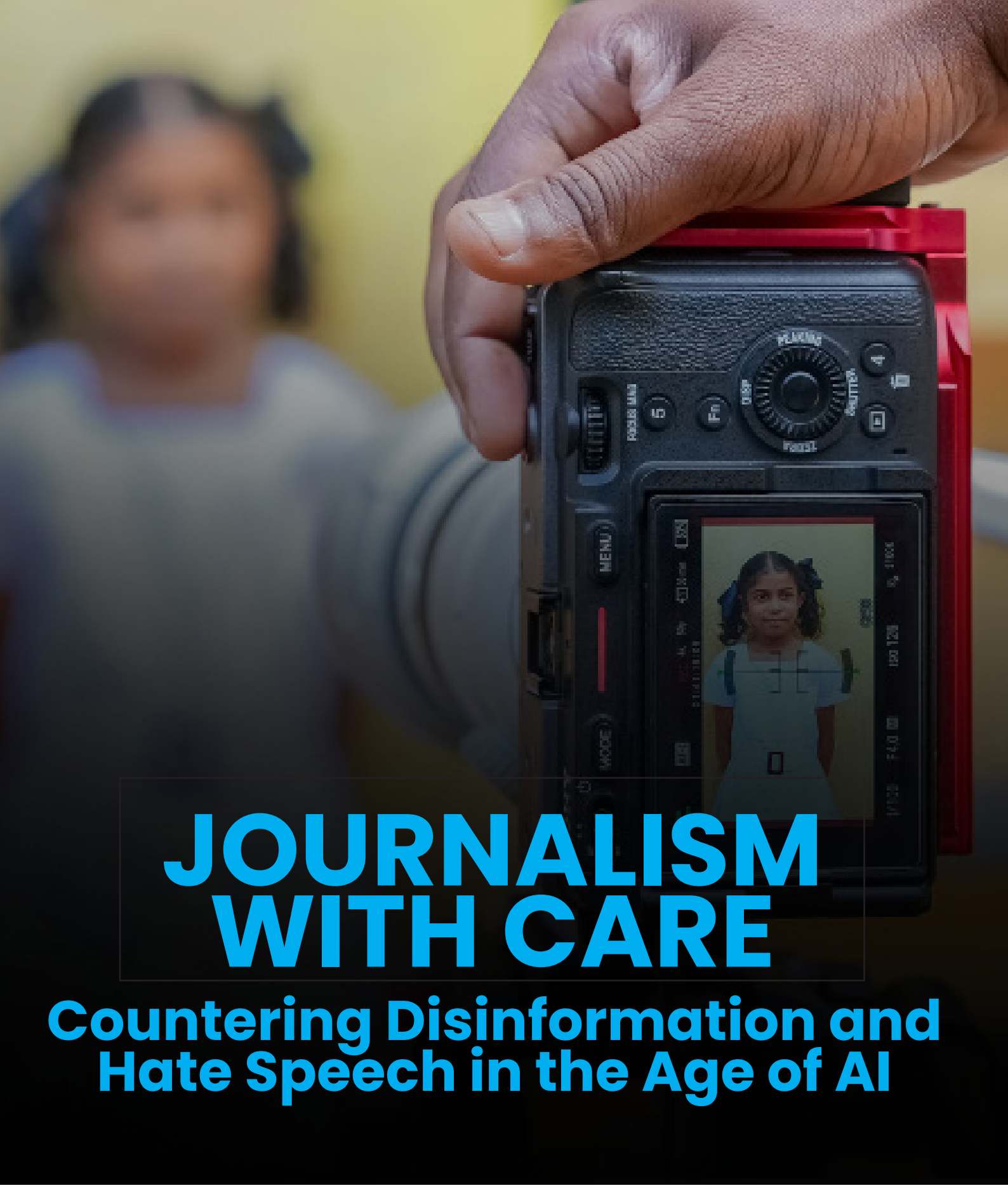

JOURNALISM WITH CARE

Countering Disinformation and Hate Speech in the Age of AI

Drafted by:

factum

UNITED NATIONS SRI LANKA

Supported by:

UNDP

unicef for every child
සෑම දරුවෙක් වෙනුවෙන්ම
ஒவ்வொரு சிறுவர்களுக்காகவும்

JOURNALISM WITH CARE

Countering Disinformation and Hate Speech in the Age of AI

Contents

EXECUTIVE SUMMARY

This training module is crafted to support journalists, content creators and any other individuals engaged in information creation/dissemination with a comprehensive framework for ethical, accurate, and safe reporting in the digital age. It places particular emphasis on addressing the risks posed by synthetic media, online hate speech, and harmful digital narratives.

The manual is structured around six core areas:

1. Digital tools to debunk claims: Fact-checking strategies, authentication of images and videos, and the use of AI-detection technologies to verify content before publication.
2. Identifying synthetic dis/misinformation: Understanding the risks of AI-generated media, including deepfakes and falsified content, and preventing the spread of harmful manipulation.
3. Hate speech coverage: Examining gendered hate, child-directed abuse, and other discriminatory narratives; avoiding stereotypes; and ensuring inclusive representation of minority voices.
4. Integrating safeguard mechanisms into newsrooms: Embedding sensitive pre-publication reviews, risk assessments, and ethical oversight in editorial workflows.
5. Social media platform policies: Navigating strengths and limitations in platforms such as Meta, X, TikTok, and YouTube regarding AI-generated content, dis/misinformation, and child safety; and techniques for preserving evidence before takedown.
6. Content and legal ethics: Applying Sri Lanka's legal and ethical standards to reporting involving children, digital media, and synthetic content.

Ultimately, this module promotes a newsroom culture centered on impartiality, harm-minimization, empathy, and professional integrity. By following these guidelines, users of this module can contribute to a safer and more inclusive digital environment for Sri Lankans.

INTRODUCTION

Artificial Intelligence (AI) is reshaping how societies function and how information is created, shared, and understood. While AI offers unprecedented opportunities, from enhancing accessibility and boosting economic productivity to strengthening public services, it also brings significant risks that affect institutions, communities, and individuals alike.

Sri Lanka's National AI Strategy highlights the transformative potential of AI across agriculture, public administration, transport, education, and environmental management. As these technologies advance, the public is increasingly exposed to algorithmic decision-making systems and personalized digital environments that influence behaviour, shape opinions, and alter civic participation.

AI is also redefining the information ecosystem. Newsrooms now operate in a landscape where AI-generated images, videos, and text can circulate globally within seconds. The same tools that support efficiency - from automated transcription to content recommendation - can also accelerate the spread of disinformation. AI-powered content farms, bots, and synthetic accounts can flood platforms with false narratives, while generative models produce highly convincing fake audio-visual materials that are difficult to distinguish from authentic reporting. This has made verification and debunking significantly more challenging, especially as journalists must produce accurate news at unprecedented speed to match the pace of digital dissemination.

A particularly harmful dimension is the rise of AI-fuelled hate speech and targeted harassment. Gendered disinformation, including deepfake explicit material, fabricated quotes, and manipulated imagery disproportionately attacks women, activists, and public figures, eroding trust, silencing voices, and distorting democratic processes. These tactics are increasingly used to intimidate or discredit individuals, making online spaces more hostile and unequal.

Children, too, are not spared: they may appear in manipulated media, be targeted by deceptive content, or be exposed to harmful narratives shaped by algorithmic amplification. Children's safety and security are also increasingly challenged, as they can be exposed to harmful AI-generated content, manipulation, and long-lasting digital traces beyond their control. Their digital footprints, once online, can influence long-term opportunities and wellbeing.

In this rapidly evolving landscape, ethical and responsible journalism is essential. This manual equips media professionals with practical tools to navigate AI-driven challenges, ranging from identifying synthetic media to mitigating harm and to uphold accuracy, integrity, and public trust in an era where truth itself is increasingly contested.

UNDERSTANDING SYNTHETIC MEDIA

Synthetic media refers to any image, video, audio, or text that is created or altered with the help of Artificial Intelligence. It is called synthetic because the content is not captured directly from real-world events. Instead, it is "made" or "synthesized" by algorithms. These systems study large collections of data and learn patterns such as how faces move, how voices sound, or how writing is structured. They then use this knowledge to generate new material that can look and sound very real.Common examples include AI chatbots like ChatGPT, image generators like Midjourney, voice-cloning tools, and deepfake software that can swap faces or recreate voices. While these tools can be used for creative or helpful purposes, they also introduce new risks.

One major concern is that synthetic media is becoming increasingly difficult to tell apart from genuine content. A manipulated video or AI-generated photo can spread across social media within seconds, and many viewers may accept it as real before journalists or fact-checkers have time to verify it. This makes it easier for false information to influence public opinion, damage reputations, or create confusion during emergencies and elections.

Synthetic media also poses serious risks to individuals. AI tools have been misused to create fake evidence, alter speeches, or produce harmful and non-consensual sexualized images often targeting women, activists, and children. Because these images or videos look real, the damage to a person's safety, mental wellbeing, or reputation can be severe and long-lasting. For children, this risk is even higher, as they may not fully understand how such content is created or how it can impact their future.

At the same time, synthetic media does have positive uses. It can support people with disabilities through AI-generated voice assistance, help storytellers visualize concepts, and assist journalists by automating simple tasks like transcription or illustration. The key is understanding both the potential and the risks.

For journalists, having a strong grasp of how synthetic media works is essential. It helps them verify content more accurately, avoid sharing misleading material, and report responsibly on incidents involving AI-generated content. Most importantly, it allows newsrooms to maintain public trust at a time when truth and authenticity are increasingly challenged.

References:

- swgfl.org.uk/topics/synthetic-media-deepfake
- techuk.org/resource/synthetic-media-what-are-they-and-how-are-techuk-members-taking-steps-to-tackle-misinformation-and-fraud.html
- cjr.org/the-synthetic-media-issue

HARMS OF SYNTHETIC MEDIA ON VULNERABLE GROUPS/ SITUATIONS

Synthetic media has transformed the information landscape. Alongside its benefits, it poses serious risks for children, women, and individuals from marginalized communities.

- Amplifying Stereotypes: AI-generated or manipulated media can reinforce harmful racial, religious, gendered, or cultural stereotypes. Because AI systems learn from biased datasets, they often reproduce or amplify existing societal prejudices, making marginalized groups more vulnerable to misrepresentation. Ethnic minorities, displaced individuals, refugees, and people with disabilities are often disproportionately targeted.

- Sexual Exploitation and Abuse: Deepfake sexual images and videos remain a critical global concern. AI can produce non-consensual sexualized content involving women and children, used to threaten, silence, or shame victims. Removal mechanisms on platforms are inconsistent and often slow, leaving victims vulnerable and exposed.

- Psychological Harm: Exposure to synthetic abuse can lead to shame, trauma, fear, and social isolation. Children may experience long-term anxiety and stigmatization based on identity attributes such as religion, ethnicity, or disability.

- Loss of Trust: The widespread presence of manipulated media undermines public confidence in authentic stories shared by vulnerable groups. Survivors' testimonies may be discredited as "fake," creating barriers to justice, assistance, and support.

- Public Incitement: During elections or periods of unrest, synthetic media can be deployed to discredit leaders, inflame divisions, and escalate hate campaigns. In Sri Lanka, where historical tensions remain sensitive, such manipulation can rapidly fuel conflict.

- Disruption During Disasters and Emergencies: Synthetic media can severely damage accurate reporting during crises. AI-generated images or videos of floods, landslides, or violence can spread faster than verified information, creating panic or false reassurance. Fabricated warnings or altered disaster footage can mislead communities, undermine emergency responses, and delay critical lifesaving decisions. Journalists face increased pressure to verify content quickly, while misinformation overwhelms official communication channels.

References:

UNICEF – Generative AI: Risks and Opportunities for Children
Case Western Reserve University Journal of Law & Technology
UNICEF Global Insight – Digital Mis/Disinformation and Children

IDENTIFYING SYNTHETIC DIS/MISINFORMATION

Synthetic dis/misinformation refers to false, misleading, or manipulated content created or enhanced using AI tools such as deepfake software, image generators, voice-cloning systems, or advanced editing techniques. Unlike traditional misinformation, synthetic content can look and sound extremely realistic, making detection harder even for trained journalists. Because these tools can fabricate entire images, voices, and events, synthetic misinformation spreads faster and is often more persuasive.

Key Definitions

- Misinformation: False or misleading information shared without intent to deceive. Example: sharing an outdated flood photo believing it is from a current disaster.
- Disinformation: False information intentionally created or spread to cause harm, influence opinion, or manipulate public behaviour. This includes deliberately edited videos of public figures or fake alerts during emergencies shared with the intention of creating panic and chaos.
- Malinformation: Genuine information shared with harmful intent. Examples include leaking private photos, exposing a child's identity, or publishing personal addresses during political or community disputes.

Common Warning Signs of Synthetic Content

Visual inconsitencies

- Uneven or impossible lighting
- Irregular facial symmetry or mismatched skin tones
- Unnatural reflections or shadows
- Distorted fingers, accessories, or background objects
- Blurred hairlines, warped edges, or overly smooth skin

Audio and video anomalies

- Lip movements that do not match spoken words
- Robotic pacing or inconsistent emotion
- Sudden background noise changes
- "Cut-off" breathing sounds or unnatural pauses

Text or language irregularities

- Repetitive or oddly structured sentences
- Overly formal, generic, or emotionless tone
- Abrupt topic shifts
- Inaccurate details about places, culture, or people

These signs are not always conclusive, especially as AI tools improve, but they help guide further verification.

Exercise

- Review a mixed set of five images or videos (real, edited, and AI-generated).
- Use verification tools such as reverse image search, metadata checkers, deepfake detectors, and source triangulation.
- Discuss the potential harm if misidentified synthetic content is published, especially during disasters, elections, or child-related reporting.

Safeguards for Journalists

- Cross-check information using independent, credible sources before publishing. Ie: verified media outlets, government agencies, official documents etc.
- Verify origin: Identify the first upload, original creator, and full context.
- Be cautious with political, nationalist, or emotionally charged content, which is often targeted for manipulation.
- Avoid assumptions about a child's identity, ethnicity, or circumstances based on appearance, name, accent, or location.
- Slow down when content evokes strong emotion; synthetic media often relies on shock, fear, or anger to spread quickly.

By embedding these practices into daily work, journalists/content creators can strengthen accuracy, avoid amplifying harmful narratives, and uphold public trust in an era of AI-driven information manipulation.

GENDERED HATE SPEECH AND SYNTHETIC MEDIA

Gendered hate speech refers to abusive, demeaning, or discriminatory language targeted at individuals on the basis of gender or gender expression. In digital spaces, gendered hate often intersects with sexism, misogyny etc. and is frequently weaponized against women journalists, activists, and public figures.
Gendered hate speech and synthetic media intersect in ways that both amplify harm and complicate accountability. When combined with synthetic media technologies, such as AI-generated images, videos, and audio, this form of abuse becomes more difficult to combat.
The combination of gendered abuse and synthetic media dramatically amplifies harm.

Generative AI makes it easy to:

- Create fabricated sexual images of women and girls,
- Manipulate videos to depict false behaviour,
- Pair synthetic visuals with misogynistic captions or slurs,
- Coordinate harassment campaigns against vulnerable targets.

These attacks create both intimate harm (shame, fear, reputation damage) and public harm (mob harassment, professional discrediting, political silencing etc.).

Groups at risk of being victimized by gendered hate speech includes women journalists and content creators, teenage girls, women in politics or public leadership roles, women from minority ethnic or religious communities and LGBTQIA+ individuals.

Synthetic content can be created in minutes, shared widely before verification, and remain online long after it is debunked. As generative AI tools become more accessible, the challenge is not only technological also societal, requiring legal, policy, and cultural responses that recognize the unique and disproportionate harm faced by women and children in digital spaces.

Journalistic Responsibilities

- Do not amplify hate speech, even when reporting on it.
- Provide contextual, rights-based coverage that avoids reproducing harmful language.
- Seek consent before naming or quoting individuals targeted by gendered harassment.
- Apply do-no-harm principles and prioritize survivors' safety.
- Correct misinformation quickly, especially when synthetic media is used to sexualize or discredit women or children.

- Journalists receive basic knowledge on the process of fast-tracking and verification, especially on issues related to children and adolescents

- Further strengthen the ability to engage in investigative journalism mainly on issues related to children

- Journalists will engage in high standards of ethical conduct in their reporting on children and adolescents

- Create an opening for the development of a multi-stakeholder driven code of conduct for

WHY IT MATTERS FOR NEWSROOMS

Gendered hate speech has real-world consequences, including self-censorship, withdrawal from public life, mental health harm, and increased risk of violence. Responsible reporting can disrupt these cycles of harm, expose manipulative campaigns, and protect vulnerable individuals.

Preventing the spread of hate speech and online harassment requires more than awareness, it also requires journalists and newsrooms to actively cultivate practices that help verify sources and stop the spread of hate speech and online harassment. Journalists and newsrooms must establish clear policies, train staff, and monitoring systems to respond quickly and ethically. When harmful misinformation or hate speech emerges, journalists can activate a fact-check and counter-narrative process of their own.

They can begin by verifying content using local fact-checking tools and trusted sources. Journalists can also verify corrections and debunk false information as soon as they can with clear evidence. It is important to engage with impacted communities to provide their perspectives as well. A newsroom can also assign a specific team member to review flagged content, especially during elections or periods that require extra sensitivity and attention. This can also help detect harmful content early.

It is necessary to consider the aftermath of the publication of sensitivity issues. It is important to track audience reactions and see if the content has fueled any form of hatred or harassment. Embedding hate speech prevention and false information identifiers into newsroom workflows is not just an ethical safeguard, it is also a protective measure against reputational damage, legal risk, and loss of public trust. Proactive monitoring, rapid response, and diverse sourcing are essential for responsible journalism in Sri Lanka's digital age.

Tools that can be used to identify synthetic media:

Category	Tools	Use
Image Verification	Hive Moderation, Google Lens, TinEye	Reverse search, detect manipulation
Video verification	DeepWare	Extract frames and find originals
AI detection	Google SynthId	Detect synthetic patterns
Geolocation	PicArta	Confirm location/time

INTEGRATING HATE SPEECH MONITORING & PREVENTION INTO NEWSROOM WORKFLOWS

Preventing the spread of hate speech requires structured newsroom policies, trained staff, and monitoring systems.

Key practices include:

- Assigning a designated reviewer to assess flagged content during high-risk periods such as elections or unrest.
- Fact-checking harmful narratives quickly and using evidence-based counter-messaging.
- Engaging affected communities to include their perspectives and ensure sensitive coverage.
- Tracking audience reactions to assess whether published content has inadvertently fueled harassment.
- Documenting harmful content (screenshots, URLs, timestamps) before it is removed, ensuring accountability.
- Embedding monitoring into newsroom workflows protects both the public and the integrity of journalism..

Tools and Best Practices of Social Media Companies to Address Harmful Synthetic Media

Social media companies have developed various policies and tools to address the growing problem of synthetic media — but their application is inconsistent, especially in smaller markets like Sri Lanka. It is imperative that journalists understand these measures both to use them effectively and to critically assess whether platforms are are meeting their responsibilities.

1. Meta (Facebook, Instagram, WhatsApp, Threads)

Policies:

- Removes synthetic and manipulated media that promotes false narratives on important issues (including elections and major crises).
- Labels content altered and distorted by AI when detected.
- Prohibits nudity or sexual activity involving minors, including AI-generated abuse images.

Tools for Journalists:

- Meta Transparency Center: Access to policy documents and enforcement reports.
- Reporting Mechanisms: In-app options to flag harmful or synthetic media for removal.

Limitations:

- Removal delays are incredibly common, especially outside of high-profile global incidents
- This delay can be reduced when working with trusted Meta partners.
- Harmful content may circulate widely before action is taken. This may exacerbate damage to the affected individual.

2. X (Twitter)

Policies:

- X's policy on manipulated media focuses on addressing content that is likely to cause harm, rather than banning all synthetic media.
- X has updated its terms of service to allow the platform to use user-generated content for training AI models, including Grok, their own AI.
- X has zero tolerance towards any material that features or promotes child sexual exploitation.
- Allows parody or satire but requires clear disclosure..

Tools for Journalists:

- Community Notes: Adds contextual corrections visible to all users.
- Advanced Search Filters: Narrow searches to track the spread of a specific video or image.
- Verification Badges: Can help audiences identify official sources (though paid verification has complicated this signal).

Limitations:

- Enforcement depends on users sending in reports and may not prioritize local language content.
- X's policies have largely shifted after Elon Musk's acquisition and the introduction of X's own AI tool, Grok.

3. TikTok

Policies:

- Prohibits "deepfake" content unless clearly labelled and used for entertainment.
- Bans depictions of minors in sexual contexts, including AI-generated material.

Tools for Journalists:

- Content Reporting: Direct flagging of harmful synthetic media.
- Transparency Reports: Includes data on content removal in Sri Lanka (though minimal).

Limitations:

- Highly visual platform where manipulated content spreads rapidly.
- Algorithm may recommend synthetic media to large audiences before moderation occurs, and the content is taken down

4. YouTube

Policies:

- Prohibits manipulated content intended to mislead voters or cause harm.
- Labels altered content where possible.

Tools for Journalists:

- Copyright Match Tool: Detects re-uploads of original content — useful for spotting manipulated reposts.
- Policy Enforcement Reports: Data on content removal.

Limitations:

- AI-generated content detection is still developing; much relies on community flagging. occurs, and the content is taken down

Platform	AI/Synthetic Media Policy	Child Safety Policy
Meta	Labels/manages AI content; slow removals	Removes CSAM but enforcement gaps are weak
X	Limited AI labelling	Weak child safety moderation
TikTok	Prohibits misleading AI	Removes harmful
YouTube	Labels AI	Moderates harassment

Journalists can also integrate their own practices when using these tools. It is important to act quickly and flag harmful content before it spreads quickly. They must notify the authorities if they believe a child is at risk. Journalists must also ensure they document evidence, report content, take screenshots and archiving links for verification and accountability. If they notice a harmful synthetic video, they must report it across platforms to ensure it gets taken down or report it to groups who actively work on flagging content. It is also necessary to track the platform response times to report on their mechanism systems.

CONTENT & LEGAL ETHICS IN CREATING CONTENT THAT COUNTERS ONLINE HARM

Creating content to counter misinformation, hate speech or harmful synthetic remains an integral part of journalism. Journalists have a responsibility for upholding ethical reporting and principles that fall in line with journalistic and legal standards. It is important to consider that even balanced writing can cause harm or promote dangerous narratives.

It is important to fact check and verify every counter-narrative. They must be independently verified and countered with evidence-based explanations. It is also important that journalists label any synthetic media they create with visible labels that show it was created by algorithmic generators.

When countering content about children or marginalized groups, apply the same checks one would when reporting on them in any form of news. A journalist must anonymize their identities and seek consent from their guardians or the individuals themselves. It is important not to promote stereotypes when countering as well. Journalists must be careful about using harmful language, even when countering, as it can help escalate the narrative even further. They can also include perspectives from those directly targeted by the harmful claims to balance the narrative.

Although there are no specific laws governing media in Sri Lanka, there are quite a list of laws in place in Sri Lanka that are pertaining to media and its conduct. Among the laws related to the media are:

- Fundamental rights (FR) laws
- Regulatory laws (including the Sri Lanka Telecommunication Regulatory Act, Online Safety Act etc.) Criminal laws (Penal Code, Obscene Publications Act, ICCPR Act, Personal Data Protection Act and others)
- National security laws (Public Security Act, Official Secrets Act, Emergency Regulations, Prevention of Terrorism Act)
- Intellectual property laws
- Defamation laws
- Election laws
- Consumer protection laws
- And other regulations such as professional codes of conduct. There are specific laws related to children and their appearance in media that include:
- Children and Young Persons (Harmful Publications) Act No. 48 of 1956
- Provisions of the Penal Code like Clause 286 (a) under Section 19
- Clause 21 of the Online Safety Act
- Clauses 11 and 20 of the Children and Young Persons Act No 48 of 1939
- Code of Conduct from the Sri Lanka Press Council

In the digital age, where harmful information can spread faster than the truth, synthetic media and AI can distort reality with ease. It remains the duty of every journalist and content creator to protect the truth and help safeguard children. For children the harm is multiplied. Every harmful stereotype and digitally manipulated image has the potential to shape truth and directly impact the safety and dignity of all young people in Sri Lanka.

Journalism with Care:
Countering Disinformation and Hate Speech in the Age of AI

Design: Haritha Dahanayaka
Edited: Factum

29A Marine Drive, Colombo 03, Sri Lanka
info@factum.lk

Factum is an organization committed to pluralism, equal opportunities and inclusivity. It respects the rights of all genders and communities, including marginalized groups.

For every child, every right

Tel: +94 112 677 550
For more information, please visit the UNICEF website at
https://www.unicef.org/srilanka/