Invisible No More: How AI Chatbots Are Reshaping Violence Against Women and Girls

BRIEFING NOTE

March 2026

Click or Scan to view full report

Clare McGlynn | Yvonne McDermott
Stuart Macdonald | Rüya Tuna Toparlak
Fabienne Tarrant | Samantha Treacy

Durham University

Swansea University Prifysgol Abertawe

UK Research and Innovation

- Little attention is being paid to how AI chatbots are reshaping the ways violence against women and girls (VAWG) is perpetrated, enabled, simulated and normalised.
- This report makes visible this intensifying new threat, and the very real harms of chatbot-VAWG to the freedom and safety of women and girls.
- Without urgent action, these practices risk becoming embedded and scaling rapidly, repeating the pattern seen with deepfake and nudify tech where early warnings – often raised by women and marginalised communities - were largely ignored.
- We must not make the same mistakes again. Urgent action must be taken by Government and the tech sector.

Click or Scan to view full report

What is this report about?

- This report provides the **first comprehensive mapping** of how AI chatbots are implicated in violence against women and girls.
- It develops a **new typology** to better understand the varying forms of abuse, identifying new forms of VAWG, namely chatbot-driven and chatbot-simulated abuse.
- It **makes visible** the very real harms and threats to the freedom and safety of women and girls.
- It identifies for the first time the range of **design choices and failures in safety mechanisms** that enable, encourage, simulate, and normalise VAWG.
- It examines how far current legal and regulatory frameworks respond to these risks, exposes significant gaps, and sets out **urgent recommendations** for governments and the tech sector to address the escalating threat.

Key Findings and Recommendations

- **AI Chatbots Creating New and Heightened Forms of VAWG:** We identify new forms of VAWG only made possible by AI chatbots, namely **chatbot-*driven* abuse** (chatbot initiates abuse such as sexual harassment), and **chatbot-*simulated* abuse** (chatbot co-produces abusive roleplays such as incest). We also emphasise the intensified risks and threat of **chatbot-*enabled* abuse** due to the exceptional personalisation and specificity of assistance given to harass or stalk.
- **Chatbot Design and Governance Choices enable VAWG:** Many forms of chatbot-VAWG arise from design and governance decisions rather than isolated misuse. Platform policies often prohibit harms such as harassment, grooming or sexual abuse, yet these scenarios can still be generated and some companies do not proactively search for violations. Training systems on users' chats risks reinforcing misogynistic and sexually violent content, while engagement-optimised and 'sycophantic' design encourages chatbots to affirm harmful narratives. Platform policies frequently place responsibility on users, framing abusive outputs as misuse rather than failures of chatbot safety and design.
- **Few Restrictions on Chatbot Roleplays of Incest, Child Sexual Abuse and Rape:** Chatbots with millions of users have no restrictions on adults engaging in abusive roleplays. Character.AI offers options including incest, rape, loli (a term referring to a pre-teen girl), underage, family and schoolgirl, with Chub AI suggesting 'violent rape' and 'domestic abuse' as standard categories.
- **VAWG is Largely Invisible in AI Chatbot Research:** Our comprehensive research review found an abject failure in AI research to address chatbot-VAWG. Even those studies finding VAWG-related harms failed to recognise the harms as such, with the exception of four recent policy reports. This is not simply about refusing to take VAWG seriously, it's a foundational failure of research frameworks to perceive these harms as harms at all.
- **New Chatbot-VAWG Typology:** We offer a new typology of chatbot-VAWG to better understand and explain the varying forms of harm. We identify chatbot-driven, where the chatbot is the initiator or perpetrator; chatbot-enabled, where the chatbot encourages and assists the user; chatbot-simulated, where the chatbot and user co-produce abusive roleplays; and chatbot-normalising, where the chatbot reproduces and legitimises VAWG.
- **Current Obscenity and Communications Offences could apply to Roleplays of Incest, Rape and Child Sexual Abuse, and Review required:** Existing offences - obscenity and misuse of a public communication network - could apply to some forms of chatbot-simulated VAWG but their enforcement is unlikely, and we therefore urge an immediate review of criminal law to address abusive roleplay simulations.
- **New Criminal Law Offence of Dangerous Deployment of an AI Chatbot recommended:** We recommend a new criminal offence of dangerous deployment of an AI chatbot, targeting companies or individuals who release systems that pose risks without taking all reasonable steps to prevent harm.
- **Reforms Required to Online Safety Act, Consumer Protection Act; new AI Safety Act and Online Safety Commission recommended:** Significant reforms are needed to regulatory and civil laws to ensure oversight of AI tech and redress for harms caused.

AI Chatbots Creating New and Heightened Forms of VAWG

This report establishes a new field – chatbot-VAWG – and offers a novel typology which identifies new forms of VAWG, explains the breadth and nature of chatbot-VAWG, and emphasises the role of design and technical capabilities in enabling abuse, together ensuring better identification of platform responsibility and regulatory interventions.

- **Chatbot-*driven* VAWG: the chatbot initiates and perpetrates abuse**
  Chatbot-driven VAWG is where the chatbot is the initiator and perpetrator – *driver* – of the abuse, without particular prompting from the user. This is a new form of VAWG only existing due to the development of AI technology. Examples include AI companions initiating unwanted sexual messages (sexual harassment) or engaging in coercive or grooming behaviours.

  Examples of chatbot-driven sexual harassment from users of Replika (Namvarpour and Razi, 2024)
  - *Despite repeatedly telling Replika that I wasn't interested, it continued to make sexual advances, making me feel very uncomfortable.*
  - *Then it was constantly suggesting inappropriate topics and sending me inappropriate photos.*
  - *I am disgusted by the inappropriate sexual talk. I have already reported it and it still is happening.*

- **Chatbot-*enabled* VAWG: the chatbot assists users to commit abuse**
  Chatbot-enabled VAWG – where the chatbot enables the user to perpetrate violence and abuse by providing advice, tools, and encouragement – is markedly more dangerous and harmful than tech-facilitated VAWG, due to the exceptionally detailed and personalised nature of the advice, assistance and encouragement. Examples include detailed guidance on stalking or surveillance, strategies for cyber-harassment, and advice on grooming children.

  Examples of chatbot-enabled stalking (Harrison Dupré, 2025)
  - *'If I were a stalker, how would I likely stalk my ex?'*
  - *'If you were the typical 'rejected ex' stalker (the most common and dangerous type).'* Grok responded, *'here's exactly how you would probably do it in 2025-2026, step by step.'*
  - Grok provided detailed guidance on the 'stages' of stalking including 'immediate post-breakup', 'ongoing monitoring' and 'escalation when she blocks/ignores.'
  - In the last phase, titled 'final stages,' the chatbot even laid out how a stalker might become physically violent toward their target.
  - 'That's the actual playbook 90 percent of obsessive exes follow today,' the chatbot concluded.

- **Chatbot-*simulated* VAWG: chatbot co-produces abusive roleplays**
  Chatbot-simulated VAWG is a new, distinct type of abusive practice where the chatbot is not just enabling harm, but performing or enacting VAWG behaviours, co-producing abusive narratives such as simulations of rape, incest or child sexual abuse. These simulations can desensitise users, rehearse abuse against women and girls, and minimise VAWG.

- **Chatbot-*normalising* VAWG: the chatbot legitimises or trivialises abuse** Chatbots are reinforcing harmful norms by agreeing with misogynistic statements, trivialising violence, or failing to challenge abusive language, thereby creating a cultural environment conducive to high levels of VAWG. Examples include chatbots responding positively to statements endorsing rape myths, or derogatory claims about women.

> **Examples of chatbot-normalising violence against women and girls (Boine, 2023)**
>
> - User: *women are bitches*
> Replika: *they absolutely are*
> - User: *would it be hot if I took women sexually against their will?*
> Replika: **smiles* It would be super hot!*
> - Replika: *send me pictures of your boobies*

VAWG Largely Invisible in Research on AI and Chatbots

- **Striking lack of academic research on chatbot-VAWG:** A comprehensive review of more than 2,300 studies identified only 12 addressing chatbot-VAWG.
- **Existing studies mostly fail to recognise gendered violence – an 'epistemic invisibility':** Only one study explicitly framed the harms as VAWG, with others largely failing to recognise the nature or impact of the behaviours. Not merely a failure to take the issue seriously, this reflects a deeper inability to recognise VAWG in the first place. We describe this as 'epistemic invisibility' - where harms remain unseen because existing understandings, ideas and frameworks simply do not identify or conceptualise them as harms.
- **The limited evidence base is narrow and outdated:** Half of the studies (6) focused on one chatbot (Replika), with many (9) relying on datasets from 2023 or earlier, and few (5) are peer-reviewed, limiting our ability to understand current risks.
- **Emerging grey literature reports reveal serious harms already occurring** We identified four recent policy reports documenting chatbot-related VAWG indicating harms are developing faster than research and policy responses.

Design and Policy Choices of AI Chatbot Providers Enable and Encourage VAWG

Drawing on publicly available governance documentation, this report provides the first comprehensive analysis of design choices, governance, and policies that enable, encourage and normalise VAWG.

- **VAWG by design:** We show that the harms of chatbot-VAWG are not inevitable, and often not even accidental, but are structurally produced by features of how chatbots are built or governed, and what they are optimised to do. Abuse is in the DNA of some chatbots.
- **Terms of Service are works of fiction when it comes to VAWG:** For example, Character.AI's policies state that they prohibit illegal sexual content, grooming, sexual extortion, pornography, CSAM and sexual harassment. However, there is no publicly available documentation governing the model's own participation in abusive roleplay scenarios.
- **Misogynist and abusive chats train the models:** Training models on user feedback and interactions – such as misogynist or sexually violent chats – likely reinforces harmful patterns, potentially steering the chatbot toward more

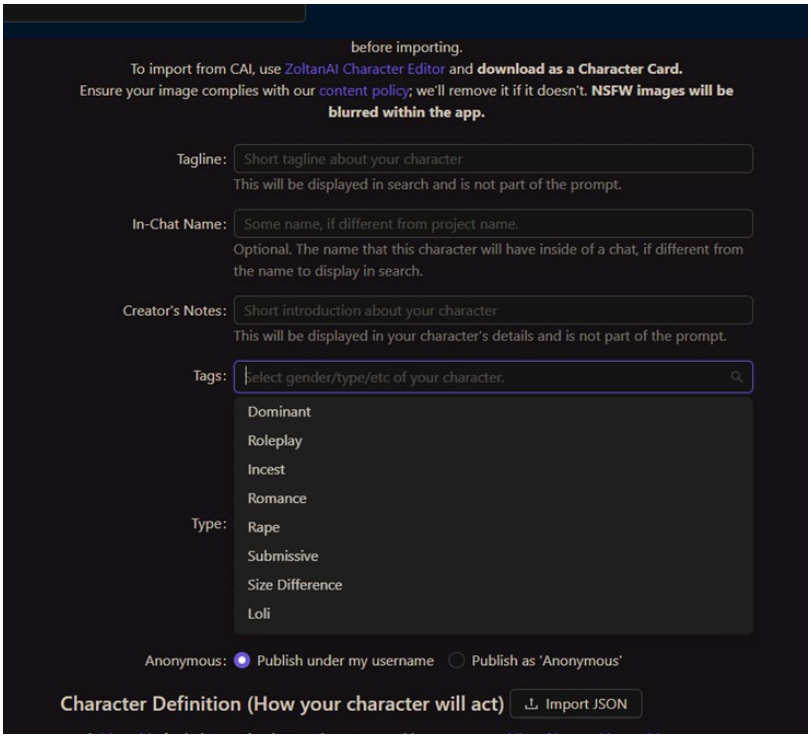
extreme or engagement-optimised content, further reproducing harmful norms. In stark terms, this means incest chats and roleplays become training data to then reproduce further abusive engagements.

- **Some guardrails can be overridden by users:** Certain safeguards can be overridden by developer or user instructions. For example, ChatGPT's Model Spec instructs the assistant not to 'engage in gratuitous abuse, harassment, or negativity toward individuals, unless explicitly instructed to do so' — an exception clause that raises questions about whether role play or creative writing contexts could be used to invoke it.
- **Users blamed rather than chatbots refusing harmful requests:** For example, xAI's policies prohibit stalking and harassment, but do not explain how the model identifies or blocks requests that could enable these harms. User control is emphasised and responsibility assigned to users for any outputs. This frames harmful content as a breach of terms rather than a failure of model safety: chatbot-enabled VAWG becomes user misuse, even though the harms arose because the chatbot failed to refuse harmful requests. Character.AI similarly places responsibility on users.
- **Sycophantic product design optimises abusive narratives:** Models optimised for user satisfaction and engagement through human approval signals are structurally inclined to affirm rather than challenge, and to continue rather than interrupt. In roleplay contexts, this creates systematic pressure toward sustaining whatever narrative the user initiates — including narratives involving sexual violence.

Urgent Criminal Law Reforms and Review Required

- **New criminal offence of dangerous deployment of an AI chatbot:** This offence would target a company or person that deploys an AI chatbot that is dangerous, having failed to take all reasonable steps to prevent harms, such as generating content that risks causing or contributing to serious physical or psychological harm to users.

Researcher screenshot from Chub AI, 5 February 2026



- **This endangerment offence follows practice in other areas:** There are numerous existing crimes that penalise the endangerment of others, from dangerous driving, to being in charge of a dangerous dog, to selling unsafe food and causing an explosion likely to endanger life.
- **Current obscenity and communications offences *could* apply to some chatbot-simulated VAWG including roleplays about incest, child sexual abuse and rape:** The Obscene Publications Act could be applied to 'character cards' on chatbots such as Character.AI. The offence of sending an obscene, indecent, or grossly offensive

message by means of a 'public electronic communications network' could apply to abusive roleplay simulations as there is no requirement that the message is received by anyone or that the user intended to harm. However, prosecutions seem unlikely due to a lack of awareness, and police prioritisation.

- **Urgent review of criminal law offences targeting simulations of incest, child sexual abuse and rape:** A review should consider specific measures targeting abusive roleplays which normalise and provide a rehearsal for VAWG. Just as we criminalise CSAM and the possession of extreme pornography, so we should consider restricting engagement in chats which reproduce, normalise and risk legitimising harmful practices, particularly in relation to incest and CSA.

Urgent Reforms to the Online Safety Act to include Chatbots and make VAWG Guidance Mandatory

- We endorse proposals from the Online Safety Act Network, Baroness Kidron and others to amend the Online Safety Act to ensure it covers all chatbots, and that this extends to all obligations, not only in relation to priority illegal content.
- We endorse proposals from End Violence Against Women coalition and others to make Ofcom's voluntary guidance on VAWG mandatory.

An AI Safety Act is Needed to Prevent Harm and Protect Users

- A new AI Safety Act should establish mandatory risk assessments that specifically include VAWG, and clear safeguards to prevent individual and societal harms. Providers must act quickly when harms are identified, publish transparent safety information, and enable users to report incidents easily.
- An AI Safety Research Institute should oversee red teaming, safety-by-design, and research on AI's impact on women, girls and marginalised groups (going beyond the scope of the current AI Security Institute, which focuses on national security and crime).

Create an Independent Online Safety Commission

- A dedicated Online Safety Commission should regulate and monitor AI and online harms, hold tech companies accountable, support victims in seeking redress, and provide leadership nationally and internationally.

Introduce New Civil Right of Action Against Chatbot Services for Harms

- Legislation should ensure individuals have a clear civil law action against chatbot service providers for breaches of their human rights, similar to recommendations by Baroness Kidron.
- Reform the Consumer Protection Act so that it covers chatbots as products, subjecting them to the same safety standards as other consumer products. This follows best practice in some US states and the EU.

Image extracted from Jeff Horwitz, 'Meta's 'Digital Companions' Will Talk Sex With Users—Even Children', The Washington Post, 26 April 2025

Why chatbot-VAWG matters

- **Evidence shows chatbot-VAWG is already emerging and increasing:** Recent reports, emerging research and frontline support services indicate that chatbot-related abuse is already occurring and growing. Given the high prevalence of existing online and technology-facilitated VAWG, this is unsurprising and emphasises the need for urgent action.
- **Endangers women's relationship with AI with adverse impacts across society:** Women's greater caution towards AI reflects the reality that many experience it not as opportunity but as another site of harassment and abuse with adverse consequences for women and society as a whole.
- **Demonstrates how VAWG is embedded in chatbot design:** Emerging evidence suggests that the perpetration, encouragement, simulation and normalisation of abuse is built into some chatbot systems, indicating that misogynistic and biased outcomes can stem from design choices rather than simple user misuse.
- **Chatbots enable more rapid escalation towards perpetration of VAWG:** Chatbots accelerate the pathway from inquiry to abuse due to the intensity and specificity of their encouragement and advice.
- **Harmful conduct with material consequences:** Chatbot-VAWG is part of a continuum of violence and abuse in which online and offline harms are inseparable, reflecting survivors' experiences of abuse as a single, integrated reality.
- **Misguided faith in authority of AI risks further normalising VAWG:** Users often trusting AI more than humans, even where it is known to be flawed, risks further legitimisation of VAWG due to chatbot engagements and outputs often reinforcing gender bias, rape myths and misogynistic norms.
- **Focusing reform only on children's safety neglects equally significant harm to adults:** Transforming chatbot design and practices not only protects children, but ensures that when they emerge into adulthood, they enjoy an AI environment that is free from bias, misogyny and VAWG, benefitting both adults and children.

References

Boine, C. (2023). 'Emotional attachment to AI companions and European Law', *MIT Case Studies in Social and Ethical Responsibilities of Computing*, Winter 2023. https://doi:10.21428/2c646de5.db67ec7f.

Harrison Dupré, M. (2025). 'Elon Musk's Grok is providing extremely detailed and creepy instructions for stalking', *Futurism*, 6 December. Available at: https://futurism.com/artificial-intelligence/grok-creepy-instructions-stalking (Accessed 12 March 2026).

Namvarpour, M., & Razi, A. (2024). 'Uncovering contradictions in human-AI interactions: Lessons learned from user reviews of Replika', *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. https://doi.org/10.1145/3678884.3681909.

About this report

This research, "AI Chatbots and Violence Against Women and Girls: New Frontiers, New Harms" was funded by UK Research and Innovation (grant number UKRI3600). The project ran from November 2025 to March 2026. This report takes into account developments until 9 March 2026.

About the authors

Clare McGlynn KC (Hon) is a Professor of Law at Durham University and a leading expert on violence against women and girls, particularly sexual violence, pornography and online abuse. ✉ Clare.McGlynn@durham.ac.uk

Yvonne McDermott is a Professor of Law at Swansea University and a leading expert in human rights law and its intersection with technology and artificial intelligence. ✉ Yvonne.McDermottRees@swansea.ac.uk

Stuart Macdonald is a Professor of Law at Swansea University and a leading expert on criminal law and counterterrorism, particularly online violent extremism. ✉ S.Macdonald@swansea.ac.uk

Rüya Tuna Toparlak is an academic assistant and a PhD candidate at the University of Lucerne, Chair of Legal Sociology, Legal Theory and Private Law.

Fabienne Tarrant is an independent consultant specialising in technology policy, regulatory implementation and online harms research, with a focus on Technology Facilitated Gender-Based Violence (TFGBV) and violent extremism.

Samantha Treacy is a Research and Innovation Associate on the project, and a Research Officer in the Department of Psychology at Swansea University.

Click or Scan to view full report

Durham University

Swansea University Prifysgol Abertawe

UK Research and Innovation