SPECIAL PROCEDURES
UNITED NATIONS
HUMAN RIGHTS COUNCIL

Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Ben Saul

Inter-Parliamentary Union High Level Group on Countering Terrorism and Violent Extremism

152nd IPU Assembly, Istanbul, 15 April 2026

Item 6 – Online radicalization and digital platforms: Emerging threats, counter-terrorism challenges and parliamentary responses while protecting human rights

Excellencies, distinguished participants

I am pleased to join you from Sydney. My mandate as United Nations Special Rapporteur recently published a Position Paper[1] warning of the many human rights risks of using artificial intelligence in countering terrorism, and making recommendations to prevent them.

One major issue it addresses is the use of AI in online platforms. **Some social media and gaming platforms have been misused by terrorists and violent extremists** to incite violence and disseminate incendiary messages which threaten the human rights of others. The anonymity and decentralized architecture of online platforms can amplify terrorist content, while weakening preventive and remedial measures.

Generative AI can enable malicious actors to automate content production, creating sophisticated propaganda materials that adapt messaging strategies to target specific human vulnerabilities and cultural sensitivities across populations. It can also facilitate hyper-personalized radicalisation pathways through algorithmic analysis of individual behavioural patterns, social media interactions, and demographic characteristics.

While AI systems can assist States and technology companies to counter terrorism and violent extremism online, their use can also seriously endanger human rights. Automated content moderation systems can assist in detecting explicit terrorist propaganda, but they are ill-suited to detecting and moderating implicit terrorist or violent extremist content, namely content that "conveys harmful messages through coded language, irony, humour, or cultural references".[2] Automated systems tend to over-moderate political expression in sensitive contexts, including conflict situations, or fail to remove harmful content. In some instances, algorithms have even amplified harmful content to maximise user engagement. The problems

[1] https://www.ohchr.org/en/documents/position-papers/protecting-human-rights-while-using-artificial-intelligence-counter.
[2] Bibi van Ginkel, Yael Boerma and Tanya Mehra, "Reading Between the Lines: The Importance of Human Moderators for Online Implicit Extremist Content Moderation", International Centre for Counter-Terrorism, December 2025.

of false positives and false negatives are aggravated by moderation in multilingual, under-resourced contexts where training data is sparse and moderation systems are underdeveloped. There is also a tendency to automatically remove content, instead of taking a proportionate approach, such as by "shadow-banning".

Excessive content moderation risks fuelling social division and alienation from democratic political processes, rendering populations more vulnerable to terrorist narratives. Interventions must focus not merely on suppressing harmful content but on promoting pluralism of political participation and view-points and the free flow of public interest information.

Where online platforms are regulated by State authorities, vague definitions and criteria may not give platforms sufficient guidance and can allow private actors too much discretion. While human moderation is essential for interpreting context, intent, and nuance, staff and resources devoted to moderation have been scaled down by major online platforms since 2022, including in pursuit of maximalist "free speech", and investment in AI moderation has expanded, along with reliance on unreliable "community" or crowd-sourced approaches.

States must regulate online content moderation in accordance with human rights law, particularly articles 19(3) and 20(2) of the International Covenant on Civil and Political Rights. They must ensure that expression is restricted, whether by States authorities or private actors, only in accordance with the requirements of legality (thus requiring precision in the legal criteria for moderation), necessity, proportionality and legitimate aim, consistent with the right to freedom of expression, the prohibition on hate speech (including the Rabat Plan of Action against hate speech), and the recommendations on defining incitement to terrorism of my mandate.[3]

Further, States must ensure that technology companies undertake and disclose human rights due diligence assessments, with heightened obligations in conflict and high-risk situations. Companies should also provide adequate, well-trained, unbiased human content moderation and human oversight of AI content moderation; ensure algorithmic transparency; monitor the effectiveness of mitigating actions; publish their moderation policies and detailed moderation reports; maintain adequate and well-trained in-house human rights staff and multilingual and culturally diverse expertise; and provide effective internal appeal mechanisms.

Other uses of AI online by State authorities can jeopardize human rights. Firstly, AI systems risk intensifying unjustified mass surveillance online, enabling authorities to target political dissent, human rights defenders, minorities and journalists. Such uses violate rights to privacy and liberty, and civil and political freedoms, and ultimately undermine democracy. AI algorithms can aggregate and analyse highly personal or sensitive data to profile the alleged terrorist risks of individuals and groups and prompt invasive security measures.

Secondly, AI-powered tools can be used by the authorities and technology companies to deliberately censor, block or "take down" legitimate online expression, and identify individual online accounts for restriction, suspension or disabling.

[3] A/HRC/16/51, paras. 30-32; A/HRC/40/52, para. 37.

Thirdly, authorities can deploy AI-generated misinformation, disinformation, "deepfakes" and deceptive counter-narratives to undermine and neutralise legitimate expression, association, peaceful assembly and civic and political participation, including during elections.

In the experience of United Nations human rights mechanisms, the use of new technologies in counter-terrorism has unleashed new waves of human rights violations targeting civil society. The uses of AI in counter-terrorism contexts have the potential to violate numerous rights, including privacy and data protection; equality and non-discrimination; freedoms of expression, association, assembly, religion; political participation; liberty; security of person; life; freedom of movement and non-refoulement; and the rights of vulnerable, marginalised and disadvantaged groups. The work and safety of human rights defenders is also at risk.

AI systems are far from infallible. Their datasets and algorithms may be biased, unrepresentative, illegally sourced or inaccurate, potentially leading to discriminatory profiling, violations of the right to privacy, and counter-productive false positives and failures to detect risks. AI systems may lack adequate human control and analysis, resulting in over-reliance on automated decision-making lacking in nuance, context and responsiveness to human rights considerations. Many AI systems are still in developmental, experimental or testing phases, and their capabilities, including in real-world conditions, remain to be proved.

The General Assembly and Human Rights Council have affirmed that AI systems must be developed and used in ways that respect, protect, and promote human rights. Contemporary global processes for inclusive, transparent, and accountable AI governance must address the particular risks in counter-terrorism and security contexts. Laissez-faire industry self-regulation is not sufficient.

My Position Paper makes other recommendations for the responsible use of AI across the spectrum of counter-terrorism measures, from border security to military operations. Essential elements of adequate legal frameworks to regulate AI systems to protect human rights in counter-terrorism include:

- Clear and explicit legal regulation:
    - No blanket exceptions for AI systems used in security, military, law enforcement, and border management purposes;
    - Prohibitions on the use or transfer of AI systems that cannot meet human rights standards or pose an unacceptable risk;
    - Heightened regulation of permissible AI systems used in security applications;
- Thorough human rights due diligence assessments by developers and deployers of AI systems throughout their lifecycle, with effective penalties for non-compliance;
- Stringent safeguards on data quality, testing and validation;
- Ensuring AI systems are transparent and explainable;
- Rigorous transfer and export controls;
- Effective and meaningful human control in decision-making; and
- Personal data protection and data security, including safeguards on cross-border information sharing.

AI systems must also be subject to effective oversight by internal and independent bodies, to monitor their use, investigate complaints, and ensure compliance with the law and human rights. Independent bodies must be empowered to evaluate training data quality, machine learning model performance, and potential discrimination.

Finally, there must be effective and accessible accountability mechanisms to provide remedies to victims of rights violations and other illegal acts. Procedures must ensure due process, including sufficient disclosure concerning algorithms, data sources and potential biases.

Thank you.