



# Protecting the Public from Abusive AI-Generated Content

# Table of Contents

Foreword	3
Part I: Diagnosing the problem of abusive AI-generated content	8
Part II: Microsoft’s approach to combating abusive AI-generated content	18
Part III: Microsoft’s policy recommendations to combat abusive AI-generated content risks	28
Protect content authenticity	29
Detect and respond to abusive deepfakes	36
Promote public awareness and education	48

# Foreword



**Brad Smith**  
Vice Chair and President,  
Microsoft

*"The greatest risk is not that the world will do too much to solve these problems. It's that the world will do too little. And it's not that governments will move too fast. It's that they will be too slow."*

Those sentences conclude the book I coauthored in 2019 titled "Tools and Weapons." As the title suggests, the book explores how technological innovation can serve as both a tool for societal advancement and a powerful weapon. In today's rapidly evolving digital landscape, the rise of artificial intelligence (AI) presents both unprecedented opportunities and significant challenges. AI is transforming small businesses, education, and scientific research; it's helping doctors and medical researchers diagnose and discover cures for diseases; and it's supercharging the ability of creators to express new ideas. However, this same technology is also producing a surge in abusive AI-generated content, or as we will discuss in this paper, abusive "synthetic" content.

**Five years later, we find ourselves at a moment in history when anyone with access to the Internet can use AI tools to create a highly realistic piece of synthetic media that can be used to deceive:** a voice clone of a family member, a deepfake image of a political candidate, or even a doctored government document. AI has made manipulating media significantly easier—quicker, more accessible, and requiring little skill. As swiftly as AI technology has become a tool, it has become a weapon. As this document goes to print, the U.S. government recently announced that it successfully disrupted a nation-state sponsored AI-enhanced disinformation operation. FBI Director Christopher Wray said in his statement, "Russia intended to use this bot farm to disseminate AI-generated foreign disinformation, scaling their work with the assistance of AI to undermine our partners in Ukraine and influence geopolitical narratives favorable to the Russian government." While we should commend U.S. law enforcement for working cooperatively and successfully with a technology platform to conduct this operation, we must also recognize that this type of work is just getting started.

**The purpose of this white paper is to encourage faster action against abusive AI-generated content by policymakers, civil society leaders, and the technology industry.** As we navigate this complex terrain, it is imperative that the public and private sectors come together to address this issue head-on. Government plays a crucial role in establishing regulatory frameworks and policies that promote responsible AI development and usage. Around the world, governments are taking steps to advance online safety and address illegal and harmful content.

The private sector has a responsibility to innovate and implement safeguards that prevent the misuse of AI. Technology companies must prioritize ethical considerations in their AI research and development processes. By investing in advanced analysis, disclosure, and mitigation techniques, the private sector can play a pivotal role in curbing the creation and spread of harmful AI-generated content, thereby maintaining trust in the information ecosystem.

Civil society plays an important role in ensuring that both government regulation and voluntary industry action uphold fundamental human rights, including freedom of expression and privacy. By fostering transparency and accountability, we can build public trust and confidence in AI technologies.

The following pages do three specific things:

- 1. Illustrate and analyze the harms arising from abusive AI-generated content**
- 2. Explain Microsoft's approach**
- 3. Offer policy recommendations to begin combating these problems**

Ultimately, addressing the challenges arising from abusive AI-generated content requires a united front. By leveraging the strengths and expertise of the public, private, and NGO sectors, we can create a safer and more trustworthy digital environment for all. Together, we can unleash the power of AI for good, while safeguarding against its potential dangers.

## Microsoft's responsibility to combat abusive AI-generated content

Earlier this year, we outlined a [comprehensive approach](#) to combat abusive AI-generated content and protect people and communities, based on six focus areas:

1. **A strong safety architecture**
2. **Durable media provenance and watermarking**
3. **Safeguarding our services from abusive content and conduct**
4. **Robust collaboration across industry and with governments and civil society**
5. **Modernized legislation to protect people from the abuse of technology**
6. **Public awareness and education**

Core to all six of these is our responsibility to help address the abusive use of technology. We believe it is imperative that the tech sector continue to take proactive steps to address the harms we are seeing across services and platforms.

We've taken concrete steps, including:

- **Implementing a safety architecture** that includes red team analysis, preemptive classifiers, blocking of abusive prompts, automated testing, and rapid bans of users who abuse the system.
- **Automatically attaching provenance metadata** to images generated with OpenAI's DALL-E 3 model in Azure OpenAI Service, Microsoft Designer, and Microsoft Paint.
- **Developing standards for content provenance and authentication** through the Coalition for Content Provenance and Authenticity (C2PA) and implementing the C2PA standard so that content carrying the technology is automatically labeled on LinkedIn.
- **Taking continued steps to protect users from online harm**, including by joining the Tech Coalition's Lantern program and expanding PhotoDNA's availability.
- **Launching new detection tools like Azure Operator Call Protection** for our customers to detect potential phone scams using AI.
- **Executing our commitments to the new Tech Accord** to combat deceptive use of AI in elections.

## Protecting Americans through new legislative and policy measures

This February, Microsoft and LinkedIn joined dozens of other tech companies to launch the [Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#) at the Munich Security Conference. The Accord calls for action across three key pillars that we utilized to inspire the additional work found in this white paper: addressing deepfake creation, detecting and responding to deepfakes, and promoting transparency and resilience.

In addition to combating AI deepfakes in our elections, it is important for lawmakers and policymakers to take steps to expand our collective abilities to (1) promote content authenticity, (2) detect and respond to abusive deepfakes, and (3) give the public the tools to learn about synthetic AI harms. We have identified new policy recommendations for policymakers in the United States. As one thinks about these complex ideas, we should also remember to think about this work in straightforward terms. These recommendations aim to:

- **Protect our elections**
- **Protect seniors and consumers from online fraud**
- **Protect women and children from online exploitation**

Along those lines, it is worth mentioning three ideas that may have an outsized impact in the fight against deceptive and abusive AI-generated content.

- **First, Congress should enact a new federal “deepfake fraud statute.”** We need to give law enforcement officials, including state attorneys general, a standalone legal framework to prosecute AI-generated fraud and scams as they proliferate in speed and complexity.
- **Second, Congress should require AI system providers to use state-of-the-art provenance tooling to label synthetic content.** This is essential to build trust in the information ecosystem and will help the public better understand whether content is AI-generated or manipulated.
- **Third, we should ensure that our federal and state laws on child sexual exploitation and abuse and non-consensual intimate imagery are updated to include AI-generated content.** Penalties for the creation and distribution of CSAM and NCII (whether synthetic or not) are common sense and sorely needed if we are to mitigate the scourge of bad actors using AI tools for sexual exploitation, especially when the victims are often women and children.

These are not necessarily new ideas. The good news is that some of these ideas, in one form or another, are already starting to take root in Congress and state legislatures. We highlight specific pieces of legislation that map to our recommendations in this paper, and we encourage their prompt consideration by our state and federal elected officials.

Microsoft offers these recommendations to contribute to the much-needed dialogue on AI synthetic media harms. Enacting any of these proposals will fundamentally require a whole-of-society approach. While it's imperative that the technology industry has a seat at the table, it must do so with humility and a bias towards action. Microsoft welcomes additional ideas from stakeholders across the digital ecosystem to address synthetic content harms. Ultimately, the danger is not that we will move too fast, but that we will move too slowly or not at all.



**Brad Smith**  
Vice Chair and President,  
Microsoft









# 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 so many ways, AI will 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 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](#), the same year that a fake lip-sync video of former President Obama was released. Since that video came out, deepfake images, videos and audio, all of varying degrees of sophistication, have flooded our discourse. Yet, media manipulation is not new. It dates back to well before the digital age.

Timeline of deepfake examples making headlines (not exhaustive)

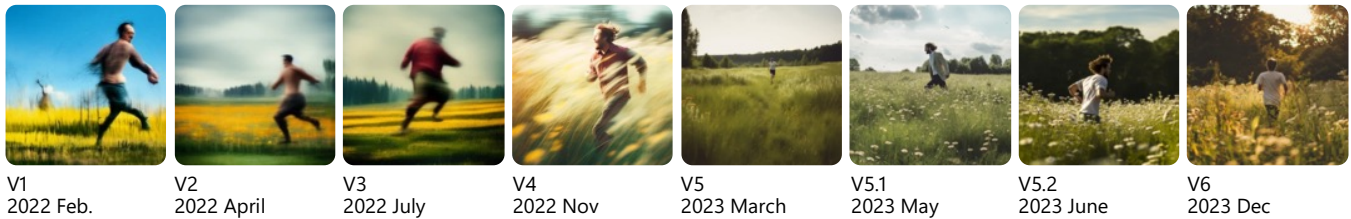
 <p>2017 July</p> <p>Lip-syncing Obama: New tools turn audio clips into realistic video</p> <p>Source: <a href="#">UW News</a></p>	 <p>2019 August</p> <p>Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case</p> <p>Source: <a href="#">WSJ</a></p>	 <p>2021 August</p> <p>How a deepfake Tom Cruise on TikTok turned into a very real AI company</p> <p>Source: <a href="#">CNN</a></p>	 <p>2023 June</p> <p>DeSantis campaign shares apparent AI-generated fake images of Trump and Fauci</p> <p>Source: <a href="#">NPR</a></p>	 <p>2023 Sept.</p> <p>Naked deepfake images of teenage girls shock Spanish town: But is it an AI crime?</p> <p>Source: <a href="#">Euronews</a></p>	 <p>2024 May</p> <p>consultant faces charges and fines for Biden deepfake robocalls</p> <p>Source: <a href="#">NPR</a></p>
---	---	---	--	--	---



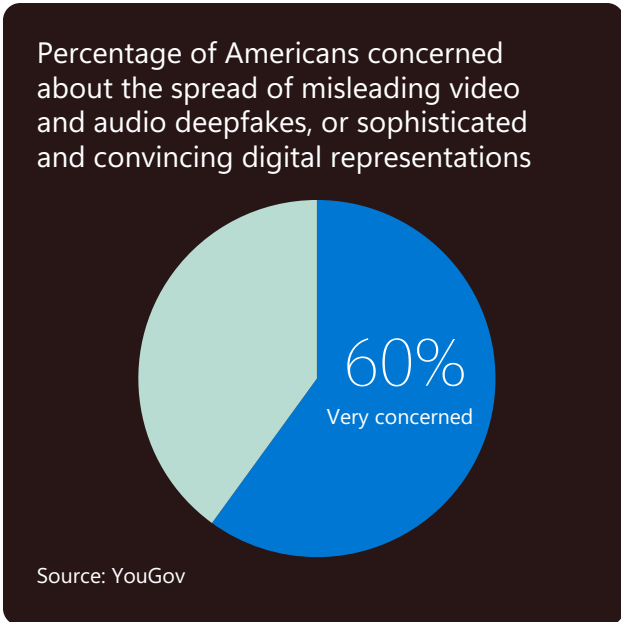
[In the 18th century, photographers and artists manipulated photos to create deceptive content.](#) Totalitarian rulers such as Stalin and Hitler notoriously used such techniques to alter photographs for propaganda purposes. The introduction of photo editing software in the 1990s led to a [surge in doctored images.](#)

While this manipulation is not new, the development of generative AI technology has [increased the risk](#) of abusive content. With more advanced technology, we now have AI-generated content that is difficult to distinguish from real images, videos or audio.

Timeline of Midjourney versions (Prompt: a man running in the meadow photography)

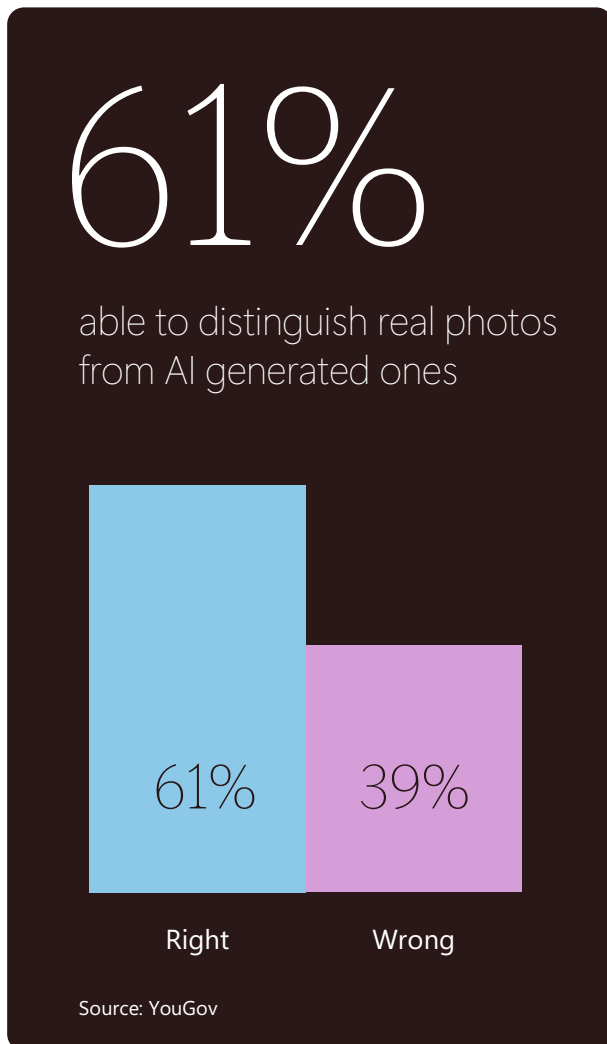


And the technology has become easier to access, learn, and use, making the creation of a realistic deepfake more convenient for cybercriminals and for other bad actors. And, as we have seen over time, technology has also facilitated the broad distribution and weaponization of this harmful content. It is no surprise that in [a study](#) from 2023, 60% of Americans said they were very concerned about the spread of misleading video and audio deepfakes, or sophisticated and convincing digital representations. And this concern increases with age, with senior citizens representing the most concerned demographic.



Coupled with this concern about abusive AI-generated content is difficulty in identifying it as fake. In [a recent study](#) funded by the National Science Foundation, investigating the vulnerability of different groups to deepfake videos, results showed that the general adult population was only 46% likely to correctly identify a deepfake video as inauthentic.

This rate was lower than middle school students (58%), and substantially lower than Carnegie Mellon University (CMU) students (80%), all of whom fared better in their identifications. The study authors noted that CMU students were the only population more likely to correctly identify deepfake videos than the authentic videos across all groups, likely because of their experience and expertise in AI and machine learning. Another [research paper](#) confirms similar results for AI-generated images, finding that participants on average were able to correctly distinguish only 61.3% of the images.



(a) Are these real photos?



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. But we must also recognize that the same tools can be used as weapons against the public.

In the following examples, we identify four types of harms that illustrate the need for a robust public policy response from technology companies and policymakers: (1) AI-generated fraud; (2) synthetic child sexual abuse material; (3) AI-generated election content; and (4) non-consensual intimate imagery.

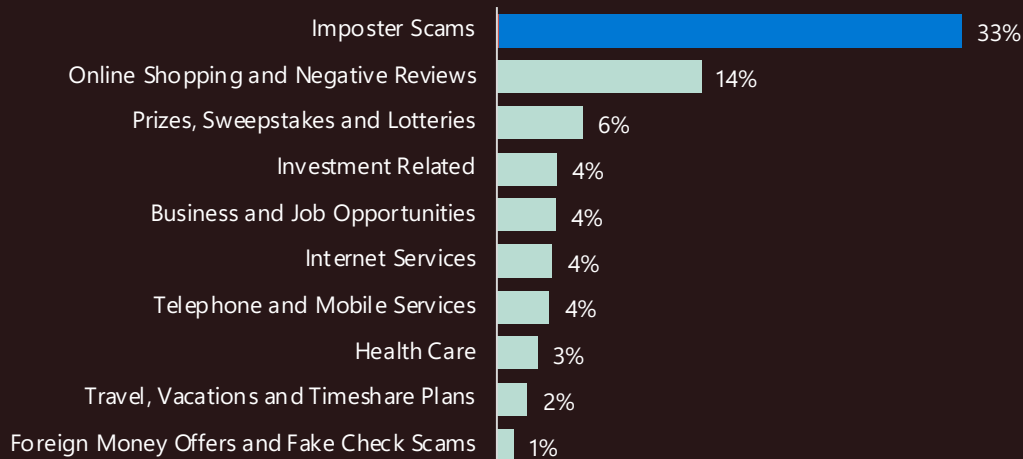
## Deepfake fraud in Hong Kong

In January 2024, an employee working at the Hong Kong branch of a multinational company received a message from someone claiming to be the company's UK-based chief financial officer. The employee then had a video call with this "CFO" and other employees, [all of whom turned out to be deepfake recreations of his colleagues](#), based on publicly available video. Unfortunately, the employee did not realize the deception at the time, followed their criminal instructions, and transferred millions of dollars to various bank accounts.

As a result of the scam, the company lost \$25 million.

While the sophisticated nature of this incident and its details may not be the norm, imposter scams over email, text message and phone are much more common. The Federal Trade Commission (FTC) currently [ranks imposter scams](#) as the most reported type of fraud. The losses from these scams have been increasing since 2019. For 2023 only, the scams resulted in losses of \$2.7 billion. The median loss per scam was \$1,000.

### Top 10 fraud categories



Source: FTC

## Cybercrime experiences in adults age 50+



Source: AARP

However, the [median loss increases with age](#), with the 80-plus population suffering the largest losses. And as the Hong Kong example demonstrates, it is a global issue. [A 2023 survey](#) found that 37% of organizations globally have experienced some form of voice deepfake fraud attempt. This trend is particularly concerning since AI has the potential to enable more accurate and misleading imposter scams. In a recent [AARP survey](#), 60% of respondents were undecided about the impact of generative AI, and only 9% had reported using it. The AARP researcher noted that the hesitancy could be linked to concern about online scams since nearly 75% of older Americans report being targets of cybercrime, with 19% having been a victim, and 43% personally knowing a victim of cybercrime. The concern about how AI may impact financial scams has become so acute that the FTC has already issued a [consumer alert](#) for it.

## DOJ brings first synthetic child sexual abuse material charges

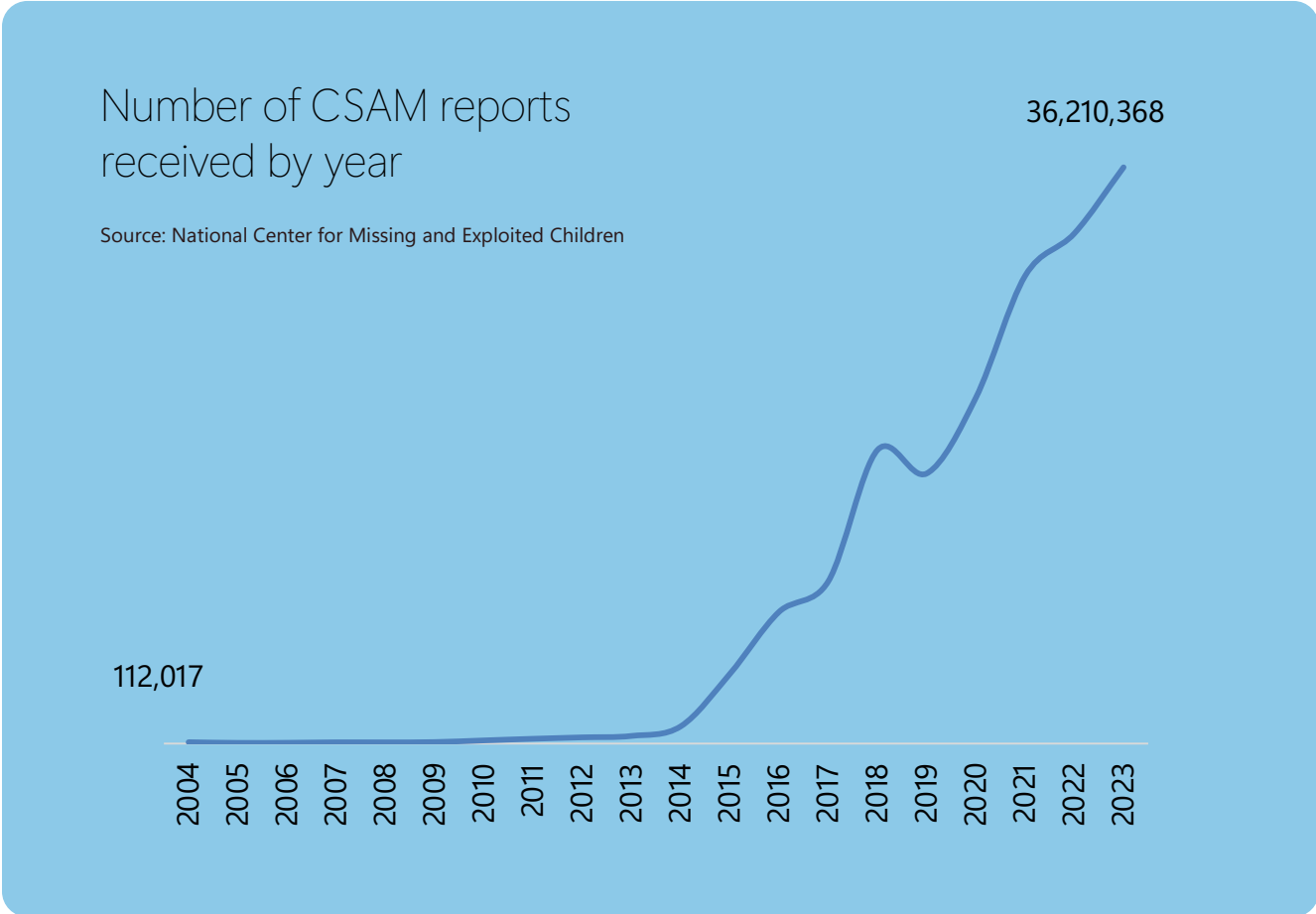
According to a [brief](#) filed by the U.S. Department of Justice (DOJ), Steven Anderegg, a 42-year-old man in Wisconsin, used an AI image generator to produce thousands of realistic nude or partially-nude images of prepubescent minors. [According to the DOJ](#), evidence recovered from Anderegg's devices revealed that he generated these images using specific, sexually explicit text prompts related to children. Additionally, Anderegg communicated with an underaged boy and described how he used the technology and then sent the child several synthetic images over a messaging platform. Law enforcement was alerted to Anderegg through a CyberTip from the National Center for Missing and Exploited Children (NCMEC) after the messaging platform reported Anderegg's account for distributing these images. Federal prosecutors have now charged Anderegg for creating synthetic child sexual abuse material, [the first federal case involving images produced entirely through AI](#).

While this case represents the first federal indictment for synthetic child sexual abuse material (CSAM), NCMEC is already seeing the impact of generative AI on reports into its CyberTipline. In 2023, it received [4,700 reports](#) related to synthetic CSAM.

This number is a fraction of the overall number of reports that NCMEC received in 2023 ([36 million reports](#)), but the [misuse of AI](#) has the potential to exponentially increase the production of this exploitative content and to accelerate this harm. For example, in 2004, the number of reports into the CyberTipline was around [112,000](#). Reporting numbers have grown year-on-year and risk accelerating still further as abusive AI-generated imagery spreads.

Synthetic CSAM cannot be disregarded because it creates real harm. Hundreds of thousands of reports of AI-generated CSAM could easily overload an already strained reporting ecosystem. This influx may delay the rescue of child victims or divert law enforcement resources from active investigations by creating uncertainties about which images depict real children.

Additionally, the Internet Watch Foundation has reported on perpetrators using AI to alter existing CSAM to generate new content, re-victimizing survivors. And recent [research](#) from Thorn and NCMEC highlights that generative AI may 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.



## An election deepfake in Slovakia has impact

In fall 2023, two days before Slovakia's elections, an [audio recording spread online](#) in which one of the top candidates, Michal Šimečka, boasted about rigging the election. Although he and the other party to the recording 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 stay quiet. And although some platforms removed or placed warnings on the post, [it did not stop the spread](#) of the recording which went viral quickly. 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](#) 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](#). In the U.S., [58% of Americans](#) believe that AI will increase the spread of misinformation in the 2024 presidential election.

Another cause of concern [raised in a study](#) is that as the increase in the number of deepfakes goes up, so does uncertainty among the population regarding authentic content. Indeed, 40% of respondents indicated a sense of skepticism or a sense of being misled or misinformed.

Yet, there is reason to be optimistic. [India recently concluded the largest election](#) in history in June 2024 with over 640 million votes tallied, and the campaigns extensively used AI. Political parties creatively used it to conduct outreach to voters, from making a video of Modi dancing to a Bollywood song to resurrecting Muthuvel Karunaanidhi, an iconic Indian actor-turned politician, who died in 2018, for an endorsement video.

Similarly, the European Parliament elections and the snap elections in both France and the United Kingdom in June and July of 2024, did not see a surge in deceptively realistic AI-generated content going viral and influencing voting behavior. Nevertheless, the consensus was clear. Despite fears similar to what Americans have expressed, AI was used in typical political ways—some negative campaigning— but often to better connect with voters.

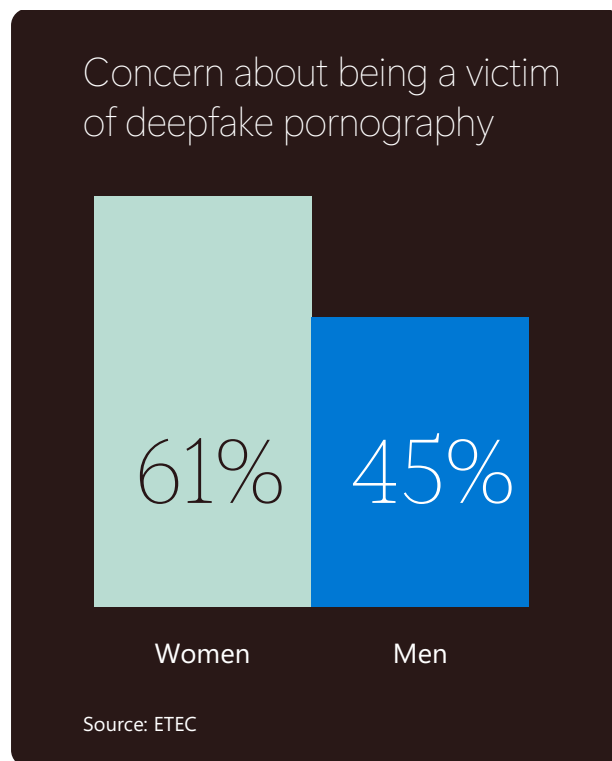


## Synthetic non-consensual intimate imagery is weaponized against women

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](#). 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](#), humiliating insinuations, and was even [sexually propositioned](#) 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 is not a new risk—but it is one that is vastly exacerbated by generative AI. In 2019, even before the advent of generative AI, [a report by Sensity AI](#) found that 96% of so-called “deepfakes” were pornographic, and of those, 99% were made of women. Such images have long been used to shame, harass, and extort the person depicted, affecting not only individuals with a public profile, but also private individuals, including teens.

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 distributed widely.



This harm is also deeply gendered, with women most often targeted, and facing consequences ranging from fear and pain to long-lasting reputational damage. In a March 2024 [ESET](#) survey into the prevalence of deepfake pornography, 61% of women from the United Kingdom reported concerns about being a victim of this harm, in comparison to less than half (45%) of men.

Microsoft's own consumer research, released for [Safer Internet Day](#) 2024, shows that teen girls are more likely to experience risks online (72% of teen girls, versus 68% of teen boys) and that 69% of respondents globally are worried about the potential use of AI for "deepfakes". This is also not a theoretical risk: [research from Graphika](#) suggests that in September 2023 alone, there were 24 million unique visitors to synthetic NCII websites. The same report found that the number of links advertising synthetic NCII services increased more than 2,400% on social media from 2022 to 2023, and many of the services [only work on women](#). In other words, this harm is on the rise, is deeply gendered, and the consequences are significant and long-lasting.

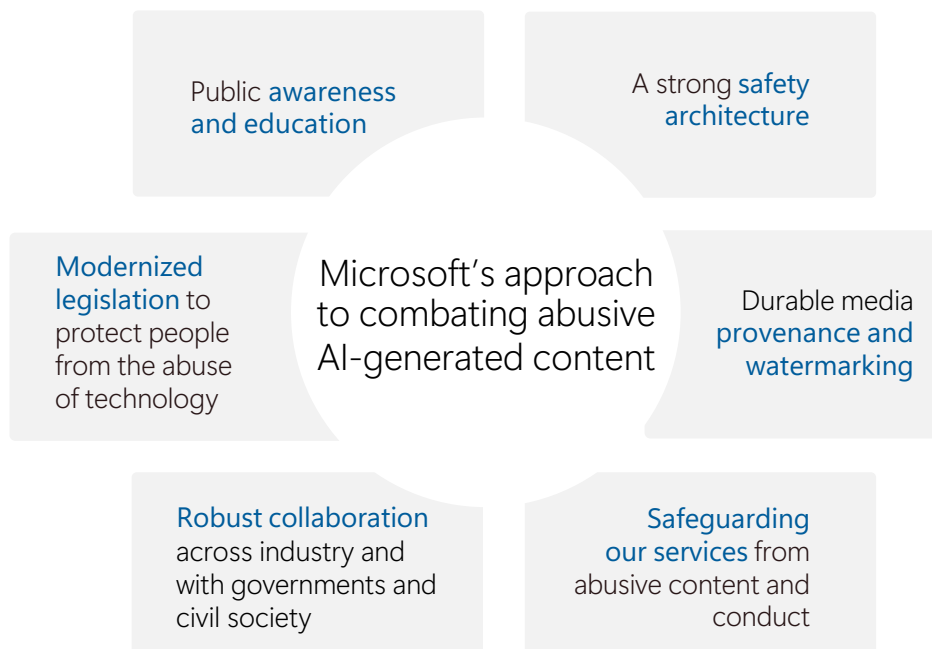
# Part II: Microsoft's approach to combating abusive AI-generated content

Throughout the United States, policymakers, academics, civil society, and others are grappling with how to address the challenges associated with abusive AI-generated content. Microsoft is committed to taking a responsible, balanced approach that protects the public from the harms of abusive AI-generated content while promoting innovation and creativity.

In February 2024, Microsoft's Vice Chair and President Brad Smith published a blog [post](#) 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 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, as part of that approach laid out in February, here are some examples of how Microsoft has been approaching synthetic content risks.



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

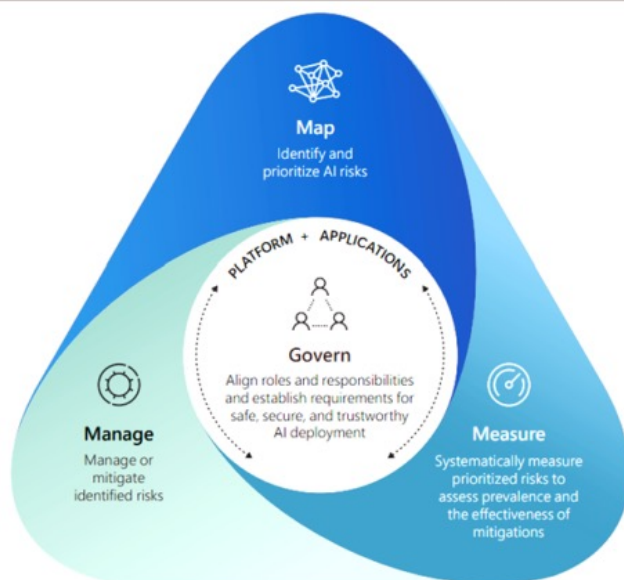
It should include aspects such as ongoing red team analysis, preemptive classifiers, the blocking of abusive prompts, automated testing, and rapid bans of users who abuse the system. At Microsoft, we understand that this is a multi-faceted process and that it is also iterative. Part of our safety architecture includes prepared responses to offensive, inappropriate or otherwise harmful prompts. We also display information sources as part of Copilot, to help people understand where the AI-generated content is coming from.

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 they're

detecting as much relevant content as possible and have expanded 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 and that 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](#), an organization of trained journalists, to evaluate Microsoft Designer. We have shared all this information recently in our [2024 Responsible AI Transparency Report](#), 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](#) so that stakeholders can better understand our risk management process.

Govern, map, measure, manage: An iterative cycle



## 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 still exist, including that no disclosure method is perfect and all will be subject to adversarial attacks. This includes stripping or removal of the disclosure method and attempts to add fake disclosure signals. More research and study, such as conducting technical assessments and understanding the impact and benefits of combining disclosure methods (e.g., provenance, watermarking, and/or fingerprinting) in the face of adversarial attacks, will be necessary to achieve durable provenance and watermarking.

With industry partners, Microsoft has led significant progress in 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 \(C2PA\)](#) 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 specifications 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.

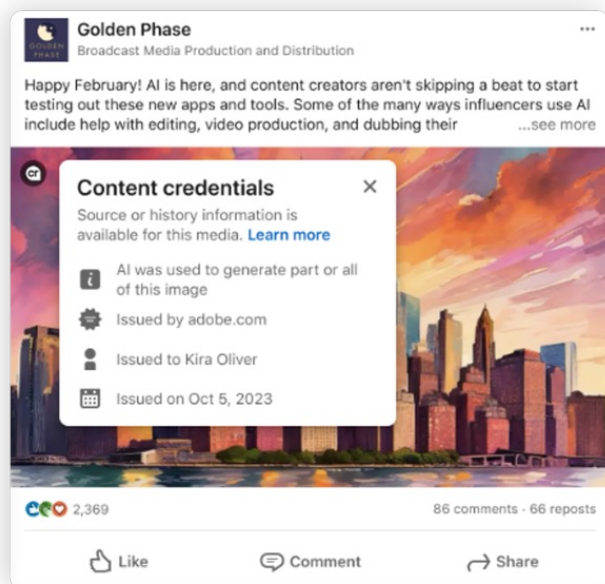
Since the end of 2023, we automatically attach provenance metadata to images generated with OpenAI's DALL-E 3 model in our Azure OpenAI Service, Microsoft Designer, and Microsoft Paint. This provenance metadata, referred to as Content Credentials, includes important information such as when the content was created, and which organization certified the credentials. 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, as well, implemented C2PA so that content carrying the technology is automatically labeled 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 coverage to other surfaces in addition to its LinkedIn Feed, including ads. Incorporating this feature provides for 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, Microsoft is mindful that relying on one approach alone will be insufficient. This is why Microsoft continues to play an important role on the C2PA Steering Committee, developing guidelines and helping to ensure collaboration among peers. We are also continuing to test and evaluate combinations of techniques in addition to new methods altogether to find effective provenance solutions for all media formats.





## Safeguarding our services from abusive content and conduct such as synthetic non-consensual intimate imagery, fraudulent AI-generated content, or AI-generated CSAM is also 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](#) governs what content and conduct is permitted, and we will take steps to enforce our [policies](#) against abusive content, including AI-generated content that violates those policies.

LinkedIn also has a robust trust and safety structure and [policy framework](#) prohibiting 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, trusted, and professional experience.

GitHub has also been considering how to evolve its [policies](#) to address abusive AI-generated content challenges, including by consulting on [proposed changes](#) to address potential tools for the creation of NCII and 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](#) is a robust hash-matching technology that enables the detection of previously identified harmful content, supporting tech companies to address harm at scale. Microsoft donated PhotoDNA to NCMEC, which has been able to make this technology widely available across the industry. We have also [recently donated](#) an updated version of PhotoDNA to [StopNCII](#), a service developed with support from Meta that enables people to protect themselves from having their intimate images shared online without their consent. Integrating PhotoDNA supports StopNCII's efforts by enabling people to report and hash content without it leaving their device and supporting a cross-industry approach to addressing non-consensual intimate imagery, including synthetic imagery that has been reported. Similarly, NCMEC's [Take It Down](#) initiative helps people under age 18 remove or stop the online sharing of their imagery.



Microsoft has continued to invest in improvements to PhotoDNA. In addition to the device-level hashing capability leveraged by StopNCII, we have also continued to update the algorithm to improve performance and reduce the cost of this process with no loss of accuracy. These enhancements will enable companies to continue to deploy PhotoDNA as a core technology in the detection and removal of identified CSAM at an increasing scale. This is an area where continued industry innovation and tool-sharing is critical: other examples include [Google's Content Safety API](#) and [CSAI Match](#) and Meta's [PDQ and TMK+PDQE](#), as well as Discord's [recent efforts leveraging AI](#).



Reflecting on our ongoing commitment to tackle this harm as it evolves, in April 2024, Microsoft joined other major AI companies in announcing our support for new [Safety by Design principles](#) to address risks related to online child sexual exploitation and abuse (CSEA) in AI models and services. Led by NGOs [Thorn](#) and [All Tech is Human](#), the principles comprise a set of high-level commitments to reduce CSEA-related risks in the development, deployment and maintenance of AI models and services. The principles will guide us as we continue to enhance our robust safety and responsible AI infrastructure and the safeguards on our services.

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](#) 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. Azure Operator Call Protection uses AI to analyze call content to determine whether a call is likely to be a scam. It listens for language patterns that are commonly used by fraudsters, such as asking for your credit card number, your Medicare information, or your Amazon account details. It can then recognize if the caller is using an AI-generated voice, which is illegal, and then it will alert the subscriber by text message. The service, which is an opt-in choice, does not automatically end the call for the subscriber, and it does not save or use the data from the call to train AI models.

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

Addressing complex online harms requires a whole-of-society approach and cannot be addressed by any one sector. We have a range of longstanding digital safety partnerships and 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](#), [The Christchurch Call](#), and beyond. We have also been at the table for critical conversations on NCII since roundtable discussions were convened in partnership with the Cyber Civil Rights Initiative in 2015.

These collaborations are already evolving to meet the AI moment. 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. Microsoft is proud to have been a founding member of this industry coalition. We welcome this ongoing partnership and engagement to ensure ongoing information-sharing with critical stakeholders, such as with NCMEC.

We also recognize that addressing the potential acceleration of harm in the AI era will require new collaborative measures. To that end, we are joining the Tech Coalition's flagship [Lantern program](#). Announced in November 2023, Lantern is the first cross-industry signal-sharing program that enables technology companies to more effectively collaborate and better enforce their child safety policies.

Continuing these collaborations to address harms associated with generative AI is vital to Microsoft's commitment to responsible AI. This most recently came together at the Munich Security Conference in February 2024 when 20 companies, including Microsoft and LinkedIn, announced a new [Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#), with a straightforward but critical goal to combat video, audio, and images that fake or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders. This cross-tech sector agreement contains several essential commitments, including (1) developing and implementing technology to mitigate risks related to deceptive AI election content; (2) assessing models in scope of the Accord to understand the risks they may present regarding deceptive AI election content; (3) seeking to detect the distribution of deceptive AI election content; (4) seeking to appropriately address deceptive AI election content detected;

(5) fostering cross-industry resilience to deceptive AI election content; (6) providing transparency to the public; (7) continuing to engage with a diverse set of global civil society organizations, academics, and other relevant subject matter experts; and (8) supporting efforts to foster public awareness and all-of-society resilience. Since the announcement, Microsoft has worked to implement the commitments in the Accord within our own company. We have released new tools for political campaigns that attach C2PA content credentials to positively assert authentic images, video, and audio. We have also created a reporting [portal](#) for deceptive AI election content and are continuing to roll out more services and announcements.

We have also implemented the European Commission's '[election guidelines](#)' as part of the European Union's Digital Services Act, which regulates online intermediaries and platforms to provide a safe and accountable online environment. In addition, we continued our efforts to tackle disinformation, including with respect to AI-generated content, in the context of our commitments under the European Union (EU) Code of Practice on Disinformation, and regularly [publish detailed reports](#) on these efforts, with our next report coming out in September, which will have a particular focus on the recent European Parliament elections.

We are also pursuing additional collaborations across the industry, with civil society and governments in other critical spaces. Microsoft's Digital Crimes Unit

(DCU), which works collaboratively to fight cybercrime, is co-leading a project as part of the [European Multidisciplinary Platform Against Criminal Threats \(EMPACT\)](#) with the US Secret Service (USSS) and the German Federal Criminal Police (BKA), funded by Europol, to evaluate and address the threat caused by cybercriminals' misuse of AI services, including synthetic media and fake content.

The main objectives of this project, which brings together representatives from international law enforcement and private sector companies, are to map out the threat landscape concerning criminal actors' use of AI services, based on the analysis of available data and intelligence, as well as the input from relevant organizations and experts from both the public and private sectors.

Microsoft's DCU has also partnered with the Department of Homeland Security and the Federal Bureau of Investigation (FBI) to form a working group to study the "Impact of AI on Criminal and Illicit Activities." The working group primarily focuses on generative AI, and specifically how those text, audio, and visual outputs can be used to facilitate criminal activities and what strategies and tools are available to mitigate criminal use of AI, including government-private sector collaboration. The project has three subsections: current state of AI technologies, the current and future AI-enabled threat landscape, and mitigation approaches for U.S. government and industry partners.

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

As part of Microsoft's commitments in the Tech Accord, we have been developing 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. For example, in advance of the European Parliament elections in June 2024, Microsoft organized briefings in Brussels and across the 27 EU Member States with political parties and candidates, providing them with information on the risks of deepfakes, and solutions to protect themselves and react effectively. In addition to the training, Microsoft also ran a broad public awareness campaign across the EU. This campaign drove voters to trusted sources of election information as well as media and information literacy resources to help combat any possible attempts to use deceptive AI to impact the election. The campaign garnered millions of impressions driving millions of voters to the EU's election resources.

In May 2024, [Microsoft and OpenAI announced the launch of a \\$2 million Societal Resilience Fund](#) 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 better understanding of AI capabilities.

For example, OATS and AARP plan to use the grant to develop and deploy training programs focused on educating older adults on the foundational aspects of AI, including in-person and virtual trainings and guides so that older adults can learn more about the opportunities of the technology, as well as the risks and potential for misuse. Together, we will promote whole-of-society resilience against the use of deceptive AI content.

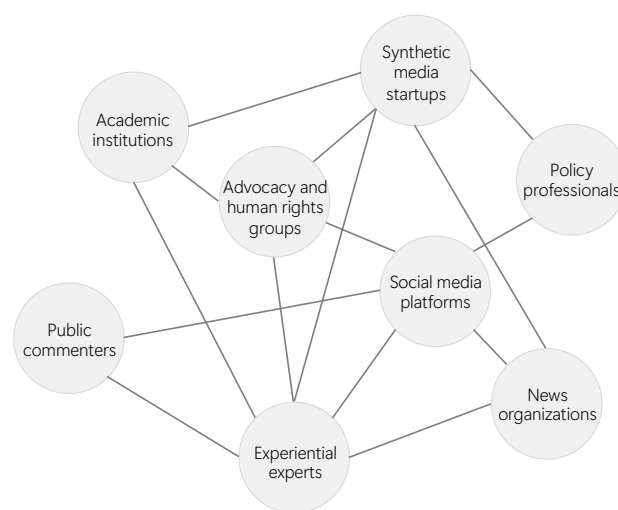
As a co-founder of C2PA, Microsoft has also been involved in the public awareness and education work that C2PA has been conducting through public events and with policymakers about the importance of provenance. And, since its inception, we have been a part of the Partnership on AI's [AI & Media Integrity Steering Committee](#) which has advocated for greater awareness among the public and with policymakers on rising challenges for media integrity presented by generative AI, as well as potential best practices and mitigations. Microsoft has also collaborated with others from the tech industry and civil society on the development of [PAI's Responsible Practices for Synthetic Media](#), such as Adobe, Witness, and the other [Framework supporters](#).

We will continue to work together to share learnings from our experience implementing the framework to support its evolution over time as part of a community of practice. We recognize there is more work to do and 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](#) for educators, as well as guidance for parents in our [Family Safety Toolkit](#).

To meet young people where they are, we have also released "[The Investigators](#)", 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.

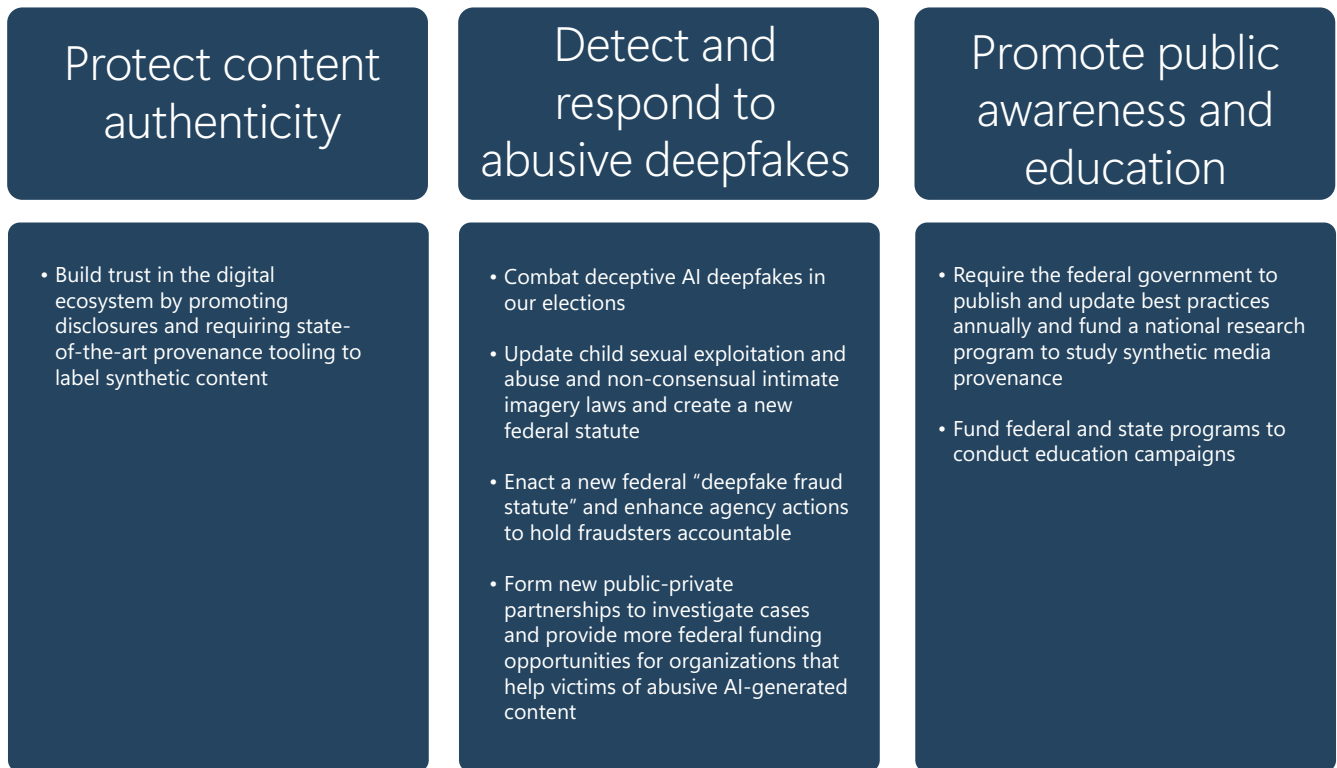
Partnership on AI has worked with more than 50 organizations



Source: Partnership on AI

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

We are sharing new recommendations for policymakers in the United States to consider as they work on advancing legislation to protect the public. The recommendations address three fundamental pillars we believe are essential to a robust policy framework for combating abusive synthetic content risks:



At Microsoft we recognize that this conversation will continue to evolve, and we look forward to being a part of those conversations. However, every organization that creates or uses advanced AI systems also has a responsibility to think broadly about the potential impact of AI on individuals and society.

This white paper is our attempt to put forward our legislative and policy ideas to address abusive AI-generated content risks. We look forward to receiving feedback and continuing to work with civil society, policymakers, and stakeholders across the tech sector and beyond on effective policy measures.

## Protect content authenticity

The ability of AI systems to create compelling audio and visual content has undergone rapid improvements in recent years, with the rise of highly capable text to image models like Dall-E, Stable Diffusion and Midjourney. These technologies are supercharging people's creative expression, allowing anyone to create a wide range of audio and visual content, including highly lifelike media depicting real people or scenes. These tools also increasingly provide easy to use editing functionality allowing people to do everything from touching up a photo to dramatically reimagining entire scenes. This technology will continue to improve rapidly, with powerful text-to-video models, capable of generating entire videos from text prompts, likely to soon be accessible broadly. Increasingly autonomous systems, able to converse with people using synthetic audio, will offer the potential of virtual assistants able to assist across a range of issues.

The increasing prevalence of AI-generated content is creating concern around whether people can trust the information they are interacting with online. In [Microsoft's 2024 annual Global Online Safety Survey](#), there was a particular focus on how people of all ages perceive the opportunities and risks posed by generative AI. While the survey showed that young adults see the use of AI as exciting and as a practical tool for translation purposes, work and school, they also expressed concern about at least one potential risk, including deepfakes.

[Only 11% of respondents to a different poll](#) believed they could accurately identify AI content, and the recent coverage of altered images of public figures has further heightened concerns about the impact of synthetic content on trust in the information ecosystem.

Do you think you would be able to tell if an image, video or audio clip was generated using artificial intelligence?



Source: Data for progress



Beyond grappling with a flood of AI-generated content, the rising tide of synthetic media raises questions and challenges peoples' ability to detect and trust authentic content. It is becoming increasingly easy for malicious actors to claim authentic content, such as imagery of atrocities, are "fake" or AI-created. We must therefore leverage provenance tools both to help people to understand when content comes from a trusted source and to label and recognize AI-generated content. Not all AI-generated content is abusive—indeed, we want people to make the most of this technology and their creativity, but we need measures to support information integrity.

As with other transformative technologies, society will need new rules to guide responsible approaches to synthetic content. Already, our federal and state governments are taking steps and thinking about how to address this complex challenge.

At the federal level, there have been bills introduced in Congress to require the identification and labeling of online images, videos and audio generated using AI, including through metadata and watermarking. There are also efforts underway to introduce federal legislation that would prohibit the removal of provenance labels, the generation and distribution of false provenance information, and the development of products primarily intended to disable provenance information.

Similar efforts are also underway at the state level. Legislators in [California](#), for example, have put forward legislation that would standardize the specifications of provenance metadata. In Connecticut, [legislation](#) that would require the developer of an AI system to ensure that audio, image, or video outputs are marked in a machine-readable format and detectable as synthetic digital content, passed the state Senate.

Building trust in the digital ecosystem will require a range of interlocking, complementary policy measures, with industry, government and civil society all playing their part. No one measure alone will suffice. Underlying all these efforts, however, is the objective of building public understanding that differentiates authentic, non-AI generated content from AI-generated or AI-edited content. The following are important measures to achieve that objective.

*Providers of AI systems designed to interact with people should be required to provide notification to users that they are interacting with an AI system.*

Transparency and accountability obligations are at the core of protecting people from the abuse of any technology, including AI. At Microsoft, they are central to our [responsible AI approach](#) along with other principles, including fairness, reliability and safety, privacy and security, and inclusiveness.

## Fairness

How might an AI system allocate opportunities resources, or information in ways that are fair to the humans who use it?

## Reliability and safety

How might the system function well for people across different use conditions and contexts, including those it was not originally intended for?

## Privacy and security

How might the system be designed to support privacy and security?

## Inclusiveness

How might the system be designed to be inclusive of people of all abilities?

## Transparency

How might people misunderstand, misuse, or incorrectly estimate the capabilities of the system?

## Accountability

How can we create oversight so that humans can be accountable and in control?

AI systems are becoming more capable and interactive, helping people to more quickly complete tasks or search for information in convenient and intuitive ways, for example by allowing people to converse with a system in natural language. As these interactive systems become more commonplace, it will be critical that users know when they are interacting with an AI system, rather than with another human being.

Providers of AI systems intended to interact with people should be required by law to notify users they are interacting with AI, unless this would be obvious to a reasonably well-informed person, considering the circumstances and the context of use. [The EU AI Act](#) includes such a requirement, stipulating that “AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system.” The recently enacted [Colorado AI Act](#) also requires developers or deployers of any AI system that is intended to interact with consumers to inform each consumer who interacts with the system that the consumer is in fact interacting with an AI system. This requirement is designed to foster trustworthy AI. The U.S. should pass federal legislation that requires a similarly straightforward duty on providers of systems intended to interact with people. A single federal standard would help simplify disclosures to users and increase broader public awareness. It is already included as a provision in at least one bipartisan federal

bill, the [Artificial Intelligence \(AI\) Research, Innovation, and Accountability Act](#), which among other things, requires large internet platforms to provide clear and easy to understand notice to users when a platform is using generative AI to create content the user sees. Passing legislation with this requirement would go a long way in promoting trust in people’s interactions with technology.

***We should also promote the use of provenance information for authentically captured media so that we accelerate the government’s adoption of provenance technologies that can help the public better understand whether media comes from a government source.***

Amidst a rising tide of AI-generated deceptive content, it is becoming increasingly valuable to provide signals of “authenticity,” meaning content that is authentically captured or composed by a given non-AI source. To help the public differentiate between deceptive or manipulated content and authentically captured media, provenance information should first and foremost be added to authentic media at its origin. Greater use of provenance information for authentic media will enable the public to more effectively assess any given piece of media.

Although bad faith actors may remove or fail to apply labels to synthetic content in an attempt to deceive the public, good-faith actors can deploy tamper-evident provenance tools that attest to authenticity back to the content's source of origin—and the public can give greater weight to content with authenticity provenance information present. This will be important for reinforcing the value of synthetic and authentic content labeling.

Tooling based on the C2PA standard demonstrates the promise of these types of measures: it attaches cryptographically signed metadata to audio and visual files that allows someone to see who created the file and if and how the file has been edited through the course of its existence. Legislation should not, however, mandate the C2PA standard or any specific tooling or standard; instead, legislation should more generally point to industry standards and require use of state-of-the-art tooling.

Government has an important role in adopting these tools, enabling their wide use, and supporting public education. The [White House Executive Order](#) took a critical step forward, tasking the Office of Management and Budget with issuing guidance to agencies for labeling and authenticating content that they produce or publish by June 2025. This guidance will inform government agency use of provenance metadata on the authentic images, audio, and video they distribute, and will show, for example, if files were indeed captured by a camera and when.

We applaud this work but recommend accelerated adoption—much as the Department of Defense has done with its [pilot](#) to explore adding provenance to media content it produces and owns. Government should take steps to help people identify authoritative government outputs as authentic.

To further mitigate the risks that content is misused for deception, impersonation, and fraud, the federal government should support awareness and use across the media ecosystem, by journalists, enterprises, and the public at large. Already, camera manufacturers like Sony, Leica, and mobile applications like Truepic include these capabilities. Microsoft also recently announced [Microsoft Content Integrity](#) to support election candidates, political parties and journalists with authentic capture and provenance signing of photo, video, and audio files. At the same time, it will remain important to ensure that use of these tools respect privacy and civil liberties. Importantly, C2PA has developed methods for handling [anonymity and privacy](#), which have already been used to provide protections to citizen reporters who capture images of war crimes and transmit photos signed with provenance information. Public awareness campaigns on the risks posed by abusive AI-generated content, outlined later in this white paper, should expressly include information on verifying authentic content to support widespread adoption of these solutions.

***Finally, policymakers should examine requiring system providers to use state-of-the-art provenance tooling to label synthetic content and prohibit the stripping, tampering with or removal of provenance metadata.***

The U.S. government should ensure the National Institute of Standards and Technology (NIST) and the U.S. AI Safety Institute and AI Safety Institute Consortium prioritize work to build out further authenticity and provenance techniques. This work should be done with AI Safety Institutes in like-minded partner countries, helping develop techniques and guidance to support information integrity on a global scale. Providers of AI systems that can create sophisticated audio and visual content should be required by law to utilize state-of-the-art provenance tooling so people can understand whether a piece of content is AI-generated or manipulated. Alongside this provider-focused requirement, and to reinforce the value of synthetic content labeling, policymakers should 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.

In addition to promoting the use of provenance for authentically captured or produced media, federal legislation should also require system providers to use state-of-the-art provenance tooling to label synthetic content. Because significant work remains actively underway at NIST and in other research settings to understand the best technical approaches for implementing provenance metadata for synthetic content, requirements should specify that these measures be implemented as far as technically feasible and as reflected in any relevant technical standards (for example, the C2PA provenance specification). Furthermore, requirements should account for the specificities and limitations of different types of synthetic digital content, implementation costs, and the generally acknowledged state-of-the-art requirements should specify and respect any applicable accessibility requirements.

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 automobiles, policymakers should prohibit intentionally deceptive tampering with, stripping or removal of provenance metadata indicating if content is authentic or synthetic.

To protect privacy, legislation should support the ability of people and organizations to redact personal information from provenance information and simply retain authentication of the digital source type (i.e., the source from which media was created)—which is ultimately the most essential piece of information indicating whether a media file was authentically captured or AI-generated or manipulated.

Legislation should also protect the identity of whistleblowers or journalists and enable researchers to test the rigor of these systems.

Congress is currently exploring legislation, such as through Section 511 of Senator Warner’s Intelligence Authorization Act, [S. 4443](#). This section would establish penalties for bad actors working to intentionally remove, strip or tamper with authenticity or provenance metadata of AI content. It is a common-sense effort to protect responsible AI efforts and to hold bad actors accountable. Microsoft encourages Congress to work together to pass Section 511, or a standalone version of this section, if introduced.

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.)

## Detect and respond to abusive deepfakes

*New laws and actions are needed to protect against deceptive AI content in our elections and prohibit fraudulent misrepresentations created and distributed using AI tools.*

More people and countries will vote for their elected leaders in 2024 than in any year in human history. At the same time, AI presents new challenges to our elections, and in the United States we have already seen attempts by [bad actors to deceive voters using this new technology](#). While there has been progress to address this issue, including 20 companies from the tech sector coming together at the Munich Security Conference in February 2024 to announce a new Tech Accord to Combat Deceptive Use of AI in 2024 Elections, more action is needed to protect our elections from AI-based manipulation.

Microsoft recommends as a next step that Congress pass the bipartisan [Protect Elections from Deceptive AI Act](#), sponsored by Senators Klobuchar, Hawley, Coons and Collins. This important piece of legislation prohibits the use of AI to generate materially deceptive content falsely depicting federal candidates in political ads to influence federal elections, with important exceptions for parody, satire, and the use of AI-generated content by newsrooms.

Such legislation is needed to ensure that bad actors cannot exploit ambiguities in current law to create and distribute deceptive content, and to ensure that candidates for federal office have meaningful recourse if they are the victim of such attacks. Several [states](#) have proposed or passed legislation similar to this federal proposal. While the language in these bills vary, we recommend states adopt prohibitions or disclosure requirements on “materially deceptive” AI-generated ads or something akin to that language and that the bills contain exceptions for First Amendment purposes.

Microsoft is also supportive of the actions that the Federal Election Commission (FEC) began in August 2023 to potentially regulate AI-generated deepfakes in political ads ahead of the 2024 election. [The FEC solicited public comments on a petition](#) seeking amendment of a regulation that prohibits a candidate or their agent from fraudulently misrepresenting other candidates or political parties. The amendment would make clear that the related statutory prohibition applies to deliberately deceptive AI campaign ads. Microsoft urges the FEC to issue guidance promptly to safeguard campaigns, voters and the 2024 election.



Existing robocall provisions are another means of addressing the fraudulent use of synthetic content. These provisions have historically restricted the use of artificial or prerecorded voices and allow for enforcement actions when these rules are violated.

More recently, the Federal Communications Commission (FCC) updated [federal rules](#) to make clear they apply in scenarios in which human voices are generated through artificial intelligence and then used with an intent to defraud. Enabling platforms to self-police is an important tool advanced by clear rules and prohibitions aligned across jurisdictions. This counsels for vesting enforcement responsibilities with regulators and attorneys general to avoid conflicting or extreme outcomes that can arise from using class action litigation as an enforcement tool. Regulators could subject violators to fines, injunctive relief and a requirement to block the illegal calls. In addition to federal enforcement, state enforcement can help to provide attorneys general with their own enforcement tools to address the issue under state law. Policymakers have included exemptions when the customer has provided prior express consent. An exemption should also be available for consumers using their AI-generated voice due to a disability. However, to detect abuse of synthetic voice technology, one needs to be able to identify that the call was AI-generated.

To do so, we recommend that the federal government explore standards for future mobile devices and their hardware to allow for provenance information to be readily conveyed and displayed in real time.

***Child sexual exploitation and abuse and non-consensual intimate imagery laws must also be updated, and Congress should pass a new federal statute to address non-consensual intimate imagery.***

Today, many existing laws in the states to combat child sexual abuse material and non-consensual intimate imagery do not reflect AI-generated content, and on the federal level new laws are required to fight the creation and dissemination of non-consensual intimate imagery. Child sexual exploitation and abuse imagery is near-universally criminalized, given the global recognition that this is an abhorrent crime. 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.

Reports of online child sexual exploitation and abuse content have already been growing year to year: in 2022, NCMEC analyzed just over 32 million reports of CSAM received from across the globe. [This is an 87% increase on the number processed in 2019](#)—with the true scale of child sexual exploitation and abuse content online likely still greater. These numbers likely do not yet incorporate the full scale of the synthetic CSAM risk, but leading child safety organizations such as the [Internet Watch Foundation](#) have reported that AI is already being used to generate CSAM that is indistinguishable from real images, including revictimizing survivors by generating new imagery of known victims.

CSAM is not only inherently harmful but also may be used to facilitate other harms, such as financially motivated extortion, grooming, or trafficking. [Large volumes of synthetic content may also hinder efforts to address real-world harm](#) by overwhelming law enforcement with synthetic content that is indistinguishable from real content, impeding victim identification, and fueling demands from bad actors for new content. Exposure to CSAM may also lead to an increased risk that offenders seek contact with children offline. 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 as well as to 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.

## **At the state level: modernize existing CSAM laws**

As a first step here, state policymakers must modernize existing criminal law so that any attempts to generate synthetic CSAM are criminalized. Existing law should make it clear that the knowing creation, generation, distribution, and/or dissemination of real or realistic CSAM should be criminalized, including where such content is AI-generated. A range of state-level proposals have already been introduced this legislative session, some of which have been signed into law or are gaining traction in their legislative bodies, including [South Dakota's SB 79](#) (signed in February 2024), [Washington's HB 1999](#) (signed in March 2024), and [California's AB 1831](#) (passed Assembly, pending in Senate as of July 2024). More states need to follow this lead and update their laws accordingly. Such measures may help deter its creation, reducing harm and the risk of overwhelming the current child safety ecosystem. On the federal level, the FBI has already [warned the public](#) that CSAM created with content manipulation technologies, including AI, is illegal. Microsoft welcomed this clarity.

## Unlock innovation to detect CSAM and enable responsible AI best practices

Service providers play a critical role in tackling CSAM risks.\* Microsoft has had a longstanding commitment to addressing this harm, and recognizing how emerging technologies might be subject to abuse, we [recently announced our support](#) for new safety by design principles to tackle CSAM risks in AI models and services. These commitments include taking steps to build services responsibly, but equally to continue to innovate. The era of generative AI has accelerated the need for innovative tools and partnerships to address this issue. While hash matching technologies like PhotoDNA will remain critical to detect existing material at scale, new tools will be required to prevent and detect novel child sexual abuse exploitation and abuse imagery (CSEAI), including newly created synthetic imagery. Until recently, federal law required online service providers to [preserve CSAM for only 90 days](#). This was [often insufficient, given the timelines for law enforcement investigations](#). Thankfully, President Biden signed the bipartisan [Revising Existing Procedures on Reporting via Technology \(REPORT\) Act](#) into law, which among other provisions, extends the preservation period to one year.

To clearly support further technical innovation, we also recommend Congress provide express authority for technology companies to use lawfully retained CSAM for the sole purpose of training

technologies to detect child sexual exploitation and abuse material. This would advance industry efforts to detect, address and report all CSAM and enhance existing safeguards in AI models and services.

Microsoft, along with a range of leading AI companies, has been developing and refining best practices in responsible AI, intended to mitigate potential risks in the development and deployment of AI models. This includes red teaming to test a model's propensity to produce harmful content, as well as the systematic measurement and mitigation activities required to reduce risk on an ongoing basis. However, the current U.S. federal framework addressing CSAM does not provide industry with a clear legal basis on which they can safely undertake this kind of testing to ensure AI models cannot produce synthetic CSAM. Microsoft recommends that Congress provide technology companies with appropriate legal clarity so that the necessary systematic measurement and mitigation activities can move forward, and we can better protect and reduce the risks of AI models and services generating synthetic CSAM, in keeping with our commitments.

1

---

\* Globally, a variety of regulatory approaches have already emerged that require online service providers to have systems and processes in place to tackle online safety risks such as CSAM and NCII. Measures such as the United Kingdom's Online Safety Act, the EU's Digital Services Act, and Australian Online Safety Act have also been drafted in a technology neutral fashion, enabling the measures to address AI risks arising from in-scope services. The EU has also taken critical steps to clarify the illegality of synthetic CSAM and NCII through proposed amendments to the Directive on CSAM, and the newly adopted Directive to combat violence against women.

## **Establish an expert commission to study the issue**

In September 2023, the [National Association of Attorneys General sent a letter to Congress](#) signed by attorneys general of the 54 states and U.S. territories requesting that Congress establish an expert commission. They asked for the commission to study the means and methods of AI used to exploit children and to propose solutions to deter and address such exploitation. Although Microsoft recognizes that we must act now and take steps to prevent the harms that synthetic CSAM currently pose, we also agree with the 54 attorneys general that this is an issue that must be studied and better addressed because we may not currently know all the ways synthetic CSAM can manifest and how to best prevent it. Microsoft recommends that Congress develop such an expert commission, with public and private representation on it, to propose solutions that Congress and the states can then evaluate and consider. We also welcome the Tech Coalition's recently announced [generative AI research](#).

## **Pass new state and federal legislative laws to ensure efforts to develop and disseminate synthetic and other non-consensual intimate imagery is appropriately criminalized**

One of the most likely risks arising from the widespread availability of generative AI is the development of highly realistic "deepfaked" images of real individuals. While concerns often center on risks related to democratic processes or political candidates, the vast majority of deepfakes are nude, sexual or pornographic. Images may be taken from social media or other public profiles without the knowledge of the person depicted. We therefore recommend that policymakers pass measures to address the risk that these tools are misused to develop and disseminate AI-generated intimate images without the consent of the subject. We also recommend measures to close existing gaps in the law related to the non-consensual distribution of any intimate imagery.

Over the last decade, most states have enacted measures criminalizing this conduct. However, these bills differ considerably in their definitions, classifications, and remedies. The bills also vary in terms of their *mens rea* requirements, some needing to show that the perpetrator is motivated by a desire to hurt the victim, and many of these laws do not adequately capture the potential for deepfakes. And yet the states, particularly in this most recent legislative session, have introduced several bills to modernize their statutes with the ongoing development of AI.

These bills, including [New Jersey's A3540](#) (passed the Assembly), [Indiana's HB 1047](#) (signed into law), [Idaho's HB 575](#) (signed into law) and [Virginia's HB 1525](#) reflect efforts to ensure that the development, creation, and/or dissemination of synthetic non-consensual intimate imagery is appropriately criminalized.

More states should follow their lead and expand existing laws or draft new ones to include images generated by AI. To take as victim-centred approach as possible, Microsoft recommends that such activity must be done knowingly, but should not require evidence that content was shared or produced with the intent to cause distress to the victim.

On the federal level, Congress should address this issue for both non-synthetic and synthetic content. Although Congress created a new, private right of action for victims of non-consensual intimate imagery in the [Violence Against Women Act Reauthorization Act of 2022](#), at the federal level there is no equivalent criminal offense specifically aimed at combating this harm. While this white paper is focused on synthetic content, we would be remiss to ignore this issue in the non-AI context when the impact is so devastating and the need for legislation so urgent.

Therefore, Microsoft endorses and encourages Congress to pass Senators Klobuchar's and Cornyn's bipartisan [Stopping Harmful Image Exploitation and Limiting Distribution \("SHIELD"\) Act](#), the first federal criminal law introduced to prohibit non-consensual distribution of intimate images dissemination. The Senate passed the SHIELD Act in early July by unanimous consent, and we urge the House to follow suit.

Fortunately, there is also bipartisan support in Congress to address the spread of AI-generated imagery. The bipartisan [Preventing Deepfakes of Intimate Images Act](#), introduced by Representative Joseph Morelle (D-NY) in the House and by Senators Hassan, Cornyn, Butler and King in the Senate, prohibits the non-consensual disclosure of digitally altered intimate images. The legislation would make the sharing of these images a criminal offense and would create a private right of action for victims to seek relief. Both pieces are necessary in legislation to deter bad actors and to ensure justice for victims. The bill also seeks to hold AI services accountable but specifically carves out liability when voluntary good-faith efforts to restrict deepfakes while providing recourse for victims are made. These provisions are common-sense and encourage providers and applications to take reasonable steps to protect users while also enabling liability and charges to be made against bad actors, such as "nudifying" apps or other services marketing the production of deepfake intimate imagery.

*Congress should enact a new federal “deepfake fraud statute”, and federal agencies should take action to hold fraudsters accountable.*

Like other areas that are evolving because of generative AI, the fraud landscape is fraught with new challenges, making it difficult to discern genuine content from deceptive schemes. United States law enforcement officials and industry leaders recognize that we are at an inflection point concerning criminal use of AI and synthetic media. Synthetic content provides cybercriminals with the capability to enhance and scale existing fraud schemes while enabling new fraud types.

Financial fraud scams had already been growing exponentially over the years, even before and without AI, overwhelming police and prosecutors. Online and telephone scams are particularly commonplace, and [the most frequent targets are older Americans who hold more wealth as a group and are often seen as ripe targets by scammers](#). Elder fraud complaints to the FBI’s Internet Crime Complaint Center [increased by 14% last year](#), and according to the AARP, [Americans over 60 lose \\$28.3 billion each year to fraud](#). Yet, as the U.S. population ages and with new technology, such as generative AI, those numbers are expected to grow.

Deputy Assistant General Monaco recently discussed the growing threat posed by criminal use of AI and synthetic content emphasizing that AI lowers the barriers to entry for criminals, changes how online crimes are committed, and supercharges the threat posed by the most sophisticated cybercriminals.

## **Enact a new federal deepfake fraud law**

Although there are current existing federal fraud statutes that could be revised and enhanced to address synthetic content, the most comprehensive way to approach this issue would be to enact a new federal synthetic content fraud statute to encompass both civil and criminal provisions. The statute could also provide for criminal penalties, civil seizure and forfeiture, as well as injunctive and other equitable relief. While there is no pending legislation in this precise area, there is a useful, albeit imperfect template for Congress to consider. In 2010, [the Truth in Caller ID Act](#) was enacted which makes it a crime “to cause any caller identification service to knowingly transmit misleading or inaccurate caller identification information with the intent to defraud, cause harm, or wrongfully obtain anything of value.” The Act provides for civil forfeitures as one of the penalties for violations, enforceable by the FCC, and the possibility of criminal fines and imprisonment. It also allows enforcement by state attorneys general, who may bring civil suits on behalf of the residents of their states.



Congress should consider legislation patterned after this bill and prohibit similar deceptive content in the AI context, for example, “the knowing transmission of synthetic content with the intent to defraud, cause harm, or wrongfully obtain anything of value.” Such a statute could authorize enforcement by the FCC or FTC, as well as prosecution by the DOJ.

While we recognize that passing legislation is no easy feat, this is an area where we believe there could be bipartisan consensus. There are also advantages to this comprehensive approach. A bill of this nature would provide substantial flexibility for law enforcement authorities to address violations across a spectrum of seriousness. State attorneys general can also leverage the federal framework’s rights of action to state-level priorities while simultaneously enabling federal oversight and control through an administrative right of intervention. Lastly, the statute could address the question of state preemption in a manner that allows state legislatures to pass laws targeting fraudulent use of synthetic content, providing additional protection to citizens.

### **Enhance federal agency action**

Multiple federal agencies already possess the authority to address fraudulent synthetic content and should exercise that authority by publishing guidance and initiating enforcement actions. While this is not an exhaustive list, here are some examples of actions that agencies can take now.

1. The United States Sentencing Commission can revise the nonbinding federal sentencing guidelines for existing fraud-related offenses to add sentencing enhancements for the fraudulent use of synthetic content during the commission of a crime. The current sentencing guidelines for fraud include enhancements for a wide variety of aggravating factors, each of which increases the “level” of the offense for the purpose of sentencing. Use of synthetic content to commit a crime should likewise be an aggravating factor that federal judges should consider.
2. The United States Deputy Attorney General (DAG) is empowered to prioritize enforcement of particular crimes by U.S. Attorneys. The DAG can issue a memorandum prioritizing synthetic content fraud enforcement. This action is consistent with prior directives of the DAG who regularly provides DOJ personnel with guidance relating to the investigation and prosecution of unlawful conduct.
3. The FTC is authorized to seek penalties from perpetrators of unfair and deceptive trade practices where the FTC has already issued a written decision that the conduct at issue is unfair or deceptive, and the enforcement target knew that the conduct was unfair or deceptive. The FTC can exercise its authority by serving “[Notice of Penalty Offenses](#)” describing conduct that the FTC considers to be unfair or deceptive—in this case fraudulent use of synthetic content.



Failure to comply with such a Notice subjects the offender to substantial civil penalties. The FTC could scale enforcement by issuing Notices of Penalty Offenses to any entities knowingly involved in fraudulent or otherwise unfair and deceptive trade practices predicated on the creation or use of deepfakes.

Indeed, the FTC is already addressing fraud committed using synthetic content with its authority to promulgate a new rule prohibiting certain forms of impersonation. On March 1, 2024, the FTC published the [text of the final rule](#), which reads as follows:

*It is a violation of this part, and an unfair or deceptive act or practice to (a) materially and falsely pose as, directly or by implication, a business or officer thereof, in or affecting commerce as commerce is defined in the Federal Trade Commission Act (15 U.S.C. 44); or (b) materially misrepresent, directly or by implication, affiliation with, including endorsement or sponsorship by, a business or officer thereof, in or affecting commerce as commerce is defined in the Federal Trade Commission Act (15 U.S.C. 44). To be codified at 16 C.F.R. § 461.3.*

The FTC impersonation rule presents an opportunity for future enforcement to address synthetic content fraud.

***Form new public-private partnerships to investigate cases and provide more federal funding opportunities for organizations that help victims of abusive AI-generated content.***

[Microsoft's Digital Crimes Unit \(DCU\)](#) is an international team of technical, legal and business experts that fights cybercrime, protects individuals and organizations, and safeguards the integrity of Microsoft services. Its expertise and unique insights into online criminal networks enable it to uncover evidence used in Microsoft's criminal referrals to law enforcement. The DCU also works to increase the operational cost of cybercrime by disrupting the infrastructure used by cybercriminals through civil legal actions and technical measures. No single entity can fight cybercrime alone; the DCU has developed deep relationships with security teams across Microsoft, and with law enforcement, industry partners, security firms, researchers, nongovernmental organizations and customers to increase both scale and impact when fighting cybercrime. This model could serve as a template for combating abusive AI-generated content.

Other partnerships, such as ones with NCMEC, will also be critical in fighting these harms, especially when it comes to supporting NCMEC to distinguish AI-generated content. Unless apparent CSAM carries provenance information, it is likely that NCMEC will continue to grapple with huge volumes of content, some of which may be indistinguishable from "real" CSAM. Making this distinction is critical for law enforcement for many reasons, including victim identification. Attempting to verify if content contains provenance metadata or watermarks should always be a first line of defense. However, if this information is missing, detection tools can be used to try to assess the probability that a piece of content is AI-generated or modified.

While experts are skeptical that such detection tools will be a viable solution in the long-term as deepfake capabilities become increasingly sophisticated, we see potential to use them in the near-term as one of many methods that forensics experts can use to assess the authenticity of high-stakes content, such as CSAM. Microsoft welcomed the passage of the REPORT Act, which will better enable NCMEC to leverage cutting edge technology in its work. Additionally, work on options to combat this challenge could be undertaken through [NIST's process to respond to the AI Executive Order](#).

Furthermore, the federal government should invest more into training law enforcement to identify deepfakes and into developing better and more resilient technology to analyze potential deepfakes. [The Bureau of Justice Assistance](#) (BJA), which issues grants for law enforcement priorities, should prioritize synthetic content fraud enforcement. While some BJA grants fund training and operational costs for law enforcement, others are directed to researchers and other non-governmental actors. The BJA could initiate a grantmaking proceeding focused on training law enforcement personnel to identify, investigate, and prosecute synthetic content fraud. The BJA is also authorized to consider funding research which could include research focused on synthetic content verification and detection technology and related AI investigative tools that would benefit local, state, and federal enforcement in this area.

Victims of synthetic non-consensual intimate imagery may also have concerns about reporting to law enforcement agencies, who may not be appropriately resourced to address this accelerating category of harm. The federal government should ensure that funding is available for law enforcement training programs specific to this harm, and law enforcement should seek to take forward cases where possible, for deterrent effect. Technology companies may also wish to consider partnering with law enforcement agencies to offer training on the kinds of evidence that may be available to support investigations and prosecutions. Equally important will be to ensure that judges are well-educated on the harms arising from the generation and distribution of any non-consensual intimate imagery. We recommend that the government explore grants to advance judicial education on AI-generated content in legal proceedings where it can produce particularly consequential effects. Stakeholders can also work with government organizations such as the Federal Judicial Center and industry organizations, such as the American Bar Association, to drive forward these efforts.

**As the volume of cases involving synthetic content rises, so does the need to provide resources to those impacted by it, but many of these services are not adequately resourced and more federal support is required.**

In the context of CSAM, NCMEC is the linchpin not only for efforts in the United States but also globally. NCMEC's workload has risen exponentially in recent years—even before the advent of generative AI, NCMEC was already overwhelmed by incoming CSAM reports (as are the law enforcement agencies to which it routes reports). For example, in 2023 the [NCMEC CyberTipline](#) received 36,210,368 reports, and most of these reports related to victims or offenders outside the United States. By contrast, fifteen years earlier 54% of tips related to victims or offenders in the United States, and there was a total of 102,029 reports received by the CyberTipline that year. Despite this increase in volume, the amount that NCMEC has received in federal funding has stayed somewhat stagnant over time.

Microsoft, along with a range of other private sector entities, provides voluntary funding to NCMEC, but additional governmental funding is needed to ensure NCMEC can continue to expand, adapt its technology, and meet the moment.

While there are legislative proposals to increase funding to NCMEC, such as through Senator Wyden's and Representatives Eshoo's and Fitzpatrick's bipartisan [Invest in Child Safety Act](#), and the [Missing Children's Assistance Reauthorization Act](#), which Chairman Durbin and Ranking Member Graham advanced through the Senate last year, more can be done without the express need for legislation.

We recommend that Congress request and the administration award more funding to NCMEC so that it can carry out its vital functions, such as the operation of its CyberTipline. In Fiscal Year 2023, the Office of Juvenile Justice and Delinquency Prevention at the DOJ gave NCMEC approximately [\\$41 million](#) to fund all its critical work on behalf of missing and exploited children. More will be needed with the advent of generative AI, as NCMEC expects the rate of tips into the CyberTipline to grow exponentially, and staff will need more training on AI, as well as more tools to assist in reviewing documents and images, managing caseloads, and implementing more reforms and updates to the CyberTipline.

Similarly, as the volume of cases involving synthetic non-consensual intimate imagery rises, so does the need to ensure that support services are readily available. This funding must include helplines such as the one at the [Cyber Civil Rights Initiative](#) (CCRI)—the first-ever national helpline for survivors of image-based sexual abuse. As of now, they do not have the resources to handle the volume of calls they receive and need more funding to assist with everything from initial call intake to more specific research, image removal, and counselling needs.

To date, the CCRI Image Abuse Helpline has assisted over 26,353 individuals. In 2018, it responded to 2,670 callers. In 2023, that call volume increased to 6,600 calls, representing a nearly 150% increase in number of calls. Because laws in this area are relatively “new,” organizations in this space do not have access to many federal grants, and they have struggled to keep up with demand for their services. Indeed, the [Department of Justice Office for Victims of Crime](#) (OVC) just recently funded for the first time the the CCRI’s national helpline. Therefore, more funding needs to be given to OVC to support this national helpline and for additional support to organizations to assist victims of these crimes. For example, CCRI is hoping to expand its [Safety Center](#) so that it can become a one-stop shop for victims of image-based sexual abuse.

## Promote public awareness and education

The ways synthetic content harms manifest will evolve, and new harm areas will likely emerge, as bad actors seek to create and share deceptive AI-generated content. Considering this, providing provenance data for both AI-generated and user-generated content will become increasingly important as a means to provide information about the history and origin of content, including how it was made and whether it has been edited. While providing this type of transparency will help build societal resilience to deceptive AI-generated content, no disclosure method for AI-generated content is perfect and all will be subject to attacks. These attacks will include bad actors removing provenance information from AI-generated content to deceive the public into thinking it is authentic, as well as forging watermarks to mark authentic content as AI-generated. It will be critical to continually assess and improve the efficacy of disclosure approaches for AI-generated and manipulated content, to ensure that the transparency they offer is meaningful to content consumers, and to make sure that the capabilities and limitations of these approaches are well understood by the public. Without this, we run the risk of individuals [distrusting all digital content and dismissing even the authentic as manipulated](#); this would have grave consequences for our economy, court rooms, the state of elections, and even national and global security.

*Require the federal government to publish and update best practices annually and fund a national research program to study media provenance.*

The technological capabilities of AI will continue to improve, which will require the federal government to regularly update best practices and standards for helping the public understand how to navigate synthetic content. Congress should also secure funding for a dedicated national research program supported by the National Science Foundation in partnership with AISIC to ensure NIST's work of promoting best practices continues as deepfake technology evolves.

NIST, through its [AI Safety Institute and Consortium \(AISIC\)](#), has begun the critical work of assessing best practices for synthetic content labeling, verification, and detection. It will be important to update these best practices annually as the sophistication and complexities of synthetic content increase, methods and tools for labeling and detection progress, adversarial attacks to deceive the public about provenance evolve, and as the public's understanding on labeling and detection approaches grows. Congress should ensure NIST has sufficient appropriations to continue this work in the long-term and should require that these best practices be reported publicly on an annual basis.

A national research program to study synthetic content harms that can reach across common AI research resources for academic communities and share information and best practices related to key topics is essential as deepfake technology evolves. We recommend that Congress secure this funding supported by the National Science Foundation and in partnership with AISIC. Such a research program should explore existing and emergent synthetic content harms, building an evidence base of where harms are manifesting, and assessing how to best measure and mitigate them. In addition to harms directly related to synthetic content, this should include core methods, designs and signals for consumers and an assessment of any harms resulting from loss of trust in authentic content. Research should also assess how well tools to label and detect synthetic content and display provenance are working in practice, including sociotechnical analyses of how they are used and perceived. Evidence on how well authenticity and provenance infrastructure is working in practice should inform ongoing public education campaigns and best practices for synthetic content disclosure.

***Fund federal and state programs to conduct education campaigns.***

Governments are in a unique position to deliver tailored education campaigns to the public around safety and harms, just as they do every day for traffic, weather and more. Congress and states should use existing funding programs and create new programming to help educate the public.

Federal and state governments should use existing funding programs and create new programming that would help educate the public about deceptive uses of synthetic content presenting safety risks and harms, as well as approaches they can use to discern amongst digital content. This includes how to assess signals about whether content was authentically captured, or AI-generated or manipulated, what signals and tools can be used to see if it came from a source the content consumer trusts, as well as recommended practices to address the latest scams employing synthetic content. Education campaigns can also be targeted across vulnerable demographics such as older adults and young people. [According to the AARP](#), when it comes to new technology, most older adults are later adopters and have lower confidence in their digital literacy. Similarly, Microsoft's [research](#), conducted in partnership with National 4H, shows that 72% of young people seek support from adults in learning how to use AI tools correctly and with confidence.

[The National Artificial Intelligence Advisory Committee \(NAIAC\) recommended](#) creating a National AI Literacy Campaign that would foster AI literacy, leveraging the Biden administration's digital equity campaign as an AI literacy framework, investing in formal educational or existing learning frameworks to advance AI literacy, and investing in informal learning opportunities such as standalone public sessions and messaging efforts.



This campaign could also build off valuable work already begun by federal agencies. We recommend that NIST, in collaboration with other agencies, leverage AISIC findings to foster media literacy, so that Americans learn about both the risks of synthetic content and tools available to help protect themselves from being deceived when such content is misused. This would help train the public to become critical content consumers. It would also help ensure that as provenance and other complementary disclosure methods are deployed at scale, they are easily digested and comprehended, including what they mean and do not mean, their strengths and limitations, and how to use them. Such a campaign could elevate [guidance from the FTC](#) to protect consumers by increasing awareness of best practices such as how to avoid scams leveraging AI-generated content.

The federal government can also invest in and help build partnerships between it, industry and civil society that accelerate work to educate people about authenticity and provenance tooling. This partnership would not need to start from scratch; there is already a good foundation in the success of projects, such as [BBC Verify](#), which could be a key part of the effort.

We also recommend that any education campaign reflects the input of civil society groups and is disseminated in coordination with groups trusted by local communities. Beyond achieving broad public awareness, education campaigns should specifically reach frontline actors, including local media and journalists, community leaders,

and civil liberties and human rights groups who will need to assess potential deepfakes and educate others as part of their work. Education campaigns for these audiences should be complemented by access to forensics experts and leading-edge tools validated by the NIST AISIC.

Some education campaigns can focus on areas of civic importance, such as election integrity. In March 2024, the Commissioners of the U.S. Election Assistance Commission (EAC) [approved the use of Election Security grant funding](#) authorized by the Help America Vote Act (HAVA) to counter disinformation generated with AI. Grantees may use HAVA Election Security Grant Funds to counter foreign influence in elections, election disinformation, and potential manipulation of information on voting disseminated and amplified by AI technologies.

This could include access to tools like content provenance signatures as a service for election officials, and public information campaigns about provenance use to help the public understand what content can be traced back to election bodies versus what may have been seeded by someone else. Grantees should consider using these funds for public education campaigns, and Congress should also consider leveraging this funding and other existing funding for such purposes.



Lastly, we recommend continued efforts to support online safety and media literacy education for young people and older adults, including through specialized curricula.

For young people, developing these skills will be critical for their digital futures, including understanding how to engage with complex online information ecosystems, as well as the safe and responsible use of AI technology. Education is critical to ensure young people also understand the real harms that can arise from the misuse of technology and can take steps to protect themselves and others. For older adults, digital literacy can help them thrive in an ever increasing digital world and improve their social engagement, financial security and overall participation in their communities. AARP's partnership with OATS is a template for ensuring better education and access for this demographic and can be supported and modeled more broadly.

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](https://aka.ms/protectthepublic)**



©2024 Microsoft Corporation. All rights reserved. The example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious. No association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred.

Microsoft, list Microsoft trademarks used in your white paper alphabetically are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.