Radicalisation Awareness Network
RAN Practitioners

Webpage: ec.europa.eu/ran

twitter | facebook | linkedin | youtube

RAN C&N

CONCLUSION PAPER
RAN C&N Working Group meeting
27-28 February 2024, Prague, Czech Republic

AI: Understanding and opportunities for P/CVE practitioners

Key outcomes

Artificial intelligence (AI) technologies are developing rapidly and becoming increasingly advanced. It is known that AI is being exploited by violent extremists, for example by using generative AI tools in order to create propaganda and develop personalised messages on a large scale, facilitating online recruitment efforts. However, AI can be used for positive purposes in preventing and countering violent extremism (P/CVE) as well, such as for detecting and preventing online hate speech, identifying and supporting vulnerable individuals by building resilience, and enhancing the effectiveness of practitioners' work.

On 27 and 28 February 2024, the RAN Communication and Narratives (RAN C&N) Working Group held a meeting on **'AI: Understanding and Opportunities for P/CVE practitioners'**. The main objective of this meeting was to obtain a deeper understanding of how AI could be used in a positive way within the realm of P/CVE and to identify practical opportunities to do so. The meeting brought together experts and first-line practitioners who have experience with integrating AI in both online and offline P/CVE work.

The meeting focused on how, within the sphere of P/CVE, AI can be employed in the best way possible. This means that human rights will be respected and ethical considerations will be part of daily usage of AI within P/CVE. Starting off with speaking about current possibilities of AI for current challenges in P/CVE, the group moved on to discuss wishes and possibilities for the future, including critical reflection about ethics and human rights.

Key outcomes of the meeting are:

- When further developing and implementing AI within P/CVE, important **ethical and human rights aspects** will have to be considered. For instance: the importance of taking into account bias, fairness of using data and quality of the data, and privacy and data security. Moreover, transparency and explainability will increase the trust in these tools and products.
- **Experiment** with the technology, while keeping the **"human in the loop"** and remaining critical towards the outcomes of this technology. This will help P/CVE practitioners to keep up with developments.
- **Funding** might be needed to develop specific tooling for the tasks of P/CVE practitioners in a rapidly evolving and expanding field of AI technology.

- Some of the current vulnerabilities within (online) P/CVE work could potentially be (partly) resolved by using AI tools. Wishes of the participants in this respect were, for instance: monitoring across multiple platforms, using multimodal AI tools (analysing multiple types of content, i.e. text and image, simultaneously), identifying the spread of extremist content in closed environments, and analysing data in real time (i.e. live broadcasts/streams with an immediate risk assessment).

- The most important **recommendations** made during the meeting were:

    - The different perspectives of practitioners, tech companies and policymakers should be combined and cross-sector cooperation is needed to deal with the challenges of AI in the realm of P/CVE.

    - As the policy and law-making process is not as fast and adaptable as AI technologies or the daily work of practitioners, a lot of effort should be directed at working based on ethics and privacy, aside from focusing on the current state of law.

    - Another important recommendation is to work towards "AI alphabetisation", for practitioners as well as the people they are working with. In P/CVE work as well as in digital literacy training, the aspects of AI should be incorporated.

    - Policymakers are advised to increase the level of inclusion of practitioners and tech partners in their decision-making processes.

    - For the tech sector, the call is to apply "transparency by design" in its AI development in the short term.

Highlights of the discussion

The meeting's structure enabled participants to initially obtain a comprehensive overview from three perspectives on current AI developments and examples of how AI tools can be used for good in P/CVE. After that, P/CVE challenges were identified that practitioners are currently facing in their work, and testing how generative AI tools could address these challenges. This enabled obtaining a deeper understanding of what is possible right now. On the second day of the meeting, participants got the chance to "dream big" and brainstorm about potential future scenarios. Group dialogue on ethics and human and legal rights enabled the identification of the most important boundaries when using AI in P/CVE work as well. This papers roughly follows the same structure as the meeting.

Practitioner, research and big tech perspectives on AI and P/CVE

In order to set the scene, three overarching perspectives were shared on the role that AI might play in P/CVE: the local safety perspective, the research perspective, and the big tech perspective.

Local safety perspective – Project in European cities

Democratic dialogue and debate are increasingly moving to the online sphere, which constitutes a very important infrastructure for today's democracy. However, harmful content is increasingly mainstreamed into public debates, and the presence of online hate speech is influencing democracy in a negative way. Therefore, societies need to think about how to create an inclusive and democratic online space, enabling safe and open dialogue for its users. Currently, a lot of hate speech and polarised and racialised conversations can be observed online, which influences

how citizens are using online platforms. Some users do not upload or take part in online conversations, or even stop using online platforms, because of the high level of harassment and the fear of being threatened.

In reaction to this development, a local network started to analyse digital communication in several European cities by using AI. The main objective is to detect and analyse hate in online discussions present in these cities. Through a trained algorithm, hate speech on relevant social media pages is tracked and analysed. A great advantage of using such an AI tool is that it can detect online hate very quickly and provides the individual responsible for moderating the page with suggested responses. The tool can monitor 24/7 and is continuously improving and adaptable to local contexts. One of the constant dilemmas while using this tool relates to definitional clarity. On the one hand, it is important to ensure that the algorithmic definition of hate speech is not too broad, as this may infringe on freedom of expression. On the other hand, a narrow and clear definition leads to a situation wherein comments that are considered as hate by some users are not always taken down.

In order to improve digital prevention, the information stemming from the tooling is shared with individuals working within the preventive sphere. Through the help of this AI analysis, it is determined where digital "street presence" is needed that fosters awareness among local users. Additionally, digital democracy supporters, consisting of civil society volunteers, are trained to minimise polarising debates online, and a digital security team is used to train prevention workers on what is happening online.

Research perspective – How AI can help in evaluating P/CVE programmes

Extremists are using AI for the development of propaganda, disseminating hate speech, using it as a recruitment tool and possibly even as operational means. One example is the use of certain AI tools that can provide instructions on how to produce bombs or other weapons. In addition, a crime–terror nexus is observable wherein AI is used for financial gain and scamming schemes. While extremists are finding their way in implementing AI to further their goals, there seems to be a lack of understanding among practitioners regarding the use of AI. Information is lacking about available tools they can implement in their day-to-day work and the ethics around the use of these technologies. An extra complication is the need for human oversight, since not all AI-generated information can be trusted.

When focusing on the evaluation of P/CVE programmes and testing their effectiveness, it is being observed that this is of great importance, but often a challenge. Intelligent conversation chatbots, such as OpenAI's ChatGPT and Microsoft's Copilot, can support such evaluative research. While these tools can help with certain steps, they are not fully ready to accompany practitioners in every step of evaluation. Some of the mentioned limitations were:

- Currently not all AI tools are able to translate information into other languages. The absence of certain languages poses a challenge, as practitioners are coming from different countries.
- The interdeterministic characteristics of AI tools means that different answers are provided to the same questions.
- The lack of credible sources makes it challenging to validate AI-generated information.
- The development of a tailor-made AI tool that perfectly aligns practitioners' needs in their everyday tasks is often difficult due to the lack of funding.

Within the P/CVE field, the use of AI seems to be mostly securitised, commonly considered as a technology exploited by extremists for harmful purposes rather than a technology that practitioners can benefit from in their work. Therefore, there is a need to de-securitise the usage of AI, facilitating its integration into the day-to-day practices of P/CVE work. Other important aspects to make AI useful for practitioners are the development of toolboxes for P/CVE in which different tools are combined, the development of training and capacity building. In all of this, a multidisciplinary approach is needed, combining different angles.

Big tech perspective – The use of AI on social media platforms

AI is not new, and people have already been working together with AI for quite some time; in 1956 the first "AI period" began, with computers solving problems like humans ([1]). Right now we are in a world of "AI 2.0", consisting of generative AI, with the ability to create new outputs such as text, art and music. Big tech platforms are now able to create and assess data on a massive scale.

As hate speech, deepfakes, misinformation and extremist content are not always directly observable, but sometimes concealed in humour or only understandable for those familiar with an extremist narrative, it is often challenging to detect certain content. Nevertheless, AI makes it possible to have a model that brings together multiple types of data and is continuously trained, which offers new possibilities. The benefit of AI lies in its multimodal understanding, facilitating a comprehensive understanding of complex content, grasping the bigger picture. This capability is particularly important as text and image, when analysed in isolation, may result in varying interpretations. AI offers the possibility to facilitate deeper understanding of the bigger picture, countering the presence of extremist networks online. Although many new possibilities are being offered, it is essential for both big tech companies and those using AI in P/CVE to consider the following as well:

- The commitment to responsible AI is of great importance, making sure that it does not impede on the rights of users. The use of AI also comes with the risk of accidently infringing on human rights, such as freedom of expression. This can be partly addressed through the use of protection mechanisms.
- A focus on reduction of the likelihood of abuse by being aware of how the model is fed (input filters) and what is filtered out of the data (output filters) is needed.
- Fairness and inclusion are important, by being transparent about the platform's policy and implementation behind used AI products/tools. Model cards that provide information about how the model works can be used to foster this transparency.
- The collaboration between partners and institutions is of great importance while innovating AI. This innovation should be based on the underlying principles of safety, responsibility and security. A relevant initiative in this respect is the AI Alliance, consisting of a partnership between companies, universities, research institutes, non-profit foundations and government organisations.

([1]) European Commission (2020): AI Watch Historical Evolution of Artificial Intelligence: Analysis of the three main paradigm shifts in AI (p. 7).

The potential use of AI in tackling current P/CVE challenges

After discussing the current state of affairs of the use of AI in the context of P/CVE from three different angles, participants discussed the biggest challenges they currently encounter in their work and explored how AI could help to reduce these challenges. Possible AI tools for four different P/CVE challenges were discussed and both their advantages and disadvantages for implementation were highlighted. An overarching limitation addressed by participants is the lack of funding, which all these tools need in both the testing phase and during the necessary further development of these tools while in use. A commonly mentioned advantage is the enhanced efficiency when using these tools and the increased impact in terms of P/CVE.

1. Knowledge drain

P/CVE Challenge: Losing valuable knowledge and skills due to the turnover of employees in P/CVE and the presence of segregated working environments within organisations.

AI Tool that might be of use: A knowledge-based chatbot that documents internal knowledge, enabling organisations to build upon previously acquired organisational knowledge. The knowledge base provides staff training through scenario building using large language models (LMMs) and generative videos as well the possibility to evaluate certain measures, since it is clear what has already been done and how it worked out within the data.

- Shows the evolution of internal knowledge on certain P/CVE topics
- Helps in addressing the specific target audience
- Helps to retrieve used definitions of concepts and prevents employees doing the same work in different projects

- Safety concerns due to the use of sensitive data
- The risk of biases when using LLMs, identification asks for additional knowledge and training
- The question of responsibility and maintenance of data hygiene
- Costs to develop an organisational knowledge base chatbot

2. The lack of shared language and tools

P/CVE Challenge: There is a lack of understanding across sectors/different "tribes" of practitioners. The translation and facilitation between different types of practitioners, for example tech companies, civil society, security services and the private sector, is often complicated due to different languages and multiple frameworks. Therefore, best practices and other resources are not being used to the fullest.

AI Tool that might be of use: A tool that facilitates communication between these different tribes, creating a form of mutual language, shares information and has tools in place that are not tailored to specific sectors.

- Reduces workload since individuals can more easily acquire and share their knowledge
- Increases multifaceted solutions, as a more diverse group of people is involved, whereas previously everyone worked more isolated
- Speeding up exchange and accelerating impact

- Could be exploited by extremists or other malicious actors as well, in order to connect with each other
- This AI tool itself does not have morals or ethics, but the way it needs to be used needs these

3. Track cross-platform extremist activity

P/CVE Challenge: The current inability to track cross-platform extremist activity. Therefore, practitioners are having great difficulty to identify and link online presence of these actors.

AI Tool that might be of use: An OSINT investigator that is able to connect the dots. Enabling practitioners to easily detect extremist activity online. This tool is able to identify certain slang and names used by extremist groups and individuals.

- Enhanced efficiency
- Increased leads and ability to respond to existing extremist threats
- Helps trust and safety works and fosters platform cooperation

- Practitioners need AI training in order to work with this tool and interpret data to identify potential biases
- Data overload
- Ethical and legal issues

4. Reintegration programmes for returnees

P/CVE Challenge: A current challenge is centred around reintegration programmes of returnees, and the stigmatisation upon return of these individuals back into society. Additionally, sometimes the safeguarding of human rights can be challenging when working in these programmes and there seems to be a lack of gender-sensitive approaches in this field, for instance a comprehensive perspective on how to reintegrate women and their children.

AI Tool that might be of use: A customer support AI tool that supports practitioners working in reintegration programmes. Such an AI tool will be able to create recommendations and strategies on integration. Including housing, finance, mental health care possibilities, and based and trained on the knowledge of previous reintegration programmes.

- Enhances time efficiency, enabling more time for human-to-human interaction and less workload for social services
- Can be designed in a user-friendly way facilitating use for a large population of practitioners
- Can make predictions of individuals' needs
- Facilitates connections with existing educational programmes
- It could work offline

- Questions on how to deal with privacy concerns as data from real returnees is necessary
- The AI tool could replace a part of human jobs, which has its ups and downs
- It may be challenging to maintain the distinction between AI recommendations versus letting AI make actual decisions instead of the practitioners
- Funding is needed, also for testing and updating the tool throughout its use

Future scenarios and ethical considerations

During the next session of the meeting, participants were asked to "dream big" and come up with project plans regarding the potential future use of AI in P/CVE efforts. Participants were asked how they would utilise AI in P/CVE if everything is possible, no limitations exist and all possibilities are on the table. This was done in order to encourage creative thinking and the exploration of more comprehensive ideas on the opportunities AI has for P/CVE.

The ideas coming out of this open brainstorming session were subsequently addressed from an ethical and human rights perspective. This was in order to identify the most important boundaries when using AI in P/CVE. Participants were encouraged to think about ethical and human rights considerations in three stages: 1) development, 2) deployment, and 3) launching/using the AI technology.

Key elements (wishes and opportunities) from the "dream big" brainstorm are:

- **Cross-platform** analysis and tracking of potentially harmful and/or extremist narratives. Identifying and monitoring "common" extremist and terrorist narratives, potentially using machine learning to identify the most harmful narratives or even predict what narratives will emerge in the future.

- Using **multimodal AI tools**: analysing text, image, audio and video simultaneously. Real-time monitoring and analysis on (social media) platforms could then be used to identify potentially dangerous situations before they take place (for example, analysing texts like manifestos and live video feeds to anticipate the livestreaming of a shooting and notifying authorities in an early stage).

- Building **chatbots** that can identify and gather users spreading extremist content in a closed environment.

- Providing practitioners with a **customisable P/CVE AI tool** to cater to their needs. Based on the needs of different practitioners, the AI tool can offer guidelines and recommendations and connect practitioners to relevant stakeholders.

- As for **data sources** and **AI models** used, multiple suggestions for innovations were made:

    - using live broadcasts to analyse sentiment/emotion;

    - using street view and geolocating to determine the location of a livestream;

    - combining social media/online data with "real-world" data such as open (government) registers;

    - using social media platform user reporting data (i.e. of alleged hate speech) to train AI models.

It is important to reiterate that the above ideas were formulated by the participants as part of an open brainstorming exercise. These ideas do not reflect any current or future existing projects of using AI and did not take ethics into consideration. The goal of this exercise was to first "dream big", and subsequently think about the ethical and human rights implications that these big ideas would incur. Key **ethical and human rights considerations** that were discussed in light of the ideas above are:

- Privacy and GDPR-related considerations include:

    - Who are the stakeholders/owners of an AI tool? And who are the users? In other words: who can access the data?

    - Do users have access to raw data, or only the analysed/interpreted/anonymised data? Giving access to all data improves transparency and enables human quality checks but would also mean giving access to more privacy-sensitive data.

- An AI tool predicting which narratives will have the highest risk of spiralling into violent extremism might lead to suppression of legitimate thought and criticism, hereby impairing freedom of speech.

- Using different data sources leads to the issue of data normalisation: how do you want to structure the data you use? The ethical consideration here relates to transparency: a unified approach would be preferred, but the original or raw data also needs to be requestable by other organisations. Automating the normalisation process using AI without retaining the raw data could lead to issues.

- AI is not ethical by default, so we should not expect AI to act ethically by itself. This implies it needs regulation and steering. However, this could prove to be difficult, as efforts for sensitising AI tools can also lead to a distortion or misrepresentation of reality. For example, requesting an image-generating AI to generate an image of a famous historical figure, combined with it being programmed to depict racially diverse people in its results, has led to false results ([2]).

- In relation to this, AI will always be biased based on the data it is trained on. An undesirable result could be that harmless content is being identified as terrorist content by AI (i.e. false positives), in combination with blindly trusting the judgement of AI while it is based on biased data (i.e. false confidence). This could lead to measures being taken against harmless individuals and infringing on the freedom of expression.

([2]) See for instance: https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

In short, important ethical and human rights aspects to consider when applying AI in P/CVE are: the importance of taking into account bias, fairness of using data and quality of the data, and privacy and data security. Moreover, transparency and explainability will increase the trust in these tools and products. A final observation was that in the near future, AI tools themselves might be able to help in creating new AI-based tools taking into account ethical and human rights aspects. Experimenting with this could be a promising follow-up.

Recommendations

After discussions on the potential usage of AI for P/CVE as described above, the meeting focused on gathering practical and tangible recommendations for practitioners and tech companies as well as policymakers. The following ideas were put forward:

- It is important to reiterate that **ethical and human rights considerations** were at the core of the discussions during the meeting.
    - A key insight that was discussed is that policy and law-making processes are not as fast as tech and as the daily work of practitioners are developing. Therefore, a lot of effort should be directed at working on the basis of ethics and privacy, aside from focusing on the current state of law. This means for instance thinking about "open source" versus "closed source" and anonymity versus user identity disclosure.
    - For tech companies to comply with ethical guidelines, these guidelines need to be well structured. There are currently no common ethical standards, the EU and the United States are doing things very differently here. And if tech companies are not provided with a clear ethical framework, they will have their own interpretation. Such a framework will have to be put in place by civil society organisations (CSOs) working together with governments. During the meeting, it was suggested to create a roadmap, a set of milestones on how to reach these standards over the coming years.

- Although some specific recommendations were made for practitioners, tech companies and policymakers, the most frequently given advice coming from the participants was that these different angles should be combined and **cross-sector cooperation is needed to deal with the challenges of AI in the realm of P/CVE**.
    - The sectors need to understand each other and speak to each other on a regular basis. Between them, there is a need to work towards definitions and classification of AI and reach a common understanding of terrorism/violent extremism. Also, sound oversight on and accountability of AI companies is a priority. Designing proper mechanisms for this that aren't easily evaded or have biases that can be used for violations of human rights is necessary. Practitioners and CSOs could be well placed to initiate cooperation from a "neutral" position and come up with suggestions around ethics, for example.

- Another important recommendation is to work towards "**AI alphabetisation**", for practitioners as well as the people they are working with. In P/CVE work as well as in digital literacy training, the aspects of AI should be incorporated, and the public should be made aware of the possibilities as well as the limitations of AI. For practitioners who want to employ AI in their daily practice (procedural, automation, knowledge management, etc.), a knowledge base on existing and available applications and on current and upcoming regulation of AI is crucial. Frontrunners in the field could provide such an overview to their colleagues, possibly combined with a framework for sharing good practices. For practitioners in general, the advice is to start using AI and experiment, to see what's out there and how you can work with it.

- Policymakers are advised to invest in **including practitioners and tech partners in their decision-making** processes. For the tech sector, the call is to apply "transparency by design" in its AI development in the short term. Another measure that could be taken by the tech sector is to ensure user authenticity combined with a level of public anonymity, to improve oversight on the use of AI tools.

Relevant practices

- **Safe Digital City** project by Nordic Safe Cities. The aim of the project is to give local professionals and municipalities a deeper understanding of the problems in their city and to give them new tools to strengthen their digital prevention work locally and to safeguard their residents from harmful content online — ultimately creating a safer digital democracy.

Follow-up

In general, the most tangible follow up to this meeting is that P/CVE practitioners will engage in employing AI in their daily work and experimenting with its possibilities while always being mindful of the possible biases and other negative aspects that might come with it.

Further exploration of the usage of AI for P/CVE and the pros and cons that come with it should be on the agenda of P/CVE networks, NGOs and think-tanks. As the technology will advance, new possibilities and downsides will present themselves and offer ample opportunities for experimentation and debate. Specifically, triangulating AI, ethics and human rights, and the P/CVE realm will be an important way of structuring the debate, building on the outcomes of this meeting.

Further reading

European Parliament (2023): Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI

Horiachko, A. (2023): NLP vs LLM: A Detailed Comparison Guide

Kasneci, E. et al. (2023): ChatGPT for good? On opportunities and challenges of large language models for education

Vaswani, A. et al. (2017): Attention Is All You Need