ETSI TR 104 137 V1.1.1 (2026-03)

TECHNICAL REPORT

Cyber Security (CYBER);
Human-to-Human Online Preventative Security

Reference
DTR/CYBER-00156

Keywords
end-user, privacy, public safety, user-centric

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Cyber Security (CYBER).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

The purpose of this present document is to provide information about the threats and harms from online services that affect people not just online, but also how those harms can impact their daily lives. The aim is to provide an understanding of how this complex environment of apps, devices and social networks can be used for harm, along with presenting and discussing the steps that can be taken to mitigate harm and risks.

The present document builds upon and connects to previous work in ETSI relating to users, services and devices. These include, but are not limited to:

- TC Cyber ETSI TR 103 936 [i.79] Implementing Design practices to mitigate consumer IoT-enabled coercive control.
- TC Human Factors ETSI TR 104 077 [i.80] Series Age Verification Pre-Standardization Study: Stakeholder Requirements; Solutions and Standards Landscape and Proposed Standardization Roadmap.
- TC Human Factors ETSI TR 102 133 [i.81] Access to ICT by young people: issues and guidelines.

- TC Securing Artificial Intelligence ETSI TR 104 159 [i.82] Understanding and Preventing Harm from Generative AI.

The present document provides information and guidance relevant to the current legislative and regulatory landscape. These include, but are not limited to, the:

- EU Digital Services Act.
- UK Online Safety Act.
- Australia Online Safety Act 2021:
    - Online Safety Amendment (Social Media Minimum Age) Act 2024 (Cth).

1 Scope

The present document will aim to provide guidance and information about social media networks and messaging applications, and their many issues regarding user safety, privacy, and content. When tackling these issues, one tends to think of a single service or application. The present document moves to a holistic approach. This is the same way in cybersecurity when helping companies understand their attack surface from their systems and devices to the suppliers and vendors they rely on.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] Us Department Of Homeland Security: "Swatting Calls And Hoax Threats".

[i.2] Federal Bureau of Investigation: "Recent Swatting Attacks Targeting Residents with Camera and Voice-Capable Smart Devices"; 2020; Public Service Announcement.

[i.3] Gary Shelton, Journal of the Advanced Practitioner in Oncology: "Appraising Travelbee's Human-to-Human Relationship Model", 2016.

[i.4] Preventative Control.

[i.5] How the world changed Social Media; 2016; UCL Press.

[i.6] The Dark Figure of Crime and the Reporting of Crimes.

[i.7] The Dynamics of Racism, Antisemitism and Xenophobia on Social Media in South Africa.

[i.8] The Threat of Online Hate to Religious Freedom is Too Great to Ignore.

[i.9] Religious Intolerance: A Challenge For The European Media.

[i.10] Addressing misogyny, toxic masculinity and social media influence through PSHE education.

[i.11] SAFER Scrolling: "How algorithms popularise and gamify Online hate and misogyny for young people".

[i.12] Online Hate Speech: "A Guide for General Public".

[i.13] Facebook: "Telegram, and the Ongoing Struggle Against Online Hate Speech; Caroline Crystal; 2023; Carnegie Endowment for International Peace".

[i.14] A postmodern Pandora's box: Anti-vaccination misinformation on the Internet; Anna Kata; Vaccine 28 (2010) 1709–1716.

[i.15] COVID-19 vaccine misinformation.

[i.16] Poll Vaulting: Cyber Threats to Global Elections.

[i.17] Science hostility: "What we know and what we can do about it".

[i.18] 'Emotivism' in social media and the rule of law.

[i.19] What Is Cyberbullying; stopbullying.gov.

[i.20] Cyberbullying among young people: Laws and policies in selected Member States; 2024; European Parliamentary Research Service.

[i.21] Cyberbullying: What is it and how to stop it - What teens want to know about cyberbullying; UNICEF.

[i.22] Doxing: what is it and what are your rights?; 2022; Scottish Women's Rights Centre.

[i.23] Grooming; Metropolitan Police.

[i.24] Online Grooming, Considerations for Detection, Response, and Prevention of Online Grooming; Tech Coalition.

[i.25] Algorithms in Social Media Platforms.

[i.26] Misinformation on Social Media; Social Media Algorithms.

[i.27] Unveiling The Dark Side Of Social Media Algorithms – Harmful Effects On Mental Health.

[i.28] ChatGPT and large language models: what's the risk?; 2023; NCSC.

[i.29] AI Privacy Risks & Mitigations Large Language Models (LLMs); 2025; Isabel BARBERÁ; EDPB.

[i.30] Let's Chat: Examining Top Risks Associated With Generative Artificial Intelligence Chatbots.

[i.31] Polymedia: "Towards a new theory of digital media in interpersonal communication; Mirca Madianou and Daniel Miller"; 2012.

[i.32] Using digital at primary school; internet matters.org.

[i.33] Moving to secondary school; internetmatters.org.

[i.34] Student Generative AI Survey 2025.

[i.35] Using AI in the workplace: Opportunities, risks and policy responses; 2024; OECD.

[i.36] European Commission, Code of Conduct on Disinformation 2025.

[i.37] European Commission, Action Plan against Disinformation (Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, 2018).

[i.38] European Commission, Communication tackling online disinformation: A European Approach (Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions COM/2018/236).

[i.39] Journalism Trust Initiative; Workshop Agreement CWA 17493:2019.

[i.40] Age assurance for the Children's code.

[i.41] A Simplified Guide to Digital Evidence.

[i.42] EU Age Verification Solution.

[i.43] Operational, Security: "Product, and Architecture Specifications".

[i.44] The impact of the Digital Services Act on digital platforms.

[i.45] Ofcom: "Illegal content duties under the Online Safety Act".

[i.46] awesome-threat-intelligence.

[i.47] Cybersecurity: "Databases & Recommended Resources".

[i.48] NIST SP 800-160v1r1: "Engineering Trustworthy Secure Systems".

[i.49] NIST Cybersecurity Framework.

[i.50] Secure by Design Principles; Government Security Group.

[i.51] Ministry of Defence: "Secure by Design: Design for security from the start".

[i.52] ISO/IEC 27000 series: "Information security management".

[i.53] ETSI TS 103 645 (V3.1.1): "CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements".

[i.54] ETSI EN 303 645 (V3.1.3): "CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements".

[i.55] CISA: "Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software".

[i.56] Department for Science, Innovation and Technology and Department for Digital, Culture, Media & Sport; 2021: "Principles of safer online platform design; Understand how preventative design measures can reduce the risk of harms happening on your online platform".

[i.57] GDPR: "Privacy by Design".

[i.58] iapp.org: "What is Privacy?".

[i.59] UK Information Commissioner's Office (ICO): "Data protection by design and default".

[i.60] Cybersecurity for Children.

[i.61] Cyberfit Nation.

[i.62] Cyber A.C.E.S: "Activities in Cybersecurity Education for Students".

[i.63] Cyberlite.

[i.64] ETSI Cybersecurity Conference 2024: "Impersonation & Psychological Warfare Protecting The Children".

[i.65] Open Technology Institute; 2024: "Age Verification: The Complicated Effort to Protect Youth Online".

[i.66] CNIL: "Online age verification: a complex issue with significant privacy risks".

[i.67] "Digital Regulation Cooperation Forum; 2022: "Online safety and data protection: A joint statement by Ofcom and the Information Commissioner's Office".

[i.68] GOV.UK: "Users' account details and activity visible to others; improve the safety of your online platform".

[i.69] GOV.UK: "Search functionality; improve the safety of your online platform".

[i.70] NSPCC Learning: "Online harms: protecting children and young people".

[i.71] CLEMi: "Le numérique en famille!; Outils et ressources pour les parents et les professionnels".

[i.72] NCSC: "Stay safe online top tips for staff infographic".

[i.73] NCSC: "Training; Top Tips For Staff".

[i.74] ageUK: "Staying safe online".

[i.75] Tremau T&S Research Team; 2022: "Content moderators: How to protect those who protect us?".

[i.76] gearinc; 2025: "How to Protect Employees Involved in Content Moderation".

[i.77] ETSI TS 102 232 (all parts): "Lawful Interception (LI); Handover Interface and Service-Specific Details (SSD) for IP delivery".

[i.78] ETSI TS 103 643 (V1.2.1): "Techniques for assurance of digital material used in legal proceedings".

[i.79] ETSI TR 103 936 (V1.1.1): "Cyber Security (CYBER); Implementing Design practices to mitigate consumer IoT-enabled coercive control".

[i.80] ETSI TR 104 077: "Human Factors (HF); Age Verification Pre-Standardization Study; Part 1: Stakeholder Requirements".

[i.81] ETSI TR 102 133: "Human Factors (HF); Access to ICT by young people: issues and guidelines".

[i.82] ETSI TR 104 159: "Securing Artificial Intelligence (SAI); Understanding and Preventing Harm from Generative AI".

[i.83] European Convention on Human Rights.

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

age assurance: techniques for estimating, inferring or verifying the ages of a person

algorithm: in social media networks and services, algorithms are rules, signals and data that govern the platform's operation. These algorithms determine how content is filtered, ranked, selected and recommended to users

AI: Machine Learning (ML) applications that involve software components (models) that allow computers to recognize and bring context to patterns in data without the rules having to be explicitly programmed by a human to generate predictions, recommendations, or decisions based on statistical reasoning

antisemitism: a certain perception of Jews, which is expressed as hatred toward Jews

NOTE: Rhetorical and physical manifestations of antisemitism are directed towards Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities. However, criticism of Israel, similar to that levelled against any other country (for example, not liking a country's food or weather), cannot be regarded as antisemitic. Antisemitic acts are criminal when they are so defined by law (for example, denial of the Holocaust or distribution of antisemitic materials and acts of violence in some countries).

cyberbullying: is bullying with the use of digital technologies

NOTE: The places it can happen include social media, messaging platforms, gaming platforms and mobile phones. It is a repeated behaviour, aimed at scaring, angering or shaming those who are targeted.

DEI: organizational frameworks that seek to promote the fair treatment and full participation of all people, particularly groups who have historically been underrepresented or subject to discrimination based on identity or disability

demography: study of people in a particular area

digital literacy: confident and critical use of a full range of digital technologies for information, communication and basic problem-solving in all aspects of life

disinformation: misleading content deliberately spread to deceive people, or to secure economic or political gain and which may cause public harm

disinformation control: measure that treats disinformation risks by reducing their likelihood or their consequences.

disinformation stakeholder: natural or legal person, public authority, agency or any other body that can affect, be affected by, or perceive themselves to be affected by a decision or activity related to disinformation

grooming: action of attempting to form a relationship with a child or adult, with the intention of sexually assaulting them or inducing them to commit an illegal act such as selling drugs or joining a terrorist organization

hate crime: crime, typically one involving violence, that is motivated by prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds

hate speech: offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion or gender) and that may threaten social peace

illegal content: under the UK's Online Safety Act, refers to content that constitutes a "relevant offence," which includes priority offences like child sexual exploitation and abuse, terrorism, fraud, and human trafficking

NOTE: It also covers content related to extreme pornography, hate offences, harassment, and assisting or encouraging suicide.

internet: applications and services, most prominently the World Wide Web, including social media, electronic mail, mobile applications, multiplayer online games, direct messaging, voice calls, file sharing, and streaming media services that can be found or used on it

Large Language Model (LLM): category of foundation models trained on immense amounts of data, making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks

misinformation: piece of information that lacks veracity and that could mislead people

misogyny: dislike of, contempt for, or ingrained prejudice against women

racism: prejudice, discrimination, or antagonism by an individual, community, or institution against a person or people on the basis of their membership of a particular racial or ethnic group, typically one that is a minority or marginalized

street aware(ness): ability to manage or succeed in difficult or dangerous situations

xenophobia: dislike of or prejudice against people from other countries

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

2FA	Two Factor Authentication
AI	Artificial Intelligence
API	Application Programming Interface
AppDev	Application Developer
ARF	Architecture and Reference Framework
AVMSD	Audio-Visual Media Services Directive
BBM	BlackBerry Messenger
CD	Compact Disc
CISA	Cybersecurity and Infrastructure Security Agency
CLEMi	Le centre pour l'éducation aux medias et à information
CSAM	Child Sexual Abuse Materials
CV	Curriculum Vitae
DDoS	Distributed Denial of Service
DEB	Digital Evidence Bag
DEI	Diversity, Equity, Inclusion
DMA	Digital Market Act
DMCA	Digital Millennium Copyright Act
DNA	Deoxyribonucleic acid
DoB	Date of Birth

DSA	Digital Services Act
EC	European Commission
EU	European Union
EUDI	European Digital Identity
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
HW	Hardware
ICT	Information and Communication Technology
ID	Identification
IEC	International Electrotechnical Commission
IM	Instant Messaging
IoT	Internet of Things
IP	Internet Protocol
IRTF	Internet Research Task Force
ISD	Institute for Strategic Dialogue
ISO	International Organization for Standardization
IT	Information Technology
LEA	Law Enforcement Agency
LGBT+	Lesbian, Gay, Bisexual, and Transgender,

NOTE: With the plus (+) signifying inclusivity for other sexual orientations and gender identities like Queer, Questioning, Intersex, Asexual, and others, representing a diverse community whose members are not heterosexual or cisgender.

LLM	Large Language Model
MFA	Multi-Factor Authentication
ML	Machine Learning
MoD	Ministry of Defence
NCSC	National Cyber Security Centre
NIST	National Institute of Standards and Technology
PIN	Personal Identification Number
PTSD	Post-Traumatic Stress Disorder
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning with Human Feedback
SIM	Subscriber Identity Module
SMS	Short Message Service
SP	Special Publications
SSD	Service-Specific Details
UK	United Kingdom
URL	Uniform Resource Locator
USA	United States of America
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine
VOIP	Voice Over Internet Protocol
VPD(s)	Vaccine-preventable disease(s)
VPN	Virtual Private Network
WWW	World Wide Web
ZTA	Zero Trust Architecture

4 Harmful Actions, Attacks, and Societal Risks

4.0 Overview

A significantly diverse array of analyses and legislative findings makes plain that social media networks, messaging applications, and associated support services, including search engines, give rise to a broad array of harmful actions and attacks on individual users, institutions, and society, including fundamental human rights and values. These developments and actions constitute a form of serious cybersecurity threat. This is not a new problem, as the issues began to emerge when the first social networks arrived in the form of Dial-in Bulletin Board Systems. However, the rapid evolution of social networks and associated services today, especially when combined with manipulative Social Network algorithms and Artificial Intelligence, profoundly increases the potential harm and attacks.

When treating these issues, there is a tendency to think of social media as a single service or application. The present document provides a holistic approach to understanding the relations between users, harms and the platforms they use, similar to understanding the concept of an attack surface to improve cybersecurity. In the present document, the Human-to-Human Relationship is established as an interactive process. The inaugural meeting or encounter may immediately establish a connection. Unfortunately, this connection may not be positive. Through the emergence of various personal identities, both humans attempt to relate to or find meaning in their encounters [i.3]. Examples of negative interaction are explored in clause 4.1. An understanding of the demographics of the people using social networks and the types of risks and harms they can experience at different ages is discussed in clause 6. Mitigating these negative interactions can be achieved through online preventive security, which includes privacy, safety and security mechanisms, tools or practices that can deter or mitigate the undesired actions or events from occurring [i.4]. These preventive and mitigation measures are discussed in clause 6.

4.1 Hate Speech and Hate Crime

4.1.1 Xenophobia, Racism, Antisemitism

Various discussions on social media are driven and dominated by extremist and/or fringe views [i.7]. Societal fissures along national, racial and economic lines are exploited to expand certain views for social media footprint and amplify their message. These and other users often seek to create a crude 'us'/'them' binary. Those portrayed as 'others' are stigmatised. Those social media users perceived not to tow the appropriate line are routinely described as sell-outs and race traitors and positioned as having no standing to have their opinions heard.

In these cases, rather than galvanising healthy debate through exposure to different viewpoints, social media instead fosters small, polarized communities of like-minded individuals (echo chamber). The intemperate and alienating nature of online political discussion hardens boundaries. Those attempting to cross these lines are routinely attacked in hateful terms.

Paradoxically, the hateful content this produced typically spawned more content as a chorus of supporters and detractors chimed in. By generating attention, the production and publication of hate speech, thus, in turn, likely increased the amount of time users spent on online platforms. Given that user engagement is the desideratum of social media companies, and that content moderation is complex, costly and time-consuming, platforms at times struggle to provide adequate and timely moderation, particularly that which is sensitive to local languages and cultures. Also, an aspect of social media algorithms is their tendency to amplify negative or controversial content. Because negative emotions often drive higher engagement, the algorithm may prioritize this type of content in users' feeds.

So, while content moderation can limit the spread of hateful content, as it is a mainly reactive method, malicious actors can circumvent it by sharing content that does not automatically get flagged. This includes the deliberate resort to derogatory terminology drawn from local languages to evade known moderation terms, provocations designed to draw attention and traffic, the use of fake accounts, and other techniques.

4.1.2 Religious Intolerance

Religious intolerance can be understood as an intolerance of a person's religion, religious beliefs, or practices. It is rooted in negative attitudes, values, and beliefs toward those of a particular religion. While attitudes of religious intolerance can be challenging to measure and quantify, actions are more recognizable.

Religious intolerance can materialize in many ways, from microaggressions to a lack of accommodation and acceptance of religious practices, to vandalism of religious buildings, hate speech, and physical violence. All acts of religious intolerance are forms of discrimination based on religion. The psychological, economic, and societal impacts that these intolerant behaviours and attitudes have on people can be very detrimental and long-lasting.

Due to social media becoming intertwined with our lives offline, threats to religious freedom are no longer confined to the physical world. This has led to harmful content on social media manifesting in physical acts of violence targeting vulnerable communities [i.8]. On social media, religious intolerance content has been steadily increasing, with it presented as opinions or comments posted on micro-blogging sites, as memes on platforms, or as fake news circulating on various websites [i.9].

4.1.3 Misogyny and Sexual Orientation

Misogyny can be defined as a hatred of, aversion to, or prejudice against women and Sexual Orientation, also known as sexuality, refers to a person's sexual attraction to other people, or lack thereof [i.10]. This includes emotional, romantic, sexual or affectionate attraction to other people. On social media, these two topics intersect with each other as misogynistic rhetoric can be permissive of discriminatory behaviours and attitudes (including racist, antisemitic, homophobic, transphobic, and misogynistic attitudes); sexual harassment and abuse; abuse in relationships; and victim-blaming narratives. While people of all ages can be affected by this, the entry point for this behaviour starts at a young age, often during secondary education, which is a time of great change for children as they go through puberty, social change, exams, etc.

This all comes at an age when young people may be particularly insecure and vulnerable to persuasive narratives. For example, much of this content taps into insecurities about body image and agency. The focus on money, success and power also plays on financial and status insecurities that may lead to risky and even illegal behaviours.

Women and girls in particular are put at risk by narratives that normalize sexual harassment and abuse, promote unhealthy relationship behaviours and victim-blaming. Similarly, the sharing of homophobic and transphobic content by some influencers can inform behaviour and attitudes towards LGBT+ young people in schools. Toxic masculinity is also harmful and restrictive to boys and men. Promoted stereotypes contribute to existing narratives that restrict and undermine help-seeking, especially concerning mental health and emotional well-being. Some online spaces even direct young people towards other harmful content, including content that promotes self-harm and suicide.

In some extreme circumstances, mainly boys can also become radicalised via involvement in forums and communities that promote and celebrate violent behaviours, including rape, child abuse and terrorist acts.

Part of the reason is that people are shown this content because social media algorithms show users content they interact with and/or recommend them content that peers have also engaged with. This is because an algorithm is a set of rules and signals that social media sites use to show their users the videos, pictures or articles that they are most likely to interact with [i.11]. Harmful content is presented as entertainment through the algorithmic processes of social media platforms, which can amplify negative materials to young people and vulnerable groups. In this way, toxic, hateful or misogynistic material is pushed to young people, exploiting adolescents' existing vulnerabilities. Boys who are suffering from poor mental health, bullying, or anxieties about their future are at heightened risk. As a result, ideologies, such as sexism and misogyny, are normalized amongst young people and seep into their everyday interactions. Young people increasingly exist within digital echo-chambers, which normalize this rhetoric. This seeps into everyday interactions between young people, with boys lacking awareness of its impact on their female peers.

Different measures can be taken for safeguarding against toxic masculinity and misogynistic narratives (as well as the influencers and algorithms that deliver them so effectively). These include preventative education, providing a foundational understanding from an early age about healthy relationships, respect, self-esteem, digital literacy, economic wellbeing, critical thinking, and recognizing and challenging negative influences. And understanding the links between these areas.

4.2 Disinformation and Misinformation

4.2.0 Overview

The following clauses are not the only types of disinformation and misinformation found online. They are non-exhausted list of examples of how disinformation and misinformation are created and spread online.

4.2.1 Abetting hate speech

Social media is used by members of the public to publish content online. The majority of posts on social media are ordinary and harmless (such as posts on daily activities, comments on TV shows or news stories, sharing photos and videos of friends and family) [i.12].

However, a small number of social media users post comments that might be viewed as offensive or hateful towards individuals and communities. Hateful posts sent through social media targeted at individuals because their identity, such as race, religion, sexual orientation, disability and transgender identity, can be regarded as criminal offences depending on location.

Hateful social media posts, whether they are regarded as criminal or not, can cause a great deal of harm to individuals and communities. The law around this varies from country to country.

Hate crime generally covers offences that are aggravated because of hostility towards the victim's race, religion, disability, sexual orientation or transgender identity. Hateful social media posts (other than those which amount to specific offences in their own right, such as making threats to kill, blackmail, stalking, etc.) can be considered to be criminal if, for example:

- Their content is grossly offensive.
- Their content is threatening or abusive and is intended to or likely to stir up racial hatred.
- Their content is threatening and is intended to stir up hatred on the grounds of religion or sexual orientation.

Online hate speech on the Internet has been around since the beginning, with it present in one form or another. Before social media, hate speech was largely confined to being used on static web pages, usually set up by hate groups. These websites promote the goals of these groups, aiming to recruit like-minded people. Social media enabled them to more widely spread their message and reach a larger audience than previously.

Websites that promote the incitement of hatred towards minorities are not permitted in the UK or EU, but many still exist on the Internet as they are hosted in countries where the laws are different. For example, in America, websites that contain information that may be considered hateful by people in the UK or EU are allowed because there are specific laws that protect 'free speech' in the US Constitution, as freedom of expression in the US is protected but not fundamental, and it applies to strictures by government. Contrary to the USA, freedom of expression is a qualified right in the EU, and as such it is not absolute and it can be limited in certain circumstances, such as, for example, to protect other rights, public safety or national security. Examples of limitations on freedom of speech might include laws against hate speech, defamation, or incitement to violence (see Article 10 of the European Convention on Human Rights [i.83]).

Different measures can be taken to prevent or minimize the sharing of hate speech, including content moderation, education, digital literacy, reporting mechanisms and blocking/banning malicious users [i.13]. How effective these measures are depends on the amount of resources allocated to implementing them.

4.2.2 Anti-vaccination promotion

This term refers to efforts or campaigns that seek to discourage or prevent people from receiving vaccines, oftentimes by spreading misinformation, misleading arguments or unfounded fears. Such efforts can take place through various channels, such as social media, websites and sometimes through public figures, including influencers or celebrities.

With morbidity and mortality from Vaccine-Preventable Diseases (VPDs) having reached record lows [i.14], vaccines are among the most successful tools for biomedical science and public health. Yet paradoxically, the effectiveness of vaccination has led to the reemergence of anti-vaccination sentiments. Vaccines may be seen as unnecessary or dangerous because incidence rates of VPDs in developed countries have plummeted. Vaccine "reactions" – negative health events following vaccination, attributed to the vaccine, then appear to be more common than the diseases themselves. In this way, vaccines can be considered victims of their success.

The media plays a large role in disseminating and sensationalising vaccine objections. Such objections are part of what has been called the "anti-vaccination-movement", which has had a demonstrable impact on vaccination policies, individual and community health. A common sequence to vaccination scares involves scientific debate about potential vaccine risks, which communication technology transmits via a rhetoric of doubt; parents incorporate this with personal experiences and spread their views to their social groups (offline and/or online). These social groups exert considerable pressure on vaccination decisions by creating a "local vaccination culture". With the prominence of the Internet in today's world, the attitudes, beliefs, and experiences of that local culture can quickly become global.

Currently, the main source of vaccine misinformation is social media [i.15]. Social media often amplifies misinformation and allows it to spread quickly to a large number of people, as misinformation spreads much faster than factual information online. The rapid spread of misinformation online is potentially fuelled by the algorithms that underpin social media platforms, which display content that is likely to receive a high amount of engagement. Social media algorithms also tailor the content that a user sees to display information that aligns with their existing beliefs and values, creating what is known as 'echo chambers'. Echo chambers can cause certain beliefs to be amplified because users do not get shown information or opinions from an alternative perspective that may challenge their attitudes.

A key characteristic of anti-vaccination promotion is misinformation and disinformation. For example, it oftentimes involves incorrect or distorted information about vaccines, including, but not limited to:

i) False claims about vaccines, such as, that they cause autism, despite scientific evidence disproving this link.

ii) Unfounded or exaggerated fears about the side effects of vaccines, which can make people hesitant to get vaccinated.

iii) Claims that vaccines are "unnatural" or that the body is better off without them due to natural immunity.

Various steps can be taken to tackle misinformation on social media platforms, including removing or demoting misinformation, directing users to information from official sources, and banning certain adverts.

4.2.3 Election Integrity Attacks

These refer to efforts or activities designed to undermine or disrupt the fairness, transparency, or accuracy of an election process. These can take various forms, including disinformation campaigns aiming to confuse or manipulate voters. The overall objective of such attacks is to influence the outcome of an election, erode public confidence in democracy, or disrupt the functioning of the electoral system.

Disinformation and misinformation campaigns aiming at attacking election integrity can take the following forms:

i) Disinformation campaigns often at times have the objective of confusing voters by spreading false or misleading information. For example, this could entail fabricated claims about voting procedures, fake reports of voter fraud, or exaggerated claims about the outcome of elections.

ii) Social media platforms can be manipulated in such a way to amplify disinformation, create confusion among voters, and promote polarizing narratives. Oftentimes, bot accounts, fake profiles on social media or paid advertisements are used to mislead voters or dissuade them from voting.

iii) There is an increasing use of deepfake videos or fake news websites, which show misleading or fabricated content. Deepfake videos can be used to distort the reputation of political candidates by showing them making controversial statements.

In democratic countries, elections are a crucial part of a country's democratic process, and organizations with a role coordinating them may face threats from criminals and nation-state actors who wish to disrupt the electoral process [i.16]. Integrity attacks targeting election-related infrastructure can combine cyber intrusion activity, disruptive and destructive capabilities, and information operations, which include elements of public-facing advertisement and amplification of threat activity claims. The attack surface of an election involves a wide variety of entities beyond voting machines and voter registries. Election integrity attacks also target the key people involved in campaigning, political parties, news and social media more frequently than actual election infrastructure. This means securing elections requires a comprehensive understanding of many types of threats and tactics, from Distributed Denial of Service (DDoS) to data theft to deepfakes, that have the potential to impact elections.

4.2.4 Attacks on established science

These refer to efforts to undermine or distort well-established scientific research. One of the forms that such attacks can take are public misinformation campaigns with an aim to create confusion and doubt or weaken public trust in scientifically established conclusions. For example, this can consist of spreading incorrect information on social media about climate change, the safety of vaccines, and the effectiveness of certain medical treatments. This can be done by highlighting data that supports a certain view by ignoring a large body of evidence that contradicts it.

Attacks on established science can lead to the erosion of public trust in scientific institutions and experts. Spreading misinformation about vaccines or certain medical treatments can lead to the resurgence of preventable diseases and can result in increased mortality or morbidity.

Social media networks are increasingly relevant as a platform for science communication and the discussion of scientific topics, both because scientists use them to collaborate and distribute research findings and because laypeople turn to the Internet to search for information [i.17]. They allow for a far-reaching distribution of scientific findings and comments by the general audience. But such contributions are often short, pointed, and invite oversimplification.

Public communication on controversial issues like climate change or global health crises has meant that science is more relevant than ever before. At the same time, it has made researchers vulnerable to attacks that undermine their credibility, potentially silencing them and prompting them to withdraw from the public sphere. Attacks arise from both scientific and non-scientific actors, manifesting as online harassment, verbal threats and even physical attacks.

It is worth noting that hostility toward science is not limited to communication between academia and society; it also exists within scientific communities, arising from debates over methodologies, data quality, or openness.

There are different strategies for supporting researchers who face attacks or are cautious about them. For example, is institutional backing, where the organization concerned issues public statements or assists with moderating social media discussions. This includes documenting incidents or reporting abusive behaviour.

4.2.5 Attacks on the rule of law and societal institutions

The rule of law is the foundation that supports a country's democracy, and the societal institutions required to maintain it. It creates a social contract and is the arbiter of disputes as well as the insurer of basic human rights. The size and reach of online social networks can affect trust in the rule of law [i.18].

These refer to efforts to undermine or weaken the foundational systems and structures that ensure fairness, justice and social order within a society. Such attacks often target the principles sustaining democracy, human rights and civil liberties, which in turn damage public trust in institutions, like the judiciary, LEAs, government bodies and the media. These in turn can lead to political, social and economic instability.

This can occur because social media influencers and other prominent figures online, often with thousands or even millions of followers, tend to propagate biases. The issue of bias is that online platforms reinforce users' views and drive confirmation bias: the tendency to look for information consistent with already-held beliefs. Also, a lack of public understanding of the law and politicians criticizing judges and/or lawyers on social media weakens trust in the rule of law.

Disinformation can be used to attack the rule of law and societal institutions. Governments, political groups or other entities oftentimes deliberately spread false or misleading information to control public opinion and undermine the integrity of societal institutions. Disinformation regarding societal institutions instils confusion, division and distrust in the public. Such attacks can also be directed against independent organizations, such as human rights organizations or electoral commissions, when they hold powerful individuals or governments accountable.

Improved education and digital literacy in this area to improve understanding may be a way to ensure continued trust in the rule of law and societal institutions.

4.3 Personal safety and health

4.3.1 Harassment and Cyberbullying

Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming, where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else, causing embarrassment or humiliation. Some cyberbullying crosses the line into unlawful or criminal behaviour [i.19].

The most common places where cyberbullying occurs are:

- Social Media.
- Text messaging and messaging apps on mobile or tablet devices.
- Instant messaging, direct messaging, and online chatting over the internet.
- Online forums, chat rooms, and message boards.

- Email.
- Online gaming communities.

This can be a displacement of problems onto social media. Social media has changed, but may not increase this problematic behaviour. A consequence of social media, such as the use of 'indirect messages' (messages that do not specify the person they are directed at), the 'hiding behind a screen', and the expansion of such behaviour from the playground to the home. This means people hide behind "anonymity" to harass people they may know in real life or strangers, or high public figures, believing they will not be caught and will not suffer repercussions. There are many countries which have which have either laws that explicitly focus on fighting cyberbullying, or it is covered by other legislation which targets harassment in general. The EU revised Audio-visual Media Services Directive (AVMSD) was adopted in 2018. It sets out requirements in relation to the provision of audio-visual media and addresses aspects such as the prohibition of hate speech, discrimination based on disability and the protection of minors. The Digital Services Act (DSA) regulates online platforms and intermediaries. The act intends to safeguard users by placing obligations on providers to address illegal content and harmful activities online and protect consumers' fundamental rights online [i.20].

In recent times, it has been observed that Generative AI allows for both the automatic creation of harassing or threatening messages, emails, posts, or comments on a wide variety of platforms and interfaces, and its rapid dissemination. Furthermore, generative AI algorithms can take online attacks to a higher level by learning from granular-level data available about a person. It can analyse a target's social media posts, online activities, or personal information to generate highly specific and threatening messages or content. It can create output that references specific locations, recent events, or private details about the target's life after learning as much as possible about them, making the harassment much more personal and intimidating.

There is an overlapping connection between disinformation, online harassment and cyberbullying. It is often at times challenging to distinguish between freedom of expression, criticism, and harassment, especially when disinformation takes place. For example, disinformation can be used as a harassment tactic: through spreading fake or manipulated content (e.g. deepfakes), by damaging someone's reputation or using fake information to incite mob harassment.

Furthermore, there is the phenomenon of trolling, which is connected to misinformation campaigns. For example, troll farms or organized groups may spread disinformation along with bullying and harassment to silence dissent or to manipulate public opinion. Targets of such campaigns are oftentimes journalists or activists. Additionally, groups may coordinate to overwhelm a target with fake information, threats or insults.

Ideally, platforms and networks should offer tools that allow the user to restrict who can comment on or view their posts or who can connect automatically as a friend, and to report cases of bullying. This should involve simple steps to block, mute or report cyberbullying [i.21]. Also, platforms should signpost users who are reporting this behaviour or being affected by it to resources such as https://findahelpline.com/.

4.3.2 Doxing/Doxxing

Doxxing (alternative spelling doxing) is the action or process of searching for and publishing private or identifying information about a particular individual on the internet, typically with malicious intent, without that individual's consent. People are 'doxed' for a variety of different reasons, and doxxing is often a way that online conflicts escalate into having real-world consequences [i.22]. Sometimes this violation of privacy is done as a form of revenge.

The publication of private or personal information about someone takes place without their consent.

This can be perpetrated against public figures or against private citizens and is usually done to harass, expose, or cause financial harm to the target. The private information shared without consent could include the victim's/survivor's real name, their email address, home address, or telephone number, or it could include documents/files, such as bank statements, ID documents, or even images/videos of the victim/survivor (including those of an intimate nature).

Sharing this information could result in wide-ranging consequences. For example, it could include someone signing another person up for different mailing lists to fill their email inbox with spam or encouraging strangers to send intimate images to the victim/survivor.

In the context of gender-based violence, doxing can be used as a way to perpetrate abuse and cause fear and alarm. Gender-based violence relies on power inequalities, which allow perpetrators to exert control. If an abusive relative, friend, or ex-partner no longer has access to someone after a relationship has broken down, they can look for ways to continue to exert control and continue the abuse, and they can use the victim's/survivor's private information as a way to do this.

For example, an abuser could create a fake profile or post on social media posing as their ex-partner, encouraging strangers to send inappropriate sexual communications to their ex-partner's real phone number or email address.

Doxing and disinformation are frequently used in tandem to cause harm. For example, disinformation can be used to justify doxing. False information about someone can incite outrage, which in turn is used to justify divulging that person's personal information (e.g. telephone number, home address, family details).

Disinformation renders doxing more dangerous and contributes to its more effective spreading. When divulging personal data is paired with false information, the risk of harassment and social or professional exclusion of the targeted person increases. Conversely, doxxing can be used to "prove" disinformation by releasing fabricated information to support a false claim, such as fake chat logs, manufactured emails, and misleading photographs.

There are limited options that a victim of doxxing can take. They rely on the victim or a trusted person acting on their behalf to take steps to stop or block the spread of the doxed information. While countries generally do not have a specific criminal offence of doxxing. However, depending on the facts and circumstances of the case, doxxing may amount to an existing criminal offence under laws covering offences, such as:

- Threatening and abusive behaviour
- Stalking
- Abusive behaviour towards a partner or ex-partner
- Improper use of a public electronic communications network

The removal of personally identifiable information or doxxing content from the internet involves contacting search engine providers, online platforms and social networks and requesting they remove the content, block access to it, or remove URL links to it. The use of GDPR's right to be forgotten' and using copyright law, for example, DMCA notice, and takedown process can aid in removing the doxed content.

4.3.3 Swatting

Swatting is the act of harassing or deceiving an emergency service (via such means as hoaxing an emergency services dispatcher) into sending a police or emergency service response team to another person's address. It is often linked to doxing. Swatting occurs when a hoaxer gathers information about their intended target, calls emergency services, and makes a false, urgent report. If the caller sounds credible and describes a dangerous scenario, dispatch may send a response to the target's location, leading to a potentially traumatic and dangerous interaction with the victim [i.1].

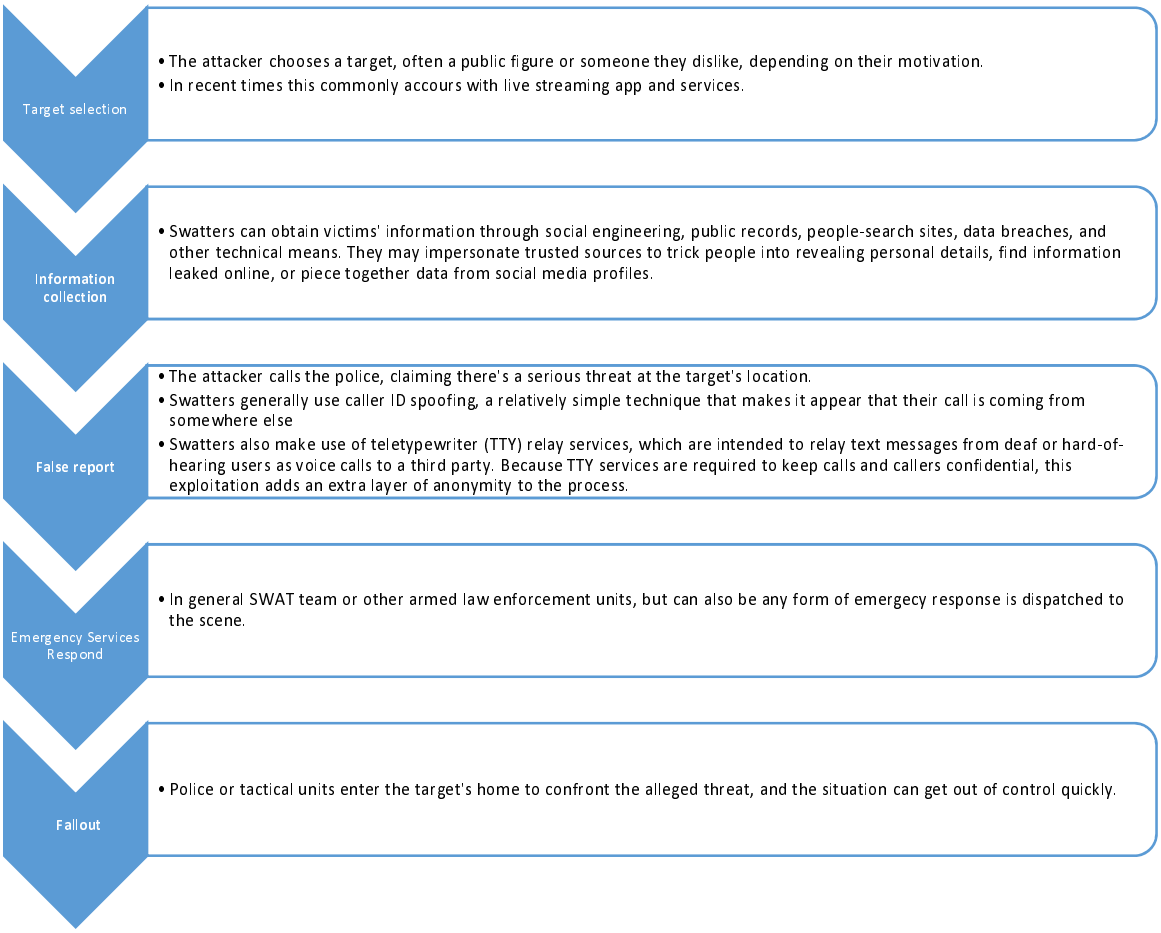

Figure 1: Outline of Swatting Steps

People typically swat others as a form of harassment, revenge, or intimidation [i.2]. Motivated by grudges, online disputes, or attempts to gain notoriety, swatters intend to frighten their victim and cause maximum disruption to their home life.

There are different reasons as to why swatters call in fake threats, a non-exhaustive list:

- Revenge: Swatting can be a form of retaliation against someone the attacker dislikes or feels wronged by.
- Intimidation: The goal may be to scare or intimidate the target, causing them to feel unsafe or uncomfortable.
- Thrill-seeking: Some individuals may like the adrenaline rush that comes with causing chaos in someone's life.
- Attention-seeking: Swatting can be a way to gain attention or notoriety, especially if the incident is widely reported in the media.
- Mental health issues: In some cases, the wrongdoer may have a mental health condition that contributes to their behaviour.

There is a connection between swatting and disinformation which is rooted in manipulation, malice and abuse of systems of trust, especially those involving emergency services.

Disinformation is embedded in the act of swatting as it involves falsely reporting a serious emergency, such as a hostage situation or an active shooter, prompting an aggressive law enforcement response to a person's home.

For example, disinformation can be used to make the swatting appear credible, such as telling the authorities that the attackers are holding hostages or using spoofed caller IDs to back up the false information. Conversely, individuals are often targeted for swatting after disinformation campaigns portray them as criminals, terrorists or extremists based on false information.

Furthermore, there is a link between doxing, harassment, swatting and disinformation. In such cases, swatting is an extreme escalation, following disinformation to build a narrative, doxxing to find the location of the victim and then swatting to cause terror or harm.

At the moment, the majority of prevention of swatting is placed on the targeted victim to take measures to minimize what information can be found online and/or used against them. As well as to contact if they are swatted or believe they are at risk of being swatted, to also proactively work with local law enforcement/police to prevent future escalation.

4.3.4 Online Grooming

It is an action or behaviour used to establish an emotional connection with a vulnerable person and sometimes the victim's family. This can be, for example, to extort them financially (more likely adults) or sexually abuse (more likely children). Groomers are skilled at deceiving others about their identity, particularly online, where they can create a false persona and pretend to be younger than they are [i.23]. People can be groomed online through:

- social media networks;
- text messages and messaging apps;
- email;
- text, voice and video chats in forums, games and apps.

The ways to prevent or stop it involve being able to identify the patterns of grooming behaviour. These are typically adults who are parents with children or teachers who have a duty of care. These include, but are not limited to:

- Are they being secretive about how they are spending their time?
- Do they have money or new things like clothes and mobile phones that they cannot or will not explain?
- Do they seem upset or withdrawn?
- Are they spending more time away from home or going missing for periods of time?

A person will not know they are being groomed, they will trust their abuser, who is giving them lots of attention and gifts. Also, their groomer may have warned them not to talk to anyone about it. Online grooming is linked to catfishing when a person uses false information and images to create a fake identity online with the intention to trick, harass, or scam another person.

There is an inextricable link between sexual grooming and disinformation because perpetrators use false information to manipulate, deceive, isolate and control their victims.

For example, disinformation can be used to manipulate the victim's perception through false promises to build trust and emotional dependence, or even to silence and control the victim. Additionally, disinformation can be used to create false personas, with groomers impersonating someone else by using fake profile information and photos. Another example of the link between disinformation and sexual grooming is when a victim tries to report the abuse, the perpetrators may spread disinformation to discredit them.

There are several methods that online services can take to prevent, detect and respond to online grooming [i.24].

Examples of Prevention include:

- Safety by Design:
  - A concept that describes how companies can build for user safety by anticipating how products and features might be used for abuse and incorporating safety features before product launch.
- Age Assurance:
  - The process by which a company estimates or verifies a user's age, including children. Companies could then offer custom experiences to users based on their age, for example, limiting a child's access to features that would introduce them to people they do not know or providing a default privacy setting that minimizes the reach of a child's post.

Example of Detection:

- User Reporting:
  - Enables users to inform a company when they encounter harmful behaviour, content, or other issues that may violate a company's policy or make the user feel unsafe. This can be a helpful channel for companies to gain insight into new trends that help improve detection. In some instances, a user report can be a powerful tool to help identify a groomer who has evaded detection and prevent further abuse. The challenge with user reporting is that it requires the user to recognize the abuse and feel comfortable reporting it. For grooming, it can be challenging to get minors to report online sexual interactions for various reasons.

Example of Response:

- Enforcement and Reporting:
  - Once a platform confirms suspected grooming, the company should take steps to prevent additional harm, for example, by disabling accounts, and possibly sending a report to the relevant law enforcement or other body (depending on the company headquarters or the location where the abuse occurred) if required due to the nature of the account being reported.

4.3.5 Sextortion

'Financially Motivated Sexual Extortion', a type of online blackmail often referred to as 'sextortion'. It is a type of online blackmail where malicious actors threaten to share sexual pictures, videos, or information about someone. They may be trying to take money from them or forcing them to do something else they do not want to. It is also linked to scam emails that send threatening messages but do not have real material; they rely on the threat online. Also, the rise of deepfakes means malicious actors can target someone with manipulated media.

A sextortion attack often has these steps:

- contacted by an online account the user does not know, or a hacked account of someone they do know, where the communication feels unfamiliar;
- quickly engaged in sexually explicit communications, which may include the offender sharing an image first;
- moved from a chat on social media, an online platform or a game to a private platform such as an end-to-end encrypted messaging app;
- manipulated or pressured into taking nude or semi-nude photos or videos;
- told they have been hacked, and the offender has access to their images, personal information and contacts (whether this is true or not);
- blackmailed into sending money or meeting another financial demand (such as purchasing a pre-paid gift card) after sharing an image, or the offender sharing hacked or digitally manipulated/AI-generated images of the victim to make the threat of sharing them wider.

There is a connection between sextortion and disinformation insofar as false information is used to deceive, coerce, manipulate or silence victims. For example, perpetrators use disinformation to trick victims into sending explicit content, by using fake identities, deepfakes or AI-generated profiles.

Another example is blackmail through false claims, such as the perpetrator threatening the victim that he has sent photos to her parents. Yet another example is when perpetrators spread disinformation to isolate or shame, such as by gaslighting victims that nobody will help them.

Victims of any age are potential targets; however, teenage males aged 14-17 and male adults aged 18-30 are particularly at risk.

Often, the way to prevent this relies on users knowing how to identify the patterns of what sextortion look likes before they become victims. If a user becomes a victim, they may feel distressed or blame themselves. It is important to remember they have been tricked or deceived; it is not their fault.

The applications and platforms where this occurs should have tools and measures to aid the victim. This includes reporting and blocking measures. The applications or platform should signpost the user to tools to help them remove any images or videos.

Examples of these include:

- https://stopncii.org/
- https://www.childline.org.uk/info-advice/bullying-abuse-safety/online-mobile-safety/report-remove/
- https://takeitdown.ncmec.org/

4.4 Dark Number/Dark Figure

The dark number or dark figure, which comes from criminology, is a term that is used by crime experts and sociologists to illustrate the number of committed crimes that are never reported or are never discovered, and this puts into doubt the effectiveness and efficiency of the official crime data [i.6]. Among the crimes that take place in any given place at a given period of time, some of them are never reported to the police, and some are reported but never recorded by the police officers.

There are different reasons that contribute to a dark number/figure for online incidents. One important fact that contributes to this is that members of the public fail to report incidents that they have witnessed being committed, as they believe that they are insignificant and do not regard them as significant enough to report. Others fail to report cases because the victims may consider the crimes embarrassing, for example, being scammed. They therefore choose to keep to themselves and suffer in silence. One key reason as to why incident is not reported when they are perpetuated is that the victim may not know that they are a victim. For example, in the case of fraud, individuals fail to recognize that are defrauded. Also, there are cases where those who are victimised may not be in a position to report or even realize it. This is especially true in the cases of crimes that are perpetrated against children or the elderly. They either fail to recognize the fact that they are being victimised or cannot report it. For example, when children are victims of abuse, they usually fail to report it as they are scared. Other reasons for failure of reporting of cases are lack of trust with the official bodies due to previous experiences, fear of victimisation and rough justice, where the victims may decide to take the law into their own hands.

It can be useful for companies to understand dark numbers/figures as they indicate how effective their reporting and recording systems are. Also, by understanding differences in the known reported numbers compared to dark numbers, one can show how effective their measures are in preventing things from occurring or if there is a potential gap in their measures.

5 Factors Affecting Harms and Risks

5.1 Scale, Dominance and Presence

There are two key scales. The first is the scale from the most private to the most public. The second is the scale from the smallest group to the largest group. At one end of both of these scales, there are private dyadic conversations and at the other end, there is fully public broadcasting [i.5].

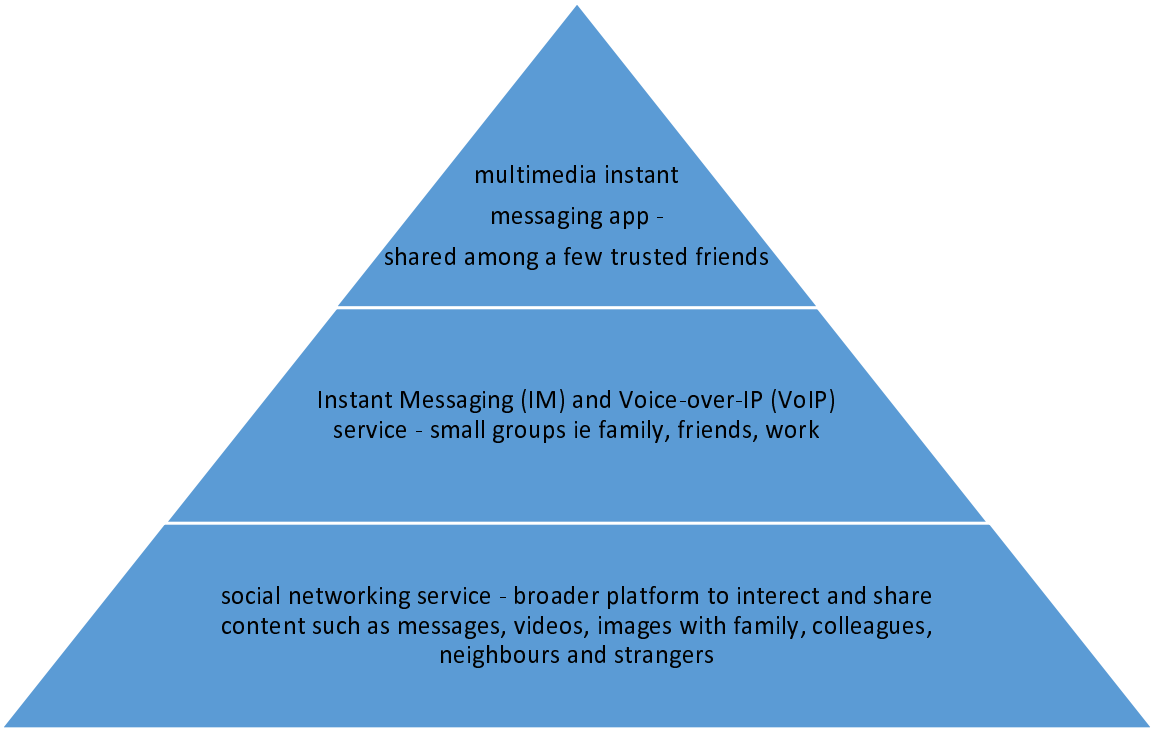

Figure 2: The scales of social media

Social networking and messaging platforms have various uses, some of which are gained through user actions and may not have been the original intent of the designers. Content migrates easily between platforms. For example, the genre of interaction among school children started as BBM on BlackBerry® phones, then moved to Facebook®, and is now on X®, formerly Twitter®, and will likely move on to different platforms in the future.

Social media is a technology that affords 'scalable sociality'. This means that social media provides greater control in communication over both the degree of privacy and size of the group when compared with previous forms of communication media.

5.2 Devices and Apps

The number of internet-connected devices a person or household uses can range from the single digits to potentially the hundreds if they have a smart home with many IoT devices. Though these will be part of the furniture and fittings of the house. A small number of devices would be used daily, such as smart TVs and speakers, laptops and tablets etc. How people use these devices will vary by the role they serve and the apps they choose to use on them. For example, one person uses a laptop mainly for work and e-mails, leaving direct messaging and social media on their smartphone. Another person does all their communications, emails, social media, and direct messaging on their smartphone while using a tablet for media consumption.

There are different and a broad range of risks which are affected by user behaviour of their devices and apps. For children, this could mean safety and privacy risks through social media and direct messaging apps. Which can be mitigated through the application of parental controls.A working adult could be at risk from phishing or social engineering, which can be mitigated through training.

There are different attitudes to how people share devices and apps within families, and generally, differences between different social and cultural groups. This means some people happily share passwords and accounts while others do not. Or people divide the role of managing services online. For example, one person in a family has control over all banking and money and another person has control over messaging and social media. This can either be an acceptable arrangement within a family or it can be a sign of coercive control.

To keep people safe and secure requires them to have awareness of risks about privacy, personal data and cybersecurity. This may include alerts and news, but at the same time, it cannot be so constant that people experience fatigue from it and tune it out. Those measures should be the final step, with secure by design and privacy by design being the first steps for user devices and apps.

5.3 Algorithms

Algorithms in social media platforms can be defined as technical means of sorting posts based on relevance instead of publication time, and to prioritize the content a user sees first according to the likelihood that they will engage with such content [i.25].

The function of an algorithm is to deliver relevant content to users. A reason why social media platforms use algorithms is to more organically filter through the large amount of content that is available on each platform. Algorithms prioritize delivering content that is potentially more "interesting" to a user, potentially at the expense of posts deemed irrelevant or of low quality, either in general or to a specific user.

Algorithms are designed in a way that considers different aspects. Some of these aspects are content-based, meaning that this kind of algorithmic design seeks to match a user's taste, based on their profile, to specific posts that the system guesses the user will like. Once users show interest in a specific tag or category, they are directed to other items in the same category. These are designed to keep the user engaged within the social media platform.

This means social media platforms are generally proficient at holding users' attention with hyper-personalized content. This is a reason why algorithms contribute to the spread of misinformation and extreme content. This can create negative cycles affecting a person's mental health [i.26]. The constant bombardment of curated content can have severe repercussions on people's mental health. From increasing levels of anxiety and depression to fostering body image issues, the algorithmic choices are not always in the best interest of the users. The more they engage with content that triggers negative emotions, the more such content they'll see.

This also contributes to the spread of misinformation and/or disinformation as algorithms prioritize content that garners high engagement, not necessarily content that is factual or beneficial. This can lead to the creation of echo chambers where users are only exposed to opinions and information that align with their existing beliefs, which leads to polarized viewpoints, meaning they cannot or will not accept differing views and impacts their cognitive processes as it leads to users having a confrontational view of the worlds either people are with them or against them.

Hate and disinformation algorithms refer to the way algorithmic systems amplify or prioritize harmful, hateful or false content. These algorithms incentivise and spread hate and disinformation because they are designed to optimize engagement, attention and profit.

For example, false content and hate speech tend to generate more clicks, more comments and more shares and algorithms reward this engagement, even though the content is harmful or false.

Moreover, algorithms show users more of what they already engage with, so if they are interacting with hateful or false content, they are likely to be shown more of that content, which can lead to progressive radicalisation, especially when it comes to young or vulnerable users.

Given that platforms rely on automation, harmful content can go viral before it is flagged and removed. By the time it is removed, algorithms may have already caused harm, such as election interference, vaccine misinformation or mob harassment.

A measure to mitigate the harm from algorithms is ensuring the digital literacy of users [i.27]. This includes user awareness and control. Knowing and understanding how these algorithms work can help people make more informed choices about their social media use. As well as encouraging the prioritization of their mental well-being by customizing their feed, setting time limits, and being mindful of their emotional responses to the content they consume.

5.4 AI LLM Risks

An LLM is where an algorithm has been trained on a large amount of text-based data, typically scraped from the open internet, and so covers web pages and, depending on the LLM, other sources such as scientific research, books or social media posts [i.28]. This covers such a large volume of data that it is not possible to filter all offensive or inaccurate content at ingest, and so 'controversial' content is likely to be included in its model.

The algorithms analyse the relationships between different words and turn them into a probability model. It is then possible to give the algorithm a 'prompt' (for example, by asking it a question), and it will provide an answer based on the relationships of the words in its model.

LLMs have the ability to generate a huge range of convincing content in multiple human and computer languages. However, there are different risks associated with LLMs [i.29], [i.30]. These include:

- Risk of Indirect Prompt Injection: the insertion of malicious information into the data sources of a GenAI system by hiding instructions in the data it accesses, such as incoming emails or saved documents. Unlike direct prompt injection, it does not require direct access to the GenAI system, instead presenting a risk across the range of data sources that a GenAI system uses to provide context. Mitigation of the risk of indirect prompt injection includes data quality controls, conscious management of access to data, clear user education on the safe use of tools and continuous monitoring to detect suspicious behaviour.

- Chat Bots: use AI and natural language processing to understand customer questions and generate natural, fluid, dialogue-like responses to their inputs. Generally, chatbots are sensitive to the form and choice of wording. Thus, depending on how a prompt is phrased, the results might vary. Similarly, if questions or prompts are posed that have no definitive answers, the technology's responses might fluctuate. This risks the chatbot providing hallucinations, which are answers which are either nonsensical, misleading or wrong which, depending on the context, may give a user dangerous information if acted upon.

- Risk of personal data being exposed: Each stage of an LLM's development lifecycle could introduce potential privacy risks, as the model interacts with large datasets that might contain personal data, and it generates outputs based on that data. Some of the key privacy concerns may occur during:

    - The collection of data: The training, testing and validation set could contain identifiable personal data, sensitive data or special categories of data.

    - Inference: Generated outputs could inadvertently reveal private information or contain misinformation.

    - Retrieval-Augmented Generation (RAG) process: It might use a knowledge bases containing sensitive data or identifiable personal data without implementing proper safeguards.

    - Feedback loops: User interactions might be stored without adequate safeguards.

Manipulated AI LLM Information Ingestion refers to the intentional feeding of false, biased or malicious information into the data pipelines or environments that Large Language Models (LLMs) learn from or are influenced by either during training or post-deployment.

If the input data includes manipulated or misleading content, the model learns and reproduces falsehoods or amplifies biases. An example of a dataset that might feed an algorithm are fake news websites designed to look like reputable sources.

Moreover, LLMs can absorb manipulated information post-training, from user interactions, web-browsing tools, APIs or plugins, fine tuning or RLHF (Reinforcement Learning with Human Feedback) pipelines. This can influence the model's factual reliability.

Safeguards (or guardrails) in LLMs are mechanisms implemented to ensure that the models operate in a safe, ethical, and reliable manner and have the potential to reduce risk. They can be applied to various stages of the LLM pipeline (preprocessing, training, and output) and are focused on addressing different risks. For instance, some safeguards aim to avoid the generation of unethical, harmful or inappropriate content (so the behaviour of the model), while others focus on preserving the privacy of the owners of the data (or other stakeholders).

Different types of behavioural guardrails aim to moderate the LLM's output and mitigate harm that could be caused by the output without intervention. These include, but are not limited to:

- Content filters: moderate outputs by blocking or flagging harmful or toxic content.

- Prompt refusals: prevent responses to dangerous or unethical prompts (like a request for instructions to a successful robbery).

- Bias mitigation: Reduce stereotypical or unfair outputs during inference.

- Human-in-the-Loop approaches: human oversight for high-risk applications, to not leave important decision-making fully in the 'hands' of an automated system, which cannot truly comprehend what is at stake.

- Post-processing detoxification: filter or rewrite outputs to remove harmful content.

- Adversarial testing (red teaming): evaluate and stress-test the model's ability to successfully deal with harmful prompts.

5.5 The Theory of Polymedia

In the past, the reason people chose one media rather than another was usually a question of either cost or access. Today, societies have a much wider range of communicative tools at their disposal, for example, smartphones, fixed telephones, text messages, broadband internet, email and all kinds of social media platforms. Also, the choice now has more to do with whether the user wants to avoid a quarrel, use the platform they feel more confident about, or avoid a certain audience. But as a result, they may be judged on which platform they choose. So, Polymedia is a theory of how media choice has become more of a social and moral issue instead of one of cost and access [i.31].

Until not so long ago, most people wishing to communicate at a distance had a limited choice of media at their disposal, mainly expensive international phone calls or letters. As a result, the choice of medium was largely the result of constraints of access and cost. The proliferation of new communication technologies and the increased convergence that society has witnessed in the last few years are radically transforming interpersonal communication at a distance. This coincides with an increased demand for mediated communication, given the rise in global migration and flows of human capital. Once users have obtained either a computer or a smartphone, and once the hardware and connection costs are met, then the cost of each individual act of communication itself becomes largely inconsequential. So today a typical urban young adult of a lower to middle-class income in many parts of the world can choose between calling though a landline, mobile phone or Voice over Internet Protocol (VoIP) through applications, with or without webcam; alternatively, users can send a text or an email, use Instant Messaging (IM) or a variety of social networking applications.

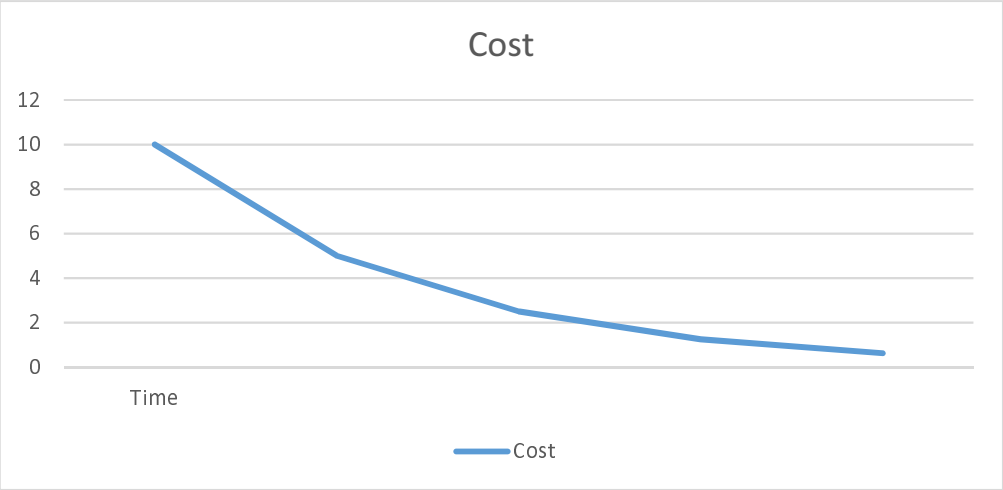

Figure 3: Cost and barriers to access decrease

The theory of polymedia tries to show how, in a situation where people have the option of many different media to communicate with, the platform people choose can have significant consequences for their social relationships. It means that to understand any single communication channel, organizations should think about the whole media package available to the person, typically the end-user.

The cost of computing has also contributed to increased access, as over the past 30 years, the average cost of a computer has significantly decreased due to advancements in technology and increased competition. A computer that cost hundreds or thousands of euros in the early 1990s can now be purchased for a fraction of that price. This also meant most households might have only a single computer. During the 2000s, low-cost but still capable computers began to emerge, costing in the low hundreds, coupled with increased digitization and common use of computers at schools and the workplace, led to more households beginning to have multiple computer devices, especially as laptops became popular and affordable. This was coupled with the low-cost but also increasing speed of broadband internet access, which became commonplace, allowing multiple people in the same household to access online services and communities. A key part of this trend is the rise of the smartphone as an affordable, convenient platform to access online services. This was also aided by the rise of free applications and free access to platforms which provided messaging and social networking functions. The smartphone also saw the rise of cheap data access to the internet through mobile networks.

This means a user typically makes use of multiple devices and services. For example, a family might have their own devices for work and home use, shared devices like a tablet for consuming media, and each adult likely has at least one smartphone and maybe more if they also have a dedicated work phone. The children likely have a computer for homework, smartphones, tablets for media consumption, game consoles, etc. So, risk and harm can come from any device, such as phishing emails on computers, grooming through direct messaging apps on smartphones or tablets, accessing inappropriate content, or anything with an internet connection and a screen. Keeping people safe requires understanding what types of devices and services they are using to provide the best advice to them, but also understanding where they are using them. For example, in a home environment, people may act differently and take different actions compared to being out in public. For example, a person who only accesses internet banking at home and not in public, as they believe all public Wi-Fi® is unsafe.

5.6 Demography - Understanding how different people use apps and devices

5.6.1 Introduction

This present clause aims to provide a broad understanding of how different people, divided into age ranges, use apps and devices. This is a constantly changing area, and what is true today might not be true tomorrow. Also, this clause does not cover the other factors that can affect how different users use apps and devices, as there are differences for genders/sexes, ethnicities, nationalities, etc. It is also important to note that different groups are affected differently by the harms that the present document will include.

5.6.2 Ages 6 and Under

This includes children who are in pre-school and at the beginning of the primary education level. Preschoolers learn by interacting with the world around them. Parents are using digital devices with their young children both to support learning and to keep them quiet. They are also using the internet and social media to get ideas for learning and play activities. If very young children are supported by adults when they use digital devices, then they soon become able to use them independently for some tasks, including talk and literacy-related activities, such as recognizing and purposefully selecting screen symbols and icons. They develop the technical and motor skills needed to operate touch screens - to swipe, locate, and tap to select and use apps such as the camera and photo gallery to take and view photos and videos, or to access online media. Unsupervised use has a risk of young children being exposed to material they are not developmentally ready for. Or when using a parent's device unsupervised, making accidental unauthorised purchases either via touchscreen apps or voice-controlled smart audio devices.

5.6.2 Ages 7 to 11

This includes children who are at the primary education level. At this age, most children use apps and websites for entertainment and begin to use devices for schoolwork. Typically, children will have their first user experience with a parent's computer, tablet, smartphone, etc. designed and sized for adult users. At this stage in life, there is a risk that if a child has unsupervised internet access, there is a chance of them being exposed to themes they are not developmentally ready for. Information received by children beyond their experience will be interpreted within the framework of their experience, sometimes causing great anxiety and distress. For example, overexposure to themes such as war and sex may cause great difficulty later on, as children have to reinterpret knowledge [i.32]. Online activities include watching multimedia content on video-sharing platforms, playing video games online, and some are beginning to use social media platforms, even though most have a minimum age of 13. Along with using education apps to supplement school learning.

5.6.3 Ages 12 to 18

These are children who are in secondary school, going through their teenage or adolescent years, which brings many changes not only physically, but also mentally and socially. As children transition from primary education to secondary, they experience many digital firsts, with many getting their first smartphone [i.33]. During these years, adolescents increase their ability to think abstractly and eventually make plans and set long-term goals. Each child may progress at a different rate and may have a different view of the world. This can be a highly emotional time. They are coping with hormonal changes brought about by puberty. They are also hitting a time when peers will have the most influence on them. This is a period of life when some kids start dabbling in riskier behaviour (self-harm, smoking, drug use, sex, etc.). This can be exacerbated by access to certain types of internet platforms and services, which can include social media, photo/video sharing platforms and instant messaging services. Teenagers essentially communicate as adults, with increasing maturity throughout their secondary education. As teenagers seek independence from family and establish their own identity, they begin thinking abstractly and become concerned with moral issues. All of this shapes the way they think and communicate. During this time, they will begin to interact and consume social media along with online content, which may influence the development of their identity.

5.6.4 Ages 19 to 29

This age range includes people in tertiary education, who are beginning to start their working life. During this time, people are now legally adults (the age of majority), though the age varies from country to country. In the majority of places, the age is 18, but in some countries, it is younger, while in others, it is older. Also, some countries have different ages within them and are often separate from the legal age to carry out certain activities, such as buying alcohol or gambling, etc. This age range is considered digital natives; they grew up with smartphones and the internet. Generally, they prefer apps that provide visual stimulation, and this group consumes the most video content. Within this age group, but not limited to, there is a trend of increasing use of generative AI tools, with the main reasons being as a way to save themselves time and to improve the quality of their work. Though there is an acknowledgement of the risk of getting false or biased results [i.34].

5.6.5 Ages 30 to 55

This age range is broad and diverse. Generally, people in this age will have their careers and have their own families. Generally, use a wide variety of apps and services for work and personal use. They will be the likeliest to respond to push notifications. They may prefer personal interaction rather than visual stimulation when compared to younger users of apps. They may also have a burden of responsibility of raising their children while at the same caring for older relatives and parents. There is also the rapid change and use by employers to use AI applications to sift through CVs, interact with customers, allocate, direct, and evaluate work, and identify and provide training. Also, workers are using AI in an increasing number of tasks. Depending on the type of work, it is being used to improve productivity and reduce the burden of certain tasks. It is important to note that AI may bring increased work intensity and the collection and use of data, amongst others plus the risks include: bias and discrimination, unequal impact on workers, lack of human oversight, as well as lack of transparency, explainability and accountability, will not just this age groups but all the other age groups as well to different degrees [i.35].

5.6.6 Ages 56 to 75

This age range tends to be nearer the end of their working life and beginning to look towards retirement. They may be more financially secure, for example, their mortgage is paid off, and their children have left home. They generally use apps and services that help maintain a better quality of life, but not all in this age range may be adept at adjusting to new interfaces with the continuing digitalisation of key services necessary to live life, so they are potentially more vulnerable to being left behind as services become harder to access when locked behind a digital wall. Online risks for this age group will differ from others, such as coming up to retirement, they may be targeted in pension scams, while at the same time, they face the same risks of being targeted for romance or sextortion attacks that can affect all ages.

5.6.7 Ages 76 and Older

While this age group is increasingly online, there are still people who do not use the internet and do not have a smartphone. Digital technology is playing an increasing role in our lives, and for many people, it is essential to the way that they socialise, work, shop, manage their finances, access services, and get entertainment. For example, smart speakers are gaining popularity with this demographic. They can perform tasks like playing music, reporting the weather, and making calls via simple voice commands, which can be useful for those with mobility issues. However, not everyone is online, while others only use the internet in limited ways. Although many older people fully embrace the digital world, digital exclusion increases with age. There is also a divide in the skills of this age group. So, some may be completely comfortable and know how to stay safe online, for example, recognize suspicious links and know that clicking on these links or downloading unfamiliar attachments is a risk (e.g. spam/phishing emails, texts, pop-ups) others do not have this knowledge or only a limited amount. Most older adults require help setting up and learning new devices. Access to training and support, whether from family, friends, or dedicated programs, is a critical factor for ensuring they are comfortable using smart devices and apps. This age group will also have a higher prevalence of memory disorders, for example, mild cognitive impairment, Alzheimer's. age-related forgetfulness and dementia, etc. This puts them at greater risk of being easy targets for online attacks such as phishing, scams, fraud, and identity theft.

6 Preventive and Mitigation Measures

6.1 Regulatory Conformance

6.1.1 Introduction

There are different regulatory pieces at the time this present document is published that companies, organizations and service providers have to conform to. See Annex A. These relate to online safety, disinformation and harms that affect users online. They have to take action to mitigate the impact of disinformation, prevent minors from accessing inappropriate content, with age assurance being a primary tool to achieve this and remove harmful content quickly from their platforms.

6.1.2 Conforming to The Code of Conduct on Disinformation

Platform services and providers of social networks, messaging apps and content-sharing platforms can potentially reduce the spread of disinformation by conforming to the measures in The Code of Conduct [i.36], [i.37], [i.38], which aims to combat disinformation risks while fully upholding the freedom of speech and enhancing transparency under the Digital Services Act (DSA). The key areas the code of conduct covers include:

- Demonetisation: cutting financial incentives for purveyors of disinformation.
- Transparency of political advertisement by recognizing the importance of political advertising in shaping public opinion.
- Ensuring the integrity of services by reducing manipulative behaviour used to spread disinformation, e.g. fake accounts, bot-driven amplification, impersonation, malicious deep fakes.
- Empowering users to navigate safely and make informed decisions, including by providing them with more context and information about the content they are seeing.

- Empowering Researchers enabled by online platforms provides better support to research on disinformation, including by giving researchers better and wider access to the platform's data.
- Empowering the fact-checking community on platforms by extending fact-checking coverage across all EU Member States and languages, and to make a more consistent use of fact-checking on their services.

The Code of Conduct on Disinformation 2025 [i.36] sets out 43 commitments and 128 specific measures relating to disinformation. The different measures that are recommended include:

- **Commitment 22:** Relevant Signatories commit to provide users with tools to help them make more informed decisions when they encounter online information that may be false or misleading, and to facilitate user access to tools and information to assess the trustworthiness of information sources, such as indicators of trustworthiness for informed online navigation, particularly relating to societal issues or debates of general interest.
- **Measure 22.6**: Relevant Signatories providing trustworthiness indicators by means of voluntary, self-regulatory and certifiable European standards or European standardization deliverables as defined by European law ('technical standards'), such as the CWA17493:2019 (Journalism Trust Initiative) [i.39] will:
  - develop and revise them based on internationally accepted best practices and ethical norms;
  - make them publicly available and accessible in a non-proprietary, neutral way;
  - govern their implementation in line with European Accreditation and EU Regulation (EC) No 765/2008.

6.1.3 Age Assurance Conformance

6.1.3.1 Introduction

Age assurance methods are used to determine the age or age range of an individual, including age verification, estimation, and self-declaration. They are required for online platforms and services to comply with regulations like the UK Online Safety and the EU Digital Service Act, to ensure users, particularly underage users (children), are not exposed to or can access inappropriate content. There are different methods and schemes available to conduct age assurance. There are also expectations regarding data protection and privacy [i.40].

6.1.3.2 Age Verification

Age verification is any method designed to verify the exact age of users or confirm that a user is over 18.

There are different approaches to age verification:

- Verifying the user's age through scanning a 'hard identifier' such as a driver's license or passport.
- Verifying a person's age through a third-party provider, which can use a range of information sources (e.g. credit card information, banking information or voter registration records).

Platforms should ensure that the amount of personal information they collect about a person to verify their age is proportionate to the risks that their service poses.

Age verification does not always require platforms to collect and store large amounts of personal information. They may be able to verify a user's age without directly collecting their actual age or date of birth.

6.1.3.3 Age Estimation

Age estimation is any method designed to estimate the age, or age range, of a user, often by algorithmic means.

Platforms could use age estimation approaches for initial onboarding or account creation, or for ongoing monitoring. These approaches estimate the age of a person, rather than confirming whether someone is a specific age (e.g. through documentary evidence or a trusted third party). As they do not require documentary evidence, platforms could find this a more privacy-friendly method than using hard identifiers.

Age estimation systems use a mix of methods, including:

- A computer vision-based approach - this estimates age from an image of the person. The image may be captured in real time by a mobile device camera or webcam. Facial age estimation has seen significant progress and is now the most widely used age estimation approach. It has high levels of reported accuracy and efficacy, albeit with variances in relation to skin tone, sex and age.
- Other biometric approaches - such as voice analysis to estimate a person's age. This area is continuing to develop, with other biometric approaches launched to market recently and achieving accreditation. Whilst the efficacy of these products is improving, currently they tend not to reach the higher levels of accuracy that would make them appropriate for high-risk scenarios.
- Analysing account profiling or information - information derived from the person's activity on the platform. This may include analysing their digital footprint, which looks at their interaction or accounts across many different sites. This may be via a person's email address or mobile phone number, for example. It can also include analysing on-site behaviour once a person is using a service, such as activities, content choices, or friends that suggest the person is below the minimum age of the terms of service. The efficacy of these methods varies.

6.1.3.4 Self-declaration

Self-declaration is a method where a user states their age but is not required to provide evidence to confirm it. It is a popular approach because there are relatively few steps to follow, and because it requires minimal personal information. It often takes the form of a tick box to self-affirm that the person meets the age requirements in the terms of service.

It is not considered a reliable method under the Online Safety Act or the Digital Services Act. This is because it is based entirely on trust and can be easily circumvented and therefore does not significantly mitigate risk. Platforms should avoid using a self-declaration age assurance method as it is unlikely to be accurate and effective if:

- there are significant risks to children from the data processing on their site; or
- they are choosing to restrict access to underage users from an adult site.

Self-declaration can be minimally intrusive, and platforms could consider using it for online activities which do not pose a high risk to children, or in conjunction with other methods.

6.1.3.5 Waterfall techniques and age buffers

The waterfall technique combines different age assurance approaches. Waterfall techniques build on the output of successive age assurance approaches to provide a cumulative result with a greater level of confidence than any of these approaches in isolation.

Waterfall techniques have the potential to offer high levels of confidence while providing a privacy-respecting approach for users.

A common example is if platforms combine an age estimation method with a secondary age verification method when they require a high level of assurance.

Some age estimation methods can provide a high level of assurance where the person is clearly over the age threshold. For example, when someone over 40 is looking to access a service for only those over 18 years of age.

The potential for errors may increase for people who are closer to a set threshold (i.e. the risk of a 16-year-old receiving an estimate they are 18, or a 19-year-old receiving an estimate they are 17).

Platforms could apply an age buffer. This means that a person who is close to the minimum age required to access the service would be required to complete a further age check, using an age verification method.

A use-case scenario for a waterfall technique requiring people to establish they are 18 or over could involve the following:

- An age estimation method is deployed with a buffer of plus seven years.
- All people reported as over 25 passed without further checks.

- All people identified as being under 25 are referred to a secondary age assurance method (i.e. a choice of credit card check or production of official ID, or mobile phone check).

If platforms choose to use a waterfall technique, they should allow people to challenge the decision.

If platforms are relying solely on automated decision-making, depending on the impact of that decision on the person, there may be additional data protection requirements.

Platforms need to carefully design waterfall techniques to ensure they achieve increased accuracy whilst preserving privacy. A poorly designed waterfall technique risks collecting unnecessary information, which provides little additional age assurance. This may result in an unjustified level of privacy intrusion, which risks non-compliance with the data minimization principle of GDPR.

6.1.3.6 EU Age Verification Solution

The European Commission is developing a harmonised, EU-wide approach to age verification, accompanied by a comprehensive age verification blueprint that is intended to facilitate practical adoption across all Member States and can be customized to the national context [i.42]. This approach defines the structural and functional aspects of the age verification solution for online services. It details the key components, interfaces, and interactions within the age verification ecosystem, providing a technical foundation to ensure interoperability, security, privacy and compliance with applicable regulatory and industry requirements [i.43].

It includes:

- The end-to-end process for issuing, presenting, and verifying Proof of Age attestations for access to online services, including age-restricted content, products, or services.
- The operational, security, product, and architectural requirements necessary for a harmonised and scalable solution that can be adopted and extended by Member States or other actors.
- The definition of system boundaries includes the roles of Attestation Providers, Relying Parties, and Users, as well as the interactions between these entities.
- The technical requirements for device compatibility focus on enabling Users to present Proof of Age attestations across commonly used devices such as mobile phones, tablets, laptops, and desktop computers.
- The specification of interoperability mechanisms to support integration with existing and future digital identity solutions, including the planned incorporation into EU Digital Identity Wallets.

The EU Age Verifcation solution does not cover the internal implementation details of third-party Relying Parties, the development of alternative age verification methods beyond those explicitly described, or the detailed integration of operational systems required for an enrolment, as this is subject to country-specific requirements. It serves as a reference for the harmonised implementation of age verification solutions across the European Union.

The EU Age Verifcation solution builds upon the Architecture and Reference Framework (ARF) for the European Digital Identity (EUDI) Wallets, adopting the same foundational technical standards and design principles to ensure interoperability, security, and privacy within the EU digital identity ecosystem. However, it defines only a subset of the requirements and functionalities compared to a full EUDI Wallet implementation, focusing specifically on the needs of age verification solutions. The objective is to provide a compatible architecture that facilitates seamless integration with EUDI Wallets as the ecosystem evolves.

6.1.4 Digital Content and Services Conformance

6.1.4.1 Introduction

Recent legislative changes, the EU Digital Services Act [i.44]. and the UK Online Safety Act [i.45], have introduced stronger legal obligations on online service and content providers to provide a safer online experience for all users. This includes removing harmful and illegal content. In the EU, there is a stronger public oversight mechanism for online platforms, in particular for those Very Large Online Platforms (VLOPs) which reach more than 10 % of the EU's population.

6.1.4.2 EU Digital Service Act (DSA)

6.1.4.2.1 Reporting of illegal Content

The DSA requires platforms to put in place measures to counter the spreading of illegal goods, services or content online, such as mechanisms for users to flag such content and for platforms to cooperate with "trusted flaggers".

6.1.4.2.2 Transparency in content moderation and options to appeal

Online platforms are a digital space where users express themselves, showcase their work, and are in contact with friends or customers. This is why it is particularly frustrating when their content gets removed or the reach of their posts is inexplicably reduced. With the DSA, providers of intermediary services, including online platforms, have to communicate to their users why they have removed their content or why access to an account has been restricted. Providers of hosting services, including online platforms, now have an express legal obligation to provide clear and specific statements of reasons for their content moderation decisions. The DSA also empowers users to challenge such decisions through an out-of-court dispute settlement mechanism.

6.1.4.2.3 Control on personalization options

The DSA obliges providers of online platforms to guarantee greater transparency and control over what users see in their feeds. This should allow them to discover on what basis online platforms rank content on their feeds and to decide whether users want to opt out of personalized recommendations, since VLOPs have to offer an option to turn off personalized content. Similar obligations apply to ads: in addition to further transparency and control on why users see a certain advertisement on their feed, platforms need to label ads, and VLOPs have to maintain a repository with details on paid advertisement campaigns run on their online interfaces.

6.1.4.2.4 Zero tolerance on targeting ads to children and teens and on targeting ads based on sensitive data

The DSA bans targeted advertisements to minors on online platforms. Targeted advertisement on online platforms is also prohibited when profiling uses special categories of personal data, such as ethnicity, political views, and sexual orientation.

6.1.4.2.5 Protection for minors

Under the DSA, online platforms that are accessible to minors should protect the privacy and security of those users, as well as their mental and physical well-being.

The European Commission has adopted guidelines for protecting children and created a blueprint for an age verification solution, see clause 6.1.3.6.

The guidelines address issues such as addictive design, by disabling features such as 'streaks' and 'read receipts' to curb excessive use and combating cyberbullying by empowering minors to block users and preventing unwanted content downloads. They also aim to mitigate the impact of harmful content by giving minors more control over recommendations and promoting private-by-default accounts to prevent unwanted contact from strangers.

The guidelines also recommend the use of effective age assurance methods online, provided that they are accurate, reliable, robust, non-intrusive, and non-discriminatory. In particular, they recommend that age verification measures be put in place for adult content, such as pornography or gambling, or when national laws set a minimum age for social media.

6.1.4.2.6 Integrity of Elections

The DSA requires VLOPs and VLOSEs to identify, analyse, and mitigate with effective measures risks related to the electoral processes and civic discourse, while ensuring the protection of freedom of expression.

6.1.4.2.7 Obligations on traceability of business users in online marketplaces

The DSA introduces obligations for providers of online marketplaces to counter the spread of illegal goods. In particular, such providers have to ensure that sellers provide verified information on their identity before they can start selling their goods on those online marketplaces. Such providers have to guarantee that users can easily identify the person responsible for the sale. Moreover, if a provider of an online marketplace becomes aware of the selling of an illegal product or service by a seller, it has to inform the users who purchased the illegal good or product, as well as the identity of the seller and the options for redress.

The EC, along with Member States, will supervise and enforce compliance with the obligations against the spread of illegal goods on online marketplaces.

6.1.4.3 UK Online Safety Act

6.1.4.3.1 Illegal content safety duties for user-to-user services

"The provider of a user-to-user service shall":

- *take proportionate steps to prevent their users from encountering illegal content*
- *mitigate and manage the risk of offences taking place through their service*
- *mitigate and manage the risks identified in their illegal content risk assessment*
- *swiftly remove illegal content when the provider becomes aware of it, and minimize the time it is present on their service*
- *explain how, as the provider, this will be done in their terms of service*
- *allow people to easily report illegal content and operate a complaints procedure*

6.1.4.3.2 Illegal content safety duties for search services

"The provider of a search service shall":

- *take proportionate steps to minimize the risk of their users encountering illegal content via search results*
- *mitigate and manage the risks identified in their illegal content risk assessment*
- *explain how they will do this in a publicly available statement*
- *allow people to easily report illegal content and operate a complaints procedure*

6.2 Threat, Vulnerability and Intelligence

6.2.1 Overview

A threat database is often a structured collection of information about threats to a system or organization, frequently used to identify and mitigate vulnerabilities. More narrowly, it could also refer to a specific security application that maintains a database of known threats, such as malware or malicious IP addresses, with the goal of protecting a service and/or organization. In a broader national security context, it can refer to an intelligence dynamic database containing information on adversaries, potential attacks and malicious URLs. This may also include the ability to identify zero-day attacks. A vulnerability database is a centralized, curated collection of information on known security flaws in software, hardware, and systems. These databases describe the vulnerability, assess its potential impact, provide affected product details, and suggest mitigations like patches or updates. They are resources to track and correct security weaknesses, prioritize vulnerabilities, and prevent cyber-attacks. Intelligence for cybersecurity is the process of collecting, analysing, and applying data on cyber threats, adversaries, and attack methodologies to enhance an organisation's security posture. It involves taking raw threat data from various sources and transforming it into actionable insights that enable organizations to anticipate, detect, and respond to cyber risks.

6.2.2 Application to Online Preventive Security to Protect Users

There are threat, vulnerability databases and intelligence sources that apply to online preventive security, which monitor and track threats and activities that can affect ordinary users [i.46], [i.47]. The types of threat information which can aid in protecting users, either directly or the services they use, include, but are not limited to:

- collections of anonymous or disposable email domains commonly used to spam/abuse services
- data of honeypot activity
- list of malicious domains that also perform reverse lookups and list registrants, focused on phishing, trojans, and exploit kits
- email addresses used by malware
- law enforcement resources on cyber threats and internet crime
- cybersecurity data and visualizations, such as tracking ongoing DDoS attacks

6.3 Online Preventative Security by Design

6.3.1 Introduction

In the present document, Human-to-Human Online Preventive Security by design is a proactive approach to security and safety that integrates threat identification, risk assessment, and security controls directly into the design and development process of a service and/or a product. This includes putting user safety and rights at the centre of the design and development of online products and services. This requires applying the appropriate technical and organizational measures to implement, for example, the data protection principles effectively and safeguard individual rights. Secure by design is the foundation that integrates safety and privacy design into the product or service. However, the type of product or service being provided or used by the user will affect the balance between safety measures and privacy measures. This requires risk assessments and depends on what type of user is expected. For example, if children are going to use it or it has user-to-user messaging that requires additional measures. But if it is not likely to be used by children and messaging is not part of the product, then privacy measures can be increased. Because implementing certain types of safety measures requires collecting additional information about users.

6.3.2 Secure by Design

6.3.2.1 Introduction

Secure by design is a principle of cybersecurity and systems engineering that requires systems to be built with security as a foundational property rather than as an afterthought. It is concerned with embedding protections at the earliest design stages of hardware, software, and services, so that security requirements shape the architecture itself, rather than being retrofitted later through patching or external controls. This clause provides an overview of the principles, and it is worth noting that they overlap and complement other design paradigms, such as Zero Trust Architecture (ZTA), privacy by design, and safety by design. There are various standards and guidelines available for secure by design, including but not limited to:

- National standards such as the USA's NIST SP 800-160v1r1 [i.48] (Systems Security Engineering) and the NIST Cybersecurity Framework [i.49].
- Government policies, including the UK's Secure by Design Policy for digital services [i.50] and the UK's MoD's 10 Secure by Design principles [i.51].
- International standards, including the ISO/IEC 27000 series [i.52] and ETSI TS 103 645 [i.53]/ETSI EN 303 645 [i.54] (IoT security).
- International Guidance, including the USA's CISA Secure by Design documentation [i.55].

6.3.2.2 Principles

There are different design principles depending on the requirements for different types of organizations and the type of product or service. The following list gives examples of some of these principles, which are relevant to the present document:

- Create responsibility for cybersecurity risks. Assign risk owners to be accountable for managing cybersecurity risks for a service throughout its lifecycle. These should be senior stakeholders with the experience, knowledge and authority to lead on security activities.
- Design usable security controls. This requires performing regular user research and implementing findings into service design to make sure security processes are fit for purpose and easy to understand.
- Take ownership of customer security outcomes. This includes steps such as eliminating default passwords and ensuring that the burden of security is not placed on the end user.
- Minimize the attack surface. Use only the capabilities, software, data and hardware components necessary for a service to mitigate cybersecurity risks while achieving its intended use.
- Embed continuous assurance. Implement continuous security assurance processes to create confidence in the effectiveness of security controls, both at the point of delivery and throughout the operational life of the service.

6.3.3 Safety by Design

6.3.3.1 Introduction

Safety by design (online) is the process of designing an online platform to reduce the risk of harm to those who use it. Safety by design is preventative [i.56]. It considers user safety throughout the development of a service, rather than in response to harms that have occurred.

By understanding how a platform's services expose users to risk, designers can put in place safety measures to protect users from harm. The platform's users may be at increased risk of online harms if the platform allows them to:

- Interact with each other, such as through chat, comments, liking or tagging.
- Create and share text, images, audio or video (user-generated content).

Safety by design can reduce online harms by preventing them before they happen. Platforms can do this by taking a safety by design approach, which can benefit their business or organization by:

- Creating an environment where users feel safe and make safer choices.
- Preventing problems that might be difficult or costly to solve later on.

It is not foolproof and ideally should be regularly reviewed and updated when new risks are identified or incidents occur.

6.3.3.2 Principles

There are a few key principles that can be implemented to create a safer online environment for people who use online services and platforms:

- Users are not left to manage their own safety. Anyone who owns or manages a platform should take preventive steps to make sure their service reduces a user's exposure to harm. For example:
  - makes a user aware when they do something that might harm themselves or others;
  - helps a user report content or behaviour that they think is harmful;
  - makes it harder for users to upload or share content that is illegal or breaks your terms of service.

- Platforms should consider all types of users. Many factors can increase a user's risk of being a victim of harm. Platforms should aim to understand the people who use their service so they can be aware of the risks their service might present to them. This requires an inclusive design that considers the needs of all users. Platforms should consider users:
  - with protected characteristics which may make them victims of discrimination, for example, race, disability or sexual orientation;
  - with low levels of media literacy, which can be a result of age, background or level of education;
  - of different levels of ability, for example, children, or learning-disabled people;
  - who have accessibility needs, for example, visually impaired people;
  - who cannot speak or speak a limited amount of the primary language of the platform.
- Users are empowered to make safer decisions. Platforms should give users the tools and information they need to make safer choices online. They can do this by designing their platform to promote safer decisions. For example:
  - warning users about excessive screen time;
  - prompting users to review their privacy settings on a regular basis;
  - highlighting when content has been fact-checked by an official or trustworthy source.

It is important to note that the platform design does not limit a user's ability to make informed choices. For example, using algorithms to recommend content that is harmful to a user, over which they have no or limited control. Ideally, a platform design helps users understand the reliability and accuracy of the content they are interacting with and how their online activity is seen by others, and how to manage that, such as by changing privacy settings or blocking a user:

- Platforms are designed to keep children safe. Children are generally less able to understand risk, which makes them particularly vulnerable to online harms. Even if the online platform is not targeted at under-18s, the platforms should take steps to ensure that only users who are old enough are able to access their service. This could mean undertaking measures to identify child users, using age assurance or verification solutions and tailoring users' online experience in line with their age. Designing for children should include:
  - Ensure information like terms of service, or tools used to report harms, is prominent and easy to understand for children of different ages.
  - Encourage safe, positive interactions and ensure users can interact free from abuse, bullying and harassment.
  - Limit access to certain features, functions and content which pose a greater risk of harm to them.
  - Give responsible adults the ability to shape a child's online experience using safety settings.
  - Set safety, security and privacy settings to high by default.
  - The platform should establish a clear process for tackling abuse and regularly check that reporting and moderation processes are being enforced effectively.

6.3.4 Privacy by Design

6.3.4.1 Introduction

The topics of "privacy by design" and "privacy by default" are related to data protection. The term "Privacy by Design" can mean "data protection through technology design." [i.57] This means there are different ways to achieve "privacy by design", which depends on what is meant by privacy itself. Privacy can be defined as the right to be let alone, or freedom from interference or intrusion. Information privacy is the right to have some control over how personal information is collected and used.

Most people (the end-users), when it comes to privacy, know about it or understand it from conversations about data breaches, wearable tech, social networking, targeted advertising miscues, etc. [i.58]. Also, various cultures have widely differing views on what a person's rights are when it comes to privacy and how it should be regulated.

Privacy and security, while they complement each other, are not the same thing. Data privacy focuses on the use and governance of personal data, for example, by implementing policies to ensure that consumers' personal information is collected, shared, and used in appropriate ways. Security focuses more on protecting data from malicious attacks and the exploitation of stolen data for profit. While security is necessary for protecting data, it is not sufficient for addressing privacy.

There are seven 'foundational principles' of privacy by design [i.59]. Although privacy by design is not necessarily equivalent to data protection by design, these foundational principles can nevertheless underpin any approach that is taken.

6.3.4.2 Principles

6.3.4.2.1 Proactive not reactive; preventative not remedial

A proactive approach should be taken to data protection, anticipating privacy issues and risks before they occur, rather than waiting until after the fact. This does not just apply in the context of systems design; it involves developing a culture of 'privacy awareness' across an organization.

6.3.4.2.2 Privacy as the default setting

When designing any system, service, product, or business practice, it should protect personal data automatically. With privacy built into the system, the individual does not have to take any steps to protect their data; their privacy remains intact without them having to do anything.

6.3.4.2.3 Privacy embedded into design

Embed data protection into the design of any systems, services, products and business practices. It should ensure data protection forms part of the core functions of any system or service, meaning it becomes integral to these systems and services.

6.3.4.2.4 Full functionality - positive sum, not zero sum

Also referred to as 'win-win', this principle is essentially about avoiding trade-offs, such as the belief that in any system or service, it is only possible to have privacy or security, not privacy and security. Instead, it should be looking to incorporate all legitimate objectives whilst ensuring compliance with all legal and regulatory obligations.

6.3.4.2.5 End-to-end security - full lifecycle protection

Putting in place strong security measures from the beginning and extending this security throughout the 'data lifecycle', i.e. process the data securely and then destroy it securely when it is no longer needed.

6.3.4.2.6 Visibility and transparency - keep it open

Ensuring that whatever business practice or technology is used operates according to its premises and objectives and is independently verifiable. It is also about ensuring visibility and transparency to individuals, such as making sure they know what data is processed and for what purpose(s) it is processed.

6.3.4.2.7 Respect for user privacy - keep it user-centric

Keep the interests of individuals paramount in the design and implementation of any system or service, e.g. by offering strong privacy defaults, providing individuals with controls, and ensuring that appropriate notice is given.

6.3.5 Balancing the Different Requirements for Online Preventative Security by Design

6.3.5.1 Introduction

This clause discusses the balancing of the different requirements of online preventative security by design. The foundation required for online preventative security by design is making use of secure by design principles for a product or service. When it comes to privacy and safety, while some measures for them are complementary, other measures can be in opposition. Privacy is intended to grant individuals autonomy over their personal information against unwarranted surveillance and intrusion into personal lives. It should establish boundaries that protect against overreach. Privacy is vital for building trust in the digital landscape, which fosters confidence in online engagements such as e-commerce or social media, where users rely on the assurance that their privacy will be respected.

However, there is a balancing act between safety and privacy in social media. A social media platform utilizes algorithms to combat online harassment and cyberbullying, aiming to create a safer digital environment. However, this strategy requires ongoing monitoring of user interactions and content. Users may applaud the platform's commitment to safety but be apprehensive about the extent of data collection and the risk of personal information misuse.

Balancing different requirements for Online Preventative Security by Design is about understanding the users' behaviour and how to protect them from harm while also allowing them to participate in online activities.

Different examples of this balancing act are found in the areas of age verification, signing in or up to online services and user participation in online services. These examples are not comprehensive, and other interactions between online safety and privacy will occur.

6.3.5.2 Age Assurance

Online age assurance is a complex issue regarding privacy, access to content (often to stop children accessing age-inappropriate content) and data protection [i.65]. Verifying the age of an Internet user is hampered by the difficulty for the various stakeholders on the Internet to really know who the person is behind the computer or smartphone. The need to identify Internet users is an issue for privacy and personal data protection, since knowledge of an individual's identity can then be linked to their online activity. Yet, this contains particularly sensitive, private information.

In order to visit certain sites or for certain online activities, it is by nature necessary to identify oneself (e.g. to buy an age-restricted good like alcohol or cigarettes on an e-commerce site) [i.66]. Age verification, when necessary, takes place in a context where the site publisher already knows certain elements of the identity or personal data of the Internet user (including banking details). On the other hand, where access to the site or an online service does not necessarily require identification, age verification is likely to alter the protection of the Internet user's privacy, by preventing them from visiting the site if they do not provide the publisher with information on their identity.

An age assurance method may be at odds with data minimization, posing risks to user data privacy and security. Best practises require that online operators cannot knowingly retain users' Personal Identifiable Information (PII), but the act of verifying user ages itself can put personal and sensitive data at risk. For instance, operators verifying users' ages through government-issued ID or credit card information put data at risk if secure processes are not in place for use, collection, processing, storage, or deletion of PII. This, in turn, increases the risk that such sensitive data could be merged, stolen, sold, or turned over as part of legal proceedings.

At the same time, operators who choose to verify ages through estimation or inference models may increase surveillance and monitoring of users' online activity, such as their content, engagement, social networks, geographic location, screen time, linked accounts, and browsing history. Subjecting users to such intrusive practices may result in an effect that suppresses online speech and enables the potential collection, use, or sale of user activity data.

Age assurance methods are not foolproof, and nor will they completely stop underage users from intentionally or unintentionally accessing age-inappropriate content. Users can still use tools like Virtual Private Networks (VPNs) to bypass age verification. Also, users evade restrictions through a variety of methods, for example, users can evade age assurance by borrowing the device of a parent or adult or by buying, renting, or trading verified adult accounts.

Further, as the technology has developed, users have been able to use generative AI to circumvent age verification methods. For example, users could use realistic filters that can alter the age a person is perceived as in images and videos, or they could generate an image of an accepted identification document. Or even simply use near photo-realistic character creator tools from video games to fool age assurance methods that use facial age estimation.

6.3.5.3 Signing in or Up to Online Services

The type of personal information that users could provide to access platforms varies. Some online services offer access to content without the need to subscribe or create a profile, while others might require users to create a profile or verify certain information about themselves (for example, their name, age and contact email address) before they can access services or specific types of content [i.67]. Also, users might have anonymous or multiple accounts. A user may wish to protect their identity, for example, those in LGBT+ communities, whistleblowers and victims of domestic violence.

There are also valid reasons for creating multiple accounts, such as for work and personal activity. But some users may use these functions to harm others online. For example, if users are able to hide their true identity from others, some may be encouraged to engage in harmful behaviour, for example, by creating multiple accounts, they can contact others even if they have been asked to stop or have had a previous account disabled [i.68]. Harms that can occur as a result of anonymous or multiple accounts include:

- cyberbullying and cyberstalking
- child sexual exploitation and abuse
- terrorist content
- hate crime
- disinformation

In general, there are different steps to reduce harm, including but not limited to:

- restricting the ability to reduce safety or privacy levels for child or unverified accounts
- prompting users when they message an unverified account, asking them to confirm they understand the risks before allowing them to continue
- prompt users to additionally verify their accounts during account creation, for example, using two-Factor Authentication (2FA)

6.3.5.4 User Participation in Online Services

One way in which online service providers present content is by using content recommendation algorithms, or systems which aim to rank or return content that matches a user's perceived interests [i.67]. A large number of factors are used to do this, including information that constitutes the user's personal data. How algorithms work can lead to a range of different outcomes for users, including directing users to content that they will interact with for longer, or encouraging users to engage with content posted by their friends. But they can also lead to harmful content being amplified.

Where personal data is used to determine the content served to users, data protection law will apply [i.69]. For example, services need to be transparent about the way in which they process personal data to make content recommendations. This extends to data processing in connection with the online safety measures that are incorporated into their content recommendation systems, such as identifying vulnerable users who could be placed at risk by certain types of content.

There are different steps to reduce the risk of harm when users participate in online services, including but not limited to:

- blocking high-risk search terms relating to illegal harms and activity
- ensuring that auto-suggestion functions and algorithmic recommendations do not lead people to harmful content
- ensuring that harmful content does not appear at the top of search results
- providing links to resources and support to users who search for high-risk search terms
- users' content, contacts and activity are only visible to friends
- users cannot share their location with strangers
- automatic face recognition is turned off for photos

6.4 Education and Digital Literacy

6.4.1 Introduction

A key part of reducing and preventing harm is for users to have the knowledge to identify the potential risks and to be conscious of personal security while browsing, sharing or surfing the internet through different apps, services and devices they may use. This includes education that enables children (and parents), akin to teaching children to be street aware, but online. The difficulty is that many adults were not brought up with technology, and their children understand more than they do. It is important for parents to be acutely aware of the dangers posed to children while they are exposed to the internet. Practically all parents understand the dangers of unsupervised youngsters talking to strangers, and/or accepting some presents from them.Parents and teachers across the globe need to realize that nowadays, allowing children to navigate the WWW is like placing those children into an unknown geographical location, completely on their own, unsupervised. Parental controls only go so far, but it does not help them as they get older, as children, especially during their teenage years, when children begin to explore and seek greater independence.

One example of teaching adults is that parents also need to understand the risk of providing incorrect Dates of Birth (DoBs) to allow their children access to social media. If a parent provides an incorrect DoB for their child, that child could be placed at greater risk of encountering age-inappropriate or harmful content online. Once a user reaches age 16 or 18, some platforms, for example, introduce certain features and functionalities not available to younger users - such as direct messaging and the ability to see adult content.

As Figure 4 illustrates, there are many influencers for children's cyber education and awareness, in the centre of which are regulations and standards. The several pillars stemming out of regulations and standards are: parents, children, schools, network/transport service providers, content providers, Applications Developers (AppDev), and Hardware (HW) vendors. All the pillars should work together on building Awareness through Education on Cyber Threats and Security.

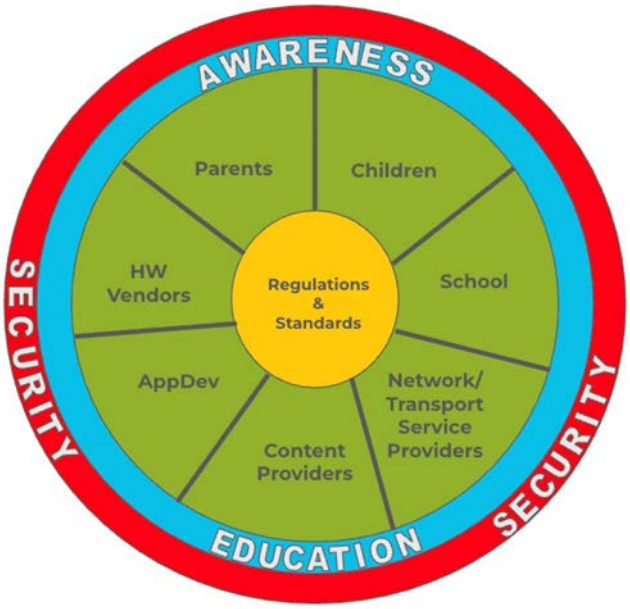

Figure 4: Pillars involved in Cyber Education and Digital Literacy [i.64]

Some cybersecurity vendors go as far as working together with various governments on various education programs for children. One such example is Palo Alto Networks Inc [i.62] where the "Lesson Roadmap" table presents an illustration of suggested education levels in cybersecurity for various age groups.

NOTE: When it comes to school-age children, what they are taught will largely be determined by their country's curriculum and the resources allocated to schools and by schools to teach them about online security, safety and privacy. This does not stop them from independently learning or parents teaching their children about these issues. The following clauses are to give broad advice to ensure appropriate knowledge of digital literacy.

The following clauses build on each other. So, ideally, the older a person, the knowledge they have to stay safe and secure in a digital world, which always grows and is reinforced by refreshing core lessons. Additionally, information targeted at one group can be just as relevant for another, regardless of age group. For these age groups, people within them will have different knowledge and confidence levels with technology, which will affect their ability to solve problems and mitigate risks.

6.4.2 Ages 6 and under

The age at which digital literacy needs to start is rapidly diminishing. Children, as young as 1 year old, notice their parents' smartphone devices and are interested in pressing the touch-sensitive screens, where displays change rapidly. Before these very young children can read or write, they know how to turn on various applications on these smart hand-held devices and manipulate them. Although at this young age, parents are typically in control of what applications their young child can open and play with on their smart handheld device, it is advisable to start cyber education with children as young as possible.

Many ways and methods could be used to educate children. Using the example of Palo Alto Networks Inc. Cyberfit Nation [i.61], [i.62] education plan, which was launched in Australia and New Zealand with respective governments' collaboration, Cyber Safe Kids program provides free educational resources for students of various ages, starting with age 5. The program includes cartoon-type videos providing education through activities in cybersecurity - Cyber A.C.E.S. (Activities in Cybersecurity Education for Students) [i.60].

Another example is Cyberlite [i.63] offering "*... online safety education, working at the intersection of cyber safety, digital wellbeing, and generative AI literacy*" for children.

6.4.3 Ages 7 to 11

For this age group, children are becoming aware of themselves and group dynamics [i.70]. This is also an age when they may start to use the internet and access online content unsupervised. They should be taught through lessons and exercises. For example's passwords should not be shared with strangers. To start teaching them to understand not to share personal information online. Also, to be socially responsible, such as online bullying is just as bad as physical bullying.

There is a chance they will see or experience harmful content online either accidentally or intentionally. They should always be offered support. What this entails will depend on the nature of the harmful content, the context of the situation and the individual child or young person's needs. Remember, if a child or young person has actively sought out the content, this may indicate there's another concern in the young person's life. These measures are targeted more at schools and teachers.

Depending on the type of harmful content a child or young person has seen, different steps could be taken, including but not limited to:

- Help them understand or process what they have seen.
- Answer questions or allay fears or worries.
- Make sure they have a safe space to talk in the future, if needed, rather than sharing harmful content with friends.
- Support them to block or report content to stop them from encountering it again.
- Consider what steps can help them to stay safe online in the future.
- Make sure they know who to get help from if they ever see anything that's worrying or upsetting.
- Provide support for any underlying issues.

6.4.4 Ages 12 to 18

This is an age when a child is becoming a teenager and will seek greater independence in their life and online. This can expose them to risk and harm; therefore, it is important to teach them to understand a range of ways to use technology safely, respectfully, responsibly and securely, including protecting their online identity and privacy; recognizing inappropriate content, contact and conduct; and knowing how to report concerns.

They will be encouraged to talk to their parents and teachers about concerns or incidents they may experience. There is a chance they will feel embarrassed or worried about taking that step. So, they should be aware of resources, for example 'nofiltr.org', which seek to provide a non-judgemental space to get support and to help people learn to be safe online. This allows teenagers to find resources to learn more about the potential risks they may face online, learn how to navigate them, and get help when they need it. Give or seek advice from their peers or a professional, knowing when to block or report someone or identify an online grooming situation. As well as test their knowledge by exploring quizzes around topics like digital boundaries, red flags, online self-care, and their online persona.

During this time, they should learn:

- The importance of using unique passwords and multifactor authentication. This can also include methods such as passkeys.
- Understand what personal identifiable information is and how it is used by companies online, and if shared without due consideration, how it can affect them.
- Understand what social engineering is and how to spot the signs, along with being able to identify an online scam and how to identify what is presented as fact vs fiction online.
- Understand the dangers in spreading misinformation online, not just in general, but also the harm when it happens to a person.
- Understand the dangers of downloading files from untrusted sources and avoid clicking on unknown, suspicious links.

6.4.5 Ages 19 to 29

This is an age where people are considered adults, and they are leaving their parents' home, getting jobs, going to university and are on the road, becoming independent adults. They should be aware of different types of online threats. These include malware, phishing attacks, and scams. Understanding how these threats work helps users to detect and reduce the risk of being affected by them. At this stage in life, they should have been taught different good practices to stay safe and secure. Including but not limited to:

- Safeguard Passwords: Use a strong password for each account. Enable two-factor authentication when available.
- Exercise Caution with Suspicious Messages: Be wary of phishing attempts. Avoid clicking links or downloading attachments from unknown sources.
- Reduce Public Wi-Fi Usage for Secure Activities: Public Wi-Fi networks can be tempting to use, but they can also be vulnerable or fraudulent. Limiting their usage for sensitive tasks like e-banking or online shopping is advisable.
- Practice Discretion in Social Media Sharing: Be mindful of what is shared on social media. Also, knowing how to adjust privacy settings to control the visibility of personal content.
- Enhance Security with Two-Factor Authentication: Enable two-factor authentication whenever possible. Understand, this step is necessary to add an extra layer of protection to accounts.
- Maintain Updated Apps and Software: Keep their devices, applications, and software current.
- Safeguard Their Personal Information: Be cautious about sharing sensitive information online. Provide data only on secure, reputable websites.
- Shop Smart and Securely Online: Be careful when purchasing online. Ensure the website is trustworthy and uses secure payment methods.
- Stay Vigilant and Alert: Have a critical mindset, and question suspicious offers/requests. Also, think twice before sharing personal information or engaging in risky virtual activities.

Though this age group is considered to be digital natives, that does not mean they are naturally better at understanding risks and knowing how to stay safe online and problem-solving issues with technology. Part of it is due to how the majority of devices and services have a friendly user experience-focused design, which requires little input from the user to set up and use.

6.4.6 Ages 30 to 55

Generally, this is a time when people have careers and families. Often, the companies will be responsible for ensuring their employees receive cybersecurity training, and the depth and knowledge will vary for a job's role and responsibilities. Training resources from organizations such as the UK's NCSC [i.72], [i.73] are good starting places. They give advice about passwords, devices, phishing and reporting:

- Creating strong passwords:
  - Use three random words. Ideally, trying to make the words as random and unrelated as possible. Avoid using predictable passwords, such as dates, family and pet names.
  - Two-step verification is enabled.
- Keeping devices secure:
  - Secure a device with a screen lock. This can be a PIN, password, biometric or pattern. A user should pick one of these that they can stick with.
  - Be aware of surroundings: Consider who's around them. Be aware of others around them who might be overlooking their screen or listening in on their conversations. For example, consider using privacy screens, particularly if they are regularly using devices on the move.
- Defending against phishing:
  - Manage digital footprint: review privacy settings and keep personal information to a minimum.
  - Remember the key identifiers: always check emails for signs of urgency, using the authority of the sender, imitation attempts and spelling or grammatical errors.
  - Know policies and processes: read an organisation's policies and processes to help know what to do.
  - If in doubt, check it out: if a user makes a mistake, or they are unsure about something, then they should always report it within their own organization and forward it to the relevant country's cybersecurity or online fraud organization to investigate.
- Reporting incidents promptly:
  - Cyberattacks can be difficult to spot, so people should not be hesitant to ask for further guidance or support when something feels suspicious or unusual.
  - Report attacks as soon as possible; do not assume that someone else will do it. Even if they have done something (such as clicked on a bad link), they should not be afraid to always report what's happened.

There are resources available for parents and guardians to help them have dialogues with their children about online risks, privacy and cybersecurity. For example, 'Le centre pour l'éducation aux medias et à information (CLEMi)' [i.71] has themed guides on cyberbullying, information and disinformation, social networks, screen time and content. They are designed to provide activities to educate and enable media literacy regarding online risks and harm. They cover issues including, but not limited to:

- Identify, understand, and talk about cyberbullying.
- How to recognize misinformation and talk about it with children.
- How to protect personal data.
- Advice on talking about social media with teenagers.
- Managing children's screen time.

This age group experienced the analogue-to-digital transition and the development of current internet-connected information technology. This can mean they have better knowledge and experience in problem-solving issues with devices and a better understanding of how to be safe online and the importance of good cybersecurity practices.

6.4.7 Ages 56 to 75

Being online can make life easier for older people in many ways, but it also comes with the increased risk of scams and fraud. Online scams are becoming increasingly targeted at older, potentially less tech-savvy users, but they can protect themselves by knowing what to look out for and what to do if they suspect a scam [i.74]. This includes but is not limited to:

- Identify scam emails:
  - errors in spelling or grammar, or an unusual style of writing;
  - unofficial or suspicious email addresses or senders;
  - requests for personal information, such as their username, full password or bank details, a genuine organization will never ask for this;
  - threats that unless they act now, a deal will expire, or their account will be closed.
- Identify a fake website:
  - Check the URL. Does it start with HTTP or HTTPS? Are there spelling mistakes or typos in the name of the company? Is it a recognizable URL? Does it end in .net or .org, as these are rarely used for online commerce sites?
  - Look out for spelling and grammar issues. Watch out for spelling and grammar issues, as well as any odd phrases that may suggest the site was created in a hurry.
  - Check if there's a 'Contact us' page. Legitimate businesses and websites will provide a way to get in contact with them.
- Understand the risks of computer viruses:
  - Users may be sent an email with an attachment which, when they click on it, will release a virus.
  - Criminals can then use this to take control of their computer, or the virus may scan their computer for personal information. It can also slow their computer down, send out spam emails or delete files.
  - They may even get a phone call from someone claiming to be from a well-known software company, saying there's a problem with their computer and that they need to get access to it, including asking for personal details. Legitimate IT companies never contact customers in this way. This is a common phone scam; they should know to hang up straight away.
- Relationship scams:
  - Know that scammers can use social networks or dating websites. Once they have gained a person's trust, they might start asking them for money, often by telling them an emotional story about their life.
  - Never send the person money or give them account details. It is always worth talking to a friend or relative about it, especially if things seem to be moving fast. Be wary of the person moving away from the chat room or dating site to communicating by email or text message.
  - If arranging to meet, make sure it is in a public place, tell someone else where the meeting is happening and do not give away information too quickly.

Within this age group, there will be a wide range of people with varying levels of knowledge and confidence with the internet and ICT in general.

6.4.8 Ages 76 and Older

This is an age where the likelihood of cognitive decline increases, which can make people more vulnerable to being scammed or manipulated. A solution to this is for them to organize a mechanism, often a power of attorney, where a trusted (younger) family member, friend or solicitor/attorney has joint control or shared access to their digital assets and/or services. This should be organized before any serious cognitive decline, but it can be done afterwards. This legal process will vary from country to country.

This involves identifying digital assets/services which commonly include:

- Financial accounts.
- Email and social media.
- Digital files: photos, documents stored in cloud services.
- Online subscriptions.
- Business accounts.
- Access to devices.

It is useful to list each asset with login details and provider contact information. This helps to find and manage them when needed quickly.

This can reduce the risk of the person being targeted or affected by scams, identity theft, etc. This should also be done with the consensus of not just the person granting autonomy, but also with other family members, relatives, and friends, as needed, as there is a risk that someone may abuse this for their own gain. Open dialogue about this emotive topic is essential for trust and hopefully prevents misuse.

It should be noted that these issues can also apply to people of any age due to circumstances such as suffering a serious injury or having a cognitive disability, which can leave people at greater risk of being successfully targeted by malicious or bad actors.

6.5 Support for Users and Content Moderators

6.5.1 Introduction

Ensuring online preventative security for users of ICT requires companies and services to take a proactive approach in support of their users, but also the customer-facing staff, along with content moderators who are responsible for removing and preventing harmful content from spreading on social networks and services.

6.5.2 Users

There are different ways and means to provide support to users, and part of this is the user experience and design. A part of this is clear signposting of users to report problems and to find the appropriate help and support. This should not be buried or in an unclear place within submenus or kept within categories which are not obvious to find. It means creating a clear structure with intuitive navigation that helps users find what they are looking for without getting lost and using clear descriptive titles for links or options to avoid confusion and set the right expectation for the user. This can also provide links to resources and information about other organizations which can help with problems beyond their capability.

For example, pointing to sites like 'haveibeenpawned.com' allows users to identify whether their email address or password has been exposed in a data breach, (past or recent). This can help users to better protect their own security and privacy. For example, taking steps to ensure they have MFA enabled on accounts and use unique passwords or passkeys for sites, apps and services, which can be aided by a password manager.

6.5.3 Content Moderators

Content moderators have become indispensable for online platforms' everyday operations. Their role is to remove reported and flagged harmful and illegal content. Companies have legal responsibilities for this under laws such as the UK Online Safety and the EU's Digital Services Act [i.75]. The turnover for content moderators' sites is high, as most moderators cannot continue for more than 2 years on average.

Poor mental health is one of the major reasons behind moderators leaving their positions, as their jobs require them to review large volumes of texts, pictures, and videos containing highly disturbing content around violence, extremism, drugs, Child Sexual Abuse Materials (CSAM), self-harm, and many more. Long-term exposure to such harmful content has triggered serious mental health issues among moderators, including depression and anxiety. With deteriorated mental health conditions, more severe issues like PTSD and addictions to drugs and alcohol have also been noted to emerge.

Certain measures can be implemented to better support and protect content moderators, [i.76] including but not limited to:

- Mental Health Support:
  - Content moderators are often exposed to traumatic imagery and language. Employers should provide access to mental health resources, such as counselling, therapy sessions, and wellness programs. Regular check-ins, anonymous reporting channels, and resilience training can help employees cope with the psychological challenges of content moderation.
- Rotate Workloads and Limit Exposure:
  - Moderators should not continuously review harmful content; organizations should rotate tasks to balance workloads. Alternating between routine reviews and potentially distressing content reduces prolonged exposure and minimizes emotional fatigue. There should be daily or weekly caps to ensure moderators do not spend excessive hours on high-risk material.
- AI and Automation:
  - AI-powered tools can potentially handle much of the repetitive, high-volume screening in content moderation, leaving human moderators to focus on nuanced decision-making. Reducing the sheer number of harmful posts that employees could view, it can also protect moderators' mental health.

7 Digital Evidence and Social Media

7.1 Overview

Given the massive amount of data available, social media has become one of the most fertile sources of information for background checks and intelligence gathering, as well as for various types of investigations, including criminal, regulatory, insurance and civil matters such as cyberbullying and defamation.

However, these are a volatile and dynamic source of data. Posts, including photos, comments, memes, and other potentially relevant information, can disappear in a matter of minutes, if not seconds, nullifying any chance of using them as probative elements.

Also, depending on the country where a platform or service operates and the user's geographic location, there may be additional requirements to report to the country's law enforcement agencies or police. This should occur if a victim is under 18 years of age, and their dedicated organizations, such as child protection organizations, should also be reported to.

If required, companies may have to record and retain data when identifying such material if served by warrants. Requirements for this can be found in ETSI TS 102 232 [i.77] Series Lawful Interception: Handover Interface and Service-Specific Details (SSD) [i.77] for IP delivery and ETSI TS 103 643 [i.78]. ETSI TS 103 643 [i.78] defines a process called the Digital Evidence Bag (DEB) of receiving, transforming and outputting material that can be assured digitally. It identifies the ways that a DEB can be used to ensure the material used in legal proceedings. Specifically, the assurance of the material is not dependent on the process having been carried out by a qualified or trained human expert. It is designed to be used in situations where a risk assessment of the handling of digital material has identified that extra assurance of the integrity, provenance, continuity and validity of the digital data is required.

7.2 Digital Evidence Gathering and Processing

7.2.1 Overview

A key requirement for investigations, prosecutors and defenders is that the digital evidence collected and used is:

i) is identical to the source;

ii) is not altered; and

iii) is traceable, leaving no room for questioning when the evidence is most needed.

7.2.2 Gathering and Processing

Computer documents, emails, text and instant messages, transactions, images and Internet histories are examples of information that can be gathered from electronic devices and used as evidence [i.41].

In addition, many mobile devices store information about the locations where the device travelled and when it was there. For example, photos posted to social media may contain location information. Photos taken with a Global Positioning System (GPS)-enabled device contain file data that shows when and exactly where a photo was taken.

The gathering and processing of evidence from devices and online services will vary from country to country due to the different judicial and policing regulations. Though there are general practices that can be employed to ensure the requirements of:

i) is identical to the source;

ii) is not altered; and

iii is traceable.

These include, but limited to:

1) Prevent contamination: Before analysing digital evidence, an image or a work copy of the original storage device is created. When collecting data from a suspect device, the copy need to be stored on another form of media to keep the original pristine. Analysts need to use "clean" storage media to prevent contamination or the introduction of data from another source. For example, if the analyst were to put a copy of the suspect device on a CD that already contained information, that information might be analyzed as though it had been on the suspect device. Although digital storage media such as thumb drives and data cards are reusable, simply erasing the data and replacing it with new evidence is not sufficient. The destination storage unit need to be new or, if reused, it need to be forensically "wiped" before use. This removes all content, known and unknown, from the media.

2) Isolate Wireless Devices: Smartphones and other wireless devices should be initially examined in an isolation chamber, if available. This prevents connection to any networks and keeps evidence as pristine as possible. The Faraday bag to collect and store devices can be opened inside the chamber, and the device can be exploited, including phone information, SIM cards, etc. The device can be connected to analysis software from within the chamber. If there is no isolation chamber, investigators can place the device in a Faraday bag and switch the phone to aeroplane mode to prevent reception.

3) Install write-blocking software: To prevent any change to the data on the device or media, there should be installed a block on the working copy so that data may be viewed, but nothing can be changed or added.

4) Select extraction methods: Once the working copy is created, determine the make and model of the device and select extraction software designed to most completely "parse the data," or view its contents.

5) Submit device or original media for traditional evidence examination: When the data has been removed, the device is sent back into evidence. There may be DNA, trace, fingerprint, or other evidence that may be obtained from it, and the digital analyst can now work without it.

6) Proceed with investigation: At this point, the digital analyst will use software to view data. They will be able to see all the files on the drive, can see if areas are hidden and may even be able to restore the organization of files, allowing hidden areas to be viewed. Deleted files may also be visible, as long as they have not been overwritten by new data. Partially deleted files can be of value as well.

Files on a computer or other device are not the only digital evidence that can be gathered. Digital evidence resides on the Internet, including chat rooms, instant messaging, websites and other networks of participants or information. By using the system of Internet addresses, email header information, time stamps on messaging and other encrypted data, the analyst can piece together strings of interactions that provide a picture of the activity of an individual online and offline.

8 Conclusion

To begin to provide Human-to-Human Online Preventative Security requires an understanding of the digital devices and online services landscape used by the user. Along with the potential threats/harms that emerge from these, which affect the typical end user. This should provide a knowledge base to be able to implement design choices to mitigate them. This could be online preventative security formed of three parts, first a foundation of secure by default, then implementing safety by design and privacy by design. If required by regulation due to the type of service or features that a service, e.g. direct messaging between users, provides, it may require additional measures, for example, age assurance to be implemented, which will require its measures to prevent loss or leakage of personal data. If and when things go wrong, there should be appropriate steps and actions that can be taken to address the harm and remedial action. This may include signposting to helpful resources and measures to preserve evidence of malicious material. Remedial action should be taken to learn from the lesson of when harm is committed, to provide feedback to implement fixes and improve protections against harm, intending to reduce the likelihood of them occurring again.

Annex A:
Online Safety Landscape

A.1 Global Online Safety Regulators Network

The Global Online Safety Regulators Network is a collaboration among the pioneers in online safety regulation. The Network paves the way for a coherent international approach to online safety regulation, by enabling online safety regulators to share insights, experience and best practices.

Current Network members include:

- eSafety Commissioner - Australia
- Online Safety Commission - Fiji
- Arcom - France
- Coimisiún na Meán - Ireland
- Korea Communications Standards Commission - Republic of Korea
- Council for Media Services - Slovakia
- Film and Publication Board - South Africa
- Ofcom - United Kingdom

Members share a commitment to act independently of commercial and political influence and adhere to objective criteria for respect for human rights, democracy, and the rule of law. The Network is also open to observers - specifically organizations that have expertise and interest in online safety regulation and who wish to follow and engage with the Network.

Current observers include:

- 5Rights - Global
- Canadian Centre for Child Protection - Canada
- Department of Canadian Heritage - Canada
- European Parliament Intergroup on Children's Rights - EU
- Family Online Safety Institute - Global
- Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM) - Germany
- INHOPE - Global
- Netsafe - New Zealand
- Te Mana Whakaatu | Classification Office - New Zealand
- WeProtect Global Alliance - Global

A.2 European Union

EU Digital Services Act - Aims to improve reporting of illegal content, goods or services on online platforms. Due to diligence obligations for platforms and stronger obligations for very large platforms, where the most serious harms occur.

EU Digital Market Act (DMA) - establishes a set of clearly defined objective criteria to identify "gatekeepers". Gatekeepers are large digital platforms providing so-called core platform services, such as online search engines, app stores, and messenger services. Gatekeepers will have to comply with the do's (i.e. obligations) and don'ts (i.e. prohibitions) listed in the DMA.

France - Law LOI n 2022-300 du 2 mars 2022 visant à renforcer le contrôle parental sur les moyens d'accès à internet (LAW n 2022-300 of March 2nd, 2022, aimed at strengthening parental control over the means of access to the Internet)

A.3 Commonwealth Nations

UK Online Safety Act - Advances Safety by Design for online and digital products and services. It pushes for greater transparency and accountability, along with inclusivity and resilience.

Australia Online Safety 2021 - capture harms that occur on services other than social media. Online service providers to be proactive in how they protect people from abusive conduct and harmful content online.

Australia - Online Safety Amendment (Social Media Minimum Age) Act 2024 (Cth) is an Australian act of parliament that aims to restrict the use of social media by minors under the age of 16. It is an amendment of the Online Safety Act 2021 and was passed by the Australian Parliament on 29 November 2024. The legislation imposes monetary punishments on social media companies that fail to take reasonable steps to prevent minors from creating accounts on their services.

Annex B: Bibliography

- European Parliamentary Research Service: "Hate speech and hate crime: Time to act?"
- European Parliament: "Artificial intelligence (AI) and human rights: Using AI as a weapon of repression and its impact on human rights".
- Europol: "Europol spearheads largest referral action against online hate speech".
- European Parliament: "Victims of online hate speech".
- European Commission: "Commission is gathering views on draft DSA guidelines for election integrity".
- European Union Agency on Fundamental Rights: "Online content moderation - Current challenges in detecting hate speech".
- European Commission: "Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act".
- EU Disinfo Lab: "The Problem".
- Tech Transparency Project: "Dangerous by Design, Two Studies: Social Media Algorithms Fuel Online Hate".
- Australia eSafety Commissioner: "Time to take a look under the virtual hood at how algorithms might be harming our kids".
- Australia eSafety Commissioner: "Safety by Design puts user safety and rights at the centre of the design and development of online products and services".
- University College London: Social media algorithms amplify misogynistic content to teens.
- IETF Human Rights as a Service (HRaaS): draft-rutkowski-hrpc-hraas-00.
- National Library of Medicine: "The algorithm will screw you": Blame, social actors and the 2020 A Level results algorithm on Twitter.
- Institute for Strategic Dialogue (ISD): Online Civil Courage Initiative France
- Institute for Strategic Dialogue (ISD): Assessing the AfD's social media strategy in the lead-up to eastern German state elections
- Carnegie Endowment for International Peace: Countering Disinformation Effectively: An Evidence-Based Policy Guide.
- Global DisInformation Lab: "A Guided Tour of Disinformation Policy".
- Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.
- 96/443/JHA: Joint Action of 15 July 1996 adopted by the Council on the basis of Article K.3 of the Treaty on European Union, concerning action to combat racism and xenophobia.
- European Commission: The EU Code of conduct on countering illegal hate speech online.
- Council of Europe: Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189).
- Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance).

- European Commission: Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines.
- Council of Europe: Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
- Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) (Text with EEA relevance).
- UK Parliament: Social media, misinformation and harmful algorithms.
- Australian Public Service Commission: Social media: Guidance for Australian Public Service Employees and Agencies.
- France Premier Ministre: Mobilizing France Against Racism and Anti-Semitism.
- France Premier Ministre: Creating a French framework to make social media platforms more accountable: Acting in France with a European vision.
- European Commission: European Centre for Algorithmic Transparency.
- European Commission: Report from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions Protecting Fundamental Rights in the Digital Age - 2021 Annual Report on the Application of the EU Charter of Fundamental Rights.
- European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions EU strategy on the rights of the child.
- Consolidated text: Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (Text with EEA relevance)Text with EEA relevance.

History

Version	Date	Status
V1.1.1	March 2026	Publication