Radicalisation Awareness Network
RAN
Practitioners

Webpage: ec.europa.eu/ran

twitter | facebook | linkedin | youtube

CONCLUSION PAPER

RAN C&N Working Group meeting
16-17 April, Rotterdam, the Netherlands

Dealing with borderline content from the perspective of public trust

Key outcomes

When working (online) on preventing and countering violent extremism (P/CVE), the influence of online dynamics is omnipresent. One of the biggest challenges in safeguarding the web from extremism is the fact that borderline content is hard to tackle. As this content does not clearly violate laws or platform policies, it is not clear cut what to do with it. On the one hand, it can really add to creating a toxic online environment for those vulnerable to radicalisation. On the other, overreacting by deleting content or taking down accounts can be counterproductive as it undermines trust in tech platforms and institutions.

On 16 and 17 April 2024, the RAN C&N Working Group held a meeting on **'Dealing with borderline content from the perspective of public trust'**. The main goal of this meeting was to gather insights from practitioners, researchers and social media experts on how to approach and deal with borderline content from a bottom-up perspective, while maintaining a focus on protecting public trust.

During the meeting, participants discussed: what borderline content is and how it is used by extremists, especially through hate speech, memes and symbolism; how the spread of borderline content by extremists can foster radicalisation and how the responses by authorities and platforms can erode (public) trust; and how practitioners can address borderline content in a way that also contributes to the (re)building of public trust.

Key outcomes of the meeting are:

- **Collaboration between different stakeholders** is key for a holistic and coordinated response to borderline content. Researchers could work together more with practitioners, but authorities and tech companies should also be involved.
- A **new legal approach and international cooperation** is imperative in dealing with borderline content. Dealing with borderline content is already complex, but even more so due to local differences in what is legal and illegal. International cooperation is needed to work towards a unified approach.
- **Invest in different approaches** when addressing borderline content and engaging in alternative- or counter-speech efforts. Examples from the participants range from making use of gamification to countering "troll bots" with "love bots", hate with love, while using humour.
- **Having a backup plan** before engaging with borderline content is important in mitigating the risk of a backlash. Communication is key in this. Consider beforehand what your line of communication is in case your efforts go sideways, for example if you inadvertently become targeted by the groups you are trying to address.

- More attention should go to **implementing security parameters** to ensure safety of both moderators addressing borderline content as well as platforms themselves, for example to protect them against cyberattacks.
- While **artificial intelligence (AI)** offers opportunities to respond to borderline content (for example, in automated content moderation), a lot of limitations exist in the ability of AI tools to respond effectively. Therefore, it is recommended to always keep a human in the loop when using AI to respond to borderline content in a P/CVE context.

This paper first covers the highlights of the discussions and presentations that were given during the meeting, followed by an outline of the case study action plans that the participants worked on during the meeting. Then, the main conclusions and insights from the meeting are formulated in a "do's and don'ts" overview as well as a list of general recommendations.

Highlights of the discussion

A Setting the scene

The meeting was kicked off by three presentations from practice (viewpoint of an NGO), research (the perspective of a researcher on content moderation) and a social media/tech company.

The NGO perspective focused on a (case) study of online channels about Islam, and the prevalence of jihadist borderline content there, in a European country. It was noted that most online channels about Islam in the country are spreading non-violent messages and mostly rejecting jihadi organisations. Some channels seem to function as a gateway into an information bubble, with an echo chamber effect. The information in this bubble refers a lot to a radical rejection of democracy, the West, other religions, mainstream media, science, reforms and LGBTQ. It also links to Islamic content in the English language a lot, as opposed to channels in the country's local language. The same actor within this realm can appear in different styles and settings, dependent on the target audience and topic. Based on what is being seen online, a set of elements become clear of what could indicate the deliberate spread of borderline content: 1) Messenger/Channels (biography, core messages, network); 2) Language & Style (toxic, hate speech, autocratic, polarisation, etc.); 3) Persuasive Methods (propaganda, populist or demagogic elements); 4) Messages & Narratives (analysis of a crisis – diagnosis/guild – one single solution – call to action, conspiracy narratives); 5) Iconography & Symbols (in this case jihadistic/Islamistic, anti-Semitic, anti-democratic, anti-modernism, pro-caliphate, -ummah, etc.); 6) Visions & Ideals; and 7) Campaigns. Based on these characteristics, NGOs and researchers can reveal, in a transparent way, the workings of the spreaders of borderline content and raise awareness within government and the society as a whole.

A researcher spoke about the challenges and possibilities of using AI-powered content moderation in dealing with borderline content. Social media companies are now leveraging the power of AI, aiding in the challenge of identifying problematic content quickly, accurately, based on limited information and at scale. This is leading to the assessment of situations without any contextualisation. And as language is continuously evolving, this also leads to challenges regarding flexibility and adaptability of these content moderation tools. In an article ([1]) it is suggested that, to date, the evidence about the effectiveness of AI tools remains partial, but also that their use raises several important questions about the reproducibility of results, about privacy and copyright issues, and about the primacy of the English language. Several studies ([2]) have shown that the use of content moderation had an adverse effect on those with legitimate voices and is leading to discrimination against marginalised voices. It is clear that human oversight is still needed when it comes to detection and decision-making regarding borderline content.

([1]) Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023, October 4). *ChatGPT for text annotation? Mind the hype!* https://doi.org/10.31235/osf.io/x58kn

([2]) For instance: Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

A representative from a social media company explained that the company's approach to issues around content moderation is first of all to have strong and transparent policies. Those community standards are accessible to everyone and are constantly evolving, making them as comprehensive, detailed and so on as possible.

- When it comes to content moderation, having definitional clarity is absolutely essential; if you want to operationalise this to AI classifiers, you need to be able to train them on a data set where everything is labelled accurately. Your goal is to have maximum accuracy and to reduce over-enforcement/false positives; especially since the amount of extremist content on the platforms is really low. 99.8 % Of the content is ok, and we should not limit people in connecting and in building communities.

- Dealing with borderline content starts with understanding the space and how the platforms are misused. It is extremely difficult to approach borderline content from a technology perspective. It starts with the understanding of the problem, understanding the space, and how actors are using technology, messages, mediums, etc. This is the understand phase, before the actual policy development starts. After that, the key point is consulting with the company's stakeholders – experts, civil society organisations, researchers, marketing teams, officials, etc. Asking them questions like "Are we on the right track?", "What are the flags?", etc.

- This is important, because the borderlines are really thin and blurry. For instance, in pop culture and music, if someone is using images from a drug lord, glamourising this lifestyle, how do we treat that content? Do we want to enforce limits on this content, as it might inspire someone into that lifestyle? There are artists who are aligned with criminal groups, which they use to build an image for themselves. Is this artistic freedom or not? On the platforms there are also individuals, not part of any extremist group, who enjoy developing and spreading content that is ambiguous. They seem to be part of some kind of a contest about "who is going to create the most funny/outrageous image". How does that fit within the traditional understanding of the violent extremist landscape? Other examples of the difficulty in distinguishing borderline content are for instance AI-generated subliminal images, influencers who use humour and sarcasm, pop culture and music (glamourising criminal lifestyle), and conspiratorial groups (freedom of opinion and where to draw the line).

- There is a clear difference between borderline content and straightforward extremist propaganda. Extremist propaganda is owned and pushed by an organisation, with a vision and agenda, there is a claim of responsibility. The intention is to persuade people to join the movement and recognise their ideas. Borderline content can be seen as part of "cultural warfare"; it doesn't directly radicalise you but tweaks your mindset and opens you up to what is more extreme. It's not about content control and content removal, it's about behaviour and not on an individual level only, it's about networks, networks of networks. If we want to make progress, we need to go beyond moderation and understand the complex strategies and interrelated behaviours.

B Case study action plans and red/blue teaming

After the scene setting panel, two practitioners presented a case study of borderline content they faced in their day-to-day work. These presentations served as an inspiration, as participants needed to develop their own case studies of borderline content and develop a solution action plan, aimed at providing practitioners solutions for the developed case study. On the second day of the meeting, participants took part in a red team/blue team exercise. During the first round of this exercise, teams were stepping into the shoes of a borderline content spreader (red team) thinking about strategies to "undermine" the other group's action plan. In the second round, teams switched back to the perspective of a P/CVE practitioner (blue team) and brainstormed about how to respond to the opposing team's ideas. This exercise form serves to think critically about your ideas and gain perspectives you did not think about before.

Please note that the "red team" and "blue team" ideas presented below are entirely hypothetical and were developed during a brainstorming session. They are intended to illustrate potential ideas and do not represent current actions or real recommendations.

Case 1 - Protest slogans

The use and spread of protest slogans that are ambiguous in their meaning and considered as borderline content. It is important to note that certain protest slogans are considered illegal in some countries and legal in others.
Action plan:

1. **Structure public discussion** on the local level, fostering dialogue in, for example, schools, communities, religious institutions, about the question why individuals use a certain protest slogan.
2. Besides the presence of content moderation, users need to have the **opportunity to explain** themselves, as the spread of borderline content (protest slogan in this case) may be not intended to be harmful.
3. Inform users through **referrals to information pages** about why a certain protest slogan can be considered as borderline content.
4. **Human in the loop** is of great importance, given the potential for a protest slogan to be interpreted differently in various contexts, and relying solely on an AI tool may lead to biases.

Red team

- Strategically **spread disinformation** during public discussions to foster mobilisation and recruitment efforts, while amplifying the narrative of oppression, infringement on freedom of expression and double standards imposed by the government
- **Use coded language, humour and non-English discourse** to undermine content moderation efforts
- Avoid detection by sharing slogans through **non-text-based media**
- Exploit the increased attention for a **counter-messaging campaign**
- **Challenge content moderation** by creating an overload of content through automated messages or bot accounts

Blue team

- Proactively minimise the spread of disinformation by **carefully managing attendance** to public discussions
- **Train individuals and professionals**, such as teachers, to lead discussions in a constructive manner and develop discussion guidelines
- Focus on diversity and inclusion, by involving **different perspectives**, making sure there are no double standards
- The use of **ex ante content moderation**, meaning reviewing content before it is made available to the public

Case 2 - Polarising content posts from influencers

Influencers exploiting their online popularity in order to fuel existing societal tensions. For example, the scapegoating of minority groups based on the spread of disinformation.
Action plan:

1. Organise **community counter events** that foster community contact and build social cohesion through a form of positive dialogue between present communities, with different perspectives. Community leaders play a crucial role here.
2. Develop **online alternative narratives** through the use of humour that indirectly shows that the initial narrative is wrong.
3. The **creation of preventive content** in order to address societal issues before they become problematic, aimed at educating and informing individuals to prevent them from spreading this harmful content or narrative.
4. The **monitoring of social media posts** that fuel societal tensions.

Red team	Blue team
• **Infiltrate community counter events** in order to undermine the development of social cohesion • **Reinforce the idea of us versus them**, by focusing on the government or elite as protectors of a misconception and the dissemination of fake news • **Discredit community leaders and institutions** that develop preventive measures by spreading harmful rumours within the online realm • **Make use of bots** in order to disturb the online monitoring of content	• Make **use of humour** in order to counter online harmful content • Development of a moderation tool that **flags fake news** • Use influencers to disseminate positive alternative narratives that **debunk an "us versus them" narrative** • **Implement "love bots"** that subtly challenge borderline content through AI-generated human-like positive conversations • **Report** instances where community leaders and institutions are discredited to officials

Case 3 - Far-right extremist video game

Individuals playing a far-right extremist video game, wherein controversial European "heroes" are fighting against representatives of "woke" culture, often consisting of minority groups present in several societies.

Action plan:

1. Infiltrate gaming groups to **disseminate counter- or alternative narratives**.
2. Development of a more **comprehensive understanding** of these groups, by decoding and documenting activities and behaviour on these platforms, serving as input for research.
3. **Reporting trends** to practitioners and other actors within the field.

Red team	Blue team
• Use **brigading** as a tactic consisting of mass posting and commenting aimed at harassing or silencing the adversarial community • **Cyber swarming:** using bots to spam comment sections to sabotage the spread of alternative and counter-narratives • **Doxing** of individuals disseminating these counter- or alternative narratives • **Swatting:** law enforcement actions against actors involved within action plan under false pretences • Using **politicians**, social media **influencers**, media and other powerful individuals to normalise these far-right extremist narratives	• Creating an environment that **hides the identity** of actors spreading alternative and counter-narratives to prevent doxing and other tactics • Avoid swatting through the involvement of **police forces** in their network • Use of **strategic public relations** campaigns enhancing credibility and improve awareness • **Creating a network and community** of knowledge to effectively monitor and prevent escalation

Case 4 - Ambiguous and contradictory borderline content

Online content that lies at the intersection of historical discussion and the glorification of perpetrators involved in these events, making it complex to determine whether content violates community guidelines.
Action plan:

1. The implementation of **moderation and regulation** based on transparency principles to change the platform culture and leverage AI. For example, for auto moderation.
2. **Educational programmes** involving users, educational institutions and other stakeholders driven by common platform themes that resonate with the target audience.
3. **Offering offline support** users can voluntarily engage with.

Red team

- Moving to **alternative platforms or duplicate channels**
- **Hacking platforms** implementing own content moderation tools
- **Threatening content moderation** teams and frame moderators' presence as undesirable
- **Disseminating fake information** to offline support initiatives
- **Copying language and formats** used within educational programmes to spread extremist narratives

Blue team

- **Ensuring privacy protection** and mental health support for practitioners involved
- **Explore manipulative tactics** used by platform users, **inoculation games** to build resilience and use **trolling in a positive way**
- **Protection against hacking** by putting parameters in place
- **Specific your target** audience by focusing on users who are "moderate" and "influenceable" rather than extreme
- **Involve regional representatives** to support offline support initiatives

Do's, don'ts and recommendations

The meeting proved to be very insightful for participants, who were asked to formulate do's and don'ts and their key takeaways at the end of the meeting. This section builds on the insights of the participants to draw conclusions on do's and don'ts and recommendations for different stakeholders.

Do's and don'ts

Based on the experiences of the participants during the meeting, a list of do's and don'ts can be formulated. These are general do's and don'ts to keep in mind when working with borderline content. They are not necessarily focused on a certain type of practitioner or stakeholder.

Do’s

- ✓ Connect with other organisations. Create a network of stakeholders and involve government/authorities, researchers and practitioners.
- ✓ Use agile mitigation strategies, rather than just focusing on removal. Removal can backfire and is tricky to enforce consistently with this type of content, so it should be treated as a last resort.
- ✓ Ensure protection of researchers, practitioners and moderators who work on platforms where borderline content is spread. Those who face risks of being doxed, threatened, etc.
- ✓ Be aware of the context of borderline content, and that some responses can be counterproductive.
- ✓ Determine the legal boundaries.
- ✓ Use humour and love to counter hate.
- ✓ Use available technology to your advantage.
- ✓ Debunk false facts.
- ✓ Address the silent majority.
- ✓ Be transparent about moderation decisions.

Don’ts

- ✗ Don’t stifle debate and free speech.
- ✗ Don’t be insensitive to target audience, context and platform culture. Each case is unique, so there is no general solution.
- ✗ Don’t unintentionally amplify the problematic content.
- ✗ Don’t unintentionally stigmatise individuals/groups.
- ✗ Don’t attack the ideology, but the consequences of it.
- ✗ Don’t expose people to harm, or the identities of your team.
- ✗ Don’t use divisive narratives.
- ✗ Don’t stay silent.
- ✗ Don’t underestimate the effects of certain borderline content.
- ✗ Don't feed the trolls.

General insights and recommendations

Besides the do's and don'ts formulated above, there are some general insights and recommendations that came out of the meeting.

Understanding the problem

A first set of insights and recommendations regards **understanding the problem**:

- Involve the **right stakeholders** and **be critical** on what does and does not fall within your responsibilities as a P/CVE practitioner. Not everything needs to be approached from a P/CVE perspective. This is especially relevant in dealing with borderline content — which is not illegal content. Be aware that the risks of dealing with it from a P/CVE perspective are possible over-reacting and over-securitising.

- **There is no universal solution.** Borderline content is fluid and always needs to be put into context. For example, the definition of borderline can change a lot depending on whether we are looking at it from a legal standpoint or from a content moderation perspective.

- Moreover, our own **identity and personal beliefs** also shape the perception of what is considered borderline and what is not. Practitioners, therefore, should be careful about this, as there is a risk of polarising even more. The element of general public trust is also relevant here: acting from your own biased perspective without proper consideration can result in a backlash.

- Related to this, making sure there is a **proportional response** that doesn't inflame extremist beliefs is highly important.

- Therefore, **understanding the context** is key. Different communities, for example, use different coded language when spreading borderline content. Moreover, the intent and goals behind the spread of borderline content also impact the appropriate response to it.

- Lastly, it is important to always keep in mind that you are mostly dealing with **vulnerable audiences** when addressing borderline content.

Solution-oriented

A second set of insights and recommendations is more **solution-oriented**:

- **Collaboration** between different stakeholders is key for a holistic and coordinated response to borderline content. Researchers could work together more with practitioners, but authorities and tech companies should also be involved.

- Being **transparent** and approaching borderline content from the root up is generally considered to be a good approach.

- A **new legal approach** and international cooperation is imperative in dealing with borderline content. Dealing with borderline content is already complex, but even more so due to local differences in what is legal and illegal. International cooperation is needed in order to work towards a unified approach.

- **Try different approaches** when addressing borderline content and engaging in alternative- or counter-speech efforts. Examples from the participants range from making use of gamification to countering "troll bots" with "love bots", hate with love, while using humour.

- **Implement security parameters** in order to ensure safety of both moderators addressing borderline content as well as platforms themselves, for example to protect them against cyberattacks.

- **Having a backup plan** before engaging with borderline content is important in mitigating the risk of a backlash. Having a communications strategy is key here. Discuss beforehand what your line of communication is in case your efforts go sideways, for example if you inadvertently become targeted by the groups you are trying to address.

- While AI offers opportunities to respond to borderline content (for example, in automated content moderation), a lot of limitations exist in the ability of AI tools to respond effectively. Therefore, it is advised to always keep a **human in the loop** when using AI in a P/CVE context.

- Using **offline initiatives to address online borderline content** can be effective. An example given was a campaign that proves indirectly that certain narrative and related content is false and organises community gatherings to foster dialogue between different individuals referred to as "Meeting the hate with love and cake".

Possible follow-up

The debate around dealing with borderline content will be ongoing over the coming years. This is also a relevant debate in the realm of P/CVE. Maybe the suggestion of defining a **new international legal approach** to dealing with borderline content could be a suitable topic for a follow-up to this meeting.

Further reading

Farrand, B. (2023). 'Is this a hate speech?' The difficulty in combating radicalisation in coded communications on social media platforms. *European Journal on Criminal Policy and Research, 29*, 477-493. https://doi.org/10.1007/s10610-023-09543-z

Macdonald, S., & Vaughan, K. (2023). Moderating borderline content while respecting fundamental values. *Policy & Internet*, 1-15. https://doi.org/10.1002/poi3.376

Mattheis, A. A., & Kingdon, A. (2023). Moderating manipulation: Demystifying extremist tactics for gaming the (regulatory) system. *Policy & Internet, 15*(4), 478-497. https://doi.org/10.1002/poi3.381

Ocampo, N. B., Sviridova, E., Cabrio, E., & Villata, S. (2023). An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1997-2013). Association for Computational Linguistics. https://aclanthology.org/2023.eacl-main.147

Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023, October 4). *ChatGPT for text annotation? Mind the hype!* https://doi.org/10.31235/osf.io/x58kn

Saltman, E., & Hunt, M. (2023). *Borderline content: Understanding the gray zone*. Global Internet Forum to Counter Terrorism (GIFCT). https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf

Weng, L., Goel, V., & Vallone, A. (2023, August 15). *Using GPT-4 for content moderation*. OpenAI. https://openai.com/index/using-gpt-4-for-content-moderation/

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review, 10*(2). https://doi.org/10.14763/2021.2.1565

Product of the Radicalisation Awareness Network **(RAN)**,
Authored by RAN Staff **Rik Scheele**, **Alwyn Voogd** and **Alek Berends**,
after consultation with the RAN C&N Co-chairs **Fabian Wichmann** and **Ross Frenett**.