Headline

생성형 AI 콘텐츠 진위 검증을 위한 워터마크 기술의 현황

컨설팅사업그룹 기업컨설팅 1 팀 권순철 책임

■ 생성형 AI 의 확산과 긍정적 영향

최근 생성형 인공지능(Generative AI)이 우리 일상과 산업에 빠른 속도로 스며들고 있다. 텍스트 작성, 이미지 생성, 음성 합성, 영상 편집 등 다양하게 활용돼, 인간의 창작 능력을 보완하거나 새로운 가능성을 높이는데 활용되고 있다. 기업은 마케팅 자료, 고객 응대 콘텐츠, 디자인 시안 등을 빠르게 제작해 시간을 절약하고, 개인은 전문 지식이 없이도 수준 높은 결과물을 손쉽게 얻을 수 있게 됐다.

이처럼 생성형 AI 는 생산성 향상과 창의성 증대, 맞춤형 서비스 확대 등 긍정적 효과를 가져오고 있으며, 디지털 전환을 가속화하는 핵심 동력으로 자리매김하고 있다.

■ 악용되는 생성형 AI 콘텐츠

생성형 AI 가 확산되는 만큼 악용 사례 증가도 발생하고 있다. 대표적인 것이 딥페이크(Deepfake) 기술이다. 특정 인물의 얼굴과 음성을 정교하게 합성해 허위 영상이나 음성을 만들어 명예훼손, 사생활 침해, 금융사기 등 사회 전반에 심각한 위협을 끼치고 있다. 실제로 SNS 와 유튜브에는 AI 합성 영상을 활용한 허위 정보 유포 사례가 꾸준히 보고되고 있다.

최근에는 이미지 편집에 특화된 생성형 AI 모델도 속속 등장하고 있다. 구글 딥마인드가 선보인 '나노바나나(Nano Banana)' 역시 큰 화제를 모으고 있다. 이 서비스는 단순한 지시만으로 배경을 변경하거나 인물을 합성하고, 심지어 가상의 의상을 입히는 작업까지 가능하다. 이러한 기능은 편리해 보이지만, 사실상 딥페이크와 유사한 결과물을 만들어낼 수 있어 초상권 침해, 허위 이미지 제작, 불법 합성물 확산으로 이어질 수 있다는 우려를 낳고 있다.

또한 가짜뉴스(Fake news) 전파로 선거 등 정치 영역에서도 문제가 발생하고 있다. 특정 후보자의 발언이나 사진을 불법적으로 조작·합성해 유권자를 혼란에 빠뜨리는 등 민주적 절차에 대한 신뢰를 무너뜨린다. 실제로 국내는 물론 해외에서, 선거 기간 중 딥페이크 영상이 대량 확산되어 사회적 파장을 일으킨 바 있다.

사이버 범죄와 결합하는 양상도 뚜렷하다. 가짜 음성을 활용한 보이스피싱, 위조 문서나 신분증 생성 등은 기존 보안 체계를 위협한다. 여기에 대화형 AI 를 통한 개인정보 및 민감정보 탈취 문제까지 더해지면서 피해 범위는 한층 확장된다. 공격자가 챗봇과의 일상정인 대화 과정에서 유출한 중요한 정보를 활용한다면 맞춤형 피싱 또는 정교한 스팸 공격 설계가 가능하다. 특히 기업 내부에서는 직원이 챗봇에 업무 관련 기밀을 입력할 경우, 영업비밀 유출로 이어질 수도 있다.

마지막으로 저작권 침해 문제 역시 간과할 수 없다. 작가의 화풍을 모방한 이미지 생성이나 음악 패턴을 학습한 결과물은 창작자의 권리를 직접적으로 위협한다. 더 나아가 AI 가 만들어낸 허위 리뷰와 광고 콘텐츠는 소비자 보호와 시장의 공정성을 무너뜨린다.

진짜와 가짜의 경계가 흐려지는 현상은 단순한 기술적 문제를 넘어 사회적 혼란, 법적 분쟁, 국가 안보 위협으로 이어질 수 있다.

■ 생성형 AI 콘텐츠 워터마크의 필요성

앞서 언급했듯이 생성형 AI 가 만들어내는 콘텐츠는 점차 사람들의 일상과 산업 전반으로 확산되는 동시에 위·변조 이미지, 허위 정보, 딥페이크와 같은 심각한 사회적 문제를 유발할 수 있다. 이에 따라 AI 생성물에 워터마크를 적용해 출처를 명확히 하고 위·변조 여부를 검증할 수 있는 기술적 장치의 필요성이 커지고 있다.

2026 년 1 월 국내에서 시행 예정인 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 "AI 기본법")」에서도 합성 영상 및 이미지에 워터마크 삽입을 의무화하는 내용이 포함되어 있다. 또한, 고위험 AI 개념 도입을 통해 인권과 안전 관련 기술에는 강화된 관리와 책임을 부여하는 등 구체적인 컴플라이언스 요구사항이 제시되고 있다. 이는 기술적 대응이 단순한 선택지가 아니라, 기업이 고려해야 할 제도적 흐름과 맞물려 있다는 점을 시사한다. 결국 기업들은 워터마크 도입을 기술적 신뢰 확보 차원을 넘어, 미래 규제 환경을 대비하는 전략적 대응이 마련되어야 한다.

생성물 형태	기업명		상용 서비스 현황	워터마크 유형
이미지	국내	SAMSUNG	갤럭시 AI(갤럭시 스마트폰)	인지 가능
		SK telecom	에이닷(A.)	인지 가능
	국외		Porme AI(AI Tool)	인지 가능
		Meta	AI 챗봇(AI Chatbot)	인지 가능, 인지 불가능
		Adobe	파이어플라이(Firefly)	인지 가능, 인지 불가능
		OpenAI	DALL·E 3	인지 불가능
		Microsoft	코파일럿(Copilot)	인지 불가능
			이미지 크리에이터(Image Creator)	인지 불가능
		Google	제미나이(Gemini)	인지 불가능
동영상	국외		드림 머신(Dream Machine)	인지 가능
		Canva	매직 미디어(Magic Media)	인지 가능
		TikTok	심포니 아바타(Symphony Avatars)	인지 가능
		Meta	무비 젠(Movie Gen)	인지 가능
		DeepMind	비오(Veo)	인지 불가능
오디오	국내	NAVER	클로바 더빙(CLOVA Dubbing)	인지 가능
	국외	OpenAI	보이스 엔진(Voice Engine)	인지 가능
		Azure	AI 음성(AI Speech)	인지 불가능
텍스트	국외	OpenAI	챗GPT(ChatGPT)	인지 불가능
		Google	제미나이(Gemini)	인지 불가능

* 출처 : 한국정보통신기술협회

그림 1. 국내외 인공지능 생성물 워터마크 도입 현황('24.12)

■ 생성형 AI 콘텐츠 워터마크의 유형

기존의 디지털 워터마크는 문서, 이미지나 영상 중심으로 출처를 표시하고 무단 사용을 방지하는 역할을 했다. 이러한 워터마크는 단순한 표시를 넘어, AI 생성물의 출처와 무결성을 확인할 수 있도록 특정 정보를 삽입하는 기술적 장치로 발전하고 있다.

하지만 AI 기반 콘텐츠가 폭발적으로 증가하면서 기존 방식만으로는 생성 출처를 충분히 확인하기 어려워졌다. 이에 따라 텍스트, 이미지, 음성, 영상 등 다양한 AI 생성물의 눈에 보이거나 보이지 않는 정보를 삽입하여 사람이 만든 것인지 AI 가 만든 것인지를 추적하고 식별할 수 있는 생성형 AI 콘텐츠 워터마크가 등장하고 있다.

생성형 AI 워터마크는 눈에 보이는 인지 가능(Visible) 형태와 눈에 보이지 않는 인지 불가능(Invisible) 형태로 구분되며, 목적과 활용 환경에 따라 적절히 적용된다.

유형	설명	장점	단점
인지 가능 워터마크	콘텐츠 위에 눈으로 확인 가능한 표시	출처 확인 용이	시각적 품질 저하, 워터마크 제거에 취약
인지 불가능 워터마크	콘텐츠에 디지털 신호나 메타데이터(C2PA 표준 등) 등으로 삽입되어 눈에는 보이지 않지만 검증 가능한 데이터 삽입	시각적 품질 유지, 강화된 보안	즉각적인 출처 확인 제한
동시 적용 (인지 가능/인지 불가능 워터마크)	인지 가능 + 인지 불가능 워터마크를 함께 사용	출처 표기와 위조 방지 동시에 가능	구현 복잡도와 리소스 부담 증가

표 1. 식별 워터마크와 비식별 워터마크의 유형 비교

■ 생성형 AI 콘텐츠 워터마크의 활용 사례

생성형 AI 콘텐츠가 이미지, 음성, 영상 등 다양한 형태로 확장되면서 주요 국내·외 서비스들은 각 콘텐츠 특성에 맞춰 인지 가능(Visible) 또는 인지 불가능(Invisible) 워터마크를 적용하고 있으며, 그 활용 방식은 다음과 같다.

콘텐츠 유형	주요 서비스	워터마크 유형	활용 사례
이미지	에이닷(SKT)	인지 가능 워터마크	생성형 AI 사진 및 프로필 이미지에 좌측 하단 로고 삽입으로 식별 가능한 워터마크 제공
	제미나이(Google)	인지 불가능 워터마크	텍스트 기반 이미지 생성 시 SynthID 를 통해 픽셀 단위로 인지 불가능한 워터마크 삽입
음성	클로바 더빙(Naver)	인지 가능 워터마크	생성된 음성 콘텐츠에 출처 표기 자동/직접 삽입 기능 제공
	AI 음성(Microsoft Azure)	인지 불가능 워터마크	에코은닉기술, 양지화 지수 변조, 확산 스펙트럼 기술을 이용해 96 비트 키 기반 워터마크를 오디오에 삽입
영상	매직미디어(Canva)	인지 가능 워터마크	AI 생성 영상 우측 하단에 특정 이미지 삽입 방식으로 시각적 워터마크 적용
	비오(Google DeepMind)	인지 불가능 워터마크	SynthID 를 활용해 영상 프레임마다 인지 불가능한 디지털 워터마크 삽입

표 2. 국내외 주요서비스의 AI 생성형 콘텐츠 유형별 워터마크 활용 사례

* 출처 : 과학기술정보통신부 블로그 콘텐츠 '빠르게 발전하는 AI 워터마크 기술, 어디까지 왔을까?'

그림 2. SK 텔레콤 에이닷(A.)으로 생성한 이미지(AI 프로필) 및 좌측 하단 이미지(AI 프로필)인지 가능 워터마크

* 출처 : 과학기술정보통신부·한국정보통신기술협회《인공지능(AI) 워터마크 기술 동향 보고서》

그림 3. 구글 제미나이로 생성된 원본 이미지(왼쪽)와 인지 불가능 워터마크(신스 ID)를 적용한 이미지(오른쪽) 비교

■ 생성형 AI 콘텐츠 워터마크의 기술적 한계

생성형 AI 컨텐츠의 진위를 구별하고 투명성을 확보하기 위해 워터마크는 기술적 안전장치로서 중요한 역할을 한다. 그러나 현재의 워터마크 기술도 한계는 있다. 인터넷상에는 워터마크를 제거하거나 변조할 수 있는 상용 도구들이 광범위하게 유통되고 있어, 삽입된 식별 정보가 손쉽게 삭제될 수 있기 때문이다.

또한 워터마크 삽입 방식 자체의 취약점을 악용한 공격도 가능하다. 예를 들어, GAN(Generative Adversarial Networks)을 이용한 워터마크 제거 기술은 이미지별 워터마크 위치 차이를 학습하여 워터마크를 효과적으로 삭제할 수 있다. 더 나아가 VWGAN(Very Weak Generative Adversarial Network)을 적용하면 기존보다 약 20% 높은 성능으로 워터마크 제거가 가능하다는 연구 결과도 보고되고 있다.

인지 불가능 워터마크 역시 완벽하지 않다. 겉보기에는 사용자가 식별하기 어렵다는 장점이 있지만, 랜덤 노이즈 삽입 후 이미지 재구축, JPEG 압축, VAE 기반 공격 등으로 제거될 수 있다. 특히 RivaGAN 방식에 재생성 공격을 통해 93~99%의 제거율이 확인되기도 했다. 이러한 사례들은 현재의 워터마크의 기술적 발전이 더욱 필요하다는 것을 보여준다.

■ 생성형 AI 콘텐츠 워터마크의 보완 과제

현재의 생성형 AI 콘텐츠 워터마크는 변조, 무단 복제, 불법 배포 등 다양한 위협에 노출될 수 있으므로, 기술적 보완이 필요하다.

우선 워터마크 내구성을 강화하는 것이 중요하다. 인지 가능·불가능 워터마크 모두 JPEG 압축, 이미지 재구성, 랜덤 노이즈 삽입 등 공격에 취약하므로, 워터마크를 이미지나 영상의 다양한 주파수 영역과 구조적 특징에 분산 삽입하는 방식을 고려할 수 있다. 또한 다중 계층적 삽입과 오류 정정 코드 적용을 통해 변조나 압축에도 워터마크가 유지될 가능성을 높일 수 있다.

적응형 공격에 대응하기 위해서는 워터마크 삽입 방식을 정적 위치가 아닌 랜덤화한 동적 위치에 삽입하거나 변형 가능한 패턴으로 설계하고, 주기적 업데이트와 공격 탐지 기능을 결합하는 방법이 적용되어야 한다.

복합적 검증 체계 구축도 또한 필요하다. 단일 워터마크 방식만으로는 제거를 완전히 방지하기 어렵기 때문에, 식별 가능 워터마크와 인지 불가능 워터마크를 동시에 활용하거나, 블록체인 등 분산 원장 기술과 연계해 원본 생성 기록과 워터마크 정보를 함께 확인하는 방안이 고려될 수 있다.

마지막으로, 자동화된 탐지 및 모니터링 체계 구축이다. 워터마크 제거 시도를 실시간으로 탐지하고 경고할 수 있는 시스템을 마련하면 비정상적 이미지 변조나 손상 발생 시 신속한 대응이 가능하다. 특히 배포와 재배포가 활발히 이루어지는 SNS 플랫폼 사업자를 비롯해, 언론사, 포털 사업자, 클라우드 스토리지 서비스 제공자 등 대규모로 이미지와 영상을 다루는 기관에도 이러한 체계를 적용하는 것이 적절해 보인다.

이러한 방향들이 종합적으로 고려될 때, 생성형 AI 컨텐츠 워터마크는 단순한 표시 수단을 넘어 콘텐츠 안전성과 책임 있는 사회적 인식을 마련하는데 핵심 수단으로 발전할 수 있다.

■ 맺음말

AI 생성형 콘텐츠의 확산으로 오남용과 변조 가능성이 증가했다. 피해 위험성을 줄이기 위해서는 워터마크 기술의 필요성이 부각되고 있다. 단일 기술만으로는 충분하지 않다. 다층적이고 종합적인 기술적 대응이 필수다. 먼저, 식별 가능한 워터마크와 인지 불가능한 워터마크를 동시에 적용함으로써 생성물의 출처와 무결성을 다각도로 검증해야 한다. 여기에 블록체인 기반 기록 등을 활용하여 원본 생성 기록과 워터마크 정보를 함께 적용해야 변조 여부를 보다 정밀하게 검증할 수 있다. 또한, AI 생성물 탐지 알고리즘과 자동화 분석 도구를 활용하면 인간의 육안으로는 식별하기 어려운 변조나 조작 시도도 실시간으로 확인할 수 있으며, 지속적인 기술 개선과 업데이트를 통해 새로운 공격 유형에도 대응 가능성을 높일 수 있다.

기술적 대응과 함께, 국제적으로 통용될 수 있는 글로벌 표준을 마련하여 국가와 플랫폼 간 워터마크 반영 및 탐지 방식의 일관성을 확보하고, 기술 개발과 적용을 촉진할 수 있는 정책적 지원도 필요하다. 아울러 일반 사용자와 콘텐츠 제작자가 워터마크의 목적과 중요성을 인지하는 '인식 제고 활동'을 전개함으로써, 무분별한 제거 시도를 줄이고 안전하고 책임 있는 콘텐츠 이용 문화를 정착시키는 것도 중요하다.

무엇보다 생성형 AI 가 산업과 사회 전반으로 빠르게 확산되는 전환기적 상황에서, 기업들은 더욱 철저한 대비와 선제적 대응을 갖출 필요가 있다. 특히 사회적 파급력이 큰 고영향 AI 와 생성형 AI 를 사용하는 기업이라면, 단순한 기술 활용을 넘어 법적·윤리적 리스크를 최소화할 수 있는 컴플라이언스 체계를 정교하게 구축해야 한다. 향후 마련될 세부 법령과 지침을 면밀히 추적하고, 이를 실무 현장에 즉시 반영할 수 있는 관리 역량을 확보하는 것이 곧 기업의 지속 가능성과 직결될 수 있다.

이러한 변화 속에서 AI 기본법에 부합하는 정보 보호 체계와 신뢰할 수 있는 AI 보안은 선택이 아닌 기업의 핵심 과제가 되고 있다. 데이터 보호, 알고리즘의 투명성, 위험 대응 체계는 기업의 사회적 책임을 뒷받침할 뿐만 아니라, 글로벌 시장에서 신뢰받는 파트너로 자리매김하기 위한 필수 조건이기도 하다.

SK 쉴더스는 다년간 축적된 AI 보안 및 데이터 보호 전문성을 기반으로, 각 기업 환경에 맞춤화된 위험 관리 및 대응 체계를 설계해왔다. 기업들이 규제 변화와 새로운 위협 환경에 능동적으로 대응할 수 있도록 지원하며, 안전하고 책임 있는 AI 활용을 위한 최적의 컨설팅 서비스를 제공하고 있다. SK 쉴더스의 전문적 지원은 기업들이 단순히 규제를 준수하는 데 그치지 않고, 안전성·투명성·경쟁력을 고루 갖춘 신뢰성 있는 기업으로 도약할 수 있을 것이다.

■ 참고문헌

[1] 과학기술정보통신부/한국정보통신기술협회, “인공지능(AI) 워터마크 기술 동향 보고서”, 2025.01

[2] Cao, Z., Niu, S., Zhang, J., & Wang, X., "Generative adversarial networks model for visible watermark removal", 2019.07

[3] Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., ... & Li, L., “Invisible Image Watermarks Are Provably Removable Using Generative AI”, 2023.06

■ 참고 자료

[1] 방송통신위원회/정보통신정책연구원, “생성형 인공지능 서비스 이용자 보호 가이드라인”, 2025.06

[2] 과학기술정보통신부, 빠르게 발전하는 AI 워터마크 기술, 어디까지 왔을까?, 2025.02

[3] EY, “Identifying AI generated content in the digital age: The role of watermaking”, 2024.09