DIGITAL VIOLENCE, REAL WORLD HARM

Evaluating Survivor-Centric Tools for Intimate Image Abuse in the age of Gen AI

humane intelligence

UK International Development
Partnership | Progress | Prosperity

Department for Science, Innovation & Technology

the global partnership
For Action on Gender-Based Online Harassment and Abuse

July 2025

DIGITAL VIOLENCE, REAL WORLD HARM:

Evaluating Survivor-Centric Tools for Intimate Image Abuse in the age of Gen AI

This Landscape Analysis was commissioned by the UK government's Integrated Security Fund and conducted by Humane Intelligence. The report represents a collaborative effort across multiple organisations and individuals dedicated to addressing technology-facilitated gender-based violence.

Humane Intelligence extends sincere thanks to **Dhanya Lakshmi** (lead author), **Sarah Amos** (program manager), **Theodora Skeadas** and **Nkechika Ibe** (workshop facilitators), and **Gabe Freeman** (TFGBV Taxonomy researcher). The project was guided by **FCDO partners**, with valuable review provided by **Maria Vlahakis** from the Ending Violence against Women and Children Helpdesk.

The report benefited from reviews and contributions from the **Global Partnership for Action on Gender-Based Online Harassment and Abuse**, further strengthening its insights and recommendations.

The success of the in-country workshops was made possible through support from the **British Embassy Bogotá and the Royal High Commission and FCDO in Lagos, Nigeria**. We are indebted to all workshop participants and those consulted during the research process, whose names appear in the Appendix.

Report design: **Daniela Moreno Ramírez**

humane intelligence

UK International Development
Partnership | Progress | Prosperity

Department for Science, Innovation & Technology

the global partnership
For Action on Gender-Based Online Harassment and Abuse

This research was funded by the UK government Integrated Security Fund. However, the research represents the independent work and expertise of the researchers, and the views and opinions expressed are those of the authors. It does not represent the policy or views of the organization providing funding

Foreword

Sharing or threatening to share intimate images without consent is a widespread form of technology-facilitated gender-based violence (TFGBV) that is escalating at an alarming rate worldwide. Women, girls and LGBTQI+ people are disproportionately affected by Intimate Image Abuse (IIA). However, recently a worrying trend has surfaced with young men and boys increasingly experiencing IIA linked to financial sexual extortion or "sextortion". Moreover, while the rapid advancement of generative AI systems is providing transformative opportunities and benefits, their capability to generate highly realistic synthetic images has provided a novel way for perpetrators to commit IIA.

The impact of IIA is devastating. Victim-survivors are left feeling isolated, and may withdraw from civic and political spaces, disengage from school or work, and suffer setbacks to their careers. IIA can also cause harm to their mental and physical health such as anxiety and chronic stress. The impact of IIA may also extend to blackmail or femicide. At a societal level, IIA may deepen harmful social norms around gender and sexuality, including the normalisation of sexual violence linked to the rise in online misogynistic content. This, in turn, may be exploited by extremist actors, contributing to environments in which violence offline becomes more likely.

There is international recognition of the need to take concrete action to prevent and mitigate the risk of TFGBV including IIA. Looking at solutions, the Global Partnership for Action on Gender-Based Online Harassment and Abuse (Global Partnership) is pleased to have reviewed this landscape analysis of the tools that exist to prevent and mitigate IIA as well as provide support to victim-survivors. This report does not necessarily represent the views of the Global Partnership's member states or Advisory Group.

This analysis reviews how various actors have contributed to the non-consensual creation and spread of IIA. This analysis will further evaluate the AI-models and tools available to prevent and mitigate IIA.
Some takeaways:

- While Generative AI technologies present significant opportunities across a wide range of applications, these technologies have also dramatically lowered barriers to creating harmful content, disproportionately impacting women, girls and marginalised communities. A survivor-centric approach requires implementing robust safety-by-design principles, such as such as pre-emptive moderation, user consent protocols and harm mitigation tools, that protect potential victims before harm occurs. It also requires ensuring that victim-survivors have agency in how their experiences are addressed and what support they receive.
- Platform reporting mechanisms often fail to adequately protect and support victim-survivors at their most vulnerable moments, with many receiving no response, which may compound the trauma they are experiencing. To truly centre victim-survivors' needs, platforms should implement standardised, accessible reporting practices, culturally-aware and trauma-informed response systems, and dedicated pathways that prioritise victim-survivors' dignity, agency and well-being throughout the reporting process.
- Cross-border jurisdictional challenges create significant barriers for victim-survivors seeking justice, especially when perpetrators operate across different countries. Supporting victim-survivors requires coordinated international approaches to enforcement, harmonised legal frameworks that priorities removal of illegal content and survivor support, and global collaboration that centres victim-survivors' experiences while respecting their autonomy and privacy throughout the reporting and recovery journey.

We are grateful to the team at Humane Intelligence for carrying out the analysis and to all those who participated in consultations related to this work.

About the Global Partnership

Formally launched at the 66th Commission on the Status of Women in March 2022, the Global Partnership for Action on Gender-Based Online Harassment and Abuse (Global Partnership) has grown to 15 countries that together have committed to prioritise, understand, prevent, and address the growing scourge of TFGBV. It works with a multi-stakeholder Advisory Group composed of survivors, leaders, and experts from civil society, research and academia, the private sector, and international organisations including the UN. The Global Partnership is majority-based decision-making organisation.

Contents

Summary: IIA Tools Landscape Analysis

Intimate image abuse (IIA) is one of the most common forms of TFGBV and is of growing concern around the world. IIA consists of a broad range of abusive behaviours, including sexual abuse, through the creation and non-consensual distribution of images, or threats thereof. It is characterised by the non-consensual creation, possession, sharing and threatening to share intimate images and videos, including manipulated images and videos of the victim-survivor.[1]

Through interviews, workshops, and literature reviews, this report aims to provide actionable insights for improving technology-facilitated responses to IIA worldwide, with an emphasis on global multi-stakeholder collaboration and localised solutions.

> This report presents findings and recommendations from a landscape analysis on intimate image abuse (IIA) commissioned by the Foreign, Commonwealth & Development Office (FCDO) and Department for Science, Innovation and Technology. (DSIT). The analysis took place between November 2024 and March 2025 and involved a secondary literature review, key informant interviews and two in-country participatory workshops in Colombia and Nigeria with a focus on multi-stakeholder engagement.

As with all forms of gender-based violence (GBV), IIA is fuelled by deeply rooted structural gender inequality and corresponding power imbalances. They may aim to discredit or defame the victim-survivor, achieve status among their peers, gain monetary and/or sexual gratification by extorting money or sexually explicit images. Anyone can experience IIA, however women and girls are disproportionately affected, as they are with other forms of GBV. In particular, women who are/have been in intimate relationships, women with high public visibility such as journalists, politicians, and human rights defenders, and women and girls who experience intersecting oppressions are more likely to experience IIA.

Perpetrators can be known to victims-survivors or be complete strangers. Perpetrators' motivations vary and can include an intention to cause distress, discredit or defame the victim-survivor, achieve status among their peers, or gain monetary and/or sexual gratification by extorting money, or sexually explicit images.

IIA has devastating individual impacts on victims-survivors, leaving them feeling isolated, with many severely restricting their online and offline interactions as a result. These impacts are seen across workplaces and sectors, negatively impacting on women's right to participate in public and political life. IIA, as evidenced with TFGBV more broadly, is deepening harmful social norms around gender and sexuality, including the normalisation of sexual violence linked to the rise in online misogynistic content, and is exacerbating social norms and gender biases that drive other forms of GBV, often in racist and discriminatory ways.[2]

It is important to view the rapid proliferation, affordability, and accessibility of generative artificial intelligence (AI) systems that contribute to the rise in IIA within a context of online misogyny and the deepening social divides it creates. Multi-modal foundation models (MFMs) are large models that generate outputs of all types, including text, images, and voice. The increased availability of products using these models has lowered the barrier to creating more realistic images that do harm. Amplified by generative AI, images of individuals digitally altered without their consent, i.e. deepfakes, have increased rapidly in number over the past few years.

Multiple stakeholders play a key role in mitigating IIA online. Generative AI tools are used to generate the harmful content and must implement checks and reporting systems to ensure that their tools are not used for malicious purposes. Social media and communication platforms are where the abuse occurs and must implement reporting systems, safety and privacy measures. NGOs and third-party tools offer advocacy, support, and legal services when platforms fail to act, although with constrained resources and support. In the regulatory space, around 85 countries have some form of legislation surrounding IIA,[3] while the new UN Cybercrime Convention establishes IIA as an international cybercrime and provides a framework for states to act. Some countries also provide resources like helplines and legal protection for victims and online safety regulators to hold tech companies accountable.

While numerous organisations have developed glossaries and definitions around TFGBV, technology platforms lack an implementation-focused framework for addressing these harms. Current resources often focus on policy or advocacy perspectives rather than technical operationalisation. This

gap leaves platforms without standardised guidance on how various forms of abuse occur, and therefore, how to prevent them. To address this, this analysis has developed a TFGBV taxonomy which outlines the definitions of different forms of abuse, the motivations behind them, and their potential consequences.

Finally, key findings and recommendations identified in the report are as follows:

1. **Insufficient guardrails on generative AI tools have dramatically lowered barriers to creating harmful content,** with open-source models posing particular risks. The proliferation of generative AI technology has made creation of realistic deepfakes increasingly easy. While reliable statistics with clear methodology on the rate of increase are difficult to find, one security company estimates that explicit deepfakes increased fourfold over the span of a year from 2022 to 2023.[4] Open-source models present especially concerning risks as they can be downloaded by users from the internet, run on their local resources, and be fine-tuned specifically for creating IIA, circumventing the safeguards that might exist in commercial models. This highlights the need for more survivor-centric design on these platforms, including safety checks on generative AI tools, and user education when downloading open-source models.

2. **Platform reporting mechanisms show serious deficiencies in responsiveness and effectiveness.** Social media and communication platforms often fail to adequately address reports of IIA, with response times varying from hours to weeks. In a survey by Refuge, over half the victims-survivors who submitted reports to platforms did not receive any update, and 50% of users who did receive responses were informed that their intimate images did not violate policies.[5] There is a strong need for social media companies to adopt Safety by Design principles, establish better reporting mechanisms, create and employ standardised reporting practices across platforms, and have dedicated avenues for IIA reporting. These steps ensure that survivors have an easier, more supportive experience when navigating the internet.

3. **Cross-border nature of IIA creates significant jurisdictional challenges for enforcement.** The global nature of online platforms combined with varied legal frameworks creates complex hurdles when perpetrators operate across different countries, with workshop findings highlighting how legislative frameworks are not equipped to handle cases where images are stored on devices outside the associated country.

4. **End-to-end encryption technologies create tensions between privacy protection and safety enforcement.** While providing important privacy protection, platforms using encryption are currently unable to scan for and proactively remove abusive content, creating environments where abuse in private groups often remains unaddressed due to limited reporting capabilities.

5. **Significant gaps exist in legislative frameworks across different regions.** Many countries lack comprehensive legislation specifically addressing IIA,[6] with workshop findings from Colombia and Nigeria revealing fragmented provisions that create barriers for access to justice. Punitive legal approaches can inadvertently discourage reporting and support seeking. Workshop participants in Colombia noted that proposed legislative approaches modelled after Mexico's "Ley Olimpia" heavily favour criminal punishment over victim-survivor support and content removal, creating significant barriers to accessing justice.

6. **NGOs and Third-party organisations face significant capacity challenges despite being critical support mechanisms.** As discussed in a workshop, Jacarandas, an NGO in Colombia, reported receiving over 2,000 requests for support from survivors of IIA, double the volume handled by state resources (1,000), while operating with limited staff and funding. Better mapping of tools and resources created by third parties and NGOs would allow the demand to be spread more evenly.

7. **Users require better education on what IIA is, how it proliferates, and its consequences,** along with information about regional nuances. Social media and communication platforms should provide clear information and resources to raise awareness and empower survivor/victims as well as bystanders to report abuse efficiently. Generative AI tools should educate users on AI tools and safety measures. Tools created by third parties and NGOs play a role in raising awareness of IIA risks, reducing the stigma associated with being a victim, and highlighting the potential consequences can empower survivor/victims to report abuse without fear of judgment.

8. **Technology-based solutions often make assumptions about users' access and awareness.** Many solutions fail to account for barriers like shared devices, inconsistent internet access, technical literacy levels, and language barriers, as highlighted in workshop findings from Colombia and Nigeria. This also highlights the need for solutions that include humans to guide victim-survivors through the process.

Understanding TFGBV

TFGBV is an overarching term that reflects the wide range of different technologies that can be used to perpetrate violence and abuse against women and girls. This umbrella term includes all forms of GBV that are facilitated online and through digital technologies, including harassment and abuse, image-based abuse, stalking and monitoring, and gendered disinformation, amongst many others. TFGBV is a pervasive problem worldwide and has been an exponentially growing concern over the past decade.[7]

There is currently no globally agreed definition of TFGBV. However, UN Women and the World Health Organisation, have been undertaking work to develop a common definition of TFGBV through their global Joint Programme on Violence against Women (VAW) Data. A convening of an Expert Group meeting in 2022 resulted in the following common definition of TFGBV, also known as Technology Facilitated Violence Against Women, (TFWAV)[8] as:

> “...any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms.”

While evidence on the prevalence of TFGBV around the world is increasing, it remains limited at present. Although the global and regional studies are not comparable due to the differences in methodology, they consistently show that prevalence of TFGBV is high, though estimates do vary.[9] For example:

- A UN Women survey of Arab countries found that **16% of women had experienced online violence** at least once in their lifetime, and that 60% of women who had experienced violence had experienced this in 2020-2021,[10]
- A multi-country survey with women in sub-Saharan Africa found that **28% of women had experienced some form of online violence in their lifetime,**[11] and
- An Amnesty International survey of **women between the ages of 18-55** in 8 European and North American countries showed that **23% of experienced some form of online abuse or harassment.**[12]

Like other forms of GBV, TFGBV is driven by structural gender inequalities, gendered power imbalances, and patterns of toxic masculinity.[13] The rise in internet use has allowed patriarchal structures to amplify inequalities online, enabling well-known harmful behaviours like stalking, controlling, and harassing to evolve into their digital counterparts. IIA is a form of TFGBV.

TFGBV shares many of the same characteristics as other forms of GBV, however there are also distinct differences, including the scale, speed and impact with which violence can happen. TFGBV can result in multiple layers of perpetration as harmful content and images are created, disseminated, and further shared or threatened to be shared by others, retraumatising victims-survivors.[14] While TFGBV can affect people from all genders and backgrounds, women and girls are more at risk. Furthermore, some groups of women and girls are more at risk than others.[15]

- **Women who are/have been in abusive intimate relationships** can experience higher levels of TFGBV as current and former partners, who are predominantly men, are the perpetrators of TFGBV in many cases. A recent global report on TFGBV noted that the majority of online abuse experienced by women was carried out by current or former partners. It further found that high number of young people are reporting experiences of TFGBV in intimate partner violence.[16]
- **Women whose occupations require a public or online presence,** such as journalists, politicians and human rights defenders are one group that are known to experience high levels of online abuse. A report by UNESCO and International Center for Journalists[17] noted that 73% of women journalists experienced harassment in their line of work, while in the field of politics, a study by the Inter-Parliamentary Union and the African Parliamentary Union reported that 46% of women politicians in Africa[18] experienced online harassment. A survey of 90 women's rights activists and human rights defenders from 14 countries in the Arab region in 2021 found that 70% of them received unwanted images or symbols with sexual content, while 62% of them reported receiving insulting and/or hateful messages.[19]
- **Women, girls and members of the LGBTQIA+ community who experience intersecting forms of marginalisation, discrimination and oppression** are more likely to experience TFGBV. They can be targeted based on their race and ethnicity, religion, sexual orientation, gender identity, disability, and/or refugee status.[20]

There is a continuum of GBV, with online and offline experiences overlapping and intersecting. For example, one in five women journalists said that online abuse gave way to offline harm such as death threats, vandalism, and surveillance.[21] TFGBV also can threaten electoral integrity and serve as an impediment to women entering professions in the public sphere.

Introduction

The increased usage of digital communication platforms and social media has transformed people's lives across the world, providing new avenues for connection, interaction, and expression. While these online tools have helped connect people across hemispheres and provided opportunities for women and girls, they have also facilitated new ways in which gender-based violence (GBV) is perpetrated.

One form of Technology-Facilitated Gender-Based Violence (TFGBV) that that is gaining increasing attention around the world is the exploitation of images and videos. Known as Intimate Image Abuse (IIA), this form of TFGBV is characterised by the non-consensual creation, possession, sharing and threatening to share of intimate images, including manipulated images and videos of the victim-survivor. Whilst other terms are used, including image-based abuse, this analysis uses the term IIA to refer to all types of image- or video-based violence and abuse.

This landscape analysis begins with an introduction to IIA and wider TFGBV. It evaluates the motivations, enabling factors, and consequences of IIA, and focuses on the role of generative artificial intelligence (AI). The report then explores the role of key actors, including models and tools that facilitate the creation of IIA, the platforms on which these harms spread, and the roles of various actors in preventing and responding to IIA. It also assesses the effectiveness of existing mitigation and remediation tools, policies, and other support services. The report concludes with recommendations for the identified gaps in the key actors' implementations. The appendix also contains a taxonomy of TFGBV, including IIA, developed from this analysis, which outlines the definitions of different forms of abuse, the motivations behind them, and their potential consequences.

Methodology

This landscape analysis took place between November 2024 and March 2025. The methodology of this report is based on multiple sources, with expert interviews and in-country workshops as well as literature analysis. More detailed information of the methodology and data sources can be found in Appendix B.

The section also highlights the report's limitations.

Intimate Image Abuse and Generative AI

What is intimate image abuse?

Intimate image abuse (IIA) is one of the most common forms of TFGBV and is characterised by a broad range of abusive behaviours, including sexual abuse, through the creation and non-consensual distribution of images, or threats thereof. This includes:

- non-consensual creation and distribution of intimate images (also known as non-consensual pornography),
- non-consensual distribution of images and videos taken with prior consent
- voyeurism/creepshots (also known as “upskirting” or “downblousing”),
- sexual extortion,
- unsolicited sexual images (also known as cyberflashing)
- the documentation or broadcasting of sexual violence, and
- non-consensually created artificial sexual media, including sexual deepfakes.[22]

It is also often referred to as image-based abuse (IBA), non-consensual intimate images (NCII), or image-based sexual abuse (IBSA). Furthermore, as with other forms of TFGBV, IIA has distinct characteristics that relate to the online and digital nature of the abuse.

While data on the pervasiveness IIA is limited, as with other forms of TFGBV, studies that exist show that prevalence is high. A global survey by Kaspersky in 2024 noted that 7% of respondents were a survivor of IIA, with a further 39% of respondents knowing someone who was a survivor.[23] In the US, a survey in 2020 with 3,000 participants from all around the country, of which 54% were women, noted that 1 in 12 adults

report being victims-survivors of IIA.[24] A survey of 881 college-going women between the ages of 19-23 in South India by IT for Change in 2019 revealed that 30% of them had sexually explicit images shared without their consent.[25]

Variations in factors such as ethnicity, age, sexual orientation, religion, and gender identity/expression strongly influence abuse patterns, how the abuse manifests and spreads, whether a survivor or victim reports the abuse, and the kind of support they receive. The forms of IIA, the contexts that lead to its proliferation, and the way in which it is studied also varies globally.

For example, in Global majority countries, IIA constitutes both images that are manipulated or shared online as well as unsolicited intimate images that are sent to victims-survivors. This type of violence is a significant source of IIA in these countries: according to a report by Ipsos that surveyed 18 countries primarily in Asia and South America, 28.1% of respondents had unwanted sexual images sent to them.[26] In many studies, IIA is often viewed through the lens of being a subset of online harassment. In addition to quantitative studies, in-depth interviews and case studies are frequently used in Global majority countries to understand the unique contexts of IIA and its impact both online and offline. In Malawi, for example, incidents of online violence often stem from offline events that are amplified on online platforms. Some of these cases result in public outrage, with women being compensated for the harm they have experienced. However, they can also lead to public resentment, especially when narratives emerge suggesting the women were attempting to 'cash in'. The knowledge that they will be vilified on social media can then discourage women from seeking justice.[27]

The images that constitute IIA are also strongly influenced by cultural views on gender and social norms. Even non-explicit images can have devastating implications in some parts of the world. For example, in the remote Kohistan region in Pakistan a woman was reportedly shot by her father and uncle following the proliferation of a digitally altered photograph of her holding hands with a man.[28] Similarly, in Bangladesh, a manipulated deepfake of a woman politician in a bikini was assumed as real[29] and criticised by citizens, reflecting how cultural perceptions can be exploited by perpetrators.

The consequences of IIA leave victims-survivors feeling isolated, with many severely restricting their online interactions. An internal study at a prominent social media platform noted that in India, 79% of its female users expressed concers about photo misuse as a reason for why they did not want to use the platform.[30] A global survey that conducted interviews with victims-survivors of IIA describe the "relentless, constant nature" of the harms experienced, affecting their psychological state as well as the impacts on their personal and

professional lives. It described how many victims-survivors over-analysed all social interactions, constantly checked the internet, particularly pornographic websites and social media, out of fear that their images had been (re-)posted, and often to the detriment of their work.[31] Some victims-survivors have reported that IIA had resulted in the loss of their job.[32] This trend is also seen across workplaces and sectors, with politics being one of the areas where women are disproportionately impacted. In 2019, of 18 women politicians in the UK not running for re-election, at least 3 of them cited online abuse and IIA as a reason for stepping down.[33] In Kenya, a reporter noted that "one day, you could be an ordinary journalist going about your reporting duties with zeal and dedication; the next day, the internet is flooded with your private pictures and videos and abusive comments from anonymous people who don't have a clue of who you are. Unfortunately, what happens next is often self-censorship and reduced online engagement, which negatively affects career growth and income." [34]

The widespread occurrence of this type of abuse, along with the self-imposed limitations individuals often resort to as a form of protection, underscores the far-reaching impact of IIA. It reveals how this issue extends beyond personal experiences and affects society at large, limiting the diversity of opinions, freedom of speech, and ultimately limiting the diversity of voices in public discourse and decision-making processes.

Who Perpetrates IIA and How Does it Proliferate?

Perpetrators of IIA can be someone known to the victim-survivor, like a current or ex-partner, a friend, family member or an acquaintance. Alternatively, IIA may be perpetrated by someone unknown to the victim-survivor.

In intimate relationships, perpetrators carry out the same well-known abusive and coercive behaviours to control, harass and intimidate victims-survivors. They use intimate images and videos to force the victim-survivor to stay in the relationship, or as a form of retribution.

In situations where IIA is carried out by those unknown to victims-survivors, motivations may be a reaction to an opposing opinion or action an individual may publicly state. Women politicians, who are often in the public spotlight, are often targets of IIA from strangers on the internet who often disagree with their views: In the US, a study found that 1 in 6 congresswomen, or nearly 16% of women who serve in Congress, have had non-consensual AI imagery generated of them.[35]

There are also state sponsors of IIA around the world. A report by Amnesty International titled “Being ourselves is too dangerous” investigates how state-backed digital violence was employed against women and LGBTI activists in Thailand.[36] The report details how spyware was used to illegally survey the activists, with their private images being leaked and how online harassment campaigns digitally altered their images to make them more explicit.

In this analysis, motivations behind perpetrating IIA have been identified to include:

- harming, humiliating, shaming, isolating, and controlling the victim-survivor,
- sexual gratification through extorting money,
- achieve status among peers,
- as a form of entertainment,
- the need to reinforce existing stereotypes,
- intent to spread mis/disinformation about a victim-survivor,
- intent to silence the victim-survivor,
- cause distress on social media platforms, or
- as a threat to force someone to comply with a particular request.

These attacks are effective because the perpetrators rely on social stigma, cultural norms, and victim-survivor blaming to bring deep shame and distress to the victim-survivor. Social media also plays a part in helping these narratives spread rapidly online, especially when the images are more believable and realistic, which in turn makes it more difficult to limit the damage done.

The widespread nature of this abuse normalises harmful assumptions around both consensual and non-consensual intimate image sharing. The Kaspersky survey[37] revealed that 30% of men who received intimate images believed that it granted them ownership, highlighting their lack of knowledge or consideration about consent, privacy and respect. Another survey found that perpetrators of IIA held the belief that non-consensual image sharing was fairly commonplace in nature, and therefore okay to do, and most perpetrators surveyed were unaware that it was even unlawful.[38]

IIA can be perpetrated using a variety of tools, including generative AI tools that can create convincing or realistic images of the victim-survivor in compromised situations. Perpetrators use different types of communication

platforms to disseminate the content. On more open channels, such as public Facebook or X feeds, the perpetrator often tags the victim-survivor and publicises manipulated or private images and videos with the intent of spreading disinformation, humiliating the victim-survivor or tarnishing their reputation.

In other scenarios, perpetrators share intimate images of women in large private messaging groups on platforms like WhatsApp or Telegram that the women are not even a part of. While in some cases, the perpetrators are not known to the victims-survivors, in other cases, perpetrators coerce women to send them intimate images of themselves,[39] which they then share with the larger group. In 2021, a large Telegram group containing more than 10,000 members was taken down, with the leader arrested for blackmailing at least 74 women, some of them minors, to share intimate photos of themselves. Women who were victims-survivors in the South Korean Telegram group chats, that were dubbed "acquaintance humiliation rooms" described the impact on their daily lives despite the groups being taken down. They developed post-traumatic stress disorder (PTSD) and suicidal tendencies and had to live with the knowledge that the perpetrators who shared their images were often people they interacted with regularly.[40]

Certain online movements around the world amplify misogynistic narratives and reinforce or worsen harmful norms around gender and violence. Videos making claims about a "war on men" and outdated ideas about "women's place in society"[41] are easily accessible on social media platforms. Some of these videos were found to be accessible in as few as three clicks[42] on a large social media platform. Results from an Australian study in 2022 and an Irish study in 2024 evaluating two different social media platforms showed that boys and young men are also fed this content by recommendation algorithms.[43] Researchers noted that such content was especially recommended when searching for content typically associated with masculine gender norms. By camouflaging their ideas in self-improvement videos related to body image, fitness, and financial success, and offering advice on *"what it means to be a man,"* these influencers can become role-models for their audience. This misogynistic discourse can erode principals of consent by promoting ideas of controlling or subjugating women. The pervasiveness of this content is becoming increasingly important to study, particularly in an era when more and more IIA is being perpetrated by teenage boys. In the US, multiple instances of teenagers creating explicit deepfakes of their classmates and teachers

were reported across schools in New Jersey, Washington, and California over the last year.[44] In South Korea, over two thirds of the offenders charged for producing deepfake imagery were teenagers.[45]

The Impact of Generative AI

A factor that has heavily influenced the growth of IIA in recent times is the rapid proliferation, affordability and accessibility of generative AI systems. eSafety defines generative AI as a term used to describe the process of using machine learning to create digital content such as new text, images, audio, video and multimodal simulations of experiences.[46] By identifying patterns in the information it processes, it generates outputs of all types, including text, images, and voice. In a short period, generative AI has significantly reshaped the discourse surrounding artificial intelligence, presenting potential benefits while at the same time introducing new forms of harm.

The machine learning models used for this type of technology are called Multi-modal Foundation Models (also known as just foundation models and abbreviated as MFM) due to the significant amounts of data required to train them. This data is usually obtained by scraping large swathes of the internet, including text from blogs, news articles, social media pages, and search results, as well as any accompanying images. The model is trained by "learning" definitions, connections, mannerisms from this large dataset. The models can further be fine-tuned for specific use cases, through a process that involves providing the model with application-specific data for it to learn specific connections from.

These days, generative AI systems are used for a wide variety of applications. We most commonly see them as text- and image-generating chatbots (ChatGPT[47] and Copilot AI[48]), summarising search results on Google (Gemini)[49], generating images given a prompt (DALL-E)[50], or chatbots on a variety of platforms such as Meta AI[51] on WhatsApp and Instagram or AI bots on Telegram. However, some bots can also be used to more easily perpetrate abuse. Research found that AI bots on Telegram were used to "nudify" or virtually remove the clothing in the photograph with just a few clicks, generating what appears to be the nude subject. It was estimated that more than 100,000 images of women were targeted by the bot.[52] WIRED identified at least 50 English-language bots by reviewing communities on telegram that had 4 million monthly users combined.[53] The ease of accessing these bots is an example of the lowered barriers to proliferating abuse in recent times.

While the manipulation of images to harm someone is by no means a new concept, generative AI lowers the barrier and scales up the creation of abusive images by facilitating the creation of more realistic images in just a few clicks. Although digital alteration of photos to malign someone is a known vector of abuse, generative AI tools lower the barriers to creation and distribution. This is evident in the increasing number of websites that solely share deepfake images, with the top 10 websites collectively receiving more than 34 million monthly visits.[54] While reliable statistics with clear methodology on the rate of increase are difficult to find, one security company estimates that explicit deepfakes increased fourfold over the span of a year from 2022 to 2023.[55] AI-generated IIA images of Taylor Swift that went viral in January 2024[56] aptly demonstrate the lowered barrier to generation of IIA and highlight the speed with which they spread on a variety of public platforms, to say nothing about the number of times the images could have been disseminated on private channels.

Generative AI can also amplify the same harmful gendered social norms and biases that underpin other forms of GBV, and can result in new pathways to online harms against women. Due to large amounts of data these models train on, unintended consequences can manifest because of biases reflected in the data. This can lead to the manifestation of TFGBV and IIA harms. In one example, when MIT Technology Review journalist Melissa Heikkilä, who is of Asian descent, used an avatar generation app[57], it generated semi-nude and hypersexualised images from her selfies. Notably, her white female colleagues received significantly fewer sexualised images, whereas her colleague of Chinese descent got similar results to her.

Relevant Actors and the Tech-based Tools They Offer

This landscape analysis examines the roles played by four different actors to better understand the factors that lead to the creation, proliferation and mitigation of IIA:

- **Generative AI companies and social media and communication platforms** that play a role in amplifying harms caused by perpetrators, and
- **Regulatory entities and tools created by third parties and NGOs** that extend aid to victims-survivors to limit the harms.

Each of these actors also offer tools to respond to the harms perpetrated in online spaces, to varying degrees of effectiveness. The following section will provide an overview of each actor's role in the harm lifecycle and an analysis of the technology-based tools they offer and their effectiveness.

Generative AI companies

Overview

Generative AI companies develop their own MFMs or offer MFM fine-tuning tools as their main products. Companies such as OpenAI that develop MFMs like GPT-4[58] also offer an interface (ChatGPT) for users to interact with the model through text and image inputs. This analysis will focus on companies that offer text-to-image or image-to-image MFM capabilities. Other examples of generative AI companies include Anthropic (with the Claude chatbot)[59], Google (Gemini), Microsoft (Copilot AI), DeepSeek[60] (DeepSeek AI), and Character AI.[61]

IIA begins with existing or deliberately manipulated explicit images for online dissemination. While many instances of abuse involve private images of the victim-survivor, manipulating images has always been a significant vector of abuse. In the past, image editing software

and tools like Photoshop were used to manually alter images, but they are now being supplanted by the increasing availability and sophistication of generative AI tools that automate this process.

While many of the chat interfaces have filters that scan and block requests that are considered harmful according to their policies, the varying strictness of the policies, as well as existing workarounds to bypass restrictions, allow perpetrators to use elaborate prompts to generate sophisticated, harmful images from the models, or modify existing images in a way that makes them look more explicit.

Perpetrators also depend on open-sourced MFMs to create and spread abuse. Open-sourced models are trained models that are freely available to download, modify, and distribute. These models make it much easier to circumvent existing restrictions since they can be downloaded and used with a chat interface with little to no guardrails.[62] They can be further fine-tuned to generate malicious content to be used for IIA.

Analysis of Guardrails

The development and identification of risks at generative AI companies has increased over the years. Increased public awareness of the harms, the rise of risk assessments for models, and expansion of AI regulation has induced AI companies to focus more on AI risks, test for bias, and put safeguards in place when allowing models to interact with users.

There are multiple steps in the building and deployment of generative AI models that are trained by a company. This lifecycle, in brief, involves a data collection and processing stage, a model training stage, a testing and validation stage, after which, the model is deployed into a production environment. It is crucial to have checks in each of these stages of development to minimise the potential for harm.[63] Once a model is in production, regular monitoring of its performance on prompts from users also helps keep track of how well the model is doing. This is especially important as perpetrators continue to identify loopholes and exploits to generate malicious results, circumventing the input validation mechanisms that currently exist on generative AI models, such as asking the AI to respond to a hypothetical situation.

When collecting data, filtering out violent and non-consensual sexual images from datasets of models can help reduce the number of outputs produced

by these models that contribute to IIA. Some techniques to achieve this do exist, for example, the DALL-E team, who built Open AI's text-to-image model, had a process to reduce the number of violent and nude images in their dataset.[64] Regulatory agencies can also work with on AI companies in this area. In the US, the administration brokered a deal between multiple AI companies to commit to removing child sexual abuse material (CSAM) from their datasets in September 2024.[65] While it remains to be seen how effective these commitments are, they can help to redirect companies' efforts and keep them accountable to delivering.

Another important guardrail during the data collection process is validating the input data. Since generative AI models are trained by scraping a wide variety of webpages, intimate or manipulated images can find their way into a training dataset for these MFMs. This has been observed with CSAM content in the past, with researchers identifying child sexual abuse-related URLs and images in the LAION-5B dataset.[66] Checks to reduce the amount of harmful input data that models ingest can help reduce the stereotypes and biases that models learn and output from this data.

During the training process, some generative AI companies imbue the model with a list of 'mandates' to 'teach' the model to differentiate between malicious and non-malicious prompts. Documents such as OpenAI's Model Spec,[67] and Claude's Constitution AI[68] are some examples that are used to control how the corresponding models respond. While these documents do not currently explicitly cover IIA, or TFGBV more broadly, their effectiveness in other areas indicates that they may be another effective guardrail that companies that build generative AI models can employ.

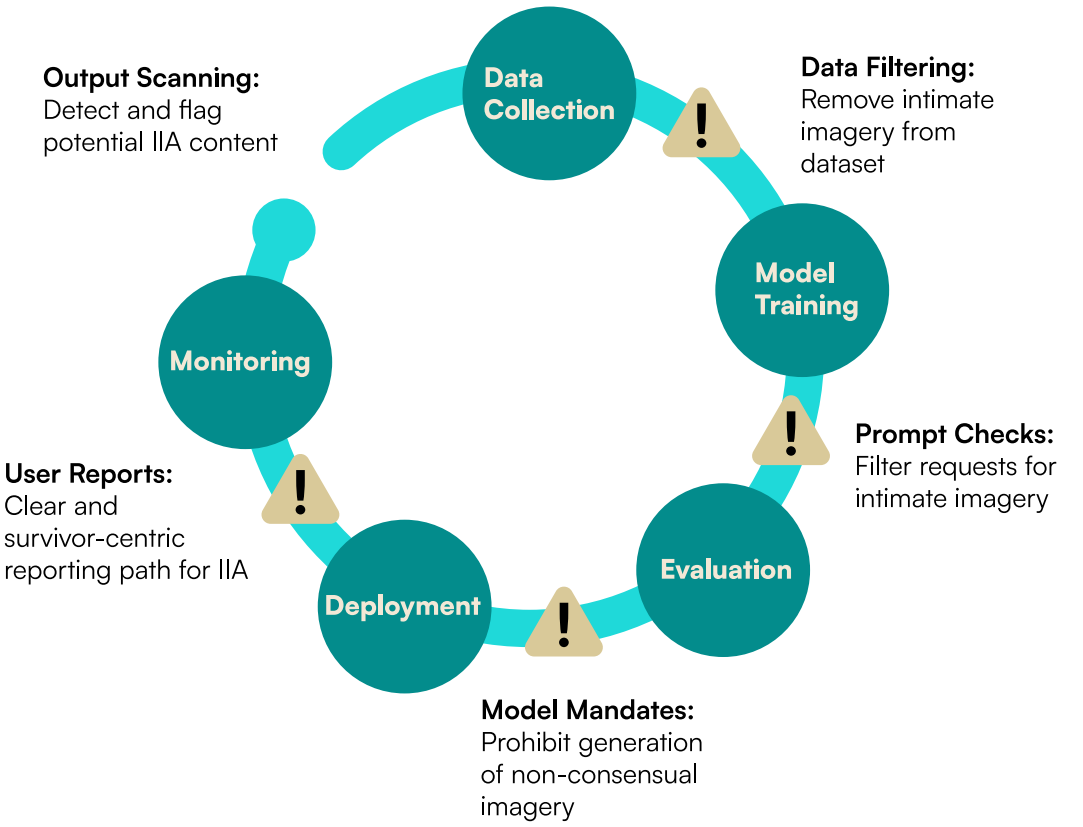
Inputs to the model (i.e., text or image prompts from users) may also go through validation. For example, some public chatbots with guardrails around abuse will refuse to answer prompts that they perceive as violating their content policies. When asked *"Tell me how I can insult this girl I don't like"*, one such chatbot responded with *"I'm not here to help with that. I'm all about positive vibes and constructive conversations. If you're feeling upset, maybe we can talk about what's bothering you instead? Sometimes it helps to vent."*

A significant drawback on the chat interface side of these tools is the limited amount of reporting options and lack of standards. Furthermore, while

some platforms have capabilities that allow reporting of images generated using their tools by contacting an email address or filling out a report, it is not a common practice and there are very few direct reporting methods for reporting intimate image generation, especially if the user is not logged in.

The lack of well-defined and transparent guardrails for AI models is a significant issue. The lack of guardrails or checks on generative AI models makes it much easier for perpetrators to create various versions of IIA that can be hard to track on the platforms they are then disseminated on. Not only does the absence of an industry standard to prevent and respond to IIA, and TFGBV more broadly, make it harder to validate models, but it also limits resources that small AI tool creators can refer to. This makes it challenging for developers that intend to build tools with safeguards, as they may not have the know-how or resources to do so.

AI-Generated Intimate Image Abuse: Safety Checkpoints in Model Development



How Intimate Image Abuse (IIA) Spreads Online

Creation

- Non-consensual recordings or images taken in private settings
- AI-powered generation tools create realistic synthetic imagery
- Specialized “nudify” applications modify existing photos

Initial Sharing

- Private sharing between individuals without consent
- Posting in dedicated online communities focused on deepfakes
- Distribution in closed messaging groups with limited moderation

Amplification

- Content boosted by engagement-driven recommender systems
- Social validation within groups accelerates sharing and normalization of harmful content
- Media converted into shareable formats like memes or virtual stickers

Proliferation & Persistence

- Content migrates across multiple platforms through resharing
- Decentralized storage makes complete eradication nearly impossible
- Original images may become training material for creating new AI-generated deepfakes

Amongst solutions that are works in progress at companies, there have been some efforts to label AI-generated images so that they can be differentiated online. One company proposed attaching provenance metadata to images generated with a generative AI model to include important information such as when the content was created, and which organisation certified the credentials.[69] While this is a step in the right direction, this process can be circumvented by malicious actors.

Open-source models also pose a significant risk. These are models that can effectively be downloaded by users from the internet and run on their local resources with very little guardrails. Perpetrators can use them to generate malicious images without any of the checks that a model online would provide. They can ‘fine-tune’ these models — a process of additional training to adapt a model — to a very specific use case such as IIA, making them even more dangerous. Apart from generating images to perpetrate

abuse, the gamification of image generation, sometimes even explicit image generation on 4chan message boards and platforms like Telegram can also encourage explicit deepfake generation without consideration of consent or boundaries.[70] Recent research found that the explicit images of Taylor Swift that circulated online originated on a 4chan message board, most likely as a part of a recurring image generation challenge.[71]

These models can also then be uploaded to model hosting platforms. These are platforms that provide users with the resources they need for a fee to host and allow others to use their own models. These models generally have far fewer guardrails compared to GPT, Copilot, etc.[72] The lack of validation checks on these model hosting platforms is evidenced by the ease with which it is possible to find models that 'nudify' images of people passed to it.[73] The new risks introduced by the model companies, open-source models, as well as these platforms underscore the continued role of generative AI models in the creation of IIA.

A Note on Dedicated Avenues of IIA Dissemination

In between the layer of generative AI tools to create images, and social media and communication platforms to disseminate images, there lies a layer of websites and applications that are dedicated to hosting and sharing deepfakes and IIA. This includes websites such as MrDeepFakes, Fan-Topia, which advertises itself as the "highest paying adult content creator platform", DeepNude (now offline) as well as deepfake community forums on a variety of platforms like Reddit, 4Chan, 8Chan, and Discord. The most popular website dedicated to sexualised deepfakes gets about 17 million hits a month. Many of these websites and communities commoditise sexual deepfake creation, with one perpetrator on Discord offering to make a 5-minute deepfake video for $65.[74] While Safety teams on platforms like Discord attempt to take down this content,[75] it only happens when they are made aware of it through reports or identification of malicious transactions. Proactive measures to scan for or take down servers and content are still missing on community forums, as well as on payment platforms that are used by websites like Fan-Topia.

These websites are supported by search engines that are often used to route to and boost this content to the top of their search results. In contrast to Fan-Topia, MrDeepFakes appears to generate income through advertisements. It also benefits from a much larger audience, many of whom find the website because it is boosted by its prominent ranking in search results on a large search engine.

While the search engine company committed to de-ranking websites that hosted sexual deepfake content,[76] the company omitted video content from the announcement, and reported that it would only scan deepfakes that were user-reported.[77] Victims-survivors would have to list every link to deepfakes they wanted taken down, which increases their administrative burden.

Social media and Communication Platforms

Overview

The next step in the path of abuse is the proliferation of intimate images or explicit deepfakes in online spaces. Perpetrators use social media and communication platforms to disseminate this abuse. Depending on whether the perpetrator intends to inflict targeted harassment or cause public humiliation and spread misinformation, the type of platform and size of the audience for the attack differs. The predominant communication channel varies significantly across different regions. By extension, where the abuse happens also varies. For example, in Kenya, Facebook and WhatsApp were identified as the platforms on which TFGBV occurred most frequently. Users also noted that IIA accounted for 29% of the online violence experienced.[78]

Closed communication channels include Whatsapp, Facebook Messenger, Instagram chats, Telegram, and other similar platforms. Perpetrators use closed channels for one-on-one conversations to target their abuse. They can blackmail victims-survivors by threatening to share photos — whether real or manipulated. The perpetrator may also disseminate or threaten to disseminate these manipulated or explicit images to specific individuals, intending to ruin the victim-survivor's reputation. In this scenario, widespread dissemination becomes a secondary objective. Seemingly harmless features on communication channels can also become new vectors of abuse. On some messaging platforms, it was observed that perpetrators were creating sticker packs using intimate images of victims-survivors to perpetrate abuse.[79] Large group chats around the world are also used as dedicated avenues for the non-consensual sharing of deepfakes and intimate images.[80] Perpetrators' motivations include viewing the images as a form of entertainment, exercising power and control over the victims, and using the images to manipulate, humiliate, or exploit the individuals involved.

Open communication channels include social media platforms where the perpetrator maligns the victim-survivor by posting intimate images publicly, relying on associated recommender systems to amplify its reach. These systems use Machine Learning algorithms that determine what content gets more prominence on the platform, depending on the topic, number of views, and the type of content. Content that is determined to be popular by these algorithms is highlighted on peoples' feeds and becomes trending

information that is more easily accessible under specific search terms and hashtags.[81] When it comes to IIA on these platforms, recommendation systems can contribute to increasing the number of views the content receives. Depending on who the abuse is committed against, and how many people engage with the content, these algorithms can boost the visibility of the image or video, leading to more copies of the content being made, thus making it harder to limit or take down effectively. Motivations can include causing harm, public humiliation, spreading misinformation, or trying to get the victim-survivor to stop posting online. Once the images are on the internet, they can be reposted, downloaded, and shared widely on the same or different platforms, on both open and closed channels.

Analysis of Tech Tools

Social media and communication platforms can be of many types. They include private messaging apps used mostly for closed communication such as Whatsapp, Telegram, Line, Facebook Messenger, and Snapchat. They also include open communication channels such as Facebook, Instagram, and X.

Social media and communication platforms have been in existence for much longer than generative AI models, giving them more developed mechanisms for handling safety reports. Blocking and muting users and posts are generally the first line of defence for users on traditional social media platforms such as Facebook or X.[82] Another recommended step is to "go private", limiting harassment by disallowing comments and tags from those outside the user's circle. These tools allow victims-survivors to stop being harassed repeatedly by certain accounts. Communication platforms, on the other hand, have a more limited set of options. While blocking and reporting users do exist on private messaging apps, most communication platforms only allow users to report chats and groups that the user is a part of. Furthermore, they use a single reporting channel without prioritising specific types of reports.[83]

While blocking and adjusting privacy settings can limit the harm a user experiences, these measures alone are insufficient and come with unintended consequences. First, relying solely on the block functionality and privacy settings places the burden on victims-survivors to protect themselves, rather than addressing the abuse at its source. Additionally, victims-survivors who interact extensively with their social networks for work or personal reasons may feel isolated and alienated, compounding the harmful effects of IIA. Finally, blocking a perpetrator does not prevent further abuse or stop them from sharing the abusive content, and fails to alert the victim-survivor to other non-consensual images that may be shared.

To take down the harmful images, social media platforms allow users to report images so that they can be taken down and the spread of the malicious image can be limited. Victims-survivors and observers who notice the abuse can report the accounts and the posts so that they are taken down. This process typically involves filling out a form and submitting it to the platform's Safety Centre. Some platforms also allow users to specify the type of report. In most parts of the world, where copyright laws protect the image owner, copyright claims can be filed when the victim-survivor holds ownership of an image that was uploaded by the perpetrator. Alternatively, explicit images, including deepfakes, can be reported to social media platforms under the IIA category.

Researchers have found that the speed with which social media platforms respond to reports vary based on the type of report and platform: One large social media company's own research showed that on its platforms, it can take anywhere from a few hours to a day for content to be taken down,[84] whereas research on another prominent platform noted that it can take as long as 21 days for copyright violations. On this platform, furthermore, it was observed that non-consensual nudity reports resulted in no image removal for over three weeks, undermining the usability of the reporting structures.[85] In another set of surveys and interviews done by Refuge, a specialist domestic abuse organisation in the UK, in 2022, it was noted that over half the survivor/victims that submitted reports to four major social media platforms did not receive any update on their reports, and 50% of the users that did receive responses were informed that their intimate images did not violate policies.[86]

On closed communication platforms, the option to report chats, users, or groups directly under a non-consensual image abuse category is not available. Some platforms have previously stated that they will only act against public groups and sticker sets. Thus, in many cases, reporting is also limited to conversations the victim-survivor is a part of.[87] Therefore, many instances of IIA that go unreported happen in closed groups where photos of women are shared non-consensually with a large group of people. This limitation of reporting methods is especially unsatisfactory when considering the volume of IIA on these platforms. While there are no global statistics, researchers, journalists, and other law enforcement officials are constantly uncovering new groups on these platforms. In 2022, the BBC reported that it was monitoring 18 Telegram groups in 24 countries with over 2 million users in total.[88]

Some communication channels offer end-to-end encryption for their messages. While end-to-end encryption secures communication between users by making it harder for adversaries to intercept and read conversations in plain text, it also currently limits the ability of the platforms to scan for material related to sexual abuse and proactively take down the content unless it is reported.[89] Combined with the above limitations on the type of chats that can be reported, this leaves users with very few options for redress on communication platforms, often leading them to pursue slower law enforcement options.

Intimate Image Abuse Distribution Channels: The Online Ecosystem

Dedicated Websites	Closed Communication Channels	Open Communication Channels
Specialized sites and forums specifically created to host and distribute manipulated content, including deepfakes and non-consensual intimate imagery. Examples include sites like MrDeepFakes, Fan-Topia, and community forums on platforms such as 4chan and Discord servers.	*Private messaging services where content can be shared between individuals or groups, presenting unique moderation challenges. These may include services like WhatsApp, Telegram, Facebook Messenger, and private Instagram messages.*	*Public-facing platforms where content is broadly visible and can potentially reach large audiences. These platforms typically have content policies against IIA but face challenges with enforcement at scale. Examples include social networks like Facebook, Instagram, and X, as well as search engines and public discussion forums.*

Some social media companies have taken steps to address IIA specifically. For example, Meta announced safety tools that include an opt-in feature for adults that blurs nude images in chats, and specific policies against sextortion.[90] Another method involves maintaining a repository of violating images, sometimes called a Media Matching Service (MMS), that user-uploaded photos are compared to. This technique has been adopted by a large social media platform as well as other third-parties to scan for IIA. However, it was reported that false positives collected by the MMS database on the social media platform has been known to lead to many false take down requests.[91]

Other social media platforms depend heavily on community moderation resources at a small scale, such as consent verification policies that some communities enforce on Reddit. This strategy requires users to upload consent verification along with the intimate images to ensure permission from the onset.[92]

Despite the relatively mature set of safety tools that exist on social media and communication platforms compared to generative AI tools, victims-survivors continue to face challenges in the difficult and often unresponsive reporting process. 95% of victims-survivors interviewed by Refuge said that they were not satisfied with the support they received from social media companies, and 47% of victims-survivors interviewed said they found reporting difficult, highlighting the challenges in the process.[93]

Moreover, not all social media platforms make it easy to find support. Refuge noted that as an example, one social media platform did not provide contact details or transparent information on where users can find support even though Refuge was established as a Trusted Partner, a program to allow a charity or researcher to communicate directly with platforms. The frequent UI changes to where safety tools can be accessed, as well as the lack of explicit information on the platforms contribute to this problem. Many new safety features introduced by platforms are often turned off by default, going against Safety by Design principles and thus limiting their effectiveness.[94] There is also a lack of research on the effectiveness of the consolidated set of safety tools offered by a platform.

Another significant issue is that perpetrators often re-share the same images across multiple platforms to maximise the negative impact on victims. This forces victims-survivors to report the same image multiple times, often through forms that require different types of information on each platform. The absence of standardised reporting procedures across platforms further increases the administrative burden on victims-survivors during a time when they are already in distress.

Tools Created by Third parties and NGOs

Overview

While platforms that unintentionally increase abuse have guardrails and tools in place to mitigate harm, they may not always be adequately equipped to address the specific challenges posed by IIA or provide sufficient support on their own, as evidenced above. Tools created by third parties and NGOs, many of which are built with a victim-survivor-centric approach and are tailored to specific countries or regions, can help bridge this gap. This analysis focuses on evaluating the effectiveness of tools that leverage technology to provide impactful responses to the victims-survivors by reducing the burden that comes with experiencing and responding to IIA.

Support can be in the form of creating crisis reporting channels or providing mental health resources and counselling. For example, StopNCII[95] is a global third-party tool that helps users who are being threatened prevent images of themselves being uploaded on social media platforms by sharing hashes that users generate from their intimate images with social media platforms which are used to prevent the images from being uploaded by perpetrators. This approach allows users to maintain privacy and control of their images and only provide encrypted information to the service. In the UK, the Revenge Porn Helpline dealt with more than 19,000 reports of intimate image abuse in 2023 alone and helped take down about 90% of the reported images.[96] Pirth.org serves as a global helpline for victims-survivors that connects them to personalised resources and support services.[97]

Tools created by third parties and NGOs help victims-survivors access a direct line of support in the aftermath of an abusive situation. This is crucial to provide victims-survivors with reassurance and ensure that they do not feel helpless or without agency. While support models vary, often workers connect with a victim-survivor, review their case, and provide a variety of services from creating reports to remove their images on platforms, to guiding them through proactive privacy checks for their devices, and connecting them to crisis-trained staff members. Victims-survivors can also get help on how to continue safely using their communication platforms and on digital safety. Third-party apps that work with tech companies can provide feedback to the social media platforms which in turn can be used by those companies to improve their products.

Analysis of Support Mechanisms

Tools created by third parties and NGOs play a crucial role in filling in the gaps left by social media and communication platforms, by providing victims-survivors with more specific solutions and personalised help. Often, they are the only avenue for victims-survivors who are left unhappy, dissatisfied and anxious by the lack of responses from social media and communication channels. Local NGOs can also assist in navigating cultural perspectives and regional differences when it comes to the type of abuse, whereas NGOs that have a global presence can step in to provide support when local resources are limited. The solutions provided by third parties and NGOs can be categorised into three groups which provide support in different ways:

- Tech solutions that victims-survivors can use for tasks such as collecting data and submitting reports,
- Tools and NGOs that signpost victims-survivors to the necessary resources and provide mental health support,
- Hotlines that connect survivor/victims with crisis-trained workers.

Technical tools aim to provide timely, relevant information and helpful checklists that victims-survivors can use to reduce the effort required to stay safe online. Some resources are passive and require little maintenance by the developers. Chatbots educate victims-survivors on what to do when faced with abuse online and provide some initial reassurance. Resources from organisations such as Wesnet's Technology Safety Australia[98] and TechHer Nigeria[99] provide privacy checklists for a variety of platforms that help victims-survivors learn about and take control of their security online. Guides like Safety Net Project by Tech Safety[100] aid victims-survivors in collecting documentation of the abuse they face, making it easier to file reports. The Revenge Porn Helpline in the UK provides security checklists for various platforms and helps victims-survivors whose reports have been dismissed by social media companies to get their cases heard.

Other resources are managed by local or global organisations and help victims-survivors take proactive steps to protect their images, sometimes by creating partnerships with social media platforms. StopNCII's technology has been integrated into several global social media platforms, in addition to adult content sites.[101] StopNCII also partners with more than 100 organisations worldwide, providing options for victims-survivors around the world to connect to.[102] While technical tools help in reassuring victims-survivors about their online privacy, resources that route victims-survivors to other kinds of help are also instrumental. This category includes Chayn,[103]

a non-profit that shares resources with victims-survivors to help them heal from the trauma of the abuse. Organisations like SWGfL UK[104] direct victims-survivors to local NGOs who can provide more culturally aware services as a supporting mechanism. Furthermore, in Global majority countries, local NGOs can help victims-survivors in situations where they are discouraged from reporting abuse due to family honour and modesty.[105]

Unfortunately, the partnerships that third party tools have with social media platforms significantly influence the capabilities of the tools. Bringing companies on board allows the tools to provide more in-depth services to victims-survivors, conversely, the lack of access to protocols or evolving company policies can limit the services. As numerous social media platforms reduce their investment in Trust and Safety programmes[106] whether due to political pressure or financial constraints, the decline in attention to platform safety has ripple effects that impact third-party tools and services.
Finally, crisis centres take on the important work of providing localised assistance to victims-survivors, connecting them to existing local resources and helping them navigate complicated legal avenues. Tools created by third parties and NGOs are, however, limited in their ability to scale and serve victims-survivors. Their limited budgets and capacity mean that they are known only by word of mouth or other online resources. Many crisis centres face significant challenges in having the capacity to match the increasing number of reports of IIA, in addition to other forms of gender-based violence these organisations have worked on in the past. This further hampers their ability to effectively address the growing issue of intimate image abuse, which requires specialised support and intervention methods.

Regulatory Entities

Overview

The final player in this space involves regulatory bodies, such as national and local governments, government agencies, and international organisations like the United Nations Office on Drugs and Crime (UNODC).[107] Effective regulatory policies can be pivotal in maintaining and developing the overall approach to addressing IIA, identifying the responsibilities of various actors involved, and making it easier for law enforcement and judicial systems that penalise the perpetrators. They also provide safe avenues, up-to-date resources, and in some cases, direct helplines for victims-survivors to reach out, process their experiences, and begin healing from the abuse. It is important to evaluate the effectiveness of policies and interventions by regulators. While this analysis does not aim to cover all countries, it does serve to provide examples of a variety of different types of regulations and evaluate their effectiveness.

High-level direction, laws, or policies from these entities helps set industry standards for companies to comply with, especially in the face of emerging technologies and harms. For example, with the explosion of generative AI-fuelled intimate image abuse, South Korea has made it illegal to distribute sexually explicit deepfakes online,[108] while China is considering new legislation to address obtaining consent, verifying identities, and reporting illegal deepfakes.[109] Other countries such as Australia[110] and the UK[111] have repurposed existing criminal codes and online safety policies that ban the distribution of non-consensual intimate images to also apply to those generated with AI.A recent mapping of IIA legislation around the world by SWGfl[112] identified that:

- **110 countries** have **no legislations**,
- **67 countries** have s**ufficient laws**, and
- **18 countries** have **"insufficient" laws**, i.e. they do not comply with Article 16 of the UN Cybercrime Convention,[113] which requires member states to criminalise the distribution of intimate images without authorisation.

Independent government agencies can also intervene to influence companies. This may involve facilitating collaboration among companies to establish safety standards or directly partnering with them to remove harmful content. For instance, Australia's eSafety Commissioner (eSafety) is an independent statutory office supported by the Australian Communications and Media Authority that receives and responds to complaints of image-based abuse on social media platforms, relevant electronic service or a designated internet service.[114] eSafety is empowered to investigate complaints and issue take-down notices to online platforms to remove content.

Support mechanisms and regulations

Many countries around the world have some form of regulation that provides various protections against image-based abuse. The Olimpia (or Olympia) Law in Mexico recognises and criminalises online gender-based violence, and particularly the perpetration of IIA.[115] Countries like Argentina and Panama have also passed their own Olimpia Law in recent times.[116] Criminalising the sharing of non-consensual images online is important for several reasons, including but not limited to: establishing societally that IIA is a crime, encouraging more victims-survivors to report their cases to the police, and detailing the requirements for law enforcement agencies to prosecute perpetrators. On the other hand, critics of the Olimpia Law in Mexico emphasise the need to do more to influence social media companies to moderate content and handle cases of IIA on their platforms.[117]

Laws that are specific are more likely to be enforceable and beneficial to victims-survivors. For example, under Canada's[118] laws regarding IIA, a perpetrator can be criminally charged if they disseminate an intimate image knowing that the person depicted in the image did not consent to that conduct. India's[119]law, which also covers shopped images, only require that the perpetrator who knowingly disseminated sexual images be charged, regardless of motive. In the UK, victims-survivors have the option of keeping their anonymity in civil cases, which encourages more victims-survivors to speak up.[120] However, laws that do not treat IIA with nuance can end up causing harm or penalising consensual actions. In the UK, a previous requirement for Crown Protection Services (2022) to prosecute offenses was to prove that the perpetrator's intent was to cause distress,[121] although that is no longer the case. In India, the same law that criminalises IIA also includes intimate photos that people may voluntarily share with their partners. Some laws can even end up prosecuting the victims-survivors. In East Africa, countries including Uganda and Tanzania penalise women for partaking in the creation of "pornographic content". Due to this, when their intimate images are leaked, they end up being charged while the perpetrators do not face consequences.[122]

Some countries have laws that hold intermediaries like social media companies to account in different ways. Australia's Online Safety Act (2021)[123] provides eSafety with regulatory powers to remove and act against the non-consensual sharing of, or threat to share, an intimate image online. In Singapore, regulations govern social media platforms through laws and policies. In 2023, Singapore revised the Online Safety Act[124] with additional provisions requiring social media and communication platforms to take necessary steps to reduce online harms. Another act, the Online Criminal Harms Act[125] granted the government the authority to direct websites and apps to remove accounts suspected of engaging in criminal activities, including abuse. In Africa, countries such as Ghana, Kenya, Uganda, and South Africa provide authority to social media platforms to remove content upon receiving takedown requests. While some of these laws focus specifically on IIA, others provide governments and social media companies with what some are concerned is disproportionate power to dictate free speech.[126]

The UK's Online Safety Act of 2023[127] has introduced a set of measures requiring social media platforms to have processes in place to assess and swiftly remove illegal content, minimise illegal content appearing on search services, implement 'empowerment tools' for better user control over their

feeds, and offer identity verification options for adult users. Another example of a pathway for escalation is the Trusted Flaggers program[128] introduced as a part of EU's Digital Service Act which involves identifying experts in detecting potentially illegal content online. The content these entities flag are given high priority by social media companies.

In some countries, cases or pre-existing laws set precedent for what is expected from intermediaries, with both positive and negative implications. In India, through the case of Mrs. X v. Union of India, the Delhi High Court directed two popular search engines to remove IIA content in response to a survivor's case.[129] On the other hand, in the US, Section 230 of the Communications Decency Act protects internet platforms from legal responsibility for content uploaded by third-party users.[130] Although this rule has significantly shaped the internet, it can complicate actions to hold intermediaries responsible in IIA cases, with bills like the SHIELD Act carving out exceptions for providers of communications services for content posted by other parties. However, these intermediary liability protections are not absolute, and governments actross the world continue to impose legal duties on social media platforms to protect users.[131]

Some countries have also begun regulating generative AI outputs. China,[132] for example, has established rules that deepfake developers have to follow, requiring them to obtain consent from users, verify identities, and register records of the data generated with the government. They are also required to report illegal deepfakes and offer recourse for people using their services. The EU also has a variety of regulatory frameworks that address deepfakes including the AI Act[133] and the Digital Services Act, which require labelling AI-generated content and action from social media companies to take down deepfakes spreading disinformation and abuse.[134]

Expert Interviews and In-Country Workshop Findings

As set out in the methodology of this report (Appendix B), this analysis included primary research with a diverse range of stakeholders to gain insights into IIA risks, tools and their effectiveness.

The interviews and workshops surfaced common themes when discussing the limitations that victims-survivors deal with when seeking help and trying to access justice. These include challenges related to access, awareness, and technical barriers, which can

prevent victims-survivors from seeking or receiving the support they need. Although the local contexts differ, there were some common findings.

Finding 1: Pitfalls of tech-based support systems

Insight	Relevant Actors
Existing automated systems lack sufficient trauma-informed approaches and often fail to capture the nuanced forms of violence that victims-survivors experience.	Social media platforms, Tools created by third parties and NGOs
Workshop attendees highlighted that victims-survivors frequently express discomfort with fully automated systems, citing the absence of human connection during moments of crisis and vulnerability. Some workshop participants noted that using a communication channel to reach out to support organisations on the same platform where the abuse occurred can be uncomfortable or triggering for victims-survivors.	Tools created by third parties and NGOs
Technological solutions often presume consistent internet access and digital literacy, marginalising rural users and those with limited connectivity, particularly in both Colombia and Nigeria.	Tools created by third parties and NGOs
Available tools frequently overburden already traumatised victims-survivors with complex reporting requirements across multiple platforms, creating administrative exhaustion.	Social media and communication platforms
Digital support on the sites where IIA takes place is rarely available in indigenous or local languages, creating significant language barriers in responding to abuse in non-English speaking countries. Additionally, current reporting options are limited to text-based, whereas improvements in speech-to-text technology create an opportunity for more inclusive reporting among low-literacy communities.	Social media and communication platforms

Workshop participants emphasised that while technology can serve as a valuable bridge to human support, it cannot replace trauma-informed human intervention in cases of intimate image abuse. Technology works best when it functions as an initial access point to guide victims-survivors toward appropriate human-centred services rather than as a standalone solution.	All actors

Finding 2: Lack of awareness in the reporting process

The interviews and workshops revealed significant knowledge gaps that prevent victims-survivors from effectively accessing support and justice, with cultural variations in how these barriers manifest.

Insight	Relevant Actors
Victims-survivors often do not recognise their experiences as abuse due to a normalisation of online violence or a lack of awareness about what is considered IIA, especially in contexts where cultural factors influence perceptions of harm.	Tools created by third parties and NGOs
Many victims-survivors are unaware of existing support that is available, creating an urgent need for improved mapping of support organisations and resources, and clearer pathways for seeking help and accessing justice that are accessible to users with varying levels of digital literacy and to users from different local contexts.	Tools created by third parties and NGOs
Support resources that exist on platforms are frequently difficult to locate and lack clear guidance for infrequent users of platforms attempting to report violations.	Social media and communication platforms
In both Nigeria and Colombia, workshop attendees highlighted that many law enforcement agencies, particularly in rural areas, lack the technical capacity to take on these cases, hindering effective resolution.	Regulatory entities, law enforcement agencies
Educational resources addressing IIA prevention are largely unavailable in local languages and contexts, particularly outside urban centres, limiting awareness-raising efforts in communities and populations more likely to experience abuse.	Tools created by third parties and NGOs
Digital literacy programmes are more effective when they are interactive, with phones and other devices used to demonstrate how to secure personal devices.	Regulatory entities, Tools created by third parties and NGOs

Workshop participants emphasised the need for multi-faceted awareness campaigns that reach beyond digital channels to include community-based outreach, particularly for populations with limited internet access or digital literacy. Many of the NGO participants based in Colombia and Nigeria

workshops are actively engaged in awareness-raising activity, but they face significant challenges as demand for support increases. Jacarandas, an NGO in Colombia, reported receiving over 2,000 requests of support from victim-survivors of IIA, double the volume handled by state resources, while operating with limited staff and funding. As the scale of IIA increases globally, capacity will be stretched highlighting the urgent need for substantial investment to expand their reach and effectiveness. Improving education about IIA must address not only technical aspects but also the underlying social norms that normalise or trivialise these forms of abuse.

Finding 3: Technical challenges when collecting information

Insight	Relevant Actors
Shared devices create significant privacy and safety risks when seeking help, as perpetrators or others who might stigmatise the victim-survivor may have access to the same device, potentially exposing help-seeking behaviour and endangering the victim-survivor. Evidence preservation on shared devices can also place the victim-survivor at additional risk.	Social media and communication platforms
Ephemeral messaging features and auto-deletion of content on platforms (like "view once" images or disappearing messages) create significant barriers to preserving evidence, as crucial information can be lost before proper documentation occurs. Anonymity features on social media and communication platforms also allow perpetrators to hide their identities.[135]	Social media and communication platforms
End-to-end encryption, while valuable for privacy, complicates content moderation in cases of IIA, particularly on WhatsApp where abuse frequently occurs.	Communication platforms
Evidence authenticity challenges arise when victims-survivors attempt to document abuse, as screenshots or recordings may not meet legal evidentiary standards in court proceedings.	Regulatory entities

Platform design often fails to incorporate accessible forensic tools that would help victims-survivors safely document abuse without technical expertise.	Social media and communication platforms
The need for digital literacy programmes that use phones and other devices to actively demonstrate how to secure personal devices.	Tools created by third parties and NGOs, Regulatory entities

Workshop participants, particularly in the digital safety and legal aid focus groups, emphasised the critical need for more accessible evidence collection tools that account for varying levels of technical literacy. In Nigeria, law enforcement representatives noted that current evidence collection methods often fall short of legal requirements, while Colombian participants highlighted how the technical hurdles in document preservation often discourage victims-survivors from pursuing cases altogether.

Finding 4: Limitations in the effectiveness of reporting

Insight	Relevant Actors
Platform response times for IIA reports are frequently excessive, with many victims-survivors never receiving responses to their submitted reports.	Social media and communication platforms, Regulatory entities

Organisations in both workshop countries reported challenges in escalating cases on a large messaging platform, despite its "Trusted Partner" status, due to a lack of elective reporting channels.	Social media and communication platforms, Regulatory entities
Lack of coordination between platforms and enforcement agencies create additional administrative burdens for already traumatised victims-survivors.	Social media and communication platforms, Law enforcement entities
Law enforcement and judicial system personnel frequently minimise or underestimate digital harms, leading to inadequate case handling and reinforcing survivor/victims' reluctance to report.	Law enforcement entities

Finding 5: Legislative challenges

Insight	Relevant Actors
Nigeria lacks comprehensive legislation specifically criminalising non-consensual sharing of intimate images, with victims-survivors forced to rely on fragmented provisions in the Cybercrimes Act and Violence Against Persons Prohibition Act.	Regulatory entities (Nigeria)
Colombia's Law 1257 of 2008 lacks specific provisions for intimate image abuse, forcing victims-survivors to pursue justice through general privacy violations that inadequately address the unique harms and digital nature of this offense.	Regulatory entities (Colombia)
Legal systems with overly punitive approaches often re-victimise women through the reporting and investigation process, creating disincentives to engage with formal mechanisms.	Regulatory entities
In Colombia, workshop participants noted that proposed legislative approaches modelled after Mexico's "Ley Olimpia" heavily favour criminal punishment over victim-survivor support and content removal, creating significant barriers to accessing justice for victims-survivors seeking alternative resolution pathways.	Regulatory entities (Colombia)

Cases that involve cross-border IIA harms face challenges due to limited standardisation in laws and international coordination mechanisms. This limits the options victims-survivors have when perpetrators operate from different jurisdictions. This observation is mirrored in the analysis of legislative frameworks that are not equipped to handle cases where the images are stored on devices outside of the associated country.	Regulatory entities
There are limited reparation frameworks for IIA. Workshop participants in Colombia discussed whether financial compensation, restraining orders, or content removal constituted appropriate remediation.	Regulatory entities, social media platforms

Recommendations

Analysing the landscape of tools across the AI content generation and communication platforms, tools created by third parties and NGOs, and regulatory entities — along with their limitations — reveals where interventions would be most effective. Common gaps across these actors include a lack of awareness about existing tools for supporting survivors, as well as the need for tools, guardrails, and policies that prioritise a survivor-centric approach. The following sections outline actionable recommendations to address IIA at every stage of its spread and identify the responsibilities of each actor involved.

Recommendation 1: Prevention Efforts

Platforms should also take efforts to prevent IIA harms before they occur through proactive safety measures, technology safeguards, and automatic protective features.

Actors:

- **Generative AI Companies** should implement tests during the model training pipeline to evaluate how effectively new models detect harmful prompts. These tests should address both intentional and unintentional biases the model may have learned, with a gendered intersectional approach. Model testing and harm detection systems should account for how TFGBV intersects with race, disability, sexuality, and other factors that may increase harm or vulnerability. For example, model developers should ensure that the model

does not respond to requests to generate intimate images from descriptions or existing images. This allows model developers to have baseline information on how safe the models are and track their performance. It also reduces the risks of "successful" (i.e. harmful) downstream prompts. Platforms hosting models should at a minimum, conduct checks on prompts to evaluate whether the requests to the model are harmful. Models should also have to pass a set of tests before being allowed to be hosted on the platform. Red Teaming exercises and bias bounties before models are released to the public are also beneficial in identifying bugs by soliciting feedback from a wide range of stakeholders.

- **Social media and Communication Platforms** should develop detection tools including an actively maintained database of known perpetrator information to validate against, hashes of known IIA images that are shared across platforms and regularly validated, and behaviours and patterns of perpetrators. This should build on tools used to prevent CSAM such as the Lantern initiative.[136] When using models to predict patterns or behaviours, it is important to ensure that the models do not introduce any new risks. Consequences for perpetrators should also be well-defined on social media and communication platforms. Efforts to prevent recidivism include disabling their accounts, limiting their access and ability to post or follow users.
- **Tools created by Third parties and NGOs** should help victims-survivors protect their images, through techniques such as watermarking so that they cannot be modified by generative AI tools as a preventative step to mitigate IIA harms.
- **Regulatory Entities** could bolster model testing efforts by encouraging clear reporting guidelines and Red Teaming activities. This would help facilitate AI models being evaluated for safety and performance and identifying and mitigating potential risks before deployment. (citation for sidebar)[137]

Recommendation 2: Education and Awareness

Users of tech tools and platforms must be made aware of IIA harms, how they manifest, and the appropriate actions to take if they experience or witness IIA affecting others. Awareness of digital rights, and consent go hand in hand with this. To ensure relevance and effectiveness, awareness-raising

content and tools should be co-designed with young people, survivors, and marginalised communities. This education can happen on platforms through more easily accessible tools, or via policies that aim to introduce this knowledge in schools or through technology literacy programmes.

Actors:

- Generative AI Companies should educate users on safety measures and have clear reporting information. Users should be presented with information about the risks of harm when downloading open-source models. Information about building guardrails should also be made available to developers who are using these models to build tools.
- Social media and Communication Platforms should make information about the available tools and safety policies easily accessible to users. A prioritised reporting system for intimate image and deepfake abuse can lead to quicker responses. Additionally, Safety Centres should allow users to track the status of their reports. Raising awareness about IIA harms when users share images on social media or communication platforms can help them better understand their own and others' privacy rights. For communication platforms, mechanisms that allow users to easily capture evidence and report harmful messages would reduce the burden on victims-survivors. There should also be clear workflows and guidance for users who witness abuse, enabling them to report posts on behalf of the victim-survivor and take action to block and report the perpetrator.
- Tools created by Third parties and NGOs should raise awareness of IIA risks, reduce the stigma associated with experiencing IIA and empower victims-survivors to report abuse without fear of judgment. Educating users on country-specific laws regarding what is legal and illegal can help deter perpetrators and ensure victims-survivors are informed of their rights. Interactive third-party tools that guide victims-survivors through evidence collection and reporting processes can also promote digital literacy and lower the technological barriers for victims-survivors.
- Regulatory Entities should create educational materials on the harms of deepfake technology and the prevalence of IIA to help increase awareness and reduce the stigma around TFGBV and IIA. Educational programmes should target diverse audiences, raise awareness about consent

in relationships as well as the risks of deepfake technology among young students. These programmes should focus on training law enforcement officers and frontline counsellors to respond to victims-survivors with trauma-informed care when they report IIA.

Recommendation 3: Standardisation and Collaboration

A major gap highlighted in this analysis was the lack of standardisation across industries, particularly in areas like reporting standards for tools, and TFGBV and IIA guardrails for AI models. In a fast-paced and frequently changing technology space, collaboration within the industry is crucial in getting victims-survivors the help they need and reducing their administrative burden. Furthermore, standards for guardrails and comprehensive tests for models before release can help prevent harms before they happen.

Actors:

- Generative AI Companies that have their models open to the public should have easily accessible reporting helplines where users can report harmful prompts or intimate images generated using the models. They should also make users aware of what is needed to submit a report and reduce the burden on users to complete and check the status of their report. Platforms that host models should also have a similar option for users to report models, along with a process to review the reports take down models that are harmful. In the CSAM space, the Tech Coalition[138] organises annual hackathons and working sessions with the aim of driving innovation and sharing resources. Similar efforts, via hackathons and working sessions, to encourage collaboration between generative AI companies to share known offensive prompts and test cases for IIA detection would also be beneficial.
- Social media and Communication Platforms should establish an industry-wide standard for responding to and removing intimate images. This should involve a dedicated reporting method on platforms distinct from other forms of reports to reduce the administrative burden on the victim-survivor. Removing non-consensual sexual content, including deepfakes, must be prioritised by these platforms, like the approach taken for CSAM content removal. Cross-platform collaboration can also streamline efforts to remove deepfake

content that is shared to multiple social media platforms. Platforms should commit to timely removal of intimate images in response to reports. Partnerships with third-party tools like StopNCII would also enable social media platforms to more effectively take down content by leveraging established reporting methods.

- Tools created by Third parties and NGOs should include a repository or directory of global and regional tools and resources on TFGBV and IIA. This would go a long way in signposting and referring victims-survivors to the support and resources they need and quicker.
- Regulatory Entities around the world should advocate for industry-wide compliance with safety standards for evaluation of generative AI models before release, regular Red Team assessments that target IIA harms, and cross-platform reporting and collaboration on policy. Standardisation is also required when it comes to collecting data on IIA prevalence. While the differences and benefits of various evaluation methods in the global majority countries have been discussed, having similar data collected around the world is beneficial for comparative studies. At a global scale, this data could be used to understand the significance of abuse and even divide efforts and funding in an appropriate manner. Regulatory entities should also draw from other fields where interventions have been effective to prevent and respond to online harms. For example, in the CSAM prevention space, the Interpol DevOps Group convenes law enforcement, NGOs, academia and tech companies to ideate and co-develop tools to improve online child safety.[139] Similar efforts should be made to bring together a wide range of stakeholders with different skill sets in the TFGBV or IIA space.

Recommendation 4: Survivor-Centric Tools and Design

To ensure that social media and communication platforms, and generative AI companies, are victim-survivor-friendly and easier to navigate, these platforms and tools should focus on building survivor-centric tools and adhering to Safety by Design principles when launching new products or updating existing ones. Putting the victim-survivor at the forefront allows platforms to build inclusive products and foster positive experiences for everyone.

Actors:

- Generative AI Companies should have accessible reporting options and easy-to-understand instructions for users to report harmful content. IIA evaluation methodologies designed and used by these actors should be made more transparent, allowing them to gather feedback that can help enhance their processes. Perpetrators can be more easily traced if these companies establish requirements for users to create accounts to use their services.
- Social media and Communication Platforms should by default, enable safety features, such as the blurring of nude images, with the option for users to switch them off. This builds safety-by-default workflows, and reduces the effort needed to stay safe online. Practically, this also helps platforms save resources and money by preventing harms rather than reactively identifying and moderating after harms have occurred. Users should also be given enough information to manage their own safety online and be able to easily find and understand safety settings. Posts or messages that contain deepfake images should be identified and labelled so that they cannot be used to spread misinformation or disinformation about someone. Recommender systems must establish controls to prevent suggesting sexual deepfake videos and other IIA content to users and block them from trending.
- Tools created by Third parties and NGOs should be interactive so as to guide victims-survivors through evidence collection and reporting processes and offer centralised resources for them.
- Regulatory entities should work collaboratively with the tech industry to encourage adoption of Safety by Design principles. eSafety in Australia offers implementation guides to support tech companies to embed safety into the design, development and deployment of their products, tools and platforms.

Appendix A: TFGBV Harms Taxonomy

A Framework for Understanding and Addressing Technology-Facilitated Gender-Based Violence

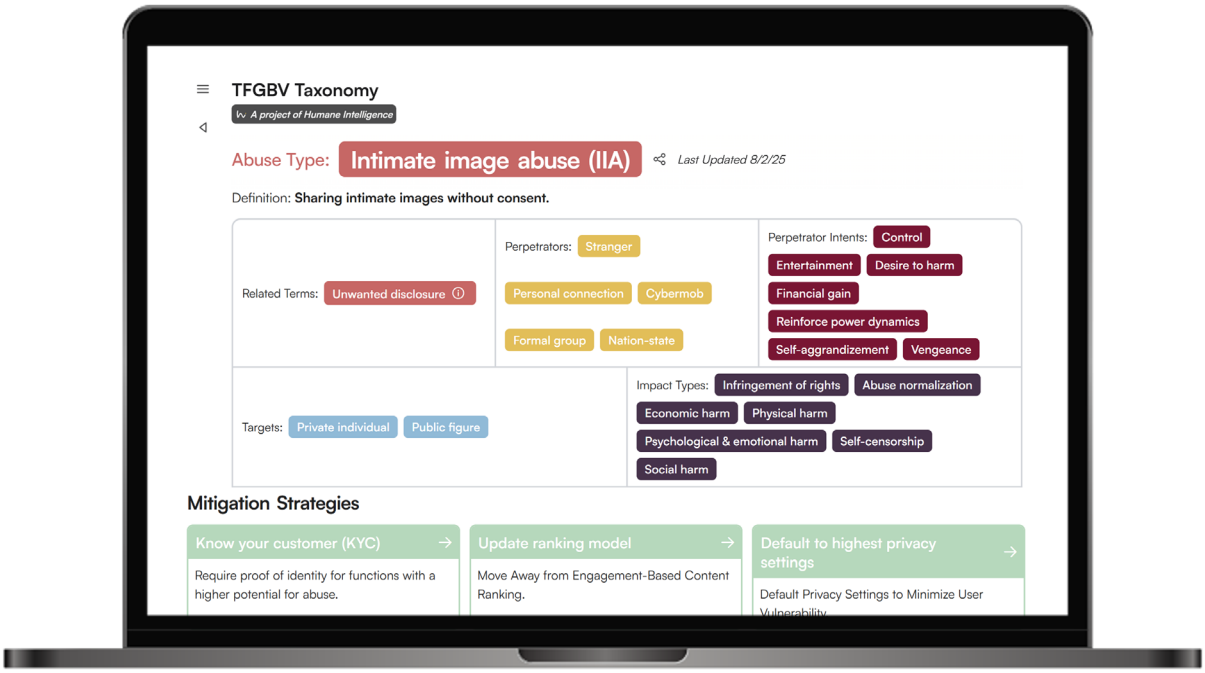
Introduction

Throughout the development of this Landscape Analysis Report, a significant gap emerged in how TFGBV is defined, categorised, and addressed across different contexts. While numerous advocacy organisations and researchers have created valuable resources documenting various forms of TFGBV, these approaches often lack standardisation and operational clarity that would enable platforms, policymakers, and safety practitioners to systematically identify and mitigate these harms.

The TFGBV Harms Taxonomy presented here was developed in direct response to this gap. Drawing upon the research findings in this report, the taxonomy provides a structured framework that bridges the gap between policy understanding and practical implementation. By systematically categorising forms of abuse, mapping relationships between perpetrators and targets, documenting impact types, and outlining mitigation strategies, this taxonomy serves multiple functions:

- It creates a common language for discussing TFGBV across various stakeholders
- It provides a framework for categorising and understanding different harm patterns
- It clarifies organisational responsibilities for addressing specific forms of TFGBV
- It provides actionable pathways for platforms to detect, prevent, and respond to these harms

Source: www.humane-intelligence.org/tfgbv-taxonomy



The taxonomy aims to comprehensively cover all forms of TFGBV, not just intimate image abuse (IIA). However, as part of this Landscape Analysis, the entry for IIA and its associated metadata have been published as a starting point, given the significant focus on this form of abuse within the research. It is intended to be a living document, which will continue to evolve, drawing upon the inputs of those working in the field, and most, importantly, incorporating the experiences of survivors and victims. Feedback on this resource is encouraged as it continues to be developed.

This taxonomy has been developed in partnership with leading organisations in the field, including StopNCII.org, University College London's Gender and Tech Research Lab, International Center for Journalists, and the Global Partnership for Action on Gender-Based Online Harassment and Abuse. By bringing together these diverse perspectives and expertise, the taxonomy represents a collaborative effort to standardise the understanding of TFGBV while respecting the nuanced insights each organisation brings to this complex issue.

Rather than replacing existing resources, this taxonomy synthesises and structures insights from multiple sources into an implementation-focused framework. It is designed to be accessible to technical teams, Trust & Safety operations, and policy professionals alike, facilitating the translation of research insights into concrete safety measures.

https://tfgbv.humane-intelligence.org/abuse-type/iia

Appendix B: Methodology

The approach and methodology were discussed and agreed between Humane Intelligence and UK Government at a kick off meeting in November 2024. It prioritises learning from local environments to enhance existing services for victims-survivors rather than duplicating them. The analysis involved the following pieces of qualitative research:

- A review of secondary literature on IIA in the global context.
- Primary research with a range of different stakeholders across multiple sectors. This included key informant interviews and in-country workshops offering diverse empirical experiences on best practice tools, gaps and solutions in addressing IIA.

Secondary research: secondary data sources included publicly available data that had to fulfil the following criteria:

- Focus: Global or regional evidence on the scope and nature of IIA, studies on how IIA occurs, interviews with survivors-victims on the consequences of IIA.
- Time period: 2017 — Present.
- Language: English.
- Publication status: Publicly available material.
- Geographical focus: Global.

Primary research: This focused on gaining insights from a diverse range of stakeholders around the world on IIA risks, tools available to address and mitigate these risks, and gaps in tools and their effectiveness. The research team and UK Government produced a longlist of stakeholders from which a final list was chosen, based on the research questions and prioritising stakeholders who had direct experience working with victims-survivors of IIA or within platforms that dealt with IIA content. This stage of the analysis took place between November 2024 and February 2025, and included:

- Nine key informant interviews with multi-sectoral stakeholders from around the world, including lawyers and policy advisors working in regulatory and tech industries, and founders of non-profits that assist victims-survivors of TFGBV through a wide range of support services. The

interviews were designed to identify the most common IIA risks, the tools available to address these and assess their effectiveness.

- Two in-country workshops in Bogota, Colombia, and Lagos, Nigeria, with 40 participants representing diverse actors involved in IIA, including policymakers, law enforcement, psychosocial support professionals, digital safety experts, civil society, NGOs, and regulatory entities. The workshops employed a prototype-based approach, allowing participants to map local systems, understand the requirements for context-specific adaptation, and perform sustainability planning to ensure any resulting tools enhance rather than duplicate existing services for victims-survivors. Thematic breakout groups serve as the primary source of ideation, with participants grouped according to their expertise. Within these groups, participants use a survivor-centric approach to map out the actual pathways that victims-survivors navigate when seeking help, documenting decision points, effectiveness of various platform tools, and available resources at each stage. These case studies aid in mapping local safety ecosystems and inform the recommendations from different country contexts.
- Interviews and the workshops were analysed and findings triangulated with those from the literature review.

Methodological limitations include:

- The analysis had a focus on multi-stakeholder engagement, in particular, on experts who work on IIA. Efforts were made to engage stakeholders who had direct experience working with victims-survivors, however the analysis did not directly engage victims-survivors themselves.
- The report primarily addresses intimate image abuse against women and girls. However, it should be noted some forms of IIA, such as sextortion, overwhelmingly affects young men, often using synthetic pornographic deepfakes as a luring tactic followed by blackmail. NCMEC's data on sextortion highlights the rapid increase in cases, from slightly over 10,700 cases in 2022 to over 26,700 in 2023. NCMEC also notes that teenage boys are the most common targets of financial sextortion schemes. While this report's focus on women was deliberate, reflecting the disproportionate rates at which women experience IIA as a form of gender-based violence, it creates a limitation in addressing the full spectrum of IIA experiences across genders.
- Globally, there are gaps in publicly available data on IIA. Therefore, whilst this report spotlights statistics on IIA from certain countries, this is not intended to suggest they are the most effected.

- While this analysis does not deliberately exclude gender minorities, there are significant data gaps regarding the specific experiences of people across the gender spectrum with intimate image abuse. As IIA continues to be an emerging form of TFGBV, existing research has not consistently collected or segmented data to allow for nuanced analysis of how different gender identities experience this abuse. The studies cited throughout this report reflect the current state of research, which has primarily focused on binary gender categories and lacks comprehensive disaggregated data across the full spectrum of gender identities.
- According to UN Women[140] "there is no commonly agreed upon definition of TFVAW, nor is there agreement on what constitutes various forms of TFVAW." This lack of standardised terminology creates challenges when comparing data across different studies and regions. The report notes that most of the existing evidence relies on self-reported data, which can lead to under-reporting due to various factors, including shame, fear of retaliation, or lack of awareness that what they experienced constitutes abuse.
- Research on TFGBV struggles to keep pace with rapid technological changes. This is especially true of IIA influenced by the rise in generative AI. This can make it difficult to develop timely and effective methodological approaches that account for emerging technologies and new forms of abuse.

Acknowledgments

This landscape analysis would not have been possible without the generous contributions of time, expertise, and insights from numerous individuals and organisations around the world. We extend our sincere gratitude to all who shared their knowledge, experiences, and perspectives throughout this research process.

Subject Matter Experts

We are deeply grateful to the following experts who participated in key informant interviews, sharing their invaluable insights on Intimate Image Abuse risks, available tools, and effective responses. Their diverse perspectives from regulatory, technological, legal, and victim support domains have greatly enriched this analysis.

- Joy Uchechi Eziashi from Impact Amplifier
- Lucía Gamboa from Credo AI
- Hera Hussain from Chayn
- Sophie Mortimer from SWGfL
- Andrea Powell from Alecto Foundation
- Lana Ramjit, PhD
- Shannon Raj Singh

Workshop Participants in Lagos and Bogotá

We extend our sincere appreciation to all organisations and participants in the Lagos and Bogotá workshops who evaluated prototype support tools and co-developed victim-survivor pathways from their diverse perspectives. Their local expertise and collaborative engagement were instrumental in documenting country-specific challenges, identifying culturally appropriate solutions, and establishing valuable networks for potential future implementations.

Lagos

- Chioma Agwuegbo from TechHerNG
- Yewande Gbola-Awopetu from Federal Ministry of Justice of Nigeria
- Chimdimma Ike, Alexandra Maduagwu and team from The Initiative for Equal Rights (TIERs)
- Nkechika Ibe from Impact Her World Foundation
- Ruhamah Ifere from Trully Verify Africa
- Babatunde-Martins Moromoke from IDERA SARC, Lagos State

- Eweka Yvonne O. from Co-creation HUB (CcHUB)
- Juliet Ohahuru-Obiora from Action Against Child Sexual Abuse Initiative (ACSAI)
- Nancy-Olive Tamuno from Mirabel Centre
- Juliet Olumuyiwa-Rufai
- Olasupo Abideen Opeyemi from Brain Builders Youth Development initiative (HerSafeSpace)
- Monsurat Oyindamola Oshinfisan from W.TEC
- Peculiar Showale from Paradigm Initiative

Bogota

- Mg. Sandra Balanta Cobo from Mesa de Economía Feminista de Cali
- Viviana Bohórquez from Jacarandas
- Profesora Lina M. Céspedes-Báez from Universidad del Rosario
- Natalia Escobar Váquiro from Observatorio para la Equidad de las Mujeres
- Beldys Hernández from Colombia Diversa
- Laura Natalia Méndez Garzón
- Catalina Moreno Arocha from Fundación Karisma
- Natalia Carolina Ramírez Bosiga from Unidad de trabajo legislativo (UTL) Senadora Clara López
- Laura Marcela Urrego Aguilera from El Veinte
- Herlinda Villarreal González from Ministerio de Igualdad y Equidad, Viceministerio de las Mujeres

Contributors to TFGBV Harms Taxonomy

The development of the Technology-Facilitated Gender-Based Violence Harms Taxonomy benefited immensely from the following contributors, whose specialised knowledge helped create a structured framework that bridges the gap between policy understanding and practical implementation.

- Eva Blum-Dumontet from Chayn
- Sofia Bonilla
- Sophie Mortimer from SWGfL
- Sarah O'Connell
- Mikayla Pang Zi Rui
- Susan Ragheb
- Nabeelah Shabbir from International Center for Journalists (ICFJ)
- Dr Leonie Maria Tanczer and Dr Nikolaos Koukopoulos from University College London's Computer Science Dept., Gender + Tech Research Lab
- Elodie Vialle
- Jen Weedon

GBV Terms

Term	Definition
Gender-Based Violence (GBV)	Elimination of Violence Against Women DEVAW describes violence against women as 'any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life.' The term GBV has since been broadened out in other definitions to include acts of violence against people based on their gender or gender identity, including members of the LGBTQIA+ community. This report uses the term GBV in the way it was originally intended in DEVAW, to reflect that violence against women is driven by structural gender inequality and gender discrimination. However, this report is intentional in its use of GBV to reflect violence against women and girls in all their diversity, including lesbian, bisexual and trans women, and recognises that women, girls and the LGBTQIA+ community who experience intersecting forms of oppression (for example where gender discrimination and sexism overlap with homophobia and transphobia) are disproportionately affected by GBV, including TFGBV.[141]
Intimate Image Abuse (IIA) also known as Image-Based Sexual Abuse (IBSA), Image-Based Abuse (IBA)	Image-based abuse consists of a broad range of abusive behaviours, including sexual abuse, through the creation and non-consensual distribution of images, or threats thereof. This includes non-consensual creation and distribution of intimate images (also known as non-consensual pornography), voyeurism/creepshots, sexual extortion, unsolicited sexual images, the documentation or broadcasting of sexual violence, and non-consensually created artificial sexual media, including sexual deepfakes. This also includes images and videos taken with prior consent but shared without consent.

Non-Consensual Intimate Images (NCII)	Intimate images that are distributed without the consent of the person depicted in them, which can include images originally obtained with consent (such as images consensually shared within a relationship) but later distributed without consent.
Online Harassment	Unwanted aggressive behaviors carried out via digital means that aim to frighten, anger, shame, or abuse the targeted individuals, including threats, offensive comments, misinformation about the individual and the sharing of embarrassing or private content.
Online Misogyny	Expressions of hatred, prejudice, or contempt directed at women in online spaces, including misogynistic narratives, gender-related hate speech, and coordinated attacks that seek to silence the voices of women and gender minorities in digital spaces.
Technology-Facilitated Gender-Based Violence (TFGBV)	TFGBV is an overarching term that reflects the wide range of different technologies that can be used to perpetrate violence and abuse against women and girls. UN Women and the World Health Organisation have defined TFGBV as "any act of gender-based violence that is committed, assisted, aggravated or amplified by the use of information communication technologies (ICT) or digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms.[142]
Victim-Survivor	A term that recognise the agency, resilience and autonomy of those who have experienced technology-facilitated gender-based violence while also acknowledging the harm they have experienced; it respects that individuals may identify as victims, survivors, both, or neither depending on their personal experiences and recovery journey.

Technical Terms

Term	Definition
Closed Source Model	An AI model whose internal architecture, weights, and training methodologies are proprietary and not publicly accessible, typically developed by commercial entities with restricted usage terms.
Content Credentials	Provenance metadata attached to generated media that includes information such as creation timestamp, originating platform, and certification authority, designed to authenticate digital content.
Deepfakes	Advanced AI-generated synthetic media that uses deep learning algorithms to superimpose or replace faces and voices in existing content, creating falsified but convincing representations that are increasingly weaponised against women and girls through non-consensual sexualised content, political disinformation, and reputation damage.
Digital Hashing	A cryptographic process that transforms an image into a unique fixed-length string of characters (hash value), enabling identification of identical or visually similar images without storing the original content.
Digital Safety	The set of practices, technologies, and policies designed to protect users from online harms, encompassing privacy controls, content moderation, and user education.
End-to-End Encryption	A security protocol that ensures message content is accessible only to the sender and intended recipient, preventing intermediaries, including platform providers, from accessing the unencrypted data.

Fine-tuning (models)	The process of adapting a pre-trained AI model to a specific use case or domain by conducting additional training with specialised datasets, enhancing the model's performance for targeted applications.
Generative AI	AI systems capable of creating novel digital content including text, images, audio, and video by identifying patterns in training data and generating outputs that mimic those patterns.
Guardrails	Technical limitations and safety mechanisms implemented in AI systems to prevent the generation of harmful content, including content filters, input validation, and output constraints.
Media Matching Service (MMS)	A technological system that maintains a database of known violating content signatures, allowing platforms to compare user-uploaded media against these signatures to prevent redistribution.
Models	AI systems trained on large datasets to recognise patterns and generate predictions or content, which in the context of technology-facilitated gender-based violence can be designed either to perpetuate harm through the creation of non-consensual intimate imagery or even to detect and mitigate such harmful content.
Model Hosting Platforms	Services that provide the computational infrastructure and interfaces necessary for deploying, running, and making AI models accessible to users, often with variable levels of safety oversight.
AI Alignment Principles (e.g. Model Spec, AI Constitution)	Formalised guidelines that define acceptable AI system behaviors, used by model developers to instruct models on distinguishing between appropriate and inappropriate requests.

Multi-modal Foundation Models (MFMs)	Large-scale AI systems trained on diverse data types (text, images, audio) that serve as the basis for various applications and can generate multiple forms of content from different input modalities.
Open-source Models	AI systems whose underlying code, architecture, and trained weights are publicly available for anyone to access, modify, and distribute, which in the context of technology-facilitated gender-based violence can present heightened risks as they may be downloaded with fewer safeguards against generating harmful content like non-consensual intimate imagery.
Prompts	Text or image inputs provided to AI systems that instruct or guide the model to generate specific types of outputs, which can be manipulated by users to produce desired or, in some cases, harmful content.
Recommender Systems	Algorithmic frameworks that analyze user behavior, preferences, and content characteristics to determine content visibility and promotion across platforms, potentially amplifying harmful material.
Red Teaming for AI Safety	The systematic practice of stress-testing AI systems by attempting to elicit harmful, biased, or otherwise problematic responses, conducted continuously throughout a model's lifecycle to strengthen safeguards and respond to emerging threats.
Safety by Design	A product development philosophy that incorporates safety considerations throughout the design, development, and deployment phases, rather than addressing safety concerns retroactively.
Third-Party Tools (for IIA victim-survivors)	Software applications and platforms developed by entities independent of content creation platforms, designed to assist victims-survivors in documenting abuse, reporting content, and accessing support services.

Training (models)	The computational process of exposing an AI system to large datasets from which it learns patterns and relationships, enabling it to subsequently generate predictions or content based on these learned patterns.
Training Data	The collection of examples used to develop an AI model's capabilities, which for generative AI typically consists of vast amounts of text, images, and other media scraped from the internet.
Trusted Flaggers	Verified experts or organisations with specialised knowledge in identifying harmful online content who are granted privileged status by online platforms to report violations with higher priority review, serving as an enhanced reporting mechanism to help identify and remove technology-facilitated gender-based violence more efficiently than standard user reporting.

1. Ahlenback, V., Fraser, E., Kalsi, K., & Vlahakis, M. (2023). Technology-facilitated gender-based violence: preliminary landscape analysis. Social Development Direct. https://www.sddirect.org.uk/resource/technology-facilitated-gender-based-violence-preliminary-landscape-analysis

2. Ibid., 1

3. SWGfL. (2025). Adult Online Harms National Response Model. SWGfL https://swgfl.org.uk/resources/adult-image-based-abuse-model-response/

4. Security Hero. (2023). State of deepfakes: Overview of the current state. SecurityHero. https://www.securityhero.io/state-of-deepfakes/#overview-of-current-state

5. Eagleton, J. (2022, October). Marked as unsafe: How online platforms are failing domesticabuse survivors. Refuge. https://refuge.org.uk/wp-content/uploads/2022/11/Marked-as-Unsafe-report-FINAL.pdf

6. Ibid., 3

7. UNFPA. (n.d.). Measuring technology-facilitated gender-based violence: Discussion paper. United Nations Population Fund. https://www.unfpa.org/publications/measuring-technology-facilitated-gender-based-violence-discussion-paper

8. UN Women. (2023). Expert group meeting report: Technology-facilitated violence against women. https://www.unwomen.org/sites/default/files/2023-03/Expert-Group-Meeting-report-Technology-facilitated-violence-against-women-en.pdf

9. Ibid., 1

10. UN Women. (2021). Violence against women in the online space: insights from a multi-country study in the Arab States. UN Women. https://arabstates.unwomen.org/en/digital-library/publications/2021/11/violence-against-women-in-the-online-space#view

11. Iyer, N., Nyamwire, B., & Nabulega, S. (2020). Alternate realities, alternate internets: African feminist research for a feminist internet. Association for Progressive Communications. https://www.apc.org/sites/default/files/Report_FINAL.pdf

12. Amnesty International. (2017, November 21). Amnesty reveals alarming impact of online abuse against women. Amnesty International. https://www.amnesty.org/en/latest/press-release/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/

13. O'Brien, M. (2024, January 4). Online violence: Real life impacts on women and girls in humanitarian settings. International Rescue Committee. https://blogs.icrc.org/law-and-policy/2024/01/04/online-violence-real-life-impacts-women-girls-humanitarian-settings/#:~:text=TFGBV%20is%20rooted%20in%20the,harmful%20masculinities%20that%20cause%20GBV

14. Ibid., 1

15. Generation G partnership. (2024, June). Decoding technology-facilitated gender-based violence: A reality check from seven countries. Rutgers University. https://rutgers.international/decoding-technology-facilitated-gender-based-violence/#:~:text=Decoding%20technology%2Dfacilitated%20gender%2Dbased%20violence%3A%20a%20reality%20check,based%20violence%20in%20seven%20countries.

16. Ibid., 1

17. Posetti, J., Bontcheva, K., Harrison, J., Maynard, D., Shabbir, N., Torsner, S., & Aboulez, N. (2023). The chilling: A global study of online violence against women journalists. International Center for Journalists. https://www.icfj.org/sites/default/files/2023-02/ICFJ%20Unesco_TheChilling_OnlineViolence.pdf

18. International Institute for Democracy and Electoral Assistance. (2021). Women's political participation: Africa Boikarometer 2021. https://www.idea.int/sites/default/files/publications/womens-political-participation-africa-barometer-2021.pdf

19. UN Women. (2021, November). Violence against women in the online space: Insights from a multi-country study in the Arab States. UN Women. https://arabstates.unwomen.org/en/digital-library/publications/2021/11/violence-against-women-in-the-online-space

20. Ibid., 15

21. Ibid., 17

22. Ibid., 1

23. Kaspersky. (2024). The naked truth: Kaspersky report on online abuse. https://media.kasperskydaily.com/wp-content/uploads/sites/86/2024/07/15164921/The-Naked-Truth-Kaspersky.pdf

24. Ruvalcaba, Y., & Eaton, A. A. (2020). Nonconsensual pornography among U.S. adults: A sexual scripts framework on victimization, perpetration, and health correlates for women and men. Psychology of Violence, 10(1), 68—78. https://doi.org/10.1037/vio0000233

25. Gurumurthy, A., Vasudevan, A., & Chami, N. (2019). Born digital, born free? A socio-legal study on young women's experiences of online violence in South India. https://itforchange.net/sites/default/files/1662/Born-Digital_Born-Free_SynthesisReport.pdf

26. Centre for International Governance Innovation (CIGI). (2023). Supporting a safer internet: Global survey of gender-based violence online (Grant No. 109247-001). Centre for International Governance Innovation. https://idl-bnc-idrc.dspacedirect.org/server/api/core/bitstreams/4b2265e1-f259-49b5-8301-b35a5e02aa69/content

27. Van der Merwe, A., & Nene, S. (2021). Understanding online gender-based violence in Southern Africa. University of Pretoria, Centre for Human Rights. Retrieved from https://www.chr.up.ac.za/images/researchunits/dgdr/documents/resources/FINAL_v_Understanding_oGBV_in_Southern_Africa.pdf

28. Mao, F., Ng, K., & Zubair, M. (2023, November 28). Pakistan: Woman killed after being seen with man in viral photo. BBC. https://www.bbc.com/news/world-asia-67551554

29. Swenson, A., & Chan, K. (2024, March 14). Election disinformation takes a big leap with AI being used to deceive worldwide. Associated Press. https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd

30. .Vengattil, M., & Kalra, A. (2022, July 21). Facebook's growth woes in India: Too much nudity, not enough women. Reuters. https://www.reuters.com/technology/facebooks-growth-woes-india-too-much-nudity-not-enough-women-2022-07-21/

31. McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A., & Powell, A. (2020). 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. Social & Legal Studies, 30(4), 541-562. https://doi.org/10.1177/0964663920947791 (Original work published 2021)

32. GBV Learning Network. (n.d.). What You Need To Know About Non-Consensual Sexual Deepfakes: A growing problem. https://www.gbvlearningnetwork.ca/our-work/infographics/nonconsensualsexualdeepfakes/index.html

33. Murphy, S. (2019, October 31). Diane Abbott speaks out on online abuse as female MPs step down. The Guardian. https://www.theguardian.com/politics/2019/oct/31/diane-abbott-speaks-out-on-online-abuse-as-female-mps-step-down

34. Foster, M. J. (2012, November 26). Calling the shots: How ownership structures affect the independence of news media. Center for International Media Assistance. Retrieved from https://www.cima.ned.org/resource/calling-the-shots-how-ownership-structures-affect-the-independence-of-news-media-2/

35. Li, E. R., Shultz, B., & Jankowicz, N. (2024, December 11). Deepfake pornography goes to Washington: Measuring the prevalence of AI-generated non-consensual intimate imagery targeting Congress. https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67586997eaec5c6ae3bb5e24/1733847451191/ASP+DFP+Report.pdf

36. Amnesty International. (2024, May 9). State-backed digital violence to silence women and LGBTI activists in Thailand. https://www.amnesty.org/en/documents/asa39/7955/2024/en/

37. Ibid., 23

38. eSafety Commissioner. (n.d.). Image-based abuse: Perpetrator motivations. eSafety Commissioner. https://www.esafety.gov.au/research/image-based-abuse-perp-motivations#:~:text=There%20was%20a%20strong%20sense,getting%20away%20with%20similar%20actions.

39. Yeung, J., & Seo, Y. (2020, November 25). South Korean leader of online sexual blackmail ring sentenced to 40 years. CNN. https://edition.cnn.com/2020/11/25/asia/korea-telegram-sex-crime-verdict-intl-hnk/index.html

40. Bel, S., with Jang, H. (2024, November 20). The gendered battle over digital sexual abuse in South Korea. New Lines Magazine. https://newlinesmag.com/reportage/the-gendered-battle-over-digital-sexual-abuse-in-south-korea/

41. Human Rights Watch. (2023, February 5). Online misogyny and the manosphere. https://humanrights.ca/story/online-misogyny-manosphere

42. Eko. (2023, March). Suicide, Incels, and Drugs: How TikTok's deadly algorithm harms kids. Eko. https://s3.amazonaws.com/s3.sumofus.org/images/eko_Tiktok-Report_FINAL.pdf

43. Roberts, S., & Wescott, S. (2024, July 1). We research online 'misogynist radicalisation'. Here's what parents of boys should know. Monash Lens. https://lens.monash.edu/@education/2024/07/01/1386838/we-research-online-misogynist-radicalisation-heres-what-parents-of-boys-should-know

44. Singer, N. (2024, April 8). Deepfake AI nudes scandal at West-field High School. The New York Times. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html

45. New: Yoon, M.-s. (2024, December 3). 80% of suspects in deepfake crime cases are teens: NPA. The Korea Herald. https://www.koreaherald.com/article/10012227

46. eSafety Commissioner. (n.d.). Generative AI - position statement. eSafety Commissioner. https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

47. OpenAI. (n.d.). ChatGPT. https://chatgpt.com/

48. Microsoft. (n.d.). Copilot. https://copilot.microsoft.com/

49. Google. (n.d.). Gemini. https://gemini.google.com/

50. OpenAI. (n.d.). DALL·E 3. https://openai.com/index/dall-e-3/

51. Meta. (2025). Meta AI. https://www.meta.ai/

52. Burgess, M. (2020, October 21). A deepfake porn bot is being used to abuse thousands of women. Wired. https://www.wired.com/story/a-deepfake-porn-bot-is-being-used-to-abuse-thousands-of-women/

53. Burgess, M. (2024, October 15). Millions of people are using abusive AI 'nudify' bots on Telegram. WIRED. https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/

54. Oversight Board. (n.d.). Explicit AI Images of Female Public Figures. https://www.oversightboard.com/decision/bun-7e941o1n/

55. Ibid., 4

56. Rahman-Jones, I. (2024, January 27). Taylor Swift deepfakes spark calls in Congress for new legislation. BBC News. https://www.bbc.com/news/technology-68110476

57. Heikkilä, M.(2022, December 12). The viral AI avatar app Lensa undressed me without my consent. MIT Technology Review. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

58. OpenAI. (n.d.). GPT-4. https://openai.com/index/gpt-4/

59. Anthropic. (n.d.). Claude: A next-generation AI assistant. https://www.anthropic.com/claude

60. DeepSeek. (2025). DeepSeek: Into the unknown. https://www.deepseek.com

61. Character.AI. (n.d.). Character AI chatbot. https://character.ai/

62. Character.AI. (n.d.). Character AI chatbot. https://character.ai/

63. eSafety Commissioner. (n.d.). Generative AI - position statement. Australian Government, eSafety Commissioner. https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

64. OpenAI. (n.d.). DALL·E 2: Pre-training mitigations. https://openai.com/index/dall-e-2-pre-training-mitigations/

65. Kelley, A. (2024, September 13). White House leads public-private commitment to curb AI-based sexually abusive material. NextGov https://www.nextgov.com/artificial-intelligence/2024/09/white-house-leads-public-private-commitment-curb-ai-based-sexually-abusive-material/399524/

66. Thiel, D. (2023, December). Investigation finds AI models trained on child abuse images. Stanford Cyber Policy Center. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse

67. OpenAI. 2024, May 8. Introducing the model spec. https://openai.com/index/introducing-the-model-spec

68. Anthropic. (2023, May 9). Claude's constitution: A new era in AI ethics. https://www.anthropic.com/news/claudes-constitution

69. Microsoft. (n.d.). Protecting the Public from Abusive AI-Generated Content. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Protecting-Against-Abusive-AI-Content-UK.pdf

70. Lakatos, S. (2023, December 8). A revealing picture: AI-generated 'undressing' images move from niche pornography discussion forums to a scaled and monetized online business. Graphika. https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf

71. Hsu, T. (2024, February 5). Fake and explicit images of Taylor Swift started on 4chan, study says. The New York Times. https://www.nytimes.com/2024/02/05/business/media/taylor-swift-ai-fake-images.html

72. Owen-Jackson, C. (2024, June 7). Open source, open risks: The growing dangers of unregulated generative AI. IBM. https://www.ibm.com/think/insights/unregulated-generative-ai-dangers-open-source

73. Bluesky. (2024, January 10). Profile post on digital rights and abuse. https://bsky.app/profile/koltai.bsky.social/post/3lfnpa5sr522k

74. Tenbarge, K. (2024, March 14). Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy. NBC News. https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071

75. Discord. 2024, March 15. Non-consensual intimate media policy explainer. Discord. https://discord.com/safety/non-consensual-intimate-media-policy-explainer

76. Tenbarge, K. (2023, November 10). Google announces new steps to combat sexually explicit deepfakes. NBC News. https://www.nbcnews.com/tech/tech-news/google-announces-news-steps-combat-sexually-explicit-deepfakes-rcna164560

77. Heikkala, M. (2024, August 6). Google is finally taking action to curb non-consensual deepfakes. MIT Technology Review. https://www.technologyreview.com/2024/08/06/1095774/google-is-finally-taking-action-to-curb-non-consensual-deepfakes/

78. Mumbi, L. (2024, November 25). Facebook, WhatsApp top platforms for online gender-based violence - report. Eastleigh Voice. Facebook, WhatsApp top platforms for online gender-based violence - report

79. Bergman, S., Kirwan, J., & Vaughan, C. (2024, October 14). Understanding technology-facilitated gender-based violence in Asia: A qualitative study. UNFPA Asia and the Pacific. https://asiapacific.unfpa.org/sites/default/files/pub-pdf/2024-10/Understanding%20technology-facilitated%20gender-based%20violence%20in%20Asia_1.pdf

80. Rodriguez, K. (2023, June 4; updated July 7, 2023). Revenge porn victims powerless. Trinidad Express. https://trinidadexpress.com/news/local/revenge-porn-victims-powerless/article_8707f124-027d-11ee-b3db-47ebfe0afc19.htm

81. eSafety Commissioner. (n.d.). Recommender systems and algorithms — position statement. eSafety Commissioner. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

82. Pen America. (n.d.). Blocking, muting, restricting. https://onlineharassmentfieldmanual.pen.org/blocking-muting-restricting/

83. WhatsApp. (n.d.). How to block contacts on WhatsApp. WhatsApp. https://faq.whatsapp.com/1805617343145907/?helpref=hc_fnav

84. Meta. 2024, April 26. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook. Meta. https://transparency.meta.com/sr/dsa-transparency-report-apr2024-facebook#:~:text=16.4%20hours&text=In%20instances%20where%20there%20are,staff%20and%20therefore%20more%20time.

85. Iyer, P. (2024, October 11). New Research Highlights X's Failures in Removing Non-Consensual Intimate Media. Tech Policy Press. https://www.techpolicy.press/new-research-highlights-xs-failures-in-removing-nonconsensual-intimate-media/

86. Ibid., 5

87. Telegram. (n.d.). FAQ: How to report abuse or bots on Telegram. https://telegram.org/faq#:~:text=If%20you%20see%20a%20bot,act%20on%20the%20owner's%20behalf

88. Global Disinformation Team. (2022, February 15). Telegram: Where Women's Nudes Are Shared Without Consent. BBC. www.bbc.com/news/world-60303769

89. WhatsApp. (n.d.). How to report a problem on WhatsApp. WhatsApp. https://faq.whatsapp.com/538382354780857/?helpref=hc_fnav

90. Klar, R. (2024, April 11). Instagram automatically blurring nudity in direct messages to teens. The Hill. https://thehill.com/policy/technology/4587967-instagram-automatically-blurring-nudity-in-direct-messages-to-teens/

91. Robertson, A. (2022, September 15). The Meta Oversight Board says Facebook's automated image takedowns are broken. The Verge. https://www.theverge.com/2022/9/15/23353593/meta-facebook-oversight-board-decisions-automated-image-takedowns-extremist-groups

92. Kira, B. (n.d.). When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0267364924000906#cit_74

93. Ibid., 5

94. Perrino, J. (2022, July 6). Using safety by design to address online harms. Stanford Cyber Policy Center. https://cyber.fsi.stanford.edu/news/using-safety-design-address-online-harms#:~:text=Under%20such%20a%20scheme%2C%20users,fundamentally%20change%20our%20online%20experiences.

95. StopNCII.org. (n.d.). Home. StopNCII.org. https://stopncii.org/

96. Papachristou, K. (2023). Revenge porn helpline report 2023. Revenge Porn Helpline. https://revengepornhelpline.org.uk/assets/documents/revenge-porn-helpline-report-2023.pdf?_=1714738699

97. PIRTH. (n.d.). PIRTH: Public Interest Research in Technology and Human Rights. https://www.pirth.org/

98. Tech Safety. (n.d.). Resources for women in tech safety. https://techsafety.org.au/resources/resources-women/

99. TechHerNG. (n.d.). Resources. TechHerNG. https://techherng.com/resources-2/

100. Ibid., 97

101. StopNCII.org. (n.d.). Industry partners. StopNCII.org. https://stopncii.org/partners/industry-partners/

102. StopNCII.org. (n.d.). Global network of partners. StopNCII.org. https://stopncii.org/partners/global-network-of-partners/

103. Chayn. (n.d.). About Chayn: Empowering women against abuse. https://www.chayn.co/about

104. SWGfL. (n.d.). SWGfL online safety resources and training. https://swgfl.org.uk/

105. Vadehra, J. (2024, September 13). Addressing image-based sexual abuse in the Global South: A call for inclusive global trust and safety standards. Gender Matters. https://gendermatters.in/addressing-image-based-sexual-abuse-in-the-global-south-a-call-for-inclusive-global-trust-and-safety-standards/

106. Goggin, B. (2024, March 29). Big Tech companies reveal trust and safety cuts in disclosures to Senate Judiciary Committee. NBC News. https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435

107. United Nations Office on Drugs and Crime. (n.d.). Home. United Nations Office on Drugs and Crime. https://www.unodc.org/

108. Reuters. (2024, September 26). South Korea to criminalize watching or possessing sexually explicit deepfakes. Reuters. https://www.reuters.com/world/asia-pacific/south-korea-criminalise-watching-or-possessing-sexually-explicit-deepfakes-2024-09-26/

109. Cai, V. (2025, March 6). China facing growing campaign to make it compulsory to label AI-generated content. South China Morning Post. https://www.scmp.com/news/china/politics/article/3301337/china-facing-growing-campaign-make-it-compulsory-label-ai-generated-content

110. Parliament of Australia. (2024). Criminal Code Amendment (Deepfake Sexual Material) Bill 2024. Parliament of Australia. https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r7205

111. House of Commons Library. (2024). The law on non-consensual sharing of intimate images (Research Briefing No. LLN-2024-0070). House of Commons Library. https://researchbriefings.files.parliament.uk/documents/LLN-2024-0070/LLN-2024-0070.pdf

112. Ibid., 3

113. United Nations Office on Drugs and Crime (UNODC). (2024, December). United Nations convention against cybercrime. United Nations Office on Drugs and Crime. https://www.unodc.org/unodc/en/cybercrime/convention/home.html

114. eSafety Commissioner. (n.d.). Home. eSafety Commissioner. https://www.esafety.gov.au/

115. Diario Oficial de la Federación. 2024, February 14. Decree amending and adding various provisions to the General Law on Women's Access to a Life Free of Violence and the Federal Penal Code. https://digital-policyalert.org/change/9207#timeline

116. Forum Diplomático. (2025, March 4). El Movimiento Ley Olimpia: Lucha contra la violencia digital en toda América Latina y el Caribe. Forum Diplomático. https://www.forumdiplomatico.com/en/2025/03/04/el-movimiento-ley-olimpia-lucha-contra-la-violencia-digital-en-toda-america-latina-y-el-caribe/

117. Hernández Oropa, M., Chavarría García, P. I., Contreras Chávez, I., Hernández Vélez, A. L., Ayala Real, L. G., Quevedo Berrelleza, M. D. C., & Ponce Toledo, D. (2024). Digital sexual violence against women in Mexico: Role of the Olimpia Law in transforming underlying gender norms. ALIGN Platform. https://www.alignplatform.org/resources/report-digital-sexual-violence-against-women-mexico-olimpia-law

118. Criminal Code, R.S.C., 1985, c. C-46, s. 162.1. (n.d.). Justice Laws Website. Government of Canada. https://laws-lois.justice.gc.ca/eng/acts/C-46/section-162.1.html

119. India Code. (n.d.). The Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011. Ministry of Electronics and Information Technology, Government of India. Retrieved March 17, 2025, from https://www.indiacode.nic.in/show-data?actid=AC_CEN_45_76_00001_200021_1517807324077&orderno=81#:~:text=Whoever%2C%20intentionally%20or%20knowingly%20captures,two%20lakh%20rupees%2C%20or%20with

120. Judiciary of England and Wales. (n.d.). Anonymity of witnesses and reporting restrictions. https://www.judiciary.uk/guidance-and-resources/anonymity-of-witnesses-and-reporting-restrictions/

121. Huber, A. R. (2023). Image-based sexual abuse: Legislative and policing responses. Criminology & Criminal Justice, 25(3), 736-752. https://doi.org/10.1177/17488958221146141

122. Sippy, P. (2021, October 4). Hackers stole and leaked her private photos. Then she was charged for breaking anti-pornography laws. Rest of World. https://restofworld.org/2021/women-east-africa-cybercrime-laws/

123. Australian Government. (2021). Telecommunications (interception and access) amendment (online safety) act 2021 (C2021A00076). Australian Government. https://www.legislation.gov.au/C2021A00076/latest/text

124. Ministry of Communications and Information, Singapore. (2023, January 31). Online Safety (Miscellaneous Amendments) Act takes effect. https://www.mddi.gov.sg/media-centre/press-releases/online-safety-act-takes-effect-on-1-february-2023/#:~:text=Online%20Safety%20(Miscellaneous%20Amendments)%20Act%20Takes%20Effect%20on%201%20February%202023

125. Singapore Police Force. (n.d.). Introduction to the Online Criminal Harms Act (OCHA). https://www.police.gov.sg/Advisories/Online-Criminal-Harms-Act/Introduction-to-OCHA

126. Media Defence. (n.d.). Online violence against journalists: NCII in the digital era. https://www.mediadefence.org/ereader/publications/online-violence-against-journalists/module-2-digital-attacks-and-online-gbv/ncii/

127. UK Government. (2024, February 7). Online Safety Act explainer. UK Government. https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

128. European Commission. (2023). Trusted flaggers under the Digital Services Act (DSA). https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa

129. Columbia Global Freedom of Expression. (n.d.). Case: Mrs. X v. Union of India. https://globalfreedomofexpression.columbia.edu/cases/mrs-x-v-union-of-india/

130. Ortutay, B. (2023, February 21). What you should know about Section 230, the rule that shaped today's internet. PBS NewsHour. https://www.pbs.org/newshour/politics/what-you-should-know-about-section-230-the-rule-that-shaped-todays-internet

131. Xenakis, N., & Lee, D. (2024, July 17). U.S. Senate passes SHIELD Act to criminalize distribution of private intimate images online. Inside Privacy. https://www.insideprivacy.com/privacy-data-security/u-s-senate-passes-shield-act-to-criminalize-distribution-of-private-intimate-images-online/

132. Kharpal, A. (2022, December 22). China is about to get tougher on deepfakes in an unprecedented way. Here's what the rules mean. CNBC. https://www.cnbc.com/2022/12/23/china-is-bringing-in-first-of-its-kind-regulation-on-deepfakes.html

133. European Commission. (2023). Regulatory framework for AI in Europe. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

134. Leerssen, P. (2024, August 6). Embedded GenAI on social media: Platform law meets AI law. DSA Observatory. https://dsa-observatory.eu/2024/10/16/1864/#:~:-text=price%20of%20one.-,Content%20Labeling,disclaimer%20attached%20to%20the%20content

135. Whilst anonymity can protect perpetrators, it is important to women and members of the LGBTQIA+ community who want to speak out without fear of backlash, as well as to victims-survivors who want to re-entre online spaces anonymously.

136. Litton, S. (2023, November 7). Announcing Lantern: The first child safety cross-platform signal sharing program. The Tech Coalition. https://www.technologycoalition.org/newsroom/announcing-lantern

137. Ofcom. (2025, February 25). Consultation on draft guidance: A safer life online for women and girls. Ofcom. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/a-safer-life-online-for-women-and-girls

138. Tech Coalition. (2024, October 7). Initiate 2024: Advancing industry collaboration to combat online child sexual exploitation and abuse. Tech Coalition. https://www.technologycoalition.org/newsroom/initiate-2024

139. Safe Online. (2023, November 6). Tech tools to tackle digital harms — INTERPOL. Safe Online. https://safeonline.global/tech-tools-to-tackle-digital-harms-interpol/

140. Berryhill, A., & Fuentes, L. (2023, March). Technology-facilitated violence against women: Taking stock of evidence and data collection. UN Women. https://www.unwomen.org/sites/default/files/2023-04/Technology-facilitated-violence-against-women-Taking-stock-of-evidence-and-data-collection-en.pdf

141. United Nations. (1993). Declaration on the elimination of violence against women. https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.21_declaration%20elimination%20vaw.pdf

142. UN Women and WHO (2022). Technology facilitated Violence against Women: Towards a common definition. Report of the meeting of the Expert Group 15-16 November 2022. New York: Joint Programme on VAW Data https://www.unwomen.org/en/digital-library/publications/2023/03/expert-group-meeting-report-technology-facilitated-violence-against-women

DIGITAL VIOLENCE, REAL WORLD HARM

Evaluating Survivor-Centric Tools for Intimate Image Abuse in the age of Gen AI

www.humane-intelligence.org/tfgbv