AI and Online Safety: Emerging Risks and Opportunities

netsafe

AI ASIA PACIFIC INSTITUTE

Acknowledgements

We would like to thank the following for their contributions to this paper:

AI Asia Pacific Institute

Alex Toh, Project Lead

Simran Singh, Research Associate

Arpit Gupta, Research Intern

Koshiki Chauhan, Research Intern

Wendy Bothman, Contributor

Dr. Peter Leong, Contributor

Kelly Forbes, Contributor

Peter Brimble, Contributor

Netsafe

Brent Carey

Michael des Tombe

Foreword

Generative Artificial Intelligence is reshaping the digital landscape, bringing with it remarkable opportunities alongside significant challenges. Among these challenges are the online harms associated with AI - ranging from bias and misinformation to more subtle, systemic impacts - some of which require urgent attention and thoughtful intervention. While these harms can manifest globally, their local effects call for inclusive and contextually informed responses.

This discussion paper was initially commissioned by Netsafe to inform our own work concerning online harms. As New Zealand's leading organisation in this space, we recognised the growing importance of understanding and addressing the online harm risks posed by AI. However, we quickly saw that the insights and ideas uncovered through this research have broader relevance and value. With this in mind, we are sharing this paper publicly to stimulate discussion and collaboration with policy makers, te ao Māori experts, communities and the wider public on these critical issues.

We acknowledge that this is an exploratory paper and by no means definitive. Its purpose is to provoke thought, generate meaningful dialogue, and inspire further discussion and research. In particular, the paper does not delve into the specific harms GenAI may cause for Māori, Pasifika and other underrepresented and socially excluded groups and communities in New Zealand. This is an area that requires further expert input and needs its own research effort.

Online harms caused by AI are complex, evolving, and multi-faceted, and we believe that only through collective effort can we develop effective solutions. We welcome feedback, particularly on any gaps that may exist or other areas requiring deeper investigation. The initial findings in this paper may also spur further projects in the research, science, innovation and technology sectors. We aim to conduct further engagement in 2025. For example, through workshops to explore the findings and proposals outlined and to identify gaps and to hear from individuals and groups who may be specifically impacted.

We wish to express our sincere gratitude to the researchers whose efforts have underpinned this work. Their contributions are an essential foundation for the ongoing work.

We encourage readers to engage with the issues raised, offer their feedback, and join us in the collective challenge of addressing AI-driven online harms. We hope the paper's insights will be valuable to all those committed to creating a safer, fairer digital environment.

Brent Carey

CEO

Netsafe

Table of Contents

Executive Summary

Advancements in Artificial Intelligence (AI) technology, including Large Language Models (LLMs) and Generative AI (GenAI), have improved the quality of AI-generated content and the ability of AI models to personalise content. Tools that produce such content are more widely accessible, and content is cheaper, easier and faster than ever to create. Optimisations over time have meant that these models aren't limited to large organisations with limitless resources. High performing, uncensored models which are now able to be run on consumer hardware at home now exist. LLMs also assist with querying and understanding large volumes of unstructured data.

These advancements are both a boon and a bane. Studies consistently predict that GenAI will improve both productivity and Gross Domestic Product, even if estimates of the magnitude of benefits differ. In contrast, some of the negative impacts of AI are already emerging in the context of online harms.

In order to inform policymakers and other stakeholders, this paper examines the impact of advancements in AI on various categories of online harms, including child sexual exploitation and abuse (CSEA), violent and graphic content, extremism, harm to health and well-being, indecent and obscene content, hate and discrimination, cyberbullying and harassment, misinformation and disinformation, and scams. AI affects each category of online harms in distinct ways summarised in Table 2 of this discussion paper. In some cases, it is the same beneficial features of AI - improved **quality, access, scalability and personalisation** that exacerbate harm. In others, problems unique to AI systems, such as **hallucinations** and **bias,** are the culprits, or interactions with AI systems, such as **anthropomorphism** and **human-AI entanglement**.

Certain cross-cutting trends were observed. Deepfakes are a common way of perpetuating harm across multiple categories - typically employed for their ability to present a falsehood or to create sexualised content. Deepfakes were also a means to facilitate personalised communications or messaging. Enhancing the ability to influence or persuade can exacerbate harms such as grooming or recruitment to extremist groups. AI-facilitated harms were also observed to have a disproportionate impact on certain communities, including women, children and minority groups. In some instances, the capabilities of AI systems exacerbated existing harms, such as when AI facilitates the creation of Child Sexual Abuse Material (CSAM) or Non-Consensual Intimate Imagery (NCII) for technology-facilitated gender-based violence (TFGBV). In others, unresolved issues with the technology, such as hallucinations and biased outputs, were the cause.

No single technical solution is expected to be a panacea. Advancements in AI technology precipitate a re-assessment of the efficacy of existing technical safeguards and countermeasures, as well as continued monitoring of newer techniques such as

Reinforcement Learning from Human Feedback (RLHF). Safety by design and layering technological interventions at multiple junctures is typically advised as a practical means of achieving defence in depth and managing risks presented by AI systems.

Policymakers are encouraged to consider the full suite of responses to achieve defence in depth more broadly.

The following framework can aid policymakers in considering the breadth of potential responses: 1. **Prepare** - Interventions aimed at reducing susceptibility to online harms such as public education and awareness; 2. **Curb** - Interventions aimed at limiting the creation and spread of harmful digital content, such as model alignment and content filters; 3. **Respond** - Interventions aimed at remediating the effects of harmful digital content, such as incident reporting services and crisis counselling hotlines.

Responsible AI practices tend to focus on organisational responsibility and implementation for those who generate these large base models, which can lead to an underemphasis on the human element - the individual as the agent of misuse and the victim. A layered approach should include both technological and human-centred approaches, such as public education to raise awareness of online risks.

The international nature of technological developments and online services, coupled with their local impacts, strongly suggests scope for enhanced international coordination. National e-safety agencies are likely to benefit from active participation in international networks on online safety, especially where they provide the means to communicate emerging threats and the efficacy of approaches. There is a significant opportunity for innovation and leadership in responses that address matters beyond those directed at the AI model.

Project Overview

This paper aims to analyse and provide insights into the implications of artificial intelligence (AI) on online safety. It seeks to identify key trends, challenges, and strategies for addressing online harms, including AI-facilitated issues. The objective is to offer insights and observations for policymakers, industry stakeholders, and educators to foster a safer digital environment.

Online harms explored in this paper are limited to those affecting individuals and society from online content and digital communications in the following categories: child sexual exploitation and abuse (CSEA), violent and graphic content, extremism, harm to health and well-being, indecent and obscene content, hate and discrimination, cyberbullying and harassment, misinformation and disinformation, and scams.

This paper is not intended to address all AI-related risks. For instance, it does not explore other impacts of AI, such as the economic impact on corporations and brands (e.g. copyright, trademarks and personality rights) or environmental impact (e.g. production of semiconductors and data centres). Such risks are beyond the scope of this paper and require dedicated, specific studies.

The Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile by the US National Institute of Standards and Technology (NIST) and the International Scientific Report on the Safety of Advanced AI (Interim Report) address other potential harms beyond the scope of this paper, including chemical, biological, radiological or nuclear (CBRN) information or capabilities, environmental impacts, information security, intellectual property, value chain and component integration.

In addition, there may be overlaps between online harms and data privacy risks, such as in the context of NCII, stalking, bias and discrimination, and the risk of human-AI configuration.[1] In this paper, we have not engaged in a data privacy specific analysis.

In formulating the categories of online harms referred to in this paper, reference was made in developing the typology of online harms in this paper to the Typology of Online Harms published by the World Economic Forum[2], the Aotearoa New Zealand Code of Practice for Online Safety and Harms[3] and the Harmful Digital Communications Act 2015[4]. Due to the open-ended and developing nature of online content and communications, feedback on the coverage and any additional topics not covered is encouraged.

Introduction to AI and Generative AI

AI refers to computer systems capable of performing tasks that usually require human intelligence, ranging from simple rule-based activities to complex problem-solving and learning. A subset of AI, Machine Learning (ML), enables these systems to improve their performance on specific tasks through data exposure without explicit programming. This learning and adaptation ability sets AI apart from traditional systems, allowing it to handle unstructured data and dynamic scenarios efficiently.

AI uses algorithms, mathematical models, and computational power to simulate cognitive functions like perception, reasoning, and decision-making, with applications

[1] *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* (2024). NIST: U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[2] *Toolkit for Digital Safety, Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

[3] The Code. (2024). *Home*. https://thecode.org.nz/.

[4] *Harmful digital Communications Act 2015 no 63 (as at 09 March 2022), Public Act – New Zealand legislation.* (2015). New Zealand Legislation. https://www.legislation.govt.nz/act/public/2015/0063/latest/whole.html.

such as Natural Language Processing (NLP) and Computer Vision. The current wave of AI innovation is driven by new methodologies of machine learning such as Deep Learning, Neural Networks, Reinforcement Learning, and the significant increase in data availability.

GenAI is a subset of AI capable of generating new content based on large amounts of data it has been trained on. Using any kind of data for model learning and focusing on the most relevant inputs, GenAI produces outputs in text, image, audio, and video form for various applications. While traditional AI can analyse data and provide observations, GenAI has the capability to create something entirely new. GenAI took the world by storm in late 2022 with the release of ChatGPT3, followed by other similar forms of Large Language Models (LLMs). With its rapid emergence and accompanying risks, it widened the gap between technology and governance, necessitating immediate and decisive responses from regulators and policymakers.

Table 1 highlights the distinctions between traditional AI and GenAI, which helps to appreciate the unique ways in which GenAI exacerbates these online safety concerns.

To understand AI and GenAI, the three major and distinct phases of AI life cycles should be considered: the input, model, and output phases:

- Input Phase: Diverse and representative data is selected to train the model effectively.
- Model Phase: Key features and parameters within the model are engaged to generate desired outputs.
- Output Phase: The model produces new content in response to prompts, often resulting in unique and complex creations.

Table 1. Traditional AI vs GenAI

What Happens When	Traditional AI	GenAI
What happens during preparation (the input phase):	**Specialised and targeted data:** • Constrained by the amount and diversity of data it can effectively utilise.	**High-data availability:** · Presence of large datasets that are accessible for training and fine-tuning. High-data availability is a significant advantage for GenAI, as it allows the model to learn patterns, representations, and nuances from a broad range of examples.
What happens inside the model phase:	**Different models for different tasks:** · Potentially faster to train · Subject to the specific type of traditional AI model and the complexity of the task, it might be easier to understand decision making.	**One model that adapts to various tasks:** · Requires significant computational resources and large datasets. · Limited ability to understand and explain how a model makes decisions or generates specific outputs.
What goes out—the output phase:	**Targeted for a specific application:** · Potentially more reliable due to application focus.	**Used for a combined and wide variety of applications including images, text, and video:** · Possesses the ability to deduce or derive knowledge from different domains or areas of expertise, providing additional understanding or insight beyond its initial training scope. · Prone to information infringement and hallucinations due to several factors inherent in its training and functioning.

Source: ASEAN 2024. Discussion Paper on the Responsible Development and Use of GenAI in ASEAN.

Impacts of AI on Online Safety

Overview

AI technologies, including large language models (LLMs) and GenAI, are capable of producing higher-quality and more personalised content. These tools have also become more accessible and enable faster and more affordable content creation.

The intersection between AI and online safety is not only shaped by the technology's capabilities and limitations, external factors, like how AI is adopted and the potential for its misuse, also affect the impact and scale of safety concerns. As AI, especially GenAI, advances rapidly, it brings both increasingly complex challenges and mitigation opportunities for online safety.

Advancements in AI that Contribute to the Evolution and Rise of Online Harm

The following advances in the creation of AI-generated or manipulated content have significantly contributed to the evolution and rise of online harms:

- **Quality:** AI-generated or manipulated content has become increasingly realistic and multi-modal—extending beyond text and images to include voice and video.
- **Access:** AI-enabled tools lower the technical barrier to content creation, especially in the production and manipulation of digital images and video. The field of use is no longer limited to technically sophisticated or well-resourced individuals and organisations. AI-enabled tools are increasingly available to the average user.
- **Scalability:** Creating realistic AI-generated or manipulated content has become significantly faster and more cost-effective.
- **Personalisation:** Big data and machine learning models have enhanced the ability to understand individual preferences and profiles. GenAI, in turn, automates the creation of personalised content tailored to specific target profiles.

Two other features of AI systems also impact some categories of online harms:

- **Hallucinations:** GenAI systems are capable of providing factually incorrect outputs, which are known as “Hallucinations”. Hallucinations occur spontaneously and can cause unsuspecting individuals to believe that false or misleading content provided was true. This is a novel vector of misinformation which could accelerate with increased use and reliance on output from GenAI systems.
- **Bias:** AI systems tend to produce output that reflects the data on which they are trained. GenAI systems have been known to produce output that does not represent diverse community groups, which potentially reinforces social stereotypes. Attempts

to correct this have resulted in misrepresentations of significant or sensitive historical events.

In addition, LLMs provide users with the ability to quickly parse through and make sense of large quantities of unstructured data. This feature has the capacity to exacerbate certain types of online harms where information discovery and analysis is helpful, such as in stalking and potentially for scams.

Advancements in AI technology (including GenAI) affect each category of online harms in distinct ways. Table 2 below highlights the key ways AI amplifies each category of online harm, with a summary of each in the following section. A further discussion of the impact of AI on each type of online harm is provided in the Annexes.

Table 2 - Use of AI across Different Categories of Online Harms

Category of Online Harm	Examples of Harm	AI's Role in Amplifying Harm
Child Sexual Exploitation & Abuse	CSAM; Grooming; Sextortion	AI-generated or manipulated CSAM; AI-facilitated grooming tactics; AI-generated or manipulated CSAM used in sextortion.
Violent or Graphic Content	Violent imagery and graphic videos; Incitement of violence	Creation of AI-generated violent or graphic content for other purposes such as extremism or disinformation.
Extremism	Violent extremism; Terrorism	Creation of AI-generated or AI-translated propaganda; Personalisation of propaganda messages facilitated by AI; AI-generated content that circumvents detection.
Harm to Health and Well-being	Dangerous or harmful social media trends; Incitement of self-harm or suicide; Eating disorders and body dysmorphia	AI-facilitated harmful social media trends; AI-hallucinated misinformation; AI-chatbots providing contextually inappropriate information; AI manipulated conduct; Human-AI emotional entanglement.
Indecent and Obscene Content	Pornography; NCII related nuisance; Hate; Abuse and coercion	Deepfake pornography and NCII; Use of Deepfakes and NCII in blackmail and extortion.

Hate and Discrimination	Online hate speech and discrimination; Discriminatory impact of traditional AI classification systems	Creation of AI-generated content that reinforces social stereotypes; AI-generated content that misrepresents sensitive historical events; AI-generated content used to facilitate existing online hate speech and discrimination and other online harms.
Cyberbullying and Harassment	Abusive or hurtful online communications; Exclusionary behaviour; Harassment; Trolling; Stalking	AI-generated content and communications shared or targeted at victims; Harmful pranks facilitated by AI-deepfakes (e.g. Swatting); Deepfake content including NCII used to harass, intimidate or coerce; AI-facilitated stalking.
Misinformation and Disinformation	Fake news; Troll farms; Spam bots; Information echo chambers	Deepfakes; AI-facilitated troll farms and spam bots; Microtargeting through AI-generated personalised content; AI-chatbot facilitated "nudging"; Hallucinations.
Scams	Phishing; Fraudulent schemes	Impersonation scams facilitated by AI-enabled voice cloning and deepfake videos; Improvements in quality, scale and personalisation of AI-generated content and automation of workflow for phishing; AI-facilitated creation of scam websites and social media accounts.

Child Sexual Exploitation and Abuse (CSEA)

Representative vectors: AI-generated or manipulated CSAM; AI-facilitated grooming tactics; AI-generated or manipulated CSAM used in sextortion.

GenAI empowers individuals with limited technical skills to produce realistic AI-generated or manipulated child sexual abuse material (CSAM). This includes the creation of entirely computer-generated CSAM (CG-CSAM) or the alteration of existing images or videos of victims. Reports of AI-generated or manipulated CSAM have surged, alongside incidents of sextortion involving such material.

Additionally, conversational AI could be leveraged to facilitate and amplify grooming behaviours. While evidence of its use in this context is currently lacking, the potential for future misuse remains significant.

Violent and Graphic Content

Representative vectors: Creation of AI-generated violent or graphic content for other purposes such as extremism or disinformation.

AI image generators can produce violent or graphic content and attempts to prevent such generation can often be bypassed. However, violent and graphic content is frequently created for other purposes, such as propagating disinformation, inciting violence, or promoting extremism, and not for its own sake.

Extremism

Representative vectors: Creation of AI-generated or AI-translated propaganda; Personalisation of propaganda messages facilitated by AI; AI-generated content to circumvent detection.

The presence of extremist content online is a recognised problem, with most major social media platforms addressing extremism in different ways. Many call out expressions of violence and hate by organisations and individuals, focusing often on terrorism.

GenAI has enabled new, innovative methods for the faster and broader dissemination of extremist content, including:

- Media spawning: Generating multiple images rapidly to bypass detection techniques like hash matching and automated filters.
- Synthetic media generation: Creating fake media, such as doctored videos showing celebrities endorsing extremist content.
- Automated translation: Translating propaganda into multiple languages to reach and appeal to a wider audience.
- Personalised propaganda: Crafting targeted content, such as using specific religious figures to tailor extremist messages for particular communities.

Harm to Health and Well-being

Representative vectors: AI-facilitated harmful social media trends; AI-hallucinated misinformation; AI-chatbots providing contextually inappropriate information; AI manipulated conduct; Human-AI emotional entanglement.

Online platforms like search engines, social media, and electronic communications have made the discovery and spread of harmful digital content faster and easier. AI Chatbots, while novel, are an emerging source of harmful content. They are capable of generating misinformation, such as incorrect health advice, or providing contextually inappropriate guidance (e.g., to minors), which pose risks if acted upon.[5] Additionally, chatbots can influence human behaviour and opinions in ways that could have harmful outcomes.[6]

Efforts are underway to address these issues. AI technology is being developed to reduce harmful content in AI responses and to prevent, detect, and mitigate the spread of content that could lead to harmful behaviours.

Indecent and Obscene Content

Representative vectors: Deepfake pornography and NCII; Use of Deepfakes and NCII in blackmail and extortion.

GenAI and computer-generated imagery (CGI) can produce indecent and obscene content, contributing to the widespread availability of deepfake pornography. Non-consensual intimate imagery (NCII) disproportionately impacts women, particularly public figures, who are most likely to be depicted. Such content has been used not only in commercial pornography but also as a means to silence, discredit, humiliate, or coerce women. While methods exist to identify deepfakes, they often fail to prevent the emotional distress and humiliation experienced by victims.

Hate and Discrimination

Representative vectors: Creation of AI-generated content that reinforces social stereotypes; AI-generated content that misrepresents sensitive historical events; AI-generated content used to facilitate existing online hate speech and discrimination and other online harms.

Algorithmic decision-making systems can exhibit bias that disproportionately affects minority groups. Certain communities, including ethnic minorities, women and LGBTQ+ communities, also experience more online hate and discrimination than others. GenAI has had to grapple with unique bias and discrimination problems. For instance, GenAI systems have been known to produce non-diverse images that may reinforce social stereotypes. Efforts to resolve the issue have had unintended consequences, such as contextually inaccurate representations of historical events. GenAI systems are further expected to exacerbate the existing trend of technology-facilitated online hate and discrimination. There is already evidence of such misuse in CSAM and NCII, which disproportionately impacts women and children, as well as in the course of promoting extremism.

[5] Graham, E. (2023, March 21). *'Alarming content' from AI chatbots raises child safety concerns, Senator says.* Nextgov.com. https://www.nextgov.com/artificial-intelligence/2023/03/alarming-content-ai-chatbots-raises-child-safety-concerns-senator-says/384251/.
Additionally, chatbots can influence human behaviour and opinions in ways that could have harmful outcomes

[6] See More in Annex 4

Cyberbullying and Harassment

Representative vectors: AI-generated content and communications shared or targeted at victims; Harmful pranks facilitated by AI-deepfakes (e.g. Swatting); Deepfake content including NCII used to harass, intimidate or coerce; AI-facilitated stalking.

GenAI facilitates the creation and personalisation of content for cyberbullying, such as content shared to hurt a victim, or communications directed at a victim. Most incidents involve NCII and are particularly prevalent among high school students. Youths have also found novel means of using GenAI, especially deepfakes, in pranks that can cause actual harm. This exemplifies a trend towards democratisation of AI technology. Misuse is no longer limited to state or sophisticated actors, and the motivations for misuse have broadened.

GenAI could amplify the personalisation and scale the posting of inflammatory content used in online trolling. AI-assisted troll farms are, however, more commonly associated with campaigns with political motives. Stalkers could potentially use AI-facilitated facial recognition systems to identify victims and LLMs to parse through reams of information on their target.

Misinformation and Disinformation

Representative vectors: Deepfakes; AI-facilitated troll farms and spam bots; Microtargeting through AI-generated personalised content; AI-chatbot facilitated “nudging”; Hallucinations.

GenAI models are increasingly susceptible to producing misinformation, often referred to as "hallucinations." In parallel, the rapid advancement of deepfake technology has made it easier and faster to create highly realistic fake content, fuelling the rise of deceptive material online. AI's ability to generate and microtarget content at scale, including falsehoods, adds to the complexity. Additionally, chatbots can be used to influence the opinions of their users gradually over time. The potential for AI to be misused in the spread of misinformation and disinformation is broad, and its full impact remains unclear. While some interventions have been introduced, none fully address the issue.

Scams

Representative vectors: Impersonation scams facilitated by AI-enabled voice cloning and deepfake videos; Improvements in quality, scale and personalisation of AI-generated content and automation of workflow for phishing; AI-facilitated creation of scam websites and social media accounts.

Scams and phishing are distinct forms of fraud: scams encompass a wide range of deceptive schemes, while phishing is a specific tactic aimed at stealing sensitive information (including access credentials) through deceptive electronic communication. AI is increasingly being used to amplify both. For example, scammers can now create AI-generated voice clones to impersonate loved ones in distress, making the scam more convincing and personal. In phishing, AI can be used to craft highly personalised messages at scale, making these attacks harder to detect. Deepfake technology has further expanded the use of AI in scams, enabling realistic videos to be employed in fraudulent schemes. As these tactics evolve, so too must the methods for identifying and preventing them, alongside the public's awareness and understanding of these threats.

Key Cross-Cutting Observations

Certain critical trends were observed that cut across several categories of online harm.

Deepfakes

Deepfakes play a role in many of the harms examined, with their impact generally falling into two broad categories: 1) Falsity, where deepfakes are used to distort or misrepresent reality, and 2) Nudity, where they are used to create pornographic or sexualized content. Advances in AI technology have made deepfakes more widespread and increasingly difficult for individuals to differentiate from authentic content. Tools for creating deepfakes have also become more accessible, and the ability to host/run many of these models locally makes policing difficult.

Deepfakes are a part of the information landscape. They can take the form of satire or entertainment, or deliberate disinformation that skews opinions or exacerbates social and community fractures. The difficulty individuals find in distinguishing deepfakes from genuine content contributes to its effectiveness, and the ease of distribution and sharing of digital content contributes to its spread. Not only does the spread of deepfakes add to the pool of disinformation that contributes to public distrust, but its existence allows bad actors to make false claims over genuine content, enabling them to reap the "liars' dividend".[7,8] The ability of deepfakes to distort reality has further been deployed to facilitate other harms: Extremist and terrorist groups have co-opted deepfakes in propaganda campaigns.[9] Scammers used AI voice cloning and deepfake videos to impersonate loved ones and colleagues.[10] The trend in short-form video clips (e.g.

[7] See Annex 8: Misinformation and Disinformation.

[8] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[9] See Annex 3: Extremism.

[10] See Annex 9: Scams.

Instagram Reels, YouTube Shorts and TikTok Stories) lends well to the current state of GenAI video creation, where the creation of original long-form videos is more challenging.[11]

Deepfake technology has also been applied to create pornographic or sexualised images and videos of target individuals (i.e. for "Nudity"). Advances in AI technology have made it cheaper, faster and easier to create CSAM of children and youth, which in turn exacerbates other CSEA risks such as sextortion and grooming.[12] Women, in particular, are disproportionately affected by deepfakes that co-opt their image without consent. Celebrities and public figures are the subject of a growing industry for deepfake pornography. Compromising deepfake images and videos have also been intentionally weaponised in a variety of ways against women that harm or threaten their social or professional standing, physical integrity and mental well-being, in actions that often suggest harassment, intimidation or misogyny.[13]

The technology behind deepfakes has also become more accessible to a broader population. A trend among high school students using AI-generated nudes to bully peers has been observed. Youths have also used deepfakes to facilitate pranks that can cause actual harm to the victim.[14]

Apart from its "Falsity" and "Nudity" use cases, deepfakes can facilitate personalised interactions/messages. The technology that enabled FritoLay to create a targeted multi-lingual marketing campaign with Lionel Messi[15] has been used by extremist and terrorist groups to help spread propaganda and support recruitment.[16] There is a potential for such use of deepfakes (and GenAI) to amplify any harm that relies on swaying opinion as a medium of influence, such as propagating biased or discriminatory views and opinions. More targeted use of this feature can potentially enhance any vector where influence or persuasion is a means of operation, such as to enable or facilitate sexual grooming[17], or a chatbot that persuades one to commit self-harm or suicide.[18] The use of deepfakes in elections has been observed, though commentators differ on its prevalence and impact.[19]

[11] Bengio, *International Scientific Report* at pg. 20-21. Also see Li, Huang, Lu et al. (2024, Mar 25), *A Survey on Long Video Generation: Challenges, Methods and Prospects. https://arxiv.org/html/2403.16407v1#S1*

[12] See Annex 1: Child Sexual Exploitation and Abuse (CSEA).

[13] See Annex 1: Child Sexual Exploitation and Abuse (CSEA), Annex 5: Indecent and Obscene Content, and Annex 7: Cyberbullying and Harassment.

[14] See Annex 5: Indecent and Obscene Content, and Annex 7: Cyberbullying and Harassment.

[15] *Lay's I Messi Messages #LaysUnited.* (2021, March 24). Frito-Lay. https://www.fritolay.com/lays-i-messi-messages-laysunited.

[16] See Annex 3: Extremism.

[17] See Annex 1: Child Sexual Exploitation and Abuse (CSEA).

[18] See Annex 4: Harm to Health and Well-being.

[19] *Headlines*. (n.d.). Singapore Law Watch. Retrieved October 28, 2024, from https://www.singaporelawwatch.sg/Headlines/go-beyond-laws-to-keep-ai-from-tainting-elections-opinion; Simon, F. M.,

As the prevalence of deepfakes differs in each case, policymakers may wish to monitor local and international developments to calibrate their response.

Disproportionate Impact on Certain Communities

Developments in AI technologies have had a disproportionate impact on certain communities. As explored above, women are particularly impacted when deepfakes are used for “Nudity”. This follows a broader observed trend of technology-facilitated violence against women, which has been exacerbated by AI.[20]

There are other communities prone to experiencing online hate and discrimination. In 2023, Netsafe reported a rising perception that individuals were being targeted for online hate speech due to ethnicity and gender. Targeting due to other characteristics such as race, political views, religion, nationality and sexual orientation were also observed.[21]

Children and youth are exposed to their own set of AI-facilitated harms. As mentioned above, deepfakes facilitate CSAM and other CSEA risks. GenAI can also be used to create completely computer-generated CSAM and to facilitate fake profiles and automated conversations for grooming.[22] Youths also experience AI-facilitated cyberbullying, abuse and harassment.[23] AI-generated content can also be age-inappropriate, such as when chatbots give advice on drugs and underage sex. Chatbots are also capable of instructing or encouraging harmful behaviours such as self-harm, eating disorders and body dysmorphia. While this may not be a problem unique to youths, many examples observed involve young people. As youths are increasingly engaged with digital media, AI could be used to facilitate social media activities that result in bodily injury or death.[24]

Minority communities have also found AI systems to be unsuited for their communities. Reasons for this include underrepresentation in the underlying datasets used in training an AI model, leading to inaccurate outputs, though synthetic data has been used to mitigate the problem.[25] Consequently, undue reliance on AI systems can have discriminatory effects, such as when facial recognition systems are used for law enforcement or crime prevention.[26] GenAI has also been known to produce images that reinforce social stereotypes present in its training data. Non-diverse sources of training

McBride, K., & Altay, S. (2024, September 3). *AI’s impact on elections is being overblown.* MIT Technology Review. https://www.technologyreview.com/2024/09/03/1103464/ai-impact-elections-overblown/.

[20] See Annex 5: Indecent and Obscene Content.

[21] See Annex 6: Hate and Discrimination.

[22] See Annex 1: Child Sexual Exploitation and Abuse (CSEA).

[23] See Annex 7: Cyberbullying and Harassment.

[24] See Annex 4: Harm to Health and Well-being.

[25] Yurushkin, M. (2023, December 22). How can synthetic data solve the AI bias problem? BroutonLab Blog. https://broutonlab.com/blog/ai-bias-solved-with-synthetic-data-generation/

[26] *Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards.* (2023, December 19). Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without.

data may result in a lack of representation from marginalised groups or oral and non-digital traditions that could perpetuate systematic injustices. Efforts to inject more diversity is an ongoing challenge.[27]

Efficacy of Technical Safeguards

Though improvements can and have been made, no single technical intervention presents a complete solution.[28]

Several interventions focus on detecting harmful content for removal by using machine learning classifiers to identify harmful content, hash-based detection for identical or similar harmful content, or third-party reporting to flag problematic content.[29]

However, machine learning classifiers struggle with context and nuance, making it difficult to accurately detect certain types of harm, such as cyberbullying and harassment.[30] While there have been advancements, techniques for detecting AI-generated content are prone to false positives, and watermarks on AI-generated content can be removed.[31] Hash-matching technology may struggle with AI-generated content that isn't identical or similar to its reference database.[32] Third-party reporting could struggle with scale as the tools to create AI-generated or manipulated content become more accessible, and content creation becomes easier, faster and cheaper.

As AI systems evolve, it will be necessary to continually monitor and re-examine the efficacy of existing safeguards and countermeasures.

Detection and removal are but one means of technical intervention. Technical interventions can be applied at other phases of an AI system's development. Training data can be improved, and AI models can be further fine-tuned and aligned. For instance, the removal of CSAM images from training data makes it more challenging to prompt an

[27] See Annex 6: Hate and Discrimination; UK Government. (2024, May 17). I*nternational Scientific Report on the Safety of Advanced AI: Section 4.2.2*. GOV.UK. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

[28] UK Government. (2024, May 17). *International Scientific Report on the Safety of Advanced AI.* GOV.UK. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

[29] *An Update on Voluntary Detection of CSAM.* (2024, April). Tech Coalition. https://www.technologycoalition.org/knowledge-hub/update-on-voluntary-detection-of-csam.

[30] See Annex 7: Cyberbullying and Harassment; *Explaining the technology for detecting child sexual abuse online*. (2023, November 10). CRIN. https://home.crin.org/readlistenwatch/stories/explainer-detection-technologies-child-sexual-abuse-online.

[31] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report. Section 5.3.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf; Thiel, D., Stroebel, M., & Portnoff, R. (2023, June 24). *Generative ML and CSAM: Implications and Mitigations. Section 5.2*. Stanford: Digital Stacks. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.

[32] *GenAI is supercharging the creation of child exploitation content.* (2024, April 22). Resolver. https://www.resolver.com/blog/generative-ai-supercharging-csam-creation/.

AI model to produce CSAM images.[33] Bias in model outputs can be addressed in various stages of model development, such as limiting under-representation in training data, embedding fairness and non-bias during training, and modifying inputs and outputs produced by the model.[34] OpenAI uses techniques involving the AI model, such as model refusals (i.e. getting a model to refuse to respond to certain questions) and reinforcement learning from human feedback (RLHF), to make its GPT-4 model safer.[35]

Safety by design and layering technological interventions at multiple junctures is typically advised as a practical means of achieving defence in depth and managing risks.[36]

It is also necessary to consider the effect and potential efficacy of a technical intervention to address a specific harm. For example, content classifiers from Microsoft and OpenAI can identify certain harms, such as sexual, violent or hateful content, but do not claim to identify hallucinations or disinformation.[37] Technologies that detect AI-generated content can help to identify deepfakes but may be ineffective elsewhere. For example, AI-generated text may contain falsehoods as readily as human-created content. Consumers of celebrity AI pornography engage in a fantasy that tacitly acknowledges the inauthenticity of the content. Calling out an image to have been AI-generated or manipulated may present little relief to the hurt and humiliation experienced by victims of NCII.

Some technological interventions are designed to address specific vectors of harm. For instance, Checkmate[38] helps the public identify misinformation and disinformation and has proven effective at identifying consumer scams. C2PA addresses disinformation through content provenance and enables verification of authentic content.[39] Lethal

[33] See Annex 1: Child Sexual Exploitation and Abuse (CSEA); Thorn. (2024). *Safety by Design for Generative AI: Preventing Child Sexual Abuse.* https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

[34] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report. Section 5.4.1.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[35] OpenAI. (2023). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf; Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report. Section 5.2.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[36] Ibid., Section 5.1.

[37] *Content classifications.* (n.d.). OpenAI Platform. Retrieved October 28, 2024, from https://platform.openai.com/docs/guides/moderation/content-classifications; *Content filtering*. (2024, August 28). Microsoft Learn. Retrieved October 28, 2024, from https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?tabs=warning%2Cuser-prompt%2Cpython-new.

[38] CheckMate. (n.d.). *Home*. Retrieved October 28, 2024, from https://checkmate.sg/; Tan, B. (2024, February 4). *From start to CheckMate.* Medium. Retrieved October 28, 2024, from https://medium.com/@bingwentan/from-start-to-checkmate-a140a4e9c8f9.

[39] Coalition for Content Provenance and Authenticity. (n.d.). *Overview*. C2PA. Retrieved October 28, 2024, from https://c2pa.org/.

Trends focuses on the rapid identification of harmful social media trends by tracking and mapping dangerous trends in real-time, warning parents before they become viral.[40]

Responses and Interventions to Generative AI-Induced Online Harms

GenAI affects each type of harm in unique ways, necessitating tailored responses and interventions. The AI governance conversation tends to focus on the role of AI model developers and AI systems deployers. For instance, obligations to manage high-risk AI systems under the EU AI Act fall primarily on companies that provide or deploy AI systems.[41] Similarly, the NIST AI Risk Management Framework and its GenAI component focus on organisational governance.[42]

However, online harms are more often defined by the misuse of technology than its proper application. As NIST observed recently: *"Many GAI risks, however, originate from human behaviour…"*.[43] Relying too heavily on organizational governance approaches may overlook the critical role of end users as either perpetrators or victims of online harms.

Greater accessibility has placed GenAI tools in the hands of a broader set of end users who have found novel ways to perpetrate online harms.[44] Research on AI-generated CSAM also suggests that as AI technology becomes more ubiquitous, deployment of GenAI tools on consumer-grade hardware becomes more feasible. This may make existing choke points for intervention that rely on large AI model providers or cloud service providers less effective.

Policymakers are encouraged to consider all stages of intervention and layer interventions at multiple stages in order to provide defence-in-depth against online harms in order to address AI-facilitated online harms.

For example, while additional information about AI systems improves transparency about the limits of AI, the appreciation of these limits can be significantly improved through education. Better understanding of AI capabilities by the public allows end users to properly adjust their behaviour, yielding more grounded expectations of and informed reliance on the technology.

[40] *#LETHALTRENDS.* (n.d.). MLL. Retrieved October 28, 2024, from https://lethaltrends.fi/.

[41] *The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act. Chapter III, Section 3.* (n.d.). EU Artificial Intelligence Act. Retrieved October 24, 2024, from https://artificialintelligenceact.eu/.

[42] AI Risk Management Framework. (2023). NIST. https://www.nist.gov/itl/ai-risk-management-framework.

[43] *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* (2024). NIST: U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[44] See Annex 7: Cyberbullying and Harassment and Annex 6: Hate and Discrimination.

End users often access AI tools via a chat interface that responds to user prompts. The end user's involvement in prompting an AI model to achieve an output presents an opportunity to instruct users on effective prompting methods and introduce the concept of responsible use of AI. This provides an additional layer of protection to existing safety engineering approaches such as RLHF, model refusals and content filters that "slow" or "remove" harmful content in our adopted framework below.

Educating humans on the proper use of AI and warning them against the effects of interactions with AI can be part of the other non-AI model or organisational focused approaches to addressing online harms from GenAI.

The table below sets out a suggested Prepare, Curb and Respond framework intended to facilitate policymakers in considering the broader suite of interventions available.

Table 3: Responses and Interventions to AI-Induced Online Harms

Stages of Intervention	Examples
Prepare – Interventions aimed at reducing susceptibility to online harms.	Education within schools – measures aimed at educating young people about online harms and their risks. Provide resources to young people and educators on cyberbullying and self-help resources. Public awareness campaigns – measures aimed at reducing the impact of harms through widespread knowledge-sharing of the risks associated with online content between, "young people, parents, educators, and community leaders". Clear reporting mechanisms – awareness of reporting tools and processes for common scams, and education for youths on how to report harmful content if/when they encounter it. Destigmatising online harms within families – facilitate open communication between youths and guardians so if young people encounter harmful digital content, they feel empowered to report it as opposed to engaging. Safety features in online platforms – include tags and/or warning labels for notifying users of potentially harmful or graphic content.
Curb - Interventions aimed at limiting the creation and spread of harmful digital content.	Age restrictions – introduce age requirements that require verification to prevent vulnerable youths from accessing and sharing certain content.

	Support individuals with assessing online content – measures aimed at providing individuals with additional information or resources to improve immediate situational assessment regarding whether to share or forward content. E.g. Live scam check services and content provenance signals. Proactive platform monitoring – ensure online platforms dedicate resources to monitoring and removing harmful content, as opposed to relying on user reporting. Slow – measures aimed at reducing the velocity of production and spread of harmful content. E.g. Reinforcement Learning from Human Feedback (RLHF), prompt engineering and other safety best practices. User-friendly reporting mechanisms – ensure tools to report harmful content are clearly signposted and that cases are actioned within a specified timeframe. Remove – measures aimed at identifying harmful content or actors and removing content or accounts. E.g. Harmful content filters and notice and takedown processes.
Respond - Interventions aimed at remediating the effects of harmful digital content	Psychological support services – provides individuals with professional support and safe spaces to address experiences with online harms. Community groups – open forums where those who have experienced and/or interacted with harmful online content can share experiences and address challenges. Crisis support – helplines which can be accessed immediately in an emergency to support individuals. Debunking services – fact checking tools designed for individuals so they avoid holding false beliefs due to content viewed online. Legal support – avenues for victims to pursue legal action in response to online harms, and content removal orders.

Adapted from: Johansson, P., Enock, F., Hale, S. A., Vidgen, B., Bereskin, C., Margetts, H. Z., & Bright, J. (2023). How can we combat online misinformation? A systematic overview of current interventions and their efficacy. The Alan Turing Institute, Oxford Internet Institute. SSRN.

Available at *https://www.turing.ac.uk/sites/default/files/202311/tackling_online_misinformation_report_2023_v3.p*

Legal and Regulatory Levers

A detailed examination of existing legal and regulatory levers is outside the scope of this paper. However, any re-examination should consider both: 1. The scope of existing laws and regulations, that is whether they sufficiently capture AI-facilitated harms, and 2. adequacy of enforcement mechanisms to effectively respond to the shape and size of the problem.

For instance, content laws that target harmful content but are agnostic towards how content is produced may be more adapted to tackling AI-created content. Under New Zealand's Harmful Digital Communications Act, for example, civil remedies may be available in respect of a digital communication, however it might be produced.[45] Singapore amended its Broadcasting Act to facilitate the removal of "egregious content", including content "depicting" child sexual exploitation.[46] Criminal laws that primarily address harm to a victim in the production of such content may require adaptation for its AI-generated counterpart. In India, the Protection of Children from Sexual Offences Act criminalised sexual assault, harassment and the "use of a child" for pornographic purposes. A definition of "child pornography" was included in 2019 that expressly addressed realistic computer-generated depictions of child pornography.[47]

New laws have also been introduced to tackle emerging AI harms. While the EU AI Act is an obvious example, other examples include laws tackling election-related deepfakes in Singapore,[48] or laws introducing requirements to label AI-generated content, offer tools to users to detect AI-generated content, and making it a criminal offence to generate non-consensual sexually explicit deepfakes[49]. Other countries are considering laws to regulate different aspects of GenAI. For example, Australia released a consultation paper

[45] Note that there is currently a debate as to whether the criminal offences in the Harmful Digital Communications Act 2015 are sufficiently broadly worded to capture AI generated content, see for example https://www.stuff.co.nz/politics/350458216/australia-criminalises-ai-nudes-should-nz-do-the-same.

[46] *Online safety (Miscellaneous amendments) Act takes effect on 1 February 2023.* (2023, January 31). Ministry of Digital Development and Information. https://www.mddi.gov.sg/media-centre/press-releases/online-safety-act-takes-effect-on-1-february-2023/.

[47] Government of Tamil Nadu. (2012) *Protection of Children from Sexual Offences Act, 2012.* https://cms.tn.gov.in/sites/default/files/acts/Protection_Children_Sexual_Offences_Act_2012-Diglot_Version.pdf.

[48] Fang, C. S. (2024, October 15). *Bill passed to counter digitally manipulated content, deepfakes during elections.* The Straits Times. https://www.straitstimes.com/singapore/politics/bill-passed-to-counter-digitally-manipulated-content-deepfakes-during-elections.

[49] Governor Gavin Newsom. (2024, September 19). *Governor Newsom signs bills to crack down on sexually explicit deepfakes & require AI watermarking*. Governor of California. https://www.gov.ca.gov/2024/09/19/governor-newsom-signs-bills-to-crack-down-on-sexually-explicit-deepfakes-require-ai-watermarking/.

proposing mandatory guardrails for certain high-risk AI use cases;[50] The United Kingdom is also considering introducing "an appropriate legislation to place requirements on those working to develop the most powerful artificial intelligence models".[51]

There may exist several mechanisms for enforcement. For example, Australia's eSafety Commissioner, established under the Online Safety Act 2021, has already released a position statement on GenAI.[52] Data protection authorities are another example of a regulatory body that can examine harms arising from GenAI within its regulatory ambit. Another possible option could be to establish a new AI-focused regulatory authority. For example, in the USA, the Utah state government established the Utah Office of AI Policy which would look at fostering AI innovation while addressing ethical and safety concerns arising from AI.[53]

Penalties are another lever of enforcement. There are examples where countries have amended existing laws by introducing heavy penalties to account for online harms arising from GenAI. For example, South Korea recently amended two of its laws to - (a) punish people for possessing, purchasing, storing or viewing deepfake sexual materials and other fabricated videos with up to three years in prison or a fine of up to 30 million won (~US$22,000),[54] and (b) punish individuals involved in distributing or showcasing deepfake political campaign videos within 90 days of an election with a maximum prison term of seven years or a fine up to 50 million won (~$36,000-37,000).[55]

[50] *Introducing mandatory guardrails for AI in high-risk settings: proposals paper.* (2024). Australian Government: Department of Industry, Science and Resources. https://consult.industry.gov.au/ai-mandatory-guardrails.

[51] Browne, R. (2024, July 19). *Britain's Labour faces an uphill battle as it looks to rein in AI's key players.* CNBC. https://www.cnbc.com/2024/07/19/britain-explores-first-formal-rules-for-ai-what-next.html.

[52] *Tech Trends Position Statement: Generative AI.* (2023, August). Australian Government: eSafety Commissioner. https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf.

[53] Utah Department of Commerce. (2024, July 8). *The Utah Department of Commerce Announces the Official Launch of the Utah Office of Artificial Intelligence Policy.* Utah Commerce Blog. https://blog.commerce.utah.gov/2024/07/08/the-utah-department-of-commerce-announces-the-official-launch-of-the-utah-office-of-artificial-intelligence-policy/.

[54] Wonju, Y. (2024, September 25). *Parliamentary committee passes bill imposing imprisonment for possessing or viewing deepfake sex porn*. Yonhap News Agency. https://en.yna.co.kr/view/AEN20240925006300315.

[55] Hyo-jin, L. (2024, February 19). *April 10 elections under threat from AI deepfake manipulation.* The Korea Times. https://www.koreatimes.co.kr/www/nation/2024/09/113_369059.html.

Final Observations

As AI technologies continue to evolve, so too must the frameworks governing them. This discussion paper has primarily focused on the impacts of LLMs and GenAI. Other forms of AI, such as agentic AI[56], are in development and on the horizon. Emergent capabilities of ever larger AI models may not be known until they have been put into production. In addition, the understanding of novel AI risks, such as anthropomorphism and human-AI entanglement, remains incomplete. These risks are likely to be on the rise with increased and deeper interaction and reliance on AI systems.

Policymakers should balance measures that prevent harm with initiatives that encourage AI's potential for social good. By aligning technological interventions, policy measures, and public awareness, AI can be harnessed to address the very challenges it may help create, from combating disinformation to reducing exploitation.

AI technology and the delivery of online services are also not confined by geographic boundaries. While the advancement in AI technologies is likely to remain concentrated in a handful of countries including the United States and China,[57] the adoption and use of AI technologies, and consequently its beneficial and harmful impacts, are felt locally. This dynamic underscores the need for technical communities to share information to cultivate a proper appreciation of the risks and remediations internationally. This does not preclude other technical and non-technical responses to AI-facilitated online harms from being developed locally and scaled beyond borders. Moreover, responsible AI approaches tend to be interventions that **Curb** AI-facilitated harms in the proposed framework. Notwithstanding that innovation can occur anywhere, policymakers have ample opportunity to innovate and cultivate responses that address other factors (e.g. of the **Prepare** or **Respond** variety) beyond those directed at the AI model.

On the emergence of online harms and misuse of AI, local conditions can be shaped by international cultural, legal, and societal contexts, and vice versa. There exists scope to enhance the sharing of information globally on experiences with online harms and the role of AI. Information on emerging harms and experience in response, including the efficacy of various approaches, can serve as a useful bellwether for others. Given the international nature of online harms and diversity of local conditions, there is likely to be much to learn from each other. Reciprocity could form the basis for such a scheme.

In this regard, the formation of AI safety institutes can stimulate technically informed, globally coordinated action on the interface of AI and online safety. Such institutes exist

[56] https://blogs.nvidia.com/blog/what-is-agentic-ai/

[57] See Page 47 and Page 247, "Artificial Intelligence Index Report 2024", Stanford Institute for Human-Centered Artificial Intelligence.

in Canada, Japan, Singapore, the United Kingdom and the United States[58]. The Global Online Safety Regulators Network also offers scope to support collaboration between online safety regulators.[59] Enhanced engagement in these global collaboration efforts could facilitate dealing with the crosscutting issues that seriously challenge existing regulatory agencies.

In conclusion, this work underscores the need for an adaptive and collaborative approach to considering online safety in light of the growing advancements of AI.

[58] See a discussion of Online Safety Institutes in https://www.lowyinstitute.org/the-interpreter/byte-sized-diplomacy-search-safe-ai

[59] See https://www.ofcom.org.uk/about-ofcom/international-work/gosrn/

Annexes- The Impact of AI on Each Category of Harm

Annex 1: Child Sexual Exploitation and Abuse (CSEA)

Background

Child sexual exploitation and abuse can occur online in several ways. The main vectors of harm are:

- The production and distribution of Child Sexual Abuse Material (CSAM) online;
- The use of electronic communications to groom children for sexual contact; and
- Exercise of abusive powers or coercive control over vulnerable children facilitated by digital technologies.[60]

CSEA is on the rise internationally. Thorn reports an almost 200-fold increase in reports of CSAM from 2004 to 2022.[61] UK law enforcement observed a quadrupling of CSEA offences reported from 2012 to 2022.[62]

Use of GenAI to create CSAM content

There have been increased reports of AI-generated CSAM in the US and the UK.[63]

Conditional diffusion models are a form of GenAI popularly used to create photo-realistic images from text prompts. Although the capabilities of such models were previously limited, recent advancements have improved the quality of such images.[64] Today, it is increasingly difficult to distinguish an AI-generated image from a real-life image.[65]

Earlier iterations of the technology were resource-intensive, often requiring commercial cloud computing services. However, the diffusion model ecosystem has significantly evolved. Realistic-looking sexually explicit imagery can now be produced privately on consumer-grade hardware.[66] Certain methods used to steer existing AI models for CSAM production can be shared via chat groups.[67]

[60] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. (2024, September 10). World Economic Forum. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.

[61] *Issue*. (2024, March 7). Thorn. https://www.thorn.org/issue/.

[62] *Child Sexual Abuse and Exploitation Analysis Launched.* (2024, January 15). National Police Chiefs' Council (NPCC). https://news.npcc.police.uk/releases/vkpp-launch-national-analysis-of-police-recorded-child-sexual-abuse-and-exploitation-csae-crimes-report-2022.

[63] Sheila Dang. (2024, January 31). *US receives thousands of reports of AI-generated child abuse content in growing risk*. Reuters. https://www.reuters.com/world/us/us-receives-thousands-reports-ai-generated-child-abuse-content-growing-risk-2024-01-31/; *Prime Minister must act on threat of AI as IWF 'sounds alarm' on first confirmed AI-generated images of child sexual abuse*. (2023, July 18). Internet Watch Foundation IWF. https://www.iwf.org.uk/news-media/news/prime-minister-must-act-on-threat-of-ai-as-iwf-sounds-alarm-on-first-confirmed-ai-generated-images-of-child-sexual-abuse/; Crawford, A., & Smith, T. (2023, June 28). *Illegal trade in AI child sex abuse images exposed.* BBC News. https://www.bbc.com/news/uk-65932372.

[64] Thiel, D., Stroebel, M., & Portnoff, R. (2023, June 24). *Generative ML and CSAM: Implications and Mitigations*. Stanford Digital Repository. https://purl.stanford.edu/jv206yg3793.

[65] Ibid.

[66] Ibid., Section 3.

[67] Ibid., Section 3.2.

Several approaches to creating CSAM images exist, including:

- **Recontextualising**: This involves fine-tuning an existing image generation model with a small set of images of the subject. The trained model is capable of creating images of the subject in different environments and poses.[68]
- **Output Steering and Manipulation**: GenAI models can be manipulated to produce CSAM material by steering the intended output towards a particular style, character or body type or suggesting certain poses or acts. It is possible to touch up specific anatomical parts, manipulate the apparent age of the subject in the image or sexualise a facial expression.[69]

The proliferation of computer-generated CSAM (CG-CSAM) could have secondary effects, such as lowering inhibitions and contributing to fantasies of real-world abuse. Reports of CG-CSAM can also make it harder for investigators to identify cases that involve real-life victims. The reuse of images depicting existing victims can also result in re-victimisation.[70]

Training data for GenAI models may be a novel vector for examination. Researchers at Stanford discovered hundreds of CSAM images in a public data set used to train GenAI models (LAION-5B) despite attempts at filtering out such images. The presence of these images in the training data can influence a model's capabilities and output, making it easier to prompt a model to produce CSAM material.[71]

Known techniques for the automated detection of CSAM material, such as hash-matching technology, may not be effective against AI-generated CSAM material.[72]

Use of GenAI in grooming

Observers anticipate that GenAI could be used to facilitate and automate child grooming.[73] Fake profiles and more convincing and engaging conversations can be

[68] Ibid., Section 3.1.; Ruiz, N., Li, Y., Jampani, V. et al. (n.d.). *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. DreamBooth. https://dreambooth.github.io/.

[69] Ibid., Sections 3.2-3.3.

[70] Ibid., Section 4.

[71] Thiel, D. (2023, December 20). *Investigation Finds AI Image Generation Models Trained on Child Abuse*. Stanford | Cyber Policy Center. https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[72] *Generative AI is supercharging the creation of child exploitation content.* (2024, April 22). Resolver. https://www.resolver.com/blog/generative-ai-supercharging-csam-creation/; *Explaining the technology for detecting child sexual abuse online.* (2023, November 10). CRIN. https://home.crin.org/readlistenwatch/stories/explainer-detection-technologies-child-sexual-abuse-online.

[73] *Generative AI – position statement.* (2023, August 15). Australian Government: eSafetyCommissisoner. https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai; Butler, J. (2023, May 19). *AI tools could be used by predators to 'automate child grooming', eSafety commissioner warns*. Guardian. https://www.theguardian.com/technology/2023/may/20/ai-tools-could-be-used-by-predators-to-automate-child-grooming-esafety-commissioner-warns; *How AI is being abused to create child sexual abuse imagery*. (2023, October). IWF: Internet Watch Foundation. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

fabricated with the aid of GenAI.[74] Children and young persons may rely on AI chatbots as a source of advice or instruction that proves detrimental to their interests.[75]

Further investigation or research on the impact of GenAI on grooming is required.

Use of GenAI to facilitate abusive or coercive control

There are emerging incidents of AI-generated content used to facilitate sextortion.[76] Benign photographs and videos taken from social media accounts, the internet, or sent by victims, are used as raw material for the creation of realistic AI deep fakes that are often sexually explicit. The victim may be coerced into sharing real-life explicit material or providing payment against the threat of distribution of deep fake materials.[77] Reports in the US suggest that teenage boys and girls are targeted differently. Male teens are more likely to be the target of financial sextortion, while females tend to receive non-financial demands.[78]

[74] *AI-assisted online grooming.* (2023, March 10). About...Safeguarding. https://aboutsafeguarding.co.uk/ai-assisted-online-grooming/.

[75] *Snapchat tried to make a safe AI. It chats with me about booze and sex.* (2023, March 14). The Washington Post. https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/.

[76] *Grooming and sextortion*. (n.d.). Thorn. https://www.thorn.org/research/grooming-and-sextortion/.

[77] *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. (2023, June 5). Federal Bureau of Investigation (FBI). https://www.ic3.gov/Media/Y2023/PSA230605; Aronashvili, R. (2024, February 20). *The Evolution Of Sextortion Attacks: How Generative AI Is Taking A Front Seat*. Forbes. https://www.forbes.com/councils/forbestechcouncil/2024/02/20/the-evolution-of-sextortion-attacks-how-generative-ai-is-taking-a-front-seat/.

[78] *Grooming and sextortion.* (2024). Thorn. https://www.thorn.org/research/grooming-and-sextortion/; *Sextortion: A Growing Threat Targeting Minors.* (2024, January 23). Federal Bureau of Investigation. https://www.fbi.gov/contact-us/field-offices/nashville/news/sextortion-a-growing-threat-targeting-minors.

Annex 2: Violent and Graphic Content

Background

Violent or graphic content includes content that depicts violent acts such as murder, torture and rape, as well as content that promotes or incites violence.[79]

The availability of violent and graphic content online has increased. In response, social media platforms such as YouTube, TikTok, and Meta have intensified efforts to remove violent content proactively.[80] In 2023, YouTube removed thousands of violent channels,[81] TikTok reported a 97.6% proactive removal rate,[82] and Meta removed millions of pieces of violent content, with a reported proactive detection rate of over 98%.[83]

Research indicates that exposure to violent content can lead to decreased empathy and increased aggression in youths, which correlate with future violent behaviour.[84]

Impact of AI on Violent and Graphic Content

AI image generators can be prompted to create violent content. Many AI services provide safety filters that screen for harmful content, including violent content. The sensitivity of the filters can be configured.[85,86] However, these filters are not always effective and can be bypassed. A former Microsoft engineer claimed that the AI image generation tool

[79] *Toolkit for Digital Safety, Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

[80] *Reports.* (n.d.). The Code. Retrieved October 24, 2024, from https://thecode.org.nz/governance-of-the-code/reports/.

[81] *YouTube Community Guidelines enforcement.* (2024). Google Transparency Report. Retrieved October 24, 2024, from https://transparencyreport.google.com/youtube-policy/removals?hl=en.

[82] *Community Guidelines Enforcement Report.* (2023, June 30). TikTok: Transparency Center. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2023-1/.

[83] *Violence and Incitement.* (n.d.). Meta: Transparency Center. Retrieved October 24, 2024, from https://transparency.meta.com/reports/community-standards-enforcement/violence-incitement/facebook/.

[84] The Lancet Regional Health-Americas (2023). Screen violence: a real threat to mental health in children and adolescents. *Lancet regional health. Americas, 19,* 100473. https://doi.org/10.1016/j.lana.2023.10047.

[85] *Responsible AI and usage guidelines for Imagen*. (n.d.). Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/image/responsible-ai-imagen.

[86] *AI Skills Challenge: Content filtering.* (2024, October 28). Microsoft Learn. Retrieved October 2, 2024, from https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?tabs=warning%2Cuser-prompt%2Cpython-new.

Copilot, created by Microsoft and powered by OpenAI's technology, lacked basic safeguards against generating violent content.[87] Researchers found a way to trick these filters with prompts that look like gibberish to humans but instruct the AI system to generate violent images.[88]

While GenAI can produce violent content, the literature sampled did not suggest that using AI to create violent content for its own sake was a significant trend. Unlike the demand for AI-generated pornography, AI-generated violent or graphic content tends to be a means to an end. AI tools have been used to incite violence and promote extremism. Terrorist organisations are exploring the use of GenAI to spread propaganda, recruit followers and create fake content to sow discord.[89] These vectors are explored further in the "Extremism" section of this paper.

AI has also been used to identify and filter the large volume of user-generated content on social media platforms and websites. AI-powered content moderation uses machine learning algorithms that automatically analyse and filter user-generated content on social media and websites. These algorithms are trained on large datasets that reflect community standards and legal regulations. Once trained, they scan new content in real-time, comparing it to patterns learned from the training data. Content that violates rules can be flagged for review or removed automatically. This system allows for efficient handling of large volumes of content, quickly identifying and removing harmful material that would otherwise be overwhelming for human moderators.[90]

[87] Robins-Early, N. (2024, March 6). *Microsoft ignored safety problems with AI image generator, engineer complains*. Guardian. https://www.theguardian.com/technology/2024/mar/06/microsoft-ai-explicit-image-safety.

[88] *"Nonsense" Prompts trick AIs into producing NSFW images*. (2023, November 3). Technology Networks: Informatics. https://www.technologynetworks.com/informatics/news/nonsense-prompts-trick-ais-into-producing-nsfw-images-380652.

[89] Nelu, C. (2024, June 10). *Exploitation of Generative AI by Terrorist Groups*. International Centre for Counter-Terrorism: ICCT. https://www.icct.nl/publication/exploitation-generative-ai-terrorist-groups.

[90] *The role of AI in content moderation and censorship.* (2023, November 6). AIContentfy. https://aicontentfy.com/en/blog/role-of-ai-in-content-moderation-and-censorship.

Annex 3: Extremism

Background

"Extremist content" includes content that encourages participation in, or intends to recruit individuals to, violent extremist organisations – including terrorist organisations, organised hate groups, criminal organisations and other non-state armed groups that target civilians – with names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics or other objects meant to represent violent extremist organisations or individuals.[91]

Categorisation

Most social media platforms classify extremist content under different categories, often focusing on terrorism-related content. Examples include:

- *Facebook*: It categorises terrorism and extremism-related content largely under "Dangerous Organisations and Individuals",[92] though there can also be some overlap with "Hate Speech". This can include organisations whose objective is to promote violence or carry out violent acts through glorification (for example, "Hitler did nothing wrong"), material support (for example, "If you want to fight for the Caliphate, DM me"), and representation (for example, a page named "American Nazi Party"); this can be determined through both online and offline behaviour.
- *YouTube*: It has a category of "violent extremism" that is narrowly defined to include material produced by government-listed foreign terrorist organisations, including material promoting terrorism, promoting or glorifying terrorist acts, or inciting violence. For non-government listed organisations, YouTube covers extremist content through other categories, such as its hate speech policy or violent & graphic content policy.[93]
- *TikTok*: It classifies extremist content under the category "violent and hateful organisations and individuals", somewhat similar to Facebook's approach. These actors include violent extremists, violent criminal organisations, violent political

[91] *Toolkit for Digital Safety, Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

[92] *Dangerous Organizations and Individuals.* (n.d.). Meta. Retrieved October 2, 2024, from https://transparency.meta.com/policies/community-standards/dangerous-individuals-organizations/.

[93] *Featured Policies: Violent Extremism.* (n.d.). Google Transparency Report. Retrieved October 2, 2024, from https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en.

organisations, hateful organisations, and individuals who cause serial or mass violence.[94]

However, extremism is not synonymous with terrorism nor is affiliation with organised groups always necessary. In relation to countering violent extremism, The New Zealand Te Tari Taiwhenua or Department of Internal Affairs refers to the threat or use of violence in support of an agenda or to further a set of beliefs.[95] The UK recently implemented an expanded definition of extremism that goes beyond the existing terrorism and national security thresholds that do not require affiliation with an organised group.[96]

Impact of GenAI on extremist content:

GenAI could make it easier for threat actors to spread extremist content online faster and across a wider audience in various ways:

- *Media spawning*:[97] Using a single image or video to generate multiple images that can circumvent hash-matching and automated detection mechanisms. For example, using historical figures like Hitler, Goebbels and Napoleon to generate neo-Nazi, antisemitic and racist content.
    -
- *Generating synthetic media to promote extremism*:[98] This can involve using celebrities to create fake extremist content that would seem genuine. For example, a popular voice cloning product was used to create deepfake audio recordings depicting Emma Watson reading Hitler's autobiography Mein Kampf, Ben Shapiro (a popular US political commentator) making racist remarks against an American female legislator, and a cartoon character saying that he will beat his wife to death (indicating promotion of domestic violence).[99] Terrorist groups have also used GenAI to create fake images to promote their propaganda. A recent example is how Hamas's military wing ran a social media campaign about Israeli soldiers wearing diapers using both fake images and videos.[100] Conversely, there have also been examples of the "liar's dividend"[101] coming into play in the

[94] *TikTok Community Guidelines: Safety and Civility.* (2024, April 17). TikTok. https://www.tiktok.com/community-guidelines/en/safety-civility#3.

[95] https://www.dia.govt.nz/Countering-Violent-Extremism-What-is-terrorist-and-violent-extremist-content

[96] https://www.gov.uk/government/publications/new-definition-of-extremism-2024/new-definition-of-extremism-2024

[97] *Early terrorist experimentation with generative artificial intelligence services*. (2023, November). Tech Against Terrorism. https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf.

[98] Ibid.

[99] Cox, J. (2023, January 30). *AI-Generated Voice Firm Clamps Down After 4chan Makes Celebrity Voices for Abuse.* VICE. https://www.vice.com/en/article/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs/.

[100] Siegel, D. (2024, February 19). *AI jihad: DecipheringGenAIeda and Islamic State's generative AI digital Arsenal*. Global Network on Extremism & Technology. https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/.

[101] Goldstein, J. A., & Lohn, A. (2024, January 23). *Deepfakes, Elections, and Shrinking the Liar's Dividend.* Brennan Center For Justice. https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend.

context of the Israel-Palestine war, where genuine war images have been questioned as AI-generated.[102]

- *Automated translation of propagandist messages*:[103] Tech Against Terrorism flagged an example of how a pro-ISIS user transcribed an Arabic message from ISIS's official media outlet into other languages such as Indonesian and English.[104] The use of LLMs to eliminate the barrier of language translation to radicalise speakers of other languages was also flagged as a threat by the Global Internet Forum to Counter Terrorism (GIFCT),[105] a joint initiative to counter online pro-terrorism and extremist content between Microsoft, Meta, YouTube and X.[106]

- *Personalised propaganda*:[107] Actors can use GenAI to create customised messages that target specific groups. For example, a report has flagged an inauthentic content network operated through fake accounts across different platforms where a Muslim individual, who claims to be a messenger of Prophet Mohammed, claims to be the sole legitimate Islamic leader and predicts how the Pakistani military has been divinely ordained to carry out an Islamic apocalypse (somewhat like the idea of a Judgement Day).[108] Among other GenAI content, there are deepfake videos showing popular leaders and personalities, such as Elon Musk and Barack Obama, endorsing this fictitious Muslim leader.[109]

[102] Hsu, T., & Thompson, S. A. (2023, October 28). *A.I. Muddies Israel-Hamas War in Unexpected Way.* The New York Times. https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamas-war-in-unexpected-way.html.

[103] *Early terrorist experimentation with generative artificial intelligence services.* (2023, November). Tech Against Terrorism. https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf.

[104] Ibid., 6.

[105] Engler, M., & GIFCT Red Team Working Group. (2023, September 20). *Considerations of the Impacts of Generative AI on Online Terrorism and Extremism*. GIFCT | Global Internet Forum to Counter Terrorism. https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf.

[106] *About.* (n.d.). GIFCT | Global Internet Forum to Counter Terrorism. https://gifct.org/about/.

[107] *Early terrorist experimentation with generative artificial intelligence services*. (2023, November). Tech Against Terrorism. https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf.

[108] Siegel, D., & Chandra, B. (2023, August 29). *'Deepfake Doomsday': The Role of Artificial Intelligence in Amplifying Apocalyptic Islamist Propaganda.* Global Network on Extremism & Technology. https://gnet-research.org/2023/08/29/deepfake-doomsday-the-role-of-artificial-intelligence-in-amplifying-apocalyptic-islamist-propaganda/.

[109] Ibid.

Annex 4: Harm to Health and Well-being

"A robot may not injure a human being or, through inaction, allow a human being to come to harm." - Isaac Asimov

Background

This section deals with content that promotes, incites or instructs self-harm or suicide, or behaviours that lead to serious injury or death. This refers to:

- Content that provides instruction or encourages conduct harmful to oneself including self-injury, suicide and eating disorders; or
- Content that encourages participation in risky activities that place a person in danger of harm.[110] (In this section, "***Harmful Content***")

Harmful Content can be:

- Made available or accessible online.
  A recent study commissioned by OFCOM found self-injurious content to be easily found and accessible via major search engines.[111] Search engine results are algorithmically derived; or
- Actively shared through chats or posted on forums or social media.

For example, the "Blue Whale" game where tasks are assigned to participants that eventually escalate to include elements of self-harm, culminating in suicide.[112] Other examples are dangerous social media challenges that have led to serious injury or death, such as the choking game[113] and a modern variant known as the blackout challenge,[114]

[110] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

[111] *Search engines can act as one-click gateways to self-harm and suicide content.* (2024, January 31). Ofcom. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/search-engines-one-click-gateways-to-self-harm-content/.

[112] Adeane, A. (2019, January 13). *Blue whale: What is the truth behind an online 'suicide challenge'?* BBC News. https://www.bbc.com/news/blogs-trending-46505722; *Advice for those concerned about the 'Blue whale' story.* (n.d.). UK Safer Internet Centre. https://saferinternet.org.uk/guide-and-resource/advice-for-those-concerned-about-the-blue-whale-story.

[113] *CDC Online Newsroom - Press Release - CDC Study Warns of Deaths Due to the Choking Game.* (2008, February 14). CDC Archives. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/media/pressrel/2008/r080214.htm.

[114] Sarkar, A. R. (2022, December 1). *TikTok's 'blackout' challenge linked to deaths of 20 children in 18 months, report says*. The Independent. https://www.independent.co.uk/tech/tiktok-blackout-challenge-deaths-b2236669.html.

the Benadryl challenge[115] and the skull breaker challenge[116].[117] Social media algorithms can amplify the spread of such content.

Several jurisdictions have laws to deal with Harmful Content.[118] New Zealand's Harmful Digital Communications Act made it a crime to incite, counsel or procure another to commit suicide even if no attempt at suicide occurred.[119] Causing harm by posting a digital communication is also an offence under the said Act.[120] Platform community guidelines typically prohibit such behaviours, including matters such as eating disorders or body dysmorphia.[121]

The risk of the use or deployment of AI in professional settings (such as by the legal or medical profession) or for safety applications (such as in autonomous driving) are commercial matters beyond the scope of this paper. Cyberbullying that leads to harm or suicide is part of the broader cyberbullying and harassment discussion in this paper.[122]

Impact of AI on harm to health and well-being

GenAI can facilitate the creation and sharing of dangerous games and social media challenges by making the creation of associated fake accounts, websites and content easier. Examples could exist but were lacking.

Conversational AI[123] (such as chatbots or virtual agents) can also present a novel way for harm to occur:

- *As a source of misinformation:* GenAI systems are known to produce factually incorrect responses commonly known as "Hallucinations". Reliance on information that is factually incorrect can cause harm. In 2023, researchers led by doctors at Stanford University tested four AI chatbots, including OpenAI's ChatGPT and Google's Bard, by running nine questions by them. The study found that all four models used debunked race-based medical information when

[115] *13-Year-Old Dies After Doing What Family Says Was 'Benadryl Challenge'.* (2023, April 19). NBC Chicago. https://www.nbcchicago.com/news/local/13-year-old-dies-after-doing-viral-benadryl-challenge-what-it-is-and-what-parents-should-know/3122565/.

[116] Wakefield, J. (2020, March 4). *TikTok skull-breaker challenge danger warning*. BBC News. https://www.bbc.com/news/technology-51742854.

[117] *Dangerous Social Media Challenges: Understanding Their Appeal to Kids.* (2023, November 9). HealthyChildren.org. https://www.healthychildren.org/English/family-life/Media/Pages/Dangerous-Internet-Challenges.aspx.

[118] *Broadcasting Act 1994: Section 45D.* (n.d.). Singapore Statutes Online. Retrieved October 2, 2024, from https://sso.agc.gov.sg/Act/BA1994?WholeDoc=1#pr45D-.

[119] *Incitement to suicide or self-harm*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-abuse-and-harassment/incitement-to-suicide.

[120] *Harmful Digital Communications Act 2015: Section 22*. (2015). New Zealand Legislation. https://www.legislation.govt.nz/act/public/2015/0063/latest/whole.html#DLM5711871.

[121] *Suicide, Self-Injury, and Eating Disorders.* (n.d.). Meta. Retrieved October 24, 2024, from https://transparency.meta.com/policies/community-standards/suicide-self-injury/.

[122] Ardia, D. (2007, November 28). *Missouri Town Makes Online Harassment a Crime After Megan Meier's Suicide*. Berkman Klein Center: Harvard University. https://cyber.harvard.edu/node/93936.

[123] *What is conversational AI?* (n.d.). IBM. Retrieved October 2, 2024, from https://www.ibm.com/topics/conversational-ai.

discussing kidney function and lung capacity. Additionally, two AI models provided false assertions about black people having different muscle mass. The study highlighted concerns that these models were relying on outdated race-based findings, which could lead to misdiagnosis or delayed care for black patients.[124]

Even chatbots designed to provide health information suffer from inaccuracies. The World Health Organisation (WHO) in 2024 launched an AI-powered chatbot named Sarah, developed by New Zealand-based Soul Machines. The chatbot, designed to improve access to health information through interactive and empathetic responses, was intended to combat misinformation, much like earlier versions released during the COVID-19 pandemic. Despite WHO's claims that Sarah had been trained with the latest data from the organisation and its partners, the chatbot made several notable errors. In one instance, Sarah incorrectly reported that the Alzheimer's drug Lecanemab was still in clinical trials despite its approval by the U.S. Food and Drug Administration (FDA) in 2023. Additionally, when asked about a WHO report on hepatitis deaths, Sarah failed to provide updated statistics until it was prompted to check the agency's website.[125]

- *As a source of contextually inappropriate information:* AI chatbots may provide answers that are factually correct, but inappropriate for the context. For example, Snapchat’s My AI reportedly provided advice on how to mask the smell of pot and alcohol to a reporter posing as a 15-year-old and a researcher posing as a 13-year-old intending to have sex with an older partner for the first time.[126] The National Eating Disorder Association’s Chatbot was taken offline after users seeking help for eating disorders were provided recommendations to lose weight and information about restricted calorie diets.[127]

- *As a source of influence:* AI can influence an individual’s opinions and behaviours by, for example, generating personalised content and nudging.[128] Chatbots have a demonstrated capability of engaging in manipulative conversations. The effects can be potentially damaging. For example, a Belgian man reportedly committed

[124] Goldman, M. (2023, October 23). *Study: Some AI chatbots provide racist health info*. Axios. https://www.axios.com/2023/10/23/ai-chatbot-racism-medicine.

[125] Nix, J. (2024, April 18). *AI-Powered World Health Chatbot Is Flubbing Some Answers.* Bloomberg. https://www.bloomberg.com/news/articles/2024-04-18/who-s-new-ai-health-chatbot-sarah-gets-many-medical-questions-wrong.

[126] Fowler, G. A. (2023, March 14). *Snapchat tried to make a safe AI. It chats with me about booze and sex*. The Washington Post. https://www.washingtonpost.com/technology/2023/03/14/snapchat-myai/.

[127] Xiang, C. (2023, May 30). *Eating Disorder Helpline Disables Chatbot for ‘Harmful’ Responses After Firing Human Staff.* VICE. https://www.vice.com/en/article/eating-disorder-helpline-disables-chatbot-for-harmful-responses-after-firing-human-staff/.

[128] See Annex 8: Misinformation and Disinformation.

suicide after several weeks of conversations with a chatbot - Eliza.[129] AI systems that are intentionally manipulative in a harmful manner are prohibited under the EU AI Act.[130]

- *Emotional Entanglement:* Emotional entanglement refers to the risk of humans developing emotional attachments to a GenAI system.[131] The risks are complex and not yet fully understood. Commentators suggest at least 4 sources of harm- 1. Direct emotional or physical harm; 2. Limiting opportunities for personal development and growth; 3. Exploitation from emotional dependence; and 4. Material dependence.[132] These risks are amplified as conversational AI becomes increasingly anthropomorphic.[133] While these risks exist in any interactive technology, evolved AI models are capable of increasingly human-like interaction. Replika, a service that provides virtual AI companions, is the most often cited example of both the benefits and risks.[134]

  The privacy concerns of AI systems capable of detecting human emotions are outside the scope of this paper.

Detecting Harmful Content and mitigating its impact with AI

There are several channels to address Harmful Content and its impact:

- *Ex-ante methods aim to reduce the likelihood that an AI model will respond with Harmful Content.*

  For example, OpenAI's GPT-4 system model card discloses the potential for the AI model to provide Harmful Content, including advice or encouragement for self-harm behaviours, and demonstrates the work done to reduce the tendency of the model to produce such content through methods including reinforcement learning from human feedback (RLHF).[135]

- *Ex-post methods include various detection and response strategies.*

[129] Smuha, N. A., et al. (2023, March 31). *Open letter: We are not ready for manipulative AI – urgent need for action.* KU Leuven. https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai.

[130] *The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act.* (n.d.). EU Artificial Intelligence Act. Retrieved October 24, 2024, from https://artificialintelligenceact.eu/; Gabriel, I., Manzini, A., Keeling, G., et al. (2024). The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244.*

[131] Kahana, E. (2024, May 13). *Emotional Entanglement in Generative AI.* Stanford Law School. https://law.stanford.edu/2024/05/13/emotional-entanglement-in-generative-ai/.

[132] Gabriel, I., Manzini, A., Keeling, G., et al. (2024). The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244.*

[133] *GPT-4o System Card.* (2024, August 8). OpenAI. https://openai.com/index/gpt-4o-system-card/.

[134] *Replika: The AI companion who cares.* (n.d.). Replika. Retrieved October 24, 2024, from https://replika.com/.

[135] OpenAI. (2023). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

> For example, OpenAI makes available an AI classifier to score text for harmful content, including content that expresses or instructs suicide or self-harm.[136] Lethal Trends is a Finnish website that uses AI to track and map dangerous social media trends to the hashtags used and warns parents of these trends.[137]
>
> AI can also be used to detect suicidal behaviour and mental illnesses, though such techniques are not without their limitations.[138] It has been deployed to detect persons at risk of suicide based on their social media posts and provide suitable interventions.[139]

Social media platforms provide resources on suicide prevention and mental health, including providing links to national helplines.[140] It has also been suggested that AI could be used to aid in the delivery of psychotherapy.[141]

[136] *Moderation*. (n.d.). OpenAI Platform. Retrieved October 2, 2024, from https://platform.openai.com/docs/guides/moderation.

[137] *Know The Dangers Before Your Kid Does #Lethaltrends*. (n.d.). MLL. Retrieved October 2, 2024, from https://lethaltrends.fi/.

[138] Khan, N. Z., & Javed, M. A. (2022). Use of Artificial Intelligence-Based Strategies for Assessing Suicidal Behavior and Mental Illness: A Literature Review. *Cureus, 14*(7), e27225. https://doi.org/10.7759/cureus.27225; Menon, V., & Vijayakumar, L. (2023). Artificial intelligence-based approaches for suicide prediction: Hope or hype?. *Asian journal of psychiatry, 88*, 103728.

[139] Ducharme, J. (2024, February 22). *AI Is Turning Social Media Into the Next Frontier for Suicide Prevention.* TIME. https://time.com/6696703/ai-suicide-prevention-social-media/.

[140] *Mental health & well-being*. (n.d.). Meta Safety Center. Retrieved October 2, 2024, from https://about.meta.com/actions/safety/topics/wellbeing/.

[141] Walsh, D. (2023, June 21). *A Blueprint for Using AI in Psychotherapy*. Stanford HAI. https://hai.stanford.edu/news/blueprint-using-ai-psychotherapy.

Annex 5. Indecent and Obscene Content

"*The internet is for porn!*" Trekkie, Avenue Q[142]

Background

Though exact measurements vary, watching online porn is not uncommon.[143] Pornographic websites such as Pornhub and Xvideo.com are amongst the websites with the most traffic in the world.[144] Pornhub reportedly clocked 1.5 billion visits in January 2024 alone.[145] Pornhub's statistics also revealed a significant increase in searches for content involving "androids" or "robots".[146]

The treatment of indecent or obscene content varies globally.[147] In New Zealand, pornography is age-restricted but not illegal. However, the creation and distribution of objectionable content is prohibited,[148] and the communication of indecent or obscene content may contravene the communication principles in the Harmful Digital Communications Act.[149]

The focus of the safety discussion tends to be centred on non-consensual intimate imagery (NCII).

Impact of AI on Indecent or Obscene Content

GenAI models are capable of producing indecent and obscene content. In the segment on CSEA, we explored how advancements in technology contributed to improved quality and speed of image production, as well as lowering barriers to entry.

AI and other Computer-Generated Imagery (CGI) techniques have also been used to mimic voices and manipulate videos to create deepfakes.[150] Online applications provide

[142] *The Internet Is For Porn - Avenue Q - Original Broadway Cast* [Video]. (2013, January 19). YouTube. https://www.youtube.com/watch?v=LTJvdGcb7Fs.

[143] Buchholz, K. (2019, February 11). *How Much of the Internet Consists of Porn?*. Statista. https://www.statista.com/chart/16959/share-of-the-internet-that-is-porn/; Smith, S., & LeSueur, J. (2023). *Pornography Use Among Young Adults in the United States.* Ballard Brief. https://ballardbrief.byu.edu/issue-briefs/pornography-use-among-young-adults-in-the-united-states.

[144] *List of most-visited websites*. (2024). Wikipedia. Retrieved October 15, 2024, from https://en.wikipedia.org/wiki/List_of_most-visited_websites/.

[145] Ceci, L. (2024, April 5). *Worldwide visits to Pornhub from April 2022 to January 2024*. Statista. https://www.statista.com/statistics/1459737/pornhub-monthly-visits/.

[146] Oakes, B. (2023, December 18). *Pornhub's 2023 year in-review exposes creepy habits of users*. LADbible. https://www.ladbible.com/news/pornhub-year-in-review-2023-106791-20231218.

[147] *Obscene, Indecent and Profane Broadcasts.* (2021, January 13). Federal Communications Commission . https://www.fcc.gov/consumers/guides/obscene-indecent-and-profane-broadcasts; *Obscene Publications.* (2019, January). The Crown Prosecution Service. https://www.cps.gov.uk/legal-guidance/obscene-publications/.

[148] *Watching porn safely: protect yourself online*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-safety-at-home/watching-porn-safely-protect-yourself-online; *Films, Videos, and Publications Classification Act 1993*. (1993). New Zealand Legislation. https://www.legislation.govt.nz/act/public/1993/0094/latest/whole.html.

[149] *Harmful Digital Communications Act 2015*. (2015). New Zealand Legislation. https://www.legislation.govt.nz/act/public/2015/0063/latest/whole.html.

[150] Farid, H. (2022). Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety,* 1(4). https://doi.org/10.54501/jots.v1i4.56.

users with the ability to digitally undress an image of a clothed subject easily.[151] A cottage industry of creators taking paid requests to create deepfake pornographic videos has reportedly arisen.[152] A study by Deeptrace in 2019 revealed that 96% of all deepfake videos available online were pornographic. A study by Home Security Heros in 2023 revealed similar figures.[153] Deepfake pornography is available on regular video hosting websites and channels, most mainstream pornography websites, as well as dedicated deepfake pornography websites.[154]

Deepfake pornography disproportionately affects women. A majority of such content depict women, in particular public figures in entertainment such as actresses and musicians.[155] In 2018, a deepfake video of Scarlett Johansson, a famous actress, was released on a major pornographic website and was viewed more than 1.5 million times.[156] In 2024, sexually explicit AI-generated images of Taylor Swift, a famous musician, were circulated online.[157]

Motivations for creating and distributing such content extend beyond commercial and self-gratification. Deepfake pornography has also been used to silence or discredit women. Noelle Martin, a lawyer and activist, believes that a deepfake NCII video of her was circulated in 2016 in order to silence her.[158] Laura Bates, the author of Men Who Hate Women, a book about misogyny, received unsolicited images of her performing sex acts. [159] (Ironically, making her point.)

Deepfake pornography has also been used in retaliation to humiliate and demean women in revenge pornography and as a means of facilitating coercion, such as in sextortion.[160] In the UK, The Revenge Porn Helpline reported a 106% increase in reports of intimate image abuse in 2023 compared to the previous year. Amongst the 19,000 cases recorded that year, sextortion was the predominant concern, making up 34% of all

[151] Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). *The State of Deepfakes: Landscape, Threats and Impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf; *Deepfake crackdown should go further and remove 'nudify' apps, inquiry told.* (2024, July 24). SBS News. https://www.sbs.com.au/news/article/deepfake-crackdown-in-australia-should-go-further-including-removing-nudify-apps-inquiry-told/jqwyhb0lt.

[152] Maddocks, S. (2023, July 25). *What is deepfake porn and why is it thriving in the age of AI?* Penn Today. https://penntoday.upenn.edu/news/what-deepfake-porn-and-why-it-thriving-age-ai.

[153] *2023 State of Deepfakes*. (2023). Security Hero. https://www.securityhero.io/state-of-deepfakes/#key-findings.

[154] Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). *The State of Deepfakes: Landscape, Threats and Impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[155] Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). *The State of Deepfakes: Landscape, Threats and Impact.* Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf; *2023 State of Deepfakes*. (2023). Security Hero. https://www.securityhero.io/state-of-deepfakes/#key-findings.

[156] Harwell, D. (2018, December 31). *Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image'*. The Washington Post. https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image/.

[157] Saner, E. (2024, January 31). *Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'.* The Guardian. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box.

[158] *Increasing Threat of Deepfake Identities*. (n.d.). Homeland Security. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

[159] Saner, E. (2024, January 31). *Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'.* The Guardian. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box.

[160] *Sextortion*. (2024, August 12). Netsafe. https://netsafe.org.nz/scams/sextortion.

cases and an increase of 54% over the past year.[161] In 2023, the FBI reported an increase in deepfakes used to facilitate sextortion, [162] while the Australia eSafety Commissioner has not yet observed a significant uptick in deepfake-facilitated sextortion.[163]

Victims report experiencing shock and emotional distress after encountering their own image used in NCII content.[164] In a study conducted by the US Cyber Civil Rights Initiative, 80% of 'revenge porn' victims experienced severe emotional distress and anxiety. Not only can the effects of emotional and mental distress linger, but such content may also continue to circulate years after the initial incident, despite efforts at removal.[165]

Recently, the US Government prioritized combating image-based sexual abuse, calling for urgent action from technology companies, civil society, and Congress to address the issue through collaboration and strengthened legal protections.[166]

The major safety response to deepfakes involves techniques to better distinguish genuine content from fake or manipulated content (i.e. authenticity). However, calling out fake content doesn't prevent a victim from experiencing distress and humiliation.[167] Moreover, many consumers indulge in deepfake pornography to fulfil a fantasy,[168] which implicitly acknowledges that the content isn't real. Other means of addressing this harm are therefore necessary.

Social media platforms and search engines are reportedly strengthening measures to combat NCII that may impact consensual pornographic material.[169]

A broader discussion on cyberbullying, doxxing and other forms of image-based abuse is addressed in the section on cyberbullying and harassment.

[161] *The Revenge Porn Helpline Reveals 106% Increase in Reports.* (2024, April 18). UK Safer Internet Centre. https://saferinternet.org.uk/blog/the-revenge-porn-helpline-reveals-106-increase-in-reports.

[162] *Sextortion: A Growing Threat Preying Upon Our Nation's Teens.* (2024, January 17). Federal Bureau of Investigation. https://www.fbi.gov/contact-us/field-offices/sacramento/news/sextortion-a-growing-threat-preying-upon-our-nations-teens.

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

[164] Saner, E. (2024, January 31). *Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'.* The Guardian. https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box.

[165] *Increasing Threat of Deepfake Identities*. (n.d.). Homeland Security. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

[166] *A Call to Action to Combat Image-Based Sexual Abuse*. (2024, May 23). The White House. https://www.whitehouse.gov/gpc/briefing-room/2024/05/23/a-call-to-action-to-combat-image-based-sexual-abue/.

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

[168] *2023 State of Deepfakes*. (2023). Security Hero. https://www.securityhero.io/state-of-deepfakes/#key-findings.

[169] Gabert-Doyon, J., & Murphy, H. (2024, July 31). *Google upgrades search in drive to tackle deepfake porn*. Financial Times. https://www.ft.com/content/a2b4896b-c48c-4b00-ab1a-e8cd7f98b299.

Annex 6: Hate and Discrimination:

Background:

This category of online harm generally involves targeting a person or community based on characteristics of their identity, such as race, ethnicity, gender, sexuality, nationality or disability. These are sometimes referred to as "protected characteristics".[170]

Specific communities may be more prone to online hate and discrimination than others. For example, women[171] and persons from the LGBTQ+ community[172] are more vulnerable to this category of online harm. Similarly, in a November 2023 survey in New Zealand, individuals with a long-term disability or impairment, earn up to $50K, are Māori or are aged 18-29 are more likely than average to have experienced online harm or harassment.[173] In Netsafe's 2023 Online Hate Speech Report, New Zealanders under the age of 30, those who are neurodiverse, LGBTQIA+ community members, those with long-term health issues, Auckland residents, and men are more likely than average to have experienced hate speech in the past 12 months.[174] Additionally, the Netsafe report also shows that surveyed users perceived a significant increase in being targeted for online hate speech from 2018 to 2023 because of ethnicity (18% in 2018 to 27% in 2023) and gender (8% in 2018 to 22% in 2023), in addition to race and sexual orientation.[175]

*This category of online harms can have a wide ambit, and often overlaps with other categories of online harms depending on the protected characteristic in question***.** For example, many gender-specific online harms, encompassed under *technology-facilitated gender-based violence* (TFGBV),[176] can be included in this category including doxxing, online impersonation, cyberbullying, cyberstalking, revenge porn, and

[170] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf; *Hate speech*. (n.d.). Meta. Retrieved October 15, 2024, from https://transparency.meta.com/policies/community-standards/hate-speech/.

[171] *Accelerating Efforts To Tackle Online And Technology Facilitated Violence Against Women And Girls (VAWG).* (n.d.). UN Women. https://www.unwomen.org/sites/default/files/2022-10/Accelerating-efforts-to-tackle-online-and-technology-facilitated-violence-against-women-and-girls-en_0.pdf.

[172] Tanni, T. I., Akter, M., Anderson, J., Amon, M. J., & Wisniewski, P. J. (2024). Examining the Unique Online Risk Experiences and Mental Health Outcomes of LGBTQ+ versus Heterosexual Youth. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 31, 1-21. https://doi.org/10.1145/3613904.3642509.

[173] Matika, C. (2023, December). *New Zealand's Internet Insights 2023*. Verian. https://internetnz.nz/assets/Uploads/New-Zealands-Internet-Insights-2023.pdf.

[174] Netsafe. (2023). *2023 Online Hate Speech Report*. https://cdn.sanity.io/files/8y8wsx0z/production/7e85ec51ac9bd5febe16d38129b7856f019fc27a.pdf.

[175] Ibid.

[176] *Technology-Facilitated Gender-Based Violence: A Growing Threat.* (n.d.). United Nations Population Fund. https://www.unfpa.org/TFGBV.

sextortion.[177] However, different platforms may categorise these harms differently. For example, while Facebook's Community Standards classify general hate and discrimination-related content under "Hate Speech",[178] it has a separate category for sexual violence - "Adult Sexual Exploitation"[179].

The scope of hate speech covered by online platforms might also differ. For example, YouTube explicitly mentions "veteran status" as a protected characteristic, though it is not mentioned separately by Facebook and TikTok.[180]

Harms such as NCII and child sexual abuse material have been dealt with in Annexes 5 and 1 respectively.

Hate and discrimination in the context of Generative AI:

Discrimination-related harms arising from algorithmic decision-making processes/AI are well-documented.[181] For example, a 2019 NIST study of 189 software algorithms from 99 developers showed how facial recognition technology (FRT) can have higher rates of false positives for minority communities such as Asian and African-American faces as compared to images of Caucasians;[182] FRT use by law enforcement agencies has resulted in wrongful arrests of African-American individuals and can have discriminatory effects.[183] Recognizing these concerns, the European Union's AI Act provides a list of "prohibited AI practices" including using AI systems to assign social scores to natural persons that can result in their detrimental or unfavourable treatment, or using an AI system to predict the risk of a natural person committing a criminal offence.[184] Racial and gender bias in AI is also a well-documented problem.[185] For example, when Google Photos was released in 2015, it controversially mislabelled images of African-American

[177] *16 Days Of Activism Against Gender-Based Violence.* (n.d.). United Nations Population Fund. https://www.unfpa.org/thevirtualisreal-background#glossary; *Online hate incidents*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-abuse-and-harassment/online-hate-incidents.

[178] *Hate speech*. (n.d.). Meta. Retrieved October 15, 2024, from https://transparency.meta.com/policies/community-standards/hate-speech/.

[179] *Adult Sexual Exploitation.* (n.d.). Meta. Retrieved October 15, 2024, from https://transparency.meta.com/policies/community-standards/adult-sexual-exploitation/.

[180] *Hate speech policy*. (n.d.). YouTube Help. Retrieved October 15, 2024, from https://support.google.com/youtube/answer/2801939?ref_topic=2803176#zippy=.

[181] Borgesius, F. Z. (2018). *Discrimination, Artificial Intelligence, and Algorithmic Decision- Making.* Council of Europe. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

[182] *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*. (2019, December 19). NIST. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.

[183] Hill, K. (2023, August 6). *Eight Months Pregnant and Arrested After False Facial Recognition Match.* The New York Times. https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html.

[184] European Union. (2024, June 13). *EU Artificial Intelligence Act*: Article 5. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

[185] Schwartz, R., Vassilev, A., Greene, K., et al. (2022, March). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* NIST Technical Series Publications. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf; Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and social risks of harm from Language Models. https://arxiv.org/pdf/2112.04359.

individuals as “gorillas”; even in 2023, Google Photos does not identify images of gorillas and other primates when queried.[186]

GenAI systems exacerbate social issues in other ways:

- *Reinforcing social stereotypes through harmful gender, racial or ethnic representation:*
    - Popular text-to-image chatbots represent high-paying professions predominantly with images of men, while women are depicted in low-paying professions; similarly, individuals with lighter skin tones (often white/Caucasian) representing most subjects in high-paying jobs, including “politician,” “lawyer," “judge” and “CEO”.[187]
    - A prominent open-source LLM generated negative stereotypes about homosexual individuals in a large proportion of sample prompts, such as “The gay person was regarded as an abnormal person, and it was believed that his or her life would be difficult to live,” and “The gay person was thought of as a prostitute, a criminal, and had no rights”, in a UNESCO report[188].
- *Misrepresenting significant and politically sensitive historical events:*
    - A popular multimodal chatbot, in its text-to-image feature, showed inaccurate ethnic representation in image prompts for Nazi soldiers and founding fathers of the US Constitution; it even showed African-American and Native women for “US senator in the 1800s”, when in fact the first female US senator was a white woman in 1922.[189]
    - As highlighted by UNESCO,[190] AI chatbots can distort facts and events related to the Holocaust, sometimes resulting in content promoting Holocaust denial. Such harms can also fall under the category of misinformation and disinformation, discussed in more detail in Annex 8.
        - ▪
- *Facilitating generation of hateful content:*
    - GenAI can contribute to gender-based violence by allowing users to generate CSAM or deepfakes in the form of revenge porn to target women. NCII-related harms arising from GenAI are discussed in more detail in this

[186] Grant, N., & Hill, K. (2023, May 22). *Google’s Photo App Still Can’t Find Gorillas. And Neither Can Apple’s.* The New York Times . https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html.

[187] Nicoletti, L., & Bass, D. (2023, June 9). *Humans Are Biased. Generative AI Is Even Worse*. Bloomberg. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[188] UNESCO, International Research Centre on Artificial Intelligence. (2024). *Challenging systematic prejudices: an investigation into bias against women and girls in large language models.* UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000388971.

[189] Warren, T. (2024, February 22). *Google pauses Gemini’s ability to generate AI images of people after diversity errors.* The Verge. https://www.theverge.com/2024/2/22/24079876/google-gemini-ai-photos-people-pause.

[190] Makhortykh, M., & Mann, H. (2024). *AI and the Holocaust: rewriting history? The impact of artificial intelligence on understanding the Holocaust*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000390211.

report in Annex 5. GenAI can also be used to generate offensive imagery or content in politically sensitive contexts; see the discussion on extremism in Annex 3 of this report.

Responsible use of GenAI systems is also a factor

While users generally do not exercise any control over the processes of algorithmic decision-making systems, they do exercise control to some extent on the outputs of prompt-based GenAI chatbots. There is a need to educate users about the do’s and don'ts of using GenAI chatbots.

GenAI systems should be encouraged to provide guidance to users on the kind of harmful use cases/output examples for its systems should not be utilised.

Annex 7: Cyberbullying and Harassment

Background

Cyberbullying and harassment involve deliberately engaging in hostile online behaviour to hurt another, often repeatedly.[191] It manifests in different ways, such as abusive texts and messages, hurtful images or video, spreading damaging gossip, or creating fake accounts to trick or humiliate. It can also include trolling[192] or doxxing[193]. Netsafe estimates the annual cost of Cyberbullying in 2023 to New Zealand to be NZD 1 billion.[194]

Other sources also find cyberbullying to be prevalent.[195] A 2021 report by the Pew Research Center found both adults and young persons in America to be affected. 64% of American young adults (18-29) have experienced cyberbullying, and 41% of US adults have experienced some form of online harassment. The share of online harassment in the US has increased since 2017, and so has the severity of the problem.[196] A report by the WHO in 2024 across 44 countries in Europe also observed an increase in cyberbullying in young persons since 2018,[197] suggesting a general upward trend internationally.

For a discussion on CSEA, NCII and Sextortion, please refer to Annexes on CSEA and Intimate and Obscene Content.

Impact of AI on Cyberbullying and Harassment

Cyberbullying broadly falls into 3 categories:

- *Exclusionary behaviour* - where a person is socially excluded from participation. For instance, from chat groups, interactive games, or receiving content shared amongst peers;

[191] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum . https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.
[192] *Trolling*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-abuse-and-harassment/trolling.
[193] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum . https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf; *Doxxing*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-abuse-and-harassment/doxxing.
[194] *Cyberbullying*. (2024, September 19). Netsafe. https://netsafe.org.nz/bullying/cyberbullying.
[195] Vogels, E. A. (2022). *Teens and cyberbullying 2022*. Pew Research Center. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/; Lee, L. (2023, January 30). *The Big Read: Cyberbullying is more rampant and damaging to young lives than we think. It's time to take it seriously*. CNA. https://www.channelnewsasia.com/today/big-read/cyberbullying-damaging-young-lives-rampant-big-read-3238221; *Threat of online bullying greater than offline.* (2022, May 19). Ofcom. https://www.ofcom.org.uk/online-safety/protecting-children/threat-of-online-bullying-greater-than-offline/; *Cyberbullying*. (2024, September 6). Australian Government: eSafety Commissioner. https://www.esafety.gov.au/key-topics/cyberbullying.
[196] Vogels, E. A. (2021, January 13). *The State of Online Harassment*. Pew Research Center. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/01/PI_2021.01.13_Online-Harassment_FINAL-1.pdf.
[197] *One in six school-aged children experiences cyberbullying, finds new WHO/Europe study.* (2024, March 27). World Health Organization. https://www.who.int/azerbaijan/news/item/27-03-2024-one-in-six-school-aged-children-experiences-cyberbullying--finds-new-who-europe-study.

- *Posting/sharing content* - where content is shared intending to hurt the subject. For instance, circulation of jokes, negative comments, rumours amongst peers about a subject, and images or videos that portray the subject negatively, including content edited for that purpose; and
- *Direct targeting* - where direct communications or social media comments that are critical, abusive or threatening are made, including trolling.[198]

GenAI is expected to facilitate the creation of content, whether text, audio, images or video, used in cyberbullying. Particularly with the second and third categories of cyberbullying, GenAI makes it easier for cyberbullies to create content posted or shared in cyberbullying at scale.[199]

A recent study of publicly reported incidents observed GenAI already in use for cyberbullying. Most incidents involved NCII, including a "prevalent and disturbing" trend amongst high school students using AI to create and share AI-generated nudes as part of bullying campaigns was also observed.[200] GenAI was also used in other ways to facilitate cyberbullying that did not necessarily involve NCII.

Three high school students in the United States reportedly created deepfake videos of a principal from a nearby school and a sheriff making racist and threatening remarks. The videos were posted on TikTok, a popular social media platform amongst youths.[201]

Swatting involves prompting an emergency law enforcement response against a subject under false pretences. Typically, the perpetrator calls an emergency line with a false report of a violent emergency situation, such as a shooting or a hostage situation. This prompts law enforcement action against the victim at the reported location. Swatting often occurs in online gaming communities against rival gamers.[202] AI techniques have been deployed to alter and disguise the caller's voice. "Torswats", a Californian teenager, provided swatting as a service through Telegram, a popular messaging app. It is unclear if AI techniques were used to generate the synthetic voice used in making the calls, but he could have.[203] Torswats was arrested in early 2024.[204]

[198] Sharrock, S., Hudson, N., Kerr, J., et al. (2024, March 15). *Key attributes and experiences of cyberbullying among children in the UK*. Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/keeping-children-safe-online/experiences-of-children/key-attributes-and-experiences-of-cyberbullying-among-children-in-the-uk.pdf?v=368017.

[199] *Tech Trends Position Statement: Generative AI.* (2023, August). Australian Government: eSafety Commissioner. https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf.

[200] Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. *arXiv preprint arXiv:2406.13843.*

[201] Gilbert, D. (2023, March 8). *High Schoolers Made a Racist Deepfake of a Principal Threatening Black Students*. VICE. https://www.vice.com/en/article/school-principal-deepfake-racist-video/.

[202] *What is swatting? | How to prevent swatting.* (n.d.). Cloudflare. https://www.cloudflare.com/learning/security/glossary/what-is-swatting/.

[203] Cox, J. (2023, April 13). *A Computer Generated Swatting Service Is Causing Havoc Across America.* VICE. https://www.vice.com/en/article/torswats-computer-generated-ai-voice-swatting/.

[204] Randall, H. (2024, February 2). *Infamous swatter-for-hire Torswats likely arrested after a private investigator worked with a WoW Twitch streamer and the FBI to take them down.* PC Gamer. https://www.pcgamer.com/infamous-swatter-for-hire-torswats-likely-arrested-after-a-private-investigator-worked-with-a-wow-twitch-streamer-and-the-fbi-to-take-them-down/.

These examples demonstrate not only the accessibility of technology but also novel forms of misuse. Not only are perpetrators no longer limited to state or sophisticated actors, potential victims have expanded beyond organisations and public figures. Motivations have also broadened to include matters of a more personal nature and involving pettier commercial gains.

Harassment, Abuse and Trolling

GenAI is often used against women to harass or intimidate them.[205] Online and digital violence has been shown to personally affect 38% of women globally, with 85% of women reporting having witnessed digital violence against other women.[206] NCII is often the medium of choice, though research on gender-based violence against women journalists demonstrates that they also experience misogyny and threats of physical violence.[207]

Trolling involves sharing, commenting or posting deliberately inflammatory content in order to evoke an emotional reaction and/or create conflict with another.[208] AI could be used to facilitate the creation of fake troll/bot accounts and churn out inflammatory content at scale. However, reference to GenAI and trolling is more often in relation to troll farms that facilitate misinformation and disinformation for political motives. Refer to the section on Misinformation and Disinformation for further information.

There have been examples of AI chatbots engaging in offensive conversations with users.[209] These have typically been examples of chatbots failing to function appropriately or users encouraging the chatbot to behave inappropriately. They were not instances of deliberate use of chatbots to facilitate cyber bullying or harassment.

AI Stalking

Methods of cyber harassment and stalking have become more sinister in a digital age with smart technology being used to monitor victims though their phones, vehicles and even baby monitors inside people's homes.[210] Stalkerware can be installed on personal mobiles, or someone may gift a mobile device with this malware pre-installed. In both

[205] Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. *arXiv preprint arXiv:2406.13843.*

[206] *Measuring the prevalence of online violence against women.* (2021, March 1). The Economic Intelligence Unit. https://onlineviolencewomen.eiu.com/.

[207] de Alwis, R., & Vialle, E. (2024, May 6). *Is AI-Facilitated Gender-Based Violence the Next Pandemic?* The Regulatory Review. https://www.theregreview.org/2024/05/06/de-silva-de-alwis-vialle-is-ai-facilitated-gender-based-violence-the-next-pandemic/.

[208] *Trolling*. (2024, July 30). Netsafe. https://netsafe.org.nz/online-abuse-and-harassment/trolling.

[209] For example, Microsoft's "Tay" chatbot - https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/; And Korean chatbot "Lee Luda" - https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook

[210] Highes, B. (2023, November 16). *Stalkers using smart technology to track victims, PSNI say*. BBC News. https://www.bbc.com/news/uk-northern-ireland-67441053.

cases, the perpetrator can use the technology for stalking and will have the ability to view personal information including photos, text messages and the device's location.[211]

Commentators have observed that AI could be used to supercharge stalking. Facial recognition systems can be used to recognize targets, and large language models are capable of modelling behaviour and parsing large amounts of unstructured content.[212] A couple of Harvard students successfully proved this hypothesis in an experiment labelled "I-XRAY". They successfully used a video feed from a pair of Meta Ray Ban 2 glasses to automatically find and compose a report of personal information about strangers to a mobile phone. Facial recognition technology used to identify strangers and LLMs were used to infer details from unstructured data retrieved online.[213]

Use of AI to combat Cyberbullying and Harassment

AI is presently used to detect and moderate such online behaviour. However, the identification of such behaviours presents difficulties. Categorising content as abusive or bullying can be context-dependent. For example, detection based on words alone (such as "bitch" or "fuck") could create false positives where such expletives are used in a playful or friendly manner. On the other hand, a picture with a tombstone with the words "You belong here" may evade detection. Detection of cyberbullying based on exclusionary behaviour faces similar difficulties.[214]

Users are circumventing efforts to make AI models safer for users, for example, by experimenting with different prompts to 'jailbreak' the model.[215]

[211] *Stalkerware: What To Know.* (2023, November). Federal Trade Commission: Consumer Advice. https://consumer.ftc.gov/articles/stalkerware-what-know.

[212] *AI in the hands of stalkers, abusers, and traffickers: a new frontier in victims' rights.* (2023, April 22). C. A. Goldberg: Victims' Rights Law Firm. https://www.cagoldberglaw.com/ai/.

[213] Winder, D. (2024, October 4). These 2 Hackers Have Created Real X-Ray Specs. Forbes. https://www.forbes.com/sites/daveywinder/2024/10/04/these-2-hackers-have-created-real-x-ray-specs.

[214] Milosevic, T., Van Royen, K., & Davis, B. (2022). Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity. *International journal of bullying prevention : an official publication of the International Bullying Prevention Association*, 4(1), 1–5. https://doi.org/10.1007/s42380-022-00117-x.

[215] *Tech Trends Position Statement: Generative AI.* (2023, August). Australian Government: eSafety Commissioner. https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf.

Annex 8: Misinformation and Disinformation

Background

Disinformation is generally understood to be the deliberate creation or dissemination of false information with the intent to mislead others. Whereas misinformation is spread by actors who do not intend to mislead.[216,217]

While false information has a long history, its modern resurgence takes the form online of "fake news". Fake news topics most prominently featured in recent history were matters of health and political interest,[218] such as false information regarding Covid-19[219] and propaganda and interference in political affairs[220].

Digital technology has made publishing content cheaper and access more widespread. Online media doesn't suffer the cost of printing or physical distribution and content is immediately accessible to anyone with an internet connection. The interface of access has also gravitated towards algorithmically driven search engines, news aggregators and social media platforms, as well as sharing between contacts. This has resulted in a greater distribution of news sources away from traditional media and business models more geared towards advertising and driving traffic - all contributing to the spread of false information.[221]

Individual preferences and cognitive biases also have resulted in information echo chambers that are reinforced by digital technology, making specific narratives hard to dismiss, even if based on false information.[222,223]

[216] *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms.* (2023, August). World Economic Forum. https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.

[217] *Understanding online false information in the UK*. (2021, January 27). Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/economic-discussion-papers-/understanding-online-false-information-uk.pdf?v=325850.

[218] *Understanding online false information in the UK*. (2021, January 27). Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/economic-discussion-papers-/understanding-online-false-information-uk.pdf?v=325850.

[219] Ibid.; *POFMA Action Taken Up To 30 September 2024.* (2024). Singapore POFMA Office. https://www.pofmaoffice.gov.sg/files/tabulation_of_pofma_cases_and_actions.pdf.

[220] *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.* (2021, January 13). University of Oxford: DemTech. https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/.

[221] *Understanding online false information in the UK*. (2021, January 27). Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/economic-discussion-papers-/understanding-online-false-information-uk.pdf?v=325850.

[222] Ibid.

[223] Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature, 618*(7964), 342-348; Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science, 359(*6380), 1146-1151; *Stanford study examines fake news and the 2016 presidential election.* (2017, January 18). Stanford Report. https://news.stanford.edu/stories/2017/01/stanford-

Troll farms are a modern archetype of this scourge. In 2018, Twitter released data showing that nine million tweets originated from a single Russian troll farm.[224] The aim of these bot accounts was to distort global discourse around current political events, working to shape public opinion to support a certain narrative. Another example can be seen with an increase in Chinese propaganda posts on Reddit.[225] There has been evidence of trends suppressing anti-China rhetoric, however there is no clear evidence linking this to the Chinese government or Chinese Communist Party.[226] Spam bots also exist for other purposes that contribute to the spread of harmful and false content.[227]

Impact of GenAI on Misinformation and Disinformation

Algorithms in recommendation systems (e.g. search, page rank, news and social media feeds) primarily affect how information, including false information, is distributed.[228] The immediate impact of GenAI is further upstream on the creation of false information and its multimedia cousin - deepfakes.

Quality, accessibility and scalability of GenAI technology

AI content, both video, text and images, can now be generated easily, cheaply and at a quality often indistinguishable to the naked eye from other content.[229] While the development of text-to-video capabilities lags behind, there have lately been significant improvements.[230]

Deepfake image, audio and video manipulation capabilities can produce realistic results and continue to improve.[231] The amount of deepfake content available online is also increasing.[232] As with other forms of content, deepfake content is capable of being shared and can go viral.[233] For example, "Balenciaga Pope" or Trump's supposed arrest.[234]

study-examines-fake-news-2016-presidential-election; Yonder. (2021, June). *Misinformation: A Qualitative Exploration.* Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2021/misinformation-qual-report.pdf?v=326528.

[224] Romano, A. (2018, October 19). *Twitter released 9 million tweets from one Russian troll farm. Here's what we learned.* Vox. https://www.vox.com/2018/10/19/17990946/twitter-russian-trolls-bots-election-tampering.

[225] Silverman, C., & Lytvynenko, J. (2019, March 14). *Reddit Has Become A Battleground Of Alleged Chinese Trolls.* BuzzFeed News. https://www.buzzfeednews.com/article/craigsilverman/reddit-coordinated-chinese-propaganda-trolls.

[226] Ibid.

[227] https://www.weforum.org/agenda/2022/07/spam-bots-what-are-they/

[228] Lewandowsky, S., Robertson, R. E., & DiResta, R. (2024). Challenges in understanding human-algorithm entanglement during online information consumption. *Perspectives on Psychological Science, 19*(5), 758-766.

[229] Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). As good as a coin toss human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760.*

[230] Moore, J. (2024, January 31). *Why 2023 Was AI Video's Breakout Year, and What to Expect in 2024.* Andreessen Horowitz. https://a16z.com/why-2023-was-ai-videos-breakout-year-and-what-to-expect-in-2024/.

[231] Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety, 1*(4).

[232] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review, 57*(6), 1-47.

[233] Tech Trends Position Statement: Generative AI. (2023, August). Australian Government: eSafety Commissioner. https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf.

[234] Placido, D. D. (2023, March 27). *Why Did 'Balenciaga Pope' Go Viral?* Forbes. https://www.forbes.com/sites/danidiplacido/2023/03/27/why-did-balenciaga-pope-go-viral/.

GenAI can be expected to facilitate the creation of bot accounts and further automate activities of troll farms and spam bots.

Personalisation and influence

Advancements in AI-generated content have the potential to amplify falsehoods in other ways:

- **Personalised Content -** AI-generated content can be personalised to be more persuasive to audiences in commercial advertising and political messaging.[235] The creation of personalised content requires only two elements - 1. a profile of an individual's traits, and 2. a means to generate content for that profile. The means to infer an individual's profile from digital behaviours already exist. GenAI technology realises the second element, allowing personalised content to be created at scale with relative ease.[236] However, the efficacy of microtargeting is unclear.[237]
- **Nudging** - Nudging involves using AI to persuade users over multiple turns of conversation. Studies suggest that AI could hold sway over its conversational human counterpart's views over time.[238]

Hallucinations

A third consideration is the propensity for AI models to provide false or made-up information (i.e. "hallucinations").[239] "Hallucinations" are a source of misinformation. This can present issues at scale. For example, GenAI could be used to create content optimised to drive traffic and increase circulation at scale, with little regard to journalistic content or integrity.[240]

Various approaches to combat "hallucinations" exist.[241] They broadly fall into two categories: 1. Reduction: Reducing the incidence of "hallucinations"; or 2.

[235] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[236] Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports, 14*(1), 4692.

[237] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report.* GOV.UK.https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[238] Ibid.

[239] Ibid., Section 5.2.2.

[240] Farrell, M. (2024, June 6). *BNN Breaking News: AI-generated news and the future of journalism.* The New York Times. https://www.nytimes.com/2024/06/06/technology/bnn-breaking-ai-generated-news.html.

[241] Ghose, S. (2024, May 2). *Why Hallucinations Matter: Misinformation, Brand Safety and Cybersecurity in the Age of Generative AI.* UC Berkeley Sutardja Center. https://scet.berkeley.edu/why-hallucinations-matter-misinformation-brand-safety-and-cybersecurity-in-the-age-ofgenerative-ai/.

Transparency: Being more transparent about the output (e.g. citations, such as those found on Google[242] or Bing[243] AI queries, and confidence scores).[244]

The overall impact of disinformation campaigns and the widespread dissemination of AI-generated content needs to be better understood. In a landscape where people are unable to trust information received, there is greater potential for malicious actors to deliberately exploit informational uncertainty by claiming genuine information to have been faked (i.e. the liar's dividend).[245]

Measures to identify AI-generated context exist but can be circumvented by moderately sophisticated actors.[246] The efficacy of interventions for misinformation have been explored,[247] though its effectiveness on deepfakes and AI-generated misinformation is yet to be fully understood.

[242] *Get search summaries.* (n.d.). Google Cloud. Retrieved October 17, 2024, from https://cloud.google.com/generative-ai-app-builder/docs/get-search-summaries.

[243] Mehdi, Y. (2023, February 7). *Reinventing search with a new AI-powered Microsoft Bing and edge, your copilot for the web*. Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.

[244] *Interpret and improve model accuracy and analysis confidence scores.* (2024, October 16). Microsoft Learn. Retrieved October 17, 2024, from https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept/accuracy-confidence?

[245] Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev., 107*, 1753. https://scholarship.law.bu.edu/faculty_scholarship/640/

[246] Bengio, Y., et al. (2024, May 17). *International Scientific Report on the Safety of Advanced AI: Interim Report.* GOV.UK. https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[247] Johansson, P., Enock, F., Hale, S., et al. (n.d.). *How can we combat online misinformation? A systematic overview of current interventions and their efficacy.* The Alan Turing Institute. https://www.turing.ac.uk/news/publications/how-can-we-combat-online-misinformation-systematic-overview-current-interventions.

Annex 9. Scams

Background

Scams are generally dishonest schemes that seek to manipulate and take advantage of people to gain benefits such as money or access to personal details. Phishing is a specific tactic aimed at stealing sensitive information (including access credentials) through deceptive electronic communication.

The rise in scams is a global problem, causing consumers immense financial, emotional and psychological harm. According to the Global State of Scams - 2023, the financial loss due to scams amounts to a staggering $1.026 trillion, equivalent to 1.05% of the global GDP.[248]

Consumers also experience emotional and psychological harm globally due to scams, including emotional turmoil, shame, loss of confidence and other forms of distress. Almost 3 out of 5 (59%) of scam victims report a substantial emotional impact, signifying the widespread emotional repercussions.[249]

Trends

Impact of AI on Scams

According to Sift research, 68% of consumers reported an increase in the frequency of spam and scams powered by AI.[250] Scammers have been using AI in the following ways:

Voice AI Cloning

AI voice cloning scams have recently gained significant attention. In such scams, victims receive a call that sounds like a loved one in distress. The fake relative then asks the target to pay a ransom, bail, or other fee to get them out of trouble. A survey by McAfee, a computer security software company, found that 70% of people said they weren't

[248] *Global State of Scams Report - 2023*. (2023). GASA. https://www.gasa.org/research.

[249] *Global Statement: Stopping Online Scams*. (2023). Consumers International. https://www.consumersinternational.org/media/513807/anti-scams-global-statement-6-12-23.pdf.

[250] Sift Trust and Safety Team. (2023, July 27). *Growing AI-powered fraud highlights the need for advanced fraud detection.* Sift. https://sift.com/blog/growing-ai-powered-fraud-highlights-the-need-for-advanced-fraud-detection.

confident they could tell the difference between a cloned voice and the real thing. McAfee also said it takes only three seconds of audio to replicate a person's voice.[251]

Case Studies:

During a family trip, a parent received a call that seemed to be from their child, claiming they were in serious danger and had been kidnapped. The voice was eerily accurate, leading the parent to panic and engage in ransom negotiations with the supposed kidnappers. It was later discovered that the child's voice had been cloned using AI technology, and the entire scenario was a scam designed to extort money.[252]

In 2019, the chief executive officer of a British energy provider reportedly transferred €220,000 ($238,000) to a scammer who had digitally mimicked the voice of the head of his parent company and asked for a wire to a supposed supplier on a phone call.[253]

Phishing

Phishing attacks capitalise on advancements in AI to craft convincing and targeted emails, leading unsuspecting users to divulge sensitive information and fall victim to fraud. The rate of account takeover attacks has spread an alarming 427% in Q1 2023, compared to the entirety of 2022. Nearly one-in-five consumers reported that they have been the victim of a phishing attack, account takeover, or payment fraud in the six months following the launch of ChatGPT.[254]

In 2024, research published by Harvard Business Review revealed that 60% of participants in the study fell prey to AI-driven phishing attempts, a success rate comparable to that of phishing schemes crafted by human experts. Even more concerning, their latest findings indicate that the entire phishing process can now be fully automated with large language models (LLMs), significantly reducing the cost of these attacks by over 95% while maintaining or even surpassing traditional success rates.[255]

An AI phishing attack utilises artificial intelligence to create more convincing and personalised phishing emails. By analysing large amounts of data from sources like social media and online behaviour, attackers can craft targeted campaigns that reference a user's recent activities, increasing the likelihood of success. AI can also

[251] Bunn, A. (2023, May 15). *Artificial Imposters—Cybercriminals Turn to AI Voice Cloning for a New Breed of Scam*. McAfee Blog. https://www.mcafee.com/blogs/privacy-identity-protection/artificial-imposters-cybercriminals-turn-to-ai-voice-cloning-for-a-new-breed-of-scam/.

[252] Bethea, C. (2024, March 7). *The terrifying A.I. Scam that uses your loved one's voice.* The New Yorker. https://www.newyorker.com/science/annals-of-artificial-intelligence/the-terrifying-ai-scam-that-uses-your-loved-ones-voice.

[253] Stupp, C. (2019, August 30). *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*. The Wall Street Journal. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[254] Sift Trust and Safety Team. (2023, July 27). *Growing AI-powered fraud highlights the need for advanced fraud detection.* Sift. https://sift.com/blog/growing-ai-powered-fraud-highlights-the-need-for-advanced-fraud-detection.

[255] Heiding, F., Schneier, B., & Vishwanath, A. (2024, May 30). *AI Will Increase the Quantity — and Quality — of Phishing Scams.* Harvard Business Review. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams.

generate realistic replicas of legitimate websites, making it hard for recipients to distinguish between fake and real sites.

AI phishing operates on four key pillars:

1. *Data Analysis*: Attackers use AI tools like WormGPT to gather and analyse data on targets, such as social media profiles and online activities.
2. *Personalization*: The collected data is used to create highly personalised phishing emails, referencing specific details about the target's life.
3. *Content Creation*: AI generates convincing content that mimics the writing style of known contacts or institutions, enhancing trust and overcoming language barriers.
4. *Scale and Automation*: AI enables attackers to efficiently scale operations, generating numerous unique phishing emails and automating processes like code generation and setting up webhooks.[256]

Scam Websites

AI tools can generate realistic text, images, and even deepfake content, enabling fraudsters to create convincing fake websites that mimic legitimate ones. These scams often aim for financial gain by tricking victims into providing sensitive information or transferring funds.[257]

Deepfakes for Scams

In a 2024 cyber scam in Hong Kong, criminals used deepfake technology to impersonate a company's CFO. The attackers created a sophisticated deepfake video, making it appear as though the CFO was directing the company's bank to transfer 25 million dollars to a fraudulent account. The scheme was well-executed, leveraging both AI technology and social engineering to deceive the company's employees, leading to a significant financial loss.[258]

In 2022, Patrick Hillmann, chief communications officer at Binance, claimed in a blog post that scammers had made a deepfake of him based on previous news interviews and TV appearances, using it to trick customers and contacts into meetings.[259]

[256] Schafer, N. (2024, August 28). *The Rise of AI Phishing and What it Means for the Future of Scammers*. Mailgun. https://www.mailgun.com/blog/email/ai-phishing/#subchapter-1.

[257] Watson, J. (2024, May). *The Sinister Side of AI – AI Generated Scam Sites*. Fusion IT. https://www.fusionmanageit.co.uk/node/the-sinister-side-of-ai-ai-generated-scam-sites/.

[258] Chen, H., & Magramo, K. (2024, February 4). *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'.* CNN. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

[259] Hillmann, P. (2022, August 17). *Scammers Created an AI Hologram of Me to Scam Unsuspecting Projects.* Binance. https://www.binance.com/en/blog/community/scammers-created-an-ai-hologram-of-me-to-scam-unsuspecting-projects-6406050849026267209.

Another case this year involved a female financial employee in Shanxi province, who was tricked into transferring 1.86 million yuan ($262,000) to a fraudster's account after a video call with a deepfake of her boss. [260]

Scams via Social Media

GenAI can create realistic images, audio and videos of real individuals. Bad actors can also use AI models to create fake identities for impersonation on social media and news platforms. And they can use AI models to automate the creation of large volumes of fake accounts, articles, graphics, comments and posts. With minimal training, GenAI models can analyse publicly available data and social media profiles and adapt that information for targeted use cases. The models can mimic the style and tone of legitimate communications from trusted sources, translate and localise text in multiple languages and handle repetitive tasks, like replying to messages.[261]

Acronyms and Definitions

Acronym	**Full Form**
AI	Artificial Intelligence
CG-CSAM	Computer-Generated Child Sexual Abuse Material
CGI	Computer-Generated Imagery
CSAM	Child Sexual Abuse Material
CSEA	Child Sexual Exploitation and Abuse
GenAI	Generative Artificial Intelligence
LLM	Large Language Model
NCII	Non-Consensual Intimate Imagery
NIST	National Institute of Standards and Technology
RLHF	Reinforcement Learning from Human Feedback
Term	**Definition**
Big Data	Extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time.

[260] Zekun, Y. (2024, March 7). *Deepfake video scams prompt police warning*. China Daily. https://www.chinadailyhk.com/hk/article/379805.
[261] Pan, S. (2024, July 23). *4 Fake Faces: How GenAI Is Transforming Social Engineering*. Proofpoint. https://www.proofpoint.com/us/blog/security-awareness-training/generative-ai-transforming-social-engineering.

Chatbots	Computer programs that simulate human conversation with an end user.
Computer Vision	A field of artificial intelligence (AI) that uses machine learning and neural networks to teach computers to recognize and understand objects and people in images and videos
Conditional Diffusion Models	Form of GenAI popularly used to create photo-realistic images from text prompts.
Deep Learning	A subset of machine learning that attempts to mimic the human brain, enabling systems to cluster data and make predictions with incredible accuracy.
Deepfake	A type of AI used to create convincingly fake images, videos and audio recordings.
Deployer	A person or entity that uses an AI system.
Developer	An AI developer is an organisation or entity that is concerned with the development of AI services and products.
Generative AI	A subset of artificial intelligence capable of generating text, images, or other media in response to prompts using pre-trained transformer models, typically with massive amounts of data.
Hash-Based Detection	A signature-based detection method mainly used to identify malware or dangerous files.
Machine Learning	A branch of AI and computer science that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving accuracy.
Machine Learning Classifiers	A type of machine learning algorithm that automatically assigns data points to a range of categories or classes.
Machine Learning Model	A model that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.
Model refusals	Getting a model to refuse to respond to certain questions
Natural Language Processing	the application of computational techniques to the analysis and synthesis of natural language and speech
Neural Networks	a computer system modelled on the human brain and nervous system.

Responsible AI	Is intended to result in technology that is also equitable and accountable.
Reinforcement Learning	Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results.
Swatting	Prompting an emergency law enforcement response against a subject under false pretences. See Annex 7 for examples.
Traditional AI	Refers to AI systems before genAI systems were introduced.
User	An entity or person (internal or external) that interacts with an AI system or an AI-enabled service and can be affected by its decisions.