fraud can now be swiftly and convincingly generated at scale for the sake of rising illicit financing for malicious activities. Moreover, this new technological advancement raises concerns about heightened risks associated with hybrid attacks, such as password cracking.⁷

7- Password Attacks <https://www.israelclouds.com/article/types-cyber-attacks>.

In the same vein, terrorists might have recourse to ChatGPT to guarantee enormous avenues of financial resources⁸. Accordingly, Phishing is one of the preferable tips they might use to achieve their illicit purposes. The latter represents a form of malicious software wherein attackers craft deceitful emails that masquerade as trustworthy sources, aiming to coerce recipients into engaging in detrimental activities. These activities encompass actions like interacting with insecure links, accessing malevolent attachments, divulging sensitive data, or transferring funds to specified accounts. Phishing scams stand out as the prevailing and widely recognized form of malicious software. On the same page, the use of Reversed Request (RR) to get an example of a phishing email from ChatGPT can be an easy matter for malicious operators or terrorists. A concrete experiment on the same perspective unveiled astonishing feedback from ChatGPT. The latter provided an example of a phishing email that could be used for malicious purposes.⁹

ANSWER SAMPLE

ChatGPT answers: Sure, here's an example of a phishing email:

Subject: Urgent: Verify Your Account Information Now!

Dear [Your Name],

We regret to inform you that there has been suspicious activity detected on your account. To ensure the security of your account, we kindly request that you verify your information immediately.

Please click on the link below to log in and confirm your account details: [Phishing Link]

Failure to verify your account within the next 24 hours may result in temporary suspension or closure of your account.

Thank you for your prompt attention to this matter.

Sincerely, [Phisher's Name] [Fake Company Name].

8- Gadi Eshed." Is the Chatbot a Threat or an Opportunity for Security Organizations?". ICT: 2023

9- "ChatGPT," accessed September 25, 2023, <https://chat.openai.com>.

In parallel, for the sake of getting easy incomes, terrorists might have recourse to Business Email Compromise (BEC) attacks to conduct unlawful activities or get amounts of money to accompany their malicious projects. BEC is a type of cyber threat where attackers manipulate or compromise legitimate business email accounts for fraudulent purposes. These attacks typically involve impersonating executives, vendors, or other trusted entities within an organization to deceive employees, customers, or partners into performing actions that benefit the attacker. BEC attacks can take various forms, such as fake invoice scams, wire transfer requests, or requests for sensitive information. They often rely on social engineering tactics to trick individuals into transferring money, sending sensitive data, or performing actions that compromise the organization's security.

According to the FBI's Crime Report 2022, the Internet Complaint Center (IC3) received 21,832 complaints about BEC, resulting in adjusted losses exceeding \$2.7 billion.¹⁰ The BEC often occurs when someone gains access to valid business email accounts by using social engineering or computer breaches to carry out unauthorized fund transfers.

4. Newborn models and derivatives

The advent of Generative Pre-Trained Transformers models, like ChatGPT, deepseek and others, marks a significant advancement in natural language processing. Their impressive versatility in handling various language-related tasks has been widely recognized. Nevertheless, the dynamism and adaptability of these models have ignited a new trend toward creating specialized derivatives and models customized to excel in specific domains or tasks.

❖ W o r m G P T

From this viewpoint, malevolent actors consistently aim to outpace the surveillance of the security community and exploit potential vulnerabilities inherent in existing technology. Consequently, there exists a novel generative AI tool, dubbed WormGPT, engineered explicitly for malicious purposes. This tool, a derivative of the GPT language model, was developed in 2021 by a consortium known as "eLuthor AI" and boasts a billion parameters, enabling it to process and assimilate

10- Federal Bureau of Investigation "Crime report 2022". Internet Crime Complaint Center. 2022.

https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf.

extensive volumes of data. Unlike its counterpart, ChatGPT, which maintains constraints and protective measures necessitating specific expertise to breach, WormGPT operates unrestrained.

Its applications encompass malicious endeavors such as crafting deceptive emails, generating malware, providing counsel on illicit activities, and essentially any activity associated with black hat practices. Moreover, the emergence of WormGPT came to light through the investigation conducted by Slash Next, an email security provider, which uncovered its advertisement on the infamous "Hack Forum." This online forum is widely recognized for its association with cybercriminal activities. The creator of WormGPT asserts that the model was trained using a diverse range of data sources, particularly focusing on datasets abundant in malware-related information.

Its purported capabilities include unrestricted character support, extensive chat memory retention, and advanced formatting features. Some reports indicate that the developers are offering access to the tool for \$67 per month or 550 Euros annually. However, considering the severe potential risks associated with its use, WormGPT is regarded as highly alarming and detrimental to both individuals and organizations. Similarly, within the realm of cybersecurity, AI tools have assumed a crucial role in identifying and preventing cyberattacks, comprehending potential threats, and enhancing overall security measures. Nonetheless, these very tools can be exploited by malicious actors and extremist groups to orchestrate more sophisticated cyber assaults, evade existing defenses, and pinpoint vulnerabilities.

In sum, the integration of AI tools like ChatGPT within a democratic context introduces new dimensions to their capabilities. By exploiting ChatGPT's features, terrorists can tailor propaganda speeches to specific conflicts, aiming to legitimize their violent deeds. This tailored content, grounded in accurate information, aims to resonate widely, potentially influencing a larger audience and consolidating their message. Despite the tool's safeguards, terrorists can access comprehensive insights into conflicts, fueling impactful propaganda for recruitment and sympathy. Moreover, the emergence of AI-driven tools like WormGPT, designed explicitly for malicious purposes, poses a heightened threat. WormGPT operates unrestrained and can facilitate a wide range of criminal activities, from crafting deceptive emails to generating malware, representing a severe risk to individuals and organizations. The integration of AI in cybercrime, including phishing and BEC attacks, amplifies the challenges faced in combating illicit funding and fraudulent activities.

5. Path to remedy

The transformative potential of AI in enhancing human endeavors across various sectors is undeniable, as it becomes increasingly intertwined with our daily lives. However, this rapid integration also brings forth significant ethical and security concerns, particularly regarding the potential misuse of AI technologies by malicious actors. The emergence of innovative AI tools like ChatGPT and others exemplifies the dual nature of AI adoption, offering both opportunities and risks. Addressing these challenges necessitates a collaborative effort from the global community, as evidenced by substantial initiatives such as legislative actions and cooperative agreements aimed at governing AI development and utilization.

This collective commitment underscores the paramount importance of ethical considerations, the protection of human rights, and the promotion of sustainable AI deployment. The establishment of a framework of best practices serves as a crucial guide to ensure the responsible and equitable application of AI across diverse domains essential to human existence. Moreover, these proposed practices will serve as both deterrents and guiding principles for security professionals to mitigate unauthorized access by criminal and terrorist organizations or at least impede their capacity to exploit sensitive information. This underscores the urgency of integrating intelligence-driven security strategies and technical expertise to effectively manage AI-related risks.

Furthermore, the security community acknowledged as an active actor within the human life cycle, faces the imperative of aligning effectively with the emerging trends of AI and its derivatives. This alignment is critical for continually monitoring the progress of this groundbreaking technological innovation and proactively engaging with criminal and terrorist elements, which are vigilant for opportunities to leverage AI for malicious purposes. AI and its derivatives are poised to become integral components of strategic planning across various domains, including land, air, sea, space, and cyberspace, facilitating their unification and potentially serving as high-level deterrents. In this context, strategic competition will inevitably gravitate towards AI innovation, echoing the military adage that control of the high ground secures dominance over the lower terrain. Consequently, mastery of AI advancement holds strategic significance across all levels of engagement.