

The Challenge and the Potential Promise of AI in Preventing and Countering Violent Extremism

Alistair Millar

3 September 2026

A range of experts, including those with [tech](#) and [security policy](#) backgrounds and [historians](#), have all recently sounded the alarm bell about the need for regulatory measures to address the potential harms of Artificial Intelligence (AI) “[before it’s too late](#).” One of those harms lies at the intersection of AI and violent extremism. While there is a growing body of [research](#) on the implications of AI when preventing and countering violent extremism (P/CVE), significant work remains to be done in two areas: understanding and countering how extremists exploit AI and developing safeguards that allow prevention practitioners to use AI responsibly.

The Threat

Violent extremist groups have proven rapid and creative adopters of AI capabilities. Extremist recruitment can now be directed at individuals vulnerable to radicalization through behavioral profiling, which uses machine learning to detect data patterns associated with psychological susceptibility. During the early 2010s, Islamic State recruiters pioneered a labor-intensive [version](#) of this targeting by identifying isolated, grievance-laden young men through their social media activity and engaging them in one-to-one conversations designed to exploit specific vulnerabilities.

Two years ago, researchers at the Global Network on Extremism and Technology [documented](#) AI-generated content circulating across dozens of platforms at once, with each version subtly adapted to the cultural and linguistic context of its intended audience.

White nationalist networks have [used generative image tools](#) to create large volumes of memes, recruitment posters, and fabricated news graphics. These materials can look professionally designed even when produced by individuals with no design skills or organizational support.

AI-powered “[recommendation systems](#)” are designed to [maximize](#) user engagement by serving content that provokes strong emotional responses. In practice, these systems can push users toward increasingly extreme material by repeatedly rewarding content that captures attention, outrage, or grievance. A 2019 [investigation](#) by researchers at the University of Pennsylvania described a “rabbit hole” dynamic on YouTube and [on TikTok](#). Users who began with mainstream political content were gradually served more extreme material by the platforms’ recommendation algorithms, including by offering content from incel and white nationalist communities.

Seven years ago in New Zealand the [Christchurch](#) shooter’s [manifesto](#) explicitly framed the attack as online content engineered for algorithmic amplification. This [gamification](#) of mass violence was intended to generate the engagement metrics needed to spread the message farther than any conventional propaganda campaign could.

AI has become automated and vastly increased in scale. Using the same tools advertisers use to target consumers, contemporary extremist actors [can use commercially available data analytics](#) to identify users whose behavioral data suggests isolation, economic frustration, or a desire for belonging, then deliver recruitment content calibrated to those emotional states.

The United Nations is [struggling](#) to implement a unified approach to this issue: its Office of Counter-Terrorism has [identified](#) the growing volume of AI-generated content as a structural problem with consequences that extend beyond any individual extremist group. These conditions give extremist narratives more room to gain traction because they often offer simple explanations for complex grievances. The problem, then, is twofold: extremists are using AI directly, and AI is also reshaping the broader information environment in ways that make it more receptive to extremist messaging.

AI as Part of the Prevention of and Resilience to Violent Extremism

The same technologies that amplify the threat also offer [tools for addressing it](#). AI can support prevention efforts in several ways. Machine learning models trained on large datasets of known extremist material can help detect new content before it spreads widely. Sentiment analysis can also flag online conversations [that show](#) emotional [signatures](#) associated with radicalization, such as grievance, isolation, and the search for belonging or purpose. Used responsibly, these signals may help prevention practitioners intervene before radicalization progresses toward violence.

Some proven [AI-driven counter-narrative initiatives](#) offer significant potential by delivering tailored messaging to individuals [identified as at risk](#), connecting them directly with constructive alternatives to extremist content. For example, in Alberta, Canada, they are deploying [Unison](#), a public safety deployment technology to help ensure the “right resource is at the right place at the right time” at the community level.

The Human Rights Implications

The civil liberties implications of these applications deserve emphasis that the enthusiasm for AI-driven prevention sometimes obscures. Mass behavioral surveillance is not a neutral tool even when directed toward prevention rather than prosecution. The same profiling systems that identify vulnerable individuals for intervention can be [turned](#) toward the surveillance of political dissent, the targeting of minority communities, and the suppression of legitimate expression. Decades of counterterrorism policy and practice provide [abundant evidence](#) that prevention powers designed for one purpose will be applied to others when political conditions change. AI-powered P/CVE, precisely because of its power and reach, must be subject to robust human rights safeguards, democratic oversight, and meaningful accountability mechanisms which are requirements that few existing frameworks adequately address.

Managing AI’s Dual-Use Challenge

AI presents a profound dual-use challenge for P/CVE. The same capabilities that extremist movements can exploit for propaganda, recruitment, and evasion [can also support prevention](#), resilience, and response. The task is not to reject AI outright or treat it just as a technical issue. The key will be to govern its use carefully so that its

benefits do not come at the expense of rights, trust, or democratic accountability. This effort will require coordinated action among governments, civil society, technology companies, and nongovernmental experts, supported by durable safeguards and sustained responsible oversight so that AI is a tool to prevent, rather than stoke political violence.

[Alistair Millar](#) is President of the Fourth Freedom Forum. He has taught graduate courses, written extensively on preventing and countering violent extremism, and helped more than a dozen countries develop national strategies to prevent and counter violent extremism.