Protecting the Public from Abusive AI-Generated Content across the EU

In consultation with

Microsoft

WPL Women Political Leaders

meNAble against gender-based violence online

Co-funded by the European Union

IWF Internet Watch Foundation

weprotect Global Alliance

esu european seniors' union

Table of Contents

Foreword

Nanna-Louise Linde

Vice President, European Government Affairs, Microsoft

As a long-standing technology partner to European governments, businesses and citizens, Microsoft seeks to ensure that the continent benefits from digital technologies and artificial intelligence (AI), while continuing to respect the rights of EU citizens online.

AI is no longer a distant prospect but a present reality, reshaping the business landscape, revolutionizing healthcare, and accelerating scientific discovery across the EU. Yet, as with any transformative technology, AI brings potentially significant challenges as well as immense opportunities. As a technology company providing AI services, we bear a responsibility to make sure that the solutions we deliver are deserving of public trust.

The start of the new EU mandate offers an opportunity to reflect on how best to leverage new technologies for the benefit of people across the continent – driving innovation and competitiveness – as well as to take proportionate steps to protect people from potential abuses of the same technology. At Microsoft, we are looking forward to working with the new decision makers in the European institutions as they embark on the 2024-2029 mandate.

Strong political leadership is all the more necessary as we stand at the beginning of a new age of technological innovation. As President of the European Commission, Ursula von der Leyen, said, *"Europe is leading the way in making AI safer and more trustworthy, and on tackling the risks stemming from its misuse"*. In this pursuit however, the EU should not lose sight of AI's central role in driving the continent's digital transformation and potential for economic growth. Indeed, the EU should therefore *"focus on becoming a global leader in AI innovation"* as emphasized by President von der Leyen in her political guidelines. In her commitment to protecting democracy, President von der Leyen also expressed her intention to continue strengthening the EU's approach to AI-produced content in the current mandate.[1]

Advancing innovation and safety will require a balanced, whole-of-society approach that recognizes the respective roles of government, civil society, and industry. The EU is already at the forefront of creating a robust legal and

regulatory frameworks, making industry players accountable for the development of safe online products, including AI. Microsoft recognizes the legislative developments undertaken during the 2019-2024 mandate of the Commission and stands ready to engage in dialogue withEU stakeholders on implementing these in an effective and proportionate way. We also see a need for modernized criminal and other laws to help address the misuse of AI. The pace of innovation calls for a continued focus on these challenges as the AI revolution unfolds.

Our annual safety research[2] reveals the scale of the potential challenge. Certain societal groups are disproportionately at risk from deliberate misuse of this technology. We therefore see a need for practical steps to protect people - most notably children, women, and older adults - from the harms which arise from abusive AI-generated content.

In this white paper, we outline steps that Microsoft is taking to address this harm, as well as policy recommendations to build on the existing efforts and rules that address these issues head-on.

Central to our recommendations is the need to establish clear and proportionate rules that protect individuals while enabling Europe to continue innovating. In our paper, we advocate for the EU to integrate provenance tools, strengthen appropriate existing legal frameworks, and enhance measures that put victim-based decision making at the forefront.

As a company, we know we need a strong safety architecture for our services, grounded in safety by design, and incorporating durable media provenance and watermarking. Equally, we must continue to safeguard our services from abusive content and conduct (whether synthetic or not), through robust collaboration across industry and with governments and civil society, supported by ongoing education and public awareness efforts. It is crucial that we build trust in AI across society for its benefits to be fully realized.

In the context of the EU's mature regulatory landscape, we center our recommendations on enhancing the response to the misuse of AI, through the lens of three key risk areas:

1. Protecting children from online exploitation.

2. Safeguarding women from non-consensual intimate imagery.

3. Safeguarding older adults, especially against AI-enabled fraud.

Table 1: Policy recommendations for protecting vulnerable groups from AI-related harms.

 **Protecting Children**	 **Safeguarding Women and Girls**	 **Safeguarding Older Adults**	**Cross-harm Recommendations**
Drive adoption of Child Sexual Abuse (CSA) recast Directive Continue pushing for compromise on CSA Regulation Establish taskforce to study AI-enabled CSA Drive Annual Youth Policy Dialogue with stakeholders	Adopt Directive on Violence against Women and Domestic Violence[3] across EU and ensure strong implementation mechanisms Expand scope of the Directive to include all non-consensual content Ensure existing DSA rules are leveraged to tackle online gender-based violence (GBOV)	Strike a proportionate balance between the need to detect AI-generated fraud and the EU data protection instruments Ensure better intergenerational solidarity in the new EU mandate	Prohibit the stripping, tampering with or removal of provenance metadata Support and enhance cross-EU public education campaigns on AI and synthetic content

The challenges we face are significant, but so is the opportunity. By proactively addressing these issues, we can build a future where AI enhances human creativity, protects individual privacy, and strengthens the foundations of our democracy.

At Microsoft, we are committed to playing our part, but we recognize that we cannot do it alone. We welcome engagement and feedback from stakeholders across the EU's digital ecosystem. It is essential that we get this right, and that means working together.

Microsoft stands for technology that is a positive force in society and people's lives, in line with our mission to empower every person and organization on the planet to achieve more. The time for action is now.

Part I: Diagnosing the problem of abusive AI-generated content

Each day, millions of people use powerful generative AI tools to supercharge their creative expression. In many ways, AI is going to create exciting opportunities for all of us to bring new ideas to life. But, as these new tools come to market from Microsoft and across the tech sector, we must take steps to ensure that these new technologies are resistant to abuse and maintain trust in the information ecosystem.

In recent years, the term "deepfake" has become part of our everyday jargon. It was coined in 2017[4], the same year that a fake lip-sync video of former US President Barack Obama was released. Since that video came out, deepfake images, videos and audio, varying in degrees of sophistication, have flooded the discourse.

Yet, media manipulation is not new. It dates back to well before the digital age. Photographers and artists have long manipulated photos to create deceptive content.[5] Totalitarian rulers such as Stalin and Hitler notoriously used such techniques to alter photographs for propaganda purposes. In the second half of the century, the introduction of photo editing software in the 1990s led to a surge in doctored images[6]. While this manipulation is not new, the development of generative AI technology has increased the risk[7] of abusive content. Thanks to more advanced technology, we are now dealing with AI-generated content that is difficult to distinguish from real images, videos or audio.

Figure 1: Non-exhaustive timeline of deepfake examples making headlines**.**

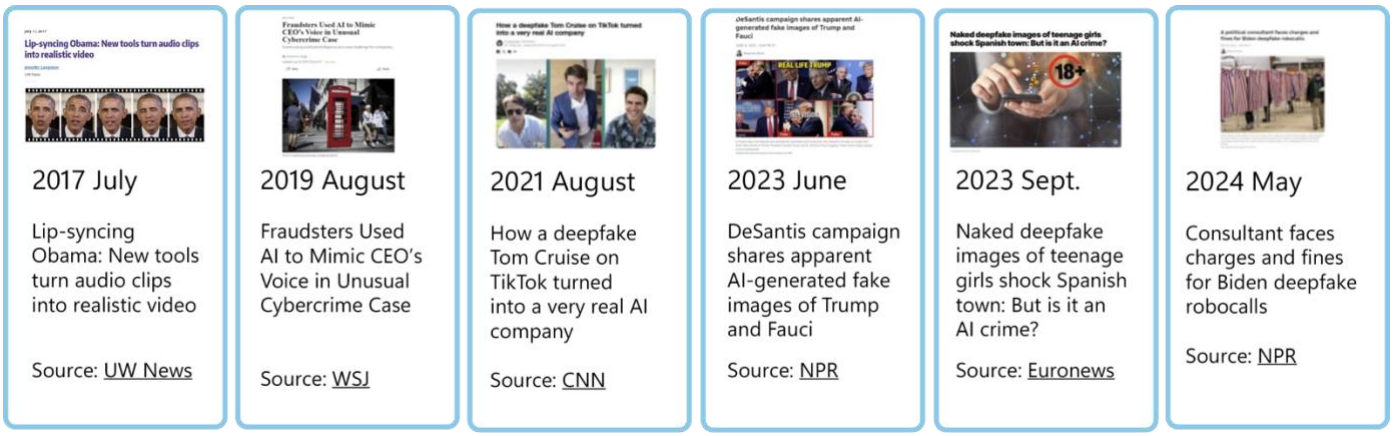


Figure 2: Timeline of Midjourney versions (Prompt: a man running in the meadow photography)**.**

The technology has become easier to access, learn, and use, making the creation of realistic deepfakes more convenient for cybercriminals and other bad actors. Additionally, as we have seen over time, the same technology has facilitated the broad distribution and weaponization of this harmful content. It is no surprise that in our most recent Global Online Safety Survey, 72 % of people were worried about deepfakes. Research shows that women are far more likely to report experiencing such fears than men. According to a study by Deeptracelabs[8], 96% of deepfake videos online concern material depicting nudity or sexually explicit activities, most of them targeting women. In the World Economic Forum's Global Risk Report[9], they list misinformation and disinformation, including those caused by AI, as the top short-term risk for 2024.

Coupled with this concern about the spread of abusive AI-generated content is increasing difficulty to identify it as fake. A recent study found that only 17% of adults reported feeling confident about spotting deepfakes, with most people (66%) unsure if they would be able to spot them.

Malicious AI-generated content is not just cause for concern in the future—today, we see AI tools being abused by bad actors to cause real world harms that will require a whole-of-government and whole-of-industry response. The promise of AI is great, and AI technologies are already delivering public benefits. However, we must also recognize that the same tools can be used as weapons against the public.

In the following examples, we identify four manifestations of synthetic AI content already causing real life harms to vulnerable groups in Europe – which form the basis of our recommendations in Part III.

Europe remains a hub for online child sexual abuse and exploitation

As highlighted in Commission President von der Leyen's political guidelines, and Executive Vice-President (EVP) Henna Virkkunen's mission letter, protecting children remains one of the EU's top priorities.

We share this commitment and recognize AI's potential both to exacerbate, and to address, the harm of online child sexual abuse and exploitation.

In 2023, more than half (51%) of the content the Internet Watch Foundation (IWF) actioned for removal[10] was traced to hosting services in EU Member States. The IWF additionally reports[11] on how AI is being used to create child sexual abuse imagery, posing new threats both towards existing victims of child sexual abuse, but also to new potential victims. Many of the AI-generated images and videos of children being hurt and abused are so realistic that they can be incredibly difficult to tell apart from real imagery of children.

The U.S. National Center for Missing and Exploited Children (NCMEC) is already seeing the impact of generative AI on reports into its CyberTipline. In 2023 alone, NCMEC received 4,700 reports[12] related to synthetic child sexual abuse material (CSAM). While this number is a fraction of the overall number of reports that NCMEC received in 2023 (36 million reports), it is telling of the potential misuse of AI to exponentially increase the production of this exploitative content.

WeProtect Global Alliance and Thorn have also published research[13] reviewing certain technologies' risks and opportunities in the fight against online child sexual exploitation, where 'Generative AI' is listed as a key risk area. It notes its harm for creating new ways to sexually exploit and revictimize children, enable the misuse of children's benign imagery in training models, reduce the social and technical barriers to sexualizing minors, and impede victim identification.

Additionally, the availability of AI "nudifiying" tools has resulted in children making, or attempting to make, indecent images of one another – images of children created with the intent to bully or shame classmates may in fact constitute child sexual abuse material. We also know that child sexual exploitation is a gendered harm that disproportionately impacts girls, and nudifying tools are no exception. For example, last year in Spain a prosecutor received 20+ complaints from families who claim their daughters' social media pictures were used by teenage boys to create AI-generated nudes. The artificial images were then widely distributed on social media and communication applications, causing significant psychological harms to the victims. The verdict has yet to be announced and could form the first official sentencing for creation of AI-generated child sexual abuse in the EU.

According to Europol[14], the misuse of AI goes beyond the creation of synthetic content to helping perpetrators create fake identifies for the purposes of grooming children, or to depict a sexual encounter to financially exploit or threaten a child. The WeProtect Global Alliance Global Threat Assessment from 2023[15] also finds that

perpetrators may use AI to mask child sexual abuse material to evade detection, and pool information on how to destroy evidence and evade law enforcement. AI may also increasingly be used to target young people for financial sextortion, a risk that has risen alarmingly in recent years. This risk, predominantly targeting boys and young men, sees perpetrators deliberately play on fears of nude imagery being shared to demand money, sometimes with tragic consequences.

Synthetic CSAM cannot be disregarded because it creates real harm, including to the viewer. Hundreds of thousands of reports of AI-generated CSAM could also easily overflow an already strained reporting ecosystem. Europol warns that this influx may make it harder to identify and rescue 'real-life' victims, or divert law enforcement resources from active investigations, by creating uncertainties about which images depict real children. In 2022, NCMEC had already seen a 7200% increase in financial sexual extortion from 2021-2022.

Additionally, the intersectionality of this crime must be considered in order to coordinate holistic responses. In particular, this crime is deeply gendered. In 2023, INHOPE[16] found that 95% of the victims in child sexual abuse material are female.

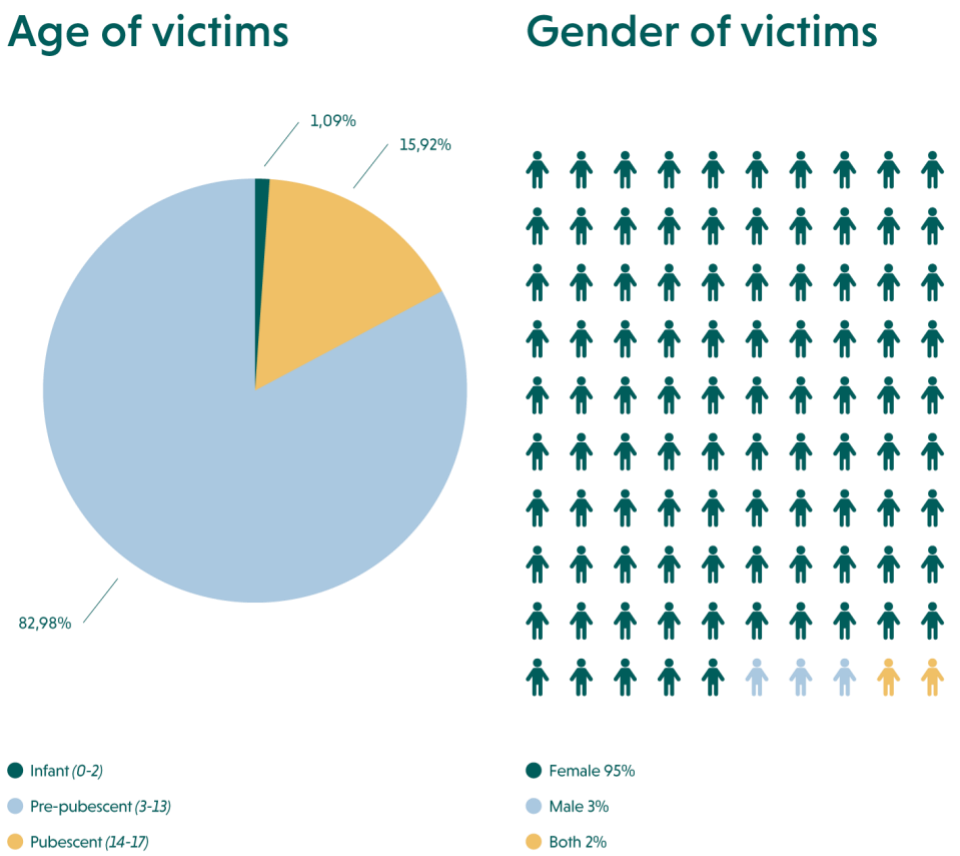


Figure 3: The number of prepubescent victims of CSAM increased from 11% to 16% between 2022 and 2023, with roughly 95% of the victims being female.
Source: InHope, Annual Report 2023[17]

Older adults are increasingly targeted by AI-powered scams

A Europol investigation has reported an increasing number of scams that leverage AI-powered technologies to deceive older adults that are particularly vulnerable due to limited familiarity with technology.

This past year, Princess Leonor of Spain was impersonated online, and her images were used to target victims globally. The fraudulent profiles, many of which were created by AI, falsely presented themselves as the princess offering financial aid. An analysis done by El Pais[18] suggests that the scammers primarily targeted older adults, claiming to offer special assistance to those over 60 years old and those in vulnerable circumstances. Most recently, a French woman was duped out of hundreds of thousands of euros[19] by scammers posing as actor Brad Pitt. The scammers used AI to create images of the actor undergoing cancer treatment to illicit the victim to send money over the span of a year.

A cross-European study by Signicat[20] revealed that 76% of respondents believe that fraud is now a bigger threat than three years ago, and that 66% of those view AI-driven identity fraud as an even greater concern.

A study[21], conducted by Censuwide in 2024, found that not only is fraud increasingly present, but that its AI manifestations, especially deepfakes for the purpose of impersonation, is emerging as the top type of identify fraud reported by European financial institutions.

An estimated 42.5% of detected fraud attempts were now found to have been using AI. Deepfake frauds now account for 6.5% of all fraud attempts, representing a 2137% increase over the last three years.

A recent project by the Palacký University in Olomouc, Czechia, in cooperation with CEDMO[22], found that older adults are especially susceptible to such attacks. Testimonies pointed to seniors losing their savings following cases of deepfakes of high-profile individuals recommending investing in 'guaranteed' investment products.

The Commission's 2023 report[23] on the state of the Digital Decade 2030 objectives highlighted that 46% of Europeans, especially the older demographic, lacked the basic digital skills necessary for both the use, and safeguarding against, technological tools like AI.
Looking to the new EU mandate, we have a unique opportunity to reinforce upskilling and reskilling efforts across the EU's older adult population, so that they are better equipped to protect themselves against potential misuses of technology – including AI - and can continue to reap its benefits.

With the right diffusion, or spread, of transformative technologies like generative AI, older adult populations in the EU could benefit from improved healthcare[24] – with earlier diagnosis of diseases – more independent living[25], and reduced loneliness[26].

Figure 4: 75% of adults aged more than 50 years old who have experienced cybercrime believe that they have been a target of a scam using technology.

Deepfakes could pose risks to electoral processes

In fall 2023, two days before Slovakia's elections, an audio recording spread online[27] in which one of the top candidates, Michal Šimečka, and journalist Monika Tódová appeared to boast about rigging the election. Although they immediately denounced the audio as fake, it was posted during a 48-hour moratorium ahead of polls opening, which under the country's election rules meant that politicians and media outlets were supposed to refrain from political campaigning. Although some platforms removed or placed warnings on the post, it did not stop the spread[28] of the recording which quickly became viral. The election had already been a tight race between Šimečka and his opponent, and when the race was eventually called, it was a five-point win[29] for Šimečka's opponent. While it is impossible to credit the deepfake for the result, the spread is within the typical statistical error rate, and its impact cannot be easily dismissed.

Slovakia is not the only country to have AI impact its elections. Election deepfakes also played a role in Turkey's elections in 2023[30]. Another cause of concern raised in a study is that as the number of deepfakes increases, so does uncertainty among the population regarding authentic content. 40% of respondents to the study indicated a sense of skepticism, being misled or being misinformed.

Yet, there is cause to be optimistic. For example, despite initial concerns, experts found there to be little evidence that AI-generated disinformation meaningfully impacted the European Parliament elections.[31] Likewise, electoral processes in Member States, including the snap elections in France, along with elections in Austria, Belgium, Ireland, Lithuania, Portugal, and Romania, were not impacted by a surge in deceptively realistic AI-generated content going viral and influencing voting behavior. While Romania's Constitutional Court annulled the country's presidential election citing multiple irregularities and violations of electoral legislation, there were no reports around widespread use of deceptive AI-generated content.[32] As a result, the consensus was clear. AI has thus far been used in typical political ways—some negative campaigning—but often, to better connect with voters. Nevertheless, a focus on addressing the challenges that deceptive and abusive AI-generated content can pose to the information environment, specifically during elections, is a whole-of-society priority which necessitates

continued attention and joint action by industry, the public sector and civil society.

As Commissioner for Justice, Democracy and the Rule of Law, Michael McGrath put in his Confirmation Hearing[33] in November 2024, the harmful use of AI has the potential to serve as a disinformation tool to manipulate elections and democratic processes. This is why the announcement of the upcoming European Democracy Shield comes at a critical time, presenting a unique opportunity for the EU to continue countering disinformation and foreign manipulation powered by AI in the years to come. This initiative was also highlighted and reinforced by Executive Vice-President for Tech Sovereignty, Security and Democracy, Henna Virkkunen.

Synthetic non-consensual intimate imagery (NCII) is weaponized against women

Shortly before the Northern Irish legislative elections in 2022, a 24-year-old local politician, Cara Hunter, was attending her grandmother's 90th birthday when she received a message on her cellphone from an unknown number. The message was from a man inquiring if she was the woman in an explicit video[34]. The man then shared the 40-second video clip—an AI-generated deepfake of Hunter performing a sexual act—which quickly spread around the world. Hunter was subsequently bombarded by sexual and violent messages[35], humiliating insinuations, and was even sexually propositioned[36] on the street. She lost trust within her community after having spent years building it. While Hunter went on to narrowly win her election, she felt that the video tarnished her reputation in a way that will have repercussions for the rest of her life. Such synthetic non-consensual intimate imagery (NCII) is not a new risk—but it is one that is exacerbated by generative AI.

Before becoming Italy's Prime Minister, in 2020, Georgia Meloni also saw a deepfake pornographic video of herself[37] surfacing on the internet. Meloni brought a civil lawsuit against the two men allegedly involved in creating and disseminating the video, and is seeking compensations for the damages produced. In her testimony in court[38] Meloni stated *"If we let it pass that the face of any woman can be mounted on the body of another woman, with the advent of artificial intelligence we will find our children in these situations, which is exactly why I believe it's more than legitimate to fight this battle".*

Online violence has a chilling effect on the political ambitions and engagement of women and girls, diminishing their presence and agency in public life. Women are frequently and disproportionately subjected to abusive online content, aimed at silencing or delegitimizing them. In the words of Professor McGlynn, "*more women and girls are going to be reluctant to go into politics. Those who are in politics have to constantly second-guess what they're seeing online and worry about how even innocent interactions might be manipulated and weaponized against them."*

This risk not only affects women in public life, but women and girls all over the world - as online services continue to be exploited as an instrument for numerous forms of violence, including intimate image abuse. A recent research paper[39] found that in 2020, over 100,000

images of Irish women[40] and girls were leaked online, leading to the introduction of new legislation criminalizing all forms of image-based sexual abuse.

The creation and dissemination of synthetic NCII predates the advent of generative AI. In 2019, a report by Sensity AI found that 96% of so-called "deepfakes" were pornographic, and of those, 99% were made of women. As reported in the 2024 European Women's Lobby Report on Cyber Violence Against Women[41], such content may have negative repercussions at the individual, organizational and societal levels, causing psychological, financial, and societal harm.

Such content has long been used to shame, harass, and extort the person depicted. Whether real or synthetic, the release (or threat to release) of such imagery has real and lasting impacts for the victims, including emotional and reputational consequences. The harm is virtually irreparable — once images have been shared, they can be re-distributed widely across the web.

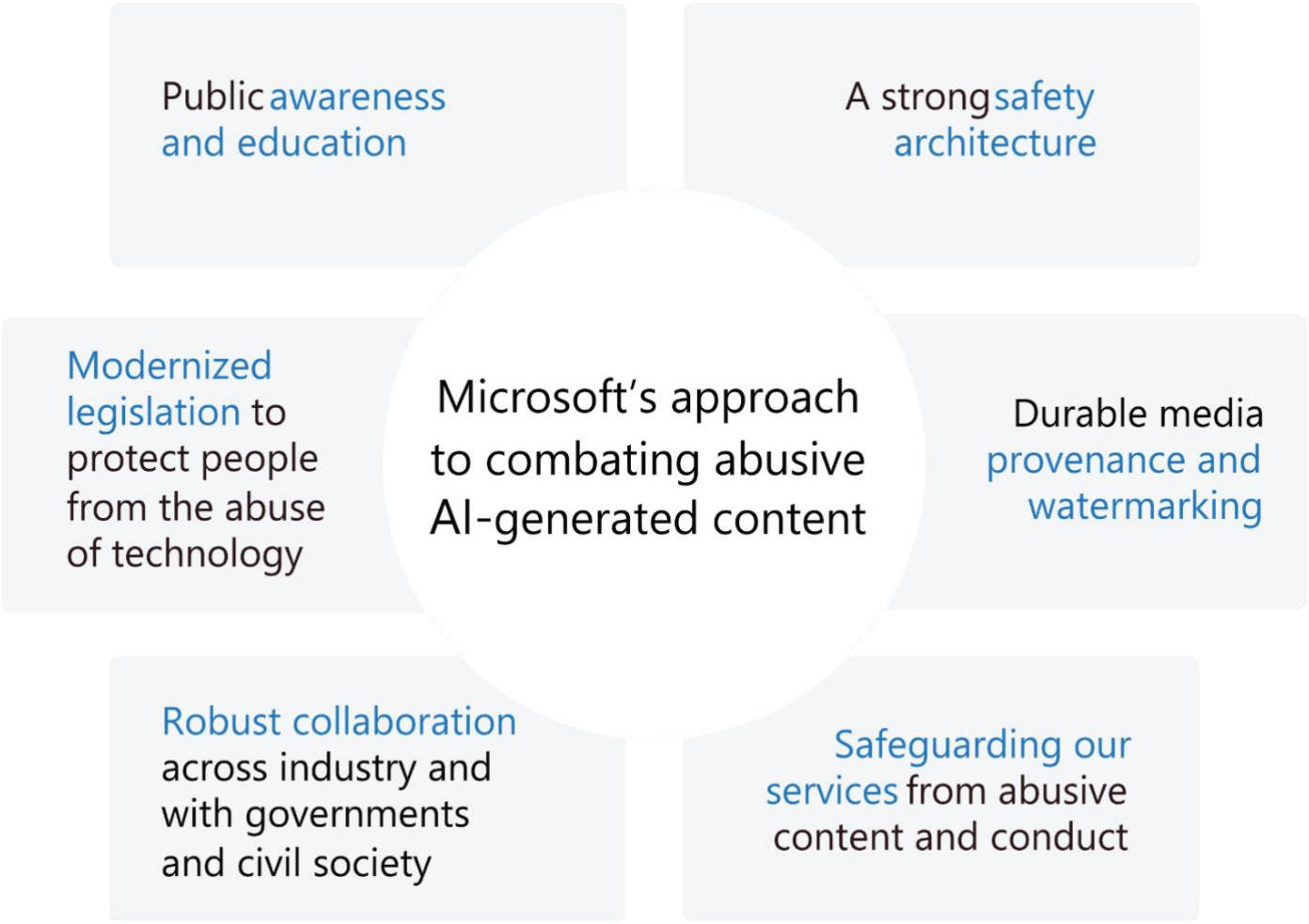
Part II: Microsoft's approach to combating abusive AI-generated content

Stakeholders across the EU are grappling with how to address the challenges associated with abusive AI-generated content. It is no coincidence that during the recent *Commissioner Hearings*[42], Commissioner for Democracy, Justice and the Rule of Law, Michael McGrath expressed his intention to strive for the right balance between achieving the highest potential of AI and protecting citizens from the damages caused by it. Commissioner for Preparedness, Crisis Management and Equality, Hadja Lahbib also expressed her commitment to continue to combat the online manifestations of gender-based violence, and Commissioner for Internal Affairs Magnus Brunner underlined his support for an agreement on the EU's Child Sexual Abuse Material Regulation.

Microsoft is committed to taking a responsible, balanced approach that protects the public from harm while promoting innovation and creativity. In February 2024, Microsoft's Vice Chair and President Brad Smith published a blog post[43] acknowledging that powerful AI tools will lead to exciting opportunities for creative expression, but also become weapons for those with bad intentions. In the blog, he called for Microsoft and others to act with urgency to combat abusive AI-generated content, and laid out six focus areas as part of a robust and comprehensive approach to addressing this critical issue.

While the recommendations in this whitepaper are focused specifically on one of those areas—modernized policy and legislation to protect people from the abuse of technology—Microsoft recognizes that solving this problem will take a whole-of-society approach. As a technology company and AI leader, we have a special responsibility to lead here, but also to continue to collaborate with others. While not an exhaustive list, this section of the paper lays out some examples of how Microsoft has been approaching synthetic content risks across each of the six focus areas.

Figure 5: Microsoft's approach to combat abusive AI-generated content.



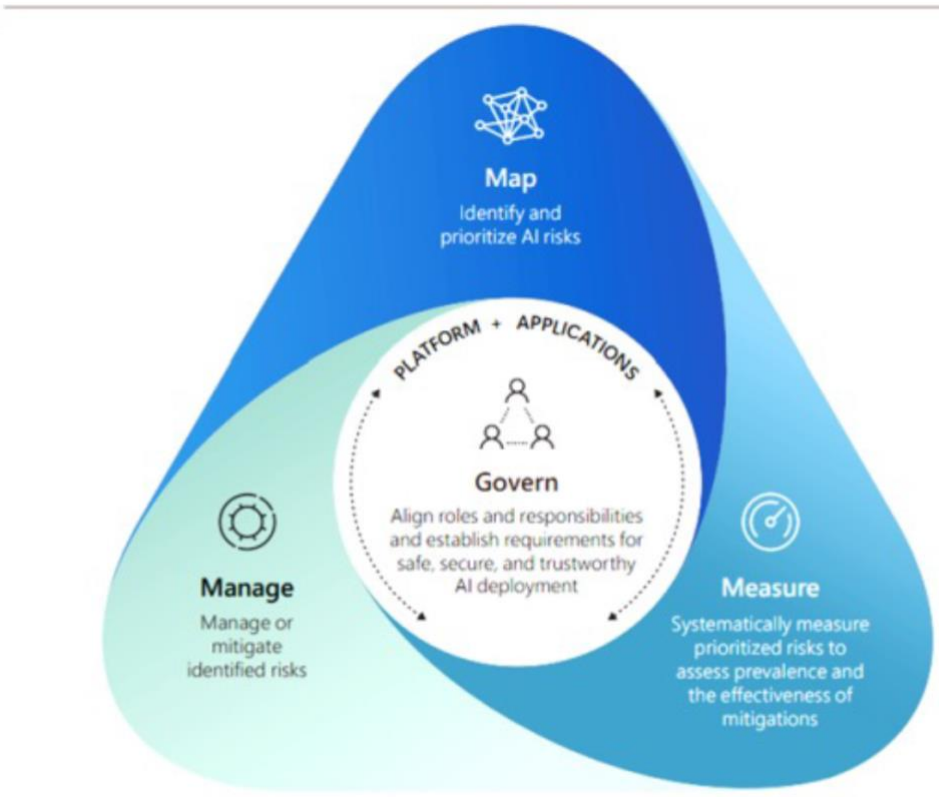
A strong safety architecture needs to be applied at the AI platform, model, and applications levels.

Strong safety architectures should include aspects like ongoing red team analysis, pre-emptive classifiers, the blocking of abusive prompts, automated testing, and rapid bans of users who abuse the system. At Microsoft, we understand that this is a multi-faceted and iterative process. Part of our safety architecture is prepared responses to offensive, inappropriate or otherwise harmful prompts.

As part of our commitment to build responsibly and help our customers do so as well, we integrate content filtering within the Azure OpenAI Service. We regularly assess and update our content filtering systems to ensure that they're detecting as much relevant content as possible, as well as expanding our detection and filtering capabilities over the last year.

We also understand that the work of AI risk management cannot be done by companies alone, as civil society and outside stakeholders provide important perspectives to consider when evaluating our products, which is why we regularly partner with them for additional feedback. For example, to better understand the risk of misleading images, Microsoft partnered with NewsGuard[44], an organization of trained journalists, to evaluate Microsoft Designer. Further information on this partnership is available in our 2024 Responsible AI Transparency Report[45], which details the steps we take to map and measure risks, and then manage or mitigate the identified risks at the platform or application levels. We also make publicly available our Responsible AI Standard[46], so that stakeholders can better understand our risk management process.

Figure 6: An iterative cycle**:** Govern, map, measure, manage.

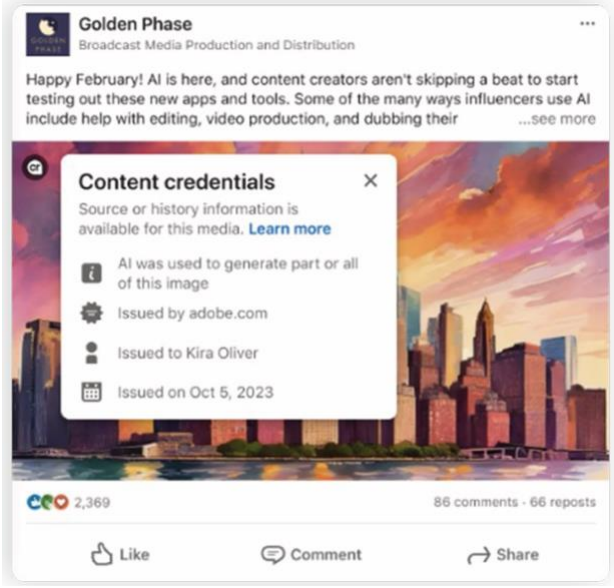


Durable media provenance and watermarking are essential to build trust in the information ecosystem.

As more creators use generative AI technologies to assist in their work, the line between synthetic content created with AI tools and human-created content will increasingly blur. While considerable progress has been made to develop and deploy disclosure methods for generative AI media, several challenges persist. This includes stripping or removal of the disclosure method, and attempts to add fake disclosure signals.
With the aim of advancing disclosure methods to help consumers understand whether digital content was created or edited with AI, in 2021 Microsoft co-founded the Coalition for Content Provenance and Authenticity (C2PA)[47] alongside Adobe, Arm, BBC, Intel, and Truepic.

C2PA is a standards-setting body with a mission to develop an end-to-end open standard and technical specification on content provenance and authentication. Because of this commitment, in 2023, we were able to announce media provenance capabilities that use cryptographic methods to mark and sign content, including that generated by AI, with metadata about its source and history. We are also actively exploring watermarking and fingerprinting techniques that help to reinforce provenance techniques. We are committed to ongoing innovation that will help users quickly determine if an image or video is AI-generated or manipulated.

LinkedIn has also implemented C2PA so that content carrying the technology is automatically labelled on the platform. Starting with content on the LinkedIn Feed, users can click on an icon in the upper left corner, which then reveals source history information, including whether the material was generated in whole or in part by AI:



LinkedIn is currently working to expand the coverage of feature to other surfaces in addition to its LinkedIn Feed, including ads. This feature provides a verifiable trail of where the content

originates from and whether it was edited, creating a more transparent and secure environment for LinkedIn members.

Beyond Microsoft, we continue to advocate for increased industry adoption of the C2PA standard. There are now more than 180 industry members of C2PA, including Google, BBC, Intel, Sony, and AWS. While the industry is moving to rally around the C2PA standard, we are mindful that relying on one approach alone will be insufficient. This is why we are also continuing to test and evaluate combinations of techniques, in addition to new methods, to find effective provenance solutions for all media formats.

Safeguarding our services from abusive content and conduct, whether real or synthetic, is critical to reduce the potential for harm.

At Microsoft, we have long recognized our responsibility to keep our users safe, especially young people, and to contribute to building a safer online ecosystem. To achieve that, we take steps to protect our users from illegal and harmful online content, while respecting critical human rights such as privacy, freedom of expression, and access to information. Across Microsoft's consumer services, the Code of Conduct in the Microsoft Services Agreement[48] governs what content and conduct is permitted, and we plan to take steps to enforce our policies[49] against abusive content, including AI-generated content that violates those policies.

LinkedIn also has a robust trust and safety structure and policy framework[50] which prohibit all forms of false and misleading content, scams, fraud, and other forms of abuse, as well as fake profiles. LinkedIn combines human reviewers and investigators with automated solutions for a safe, trustworthy, and professional experience.

GitHub has also updated its policies to prohibit the sharing of software tools that are designed for, encourage, promote, support, or suggest in any way the use of synthetic or manipulated media for the creation of non-consensual intimate imagery, or any content that would constitute misinformation or disinformation.

In addressing abusive AI-generated content, we are building on existing frameworks, policies, and partnerships that support our ongoing efforts to safeguard our services. In perhaps the best-known example, in 2009, Microsoft collaborated with Dartmouth College to develop PhotoDNA, which was a landmark step forward in our collective ability to detect and address CSAM across the online ecosystem. PhotoDNA[51] is a robust hash-matching technology that enables the detection of previously identified harmful content and has been widely adopted, supporting tech companies to address harm at scale.

Recognizing the evolving nature of online harms, we have also donated[52] an updated version of PhotoDNA to StopNCII.org, a service that enables people to protect themselves from having their intimate images shared online without their consent by reporting to StopNCII.org and hashing their content without it leaving their device. We have recently announced[53] that we are partnering with StopNCII to pilot efforts to detect and remove this victim-reported imagery from Bing's image search index: a step we believe will make a significant difference to reduce the availability of NCII across the ecosystem, in addition to addressing reports we receive directly from victims.

In addition to our work in these spaces, Microsoft is also innovating to address widespread problems, such as spam calls, that are increasing with the rise of advanced technology. In order to address this growing problem, Microsoft has developed Azure Operator Call Protection[54] for our customers, which is a fraud detection service for voice network operators that performs real-time analysis of consumer phone calls to detect potential phone scams and alert subscribers when they are at risk of being scammed.

Robust collaboration across industry and with governments and civil society is critical to create a safer digital ecosystem.

Addressing complex online harms requires a whole-of-society approach, unable to be addressed by any one sector. We have a range of longstanding digital safety collaborations through which we receive vital multistakeholder feedback and can advance shared goals, including through the Global Internet Forum to Counter Terrorism, WeProtect Global Alliance, Internet Watch Foundation, Tech Coalition, and beyond.

These collaborations constantly evolve to keep up with developments in AI. For example, the Tech Coalition, which is dedicated to facilitating cross-industry cooperation to address CSEA risks, has been leading cross-industry collaboration on best practices to address a range of generative AI issues and briefing stakeholders on the issue. For example, Microsoft was proud to host a Tech Coalition briefing on generative AI and CSAM in Brussels in November 2024.

Continuing collaboration to address harms associated with generative AI is vital to Microsoft's commitment to responsible AI. For example, at the Munich Security Conference in February 2024, 20 companies, including Microsoft and LinkedIn, announced a new Tech Accord to Combat Deceptive Use of AI in 2024 Elections[55], with a straightforward but critical goal to combat video, audio, and images that impersonate or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders. The cross-tech sector agreement contains several essential commitments, which fall into three main areas of action: (1) protecting content authenticity (for example through watermarking), (2) detecting and responding to deceptive deepfakes and (3) promoting public awareness and resilience.

In addition, since 2023, Microsoft has been partnering with Women Political Leaders, to help address the specific challenges that women in public life are facing. This partnership has explored the challenges that women in politics face online, as well as positive, constructive uses of AI, including for political campaigning.

Public awareness and education are necessary to ensure a well-informed public that can discern the difference between legitimate and deceptive content.

As part of Microsoft's commitments in the Tech Accord, we have developed training materials and public campaigns to drive awareness of the issue of deepfakes in elections and increase

understanding of the tools available to protect against deceptive AI-generated content. Ahead of the EU Parliamentary Elections we organized a number of briefings in Brussels, and across the 27 Member States, with political staffers and candidates to provide them with information on the risks of deepfakes, as well as with solutions to protect themselves and react effectively. In an effort to raise awareness and increase voter education, we supported the European Parliament's *Use Your Vote*[56] campaign, which encouraged citizens to go to the polls. We were proud to support this campaign on Bing Search by including search results tailored to the 24 official EU languages with links to official European Parliament sources, as well as banners dedicated to the EU elections, making sure that EU citizens had the right information on when, where, and how to vote.

In May 2024, Microsoft and OpenAI also announced the launch of a $2 million Societal Resilience Fund[57] to further AI education and literacy among voters and vulnerable communities. Grants from the fund will help several organizations, including Older Adults Technology Services from AARP (OATS), the C2PA, International Institute for Democracy and Electoral Assistance (International IDEA), and Partnership on AI (PAI) to deliver AI education and to support their work in creating a better understanding of AI capabilities.

We will continue to work together to share learnings from our experience, implementing a the framework that supports the evolution of the fund over time as part of a community of practice. We recognize that there is more work to do, and we look forward to playing an important role in it.

Finally, we also recognize the importance of education for young people to help build critical media literacy and digital citizenship skills, including the safe and responsible use of AI. We have made available a range of AI resources[58] for educators, as well as guidance for parents in our Family Safety Toolkit[59].

> ***I think the school has a big part to play [in teaching] AI literacy because we spend most of our time there, so it would be very nice to have activities on this topic"***
>
> *Young person from Romania, Youth Focus Group convened by menABLE*

To meet young people where they are, we have also released "The Investigators"[60], a Minecraft Education media literacy game that teaches young people some of the most critical digital skills— the ability to find, consume, and share authoritative information. Similarly, we recently launched a new Minecraft Education feature called AI Foundations[61] which is designed to help students, educators, and families understand and use AI tools responsibly.

Part III: Microsoft's policy recommendations to combat abusive AI-generated content risks

At the time of writing, the EU is embarking on a new political mandate. We acknowledge and recognize that the 2020-2024 mandate represented a significant regulatory leap forward, which saw several major steps taken to protect consumers key to incentivizing technology companies in deploying safe and trustworthy AI.

The EU's Digital Services Act (DSA), adopted in 2022, marked a significant shift in regulatory engagement on digital safety. Its "systemic" approach to risk assessment and mitigation raises the bar for online intermediaries. The framework also aims to foster dialogue between service providers, regulatory bodies, and civil society to promote continual improvements that will ultimately better protect Europeans online. The EU Artificial Intelligence Act (AI Act) also represents a key regulatory milestone. It prohibits the use of AI systems for certain practices that can lead to harm, such as manipulating people's behaviour through subliminal techniques and exploiting vulnerabilities, for example those related to older age. It classifies certain AI systems as high-risk, such as systems used for influencing elections or voting behaviour, as well as systems used for biometric or emotion recognition, requiring providers to abide by stringent safeguards. Finally, it places transparency requirements on generative AI systems, as well as risk assessment, mitigation, and transparency measures on providers of general-purpose AI models.

While the DSA and AI Act are by no means the only relevant pieces of EU legislation that will govern AI systems and their possible effects on end users, and much of their impact on the ground will be determined by how these regulatory acts are implemented and enforced by the European Commission and Member States.

However, AI-generated content risks are also whole-of-society risks, requiring holistic action. So, it is alongside these critical frameworks for online service providers that we offer the following recommendations for the EU and its Member States to consider, aiming at establishing additional guardrails to protect the public from abusive AI-generated content. Our recommendations highlight where gaps may still exist, and suggest concrete steps to protect and safeguard children, women and older adults. We also highlight additional policy recommendations to horizontally address abusive AI-generated content risks.

Protecting Children from Online Exploitation

Young people are both the most excited about and most at risk from AI and its potential for abuse. Microsoft's 2024 Global Online Safety Survey found that young adults and children see the use of AI as exciting and as a practical tool for translation purposes, work and school, but also expressed concern about at least one potential risk, including deepfakes. In a 2024 youth focus group Microsoft held with young Europeans, facilitated by menABLE - project co-funded by the European Union - participants were deeply excited about the applicability of AI in their everyday lives.

> **"I've seen peers become overly reliant on AI for social interactions"**
>
> *Young person from Portugal, Youth Focus Group convened by menABLE*

This focus group brought together youth representatives from five EU Member States. During the focus group, we addressed the youth representatives' relationship with AI, as well as their concerns. Their input and perspectives are reflected in the present paper and have guided our policy recommendations in this space. Amongst other worries, the focus group session elicited concerns from young respondents over their peers becoming increasingly reliant on AI in social settings.

In terms of risk, the 2024 survey results highlight that 66% of respondents experienced an online risk in the past year, with the number jumping to 72% for teens aged 13-17. Teen girls were chiefly worried about cyberbullying, sexual solicitation, and exploitation while teen boys showed more concern around graphic and violent content exposure.

> ***"I use AI a lot for coding. I try to program some stuff and it's very, very useful. Very impressive actually."***
>
> *Young person from France, Youth Focus Group convened by menABLE,*

> ***" I write something in English and then give it to AI to check what I did wrong and then I learn from my mistakes."***
>
> *Young person from Croatia, Youth Focus Group convened by menABLE*

One of the most likely risks arising from the widespread availability of generative AI is the development of highly realistic "deepfake" images of real individuals, including children. Multiple studies have shown that the vast majority of deepfakes are nude, sexual or pornographic. Unfortunately, this includes child abuse material – and, as the IWF has found, AI has been used to create new images based on existing material.

As outlined earlier in this paper, CSAM (whether synthetic or genuine), is not only inherently

harmful but also may be used to facilitate other harms, such as financially motivated extortion, grooming, or trafficking. Additionally, large volumes of synthetic content may hinder efforts to address real-world harm, fuel demands from bad actors for new content, and reduce social and technical barriers to sexualizing minors. Exposure to CSAM may also lead to an increased risk that offenders seek contact with children offline. We know that accessing CSAM is often the first step towards hands-on abuse, a factor that may be aggravated by realistic AI depictions.

However, we must not lose sight of the harms that arise from the abuse and exploitation of real children—our goals must be to minimize harm and ensure law enforcement can take steps to rescue children in danger. Our recommendations below are therefore intended to address known challenges in tackling CSAM and to mitigate additional risks that may arise because of AI.

The DSA already provides a comprehensive framework to address systemic risks to EU internet users, including child sexual abuse and exploitation. Microsoft stands ready to continue to assist the Commission and the Digital Services Coordinators in ensuring its proportionate implementation. While the DSA will play a key role in combatting a range of harms and their different manifestations, including tackling CSAM risks (whether synthetic or otherwise), there are opportunities to clarify Member State laws on the AI-generated manifestations of this crime which are intended to help appropriately deter and respond to criminal activity.

Reflecting the abhorrent nature of the abuse, CSAM has long been criminalised across all Member States. It is also singular among online harms, in that the content is regarded as inherently harmful, regardless of context. As new technologies have emerged, predators and bad actors have consistently evolved their tactics and found new ways to misuse technology to exploit children—generative AI, unfortunately, is no exception. While a majority of EU Member States have updated their penal codes to ensure that this harm is criminalized even when content is AI-generated or digitally altered, the legal basis is not fully harmonized in the EU.

Recommendation 1: Modernize legislation through the adoption of the Recast of Directive 2011/93/EU 'on combating the sexual abuse and exploitation of children and child sexual abuse material'

While the recent Penal Code amendments in many Member States represent a strong step forward in the appropriate criminalization, we welcome enhanced legal certainty and harmonization of a legal definition of CSAM that includes "realistic images, reproductions, or representations".
The Recast's clear amendments to Article 2(3)(d) with regards to artificial representation and reproductions would be an advance on the status quo and create a strong baseline across all EU Member States.
The swift adoption of the Recast, as well as the consistent ratification across EU Member States, will be a key contributor to the success of all relevant stakeholders, including industry, government, civil society, and law enforcement.

Recommendation 2: Start discussions to extend the ePrivacy interim derogation (April 2026) and continue to develop a long-term framework to address CSAM

The 2022 proposal for a Regulation laying down rules to prevent and combat child sexual abuse is a welcome effort, serving to raise the bar for addressing the online manifestations of this harm. We stand ready to assist policy makers - especially the Council, which has faced years of deadlock - to help ensure a meaningful compromise is struck before the end of the ePrivacy derogation extension (set for April 2026).

More specifically, we encourage the Council to ensure that their compromise features an explicit legal basis that would allow Interpersonal Communication Service (ICS) providers to continue proactively detecting CSAM on their services. As of now, neither the Commission's Proposal, nor the Parliament's text have proposed appropriate derogations from the ePrivacy Directive, which could inadvertently restrict companies' ability to use the technology to detect and address this harm at scale.

We recommend that EU lawmakers enable providers to continue their proactive efforts to detect and disrupt online child sexual exploitation– given the size and scope of the problem, European law must not unduly restrict the use of vital detection technologies. We also recommend EU lawmakers continue to consider how to future-proof this legislation: as AI technology evolves, so too will our ability to detect and reduce sexual exploitation and grooming risks to children. Detecting previously identified CSAM (whether real or synthetic content) is critical, but the long-term legal framework must also address the potential for high volumes of novel synthetic content and enable innovation using cutting-edge AI technologies.

Lastly, in the event that negotiations continue in Council, we recommend that co-legislators engage in discussions over a possibility to extend the ePrivacy derogation once more, ensuring that Interpersonal Communications Services retain a clear legal basis for voluntary activities to detect CSAM. While we recognize that an additional extension is by no means a long-term legal solution, the alternative could mean a significant step backwards in our collective efforts to tackle this harm.

Recommendation 3: Establish an expert taskforce to study AI-enabled child sexual exploitation

Microsoft recommends that the EU establish a dedicated expert taskforce to study the means and methods of AI used to exploit children, and to propose comprehensive solutions to deter and address such exploitation.

This taskforce should build upon the valuable work already undertaken by organisations such as the Tech Coalition and IWF, in addition to incorporating expertise from both the public and private sectors. It should include representatives from law enforcement, child protection agencies, technology companies, academic institutions, and relevant government departments.

The taskforce's mandate could include conducting in-depth research on current and potential future manifestations of AI-enabled

child exploitation; evaluating the effectiveness of existing legal and technological measures in preventing synthetic CSAM; proposing privacy-preserving solutions that leverage AI for detection and prevention of child exploitation; and recommending policy and legislative updates to emerging challenges.

In the United States, Microsoft has supported a similar initiative proposed by 54 Attorneys General requesting that Congress establish an expert commission to study the means and methods of AI used to exploit children, and to propose solutions to deter and address such exploitation.

Recommendation 4: Meaningfully adopt the Annual Youth Policy Dialogue mentioned in the mission letter of Executive Vice-President (EVP) for Tech Sovereignty, Security and Democracy, Henna Virkkunen

Citizen participation is key to develop proportionate, meaningful, and fit-for-purpose legislation. As President von der Leyen highlighted in her mission letter to EVP Virkunnen, embedding citizen participation into decision-making has the potential to instil a lasting culture of participative democracy in the EU.

This rings particularly true for children, especially for those born 'digital natives', who stand to benefit the most from innovative technologies such as generative AI. They are, however, also some of the most at risk groups for online harms. Ensuring their participation will be central to earning their trust and developing rules that empower them online, as well as meet their needs and expectations.

Microsoft stands ready to assist the Commission in establishing these links. We also commit to ensuring that our work, especially on the development of safe AI, reflects youth perspectives.

> ***"I think every country could have a safer internet center that [could teach] media literacy and I think they need to come into schools or do some non-formal workshops for everyone."***
>
> *Young Person from Croatia, Youth Focus Group convened by menABLE,*

Safeguarding Women from Non-Consensual Intimate Imagery

Even before the onset of AI, technology has been misused to facilitate gender-based violence. One in ten[62] women in the European Union have experienced cyber harassment since the age of 15, and 51% of girls online have reportedly[63] experienced some form of gender-based online violence at a personal level, and of those, 85% said they have experienced multiple forms of harassment. In 2021, the United Nations declared[64] the rising trend of violence against women and girls as a 'shadow pandemic'.

Studies have consistently shown that not only are the vast majority of AI-generated deepfakes nude, sexual, or pornographic in nature, but that these artificial images are often created without the consent or knowledge of the person depicted. Additionally, data[65] also points to the fact that such non-consensual deepfakes impact women and girls at vastly higher rates than men and boys.

NCII, including synthetic NCII and deepfake pornography, is used to shame, harass, and even extort the person depicted. The harm can be emotional, economic, and reputational, and in some instances, victims also find their physical safety at risk.

The EU has spearheaded efforts to address this harm, including through the development of a thorough legal framework. As early as 2011, the Istanbul Convention, ratified in recent years by 22 European Member States, aimed to address this societal harm, including 'image-based sexual abuse'. The DSA also identifies several systemic risks that online services must assess and mitigate, including in relation to gender-based violence. Lastly, in March 2022, the European Commission proposed the landmark proposal for a Directive on Combating Violence against Women and Domestic Violence. The Directive, now adopted, established minimum rules on the definition of relevant criminal offences and penalties, access to justice, and victim support. Most relevant to this paper, the Directive also clarified the criminal nature of 'cyber' manifestations of violence against women, including the misuse of AI technologies.
The swift and thorough ratification and adoption of this Directive by all EU Member States will be a central part of addressing abusive AI-generated content risks to women and girls.

Recommendation 1: Drive wide adoption of the Directive on Violence against Women and Domestic Violence and ensure strong implementation mechanisms

The Directive is a strong baseline principle to ensure there is a harmonized and consistent understanding across Member States' penal codes, especially with respect to the dissemination of AI-generated NCII. The text clearly calls for the criminalization of the non-consensual distribution of images or videos, where: the conduct is intentional; the content is made accessible to the public; the content showcases sexually explicit activities or the intimate parts of a person; and where such conduct is likely to cause serious harm to that person.

Most notably, the Directive's Article 5 explicitly calls on Member States to amend their penal codes and formally criminalize non-consensual sexually explicit activities, including the creation/production of deepfakes, as well as the manipulation or alteration of images without the consent of the person depicted.

The explicit reference to synthetic media will be a valuable tool in addressing the lacking legal gaps for Member States that have not adopted the Istanbul Convention and setting up a clear and harmonized legal understanding around the criminality of such activities.

We encourage all Member States to adopt the Directive in their penal codes swiftly and thoroughly by the June 2027 deadline. We also encourage the European Commission to set out clear rules to ensure the transposition process remains on track and that strong implementation mechanisms are put in place. Doing so ensures

that the Directive is followed, particularly with regards to criminalizing non-consensual AI-generated intimate imagery.

Beyond the scope of the Directive, there are opportunities for enhanced research to further our collective understanding of the scale and impact of AI-driven harassment of women, especially public figures. Such research would serve to help develop and deliver targeted policy, and other interventions across diverse manifestations of the harm, across EU Member States.

Recommendation 2: Expand, where possible, the scope of the Directive to explicitly include deepfake 'nudes'

As is written currently, the Directive's main references to AI-generated non-consensual intimate imagery sharing could be interpreted more narrowly than was intended. This is because Article 5 only mandates that Member States criminalize the sharing of such content – whether synthetic or not – if it depicts a "person engaged in a sexually explicit activity". Such a definition could inadvertently create a loophole and fail to criminalize the creation of nude imagery without the consent of the person depicted.

While the Directive makes wider references to "intimate parts" in early recitals, we recommend that in ratifications, Member States consider include the word "nude" or similar as part of the criminal offense.

This is because not all nudity is considered sexually explicit, and may thus rule out the deep psychological harm that can result from content created in 'nudification' or similar apps without consent then being disseminated. The provision's applicability to a person "engaging" in sexual activities may also cause confusion over the use of such nudification apps and hamper a victim's ability to seek justice. Taking this approach may also have a deterrent effect.
As a result, we encourage all Member States ratifying this law to apply the same broad definition of NCII sharing (whether AI-altered or not) and ensure that "intimate" imagery includes nude as well as sexual activities. Therefore, we recommend Member States to introduce more precise language ensuring both sexual and non-sexualized non-consensual AI-generated content is covered in their ratifications, alongside practical examples for clarity.

Recommendation 3: Prioritize legal clarity and implementation of the EU's legal framework, including the DSA, so existing tools are appropriately leveraged to tackle NCII

The DSA outlines certain obligations for online intermediaries to address systemic risks related to illegal content, which include elements of gender-based online violence, such as non-consensual intimate imagery sharing.

For example, the DSA requires Very Large Online Platforms and Very Large Online Search Engines to assess the systemic risks that could stem from the design or functioning of their service – including any actual or foreseeable negative effects relating to gender-based violence. The Commission has also since issued a more extensive list of categories of illegal content pertaining to 'cyber violence' and 'cyber violence against women'. The extensive detail of these categories recognizes the multifaceted ways in

which this harm may manifest online. We welcome this clarification and look forward to the continuation of the DSA's framework to tackle this multifaceted harm.

Through industry and civil society engagement, the Commission can facilitate a greater understanding of potential steps to tackle online gender-based violence, as well as how to understand this as a systemic risk across diverse Very Large Online Platforms and Search Engines.

We stand ready to continue to support Commissioner Lahbib's and President von der Leyen's efforts to end gender-based violence, a central priority to the European Commission's gender equality strategy for 2020–2025, which was renewed in September 2024 by President von der Leyen, in her mission letter[66] for the Commissioner-designate for Equality. We particularly welcome the intention to continue to uphold the EU's commitment to gender equality in the form of a new, post-2025, gender equality strategy.

Safeguarding Older Adults, especially against AI-enabled fraud

As generative AI technologies evolve, so too do their capacity for fraudulent misuse. Synthetic content provides cybercriminals with the capability to enhance and scale existing fraud schemes, while enabling new forms of deception. For example, generative AI can be used to generate and maintain dialogues with potential victims, while image creation capabilities can be misused to create convincing fake identity documents to facilitate identify fraud.

Financial fraud scams have been growing exponentially in recent years, even before the widespread adoption of AI, overwhelming police and prosecutors. Online and telephone scams are particularly prevalent, with older adults often targeted due to their perceived vulnerability and accumulated wealth. Older adults are statistically more likely to face challenges with rapidly evolving technology, making them prime targets for AI-enabled fraud. In a 2025 focus group - convened by AGE Platform Europe and held by Microsoft - with European older adults, participants highlighted both their optimism in generative AI's potential for older adults' social inclusion, while pointing to their concerns about being able to adapt.

> ***" AI for us is like trying to jump on an express train going full speed. It is not easy"***
>
> *Older Adult from Finland, Older Adults Focus Group convened by AGE Platform Europe*

They may struggle to recognize phishing emails, deepfake videos, or other AI-generated scams, due to a lack of familiarity with the techniques being used. Additionally, in some instances cognitive decline, social isolation, and loneliness can put older adults at higher risk of being victims of AI-generated fraud or scams.

Part of the challenge in addressing this risk is that fraud is happening outside of traditional payment eco-systems, at an unprecedented scale, and is moving to other types of online services, such as online platforms, which

inherently have fewer in-person interactions. The fast changes in the eco-system of this harm may cause additional and unexpected challenges for older adults.

Additionally, it is worth recognizing that the concept of "AI-enabled fraud" has no specific and uniformly agreed definition in the EU. There are many definitions of fraud under national laws, and international initiatives such as the Council of Europe Convention on Cybercrime (the Budapest Convention) which introduce the notion of "computer-related fraud", but they are not linked to AI. Other legislative initiatives in the EU that tackle AI content reference fraud in passing and recognize the role that AI can serve to further enable crime, but no horizontal concept exists.

EU privacy laws also set strict requirements on how personal data can be collected, used, or otherwise processed. However, this also means that both the General Data Protection Regulation (GDPR) and the ePrivacy Directive limit the action that online service providers can undertake to proactively detect, block, or otherwise combat AI-enabled fraud. Most detection tooling necessary to combat and prevent fraudulent activities would risk infringing EU privacy law.

Recommendation 1: Recognize that detecting AI-generated fraud is a legitimate purpose under existing and new EU data protection legislation

Preventing any kind of fraud requires providers to be able to leverage trends, patterns, and insights based on a sufficiently representative sample of signals, many of which fall under the scope of either or both of the GDPR and the ePrivacy Directive. Efficient prevention of AI fraud firstly requires that these legal instruments are implemented in a way that is forward-looking and recognizes combatting AI fraud as a laudable goal. A goal which may in turn require additional collection of data for the purposes of detecting, removing and reporting such illegal activity as fraud. More consistent recognition and interpretation of such a goal in existing European privacy frameworks could serve to enable more thorough detection. In the event that the ePrivacy framework is reviewed, in the 2025-2030 Commission mandate, we recommend introducing targeted changes that recognize fraud prevention as a proper legal basis for data processing.

There are two legal instruments that govern companies' ability, or lack thereof, to deploy detection technology for the purposes of detecting, removing and reporting fraudulent activities, including those generated by AI. Firstly, the EU General Data Protection Regulation (GDPR), adopted in 2016 (2016/679), is Europe's comprehensive law governing data protection and privacy. It mandates strict rules on how personal data can be collected, processed, and stored. In Recital 47, fraud prevention is explicitly recognized as a legitimate interest of data controllers, by which processing of personal data is considered lawful. However, we see in practice that overly conservative interpretations by Data Protection Authorities (DPAs) can hamper the adoption of effective fraud detection and prevention measures that would serve to protect groups such as older adults. Therefore, we call on the relevant EU stakeholders, including the European Data Protection Board (EDPB) and the national DPAs, to adopt a forward-looking approaches and understandings while implementing the GDPR and the ePrivacy instruments. Particularly, they should consider fraud detection and prevention to constitute a legitimate interest that is inherently linked with

the provision of the service. The upcoming EDPB guidance on legitimate interests[67] provides a first opportunity to do so.

Secondly, the EU ePrivacy Directive (2005/58), serves to compliment the GDPR which governs data protection specifically in electronic communications. It regulates cookies usage, email marketing, data minimization, and data confidentiality. Many of its provisions , especially on cookies and traffic data processing, are outdated and do not give enough flexibility to properly address AI-enabled fraud. While some Member States, such as Finland, allow for data processing under the ePrivacy Directive for the purposes of fraud detection and prevention, this option is not consistently applied across all Member States. Therefore, we call on EVP Virkunnen and Commissioner McGrath to prioritize tools to help online services better protect consumers against fraud, in accordance with data protection law. As the case may be, this should also include a modernization of the ePrivacy Directive, to include clear legal bases – subject to additional safeguards and duties – for fraud detection and prevention of all data subjects, including older adults, against AI-powered scams. In such a review of the ePrivacy framework, inspiration can be drawn from the conditions and safeguards included in the temporary derogation to the ePrivacy Directive for CSAM detection.

Recommendation 2: Ensure better intergenerational solidarity in the new EU Mandate

Expected policies from the new EU mandate related to demographic change, intergenerational fairness or equality, as reflected in the Mission Letters[68] of the relevant Commissioners covering these portfolios, could benefit from a more explicit mention of older adults, especially in regards to their use of AI. As was laid out in a recent letter[69] from the Europe Seniors' Association (ESU), as well as by AGE Platform Europe, to President von der Leyen, young people are allocated more political prioritization across the Commissioner portfolios.

As Europe's ageing population continues to grow, so does the need to pay attention to the challenges of older adults in today's digital economy. Responses to the Microsoft focus group held with older adults from Europe, also revealed their enthusiasm for being included in policy debates.

> ***" Lifelong learning is key to integrating older adults into the policy discussion"***
>
> *Older Adult from France, Older Adults Focus Group convened by AGE Platform Europe*

Building on the achievements from the previous mandate concerning this demographic, we recommend the Commission to continue to look into action plans and mechanisms to ensure ageing continues to be mainstreamed across policy areas, not least in the digital space, and in particular looking into the protection of older adults against abusive AI-generated fraud online. The upcoming Intergenerational Fairness Strategy, led by Commissioner Micallef and expected by early 2026, presents an opportunity to ensure an age equality strategy[70].

The creation of a European Parliament Intergroup or Interest Group in this area may help facilitate these discussions and contribute to bringing about meaningful change. We therefore support the creation of a European Parliament Intergroup:

for Europe of all ages, as put forward by AGE Platform Europe, to address, amongst other things, the opportunities and challenges of the digital economy for older adults. In a similar vein, the new Commission may wish to engage in similar consistent dialogue with older adults, as is planned for young people. Further embedding citizen participation into decision-making has the potential to instill a lasting culture of participative democracy in the EU, and older people should not inadvertently be excluded from such participatory initiatives.

As pointed out by AGE Platform Europe[71], there is a research gap on how older adults are affected by new technologies, AI especially. Stakeholders across the eco-system, whether it be government, civil society, or industry, may wish to seek, where possible, further data collection and research, especially for people above the age of 74 years old, who are particularly unrepresented.

Cross-harm recommendations

As we have highlighted in prior sections of this paper, malicious actors are using AI to create abusive content, for harmful purposes against women, children, and older adults. Moreover, abusive-AI-generated content also creates risks to information integrity through raising concerns about what information people can trust online.

Similarly, it is becoming increasingly easy for malicious actors to claim authentic content, such as images of atrocities, are "fake" or AI-created, the so-called Liar's Dividend (a term coined by legal scholars Bobby Chesney and Danielle Citron). We must therefore leverage provenance tools both to help people understand when content comes from a trusted source, by labelling AI-generated content, and build information and AI literacy.

The EU's AI Act, the world's first horizontal legislation governing AI, already introduces transparency obligations, including in Article 50, that serve to guide responsible approaches to transparency, including with respect to interaction with AI systems and identification of synthetic content created or edited by AI. Specifically, the Act requires providers of AI systems to ensure that they inform users that they are interacting with an AI system. In addition, providers and deployers of AI systems that generate synthetic content will need to ensure outputs are detectable and disclosed as such using robust and reliable technical solutions.

An important step in building complementary, state-of-the-art provenance standards that can help users differentiate authentic content from its AI-generated or AI-edited counterpart, is going to be the development of a code of practice. As foreseen in Article 50 of the AI Act, a code of practice facilitates effective implementation of such detection and labelling requirements. This process provides an opportunity to establish clear standards for AI content provenance, ensuring that as these technologies evolve, there are robust mechanisms in place to maintain transparency and trust in digital information. It will be important to ensure this process is started in a timely manner and draws upon international developments, best practices and standards, including work done by cross-industry groups like the Coalition for Content Provenance and Authenticity, as well as the AI Safety Institutes in likeminded countries and work by the National Institute of Standards and Technology (NIST). Such an approach will help develop techniques and guidance to support information integrity on a global scale.

Recommendation 1: Policy makers should examine prohibiting the stripping, tampering with or removal of provenance metadata

The AI Act's transparency requirements are expected to further drive adoption of state-of-the-art provenance tooling, such as C2PA, so people can understand whether a piece of content is AI-generated or manipulated.
Alongside this provider-focused requirement, we recommend that, in order to reinforce the value of synthetic content labelling, policy makers across the EU consider updating penal codes in a harmonized way to prohibit the intentional and deceptive stripping, tampering with or removal of provenance metadata from AI-generated or edited content indicating if content is authentic or synthetic.

This is particularly important for large content distribution platforms, given the important role they play in sharing and facilitating access to online content.
Distribution platforms, such as social media companies, must also play their part in advancing a robust authenticity ecosystem. These platforms are often where AI-generated or edited content is most widely spread. A requirement for system providers to attach provenance information to content is ineffective if that information is then stripped by the platforms through which that content is shared.

Just as it is against the law today to tamper with or remove the identification number on physical assets, like cars, policy makers should prohibit intentionally deceptive tampering with, stripping or removal of provenance metadata indicating if content is authentic or synthetic.
At the same time, to protect privacy, such legislation should support the ability of people and organisations to redact personal information from provenance information and simply retain authentication of the digital source type (i.e., the source from which the media was created)—which is ultimately the most essential piece of information indicating whether a media file was authentically captured, compared to being AI-generated or manipulated.
Legislation should also protect the identity of whistleblowers and journalists, as well as enabling researchers to test the rigor of these systems.

We support legislation to establish penalties for bad actors working to intentionally remove, strip or tamper with authenticity or provenance metadata of AI content. This would be a common-sense measure to protect responsible AI efforts and hold bad actors accountable.
It will also be important to implement stronger controls for the subset of generative AI content that will pose the highest degree of risk. While carrying provenance information will be an important baseline mitigation for all synthetic content, more controls are appropriate for advanced deepfake capabilities on the horizon that pose a heightened risk of deceptive impersonation (i.e., for fraud.)

Recommendation 2: Support and enhance public education campaigns on AI and synthetic content as laid out in the 2023 Council Recommendations on digital education and training

European governments are uniquely positioned to deliver tailored education campaigns to the public around online safety and abusive AI-generated content harms, much as they do on

other critical issues. Departments of education can, and should, use existing funding programs at their disposal to help the public build digital and media literacy skills.

The European Commission also has several initiatives aiming to harmonize efforts to upskill the European population by leveraging Member States, companies, social partners, and education providers. Programs such as Horizon Europe are central to fostering AI innovation in Europe, but can only come to fruition with widespread digital literacy initiatives. According to a 2023 report on the EU's 2030 Digital Decade[72] project, 32% of Europeans still lack basic digital skills, a trend particularly present in older age brackets.

New education campaigns should be developed to help the public, especially at-risk groups, to understand potential deceptive uses of synthetic content, the associated safety risks and harms, and approaches for discerning authentic digital content.

Such literacy campaigns would serve to teach Europeans not only how to use and benefit from AI, but also how to assess whether content was authentically captured or AI-generated, identifying trusted sources, and recognizing the latest scams employing synthetic content. The campaigns should specifically aim to target vulnerable demographics, such as older adults and young people. There is merit in such campaigns also being developed at local level, by all members of society able to both create teaching material and widely disseminate it across our diverse societies, making sure to reach vulnerable populations, and to cater content to their age, region, language, existing digital literacy, and experience.

> ***" AI education should start in schools and continue throughout lifelong learning initiatives. Education should be diversified, not just from technology or government, but all society members"***
>
> *Older Adult from France, Older Adults Focus Group convened by AGE Platform Europe*

While we recognize and support the importance of the AI literacy requirement for providers of AI systems as laid down in Article 4 of the AI Act, as well as the vast number of upskilling initiatives already put in place by the EU and locally in Member States, we support the creation of a cross-EU AI Literacy Campaign aimed at the general public. This could leverage the EU government's digital skills frameworks and stimulate investment in both formal educational structures and informal learning opportunities to advance AI literacy across the EU, as per the objectives laid out in the EU's Digital Education Action Plan (2021-2027)[73].

> ***" I am concerned that once people leave their working environment, they are gradually more out of touch with what is happening in the sphere."***
>
> *Older Adult from the UK, Older Adults Focus Group convened by AGE Platform Europe*

We recommend that any education campaign incorporates input from civil society groups and is disseminated in coordination with trusted local community organizations. Beyond achieving

broad public awareness, these campaigns should specifically target frontline actors such as local media, journalists, community leaders, and civil liberties groups who will need to assess potential deepfakes and educate others.

In areas of civic importance, such as election integrity, we recommend that the European Commission, in collaboration with EU Member States' electoral commissions and media regulators, develop targeted education campaigns. These could include information about content provenance tools and how to distinguish official election-related content from potentially misleading synthetic media.

Additionally, we recommend continued efforts to support online safety and media literacy education for both young people and older adults. For young people, these skills are crucial for navigating complex online information ecosystems and using AI technology safely and responsibly. For older adults, improved digital literacy can enhance their social engagement, financial security, and overall participation in an increasingly digital society.

Cross-sectoral collaboration on public awareness and education campaigns will remain a key element in combatting this shared challenge. Industry and civil society have a key role to play in providing resources and training to help women in leadership learn about AI risks, and how best to recognize and respond to AI threats.

Lastly, we note the need for varied stakeholder participation for the upcoming European Democracy Shield initiative. EVP Virkkunen and Commissioner McGrath, who will lead the work designed to address the most severe risks to democracy must ensure that meaningful stakeholder discussions are held to maximize its effectiveness.

The Democracy Shield, which will seek to counter foreign information manipulation and interference online by enhancing media literacy, and enforcing the Digital Services Act and the AI Act, is uniquely positioned to serve as a vessel to drive the objectives we have mentioned in this paper.

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. Microsoft makes no warranties, express or implied, in this document.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in, or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property. aka.ms/protectthepublic

References

[1] Political Guidelines 2024-2029 | European Commission

[2] Global Online Safety Survey Results | Microsoft

[3] First ever EU rules on combating violence against women: deal reached

[4] Deepfakes, explained | MIT Sloan

[5] Fake News in an Era of Social Media: Tracking Viral Contagion | Yasmin Ibrahim

[6] Deep Fakes Are Merely Today's Photoshopped Scientific Images | Forbes

[7] How can we combat the worrying rise in deepfake content? | World Economic Forum

[8] The State of Deepfake | Deeptrace

[9] The Global Risks Report 2024 | World Economic Forum

[10] Geographical Hosting: URLs | IWF 2023 Annual Report

[11] How AI is being abused to create child sexual abuse material (CSAM) online | International Watch Foundation

[12] Generative AI CSAM is CSAM | National Center for Missing and Exploited Children

[13] WeProtect Global Alliance and Thorn release report on new and emerging technologies impacting online child safety | WeProtect Global Alliance

[14] Malicious Uses and Abuses of Artificial Intelligence | Europol

[15] Global Threat Assessment 2023: Assessing the scale and scope of child sexual abuse online | WeProtect Global Alliance

[16] Annual Report 2023 | INHOPE

[17] Annual Report 2023 | INHOPE

[18] Scammers impersonate Spain's Princess Leonor on TikTok to deceive victims worldwide | EL PAÍS English

[19] French woman duped by AI Brad Pitt faces mockery online | BBC News

[20] The Battle Against AI-driven Identity Fraud | Signicat

[21] AI-Driven Fraud Soars Across Europe Financial Sector, Driven by Deepfakes and Identity Theft | Fintech Switzerland

[22] Risks associated with artificial intelligence for the elderly | CEDMO

[23] 2030 Digital Decade | Publications Office of the EU

[24] New Horizons in artificial intelligence in the healthcare of older people | Oxford Academic

[25] AI can help us live longer and better lives - Edinburgh Impact | The University of Edinburgh

[26] Could tailored AI robots help alleviate the loneliness epidemic? | Euronews

[27] A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning | CNN Politics

[28] Slovakia's Election Deepfakes Show AI Is a Danger to Democracy | WIRED

[29] POLITICO Poll of Polls — Slovakian polls, trends and election news for Slovakia | POLITICO

[30] Turkey's deepfake-influenced election spells trouble | Fortune Europe

[31] AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections | The Alan Turing Institute

[32] https://foreignpolicy.com/2024/12/06/romania-presidential-election-annulled-russia-interference-georgescu/ Why Romania's election was annulled – and what happens next? | THE CONVERSATION

[33] Confirmation hearings for the European Commission | European Parliament

[34] Deepfakes, distrust and disinformation: Welcome to the AI election | POLITICO

[35] Deepfake porn is political violence | POLITICO

[36] The real women whose faces have been used by AI to create deepfake X-rated images | Mirror Online

[37] Italy's Giorgia Meloni called to testify in deepfake porn case | POLITICO

[38] Giorgia Meloni al processo per i video porno con la sua faccia: "Intollerabile violenza" | TODAY

[39] Towards an EU criminal law on violence against women: The ambitions and limitations of the Commission's proposal to criminalise image-based sexual abuse | New Journal of European Criminal Law

[40] Gardaí looking into allegations that large number of images of women were shared online without their consent | The Journal

[41] NEW PUBLICATION: Report on Cyber Violence Against Women | European Women's Lobby

[42] Confirmation hearings for the European Commission | European Parliament

[43] Combating abusive AI-generated content: a comprehensive approach | Microsoft On the Issues

[44] Transparent Reliability Ratings for News and Information Sources - NewsGuard

[45] Microsoft Responsible AI Transparency Report | Microsoft CSR

[46] Microsoft Responsible AI Standard v2 General Requirements | Microsoft

[47] Overview - The Coalition for Content Provenance and Authenticity (C2PA)

[48] Microsoft Services Agreement | Microsoft

[49] Digital Safety Policies | Microsoft

[50] Políticas para la comunidad profesional de Linkedin | Linkedin

[51] PhotoDNA | Microsoft

[52] StopNCII Announces PhotoDNA Integration and Niantic as New Industry Partner | SWGfL

[53] An update on our approach to tackling intimate image abuse | Microsoft On the Issues

[54] Solutions pour l'industrie des télécommunications | Microsoft Industry

[55] A Tech Accord to Combat Deceptive Use of AI in 2024 Elections | AI Elections Accord

[56] (video) Use your vote or others will decide for you | European Parliament

[57] Microsoft and OpenAI launch Societal Resilience Fund - Microsoft On the Issues

[58] Smart learning: AI resources every educator should know | Microsoft Education Blog

[59] Microsoft Family Safety Toolkit | Microsoft

[60] InvestiGators | Minecraft Education

[61] AI Foundations | Minecraft Education

[62] Facts and figures: Ending violence against women | UN Women – Europe and Central Asia

[63] State of the World's Girls 2020: Free to Be Online? | Plan International

[64] The Shadow Pandemic: Violence against women during COVID-19 | UN Women – Headquarters

[65] Nonconsensual pornography among U.S. adults: A sexual scripts framework on victimization, perpetration, and health correlates for women and men. | Psychology of Violence

[66] Mission Letter | Commissioner-designate for Equality

[67] Guidelines 1/2024 on processing of personal data | European Data Protection Board

[68] Commissioners-designate (2024-2029) | European Commission

[69] European Commission priorities for 2024-2029: The perspective of the European Senoirs' Union | ESU

[70] AGE Manifesto for the European Parliament Elections 2024 | AGE Platform Europe

[71] Digitalisation and older people our call to EU Policy Makers | AGE Platform Europe

[72] Europe's digital decade: 2030 targets | European Commission

[73] Digital Education Action Plan (2021-2027) | European Commission