

Emerging
Trends
in PCVE

THE HUB. INSIGHTS

JUNE 2026



European
Commission



EU Knowledge Hub on
Prevention of Radicalisation



THE TRADE-OFF: ALGORITHMIC RADICALISATION FOR ECONOMIC GAINS



“From a preventing violent extremism (PVE) lens, treating algorithmic radicalisation as a trade-off for economic gains could undermine the core principles of early intervention, resilience-building, and structural prevention that define effective counter-extremism strategies.”

Algorithmic radicalisation—the subtle steering of digital citizens toward increasingly polarising or extreme content through engagement-driven systems—is a structural and predictable contributor of radicalisation. While platforms do not officially endorse algorithmic radicalisation as a trade-off—their (in)actions might effectively normalise it. From a preventing violent extremism (PVE) lens, treating algorithmic radicalisation as a trade-off for economic gains could undermine the core principles of early intervention, resilience-building, and structural prevention that define effective counter-extremism strategies. This insight argues that platforms might be aware of the dangers of radicalisation but deprioritise safety to protect profits. Finally, the article provides recommendations to address algorithmically driven pathways to radicalisation.

Engineered Engagement and (Algorithmic) Radicalisation

Most platforms are economically incentivised to be addictive and their algorithms are engineered for engagement. Researchers observed, YouTube’s recommender systems “paint an alarming picture of online radicalisation: not only Incel activity is increasing over time, but platforms may also play an active role in steering [digital citizens] towards such extreme content.”



Researchers also indicate that, [TikTok's recommender system](#) is culpable of “pushing increasingly extreme content into the feed of new [digital citizens] after they've used the app for just 10 minutes.” These examples reflect a structural pattern consistent with the definition of algorithmic radicalisation. Most concerning, this *process mirrors established radicalisation pathways* and individuals with pre-existing grievances, identity insecurities, or social isolation are particularly vulnerable. Algorithms are influencing individuals, as platforms reinforce and normalise narratives, while isolating digital citizens from dissenting views and fostering echo chambers where extremist ideologies can take root and be exploited. Empirical studies show social media platforms could quarantine legal-but-harmful content, but are reluctant to deploy this type of intervention, as it could [reduce engagement at the expense of commercial losses](#). Platforms' prioritisation of financial incentives over safety leaves individuals trapped on [addictive content loops](#) that enable radicalisation. So, if engagement and profiteering is prioritised over safety, then algorithmic radicalisation is a feature which can and should be appropriately addressed by platforms.

Reactive Moderation: Limitations and Consequences

Across both AI and social media platforms, “free speech” and “freedom of expression” are sometimes used to justify limited intervention, often serving as armor to shield algorithmic radicalisation. Harmful content is moderated only after dissemination, prioritising engagement and revenue over proactive safeguards

“So, if engagement and profiteering is prioritised over safety, then algorithmic radicalisation is a feature which can and should be appropriately addressed by platforms.”

This creates a structural bias toward late-stage intervention, resulting in a conflict of interest: recommender systems fuel radicalisation, while transparency and accountability are treated as economic liabilities rather than ethical obligations. Institutional oversight is weakened, reducing content moderation to a trade-off between public safety and platform growth. For instance, some platforms participate in the Global Internet Forum to Counter-Terrorism - GIFT's hash-sharing databases to manage Terrorist and Violent Extremist Content (TVEC), as part of a crisis response protocol, but these efforts focus on overt explicitly designated terrorist materials rather than a broader range of TVEC content that fuels upstream radicalisation. Scholars have documented, “The hash-sharing database doesn't regulate 'TVEC' by trying to change harmful online behaviour—it intervenes in global data flows infrastructurally to make sure ‘things don't even hit the surface’.” From an optics perspective, this visually manages enforcement by targeting symptoms; however, it is not effectively addressing the normalisation of violent and extreme content stemming from algorithmic radicalization. From a PVE lens, reactive moderation fails to address the upstream drivers of radicalisation, such as: [misogyny](#), conspiracy theories, and grievance narratives. It does not build resilience, promote critical thinking, nor provide alternative narratives. Instead, it reinforces a cycle of suppression and backlash, and does little to prevent individuals from entering extremist pipelines in the first place.

“From a PVE lens, reactive moderation fails to address the upstream drivers of radicalisation, such as: misogyny, conspiracy theories, and grievance narratives.”

Power Asymmetry: Reduced Agency is Operationally Convenient

From a structural standpoint, platforms rarely treat the risk of radicalisation as a core business failure because monetising engagement is their primary objective. Platforms implement piecemeal interventions but do not challenge incentive structures or power imbalances, implicitly tolerating harms to vulnerable digital citizens as long as engagement and revenue remain strong. For platforms, the millions of people vulnerable to algorithmic radicalisation are still an *unquantifiable minority*, thus, seen as a trade-off. This estimate fails to account for the compounding harms caused to numerous individuals and communities impacted by actual real-world violence. Moreover, generative AI systems introduce similar risks, where researchers have raised concerns about how AI chatbots respond to self-harm or violence-related prompts, particularly during extended interactions where safety guardrails degrade. Since billions of people use social media and hundreds of millions interact regularly with generative AI systems, even a small fraction exposed to radicalisation pathways translates into *millions of at-risk individuals*. Yet, these digital citizens are often treated as *statistically negligible*—rendered invisible within profit-driven calculations.

Platforms create asymmetric dynamics due to structural imbalances. For example, digital citizens rarely see how algorithmic radicalisation functions in real-time or how their data influences exposure to harmful content. By controlling recommender systems that are responsible for algorithmic radicalisation, platforms gain significant power to influence the information digital citizens see and interact with online. On the other hand, digital citizens have no-to-low transparency or control over how their data is used, creating an imbalance between platforms' decision-making and individual autonomy. Regrettably, this is further exacerbated as “opt-out” mechanisms, (when they exist) are buried or require substantial digital literacy.

This power imbalance is further heightened by two interrelated factors: first, algorithmic radicalisation is structural and second, algorithmically driven systems contribute to making radicalisation feasible, as it continuously exposes digital citizens to influence-oriented information, undermining space for genuine deliberation and reflection. Here, the power dynamic is subtle: platforms do not assume digital citizens lack agency, but where agency exists, it is systematically weakened by default settings, opaque personalisation, and asymmetric knowledge. The problem is that platforms make reduced agency operationally convenient, choice, is formally preserved while being structurally sidelined. Recognising this distinction avoids technological determinism while highlighting power imbalances in curated digital ecosystems.

“The problem is that platforms make reduced agency operationally convenient, choice, is formally preserved while being structurally sidelined.”

Recalibrating Economic Trade-offs and Algorithmic Radicalisation

Platforms must recalibrate from economic trade-offs and algorithmic radicalisation to true accountability, to disrupt this cycle. For example, platforms insinuate radicalisation is a result of individual choice, susceptibility, or moral failure, when it should be recognised as a predictable outcome stemming from an algorithmic radicalisation. Because platforms transfer responsibility onto digital citizens and away from algorithmic radicalisation, this narrative preserves the legitimacy of platforms' business models by casting radicalisation as exceptional and individualised. This status quo allows platforms to be both culprit and comforter, simultaneously.

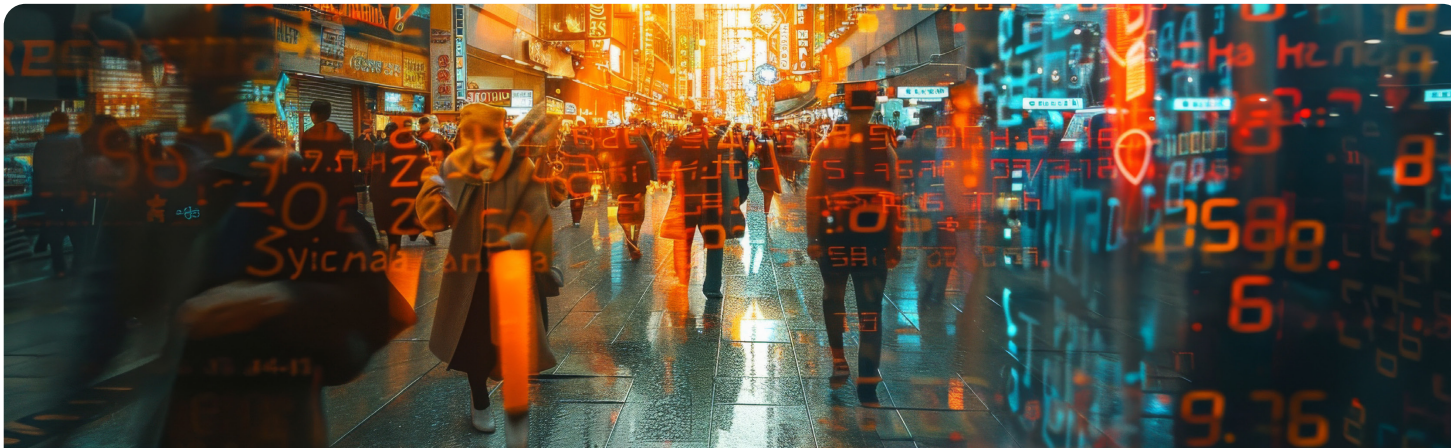
“Platforms must recalibrate from economic trade-offs and algorithmic radicalisation to true accountability, to disrupt this cycle.”

Conclusion

Economic trade-offs have real consequences: radicalised individuals, fractured communities, and weakened democracies. PVE demands rejecting the notion of algorithmic radicalisation as an unavoidable harm. It means building resilience, fostering safer communities, and embracing a future where digital citizens are endowed with agency, not stripped of it. Algorithmic radicalisation is not inevitable, and platforms can be reshaped for societal cohesion and safe(r) digital ecosystems.



Gazbiah Sans
Founding CEO, PVE Works



BUILDING RESILIENT COMMUNITIES,
TOGETHER.

