SEAN-CSO

Peace Generation

Southeast Asian Network of Civil Society Organisations

Generative artificial intelligence and countering violent and hateful extremism

Implications, risks and benefits for civil society

Matteo Vergani, Nadia Lukman, Greg Barton, Dina Zaman

Executive summary

1

Generative AI, with its potential for rapid, customised content creation, poses significant challenges for Civil Society Organisations (CSOs) focused on preventing and countering violent extremism (P/CVE), including in the Southeast Asian context. Extremist groups can employ Generative AI for purposes ranging from producing and disseminating targeted disinformation - including creating fabricated interviews or fake videos - manipulating public sentiment with tailored content, and using chatbots to befriend and groom the lonely and the vulnerable. The technology facilitates advanced propaganda distribution, sophisticated cyberattacks, and direct engagement with audiences, including seduction and ideological indoctrination via chatbots. Additionally, Generative AI has enabled new forms of online fraud, stalking and privacy invasions. It even has the potential to transcend the cyber domain and engage in physical attacks by weaponising technology such as drones and other autonomous systems. As these risks emerge, it's crucial for CSOs to adopt advanced countermeasures, redouble efforts to promote media literacy and develop critical thinking, and foster closer and more extensive collaboration with tech and security sectors.

2 Generative AI has the potential to be used to equip CSOs working on P/CVE with advanced tools to combat extremism and to promote community resilience. However, appropriate and ethical use of this tool requires advanced training and support from data scientists who can guide and train CSOs in this important policy area. By learning and leveraging the capabilities of AI apps, CSOs can multiply their capacity to effectively analyse and moderate content in diverse languages, identifying hate and extremist messages in a timely fashion. Generative AI also offers opportunities to produce more creative, effective and sophisticated counter-narrative materials, such as animated videos, automating and tailoring messaging, and opening new possibilities such as writing videogame codes. Generative AI streamlines documentation and research, and facilitates real-time digital literacy tools for the public. Ad-hoc training is required to ensure CSOs make optimal utilization of AI. In sum, Generative AI is a tool that can empower CSOs to be more efficient and effective in safeguarding communities against divisive narratives and ideologies.

3 In using Generative AI technologies, CSOs should consider key challenges, such as biases in AI models, primarily sourced from Western datasets, which can hinder efforts in non-Western contexts. Furthermore, data integrity is crucial; flawed datasets are likely to result in inaccurate projections. There are also serious privacy concerns about risks to user information. For successful deployment, CSOs should emphasise community engagement, uphold ethical standards, champion transparency, and advocate for clear, targeted AI guidelines. Addressing biases, ensuring data quality, and understanding evolving challenges, such as system vulnerabilities, is essential for responsible Generative AI usage.

INTRODUCTION

What is Generative AI?

AI, short for Artificial Intelligence, is a vast field that has been growing and evolving since the 60s. Essentially, it's about creating machines that can behave as if they were thinking and learning like us. Generative AI, a subset of this vast field, refers to tools trained to ingest, process and generate responses based on input data, generally in the form of very large data sets. They have found application across a wide spectrum of scientific and cultural production, ranging from academic authoring and software coding to music and image creation. One of the most widespread examples of Generative AI technology is ChatGPT, a machine-learning system that learns from very large data sets of textual material (such as Wikipedia pages) and can produce sophisticated and ostensibly intelligent writing after training on a massive data set of text (van Dis et al., 2023). Other similar Generative AI technology include Google's Bard, Microsoft's Bing AI, Dante AI, Chatsonic (among others). ChatGPT reached 1 million users in just five days after its release on November 30, 2022, and 100 million users in February 2023 (Sabzalieva & Valentini, 2023). This speedy growth of technology brings big changes to our culture, especially when appropriate laws and rules are yet to be framed and implemented, much less trying to keep up.

Every technology, including AI, has a 'dual use', that is, it can be used for good or bad. It's how we humans decide to use it that makes the difference. For example, in hospitals and clinics, Generative AI is helping solve problems involving big data, such as detecting cancer. But at the same time , we need to be aware that AI can be used to advance the aims of terrorist organisations, violent extremist and hate groups. In this report, we will aim to maintain a balanced view of technology and consider both its opportunities and its risks for the field of countering violent extremism.

Structure of the report?

This report discusses how Generative AI can change the work of civil society organisations working on preventing and countering violent extremism, with a focus on Southeast Asia. Our work is informed by an extensive review of the existing literature on this topic, and by discussions with experts in P/CVE and data science from Australia and Southeast Asia.

The report has three main sections.

- The first section discusses how Generative AI will change the threat landscape of CSOs working on P/CVE, and it builds on previous work on the risks of Generative AI for the spread of disinformation in the form of fake news, deep fakes, fraudulent online interactions (Siegel & Bennett, 2023).

- The second section focuses on the potential uses of Generative AI to support the work of CSOs working in the P/CVE space. This report is among the firsts to look at this topic, and will provide practical insights and suggestions for CSOs working in Southeast Asia. While some applications demand specific expertise, there's potential for the P/CVE community to equip CSOs with the necessary skills, making it a cost-effective strategy.

- In the third section, we outline the primary risks of using Generative AI in P/CVE work, emphasising the precautions CSOs must consider. As we use these helpful tools, we need to remember the risks: AI can get things wrong, and is inevitably biased – even though newer AI models are getting better by checking facts or even telling if a piece of content was made by an AI or a human. We will list what we think are the main practical suggestions and reflections of AI ethicists and regulators to help us understand how to manage the risks of using Generative AI in P/CVE.

This report is designed for a diverse global audience of researchers, practitioners and activists in countries where English is not the primary language. It employs accessible language and summarises key insights using examples, bullet-point lists and concise summaries. For questions or clarifications, please contact any of the authors or the SEAN-CSO team.

Section 1 : Generative AI and VE: the new landscape CSOs need to think about

Extremist and hate groups have found multiple nefarious uses for Generative AI, leveraging its capabilities to further their agendas. This opens new risks and a new challenging landscape for CSOs working on P/CVE. An example is the case of the Rohingya people in Myanmar, who experienced escalated real-world violence partly due to a deluge of online misinformation on Facebook (Amnesty International, 2022). This digital propaganda painted the Rohingya as threats and 'invaders', amplifying existing prejudices and influencing. Such disinformation significantly influenced public sentiment, exacerbating the atmosphere of hate. For instance, a post targeting a Muslim human rights defender was widely shared, urging violent actions against him and even his entire community. Amidst this, Senior General Min Aung Hlaing, Myanmar's military leader, stated on his Facebook page that the Rohingya did not exist in the country. This digital misinformation campaign resulted in tragic real-life consequences in 2017 when over 700,000 Rohingyas had to flee Rakhine State due to systematic acts of violence, including murder and arson, by the Myanmar security forces. Imagine what can happen in a world where Generative AI can help hate groups to create powerful, large scale and individually targeted misinformation campaigns against one or more minority groups. In a context like Myanmar, where there already are pre-existing tensions, Generative AI can potentially ignite real-world violence, in the wrong hands.

Here's an exploration of key uses, grouped by their nature, with examples relevant to the Southeast Asian context and the online environment.

Information Manipulation and Disinformation.

Dissemination of Disinformation : Generative AI is potentially a powerful tool for the rapid, automated creation of content at low cost and with low levels of technical expertise, and when misused, it can be harnessed by extremist groups to disseminate targeted disinformation. Such advanced AI systems can scan and analyse vast amounts of data to identify key narratives, biases, or susceptibilities within specific demographics. Once these vulnerabilities are recognised, Generative AI can then generate disinformation tailored to exploit them, ensuring that the fake news appears highly believable and resonates effectively with the targeted group. Prominent figures can be mimicked using deepfake technology. Imagine a video showing an influential Southeast Asian leader voicing extremist views, disseminated to cause public unrest. Such videos can be nearly indistinguishable from real footage, and, once they have gone viral, virtually impossible to retract or neutralise.

Take the 2019 Indonesian elections as an example. During the electoral process, a deluge of misinformation spread across various platforms. While traditional methods of spreading disinformation required skilled human actors to craft narratives, Generative AI has the potential to exponentially increase the volume and precision of such efforts. Consider the real risk of fabricated interviews: in the past, forging a written statement or doctoring a video clip was time-intensive. With Generative AI, however, creating a convincing, fake video interview of a political candidate making controversial remarks becomes almost trivial. Such synthetic content could sway voter sentiment, exacerbate tensions, and severely disrupt the democratic process. The transition from human-generated propaganda to AI-driven disinformation campaigns augments the challenge of discerning fact from fiction and reinforces the necessity for advanced countermeasures.

Another method of the use of Generative AI in spreading disinformation is the use of microtargeting on online users that specifically target vulnerable users and create filter bubbles (Bontridder & Poullet, 2021). Tracking methods via cookies, third party tracking or browsers fingerprinting become the source of data in which AI used to pick who sees what, which originally is crafted for online business model. This method further exacerbates the spread of disinformation to users. Furthermore, some individuals may deploy social bots on social media platforms. These social bots imitate human interactions and was capable in skewing online discourse which then contribute to disinformation.

Social Manipulation : The power of Generative AI does not just lie in its ability to create content rapidly, but also in its capability to fine-tune this content to target specific audiences. Violent extremist groups recognise the potential of Generative AI in manipulating public opinion. By leveraging this technology, they can craft highly convincing articles or speeches that subtly or overtly support radical ideologies. Generative AI's ability to convincingly mimic human-generated content provides extremist groups with a powerful weapon: the capacity to craft narratives that significantly blur the lines between fiction and reality. By leveraging Generative AI, these factions can generate timely and engaging content that appears entirely factual, making the task of debunking them more challenging.

For instance, in Southeast Asia, where platforms like Facebook and WhatsApp are incredibly popular, a Generative AI-produced article that appears to be from a credible source could be circulated widely. Such an article might draw upon regional issues, sentiments, or historical events to lend it authenticity. In a real shift of risk with the advent of Generative AI, consider the scenario where a fabricated article, purportedly from a respected regional news outlet, spreads misinformation about ethnic tensions in Myanmar. Instead of relying on human writers, who might introduce detectable biases or inconsistencies, or make grammatical or idiomatic errors, Generative AI could produce a more polished and 'neutral'-sounding piece, making the deception harder to identify. Such manipulations can fuel existing tensions, leading to potential violence and strife. The evolution from human-crafted propaganda to AI-driven manipulation underlines the urgent need for sophisticated fact-checking and media literacy initiatives.

Propagation of Extremist Propaganda. Generative AI has already brought a transformative shift in the way extremist propaganda is disseminated. While earlier propaganda relied on manual curation and distribution, Generative AI offers the capability to customize content for diverse audiences, making it more resonant and persuasive. Telegram, LINE, and similar messaging platforms, popular in regions like Southeast Asia, are frequent channels for such activities.

For example, consider a situation in the Philippines, a country that has faced jihadist insurgencies. Prior to the advent of Generative AI, extremist groups might have circulated generic messages calling for recruits or support. With Generative AI, these messages can be tailored to individual preferences and linguistic nuances. A call for support in Mindanao could reference specific local grievances, while another message targeted at urban centres like Manila could use different cultural touchpoints. As a result, the risk is magnified manifold: propaganda becomes more embedded in local narratives, harder to counteract, and more effective in radicalising individuals. The seamless integration of AI-crafted content within localised contexts underscores the increased challenge in detecting and combating extremist outreach.

Cyber Attacks and Security Breaches.

Exploitation of Systems. Generative AI enables extremist groups to model and simulate various attack strategies on digital infrastructures. For instance, in Southeast Asia, where many nations are rapidly digitising their public sectors, vulnerabilities may exist in untested or hastily-deployed systems. Consider the Philippines, where government websites have historically been targets of cyber-attacks. With Generative AI, an extremist group could simulate various strategies to identify the weakest point in a government platform promoting religious harmony, potentially causing it to spread hate instead. For civil society organisations, it becomes crucial to understand that conventional cybersecurity measures may no longer suffice in this new risk environment.

Spear-phishing. Security personnel, traditionally the last line of defense against breaches, now face sophisticated AI-generated communications that are increasingly difficult to differentiate from genuine messages. Using Generative AI, a hate group might replicate the communication style, for example, of an Indonesian NGO leader working against violent extremism. By sending a message urging immediate action with a malicious attachment to key personnel, they could gain illicit access. Civil society organisations must invest in training sessions, ensuring members are educated about the enhanced spear-phishing threats posed by Generative AI.

Data and Identity Theft. In 2017, Malaysia witnessed one of its most significant data breaches, compromising millions of personal details. Now, with the advent of Generative AI, extremist factions can deploy advanced algorithms to bypass even sophisticated authentication systems. These tools can predict patterns, passwords, and exploit overlooked vulnerabilities. For civil society organisations in Southeast Asia, where many might not yet be fully aware of Generative AI's capabilities, it is vital to conduct frequent system audits and ensure members use multi-factor authentication, treating data protection as a paramount priority.

Malicious Software Distribution. Generative AI can craft messages tailored to the reader, making malicious software distribution more insidious. For instance, a Thai government agency or NGO working against extremism might receive a seemingly legitimate email concerning a regional cooperation meeting. However, hidden within attachments or links could be software designed to steal data or disrupt operations. Civil society organisations must implement advanced threat detection systems and foster a culture of caution when receiving unexpected communications, even if they appear to come from trusted sources.

Spam Message Generation. Beyond single-target spear-phishing, Generative AI facilitates the mass creation of spam messages that seem tailored and genuine. In a region like Southeast Asia, where platforms like LINE or WhatsApp dominate, receiving a message from an 'acquaintance' sharing an article or video is commonplace. For an Indonesian or Malaysian civil society member, this could look like an article link highlighting extremist activities. But once clicked, it releases malicious software onto their device. Civil society groups need to be proactive, educating members about the heightened risks of spam and the deceptive realism that Generative AI introduces.

Direct Engagement and Propagation.

Ideological Advancement, Recruitment and Immersive training. Generative AI, when integrated with platforms similar to ChatGPT, can create realistic and coherent narratives that align with extremist ideologies. For instance, in Indonesia, where extremist online propaganda is a growing concern, these platforms can be exploited to automate the propagation of extremist ideologies at an unprecedented scale. Chatbots can engage, indoctrinate, and recruit new members, and even facilitate discussions about potential attack strategies under the guise of anonymity. AI can be used to generate for example personalised games, or immersive war games that can be used as part of training. For civil society organisations in Southeast Asia, it's crucial to understand and monitor these platforms, ensuring they can debunk or counter the automated narratives quickly and be aware that individuals might be influenced without direct human intervention from extremist groups.

Online Fraud. With the e-commerce boom in Southeast Asia, platforms like Lazada, Shopee, and Tokopedia have become central to many consumers. Generative AI can be employed by extremist factions to craft realistic product listings, reviews, or seller profiles, promoting fraudulent schemes or counterfeit products. Imagine, for example, a scenario where an AI-driven seller profile on a platform like Lazada in Thailand provides "proceeds" to a supposed charity that, in reality, funnels money to extremist activities. Civil society organisations should foster consumer awareness about these sophisticated fraudulent techniques and work closely with e-commerce platforms to detect and shut down such operations.

Invasion of Privacy. Generative AI's capabilities, especially in image manipulation, have introduced a vile form of harassment–synthetic pornography. In countries like Indonesia and the Philippines, where conservative societal norms dictate social acceptance, being a victim of such an attack can have devastating consequences. By superimposing a victim's face onto explicit content, extremists can blackmail, humiliate, or silence individuals, including activists or members of organisations working against violent extremism. It is paramount for civil society organisations to be aware of this threat and provide support, legal assistance, and digital literacy training for potential targets.

Large-scale UAV Attacks. Drones, equipped with AI, represent a dual-use technology. While they offer benefits in surveillance and logistics, they can also be weaponised. In Southeast Asia, the maritime boundaries, especially around the Philippines, are vast and challenging to monitor. Extremist groups could deploy AI-powered drones to either carry out attacks or for surveillance, scouting potential areas for illegal activities or identifying weak points for future operations. Such technology allows extremists to have a bird's eye view, potentially transforming the threat landscape. Civil society organisations should advocate for stricter drone regulations, develop early detection systems, and collaborate with security agencies to counter this emerging threat effectively.

Impact of AI on users. We cannot forget the negative impact of tools like ChatGPT on the humans who contributed (and will contribute in the future) to annotate the content to train the AI models. For example, the Wall Street Journal reported in July 2023, that contractors in Kenya say they were traumatised by effort to screen out descriptions of violence and sexual abuse during run-up to OpenAI's ChatGPT. Isolation, depression, and insecurity using the technology may impact human workers as well.

In conclusion, as Generative AI stretches into every corner of our on- and offline lives, the fight against its misuse, especially by extremist entities, becomes paramount. Civil society organisations must stay informed and prepared to counteract the malicious potential of generative AI.

Section 2. Generative AI and P/CVE: new tools to support CSOs' work against VE

The extensive capabilities of Generative AI apps have the potential to provide a significant edge for civil society organisations in Southeast Asia, especially in tackling extremism and hate groups.

We know that CSOs working on preventing and countering violent extremism in Southeast Asia are by and large defunded, have limited access to skills, resources and often rely on volunteers (Barton et al., 2022). Generative AI can help empower these civil society organisations to become more effective in their fight against violent extremism and online hate. For example, civil society organisations can use Generative AI to monitor online hate. Consider the sheer volume of posts generated every minute in social media. For a human team to monitor this content manually would be an insurmountable (and very expensive) task. Generative AI can be tailored to monitor this vast ocean of data, identify patterns, and flag potential hate speech. For instance, civil society organisations in Southeast Asia could deploy cheaply and quickly such AI tools to identify emerging hate groups or trends that might previously have gone unnoticed. Using data from Generative AI monitoring, civil society organisations can make more compelling arguments to policymakers. They could provide tangible evidence of the rapid spread of misinformation or hate speech, leading to more informed regulatory decisions.

Generative AI can also help in curating tailored educational materials. Suppose a charity organisation focusing on inter-faith harmony wants to educate different communities about each other. The AI app could generate content highlighting similarities across cultures or religions, debunking myths and stereotypes. Generative AI can assist by analysing successful counter-narratives from the past and generating content suggestions that resonate with contemporary audiences. For instance, an organisation countering Islamophobia in countries like Thailand or the Philippines might use AI-generated content that showcases stories of Muslim individuals' positive contributions to their communities. In response to trending hate topics, AI can automatically generate and spread positive, fact-based messages across platforms to counteract the surge, almost like a digital counter-movement. Using Generative AI, CSOs can, for example, create the foundational code for a video game wherein players step into the shoes of immigrants, understanding their struggles and triumphs. Such empathetic experiences, delivered innovatively, and in a timely and cost-efficient manner, can be potent tools against hate.

Here's a detailed look at how AI can make a difference:

Research and monitoring of online hate and extremism.

Recent open access articles (Ray, 2023) looked at how Generative AI is innovating research by offering advanced capabilities in data processing and analysis. For example, tools like Elicit can help to extract key information from scientific literature, reducing manual labor in literature reviews and expediting the research process. ChatGPT can discern patterns, correlations, and anomalies within vast data pools, enabling researchers to identify novel associations and generate innovative hypotheses. ChatGPT can also support research in providing advice on design, method selection, and analytical approach, ensuring high quality standards in research. In a region as linguistically diverse as Southeast Asia, monitoring online content represents considerable challenges in terms of resources.

Generative AI can process multiple languages, making content moderation for lesser-known dialects, such as those found in the islands of Indonesia or the Philippines, or in the forests of Myanmar or Laos, more streamlined and efficient with limited resources. This means extremist messages, often concealed within ordinary conversations, can be detected without exhaustive human-led translation efforts. For CSOs, this tool provides the means to swiftly identify and counteract coded extremist messages, ensuring online platforms remain safe from divisive ideologies.

Public Sentiment Analysis. Southeast Asia has witnessed an explosive growth in social media adoption. Countries like Indonesia and the Philippines see platforms like Facebook and Twitter as primary news sources. AI can sift through vast amounts of data, detecting signs of extremist sentiments or planned actions. This capability enables CSOs to potentially spot a rise in extremist chatter before a real-world incident occurs, facilitating proactive interventions.

Document Generation. CSOs can leverage AI to automate the process of collating and summarising the data generated from their research activities. For instance, an organisation monitoring online hate and extremism in Indonesia could receive concise daily summaries, allowing them to respond promptly to emerging threats. This can potentially reduce the impact of administration and reporting on the already stretched finances of CSOs, providing the opportunity to concentrate the energies and resources on actual P/CVE work.

Countering Misinformation and Building Digital Literacy.

Combating Misinformation. The digital age, while full of promise, has been marred by a surge in misinformation and disinformation. For CSOs, countering false narratives becomes paramount, especially when these narratives incite hate or division. By training AI on regional content – encompassing both genuine and misleading sources – CSOs can improve the ability to identify false narratives. Such an AI-driven tool can become an indispensable asset for CSOs, ensuring that communities remain informed and resistant to divisive content.

Digital Literacy Campaigns. Information is only as good as its understanding. CSOs can harness AI to fortify digital literacy, offering users real-time tools that discern factual content from fabricated narratives. For instance, a user in Thailand could be instantly alerted about the veracity of a news article, ensuring they're equipped to make informed decisions. Over time, such interventions will cultivate a discerning digital community, less susceptible to extremist ideologies.

Training and capacity building of personnel.

Generative AI, with its ability to provide personalized learning experiences, can transform training and education by offering tailored tutoring tailored to individual learning preferences and needs. This adaptability aids in bridging educational disparities and boosting effectiveness and efficiency. CSOs that need to train their personnel can use tools like ChatGPT to enhance their training in personalised and scalable ways.

In conclusion, the fight against extremism in Southeast Asia necessitates state-of-the-art tools, and well-trained uses. For CSOs, generative AI offers the capabilities to stay a step ahead, ensuring their communities remain informed, united, and resilient against divisive forces. It is crucial that CSOs require ad-hoc advanced training and support to fully harness the power of Generative AI. We suggest that this should be a priority for funding agencies involved in P/CVE in the region.

Section 3. Risks of using generative AI.

Imagine a scenario where CSOs use Generative AI as a co-pilot in analysing vast swathes of online communication to detect extremist propaganda. Generative AI can swiftly identify patterns, providing CSOs with insights faster than traditional methods. However, akin to using a seasoned mentor alongside teaching staff, Generative AI should complement rather than replace human expertise. Data interpretation by AI is pattern-driven, and without grounding in the specific nuances of violent extremism, the results might not always resonate with the truth.

CSOs should be wary of solely relying on Generative AI. Outsourcing everything is likely to prove perilous, particularly when it comes to sensitive areas like P/CVE. For instance, if training data for the AI contained inadvertent biases against a particular ethnic group, the AI might wrongly flag content from that group, amplifying societal divisions.

A further matter of grave concern lies in the domain of data integrity. CSOs often handle sensitive, personal information. Generative AI's usage, if not appropriately safeguarded, could inadvertently breach the integrity and confidentiality of such data, especially if it pertains to identifiable individuals. This is further compounded by the current legislative landscape, which remains nebulous regarding AI ethics and responsibilities. Its interaction, so human-like, triggers concerns about users being misled into thinking they are interacting with another human, raising questions about the ethicality of such deceptions. Additionally, the environmental footprint left behind due to the massive computing resources needed to train models like ChatGPT cannot be ignored. Collaborative efforts from AI developers, researchers, and the broader community will be instrumental in navigating these concerns. By addressing these issues proactively, we can harness the potential of AI tools like ChatGPT responsibly, ensuring that their revolutionary capabilities benefit society while minimizing unintended consequences.

As a technology capable of producing human-like text, imagery, and even code, the potential risks of Generative AI are particularly pronounced when applied in sensitive areas such as countering violent extremism.

Bias.

Foremost among concerns is bias. Virtually all AI models inherit biases from their training data, which predominantly lean towards Western datasets, skewed to white, middle-class, communities. This can lead to potential misinterpretations in non-Western contexts. For CSOs working against extremist narratives in Southeast Asia, a biased Generative AI might inadvertently bolster prejudiced views about local ethnic or religious groups, hampering efforts to foster understanding. Over-reliance on AI might curtail critical thinking, and concerns about quality control, contextual understanding, energy consumption, and privacy among others further complicates its seamless integration. A significant point of contention is also the system's model explainability: for example, understanding why ChatGPT offers a particular output remains elusive, raising concerns about transparency.

Misuses.

Misuse presents another daunting challenge. The capabilities of Generative AI can be harnessed for malicious ends. An extremist group might employ Generative AI to craft false videos or audio clips, distorting statements from community leaders to foment discord or instigate violence.

Data integrity.

Data integrity is crucial. If Generative AI attempts to complete missing information drawing from flawed datasets, the output could be skewed. Consider a CSO using AI to predict areas in the Philippines susceptible to extremist influence. A misinterpretation of local cultural nuances due to incomplete data could result in resource misallocation, leaving genuinely vulnerable regions unprotected.

Privacy.

Privacy also cannot be overlooked. Tools like ChatGPT, while powerful, may sometimes compromise user privacy. For CSOs, sharing sensitive information about their stakeholders or initiatives with such platforms might contravene ethical guidelines or even legal frameworks. Data privacy stands paramount, especially with AI's growing involvement in data processing. Intellectual property rights are blurred as AI contributes substantially to idea generation and content creation. For example, the potential misuse of ChatGPT in nefarious activities such as spreading misinformation and its vulnerability to adversarial attacks underscore the need for stringent safeguards. The opacity of ChatGPT's complex algorithms and its susceptibility to produce biased outcomes demand greater transparency and fairness.

To navigate these multifaceted challenges, CSOs can consider several strategies:

- Consistent and diligent capacity building for all CSOs involved. There is no point in reaching out to stakeholders if CSOs themselves are not trained, equipped and resilient. This needs to be multipronged: CSOs, working with P/CVE and AI experts, supported by psychologists working in conflict and security.
- Inform and Engage: It's imperative that communities, participants, or stakeholders are well-informed about the technologies in play and their potential ramifications.
- Uphold Ethical Standards: CSOs can ensure the validity and morality of their efforts by rigorously adhering to established ethical guidelines.
- Champion Transparency: Revealing methodologies, tools, and data sources can instil trust and promote inclusivity.
- Adopt an Open Science Model: Ensuring the reproducibility of results can foster accountability and clarity.

With its capacity to emulate reality, the ethics around Generative AI challenges traditional notions of fairness, data, and privacy. The National Statement on Ethics, while offering overarching directives, doesn't specifically address the unique ethical dilemmas posed by Generative AI. This absence of a comprehensive legal framework governing Generative AI poses significant risks to research credibility and raises intellectual property concerns.

While numerous AI guidelines exist today, they tend to be overly broad, unspecific, and incomplete. Deploying Generative AI in critical sectors, such as violent extremism mitigation, requires targeted training guidelines that provide clear ethical and legal directions. Addressing the inherent bias in AI systems, ensuring transparency in complex AI algorithms, understanding potential job displacements due to automation, and recognising the risks of increased tech-dependence that can lead to cyber vulnerabilities are all necessary. It is paramount for regulations to ensure that Generative AI reflects ethical standards, protects human rights, and is used responsibly.

Importantly, CSOs will need to ensure their personnel are well-versed with AI capabilities, ensuring they remain effective in this tech-driven age.

CONCLUSION

Generative AI has undeniably already revolutionised many sectors, and its potential for aiding civil society organisations (CSOs) in countering violent extremism (P/CVE) is enormous. However, navigating this transformative tool requires a judicious approach, with CSOs considering both its potential and pitfalls.

Looking ahead, the emerging field of prompt engineering hints at new avenues and careers centred on optimally utilising Generative AI. As technology evolves, future CSO professionals are likely to find themselves in roles that today's world hasn't even conceived.

As we pioneer the application of Generative AI in P/CVE, it's imperative for CSOs to acknowledge that Generative AI is now integral to their operations. It's vital to invest time in understanding and critically evaluating these tools. CSOs must continue to uphold their principles and values, discerning how to transpose them into this new digital realm. The focus should be on harnessing AI to augment, not replace, the invaluable human touch, ensuring that the fight against violent extremism remains both effective and ethical.

REFERENCES

Amnesty International (2022) Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya, Available from: https://www.amnesty.org/en/documents/ASA16/5933/2022/en/ (Accessed on 27 August 2023).

Barton, G., Vergani, M., & Wahid, Y. (Eds.). (2022). Countering violent and hateful extremism in Indonesia: Islam, gender and civil society. Springer Nature.
Bontridder, N. & Poullet, Y. (2021). The Role of Artifical Intelligence in Disinformation. Data & Policy, 3, e32.

Hao, Karen and Deepa Seetharaman. Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. July 24, 2023. The Wall Street Journal. Available at https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. Retrieved from: https://www.sciencedirect.com/science/article/pii/S266734522300024X

Sabzalieva, E. & Valentini, A. (2023) ChatGPT and Artificial Intelligence in Higher Education. Quick start guide. UNESCO Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000385146 (Accessed on 5 June 2023).
Siegel, D. & Bennett, M.D. (2023) Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda, Global Network on Extremism & Technology, Retrieved from: https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/ (Accessed on 5 June 2023).

Singh OP. Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. Indian J Psychiatry. 2023 Mar;65(3):297-298. doi: 10.4103/indianjpsychiatry.indianjpsychiatry_112_23. Epub 2023 Mar 3. PMID: 37204980; PMCID: PMC10187878.

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. Nature, 614(7947), 224-226.

SEAN-CSO

peace generation