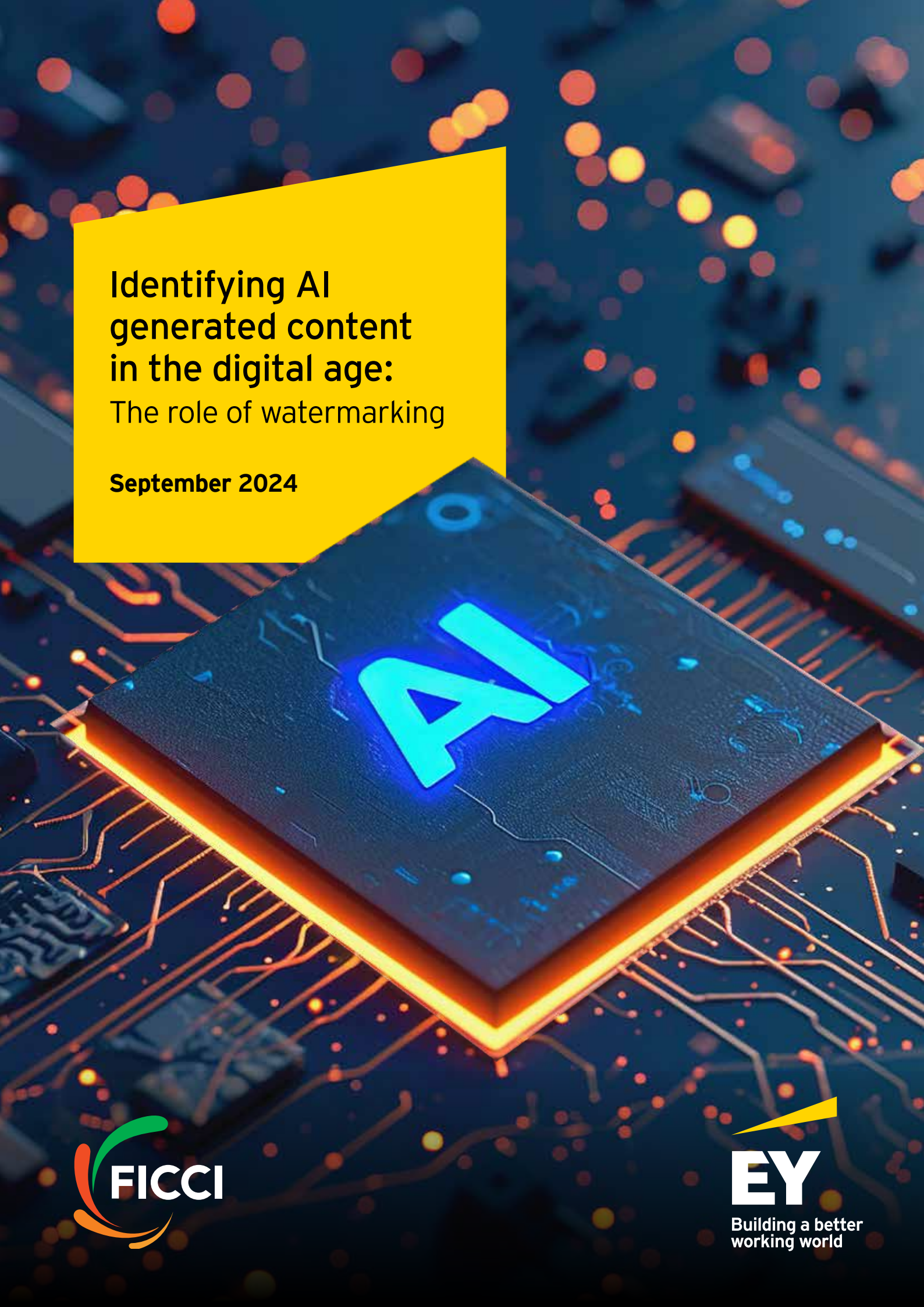


Identifying AI generated content in the digital age:

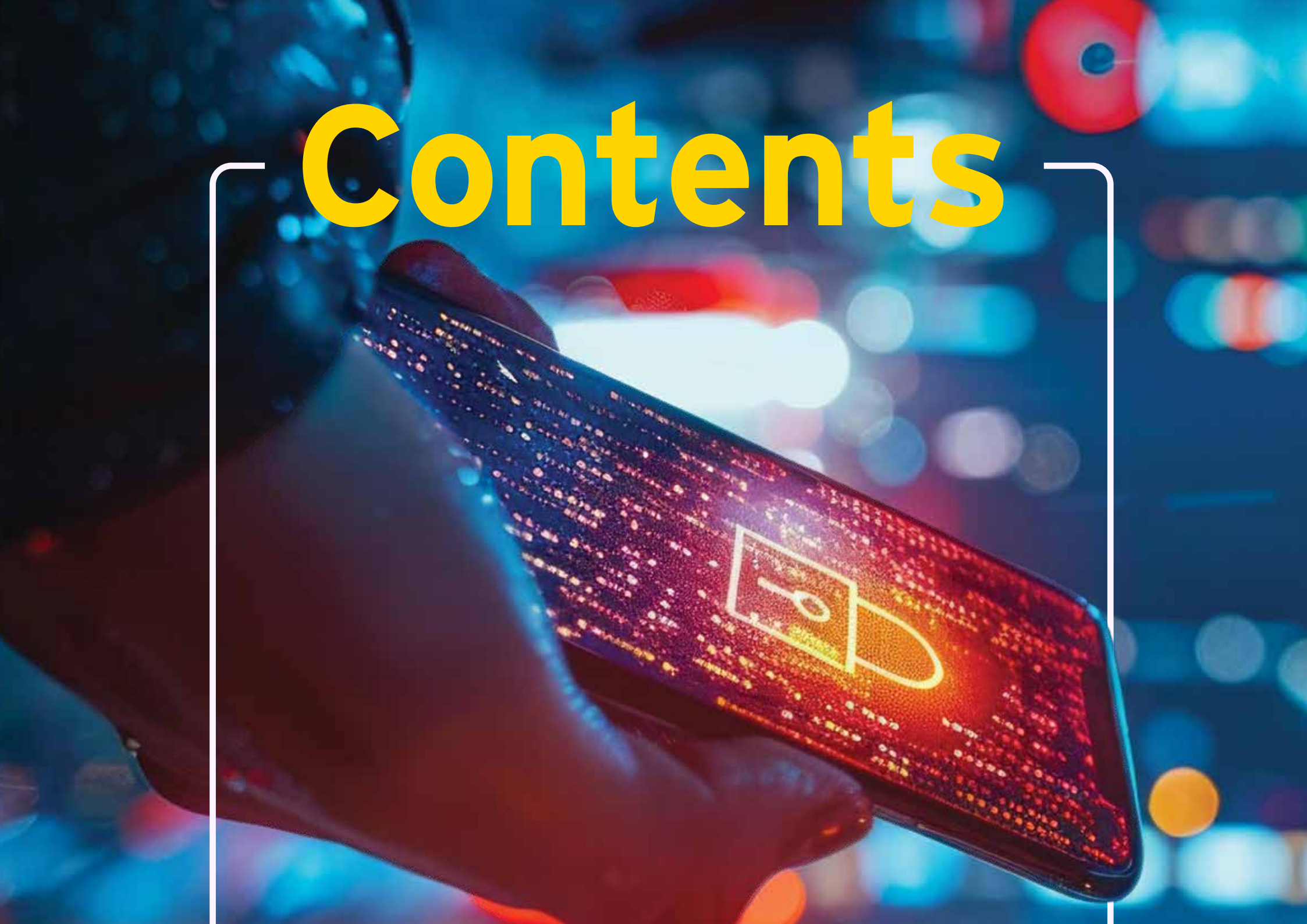
The role of watermarking

September 2024

FICCI

EY
Building a better working world

Contents





04 Foreword

08 Executive summary

12 What is AI content detection and why it matters?

18 Content detection approaches

22 Policy scenario for AI content detection mechanisms

28 Way forward

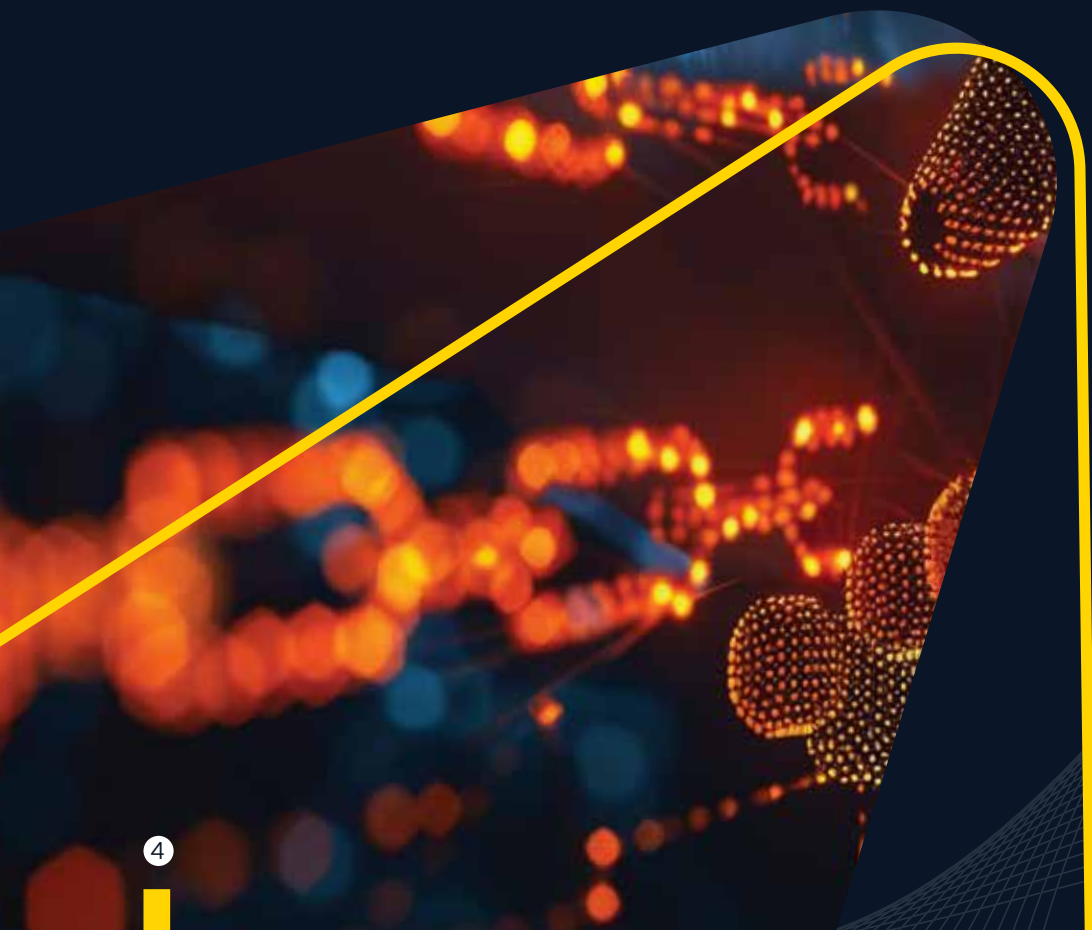
34 Annexure: Technical overview of watermarking

EY Foreword

Generative AI models have introduced a new paradigm of content generation, requiring only a few prompts and clicks. As a result, overall quantum of content generation has far exceeded the rate at which this content can be identified or traced. This makes users increasingly unaware of the source of the content, whether it is based on real events or individuals, or if it is entirely computer-generated. In the context of the private sector entities, falsely attributed content may potentially impact business interests as well as the credibility of an organization. In addition, there are new social issues such as deepfakes, fake news, misinformation, disinformation, and other forms of deceptive content that complicate the landscape of digital media. Having a way of establishing some form of identification could, therefore, address emerging concerns and help foster trust in the Artificial Intelligence ecosystem globally.

Some challenges emerging today are being mitigated through government and private sector initiatives. The governments in countries such as the US, the EU and China are taking measures that would eventually require technology companies to disclose to users whenever content is AI-generated. The executive orders issued by the government of the US and the European Union Artificial Intelligence Act (EU AI Act) are examples of such government-led initiatives. On the other hand, private sector initiatives include development of niche content detection mechanisms. Big tech players have been vocal about the importance of the role of governments in fostering trust and promoting user safety through appropriate policy and governance measures.

From a technology standpoint, various approaches are currently being considered to address this issue, including metadata verification, retrieval-based detectors, post-hoc detectors and watermarking. In announcements made by tech-giants at the Bletchley Summit in November 2023, they indicated that a combination of different AI content detection technologies would be required for effective AI content detection.





Watermarking is one of the solutions that allows for AI content detection. It gives model developers the control over the embedding of the watermarks. In an ideal scenario, effective measures around watermarking would need to be robust to prevent erasure, removal and editing. Further, the detection interface would also require a low false rate and be interoperable across the different GenAI systems being developed by the technology companies.

This report presents a high-level analysis of AI content detection and related efforts and contextualizes suggestions or measures that are relevant to India. We believe that these measures will be useful in fostering trust in the AI ecosystem in India and facilitating promotion of safe AI adoption and uptake across different sectors and industries.

Happy reading!



Rajnish Gupta
Partner, Tax and Economic Policy Group,
EY India

FICCI Foreword

The advent of generative AI models has ushered in an era of unprecedented content creation, transforming the way we conceive, produce, and disseminate information. However, with this remarkable innovation comes a pressing challenge: the need to discern the origins and authenticity of AI-generated content.

The digital landscape has become vulnerable and these challenges not only undermine public trust but also pose significant risks to the integrity of businesses, the sanctity of intellectual property, and our national security. It is, therefore, imperative that we as a country should come together, and establish robust mechanisms for tracing the provenance of digital content, thereby ensuring its authenticity and safeguarding the interests of content creators and consumers alike.

Among several other techniques, AI watermarking is one possible solution in this context. It offers a sophisticated solution that embeds indelible markers into AI-generated media, serving as a digital signature that attests to the content's origin and integrity. This report delves into the nuances of AI content detection and watermarking, highlighting its role in fostering a trustworthy AI ecosystem. It also examines the initiatives taken by governments and private sectors across the globe, including steps taken so far by our nation, to address the challenges posed by AI-generated content.

As we navigate through the complexities of the digital age, it is crucial for industry leaders, policymakers, and technology innovators to collaborate and steer the course towards a future where AI serves the greater good as a tool of socio-economic development.

Let us embrace this opportunity to lead with vision and responsibility, setting a global benchmark for the ethical use of AI. I do hope you would find the report informative and our collective efforts bear fruit for a more secure and trustworthy digital future.



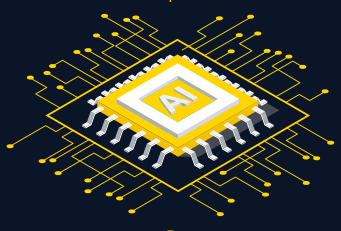
Jyoti Vij
Director General
FICCI

As Artificial Intelligence tools are used to generate digital content, there are emergent concerns relating to the source, authenticity, and accuracy of such content. To retain public trust, the labelling of content and the ability to detect whether AI tools have been used to generate content, will become increasingly relevant. It is important that all stakeholders understand AI content detection technologies, such as watermarking, so that they may start thinking about how it can be approached and incorporated in the respective AI workflows or use cases.



Mr. Abhishek Singh,
Additional Secretary, MeitY

Executive summary



Today, Artificial Intelligence (AI) capabilities are generating text and images that are often indistinguishable from those created by humans. As AI continues to increase in sophistication, it has also raised concerns about the source and authenticity of content that is available and in use on the internet, social media platforms, and other digital spaces.

AI-generated content and authenticity

The extensive use of AI-generated output and digital content as the basis or source of information often leads to humans forming opinions and making decisions based on it. Therefore, it becomes important for the consumers of such information to recognize the difference between AI-generated and human generated content, whether it's in the form of text, images, videos or audio.

Authenticity of content is important for many reasons:

- In many areas, such as education, legal proceedings, medical diagnosis and financial services, authenticity is important because decisions are based on specific knowledge about the source of inputs.
- Limiting disinformation (manipulated or biased information with the intent to mislead) and misinformation (false information spread irrespective of intent), is especially important in the context of media, areas of public interest such as law and order, and national security.
- As the digital economy develops further and leads to an increase in the number of online transactions, measures to address deceptions and financial frauds have become important.
- Ensuring protection of intellectual property.
- Ensuring that when training future AI models, one can distinguish whether the data used for training is synthetic or not.

Until now, public discourse has largely centered on issues such as spread of deep fakes and impersonation, fake news and social manipulation, false attribution, lack of AI content detection mechanisms and copyrights. However, the issue is much wider. Organizations are increasingly deploying AI to develop new products and services and to improve the efficiency of existing systems. Clarity on the source of inputs used, therefore, becomes important.

Watermarking allows identification of AI generated content

One approach to identify AI generated content is watermarking, a technology designed to embed unobtrusive identifiers into AI-generated media, to authenticate the origins and integrity of the content.

There are many precedents of the usage of digital watermarks. The roots of watermarking may be traced back to the 1950's where identifiers were first embedded in music, establishing the foundation for merging technology with intellectual property management. Recent applications include watermarking

images in healthcare by organizations. Before the advent of AI, watermarking allowed transmission and storage of medical records. With the emergence of AI and its capabilities to produce high-quality data, several new challenges have emerged for various companies. These issues can be addressed through AI content detection. Among the various AI content detection approaches that have emerged, watermarking is often discussed. AI watermarking deals with identifiers that can be later used to authenticate the origins and integrity of the content. AI watermarks can be textual or visual but are usually invisible to the human eye and can be detected by algorithms only, as it is embedded in the model and output, even before the AI generates the specific content.

Governments around the world are also taking measures to address the identified challenges and recognize the promise/potential for leveraging watermarking technologies. The EU AI Act contains provisions that require users of AI systems in certain contexts to disclose and label their AI-generated content. The Act also includes provisions that require people to be informed of when interacting with AI systems.

An executive order issued by the US government on 30 October 2023 states that the government will take measures to "Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating official content. The Department of Commerce will develop guidance for content authentication and watermarking to clearly label AI-generated content. Federal agencies will use these tools to make it easy for Americans to know that the communications they receive from their government are authentic—and set an example for the private sector and governments around the world." The executive order has requested US tech companies to develop models for voluntary commitments from major AI companies to develop "robust technical mechanisms to ensure that users know when content is AI generated," such as watermarking.[1]

In addition to the efforts of the government, technology giants with a large footprint in the AI ecosystem made announcements pertaining to their approach to "Identifiers of AI-generated material". The announcements emphasize the urgent need for a global consensus and multi-stakeholder led approach to standards, practices, and technologies integral to AI content detection.

The Indian context

An advisory issued by the Government of India on 15 March 2024 states, "Where any intermediary through its software or any other computer resource permits or facilitates synthetic creation, generation or modification of a text, audio, visual or audiovisual information, in such a manner that such information may be used potentially as misinformation or deepfake, it is advised that such information created, generated, or modified through its software or any other computer resource is labeled or embedded permanent unique metadata or identifier, in a manner that such label, can be used to identify that such information has been created, generated or the computer resource of the intermediary. Further, in case any changes are made by a user, the metadata should be so configured to enable identification of such user resource that has effected such change."[2] While this advisory is limited to intermediaries, it is likely that this may further evolve in the future.

Way forward

As AI and watermarking technologies evolve, the developments will require careful attention, so that players investing in AI products and services take the right approach.

---

1. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

2. Advisory 15March 2024.pdf (meity.gov.in)

Executive summary

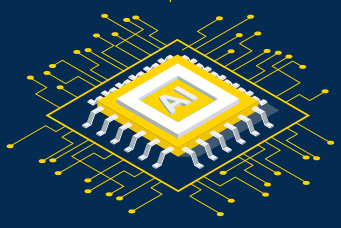
Keeping in view the complexities and challenges, the government could consider the following:

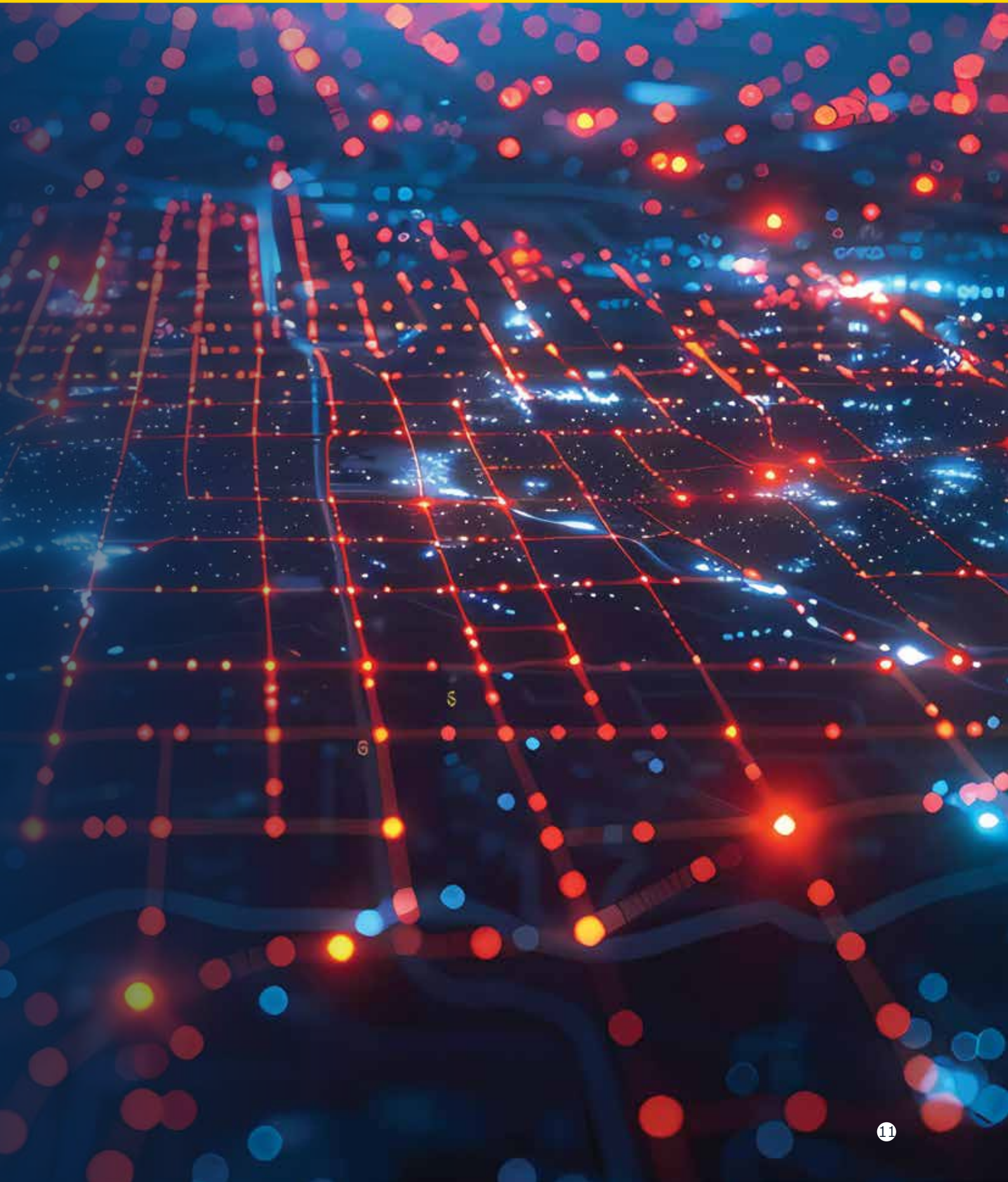
- Measures to promote the adoption of AI content detection mechanisms
- Work towards establishing standards for AI Watermarking, in line with relevant international standards, which may be reviewed/updated periodically
- Develop/promote open-source tools and digital infrastructure to support watermarking
- Promote implementation of watermark usage in the government
- Create awareness about AI content detection and its use cases

For the development of the ecosystem, the private sector (enterprises using AI generated content) may consider the following:

- Evaluate AI usage and content generation at an organizational level
- Carry out rigorous internal testing of AI watermarking systems
- Keep abreast with the technology changes and standards which are notified domestically as well as globally
- Invest in implementation

Implementing watermarking solutions will be the key to ensuring trust in the future digital economy and the broader AI landscape. Through deliberation, innovation and cooperation, India can play a pivotal role in shaping a robust domestic digital ecosystem that is secure, transparent and trustworthy.







1

What is AI content detection and why it matters?

As AI grows, it is doing more than just recognizing patterns – it is transforming how we work and creating new opportunities. But as it becomes more widespread, there could be some concerning situations arising from the malicious usage of AI. In some cases, it may even lead to implications for law and order, national security and peoples' privacy.

In the recent past, we have heard of cases where individuals used AI to fake voices, scam people, spread lies on social media, and create fake videos to attack public figures. So, while AI brings many benefits, we also need to watch out for the problems it can cause.

Such issues can further lead to large-scale implications which could hamper the development of the AI ecosystem due to a lack of trust, thus creating ethical dilemmas and presenting a host of other challenges. According to a report, concerns around Data Privacy (57%) and Trust and Transparency (43%) have significantly inhibited the adoption of generative AI technologies across enterprises globally.[3] Furthermore, the industry believes that incorporating transparency, trust and security in AI models will have far-reaching economic and business impact.

Some of the implications of not identifying AI generated content include:

Ambiguity of identity and plagiarism

- Without watermarks, it becomes challenging to attribute AI generated content to its source or system of generation or modification.
- Unwatermarked AI content is susceptible to plagiarism. Others can claim it as their own, eroding the integrity of creative works since verification is not possible without watermarks/ identifiers.
- Deepfake videos of individuals or public figures may be used by bad actors for nefarious purposes.

Misinformation leading to flawed models and poor decision making

- AI generated content may have large-scale implications for workplaces or in education as students may use it to complete assignments, write code, etc.
- Sensitive decisions, if based on information, the provenance or the source of which cannot be verified, may have damaging consequences for business decision makers.

Legal challenges

- In legal disputes, proving ownership of unwatermarked GenAI content could be challenging.
- Watermarking and AI content detection mechanisms may be a supporting mechanism for the eventual establishment of liability in the context of AI.
- The unauthorized use of proprietary data or content in a training dataset may lead to copyright concerns.

Trust erosion

- Consumers could lose trust in digital media when they encounter unattributed AI content.
- AI inputs from unverifiable sources may impact businesses adversely.

AI content detection is the process of using identifiers (visual or embedded) to identify and analyze if content has been created or modified using artificial intelligence models such as LLMs. The applications of AI content detection are wide-ranging, from combating misinformation and proliferation of deepfakes to protecting intellectual property rights.

3. IBM Global AI Adoption Index 2022

Implications of neglecting watermarking in AI-generated content

AI content detection tools are critical for identifying AI generated content and maintaining its integrity in our increasingly digital world. AI advancements make it increasingly difficult to distinguish between human-created and machine-generated content, leading to potential misuse, such as misinformation campaigns, copyright infringement, and tarnishing the credibility of digital content.

Key issues to be addressed in AI-generated content



Proliferation of deepfakes and impersonation

- Increased AI capabilities have made the manipulation of images, videos and audio recordings easier which may be used for malicious purposes, such as damaging reputations, fabricating evidence, and undermining trust.

Copyright infringement

- Since the entire internet is a large training dataset for AI models, it may use proprietary data while generating content, potentially leading to copyright infringement conflicts.

Fake news and social manipulation

- The virality of AI-generated content on social media can make it nearly impossible to distinguish between credible and faulty information. The spread of misinformation can have far-reaching implications on internal security, national security and society law and order.
- Misinformation may result in emergency responses being misdirected and may also cost public unrest.

False attribution of content

- AI will soon be capable of generating any form of information with sophistication that distinguishing it from human-generated content will be difficult.
- As the adoption of AI increases, concerns around its impact on education, work, intellectual property, etc., will become more prevalent.

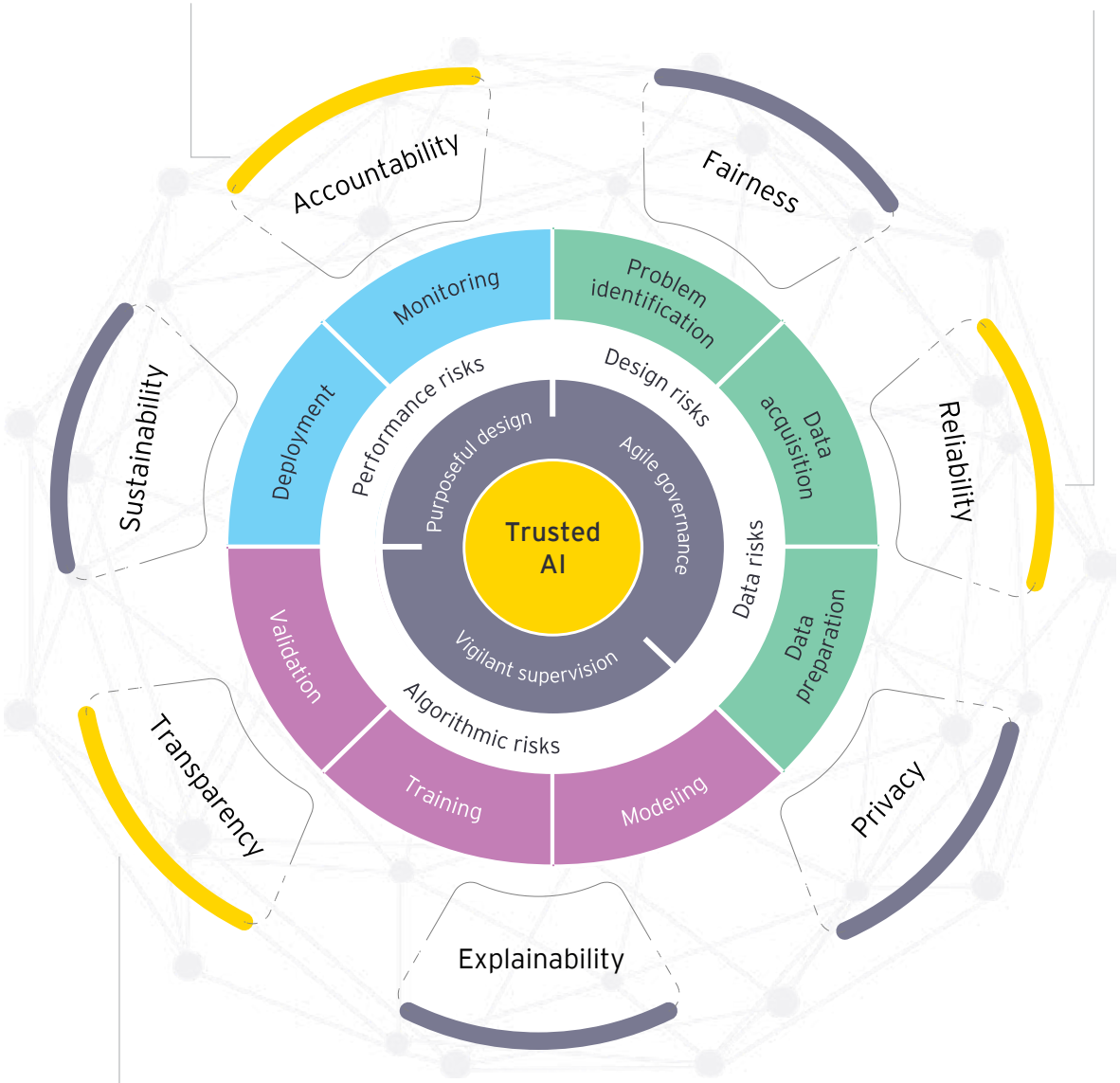
Building trust in AI through content detection mechanisms

AI content detection mechanisms become key in mitigating the risks and dangers posed by AI usage, both for the private sector and the government. It will promote safe uptake of AI across economies, industries and sectors.

AI content detection helps promote Responsible AI development as it contributes to accountability, reliability and transparency.

AI content detection promotes **accountability** by establishing the provenance of AI-generated content.

AI content detection promotes reliability by establishing the creator of digital assets.



AI content detection promotes **transparency** by providing the means to identify AI-generated content.

AI watermarking is less about restriction and more about enabling an ecosystem of accountability, transparency and reliability to foster trust in AI systems.

Source: EY Framework

Why should governments and the private sector consider content detection?

The development of AI content detection mechanisms is pertinent to the development and deployment of AI in a responsible manner. The detection mechanisms are intended to foster trust in the AI ecosystem and have several important advantages because of which it merits the consideration of all stakeholders.

01 Provenance and trust

As new AI-based tools and services emerge, it is expected that specific models become known for higher quality of outputs and accuracy. The entities developing and licensing these AI systems and services may want to establish tools that will allow users and third parties to verify whether something is AI-generated or not, thus establishing its provenance and reliability.

Establishing provenance of diagnosis in the healthcare industry:

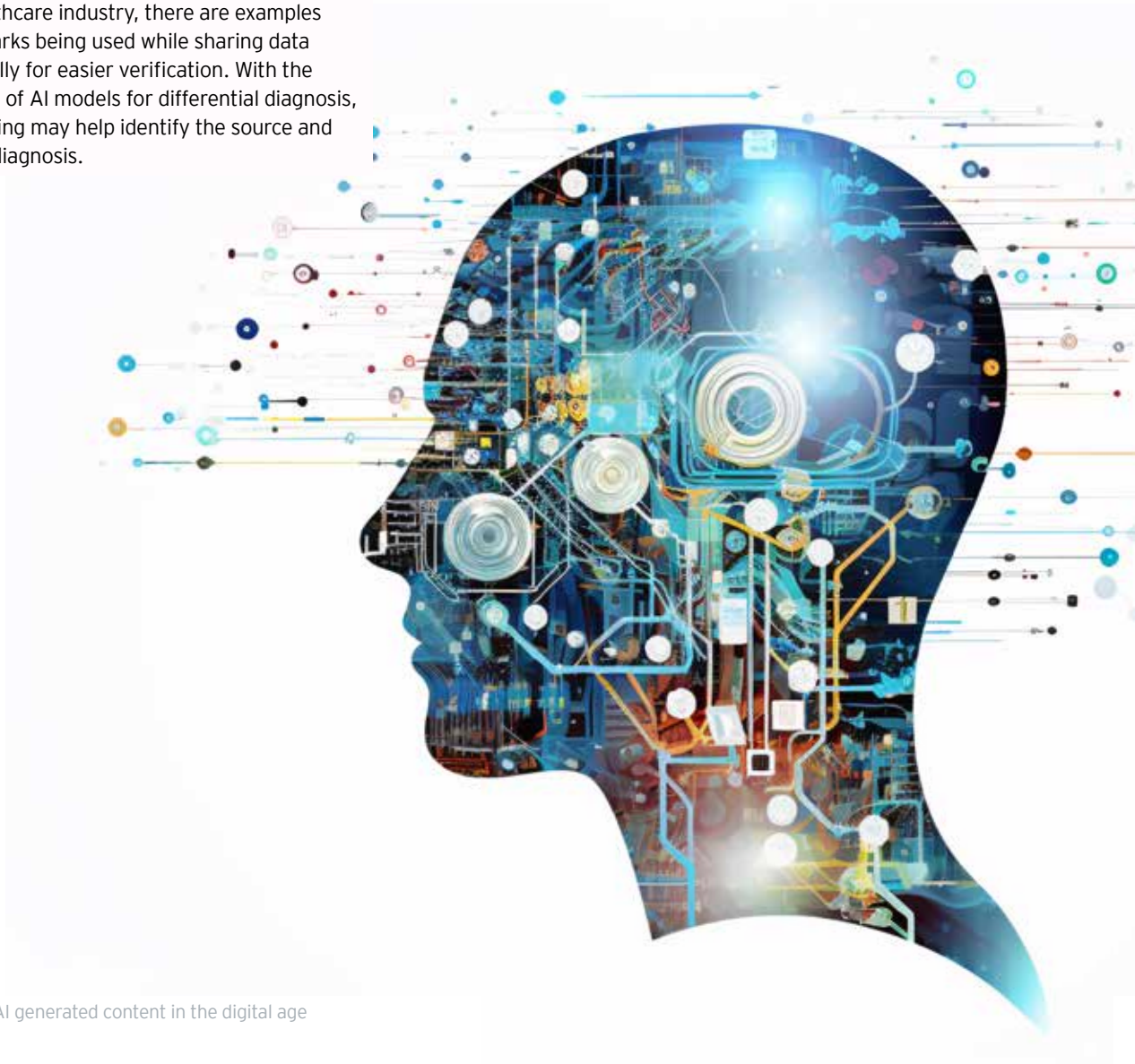
In the healthcare industry, there are examples of watermarks being used while sharing data electronically for easier verification. With the emergence of AI models for differential diagnosis, watermarking may help identify the source and verify the diagnosis.

02 Audit and review of decision making process

As numerous AI models get deployed by businesses for various purposes, it will be good to have identification mechanisms in place. In case something goes wrong and needs review, it could allow the entity to determine the source and the specific model or tool which needs improvement.

Establishing source and auditing decision making in the insurance industry:

In the future, insurance companies could use AI models trained on historical data to make decisions about claims or the insurance-worthiness of potential client/s. In case of a complaint, the insurance provider will be able to use watermarks in the output to identify/trace the specific model which was used to make the decision. This can help the company set up internal processes to address the underlying cause and will allow the insurer to modify or correct the model to avoid the same problem in the future.



03 Asserting ownership and protecting Intellectual Property (IP):

Digital content detection mechanisms may help in determining the origin and source of AI-generated content. It serves as a unique identifier that establishes content provenance.

Asserting ownership may also help businesses and corporates monetize their data. This will allow businesses to license, sell and protect their assets.

Copyrights and fairness in media:

Watermarks in photography can serve as a distinguishing feature to identify whether an image has been captured by an individual photographer or generated by artificial intelligence. Using watermarking in a photography competition would help screen photos taken by individuals and eliminate AI-generated images (especially if the use of AI is not adequately disclosed). There may also be open-ended questions on the copyrights or the fairness of use around AI-generated photos. It may not be clear whether the ownership of the generated media rests with the AI artist using the model, the AI model used to generate the content or the owners of the original images that were used to train the photo generating AI model. If AI-generated images were embedded with watermarks (visible or otherwise) it would prevent any unfair advantage or incorrect attribution.

04 Identifying AI-generated content

Watermarking helps distinguish between AI-generated and human-generated content. This is particularly important as generative AI models become increasingly sophisticated, making it harder to differentiate between human and AI-generated content.

Prevention of reputational risk and financial loss:

There are instances of fraudulent entities using deepfake videos of well-recognized public figures. Famous figures appear on video advertisements for malicious software which are then promoted en masse. Watermarking technologies could be helpful for users to recognize whether the content is AI generated and can help prevent potential losses (financial or otherwise) resulting from reputational damages or cybersecurity breaches. Proliferation of such technologies can also help high profile business entities and leaders mitigate reputational risks.

05 Combating misinformation

This will be crucial in reducing and protecting the menace of misinformation and deepfakes. For businesses, the spread of misinformation and disinformation may affect the revenue and topline of businesses.

Misinformation and fact checking:

The existing methods of AI detection often rely on manual evaluation of images, media or content. For example, for detection of deepfakes, there are certain issues or concerns with detecting generated content manually, through identifying inaccurate/unrealistic depiction of features on a frame by frame basis. With the widescale adoption of watermarking, some of these evaluation techniques can be automated, thereby reducing the time to detect deepfakes while also limiting potential damage. Instead of manual inspection, experts may either look out for visible watermarks or run content through a tool to determine its source/veracity.

Misuse of AI in the education sector:

In the education sector, there may be cases where students can use AI generated text for completing assignments, writing research papers, etc. This can potentially undermine learning outcomes as students can avoid assignments or pass off plagiarized AI generated content as their own. With the ease with which AI can generate content, this can become a crucial challenge that will need to be addressed by evaluating student assignments and submissions using AI content detection tools.

The case studies presented above highlight the need for robust AI Content detection mechanisms. These mechanisms will help promote adoption of AI systems by mitigating ethical, privacy and misuse concerns.



2

Content detection approaches

In traditional AI applications, AI systems primarily operated on a reactive basis, processing and responding to data within a well-defined framework. The scope of AI was confined to analysis, pattern recognition, and executing pre-programmed responses. The advent of GenAI and large language models (LLMs), however, marks a transformative leap in AI capabilities, expanding the purview from mere data interpretation to the creation of novel content in multiple formats vis-à-vis text, images, videos and audio. These models have also enhanced the ability and speed at which content is generated.

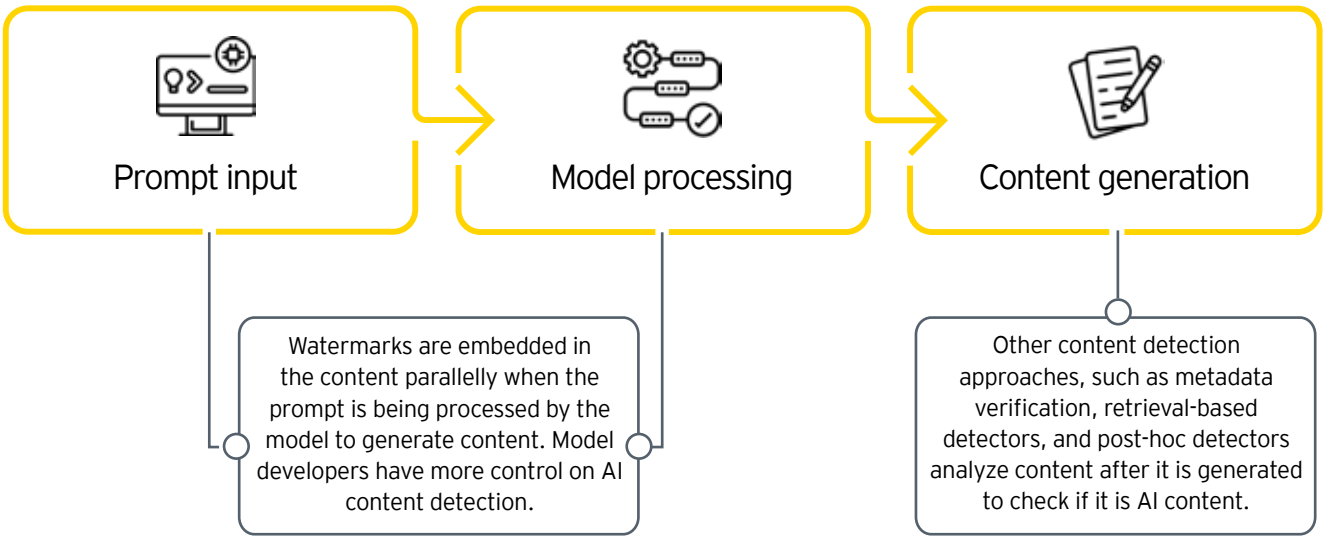
Considering the potential dangers of deepfakes, identity fraud and false campaigns, the need for robust AI content detection tools is therefore critical and urgent.

AI content generation process

AI has enabled an extremely high rate of content generation. However the tools to identify or flag AI-generated content are lagging, leaving room for malicious use of AI. Improved content detection will play a key role in addressing this challenge.

The different stages of AI content detection include prompt input, model processing, and content generation. These stages have been elaborated below and the relevant AI content detection processes are discussed below:

Basis process of AI content generation



AI content detection approaches

Different approaches cater to the different stages of AI content generation. They are:

Metadata verification

- **Definition:** In an AI context, metadata containing the source, origin and purpose of an AI generated content piece may be maintained in a shared database. Open technical standards have been suggested by the C2PA (Coalitions for Content provenance and authenticity) and IPTC for metadata storage.
- **Stage of deployment/identification:** After content generation
- **Challenge:** The database that must be maintained may violate users' privacy rights. Furthermore, it may become a challenge if the text is slightly humanized.

Retrieval-based detectors

- **Definition:** Retrieval based detectors are NLP (Natural Language Processor) detectors that can find content in a repository of content. For successful implementation, AI developers will have to keep a database of content generated.
- **Stage of deployment/identification:** After content generation
- **Challenge:** A database of AI generated content will have to be maintained, which is cost and time extensive. Despite this, the approach may be ineffective when the text is humanized.

Post-hoc detectors

- **Definition:** Post-hoc detectors retrospectively analyze machine-generated text. These detectors use heuristics such as entropy, scoring, perplexity, burstiness and other statistical methods.
- **Stage of deployment/identification:** After content generation
- **Challenge:** The scoring system may be ineffective if the text is humanized.

Watermarking

- **Definition:** Watermarking refers to embedding signals into generated content that are invisible to humans but are algorithmically detectable from a short span of tokens. Different approaches to watermarking technologies are presented in the Annexure.
- **Stage of deployment/identification:** Model Processing Stage
- **Challenge:** APIs will need to be developed and deployed for the identification of watermarks.

Among various content detection approaches, watermarking is often discussed as it can be controlled by the model developer and initiated before content generation. Though still an active area of research, AI watermarking can be incorporated into the AI model's outputs, either by directly inserting into the training data or by encoding it into the model's parameters. Once the model produces watermarked content, specialized algorithms can detect and verify the watermark to assert ownership or monitor distribution. AI engineers design this process to resist common content alterations while maintaining the quality of the AI-generated content.

Potential benefits of watermarking are as follows:

- It does not change the generated content but are machine detectable identifiers.
- It cannot be removed without appropriate technical know-how, which makes their erasure difficult i.e., they cannot be removed by minimal edits to the content.
- It allows identification of AI-generated content.
- It may be developed alongside the model, which makes the model developers liable for watermark encoding and detection.

Key concerns and considerations in AI content detection mechanisms: At present, systems that detect AI generated content—detecting anything from deepfakes to GenAI-authored articles—are significantly lacking behind and are often even approached through a manual process which may have a higher probability of false results/ alarms. Consequently, challenges such as imperceptibility, capacity, robustness, false positives and security issues highlight the limitations and vulnerabilities in these systems.

- **Imperceptibility** refers to the ability of GenAI-generated content to blend seamlessly with human-created content, making it challenging for detection systems to identify. The near-imperceptible signals of GenAI generation require increasingly complex detection algorithms, leading to a continuous arms race between generation and detection technologies.
- **Capacity** issues arise as the volume of GenAI-generated content surges. Detection systems must process and analyze vast datasets at an ever-accelerating pace. As the scale and frequency of content identification through the system grows, maintaining a detection system that can keep up without sacrificing accuracy or performance becomes difficult.
- The **robustness** of GenAI content detection systems is critical in maintaining consistent and reliable performance under varying conditions. These systems should be immune to attacks and removal while continuing to serve the purpose. A robust detection system must be adaptable, continually learning and updating its parameters to stay ahead of these threats.
- **False results/alarms** are another significant challenge. GenAI content detection systems may erroneously flag authentic content as GenAI-generated or vice versa. False alarms can have serious implications, such as unwarranted censorship, reputation damage or undue legal scrutiny.
- **Security** concerns pertain to protecting the detection systems themselves. If the integrity of these systems is compromised, the consequences can range from the propagation of false information to undermining trust in digital platforms and may even compromise the code/identifier. Secure frameworks must be in place to ensure that only verified operators can influence detection mechanisms, and robust encryption should secure data flows within these systems.

Effective watermarking systems continue to evolve, but addressing these concerns will be crucial for successful implementation and adoption in the future. This is likely to lay the necessary groundwork for the technology to protect the users of AI generated content from various adversities.

Technical challenges in AI watermarking

Watermarking is an evolving technology. Currently, there are technical challenges that companies face while implementing watermarking. Development of watermarking systems that are sophisticated enough to evade removal, and withstand various attempts of manipulation, is a complex task. Additionally, issues may arise when generated content is compressed, shared, or the file formats are changed.

The resolution of challenges pertaining to technical implementation, robustness and accuracy of watermarks would critically determine the development and adoption of the technology.

- **Technical implementation:** AI companies have found it challenging to add reliable watermarks to AI generated content without changing the essence of the same.
- **Accuracy:** Classifiers in past GenAI watermarking attempts have frequently misclassified human-generated content as AI-generated and vice versa, leading to high false positive rates.
- **Robustness :** Research indicates that both visible and invisible, text-based and audiovisual watermarks can be manipulated, removed and tampered.
- **Continuous evolution of technology:** With the rapid development and improvement in AI capabilities and deepfakes technologies across all forms of media, watermarking mechanisms will have to continue improving.
- **Increased requirement of resources:** Running detection mechanisms on all types of media in the real time will be computationally challenging. Measures will be needed to develop such mechanisms in a cost and resource efficient manner.





3

Policy scenario for AI content detection mechanisms

From a policy perspective, interoperable global standards are required to facilitate content detection mechanisms (including watermarking) across different jurisdictions. Ethical implications should be considered while assessing the balance between protecting content and potentially infringing on individual privacy. The Report further discusses some of the measures which will enable watermarking as one of the content detection mechanisms.

Key policy considerations

Interoperability

Each AI model developer can only build detectors for their own watermark. Coordination among developers is necessary for efficient identification of all watermarks.

Privacy and confidentiality

Watermarking may inadvertently leak sensitive information. Striking a balance between content authentication and privacy protection is challenging.

Open-source models

In open-source models balancing transparency and integrity can be challenging and requires careful consideration of what is made public and what is not disclosed to safeguard detection integrity.

Global standards for AI watermarking systems

There is a need to establish standards underlying global best practices for effective implementation of AI watermarking systems across the globe.

Limited adoption

Presently, there are very limited AI content detection and watermarking technologies in practice. The robustness and quality of the same is also being challenged.

Government involvement in enabling the operationalization of AI content detection mechanisms is crucial, where watermarking could be one of the measures. Primarily, governments can encourage industry to use content detection mechanisms to ensure accountability, standardization and trustworthiness across digital platforms. Government support is also vital for driving research and development in AI content detection technologies and standards, fostering technical advancements and local industry growth through resource allocation and funding. Facilitating international cooperation, fostering global norms and interoperable systems are essential for managing cross-border AI-generated content effectively for content detection mechanisms to be viable in the larger context. Lastly, government backing enables large-scale public awareness campaigns, which are necessary to engage citizens at scale.

Government involvement plays an integral role in establishing a secure, transparent and ethically aligned digital ecosystem, anchoring the operational success of AI content detection mechanisms for safeguarding our digital future. Some of the key policy measures taken by governments across the globe to adopt content detection mechanisms are highlighted ahead.



A few country-level approaches for AI content detection

Private Entity	Key initiative
People's Republic of China	**1. Provisions on the administration of deep synthesis of internet information services:** ▸ The provisions approved and passed by the Cyberspace Administrative Council of China have three broad requirements for AI-generated content. ▸ Generated content is subject to an 'explicit watermark', i.e., a prompt text indicating 'generated by AI'. ▸ AI-generated images, videos, and audio are subject to an 'implicit watermark', i.e., technical tagging which may be imperceptible to humans but technically detectable. ▸ AI-generated content saved as files should display metadata for identification.
United States of America	**1. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence:** ▸ Requires the US administration to develop effective labeling and content provenance. Mechanisms to ensure that users can be made aware if content is AI-generated. ▸ Explores the techniques, standards, tools and methods for content provenance, including but not restricted to watermarking. ▸ Legislation may also be passed on the same ground by the Congress. **2. Disrupt Explicit Forged Images and Non-Consensual Edits Act, 2024** ▸ The recently passed DEFIANCE Act recognizes the need to protect individuals from the threat posed by the use of digital forgeries and deepfakes to create sexually morphed/ nude images, The Act aims to "improve rights to relief for individuals affected by non-consensual activities involving intimate digital forgeries, and for other purposes".
European Union	**1. EU AI Act (August 2024):** Article 52 of the European Union AI Act requires providers of AI systems generating synthetic audio, image, video of text content to ensure that the outputs are marked in a machine-readable format which must be detectable as Artificially generated. **2. Code of Practice on Disinformation:** ▸ The 2022 Code of Practice on Disinformation, developed by the European Union, focuses on limiting the spread of online disinformation, including AI-generated content. ▸ Brings together global industry players to commit to combatting disinformation and includes measures like demonetizing the dissemination of false information, ensuring transparency in political advertising, empowering users, and enhancing cooperation among signatories. ▸ Highlights the importance of content provenance mechanisms to identify the source of information to combat misleading content and promote trustworthy news sources.
G7	**1. International guiding principles on Artificial Intelligence:** ▸ The provisions accepted in October 2023 require that content authentication practices such as provenance mechanisms be set up alongside advanced AI software. ▸ Provenance data should include an identifier of the service or model that created the content but need not include user information. ▸ Organizations may also create identification tools for their respective AI systems, which may be made available to the public for easier identification of AI-generated content. ▸ Appropriate disclaimers and identifiers may be put in place by AI model developers to apprise users that they are interacting with an AI model.
United Kingdom	In an attempt to crackdown on the creation of sexually morphed deepfake images, the UK government had proposed fines and criminal record against perpetrators.

Based on country-level announcements made so far, here are some of our key takeaways on how governments can approach content detection:

- **Multi-faceted approaches for AI content detection:** Successful implementation of AI content detection requires comprehensive coverage of all possible loopholes. This may be accomplished by using multiple detection mechanisms in tandem.
- **Building user confidence and trust:** Building user trust by communicating that content is AI generated through processes such as disclaimers is important to ensure customer trust.
- **Government push for AI Content Detection mechanisms:** There is a push for AI content detection mechanisms to be adopted and deployed. Globally, several jurisdictions have urged the stakeholders to promote adoption.
- **Public-private partnerships:** The government and private sector aim to work closely to mitigate the risks that emerge from the use of AI. This allows the technical capabilities of the private sector to build upon the experience of the government to develop systems that will help promote the use of AI content detection mechanisms.
- **Adoption of global best practices through standards and frameworks:** Standards and norms provide the backdrop for the adoption of "best practices" for AI content detection and ease of AI users. The AI community needs to establish standards for interoperability and reach a consensus on AI content detection mechanisms.

How is the private sector approaching AI content detection?

AI development and deployment have picked up pace. Private sector companies are using AI to automate processes, enhance productivity, and drive businesses. Trust in the ecosystem is necessary to ensure that the ecosystem continues to grow and derive the economic benefits of AI. Private companies are resorting to AI content detection mechanisms to assert ownership, protect intellectual property, identify AI generated content, combat misinformation, etc. This potentially allow businesses to hedge risks, thereby safeguarding them from financial or non-financial losses.

However, the challenge for industries with GenAI-generated content is authenticity and integrity of digital assets. As the private sector grapples with this issue, collaborative efforts among different GenAI content detection mechanisms emerge as a promising approach to address the multifaceted nature of this problem.

We are given to understand that Industry is working with PAI to develop responsible practices for synthetic media.

The Global AI Safety Summit held at Bletchley Park (UK) in November 2023 recognized shared AI safety risks and the need to develop collaborative efforts. The summit emphasized a balanced approach that respects individual national strategies while advancing research. Technology giants with a large footprint in the AI ecosystem made announcements pertaining to their approach to "Identifiers of AI-generated material". Some key takeaways from the announcements are discussed below:

- By using watermarking technology in collaboration with other AI content detection systems, companies plan to ensure accurate and robust detection of AI generated content.
- The challenges may vary across different content generation platforms. Text, for instance, is in abundance and can be easily manipulated. While watermarking visual content has been the primary focus for many companies, the emergence of AI generation tools for images and videos introduces new complexities.
- In response to these challenges, the private sector advocates for a multi-stakeholder led approach that encompasses technological innovation, self-regulation, and public awareness campaigns.

It is possible that different companies take different approaches. However, in the context of this issue, there is global dialogue and multi-stakeholder led conversations happening through forums like the Partnership for AI.

The India context

India, like many other countries, is facing a growing threat to citizen security due to the proliferation of misinformation, fake news and the misuse of AI technologies. **With greater proliferation of AI usage, these disturbing trends are also likely to affect business interests.** In recent years, the use of manipulated videos, images, and text has spread and has large-scale implications on internal law and order, and national security, making the need for robust AI content detection mechanisms more apparent. Some of the key challenges which have emerged are:



Threat to citizen security

One of the most pressing issues facing India is the widespread circulation of manipulated content, particularly during sensitive periods such as elections. India is prone to misinformation and fake news, which has significant implications on electoral processes and internal law and order. There are various videos, images of key political leaders and celebrities that are edited and circulated widely.



AI modified content, deepfakes, and AI misuse

Recognizing the severity of the situation, the Indian government has issued multiple advisories highlighting the threat posed by deepfakes and the urgent need for GenAI content detection measures. However, given the focus on the development of AI models in local languages, developers may need to create watermarking techniques for these specific purposes.



Social media manipulation

Social media manipulation is another critical issue plaguing India, with platforms being used as conduits for spreading fake news, misinformation and disinformation.



Lack of AI content detection mechanisms and copyrights

Copyright laws in India need to keep pace with the rapid advancements in AI technology. In 2019, the Delhi High Court rejected a copyright claim over a list compiled by a computer, on the grounds of, inter alia, lack of human intervention. However, in 2020, the Copyright Office had recognized an AI tool, Raghav, as an author of an artwork produced by the AI tool, along with the developer of the AI tool. A comprehensive legal framework is needed to address the intricacies of AI-generated content and ensure fair attribution and protection of intellectual property rights.

Collaboration between the government, private sector, legal experts, academia and civil society is essential to develop and deploy effective detection solutions. Investments in research and development, capacity building and public awareness campaigns will help the country stay ahead of the evolving threats posed by AI-generated content manipulation.

The current policy landscape in India

Some provisions of existing laws and policies presently address various aspects of AI generated content, namely:

- The Information Technology Act, 2000 includes Section 66E which applies to deepfake crimes involving capturing, publishing or transmitting someone's images. Sections 67, 67A, and 67B prosecutes in case of obscene or sexually explicit content.
- IT Rules also forbid impersonation content and mandate for prompt removal of morphed images from social media platforms.
- The Copyright Act of 1957, Section 51, can be invoked for unauthorized use of copyrighted images or videos in creating deepfakes.
- The proposed Digital India Act discusses the processes for the ethical use of AI based tools to protect rights or choices of users and provisions of deterrent, effective, and proportionate and dissuasive penalties may be imposed.

A set of regulations focusing on deepfakes and misinformation was under development as of November 2023. These regulations are expected to focus on identifying, preventing, reporting and creating awareness of deepfake technologies.

The industry is exploring how to set up a technical architectural framework that will enable the necessary safeguards and guardrails. But what remains of paramount importance is to ensure that there is a balance between what is being regulated or prevented while ensuring that the innovation in the ecosystem is not stifled.

> Trust and safety in AI systems can be enabled through multi-stakeholder led development of standards, guidelines and best practices concerning watermarking and AI content detection. This will enable safer adoption of such systems.





4

Way forward

To develop robust detection mechanisms for detection, there is a need for multi-layered approaches, encompassing both technological advancements and governance frameworks.

Recommendations for the government

As the use of AI increases, the exploitation and consequences that emerge are bound to increase exponentially. This may prevent economies from optimally deriving the advantages of AI. The government will need to play a key role in enabling the development of a supportive legal and governance framework. The government will also have to incentivize the development of open source and technical capabilities to build capacity and establish partnerships.

1. Establish technical and governance architecture

Measures to promote the adoption of AI content detection mechanisms	▸ Promote necessary interventions for the voluntary adoption of AI content detection mechanisms, which may include self regulation measures.
Work towards establishing global standards for AI watermarking which may be reviewed/updated periodically	▸ Bletchley Declaration - A total of 28 countries, alongside the European Union, came together to address AI safety risks. Their first objective was to build a shared scientific understanding of AI risks, which is essential for the technology's sustainable development. They also aimed to create risk-based policies that will ensure AI safety across different nations. These policies are tailored to respect the individual national circumstances and legal frameworks of the participating countries. ▸ Adoption of watermarking and other content detection mechanisms would require technical standards and universal consensus.
Ensure cyclic review of technology	▸ Ensure that technology is reviewed regularly so that the standards keep up with the industry developments and are not bypassed using watermark removal tools.

2. Support technical development and promote open-source

Develop open-source tools and digital infrastructure to support watermarking	▸ Make tools widely available so that it is easier for smaller businesses and companies to adopt established watermarking standards, along with a digital infrastructure to support it.
Fund technical research and development	▸ To improve and build upon existing water marking technologies, protocols and standards for adoption to reflect the state of the industry/ecosystem.

3. Promote capacity building, consultations and partnerships

Public Outreach and Multi-stakeholder consultations	▸ There will be a need to carry out large-scale public awareness campaigns surrounding the verification, use and adoption of watermarks. It may be expected that at some point this awareness may become as common as basic civics learning. ▸ The Government may undertake multi-stakeholder consultations to facilitate discussions on best practices for AI content detection mechanisms. This would help maintain flexibility and feasibility in the adoption and development of AI technologies.
Promote the implementation of Watermarks by the government	▸ As governments increase AI usage, it will become important to highlight any services which leverage AI generated content for delivery of such services.

Recommendations for the private sector

In the rapidly expanding frontier of AI, the responsibility of the private sector to safeguard users from potential risks has gained importance. Companies operating in this space must be proactive in adopting strategies that not only protect their users but also align with governmental initiatives aimed at ensuring the integrity of AI applications. They will have to complement the government's efforts to promote the development and adoption of AI content detection mechanisms. This will include establishing internal controls and measures while also evaluating and mitigating stakeholder concerns.

1. Establish internal control measures and foster implementation

Evaluate the usage of AI generated outputs	▸ Companies must understand what content is created using AI, and which watermarking tools or approaches will be most appropriate.
Rigorous internal testing of AI watermarking system	▸ Test AI-generated watermarks and protocols before any widescale organization wide deployment and conduct red teaming exercises to ensure that robust solutions are deployed as scale.
Keep abreast with accepted standards	▸ Companies must track the technology to ensure that the adopted practices are in line with the wider national or international standards and must also stay aware of the tools and technologies used to undermine prevalent watermarks. ▸ Private sector entities may also engage or participate with global private sector initiatives or partnerships, such as the Partnership on AI, to be aware of the latest best practises and concerns relating to AI content detection.
Evaluate implementation of watermarks for AI generated outputs	▸ Investments are necessary both in terms of infrastructure and talent to ensure that organizations meet AI Watermarking requirements.

2. Evaluate and mitigate stakeholder concerns

Beware of the risks and challenges	▸ C-suite leaders must be aware of the risks and challenges of AI watermarking given the field is nascent and discourse around the technology, its testing and adoption is still underway.
Privacy of stakeholders	▸ They must also ensure that watermarking tools do not violate customer or user privacy and adequate safeguards are in place.
Ensure compliance with the relevant legal framework	▸ As relevant national and international standards evolve, companies must be aware of what laws will apply and need to ensure adherence to the necessary legal requirements.





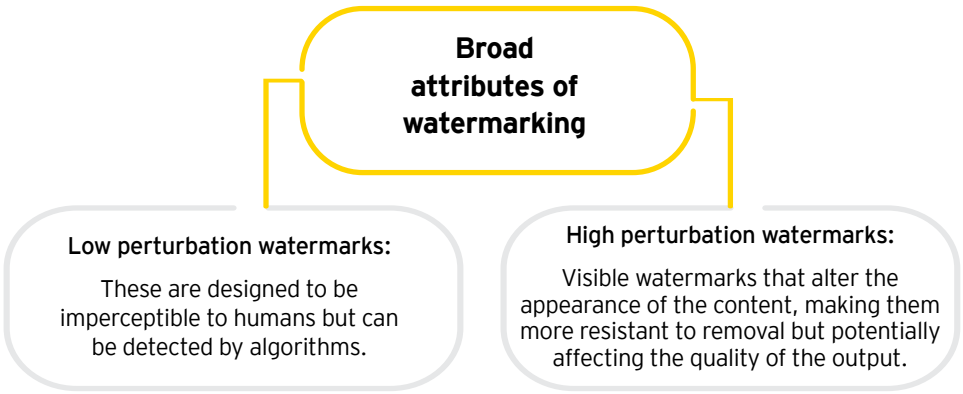
5

Annexure



Annexure:
Technical overview of watermarking

Perturbation in watermarking is about embedding information in a way that is both imperceptible to the user and resilient against attempts to remove or degrade the embedded information.



Name of the technology	Description	Advantages	Disadvantages
Steganography	Steganography is a method of hiding information within another piece of media, usually through slight modifications to the least significant bits of a file's data.	Good localization: Steganography allows precise localization both in the time and spatial frequency domain. Human perception: It achieves a higher compression ratio relevant to human perception. Robustness: Steganography is more robust to cropping.	Higher computing cost: The computational cost may be higher. Longer compression time: It requires more time for compression. Noise/blur near edges: Some steganography methods introduce noise or blur near the edges of images or video frames.
Invisible forensic watermarking	Invisible forensic watermarking is a technique used to create a unique identifier within video or image content that allows owners to track its usage and spread.	Authenticity verification: It helps determine the authenticity and provenance of AI-generated content. Transparency and ethics: Watermarking contributes to transparency and ethical practices.	Complex Implementation: Implementing invisible forensic watermarking can be challenging. Limited applicability: It may not be suitable for all types of content (e.g., text) due to its nature.
Dataset watermarking	Dataset watermarking is a method used to mark and authenticate input and training datasets for model development by linking the dataset back to its creator.	Privacy concerns: Dataset watermarking may raise privacy concerns if sensitive data is involved. Provenance tracking: It allows tracking the origin of training data. Ethical compliance: Helps ensure ethical use of datasets.	Complexity: Implementing model watermarking requires careful design. Potential Overhead: Embedding watermarks in the model may impact performance

Name of the technology	Description	Advantages	Disadvantages
Model watermarking	Model watermarking is a technique employed to monitor and authenticate ownership of an AI model or oversee its usage by incorporating information within the model's parameter and structure. Statistical bias watermarking is a type of model watermarking.	Model attribution: It associates AI-generated content with a specific model. Accountability: Helps hold models accountable for their outputs.	Complexity: Implementing model watermarking requires careful design. Potential overhead: Embedding watermarks in the model may impact performance.
Differential watermarking	Invisible forensic watermarking is a technique used to create a unique identifier within video or image content that allows owners to track its usage and spread.	Authenticity verification: It helps determine the authenticity and provenance of AI-generated content. Transparency and ethics: Watermarking contributes to transparency and ethical practices.	Complex Implementation: Implementing invisible forensic watermarking can be challenging. Limited applicability: It may not be suitable for all types of content (e.g., text) due to its nature.



Notes:

Notes:

Acknowledgments

Steering Committee:

Mahesh Makhija

Rakesh Kaul Punjabi

Rajnish Gupta

Vineet Mehta

Alexy Thomas

Core Team:

Rajnish Gupta

Ankan De

Rishi Dewan

Editorial Team:

Shweta Sharma

Design Team:

Sneha Arora



Our Offices

Ahmedabad
22nd Floor, B Wing, Privilon
Ambli BRT Road, Behind Iskcon Temple
Off SG Highway
Ahmedabad - 380 059
Tel: + 91 79 6608 3800

Bengaluru
12th & 13th Floor
"UB City", Canberra Block
No.24 Vittal Mallya Road
Bengaluru - 560 001
Tel: + 91 80 6727 5000

Ground & 1st Floor
11, 'A' wing
Divyasree Chambers
Langford Town
Bengaluru - 560 025
Tel: + 91 80 6727 5000

Bhubaneswar
8th Floor, O-Hub, Tower A
Chandaka SEZ, Bhubaneswar
Odisha - 751024
Tel: + 91 674 274 4490

Chandigarh
Elante offices, Unit No. B-613 & 614
6th Floor, Plot No- 178-178A
Industrial & Business Park, Phase-I
Chandigarh - 160 002
Tel: + 91 172 6717800

Chennai
6th & 7th Floor, A Block,
Tidel Park, No.4, Rajiv Gandhi Salai
Taramani, Chennai - 600 113
Tel: + 91 44 6654 8100

Delhi NCR
Ground Floor
67, Institutional Area
Sector 44, Gurugram - 122 003
Haryana
Tel: +91 124 443 4000

3rd & 6th Floor, Worldmark-1
IGI Airport Hospitality District
Aerocity, New Delhi - 110 037
Tel: + 91 11 4731 8000

4th & 5th Floor, Plot No 2B
Tower 2, Sector 126
Gautam Budh Nagar, U.P.
Noida - 201 304
Tel: + 91 120 671 7000

Hyderabad
THE SKYVIEW 10
18th Floor, "SOUTH LOBBY"
Survey No 83/1, Raidurgam
Hyderabad - 500 032
Tel: + 91 40 6736 2000

Jaipur
9th floor, Jewel of India
Horizon Tower, JLN Marg
Opp Jaipur Stock Exchange
Jaipur, Rajasthan - 302018

Kochi
9th Floor, ABAD Nucleus
NH-49, Maradu PO
Kochi - 682 304
Tel: + 91 484 433 4000

Kolkata
22 Camac Street
3rd Floor, Block 'C'
Kolkata - 700 016
Tel: + 91 33 6615 3400

Mumbai
14th Floor, The Ruby
29 Senapati Bapat Marg
Dadar (W), Mumbai - 400 028
Tel: + 91 22 6192 0000

5th Floor, Block B-2
Nirlon Knowledge Park
Off. Western Express Highway
Goregaon (E)
Mumbai - 400 063
Tel: + 91 22 6192 0000

3rd Floor, Unit No 301
Building No. 1
Mindspace Airoli West (Gigaplex)
Located at Plot No. IT-5
MIDC Knowledge Corridor
Airoli (West)
Navi Mumbai - 400708
Tel: + 91 22 6192 0003

Pune
C-401, 4th Floor
Panchshil Tech Park, Yerwada
(Near Don Bosco School)
Pune - 411 006
Tel: + 91 20 4912 6000

10th Floor, Smartworks
M-Agile, Pan Card Club Road
Baner, Taluka Haveli
Pune - 411 045
Tel: + 91 20 4912 6800



Ernst & Young LLP

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EYG member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is one of the Indian client serving member firms of EYGM Limited. For more information about our organization, please visit www.ey.com/en_in.

Ernst & Young LLP is a Limited Liability Partnership, registered under the Limited Liability Partnership Act, 2008 in India, having its registered office at Ground Floor, Plot No. 67, Institutional Area, Sector - 44, Gurugram, Haryana - 122 003, India.

©2024 Ernst & Young LLP. Published in India.
All Rights Reserved.

EYIN2409-012

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither EYGM Limited nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.

SA1

About FICCI

Established in 1927, FICCI is the largest and oldest apex business organisation in India. Its history is closely interwoven with India's struggle for independence, its industrialisation, and its emergence as one of the most rapidly growing global economies. A not-for-profit organisation, FICCI is the voice of India's business and industry. From influencing policy to encouraging debate, engaging with policy makers and civil society, FICCI articulates the views and concerns of industry. It serves its members from the Indian private and public corporate sectors and multinational companies, drawing its strength from diverse regional chambers of commerce and industry across states, reaching out to over 250,000 companies. FICCI provides a platform for networking and consensus building within and across sectors and is the first port of call for Indian industry, policy makers and the international business community.

ey.com/en_in

@EY_India EY YouTube EY India EY Careers India @ey_indiacareers