



Social Media and Offline Violence

Research Report

Exploring how harmful social media content may contribute to real-world violence among young people in London, to inform earlier and more effective prevention.



June 2026

VRU

MAYOR OF LONDON
VIOLENCE REDUCTION UNIT

M&CSAATCHI
WORLD SERVICES



SOUK
SHOUT OUT UK



Publication Date: June 2026.

Suggested Citation: Babic, A., Williams, M., Luker, J. (2026). Social Media and Offline Violence. Full Report. London: London Violence Reduction Unit.



Contents

Acknowledgements..... 4

Executive Summary 5

1. Introduction15

2. Approach and Methodology17

3. Defining the Hazards and harms of Interest.....24

4. Social Media as a mirror and motor of harm.....29

5. Interpersonal Violence and Violent Content43

6. Violence Against Women and Girls and Online Misogyny56

7. Who is most at risk, and what shapes vulnerability? 74

8. Concluding Implications and Recommendations.....85

Reference List94

Technical Annex.....99

Acknowledgements

This research was commissioned by the Mayor of London's Violence Reduction Unit (VRU) and would not have been possible without the contributions of many individuals and organisations.

We are deeply grateful to the young people who participated in this study, both as research participants and as members of our consultation workshops. Their willingness to share their experiences, challenge our thinking, and shape the direction of this work is reflected throughout the report. We are also grateful to the practitioners, teachers, youth workers, and other professionals who gave their time and expertise.

We would like to thank the VRU's Online Harms Steering Group and Young People's Action Group for their guidance and critical engagement throughout the project, and the organisations who supported this work: Everybody's Invited, Catch22, Safer London, and the James Ross Hunter Foundation, Islington Youth Justice Service, Lancaster Youth Club, Roadworks Media, London Youth and Islington Youth Councillors.

The findings and recommendations in this report are grounded in the experiences and expertise of the young people and professionals who took part. We hope they support more effective violence prevention for young Londoners.



ACKNOWLEDGEMENTS

We are grateful to everyone who helped shape this work through their lived and professional experience.

Credit: VRU Young People's Action Group, 2025

Executive Summary

Research headline findings and
implications for violence prevention



Executive Summary

The purpose of this research was to investigate how social media and wider online environments contribute to two forms of violence affecting young people aged 12-24 in London: interpersonal violence between young people, and violence against women and girls (VAWG).

Working backwards from these two forms of violence, the study examined whether specific forms of online content, violent content for interpersonal violence, and misogynistic content for VAWG, shaped offline violence, and in what ways. It asked three questions: *to what extent does harmful online content contribute to offline violence? How does this happen, and through what mechanisms? And which young people are most at risk, and what makes violence more or less likely?*

While previous research has established that young people's exposure to harmful online content is widespread, the nature of the link between online activity and offline violence – how it operates, for whom, and under what conditions – remains underexplored, and a more localised evidence base is needed to inform place-based prevention in London. This research was therefore designed as an in-depth, intersectional, place-based study, prioritising depth of insight over breadth of sample to generate London-specific evidence that can directly inform local prevention.

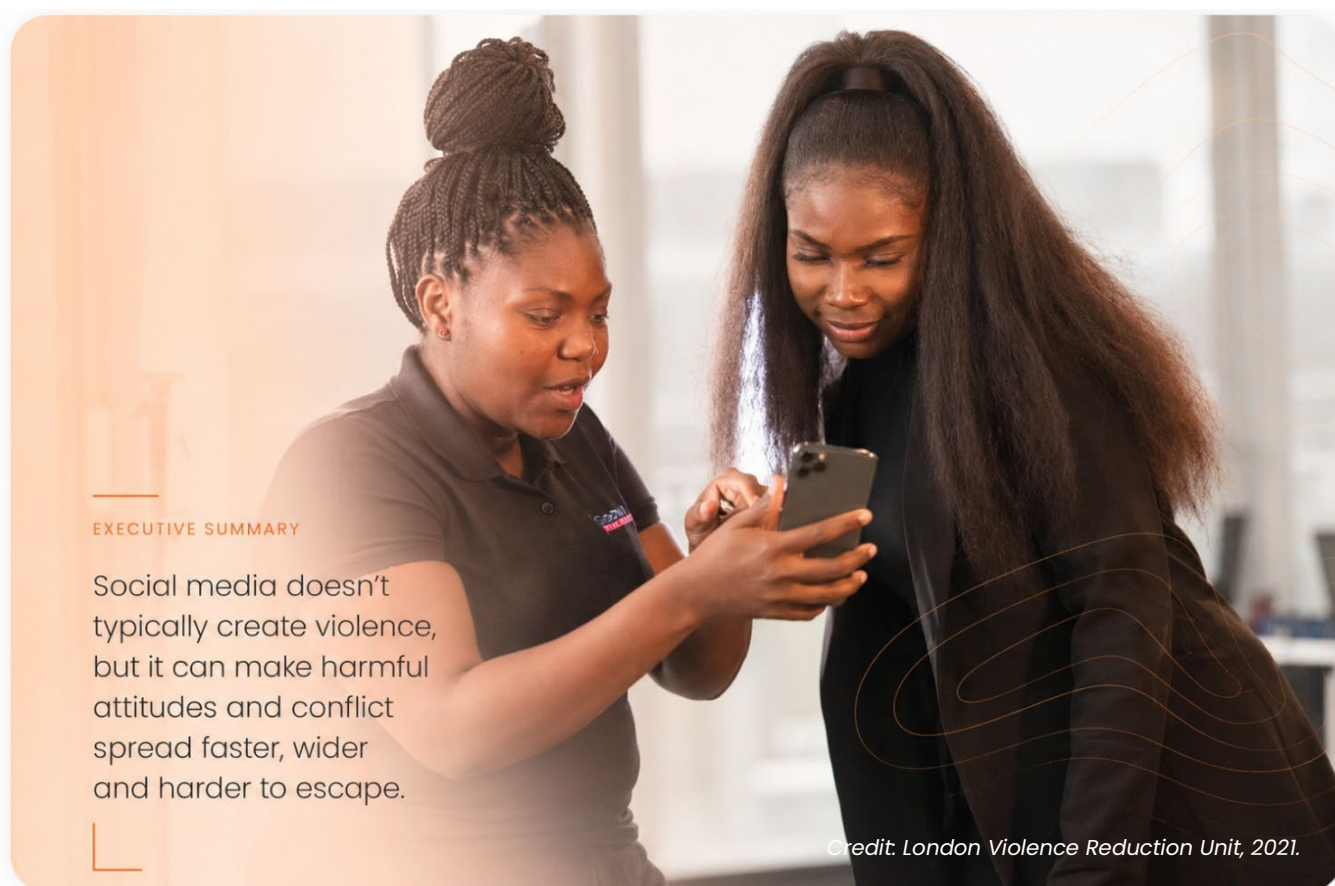
The research was a collaborative undertaking between M&C Saatchi World Services, nisien.ai, and Shout Out UK, using a mixed-methods approach. The large-scale 'big data' analysis, conducted by nisien.ai, collected more than 36,000 classified harmful social media posts on X over a 12-month period, alongside daily Metropolitan Police records of violent incidents covering the same period. This supported city-wide analysis of associations between online activity and recorded violence across London from September 2024 and April 2025, when a measure of general platform activity was available to ensure the analysis accounted for how busy X was on any given day. X was selected as the primary platform for the quantitative strand as it provides the strongest combination of accessible public conversation data, post-level metadata, and location signals available at the scale and city-level resolution this analysis required. The public conversation captured on the platform serves as a measurable proxy for the wider online environment young people inhabit, and if this volume of harmful discourse is detectable on a single public platform, the total environment across the platforms young people use most is likely to be substantially larger.

In parallel, in-depth qualitative research led by M&C Saatchi World Services and Shout Out UK drew on 22 interviews and 8 focus groups with young people, practitioners, teachers, and youth workers (n=46) to explore how online environments shaped offline violence in practice: through what mechanisms, in what peer and relational contexts, and with what consequences for the young people involved. Four anchor case studies provided detailed accounts of how online activity connected to offline harm, tracing escalation from initial online exposure through to physical

violence, relationship abuse, and exploitation. Qualitative findings are presented throughout as illustrative of the patterns and mechanisms identified rather than as statistically representative. Young people were active contributors throughout, shaping the research design and co-creating accessible outputs published alongside this report. The findings and recommendations are organised through a socio-ecological model, reflecting the study's consistent finding that harm operated across individual, relational, community, institutional, and structural levels simultaneously. An intersectional lens runs throughout: gender, race, age, neurodivergence, and socioeconomic position did not create separate vulnerabilities but intersected to shape who was exposed, how harm was experienced, and whether protection was available. Full methodological detail is set out in Section 2 and the Technical Annex.

Key findings at a glance

- 01 **Social media did not typically create violence or abuse, but it acted as both a mirror and motor,** reflecting existing conflict and harmful attitudes while making them faster, wider-reaching and harder to escape.
- 02 **On days when misogynistic content made up a larger share of public conversation on X, police-recorded violence against women and girls in London also increased.** The pattern reflected the proportion of misogynistic content, not raw post counts, and it did not appear for general violent crime, suggesting a specifically gendered association.
- 03 **Interpersonal violence and VAWG are not one generic 'online problem', and they operated through different mechanisms:** Interpersonal violence escalated quickly, driven by audience expansion, public humiliation, and peer pressure to respond physically. VAWG developed more gradually, as misogynistic attitudes were normalised through repeated exposure, image-based abuse became embedded in peer culture, and platform design facilitated the exploitation of younger users. Each requires a different prevention response.
- 04 **What young people encountered online did not stay online:** violent content desensitised young people to graphic material and lowered the threshold for is considered 'serious violence', while misogynistic content reshaped what they treated as acceptable in relationships, peer interactions, and classroom behaviour.
- 05 **The most marginalised young people were often the most digitally dependent and the least protected.** Risk was not fixed, the same young person could become more vulnerable during periods of family instability, exclusion, or isolation, and less so when support improved.
- 06 **Peers shaped whether online incidents escalated into violence or were defused.** The same group dynamics that spread conflict and normalised abuse could also interrupt them – when young people had the tools and social permission to challenge what was happening.



EXECUTIVE SUMMARY

Social media doesn't typically create violence, but it can make harmful attitudes and conflict spread faster, wider and harder to escape.

Credit: London Violence Reduction Unit, 2021.

Does social media contribute to offline violence?

Social media appears to intensify and sustain violence and abuse rather than create them in isolation. Across the research, social media was rarely described as the sole starting point for violence or abuse. More often, it intensified and sustained what already existed, changing how conflict and harmful attitudes played out by making them more visible, harder to escape, and more likely to escalate.

For Interpersonal violence, disputes, rivalries, and 'beef' typically existed before social media became central. What changed online was the scale: arguments that might have once stayed between two people became public events, drawing in wider peer groups, generating pressure to respond, and persisting long after the original disagreement was over.

For VAWG on the other hand, social media created environments where gendered abuse became difficult to contain or escape. Intimate image-based abuse, sexual shaming, controlling relationship behaviour, and the routine objectification of girls and young women circulated through group chats, algorithmically recommended feeds, and peer networks, often as part of everyday platform use rather than something that was actively sought out.

The large-scale data analysis supported what young people and practitioners described. On days when misogynistic content made up a larger share of city-wide public conversation on X, police-recorded VAWG in London also increased. This pattern held up after extensive checks: it appeared only on the same day, did not appear for general violent crime, and was driven by the relative concentration of misogynistic content rather than overall online activity. The analysis was deliberately conducted across victims and suspects of all ages to capture the full population-level picture, and the finding should therefore be read as a cross-societal city-wide association rather than one specific to young people.

Two exploratory findings add further context. Within this all-ages model, higher-count VAWG days tended to involve a younger recorded case mix, with the daily share of under-25 victims and suspects positively associated with total VAWG counts. This is a descriptive feature of the data rather than a separate test of whether the misogyny association operates differently for younger people, but it tracks with the wider qualitative finding that young women in London were experiencing the online misogyny environment as a meaningful condition of their daily lives. Exploratory analysis further showed the association was stronger in cases where the suspect was known to the victim, consistent with the qualitative finding that violence and abuse typically played out within existing relationships rather than between strangers.

How does social media contribute to offline violence?

Social media acts as a mirror and a motor for existing conflict and attitudes.

Three features of social media help explain why it amplifies existing conflict and attitudes rather than simply hosting them:

1. **Social media removes any reliable cooling-off period.** Conflict did not end when school ended, and reputational damage did not fade when a young person left the setting where it occurred. For interpersonal violence, this kept tensions active across weeks or months, while for VAWG it meant sexualised content and hostile commentary followed young women across platforms and settings.
2. **Online audiences change the stakes.** A disagreement or act of abuse that might once have involved a small number of people could now be witnessed, discussed, and recirculated by an entire peer group. It was often this audience effect, not the original incident, that drove escalation, because young people felt pressure to respond publicly or feared losing face.
3. **Harmful content was not occasional or exceptional.** It was part of everyday online life, delivered by algorithms alongside entertainment and social content. Public events, high-profile figures, and media coverage did not create this harmful environment, but they intensified it at particular moments, sharpening the volume and tone of online conversation around specific incidents and public debates, like the Gisèle Pelicot verdict and Andrew Tate trial coverage. Over the 12-month monitoring period, more than 260,000 posts were collected and analysed on X, of which 25,509 were confirmed as misogynistic and 10,978 as

violence-related. Although most young people in the research did not identify X as their main platform, this volume is consistent with broader research showing that harmful content is widespread across the platforms young people use most, including TikTok, Instagram, Snapchat, and YouTube Shorts. The total environment of harmful content young people are exposed to is therefore likely to be substantially larger than what is measurable on X alone.

Repeated exposure shapes behaviour over time. The study found that what young people see online does not stay online. Repeated exposure to harmful content appeared to shift attitudes and behaviour in ways that showed up in classrooms, peer interactions, and relationships. But the process worked differently for different types of content:

- For violent content, the main effect was desensitisation. Young people and practitioners described how graphic footage was increasingly consumed as entertainment, with little emotional response. Over time, this raised the threshold for what was considered serious and reduced young people's willingness to report what they were seeing.
- For misogynistic content, the main effect was normalisation. Rather than dulling emotional reactions, repeated exposure reshaped what young people treated as acceptable – particularly around relationships, control, and gender roles. Teachers observed boys using language and assumptions drawn from manosphere content as ordinary conversation, often without recognising it as harmful.

These two types of content also circulated differently: violence-related content was overwhelmingly reshared (80%), spreading around specific events or incidents, while misogynistic content was more often originally authored (46%), appearing continuously as everyday commentary, humour, and advice rather than spiking around particular moments. This distinction matters for prevention, the spread of violent content is more responsive to friction on sharing and platform design, while the normalisation of misogynistic attitudes requires norm disruption, critical literacy, and relational work before attitudes harden into behaviour.

Interpersonal violence and VAWG follow overlapping but distinct mechanisms. A central finding of this research is that interpersonal violence and VAWG are not one generic 'online problem'. While they share common features, they operate through distinct mechanisms that require different prevention responses.

For interpersonal violence, the research identified three main mechanisms:

1. **Sustained hostility:** disputes remained active online over days or weeks, keeping tensions alive long after the original disagreement.
2. **Rapid escalation:** group chats and public posting drew in wider audiences and increased pressure to respond physically.

3. **Filming and circulation:** incidents became widely viewed content that generated further conflict.

For VAWG, four mechanisms were identified. In contrast to interpersonal violence, these operated more gradually and cumulatively:

1. **Normalisation:** repeated exposure reshaped expectations around control, gender roles, and what counted as acceptable relationship behaviour.
2. **Image-based abuse:** a specific act triggered ongoing reputational damage that followed young women across platforms and settings.
3. **Peer sexual objectification:** routine commentary and judgement about girls' bodies and behaviour became an ordinary part of online peer interaction.
4. **Exploitation enabled by platform features:** particularly affecting younger users, where weak age verification and contact from older users created risk.

Who is most at risk, and what shapes whether violence occurs?

Risk increases when online content interacts with vulnerability, not from exposure alone. A core finding was that the same online content did not affect every young person equally, and that exposure alone did not determine whether violence followed. What mattered was the interaction between what young people encountered online, the peer and platform contexts through which it travelled, and the support and protection available around them. Risk was more concentrated where young people faced:

- greater digital dependence with fewer offline supports
- less access to trusted adults
- younger age and lower capacity to critically assess what they were seeing
- neurodivergence alongside gaps in safeguarding provision
- gendered and racial inequalities in both exposure and the responses available to them.

Vulnerability was not fixed, it emerged through the combination of circumstances a young person was facing at a given point in time, which means prevention needs to be targeted and responsive to changing conditions rather than based on static categories of risk.

Peer dynamics shape how conflict escalates, and how it can be interrupted.

Across both interpersonal violence and VAWG, peers were central to how online incidents unfolded. Sharing, public commentary, and side-taking could intensify conflict significantly. Nevertheless, peers were not only part of the problem, but the solution as well; practitioners also identified examples where peers challenged active bystander behaviour, and non-judgemental conversations helped slow situations down. This makes peer context one of the clearest points for prevention, not because peer influence is inherently harmful, but because it was one of the main ways online interactions acquired social meaning and momentum.

Implications and recommendations

The findings point to a more precise understanding of how online environments shape offline harm, and where prevention is likely to be most effective. Generic online safety approaches will not address the mechanisms through which violence develops. What is needed instead is a coordinated approach combining safer platform design, stronger relationships around young people, and interventions tailored to the different pathways identified. The full set of recommendations is set out in Section 8, with headlines outlined below.

1. **Address platform design and the wider environment, not just moderate content after the fact.** Recommendation systems, group chat structures, weak age verification, and frictionless sharing shaped what young people saw and how quickly conflict spread. The evidence points to safer platform design by default, including slower circulation of harmful content in the first 24 to 48 hours, greater user control in messaging environments, and stronger age assurance built into platform architecture rather than relied on at the point of access. These structural changes to the online environment are likely to matter more than restrictions placed on individual users, such as age-based access limits or screen time controls, which place the burden of managing harm on young people themselves rather than on the platforms that shape their exposure to it.
2. **Invest in trusted relationships and rapid response capacity as the foundation of prevention.** Early warning depended on relationships rather than monitoring, because much of the escalation took place in private channels. This points to investment in youth workers, mentors, and practitioners with the relational depth and digital literacy to recognise when harm is building, alongside coordinated response capacity across youth settings so professionals can act within the first 24–48 hours.
3. **Equip peers to interrupt harm, not just adults to respond to it.** There is a critical window between public exposure and escalation where peers are often more influential than adults. Peer-led approaches should equip young people to challenge harmful norms and de-escalate conflict, without placing the full burden of prevention on them.
4. **Align education with the realities of how young people experience online harm.** The 2025 RSHE statutory guidance strengthens the curriculum around misogyny and harmful online content. This research points to where implementation should go further: critical literacy for young people of all genders on how algorithmic recommendation delivers gendered relationship scripts and identity pressures, including how self-improvement content carries controlling relationship scripts and how oversimplified narratives about toxic masculinity can obscure the structural pressures the manosphere places on young men; education that moves beyond the stranger danger frame; and modules that recur throughout secondary school rather than being delivered as one-off sessions. Framed in public health terms, these programmes should treat the misperceptions, pressures, and norms

circulating online as a population-level health issue rather than as a problem of individual attitudes.

5. **Improve responses to image-based abuse and VAWG.** Image-based abuse was identified as a high-impact form of abuse with long-term consequences. Disclosure routes should be designed around what the young person needs, combining therapeutic support, content removal assistance, and legal guidance without re-traumatisation or requiring immediate formal reporting. Training for school staff should specifically address victim-blaming and racialised responses.
6. **Target support where cumulative risk is highest.** Commissioning should ensure that targeted provision reaches younger adolescents, neurodivergent young people, and those without trusted adult support. These groups were disproportionately affected across both types of violence and least likely to benefit from existing universal prevention.

Further research priorities

The study also identifies several areas for further work, including:

- Improving independent researcher access to social media platform data to enable systematic, longitudinal research across the major platforms young people use, including TikTok and Instagram, and where possible to more private platforms such as Snapchat.
- Tracking how the desensitisation and normalisation effects develop longitudinally, particularly during key developmental transitions such as early adolescence.
- Developing ethical approaches to studying private channel dynamics, with clear public health standards for assessing platform environments.
- Testing interventions tailored to the specific mechanisms identified in this study.
- Examining the growing role of AI-generated content, including deepfakes.

SECTION ONE

Introduction

Why this research was needed and
what it set out to answer



1. Introduction

The scale of young people's exposure to harmful online content is increasingly clear. More than 70 per cent of young people report seeing real-world violence shared on social media (Youth Endowment Fund [YEF], 2024), and more than a quarter of women who experience online abuse also report receiving threats of sexual or physical violence (Amnesty International UK, 2025). Children exposed to misogynistic influencer content have been reported to be more likely to endorse physical aggression within relationships (Women's Aid, 2023), and recent work finds that following more manosphere influencers is associated with greater dehumanisation of women (Renstrom and Back, 2024). Young people already navigating trauma, exclusion, or unstable peer environments may be more susceptible to harmful online material (Elsaesser et al., 2021).

While previous studies have established that a link appears to exist, the nature of that link, how online content interacts with the circumstances young people are navigating, remains underexplored. Understanding these mechanisms is important because it begins to surface what might work to prevent harm. Much of the work has drawn on surveys and self-reported experiences, which are valuable in identifying patterns but tell us less about how dynamics unfold in practice. Existing studies have also tended to operate at national or international scale, with the YEF's research designed to be nationally representative and Amnesty International's investigation conducted across eight countries. Localised, place-based evidence of the kind required to inform prevention activity in London remains comparatively sparse. Questions also remain about which young people are most affected, and under what conditions, with younger adolescents and intersectional differences in exposure and impact less consistently represented across studies.

In response, this study combines large-scale 'big data' analysis of social media content and police records with in-depth qualitative case studies and youth participation to examine how online environments may contribute to violence affecting young people aged 12–24 in London. Rather than treating online content as a standalone cause of violence, the research examines how online environments interact with the relationships, peer dynamics, and wider social conditions that shape young people's lives in London. The study focuses on two priority harms: 1) Interpersonal violence affecting young people and 2) Violence against women and girls (VAWG). This report addresses three central research questions:

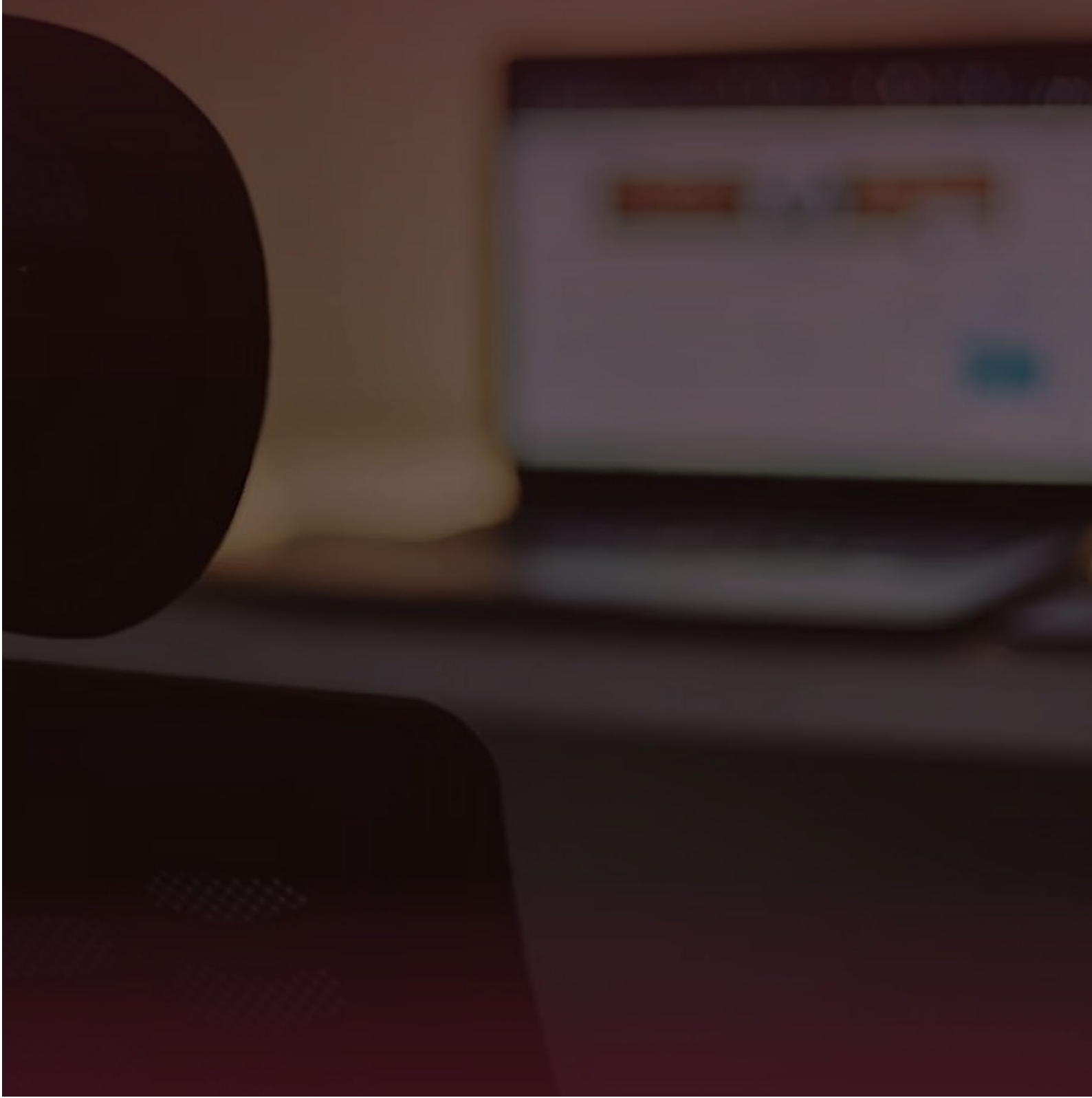
1. *How, and to what extent, does exposure to hazardous social media content contribute to, facilitate, or catalyse incidents of interpersonal violence and VAWG?*
2. *Which risk factors most influence whether such exposure translates into offline harm, and which young people are therefore most at risk?*
3. *What can organisations such as the VRU and its partners do to intervene earlier and more effectively to keep young people safe online?*

The sections that follow set out the methodology, define the two forms of violence under examination, and present the findings, first across both strands, then for interpersonal violence and VAWG in turn, and finally in relation to who is most at risk and what shapes vulnerability.

SECTION TWO

Approach and Methodology

Combining large-scale data analysis with
in-depth qualitative research across London

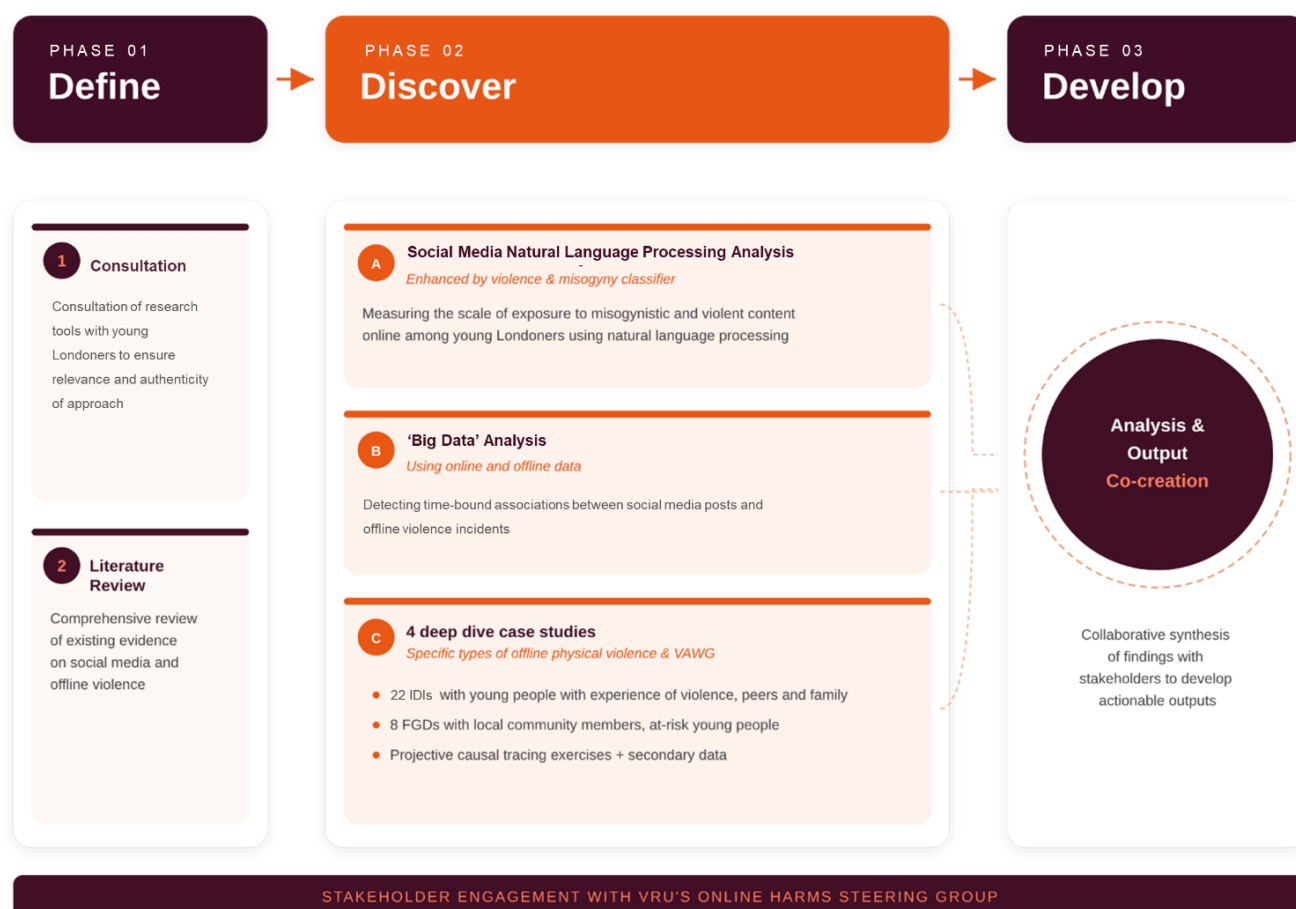


2. Approach and Methodology

The study was designed as a mixed-methods investigation combining large scale social media analysis with interviews and focus groups exploring the experiences of young people, practitioners, and educators in London.

The two strands were designed to be complementary, each answering a different part of the question. The quantitative strand explored the scale and prevalence of harmful discourse in the public online environment and tested whether shifts in the character of that environment were associated with changes in police-recorded violence across London. The qualitative strand explored how online-to-offline violence occurs: through what mechanisms, in what peer and relational contexts, and with what consequences for the young people involved. Combined, the two strands aim to move from identifying patterns to explaining how those patterns operate in practice (see Figure 1).

Figure 1. Overall Research Methodology¹



¹ The 'violence and misogyny classifier' referenced in Figure 1 are automated tools trained to read large volumes of social media posts and categorise them by content type. In this study, classifiers were used to identify which posts contained violent or misogynistic content from within the wider stream of online conversation.

2.1 Analytical framework: hazards, risks, and harms

The research is grounded in the Hazards–Risks–Harms (H–R–H) framework (DCMS, 2021), working backwards from observed harms to examine how specific forms of online content may function as hazards.

In this framework, hazards are understood as online content that has the potential to cause harm; risks are the individual, relational, and structural factors that shape vulnerability; and harms are the physical, psychological, or social consequences that may follow. Two types of online content are treated as core hazards: violent content as a hazard for interpersonal violence, and misogynistic content as a hazard for VAWG.

The framework recognises that exposure alone is insufficient to explain harm. The same content can have very different impacts depending on a young person's context, developmental stage, relationships, and wider social conditions. Following the Socio-Ecological Model (Bronfenbrenner, 1979; WHO, 2022), the study situates online harms across multiple interacting levels: the individual, relational, community, and societal/structural (see Figure 2). This approach avoids framing online harm as an individual failure and instead foregrounds systems, environments, and prevention opportunities.

Figure 2. Analytical Framework



2.2 The Social Media and 'Big Data' Analysis

The quantitative analysis (Figure 1: A and B), conducted by Nisien.ai, examined whether the character of public online discourse on a given day tracked with police recorded violence across Greater London. It brought together Metropolitan Police Service (MPS) records, London Ambulance Service (LAS) Data, social media posts collected from X across two streams (misogynistic content and violence related content, both geofenced to London, and controlled for general platform busyness).

Over a 12-month period (May 2024 to April 2025), more than 260,000 public posts were collected from X, all located within Greater London. These were sorted using automated classifier tools designed to identify violent and misogynistic content based on the language used. This produced a final dataset of 25,509 confirmed misogynistic posts and 10,978 confirmed violence-related posts. A small number of posts linked to hashtags associated with violent video content were also added manually, to capture material that the text-based tools would otherwise miss. Full detail on how these tools were developed and applied is provided in the Technical Annex.

This social media data was then analysed alongside MPS records and LAS data covering the same 12-month period across Greater London, to test whether days when harmful content made up a larger share of online conversation were also days when more violence was recorded. The core modelling drew on the period from September 2024 to April 2025, when a measure of general platform activity was available to ensure the analysis could distinguish between days when the platform was simply busier overall and days when harmful content specifically made up a larger share of what was being posted. The model also controlled for time trends and the demographic composition of recorded cases. An extensive set of robustness checks was applied. Full technical detail on classification methods, model specifications, and robustness results is provided in the Technical Appendix.

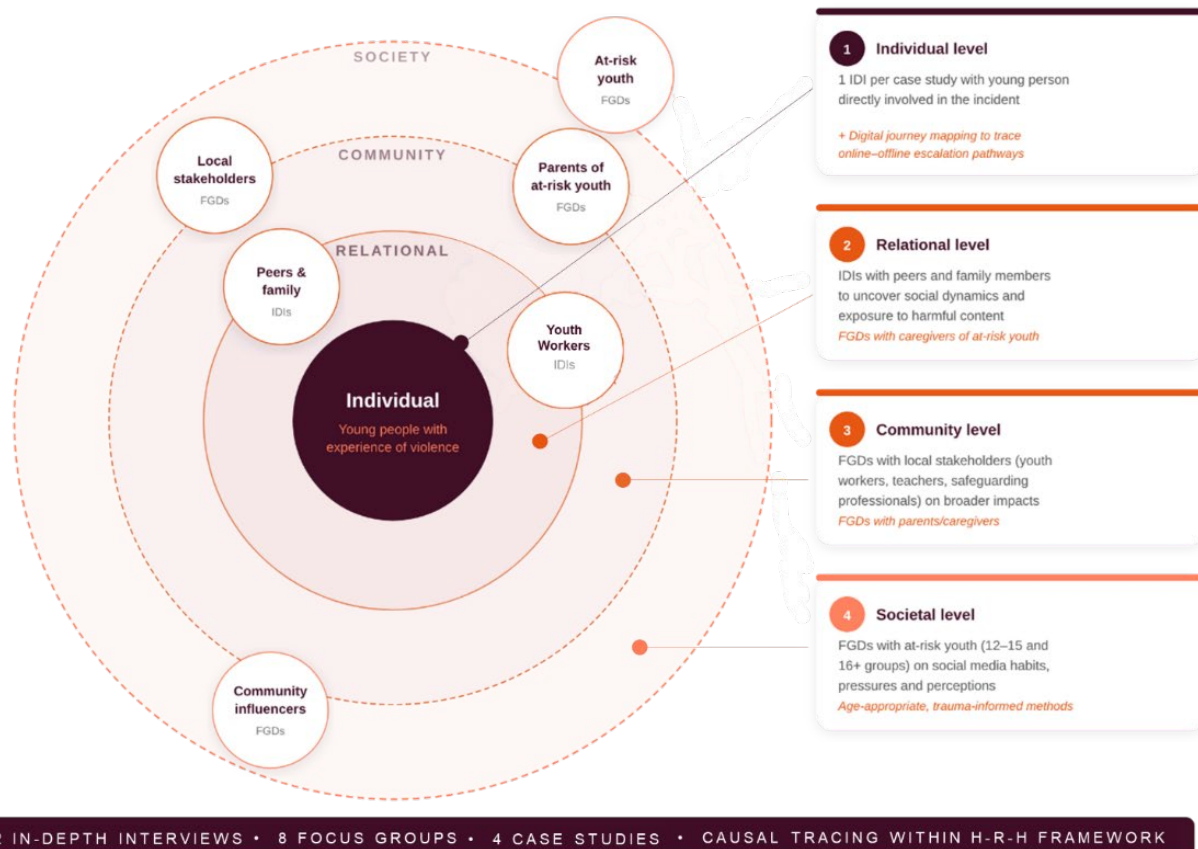
X was selected as the primary platform for the quantitative strand because it provided the strongest available foundation for the analysis this study set out to deliver: high-volume, time-resolved, city-level evidence on the relationship between public online conversation and recorded violence in London. Its combination of accessible public conversation data, post-level metadata, and location signals enabled robust, London-specific findings at a scale and resolution not currently achievable on other major platforms. The analysis cannot, however, observe private channels, visual content, or the platforms young people aged 12 to 24 reported using most, including TikTok, Snapchat, and Instagram. Its value lies in testing, at population level, whether the tone and composition of public online discourse tracks with offline harm, and the findings are interpreted throughout the report alongside the qualitative evidence about how and why those patterns may operate.

2.3 Case Studies, Interviews and Focus Groups

The qualitative research (Figure 1: C) was led by M&C Saatchi World Services and Shout Out UK. It was comprised of 22 in-depth interviews and 8 focus groups across London, involving young people aged 12-24, practitioners working in youth justice, VAWG services, education, and safeguarding, and teachers and school staff (n=46). A purposive sampling approach via youth justice and VAWG services recruited participants across the socio-ecological spectrum to ensure the research captured perspectives from those with direct lived experience of violence, those at heightened risk, peers, and the professionals working with them (Figure 3).

Four anchor case studies provided the most detailed accounts of the mechanisms connecting online environments to offline harm. These involved young people who had experienced incidents where online activity contributed directly to escalation, and their accounts are drawn upon extensively in Sections 4, 5, and 6 of this report. Each case study used a digital journey mapping exercise, working backwards from the incident of violence to trace the online and offline factors that contributed to escalation, including the content encountered, the platforms involved, the peer dynamics at play, and the wider relational and structural conditions surrounding the young person at the time. The wider qualitative sample of interviews was used to test and develop these pathways, with practitioner and peer accounts providing context for the mechanisms identified in the case studies and surfacing patterns across the dataset as a whole. The focus groups provided societal level reflections, with collaborative 'solutions labs' to explore recommendations (see Figure 3 for the full sampling strategy).

Qualitative data was analysed using reflexive thematic analysis, supported by NVivo software. Coding was developed both deductively, drawing on the research questions and the Hazards–Risks–Harms (H–R–H) and Socio–Ecological Model frameworks, and inductively, allowing themes to emerge from the data. Initial codes were reviewed and refined through discussion within the research team, with particular attention to convergence and divergence between the four anchor case studies, the wider interview and focus group sample, and the patterns identified in the quantitative strand. The pathway descriptions presented in Sections 5 and 6 are the product of this process, with each pathway grounded in evidence from multiple data sources.

Figure 3. Qualitative Sampling Strategy

2.4 Youth consultation and participation

Young people were not only research participants but active contributors to the study's design and interpretation. A consultation panel recruited through the VRU's Young People's Action Group, trusted charities, community organisations, schools, and youth justice services shaped the research direction, tools, and questions at the outset, ensuring the study reflected the realities of young people's online lives. In the final phase, the panel advised the research team on translating findings into accessible outputs including a short film, animation, infographics, and creative reflections. These co-creation outputs are published alongside this report.

2.5 Study limitations

Several important limitations of this study should be noted:

- **Platform coverage (quantitative data limited to X):** The quantitative analysis is based on public, text-based posts from X, which offered the most workable conditions for the scale, resolution, and geolocated coverage this modelling required. X is not the primary platform most young people in this study reported using, and the analysis cannot directly observe private messaging, group chats, disappearing message platforms, or visual and video content. The interpersonal violence pathway in particular operates largely through private channels,

and participants identified visual formats as the primary way they encountered harmful material. The absence of a quantitative signal for violence-related content should therefore be read as a reflection of where these dynamics take place rather than as evidence of no effect. The qualitative strand is essential in showing how patterns visible in public data operate within the platforms and channels young people predominantly use. Full detail on the data collection strategy, including how visual content was indirectly captured through keyword and hashtag queries, is provided in the Technical Annex.

- **Hazard selection and the hazard–harm distinction:** The quantitative analysis focused on two categories of online content identified as priority hazards for this study, violent content and misogynistic content. Other forms of harmful online content that may interact with offline violence, including content related to substance use, self-harm, extremist material outside the manosphere, and forms of identity-based abuse beyond gendered hostility, were not systematically captured. Two further considerations apply. First, the distinction between hazards and harms is not always clean: some forms of content, such as sexist abuse, function both as hazards contributing to offline coercion and as harms in their own right. Second, the hazard categories examined here overlap in practice, with violent and misogynistic content often appearing together. Findings should therefore be read as evidence on the specific hazards examined rather than as a comprehensive account of all online content that may contribute to offline violence affecting young people.
- **Non-representative qualitative sample:** The qualitative findings are based on in-depth interviews and focus groups with a purposive sample of young people, practitioners, and educators with close proximity to violence. The research provides detailed insight into mechanisms and lived experience but is not designed to be statistically representative of all young Londoners.
- **Association, not causation:** The study identifies associations and pathways through which online environments may contribute to offline harm, rather than direct causal relationships between specific content and specific incidents. The quantitative analysis identifies statistical associations between the composition of public online discourse and recorded violence at city level, and the qualitative analysis traces the mechanisms through which online activity appears to shape offline behaviour. Findings should be read as evidence on the conditions under which online environments and offline harm move together, consistent with the public health framing of the study and its emphasis on modifiable conditions and prevention opportunities.

- **Rapidly evolving online environment:** The platforms, features, and content trends described in this report reflect the period in which the research was conducted, and surface details may shift as platforms are redesigned or superseded. The underlying mechanisms identified here, including algorithmic recommendation, group chat escalation, the speed of content circulation, and the gap between platform access and critical capacity among younger adolescents, are likely to persist, because they reflect properties of the broader online environment rather than features of any single platform.

METHODOLOGY

Interpreting the evidence

The findings show patterns and pathways, not simple cause and effect.

Credit: VRU Young People's Action Group, 2025

SECTION THREE

Defining the Hazards and Harms

Interpersonal violence and
violence against women and girls



3. Defining the Hazards and harms of Interest

This report focuses on **two analytically distinct but overlapping harms: interpersonal violence affecting young people and violence against women and girls (VAWG)**. Though both are shaped by peer dynamics, power imbalances, and digital environments, they are examined separately in this report. The necessity for separation comes from them involving different types of online content, possibly operating through different mechanisms, and likely requiring different prevention responses. They are therefore examined through separate chapters (Sections 5 and 6), following the cross-cutting findings presented in Section 4.

The sections below define each harm and the online content treated as a hazard in this study. The Technical Annex provides full detail on how each was operationalised across the three evidence streams: social media monitoring, police and ambulance records, and qualitative fieldwork.

3.1 Interpersonal violence affecting young people

This study adapts the Department for Education/Youth Endowment Fund definition: *interpersonal harm through homicide, serious violent assault, sexual physical violence, and robbery where both victim and perpetrator are young people* (Youth Endowment Fund, Open Innovation Team, Department for Education, 2023). The definition was amended in this study to explicitly include sexual physical violence. As sexual physical violence affects all young people, excluding it from the interpersonal violence strand would mean these experiences were either overlooked entirely or visible only through the VAWG lens, neither of which reflects how violence is actually experienced.

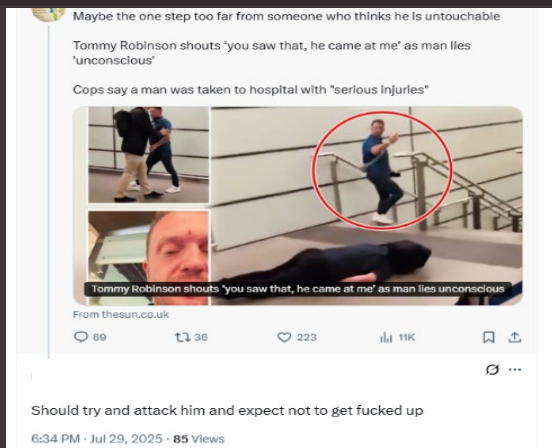
Hazard Definition

The hazard for this pathway is **violence related content and discourse**, including:

- explicit depictions or promotion of violence
- language and symbols associated with threats, retaliation, or dominance
- content that frames violence as acceptable, necessary, or reputationally valuable

In the social media strand, violence related discourse on X was captured through curated keyword search strings and classified using machine learning, producing a final dataset of **10,978 confirmed violence related posts (see Example 1)**. In the offline strand, Metropolitan Police Service crime reports meeting the interpersonal violence classification were supplemented by London Ambulance Service data (See Technical appendix for classification). In the qualitative strand, four anchor case studies traced online-to-offline escalation through interviews, focus groups, and guided digital journey mapping. Full methodological detail can be found in the Technical Annex.

Example 1: Violence-related content



Blacks decide to fight their teacher.
 #fightsvideos #fightvideos #fightingvideos #knockout #ko
 #shegotbeatup #Fights #fightvids #fightpage #fighting #FightVideo
 #fightsvideo #hoodfight #hoodfights #boyfight #girlfights #gore
 #death #schoolfights #hoodfights #girlfight #fightpage #Schoolfight
 #fightsatschool



Redacted examples from the classified violence-related dataset on X

3.2 Violence against women and girls

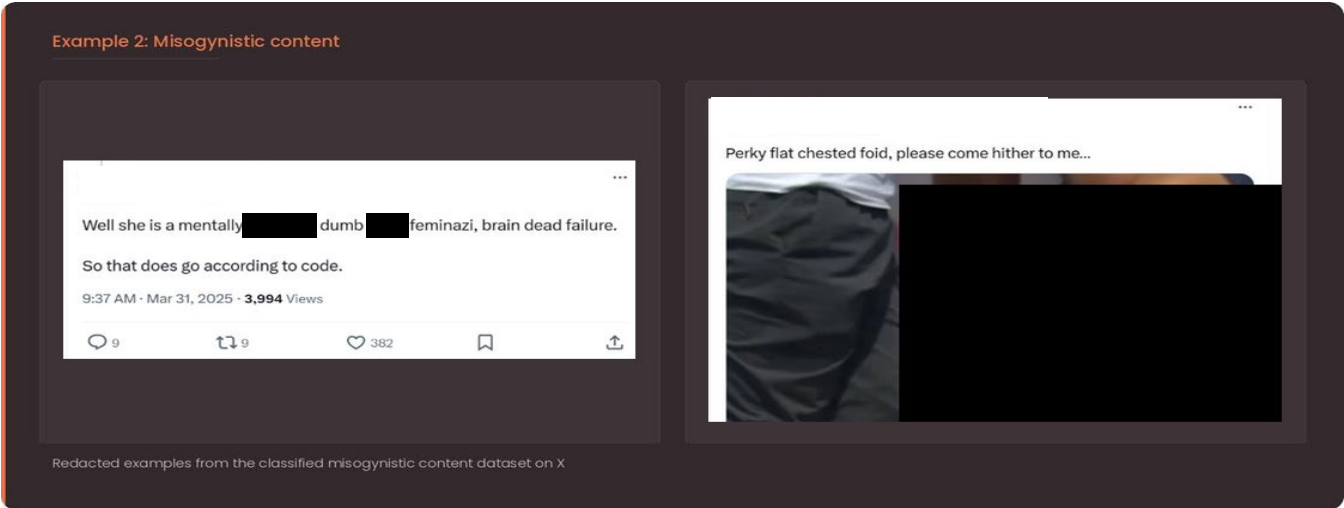
This study adopts the Istanbul Convention (2011) definition of VAWG as all acts of gender-based violence resulting in physical, sexual, psychological, or economic harm or suffering to women. Following Kelly's (1988) continuum framework, VAWG is understood not as a set of isolated incidents but as existing along a spectrum from everyday sexism to severe physical and sexual violence.

Hazard Definition

The hazard for this pathway is both *overt* and *covert* misogynistic and gender-based content, including:

- explicit misogynistic abuse and threats
- victim blaming and rape supportive language
- content promoting male dominance or female subordination
- material that normalises coercion, control, or entitlement within relationships

In the social media strand, VAWG-related discourse was captured through keyword search strings targeting misogynistic abuse, manosphere/incel terminology, and gendered slurs, producing a final dataset of **25,509 confirmed misogynistic posts (see examples in Example 2)**. In the offline strand, Metropolitan Police recorded VAWG offences (see Technical annex for classification) with compositional controls were used for the large-scale analysis. In the qualitative strand, case studies, in-depth interviews, and focus groups traced connections between online misogynistic content and image-based abuse, coercive relationship behaviour, peer sexual harassment, and exploitation. Full methodological detail can be found in the Technical Annex.



3.3 Key differences between the two strands

The table below summarises how the two harms differ conceptually and in their operationalisation.

Table 1: Key differences between the interpersonal violence and VAWG strands

	Interpersonal violence	VAWG
Conceptual focus	Violence enabling discourse related to dominance, retaliation, and ego restoration among young people	Misogynistic and gender-based abuse, harassment, victim blaming, and threats against women and girls
Key language forms	Escalatory phrases, coded threats, dominance speech, hostile memes	Gendered slurs, incel/manosphere terminology, rape myths, objectification language
Classifiers	OpenAI omni-moderation violence and violence/graphic categories combined with an OR rule	BERT+EAR Misogyny classifier, evaluated on the Evalita MISOGYNY English benchmark
Social Media Dataset	138,355 posts collected → 28,079 after removing retweets (80% reshared) → 10,978 confirmed by classifier	125,919 posts collected → 57,276 after removing retweets (54% reshared) → 25,509 confirmed by classifier
Offline data	MPS recorded homicide, serious assault, robbery, sexual offences + LAS data	MPS recorded VAWG offences with compositional controls
Qualitative focus	Group chat dynamics, reputational pressure, audience driven escalation, filming, weapon carrying	Influencer ecosystems, normalisation, image-based abuse, peer sexual harassment, grooming

SECTION FOUR

Social Media as a Mirror and Motor of Harm

Cross-cutting findings on how online environments shape offline harm



4. Social media as a mirror and motor of harm

This chapter addresses the role of social media in shaping patterns of violence and abuse, drawing on findings from across both interpersonal violence and VAWG. In contrast to previous research, which has examined online harms in general terms, this study shows how the relationship between online environments and offline harm operates differently depending on the type of harm involved, with interpersonal violence and VAWG following distinct pathways shaped by different dynamics, content patterns, and social mechanisms.

This chapter explores findings that apply across both interpersonal violence and VAWG. Sections 5 and 6 then examine how these dynamics play out within each strand, and Section 7 explores who is most at risk, and why.

Six cross-cutting findings emerged:

1. **Social media acts as both a mirror and a motor of harm**, reflecting existing inequalities and norms, while also amplifying them.
2. **Young people did not seek out harmful content but were exposed to substantial amounts of it in their everyday platform use**, delivered by algorithms rather than actively searched.
3. **On days when misogynistic content made up a larger share of public conversation on X, police-recorded VAWG in London also increased**. This same-day association did not appear for general violent crime, indicating a pattern specific to gender-based offences across the population as a whole.
4. **Violence content and misogynistic content circulate differently**: violence was overwhelmingly reshared, while misogyny was substantially originally authored, and this requires different prevention approaches.
5. **What young people encountered online did not stay online**: violent content desensitised young people to graphic material, while misogynistic content reshaped what they treated as acceptable in relationships, peer interactions, and classroom behaviour.
6. **Peers are the infrastructure through which harm travels**, but under the right conditions they can also interrupt it.

4.1 Online environments reflect and intensify existing dynamics

The central finding of this study is that social media rarely acts as a standalone cause of violence or abuse:² **Young people, practitioners, family members and peers consistently described online spaces as intensifying problems that were already present in relationships, peer groups, and**

² Previous research has identified associations between online discourse and offline harm in other contexts, including hate speech and hate crime in London, anti-refugee sentiment and assaults in Germany, and xenophobic online activity and anti-Muslim violence in the United States (Williams et al., 2020; Müller & Schwarz, 2021; 2023). This study examines how those dynamics operate specifically for youth interpersonal violence and VAWG in London.

wider social environments. In this sense, social media was less often framed as the origin of violence or abuse than as the setting in which conflict, humiliation, coercion, and exclusion became more visible, longer lasting, and harder to contain.

Three features of social media help explain why it amplifies existing conflict and attitudes rather than simply hosting them:

1. Online Platforms dissolve the boundary between private and public

A comment meant for a small circle can reach an entire year group within hours. A private disagreement, once posted or screenshotted, becomes a public event, and the audience's expectations reshape the stakes.

Practitioners, teachers, and youth workers were particularly likely to emphasise this audience effect: **social media changes the social consequences of conflict because it allows humiliation, disrespect, or threat to be witnessed, discussed, and recirculated by a far larger group than any offline interaction would typically reach.** In their accounts, the issue was often less the original post itself, than what followed: screenshots being shared, sides being taken in group chats, videos being circulated, and pressure building on young people to respond publicly or offline. This dynamic, also known as '*context collapse*' (Marwick and Boyd, 2014), was one of the most consistently identified across both interpersonal violence and VAWG.

2. Social media removes any reliable cooling-off period

Conflict did not end when school ended, and reputational damage did not fade when a young person left the physical setting where it occurred. Teachers spoke about this in immediate terms: schoolground confrontations, safeguarding referrals, and Monday morning incidents that followed a weekend of online escalation that staff had been unable to see in real time. For interpersonal violence, this kept tensions active across weeks or months until a physical encounter arose (explored in detail in Section 5). For VAWG, it meant sexualised content and hostile gendered commentary followed young women across platforms and settings, reducing any sense of escape (explored in detail in Section 6).

3. Harmful content was not occasional or exceptional.

Harmful content was part of everyday online life, delivered by algorithms alongside entertainment and social content rather than clearly separated from it.

These three features are not the product of platforms being misused. They are consequences of how platforms are designed to function – recommendation systems that prioritise engagement and monetisation, private messaging with no friction on circulation, and sharing architectures built for speed rather than safety. The harms documented in this chapter flow from those design choices, and the prevention implications follow accordingly. The next sections examine the scale, character, and distribution of this content environment in detail.

4.2 Harmful content is a routine feature of the online environment young people navigate

The social media monitoring revealed the scale of harmful content present in the public online environment. Over the 12-month monitoring period, **more than 260,000 suspected harmful posts were collected and analysed on X, of which 25,509 were confirmed as misogynistic and 10,978 as violence-related.**

These figures should be read as a conservative, single-platform indicator rather than as a comprehensive measure of exposure. The wider UK evidence base suggests that the volume of harmful content reaching young people across the platforms they use most is substantially larger. The YEF's annual survey reports that 70 per cent of 13- to 17-year-olds saw real-world violence on social media in the past year (YEF, 2024). Vodafone's research with 11- to 14-year-old boys found that 69 per cent had been exposed to misogynistic content and 79 per cent to violent content (Vodafone and Global Action Plan, 2024). Girlguiding's 2024 survey found that 77 per cent of girls aged 7 to 21 had experienced an online harm in the past year (Girlguiding, 2024). Algorithmic content analysis on visual-first platforms suggests still higher concentrations, with research at UCL and the University of Kent finding that misogynistic content on TikTok's For You page rose from 13 to 56 per cent of recommended videos within five days of an account being created (Regehr et al., 2024).

Even allowing for these limits, the X data demonstrates that **misogynistic discourse was measurably more prevalent than violence-related content in the public conversation captured, outnumbering it by more than two to one.**

Beyond the volume figures, higher-risk young people described a parallel experience with violence-related content. **Fight videos, stabbing footage, and threatening posts were encountered as part of routine platform use.** Youth workers spoke about a generation for whom graphic content had become normal, and teachers described the school-day fallout of material that had circulated overnight, often reaching hundreds of students before any adult was aware.

"We've had to address whole year groups because you can't then address things individually because of the viral movement of situations. We've got 300 kids in a year group; you can't pinpoint who exactly has reposted something." - Teacher

It is important to note that this environment was only partially visible in the data. Most young people identified TikTok, Snapchat, Instagram, and YouTube Shorts as the platforms that mattered most to them, none of which the quantitative analysis could capture with the same geolocation filters and metadata. The X data therefore provides a view of the wider public online climate in which those experiences are situated rather than a direct measure of what any individual young person saw. **If this volume of harmful discourse is measurable on one public platform, the total environment across all of the platforms young people use is likely substantially larger.**

4.2.1 Algorithms create pipelines from ordinary interests to harmful material

Harmful content did not only reach young people because they sought it out, but platform recommendation systems also actively delivered it. Boys and young men reported being served manosphere content through algorithmic feeds, often starting from unrelated interests like fitness, gaming, or music. Several tested this deliberately and observed how quickly a single interaction reshaped what they were shown. Girls and young women explained a similar experience, with sexualised and appearance-focused content appearing unsolicited across their feeds.

Practitioners highlighted that much of harmful content arrived through 'For You' pages, recommended videos, and group chat forwards, embedded within the ordinary flow of entertainment and social connection rather than separated from it. In this sense, recommendation systems do not distribute content neutrally, they help create the conditions under which harmful material becomes routine (UK Government, 2025).

This environment was not equally navigable. **Younger adolescents, neurodivergent young people, and those with weaker support networks were repeatedly described by practitioners as less able to contextualise or step back from what they were encountering.** Girls and young women were more likely to experience its gendered effects directly in how they were judged or spoken to. The algorithmic pipeline did not affect everyone equally, and neither did the capacity to resist it. The implications of this unequal exposure for risk and vulnerability are explored in detail in Section 7.

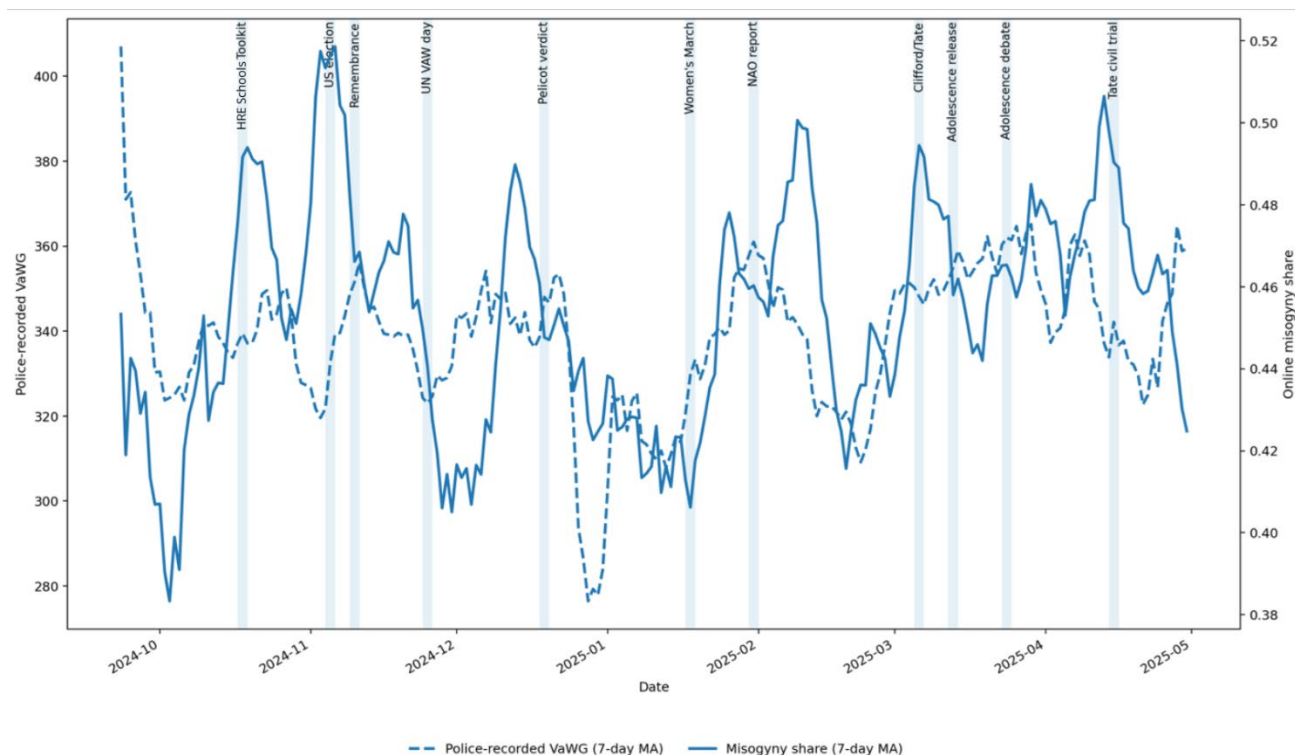
It is worth noting that practitioners frequently described young people as both digitally skilled and digitally vulnerable. **Young people often know how platforms work better than adults do, yet that operational fluency should not be mistaken for critical understanding.** Many of the same young people described as confident platform users were also described as vulnerable to manipulation, unsure of legal consequences, or unable to critically interpret the content they were encountering. This has implications for digital literacy interventions, which need to go beyond teaching young people how platforms function toward helping them engage critically with what those platforms deliver.

4.2.2 Public events and high-profile figures intensify an environment that is already primed

Public events, media coverage, and high-profile figures did not create the harmful online environment, but they appeared to intensify it. An event-shock overlay applied to the time-series data (Figure 4) showed that several visible movements in the misogyny and VAWG series corresponded to recognisable public moments, including the US Election, the launch of new toolkits in London primary schools on healthy and respectful relationships, UN VAW Day, Gisèle Pelicot verdict, the UK Women's March, the Netflix release of *Adolescence*, and renewed Andrew Tate trial coverage.

Young people and practitioners named several of these directly. Andrew Tate was the most frequently referenced figure across the qualitative data, though many young people, particularly boys, noted that his personal influence had waned since his arrest. What persisted was the wider ecosystem of manosphere content he had come to represent. **Whilst individual figures may fall out of favour, their formats, assumptions, and narratives remain in circulation.** Adolescence prompted direct discussion in several interviews, with practitioners noting it had positively opened conversations about masculinity and online influence. These moments appear to act as accelerants on an environment that is already primed, rather than triggers that create something new.

Figure 4. VAWG and online misogyny over time with event-shock windows



A critical feature of this finding is that what mattered was not how busy the platform was on any given day, but what kind of content dominated the conversation. Days that were simply busier online did not show the same relationship once overall posting volume was accounted for.

Analysis of the content on high-misogyny-share days showed that peaks were usually characterised not by a single new topic but by a denser concentration of recurring misogynistic themes: direct insults, anti-women generalisations, claims about transactional relationships, sexualised contempt, and accusations of attention-seeking or mental instability. The US presidential election was the only event where election-specific language became prominent within the misogynistic stream, though even there the content remained dominated by familiar abusive themes. Most other public events in the monitoring window showed little direct event-specific language within the misogynistic posts themselves. This suggests that peaks were more

often periods when routine misogynistic discourse became more prevalent, sometimes loosely catalysed by public events, rather than days driven by a single news moment.

This is consistent with the event-shock analysis above: on days when public discourse was more saturated with misogynistic content, whether driven by a recognisable public moment or by the cumulative weight of everyday misogynistic commentary, recorded violence against women and girls was also higher.

The analysis also tested whether this relationship built up over several days rather than appearing immediately but found no evidence that elevated misogynistic content over the preceding week predicted a subsequent rise in VAWG. A separate check confirmed that tomorrow's misogyny levels did not predict today's VAWG either. This makes a simple shared-trend explanation less consistent with the pattern, though it cannot be ruled out entirely. The safest interpretation is that **online misogyny and offline VAWG are dynamically connected, each reflecting and reinforcing the same wider gendered social climate on any given day.**

The fact that the pattern did not appear for general violent crime provides supportive, though not definitive, evidence that the association is specifically gendered: if a general underlying social condition were driving everything, we would expect to see a similar pattern for violent crime overall, which we did not. Exploratory sub-models comparing under-25 and over-25 victims also did not show statistically significant differences in the strength of the association between the two age groups, indicating that the relationship operates as a broader city-level pattern rather than as a youth-specific effect. That said, the days with the highest VAWG counts did tend to involve a higher proportion of younger victims and younger suspects in the recorded cases, suggesting that younger people were disproportionately represented in the cases recorded on the busiest days, even though the underlying relationship between online misogyny and VAWG was not stronger for them.

Exploring this further, the relationship was also statistically significant in incidents where the suspect was known to the victim. This is consistent with the qualitative evidence throughout this report, where the violence and abuse documented most often played out within existing relationships, peer groups, and social networks rather than through generalised hostility toward strangers.

Table 2. Poisson regression: daily police-recorded VAWG predicted by daily misogynistic post share. Selected predictors shown; model also includes linear trend, day-of-week indicators, and suspect ethnicity composition controls. Full specification in Technical Annex.

Predictor	B	S.E.	95% CI lower	95% CI upper	Wald X ²	df	Sig.
Misogynistic post share	0.354	0.116	0.127	0.582	9.346	1	.002

Victim proportion under 25	0.829	0.257	0.326	1.332	10.437	1	.001
Suspect proportion under 25	0.408	0.185	0.046	0.771	4.867	1	.027
Suspect proportion known to victim	-0.584	0.215	-1.004	-0.163	7.403	1	.007
Neutral Activity Index (log)	-0.046	0.038	-0.120	0.028	1.506	1	.220

Source: Nisien.ai analysis of X data and Metropolitan Police Service records, 23 September 2024 – 30 April 2025 (observed Neutral Activity Index window). Full study collection period: May 2024 – April 2025.

4.2.3 The absence of a public signal for violence reflects where escalation happens, not whether it happens

The equivalent analysis for interpersonal violence did not find a relationship between the daily share of violence-related posts on X and police-recorded violent crime. This is an important finding and should not be read as evidence that social media is unimportant in interpersonal violence. Rather, there are both methodological and substantive reasons why the analysis detected a clear pattern for VAWG but not for violence and understanding them is essential to interpreting the study’s results as a whole.

The most immediate hypothesis is that the analysis can only observe public, text-based posts on X, yet a meaningful share of violence-related online activity takes place through private group chats, direct messages, disappearing-message platforms, and circulated video. This is particularly true for the kinds of online activity connected to group-based, retaliatory, or locally embedded conflicts, as the qualitative research consistently confirmed. Violence-related content is also harder to identify through text-based analysis because it is more likely to be visual, coded, or dependent on local context rather than relying on identifiable vocabulary and framing. Platform moderation may also play a role: violent content is more likely to be removed or restricted, which can suppress original posting and concentrate spread into the resharing of content that survives takedown. Misogynistic content, often coded or framed as humour, is moderated less consistently and so persists more visibly as original commentary.

There may also be a substantive difference in how the two types of content function publicly. The study’s own data shows that misogynistic content is far more likely to be originally authored than violence-related content, which is overwhelmingly reshared. One possible explanation for this is that misogynistic attitudes lend themselves to public performance in a way that violence does not: expressing views about gender and relationships can function as identity building within peer networks, whereas violence is less commonly claimed as a public identity in the same way. If this is the case, the nature and volume of what is publicly detectable would be fundamentally different for the two strands. The evidence for this distinction is explored in full in Section 4.4.

Overall, the **absence of a detectable public pattern for violence tells us something important about where interpersonal violence escalation happens online, not whether it happens.** The qualitative research presented in Section 5 documents extensively how these dynamics operate through private and visual channels. The difference in results between the two strands is not a contradiction but part of the same finding: VAWG and interpersonal violence relate to public online environments in structurally different ways, and effective prevention needs to respond to both.

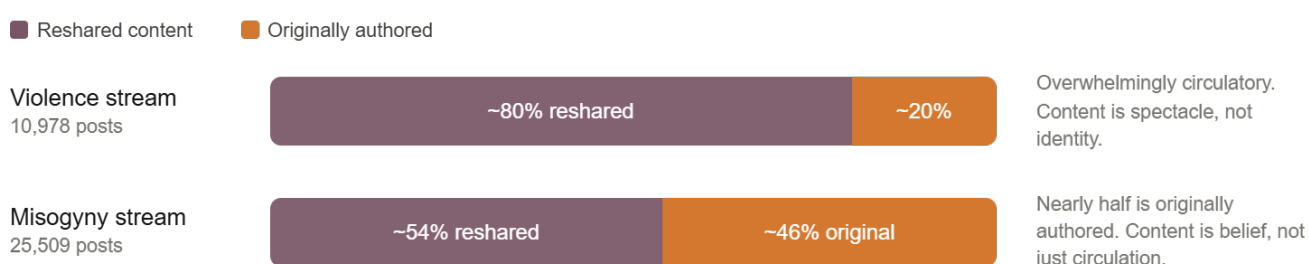
4.3 Violence content circulates; misogynistic content is more often authored – and prevention must respond differently

The previous section noted that violence-related content and misogynistic content occupy fundamentally different positions in the public online environment. This distinction is one of the most operationally significant findings in the study, because it shapes not only why the quantitative signals differ, but also what prevention needs to do.

Violence-related content is primarily circulatory. Roughly four in five suspected violence posts were reshared (see Figure 5), and this content often clusters around specific incidents before moving rapidly through peer networks. As the spread of violent content depends on sharing rather than original production, prevention is most effective when it introduces barriers to circulation: faster takedown, reduced resharing, and platform-level mechanisms that slow viral chains.

Misogynistic content, by contrast, is more often originally authored. Around 45% of the classified stream was original content, written by individual users in their own words, appearing as everyday commentary about gender, relationships, and identity. Because this content is being actively produced rather than simply passed along, prevention cannot rely on removal alone but must also involve norm disruption, critical literacy, and relational work that shifts attitudes before they harden into behaviour.

Figure 5. Circulation vs Authorship of social media tweets



Why this matters: Violence-related content spreads through recirculation and spectacle – people share what others have posted. Misogynistic content is actively produced – people author their own hostile discourse. This difference helps explain why the temporal modelling found a significant public-data signal for VAWG but not for interpersonal violence: the violence pathway operates through private channels that public monitoring cannot reach, while the VAWG pathway has a visible, measurable footprint in public online discourse.

The qualitative evidence illustrates how each of these dynamics operates in practice.

4.3.1 Young people encounter violence content and pass it on, they rarely create it

Higher-risk young people and youth workers described violence-related content, such as videos of fights, stabbings, or serious confrontations, as shared because it was shocking, reputationally informative, or relevant to ongoing disputes.

“People are especially interested when there’s weapons involved... everyone will send it around like all over social.” – Young boy, 16–18

Boys and young men framed this as content they encountered and passed on, not something they created. The Youth Endowment Fund (2024) found a similar pattern nationally: while 70% of teenagers had encountered real-world violence online, only 6% had actively searched for it. Exposure to violent content is shaped by circulation rather than deliberate production, and practitioners noted incidents being widely shared within hours, shaping expectations and responses in the days that followed.

4.3.2 Misogynistic content is not just consumed – it is reproduced as personal belief

The authorship pattern visible in the data reflected something practitioners observed directly. **Boys and young men engaged with misogynistic content in ways that went beyond consumption, actively articulating and reproducing its ideas as everyday opinion.** Ofcom’s (2025) analysis of the manosphere identifies the same participatory quality: a space where users do not just encounter content but actively engage with and reproduce it.

“A lot of the time they will just sort of drop into conversation casually what I would consider to be quite shocking views about women and girls... this idea of, you know, women have got above their station, very old-fashioned views about what women should be doing or not doing.” – Youth worker

Recent research on the manosphere helps explain this dynamic: **misogynistic ideas are packaged in socially usable ways, combining humour, self-improvement narratives, and peer validation, that make them easy to repeat in offline contexts** (Haslop et al., 2024; Lott et al., 2025; Wescott et al., 2024). Following more manosphere influencers is associated with greater dehumanisation of women, suggesting the relationship between consumption and attitude formation is measurable, not just assumed (Renstrom and Back, 2024).

The incentive to produce such content is not only ideological. Research on platforms like Telegram has documented how misogynistic material and non-consensual intimate content are also created and circulated for financial gain, with abusive content monetised through premium features and paid access (Semenzin and Bainotti, 2020). This adds a commercial dimension to the authorship pattern identified in this study’s data and means that prevention approaches focused solely on shifting attitudes are unlikely to address the full range of motivations driving content production. The distinction between circulation and authorship is not merely a technical

observation about how content spreads, but points to different mechanisms of influence, driven by different motivations, that require different prevention responses.

This speaks to a wider analytical shift in how online harm to young people is being understood. Recent work by the Council of the European Union (2025) and the Institute for Strategic Dialogue (2025) describes contemporary online harm as increasingly post-ideological in character, organised around clout, transgression, peer status, and ambient cultural production rather than around coherent ideological commitment. The findings presented here are consistent with that framing. Interpersonal violence, misogynistic content, image-based abuse, peer objectification, and platform-facilitated exploitation operate not only as discrete harms produced by individual actors, but increasingly as features of a shared cultural and algorithmic environment that young people in London navigate daily.

4.4 Repeated exposure shapes behaviour over time

Understanding how violence-related content and misogynistic content circulate differently is important because it helps explain what happens when young people are exposed to each type repeatedly. Across the qualitative data, practitioners and teachers described a consistent pattern in which sustained contact with violent and misogynistic material shifted what young people treated as normal, but the process looked different depending on the type of content involved.

4.4.1 Violence content raises the threshold for what counts as serious; misogynistic content reshapes what young people treat as acceptable

For violence, the primary effect described by both young people and practitioners was desensitisation. Graphic footage was consumed as entertainment without significant emotional response, the threshold for what counted as serious was raised, and young people became less likely to report what they were seeing. Youth workers described a generation that no longer reacted to content that would previously have alarmed them.

"I've seen so much that I'm just like desensitized to it, you know? It's kind of like, oh, it's just another thing for me." – Young boy, 16–18

For misogyny, the process was different. Rather than dulling an emotional response, repeated exposure reshaped what young people treated as acceptable. Gendered assumptions about relationships, control, and status were absorbed, often without critical engagement, encountered as common sense or humour, rather than as harmful messaging. Teachers observed boys using language drawn from manosphere content as ordinary conversation, without recognising it as ideological. Practitioners noted that by the time this material surfaced in classroom discussion, it had already been processed through peer networks and absorbed as unremarkable.

Importantly, young people did not always recognise the shift in themselves. One young man, reflecting on manosphere content, described his engagement as selective and discerning:

"I don't automatically disregard all of it... I can pick up the actual good life advice that's in there." – **Young boy, 18–19, reflecting on manosphere content**

That account illustrates a dynamic several practitioners also noted: young men who believed they were exercising critical judgement while absorbing the gendered assumptions embedded within self-improvement framing. However, the 'good advice' and the harmful norms arrive in the same package, and selective engagement does not necessarily prevent norm absorption. This is a significant challenge for prevention, because it means that awareness of harmful content does not in itself protect against its normalising effects.

Some harmful content also entered circulation through boredom and entertainment rather than ideological commitment, for example, sharing violent or humiliating material for the excitement of drama. This complicates any prevention approach that relies solely on critical literacy, because the motivation for engaging with the content is not always belief but social currency.

4.4.2 Normalised attitudes show up in relationships and classrooms, and humour is the mechanism

The effects did not stop at attitudes, youth workers and teachers described **normalised assumptions translating into observable behaviour. This ranged from controlling expectations in relationships, sexualised language in classroom speech, and misogynistic talking points repeated as casual truths.** The specific evidence for how this translation operates within the VAWG pathway is presented in Section 7.1.3.³

Across both strands, misogynistic material, violent clips, and humiliating content were repeatedly framed by young people as 'just banter' – socially usable and repeatable without full ownership. The boundary between performative repetition and genuine endorsement was difficult for anyone to draw, including the young people themselves. A young person involved in anti-VAWG peer work described this as *"a huge part of it – 'oh, it was just banter', 'oh, I didn't mean it',"* adding that it was difficult for teachers to distinguish from bullying. A practitioner working on sexual violence prevention made a similar observation:

"The fact they joke and engage with the content obviously shows that there's a small part of them that does maybe believe it... it definitely normalises it." – **Sexual violence prevention practitioner**

This matters for prevention because humour softens accountability without neutralising the effects. Prevention work focused solely on identifying overtly harmful content will miss how normalisation operates in practice – not through what young people openly endorse, but through

³ This is consistent with Flood's (2019) framework on how repeated exposure to hostile gender norms reshapes relationship expectations and lowers thresholds for recognising controlling behaviour – a process that operates through gradual erosion, not dramatic conversion.

what they stop noticing is worth challenging. The specific evidence for how this process operates within the interpersonal violence and VAWG strands is presented in Sections 5 and 6.

It is important to note, however, that not all young people described themselves as passive recipients of harmful content. Some explicitly resisted misogynistic material, mocked it as manipulative or absurd, or said they and their friends would 'scroll past.' Others described social media as a source of connection, study support, emotional management, or inclusion when offline relationships were difficult. These counter-narratives matter because they remind us that exposure does not inevitably produce the effects described above, and that the same platforms can support wellbeing, disclosure, and resistance as well as normalisation.

These counter-narratives are consistent with a wider body of academic work which argues that social media is not only an amplifier of harm but also affords novel tools for informal social control and de-escalation (Lane and Stuart, 2022). Recognising both dimensions of online environments is necessary to design prevention approaches that work with rather than against the role social media plays in young people's lives.

4.5 Peers shape whether online incidents escalate or are interrupted

Peer dynamics mattered across both strands not only because peers encouraged or observed conflict and abuse, but because they formed the social infrastructure through which it travelled. Group chats were described by participants as the spaces where disputes and abuse widened beyond the original people involved, encouraging commentary, side-taking, circulation of screenshots, and pressure to respond. What began as a disagreement or act of abuse between two people could become collective within hours once peers were drawn in.

The consequences of this peer involvement were not uniform. What was at stake for the young person involved – whether public exposure threatened physical violence, sexual reputation, or institutional sanction – depended on gender, race, age, and wider social position. Read through a contextual safeguarding lens, these dynamics point to something broader than individual online choices: they represent forms of extra-familial abuse unfolding simultaneously across peer groups, schools, neighbourhoods, and online spaces (Contextual Safeguarding Network, 2025). Addressing them effectively requires responses that account for the peer contexts in which they occur, not only the platforms on which they are visible.

At the same time, peers could also interrupt escalation. Youth workers and practitioners identified peer-led interventions – friends stepping in with non-judgemental conversation, older students modelling non-retaliatory responses, boys directly challenging misogynistic comments within their friendship groups – as among the most effective de-escalation strategies they had observed.

"They do try to have conversations with their peers... debunking some of those thoughts... and we've spoken a lot around being an active bystander and giving boys tools to challenge their male peers... that helps." – **VAWG practitioner**

Where peer-led challenge worked, it worked because young people used the same peer influence that drives escalation in a different direction: directly challenging misogynistic comments in friendship groups, stepping into arguments with non-judgemental conversation, or modelling non-retaliatory responses that gave others a way to step back without losing face. This is one of the clearest prevention opportunities identified in this study.

This aligns with Lane and Stuart's (2022) account of how social media operates as a site of 'communication visibility', the affordance by which onlookers can observe online behaviour and, in turn, exert influence over subsequent relationships, exchanges, and actions in ways that can prevent and reduce violence. Drawing on ethnographic fieldwork in Harlem and Chicago's South Side, they argue that the same visibility that enables conflict to escalate online also enables third parties to intervene. Crucially, they distinguish between insider third parties, including community members, trusted adults, and peers, whose interventions are supportive and inclusive, and outsider third parties, including law enforcement and prosecutors, whose use of the same visibility can be punitive and exclusionary. The implication for prevention is that the social media environment driving escalation is also the environment in which peer-led de-escalation can operate, but the institutional position of the third party determines whether that intervention supports or harms the young people involved.

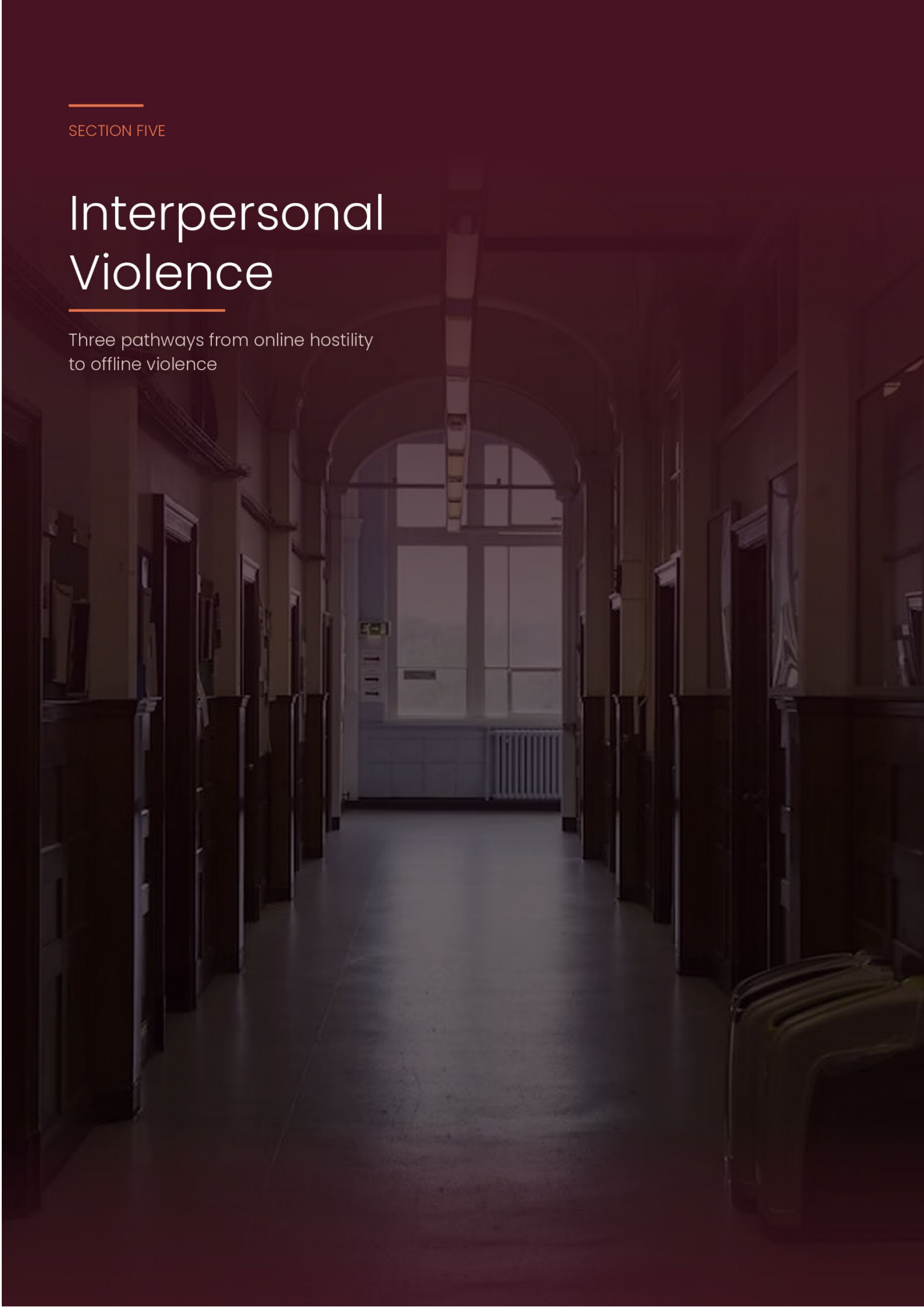
The importance of peer-led approaches is reinforced by a related finding about adult intervention, which speaks directly to this insider versus outsider distinction. Many participants stressed the value of trusted adults, schools, youth workers, and safeguarding teams, and there is strong evidence that these relationships can de-escalate conflict, validate experiences of abuse, and create routes to support. At the same time, some young people feared adults would make matters worse, increase shame, or respond punitively rather than protectively. That tension matters for prevention design: trust, not simply adult presence, appears to be the critical ingredient, and peer-led approaches may sometimes reach young people in ways that adult-led interventions cannot. The implications for peer-led prevention programmes are explored in Section 8.

The findings in this chapter point to a consistent picture: social media intensifies and sustains violence and abuse rather than creating them in isolation, through mechanisms that operate differently for interpersonal violence and VAWG. However, these findings should also be read with an awareness that social media does not operate independently of broader inequalities. Racialised surveillance, adultification, developmental compression, digital literacy gaps, and socioeconomic pressures all structure both who is exposed to harmful content and how they are able to respond (see Section 7). Social media provides environments in which existing conflicts, norms, and vulnerabilities can be intensified, circulated, and made more consequential – and it is within that understanding that the strand-specific findings in the following sections should be understood.

SECTION FIVE

Interpersonal Violence

Three pathways from online hostility
to offline violence



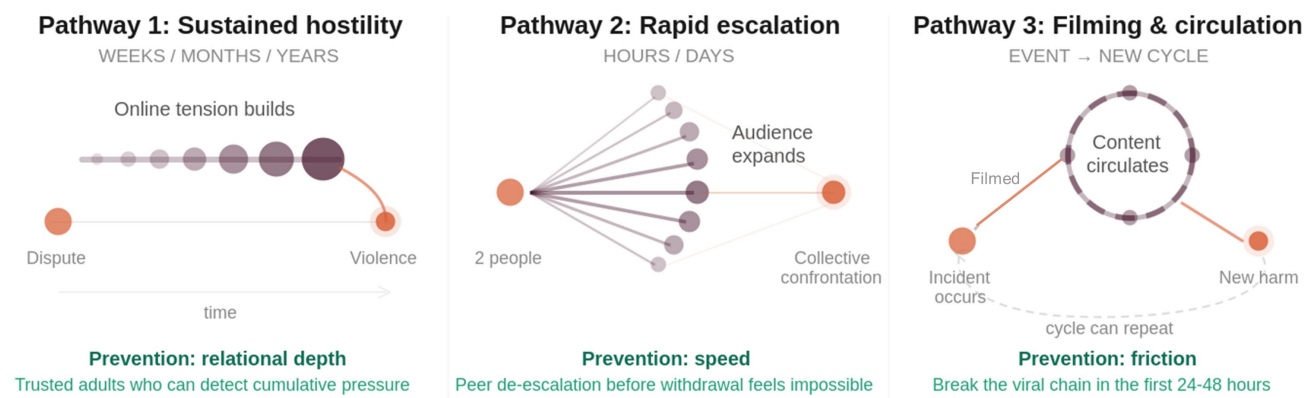
5. Interpersonal Violence and Violent Content

This chapter traces how the intensification of harm on social media operates specifically for interpersonal violence. The following evidence suggests that, compared with the VAWG pathway, the interpersonal violence pathway moves faster, is more event-driven, and operates primarily through private and semi-private channels that are not visible in public platform data. The large-scale data analysis found no city-level public-data signal for this pathway, which is consistent with the qualitative evidence that the mechanism sits in group chats, direct messages, and disappearing-message platforms rather than in public posts.

The study identified **three distinct pathways**, each with its own tempo, mechanism, and prevention logic (see Figure 6):

- Sustained online hostility that keeps conflict active over months or years.
- Rapid group chat escalation that turns private disputes into collective confrontations, increasing pressure to respond physically in ways shaped by what young people have repeatedly seen online.
- Filming that converts passing events into persistent content creating new grounds for retaliation.

Figure 6. Three Interpersonal violence pathways identified



5.1 Pathway 1: Sustained online hostility keeps conflict active until an in-person opportunity for violence arises

This pathway was characterised not by rapid escalation but by slower accumulation. Interpersonal or intergroup disputes or grievances can be kept active through persistent online visibility, even when nothing new was being actively posted. The effect was not a direct line from one post to one incident, but a state of ongoing tension in which violence became more likely whenever an in-person opportunity arose.

The pathway typically followed a recognisable pattern:

- A dispute or slight occurred, often offline.
- Online channels kept the grievance visible and emotionally charged, even during periods with no face-to-face contact. Ongoing "talking" and "instigating" sustained the dispute over time.
- Young people felt unable to disengage. Comments, tags, screenshots, and indirect references maintained a sense of being watched and judged – a reputational pressure that persisted without requiring new provocation.
- When a physical encounter eventually occurred, the accumulated hostility lowered the threshold for violence. The confrontation was experienced not as a new event but as the culmination of something building over a prolonged period.

5.1.1 Conflict stays emotionally charged even without new provocation – with young people feeling unable to disengage

Disputes in this pathway did not escalate over a weekend. They were states of hostility maintained across weeks, months, or in some cases years. **Practitioners consistently contrasted current conflicts with earlier patterns of youth violence, describing a shift from disputes that had natural endpoints to disputes that persisted indefinitely through online channels.** As one put it, in the past it *"almost stopped once you got home, whereas now it keeps following you"* across the evening and into the next day. In this pathway, conflicts were also seen to follow young people across school years and into new settings.

A recurring observation from practitioners, teachers, and young people themselves was that active posting was not required for conflicts to remain live. Young people described feeling unable to disengage even when they were not contributing – compulsively checking whether things were getting worse. Disputes did not need constant new content to remain active. Their persistence within peer attention was enough. For young people already living in contexts of instability or exclusion, this constant visibility left no room for tensions to dissipate.

Case Study 1: Long-standing online tension and serious violence involving weapons

One young boy, aged 17, who was later stabbed, linked the incident not to a recent argument alone, but to a grievance going back around five years. What kept that conflict live was not continuous face-to-face confrontation but ongoing online *"talking"* and *"instigating"*, which sustained the dispute over time and made it harder to leave behind. Even where there were no immediate tensions in the days before the attack, he described the problem as something carried forward *"from a long time ago"*, with online activity continuing to make it worse. [Full Case study 1 detail can be found below].

5.1.2 Anonymous disrespect and AI-generated content deepen the insecurity because there is no way to resolve the threat

The persistently hostile environment was not always traceable to known people. Some young people described sustained disrespect coming through people they knew only indirectly or through anonymous accounts. **Where the source could not be identified, the insecurity became harder to resolve – there was no one to confront, no dispute to settle, and no way to know when or whether it would stop.** This distinguishes the anonymity dynamic in this pathway from its role in Pathway 2. In the second pathway it triggers rapid misattribution and escalation, here it creates an unresolvable background threat that has no clear endpoint.

AI-generated content intensifies this further. Participants and practitioners identified the growing capacity to fabricate realistic content as particularly threatening because reputational damage could now be manufactured from nothing. For young people already navigating a hostile online environment, the knowledge that content about them could be created, not just shared, deepened the sense that their reputation was permanently and uncontrollably at risk.

Spotlight: An insecure online environment and carrying a weapon for protection

One young man linked his arrest for carrying a knife to a wider pattern of online disrespect, much of it coming through people he knew indirectly or from anonymous accounts. He said he *"never really got disrespected in real life... it was always people on the internet,"* and described how the pressure of having *"a lot of disrespect on your name"* contributed to his decision to carry a weapon for protection.

From his perspective, AI-generated content is particularly harmful because it could be used to *"tarnish someone's name very badly."* For this young person, the issue was not one isolated post, but the cumulative insecurity produced by online humiliation, anonymous disrespect, and the fear that damage to reputation could continue without a clear way to stop it.

5.1.3 Prevention for this pathway depends on relationship-based intelligence, not public monitoring

Given this pathway operates slowly and through private channels, the most consequential activity is invisible to public monitoring. Early warning depends on the people around the young person – those with enough relational depth to recognise when cumulative pressure is translating into risk behaviour.

The weapon carrying finding reinforces this. The young man in the spotlight above did not carry a knife in response to a specific escalation event, but due to a persistent online environment that had produced a permanent sense of insecurity. Interventions that rely on detecting a specific trigger will miss this pathway entirely. The implications for relationship-based prevention are developed in Section 8.

Long-standing online tension and serious violence involving weapons

How sustained online interaction kept a dispute active over time, lowering the threshold for serious offline violence.

CONTEXT

Across the research, interpersonal violence was often not triggered by a single online incident. Instead, disputes were sustained over time through ongoing online interaction, reputational pressure, and unresolved tension. Practitioners and young people described conflicts that remained active through “talking” and instigation on social media, even where there was no immediate offline escalation. This case reflects that wider pattern.

WHAT HAPPENED

The dispute in this case originated several years earlier, linked to a conflict between two boys. It was never fully resolved. Instead, it remained present through periodic online exchanges, comments, and “instigating”.

There was no clear escalation point. The conflict continued at a low level over time, sustained through social media. The young boy described this as people “posting” and “talking”, which kept the issue active even when there was no direct contact offline.

The incident itself occurred later, when he randomly bumped into the boy he had “beef” with on the street. The boy and his friend attempted to stab him, injuring bystanders who got involved to try and stop the fight. In his account, this violence did not arise suddenly, but from a long-standing dispute that had been maintained over time.

“

“It wasn’t even like something that just happened... it was from time ago... like five years ago... but obviously online he’s been talking... and then it just goes downhill... just people posting stuff, instigating it... that’s it really... instigating and making it worse... and then when it actually happened, they ran up and tried to stab me... other people got stabbed as well.”

”

Young boy, Interpersonal violence case study

HOW THE HARM WIDENED

This case reflects a wider qualitative pattern in which social media removed the conditions that might otherwise allow conflict to dissipate. Practitioners described disputes that “kept going” beyond school hours, carried through messages, posts, and ongoing commentary.

The young man linked this directly to reputation and ego:
“If it wasn’t for social media then we wouldn’t really have as much ego”.

A key consideration here is that the conflict did not need to escalate visibly on social media to remain risky, just be kept live long enough for the offline opportunity to present itself.

What this case shows

This case illustrates a pathway in which unresolved conflict was sustained over time through ongoing online interaction.

CORE MECHANISM

A recurring online-to-offline mechanism in interpersonal violence was the maintenance of tension over time. Social media did not need to trigger a specific incident. Instead, it kept disputes visible, sustained, reputationally significant, and unresolved. This sustained tension lowered the threshold for violence when an opportunity arose offline.

QUANTITATIVE CONTEXT

The wider analysis found no statistically reliable association between online violence content and overall violent crime. This is consistent with this case, where the most consequential dynamics developed over time and through ongoing interaction, rather than being tied to short-term changes in publicly observable online content.

RISK FACTORS

- Long-standing unresolved dispute
- High importance placed on reputation and status
- Ongoing online instigation and “talking”
- Limited trusted adult support
- Social environment where violence was normalised

WHAT WOULD HAVE HELPED:

- Earlier mediation of long-standing disputes
- Trusted adult relationships to be able to intervene over time
- Structured opportunities to disengage from conflict
- More consistent youth support outside crisis points

ESCALATION DRIVERS

- Persistence of conflict over time
- Online disinhibition (“saying things you wouldn’t say face to face”)
- Reputational pressure and ego
- Lack of resolution mechanisms
- Absence of early intervention

“ *I think if social media wasn’t around, that wouldn’t have taken place...*

It’s like when someone shoots a basketball and people cheer, it changes how you act.

”

PREVENTION LEVERAGE POINTS

- Early identification of entrenched disputes
- Long-term relationship-based youth work
- Peer mediation and conflict resolution
- Monitoring early online signals of escalation
- Support for disengagement before violence occurs

5.2 Pathway 2: Group chats and public call outs rapidly turn private disputes into collective confrontations

This pathway was faster and more event-driven than Pathway 1. Where sustained hostility produced violence through slow accumulation over weeks or months, this pathway could move from private disagreements to physical confrontations within hours or days. The trigger was audience expansion: once disputes became visible to wider peer groups through group chats or posts on stories, the social meaning of the conflict changed and the pressure to respond physically intensified.

That pressure was not only driven by the audience itself but also by what young people had come to see as a normal response to conflict. As Section 4 documented, repeated exposure to violent content raised the threshold for what young people treated as serious, making physical confrontation feel like a more ordinary way to resolve disputes. One Probation Worker and Youth Justice Practitioner described this shift clearly:

"For some girls to see other girls fighting online, it makes it seem okay. I think it's that thing of a lot of people do it, so this is how you should respond." — **Probation Worker and Youth Justice Practitioner**

What felt like an acceptable response had already been shaped by what young people had repeatedly seen online, meaning that when conflict went public, physical escalation often followed a recognisable sequence:

- A disagreement or slight occurred, usually between two people.
- It entered online space through posting, screenshots, or group-chat discussion.
- The audience widened rapidly as peers were drawn in through commentary, side-taking, and further sharing.
- Reputational pressure built. Public humiliation or challenge created an obligation to respond – particularly for boys and young men, for whom being seen as "soft" carried significant social cost.
- Physical confrontation followed, experienced not as a choice but seen as the only way to restore status in front of an audience that was now watching.

5.2.1 Once the audience widens, withdrawal becomes more dangerous than escalation

Young people repeatedly drew a distinction between something said in person to a small number of people and the same thing being posted or screenshotted in front of a much larger peer audience. In this pathway, that shift raised the reputational stakes to the point where withdrawal felt more dangerous than escalation. Boys and young men were especially likely to describe public humiliation, mockery, or challenge as creating an obligation to respond.

"It's the embarrassment of everybody in your year knowing about it, because if it was spoken in person, you and the people around you are sort of the only people that would know. But

obviously, you post it, it's their friends that see it, everyone that follows them. And it's that embarrassment. And then you have to defend yourself." – **Young girl with experience of interpersonal violence, 16-17**

Teachers described the consequences of this dynamic in immediate terms: school ground confrontations, safeguarding referrals, and Monday morning incidents that followed a weekend of online escalation staff had been unable to see in real time. The speed of this pathway regularly outpaced institutional capacity to intervene.

5.2.2 Group chats escalate the language because the space feels temporary – but nothing genuinely disappears

Group chats, primarily on Snapchat, functioned as the primary escalation infrastructure across this pathway. They pulled in people who had no part in the original dispute, through commentary and screenshot circulation, and turned private disagreements into collective confrontations within hours.

The perceived privacy and ephemerality of these spaces intensified the risk. Young people and practitioners repeatedly identified Snapchat as the platform where threats, planning, and mobilisation occurred precisely because messages were assumed to be temporary and therefore lower risk. Yet screenshots meant that nothing genuinely disappeared. The false sense of ephemerality encouraged escalatory language, while the reality of screenshotting ensured that material could be recirculated and weaponised long after it had supposedly vanished.

"Everyone thinks they can be a different person on social media... and then when it comes to real life, you're like, 'I'm going to have to live up to that person'... it's all good chatting from your bedroom, but in real life it gets way out of hand." – **Young boy, 16-18**

That gap between online persona and offline reality was where escalation happened. Disputes that might have been contained face to face became commitments that had to be honoured once they had been made public, in front of an audience, in a space that felt temporary but was not.

Case Study 2: Audience-driven escalation from online conflict to large scale offline violence

A young person recounted how a disagreement involving two people escalated over two weeks as both sides posted about each other daily with increasingly escalatory language. The audience grew rapidly through group chats, with peers drawn in through commentary and side-taking. Even after the original issue had been resolved between the two young people involved, the volume of online attention meant the dispute persisted, and eventually resulted in a large-scale brawl.

Young people were pulled into Snapchat group chats without their knowledge or consent, at which point their names were mentioned and berated. Physical altercations were later arranged through those same chats. [Full case study 2 detail can be found below].

"They always talk about group chats and being added into groups that they haven't really consented to and they kind of get sucked in and then their names get mentioned and then it kind of starts a whole thing." – **Frontline practitioner working with young people affected by serious violence**

5.2.3 Influencer "beef" culture provides a ready-made script for public escalation

Influencer culture, from drill artists, YouTubers to social media figures with public feuds, modelled a format of public conflict that young people replicated in their own peer dynamics: the call out, the public response, the audience choosing sides. Whether the beef was genuine or orchestrated mattered less than the fact that it made this sequence feel like how disputes are supposed to play out online.

Posts and videos of conflict functioned as social currency, engaging visibly with public disputes attracted recognition and reinforced status within peer groups. For young people navigating insecurity, unstable home lives, or limited offline alternatives, this performance of toughness (through public online conflict) offered a route to identity and belonging that offline environments were not providing.

"You see conflict between influencers and they're posting about each other saying really horrible things...if there's a conflict you're witnessing, everybody picks a side between two influencers, and no one ever tries to calm things down...so then in real life it feels like you want to escalate it, get involved. You don't want to try and stop it." – **Young girl with experience of Interpersonal violence, 16–17.**

This is a different kind of influencer effect from the one documented in the VAWG pathway (Chapter 7), where manosphere content shaped beliefs about relationships and gender. Here, the influence was not ideological but performative: a script for how to escalate, not what to believe.

5.2.4 There is a critical window for de-escalation, and peers may be more effective than adults at reaching it

Interviews with young people and practitioners suggest there is a particular moment – after disputes have become public but before physical responses have been fully committed to – where intervention is most likely to succeed. This is the critical prevention window for this pathway. At that stage, peer-led challenge may be especially important.

"He'd be much more likely to listen to a 15-year-old peer than to an adult." – **Youth justice practitioner, describing a boy under pressure to respond.**

The prevention logic for this pathway is fundamentally about speed: reaching young people before public attention has built to the point where withdrawal feels impossible. How that translates into peer-led and institutional response is developed in Section 8.

Audience-driven escalation from online conflict to large-scale offline violence

How peer visibility and group chat dynamics transformed a two-person dispute into a collective confrontation.

CONTEXT

This case illustrates a rapid escalation pathway, in which a private disagreement became a collective conflict within days once it entered online space. Social media did not simply provide a space for conflict to play out, but it changed its scale.

Group chats, stories, and shared posts drew in people who were not originally involved, creating pressure to respond and stay engaged. What began as a disagreement between two people could quickly become something much larger, shaped as much by the audience as by the original issue.

WHAT HAPPENED

The incident began with a disagreement between two young people. One of them posted a video about the other, making the issue visible to a wider peer group.

This marked a turning point. What had been a contained disagreement became a public event.

Others quickly became involved. Friends on both sides responded with posts, comments, and messages. Over the following two weeks, the content became more personal and more hostile, including rumours and sexualised commentary.

A group chat was created to discuss the situation. Participants were added rapidly, including young people who had not been involved in the original dispute. The group continued to grow as others joined “to see what was going on.”

“

It started off as like just small little rumours... but as it went on, as like people, like it was getting more attention, the like insults and the rumours started getting worse and worse... if people, when they post stuff, people are looking for a reaction... and as soon as you don't respond to them and you don't give them what they want, they're just going to start saying worse and worse stuff until you eventually do react.”

Young girl, 16–17, Interpersonal violence case study

Although the two individuals at the centre of the conflict resolved their issue early on, the online activity did not stop. By that point, the situation had moved beyond them.

The conflict ultimately escalated into a large offline fight involving around 50 young people.

HOW THE HARM WIDENED

What widened the conflict was not only visibility, but **audience dynamics**. As attention increased, the tone of the content shifted. With growing escalation of posts and stories, the worst material was sexualised, including false claims about girls' sexual lives. A week in, one of the girls created a group chat and added everyone involved; that group then grew “*bigger and bigger*” as more people wanted to be added “*to see what was going on.*” The atmosphere around the eventual fight was “*a lot of people just felt like pure excitement from it... it was entertaining.*”

What this case shows

This case shows how a public post can turn a private dispute into a collective conflict, with peer attention, escalating content, and group chats sustaining the pathway to offline violence.

CORE MECHANISM

The central mechanism in this case is audience-driven escalation.

A single post widened the audience. Peer attention increased pressure to respond. Group chat dynamics sustained engagement and drew in additional participants.

Escalation became less about the original issue and more about performing a response in front of others, where withdrawal risked reputational damage.

INTERACTIONAL RISK

This case illustrates how harm emerged through the interaction between:

- Public exposure (the initial post)
- Peer audience dynamics (commentary, side-taking, group chats)
- Contextual conditions (a school environment where conflict and large fights were already relatively normalised)

Exposure alone did not produce the outcome. It became consequential because of how it interacted with peer behaviour and social expectations.

RISK FACTORS

- Public posting rather than private resolution
- Rapid peer amplification
- Sexualised rumours targeting girls' reputations
- Cross-year group involvement
- Limited effective adult intervention once the scale of posting increased

WHAT MIGHT HAVE REDUCED HARM:

- Early interruption before the conflict became public
- Peer challenge to escalation and "drama" dynamics
- Support for young people under pressure to respond
- School responses that engage with online dynamics, not only offline incidents

ESCALATION DRIVERS

- Rapid audience expansion through posting and sharing
- People becoming more confident online than offline
- Pressure to respond or risk appearing weak or as if allegations were true
- Insults becoming more personal as attention increased
- Group chats drawing in spectators and allies "to see the drama"
- Conflict becoming entertainment for others
- A school context in which large fights were not unusual, making escalation easier once enough people were involved

PREVENTION LEVERAGE POINTS

- Intervene early, before audience attention builds
- Strengthen peer-led challenge at the point of escalation
- Improve visibility and response to group chat dynamics
- Provide credible alternatives to retaliation for young people under pressure
- Develop rapid-response capacity within schools (first 24-48 hours)

5.3 Pathway 3: Filming converts passing events into persistent content that creates new grounds for retaliation

The third pathway operated through the recording and circulation of violent or humiliating incidents. Unlike Pathway 1 (which sustained existing conflicts over time) and Pathway 2 (which escalated disputes through audience pressure), this pathway created new harm from events that might otherwise have ended. **The mechanism was transformation: filming converted passing incidents into persistent, shareable content that outlived original moments and generated new conflicts.**

These pathways followed a distinct sequence:

- A violent or humiliating incident occurred – sometimes minor, sometimes serious.
- Someone filmed it. The act of recording was treated as ordinary and unremarkable.
- Footage circulated rapidly through group chats and social media. Audiences expanded far beyond those who had witnessed original events.
- Footage itself became the new provocation. What people needed to respond to was no longer original incidents but the fact that they had been seen, were still circulating, and that silence could be read as weakness.
- New confrontations arose from the footage – retaliation against people in videos, or by people in videos against those who circulated them. Harm extended beyond original disputes into conflicts that original participants had not anticipated.

5.3.1 Footage becomes what people need to respond to – not the original incident

Videos recorded on a Friday could generate sides, commentary, and reputational pressure by Saturday morning, often before any adult was aware. Footage created new audiences that had not witnessed original events but could now form judgements, take sides, and apply pressure. Filming did not simply document violence, it extended it. In many cases, footage functioned as a new provocation entirely separate from whatever had started the confrontation.

The aftermath was where the damage compounded. Young people recounted how being recorded "*getting beat up*" and then having others see it and message about it produced a sense of deep disrespect that persisted long after the physical incident ended. The humiliation continued through the visibility of the clip, the commentary it generated, and the pressure on those shown to respond to what peers had now seen. Practitioners and teachers reported a **bystander culture in which recording was more common than intervening, and where the existence of footage made it harder for incidents to fade.**

5.3.2 Filming disproportionately traps those least able to escape circulation

The consequences of filming and circulation were not evenly distributed. Youth professionals suggested young people with weaker support networks, fewer stable relationships to school or

youth provision, and less capacity to manage social fallout were more likely to remain caught in cycles of retaliation driven by persistent footage, as exemplified in the spotlight below.

Spotlight: Neurodiverse young person trapped by persistent video circulation

In one case, a young girl with neurodiverse needs became recognisable from fight videos that remained online. For her, fighting was understood as a way to resolve a situation – "I'm going to fight this girl and then it's going to be resolved." But the videos stayed, and peers and friends of the original victim sought retaliation, drawing her into further violence.

"She had neurodiverse needs... for her it was always the case of, well, I'm going to fight this girl and then it's going to be resolved... But then she wouldn't realise that those videos are staying there. And then obviously some peers or friends or the alleged victim would seek retaliation, so then she'd have to fight more people... the outcome was that we had to move her away from that area because it was just getting too unsafe for her to stay there because she would be recognised on the streets from those videos." – Youth justice practitioner

5.3.3 Prevention for this pathway targets the circulation mechanism and the culture that sustains it

The priority for this pathway is friction on circulation – slowing the spread of footage in the critical first 24–48 hours (see Section 8 for specific recommendations).

Alongside platform level measures, prevention needs to challenge the culture that makes bystander recording feel routine and consequence-free. Where filming fights is treated as ordinary, and where circulating footage is treated as ordinary, the act of recording becomes part of the escalation process rather than separate from it. Practitioners pointed to the need for work that helps young people understand the consequences of sharing violent footage and reframes filming not as passive observation but as active participation in harm.

5.4 Each pathway requires a different prevention response

The three pathways identified in this chapter operate at different speeds and require different prevention responses, from relational depth for sustained hostility, to peer-led speed for rapid escalation, to circulation friction for filmed content.

All three pathways share a common equity dimension. **Young people with the least stable relationships to school, youth provision, or trusted adults are often the ones least likely to benefit from early warning and de-escalation**, despite being among those most exposed. The full implications for prevention are developed in Chapter 8.

SECTION SIX

Violence Against Women and Girls

Four pathways from online misogyny
to offline harm



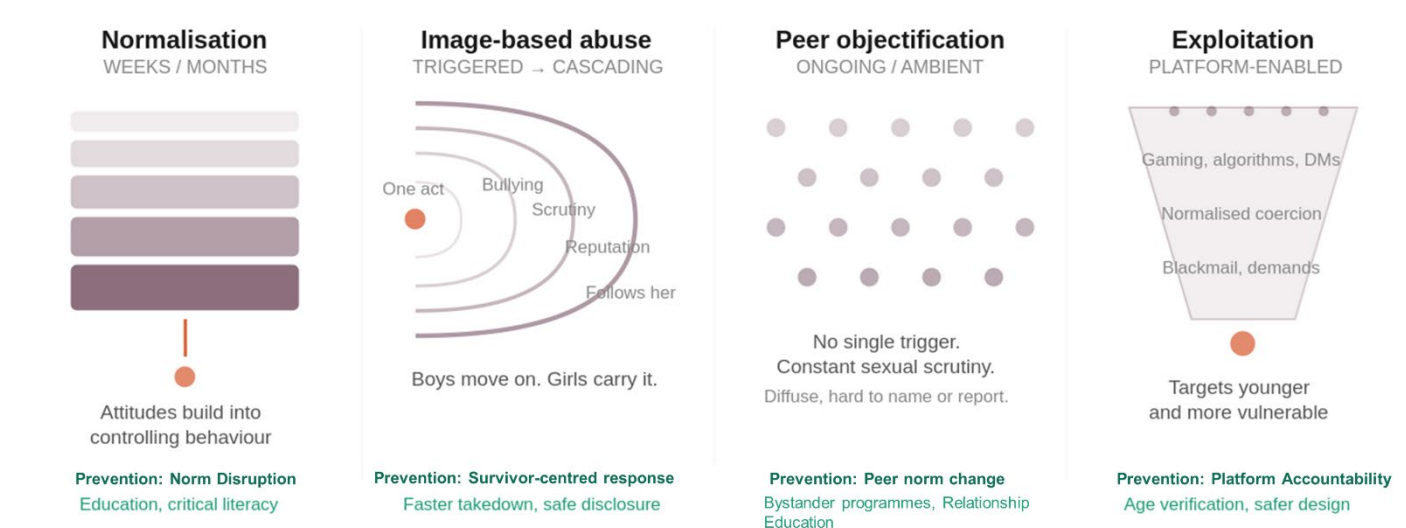
6. Violence Against Women and Girls and Online Misogyny

This chapter traces how online environments connect to VAWG, through pathways that are slower, more cumulative, and more tightly bound to the gradual normalisation of gendered assumptions than the interpersonal violence pathways documented in Chapter 6. Unlike the interpersonal violence pathway, this one has a measurable public-data signal: **the large-scale analysis found a significant same-day association between the misogynistic share of online discourse and police-recorded VAWG.**

The study identified four distinct pathways (Figure 7), each operating through a different mechanism and requiring a different prevention response:

- 1) Cumulative normalisation of controlling relationship expectations through influencer and manosphere content
- 2) Non-consensual intimate image abuse triggering cascading reputational harm
- 3) Peer sexual objectification operating as ongoing culture
- 4) Platform-facilitated grooming and exploitation targeting younger users

Figure 7. Four VAWG pathways identified



6.1 Pathway 1: Manosphere and influencer content normalises controlling relationship expectations – and those expectations translate into behaviour

This was the slowest pathway and the one most consistently described across the qualitative data. Misogynistic content – packaged as advice, self-improvement, humour, or aspiration – was absorbed through ordinary platform use over weeks and months, gradually reshaping what young men understood as normal relationship behaviour and what young women came to expect. **The harm did not arrive as a single moment of radicalisation. It operated through the gradual erosion of what registered as controlling or abusive.**

The pathway typically followed a recognisable pattern:

- Misogynistic content reached young men through algorithmic recommendation, starting from unrelated interests such as fitness, gaming, music, or motivational content.
- The material was packaged as advice, identity, and self-improvement content rather than as overt ideology. Named figures, most prominently Andrew Tate and the wider manosphere ecosystem, were the most frequently referenced sources, though practitioners noted that individual figures mattered less than the formats and assumptions that persisted after specific influencers lost prominence.
- Over time, repeated exposure shifted what young men treated as normal in relationships: jealousy was reframed as care, surveillance as concern, control as confidence, and sexual entitlement as natural.
- Practitioners, teachers, and young women described those shifted expectations translated into observable behaviour in relationships and classrooms – controlling treatment of partners, dismissal of girls' contributions, and sexualised language used as casual address.
- Girls and young women experienced a parallel normalisation environment: appearance-focused content, restrictive femininity scripts, and relationship material that framed submission and self-presentation as aspirational.

6.1.1 Content arrives through ordinary platform use, not active searching – and named influencer ecosystems drive it

The route into misogynistic content was consistently described by participants as algorithmic rather than intentional. Practitioners and young people described boys encountering manosphere material through adjacent interests: gym content, motivational videos, money-making advice, and gaming. Once that material entered a feed, it became normalised quickly because it was woven into content young people already treated as enjoyable or useful. Several young people noted testing this deliberately, observing how quickly a single interaction reshaped what they were shown.

Andrew Tate was the most frequently referenced figure across the qualitative data, though many young people, particularly boys, noted that his personal influence had waned since his arrest. As one youth professional put it: *"we don't even talk about him in schools anymore because all the kids just roll their eyes."* What has persisted though was the wider ecosystem of manosphere content he had come to represent. Individual figures fell out of favour, but the underlying formats and assumptions remained in circulation. This distinguishes the VAWG pathway from the interpersonal violence pathway, where no equivalent influencer ecosystem was identified.

This pattern is consistent with a now-substantial body of UK and international research on how recommender systems deliver misogynistic content to young men. Sock-puppet account experiments by Regehr et al. (2024) found that misogynistic content rose from 13 to 56 per cent of TikTok For You recommendations within five days of an account being created, while Baker, Ging and Brandt Andreasen (2024) found that male-coded accounts on TikTok and YouTube Shorts received manosphere content within 23 minutes regardless of whether they sought it out, with alpha masculinity, fitness and self-improvement content acting as the structural majority of toxic recommendations. UK research with 13- to 14-year-old boys (Haslop et al., 2024) further confirms that Tate-style material is now consumed indirectly through micro-influencer reposts and edits, which helps explain why the underlying ecosystem persists even when individual figures lose prominence.

What this body of research helps explain is why the content felt unremarkable to the young people encountering it: it arrived through the same channels as the gym, gaming and motivational material they were already engaging with, packaged in a register that was indistinguishable from content they considered useful or aspirational. Practitioners emphasised that much of this content did not present itself as obviously harmful. It often looked familiar, persuasive, or even useful on first encounter.

"On the surface it doesn't seem like it could promote misogyny... but once you unpack it, it's enabling unhealthy behaviours towards relationships." – **Teacher**

6.1.2 Appearance pressure and restrictive femininity scripts operate as the other side of the same process

Girls and young women described a parallel online environment shaped by appearance pressure, and highly gendered ideas about what femininity should look like. **Practitioners and young people identified tradwife videos ("trad" being short for "traditional") and dating influencers as carrying restrictive messages about how girls should look and behave in relationships.** One education practitioner noted that many young women followed content that at first viewing looked aspirational but carried *"messages underneath the kind of very surface level content"* about women being submissive and men being dominant.

Young women also described a pervasive body focused environment on which appearance and disordered eating content were difficult to avoid. One participant said TikTok was filled with "a hell

of a lot of videos promoting eating disorders and like unhealthy lifestyles," while another described girls seeing this content and thinking, *"maybe I need to start wearing makeup, maybe I need to start worrying about my appearance."* This was not limited to one platform or one type of content. It was a broader gendered online environment that participants said shaped expectations around desirability, body image, and behaviour in harmful ways.

These accounts find empirical support in research on the platforms named: a content analysis of 200 fitness and fitspo videos by Pryde et al. (2024) found that 78 per cent featured women, 76.4 per cent portrayed thin ideals, and over half of women's depictions involved sexualisation or objectification, indicating that the gendered version of the algorithmic ecosystem young men encounter operates in parallel for young women, with appearance pressure and disordered eating content embedded in the same fitness and lifestyle pipelines.

6.1.3 Normalised attitudes translate into observable behaviour in relationships and classrooms

Practitioners working in VAWG and youth justice settings consistently described boys and young men treating jealousy and surveillance as reasonable, expecting partners to be contactable on demand, and reading control as a natural part of care.

"The younger kids think that... being on the phone and asking a girl where she is all the time is completely appropriate... if that way of thinking becomes embedded while they're still 13 to 15... then it's dangerous." – **Interventions officer**

The same practitioner linked online material directly to *"harmful sexual behaviour"* and to *"a belief that it's okay to control women and to behave in a certain way."* This aligns with the growing literature on technology facilitated coercive control in young relationships, which shows that digital monitoring, compelled availability, and constant check-ins are often normalised rather than immediately recognised as abuse (Baker et al., 2025).

Schools were often where translation from online norm to offline behaviour became most visible. Teachers described boys repeating online misogynistic talking points in classrooms, corridors, and peer conversations, sometimes framed as jokes, and sometimes as obvious truths. They also described the offline consequences for girls.

One young girl reflected that *"a lot of guys think that they can stay in control"*, linking this directly to the stereotype that *"boys should be in control over girls."* Another young woman described seeing boys treat girls with *"disrespect," "a bit rough,"* and not stopping *"even when the girls say stop."* These were not simply attitudinal shifts but translated expectations that affected how girls were treated in everyday school life.

6.1.4 The large-scale data analysis supports the relational character of this pathway

The large-scale data analysis provides population level support across Greater London for the relational character of this pathway. When the analysis explored the relationship between online misogyny and recorded VAWG in more detail, the association was stronger for incidents where the

suspect was known to the victim than for incidents overall. While exploratory rather than confirmatory, this is consistent with the qualitative evidence that **normalised misogynistic attitudes tend to play out within existing relationships – between partners, ex-partners, classmates, friendship groups, and acquaintances – rather than through generalised hostility toward unknown women.** This aligns with WHO's (2021) RESPECT framework and Jewkes et al.'s (2015) findings that gender-inequitable attitudes are among the strongest predictors of intimate partner violence cross-nationally.

The analysis also found, as discussed in Section 4, that the relationship between online misogyny and recorded VAWG appeared only on the same day, with no evidence that it built up over the preceding week. This does not contradict the qualitative evidence of cumulative norm formation. The two strands of evidence are operating at different timescales. The data analysis captures a daily snapshot of the public content climate and its same-day relationship with recorded harm. The slower process of normalisation described by practitioners, in which attitudes are absorbed over weeks and months through repeated exposure, operates at a timescale that a daily analysis is not designed to detect. The two strands of evidence address different layers of the same phenomenon: the data analysis captures the daily climate, while the qualitative research explains the cumulative process through which that climate becomes consequential.

6.1.5 Prevention for this pathway requires norm disruption, not just content removal

As nearly half of misogynistic content in the public dataset was originally authored rather than reshared, this pathway cannot be addressed through takedown alone. The ideas circulate as personal belief, not just as forwarded content. Prevention therefore depends on disrupting the norms themselves, through education, critical literacy, and relational work with credible facilitators. The specific recommendations are set out in Section 8.

6.2 Pathway 2: Intimate Image abuse triggers cascading reputational harm that follows young women across settings

This pathway was faster and more event-driven than Pathway 1, though its consequences were often longer lasting. It was triggered by a specific act: an intimate image shared by an ex-partner, a screenshot taken from a private exchange, or material obtained through coercion that entered peer circulation and generated cascading secondary harms. The original act was often only the start. What followed was a widening pattern of reputational damage that could persist for months or years.

The pathway typically followed a recognisable sequence:

- An intimate or sexual image was shared without consent, usually by a known person.
- The image entered peer circulation through group chats, social media, or direct forwarding. The audience expanded rapidly beyond the original context.

- Cascading secondary harms followed: bullying, sexualised commentary, reputational damage within peer groups, exclusion, and institutional consequences.
- The harm followed the young woman across platforms, schools, and physical settings. Changing school or moving area did not necessarily end it because the content remained in circulation.
- Institutional responses could compound the harm rather than address it, particularly where responses were experienced as victim blaming, inadequate, or racialised.

6.2.1 One act of non-consensual intimate image sharing generates a cascade of secondary harms that can persist for years

Intimate image abuse was one of the most consistently reported harms across the qualitative data. Snapchat group chats were frequently named as spaces where intimate images were circulated or demanded, and where young people were added without consent to groups where material about them was already being shared. Teachers reported sexualised image sharing and coercive digital behaviour appearing among younger age groups than many schools were prepared for, including pupils in Years 7 and 8.

Case Study 3: Image based abuse widening into peer, school, and institutional harm

The most detailed case study traced a long arc from image-based abuse by an ex-partner through sustained bullying, sexualised scrutiny from male peers, and anonymous harassment. The young girl was forced to change schools but described the harm as following her regardless of where she went or what she did.

The harm did not stop at peer circulation. It was deepened by how adults interpreted and minimised what had happened, and by institutional responses she experienced as both inadequate and racialised. [Full case study 3 detail is outlined below.]

Alongside the non-consensual sharing of real images, participants also described the **growing use of digitally manipulated and AI generated content as part of this pathway**. This included images or videos in which a person's face was applied to sexualised or humiliating material, as well as edited content designed to damage reputation within peer groups.

"Deepfakes are already a problem. Children are already being deep-faked online... you can literally deepfake someone with one picture" – **Young girl with experience of VAWG, 19–20.**

Participants emphasised that these forms of harm were particularly difficult to challenge. The AI content often appeared realistic and was easily shared, generating the same patterns of bullying, sexualised scrutiny, and reputational damage as real images. In some cases, the impact was described as more severe than text-based abuse, as visual material carried greater social weight and was more likely to be believed or circulated. Practitioners also noted that the accessibility of AI tools lowers the threshold for this type of harm, increasing both its scale and reach.

These dynamics also intersect with a wider commercial ecosystem of image-based exploitation. Internet Watch Foundation (2025) data show a 127 per cent rise in reports of sexually coerced extortion in the UK, with girls' images now sold commercially in packs and box sets, indicating that peer-driven non-consensual sharing sits on a continuum with more organised forms of exploitation rather than as a separate phenomenon.

6.2.2 The harm is specifically gendered: boys move on, girls carry it

Practitioners and peer participants consistently noted that boys and girls experienced comparable incidents with markedly different consequences. Boys involved in image sharing incidents were more likely to "move on" from reputational harm, while girls continued to "carry" it. This gendered asymmetry was not incidental. It reflected the wider social conditions that attach stigma and scrutiny disproportionately to women and girls in the context of sexual content. This is consistent with emerging research on image based sexual abuse among young people, which treats these harms not as isolated digital incidents but as forms of gendered and relational violence with enduring psychological, social, educational, and relational consequences (Hellevik, Haugen and Överlien, 2025).

6.2.3 Institutional responses can compound the harm – particularly for Black girls

The case study material illustrated that image-based abuse was not only gendered but intersectionally shaped by race and institutional response. Young Black women described being removed from settings rather than supported, being disbelieved, or being treated as more culpable than their peers. This resonates with the wider adultification literature, which shows that Black girls are more likely to be read as older, less innocent, and more responsible for the harms done to them, with consequences for whether they are believed, protected, or blamed (Davis and Marsh, 2020; Epstein et al., 2017).

Disclosure remained a major barrier throughout this pathway. Practitioners working in VAWG services, schools and youth settings consistently identified shame and fear of victim blaming as the main obstacles, compounded by distrust of institutional responses. Young women who experienced image-based abuse often expected to be blamed for having shared the image at all, while others anticipated that adults would dismiss what had happened as banter or ask questions that made it partly their responsibility. Where disclosure did occur, it was most often described as happening through a trusted relationship rather than through a formal reporting route. These findings are consistent with Ringrose and Regehr's (2025) wider research on UK teen image-based abuse, which documents that non-consensual image sharing has become normalised within youth peer cultures, and that young women avoid disclosure because they fear punishment, the removal of their devices, or institutional disbelief, rather than because they lack legal protection.

Whilst the experiences described in the case study below predated the intimate image abuse offences introduced under the Online Safety Act in January 2024, the barriers participants described are not primarily legal or technical. What mattered to the young people in this research

was not whether the law was on their side but whether the people around them would take what happened seriously and respond without making it worse.

6.2.4 Prevention for this pathway requires survivor centred response and faster takedown

This pathway requires both platform-level friction – faster takedown, reporting opportunities that do not re-traumatise – and disclosure designed around what the young person needs. Regulatory and technological responses are developing in this direction. In May 2026, Ofcom strengthened its Illegal Content Codes to recommend that platforms deploy hash matching technology to detect and block the recirculation of non-consensual intimate images, alongside new legislation banning nudification tools and requiring takedown within 48 hours (Ofcom, 2026). Tools such as StopNCII.org now allow individuals to create a digital fingerprint of intimate images so that participating platforms can block recirculation without the images leaving the user's device. However, the effectiveness of these tools depends on whether young people know they exist and feel safe enough to use them, and routine signposting should be built into training for anyone working with young people who have experienced or are at risk of image-based abuse. The full recommendations are in Section 8.

Image-based abuse widening into peer, school, and institutional harm

How relationship abuse became a wider pattern of sexualised and racialised harm once intimate material entered peer circulation

CONTEXT

Across the qualitative material, image-based abuse was repeatedly described as harm initiated by a known person and then intensified through peer circulation, sexualised commentary, and weak or uneven institutional response.

This case study is anchored in the account of a young Black woman reflecting on events that began in adolescence and later unfolded across schools, peer networks, and contact with the police.

WHAT HAPPENED

The case began when an ex-partner non-consensually shared an intimate image online after the relationship ended. The image circulated through peer networks and was followed by bullying, rumours, and anonymous harassment online. In her account, the original act of sharing quickly gave way to a more persistent pattern of humiliation and sexualised attention.

When she moved schools, the image and the rumours had already travelled. Boys at the new school were aware of it and made sexual remarks to her, culminating in a later incident in which a boy sexually assaulted her at school. When she fought back, both were excluded. The exclusion did not interrupt the wider pattern of harm. Instead, it became part of it.

HOW THE HARM WIDENED

This case reflects a wider qualitative pattern in which the harm associated with intimate image-sharing was not limited to the initial act of sharing. Once images entered circulation, they could be shared across platforms, commented on anonymously, and used by peers as the basis for ongoing harassment.

“ At the time where this screenshot of me was circulating... the bullying, the rumours and all of that started...

I even moved to a new school and everyone had already seen it... Ask FM literally made my life hell... people would make statements about my appearance, my afro, my skin tone... and after that got released, I was kind of on edge because there were boys who would walk past me and have sexual remarks to make.

There was one guy in particular who... [sexually assaulted] me when he was walking past. And I got so angry, I literally fought him...

In my head, I was thinking nothing warrants this behaviour. I don't care what you've seen about me online. Nothing warrants this.

So, he got excluded... and I got excluded for fighting, which I don't think is fair because I was defending myself.”

Young woman, VAWG case study

What this case shows

This case illustrates a wider pathway identified across the research: image-based abuse initiated by a known person, amplified through peer circulation, and compounded by school and institutional response.

CORE MECHANISM

The key mechanism in this case is how an image becomes a **shared reference point** that others act on.

The visibility did not remain online. It changed how girls and young women were understood and treated offline: in school, in peer groups, and in their interactions with adults.

The harm was therefore not only that an image had been shared, but that digital circulation turned it into a public reputational marker through which girls were sexualised, judged, and held responsible for the abuse done to them.

QUANTITATIVE CONTEXT

Exploratory sub-analysis found that the link between online misogyny and recorded VAWG was also clearly present and significant in cases involving people known to each other. This is consistent with the relational pattern visible in this case.

**Overall VAWG
model**

B=0.354

P=.002

**Known-suspect
subgroup**

B=0.301

P=.018

Note: known-suspect finding is exploratory and hypothesis-generating.

RISK FACTORS

- Relationship abuse by a known person
- Rapid peer circulation of intimate material
- Gendered reputational harm
- Weak early intervention
- Racialised institutional response

WHAT MIGHT HAVE HELPED:

- Faster image removal
- Responses that focused on support rather than blame
- Clear accountability for the person who shared it
- Trusted routes to ask for help, with support beyond the initial incident
- Staff trained to handle these sensitive issues to work sensitively with both those harmed and those who caused harm
- Institutional responses free from racialised assumptions about which children require protection and which require punishment

ESCALATION DRIVERS

- Portability across platforms
- Anonymous harassment
- Offline sexualised treatment at school
- Punitive response to self-defence

“ I feel like unfortunately race does come into it... I don't feel like anyone was listening to me... they [police] basically took that as, oh, so you're being provocative. And I just don't think you should call a child provocative... calling it provocative is like the adultification of black children or ethnic children. I don't feel like it has changed much. ”

PREVENTION LEVERAGE POINTS

- Early education on image-sharing and consent
- Fast action when content is shared
- Support that does not depend on formal reporting
- Training for schools and police on victim-blaming and adultification
- Joined-up responses across school, family, and services

6.3 Pathway 3: Peer sexual objectification operates as ongoing culture, not a single triggering event

This pathway points to an ambient culture of sexual objectification running through peer groups, group chats, and school environments. The harm here was not triggered by a single act or incident but was diffuse, ongoing, and difficult for young women to name or report because no single moment constituted "the event." Practices of ranking, sexualised commentary, and peer-driven objectification are not new, but the conditions of online environments reshape them in ways that change how they operate and what they require to address.

This pathway is closely connected to the normalisation process described in Pathway 1. Where Pathway 1 traced how misogynistic attitudes are absorbed through repeated exposure to online content, what Pathway 3 describes is what those attitudes look like once they become embedded in everyday peer interaction. The distinction between the two matters for prevention: **once a culture of sexual objectification is established within a peer group, it can sustain itself through group dynamics and social currency regardless of whether any individual member is actively consuming misogynistic content online.** Therefore, addressing it requires working with peer cultures directly, not only reducing exposure to harmful content.

The pathway operated through the following set of dynamics:

- Boys produced and circulated sexualised content about girls as part of ordinary peer interaction; ranking videos, sexual commentary in group chats, appearance-based judgements shared as entertainment.
- Girls were added to group chats without consent where material about them was already being discussed or where sexual ranking was underway.
- The cumulative effect was an environment of ambient sexual scrutiny that shaped how girls navigated school and peer life, from what they wore to whether they spoke up.
- Because the harm was diffuse rather than triggered by a single act, it was difficult for young women to name and for institutions to address.

6.3.1 Ranking, rating, and sexualised commentary operate as peer currency among boys

Teachers and practitioners consistently described sexualised content production as embedded in male peer culture. Boys ranked girls in videos circulated at school. Group chats functioned as spaces where girls were sexually rated, discussed, and judged. This was not experienced as exceptional or deviant behaviour. It was ordinary peer interaction that happened to be deeply harmful to the girls it targeted.

"I think there is also a lot of videos of boys ranking the girls on how they look and what they do. At least it was the case when my daughter came and look mum, have you seen this..." -

Teacher

The sexualised commentary did not stay online. It translated into how girls were spoken to in corridors, how their bodies were discussed in peer groups, and how their participation in school life was shaped by the knowledge that they were being watched and rated. UK research on image-based sexual harassment and abuse in educational settings (Ringrose, Horeck et al., 2023) finds these dynamics operating consistently across schools and across age groups, with peer-led ranking and sexualised commentary functioning as a form of social currency that bears disproportionately on girls. For girls already navigating the appearance pressure described in Pathway 1, this peer-level objectification compounded the environment of gendered scrutiny.

6.3.2 The harm is diffuse and hard to name – which makes disclosure and institutional response more difficult

Unlike Pathway 2, where a specific image or act could be identified and responded to, the harm in this pathway was ambient. Young women described living within an environment of sexual scrutiny rather than experiencing a single identifiable incident. That made it harder to disclose, there was no 'thing' to report, and harder for institutions to address. School policies designed around identifiable incidents of bullying or harassment were poorly equipped to respond to a diffuse culture of objectification that operated through everyday peer interaction.

6.3.3 Prevention for this pathway requires challenging male peer cultures, not just responding to individual incidents

The prevention logic for this pathway is fundamentally about norm change within male peer groups. Practices of ranking, sexualised commentary, and peer-driven objectification have long operated through analogue forms such as yearbook rankings and informal talk among friends. What online environments add is intensity, scale, and visibility. They embed these practices in the daily flow of phone use, lend them vocabulary and legitimacy through manosphere language, and make peer judgement more quantified and more widely witnessed through public commentary, like counts, and shareable content. Incident-based responses will not reach a problem that operates through ongoing culture rather than discrete events. This is the pathway where peer-led challenge may be most important and where adult-led instruction alone is least likely to be effective. Full recommendations are outlined in Section 8.

6.4 Pathway 4: Platform design facilitates grooming and exploitation targeting younger users

This pathway operated through specific platform features and youth facing digital cultures rather than through the peer dynamics central to Pathways 1 through 3. Where those pathways traced how content and attitudes circulated within peer groups, **this pathway focuses on how platform design itself creates the conditions for exploitation, particularly of younger users.**

The pathway followed a recognisable pattern:

- Younger users encountered exploitative content through youth facing platforms, particularly gaming environments and their associated social media ecosystems.
- Coercive dynamics were normalised within those platforms through content that framed blackmail, image demands, and sexual compliance as part of the platform culture.
- Platform features, recommendation algorithms, private messaging, weak age verification, facilitated escalation from initial contact to exploitation.
- Young people's existing understanding of online risk did not equip them to recognise what was happening, for reasons explored below.

The following case illustrates how these dynamics operated in practice, with platform features and youth-facing digital cultures combining to expose a younger user to coercive sexual harm.

Spotlight: Roblox, TikTok, and the normalisation of coercive sexual harm

One young woman described a form of coercive sexual harm linked to Roblox and TikTok content about so-called "com boys" – older male users who offer in-game gifts or Discord perks to underage girls and then escalate to blackmail, coercion, and demands for explicit images or "cut signs" (self-harm). In her account, this was not marginal content but something "absolutely rife," with Roblox-themed TikTok posts showing "*a rich boy Roblox character dancing*" alongside captions about making girls comply.

She described the associated content as appearing "instantly" and being "so embedded within like Roblox algorithm" that "any child that likes playing Roblox is going to get exposed to that content." She also noted that comments were often full of endorsement rather than challenge, and that when users did object, comments could be filtered or removed.

"There's this thing called a com boy, which is an internet boy that you meet on Roblox... they send you coins and in return, you send cut signs and like nudes and they extort you." – **Young woman, VAWG case study** [Full case study 4 detail is outlined below].

The dynamics described in this case sit within a wider phenomenon documented across counter-terrorism, child protection and online safety research. The Council of the European Union's 2025 paper on nihilistic extremist violence identifies Roblox, Discord, Telegram, Snapchat and Minecraft as platforms used by networks loosely grouped under the label of The Com to make initial contact with younger users before moving them to more obscure channels, with gaming environments playing a particularly significant role because their young user bases provide ready access to potential victims. The Institute for Strategic Dialogue's Networks of Harm report (2025) and Argentino's GNET work (2024) further document the use of in-game gifts, mental health subcultures, and youth-facing aesthetics as recruitment routes, with sextortion and the demand for self-harm imagery as recurring features. The participant's account above aligns with this wider pattern, in which platform features and youth-facing digital cultures combine to enable exploitation.

6.4.1 Current risk education does not match how exploitation actually operates online

Practitioners and older participants in our youth cohort noted a significant gap between what young people had been taught about online safety and what they were actually experiencing.

Most young people were confident they knew not to send images to unknown adults online, but this "stranger danger" frame failed to capture the exploitation documented in this research in two important respects.

First, **platform-facilitated exploitation by unknown contacts did not look like what young people had been taught to watch out for.** As the Roblox spotlight illustrates, grooming was embedded within ordinary platform culture through in-game gifts, social currency, and content that normalised compliance rather than arriving as an obviously suspicious approach from a stranger. Young people could encounter and be drawn into coercive dynamics without ever feeling they had crossed a boundary they had been warned about.

Second, practitioners noted that young people rarely recognised similar coercive dynamics could operate "in a relationship or with someone that you trust." School-based online safety education remains too focused on anonymous online risk and not focused enough on relational abuse, coercion, and exploitation among peers and partners. The gap between what young people were taught and what they were experiencing was one of the most consistently identified problems.

These findings are consistent with a wider evidence base showing that grooming and sexual extortion increasingly take place across the online spaces young people use most, including social and gaming environments, and that technology is often used to initiate, escalate, and maintain coercive sexual harm over time (Thorn, 2025). Recent child safety research and regulation point to the role of platform features, including age assurance gaps, private messaging, and unsafe recommendation systems, in increasing children's exposure to exploitative material (Ofcom, 2026).

6.4.2 Prevention for this pathway requires platform accountability and education beyond 'stranger danger'

Prevention for this pathway sits primarily at the structural level, platform accountability for youth facing environments, alongside education that moves beyond the 'stranger danger' frame. Yet prevention also needs to recognise the role online spaces play in young people's lives beyond the harms documented here. For young people whose offline social environments are limited, including those who are neurodivergent or socially isolated, online platforms are a primary route to friendship, entertainment, and community (see Case Study 4 below). Approaches that respond to platform-facilitated harm only by encouraging disengagement risk removing a meaningful source of connection without offering an alternative. The prevention models most likely to support young people in this position are those that link online and offline support, including digital youth work, moderated online spaces, and routes that connect young people to local community provision rather than treating online and offline life as separate domains. The specific recommendations are in Section 8.

Algorithmic exposure, normalisation, and digitally facilitated exploitation

How repeated exposure to harmful content, combined with social isolation and neurodivergence, shaped expectations and increased vulnerability to harm.

CONTEXT

This case shows how risk can build when harmful material appears inside the same online spaces a young person depends on for contact with other people.

The issue highlighted here is that harmful content and risky contact became part of ordinary use, in a context where stepping away from those spaces was not straightforward.

WHAT HAPPENED

The young girl described entering online spaces at a young age and quickly encountering sexual content, including unsolicited images from adult men. She also described having limited social connection offline and relying on online platforms as a primary way of interacting with others.

Her experience of autism shaped how she navigated these spaces. She described difficulty communicating “like everyone else” and feeling that online environments offered a more accessible way to connect. At the same time, this made it harder to distinguish between different types of interaction and intent.

In some cases, this created not only normalisation, but vulnerability. Reliance on online spaces for connection, combined with differences in communication and interpretation, increased the risk of being drawn into harmful or exploitative situations, including those that moved offline. In this instance, she described an incident in which another girl attempted to “set her up”, arranging for her to be unexpectedly confronted by a group of boys.

“

You can make a new account and instantly you're seeing either dance content or violence... it just pushes you straight into it... and if you don't know how to control your algorithm, that's just what you're going to keep seeing.

It mostly just really desensitised me... I don't think it's a good thing... if I see violent things now or sexually explicit things, I'm not shocked by it... everything is so unserious... like people die and people make jokes about it because you see so much of it every day.

It's not enough for boys to be turned on with consent anymore... they're so desensitised from seeing violence online and they're exposed from it so young.

Young girl with experience of VAWG

”

HOW THE HARM WIDENED

The harm widened because exposure was repeated and ordinary. Harmful material arrived through feeds and contact with others, not through deliberate searching.

At the same time, reliance on online spaces for connection meant that disengaging from that content was not straightforward. The same spaces that provided connection also became the spaces where risk accumulated.

The later set-up was significant, but it did not come out of nowhere. It sat within a longer pattern in which harmful content, difficult interactions, and weak points of interruption had already become part of everyday use.

What this case shows

This case shows how repeated exposure to harmful content, combined with reliance on online spaces, can shape expectations over time and increase vulnerability to exploitation and offline harm.

CORE MECHANISM

The key mechanism in this case is that harmful content and risky contact are fed into a space the young person still needs to use.

Recommendation systems and platform features make sexualised and harmful material hard to avoid. Where online communication feels easier than offline interaction, stepping back carries a real cost.

Over time, this changes what feels ordinary. It also makes exploitative intent harder to read early. The result is that harmful contact can be treated as part of normal online life for longer than it otherwise would be, and becomes harder to interrupt before it moves further.

INTERACTIONAL RISK

This case brings together several conditions:

- early exposure to sexualised and violent content
- reliance on online spaces for contact with others
- differences in communication and interpretation that made intent harder to read
- limited offline support
- platforms that kept delivering similar material

No single factor explains the case. The risk came from their combination.

RISK AND PROTECTIVE FACTORS

- Heavy reliance on online platforms for social contact
- Early exposure to sexual and violent content
- Platforms that kept feeding similar material
- Limited trusted adult support
- Safeguarding approaches not well fitted to neurodivergent young people

ESCALATION DRIVERS

- Repeated exposure over time
- Normalisation of harmful content
- Difficulty recognising intent and risk
- Continued engagement due to reliance on online interaction
- Platform design reinforcing exposure
- Situations involving manipulation or “set up”

WHAT WOULD HAVE HELPED:

- Stronger offline sources of connection
- Support that recognised how important online communication was in her life and for her sense of belonging
- Earlier discussion of coercion, grooming, and manipulation beyond stranger-danger
- Tighter platform safeguards around recommendation, adult contact, and reporting

PREVENTION LEVERAGE POINTS

- Platform-level intervention on recommendation systems
- Earlier and more inclusive education
- Support for neurodivergent young people in online contexts
- Strengthening offline support networks
- Improved moderation and reporting mechanisms

6.5 Each pathway requires a different prevention response

The four pathways differ in speed, mechanism, and entry point for prevention, from norm disruption for cumulative normalisation, to platform friction for image-based abuse, to peer level culture change for objectification, to structural accountability for exploitation.

However, all four share a common feature: the harm falls disproportionately on girls and young women already least well served by existing systems. This means those who are youngest, most likely to face victim blaming, or whose previous contact with institutions has reduced trust in disclosure. The full implications for prevention are developed in Chapter 8.

6.6. Bridging the two pathways to prevention

The interpersonal violence and VAWG pathways interact with the same structural conditions but require different entry points for prevention. Figure 8 maps each pathway. The implications and recommendations that follow in Section 8 are summarised around these four levels.

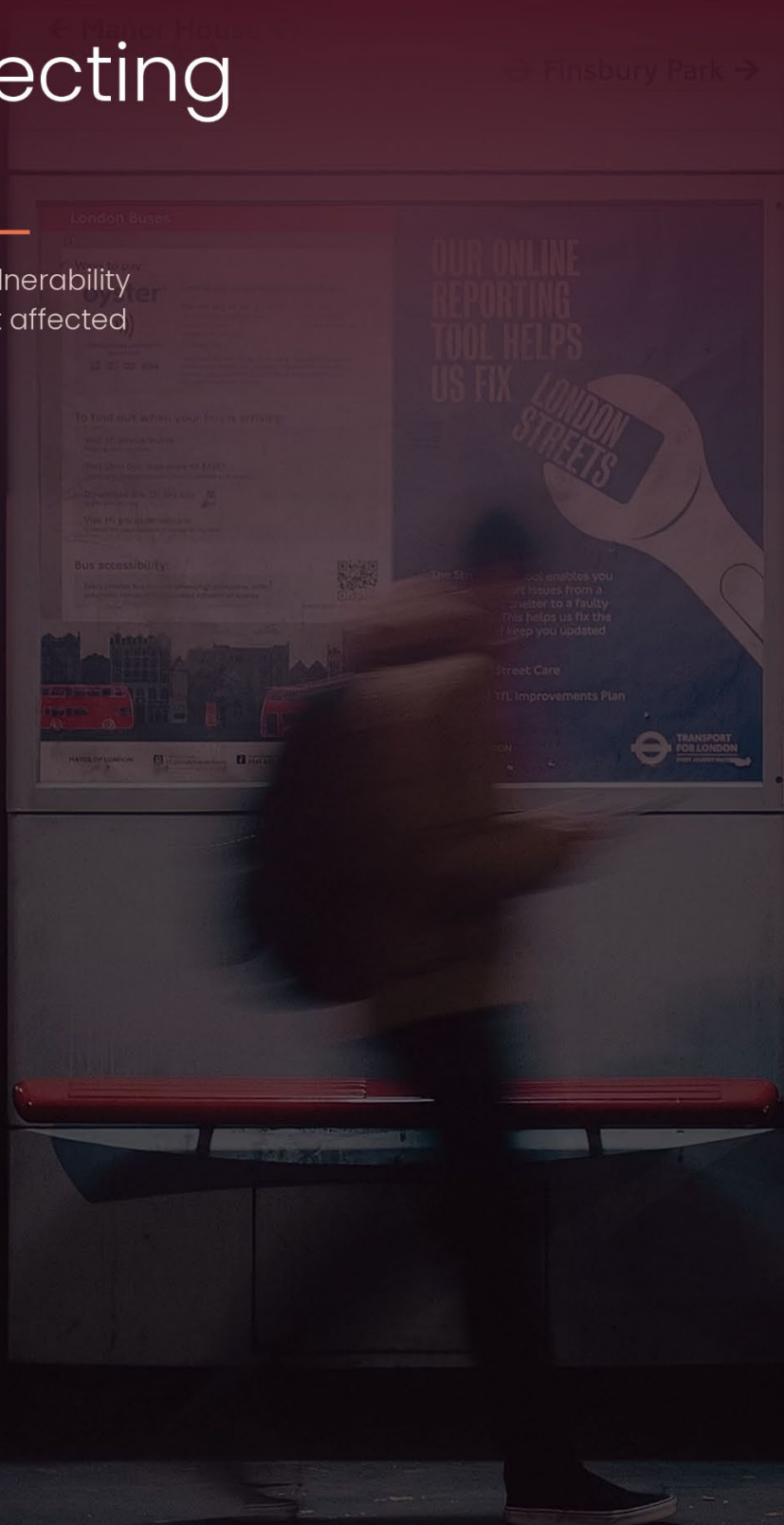
Mapping the Interpersonal Violence and VAWG pathways



Figure 8. Mapping the interpersonal violence and VAWG pathways to the socio-ecological framework. Fill opacity and border weight indicate relative strength of evidence at each level.

Intersecting Risk

What shapes vulnerability and who is most affected



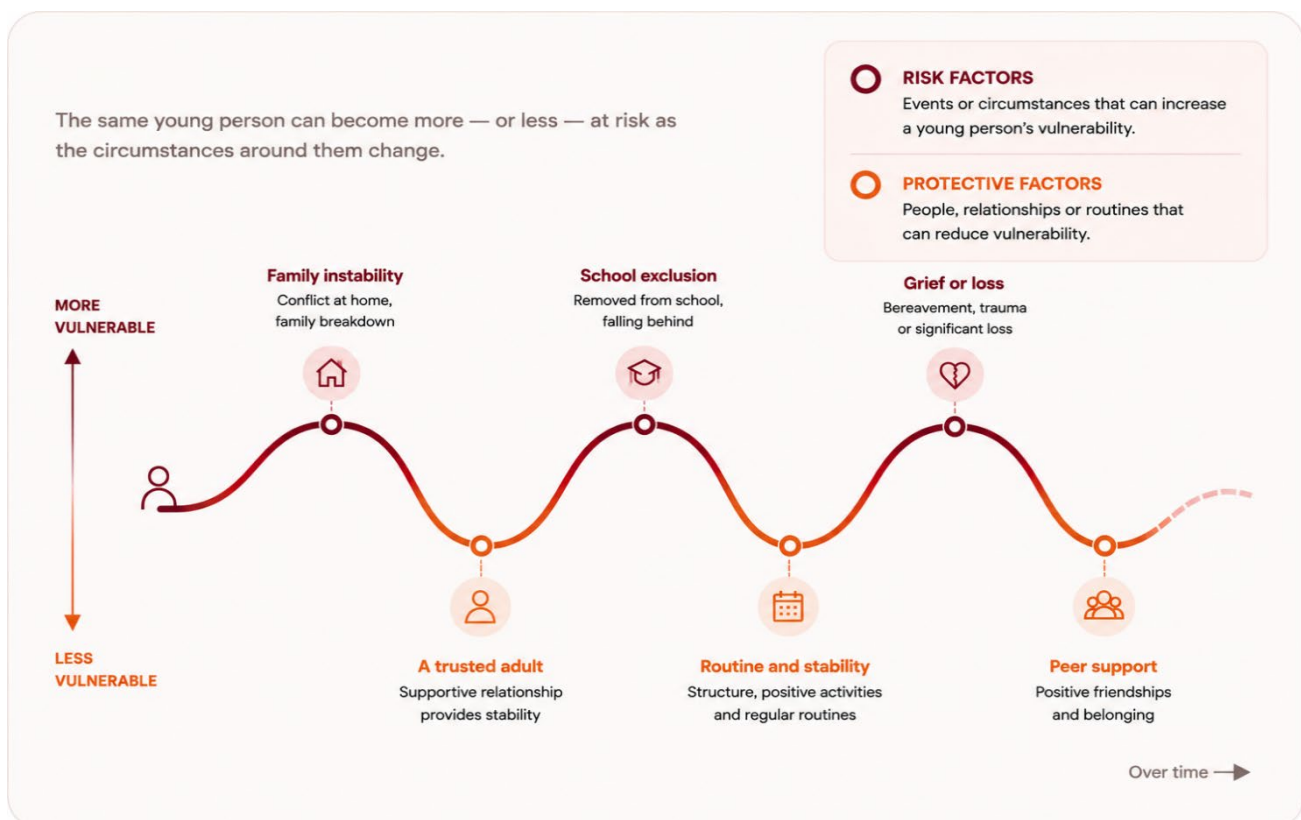
7. Who is most at risk, and what shapes vulnerability?

The same online content did not affect every young person equally, and this was one of the most consistent findings across both strands of the research. The mechanisms described in the preceding sections operated differently depending on who the young person was, what support was around them, and what else was happening in their lives.

Vulnerability was not fixed or categorical but emerged through the combination of circumstances a young person was facing at a given point in time, shifting as those circumstances changed.

Figure 9 illustrates this as a dynamic process: risk can rise during periods of instability, exclusion or loss, and ease when trusted relationships, routine and peer support are present.

Figure 9. *Vulnerability is dynamic, not fixed*



The findings in this section group around four themes:

1. **Structural and relational conditions that concentrate risk.** Practitioners and youth workers consistently identified socioeconomic inequality and the absence of trusted adults as the most consistent background factors. They described the most marginalised young people as the most digitally dependent and the least protected, with harmful content going unopposed where trusted support was absent.

2. **Individual capacity to navigate harmful content.** Teachers, practitioners and young people themselves identified younger adolescents (aged 12-14) as having the widest gap between platform access and critical capacity, compounded by the search for belonging during school transition. Neurodivergent young people faced compounding risks through literal interpretation of content, greater digital reliance, and safeguarding approaches not designed for them. Across all groups, digital fluency did not translate into digital protection.
3. **How gender and race pattern the experience.** Youth justice practitioners and young women described gender as shaping both what young people encountered and what happened when harm occurred. Race shaped exposure, interpretation, and institutional response at every level – through peer cultures, racialised surveillance, and adultification. These dimensions frequently overlapped, with the most at-risk young people facing multiple compounding factors simultaneously.
4. **What the evidence suggests about protection.** Practitioners, peers and young people with experience of violence described resistance to harmful content as possible, but as carrying social costs and requiring support. Trust, not adult presence alone, was what determined whether young people sought help.

7.1 The most marginalised young people are the most digitally dependent and the least protected

Socioeconomic inequality shaped risk in a specific way – it was visible not only in young people's wider circumstances, but in how much protection and offline alternatives were available to them. Practitioners reflected that many of the young people they worked with were *"often the most marginalised... affected by poverty and difficult living conditions,"* with fewer trusted adults and less access to stable routines. Digital spaces seemed to fill that gap for some young people, providing connection and belonging that were harder to find elsewhere.

Teachers described marked differences in how far young people's online lives were monitored or discussed at home. The implication is that the young people most exposed to harmful content are often those with the fewest resources to make sense of it or step away from it.

Where that support was absent, harmful content went unopposed. Practitioners identified a consistent pattern: isolation and the lack of a trusted adult left young people more susceptible to harmful norms and less able to disengage when situations escalated.

"It's easier for them to be influenced... because there's no one to tell them that it's wrong". – Practitioner

This was not a fixed profile, **the same young person could become more vulnerable during periods of family instability, school exclusion, grief, or isolation, and less vulnerable when relationships, routine, and support improved.** In one case, a practitioner recounted a young

woman with multiple layers of trauma, instability, and prior experiences of exploitation who became deeply entangled in group-chat dynamics and digitally organised conflict. Another youth worker contrasted those cycles with young people who had strong support at home, noting that where such support existed, it helped them disengage.

Relational isolation did not map neatly onto material deprivation – a young person from a stable economic background could still lack trusted adults, and a young person in poverty could still have strong relational support. However, **where support was absent, there was no one to offer a different frame or an off-ramp.**

7.2 Younger adolescents have the widest gap between platform access and critical capacity

Younger adolescents, particularly those aged 12–14, were identified as having the widest gap between access and critical capacity. Teachers identified early secondary years as a point of heightened vulnerability. One described the *"negative impacts of social media"* as most visible in *"key stage three, year seven to year nine"*, while also noting a spike during transition into secondary school. Younger adolescents often had broad access to adult platforms before they had developed the judgement needed to contextualise what they were seeing.

The transition into secondary school compounded this. **It was not only that younger adolescents lacked critical capacity – they were also at the point of greatest social flux, adjusting to unfamiliar schools and peer hierarchies at the same time.** Practitioners and teachers described this as a period when young people were actively searching for where they fit and who to be, making them more receptive to online content that appeared to offer answers. The manosphere material documented in Section 4.4 – self-improvement framing, confidence scripts, relationship 'advice' – is specifically designed to appeal to that search. For boys navigating new peer groups, it offered a ready-made identity. For girls adjusting to new social environments, appearance-focused and relationship-focused content arrived at precisely the moment when social comparison was most acute. The gap was not only between access and critical capacity, but between the need to belong and the ability to recognise when the content meeting that need was shaping their assumptions in ways they could not yet see.

The large-scale analysis adds a further dimension to this picture. Higher-volume VAWG days tended to involve a younger recorded case mix, with both the daily proportion of victims aged under 25 and the daily proportion of suspects aged under 25 positively associated with VAWG counts (see Table 1, Section 4.2). **On the busiest days for recorded VAWG, then, younger people were disproportionately represented in the cases recorded.** At the same time, exploratory sub-models comparing under-25 and over-25 victims did not show statistically significant differences in the strength of the misogyny-VAWG association, indicating that the **underlying relationship between online misogyny and recorded violence operates city-wide rather than as a youth-specific effect.**

These findings indicate that **younger people are not necessarily more susceptible to online misogyny than older people in the population-level data, but they are more visible in the cases recorded on the days when VAWG is highest.** This sits alongside the developmental and relational vulnerabilities described above, and it guards against framings that locate misogynistic online content solely as a youth problem. Those producing the content and those committing VAWG offences cut across age groups, even as the developmental dynamics operate with particular force at the moments younger adolescents are navigating.

7.3 Neurodivergent young people face compounding risks that current safeguarding does not wholly address

Among **neurodivergent young people, the key issue described by practitioners was often not exposure itself, but how content was interpreted.** Practitioners working with young people with SEND profiles, including autism and ADHD, identified a particular vulnerability to literal readings of content framed as humour or irony. Multiple practitioners raised this independently, and several identified it as one of the most pressing safeguarding concerns they were encountering.

"The prevalence of neurodiversity in the cohort we work with is much higher than in the general public. A lot of young people struggle to differentiate, thinking, 'I've seen this online, so it must be true and the right way to act, especially towards girls.'" - Youth Justice Practitioner

However, the risk went beyond misinterpretation. **Neurodivergent young people were also described as more reliant on digital spaces for social connection, particularly where offline social interaction was more difficult.** That increased both their exposure to harmful content and the cost of withdrawing from the platforms where it circulated – the same digital dependency dynamic documented in Section 4.7.1 but driven by social communication needs rather than socioeconomic exclusion. Practitioners also identified specific vulnerability to grooming and exploitative contact online, where difficulty reading social cues made it harder to recognise manipulative behaviour from unknown contacts.

Current safeguarding and online safety approaches were widely described as failing to account for these differences. **Generic online safety guidance assumes neurotypical processing;** it does not address how neurodivergent young people interpret content or navigate social cues online (Internet Matters, 2024). Practitioners described this as one of the clearest missed opportunities for targeted prevention identified in this study.

7.4 Gender shapes both what young people encounter online and what happens when harm occurs

Section 4.2.2 showed how online exposure and its consequences are gendered: boys absorb controlling norms, while girls experience their effects. This section addresses a different question: how gender shapes who is most at risk and who is most likely to be supported.

Girls who experienced image-based abuse, sexualised reputational harm, or controlling behaviour were consistently described as less likely to disclose and less likely to be believed when they did. The main barriers noted were fear of victim-blaming, being labelled a “snitch,” and expectations that adults would minimise what had happened to them.

Where harm involved a known person, a partner, ex-partner, or classmate, disclosure was further complicated by ongoing proximity and the social consequences of naming someone within a shared peer group. The quantitative finding that the misogyny-VAWG association appeared stronger in known-suspect incidents reinforces this: the pathway often operates within existing relationships, where disclosure is most difficult and institutional responses are most likely to fall short (see Table 2, Section 4.2).

Boys faced a different vulnerability gap. Those absorbing misogynistic content were rarely identified as at risk because the material was encountered as self-improvement rather than visible extremism. Practitioners described boys who would benefit from early intervention but were not captured by existing safeguarding thresholds. Norms that discourage boys from expressing emotional responses further reduced the likelihood that harmful attitudes would surface before translating into behaviour. This vulnerability gap sits within a wider conversation about how to engage young men with the pressures and assumptions circulating online. Recent work by the Movember Foundation in 2025 reframes these issues as a men's public health concern rather than a matter of individual attitudes to be challenged, an approach the recommendations developed in Section 8 build on.

7.5 Race shapes every level of the pathway – from what young people encounter to how institutions respond

Race did not enter the pathways at a single point. It shaped the content young people encountered online, how they and their peers interpreted harm, how their digital expression was read by adults and institutions, and what happened when they sought help.

The online environment young people navigated was not shaped by violent and misogynistic content alone, but racist, religiously hostile, and racialised content intersected with both. Participants across all groups in the study described encountering explicit racism, Islamophobia, antisemitism, and coded racialised language online, often alongside or embedded within the violent and misogynistic content documented elsewhere in this report. This ranged from highly

hostile commentary around current events to everyday racist "banter" that had become so normalised within peer culture that young people no longer registered it as exceptional.

"With young people, racism is very lightly thrown about online." — Young man, 17

A teacher described the same normalisation process: students absorbed harmful language online to the point where it felt ordinary in writing, even when it sounded shocking spoken aloud. Participants also described moderation as inconsistent, users adapted their wording to evade detection, switching letters for numbers or reordering phrases, meaning racist content persisted even when it was reported. This racist content environment did not sit alongside the violence and misogyny findings as a separate issue. It compounded them, in ways the following subsections set out.

7.5.1 Within peer groups, colourism and cultural policing shape who is believed and who is blamed

Race and colourism did not only shape young people's exposure to harmful content, they also shaped what happened when harm occurred: who was believed, who was blamed, and whose experiences were treated as serious. These are structural dynamics rooted in wider racial inequalities, but they played out most visibly through peer interactions and were amplified by social media in ways that made them harder to escape.

Participants described how colourism could shape credibility within some peer groups. In particular, several accounts suggested that lighter-skinned or mixed-race boys were more likely to be excused when accusations arose, while darker-skinned girls were less likely to be believed or recognised as victims.

As one youth justice practitioner described:

"If he was mixed race... the lighter skinned he was, the more likely he would almost be able to get away with it... People would ask 'Why would this really good looking, light skin, mixed race boy need to rape someone?' ... If the girl is a darker skinned woman... then for sure she would not be seen as a victim." — Youth Justice Practitioner

These accounts point to the role of colourism as one factor shaping how harm is interpreted and responded to within peer contexts, though the extent and consistency of this pattern may vary across different groups and settings.

When sexualised content was shared, the same dynamic applied. The person whose images were circulated was blamed rather than the person who shared them, and race compounded that blame.

"If a girl's videos had been shared, she would be demonised straight away. She would not be seen as a victim at all, and she would be called all the names. No attention would be placed on the person who shared the videos... I think something that they mentioned

something I didn't think of it, was that race played a part as well." — **Youth Justice Practitioner**

Anonymous and semi-anonymous digital spaces created a further layer of harm. Participants described platforms where racialised abuse targeting skin colour and appearance could circulate with little consequence. While the specific platforms have changed over time, the dynamic persists: low-accountability online spaces enable racialised harm that young people are unlikely to report and that moderation systems rarely catch.

What these accounts describe is not a set of isolated peer behaviours but structural racism expressing itself through the everyday interactions young people navigate. Colourism and racialised abuse online point to the same underlying dynamic: social media gives existing racial inequalities new reach and visibility. For prevention, this means that addressing race and colourism cannot be treated as a separate strand of work but must be woven into how programmes understand vulnerability, credibility, and protection across all the mechanisms described in this report.

7.5.2 Black young people's online expression is more readily read as threatening

Race also shaped how young people's online expression was interpreted by the adults and systems around them. Cultural expressions such as drill music and online self-presentation by Black young people are more readily read as threatening or as evidence of criminal intent, rather than as social commentary or identity work (Ilan, 2020; Stuart, 2020). Several practitioners in this study described similar dynamics: **digital expression that would be treated as unremarkable from a white young person being interpreted as indicative of risk when it came from a Black young person.** This narrowed the space for proportionate or protective response and contributed to a wider pattern of racialised surveillance in which Black young people were more likely to be monitored, flagged, and responded to punitively for the same online behaviours.

7.5.3 When harm occurs, institutional responses can compound it for Black children rather than address it

At the institutional level, the most clearly documented dynamic was adultification: the systemic process through which Black children, particularly Black girls, are perceived as older and less innocent than their white peers (Davis and Marsh, 2020). Participants described being removed from settings rather than supported, being disbelieved, or being treated as more culpable than their peers when harm occurred. Recent policing evidence reviews describe the same process, with impacts on whether children's needs are recognised and whether safeguarding responsibilities are fully enacted (College of Policing, 2024; Child Safeguarding Practice Review Panel, 2025).

Adultification bias has been described more broadly as a specifically intersectional bias affecting Black girls through the combined force of race, sex, and age (Epstein et al., 2017). In this study, the consequences were concrete: young people trusted disclosure routes less, were less likely to seek help, and experienced institutional responses that compounded the original harm.

"I feel like because I am black, a black female child at the time, I don't feel like anyone was actually listening to me... the police called me provocative - I don't think you should call a child provocative. It felt like the adultification of Black children, and I don't feel like it was taken seriously." – **Young woman, 22-24**

That account (explored in Case Study 3) speaks to all **three dimensions simultaneously: peer disbelief, racialised interpretation of her behaviour by police, and institutional failure to protect**. Race cannot be treated as a standalone vulnerability factor. It operates across levels – within peer groups, in how online expression is read, and in how institutions respond – and its effects compound rather than sit in parallel.

7.6 Vulnerabilities overlap and intersect – the most at-risk young people face multiple compounding factors

These were not separate vulnerabilities operating in parallel. More often, they overlapped: younger age with weak critical capacity, neurodivergence with literal interpretation of banter, poverty with greater dependence on digital spaces for belonging, racialised gender stereotypes with weaker institutional protection. The young people most at risk were those facing several of these conditions at once – conditions that digital environments interacted with but did not create.

7.7 Protective factors: resistance is possible and trust is the key condition for help-seeking

7.7.1 Young people can and do resist harmful content – but resistance requires support, not just individual willpower

Young people were not simply passive recipients of what platforms delivered. Some actively resisted harmful content – calling it out within peer groups or simply choosing to scroll past. Girls and young women were more likely to report active pushback, though that resistance could itself come at a cost, particularly if it marked them out as humourless or overly sensitive within their peer group. Practitioners observed the same tension: **resistance was possible, but it carried social costs and was rarely sustained without backing from peers or adults**. This is consistent with earlier VRU work on online harms, which highlighted the need to hold together risk, opportunity, and support rather than treating young people's digital lives as uniformly harmful (London VRU, 2023).

7.7.2 Trust, not adult presence alone, determines whether young people seek help

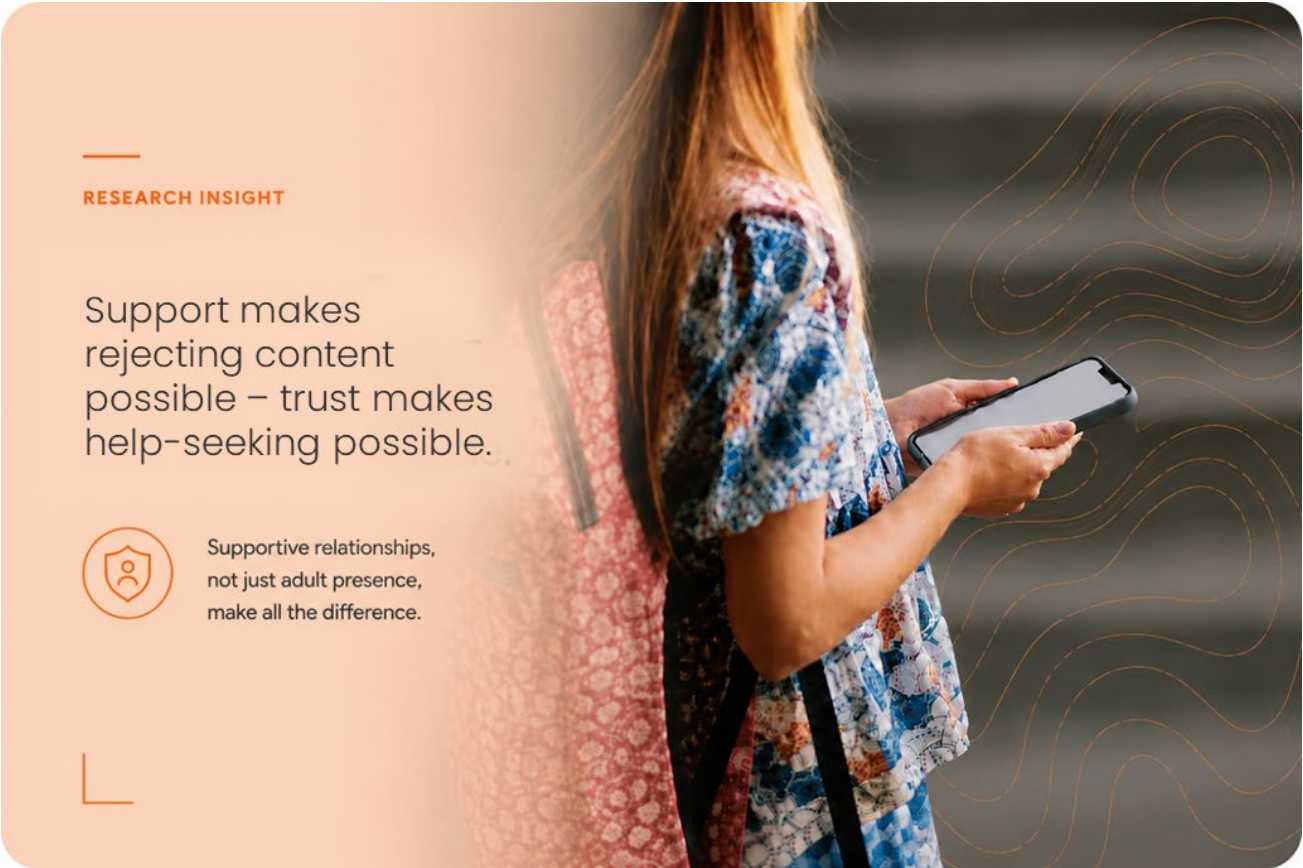
There was strong evidence that trusted adults can make a material difference. But adult presence alone was not enough, young people and practitioners repeatedly described disclosure as shaped by whether the adult was likely to respond supportively or punitively.

One practitioner working on sexual violence prevention noted that many young people were "*quite scared of authority*" and did not believe "*they're not going to get in trouble and they're not going*

to be the ones punished", adding that they often doubted an adult would *"understand or bring justice or help in any way"*. The same practitioner described parents as sometimes *"the ones punishing them instead, rather than reminding them that it's not their fault"*.

By contrast, a youth worker stressed that disclosure *"really depends on the relationships that those young people have with adults"*, describing how her son would come to her because he knew she would respond with her *"youth worker hat on"* rather than simply as a parent.

Trust emerged as the key condition – built through consistency and a perceived willingness to listen before acting. This has direct implications for how disclosure routes are designed and for who is best placed to deliver prevention work, explored further in Chapter 8.



RESEARCH INSIGHT

Support makes
rejecting content
possible – trust makes
help-seeking possible.



Supportive relationships,
not just adult presence,
make all the difference.

Implications for prevention, policy, and practice

8. Concluding Implications and Recommendations

This report identified **seven distinct pathways** through which online environments connected to offline harm among young people in London: three for interpersonal violence and four for violence against women and girls.

Each pathway operated through a different mechanism, at a different speed, and through a different digital infrastructure. A single prevention approach applied uniformly across all pathways will miss the dynamics that make each one consequential. The recommendations below are organised by pathway type, with socio-ecological levels marked within each section to show that effective prevention requires action at every level simultaneously. The recommendations also proceed from the recognition, developed in Chapter 4, that online environments are not only sites of harm but also sites of connection, support, and de-escalation. The aim is not to remove young people from online spaces, but to reform the conditions under which they encounter and engage with them.

The harms documented in this report are not primarily a product of platforms being misused. They flow from platforms functioning as designed – recommendation systems that prioritise engagement, private messaging with no friction on circulation, and sharing architectures built for speed rather than safety. Safer platform design by default, including slower circulation of harmful content in the first 24 to 48 hours, greater user control in messaging environments, and stronger age assurance built into platform architecture, would shift the conditions that produced the pathways identified. Prevention must address that design. How it does so differs by pathway, because each operates through a different part of the platform infrastructure.

8.1 For interpersonal violence

The interpersonal violence pathway operates primarily through private and semi-private channels, where disputes are sustained, pressure builds, and escalation is organised. The **most consequential online activity driving interpersonal violence among young people is therefore largely invisible to public monitoring and content-level moderation systems.**

This shapes what platform reform should focus on for this pathway. The relevant structural changes are within private messaging environments themselves: friction on forwarding, limits on group sizes, age-appropriate defaults, and safeguards built into private messaging features by default. These are design choices about how private channels are constructed, not interventions in the content that flows through them, and remain firmly within the responsibility of platforms. Alongside these, **prevention for this pathway must also operate through the relationships and institutions that can reach young people where these dynamics are unfolding**, particularly youth

workers, peer and family networks, and schools that are able to identify when conflict is escalating before it becomes physical. The recommendations below set out these elements.

INDIVIDUAL

1. **Equip younger adolescents with de-escalation skills, consequence mapping, and trauma-informed support:** Commission practical programmes that build de-escalation skills and strategies for responding to online provocation. Younger adolescents had the widest gap between platform access and the judgement needed to navigate conflict – in cases of sustained hostility, cumulative pressure translated into risk behaviours including weapon carrying, without a single clear trigger. Support should combine therapeutic input with practical help: mediation, safe spaces, and routes to disengage from ongoing conflict.⁴

RELATIONAL

2. **Invest in safeguarding processes and trusted relationships that can recognise early warning signs for online hostility:** Fund youth workers, mentors, and school-based practitioners with the relational depth to identify when online disputes are escalating. The most serious interpersonal violence in this study developed gradually: through persistent online hostility sustained over weeks or months – with no clear trigger that monitoring systems could detect. Early warning depends on adults who know the young person well enough to recognise when cumulative pressure is building.
3. **Build peer bystander capacity at the critical de-escalation point:** Expand peer led bystander and ambassador programmes that equip young people to intervene when disputes become public. There was a consistent window, after disputes became public but before physical responses were committed to, where peers were often more influential than adults in slowing escalation. Programmes should give young people practical tools to interrupt that sequence in real time.

COMMUNITY AND INSTITUTIONAL

4. **Develop rapid response protocols for online-offline escalation across youth settings:** Establish coordinated response pathways that allow schools, youth services, community organisations and safeguarding partners to act quickly when online incidents begin to escalate. Content routinely circulated over weekends and outside school hours, which generated pressure and conflict before any adult was aware. By the time it was visible, escalation had already taken place. In practice, this means building capacity across the settings where young people are known, not only schools but youth clubs, alternative provision, and community organisations, to identify and respond to online escalation within the first 24–48 hours, with staff who have the digital understanding to interpret what is happening. No single

⁴ Whilst this study included young people from aged 12, evidence shows young people are on social media from as young as 8 years old (Ofcom, 2025), so this should be considered.

institution can carry this alone, and effective response requires shared protocols and communication between the services young people move between.

5. **Fund diversionary activities that provide offline belonging and reduce reliance on online identity formation:** Invest in accessible youth provision that offers alternative routes to belonging, recognition, and identity. Where offline opportunities were limited, online environments carried greater weight for identity and status, and in some cases conflict within them became a route to recognition. Strengthening youth provision reduces that reliance and shrinks the audience for online disputes.

PLATFORM AND STRUCTURAL

6. **Introduce friction on circulation and safer group chat design:** Advocate for platform measures that reduce the speed and scale at which harmful content spreads. Violence related content was overwhelmingly reshared rather than originally created; it is this rapid circulation that amplifies conflict and increases pressure to respond. Linked to recommendation 4, this means measures that slow sharing in the first 24–48 hours, alongside stronger safeguards in group chats and messaging environments. The Online Safety Act already recognises that platforms function differently depending on their private messaging capabilities. Snapchat, for instance, is regulated under provisions that acknowledge its semi-private features alongside public content (Ofcom, 2024). The principle of regulating platforms according to how they are actually used, including their private channel dynamics, should be extended and enforced more broadly.

8.2 For violence against women and girls

Where violence related content is largely reshared, **misogynistic content is more often authored and embedded within everyday online activity**. This creates an environment in which harmful norms are produced and normalised over time.

The large-scale analysis found that what mattered was not how much was being posted on any given day, but what proportion of it was misogynistic. Days when misogynistic content made up a larger share of public conversation on X were also days when more VAWG was recorded across London. The report outlined how this operates in practice, through repeated exposure, peer reinforcement, and the gradual translation of online assumptions into relationship behaviour. Based on this, we outline six key recommendations below.

INDIVIDUAL

1. **Build critical literacy programmes that meet young people where they are, delivered by facilitators they find credible:** Commission critical literacy programmes that help young people recognise how seemingly positive content can carry harmful assumptions about

relationships and control. Boys did not typically encounter misogynistic content as overt ideology. It reached them through algorithmic feeds linked to interests like fitness and gaming, packaged as self-improvement and relationship advice that looked useful on first encounter. Girls encountered a parallel content environment of appearance pressure, restrictive femininity scripts, and relationship material framing submission as aspirational, often arriving through the same algorithmic and lifestyle pipelines. Individual influencers may lose prominence, but the underlying narratives persist. Programmes should use real platform examples, be interactive, and be delivered by facilitators credible to young people. These programmes should sit within PSHE and RSE rather than standalone online safety content.

RELATIONAL

2. **Ensure PSHE and RSE implementation addresses the realities of how young people experience online harm:** The 2025 RSHE statutory guidance, mandatory from September 2026, now includes misogyny, harmful online influencers, and toxic digital subcultures within the curriculum. This research reinforces the urgency of that update and points to where implementation should go further. Practitioners consistently described a gap not just in what was covered, but in how it was taught, one-off sessions that did not reflect how young people encountered harm. Critical literacy is needed for young people of all genders on how algorithmic recommendation delivers gendered relationship scripts and identity pressures, including how self-improvement content carries controlling relationship scripts and how oversimplified narratives about toxic masculinity can obscure the structural pressures the manosphere places on young men. Education should also address coercion, image sharing, sexual reputation, and relationship dynamics directly, using scenario-based learning introduced early in secondary school and revisited throughout. Critically, online safety education must move beyond the stranger danger frame, which this research found failed to capture either platform-facilitated exploitation or relational abuse among peers and partners.
3. **Design survivor-centred responses to image-based abuse that combine disclosure, support, and practical assistance:** Since January 2024, sharing intimate images without consent has been a criminal offence under the Online Safety Act, including deepfakes, and platforms are now required to proactively remove such material. However, this research found that the primary barrier to addressing image-based abuse was not legal protection but disclosure. Young women who experienced image-based abuse most often expected to be blamed. Fear of disclosure was the primary barrier to reporting, and where telling someone did occur it was almost always through a trusted relationship rather than a formal route. Responses need to combine accessible disclosure with specialist support: therapeutic services, content removal assistance, and legal guidance should be made available without requiring immediate formal reporting. Staff training should explicitly address victim-blaming and racialised responses.

COMMUNITY AND INSTITUTIONAL

4. **Invest in peer-led and whole-school approaches that challenge the cultures sustaining sexual objectification:** Commission bystander programmes, targeted work with boys, and whole-school strategies that address sexual objectification as a peer culture rather than a series of individual incidents. The research found that practices such as ranking, sexualised commentary, and sharing content about girls functioned as everyday social currency within peer groups, shaping how girls were treated and how they experienced school environments. Responding to individual incidents alone will not shift this dynamic because the behaviour is socially sustained rather than individually driven. Prevention needs to make these behaviours socially unacceptable within peer groups, through approaches that give young people the tools and social permission to challenge them.

PLATFORM AND STRUCTURAL

5. **Advocate for algorithmic reform targeting recommendation pipelines that surface misogynistic content.** Focus regulatory attention on how content is delivered, not only on what content exists. Young people encountered misogynistic content through algorithmic feeds without actively searching for it, often starting from unrelated interests like fitness or gaming. The study found that the relative concentration of harmful content, not overall volume, tracked in line with offline harm. Monitoring should reflect this: scrutinising recommendation systems across major platforms and tracking how prominent harmful content is within the wider feed.
6. **Hold platforms accountable for youth facing environments that facilitate harm and exploitation.** Strengthen accountability for platforms, including gaming adjacent environments like Roblox and Discord, not only mainstream social media. Harmful content was encountered across all major platforms, with specific design features that especially exploit younger children: weak age verification, private messaging, and recommendation systems contributing to risk in each case. Regulation should address everyday design features, not only the most extreme content. Recent developments reinforce the urgency of this recommendation: in March 2026, a US jury found Meta liable for failing to protect young people from exploitation on its platforms, including the facilitation of child sexual abuse and human trafficking, while a separate verdict found that Meta's products were addictive and had caused widespread harm to children. These cases underline that platform accountability is not a theoretical concern but an active area of legal and regulatory action. The harms documented in this report flow from platforms functioning as designed, and regulation should address everyday design features, not only the most extreme content.

8.3 Shared priorities across both pathways

The pathway-specific recommendations above need six cross-cutting conditions; without them, individual interventions are unlikely to gain traction.

Commission against specific pathway types, not generic online safety A single programme will not address all forms of harm. Interpersonal violence interventions should prioritise relational intelligence and peer bystander capacity. VAWG interventions should prioritise norm disruption, critical literacy, and survivor-centred disclosure.	Invest in trusted adults who understand young people's online lives Across both pathways, the most effective protective factor was a trusted relationship with an adult who understood the young person's online world. Investment in the digital awareness of adults around young people is as important as building young people's own critical literacy.
Monitor content climate, not volume Track the relative prominence of harmful content within the wider online environment, not just post counts. This study found that the share of misogynistic content, not the volume, tracked with offline VAWG. If validated across platforms, this could offer a practical early warning tool for identifying periods of heightened risk.	Support peer led approaches without placing the burden on young people alone Peer led challenge was more effective than adult instruction at key moments, but peers can also form the infrastructure through which harm spreads. Programmes must equip young people to intervene while ensuring adult support sits behind them.
Ensure targeted provision reaches the highest-risk groups Younger adolescents, neurodivergent young people, and those without trusted adult support were disproportionately affected across both pathways and least likely to benefit from universal prevention. Commissioning should ensure targeted provision reaches these groups alongside broader approaches.	Build rapid, coordinated response capacity across agencies Both pathways showed that the period immediately following high visibility online incidents, typically 24-48 hours, was when harm multiplied fastest and intervention was most effective. This requires coordinated capacity between schools, youth services, and safeguarding partners, resourced and structured rather than reliant on informal arrangements.

8.4 Further research priorities

The study identifies seven areas where strengthening the evidence base would most improve prevention and policy responses:

1. **Develop the ‘content climate’ concept as a practical monitoring tool:** The finding that relative concentration of misogynistic content, not overall volume, tracked with offline VAWG has potential beyond this research. Further work is needed to test whether this holds across platforms and over time, particularly on those most used by young people. If validated, it could provide a practical early warning tool for identifying periods of heightened risk.
2. **Improve independent research access to social media platform data:** This study worked within the limits of currently available platform data, with X selected because it offered the most workable conditions for systematic, longitudinal, city-level analysis. A stronger evidence base requires structural reform of how online environments are studied, including independent researcher access to public content data across major platforms, continuous exposure monitoring at population level, and clear public health standards against which platform environments can be assessed, in line with the environmental reform framing applied to other commercial determinants of adolescent health.
3. **Understand how exposure translates into behaviour over time:** This study captured short term patterns quantitatively and cumulative effects qualitatively, but the middle ground, how attitudes shift over weeks and months of repeated exposure, remains unmapped. This is particularly pressing for younger adolescents encountering influencer driven content for the first time, where the window between first exposure and embedded norm may be shorter than existing research assumes.
4. **Develop ethical approaches to understanding private channel dynamics:** The interpersonal violence pathway operates largely within private and semi-private channels that are not easily observable. Future work should focus on approaches that generate insight without undermining trust or privacy, through practitioner informed intelligence and participatory research with young people to reconstruct escalation sequences.
5. **Test pathway specific prevention approaches:** Peer led de-escalation for the rapid escalation pathway is the most immediately testable intervention identified in this study. Evaluation should focus on pathway specific outcomes, whether escalation slowed, whether physical confrontation was averted, rather than broad attitudinal measures that may not capture the dynamics this research identified.
6. **Understand the growing role of AI generated content:** AI generated content emerged as a concern across two pathways in particular: fabricated material deepening the insecurity that drives weapon carrying, and deep-faked intimate images compounding the harms of image-based abuse. Both lower the threshold for harm by allowing content to be

manufactured rather than captured, and both are poorly addressed by current moderation tools.

7. **Further investigation into organised online networks targeting vulnerable young people:**

This study identified platform-facilitated grooming and exploitation of younger users as a distinct pathway but did not capture the role of organised online networks, including com networks, in targeting vulnerable young people across platforms and messaging environments. Work currently underway at UCL on this question is an important contribution to the field, and further research is needed on how these networks operate, who they target, and how their dynamics interact with the platform-facilitated pathways identified here.

8.5 A note on under-16 bans

One policy question arose consistently across the research and warrants direct comment: banning social media for under-16s.

A note on social media bans for under-16s

The current debate around under-16 social media bans reflects legitimate concern about the harms documented in this report, particularly the early and routine exposure to harmful content. However, the findings suggest that bans alone are unlikely to address the mechanisms through which harm operates, and restrictions implemented in isolation places the burden of managing harm solely on young people themselves rather than on the platforms that shape their exposure to it.

The interpersonal violence pathway runs through private messaging channels that would persist on any platform young people migrate to. The VAWG pathway operates through the gradual normalisation of gendered expectations, with harmful content embedded in ordinary platform use and encountered without active searching. Participants in this study also highlighted practical limitations, noting that age restrictions are often easily bypassed and that access is likely to shift rather than reduce. The evidence suggests that restrictions may reduce exposure for some users, but are unlikely to be sufficient in isolation.

Effective prevention requires a multi-layered approach, including paying attention to the design of online environments alongside education and relational support that reflect how harm is experienced in practice. Without these wider changes, there is a risk that restrictions displace activity into less visible or less regulated spaces, whilst leaving underlying mechanisms and systemic issues uninterrupted.

8.6 Concluding reflections

The findings of this study point to a coordinated, multi-level response: one that recognises how digital environments interact with the wider social conditions of young people's lives, and that targets the mechanisms through which harm develops. Prevention must combine safer platform design with stronger relationships around young people, tailored to the different pathways identified here. The evidence in this report provides a basis for that shift in commissioning and practice. The next step is to apply it, directing effort to the points at which harm is most likely to build or become harder to stop.



Reference List

Amnesty International (2018a). *Trapped in the Matrix: Secrecy, Stigma, and Bias in the Met's Gangs Database*. London: Amnesty International UK.

Amnesty International (2018b). *Toxic Twitter: A Toxic Place for Women*. London: Amnesty International.

Amnesty International UK (2025) *Toxic Tech: New Polling Exposes Widespread Online Misogyny*. Press release, 21 March. London: Amnesty International UK.

Argentino, M.-A. (2024) '764: The Intersection of Terrorism, Violent Extremism, and Child Sexual Exploitation', *Global Network on Extremism and Technology Insights*, 19 January.

Atiénzar-Prieto, M., Baker, S. and Meyer, S. (2025). 'Young people's attitudes and perceptions of technology-facilitated coercive control: A scoping review', *Aggression and Violent Behavior*, 82, 102054.

Bailey, M. (2010). "'They aren't talking about me..." The invisibility of Black women in intersectionality research', *Meridians: Feminism, Race, Transnationalism*, 10(2), pp. 56–71.

Baker, C., Ging, D., & Brandt Andreasen, M. (2024). *Recommending toxicity: The role of algorithmic recommender functions on YouTube shorts and TikTok in promoting male supremacist influencers*. DCU Anti-Bullying Centre, Dublin City University.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Chan, J., Ghose, A. and Seamans, R. (2016). 'The internet and racial hate crime: Offline spillovers from online access', *MIS Quarterly*, 40(2), pp. 381–403.

Child Safeguarding Practice Review Panel (2025) *"It's Silent": Race, Racism and Safeguarding Children*. London: Child Safeguarding Practice Review Panel.

Citron, D.K. (2014). *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.

College of Policing (2023). *What Works: Reducing Serious Violence*. Coventry: College of Policing. [online] Available at: <https://www.college.police.uk>.

College of Policing (2024) *Adultification Evidence Review*. Coventry: College of Policing.

Contextual Safeguarding Network (2020). *What Is Contextual Safeguarding? An Overview of the Approach*. Luton: University of Bedfordshire.

Council of Europe (2011). *Convention on Preventing and Combating Violence Against Women and Domestic Violence (Istanbul Convention)*. Council of Europe Treaty Series, No. 210. Strasbourg: Council of Europe.

Council of the European Union (2025) *Nihilistic extremist violence: an emerging threat*, Document 15477/25, 19 November. Brussels: General Secretariat of the Council.

Davis, J. and Marsh, N., 2020. Boys to men: The cost of 'adultification' in safeguarding responses to Black boys. *Critical and radical social work*, 8(2), pp.255–259.

Davis, J. (2022) *Adultification Bias within Child Protection and Safeguarding*. Academic Insights 2022/07. Manchester: HM Inspectorate of Probation.

Donato, S., Kulis, S.S., Griner, S.B., Marsiglia, F.F. and Ayers, S.L. (2022).

'Social media and aggressive behavior among adolescents: The mediating role of social norms', *Journal of Youth and Adolescence*, 51(12), pp. 2390–2401.

Department of Health and Social Care (2025) *Men's Health: A Strategic Vision for England*. London: HM Government.

Elsaesser, C., Patton, D. U., Weinstein, E., Santiago, J., Clarke, A. and Eschmann, R. (2021) 'Small becomes big, fast: Adolescent perceptions of how social media features escalate online conflict to offline violence', *Children and Youth Services Review*, 122, 105898.

Epstein, R., Blake, J. and González, T. (2017). *Girlhood Interrupted: The Erasure of Black Girls' Childhood*. [online] Washington, DC: Georgetown Law Center on Poverty and Inequality. Available at: <https://ssrn.com/abstract=3000695>.

Fernández-Planells, A., Masanet, M.-J. and Figueras-Maz, M. (2021). 'Young people's consumption of political information on TikTok', *El Profesional de la Información*, 30(5), e300510.

Flood, M. (2019). *Engaging Men and Boys in Violence Prevention*. London: Palgrave Macmillan.

Frontiers in Psychology (2025) 'Outcomes of image-based sexual abuse among young people: a systematic review', *Frontiers in Psychology*

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Ging, D. and Siapera, E. (2018). 'Special issue on online misogyny', *Feminist Media Studies*, 18(4), pp. 515–524.

Girlguiding (2024) *Girls' Attitudes Survey 2024*. London: Girlguiding.

Glitch UK (2023). *The Digital Misogynoir Report: Ending the Dehumanising of Black Women on Social Media*. [online] London: Glitch.

Grunau, K., Bieselt, H.E., Gul, P. and Kupfer, T.R., 2022. Unwanted celibacy is associated with misogynistic attitudes even after controlling for personality. *Personality and Individual Differences*, 199, p.111860.

Haslop, C., Ringrose, J., Cambazoglu, I. and Milne, B. (2024). 'Mainstreaming the manosphere's misogyny through affective homosocial currencies: Exploring how teen boys navigate the Andrew Tate effect', *Social Media + Society*, 10(1), pp. 1–11. doi:10.1177/20563051241228811.

Hellevik PM, Haugen LA, Överlien C. (2025). Outcomes of image-based sexual abuse among young people: a systematic review. *Front Psychol*.

Hellevik, P., Haugen, G.M.D. and Överlien, C. (2025). 'Technology-facilitated sexual violence against children and young people', *Journal of Gender-Based Violence*, 9(1).

Hoffman, B., Ware, J. and Shapiro, E. (2020) 'Assessing the threat of incel violence', *Studies in Conflict & Terrorism*, 43(7), pp. 565–587.

HOPE Not Hate (2023) *The Manosphere: A Misogynist Threat to Young People*. London: HOPE Not Hate Charitable Trust.

Huntington, C., Pearlman, D.N. and Orchowski, L. (2022) 'The confluence model of sexual aggression: An application with adolescent males', *Journal of Interpersonal Violence*, 37(1–2), pp. NP885–NP909.

Ilan, J. (2020) 'Digital street culture decoded: Why criminalising drill music is street illiterate and counterproductive', *British Journal of Criminology*, 60(4), pp. 994–1013.

Iñigo, A. (2026) 'Navigating the manosphere: A scoping review of research on masculinist and misogynist online spaces', *The Journal of Men's Studies*, 34(1), pp. 166–192. doi:10.1177/10608265251357869.

Institute for Strategic Dialogue (2025) *Networks of Harm: A Victim-Centric Information Resource on the 764 Sextortion Network*. London: ISD

Internet Watch Foundation (2025) *Annual Data and Insights Report 2025*. Cambridge: IWF.

Jane, E.A. (2017) *Misogyny Online: A Short (and Brutish) History*. London: Sage.

Jewkes, R., Fulu, E., Roselli, T. and Garcia-Moreno, C. (2013) 'Prevalence of and factors associated with non-partner rape perpetration: Findings from the UN multi-country cross-sectional study on men and violence in Asia and the Pacific', *The Lancet Global Health*, 1(4), e208–e218.

Kwarteng, J., Coppolino Perfumi, S., Farrell, T., Third, A. and Fernandez, M. (2022) 'Misogynoir: challenges in detecting intersectional hate', *Social Network Analysis and Mining*, 12(1), p. 166.

Lane, J. and Stuart, F. (2022) 'How Social Media Use Mitigates Urban Violence: Communication Visibility and Third-Party Intervention Processes in Digital Urban Contexts', *Qualitative Sociology*, 45(3), pp. 457–475

Lewis, R., Rowe, M. and Wiper, C. (2021) 'Misogyny online: Extending the boundaries of hate crime', *Journal of Gender-Based Violence*, 5(3), pp. 351–366.

London Violence Reduction Unit (2023) *Strategic Needs Assessment: Violence in London 2023*. London: Mayor of London/Violence Reduction Unit.

Lott, K. (2026) 'Techno Music and Luxury Yachts: The Misogynousque Audiovisual Affective Codes of TikTok Manfluencers', *Social Media + Society*.

Malamuth, N.M., Linz, D., Heavey, C.L., Barnes, G. and Acker, M. (1995) 'Using the confluence model of sexual aggression to predict men's conflict with women: A 10-year follow-up study', *Journal of Personality and Social Psychology*, 69(2), pp. 353–369.

Manne, K. (2018) *Down Girl: The Logic of Misogyny*. New York: Oxford University Press.

Marwick, A.E. and boyd, d. (2014) 'Networked privacy: How teenagers negotiate context in social media', *New Media & Society*, 16(7), pp. 1051–1067.

Movember Foundation (2025) *Young Men's Health in a Digital World*. Melbourne: Movember Institute of Men's Health.

Movember Institute of Men's Health and Young Men Research Initiative (2025) *A New Map: Young Men's Digital Worlds in 2025*. Melbourne: Movember Institute of Men's Health.

Müller, K. and Schwarz, C. (2021) 'Fanning the flames of hate: Social media and hate crime', *Journal of the European Economic Association*, 19(4), pp. 2131–2167.

Müller, K. and Schwarz, C. (2023) 'From hashtag to hate crime: Twitter and anti-minority sentiment', *American Economic Journal: Applied Economics*, 15(3), pp. 270–312.

Ofcom (2024) *Online Safety Enforcement Guidance: Protecting People from Illegal Harms Online*. London: Ofcom. [online]

Ofcom (2025) *Experiences of Engaging with the “Manosphere”*. London: Ofcom. [online]

Over, H., Bunce, C. J., Konu, D., & Zendle, D. (2025). What do we need to know about the manosphere and young people’s mental health?. *Child and Adolescent Mental Health*.

Over, H., Bunce, C., Baggaley, J., & Zendle, D. (2025). Understanding the influence of online misogyny in schools from the perspective of teachers. *PloS one*, 20(2), e0299339.

Plan International (2020) *Free to Be Online? Girls’ and Young Women’s Experiences of Online Harassment*. Woking: Plan International.

Renström, E.A. and Bäck, H. (2024) ‘Manfluencers and young men’s misogynistic attitudes: The role of perceived threats to men’s status’, *Sex Roles: A Journal of Research*, 90(12), pp. 1787–1806. doi:10.1007/s11199-024-01538-2.

Renström, E.A. and Bäck, H. (2026) ‘A threatened masculinity? The role of status threats and anger in misogynistic engagement’, *Group Processes & Intergroup Relations*.

Regehr, K., Shaughnessy, C., Zhao, M. and Shaughnessy, N. (2024) *Safer Scrolling: How Algorithms Popularise and Gamify Online Hate and Misogyny for Young People*. London: Association of School and College Leaders, in partnership with University College London and the University of Kent.

Seager-Smith, L. (2025) ‘What tech-facilitated abuse looks like in teenage relationships’, *Internet Matters* (Expert Opinion). [online]

Stuart, F. (2020) *Ballad of the Bullet: Gangs, Drill Music, and the Power of Online*

Infamy. Princeton, NJ: Princeton University Press.

Sugiura, L. (2021) *The Incel Rebellion: The Rise of the Manosphere and the Virtual War Against Women*. London: Routledge.

Thiara, R. and Roy, S. (2020) *Reclaiming Voice: Minoritised Women and Sexual Violence – Key Findings*. London: Imkaan and University of Warwick.

Thorn (2024a) *Mitigating the Risk of Generative AI Models Creating Child Sexual Abuse Materials: An Analysis by Child Safety Nonprofit Thorn*. San Francisco, CA: Thorn. [online].

Thorn (2024b) ‘Will I be believed? How deepfakes are adding fears to youth experiencing sextortion’. *Thorn Blog*, 25 September. [online].

Thorn (2024c) *Youth Perspectives on Online Safety, 2023: An Annual Report of Youth Attitudes and Experiences*. San Francisco, CA: Thorn. [online]

Thorn (2025) *Deepfake Nudes & Young People: Navigating a New Frontier in Technology-Facilitated Nonconsensual Sexual Abuse and Exploitation*. San Francisco, CA: Thorn.

Thorn and All Tech Is Human (2024) *Safety by Design for Generative AI: Preventing Child Sexual Abuse*. New York: All Tech Is Human and Thorn.

UK Government (2021) *Tackling Violence Against Women and Girls: Strategy*. London: HM Government.

Vodafone and Global Action Plan (2024) *The Rise of the Aggro-rhythm: How Social Media Algorithms Target Boys with Harmful Content*. London: Vodafone UK.

Williams, M.L., Burnap, P., Javed, A., Liu, H. and Ozalp, S. (2020) ‘Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated

crime', *The British Journal of Criminology*, 60(1), pp. 93-117.

Women's Aid (2023) *Influencers and Attitudes: How Will the Next Generation Understand Domestic Abuse?* London: Women's Aid.

World Health Organization (2010) *Violence Prevention: The Evidence*. Geneva: World Health Organization.

World Health Organization (2016) INSPIRE: *Seven Strategies for Ending Violence Against Children*. Geneva: World Health Organization.

Youth Endowment Fund (2024) *Children, Violence and Vulnerability 2024: What Role Does Social Media Play in Violence Affecting Young People?* London: Youth Endowment Fund.

Youth Endowment Fund, Open Innovation Team, Department for Education (2023) *The role of systems of support in serious youth violence: evidence and gaps Deep Dive*

Technical Annex

Social Media Data Collection

The quantitative analysis is based on text-classified posts and cannot directly observe visual or video content, which qualitative participants identified as the primary format through which harmful material is encountered. To partially address this, the data collection strategy included targeted keyword and hashtag queries designed to surface posts associated with visual content, including fight-related hashtags and terms commonly used to share video material. This allowed some indirect capture within the dataset. However, this approach cannot fully represent the scale or impact of visual media as experienced on platforms such as TikTok, Snapchat, and Instagram, and is dependent on how content is labelled or described. The quantitative dataset is therefore likely to underrepresent harmful content in visual formats

Operationalisation of Hazards and Harms

This section sets out how the two harms examined in this study, interpersonal violence and VAWG, were operationalised across the three evidence streams: social media monitoring, offline police and ambulance data, and qualitative fieldwork.

Interpersonal violence

Social media strand

Violence-related discourse on X (formerly Twitter) was captured through curated keyword search strings developed from previous literature, validated through the co-creation panel with young people, and tested iteratively to exclude unrelated noise. These included terms and hashtags related to fights, weapons, threats, and street violence (e.g. #britishfights, #schoolfight, "ill pattern you," "get packed"), as well as proximity searches combining violence terms with weapon or location references.

The resulting posts were passed through OpenAI's omni-moderation model, using its violence and violence/graphic categories. A post was coded as violent if either category was returned as true. This produced a final dataset of 10,978 confirmed violence-related posts.

Figure 9. *Interpersonal Violence Search String*

```
("#britishfights" OR "#fightsatschool" OR "#fightvideos" OR "#fightingvideos" OR "#ukfights" OR "#streetfights" OR "#ghettofight" OR "#hoodfights" OR "#fightvids" OR "#shegotbeatup" OR "#schoolfight" OR "#getthofights" OR "ill pattern you" OR "i'll pattern you" OR "catching an M" OR "catching a M")1

("fights" OR "fight" OR "fighting" OR "brawl" OR "brawls" OR "punch" OR "punches" OR "beat up" OR "beat the shit") NEAR/10 ("knife" OR "knives" OR "machete" OR "machetes" OR "streets" OR "street" OR "car park" OR "gang" OR "gangs" OR "shank" OR "shanked") NOT "music" NOT "new album" NOT "new song"
```

Offline data strand

Metropolitan Police Service (MPS) records of homicide, serious assault, robbery, and sexual offences involving young people were used as the primary measure, supplemented by

London Ambulance Service (LAS) data to capture incidents that go unreported to police – research suggests 90–95% of violent injuries treated by medical services are absent from crime records. Daily counts were used in the modelling.

Qualitative strand

Four anchor case studies traced how online interactions contributed to offline violence through in-depth interviews, focus groups, and guided digital journey mapping. Journey mapping involved young people tracing online-offline escalation sequences, mapping offline incidents alongside relevant online content (comments, threats, DMs, fight videos), and documenting emotional and behavioural responses at each stage. Participants included justice-involved young people, at-risk youth aged 12–24, peers, teachers, youth workers, and practitioners, recruited through trusted charities, youth justice organisations, and schools.

Violence against women and girls

Social media strand

VAWG-related discourse was captured through keyword search strings targeting misogynistic abuse, manosphere/incel terminology, and gendered slurs – developed from previous literature and validated with the co-creation panel. These included manosphere-specific terms (e.g. “foid,” “female hypergamy,” “AWALT,” “gynocentrism”), gendered slurs in proximity to derogatory descriptors, and rape-supportive or victim-blaming language (e.g. “she was asking for it”).

Posts were passed through a misogyny classifier to identify misogynistic language. The resulting day-level modelling measure was the daily share of posts classified as misogynistic. This produced a final dataset of 25,509 confirmed misogynistic posts.

Figure 10. Excerpt of VAWG Search String

```
("she's a" OR "shes a" OR "shes an" OR "she's an" OR "she's" OR "she has" OR  
"this girl is" OR "this bitch is" OR "women are" OR "girls are") NEAR/3  
("whore" OR "whores" OR "psycho" OR "psychos" OR "roastie" OR "roasties" OR  
"hoe" OR "hoes" OR "slag" OR "slags" OR "tart" OR "tarts" OR "airhead" OR  
"airheads" OR "witch" OR "witches" OR "jailbait" OR "old hag" OR "old hags"  
OR "slapper" OR "slappers" OR "trollop" OR "trollops" OR "menopausal" OR  
"mingers" OR "munter" OR "munters" OR "mutton" OR "muttons" OR "virago" OR  
"too sensitive" OR "post-wall" OR "past her prime" OR "wall-hitter" OR "used  
goods" OR "emotional" OR "irrational" OR "hormonal" OR "hysterical" OR "chad  
chasing" OR "chad chaser" OR "an alpha widow" OR "overreacting" OR "high  
maintenance" OR "TPTT" OR "too pretty to think" OR "banshee" OR "let herself  
go" OR "a primma donna" OR "dried up" OR "dried-up" OR "airhead" OR "drama  
queen" OR "drama queens" OR "moody" OR "neurotic" OR "sensitive" OR "karens"  
OR "karen" OR "high maintenance" OR "stacys" OR "ssa" OR "b!tch" OR  
"looksmxing" OR "low-tier normie" OR "mid-tier normie" OR "high-tier normie"  
OR "low tier normie" OR "mid tier normie" OR "high tier normie")1
```

Offline data strand

MPS-recorded VAWG offences were used as the primary measure, with daily counts used in the modelling. Compositional controls included victim and suspect age profiles, suspect-victim relationship, and suspect ethnicity to adjust for whether higher-volume days systematically involved different case profiles.

Qualitative strand

VAWG was investigated through case studies, depth interviews, and focus groups tracing how online misogynistic content connected to image-based abuse, coercive relationship behaviour, peer sexual harassment, and exploitation. The focus was on influencer ecosystems, normalisation of controlling relationship expectations, gendered reputational harm, and platform-facilitated grooming.

Classifier details

Two machine-learning classifiers were used to identify harmful content within the collected social media streams. This section describes each classifier, the definitions it applied, and its measured performance.

Misogyny classifier

Posts in the VAWG stream were classified using the BERT+EAR model developed by Attanasio et al. (2022), a BERT-based classifier with entropy-based attention regularisation. The model was evaluated on the Evalita 2018 MISOGYNY English benchmark, which defines misogyny as hate or prejudice against women, encompassing stereotyping and objectification, dominance of men over women, victim-blaming (justifying abuse of women while rejecting male responsibility), sexual harassment and threats of violence, and discrediting women. The model achieved an F1 score of 0.71 on the MISOGYNY English test set. This classifier was selected over alternatives with marginally higher headline accuracy because it demonstrated stronger performance on unintended bias. This was important because the social media collection strategy was intentionally keyword-enriched, meaning the candidate stream already contained a high density of gender-related language. Conventional classifiers can overreact to identity terms such as "woman" and misclassify non-hostile posts as misogynistic; the BERT+EAR model's attention regularisation reduces this tendency. The resulting day-level modelling measure was the daily share of posts classified as misogynistic. Further detail on the model's development can be found in Attanasio et al. (2022) and the Evalita benchmark documentation (see: <https://ceur-ws.org/Vol-2263/paper009.pdf>).

Violence classifier

Posts in the interpersonal violence stream were classified using OpenAI's omni-moderation-latest model. A post was coded as violent where either the violence category or the violence/graphic category was returned as true. The violence category identifies content that promotes or glorifies violence, or celebrates the suffering or humiliation of others. The violence/graphic category identifies content that depicts death, violence, or serious physical injury in extreme graphic detail. In external validation work conducted by the project team using the Gab Hate Corpus, the violence category was tested as a simple binary classification: posts labelled as violent formed the positive class, and all remaining posts were treated as non-violent. Under this setup, the model achieved an F1 score of 0.82. A small number of additional posts (177) linked to hashtags associated with violent video content were also manually incorporated into the classified dataset, to account for material that the text-based classifier would not reliably capture. Further detail on OpenAI's moderation categories can be found at: <https://github.com/openai/moderation-api-release>.

De-duplication

Before classification, retweets were removed from both streams to reduce amplification bias and ensure the analysis reflected distinct content rather than viral circulation of a smaller number of posts. In the violence stream, approximately 80% of collected posts were retweets; in the misogyny stream, approximately 54% were retweets. This de-duplication step is analytically important: the high retweet rate in the violence stream is itself a finding about how violent content circulates and is discussed in the main report (Section 4.4). The retweet proportions for each stream were retained as a separate variable for the circulation and authorship analysis presented in the main findings.

Technical notes on quantitative analysis and modelling

The quantitative phase of the project set out to achieve two key objectives: (1) measure the prevalence and exposure to misogynistic and violent content online among young Londoners; (2) map temporal and spatial trends between online content and offline violence where suspects are under 25. It was originally proposed that online social media and offline violence records would be temporally (day) and spatially (LSOA) linked and statically modelled using fixed-effects regression, mirroring previous work that found associations between online race and religious hate speech and offline hate crime (see Williams et al. 2020). Due to limitations in obtaining geo-located social media posts introduced by API changes on the platform X, the design was changed to a city-level time-series. The inferential models were therefore estimated by day across Greater London as a whole. While the chosen design could not estimate whether online misogynistic or violent posts were concentrated in particular LSOAs, a key benefit over fixed-effects regression was the ability of a time-series design to place emphasis on temporal precision supported by a wide variety of robustness checks that would likely have become unstable under the original planned design (due to online data sparsity across LSOAs). What may have been lost in spatial granularity has been partly offset by gains in temporal refinement and robustness testing. As Metropolitan Police Service data were geocoded, a supplementary postcode level analysis aided interpretation of the models by showing whether days with elevated misogynistic social media posts coincided with greater spatial concentration of VaWG across Greater London geographies.

Conceptual Framework

Previous research has shown that an association exists between online hate speech and offline hate crime. Williams et al. (2020) linked police crime, census and Twitter data to establish a temporal and spatial association between online hate speech that targets race and religion, and offline racially and religiously aggravated crimes in London over an eight-month period. Müller & Schwarz (2021) established that antirefugee sentiment on Facebook posted by the AfD was strongly associated with crimes (especially assault) against refugees in German municipalities, finding a one standard deviation higher right-wing social media usage was associated with a 10% higher probability of an antirefugee incident (Facebook and internet outages saw the effect decrease by 18% and 53% respectively). The same authors found President Trump's tweets about Muslims predicted increases in xenophobic tweets by his followers, cable news mentions of Muslims, and hate crimes on the following days, where a one standard deviation increase in Twitter usage was associated with a 32% larger increase in anti-Muslim offline hate crimes since the 2016 presidential primaries. The result was mainly driven by anti-Muslim assault and vandalism, and not other crimes, indicating that it was not driven by increased reporting or recording (Müller & Schwarz 2023).

These studies indicate, empirically and conceptually, that online and offline harms are not part of separate social worlds. Rather, they are different manifestations of overlapping social processes. Misogynistic and violent content on social media may normalise hostile beliefs, amplify harmful behaviours, and make certain forms of expression appear acceptable beyond the fringe. At a minimum, online discourse may operate as an ambient signal of hostility in the wider environment, and at a stronger level, it may contribute to escalation by falsely justifying grievance, rewarding aggression, and increasing the visibility of harmful narratives and targets.

But these studies also warn that any observed online to offline relationship should not be interpreted simplistically as direct causation from post(s) to event(s). If daily alignment is present, it may arise through several mechanisms. While online discourse may shape norms and behaviour, it may also reflect the same underlying social conditions that drive offline harm, such as news cycles, public controversies, or localised flashpoints. For that reason, this quantitative analysis treats online activity both as a potential influence on offline phenomena and as a real-time barometer of the wider hostile climate.

Hypotheses

Based on previous research two hypotheses were formulated ahead of model building:

- Hypothesis 1: Days with a higher share online misogynistic content would be associated with higher same-day levels of police-recorded violence against women and girls.
- Hypothesis 2: Days with a higher share of online violent content would be associated with higher same-day levels of police-recorded general violence.

Data and Methods

Data collection

Three sources of day-level data for Greater London were compiled for this part of the study:

(1) data were requested from the Metropolitan Police Service, providing the police-recorded offline outcomes and the case-mix controls used in the main models. For offences related to violence against women and girls, notifiable offence, victim and suspect data were filtered on the MPS definition of VaWG as determined by the NPCC, which is: At least one of the victim on the investigation is female and aged 10 or over, and the offence either: 1. Has a domestic abuse (DA) flag; OR 2. Has an honour-based abuse (HBA) flag; OR 3. Is one of the following offences, whether or not they have been flagged as DA or HBA: Forced marriage; Violence with injury; Female genital mutilation; Public fear, alarm or distress; Homicide; Exploitation of prostitution; Rape; Stalking and Harassment; Other sexual offences; and Modern Slavery.

For violent offences, notifiable offence, victim and suspect data were filtered based on the following criteria: the investigation includes at least one suspect who meets either of these conditions: 1. Their age at the time of the incident is under 25; OR 2. Where age is estimated, the suspect "age from" value is under 25 AND any of the following offence criteria: The offence group is Robbery; OR The offence subgroup is Homicide (Violence against the Person) or Violent Disorder (Public Order Offences); OR The offence falls under one of these Home Office Classifications: 8N – Assault with Injury; 104 – Assault without injury on a Constable; 8P – Racially or Religiously Aggravated Assault with Injury; 105A – Assault without injury; 8S – Assault with Injury on a Constable; 105B – Racially or Religiously Aggravated Assault without Injury; 8T – Assault with injury on an Emergency Worker (other than constable); OR The offence meets the definition of knife crime or gun crime; OR The investigation contains a weapon-related tag as specified in the request. To avoid duplicate inflation when building the police daily series, the offence file was treated as the event spine, with linked victim and suspect records collapsed to event level before aggregation.

(2) ambulance data were obtained through SafeStats from the London Ambulance Service. Violence-related records were extracted using incident, complaint, and injury markers, then deduplicated and collapsed per day. This produced an alternative, incident-level measure of general violence for triangulation. This dataset should be interpreted as a secondary triangulation outcome rather than as a direct replication target to the police violence series (see Technical Appendix for results).

(3) online data were collected from the platform X using M&C Saatchi collection services. Posts were classified as containing misogynistic or violent content and then assembled as daily aggregates for Greater London. The online series contained raw post counts, non-quote variants, keyword measures, hashtag measures, combined keyword and hashtag measures, and engagement indicators. A Neutral Activity Index was also constructed as a daily control for general platform activity, with an observed NAI window from 23 September 2024 to 30 April 2025.

The X queries for misogynistic and violent posts were intentionally designed to maximise the amount of likely harmful content collected.

VaWG and misogynistic keywords for social media post collection: we focused on misogynistic abuse/slurs and Incel and Manosphere-specific terminology, extracted from social media platforms and previous literature:

("foid" OR "foidlet" OR "female hypergamy" OR "femoid" OR "feminazi" OR "crazy karen" OR "thots" OR "gynocentrism" OR "AWALT" OR "women belong to men" OR "all women are like that" OR "she was asking for it" OR "shes asking for it" OR "she's asking for it" OR "that's a woman's place" OR "thats a woman's place" OR "that's women's place" OR "thats women's place" OR "get back in the kitchen" OR "make me a sandwich" OR

"dumb hoe" OR "dumb bitch" OR "dumb slut" OR "dumb blonde" OR "dumb whore" OR "a stacy" OR "a karen" OR "this karen" OR "crazy bitch" OR "chad chaser" OR "chad chasing" OR "drama queen" OR "spinster" OR "minger" OR "bitchy woman" OR "bitchy women" OR "ice queen" OR "ball-breaker" OR "man-hater" OR "manhater" OR "gold diggers" OR "golddiggers" OR "golddigger" OR "gold digger" OR "gold digging" OR "beta orbiters" OR "beta oribiting" OR "emotional women" OR "emotional woman" OR "irrational women" OR "irrational woman" OR "hormonal woman" OR "hormanal women" OR "hysterical woman" OR "hysterical women") NOT "#เซนต์สกุพงษ์" NOT "#Saint_sup" NOT "#MingEr") ("she's a" OR "shes a" OR "shes an" OR "she's an" OR "she's" OR "she has" OR "this girl is" OR "this bitch is" OR "women are" OR "girls are") NEAR/3 ("whore" OR "whores" OR "psycho" OR "psychos" OR "roastie" OR "roasties" OR "hoe" OR "hoes" OR "slag" OR "slags" OR "tart" OR "tarts" OR "airhead" OR "airheads" OR "witch" OR "witches" OR "jailbait" OR "old hag" OR "old hags" OR "slapper" OR "slappers" OR "trollop" OR "trollops" OR "menopausal" OR "mingers" OR "munter" OR "munters" OR "mutton" OR "muttons" OR "virago" OR "too sensitive" OR "post-wall" OR "past her prime" OR "wall-hitter" OR "used goods" OR "emotional" OR "irrational" OR "hormonal" OR "hysterical" OR "chad chasing" OR "chad chaser" OR "an alpha widow" OR "overreacting" OR "high maintenance" OR "TPTT" OR "too pretty to think" OR "banshee" OR "let herself go" OR "a primma donna" OR "dried up" OR "dried-up" OR "airhead" OR "drama queen" OR "drama queens" OR "moody" OR "neurotic" OR "sensitive" OR "karens" OR "karen" OR "high maintenance" OR "stacys")

Violence keywords for social media post collection: we focused on violence terminology and hashtags related to violent video content, extracted from social media platforms and previous literature:

("#britishfights" OR "#fightsatschool" OR "#fightvideos" OR "#fightingvideos" OR "#ukfights" OR "#streetfights" OR "#ghettofight" OR "#hoodfights" OR "#fightvids" OR "#shegotbeatup" OR "#schoolfight" OR "#getthofights") ("fights" OR "fight" OR "fighting" OR "brawl" OR "brawls" OR "punch" OR "punches" OR "beat up" OR "beat the shit") NEAR/10 ("knife" OR "knives" OR "machete" OR "machetes" OR "streets" OR "street" OR "car park" OR "gang" OR "gangs" OR "shank" OR "shanked") NOT "music" NOT "new album" NOT "new song".

In the final social media datasets, there were 125,919 posts (54.51% retweets) in the misogyny collection, and 138,355 posts (79.71% retweets) in the violence collection. Given the targeted collection strategy employed, the online datasets should not be read as the prevalence of misogyny or violence on X in Greater London. These are theory-led outputs within intentionally harm-enriched London collection streams, designed to capture relative day-to-day shifts in harmful content composition rather than platform-wide prevalence. This design is suitable for analysing relative temporal fluctuations in harmful discourse, but it is not suitable for estimating absolute prevalence of harmful content on X in Greater London.

Dependent Measures

Two sets of dependent measures were derived from the datasets. From the MPS data, the VaWG dependent variable was the daily count of police-recorded violence against women and girls in Greater London. For the general violence analysis, the dependent variable was the daily count of police-recorded violence offences. Additional police violence subtypes were also derived for secondary analyses in order to test whether any online signal was concentrated in narrower forms of violence compared to the aggregate series.

A set of dependent measures came from the London Ambulance Service (LAS) via SafeStats, and was used as a triangulation source for general violence only. Because the LAS files did not contain the perpetrator, relationship, or contextual fields needed to distinguish violence against women and girls from general violence with sufficient validity, they were not used to construct a VaWG outcome. Instead, violence-related LAS records were extracted using incident, complaint, and injury markers, deduplicated, and collapsed to one Incidentid per day to form an incident-level general-violence series. This LAS outcome should therefore be read as a stress test across a different administrative system rather than as an outcome of equal evidential status to the police series.

Independent Measures

The key independent measures were daily social media indicators derived from the platform X. For the VaWG models, the main exposure was the daily share of posts classified as misogynistic, while for the general violence models, the main exposure was the daily share of posts classified as violent. These share measures were preferred to raw counts because they capture the composition of the discourse environment while adjusting for total activity. For clarity, each share measure was defined as the number of posts classified into the relevant harmful category divided by the total number of posts in the corresponding daily Greater London X collection stream for that measure.

For misogyny classification we used the BERT+EAR model of Attanasio et al. (2022), a BERT-based classifier shown to perform competitively on the Evalita MISOGYNY benchmark, while improving bias metrics.⁵¹ The training dataset defines online misogynistic content as: Hate or prejudice against women, including stereotyping and objectification; dominance of men over women; victim blaming i.e. justifying abuse of women while rejecting male responsibility; sexual harassment and threats of violence; and discrediting women. We prioritised a classifier with strong unintended-bias performance due to the keyword-enriched social media collection approach taken. Conventional models are known to overreact to identity terms such as “woman” and misclassify non-hostile posts as misogynistic, thereby contaminating the day-

⁵ <https://arxiv.org/pdf/2203.09192>, <https://ceur-ws.org/Vol-2263/paper009.pdf>

level signal. A model that performs better on unintended-bias metrics reduces this risk and yields a more defensible measure of genuinely misogynistic content.

For violence classification we used OpenAI's omni-moderation-latest model.⁶ We treated a post as violent if it was flagged in either the violence category, defined by OpenAI as content that depicts death, violence, or physical injury, or the violence/graphic category, defined as content that depicts death, violence, or physical injury in graphic detail.

The F1 score for the misogyny classifier was 0.71. The F1 score for the online violence classifier was 0.82 (where 1 is the highest score). Of the misogyny collection, 45% of original posts (non-retweets) were classified as containing misogynistic content. Of the violence collection, 38% of original posts were classified as containing violent content.

The classification-based measures for violence were supplemented by simpler rule-based indicators for comparison work and to capture violent video content that included specific keywords and hashtags. These included keyword-only, hashtag-only, and combined keyword plus hashtag measures, together with versions excluding quote posts.

The Neutral Activity Index (NAI) was included as a key activity control, with a log version used in the daily count models. It was constructed as a daily measure of background, non-substantive posting activity to distinguish shifts in harmful content from shifts in general platform traffic. Using the same collection parameters as the substantive online datasets, including English-language posts geofenced to Greater London, it queried a set of neutral everyday phrases designed to capture routine social media use rather than misogynistic or violent content. The search query was: ("good morning" OR "good afternoon" OR "good evening" OR "good night" OR "see you" OR "on my way" OR "at home" OR "at work" OR "back to" OR "have a look" OR "thank you"). The resulting daily series covered 23 September 2024 to 30 April 2025, which was the maximum retrospective period available based on the limitations of M&C Saatchi collection services.⁷

To account for differences in the composition of recorded offences from one day to the next, the models included a set of daily case-mix measures derived from linked offence, victim, and suspect MPS records. These variables were constructed by first identifying, for each

⁶ <https://developers.openai.com/api/docs/models/omni-moderation-latest>

⁷ The full study period ran from 1 May 2024 to 30 April 2025. However, the primary models for VaWG and General Violence were truncated to 23 September 2024 to 30 April 2025, which was the observed Neutral Activity Index window, so that the analysis relied only on directly observed online activity rather than estimated earlier values.

offence, whether particular victim or suspect characteristics were present, and then calculating the daily share of offences or linked suspects with those characteristics. In the VaWG model, these included the daily share of victims aged under 25, the daily share of linked suspects aged under 25, the daily share of linked suspects known to the victim, and the daily shares of linked suspects whose ethnicity was recorded as unknown or as Black.

In the general violence model, the case-mix controls included the daily share of victims aged under 25, the daily share of offences with any linked under-25 suspect, and the daily shares of linked under-25 suspects known to the victim, with ethnicity unknown, and recorded as Black. Importantly, these are not person-level causal predictors. Rather, they are day-level composition controls that describe the mix of cases coming through the system on a given day. For example, the daily share of offences with any linked under-25 suspect refers to the proportion of violence offences recorded on a given day that had at least one linked suspect aged under 25. It does not mean the proportion of all suspects who were under 25. Similarly, the suspect relationship and ethnicity variables refer to the proportion of linked under-25 suspect records on that day with those characteristics. These variables therefore help adjust for whether higher-volume days also involved more youth-linked violence or different linked under-25 suspect profiles.

Black ethnicity was included as a pragmatic case-mix control rather than as a theoretically privileged category. With only 220 daily observations in the observed NAI window, it was not feasible to include a full set of ethnicity proportions without over-parameterising the model and introducing instability. Black ethnicity was retained because, among the available minority ethnicity indicators, it showed sufficient prevalence and day-to-day variation to function as stable adjustment term. It is therefore treated only as a control for daily case mix and recording completeness, not as a substantive causal variable.

Additional online controls and audit variables were retained where useful, including total post volume and, for sensitivity analysis, non-quote variants of the main classifier-based measures.

Methods of Estimation

Poisson generalised linear models with a log link (using robust standard errors) were used as dependent variables were counts and because preliminary checks indicated mild to moderate overdispersion. Because the data are ordered daily series, inference was also checked against temporal dependence diagnostics and alternative covariance treatments in supplementary analysis (see Technical Appendix).

For the police VaWG models, the preferred specification used the daily police VaWG count, and controls for day of week, linear time trend, victim and suspect composition, and log NAI. The main exposure was the daily misogyny post share. Raw misogyny counts were estimated

only as secondary comparisons. A range of robustness checks was also used, including ex-quote variants, lead falsification, trimming, month fixed effects, exclusion of the first observed week, exclusion of low-post days, and exclusion of the documented X outage day (see Technical Appendix).

For the police general violence models, the same broad logic was followed. The main exposure was the daily share of violent posts, with keyword and hashtag variants retained as secondary comparators. Models were estimated first for the combined violence outcome and then, where appropriate, for selected subtype outcomes. This allowed a theory-led narrowing process in which weaker measures could be dropped and more defensible specifications retained.

For the LAS models (see Technical Appendix), the same date-based daily framework was used, but the offline controls were adapted to the fields available in the ambulance data. Daily case-mix measures included sex composition, mean age, youth shares, police-call share, alcohol-related share, weapon injury share, and sexual injury share. The final LAS exercise was treated as triangulation rather than as a replacement for the police analysis, since the ambulance data capture related but institutionally distinct aspects of violence.

Technical Appendix

This research reports model statistics in the usual way to show the size and precision of the estimated relationships. However, it does not rely on any single statistical threshold on its own, and in addition it judges the strength of the evidence by looking at whether the main findings remain consistent across a wide range of checks, including alternative model specifications, falsification tests, event-window controls, and other sensitivity analyses. This means the conclusions are based not just on whether an individual result is statistically significant, but on whether the overall pattern is robust to reasonable changes in how the analysis is conducted.

Misogyny modelling

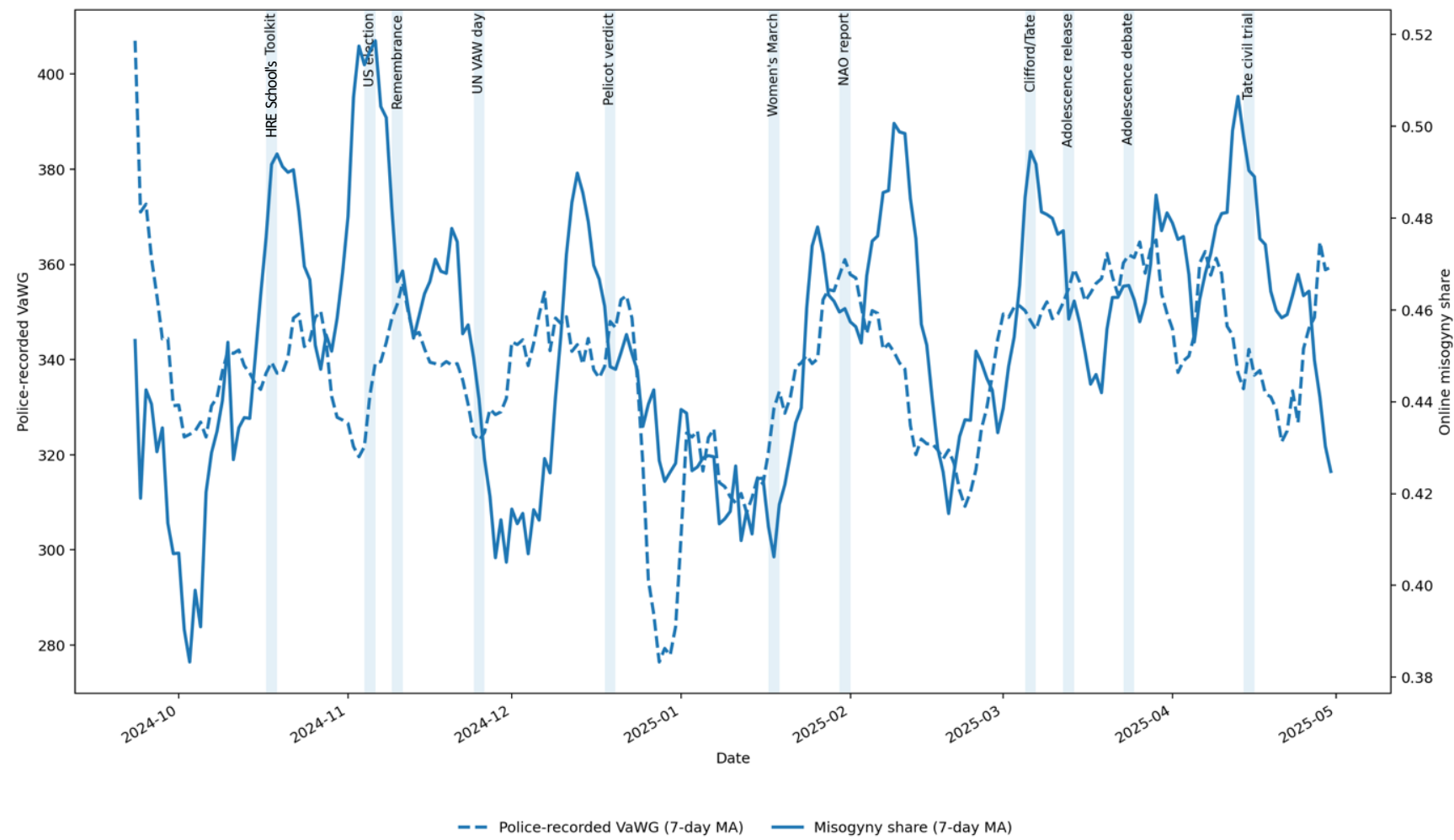
A set of robustness checks was used to test whether the same-day misogyny share held under a variety of conditions. The effect remained when quote tweets were excluded, with the ex-quote share coefficient slightly larger than the main estimate. A lead falsification test showed that tomorrow's misogyny share did not predict today's VaWG count, which is less consistent with a purely mechanical shared-trend explanation. To assess whether the association reflected a delayed rather than contemporaneous process, we estimated a distributed-lag model covering day 0 to day 7. The strongest coefficient was observed on day 0, $B = 0.300$, $p = .015$. Although one individual lag at day 2 was also positive, the lag block covering days 1 through 7 was jointly null, $p = .625$. The evidence therefore points to a predominantly same-

day association rather than a coherent delayed or cumulative lag structure. The result also held when month fixed effects replaced the linear trend, $B = 0.326$, $p = .005$, and when month fixed effects were combined with the full event-window set (see below), $B = 0.344$, $p = .004$, ruling out seasonal cycles as a driver. The result also held with the exclusion of the first observed week, exclusion of a documented UK X outage day on 10th March 2025, trimming of the top 1 percent of misogyny-share days, trimming of the top 1 percent of outcome days, and exclusion of the lowest 1 percent of post-volume days.

We acknowledge that the lead falsification and lag structure do not establish causality, and they do not rule out the possibility that both online misogyny and offline VaWG respond to the same unmeasured daily social conditions. However, they do provide useful temporal discipline, and the null lead test and the dominance of the same-day term make the pattern less consistent with two unrelated trends moving together over time. In addition, because the general violence model was null, any unmeasured same-day factor capable of explaining the VaWG result would need to be specifically gender-related rather than a generic shock affecting violence overall. If the observed association simply reflected a general underlying condition we would expect to see a similar same-day spike in aggregate violent crime, which we did not. We therefore stop short of a strong unidirectional causal claim, but view the results as suggestive of a temporally specific same-day association between online misogyny and offline VaWG that is not easily explained by simple shared trend alone.

As an additional check on omitted-variable bias from high-salience media cycles, we re-estimated the VaWG model with a set of event-window controls. These covered the Gisèle Pelicot verdict, the January 2025 UK Women's March protests, the National Audit Office report on violence against women and girls, court coverage linking the Kyle Clifford case to Andrew Tate-style misogyny, the Netflix release of *Adolescence*, the subsequent schools and Downing Street debate around the series, and the April 2025 Andrew Tate civil-trial coverage (see Figure 11). The core coefficient for daily misogyny share remained positive and statistically significant throughout (in the base preferred model, $B = 0.354$, $p = .002$; with a single any-event-window control, $B = 0.377$, $p = .001$; with all seven event windows entered separately, $B = 0.381$, $p = .002$). These checks suggest that the observed same-day association between misogynistic online discourse and police-recorded VaWG is not simply a by-product of a small number of high-profile news shocks.

Figure 11 *Online Misogyny and VAWG*



A supplementary exploratory spatial analysis examined whether higher-misogyny days were associated with changes in the distribution of incidents across postcodes, rather than simply changes in total citywide volume. The results (Table 3) suggested that any apparent spatial pattern was driven by daily incident volume rather than a substantive redistribution of VaWG across London. In naïve models, higher misogyny share was associated with lower postcode concentration, $B = -0.001$, $p = .002$, and with incidents being spread across a greater number of distinct postcodes, $B = 0.342$, $IRR = 1.41$, $p = .002$. However, both associations disappeared once the models were adjusted for postcode-coded daily volume and coverage. In the adjusted specifications, misogyny share was unrelated to postcode concentration, $B = 0.000$, $p = .806$, and to the number of distinct postcodes affected, $B = 0.002$, $IRR = 1.00$, $p = .894$. There was likewise no evidence that higher-misogyny days placed a greater share of incidents into pre-existing hotspot postcodes, either in the naïve model, $B = 0.045$, $OR = 1.047$, $p = .864$, or in the adjusted model, $B = 0.038$, $OR = 1.039$, $p = .886$. Combined, these spatial extensions suggest that the citywide misogyny signal is better interpreted as a same-day volume effect than as evidence of meaningful spatial concentration or diffusion in where VaWG incidents occurred. In plain terms, days with more misogynistic online content did not seem to change where VaWG happened across London, they just appeared to be days with more incidents overall. These exploratory results suggest that the main citywide association is better interpreted, at least in these data, as a day-level volume relationship rather than as evidence of a measurable shift in spatial concentration.

Table 3: Spatial analysis

Outcome	Naive model effect of misogyny share	Naive p	Adjusted model effect of misogyny share	Adjusted p
Postcode concentration (HHI)	$B = -0.001232$	0.0022	$B = 0.000051$	0.8063
Distinct postcodes hit per day	$IRR = 1.409$; $B = 0.342952$	0.0024	$IRR = 1.002$; $B = 0.002355$	0.8937
Hotspot share	$OR = 1.047$; $B = 0.045483$	0.8644	$OR = 1.039$; $B = 0.038453$	0.8856

We also conducted robustness checks on conditioning on features of the dependent variable (compositional controls). We conducted a sensitivity-to-inclusion analysis to assess whether conditioning on daily case composition materially distorted the focal coefficient. The misogyny-share coefficient remained notably stable across nested specifications. Relative to the full preferred model, removing the ethnicity block changed the coefficient by only 0.08 of a robust standard error, removing the relationship control changed it by 0.20 standard errors, and removing the age controls changed it by 0.18 standard errors. Even when all compositional controls were removed together, the coefficient changed by only 0.59 standard errors and remained positive and statistically significant. This suggests that, while these variables serve as useful guards against day-level case-mix volatility, they are not the primary drivers of the observed association. Furthermore, the direction of change (coefficient increases from 0.286 to 0.354 in the full specification) suggests that the case-mix variables function as

precision controls that reduce baseline noise, rather than confounders that inflate the association. These checks suggest that the focal coefficient is not highly sensitive to the inclusion of the case-mix controls, although the causal status of those controls remains open to debate.

As additional sensitivity checks, the day-level VaWG model was extended to incorporate a series of exposure- and virality-oriented measures derived from the original post-level social media data. First, a daily misogynistic impression share variable was constructed as the proportion of all daily impressions attributable to posts classified as misogynistic. When added to the main robust Poisson specification, the core misogyny-share coefficient was essentially unchanged, moving only from $B = 0.354$, $p = .002$ in the baseline model to $B = 0.351$, $p = .003$ with impression share included, while the impression-based predictor itself was not statistically significant, $B = 0.003$, $p = .884$. In a further comparison model where misogynistic impression share replaced misogynistic post share as the main online predictor, the effect also remained non-significant, $B = 0.014$, $p = .407$. The same pattern held across a broader set of measures, including non-quote misogynistic impression share, counts and shares of high-impression misogynistic posts, mean and median impressions per misogynistic post, and a retweet-weighted misogyny share. None of these additional variables was statistically significant either when added to the baseline model or when used as the main predictor in place of misogynistic post share. Across these specifications, the coefficient for daily misogynistic post share remained positive and statistically significant, varying only slightly across models, which suggests that the main VaWG finding is not explained away by simple visibility or virality proxies and that, in these data, the relative concentration of misogynistic posts remained more informative than impression-based or repost-weighted alternatives.

The VaWG model showed residual overdispersion and some serial dependence, so standard errors should be interpreted cautiously. However, the substantive result was stable across sensitivity checks. Re-estimation using a negative binomial model had a limited effect on the misogyny-share coefficient, from $B = 0.354$ to $B = 0.353$, and significant in both cases. The coefficient also remained positive and significant under HAC-style covariance estimation and in a rerun without police-derived daily case-mix controls. The aggregate violence result was likewise unchanged, with the violence-share coefficient remaining non-significant across specifications. Robust Poisson models were retained as the primary specification for consistency and comparability, with the supplementary checks indicating robustness under imperfect fit.

Violence modelling

In the aggregate violence model, the coefficient for daily online violence share was positive but not statistically significant, $B = 0.090$, $SE = 0.059$, $p = .128$. The estimate therefore does not support a clear same-day association between online violence and police-recorded violent

crime overall. While very small effects cannot be ruled out, the main result is best interpreted as null. Supplementary checks on the police general violence and LAS model files showed some serial dependence in Pearson residuals, but HAC-style covariance estimation did not change that conclusion.

As a triangulation exercise, London Ambulance Service data were processed into an incident-level daily violence series and modelled separately. These analyses did not reveal a clear positive relationship between the online violence measures and the LAS outcomes. The classifier-based violence measures were null across the combined LAS outcome and the subtype screens. A narrower final screen, using the combined LAS violence outcome with the primary violence-share measure, its ex-quote variant, and a keyword-based comparator, was also null. The LAS results are therefore best treated as a negative triangulation check rather than as a competing core analysis.

We also tested whether the null result might reflect the use of post share rather than broader exposure or diffusion measures. Adding violent impression share to the main day-level model reduced the coefficient for daily share of violent posts slightly, from $B = 0.090$, $p = .128$ to $B = 0.070$, $p = .239$, while the impression-share term itself was not statistically significant, $B = 0.018$, $p = .342$. Replacing post share with violent impression share also produced a null result, $B = 0.024$, $p = .207$. The same was true for non-quote violent impression share, counts and shares of high-impression violent posts, mean and median impressions per violent post, and a retweet-weighted violence share. None of these measures produced a reliable same-day association with aggregate offline violence. Overall, the violence results remain null across the main model, the LAS triangulation exercise, and the additional exposure and virality checks.

