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

---

From Principles to Practice: Special Dialogue on Artificial Intelligence and Preventing and Countering Violent Extremism

Co-hosted by UNOCT and the Republic of Korea

12 February 2026

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 issue it addresses is the use of AI in regulating online platforms, which is particularly relevant to today's topic. 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.

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

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

My Position Paper focuses particularly on the moderation of online content. While automated content moderation systems can often detect explicit terrorist propaganda, 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 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".

---

[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.

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.

All of this is additional to the constraints that should apply to all uses of AI, including transparency and explainability; data quality, testing and validation; data protection and security; and fair and effective external oversight, accountability and remedies.

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.

Thank you.

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