ISD | Powering solutions to extremism, hate and disinformation

ALFRED LANDECKER FOUNDATION

Online Safety Regulation & Private Online Communications

The State of Play and Ways Ahead for Addressing Terrorism, Extremism, Hate, and Disinformation

Christian Schwieter
Editor: Henry Tuck

About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of disinformation, hate speech, and extremist and terrorist content online. It is comprised of representatives of relevant ministries and regulatory bodies from liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space. We thank the Alfred Landecker Foundation for their support for this project.

About this Paper

As part of the DPL, the Institute for Strategic Dialogue (ISD) organised working group meetings in July 2024 on the topic of reviewing research on the risks and the effectiveness of mitigations in online private messaging platforms and services and reviewing the evolving regulatory landscape for this field. The working group consisted of DPL members representing national ministries and regulators from the European Commission, the Netherlands, Slovakia, the United Kingdom and the United States. Participants also included expert representatives from civil society and academia. While participants contributed to this publication, the views expressed in this paper do not necessarily reflect the views of all participants or any governments involved in this project.

About the Author

Christian Schwieter is a Fellow at ISD and a PhD candidate at the Department of Media Studies at Stockholm University, where he investigates the impact of European platform governance efforts on far-right activity online. He is also an editor of the open-access Journal of Digital Social Research (JDSR). Previously, he was a researcher at the Oxford Internet Institute and served as Specialist Adviser on Disinformation Matters to the DCMS Select Committee at the UK House of Commons. He has advised, among others, the German Ministry of Justice, the European Parliament, and the UN Office of Counter-Terrorism on issues related to extremism, disinformation, and platform regulation. He holds an MSc (Dist) in Social Science of the Internet from the University of Oxford and a BA (Hons) in World Politics from Leiden University.

About the Editor

Henry Tuck is the Director of Digital Policy at ISD, where he leads Advisory work on digital regulation and tech company responses to terrorism, extremism, hate and dis/misinformation online. Henry oversees ISD's Digital Policy Lab (DPL) and engagement on key digital regulation in Europe and Five Eyes countries, advises key governments, international organisations and major private sector tech companies, and collaborates with ISD's Digital Analysis Unit to translate research into actionable digital policy recommendations. Having joined ISD in 2013, Henry has previously worked across a variety of ISD's Analysis and Action programmes, including education, on- and offline counter-extremism interventions, and civil society networks. Henry holds a Masters in International Conflict Studies from Kings College London, and a BA in Philosophy, Politics and Economics from Durham University.

Acknowledgements

We would like to thank all participants of the working group for their contributions, including experts from governments, civil society and academia. We would like to give special thanks to the speakers and contributors to this paper for their valuable insights and feedback: Iria Puyosa (Senior Research Fellow, Democracy + Tech Initiative, Atlantic Council), Caroline Sinders (Founder, Convocation Design + Research), Jenna Omassi (International Policy Manager, Online Safety, Ofcom), and colleagues at Tech Against Terrorism. We would like to thank ISD colleagues Henry Tuck, Ellen Jacobs and Helena Schwertheim who reviewed this paper.

ALFRED LANDECKER FOUNDATION

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Executive Summary

This policy brief addresses the critical issues surrounding online harms in private communications and the role of encryption technologies, such as end-to-end encryption (E2EE), in maintaining online safety, with a focus on online terrorism, extremism, hate and disinformation. Current concerns driving discussions on breaking or circumventing encryption typically focus on its use by criminals for activities such as terrorist recruitment, planning and coordination and child sexual abuse material (CSAM) distribution. We recognise that the discussion around end-to-end encrypted messaging and tactics used to identify and mitigate instances of CSAM is very nuanced and often specific to that set of harms. While the methods for identifying and mitigating all the harmful uses of private messaging cannot fully be separated, we want to acknowledge and caveat that the scope of this brief does not fully address the conversation around CSAM content in private, end-to-end encrypted messaging spaces.

As such, the brief provides a short overview of technologies and techniques available to the online safety community to increase accountability in private, encrypted online spaces. Encryption backdoors and client-side scanning are highly criticised by security experts and digital rights activists. But less invasive techniques such as message franking and metadata analysis are already widely adopted and can often be more effective at balancing safety and privacy.

However, based on key online safety regulations adopted in recent years, particularly the UK Online Safety Act and the EU Digital Services Act, the brief argues that many of the 'online harms' or 'systemic risks' that online safety regulators are tasked with mitigating, such as terrorism, extremism, hate, and disinformation, predominantly occur in public online spaces (even if some aspects of each of these harms also occur in more private and sometimes end-to-end encrypted online spaces). Online safety regulators should consider in what scenarios privacy-by-design principles that leverage E2EE may enhance online safety for citizens. The brief further emphasises that the rise of hybrid platforms that blur the line between social media, group and private messaging functionalities makes it increasingly difficult to distinguish between 'private' and 'public' online spaces. Therefore careful case-by-case assessments about how specific functionalities relate to specific harms are required.

Based on this, the following recommendations are proposed for policymakers and online safety regulators to better mitigate online harms in private communications:

1. **Clarify the scope of online harms in private communications**. Regulators should define, on a case-by-case basis, which online harms fall within the scope of online safety frameworks and how these harms manifest in private communication spaces online. Public online spaces are often the key venues where certain online harms or risks that seek a broader audience – including terrorism, extremism, hate and disinformation – manifest. Due to data access challenges, the role of private online spaces in these harms or risks is often less clear. It is therefore crucial to specify the extent and conditions under which private communication functionalities contribute to these harms.

2. **Evaluate the impact of mitigation measures on privacy and safety**. Regulators must critically assess how different platform mitigation measures impact user safety and privacy. Measures designed to address one type of harm may inadvertently compromise user safety and privacy in other areas. Proportionality should guide the selection of appropriate design choices, functionalities or moderation tools to achieve online safety objectives without infringing on user privacy.

3. **Differentiate between the objectives of law enforcement and online safety regulators**. The goals of law enforcement agencies are focused on preventing and prosecuting criminal activity. This differs from those of online safety regulators who ensure that platforms adhere to safety standards. Although E2EE can complicate law enforcement efforts in combatting illegal terrorism or CSAM-related activity, it may play a lesser role in hindering online safety regulators from holding platforms accountable regarding the mitigation of online harms which do not always constitute criminal activity (such as extremist propaganda, online hate and disinformation campaigns).

4. **Recognise the risks of weakening encryption for online safety.** Breaking E2EE poses significant

risks to online safety, as it undermines the right to privacy upheld by entities such as the European Court of Human Rights. Access to private communications is fundamental to the exercise of democratic freedoms and human rights. Therefore, technological or regulatory measures that weaken E2EE conflict with the objectives of most online safety frameworks. Enhancing privacy through robust encryption can, in many cases, improve overall online safety.

5. **Consider how privacy-by-design in tandem with safety-by-design principles may work together to foster online safety.** Choices including making privacy-preserving technology like E2EE accessible widely, abiding by data minimisation principles for (meta)data storage, empowering users through robust reporting and blocking functionalities, and implementing privacy-by-default settings can significantly increase the safety of high-risk individuals. This includes but is not limited to journalists, human rights activists, victims of hate and harassment campaigns, and children. Safety-by-design principles, such as robust processes to report, remove and escalate illegal or harmful content, and the accounts responsible for distributing such content also reduces the risk of E2EE being exploited.

6. **Encourage service-specific mitigation measures and call out attempts by platforms to hide behind inaccurate terminology**. Platforms offering a range of functionalities, from private messaging to public forums, require tailored mitigation strategies to address different manifestations of harms across functionalities with differing levels of privacy. In the context of harms such as online terrorism, extremism, hate and disinformation, a uniform approach risks misidentifying the sources of harm. This may unnecessarily target private communications in cases where the actual issues often arise in public spaces. Similarly, platforms should be called out when misusing terms like "encrypted" or "private" to evade accountability for forms of harm that manifest through their public or unencrypted communication functionalities (such as open WhatsApp groups, or public Telegram channels).

Recent advances in quantum computing are outside of the scope of this brief. These pose a threat to current encryption standards, as well as advances in quantum cryptography or quantum encryption. In general, the online safety community should advocate for approaches that safeguard access to secure private communications as a pre-requisite to the free exercise of human rights - even in the face of future technological developments.

Introduction

The contemporary online landscape is made up of a variety of different services, ranging from big public social media platforms including Facebook, X (formerly Twitter) and YouTube to private messengers such as WhatsApp and Signal. Recently, policymakers in the Global North have primarily focused on mitigating threats to human rights based on the functionalities of the former. This focus was appropriate given that some of the most pertinent online harms or systemic risks, namely extremist propaganda and disinformation take place in public online spaces by their very nature.

When it comes to terrorism and child sexual abuse material (CSAM), however, private communication has more prominently featured in online safety discussions. In 2015, the growing threat of Islamic terrorism led to widespread demands by politicians to give law enforcement access to private, E2EE communication. In the aftermath of the 2015 Charlie Hebdo attack, the UK Prime Minister went as far as to suggest a ban on E2EE messenger services.[1] In 2024, plans by the European Commission to combat the circulation of CSAM were hotly contested by digital rights activists. They argue that it would undermine E2EE, based on an analysis by the legal service of the EU Council of Ministers and the European Data Protection Supervisor.[2]

The growth of hybrid platforms, which offer public communication functionalities alongside private communication channels, further complicate debates about what constitutes private and what constitutes public spaces online – and by extension, who should have access to those spaces. Perhaps the best example of such a hybrid platform is the messenger-turned-social media platform Telegram; it initially started out as a private messenger but added public communication functionalities over the years. These include large public groups with up to 200,000 users and one-to-many channels for broadcasting to an unlimited audience. WhatsApp also began rolling out one-to-many channels worldwide in 2023, enabling users to effortlessly switch between private chats and public fora. Conversely, the social media platform Facebook has added E2EE to its direct private messenger. These changing platform functionalities raise new challenging questions as to where regulatory intervention is appropriate.

Distinguishing between private and public activity online is therefore not an academic task: the growth of hybrid platforms poses significant challenges to a variety of stakeholders, including law enforcement and online safety regulators. What is now re-emerging is the contentious debate on whether policymakers should prioritise privacy or security; between those seeking to safeguard private online spaces through robust E2EE technologies, and those seeking to open-up those spaces for content moderation and, in some cases, law enforcement – or whether a potential middle ground exists between the two.

The goal of this policy brief is to provide a brief overview of the debate and offer potential ways forward to different stakeholders to overcome the simplified privacy versus security dichotomy. The brief starts by clarifying the evidence base on the role of private communication technologies in enabling or exacerbating online harms or systemic risks. This is followed by a brief review of global digital regulation frameworks, providing case studies on the extent to which private messaging platforms are covered under existing online safety regulation. From this follows a discussion of possible safety interventions and mitigations, distilling best practices for regulators and policymakers where possible.

Clarifying terminology – private, secure, end-to-end encrypted

The debate around online safety, private messaging and E2EE messaging has at times been further complicated by a lack of common terminology, or the use of inaccurate descriptors by stakeholders when discussing private online spaces and encryption technologies specifically. This issue is exacerbated by social media companies and communication services often using misleading terms like 'private' or 'secure' to describe their features without further explanation. Private, for example, may simply be used to describe online spaces that are not public (e.g. not accessible or visible for other users, and/or invite only online spaces such as group chats).[3] This could possibly involve features like auto-deletion of messages. However, it does not mean the conversation itself is end-to-end encrypted and therefore likely to be more 'secure'. As research by Convocation and Tech Policy Press has shown,[4] many users are unaware of how to use security features; they either falsely believed they were communicating securely or gave up on trying to enable security features altogether. This is in part due to the ambiguous terminology in user interfaces but also because of the often-lacklustre implementation of encryption technologies by companies.

When it comes to encryption, there are fundamental differences in the underlying technology available. While it is beyond the scope of this brief to describe the technology in depth, it is crucial for policymakers dealing with online harms to understand the difference between two types of encryption, transport layer security (TLS) and end-to-end encryption (E2EE). Both encrypt data to make it harder for third parties to intercept and read information sent over the internet.

However, TLS merely secures information 'in transit' as it travels between user and server. A common example is when users connect to websites like Facebook.com: the traffic between the user and the platform is encrypted in a way that only the user and platform can 'read' the information sent. It is this type of encryption that is the most widely used in digital communication as it allows platforms to collect the user information necessary for the functioning of its service. E2EE, on the other hand, encrypts data between two users. The platform functioning as an intermediary (e.g. Signal) cannot read the actual contents of the message unless it is reported to them by one of the users. Only the intended recipient of the message can decrypt the data, so the message content is only accessible on the user's end device.

What harms?

Key to the debate on how to safeguard human rights online is the identification of systemic risks and online harms, and how they may or may not be linked to private communication technologies. The below provides a brief overview of a selection key areas through which online harms or systemic risks manifest and how the relate to private online spaces.

Propaganda, disinformation and influence operations. The goal of propagandists is usually to reach the widest audience possible to influence public attitudes on specific issues (e.g. election integrity, public health or immigration). Similarly, influence operations using techniques of disinformation rely on public communication channels to be successful. Nevertheless, research has shown how private and semi-private spaces, such as invite-only (i.e. closed) groups on Telegram or Discord, are often used to organise and coordinate disinformation campaigns. One of the earliest examples is the Discord server 'Reconquista Germania', which was used for these exact purposes by far-right groups in the context of the 2017 German elections.[6] Evidence exists that propaganda and disinformation is sometimes targeted directly at vulnerable individuals through private messaging functionalities,[7] but the harm caused by influence operations typically manifests through public online spaces. A notable exemption is WhatsApp's use by diaspora communities in the US[8], former Brazilian President Jair Bolsonaro's supporters in Brazil[9] and the ruling Bharatiya Janata Party (BJP) in India[10] to spread disinformation. However, most of this activity seemingly takes place in large groups, which could effectively be considered quasi-public spaces rather than private communication channels.

Extremism and terrorism. Extremists and terrorists have always relied on means of private communication to operate outside of the purview of law enforcement and continue to do so. While a significant part of their online strategy now relies on public online spaces to radicalise and agitate through propaganda and disinformation, private channels remain key to recruiting, initiating and coordinating their activities. Prominent examples include the use of platform Telegram by the so-called Islamic State (IS) as well as by far-right extremists on the 6 January attack on the US Capitol.[11] It is important to note that radicalisation is a complex process that often takes place across both public and private platforms. Activity in public online platforms like

Public and private online spaces – a broad spectrum rather than a simple dichotomy

While public, private and closed are frequently used as distinctive descriptors for online services and their functionalities, determining the real nature of online spaces is often more complex. Many online services offer functionalities that do not neatly map onto these binary descriptors. On Telegram, a group might technically be private or closed because people require admin approval to join. However, the group may have over 10,000 members. Content posted there may therefore be seen by far more people than content posted in a public group with very few followers. Another example is YouTube's 'unlisting' function, where users can choose to make their video available only to those who have the exact link; it will not appear on the user channel or in search results.[5] The video is still technically publicly available because it does not require a password or other form of user verification. In reality, it is virtually impossible for an average user to stumble upon the video through the platform's interface. The audience of the video is therefore limited to those who have direct access to the exact URL. Similarly, content curation practices by platforms are key to making content visible in the first place – for example, showing the most trending content on the front page or algorithmically boosting specific content in search results based on user activity. The power of platforms in determining the public visibility of online content is most prominent when social media companies 'demote' or 'shadowban' content as part of the content moderation strategies. In this way, the content becomes virtually impossible to find despite the fact it still technically exists on the platform.

Facebook function as 'beacons' to direct users to unmoderated 'content archives' and funnel potential recruits into more private online spaces. This in turn may organise to flood public platforms with extremist propaganda.[12]

Hate and harassment, including gender-based violence. Hate and harassment can occur spontaneously or be premediated; women and minority groups are disproportionally affected.[13] Coordinated harassment campaigns, like influence operations and online extremism or terrorism, are often organised through private online groups. In these groups, members come together to develop strategies and discuss tactics to attack certain targets online. What distinguishes hate and harassment from the harms and risks of disinformation and terrorism, however, is that harassment does not always rely on public communication channels. While hate campaigns often take place in public online spaces, especially comment sections,[14] those seeking to harass individuals will often make use of private direct messaging functionalities to threaten, intimidate or ridicule targets.

Child sexual abuse material (CSAM).[15] A major area of concern is the spread of CSAM on private online spaces, including E2EE messenger services. Arguably, CSAM is one of the online harms that manifests most prominently via closed, private online spaces. Those engaged in sharing or consuming CSAM, and in facilitating the grooming of children, will attempt to stay out of the reach of law enforcement.[16] The harm of CSAM also does not require a broader public audience in the same way as extremist propaganda or influence campaigns. The mere possession and private consumption of CSAM content is a criminal act in most jurisdictions. Due to the nature of the harm, the proliferation of CSAM is often a key argument for those seeking greater regulatory intervention and enhanced risk mitigations in private messaging contexts to enhance child protection and online safety. Both the proposed EU Regulation laying down rules to prevent and combat child sexual abuse[17] and the US EARN IT Act[18] seek to minimise the proliferation of CSAM. Both proposals have been met with significant criticism.[19]

Fraud, scams and related cybercrime. Lastly, fraud, scams or related cybercrime can take place both in public and in private spaces. Criminals may identify vulnerable users online and then target them with personalised messages – a common example is a WhatsApp fraud where criminals impersonate family members to convince users to send them money.[20] These types of scams or fraud pre-date the internet but modern communication technologies provide these criminals with easy access to a wealth of potential victims.

The short review above shows the complex interaction between online harms or systemic risks and private online spaces. Influence operations, by their very nature, almost always ultimately take place in public online spaces. Extremist and terrorist actors, on the other hand, rely more heavily on private spaces alongside their public communication activities. Hate and harassment, the spread of CSAM as well as fraud or scams are the key online harms that can take place exclusively in private online spaces.

The recent encryption debate: the perspectives of law enforcement and advocates for end-to-end encryption

In liberal democracies, policymakers largely agree that the ability to engage in private communications, whether in-person or through digital means, is a prerequisite to the exercise of human rights. Those most threatened by government surveillance – including journalists, human rights activists or minority groups – rely on access to secure and private communications. However, the extent to which this ability may be curtailed in the context of law enforcement investigations is subject to much debate. A 2024 joint declaration by Europol and European police chiefs emphasised how the use of E2EE technology impedes the ability of law enforcement to collect evidence. It concluded that it ultimately makes the prevention and prosecution of criminal activity more difficult. The declaration calls on governments to "take action against end-to-end encryption roll-out ... to ensure public safety across social media platforms."[21] Similar calls have been made by politicians in other jurisdictions: in 2019, the US, UK and Australian governments called on Facebook to "not proceed with its end-to-end encryption plan without ensuring there will be no reduction in the safety of Facebook users and others, and without providing law enforcement court-authorised access to the content of communications to protect the public, particularly child users."[22]

The concerns raised by law enforcement may seem somewhat misplaced, given recent examples of how E2EE was successfully circumvented both by security agencies to infiltrate criminal networks (e.g. via installing malware on the servers of the encryption service EncroChat)[23] and governments spying on political opponents (e.g. the widespread use of the spyware Pegasus is a cautionary tale, emphasising the risks of abuse and the need for safeguards in liberal democracies to uphold human rights and the rule of law.)[24].

A 2016 report by the Berkman Klein Center for Internet & Society argued against many of the central claims of those seeking to weaken encryption for the sake of public safety.[25] The report challenges the two premises of the argument: that E2EE will become ubiquitous and that law enforcement will have fewer opportunities to surveil criminals. The authors argue both of these fears are unfounded, given both business interests and technological developments. For the former, the report emphasises how both the business model and key functionalities of most social media companies rely heavily on access to user data, meaning E2EE technology is only feasible for a limited set of services. For the latter, the growth of digital communication has led to the generation of a wealth of useful metadata for law enforcement; this is true even where the content of individual messages may be inaccessible. Location data, log-in times and connection records provide more surveillance opportunities than ever before for authorities.

The authors also show how the growing adoption of personal smart devices with networked sensors provide an additional growing stream of data to be intercepted, monitored and recorded by law enforcement. This is especially true when paired with data from other public surveillance technologies such as facial recognition. This can be combined with more than a decade of social media activity and personal information from open-source investigations. These options allow law enforcement to monitor and prosecute criminal activity at scale without threatening E2EE.

An additional argument in support of E2EE is advanced by cybersecurity experts, most vocally in the context of a proposal by the European Commission to combat CSAM through so-called 'client-side scanning' (see below). E2EE is key to safeguarding digital infrastructure by making it harder for third parties to gain unauthorised access to confidential information through measures such as hacking. Beyond the economic cost of such hacks, they can also be used as part of political influence operations or harassment campaigns. Examples include the Russian-linked hack-and-leak campaigns targeting US Democrats in 2016 and the UK Conservative Party in 2019[26]. In both cases, confidential or otherwise secret information was collected and later leaked to damage particular individuals and parties. While E2EE implementation itself is insufficient to completely safeguard against such hacks, it is evident that online harms or systemic risks related to influence operations and online harassment may be enabled by compromised private communications.[27] E2EE could have reduced this harm, for example by ensuring that even if private communications were intercepted, their contents would remain encrypted and unreadable, thereby limiting their potential to be weaponised in such campaigns.

In short, the overview provided above shows that two key fears voiced by law enforcement regarding E2EE do not recognise its importance in enabling rights to privacy and reducing a range of risks online. This includes many of those within the remit of online safety regulators outlined in the section above. Many digital rights groups argue that wide access and uptake of encryption technology may fact safeguard individuals against online harms and mitigate various systemic risks.

State of play: Online Safety regulators and private online communications

As illustrated above, the law enforcement position on the end-to-end encryption of private messaging services is shaped by their focus on harms they are tasked to combat, namely criminal activity. In jurisdictions where specific online safety regulation is already in place, online safety regulators have a very different mandate: the harms or risks they deal with often also involve activity in public communication spaces which do not necessarily constitute outright criminal acts. Online safety regulators like the EU Digital Services Coordinators (DSCs) or the UK's Ofcom are also tasked with regulating the online service itself while lacking powers to sanction criminal or otherwise harmful behaviour by individual users. However, different regulators have different remits in terms of how the respective legislation applies to public and private communications, and as is the case for the Australian eSafety Commissioner, whether the regulator also has remedial powers against individual users. To what extent, then, are private online communications covered under current online safety regulation frameworks?

UK Online Safety Act

The scope of the Online Safety Act includes all user-to-user services and search services but exempts email, SMS, MMS and user-to-user calling. In draft guidance, Ofcom proposed that online services should analyse user-generated content and associated metadata communicated publicly to ensure that they are complying with their legal duties as part of its risk assessment.

The distinction between private and public communication is therefore a key task for service providers under the UK's new online regulation regime as it determines when and where automated content moderation systems must be in place. In its draft regulatory guidance,[28] Ofcom outlines three factors to help service providers make this distinction:

- **Number of UK individuals able to access the content (A):** This is an estimate of individuals based in the UK who can access the content. Notably, this is distinct from the actual number of individuals who are accessing or have accessed the content, or how easy the content is discoverable. Therefore, a low number of actual accesses, or difficulty in discovering the content, does not mean the content is communicated 'privately'.
- **Access restrictions (B):** These include measures such as password-protection of specific content, invite-only access or decryption keys available to specific individuals only. They do not include mere paywalls or service-wide log-in requirements. Additionally, the lack of a search function or specific restrictions on who can interact (e.g. comment) on content do not constitute access restrictions. If no access restrictions are in place, the content should be considered to be communicated publicly. Conversely, when access restrictions are in place, services should also consider the other two factors in determining whether the content is indeed communicated privately.
- **Sharing/forwarding of content (C).** This is the ability of users to make the content accessible to additional users; for example, by reposting content, tagging users or adding new users to closed groups. The easier it is to use such functionality, the more likely the content is communicated publicly. Conversely, the more restrictions are in place in sharing/forwarding functionalities the less public the content may be communicated; this can include limits on number of forwards or access restrictions as outlined above. Notably, the ability to screenshot content and then share it should not influence decisions on whether the content is communicated privately or publicly.

In its consultation material, Ofcom specifically acknowledges that E2EE and related technologies that can facilitate online anonymity pose particular risks in regard to the spread of CSAM and terrorist content, hate speech and harassment as well as online fraud. However, Ofcom also highlights how E2EE "plays an important role in safeguarding privacy online" and how anonymity more generally is "important for historically marginalised groups".[29]

In summary, automated content moderation technologies are not expected to be applied to private or E2EE communications. However, service providers will still be required to assess and take steps to mitigate risks associated with the functionality offered, in close consultation with Ofcom.

Australian Online Safety Act

The Online Safety Act 2021 provides for mandatory codes and standards focussed on child sexual abuse material and 'pro-terror' material. It applies to eight sections of the online ecosystem, including social media services, messaging services, file storage, search, Internet Service Providers (ISPs), app stores, hosts and equipment providers.

Six industry codes drafted by industry associations in Australia were registered in 2023, including for social media services. The head terms to these codes emphasise "the desirability of not intruding upon, and otherwise maintaining the privacy and integrity of, private communications between end-users."[30] Section 6.1 of the head terms, also states that the codes do not require industry participants render methods of encryption less effective, or to monitor private communications between end-users.

Two Industry Standards, drafted by the eSafety Commissioner, are due to come in effect on 22 December 2024. Together, they cover messaging services, file storage services, websites and other services contain similar protections:

> "Providers will not be required to implement systems or technology to detect and remove material where doing so would require the provider to implement or build a systemic weakness, or a systemic vulnerability, into the service or where it would require an end-to-end encrypted service to implement or build a new decryption capability or render methods of encryption used in the service less effective."[31]

Where these exceptions apply, providers are required to take "appropriate alternative action" to detect child sexual abuse material and pro-terror material.

The Basic Online Safety Expectations apply to a wider range of "unlawful and harmful material and activity". They use a similar framework to provide protections for encrypted communications, while maintaining an expectation that providers take reasonable steps to minimise these harms.

EU Digital Services Act and the EU Terrorist Content Online Regulation

The EU Digital Services Act (DSA) and the EU Terrorist Content Online Regulation (TCO) – the core legal texts concerned with the online harms or systemic threats described above – explicitly exclude content that is not disseminated publicly. Both texts define "dissemination to the public" as "the making available of information to a potentially unlimited number of persons, namely making the information easily accessible to users in general, without requiring further action by the content provider, irrespective of whether those persons actually access the information in question."[32] They further specify that the public dissemination also applies to functionality where the sender of the communication is not specifically determining the recipient – for example, public groups or open channels. By extension, interpersonal communication, including emails and private messaging services, are not within scope of these regulations.

Encrypted services are only mentioned once in in the DSA in preamble, recital 20, where it is emphasised that services offering encryption technology to users should not be considered as facilitating illegal activity. They hence maintain their exemption from intermediary liability (as long as they abide by the notice-and-action mechanisms). However, encryption features prominently in the more consumer protection-focused 2018 European Electronic Communications Code. It explicitly promotes the use of E2EE by suggesting "where necessary, encryption should be mandatory in accordance with the principles of security and privacy by default and by design."[33]

Mitigating against online harms in private online communications

All three online safety regulation frameworks require online service providers to conduct risk assessments of their services and functionalities to understand how they may contribute to online harms or systemic risks, and what steps are being taken to mitigate such harms or risks. The UK and Australian regulation include E2EE services in such risk assessments, although currently in the UK those services are exempted from automated content moderation. It remains uncertain whether the provision in Section 121 of the UK Online Safety Act, which allows Ofcom to require the use of 'accredited technologies' to detect CSAM, could effectively result in content moderation within E2EE environments in the future. In Australia, certain exemptions apply on a case-by-case basis; however, services must still take "alternative appropriate action" which can include other forms of automated content moderation. The EU regulations described above, in contrast, appear to bracket out E2EE services entirely although it remains to be seen to what extend the European Commission may demand that Very Large Online Platforms (VLOPs) explain the role of E2EE technology as part of their wider risk mitigation measures. The core issue is how platforms can demonstrate compliance with regulatory requirements while addressing "online harms" and mitigating "systemic risks" within private and/or encrypted communications on their services.

Techniques to increase accountability in private online communications

A variety of technological solutions have been proposed in the context of online safety in E2EE communications.[34] Some of the most invasive and controversial technologies, such as encryption backdoors and client-side scanning, have not been imposed, but they raise significant concerns regarding privacy and security. Message franking and metadata analysis, on the other hand, are considered less controversial. They are already implemented by online service providers such as Meta. This range of potential mitigation measures also demonstrates that E2EE does not necessarily preclude all safety interventions.

- **Encryption backdoors** provide relevant authorities such as law enforcement with special access to encrypted communications, often through additional decryption keys (so-called key escrow systems). Encryption backdoors are essentially purposefully created vulnerabilities in online communication infrastructures – while the intent may be to empower law enforcement, such vulnerabilities can easily be exploited by cyber-criminals to gain unlawful access. They also violate the human right to privacy, as a recent decision by the European Court of Human Rights confirmed.[35]

- **Client-side scanning** is a more sophisticated approach that features most prominently in current EU proposals to combat the spread of CSAM (also known as 'Chat Control' by critics). The proposal would force companies including Apple, Google and Microsoft to 'scan' message content for potential CSAM and generate a report for authorities before the content is encrypted and sent to other users. While technically not breaking end-to-end encryption like encryption backdoors, the indiscriminate scanning of private user contents has been widely criticised by digital rights organisations.[36] There are currently no signs that EU institutions will reach a consensus on the proposal. Notably, however, several companies already scan communications on the 'client' (the device) to detect nudity and flag warnings to users, albeit without automatically reporting such content to authorities. This feature is enabled on Apple's iMessage by default for children's accounts, and Instagram has announced a similar feature.[37] Unlike the mechanisms that the EU proposal foresees, these interventions do not allow the service provider access to communications, but do provide some protection for users, and can enable or encourage user reporting.

- **Message franking** is a technological solution that preserves encryption while allowing for message verification by a specific third party. Described by Meta as part of their work on E2EE technology,[38] message franking uses cryptography to create a 'fingerprint' that links a specific message to a specific sender. This fingerprint is embedded in the message and sent to the recipient. If the recipient choses to report the message, the fingerprint is sent to Meta as part of the report. In this way, platforms can check if a reported message was indeed sent by a specific sender, without having to decrypt the actual contents of the message. This verified proof can then inform subsequent actions, such as sanctioning the sender on the platform.[39]

- **Metadata analysis** is a method rather than a specific technological solution. It focuses on the examination of metadata rather than the actual contents of private communications. As described in the 2016 report by the Berkman Klein Center, online communication generates a variety of 'trace data' beyond actual message content. Metadata analysis uses this data—such as location data, log-in times and connection records— as signals to identify potential threats or emerging risks. This activity data can be assessed and analysed at scale to detect suspicious behavioural patterns. While metadata analysis is commonly used within the cybersecurity community to identify and mitigate vulnerabilities in computer systems, it is increasingly employed by law enforcement, intelligence services, and online platforms to identify terrorist networks or influence operations online. At scale, metadata analysis can hence be a crucial method for platforms in their risk assessment work.[40]

Mitigating online harms through privacy-by-design

The mitigation measures described in the previous section are mainly useful to combat harms that manifest through activity on private communication platforms. However, legislators and online safety regulators have highlighted that in their respective legal texts and guidance, access to private communications is often a prerequisite to the free exercise of human rights. Undermining the integrity of E2EE or other privacy-preserving technologies may threaten the safety of journalists, activists and minority groups while simultaneously increasing the risk of successful hacking attempts that could be used as part of influence or harassment campaigns.

In these cases, online safety may require a privacy-by-design approach. Policymakers should consider how user privacy rights and user anonymity are key components of online safety and should guide the design of online services. Exemplary of such an approach is the 'right to encryption' included in the 2021 coalition agreement of the German Federal government. The corresponding 2024 draft bill by the German Ministry for Federal Ministry of Transport and Digital Infrastructure would, if adopted, mandate messenger, e-mail and cloud services to enable E2EE by default.[44] In the context of existing regulation, online safety regulators may already foster a privacy-by-design approach as part of their oversight role by demanding platforms:

- **Make privacy and security easily accessible for all users by:**
  - Defaulting to E2EE for any private messaging functionality.

Experimental approaches to content moderation: zero-knowledge-proofs

Zero-knowledge proofs (ZKP) are cryptographic protocols that "allows one party to prove the validity of a statement to another party without revealing any additional information beyond the truthfulness of the statement."[41] ZKP is currently used in privacy-sensitive authentication systems such as automated age verification technology. Future iterations of this technology could possibly be used to check content in private communications against external databases. This can be used for example to verify the existence of unlawful content without revealing the actual contents of specific messages. As such, the technology builds on existing classifier-based content detection and known-content hash-based detection.[42] As Derei et al. explain, through ZKP, "applications can commit cryptographically to data that are kept secret while proving that those data have certain properties, such as compliance with regulatory constraint".[43] A ZKP system for content moderation would 'check' the message for specific characteristics. It would then send a signal to relevant authorities or platforms to flag the presence of violating content, without revealing the contents itself. Such technology could prove useful in cases where regulators seek to understand the overall presence of harmful content on their private communication services, but do not seek to prosecute individual transgressions. However, ZKP technology requires a static set of rules to provide such proofs. Judging content as harmful often requires contextual understandings, which are hard to codify for automatic detection through ZKP. In this way, ZKP runs into the same problems of most automated content moderation technologies.

- Making use of clear and accessible language in user interfaces that accurately reflect and communicate the degree of privacy or security of a specific messaging function.
- Abiding by the GDPR principle of data minimisation, which may include the encryption or deletion of metadata where possible. The messenger Signal can be considered a best-practice example in this space.
- Raise awareness of how features that require additional personal information, e.g. mandatory user profiles, may undermine the safety of a broad range of users. This could range from human rights defenders to victims of stalking. These features should therefore be included in mandated risk assessment procedures.

- **Empower users to safeguard themselves against hate, harassment and grooming, and ensuring services are safe by design:**
  - Providing privacy-preserving contact request experiences, where user information (like online status and profile picture) is only shared after the user explicitly consents to this. This includes message request interfaces that hide messages from unknown users by default and ensuring that children's accounts are set to the highest level of privacy settings by default.
  - Considering mechanisms that allow users to hide the fact that they are using a specific platform or messenger service in the first place even if those attempting contact know the user's phone number.
  - Establishing easily accessible reporting mechanisms alongside rigorous blocking functionalities (see message franking technology described above).
  - Putting in place robust measures for preventing repeat offenders from continuing to use the service.

Recommendations

Based on the review above, a variety of recommendations can be made to policymakers and online safety regulators regarding mitigating online harms in private communication.

Specify on a case-by-case basis the online harms within the scope of online safety frameworks which manifest primarily through private online communications. As the review above has shown, key online harms or systemic threats – influence operations, disinformation campaigns, extremism and terrorism, as well as hate and harassment – frequently take place in public online spaces. While private communications also play a role, regulators must be specific to what extent and in what cases they expect harms to manifest through private communications.

Critically interrogate the effect of platform mitigation measures on user safety and privacy. Regulators should consider how platform mitigation measures for one harm or systemic threat may inadvertently undermine user safety down the line. Proportionality ought to be the guiding principle when assessing what design choices, functionalities or moderation tools are appropriate to achieve online safety objectives.

Distinguish between the objectives of law enforcement authorities and online safety regulators. The position of law enforcement authorities on key privacy-preserving technologies like E2EE is informed by the objectives of their work: to prevent, identify and disrupt criminal activity – and hold individuals to account. These objectives are distinct from the goals of online safety regulators, who are mainly tasked to ensure online service providers are abiding by safety standards and perform their due diligence obligations. While E2EE may complicate law enforcement efforts, the technology is not necessarily a hurdle for online regulators to hold platforms to account regarding their assessment and mitigation of 'systemic risks' such as disinformation, terrorism, extremism and hate speech.

Acknowledge that breaking end-to-end encryption can harm online safety. Breaking encryption has been found to violate the right to privacy by the European Court of Human Rights. Access to private communication is foundational to democracy and human rights. Technology or regulatory approaches that undermine E2EE hence run contrary to many of the objectives that are explicated in the online safety frameworks regulators are tasked to uphold. In many cases, online safety would also be improved through the rigorous implementation of privacy-preserving technology through a broader safety-by-design approach.

Demand that platforms develop mitigation measures specific to the different types of services offered. Platforms offer a wide variety of services, many of which include private communication functionality alongside more public means of communication. Each service requires different, targeted mitigation measures to remedy identified harms. A one-size-fits-all approach to platforms may unjustly target private communication functionality if most harms actually manifest through public communication services. Conversely, platforms may misleadingly use terms such as 'encrypted' or 'private' to avoid responsibility. Many of the functionalities that services that are often described as messengers like Telegram or WhatsApp offer, for example, are public channels or open groups that do not use E2EE. Mitigating many of the key online harms manifesting through these functionalities on such services therefore do not require breaking encryption or undermining the integrity of private communication more generally.

Conclusion

This policy brief provided a short overview of the current policy debate on encryption and private online communication. The encryption debate itself precedes the widespread adoption of the internet, culminating in the so-called 'Crypto Wars' in the 1990s. However, the current online policy debate around encryption is driven by two main concerns about how E2EE online services are being used by criminals to avoid detection: namely for terrorist recruitment and planning, and to spread CSAM.

Acknowledging the role of private and E2EE online spaces in exacerbating certain online harms, the policy brief outlined a range of technologies and techniques available to the online safety community. It argued that, in many cases, privacy-preserving options can increase accountability in these spaces. The policy brief further highlighted that many of the online harms most online safety regulators are tasked to combat – namely disinformation, extremist propaganda, and hate and harassment – rarely manifest exclusively in private online spaces. Given their distinct roles, online safety stakeholders should therefore be wary of simply adopting a law enforcement perspective on private online spaces and encryption technologies. This brief demonstrated that online safety may actually be strengthened by the adoption of privacy-by-design principles and widespread access to E2EE technologies.

Beyond the scope of this brief was a discussion of the implication of future technological developments on current encryption standards – namely quantum computing. While advancements are made to 'quantum-proof' encryption, policymakers should remain vigilant of these developments, and advocate for an online safety approach that emphasises how access to the means of private communication, even in the quantum age, is a prerequisite to the free exercise of human rights.

Further Reading

eSafety Commissioner. (2023). "End-to-end encryption trends and challenges — position statement". eSafety Commissioner.
Retrieved from https://www.esafety.gov.au/industry/tech-trends-and-challenges/end-end-encryption

Mayer, Jonathan (2019). "Content Moderation for End-to-End Encrypted Messaging". Princeton University.
Retrieved from https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf

Meta (2023). "Meta's Approach to Safer Private Messaging on Messenger and Instagram Direct Messaging". Messenger News.
Retrieved from https://messengernews.fb.com/wp-content/uploads/2021/12/Metas-approach-to-safer-private-messaging-on-Messenger-and-Instagram-DMs-Sep-23.pdf

Glover, Katlyn; Dila, Mirya; Pate, Neeley; Trauthig, Kristina Inga; Woolley, Samuel C. and Little, Kaiya (2023). "Encrypted Messaging Applications and Political Messaging: How They Work and Why Understanding Them is Important for Combating Global Disinformation". Center for Media Engagement.
Retrieved from https://mediaengagement.org/research/encrypted-messaging-applications-and-political-messaging/

Hendrix, Justin; Sinders, Caroline; Quintin, Cooper; Wagner, Leila Wylie; Bernard, Tim and Mehta, Ami (2023) "What is Secure? An Analysis of Popular Messaging Apps". Convocation & Tech Policy Press. Retrieved from https://www.techpolicy.press/what-is-secure-an-analysis-of-popular-messaging-apps/

Puyosa, Iria. (2023). "Protecting point-to-point messaging apps: Understanding Telegram, WeChat, and WhatsApp in the United States". Atlantic Council DFR Lab.
Retrieved from https://www.atlanticcouncil.org/in-depth-research-reports/report/point-to-point-messaging-apps/

Meta. (2023). "Messenger End-to-End Encryption Overview". Meta.
Retrieved from https://engineering.fb.com/wp-content/uploads/2023/12/MessengerEnd-to-EndEncryptionOverview_12-6-2023.pdf#page=17.37

Access Now, Alternatif Bilisim (Alternative Informatics Association), Asociația pentru Tehnologie și Internet Romania, Aspiration, Bits of Freedom The Netherlands, CDT Europe, Chaos Computer Club, Citizen D / Državljan D Slovenia, D3 - Defesa dos Direitos Digitais Portugal, D64 – Center for Digital Progress Germany, Danes je nov dan Slovenia, Defend Digital Me, Der Kinderschutzbund Bundesverband e.V., Digitale Gesellschaft Germany, Digital Rights Ireland, Digital Society Forum, Digital Society Switzerland, Digitalcourage Germany, ECNL, EFF, Electronic Frontiers Australia, Electronic Frontier Norway, Electronic Privacy Information Center (EPIC), Epicenter.works Austria, European Digital Rights (EDRi), European Sex Workers Rights Alliance (ESWA), Foundation for Information Policy Research (FIPR), 5th of July Foundation Sweden, Homo Digitalis Greece, ICCL Ireland, Internet Society, Internet Society Portugal Chapter, IT-Pol Denmark, Iuridicum Remedium Czechia, La Quadrature du Net France, Lobby4kids - Kinderlobby, Metamorphosis Foundation, National Association for Free Software Portugal (ANSOL), OpenMedia, Politiscope Croatia, Privacy & Access Council of Canada, SHARE Foundation Serbia, SUPERRR Lab Germany, The Commoners, The Digitas Institute Slovenia, The Tor Project, Vrijschrift.org the Netherlands, and Xnet Spain. (July 2023). "Joint statement on the future of the CSA Regulation". EDRi.
Retrieved from https://edri.org/our-work/joint-statement-on-the-future-of-the-csa-regulation/".

Riana Pfefferkorn (2022). "Content-oblivious Trust and Safety Techniques: Results from a Survey of Online Service Providers," Journal of Online Trust and Safety 1.2.
https://doi.org/10.54501/jots.v1i2.14

Endnotes

1 Kravets, D. (2015). "UK prime minister wants backdoors into messaging apps or he'll ban them". Ars Technica. Retrieved from https://arstechnica.com/tech-policy/2015/01/uk-prime-minister-wants-backdoors-into-messaging-apps-or-hell-ban-them/

2 EDRi. (2024). "Be scanned – or get banned!". EDRi. Retrieved from https://edri.org/our-work/be-scanned-or-get-banned/
Euractiv. (9 May 2023), "EU Council's legal opinion gives slap to anti-child sex abuse law". https://www.euractiv.com/section/data-privacy/news/eu-councils-legal-opinion-gives-slap-to-anti-child-sex-abuse-law/

3 For further discussion of the use of 'private' to describe relatively large online spaces with few entry requirements, for example on Telegram, see: ISD & CASM. (2023). "Researching the Evolving Online Ecosystem: Telegram, Discord & Odysee".
Retrieved from: https://www.isdglobal.org/isd-publications/researching-evolving-online-ecosystem-telegram-discord-odysee/

4 Hendrix, J. et al. (2023). "What is Secure? An Analysis of Popular Messaging Apps". Convocation & Tech Policy Press.
Retrieved from https://www.techpolicy.press/what-is-secure-an-analysis-of-popular-messaging-apps/

5 For a more detailed discussion of the private/public spectrum, and its implications for researchers as well as policymakers, please consult Tuck, H. et al. (2023) "Researching the Evolving Online Ecosystem: Telegram, Discord and Odysee ". Institute for Strategic Dialogue. Retrieved from https://www.isdglobal.org/wp-content/uploads/2023/04/Researching-the-Evolving-Online-Ecosystem_Telegram-Discord-Odysee.pdf

6 Davey, J. and Ebner, J. (2017). "The Fringe Insurgency. Connectivity, convergence, and mainstreaming of the extreme right". Institute for Strategic Dialogue. Retrieved from https://www.isdglobal.org/isd-publications/the-fringe-insurgency-connectivity-convergence-and-mainstreaming-of-the-extreme-right/

7 Trauthig, I. and Woolley, S. (2022). "Digital Disinformation Increasingly Targets the Most Vulnerable".
Centre for International Governance Innovation.
Retrieved from https://www.cigionline.org/articles/digital-disinformation-increasingly-targets-the-most-vulnerable/

8 Harry, L. (2023). "Misinformation on WhatsApp: Insights From the Caribbean Diaspora". The Center for Media Engagement.
Retrieved from https://mediaengagement.org/research/whatsapp-misinformation-caribbean-diaspora/

9 Avelar, D. (30 October 2019). "WhatsApp fake news during Brazil election 'favoured Bolsonaro'". The Guardian. Retrieved from https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests

10 Jaswal, S. (2024). "Inside the BJP's WhatsApp Machine". Pulitzer Center.
Retrieved from https://pulitzercenter.org/stories/inside-bjps-whatsapp-machine

11 Jaffe, L. and Gillum, J. (2021). "'This Is War': Inside the Secret Chat Where Far-Right Extremists Devised Their Post-Capitol Plans". Rolling Stone. Retrieved from https://www.rollingstone.com/culture/culture-news/capitol-riot-far-right-extremists-telegram-1120511/

12 Whittaker, J. (2022) "Online Radicalisation: What we know". Radicalisation Awareness Network.
Retrieved from https://home-affairs.ec.europa.eu/system/files/2023-11/RAN-online-radicalisation_en.pdf

13 United Nations. (n.d). "Targets of hate". United Nations.
Retrieved from https://www.un.org/en/hate-speech/impact-and-prevention/targets-of-hate

14 See e.g. Martiny, C. et al. (2024). "Online Gendered Abuse and Disinformation During the 2024 South African Elections". Institute for Strategic Dialogue. Retrieved from https://www.isdglobal.org/isd-publications/online-gendered-abuse-and-disinformation-during-the-2024-south-african-elections/

15 Please note CSAM is not covered in detail in the rest of the policy brief as the dissemination of CSAM falls outside of the scope of ISD's Digital Policy Lab, and was not a major focus of the Working Group discussions.

16 National Center for Missing and Exploited Children. (n.d.). "End-To-End Encryption". National Center for Missing and Exploited Children.
Retrieved from https://www.missingkids.org/theissues/end-to-end-encryption

17 European Commission. (11 May 2022). "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down rules to prevent and combat child sexual abuse".
Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A209%3AFIN

18 US Senate. S.1207 - EARN IT Act of 2023 118th Congress (2023-2024).
Retrieved from https://www.congress.gov/bill/118th-congress/senate-bill/1207

US House of Representatives. H.R.2732 - EARN IT Act of 2023.
Retrieved from https://www.congress.gov/bill/118th-congress/house-bill/2732

19 The Verge. (19 June 2024). "EU chat control law proposes scanning your messages — even encrypted ones". Retrieved from https://www.theverge.com/2024/6/19/24181214/eu-chat-control-law-propose-scanning-encrypted-messages-csam Cameron, D. (2020). "Senate Turns Shitty Orwellian Surveillance Bill Into Pointless Bill That Mostly Undermines Free Speech". Gizmodo. Retrieved from https://gizmodo.com/senate-turns-shitty-orwellian-surveillance-bill-into-po-1844252068

20 Nemchick, E. (2023). "13 WhatsApp scams to know and avoid in 2024". Norton. Retrieved from https://uk.norton.com/blog/online-scams/whatsapp-scams

21 European Police Chiefs. (2024). "Joint Declaration of the European Police Chiefs". European Police Chiefs. Retrieved from https://www.europol.europa.eu/media-press/newsroom/news/european-police-chiefs-call-for-industry-and-governments-to-take-action-against-end-to-end-encryption-roll-out

22 Patel, P., Barr, W., McAleenan, K., and Dutton, P. (2019). "Open Letter: Facebook's 'Privacy First' Proposals". Home Office, United States Department of Justice, United States Department of Homeland Security, and Department of Home Affairs. Retrieved from https://www.justice.gov/opa/pr/attorney-general-barr-signs-letter-facebook-us-uk-and-australian-leaders-regarding-use-end

23 Europol. (2023). "Dismantling encrypted criminal EncroChat communications leads to over 6 500 arrests and close to EUR 900 million seized". Europol. Retrieved from https://www.europol.europa.eu/media-press/newsroom/news/dismantling-encrypted-criminal-encrochat-communications-leads-to-over-6-500-arrests-and-close-to-eur-900-million-seized

24 Amnesty International. (2021). "Forensic Methodology Report: How to catch NSO Group's Pegasus". Amnesty International. Retrieved from https://www.amnesty.org/en/latest/research/2021/07/forensic-methodology-report-how-to-catch-nso-groups-pegasus/

25 Berkman Center for Internet & Society at Harvard University. (2016). "Don't Panic. Making Progress on the 'Going Dark' Debate". Berkman Center for Internet & Society at Harvard University. Retrieved from https://cyber.harvard.edu/pubrelease/dont-panic/Dont_Panic_Making_Progress_on_Going_Dark_Debate.pdf

26 Rawlinson, K. and Mohdin, A. (2019). "Russia involved in leak of papers saying NHS is for sale, says Reddit". The Guardian. Retrieved from https://www.theguardian.com/uk-news/2019/dec/07/russia-involved-in-leak-of-papers-saying-nhs-is-for-sale-says-reddit

27 Lack of end-to-end encryption is just one, albeit significant, cybersecurity vulnerability. Often, hackers gain unauthorized access not because of sophisticated technology, but through what is called social engineering, i.e. convincing individuals with access to give up personal information or access keys. Members of the Hillary Clinton campaign fell victim to such 'phishing' emails, giving hackers access to their Gmail inbox through a fake Google password reset alert. https://apnews.com/article/moscow-north-america-ap-top-news-hillary-clinton-phishing-addc2727b0b04c1d80ab6ca30c4dc77e

28 Ofcom. (2023). "Protecting people from illegal harms online. Annex 9: Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act". Ofcom. Retrieved from https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-9-draft-guidance-on-content-communicated-publicly-and-privately-under-the-online-safety-act/?v=330407

29 Ofcom. (2023). "Protecting people from illegal harms online. Summary of each chapter". Ofcom. p.9. Retrieved from https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/a-summary-of-each-chapter/?v=330394

30 eSafety. (2023). "Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms ". eSafety. Section 6.1. Retrieved from https://www.esafety.gov.au/sites/default/files/2023-09/Consolidated-Industry-Codes-of-Practice-Head-Terms-12-September-23.pdf

31 eSafety. (2024). "Explanatory Statement", eSafety. Retrieved from: https://www.esafety.gov.au/sites/default/files/2024-06/ExplanatoryStatement-Online-Safety-DesignatedInternetServices-Class1A-Class1BMaterial-IndustryStandard2024.pdf

32 See Preamble 14 both Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online & Regulation (EU) 2022/2065 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

33 Preamble 97 of Directive (EU) 2018/1972 establishing the European Electronic Communications Code

34 For a useful elaboration, also see Mayer, J. (2019). "Content Moderation for End-to-End Encrypted Messaging". Princeton University. Retrieved from https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf

35 Electronic Frontier Foundation. (2024). "European Court of Human Rights confirms: weakening of encryption can violate the human right to privacy". EDRi. Retrieved from https://edri.org/our-work/european-court-of-human-rights-confirms-weakening-of-encryption-can-violate-the-human-right-to-privacy/

36 EDRi. et. al. (2024). "Statement on the future of the CSA Regulation". EDRi. Retrieved from https://edri.org/wp-content/uploads/2024/07/Statement_-The-future-of-the-CSA-Regulation.pdf

37 Associated Press. (12 April 2024). "Instagram begins blurring nudity in messages to protect teens and fight sexual extortion". https://apnews.com/article/instagram-meta-nudity-sexual-extortion-social-7bea9b1244ea023fb85265672bc.

38 Meta. (2023). "Messenger End-to-End Encryption Overview". Meta. Retrieved from https://web.archive.org/web/20241217193832/https://messengernews.fb.com/wp-content/uploads/2021/12/

Metas-approach-to-safer-private-messeaging-on-Messenger-and-Instagram-DMs-Sep-23.pdf

39 Other versions of message franking also enable verification by a third party in online spaces where users operate anonymously, e.g. Signal. See Tyagi, N. et al. (2019) "Asymmetric Message Franking: Content Moderation for Metadata-Private End-to-End Encryption". Cryptology ePrint Archive. Retrieved from https://eprint.iacr.org/2019/565.pdf

40 See for example: Ibid. Meta (2023)

41 Sánchez, J.M. (2024). "What is Zero Knowledge Proof (ZKP)". Veridas. Retrieved from https://veridas.com/en/what-is-zero-knowledge-proof/

42 Mayer, J. (2019). "Content Moderation for End-to-End Encrypted Messaging". Princeton University. Retrieved from https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf

43 Derei, T. et al. (2023). "Scaling Zero-Knowledge to Verifiable Databases". Association for Computing Machinery. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3595647.3595648

44 Reuter, M. and Meister, A. (13 February 2024). „Digitalministerium plant Recht auf Verschlüsselung ". Netzpolitik. Retrieved from https://netzpolitik.org/2024/referentenentwurf-digitalministerium-plant-recht-auf-verschluesselung/

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org