Microsoft

Protecting the Public from Abusive AI-Generated Content

Table of Contents

Foreword

Hugh Milward
Vice President, External Affairs, Microsoft UK

As we stand at the beginning of a new age of technological innovation, the UK finds itself at a critical moment. Artificial intelligence (AI) is no longer a distant prospect but a present reality, transforming businesses, revolutionising healthcare, and accelerating scientific discovery across the UK. Yet, as with any transformative technology, AI brings with it both immense opportunities and significant challenges.

At Microsoft, we believe that addressing these challenges isn't just a technical imperative, but an ethical one. This white paper outlines our approach to combating abusive AI-generated content in the UK.

Our research reveals the scale of the problem: from increased risks to women and children, as well as reduced faith in information, the potential for harm is clear. However, we remain optimistic about our ability to harness AI's benefits while mitigating its risks. The UK has a long history of balancing innovation with ethical considerations, and AI presents an opportunity to build on this strong legal and regulatory framework.

We outline a series of technological solutions, policy recommendations, and proposals on how the UK can bring the public and private sectors together to address this issue head-on. Central to our recommendations is the need for clear, proportionate regulation that protects individuals without stifling innovation. We advocate for integrating provenance tools, strengthening legal frameworks, and enhancing measures to protect electoral integrity.

Modernised legislation to protect the public is one of Microsoft's six focus areas to address risks arising from abusive AI-generated content.

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

This paper offers concrete recommendations for UK policymakers, focusing on three key areas: promoting content authenticity, detecting and responding to abusive deepfakes, and educating the public about synthetic AI risks. These proposals aim to protect our democratic processes, safeguard consumers from fraud, and shield vulnerable individuals from exploitation.

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

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

The time for action is now. We must seize this moment to shape an AI future that reflects the best of British innovation, pragmatism, and responsible leadership.

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 US 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)

2017 July

Lip-syncing Obama: New tools turn audio clips into realistic video

Source: UW News

2019 August

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Source: WSJ

2021 August

How a deepfake Tom Cruise on TikTok turned into a very real AI company

Source: CNN

2023 June

DeSantis campaign shares apparent AI-generated fake images of Trump and Fauci

Source: NPR

2023 Sept.

Naked deepfake images of teenage girls shock Spanish town: But is it an AI crime?

Source: Euronews

2024 May

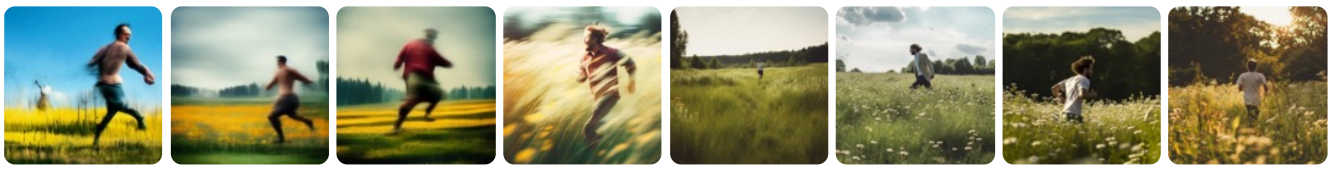
Consultant faces charges and fines for Biden deepfake robocalls

Source: NPR

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.

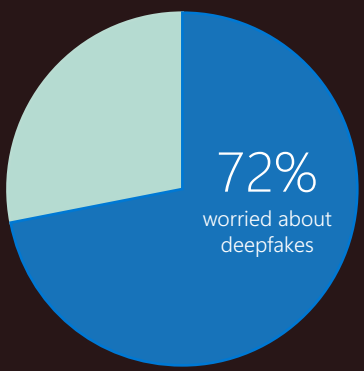
Timeline of Midjourney versions (Prompt: a man running in the meadow photography)



V1 2022 Feb. | V2 2022 April | V3 2022 July | V4 2022 Nov | V5 2023 March | V5.1 2023 May | V5.2 2023 June | V6 2023 Dec

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 weaponisation of this harmful content. It is no surprise that in our most recent Global Online Safety Survey, 72% of people were worried about deepfakes. Research consistently finds that women are far more likely to report experiencing such fears than men.

Percentage of people concerned about the spread of misleading deepfakes, or sophisticated and convincing digital representations



Source: Microsoft Global Online Safety Survey, 2024

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

(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 recognise that the same tools can be used as weapons against the public.

In the following examples, we identify four types of harms that must be addressed to protect UK citizens, including women, children, and seniors, as well as our democratic processes: (1) AI-generated fraud; (2) child sexual abuse material; (3) AI-generated election content; and (4) non-consensual intimate imagery.

Increasing number of paid-for scam ads featuring deepfake footage of celebrities

The Advertising Standards Authority (ASA) has reported an increasing number of paid-for scam ads featuring deepfake footage of high profile individuals like Elon Musk and Martin Lewis endorsing cryptocurrency and trading apps.

In the first six months of 2024, fraud prevention service Cifas reported that a record number of cases were filed to the UK National Fraud Database, with cases of identity fraud the most reported. They suggested that one of the key drivers behind the rise is the easy availability of AI, enabling lower skilled threat actors to create high quality spoof websites and brand impersonations.

Digital identity company Onfido also reported that the number of attempts to use fraudulent deepfakes to circumvent its identity solutions had increased 3000% between 2022 and 2023. They note that a small number of fraudsters are responsible for the majority of deepfake attacks.

Despite the prevalence of these incidents, imposter scams over email, text message and phone are much more common. Recent data from Starling Bank found that over a quarter of UK adults (28%) have been targeted by an AI voice cloning scam at least once in the past year. Yet almost half of UK adults (46%) do not know this type of scam exists, with less than a third (30%) of people confident in their ability to know what to look out for if they were being targeted with a voice cloning scam.

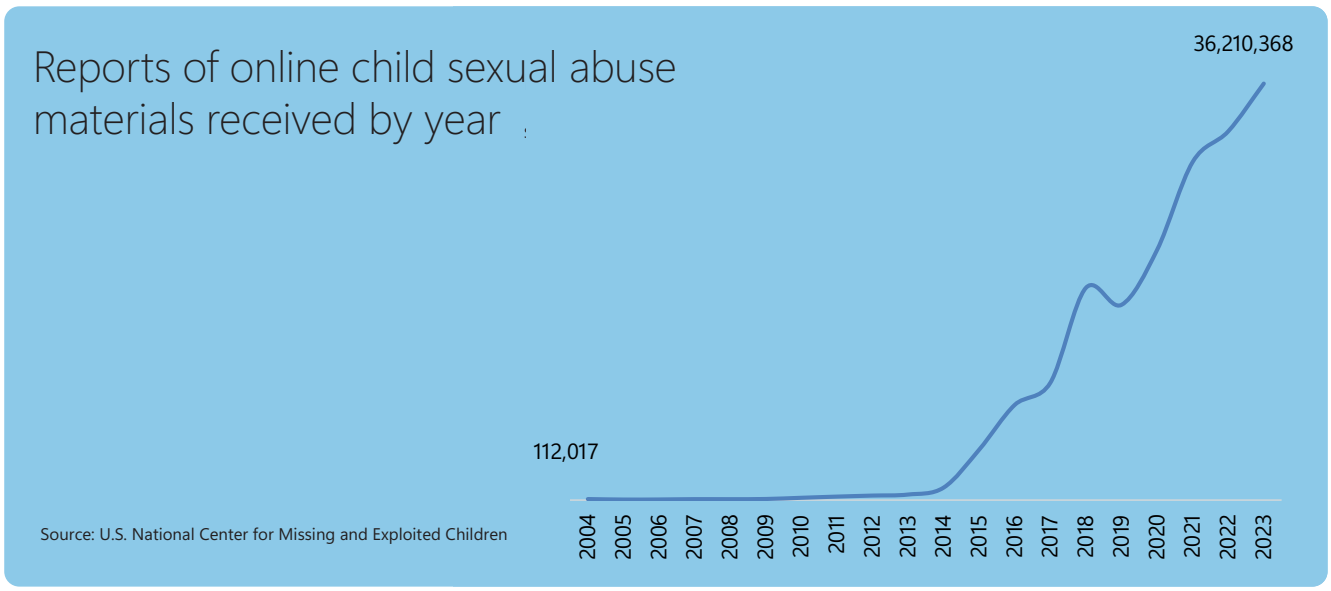
UK school children have created deepfake sexual imagery of their peers

Last year, the UK Safer Internet Centre (UKSIC) said that it had begun receiving reports from schools that children are making, or attempting to make indecent images of one another. Given their age, such imagery is regarded as child sexual abuse imagery, but with the aim of shaming or bullying. This trend has been observed in high schools worldwide and the UK is not immune.

The Internet Watch Foundation (IWF) has also found that use of AI to generate child sexual abuse material (CSAM) is increasing. Its first report in October 2023 revealed the presence of over 20,000 AI-generated images on a dark web forum in one month, where more than 3,000 depicted criminal child sexual abuse activities. Since then, the issue has escalated, with over 3,500 new AI-generated CSAM uploaded to the same forum as of July 2024. And of the images confirmed to be child sexual abuse, more images depicted the most severe kinds of abuse.

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 IWF has reported on perpetrators using AI to alter existing CSAM to generate new content, re-victimising survivors. And recent research from Thorn and the U.S. National Center for Missing and Exploited Children (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.



Deepfakes during the UK General Election

During the 2024 UK General Election several high-profile politicians found themselves targets of AI-generated content that spread rapidly across social media platforms.

Labour's then Shadow Health Secretary, Wes Streeting, became a repeated target of such deepfakes. Early in the campaign, a doctored video circulated widely, appearing to show him making disparaging remarks about his Labour colleague Diane Abbott during a BBC Politics Live appearance. Days before the election, Streeting was targeted again. This time, an audio clip purporting to capture him using profane language and expressing indifference to Palestinian casualties went viral. The clip, which Streeting promptly denounced as fake, garnered hundreds of thousands of views within hours. This incident highlighted the evolving sophistication of AI-generated content and its potential to cause disruption, especially during critical periods of an election campaign.

While some AI-generated content was clearly satirical – such as footage of the Prime Minister and Leader of the Opposition discussing policies in the context of a popular video game – other instances blurred the lines between humour and misinformation. For example, an AI manipulated TikTok video showing the Prime Minister making callous statements about energy prices, had the potential to influence perceptions.

The BBC's disinformation team found that these deepfakes created confusion among some voters, particularly those less familiar with AI technology. While many social media users were able to identify the content as fake, others expressed uncertainty about its authenticity, highlighting the challenges in distinguishing real from fabricated content.

Synthetic non-consensual intimate imagery is weaponised against women

A Channel 4 News analysis of the five most visited websites hosting pornographic or intimate deepfakes in 2024 found that almost 4,000 famous individuals were featured, including female actors and musicians.

Similarly, research by My Image My Choice found over 275,000 intimate deepfake videos on the most popular deepfake sites in 2023, with a total of more than four billion views, and with more videos uploaded to these sites than all previous years combined. The most targeted group for abusive synthetic content is women. Specific groups are also disproportionately targeted, such as high-profile female actors, social media personalities and women connected to politics.

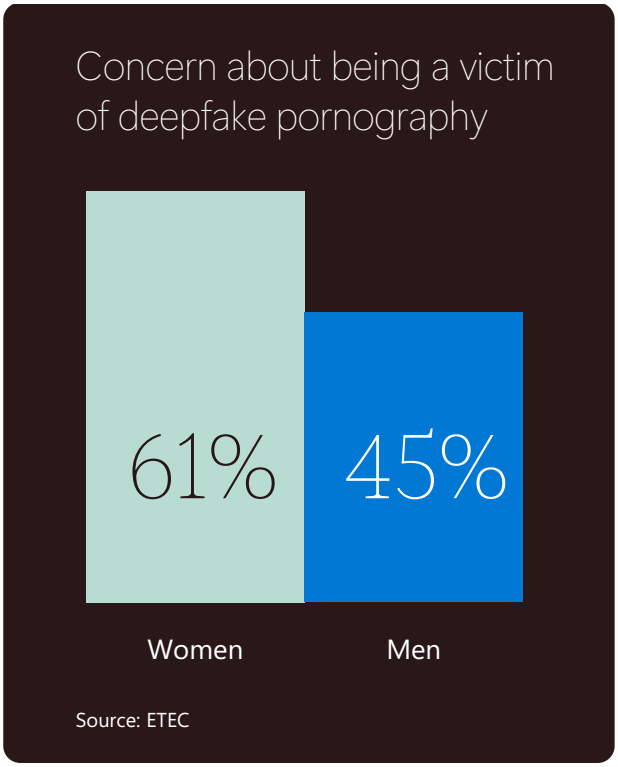
Synthetic non-consensual intimate imagery (NCII) is not a new risk—but it is one that is vastly exacerbated by generative AI. In 2019, even before the advent of generative AI, a report by Sensity AI found that 96% of deepfakes were non-consensual, and of those, 99% were made of women.

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 children.

Whether real or synthetic, the release (or threat to release) of intimate 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 and girls most often targeted, and facing consequences ranging from fear and pain to long-lasting reputational damage.

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.

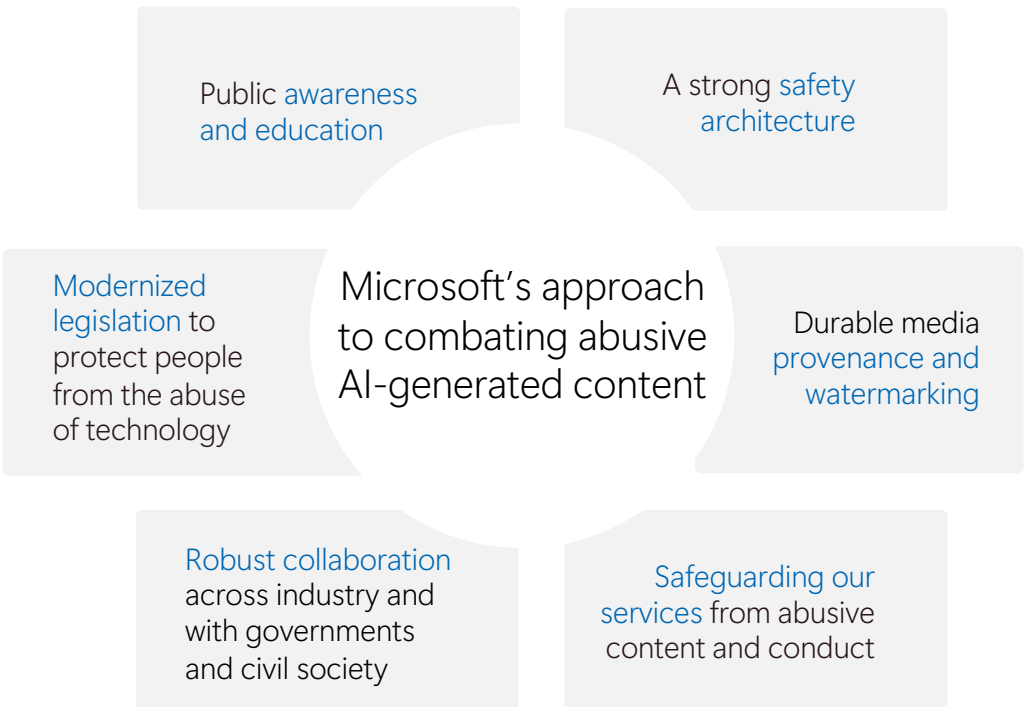


Part II: Microsoft's approach to combating abusive AI-generated content

Throughout the United Kingdom, 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 harm 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—modernised policy and legislation to protect people from the abuse of technology—Microsoft recognises 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

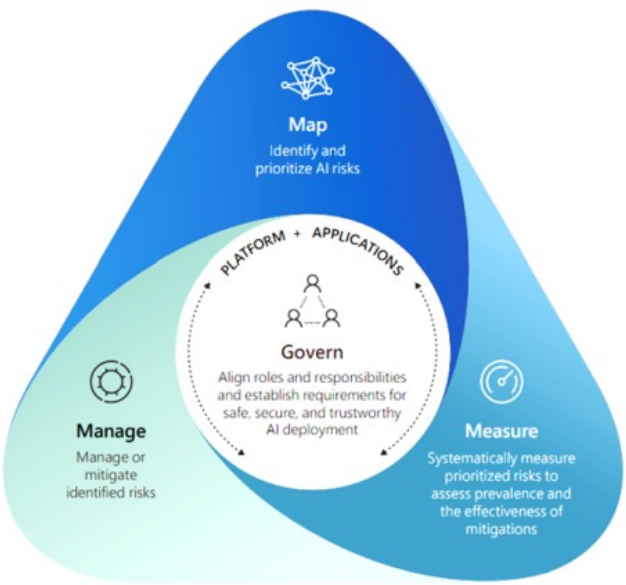
It should include aspects such as ongoing red team analysis, pre-emptive classifiers, the blocking of abusive prompts, automated testing, and rapid bans of users who abuse the system. At Microsoft, we understand that this is a multi-faceted process and that it is also iterative. Part of our safety architecture 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 organisation 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.

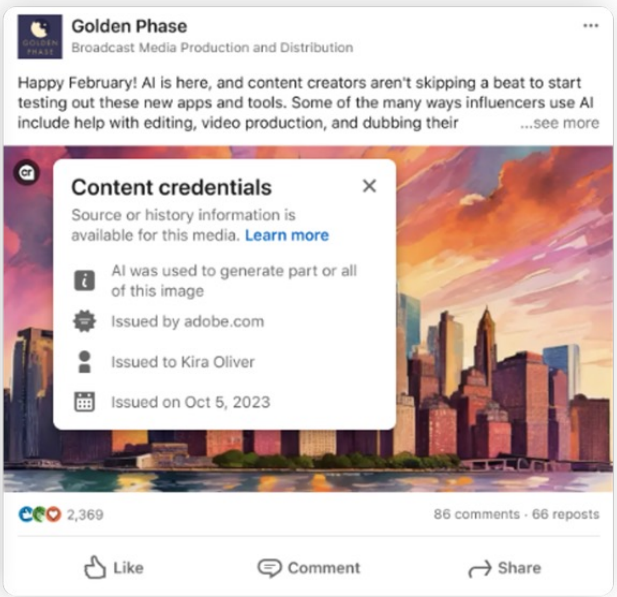
With industry partners, Microsoft has led significant progress in advancing disclosure methods to help consumers understand whether digital content was created or edited with AI.

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 organisation 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 labelled on the platform. Starting with content on the LinkedIn Feed, users can click on an icon in the upper left corner, which then reveals source/ history information, including whether the material was generated in whole or in part by AI:



LinkedIn is currently working to expand 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, whether real or synthetic, is also critical to reduce the potential for harm

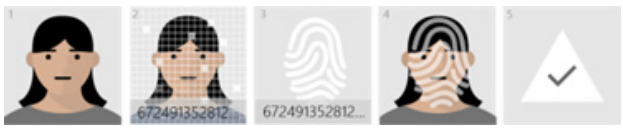
At Microsoft, we have long recognised our responsibility to keep our users safe, especially young people, and to contribute to building a safer online ecosystem. To achieve that, we take steps to protect our users from illegal and harmful online content, while respecting critical human rights such as 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 updated its policies to prohibit the sharing of software tools that are designed for, encourage, promote, support, or suggest in any way the use of synthetic or manipulated media for the creation of non-consensual intimate imagery.

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 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 synthetic non-consensual intimate imagery, including synthetic imagery that has been reported. We have recently announced that we are partnering with StopNCII to pilot efforts to detect and remove this victim-reported imagery from Bing's image search index: a step we believe will make a significant difference to reduce the availability of NCII across the ecosystem, in addition to addressing reports we receive directly from victims. To mark the 30th anniversary of the US Violence Against Women Act, we also made new voluntary commitments to address image-based sexual abuse, including as an AI model developer.

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+PDQF, 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 analyse 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 or your Amazon account details. It can then recognise 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 and recently committed to a new set of voluntary principles to address image-based sexual abuse.

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 recognise that addressing the potential acceleration of harm in the AI era will require new collaborative measures. To that end, we have joined 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 organisations, 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. Ahead of the UK General Election we created a reporting portal for deceptive AI election content and are continuing to roll out more services and announcements across Europe.

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 UK General Election in July 2024, Microsoft organised briefings with government agencies, 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 UK and the rest of Europe. 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.

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.

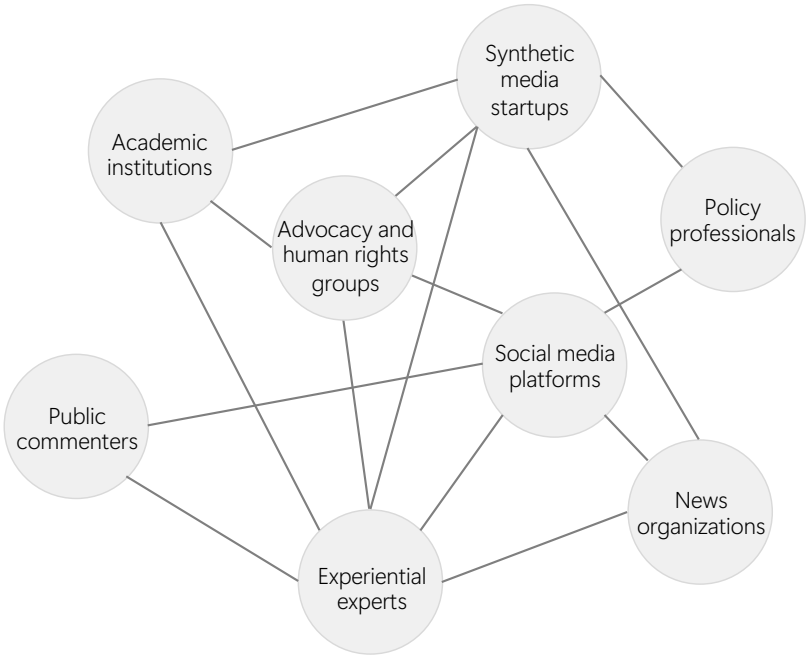
As a co-founder of C2PA, Microsoft has been involved in the public awareness and education work that C2PA has been conducting through public events and with policymakers about the importance of provenance. And, since its inception, we have been a part of the Partnership on AI's AI & Media Integrity Steering 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 recognise there is more work to do and look forward to playing an important role in it.

Finally, we also recognise 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. Similarly, we recently launched a new Minecraft Education feature called AI Foundations which is designed to help students, educators and families understand and use AI tools responsibly.

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 Kingdom 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:

Protect content authenticity	Detect and respond to abusive deepfakes	Promote public awareness and education
• Build trust in the digital ecosystem by promoting disclosures and requiring state-of-the-art provenance tooling to label synthetic content	• Combat deceptive AI deepfakes in our elections • Strengthen laws to protect women and children from online exploitation. • Strengthen the regulatory and criminal sentencing approach for AI-enabled fraud • Form new public-private partnerships to investigate cases and provide support to organisations that help victims of abusive AI-generated content.	• Enhance the NCSC's role in developing and disseminating best practices for synthetic content • Support and enhance public education campaigns on AI and synthetic content

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

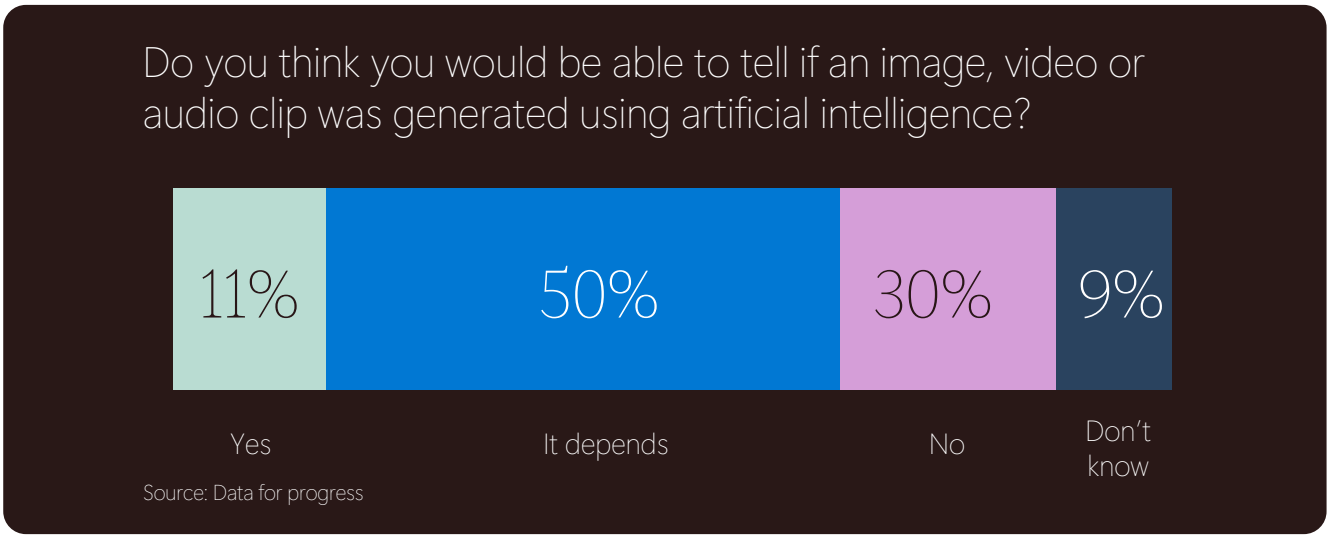
This white paper is our attempt to put forward our legislative and policy ideas to address abusive AI-generated content risks. We look forward to receiving feedback and continuing to work with civil society, policymakers, and stakeholders across the tech sector and beyond on effective policy measures.

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, which are soon likely to be broadly accessible. 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.



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 recognise 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, the UK government is taking steps and thinking about how to address this complex challenge.
At an official level, the UK government is already undertaking work to understand the role of provenance information in the Department of Science, Innovation and Technology (DSIT), Cabinet Office and National Cyber Security Centre (NCSC). Ofcom has also published discussion papers exploring strategies to combat the misuse of synthetic media, including the use of watermarking and content provenance tools. These are important first steps, but further work is needed to address the unique challenges posed by AI-generated media.

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.

There are a number of forthcoming pieces of legislation which could provide an opportunity to insert this requirement, such as the forthcoming Digital Information and Smart Data Bill. Doing so could help to simplify disclosures to users and increase broader public awareness. Working collaboratively with industry to pass 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 labelling.

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. With limited information currently available to central government on the use of provenance metadata on the authentic images, audio, and video they distribute, the Cabinet Office or NCSC has a role to play in issuing guidance on labelling and authenticating media content that they produce or publish. This would help people identify authoritative government outputs as authentic.

The UK could look to the example of the White House Executive Order issued in October 2023 which tasked the Office of Management and Budget with issuing guidance to agencies for labelling 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.

To further mitigate the risks that content is misused for deception, impersonation, and fraud, the UK 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 whitepaper, 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 UK government should ensure that NCSC and the AI Safety Institute (AISI) prioritise work to build out further authenticity and provenance techniques. This work should be done with AI Safety Institutes in likeminded countries, including the work at the National Institute of Standards and Technology (NIST), 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 utilise state-of-the-art provenance tooling so people can understand whether a piece of content is AI-generated or manipulated. This requirement could be incorporated into the forthcoming Digital Information and Smart Data Bill which provides an opportunity to establish clear standards for AI content provenance, ensuring that as these technologies evolve, there are robust mechanisms in place to maintain transparency and trust in digital information.

Alongside this provider-focused requirement, and to reinforce the value of synthetic content labelling, 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, legislation should require system providers to use state-of-the-art provenance tooling to label synthetic content. The Digital Information and Smart Data Bill or AI Bill could be opportunities to take this forward.

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 organisation 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.

We support legislation to establish penalties for bad actors working to intentionally remove, strip or tamper with authenticity or provenance metadata of AI content. This would be a common-sense measure to protect responsible AI efforts and hold bad actors accountable.

It will also be important to implement stronger controls for the subset of generative AI content that will pose the highest degree of risk. While carrying provenance information will be an important baseline mitigation for all synthetic content, more controls are appropriate for advanced deepfake capabilities on the horizon that pose a heightened risk of deceptive impersonation (i.e., for fraud.)

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.

The Government should progress plans to strengthen the UK's legal framework to protect children and women from online exploitation.

The UK's Online Safety Act 2023 (OSA) provides a comprehensive and proportionate framework to address harmful online content for UK internet users, especially children. Microsoft has welcomed the passage of the OSA and the thoughtful, evidence-based approach Ofcom is taking to its implementation. The OSA will play a key role in combatting a range of harms, including tackling child sexual abuse material, whether synthetic or otherwise, on user-to-user services and through search engines. As a priority offence, services in the scope of the OSA will be required to take a range of measures to prevent the risk that their services are misused for child sexual exploitation and abuse.

Child sexual exploitation and abuse imagery is near-universally criminalised, 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 analysed 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.

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 fuelling 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 minimise 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.

Modernise existing CSAM laws

In the UK, the creation and distribution of synthetic CSAM is already illegal under existing legislation. The Protection of Children Act 1978 (as amended by the Criminal Justice and Public Order Act 1994) criminalises the taking, distribution, and possession of "indecent photographs or pseudo-photographs of a child." This definition has been interpreted to include AI-generated images that appear to be photographs. Additionally, the Coroners and Justice Act 2009 criminalises the possession of "prohibited images of a child," which includes non-photographic depictions such as computer-generated images. The OSA further reinforced these protections by introducing new criminal offences, including those related to encouraging or assisting intimate image abuse.

While these laws provide a strong foundation, experience shows that emerging technologies will be abused by bad actors in novel ways. There is more that can be done to ensure the legal framework remains robust and effective in the face of advancing AI capabilities.

Establish an expert taskforce to study AI-enabled child exploitation

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

This taskforce should build upon the valuable work already undertaken by organisations such as the Internet Watch Foundation (IWF) and incorporate expertise from both the public and private sectors. It should include representatives from law enforcement, child protection agencies, technology companies, academic institutions, and relevant government departments.

The taskforce's mandate could include conducting in-depth research on current and potential future manifestations of AI-enabled child exploitation; evaluating the effectiveness of existing legal and technological measures in preventing synthetic CSAM; proposing solutions that leverage AI for detection and prevention of child exploitation; and recommending policy and legislative updates to emerging challenges.

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

Advance legislative measures to ensure efforts to develop and disseminate synthetic and other non-consensual intimate imagery is appropriately criminalised

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. Multiple studies have shown that 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 welcome the government's early action and manifesto commitment to ban the creation of sexually explicit deepfakes. A recent amendment to the OSA addresses this emerging risk by ensuring that sharing intimate images without consent is classified as a 'priority offence' under the new UK digital safety regime. This classification applies to sharing any photograph or film that shows, or appears to show, a person in an intimate state without their consent. While not explicitly mentioning synthetic media, the broad language encompasses AI-generated or manipulated imagery. We welcome this development and look forward to working with Ofcom on proportionate mitigation measures as the regulator finalizes the illegal content codes of practice.

Beyond efforts to help ensure online services are taking appropriate steps to address identified intimate imagery risks, we also recommend measures to close existing gaps in the criminal law related to the non-consensual distribution of any intimate imagery.

There have already been parliamentary efforts to address this challenge. Most notably, the Criminal Justice Bill, which was progressing through Parliament before the General Election, included proposals to strengthen the legal framework around deepfakes. This would have criminalised the creation of sexually explicit deepfakes, even if they were not shared. Since the election, there have been renewed efforts to put this offence into statute, with the new Government standing on a manifesto pledge to introduce legislation targeting this form of digital abuse.

However, given the rapid evolution of AI capabilities, we recommend ongoing assessment to ensure the law effectively addresses emerging challenges in this area and provides an effective deterrent to malicious actors.

Wider efforts to strengthen the legal framework include the proposals in Baroness Owen of Alderley Edge's Non-Consensual Sexually Explicit Images and Videos (Offences) Bill, which aims to criminalise the creation of sexually explicit deepfakes, even if not shared.

We encourage policymakers to continue to refine these proposals, ensuring they are comprehensive, victim-centred, and adaptable to evolving technologies. We also urge law enforcement to bring cases where possible under these strengthened laws, to establish precedent and send a clear deterrent message to potential offenders.

Strengthen measures against deepfake fraud through Ofcom's OSA implementation and new legislation

As generative AI technologies evolve, the UK faces increasing challenges in discerning genuine content from deceptive schemes. Law enforcement officials and industry leaders recognise that we are at a critical juncture concerning the criminal use of AI and synthetic media. Synthetic content provides cybercriminals with the capability to enhance and scale existing fraud schemes while enabling new forms of deception.

Financial fraud scams have been growing exponentially in recent years, even before the widespread adoption of AI, overwhelming police and prosecutors. Online and telephone scams are particularly prevalent, with older adults often targeted due to their perceived vulnerability and accumulated wealth. To address these emerging threats, we recommend a two-step approach that leverages existing regulatory frameworks while introducing new legislative measures to combat deepfake fraud effectively.

Prioritise deepfake fraud in Ofcom's OSA implementation.

The UK has already taken significant steps to address fraud through including it as a priority offence in the OSA, meaning in-scope services will be required to take a range of measures to address content relating to fraud. To support in-scope services in understanding the evolving risk that AI technologies are used to generate fraudulent content or perpetrate AI-enabled fraud, Ofcom may wish to conduct additional research on this point and develop specific guidelines on the topic. Ofcom may also wish to prioritise engagement on fraud risks, including deepfake fraud risks, as a part of its supervision regime.

The total harm of AI use to commit fraud should be considered a potential aggravating factor in sentencing

Under UK law, fraud carried out using AI should already be captured under the Fraud Act 2006. However, given the potential for AI to increase significantly the risks to society and cause a collective lack of trust, sentencing guidelines for fraud cases should encompass the total harm caused as a result of the crime, rather than only the harm to an individual victim. These wider systemic impacts of the use of synthetic content to commit a crime should therefore be considered an aggravating factor in sentencing, serving as a deterrent and reflecting the potentially severe consequences of deepfake fraud.

Form new public-private partnerships to investigate cases and provide more funding opportunities for organisations 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 organisations, 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. The UK government should look to model public-private partnerships on this template to support collective efforts to combat abusive AI-generated content in the UK.

Victims of synthetic non-consensual intimate imagery may have concerns about reporting to law enforcement agencies, who may not be appropriately resourced to address this accelerating category of harm. The 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 organisations such as the Law Commission and industry organisations to drive forward these efforts.

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.

Enhance the NCSC's role in developing and disseminating best practices for synthetic content.

As AI capabilities continue to advance, it will be crucial for the UK government to regularly update best practices and standards to help the public navigate synthetic content. We recommend that the NCSC takes a leading role in this effort, building on its existing work in AI security.

The NCSC, in collaboration with industry partners and academic institutions, should expand its efforts in assessing best practices for synthetic content labelling, verification, and detection. It is vital to update these best practices annually to keep pace with the increasing sophistication and complexity of synthetic content, advancements in labelling and detection tools, evolving adversarial attacks on provenance systems, and growing public awareness of these issues.

The NCSC should establish a dedicated program to study synthetic content harms, leveraging its existing partnerships with industry and academia. This program would 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. The scope should extend beyond direct harms to include developing core methods, designs, and signals for consumers, as well as assessing any harms resulting from loss of trust in authentic content.

As part of this initiative, the NCSC should evaluate the effectiveness of tools for labelling and detecting synthetic content and displaying provenance information. This assessment should include sociotechnical analyses of how these tools are used and perceived in practice. The insights gained from this work would inform ongoing public education campaigns and help refine best practices for synthetic content disclosure.

Collaboration with the AISI would ensure alignment with broader AI safety initiatives, creating a comprehensive approach to synthetic content challenges. By publishing and annually updating comprehensive guidance on best practices for managing synthetic content risks, the NCSC can provide invaluable resources to both industry and the public.

Support and enhance public education campaigns on AI and synthetic content.

The government is uniquely positioned to deliver tailored education campaigns to the public around safety and harms, much as it does for other critical issues. The Department for Education (DfE) and DSI should use existing funding programs and create new programming to help educate the public.

These campaigns should educate the public about the deceptive uses of synthetic content, the associated safety risks and harms, and provide approaches for discerning authentic digital content.

This includes teaching people how to assess whether content was authentically captured or AI-generated, identifying trusted sources, and recognising the latest scams employing synthetic content. The campaigns should target vulnerable demographics, such as older adults and young people, who may be particularly susceptible to AI-enabled deception.

Building on proposals in the U.S., we support the creation of a National AI Literacy Campaign. This could leverage the government's digital skills framework and invest in both formal educational structures and informal learning opportunities to advance AI literacy across the UK.

This could build upon valuable work already undertaken by public broadcasters. For example, the BBC's Verify project could serve as a foundation for broader efforts to educate the public about content authenticity and provenance tooling.

We recommend that any education campaign incorporates input from civil society groups and is disseminated in coordination with trusted local community organisations. Beyond achieving broad public awareness, these campaigns should specifically target frontline actors such as local media, journalists, community leaders, and civil liberties groups who will need to assess potential deepfakes and educate others.

In areas of civic importance, such as election integrity, we recommend that the Electoral Commission, in collaboration with the DfE and DSIT, develop targeted education campaigns. These could include information about content provenance tools and how to distinguish official election-related content from potentially misleading synthetic media.

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

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/SyntheticMediaUK