A Brief Survey of Watermarks in Generative AI

JaeYoung Hwang*
Principal Researcher, AI Trustworthiness Verification Team
Telecommunications Technology Association (TTA)
Seongnam, Gyeonggi-do, Republic of Korea
jyhwang@tta.or.kr

SangHoon Oh
Associate Researcher, AI Trustworthiness Verification Team
Telecommunications Technology Association (TTA)
Seongnam, Gyeonggi-do, Republic of Korea
tkdgns7030@tta.or.kr

***Abstract*—Generative AI technology is now capable of producing images and text at a level comparable to that of humans, showcasing its remarkable utility. However, this advancement comes with its share of challenges, such as misuse, prompting discussions on effective response strategies. Consequently, recommendations and regulations, including the adoption of watermark technology, are under deliberation on a country-by-country basis. Many companies are also integrating watermark technology into their services as a means of addressing this issue. This paper presents an analysis of the current status of watermark adoption across various countries and companies. Furthermore, it delves into further research topics that should be taken into account when implementing watermark technology. This analysis aims to provide valuable insights to those who are contemplating the implementation of watermarking in their future generative AI services.**

***Index Terms*—Artificial Intelligence (AI), Generative AI, Watermarks**

I. Introduction

The generative AI sparked by ChatGPT is currently one of the most actively researched technologies in the field of artificial intelligence. Leveraging the ability to produce text and images at a level similar to humans, it is being utilized across various industries and its significance is steadily growing. However, as generative artificial intelligence starts to have more societal impact, issues regarding biased information generation and misuse have come to the forefront. For instance, a fabricated photo claiming an explosion near the Pentagon caused a maximum 0.3% drop in the US stock market. Additionally, a fake image depicting former United States(US) President Donald Trump being arrested by the police was shared on Twitter, leading Twitter to issue a notice that the photo was fake. In response, the US government expressed concerns about the potential manipulation of public opinion in next year's presidential election. The fundamental issue arises from the inability to distinguish whether the generated content is created by humans or produced using artificial intelligence technology. As a result, each country is formulating recommendations and regulations to address these issues, and companies are working on technological developments to meet these requirements. In particular, watermark technology has recently been considered as one of the methods to mitigate the aforementioned problem.

The watermarking technique involves inserting patterns or other elements into digital image files, such as files, to identify copyright information and protect Intellectual Property(IP) rights. For instance, Getty Images, the largest commercial image website in the United States, uses the 'gettyimages' watermark on its sale images to safeguard intellectual property rights. However, in the case of generative artificial intelligence, watermarking is also being considered for preventing the misuse of content (text, images, etc.) created by artificial intelligence, in addition to protecting IP rights.

Currently, the US White House has discussed and recommended the application of watermark technology to content generated using generative AI technology, in collaboration with major big tech companies. In the cases of Europe and China, beyond the recommended level, there are regulations stipulating the application of watermark technology to generative AI content. In addition, several countries are considering applying watermark technology to generative AI content. Therefore, as each country is demanding the insertion of watermarks for generative AI, companies possessing generative models need to conduct thorough analyses to respond to this.

Therefore, this paper examines the recommendation and regulation trends concerning generative AI watermarking in key countries and how major tech companies are currently integrating watermarking features into generative AI content. Reviewing these compiled insights is expected to provide a valuable understanding of generative AI watermarking.

Consequently, in this paper, the main contributions of this paper are as follows:

- Analysis of recommendations and regulations regarding generative AI from major countries such as the United States, China, Europe, and South Korea.
- Explanation of how major big tech companies are implementing watermarking technologies.
- Based on the above analysis, additional research considerations for applying generative AI watermarking technology have been presented.

The remainder of this article is organized as follows. Section II illustrated four AI-leading countries that are carrying out recommendations or regulations. In the next section III, companies that currently provide generative AI services by introducing watermarking technology are examined. Based on the above surveys, it was suggested further research agendas should be considered to prevent misuse of the generative AI in section IV. Finally, the conclusion of this paper and future research directions are presented in section V.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

II. Status of Recommendation and Regulation to Generative AI Watermarks by Key Countries

In this section, the recommendations and regulations regarding the generative AI watermark that is currently being implemented in each country are illustrated.

A. United States

In July of this year, the US announced ***Ensuring Safe, Secure, and Trustworthy AI*** [1] a set of eight guidelines that companies providing AI services should follow to promote the safe and transparent development and use of AI services. This recommendation, demands the development of mechanisms to include the source or watermark for audio or visual content generated by AI technology, allowing the identification of whether AI created the content.

As a subsequent step, Google, OpenAI, Microsoft, and Anthropic have established the ***Frontier Model Forum***. This forum is in charge of the development of the mechanism for ensuring the safety and responsible practices of AI models. In their joint statement, they emphasize focusing on research into safety measures and the discovery of best practices for the responsible development of large-scale machine learning models.

B. European Union

Europe is the most active region in terms of regulating AI. In April 2022, the European Parliament announced the ***Digital Service Act (DSA)*** [2] to ensure the safety of European users from online misinformation, illegal content, and goods or services. In this regulation(Article 35-Mitigation of risks-(k) clause), the use of AI to generate content (images, audio, video) requires that the identity of the generating entity be indicated to consumers through prominent markings. The DSA is set to be enforced starting from August 2023, with a primary focus on Very Large Online Platforms (VLOPs) such as Meta, Apple, Amazon, Alphabet (Google), TikTok, Twitter, YouTube, and 12 other global platforms as the main targets of regulation.

C. China

The Regulations for the Deep Integration Management of Internet Information Services [3] have been announced to regulate services that utilize deep learning technology to generate or edit information such as text, images, videos, and audio. Among the specific provisions of these regulations, it is stipulated in Chapter 3, Article 17 that when service providers offer the services described in Table I, they must prominently display reasonable locations and areas for the public to identify the content's creator.

TABLE I: Type of service required to apply watermark

	Main contents
Services	• Intelligent conversation and writing services that provide text generation or editing functions mimicking human-like responses. • Voice generation including imitating human speech, and editing services that significantly alter personal identity traits. • Image and video editing services involving generation, replacement, and manipulation of human images, and gestures, with substantial modifications to personal identity traits.

D. South Korea

To promote the content industry and protect the rights of content creators, the ***Content Industry Promotion Act*** [4] has been established. Recently, there has been an opinion that to distinguish the fact that content (text, images, music, etc.) has been created using AI technology, there is a need to establish relevant laws and regulations, and therefore a revised bill has been proposed.

III. Enterprise-Specific Generative AI Watermark Support

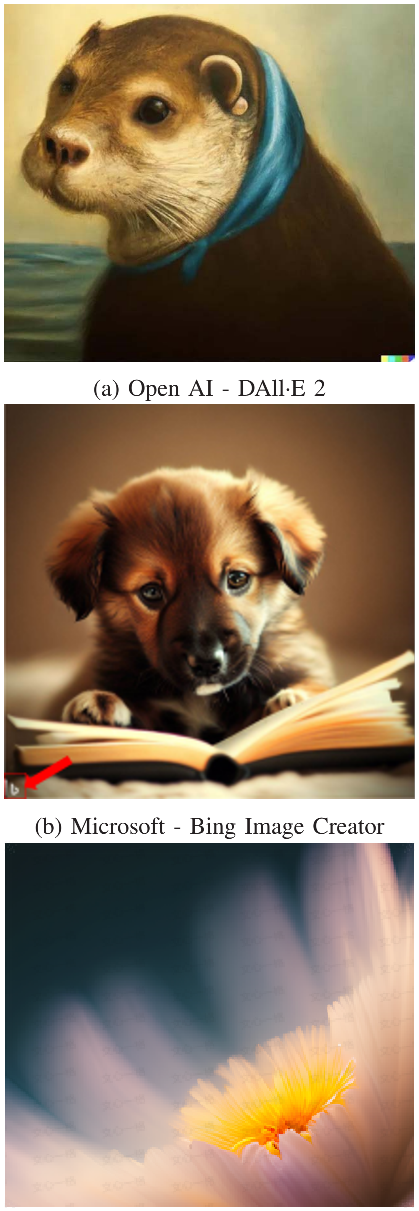
Due to concerns about the potential misuse of generative AI, the demand for continuous measures from both service users and governments is increasing. As a result, major companies providing generative AI services are incorporating watermarking technology, primarily focusing on generative image content. Additionally, they are actively engaged in open-source research, development, and standardization efforts to support the application and implementation of watermarking technology. To provide more detailed information, watermark images by leading tech companies have been included in Fig 1.

A. OpenAI

OpenAI, a leading company providing generative AI services, operates text generation service ***ChatGPT*** and image generation service ***DALL·E 2***. First, DALL·E 2, a service that creates realistic images based on natural language descriptions, inserts watermarks of five primary colors in the lower-right corner of the images. In the text data of generative AI, there exist two methods for watermarking: one applied after all words are predicted (post-hoc), and another that inserts watermarking during the next-word inference process of the text generation model [5]. According to The New York Times [6], OpenAI appears to have adopted a watermarking approach similar to the one outlined in the paper [5], although specific details have not been publicly disclosed.

B. Microsoft

Microsoft announced at the 'MS Developer Conference Build 2023' that they would be adding watermarks to all of their products that utilize generative AI technologies such as ***Bing Image Creator*** and ***Microsoft Designer***. Additionally, in collaboration with Adobe, Sony, BBC, and others, Microsoft has jointly established and is currently operating the C2PA (Coalition for Content Provenance and Authenticity) technology standard organization. This initiative aims to address the spread of false and misleading information associated with images by developing an open standard for indicating the source, authenticity, and use of generative AI in digital



(a) Open AI - DAll·E 2

(b) Microsoft - Bing Image Creator

(c) Baidu - Wenxin Yige [7]

Fig. 1: Images generated by AI

images. Furthermore, Stability AI has announced that all images generated through their APIs will adhere to the C2PA metadata standard.

C. Alibaba & Baidu

Baidu, China's largest search engine platform company, unveiled the image generation platform ***Wenxin Yige*** in August 2022. Wenxin Yige provides a watermark by inserting the 'Wenxin Yige' Chinese logo as a background when generating images. Alibaba, the world's largest e-commerce platform operator, introduced the image generation AI service ***Tongyi Wanxiang*** in July 2023. This service inserts the 'Tongyi Wanxiang' watermark in the bottom right corner and recommends the application of watermarking technology to AI-generated images when utilizing the service, following the service usage guidelines.

IV. Further research for generative AI watermarks

In an attempt to mitigate the misuse of the content generated by generative AI, several countries have recommended the adoption of watermarks and companies are researching watermark technologies. Nevertheless, additional discussion is required concerning the following matters.

A. Standardized guidelines for applying watermark technology

In the case of watermarks, various recommendations or regulations are being announced based on the specific relationships of each country, as a global consensus has not been reached. For companies providing generative artificial intelligence services, it is unclear which regulations to follow, and aligning with watermark requirements from different countries or institutions is not an easy task. Therefore, it is deemed necessary to establish collaborative bodies like C2PA for generative artificial intelligence, to develop common watermark guidelines across countries and organizations.

B. Watermarks for Open Source

Currently, requests for watermark implementation on the outputs of generative artificial intelligence are primarily being considered for major tech companies. However, the recent open-source community has also shown impressive performance, necessitating a review of this aspect. For instance, one of the Korean companies using the open-source LLaMa2 in the evaluation scores of the 'Open Large Language Model Leaderboard,' has surpassed the performance of GPT-3.5, the basis of ChatGPT. As most companies outside of the few with supermassive infrastructure are expected to utilize open-source-based generative AI models, it is essential to review strategies for applying watermarks about this.

V. Conclusion

The advancement of generative AI technology capable of producing human-level images and text has been utilized in various industries, demonstrating its usefulness. However, the problem of being unable to determine whether the content currently being used is generated by artificial intelligence or not has led to various instances of misuse. As a result, watermark technology has recently been considered as one of the methods to mitigate the aforementioned problem. In this regard, this paper examines the recommendations and regulation of watermarking technology across different countries and presents the enterprises that are applying watermarking technology in their AI-generated services. Moreover, this paper identifies topics that warrant further investigation in the context of watermarking technology deployment. The insights derived from this analysis and exposition are anticipated to provide valuable guidance to those who are seeking to incorporate watermarking technology into future AI-generated services.

The generative AI technology once considered the exclusive domain of big tech, has become accessible to various companies and users with the release of sLLM as an open source. Therefore, future work, research, and regulation are deemed necessary through the use of watermark technology.

Acknowledgment

This work was grant-funded by the Korean government (MSIT) (AI Trustworthiness Verification System Advancement Project).

References

[1] "Ensuring Safe, Secure, and Trustworthy AI," White House, [Accessed] 2023.08. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[2] "Digital Service Act," Europe Parliament, [Accessed] 2023.08. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

[3] "Regulations for the Deep Integration Management of Internet Information Services," Office of the Central Cyberspace Affairs Commission, [Accessed] 2023.08. [Online]. Available: https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm

[4] "Content Industry Promotion Act," South Korea, [Accessed] 2023.08. [Online]. Available: https://elaw.klri.re.kr

[5] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," *arXiv preprint arXiv:2301.10226*, 2023.

[6] "How ChatGPT Could Embed a 'Watermark' in the Text It Generates," The New York Times, [Accessed] 2023.08. [Online]. Available: https://www.nytimes.com/interactive/2023/02/17/business/ai-text-detection.html

[7] "Wenxin yige-Pink Flower watermarks image," Wenxin Yige, [Accessed] 2023.08. [Online]. Available: https://yige.baidu.com/galleryDetails/b9030000