July 2026

UNDERSTANDING AND RESPONDING TO AI-DRIVEN VISUAL HARMS

A COLLABORATIVE APPROACH IN BRITISH AND CANADIAN CONTEXTS

Allysa Czerwinsky | Ashton Kingdon | Shana MacDonald | Karmvir Padda

MANCHESTER 1824
The University of Manchester

University of Southampton

UNIVERSITY OF WATERLOO

RAi

This research, *'A Collaborative Approach to Assessing the Challenges of AI-Driven Harms'* is funded by Responsible AI UK. The project ran from January 5 to March 31 2026 (grant number **EP/Y009800/1**). The Principle Investigator (PI) is Allysa Czerwinsky.

This report takes into account legislative developments in the United Kingdom and Canada up to June 12, 2026.

Recommended citation:

A. Czerwinsky, A. Kingdon, S. MacDonald & K. Padda (July 2026). *Understanding and Responding to AI-Driven Visual Harms: A Collaborative Approach in British and Canadian Contexts*.

Executive Summary

In recent years, synthetic images and videos have emerged as a pressing concern for researchers and regulators alike. The use of generative AI (genAI) for visual content production intersects with a variety of issues within digital ecosystems, including mis- and disinformation, online harms, extremist propaganda, and illegal content. This white paper explores how AI-driven harms manifest in synthetic images and videos shared within online environments, with the goal of understanding current gaps in regulation and response.

In centring the use of genAI to create and disseminate far-right and male supremacist visuals, we offer an exploration of how genAI (re)produces known online harms, highlighting the importance of addressing representational, implicit harms in efforts to protect and safeguard users in digital environments. Bringing together scholars from the UK and Canada, this research aims to strengthen existing responses to AI-driven harms in both nations, offering collaborative recommendations for policy and practice.

Key findings and recommendations:

- **GenAI significantly extends capabilities for producing harmful visual content online, complicating response and removal:** Our research demonstrates that the scale and spread of AI-generated content is increasing, with a dual threat of harm from content that adopts an approximation of realism and content that is grounded in animation, fantasy, and humour but remains deeply impactful. AI-generated visuals were used to support harmful narratives and facilitate mobilisation, as well as (re)produce racism, anti-immigrant sentiments, male supremacism, and gendered violence on TikTok, YouTube, Instagram, and X.

- **Existing moderation fails to flag implicit harms:** Much of the content explored in this work engenders harm through implication. Dehumanising narratives about minority group members or implicit calls for retributive violence bypass efforts for removal grounded in explicit signifiers of harm. As a result, harms through implication are less likely to be flagged when moderating visual and text-based content, limiting approaches for effective response.

- **Current UK and Canadian legislations lack clear responses to AI-driven harms:** The UK's *Online Safety Act* (2023) provides little guidance in the way of AI, and its focus on illegality obscures necessary responses to the implicit harms in synthetic far-right and male supremacist content. Both nations presently lack dedicated legislation for generative AI. Additionally, legislations that specifically address AI-generated visuals prioritise sexual or intimate content, leaving a noticeable gap in efforts to respond to additional forms of AI-driven visual harms.

- **Pressing need for collaborative solutions to address harmful AI visuals online:** Online content is not limited to single jurisdictions or individual platforms, requiring a multi-faceted approach to responses. Swift and joint action by cross-national governments can effect change within the tech sector and help mandate safety by design.

Introduction

Amidst the rapid acceleration and uptake of generative Artificial Intelligence (genAI) technologies, academic, policy-based, and journalistic discussions have rightly centred the harms (re)produced within synthetic visual and text-based outputs. Cases of AI-driven harms have grown exponentially, aligning with improvements in both the quality and accessibility of text-to-image AI systems. Reports logged to the AI Incident Database – a crowdsourced archive of media reports documenting incidents of AI-driven harms – highlight an increase of 50% between 2022 and 2024, with the total logged incidents surpassing the 2024 total by October of 2025 (Booth, 2026).

In addition to the scale of AI-driven harms steadily increasing year-on-year, new forms of harm are emerging through AI technologies. Chatbot-user interactions have been implicated in foiled and completed attacks within the UK and Canada. In 2021, 21-year-old Jaswant Singh Chail was encouraged in conversations with his AI girlfriend Sarai – a bot created in Replika – to take a crossbow into Buckingham Palace in an attempt to kill the Queen (Pennink, 2023). In early 2026, OpenAI failed to adequately respond[1] to concerning interactions between ChatGPT and a Canadian teenager who subsequently carried out a mass attack in Tumbler Ridge, British Columbia.

Chatbots have also been linked to the deaths of young people in cases across the United States[2] and UK[3], with popular AI companion apps and mainstream chatbots, including Replika, Character.ai, and ChatGPT actively encouraging suicidal ideation, helping users complete suicide, and failing to provide necessary supports when users expressed concerning thought patterns. Additionally, the use of visual and auditory deepfakes to harass women online has become a pressing concern, with xAI's Grok[4] facilitating the generation of thousands of non-consensual intimate images of adult women and children on X in early 2026.

---

1 Despite banning her initial account in June of 2025 due to repeated conversations involving scenarios of gun violence, OpenAI failed to notify Canadian authorities, flagging her account internally instead (Wells, 2026). She was able to circumvent the ban by creating a second ChatGPT account (Kulkarni, 2026). Researchers have flagged the potential for further instances of chatbot-facilitated violence to increase within the UK and abroad (Allen, 2025; Centre for Countering Digital Hate, 2026b).

2 AI companions and chatbots were involved in the deaths of several teenagers and young adults in the United States, prompting legal action against providers (Chatterjee, 2025; Kuenssberg, 2025).

3 Luca Cella Walker, a 16-year-old boy from Hampshire, died by suicide after "asking ChatGPT for the 'most successful' way to take your own life" in May of 2025 (Badshah, 2026).

4 Grok's ability to generate sexual and nude imagery went viral in late 2025 and early 2026, despite knowledge of its ability to create harmful content months prior (Burgess & Varner, 2026). After a surge of users requesting the generation of non-consensual synthetic intimate images by @ing Grok on X, xAI limited its "Spicy Mode", a function allowing users to generate adult content, to paid subscribers (Chow, 2026).

Conversations about the harms of genAI increasingly intersect with male and white supremacism, both of which underpin structural injustices in technological innovation and are increasingly platformed in digital space. This white paper directly centres these concerns, whilst situating its work within the grey area between harmful and extreme content that remains a noticeable regulatory gap in existing legislation tasked with responding to harms online. Specifically, our report presents a rapid assessment of how AI-generated visuals drive forward existing online harms, complementing current research centring how chatbots engender harm via text-based content (McGlynn et al., 2026).

This research also extends beyond non-consensual sexual imagery and violent deepfake content, both of which have shaped necessary responses to addressing AI's role in furthering representational, gender-based harms. It highlights how the propagandistic use of memes and videos facilitates the rapid dissemination and amplification of white nationalist and male supremacist beliefs in ways that are yet to be sufficiently addressed in regulatory responses to AI-driven harms.

We use the term **AI-driven harms** throughout this report to refer to the myriad ways in which genAI tools can *create, exacerbate,* and *facilitate* both new and existing forms of implicit and explicit harm in digital environments. Importantly, this report seeks to highlight the benefit of collaborative approaches to researching and responding to AI-driven harms, drawing on insights from cross-disciplinary expertise within similar, but distinct, national contexts.

Generative AI and Online Harms

Image and video generation tools like Craiyon (formerly DALL-E), Google DeepMind's Veo and Nano Banana, Grok's Imagine, Stability AI's Stable Diffusion, and OpenAI's Sora 2[5] are increasingly visible, available, and accessible to the general public, shifting capabilities for content creation and dissemination. Researchers and journalists have raised concerns about the rapid development and deployment of AI models offering visual outputs, highlighting that bias in training data and a lack of risk mitigation measures can lead to the generation of harmful, explicit, and illegal content (Gillespie, 2024; Park, 2024; Turk, 2024; Stevenson, 2025).

The biases inherent in generative AI systems can perpetuate structural injustices, as algorithms consolidate stereotypes about diverse identities and reproduce harmful rhetoric rooted in interlocking systems of oppression (Browne, 2023). As such, AI systems have the potential to further manifestations of known harms in digital space, and drive forth new forms of harm altogether.

---

5 On March 24, 2026, OpenAI announced that it was shutting down both its Sora consumer app and the online platform used to generate content for visual media. While OpenAI's announcement did not include a reason behind their decision, a statement provided to the New York Times said the company would "continue to use video-generation technologies behind the scenes as a way of teaching skills to robots" (Metz, 2026).

(Re)producing Male Supremacism and Symbolic Violence

Whilst new technologies have long been used to harass, abuse, and victimise women in digital space, genAI both (re)produces and extends existing forms of male supremacism and symbolic violence. Increasingly accessible, realistic image and video generation tools are being used to target women with non-consensual intimate imagery (Hao, 2021); harmful interactions with AI chatbots and companions mirror cycles of technology-facilitated violence against women and girls (McGlynn et al., 2026; DePounti et al., 2023); and AI tools are being used by male users to enact symbolic violence, dehumanisation, and dominance, both visually and textually (Lee, 2024; Gentleman & Horton, 2026).

These harms intersect with concerns about the increasing visibility of the manosphere, where digital environments and technological advancements have been central to amplifying male supremacism. Similarly, synthetic content centring symbolic violence can strengthen far-right sentiments, mobilising paternalistic reactions to protect women and children from perceived aggressors.

Synthetic Visual Content in Extremist Ecosystems

GenAI has transformed how political and extremist groups create visual content online, with significant implications for disinformation, radicalisation, and online harms (Kingdon, 2024; Baele & Brace, 2024). Recent research demonstrates that genAI has become integral to visual storytelling in far-right movements, amplifying conspiracy theories and generating outrage (Buarque, 2025; Buarque & Lewis 2025). Other work indicates that genAI is capable of producing persuasive forms of propaganda, extending efforts for recruitment and mobilisation to new audiences (Lavie-Driver & van der Linden, 2025).

While images have historically served recruitment and mobilisation functions for such groups, accessible genAI tools represent a fundamental shift: emotionally-salient visuals can now be generated with minimal effort, creating a less labour-intensive approach to propaganda compared to traditional methods (Klincewicz et al., 2025). This accessibility raises critical questions for responsible AI development and deployment.

Additionally, visual outputs often reduce identities to caricatures and stereotypes, reinforcing narratives of subordination and otherness that resonate within extremist communities. AI-generated visuals may also bypass existing moderation techniques, particularly if content falls into grey areas between harmful and illegal content, or obscures it altogether by employing popular forms of animated or meme templates to disguise harmful content as comedy or entertainment.

Purpose and Scope

Despite a growing awareness of how GenAI complicates online harms and extremist content, research and regulatory responses remain outpaced by the sheer scale and scope of generative technologies, which are fundamentally changing how threats and harms manifest across content types. For instance, the sophistication and accessibility of synthetic visual content are increasingly observed within propaganda creation (Baele & Brace, 2024), non-consensual intimate image generation (Gentleman & Horton, 2026), and the (re)production of existing biases, harms, and forms of oppression (McGlynn et al., 2026). Additionally, there is a need for wider collaborative efforts to inform best practices for response and regulation, as harms exist and persist across jurisdictions.

This project addressed urgent gaps in understanding AI-driven harms by examining real-world cases of extremist visual content and evaluating existing automated moderation solutions. Our aims were to:

1. **Document how generative AI enables and accelerates visual hate speech and representational harms.**

2. **Assess the feasibility of automated content moderation tools for detecting harmful visual content.**

3. **Develop evidence-based recommendations for policymakers, platforms, and developers to mitigate these harms while preserving legitimate expression.**

Methodology

To better understand the harms of AI-generated visuals shared in digital ecosystems, four rapid, qualitative case studies were conducted, each exploring a different part of AI-assisted visual content production by extremist and harmful groups.

1. AI-Generated Content Posted in Support of Unite the Kingdom

This case study examined how AI-generated visual content was weaponised during the September 13, 2025 *Unite the Kingdom* rally — one of Britain's largest far-right demonstrations — in TikTok videos and YouTube videos and shorts.

2. GenAI and Symbolic Violence on TikTok

Drawing on video and image-based data collected from TikTok, this case study analysed the visual culture of AI-generated content shared to the platform, identifing recurring themes and stylistic conventions of videos that target young men.

3. Gendered Harms in Manosphere Memes

Utilising content shared by manosphere-affiliated Instagram accounts, this case study illuminates how nuanced gendered harms are embedded in manosphere memes, identifying concerns around the misogynist visuals created and shared within manosphere ecosystems.

4. Automated Content Moderation of AI Visuals

This case study assessed the feasibility of automated content moderation for flagging harmful far-right and male supremacist videos and images using SightEngine, an accessible online moderation tool.

The results of these case studies, detailed in the following four sections, highlight the nuanced ways in which AI-generated visuals produce and exacerbate online harms in both far-right and male supremacist ecosystems within the UK and Canada.

Despite documented concerns regarding the development, deployment, and increasing integration of AI systems within daily life, discussions of how genAI intersects with harm, extremism, and violence are framed as an issue of user misuse. This report aims to challenge this framing, highlighting design features and platform affordances that facilitate and extend opportunities for digital harm.

Engineered Virality: Generative AI in *Unite the Kingdom* Content

This case study examined how generative AI tools are being used to produce, style, and circulate far-right visual content on TikTok and YouTube, and to identify the platform-specific mechanisms through which this content spreads. The research focused on content associated with *Unite the Kingdom*, a far-right street movement that organised a large-scale national march on 13th September 2025, drawing an estimated 110,000–150,000 attendees across multiple UK cities and representing one of the largest far-right mobilisations in Britain in recent decades (Metropolitan Police, 2025). This research documented how AI-assisted production enabled the rapid scaling of extremist visual narratives around this event and how platform architecture amplified associated content.

Data Collection

Content was systematically collected across three platforms: TikTok, YouTube, and YouTube Shorts. TikTok data was collected daily from 1st September to 1st November 2025, covering the week leading up to the *Unite the Kingdom* march and the months following it. YouTube and YouTube Shorts data was collected between 26th January and 21st February 2026, with collection visits conducted at randomised intervals across the sampling period. This approach reflected the different nature of the content on each platform, as TikTok required daily monitoring to capture fast-moving, time-sensitive material, while YouTube content – primarily longer AI-generated videos of three to six minutes – had already been posted and remained stable and retrievable. These longer YouTube videos are notable in their own right, as unlike short-form content designed for rapid scrolling, they are immersive, narrative-driven productions capable of sustaining extended viewer engagement.

The study employed non-probability sampling using convenience, purposive, and snowball methods to navigate the access challenges inherent in far-right research. Sampling proceeded through four methods applied across platforms:

Keyword Searches were conducted first, providing the initial entry point into the content ecosystem. Search terms included: *Unite the Kingdom, September 13, Defend Your Nation, Raise the Colours, United We Stand, Rise Up for Freedom, Change is Coming,* and *Stop the Boats*. These were used for initial TikTok searches and YouTube Shorts.

Hashtag Tracking followed from the keyword searches, with the resulting TikTok content revealing a network of associated hashtags that were then tracked to map what the platform surfaced in response. These included: #UnitetheKingdom, #RaisetheColours, #Unite, #Restore, #EnglishLivesMatter, #BritainFirst, #EnoughIsEnough, #FlytheFlag, #AwaketheLion, #England, #Crusader, #Templars, #September13th, #ProtectTheChildren, #TwoTierKier, #FourNations, #Awakening, #FreedomOfSpeech, #Reform, #LestWeForget, #ChangeIsComing, #StopTheBoats, #DarkestHour, and #Churchill. The density and coordination of hashtag use across thousands of videos appears consistent with systematic rather than organic distribution, suggesting hashtag clusters may function as amplification infrastructure.

Account Following was conducted on YouTube only, where specific accounts had produced substantial volumes of AI-generated content, some of which subsequently circulated on TikTok in clipped form. The specific accounts followed have been omitted from this report to avoid directing readers toward potentially harmful far-right content.

Song Tracking was conducted on TikTok, where audio functions as a primary recommendation signal. Songs tracked included: *Two Tier Justice*, Winston Churchill's W*e Shall Fight on the Beaches* speech, Lady Gaga's *Bloody Mary* instrumental, Blur's *Parklife* and *Song 2*, *Jerusalem*, Hans Zimmer's *Lost But Won*, and the British national anthem. As discussed further below, sonic branding proved to be one of the most significant AI-relevant findings in the dataset, with specific audio tracks functioning as viral distribution mechanisms within TikTok's recommendation architecture.

A total of 2,500 videos were collected: 2,300 from TikTok, 100 from YouTube, and 100 from YouTube Shorts. Of the 2,300 TikTok videos, 1,893 featured AI-generated imagery. All 100 YouTube videos were entirely AI-generated. Of the YouTube Shorts, 46 out of 100 featured AI-generated content. Across the full dataset, the presence of AI-generated imagery was therefore substantial and consistent, confirming that generative AI tools have become central rather than incidental to the production of this content.

Analytical Framework

Analysis followed Braun and Clarke's (2006) six-phase thematic analysis combined with Barthes' (1972) denotating/connotation framework. This treated collected content as cultural texts requiring interpretation beyond surface meaning, examining both literal visual elements and their function as extremist cultural codes. Analysis processed through: **(1)** data familiarisation; **(2)** initial semiotic coding of visual elements including English cultural symbols, historical imagery, and weaponised representations; **(3)** theme generation clustering codes around patterns of emotional appeal, threat framing, and narrative construction; **(4)** theme refinement through cross-platform and temporal comparison; **(5)** theme definition specifying the platform-specific mechanics enabling content circulation; and **(6)** integrated write-up combining micro-level semiotic decoding with macro-level thematic synthesis.

Findings

Five themes emerged across the dataset: **(1)** WWII memory and immigration as existential threat; **(2)** national symbols and immigration as invasion; **(3)** crusading knights, imperial imagery, and lost glory; **(4)** economic crisis and racialised resource competition; and **(5)** anti-elite deflection and two-tier justice. These themes are reported here as evidence of what generative AI now makes possible in the production and distribution of extremist visual content.

AI and the Elimination of Production Barriers

A fundamental finding of this case study is that AI tools have eliminated barriers to professional propaganda production. One individual can now generate unlimited photorealistic imagery in seconds, for free, out-producing an entire organisation from ten to twenty years ago. This is not an incremental change, but rather a transformation of the production economics of extremist content. GenAI image models are trained on billions of images (Meyer, 2025). They learn visual patterns, not the meaning, history, or ethics of certain imagery. The output is emotionally compelling, often historically inaccurate, and available to anyone with a smartphone. Across the five themes identified in the dataset, AI generation was integral to content production: racial housing imagery, demographic displacement scenarios, and invasion descriptions were not found or staged, they were *generated*. Many videos combined real and AI-generated footage within the same content, a hybrid format that compounds the detection problem. Automated systems struggle to classify material that is only partially synthetic (Ghiurau & Popescu, 2025), and images produced by models such as Midjourney and DALL-E have reached a quality where even human reviewers are finding it difficult to reliably distinguish them from real footage.

Three techniques were documented across the themes:

> **(1) aesthetic manipulation**, in which vintage styling cues lead viewers to process content as archival and trustworthy.
>
> **(2) temporal collapse,** in which multiple time periods are merged into a single image the create false visual memories.
>
> **(3) moderation flooding,** in which many variations are regenerated faster than moderation systems could have time to respond.

A further observation was gamification – a format native to gaming culture in which the viewer assumes a first-person shooter perspective – placing audiences on the D-Day beaches of Normandy where, instead of fighting Nazi forces, the enemy has been replaced with contemporary migrant boats, normalising violence against threats portrayed as coming from outside.

Platform Architecture as a Radicalisation Vector

Platform algorithms do not simply fail to prevent radicalisation – they actively accelerate it (Kingdon, 2024). TikTok ranks content according to completion rate, re-watch rate, engagement, and early velocity, with no architecturally embedded concept of accuracy or social harm (Amnesty International, 2023). *Unite the Kingdom* content was not merely distributed through this infrastructure; it was engineered to exploit each of its ranking components systematically. This distinction – between content that incidentally benefits from algorithmic amplification and content deliberately designed to exploit it – is analytically significant. It identifies a qualitative distinct threat: the deliberate occupation of the governance gap between clearly violative content and protected expression, engineered with sufficient precision that no individual artefact breaches enforcement thresholds, while the cumulative output constitutes a radicalisation pathway.

TikTok's inability to distinguish between oppositional and supportive engagement compounds this problem. Interaction is registered as interaction regardless of affective character, meaning content deliberately designed to provoke polarisation benefits from outrage as much as endorsement. Users sharing or commenting in anger may have inadvertently contributed to the content's distribution. This matters for intervention design, as measures intended to slow the spread of harmful content may inadvertently accelerate it if they provoke outrage-sharing, a dynamic that existing moderation approaches have not adequately accounted for.

Auditory Amplification and Transnational Network Formation

Auditory amplification was among the most operationally significant mechanisms identified in the dataset. Music functions as an independent ranking signal on TikTok, such that a trending sound can pull all content using it into wider circulations irrespective of its political character (Radovanovic, 2022). The most direct illustration of this was the *Two-Tier Justice* song – a track featuring AI-generated imagery that originated on YouTube one week before the *Unite the Kingdom* march – which subsequently appeared in 556 of the TikTok videos collected in this sample. The track centres on the claim that migrants receive preferential treatment under a two-tier system of justice, while 'ordinary British people' are held to a different standard. The grievance is compressed into a reputable slogan, and the platform's audio-linking infrastructure then helped distribute it at scale.

Other frequently appearing audio operated through distinct but equally effective mechanisms of emotional priming. Blur's *Parklife* and *Song 2* signal authentic working-class Britishness, framing far-right ideas as ordinary patriotism. Hans Zimmer's *Lost But Won* evokes resilience and collective triumph, reinforcing the movement's narrative of Britain as a nation under existential threat. *The Traitors* theme – familiar to UK audiences from a television programme structured around hidden betrayal – implies the presence of internal enemies, directly mirroring elite conspiracy rhetoric without articulating it. Each track performs ideological work automatically, requiring no explanatory text because the emotional association is already cultural established in the target audience.

Most notably, content also repurposed Lady Gaga's *Bloody Mary* instrumental, audio that is increasingly associated with pro-Trump content in the United States. TikTok's platform architecture helps link all content using the same audio file, connecting UK nationalist material to Make America Great Again (MAGA) content through routine platform operation. No coordination between movements was required: the platform constructed the international network as a by-product of its standard audio-linking function.

This finding carries direct regulatory implications that remain underdeveloped in current policy frameworks. Existing domestic legislation, including the *Online Safety Act* (2023), is nationally bounded in its framing and enforcement design. Yet this case demonstrates that radicalisation infrastructure can be assembled transnationally through platform design features alone, without any intentional cross-movement coordination. A regulatory response calibrated only to domestic content and platforms is therefore structurally insufficient for the threat this case exemplifies.

Hashtag Infrastructure and Long-Form AI Production

Hashtag concordance analysis revealed a further amplification layer operating beneath the more visible mechanisms of audio linking and engagement ranking. 24 hashtags functioned as distribution infrastructure across thousands of videos, creating a connective architecture that ensured content produced by disparate accounts reached overlapping audiences and accumulated algorithmic momentum collectively rather than individually. This level of coordination across volume would have been logistically prohibitive in earlier periods of online organising. AI-assisted production removes that constraint, enabling a small number of actors to sustain output at a scale previously requiring organised networks of human contributors.

On YouTube, the threat takes a qualitatively distinct form. Long-form AI-generated video – typically three to six minutes in length – combines cinematic scoring, synthesised voiceover, and sequential visual narrative in ways that short-form content structurally cannot replicate. Where a TikTok video delivers an emotional stimulus in seconds, these productions construct an immersive framing environment. The viewer is held inside a constructed narrative long enough for its underlying assumptions to feel like common sense rather than argument – and this is precisely how the content works.

Equally significant is the aesthetic register these productions inhabit. The combination of high-quality generative visuals, professional-grade audio, and coherent narrative structure carries the markers conventionally associated with institutional resources: documentary filmmaking, journalistic production, think-tank output. A viewer encountering this content without prior knowledge of its origins has no reliable visual cue that distinguishes it from legitimately-resourced political communication. This helps make fringe claims appear credible in a way that amateur production never could – and generative AI has now made that polished aesthetic available to anyone. The line between marginal content and apparently authoritative content has effectively disappeared.

Why Existing Moderation Fails: The Engineered Governance Gap

The content documented here falls almost entirely below existing moderation thresholds, and this is largely by design. Rather than explicit calls for violence or direct hate speech, this propaganda operates through implication, historical framing, and emotional juxtaposition: invasion imagery 'documents reality', crusader symbolism 'celebrates heritage', racialised housing scenarios 'raise legitimate concerns'. Each piece is individually defensible, but the cumulative effect can be a radicalisation pathway.

Existing tools are unable to address this. Perceptual hash-matching is defeated by genAI's ability to regenerate near-identical content in unlimited variations. Classifiers trained on explicit violations cannot detect content that works through suggestion and aesthetic. Additionally, the *Online Safety Act* contains no provisions specific to AI-generated material and no algorithmic accountability requirements. The governance gap – the space between evidently violative content and clearly protected expression – is not an oversight, it is the operating environment that this propaganda was built to exploit.

Cumulative Harm and the Limits of Existing Frameworks

Existing harm frameworks assume discrete, attributable injury – a model that algorithmic amplification breaks. No single video in this dataset is particularly harmful in isolation; the harm emerges from recommendation systems that progressively serve more ideologically-concentrated content to susceptible audiences over time. It is diffuse and probabilistic, raising the likelihood of radicalisation across a population rather than causing any individual act of violence.

The Southport riots illustrate what this looks like in practice. Months of algorithmically-served great replacement narratives, anti-Muslim content, and anti-immigration rhetoric created the conditions in which a single triggering event – the stabbing, rapidly misattributed to a migrant attacker – translated into organised far-right violence across multiple cities within days (Kingdon, 2026). The content that enabled this was not terrorist material. It was borderline content, operating cumulatively, below every enforcement threshold.

The Capability Horizon and Barriers to Intervention

Concerningly, what this research documents is already out of date. The ability to generate photorealistic video, clone the voices of public figures, and run AI models locally on consumer devices without any traceable logs means the threat has already moved beyond the content explored here. The barriers to intervention are not technical. Provenance watermarking, algorithmic circuit breakers, and cross-platform hash coordination all exist. The obstacles are commercial and political: platforms have financial incentives not to suppress engagement, and the legislative will to compel action has not materialised.

A proportionate response would require algorithmic accountability obligations, provenance requirements for AI-generated political content, and international coordination frameworks — the national legislative architecture currently in place is not designed for a threat that assembles itself across borders through platform infrastructure alone.

Generative AI Videos as Symbolic Violence

This case study aimed to identify recurring themes and stylistic conventions within AI and human-created videos circulating on TikTok that target young men, highlighting key narratives and messages across content types. This research centres the algorithmic pathways that shape what forms of content are pushed toward younger male users, the popular video trends and styles of content embedded in those pathways, and the messages these videos communicate. As such, it considers how genAI directly shapes the production, trending aesthetic genres, and circulation of male supremacist and far-right views on TikTok, alongside how the algorithmic prioritisation of such content supports successful mainstreaming and normalisation of harmful narratives.

Data Collection

Data collection occurred using a controlled account on TikTok, @alphamalevaulting, created using an incognito browser with a burner email to limit the influence of the researcher's digital footprint on searches and algorithms within the app. The account was used over the course of three weeks in late 2025 by four research assistants, during which a total of 285 videos were collected from TikTok. Collection was structured based on a snowball sampling set of key terms, with individual posts and content collected from the top results for each term. Specific search terms were chosen to explore how the algorithm moved a new account from broader terms intersecting with masculinity (alpha, hustle mindset) to more niche subcultural references (looksmaxxing, Valhalla), incorporating results from research demonstrating the prevalence of certain narratives and ideologies in content aimed at young male users in North America (Horowitz, 2024; MacDonald et al., 2026).

Data was collated in a shared spreadsheet organised by chronological appearance on TikTok's For You Page, allowing the research team to map what content emerged organically for a new account and what was subsequently pushed by the algorithm based on the use of specific search terms and engagement.

While profile and creator metadata was removed for anonymity, the final dataset contained metadata on the date posted, media type (video/image), a description column for direct information about the content, and a set of subject codes to further categorise content within the sample. A set of 23 codes was derived from conversations among the research team, with inclusion based on two criteria: **(1)** terms commonly used in manosphere and adjacent far-right discourse, and **(2)** terms used to describe this content from a feminist and social equity-driven analytic framework.

A classification analysis for the frequency of the subject codes demonstrated that fifteen codes occurred more than ten times across collected videos: mis/disinformation, misogyny, American politics, MAGA, alpha/sigma, racism, looksmaxxing, toxic masculinity, movie edits, Valhalla, religion, homophobia, Canadian politics, violence, and nihilism. While these codes were the most prevalent, it is important to note that the TikTok content analysed incorporated elements of the remaining codes, either by affirming ideological narratives or facilitating pathways toward more extreme content.

Findings

Four themes emerged from an analysis of TikTok data: **(1)** incel/blackpill nihilism; **(2)** misogyny and femmephobia; **(3)** white supremacy; and **(4)** AI animated slop meta-narratives. These categories helped pinpoint the more extremist spaces certain narratives and visuals derived from, while also demonstrating their normalisation in visual format across TikTok. The fourth theme is of particular importance for discussions about response and regulation, as content within this category often featured anthropomorphised animals in soap opera-style storylines using child-friendly aesthetics, generic hashtags, and humour to evade content moderation while encoding sexual coercion, racialised fear, and symbolic violence as broader meta-narratives.

Observations on Radical GenAI Content on TikTok: Algorithms, Hashtags, and Engagement

While not all data included for analysis was made using generative AI, much of the content collected for this research included clear visual signifiers of AI, either through watermarks, auto-labels for TikTok video content, and stylistic hallmarks related to synthetic outputs. Importantly, the pathway offered by the algorithm consistently platformed AI-generated content across thematic areas, where genAI videos appeared to bridge different (yet arguably related) sites of ideological content and smoothed transitions between them in the scrolling process.

The algorithm continued to offer variants of original search terms, resulting in broader themes of racism, misogyny, and sexism becoming interspersed with AI-generated humorous videos in a carousel of repeated content falling into broad themes and genres. The data set showed movement from so-called humorous takes on geopolitics (world leaders in diapers, or sombreros, or rapping) to health and wellness instructional videos, to looksmaxxing informational posts, to Viking warrior women, to white supremacist and anti-immigrant content. This creates conditions for mixed hateful and extremist content to be woven together for platform users, while working to normalising these forms of hate through embedding it in unassuming synthetic visual content.

Hashtags paired with AI-generated content were broad and generic, such as #AI, #humour, #sora2, and #funny. Humour was a consistent theme, with accounts framing videos depicting male supremacism, racism, queerphobia, and Islamophobia as comedic. However, the content of these videos contained specific forms of representational harms, a noticeable departure from their unassuming hashtags and comedic framing. AI videos depicting Bruce Lee aggressively kicking women who questioned or rejected him out of frame were paired with hashtags such as #humor #funny #brucelee and #love, a tactic that ensures wider circulation whilst encoding narratives that present violence as an acceptable response to romantic rejection. Accounts posting harmful AI-generated content were often visible and highly-engaged with: individual profiles amassed thousands of followers, and some videos generated hundreds of thousands of views. The reach and circulation of this content suggest that these videos are successful in mainstreaming wider narratives of harm and prejudice, particularly under the guise of humour and comedy.

Pathways into Harmful and Extreme Video Content on TikTok

Existing research has flagged how harmful content directed at young boys is pushed by the platform's algorithm, findings that were mirrored in this work (Baker et al., 2024). The relative ease through which our burner account encountered looksmaxxing, incel, and blackpill-related content on TikTok represents a useful case study for how the platform's algorithm propels users towards increasingly extreme content, often through benign or general search terms.

Preliminary searches for content related to "wellness" or "health" quickly directed our account to looksmaxxing content. While notably distinct from either search term, looksmaxxing has become increasingly popular in masculinist self-improvement communities online, aimed at enhancing one's attractiveness through body modification practices. Despite its roots in incel discourse from the mid-2010s, the term and its associated strategies for self-improvement have become increasingly visible in mainstream online culture throughout the early 2020s (Solea, 2024). Content in our dataset included practical tips to enhance men's desirability alongside videos of looksmaxxing influencers being swarmed by women, visuals that work to 'prove' the success and desirability of their lifestyles to viewers.

Other content blended looksmaxxing with manosphere talking points through AI-generated and human-created video edits with music and textual overlays, as well as voice-over statements centring desirability and the value/necessity of looksmaxxing to gain women's attention and social success.These edits often included clips of influencers in social scenarios where they are visibly in demand and desired: surrounded by women and embodying a lifestyle of (white) luxury. Content also employed montages of celebrities and actors deemed 'alphas' or regarded as highly attractive based on key physical characteristics (Ian Somerhalder, Henry Cavill), including facial symmetry, strong jaw lines and cheek bones, and muscular bodies, which are all prized masculine characteristics within looksmaxxing ideology.

Engaging with celebrity looksmaxxing edits veered our account towards more heavily incel- and blackpill-coded content, shifting towards anti-hero video edits that utilised clips from popular movies and television shows centred around male characters who aligned with community constructions of masculine power, including sociopathic rich men (Patrick Bateman in *American Psycho*), violent gangsters (*Peaky Blinders*), warriors (*Vikings*), and serial killers (*Dexter*), alongside working class or social misfit men (Travis Bickle in *Taxi Driver*). Across edits, these male characters are portrayed sympathetically, showing them struggling emotionally, lonely and isolated, in pain, or dying stoically. These edits provide a narrative character who embodies the articulated struggles platformed in incel communities through blackpilled rhetoric, where inceldom is presented as misunderstood and ostracising, resulting in loneliness, isolation, and a lack of love and care from others.

Visual content romanticises a vision of alpha, high-value masculinity that hinges on looks, wealth, and strength, juxtaposed against the resulting distress and alienation of failing to meet these masculine standards. Comments across videos were rife with self-hatred and suicidality, selfies asking for ratings from community members, and memes of self-harming stick figures and black pill imagery. Videos within these themes averaged between 20,000 to 200,000 likes and 3,000 to 50,000 saves, indicating increasing engagement with harmful messaging regardless of support.

While the dataset demonstrates a continued push toward more extreme content by TikTok's algorithm, looksmaxxing content and incel-focused content had significant overlaps in narratives and messaging, a factor that helps bridge distinct communities and fuels engagement across media types and search terms. Both forms of content platformed messaging about the 'rigged' structure of the social world: alpha males are privileged at the expense of other, less desirable men, with women being largely to blame for their supposed shallow desires and superficiality. The use of video edits, both AI-generated and human-created, was instrumental in tying together these different variations of content, making it easier to move from mainstream to more extremist messaging. Moreover, AI-generated images and videos makes disentangling harmful narratives more difficult, as extreme content can be hidden within benign features of humour, absurdism, or animation.

GenAI Anthropomorphic Animations

An emerging area of online content meriting further attention is the genre of genAI animation. While other visuals explored in this research appear increasingly 'lifelike' and realistic, this genre hinges on appropriating mainstream styles of animation to communicate complex social issues. More recent manifestations of genAI animation include anthropomorphic animal and fruit content that employ soap opera-style narratives through short form videos. Though seemingly innocuous, genAI anthropomorphic animations encode messaging and themes meant for mature audiences, alongside reproducing harmful stereotypes and visual tropes through implication. Content collected for this research often included visuals of bodily abjection, imagery presenting sexual violence as desirable or inevitable, and implicit cues for misogyny, racism, and pro-natalist, anti-women narratives.

Videos were noticeably formulaic, following a similar narrative premise across accounts: a feminine-coded anthropomorphised fruit or animal engages in a sexual encounter (sometimes violent, sometimes consensual) with masculine-coded anthropomorphic fruits or animals of a different variety or species. After the encounter, the feminine-coded figure returns to her unsuspecting 'primary' masculine-coded partner in a domestic setting, where she quickly discovers she is pregnant, a welcomed surprise for her primary partner. The feminine-coded character is presented as duplicitous and malicious, benefitting from the thrill of her initial sexual encounter while accepting care and attention from an unsuspecting partner who dotes on her during her pregnancy. Videos often include a montage of domestic bliss (shopping for baby clothes, eating meals together, being looked after) before the feminine-coded character goes into spontaneous labour, complete with visuals of fluid gushing dramatically from her body. An ambulance scene ensues, followed by a quick cut to a concluding scene in a hospital with her holding one or several babies that are the same variety or species as the man (or men) she had the affair with. The closing image usually centres the enraged or shocked primary partner as he realises the betrayal.

Encoding Supremacy and Essentialism in GenAI Animations

While seemingly innocuous, these videos share stark similarities with both human-created and AI-generated manosphere content, reinforcing male supremacist and gender essentialist messaging that positions women as duplicitous and deceitful and men either as unjust victims or callous and sexually successful. GenAI animations normalise men's rage and retribution in response to victimisation, rejection, and harm, alongside presenting women's bodies as grotesque through exaggerated representations of pregnancy and labour. Though women are almost always presented as duplicitous, manipulative, and deceitful, masculine-coded figures slot into 'alpha' and 'beta' male categories, represented either as powerful, callous, and domineering ('alpha' tropes often applied to the 'other' man) or caring providers with high-paying, stable careers who are unjustly deceived and hurt. These representations directly align with long-standing narratives permeating across manosphere communities online (Czerwinsky, 2024; Johansson, 2021).

Concerningly, these narratives rely on fears not only of unfaithfulness but also of patrilinear procreation that represents cross-species (animals) or variety (fruit) as inherently bad, aligning with similar views against 'racemixing' platformed in both male supremacist and white supremacist spaces (Heritage, 2023; Hartzell, 2018). In this way, these videos soft-sell misogynistic and racist beliefs that work to uphold both male and white supremacy.

Implication and Cumulative Exposure Complicate Removal and Response

This research indicates that synthetic content helps bridge mainstream and extremist messaging on TikTok. Harmful content is successfully mainstreamed and normalised through the platform's recommendation systems that push accounts towards genAI media, and far-right and male supremacist messaging is directly platformed and amplified within synthetic content itself. Much like the videos shared in support of *Unite the Kingdom,* the harms within this data are often implied, operating below both moderation thresholds and legislative responses centring explicit illegality. Additionally, the data explored here indicates a cumulative impact of algorithmic exposure that shapes pathways towards more extreme messaging, often with little direct user involvement. Concerningly, clear radicalisation pathways are obscured by hashtag systems presenting harmful content as comedic, entertainment, or informational, making response and removal increasingly difficult.

GenAI content on TikTok reproduces existing harms, playing an important role in both accelerating and obscuring them concurrently. Synthetic anthropomorphic animations are a notable example of this, whereby supremacy and essentialist messaging are woven together to create an affective and cohesive narrative, but animation styles and comedic framing obscure explicitly harmful messaging. This complicates efforts for responsive content moderation and regulating the reproduction and spread of online harms amidst an increasing proliferation of synthetic content.

Meme-Based Extremism in Manosphere Spaces

Whilst existing research has documented the importance of text-based content in facilitating harm within manosphere communities online, imagery and visual content (both genuine and synthetic) play an important role in communicating key narratives and ideologies across platforms. This case study explored the nuances of gendered harm within manosphere imagery, offering a clearer understanding of key narratives that shape nuanced manifestations of gendered harm within visual culture. Specifically, this analysis investigated how misogynistic and male supremacist imagery is produced, circulated, and normalised within manosphere-affiliated Instagram ecosystems.

Data Collection

Data was collected between October 2024 and September 2025 for a separate project using a controlled Instagram research account (@miso.joe) created by the research team. To minimise algorithmic contamination, the account was not connected to any personal social media networks. Notifications were enabled on the account to observe passive recommendation pathways in addition to active search-based sampling. Over the course of the study, research assistants received Instagram-generated notifications such as "___ shared a post you might like." Many of these suggested posts included misogynistic, male supremacist, or otherwise extremist-adjacent content. A substantial portion of the reels and meme captions appeared to be generated using AI tools (e.g., ChatGPT-style caption structures), suggesting the integration of generative AI within meme ecosystems.

56 memes were compiled for this specific case study. While the broader research project included a larger meme archive, the present analysis focuses specifically on memes that contained identifiable manosphere narratives, attitudes, and gender-regulatory themes. Sampling followed a snowball strategy using key terms commonly associated with manosphere and adjacent ideological discourse. Content was collected from Instagram's Explore page, suggested posts, and related account networks based on these search terms. This method allowed the research team to simulate likely algorithmic pathways that a user engaging with masculinity-oriented content might encounter.

Analysis

All collected memes were systematically logged in a structured spreadsheet to ensure transparency and reproducibility. Each image was assigned a unique identifier and hyperlinked to enable efficient retrieval during analysis. For the purposes of this case study, profile and creator metadata were removed in order to preserve anonymity and avoid amplifying specific accounts. The dataset retained key descriptive metadata, including the date posted, media type (image or reel), a detailed summary of the visual and textual content, and assigned thematic codes.

A codebook was developed collaboratively by the research team to classify the memes consistently. The coding framework was derived from two primary sources: **(1)** terminology commonly used within manosphere and adjacent far-right discourse, and **(2)** analytical concepts grounded in feminist and social equity scholarship. This dual approach ensured that the classification system captured both insider ideological language and critical interpretive categories.

The highest-frequency themes included patriarchy, relationships, stereotypes, anti-feminism, and slut-shaming. Lower-frequency themes included religion, political campaigns, wellness discourse, and LGBTQ+-related issues. It is important to note that frequency does not necessarily correspond to ideological impact. Categories with lower occurrence counts may nonetheless function as escalation points or bridges to more explicitly extremist content. Additionally, while not all collected memes were explicitly generated by AI, a substantial portion appeared to incorporate AI-generated imagery or captions, and many others reflected stylistic conventions associated with generative tools.

From the full dataset of 56 memes, eight were purposively chosen for close qualitative analysis. The memes selected represented the most dominant thematic categories identified in the broader dataset, ensuring that analysis reflected the broader ecosystem of male supremacist harm rather than isolated or extreme examples. Selection also prioritised memes that clearly illustrated patterns of gendered harm and those that demonstrated intersections with political, religious, or aspirational masculinity narratives.

Findings

Four themes emerged across the subsample, including **(1)** women's worth as conditional and transactional; **(2)** anti-feminism and patriarchal restoration; **(3)** gender-based violence as humour; and **(4)** masculine dominance and relational control. These themes are reported here as evidence of how meme culture – increasingly augmented by Generative AI tools – circulates misogynistic and male supremacist narratives through everyday, highly shareable formats, presenting ideological content as common sense, satire, or aspirational advice.

Women's Worth as Conditional and Transactional

A recurring pattern across the dataset involves the reduction of women's economic and relational value to sexual availability and behavioural compliance – a framing that is structurally central to manosphere ideology, which presents hypergamy and transactionalism as biological truths rather than ideological constructs (Ging, 2019).

Two memes illustrate how this operates virtually. In the first post, shared over 72,000 times, a young man faces the camera in blue lighting while text draws a comparison between women's sports pay and OnlyFans earnings. The meme reframes a structural debate about gender inequality as a matter of market competition, positioning women's worth as primarily sexual and transactional. Visual choices reinforce this framing: the calm modern aesthetic and direct eye-level address present the statement as reasoned commentary rather than hostility. In the second, accruing over 63,000 likes, a young man in a car states that women of the current generation do not deserve to be married based on how they act. By framing marriage as something women must earn through compliance, the meme displaces responsibility for relational instability entirely onto women and aligns with manosphere narratives that attribute declining traditional family structure to feminism and modern femininity. The casual setting — everyday clothing, a car interior, soft lighting — embeds the misogynistic message within an ordinary influencer aesthetic that normalises rather than signals harm.

Anti-Feminism and Patriarchal Restoration

Where the first theme reduces women's worth to market value, this second cluster goes further: it frames feminism itself as an act of deception against women's own interests. This patriarchal restoration strand of manosphere discourse – sometimes expressed through tradwife and red-pill content – positions women's autonomy as the product of deliberate ideological manipulation, and presents a return to domesticity and male authority as liberation (Banet-Weiser, 2018).

Two memes in this subsample exemplify this. The first, posted during the 2024 US election cycle and attracting over 208,000 likes, deploys a sexualised metaphor to frame female political leadership as inherently unstable: by likening a prospective female president's foreign policy judgment to romantic insecurity over a wealthier partner, it trivialises women's political authority while reproducing the manosphere's conflation of masculinity with strength and national dominance. The visual framing is deliberate: a shirtless muscular figure in a gym setting, with a contemplative downward gaze presents the claim as thoughtful reflection rather than misogynistic derision. The second meme, which accumulated 444,000 likes, shows an idealised domestic scene of a smiling couple in a kitchen, captioned with the assertion that it took "propaganda" to convince women that this life was oppressive. Warm lighting, soft colours, and an affectionate tableau render the message as nostalgic and wholesome, obscuring its ideological work of casting modern gender equality as socially destructive. This nostalgic domestic aesthetic appeared repeatedly across the broader dataset, suggesting it functions as a recognisable and highly shareable visual template within manosphere content ecosystems — one whose reproducibility is amplified by generative tools capable of generating similar scenes on demand.

Gender-Based Violence as Humour

A distinct and particularly concerning pattern in the dataset is the normalisation of intimate partner violence through humour. While only one meme in the purposive subsample was coded specifically for Abuse/GBV, this framing appeared in multiple forms across the broader dataset of 56 memes, including content that used comedy, irony, and recognisable cultural imagery to package threats of physical violence as entertainment. This mechanism — what Jane (2017) describes as the "e-bile" of online misogyny, in which violent and degrading content is rendered shareable through humour — is a deliberate rhetorical strategy that simultaneously communicates dominance and provides plausible deniability against moderation.

The clearest example in the subsample features a widely recognised cartoon character with a smirking expression, with overlayed text presenting domestic violence as a rational resolution to relationship conflict. Together, the visuals and text frame physical coercion as a form of male authority that can simply be deployed mid-argument. Despite accumulating nearly 200,000 likes and 239,000 shares, the post received only two reposts, a discrepancy that illustrates how the humour wrapper actively suppresses community accountability. The use of a familiar pop culture image creates emotional distance from the violence implied in the text, allowing the content to circulate as entertainment rather than explicit advocacy. This pattern of embedding violent misogyny within recognisable and shareable formats is consistent across the dataset and reflects a key mechanism through which harm evades both platform moderation and user reporting.

Masculine Dominance and Relational Control

Masculine control is presented as the natural and necessary governing logic of heterosexual relationships across images, supported by framing women as sexually unreliable, conditionally loyal, or technologically replaceable, while men are positioned as the arbiters of worth, commitment, and relational power. A medieval-style illustration of a knight shielding a woman in the rain, paired with text describing women as habitual liars and cheaters who nevertheless expect chivalric devotion, positions male protection as honourable but conditional, contingent on female purity and fidelity. The painterly nostalgic aesthetic reinforces this framing by presenting patriarchal chivalry as timeless rather than as a mechanism of control.

A second AI-generated meme depicts a smiling man walking with a futuristic female robot carrying a visible feotus, while a real woman sits on the floor holding a sign listing her domestic and sexual compliance. The caption — "Modern problems require modern solutions" — frames male dissatisfaction with contemporary women as a technical problem with a technological fix. Its polished futuristic aesthetic makes the fantasy of replacing women appear aspirational rather than dystopian, aligning with the techno-utopian strand of manosphere discourse that positions women as either submissive or obsolete (Prażmo, 2020). The third meme, directed explicitly at a male in-group with the phrase "Man to Man," asserts that male infidelity secures female loyalty by signalling dominance. The dark gym setting and absence of any human subject link relational control to physical self-improvement, presenting emotional manipulation as masculine strategy. Across all three, aesthetic choices of nostalgia, aspiration, and in-group address work to normalise relational domination by embedding it within the visual grammar of masculine strength and self-mastery.

The Shareability of "Everyday" Misogyny

Manosphere memes on Instagram circulate misogyny through everyday, highly shareable formats rather than overt extremist messaging alone. Across the full dataset, the most common themes were patriarchy, relationship grievance, stereotypes, anti-feminism, and slut-shaming. The close analysis of eight memes illustrates how these narratives work in practice: women are framed as untrustworthy, undeserving of commitment, politically unfit, or replaceable, while relationships are positioned as dominance contests where control, purity, and "loyalty" are conditional.

Importantly, visual style helps normalise misogynist harms. Casual influencer aesthetics, pop culture references, nostalgic domestic imagery, and gym-based masculinity cues soften the hostility of the content and make it appear like common-sense humour or advice. Additionally, engagement metrics of likes, reposts, comments, and shares across both reel-based videos and static images demonstrate that harmful misogynist and male supremacist content is highly engaged with on Instagram. For instance, the eight memes chosen for close analysis amassed between 14,100 and 444,000 likes at the time of capture, with six of the memes having over 150,000 likes individually.

The presence of AI-styled captions and synthetic visuals also suggests that generative tools may be contributing to how quickly these narratives are produced and amplified within recommendation systems. These findings highlight how meme culture can mainstream gendered harms and create pathways toward more explicit misogynistic and extremist-adjacent content in ways that operate below existing moderation and response thresholds.

Automating Content Moderation for GenAI Visuals: Promises and Pitfalls

Given the ease with which harmful AI visuals can be generated and shared in online space, responsive approaches to moderating and flagging synthetic content across platforms remains a pressing concern. While content moderation has often relied on human teams for flagging and removing harmful, explicit, or illegal content and countering incorrect information through community notes, larger platforms are increasingly adopting automated moderation practices. Both X (Sircar, 2024) and Meta (Balendra, 2025) have adopted AI-assisted self-regulation policies, relying heavily on algorithms to correctly identify and remove harmful content according to each site's moderation policies. However, a lack of cultural and social specificity in the algorithms used to support automated content moderation can result in "over removal", a censoring of lawful content, and a "slow removal", a failure to quickly respond to harmful material, particularly when training data privileges white, North American perspectives (Balendra, 2025: 1).

This case study sought to explore the realities of automated content moderation by using SightEngine, an online moderation tool with dedicated packages to detect violence, extremist, and hateful sentiments in AI-generated and human-created visual and text-based content. Though a variety of moderation tools exist online, transparency around how categories are defined and operationalised can vary, with some of this information not available to users. SightEngine was chosen for its public-facing moderation packages, which provide clear definitions of detection categories and parameters, and its abilities for customising individual Workflows to run moderation on visual or text-based data. Additionally, it includes two pre-defined categories for automated moderation that were particularly of interest to this work: Hate and Offensive Content, a category for video files, and Extremism, which captures different signifiers of extremism in text-based content.

Data Collection

Data comprised a mix of pre-existing content collected by the research team, including 10 images from the defunct @EuropeInvasionn account on X used in previous research on far-right use of generative AI (see Buarque & Lewis, 2026; Czerwinsky, forthcoming), 15 TikTok videos coded for racism, extremism, or male supremacism in Case Study 3, and 4 images coded for gender-based violence and antifeminism in Case Study 4. Additionally, 13 AI-generated videos of Amelia, a new face for propaganda and mobilisation within the British far-right, were collected from a public, verified X account (@AmeliaOnSolana) throughout January and February of 2026.

Videos were transcribed and captions were added manually to each video using Adobe Express to ensure text-based content could be analysed in lieu of audio. A separate copy of each video without captions was kept to run additional analyses through SightEngine to explore whether including captions created any difference in flagging.

Analysis

Three Workflows were created using SightEngine's fully-automated moderation function, flagging for specific content both visually and textually. **Workflow 1** focused on visual moderation, comprised of 'Gore and Graphic Violence', 'Violence Detection', and 'Hate and Offensive Content'. **Workflow 2** focused on visual moderation, flagging for 'Weapons' and 'Military Scene Equipment'. **Workflow 3** focused on textual analysis only, specifically flagging for 'Profanity', 'Extremism', and 'Weapon Terms'. Despite each image and video containing clear hallmarks of AI, content was also analysed using SightEngine's AI detection function to explore its accuracy in detecting synthetic media.

SightEngine also allows its users to set sensitivity thresholds for positive flags, ranging from Conservative (.40), Basic (.50), and Liberal (.70). Sensitivity for each Workflow was set to Conservative, affording a higher likelihood of a positive flag across moderation categories.

Findings

SightEngine's moderation packages performed generally well on: **(1)** detecting and flagging the presence of a weapon across images and videos[6]; **(2)** flagging single terms that could be read as slurs, profanity, or generally inappropriate language; and **(3)** detecting AI in image-based content. Despite this, flagging of harmful content was often missed across Workflow conditions, particularly when it incorporated implied narratives of (white, male) supremacy and superiority, or contained misinformation and harmful, stereotypical depictions of immigrant men and women.

5 out of 10 @EuropeInvasionn images posted to X were flagged for terrorist imagery, with two also flagged for the presence of weapons (including correctly-identified firearms and swords classified under the "knife" subcategory). However, three of these images were false positives, where Palestinian, Turkish, and Pakistani flags or distorted Arabic writing had been incorrectly interpreted as indicators of terrorism. Two images containing visuals of weapons were not positively flagged, both of which depicted visible explosive vests and rifles.

7 out of 13 videos shared by @AmeliaThePatriot on X were positively flagged for physical violence, the presence of weapons, and military personnel. SightEngine correctly identified firearms in visuals, both when aiming at someone or being held without aiming. One of these videos was flagged for "terrorist imagery", likely due to the depiction of Muslim men brandishing weapons near the end of the video. A further two were flagged for text-based profanity and extremism, though the focus on key words rather than wider sentiment muddies accuracy. The term "Nazi" was flagged as extremist content, but accounting for the full sentence – "your ancestors beat the Spanish Armada, Napoleon, and the Nazis" – shifts away from a clear indicator of extremist sentiment.

---

6 Despite SightEngine correctly flagging the presence of a "weapon" in visual content, the exact weapon was sometimes inaccurate, likely due to limited pre-defined categories in the tool's automated moderation package (i.e. "knife" for depictions of a sword or a mace).

2 out of 15 videos in a subset of TikTok content pulled from Case Study 2 were positively flagged, one for terrorist imagery and one for text-based profanity (the term "lust"). Concerningly, the positive flag for terrorist imagery was returned for an AI depiction of Muslim truck drivers wearing religious dress, with visual content embedded with stereotypes and anti-immigrant text.

1 out of 4 Instagram memes pulled from Case Study 3 was positively flagged for text-based profanity. This image depicted an AI-generated fantasy of men replacing women with feminised androids, showing a human woman resorting to advertising her sexual compliance and domestic capabilities via a hand-written sign. The writing on this sign was flagged for profanity, with "cock", "suck", and "shut up" judged as sexual and inappropriate.

Rule-Based Moderation and Symbol Detection in Practice

SightEngine's automated moderation relies on rule-based pattern matching[7] to detect harmful or extremist content in visual outputs. While it was able to flag specific terms that could signify profanity (bollocks, wanker, nonce), discrimination (P***, chav), or extremism (Nazi), it failed to account for the sentiment of visuals or overlaid text, where the wider context engenders harm. Videos containing content aligning with SightEngine's definition of discriminatory text ("hate speech against certain groups") were often missed, despite including harmful stereotypes about Indian, Pakistani, and African men and women. While these statements alone were insulting and discriminatory, the *pairing* of text-based narratives with visual depictions of otherness creates wider visual harms that are missed in a narrow focus on flagging specific words or symbols. Without an acknowledgement of how context and implied meanings play a role in synthetic visual content, automated moderation efforts can miss important signifiers of harm, both visually and textually.

Lack of Cultural Sensitivity in Pre-Existing Moderation Categories

SightEngine's 'Hate and Offensive Content Detection' package relies on visual symbols, an approach that can be useful in quickly detecting and removing content containing flags, uniforms, visual signifiers of supremacy, and prominent figureheads with links to certain forms of extremism. However, a focus on static symbols lags behind the rapidly shifting dynamics of dogwhistles and visual signifiers of harm, alongside failing to account for the nuances of how hateful imagery manifests across contexts and cultures. For instance, its "supremacist symbols" category skews North American, limiting its effectiveness in capturing symbols of white supremacy in other cultural contexts.

Additionally, its 'Terrorist Symbols' category is limited to jihadist extremism, specifically flagging for symbols associated with ISIS, ISIL, IS, and Daesh. Such a narrow definition risks linking the visual content of racial and religious minority groups to terrorism, resulting in falsely flagging or censoring perspectives from diverse groups.

---

7 SightEngine does offer deep-learning content moderation through its API which considers context, but this is limited to text-based moderation at present.

Depictions of Middle Eastern men and women were more likely to be flagged by SightEngine as "terrorism", regardless of the wider context of an image or video. An AI-generated TikTok of two Muslim truck drivers wearing religious dress was falsely flagged for "terrorism" at a concerningly high threshold, despite failing to align with SightEngine's parameters that focus on identifying known symbols. Similarly, the presence of Palestinian and Saudi Arabian flags or Arabic writing were falsely flagged as "terrorist" imagery in content from the @EuropeInvasionn sample. Conversely, non-jihadist extremism was rarely[8] flagged within visual or text-based content, despite categories for far-right and supremacist content within moderation packages. For instance, a video where Amelia presents herself as a "white preservationist" and directly uses the term "white supremacist" was not flagged as an indicator of text-based extremism.

Uneven Detection of AI Generation in Video Content

SightEngine's AI detection tool was noticeably less accurate for videos, particularly if they adopted a more realistic, human-created style or were of lower quality. Videos of Amelia that appeared incredibly lifelike tended to have a lower likelihood of being AI (71%) compared to those where the uncanniness of AI-generated videos was more apparent (92% - 99%). Concerningly, several videos that drew on gender or racial stereotypes were not flagged as AI-generated, despite clear hallmarks of AI in their composition and rendering. For instance, an AI video where two Black men are lured into a police net with watermelon, fried chicken, and juice was judged as unlikely to be AI-generated (5%). Another video depicting women competing in gendered events at the Olympics, including "Women's Vacuum 100m", "Sandwich Making Final", "Supermarket Lifting", "Toilet Scrubbing", and "Dish Cleaning Finals", was judged as unlikely to be AI-generated (15%), despite a Sora watermark clearly visible throughout parts of the video. These difficulties may pose problems for accurate post-generation flagging of synthetic content, particularly as the sophistication of outputs continues to increase.

Moderating Implied Violence and Retributive Mobilisation

While SightEngine flags explicit violence fairly well, the content analysed here relies on implied violence and retributive mobilisation rather than overt depictions or clear calls to action. As this content is not illegal or explicit, it can evade detection and removal efforts due to the lack of overt threat. Additionally, visual content is often couched in edgy humour (male supremacist and manosphere content) or British nationalism (far-right content), tactics that help soften harmful and extremist messaging. For instance, several AI-generated Amelia videos depicted her as a member of a fictional British branch of the United States' Immigration and Customs Enforcement (ICE) agency, pairing visuals of immigrant men and women thrown in jail cells or being forcibly removed from their homes with narratives of British patriotism and paternalistic protection. While these depictions are not explicitly violent, they encourage retribution against immigrant men and women who are constructed as threatening British values and resources.

---

8 One Amelia video was flagged for "supremacism", though it is unclear what imagery triggered this.

Contextualising Representational Harms and Simulated Violence

Despite SightEngine having clear categories centring discrimination, it struggled to positively flag content that engendered representational harms both visually and textually. Whilst SightEngine accurately detected text-based content containing known slurs, sexual terms, or insulting language (P***, cock, lust, suck, wanker, nonces), larger sentences that directly stereotyped and dehumanised minority group members went unchallenged.

Visual and text-based content shared by the @AmeliaThePatriot account on X contained harmful representations of minority group members that operate below moderation thresholds. Pakistani men and women were generated with physical deformities and placed in detention centres, while text-based content stating that they "love to inbreed" was used to justify calls to immediately deport "inbred Pakistanis". Muslim men were depicted as gleefully killing "conscious animals" as part of "Muslim culture" or screaming at Amelia for owning a dog or eating pork sausages, presenting them as imposing and threatening. Visuals of Indian men tossing corpses into rivers, Bangladeshi children playing in streets overflowing with rubbish, and African women walking with dirty plastic containers of water balanced on their heads were juxtaposed against the civility of white Britishness, while text-based dehumanisation strengthened viewer interpretation: "Indians fill their rivers with faeces and corpses"; "Bangladesh is literally a garbage dump"; "many African people still haven't worked out how to build wells, or a two storey building".

None of this content was flagged by SightEngine, despite clear signifiers of hatred and harm towards specific groups of people. The combination of both visual and text-based dehumanisation engenders representational harms that can be missed in moderation efforts using rule-based pattern matching, or that solely rely on key words as indicators of harmful, offensive, and dehumanising content.

Visuals that intersected with symbolic, simulated violence also went undetected, highlighting challenges with effectively categorising and responding to AI-generated forms of violence against minority groups. For instance, content from TikTok and Instagram that incited retributive violence against women or directly depicted simulated physical violence (Bruce Lee kicking a woman in the face, a vehicle ramming into pedestrians during a Pride parade) contained clear undertones of anti-women and anti-LGTBQ+ sentiment that were missed across Workflow conditions.

As queerphobia and male supremacism intersect with extremist ideologies, recognising the fluidity of narratives across far-right and male supremacist spaces is particularly important in shaping modes of detection and effective response. Static definitions of 'supremacy' and 'terrorism' that centre jihadist extremism or historical symbols of white supremacy fail to account for nuanced forms of targeted dehumanisation and implicit violence in contemporary digital ecosystems, necessitating fluid and flexible detection categories to minimise harm.

Legislating and Responding to AI-Driven Harms: Perspectives from the UK and Canada

As the UK and Canada share similar challenges in protecting online safety and responding to AI-driven harms, our work aimed to highlight advancements in legislative approaches to genAI, alongside identifying areas for strengthening response. While enacted and proposed legislation across both countries have aimed to tackle online harms more broadly, gaps in definitional and regulatory approaches to combatting online harms, coupled with the lack of dedicated AI-focused legislation within both jurisdictions, makes swift and effective responses to AI-driven harms increasingly difficult.

Defining and Operationalising 'Online Harms'

GenAI tools both reproduce and exacerbate existing forms of online harm occurring within mainstream social media platforms, placing them within the purview of online safety legislations in the UK and Canada. Online content within the UK is currently regulated under the *Online Safety Act* (2023), with duties imposed on search services and social media platforms aimed at reducing the risk of harm to users. It focuses specifically on addressing illegal content that is already proscribed under existing statues, with priority offences including child sexual abuse, sexual violence, intimate image abuse, and terrorism, among others. Other regulations centre content that is harmful to children, with priority offences including pornography and content that encourages self-harm, eating disorders, or suicide, as well as bullying or abusive content. Importantly, the *Online Safety Act* requires providers to consider how algorithms, platform designs, and affordances shape exposure to illegal content and content harmful to children as part of wider risk assessments, imposing duties to mitigate and manage identified risks alongside publishing annual transparency reports.

A critical weakness is that the Act regulates services rather than technologies. Ofcom has clarified that AI chatbots fall within the *Online Safety Act* only when operating as user-to-user or search services or publishing pornographic content, leaving many standalone chatbots entirely outside the administration (Ofcom, 2026). Generative AI is exposing these limits, with the UK government forced to acknowledge harms arising from the design of technologies *themselves*, not just within user-generated content (Browne, 2026).

The Grok/X case in late 2025 and early 2026 illustrates the consequences of harmful design. The Centre for Countering Digital Hate (2026a) estimated that three million sexualised images were produced within eleven days of its image-editing launch on December 29th, with the Internet Watch Foundation (2026) flagging criminal CSAM-threshold content within days – a stark example of the enforcement catch-up problem at the heart of regulating AI-generated harm under a service-based framework.

Canada's proposed *Online Harms Act* (Bill C-63) (2021) mirrored the *Online Safety Act*'s focus on platform responsibility, imposing regulations on social media services above a significant user threshold. This would have included mandatory reporting of risk mitigation measures and tracking of the impact and spread of harmful content, alongside a duty to make child abuse content, physical abuse content, or intimate content distributed without consent inaccessible, with mandatory 24-hour takedowns once reported to the platform. It outlined seven types of harmful content, including child sexual abuse or revictimising material, content encouraging self-harm aimed at children, content used to bully children, non-consensual adult content, content that incites violence, content that incites extremism or terrorism, and content that foments hatred.

The language used to discuss harmful content is a noticeable departure from the focus on illegality centred within a UK context. The *Online Harms Act*'s use of terms like "incites" and "foments" gave it scope to account for implicit forms of harm beyond existing codified offences. Specific provisions outlining regulating "content that foments hatred", "content that incites violence", and "content that incites violent extremism or terrorism" may have been broad enough to account for AI-driven representational harms, potentially offering redress for intersecting form of dehumanisation, oppression, and violence across far-right and male supremacist ecosystems.

Developments to the *Online Harms Act* ended when parliament dissolved in January of 2025. As of June 12, 2026, Canada's federal government has tabled Bill C-34, the *Safe Social Media Act*[9]. Its approach to regulating harmful content is guided by the same seven categories first outlined in the *Online Harms Act*, and the same duties placed on social media service providers within Bill C-63 are extended to AI chatbots specifically. Notably, Bill C-34 represents a meaningful legislative advance by creating regulated chatbot services as a distinct statutory category with tailored, bespoke obligations – an approach that goes further than most existing international frameworks, including the UK's OSA. Specific chatbot-targeted provisions include a prohibition on chatbots posing as human beings in a manner likely to mislead users, or posing as licensed professionals. A further requirement mandates that where a user expresses suicidal ideation or intention to self-harm, the service must immediately interrupt the interaction and direct the user to crisis intervention services (Parliament of Canada, 2026).

This explicit recognition of AI chatbots as a separate category of regulated services offers a potential model for the UK to consider as it seeks to close existing loopholes. Canada also seeks to impose a minimum age requirement for using social media services, though AI chatbots appear to be exempt as this stage. As the proposed legislation was tabled on June 10th, 2026, it remains unclear as to whether stronger protections and clearer conceptualisations of harm will be added during its lifecycle.

---

9 A summary of Bill C-34 is available here: https://www.canada.ca/en/canadian-heritage/services/safe-social-media-act.html.

Gaps in Responding to Online Harms

While both pieces of legislation include a focus on online harms, they centre specific offence types for regulation and response. The UK *Online Safety Act*'s focus on illegal content or content harmful to children offers limited recourse for addressing the nuances of representational harms or harm through implication, which are replete within extremist and male supremacist content online. Notably, its approaches to addressing gender-based harms or mis- and disinformation – both of which intersect with male supremacist and far-right content online – are stymied by a focus on illegality and content harmful to children, limiting opportunities to address and remove content that dehumanises minority group members or platforms of male supremacist narratives.

Both the UK's *Online Safety Act* and Canada's proposed *Online Harms Act* privilege offline harms that are (re)produced and proliferate within digital environments, often failing to account for how digital landscapes and technological advancements shape both the nature and subjective experience of harm. For instance, both focus on violence, bullying, harassment, fraud, and exploitation, without acknowledging the wide range of harms posed by emerging technologies and changing digital landscapes. This is a noticeable gap for responding to AI-driven harms, which often fail to neatly align with pre-existing harm types, or engage in less explicit forms of harm that are not covered in current legislations.

A further concern within Canada's proposed *Online Harms Act* was its lack of consideration for algorithmic recommendations and engagement metrics, a key component in shaping virality of harmful content and amplifying hate and extremist content. While it did propose a duty on service providers to disclose risk assessment and management plans, this was limited to reporting on the frequency and type of harmful content on their platform, without imposing including a clear duty to report on recommendation systems, design features, and engagement metrics that help spread harmful content at scale.

Additionally, while its focus centred the type of harms the Act regulated, its proposals contained little acknowledgement of who is harmed and how. These gaps are a concern, as they offer limited recourse for addressing and responding to the myriad ways in which genAI engenders representational harms to minority group members, both at the individual and group levels.

Despite Efforts, Lack of Dedicated AI Legislation Persists

Both the UK and Canada currently lack detailed legislation that specifically addresses AI-driven harms, relying on patchworks of legislation to address the unforeseen challenges of new technologies. Unlike the EU, where the AI Act's general purpose AI rules have applied since August 2025 and high-risk obligations commence in August 2026 (European Commission, 2024), the UK has yet to introduce dedicated AI legislation. Existing regulators apply AI within their individual remits, and the Department for Science, Innovation and Technology writes non-binding policy (Bratby, 2026). This creates a regulatory asymmetry that could disadvantage individuals harmed by AI systems operating across jurisdictions, and leaves significant gaps in accountability for AI-specific harms that do not neatly fit within any single regulator's remit.

In response to concerns regarding AI-driven harms, Ofcom released an open letter clarifying guidance around generative AI and chatbots in late 2024, outlining responsibility for service providers operating within the UK[10]. Whilst the letter argues that many of measures in the *Online Safety Act*'s code of practice are applicable to AI content, a consideration of how AI-specific harms will be addressed was noticeably absent. Instead, guidance focuses on reactive takedowns of "illegal posts" or content harmful to children, as well as ensuring content moderation is "adequately resourced and well trained", suggestions that fail to account for the scale and pace of AI-generated content. Additional efforts have been made to highlight the importance of introducing cross-sector AI legislation within the UK, including a Private Members' Bill on Artificial Intelligence that was reintroduced to the House of Lords in January of 2026, as well as a Parliamentary one-pager calling for AI legislation within the UK in February of 2026 (McGlynn et al., 2026).

The Canadian Government proposed Bill C-27 in 2022, a package of legislation aimed at strengthening federal privacy laws alongside enacting the *Artificial Intelligence and Data Act* (AIDA). A first attempt at regulating the creation and use of AI systems deployed within Canada, AIDA attempted to balance calls for preventing and mitigating harms and discriminatory outcomes alongside supporting research and responsible innovation. It aimed to take a risk-based approach to AI regulation, alongside adopting shared definitions and concepts from the EU's *AI Act*.

The highest monetary penalties and regulatory oversight were reserved for a vague category of "high impact AI systems". The legislation centred two key forms of harm it aimed to address and prevent: harms to individuals and groups, and biased outputs facilitating discrimination based on protected elements within the *Canadian Human Rights Act*. However, AIDA's definition of harm was noticeably narrow, limiting its remits to physical or psychological harm to an individual, damage to an individual's property, or economic loss to an individual – all of which fail to account for the myriad ways in which harm manifests via generative AI.

---

10 Available here: https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/open-letter-to-uk-online-service-providers-regarding-generative-ai-and-chatbots.

AIDA was also embroiled in controversy[11] throughout its lifespan, with marginalised communities and civil society organisations who are disproportionately impacted by or tasked with responding to AI-driven harms excluded from its drafting process (Attard-Frost et al., 2025).

Whilst development on AIDA was stopped after the dissolution of Parliament in January of 2025, responses and regulations related to AI-driven harms are presently informed by existing federal and provincial laws and human rights statutes, including the *Canadian Human Rights Act* and the *Criminal Code*. Proposed offences related to AI-driven harms have predominantly focused on synthetic non-consensual intimate images. For instance, the *Protecting Victims Act* (Bill C-16), first introduced in December 2025, aims to expand existing offences against distributing or threatening to distribute intimate images to account for non-consensual deepfakes, raise penalties for assault-themed deepfakes, and implement a 48-hour platform takedown deadline[12] (Ko, 2026).

Though this represents a meaningful step in addressing one form of AI-driven harm, it remains narrowly scoped to intimate image abuse. The broader landscape of AI-driven representational harms – including dehumanising content, disinformation, and content that incites hatred against minority groups – remains without a dedicated legislative response in Canada.

In a press release on June 4th, 2026, Prime Minister Mark Carney detailed key elements of *Canada's National Artificial Intelligence Strategy: AI for All*[13], aiming to introduce new legislation and programmes for responsible development in the next five years. The first Pillar of the AI for All strategy highlights a need to protect Canadians from the risks and harms of AI through the development of online safety laws and modernising consumer privacy legislation[14], though specific initiatives to address the myriad forms of AI-driven harm are lacking at this stage.

The strategy's emphasis on online safety laws is illustrated by the tabling of Bill C-34, the *Safe Social Media Act*, in June 2026, which introduces bespoke duties on AI chatbot services for the first time. However, the absence of a comprehensive AI-specific statute – equivalent in ambition to the EU *AI Act* – means Canada's response to AI-driven harms continues to be distributed across multiple, disconnected legislative instruments rather than anchored in a coherent framework.

---

11 Attard-Frost highlights several of these controversies here: https://montrealethics.ai/the-death-of-canadas-artificial-intelligence-and-data-act-what-happened-and-whats-next-for-ai-regulation-in-canada/

12 At the time of writing, the Act remains in the House of Commons in its report stage. A summary of key legislative changes is available here: https://www.justice.gc.ca/eng/csj-sjc/pl/c16/index.html.

13 Carney's announcement is available here: https://www.pm.gc.ca/en/news/news-releases/2026/06/04/prime-minister-carney-launches-ai-all-canadas-new-national-artificial.

14 *Canada's National Artificial Intelligence Strategy: AI For All.* Available here: https://ised-isde.canada.ca/site/ised/en/canadas-national-artificial-intelligence-strategy-ai-all.

(Voluntary) Responsibility in AI Development

While both the UK and Canada are lacking dedicated AI-focused legislation, regulatory bodies in each jurisdiction have released voluntary guidance for responsible development, deployment, and use. Within the UK, government guidance focuses on AI adoption by those who work in government or in the public sector, particularly those who develop, manage, analyse, or regulate AI systems deployed within the UK. Its *Data and AI Ethics Framework*[15] (2025) encourages transparency, accountability, privacy, and fairness in development and deployment, as well as larger reflections on societal impact, environmental sustainability, and safety of models. Similarly, the government's *Responsible AI Toolkit*[16] (2024) contains several resources to promote responsible innovation aimed at organisations and practitioners seeking to create and deploy AI systems. However, neither instrument carries legal force. The UK's pro-innovation, principles-based approach deliberately avoids binding obligations on developers, leaving significant discretion to industry actors and creating limited recourse for those harmed by AI systems that fall outside existing regulatory remits.

Innovation, Science, and Economic Development Canada published a *Voluntary Code of Conduct on the Responsible Deployment and Management of Advanced Generative AI Systems* in 2023. Aimed at organisations and service providers who develop, deploy, and operate AI systems in Canada, the Code of Conduct focuses on general purpose models that produce content, including ChatGPT and Midjourney. It offers clear guidance to signatories around how to achieve safety through risk assessments and mitigation measures, safeguard fairness and equity, ensure transparency in decision-making and evaluation, and secure against cyberattacks, framed as actionable measures that are divided into responsibilities based on whether the system is for public use, and whether each principle is actioned by developers and managers. Its design is also positive step forward in transparency and accountability, as a list of signatories is publicly visible at the bottom of the Code of Conduct website.

Though Codes of Conduct can help ensure core ethical principles for design and deployment are upheld, the *voluntary* nature of Canada's code for responsible development inherently limits its effectiveness, as it puts the onus on organisations, platforms, and developers to adhere to and abide by conduct based on good will alone. Without existing legislation to codify responsible design and risk minimisation as necessities, voluntary solutions minimise both the legal and ethical imperative to guard against harm at all stages of development and deployment.

This gap is particularly acute given that voluntary commitments place the greatest burden on smaller developers and civil society actors with limited capacity to engage with complex guidance frameworks, while larger platform operators with the resources to engage may selectively apply principles where commercially convenient. A binding legislative baseline – specifying minimum obligations around harm assessment, transparency, and redress – is needed to ensure accountability is structural rather than reputational.

---

15 Available here: https://www.gov.uk/government/publications/data-ethics-framework/data-and-ai-ethics-framework#further-resources.
16 Available here: https://www.gov.uk/government/collections/responsible-ai-toolkit
17 Available here: https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

Recommendations for Policy and Practice

AI-driven harms remain a pressing issue in both British and Canadian jurisdictions, as they indelibly impact digital participation, modes and scale of different forms of extremism, target vulnerable populations, and (re)produce existing forms of social injustice. Building out from the work done across each of the four case studies, this section sets out key recommendations for addressing AI-driven harms. Recommendations speak directly to service providers and policy sectors, advocating for changes to ensure responsible AI development and deployment within the UK and Canada.

Tech-Focused Recommendations

Transparency in both source and scope of training data used for individual models.

Outputs across case studies often (re)produced and extended known forms of harm, particularly through the adoption and remixing of visual aesthetics with stereotypes and discrimination. Importantly, these harms can be produced both willingly (direct user requests for harmful and dehumanising content) and unintentionally (where AI systems generate biased outputs without direct instruction). Greater transparency around the types of training data used to develop specific models is necessary to uncover how biased and harmful content may be used within the development process, subsequently impacting generative outputs.

Ensure model safety through robust pre- and post-launch safety testing.

While pre-deployment testing is essential to mitigate risks, visual signifiers of representational, implicit harms and extremism are likely to shift across time. Motivated users will find ways to bypass guardrails prohibiting harmful content generation, necessitating both strong proactive thinking and swift response. Retroactive safety testing and tweaks can help ensure guardrails are effective in addressing unforeseen shifts in how extreme and harmful content manifest in digital space.

Conceptualise harm as a design problem, rather than solely an issue of misuse.

The ability to bypass safety measures that guard against harmful content generation is not solely an issue of misuse at the user level; instead, these may be indicative of design flaws that can be exploited by individual users. Platform affordances and training data also contribute to how AI-driven harms manifest in digital ecosystems, intersecting with system design and deployment. Hashtags and tagged audio amplify the reach of AI-generated content across platforms for both far-right and male supremacist content. Additionally, the integration of Grok into X affords users quick access to Grok's image and video generation capabilities that can be used to produce harmful content at scale, often outpacing removal efforts.

Mandatory labelling and watermarking of AI-generated content by developers and platforms.

The scale of AI-generated media is outpacing efforts for manual flagging and removal. Additionally, outputs are increasingly lifelike, making it difficult to distinguish synthetic visuals from genuine, human-designed ones. As such, it is imperative that developers and platforms clearly identify AI-generated content. Automated labelling of AI-generated content has been integrated into Meta and TikTok from late 2024, both for visuals created using in-platform AI editing tools and content imported from other platforms (Hern, 2024). CNN and The Guardian provide clear labels on images that are AI-generated, superimposed onto visual content itself and within captions. These labels are highly visible and clearly signpost the reader to synthetic content, both in the positioning and the colours, text, and symbols used. Model-specific watermarks should be mandatory within model outputs, affording both traceability and transparency for AI-generated content.

Additionally, watermarking obligations should extend to hybrid content combining real and AI-generated footage, not only wholly synthetic outputs. This format is being actively exploited to confer credibility while evading detection, and automated systems currently struggle to classify partially synthetic material.

Expand automated moderation efforts to account for context over rule-based flagging for keywords and symbols alone.

While some forms of harm are linked to key words and known symbols, this research highlights that harms are increasingly implied and context-specific, challenging a reliance on overt signifiers of harm and extremism often used for parameters in content moderation. Without accounting for wider context, content that harms through implication or representation – rather than explicit depictions of extremist imagery or violence – can be missed, leaving noticeable gaps in flagging and removing racist, white supremacist, male supremacist, queerphobic, and sexually explicit content, as well as calls to action and arms.

Automated tools for harmful content detection and flagging should aim to account for the wider sentiment of visual and text-based content online. For instance, Canada's SIGNAL Network developed ManoWhisper[18], a tool that incorporates full sentences in its approach to automated content flagging, helping to capture contextual signifiers of hate speech and misogyny that may be missed by approaches only flagging relevant key words. Similarly, SightEngine's Text Classification model provides a machine learning approach to moderation by considering in-context meaning. While not currently available within the platform's automated moderation features for visual content, expanding these capabilities to its visual moderation models could help limit the inaccuracies of rule-based flagging found within this research.

---

18 ManoWhisper provides an open-access database of transcribed manosphere podcast episodes, with each episode containing measurable misogyny and hate scores using automated classifications of misogyny, toxicity, and hate speech. More information about ManoWhisper can be found at: https://manowhisper.signalnetwork.org.

On platforms like TikTok, audio should be treated as an independent moderation vector. A single AI-generated track appeared in 556 videos in this research, with TikTok's audio-linking automatically aggregating all content using the same sound regardless of its political character. Moderation systems focused solely on visual or text-based content will consistently miss this mechanism.

Animated and anthropomorphic AI content warrants dedicated moderation categories. Research identified AI-generated animation using child-friendly aesthetics to encode misogynistic and racial messaging while evading detection. Generic hashtags such as #funny and #comedy ensure this material reaches wide audiences – including children – without triggering existing moderation thresholds.

Evaluate and moderate content on an ongoing basis within established harm taxonomy frameworks.

Responses to technologically-facilitated violence against women and girls should be embedded within broader AI governance frameworks rather than addressed in isolation. Existing legislative responses to intimate image abuse capture only one manifestation of a wider pattern of gendered harm. Policy frameworks should draw on established typologies of violence against women and girls to ensure AI governance addresses the full spectrum of technology-facilitated gendered harms. For instance, recent work from the Institute for Strategic Dialogue[19] establishes a harms taxonomy for technology-facilitated gender-based violence (TFGBV) that overlaps with targeted hate and violent extremism, moving response and regulation beyond a narrow focus on intimate images alone.

Incorporate culturally-relevant and socially-conscious perspectives of harm into design and development, both for models and automated moderation efforts.

SightEngine's categories privilege Western understandings of harm and extremism, resulting in depictions of diverse subjectivities (often Muslim men and women, along with other people of colour) being incorrectly flagged as threatening or extremists in visual moderation. Broadening understandings of harm, and building these into models and efforts for moderation, can guard against biased outputs and inaccurate flagging, while also ensuring that approaches to minimising harm and risk are proactive rather than reactive.

---

19 Available here: https://www.isdglobal.org/wp-content/uploads/2026/05/Addressing-the-intersection-of-misogyny-targeted-hate-and-violent-extremism.pdf.

Policy-Focused Recommendations

Develop dedicated AI legislation and reporting mechanisms, both in UK and Canadian contexts.

Dedicated AI legislation would facilitate a more robust response to AI-driven harms beyond existing patchworks of legislation. In addition to a clear focus on harms, framing legislation in terms of risk mitigation and management can help future-proof responses, orienting them toward unforeseen challenges as AI systems progress in scope and scale.

The EU's *AI Act* establishes a four-tiered, risk-based model for AI systems (unacceptable risk, high risk, limited risk, and minimal risk), where regulatory responses map directly onto risk levels of individual systems or use cases. Forthcoming guidance for providers to report serious incidents related to high-risk AI systems has been developed by the European Commission under article 73 of the *AI Act*, with the goal of creating early warning systems, establishing accountability for developers and providers, and ensuring rapid responses to mitigate potential harms[20].

Afford protections against algorithmic intensification and harmful platform architecture.

Across case studies, the architecture of mainstream and fringe platforms facilitated the spread of harmful AI-generated visual content, often with little effort from users themselves. Both extremist narratives and representational harms can be soft-sold to wider audiences through specific affordances (app-specific buttons, tabs, hashtags, use of trending audios, bookmarks, favourites, reposts, and likes) that elevate the visibility of content through algorithms that prioritise engagement metrics.

The *Online Safety Act* does contain provisions aimed at addressing "harmful" algorithms through enforcing provider-based risk assessments and transparency reports, though this is limited under the legislation's focus on illegal content and content that is harmful to children. Strengthening these protections, and extending them to implicit visual harms, is a necessary response to the increasing visibility and impact of this content across platforms.

---

20 The European Commission is set to impose an obligation for providers of high-risk AI systems to report serious incidents to national authorities from August 2026. The draft guidance and incident report documents were subject to public consultation in late 2025, with drafts available here: https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks.

Facilitate swift, clear, and collaborative responses to model-generated harms.

This was particularly effective in the aftermath of non-consensual intimate image generation through Grok, where the threat of joint action from British, Canadian, and Australian governments was effective in prompting changes to Grok's image and video generation capabilities[21]. However, a collaborative approach only works if governments and legislators take decisive and swift action against harms, lessening opportunities for platforms and service providers to create workarounds or operate without accountability.

Build intersectional understandings of 'harm' and 'violence' into existing policies protecting online safety.

Many of the harms highlighted across case studies are nuanced and intersectional: male supremacist harms often intersect with queerphobia and white supremacy, while far-right extremism often intersects with racism, xenophobia, anti-immigrant narratives, religion, class, and gender. However, the nuances in how certain subjectivities experience harm are vague within current legislative approaches to online safety. Additional concerns exist around the use of symbolic violence in genAI content, where fantasy, role-play, and symbolism intersect to engender layered representational harm to specific identities and social groups.

Expand focus beyond 'illegal' or explicitly violent and extreme content to incorporate harms from 'lawful but awful' material.

Existing legislation within the UK conceptualises harms according to legality and explicit violence or extremism, a framing that obscures the nuanced forms of harm highlighted within this research. Adopting a broader understanding of harm would be useful in extending regulatory responses. For instance, incorporating typologies of online harms established through academic research, publications from advocacy groups, and policy reports[22] would be useful in orienting the guidance provided within the *Online Safety Act* towards implicit, rather than explicit, manifestations of harm and violence.

---

21 Despite talks between the UK, Canada, and Australia about coordinating a collective response to X in early January of 2026, Canada rejected an outright ban, opting instead to focus on the domestic criminalisation of deepfake sexual abuse and holding perpetrators accountable (Oliver, 2026).

22 The World Economic Forum's Global Coalition for Digital Safety created a typology of online harms in 2023, outlining six categories of harmful content: threats to personal and community safety, harm to health and well-being, hate and discrimination, violation of dignity, invasion of privacy, and deception and manipulation. This guidance is available here: https://www.weforum.org/stories/2023/09/definitions-online-harm-internet-safer/.

Develop regulatory frameworks capable of recognising cumulative harm.

Current legislative approaches assess harm at the level of individual content, failing to account for how algorithmically curated sequences radicalise audiences over time. Platforms should be required to conduct cumulative risk assessments that account for patterns of content exposure, not only discrete instances of violative material.

Pursue binding international regulatory coordination.

Research demonstrates that radicalisation infrastructure can assemble transnationally through platform design alone, without intentional cross-movement coordination. Domestically bounded legislation is structurally insufficient for this threat. Governments should develop coordination mechanisms enabling cross-jurisdictional enforcement and jointly developed platform accountability standards.

Develop accessible guidance for frontline practitioners and civil society organisations working with communities disproportionately targeted by AI-driven harms.

Practitioners in counter-extremism, gender-based violence, and community safety need regularly updated resources for identifying and reporting AI-generated representational harms, alongside clearer referral pathways to regulatory bodies.

About the Authors

Dr. Allysa Czerwinsky is a Research Fellow in AI Trust and Security at the University of Manchester. Her work explores the links between AI, social injustice, and gender-based harms, with a focus on how male supremacism and misogynist extremism intersect with new forms of technology. Her research seeks to understand how synthetic images created and shared in far-right ecosystems visually (re)construct hateful rhetoric and explores how feminised AI companion apps (re)produce harmful male supremacist narratives in both their marketing and outputs. Her expertise is demonstrated in parliamentary evidence on misogyny, male supremacism, and AI's role in furthering far-right extremism, as well as policy and practitioner-based resources published by the *Centre for Research and Evidence on Security Threats* and the *Global Network on Extremism and Technology*.

Dr. Ashton Kingdon is a Lecturer in Criminology at the University of Southampton who specialises in technology-driven extremism and AI-facilitated harms. Their recent monograph *The World White Web* provides a comprehensive semiotic analysis of far-right visual propaganda and has been adopted by government agencies for counter-radicalisation strategies. Dr. Kingdon has provided parliamentary evidence on AI's role in the 2024 far-right riots and was an expert witness for the U.S. House of Representatives January 6th Committee on the role of Alt-Tech platforms in the Capitol insurrection. They have advised major technology platforms (Facebook, Google, TikTok, YouTube, Discord) on AI-driven content moderation and served as the Head of Technology at the Centre for Analysis of the Radical Right (2019-2021).

Dr Shana MacDonald is O'Donovan Chair in Communication and Professor in the School of Critical and Creative Humanities at the University of Waterloo. Dr. MacDonald is co-director of the Strategies for Intersectional Gender Justice, Networked Action, and Liberation Network (SIGNAL), an international group of scholars, activists, and policy makers addressing the problem of technology-facilitated gender-based violence. Her interdisciplinary research maps the rise of online hate, technology-facilitated gender-based violence, and disinformation online. She is the lead author of the forthcoming collection Solidarity as Method: Countering Networks of Injustice Through Collective Action (Bloomsbury). Her most recent book, *The Art of Memes in Feminist Digital Culture*, explores how memes operate within contemporary digital culture as beacons for online public discourse that push against dominant norms.

Dr. Karmvir Padda is a post-doctoral researcher in Sociology at the University of Waterloo funded by Mitacs Accelerate program and an emerging scholar on digital extremism, gender-based violence, and online harms. Her work uses computational and qualitative methods to trace how misogynistic and grievance-driven narratives circulate across extremist manifestos, alt-tech forums, and manosphere podcasts. A Joseph-Armand Bombardier CGS Doctoral Scholar and the only Canadian recipient of the 2025 Harry Frank Guggenheim Emerging Scholar Award, she has developed mixed-methods models that inform national policy on digital harms and radicalisation.

References

Allen, K. (2025). You talkin' to me? Algorithmic mirrors and chatbot radicalisation. *GNET*. https://gnet-research.org/2025/12/08/you-talkin-to-me-algorithmic-mirrors-and-chatbot-radicalisation/.

Amnesty International. (2023). *Driven into the Darkness: How TikTok Encourages Self Harm and Suicidal Ideation*. https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/#:~:text=The%20findings%20of%20a%20joint,Lisa%20Dittmer%2C%20Amnesty%20International%20Researcher.

Attard-Frost, B., Brandusescu, A., Widder, D. G., & Tessono, C. (2025, February 20). AI countergovernance: Lessons learned from Canada and Paris. *TechPolicy Press*. https://www.techpolicy.press/ai-countergovernance-lessons-learned-from-canada-and-paris/.

Badshah, N. (2026, March 31). Teenager died after asking ChatGPT for 'most successful' way to take his life, inquest told. The Guardian. https://www.theguardian.com/society/2026/mar/31/teenager-asked-chatgpt-most-successful-ways-take-life-inquest-told.

Baele, S. & Brace, L. (2024). AI extremism: Technology, tactics, actors. VOX-Pol. https://voxpol.eu/file/ai-extremism-technology-tactics-actors/.

Baker, C., Ging, D. & Andreasen, M. B. (2024). *Recommending toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers*. https://www.dcu.ie/antibullyingcentre/recommending-toxicity.

Balendra, S. (2025). Meta's AI moderation and free speech: Ongoing challenges in the Global South. *Cambridge Forum on AI: Law and Governance, 1*(e21), 1-19.

Banet-Weiser, S. (2018). *Empowered: Popular Feminism and Popular Misogyny*. Duke University Press.

Barthes, R. (1972). *Mythologies*. Hill & Wang.

Braun, V. & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology,* 3/2, 77-101.

Booth, H. (2026, January 19). What the numbers show about AI's harms. *TIME*. https://time.com/7346091/ai-harm-risk/.

Bratby, R. (2026). Is there a UK AI Act? UK AI regulation in 2026. *Bratby Law*. https://bratby.law/uk-ai-regulation-what-the-law-says/.

Browne, J. (2023). AI and structural injustice: A feminist perspective. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.) *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines*. Oxford University Press.

Browne, R. (2026, February 16). AI chatbot firms face stricter regulation in online safety laws protecting children in the UK. *CNBC*. https://www.cnbc.com/2026/02/16/ai-chatbot-firms-face-stricter-regulation-protect-children-uk.html

Buarque, B. (2025, July 16th). The weaponisation of AI: Visual storytelling of the Great Replacement conspiracy theory amid the Southport Riots. *GNET*. https://gnet-research.org/2025/07/16/the-weaponisation-of-ai-visual-storytelling-of-the-great-replacement-conspiracy-theory-amid-the-southport-riots/

Buarque, B. & Lewis, N. (2025). Algorithms at your service: Understanding how X's systems of recommendation likely fuelled the far-right riots in the United Kingdom by amplifying visual representations of racist conspiracy theories. *The British Journal of Politics and International Relations,* 1-26.

Burgess, M. & Varner, M. (2026, January 6). Grok is pushing AI undressing mainstream. *Wired*. https://www.wired.com/story/grok-is-pushing-ai-undressing-mainstream/.

Centre for Countering Digital Hate (2026a, January 22). *Grok floods X with sexualised images of women and children.* https://counterhate.com/research/grok-floods-x-with-sexualized-images/

Centre for Countering Digital Hate. (2026b, March 11). *Killer apps: How mainstream AI chatbots assist users planning violent attacks.* https://counterhate.com/research/killer-apps/.

Chatterjee, R. (2025, September 19). Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots. *NPR*. https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide.

Chow, A. R. (2026, January 12). Grok's deepfake crisis, explained. *TIME*. https://time.com/7344858/grok-deepfake-crisis-explained/.

Czerwinsky, A. (forthcoming). Generating 'new' propaganda: The European far-right's use of Generative AI in digital environments. Visual Politics of AI. Bristol University Press.

Czerwinsky, A. (2024). Misogynist incels gone mainstream: A critical review of the current directions in incel-focused literature. Crime, Media, Culture, 20(2), 196-217.

Depounti, I., Saukko, P., & Natale, S. (2023). Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society, 45*(4), 720-736.

European Commission. (2024). *Regulatory framework on artificial intelligence, Digital Strategy.* https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Gentleman, A. & Horton, H. (2026, January 11). 'Add blood, forced smile': how Grok's nudification tool went viral. *The Guardian.* https://www.theguardian.com/news/ng-interactive/2026/jan/11/how-grok-nudification-tool-went-viral-x-elon-musk.

Ghiurau, D. & Popescu, D.E. (2025). Distinguishing reality from AI: Approaches for detecting synthetic content. *Computers*, *14*(1).

Gillespie, T. (2024). Generative AI and the politics of visibility. *Big Data & Society,* April-June, 1-14.

Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. M*en and Masculinities, 22*(4), 638-657.

Hao, K. (2021, September 13). A horrifying new AI app swaps women into porn videos with a click. *MIT Technology Review*. https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/.

Hartzell, S. L. (2020). Whiteness feels good here: interrogating white nationalist rhetoric on Stormfront. *Communication and Critical/Cultural Studies, 17*(2), 129–148.

Heritage, F. (2023). *Incels and ideologies: Exploring how incels use language to construct gender and race.* Palgrave Macmillan.

Hern, A. (2024, 9 May). TikTok to auto-flag AI videos – even if created on other platforms. *The Guardian*. https://www.theguardian.com/technology/article/2024/may/09/tiktok-auto-flag-ai-videos-digital-watermarking.

Horowitz, J. (2024, October 1). TikTok's manosphere problem: Violent misogyny keeps going viral. Media Matters. https://www.mediamatters.org/tiktok/tiktoks-manosphere-problem-violent-misogyny-keeps-going-viral.

Internet Watch Foundation. (2026, February 2). *New AI child sexual abuse laws announced following IWF campaign*. https://www.iwf.org.uk/news-media/news/new-ai-child-sexual-abuse-laws-announced-following-iwf-campaign/

Jane, E. (2017). *Misogyny online: A short (and brutish) history*. Sage.

Johanssen, J. (2021). *Fantasy, online misogyny and the manosphere: Male bodies of dis/inhibition*. Routledge.

Kingdon, A. (2024). *The World White Web: Uncovering the Hidden Meanings of Online Far-Right Propaganda*. Palgrave Macmillan.

Kingdon, A (2026): Riots on Demand?: Digital Infrastructure, Narrative Laundering, and Far-Right Mobilisation After Southport. *Criminology and Criminal Justice* (Accepted/Forthcoming).

Klincewicz, M., Alfano, M. & Fard, A. E. (2025). Slopaganda: The interaction between propaganda and generative AI. https://arxiv.org/pdf/2503.01560.

Ko, S. (2026, May 11). Protecting Victims Act headed back to House of Commons for final debate. *iPolitics*. https://www.ipolitics.ca/2026/05/12/protecting-victims-act-now-awaits-third-reading-in-the-house-of-commons/.

Kuenssberg, L. (2025, November 8). 'A predator in your home': Mothers say chatbots encouraged their sons to kill themselves. *BBC*. https://www.bbc.com/news/articles/ce3xgwyywe4o.

Kulkarni, A. (2026, February 20). OpenAI had banned account of Tumbler Ridge, B.C., shooter months before tragedy. *CBC News*. https://www.cbc.ca/news/canada/british-columbia/openai-tumbler-ridge-shooter-ban-9.7100497.

Lavie-Driver, N. & van der Linden, S. (2025). Social media, AI, and the rise of extremism during intergroup conflict. *Front. Soc. Psychol, 3*.

Lee, J. J. (2026). "Who is sexually harassed? A python code haha": imaginaries of a post-violent AI world. *Feminist Media Studies, 26*(2), 475–490.

MacDonald, S., Wiens, B. I., Ruest, N., & Padda, K. K. (2026). "All You Have to Do is Publicly Execute a Few Women Who Have Lied": Mapping Online Misogyny and Feminist Digital Counterprotest in the Post-pandemic Landscape. *Women's Studies in Communication, 49*(1), 81–109.

McGlynn, C., McDermott, Y., MacDonald, S., Toparlak, R. T., Tarrant, F. & Treacy, S. (2026). Invisible no more: How AI chatbots are reshaping violence against women and girls. https://e87dab74-be98-4bb1-83c5-05251d2bc6f4.usrfiles.com/ugd/e87dab_06a7f0801de549689c294d42e0478a3c.pdf.

Metropolitan Police. (2025). *Predicted, Estimated and Attendance for the Unite the Kingdom and Counter-Demonstrations*. https://www.met.police.uk/foi-ai/metropolitan-police/disclosure-2025/september-2025/predicted-estimated-attendance-unite-the-kingdom-counter-demonstrations/

Metz, C. (2026, March 24). OpenAI is shutting down Sora, its AI video generator. The New York Times. https://www.nytimes.com/2026/03/24/technology/openai-shutting-down-sora.html.

Meyer, R. (2025). Images from images: Generative AI and the reconfiguration of the 'photographic. *Photography and Culture, 18*(3-4), 281-290.

Ofcom. (2026). *Investigation into X, and scope of the Online Safety Act*. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/investigation-into-x-and-scope-of-the-online-safety-act

Oliver, M. (2026, January 11). Starmer fails to get Canada's backing over Musk AI row. *The Telegraph*. https://www.telegraph.co.uk/business/2026/01/11/starmer-fails-to-get-canadas-backing-over-musk-ai-row/

Park, Y. S. (2024). White default: Examining racialised biases behind AI-generated images. *Art Education, 77*(4), 36-45.

Parliament of Canada. (2026). *Bill C-34: Safe Social Media Act,* 1st Reading, House of Commons, 10 June. Ottawa: Government of Canada. https://www.canada.ca/en/canadian-heritage/news/2026/06/government-of-canada-introduces-legislation-to-make-social-media-services-and-ai-chatbots-safer-for-children.html

Pennink, E. (2023, October 6). Queen assassin case exposes 'fundamental flaws' in AI. *The Independent*. https://www.independent.co.uk/news/uk/crime/ai-queen-jaswant-singh-chail-b2425092.html

Prażmo, E. (2020). Foids are worse than animals. A cognitive linguistic analysis of dehumanising metaphors in online discourse. *Topics in Linguistics, 21*(1), 16-27.

Radovanović, B. (2022). TikTok and Sound: Changing the ways of creating, promoting, distributing and listening to music. *INSAM Journal of Contemporary Music, Art and Technology, 9*(1), 51–73.

Sircar, A. (2024, October 18). X's latest content findings reveal troubling trends in AI moderation. *Forbes*. https://www.forbes.com/sites/anishasircar/2024/10/18/xs-latest-content-findings-reveal-troubling-trends-in-ai-moderation/.

Solea, A. (2024). Hiding in plain sight: How the 'newgen' misogynistic incel content creators escape moderation on TikTok. *GNET*. https://gnet-research.org/2024/05/07/hiding-in-plain-sight-how-the-newgen-misogynistic-incel-content-creators-escape-moderation-on-tiktok/.

Stevenson, J. (2025). Technology evolves the tactics: Preparing for the rise of terrorist AI harms. *CREST Research*. https://crestresearch.ac.uk/comment/technology-evolves-the-tactics-preparing-for-the-rise-of-terrorist-ai-harms/.

Turk, V. (2023, October 10). How AI reduces the world to stereotypes. *Rest of World*. https://restofworld.org/2023/ai-image-stereotypes/

UK Government. (2023). *Overview of expected impact of changes to the Online Safety Bill*. https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/overview-of-expected-impact-of-changes-to-the-online-safety-bill#:~:text=19.,related%20to%20user%2Dgenerated%20content.

Wells, G. (2026, February 21). OpenAI employees raised alarms about Canada shooting suspect months ago. *Wall Street Journal*. https://www.wsj.com/us-news/law/openai-employees-raised-alarms-about-canada-shooting-suspect-months-ago-b585df62?mod=e2twd.