www.copyright.or.kr

저작권 이슈 브리프

COPYRIGHT ISSUE BRIEF

Weekly Report

2026. 5-3

한국저작권위원회
KOREA COPYRIGHT COMMISSION

COPYRIGHT ISSUE BRIEF

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 EU AI 법 표시 의무, 복합 표시와 차등 감독으로 구체화

▶ EU 집행위원회는 2026년 5월 8일 AI 법 제50조 가이드라인 초안을 공개하며, 오는 8월 시행을 앞두고 AI 생성 콘텐츠의 표시 방식과 책임 범위를 구체화했다. 가이드라인은 워터마크나 메타데이터와 같은 단일 기술만으로는 표시의 효과성과 견고성을 동시에 확보하기 어렵다고 진단하고, C2PA 메타데이터와 워터마킹, 가시적 라벨을 결합한 복합 표시 체계를 권고하였다. 아울러 콘텐츠를 생성하는 제공자와 이를 공개하는 배포자의 의무를 단계별로 구분하고, 자율적 행동강령인 'Code of Practice' 서명 여부에 따라 감독 부담을 달리 부과하는 차등 감독 구조를 도입하였다. 이에 따라 표시 의무는 단순한 라벨 부착을 넘어 기술 도입과 검증, 운영 체계 정비를 요구하는 과제로 확대되고 있으며, 사업자 간 대응 역량의 차이가 규제 부담과 시장 경쟁력 격차로 이어질 전망이다.

산업 AI 학습 데이터 저작권 관리를 위한 블록체인 기반 컴플라이언스 체계 구축 논의

▶ 생성형 AI 확산으로 대규모 학습 데이터 활용이 일반화되면서, AI 모델이 어떤 데이터를 학습했으며 해당 데이터에 대한 이용 권한을 적법하게 확보했는지가 투자·M&A 과정의 핵심 쟁점으로 부각되고 있다. 사후 탐지 중심의 저작권 관리 방식만으로는 한계가 있다는 인식 하에, 학습 단계에서부터 데이터 출처·이용 권한·사용 이력을 기록·검증할 수 있는 관리 체계의 필요성이 커지고 있다. 이러한 흐름 속에서 블록체인은 데이터를 항목별로 시간순으로 기록하고 위·변조를 구조적으로 차단하는 특성을 바탕으로, 컴플라이언스 인프라로서의 활용 가능성이 논의되고 있다.

산업 비가시적 워터마킹 확산과 이미지 저작권·진본성 검증의 변화

▶ 디지털 이미지 유통 환경에서 포맷 변환·재인코딩 등 일상적 가공으로 메타데이터가 소실되는 구조적 한계가 부각되면서, 이미지 자체에 식별자를 내재하는 비가시적 워터마킹이 저작권 집행과 진본성 검증의 기본 관리 수단으로 확산되는 흐름이 나타나고 있다. 독일의 이미지 에이전시인 디피에이 픽처 얼라이언스가 이미지 보호 기술 기업인 아이마태그와의 계약을 통해 이미지 카탈로그에 비가시적 워터마킹을 단계적으로 적용하기 시작한 사례는 해당 기술이 선택적 보호 수단을 넘어 전문 이미지 유통 단계에서 출처 추적과 저작권 관리를 뒷받침하는 기본 관리 계층으로 편입되는 흐름을 보여준다. 나아가 촬영 단계의 C2PA 기반 출처 기록과 유통 단계의 비가시적 워터마킹을 결합하는 계층형 인프라로의 확장이 가능해지면서, 업계 공통 표준을 중심으로 한 협력 구조의 필요성이 커지고 있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 스테그닷에이아이, 비가시적 워터마킹으로 콘텐츠 유출 추적 기술 확대

▶ 스테그닷에이아이는 기업용 비가시적 워터마킹 기술과 적용 사례를 공개했다. 이 기술은 콘텐츠 내부에 사람이 알아보기 어려운 식별 정보를 삽입해, 유출이나 조작이 발생했을 때 출처와 유통 경로를 확인하는 방식으로 작동한다. 구체적으로 이미지·영상의 픽셀값을 미세하게 변형하고, 문서·오디오 등 다른 파일 형식에도 비가시적 워터마킹을 적용할 수 있다고 설명한다. 삽입된 워터마크는 자르기, 재인코딩, 스크린샷, 색상 변경 등 일반 편집 이후에도 남을 수 있도록 설계됐으며, 영상의 모든 프레임에 식별 정보를 넣어 짧은 클립이나 스크린샷만으로도 출처 식별을 시도할 수 있는 단일 프레임 워터마킹 방식도 제시됐다. 또한 C2PA 자격증명이 플랫폼 업로드나 파일 변환 과정에서 제거될 수 있는 한계를 보완하기 위해, 워터마크 내부의 원격 매니페스트 정보를 통해 출처를 다시 확인하는 구조를 제시했다.

산업 AI 클린룸 재구현과 오픈소스 라이선스 쟁점

▶ 생성형 AI가 기존 오픈소스 소프트웨어의 기능을 클린룸 방식으로 다시 구현하는 사례가 등장하면서, 저작자 표시·카피레프트 등 오픈소스 라이선스 의무를 어디까지 적용할 수 있는지가 새로운 쟁점으로 부각되고 있다. 말러스는 사람이 담당하던 사양 분석과 코드 구현 과정을 AI 봇으로 대체해 기존 라이선스 의무를 우회할 수 있다고 주장하며, AI 생성 코드의 독립성과 저작권 보호 문제를 둘러싼 논란이 이어지고 있다. 특히 코드 생성에 사용된 LLM이 어떤 자료를 학습했는지 외부에서 확인하기 어려운 구조로 인해, 기존 클린룸 법리의 핵심 전제였던 원본 코드와의 실질적 분리를 검증하기 어렵다는 지적도 제기된다. 이에 따라 기존 클린룸 관련 법리를 AI 자동화 환경에 그대로 적용하기에는 한계가 있다는 분석과 함께, AI 생성 코드의 출처와 라이선스 준수 여부를 확인할 수 있는 실무 관리 체계의 필요성이 확대되고 있다.

기술 주간 기술 동향

▶ 생성형 AI의 확산으로 디지털 워터마킹이 중요해졌지만, 기존 기술은 삽입된 정보를 제거하거나 위조(하는 공격에 취약한 한계를 보였다. 이를 해결하기 위해 제안된 PGID는 공격을 '데이터가 비정상 영역으로 이동한 상태'로 정의하는 사후 방어 프레임워크이다. 이 기술은 확산 모델의 역방향 과정에 개입하여 의도적 노이즈 제어 기법으로 왜곡된 데이터를 복원하거나 위조 여부를 판별한다. 실험 결과, PGID는 제거 공격으로 손상된 워터마크 비트 정확도를 93% 이상으로 복원했으며, 위조 공격 탐지 성능에서는 0.96 이상의 높은 AUC 값을 기록했다. 이 기술은 기존의 수동적 방어 체계의 한계를 극복하고, 훼손된 정보를 능동적으로 복원하는 새로운 방어 패러다임을 제시하며 AI 콘텐츠 저작권 보호 기술의 발전에 기여할 것으로 기대된다.

EU AI 법 표시 의무, 복합 표시와 차등 감독으로 구체화

AI 생성 콘텐츠 표시 의무 시행을 앞두고 드러난 표시 체계의 한계

• 가이드라인 초안 공개와 적용 기준 구체화

- 유럽위원회(European Commission)는 2026년 5월 8일 EU AI 법(AI Act)* 제50조에 관한 가이드라인 초안 공개를 통해, 2026년 8월 2일부터 본격 시행을 앞두고 AI 생성 콘텐츠 표시 의무의 구체적 적용 기준을 제시함
- EU AI 법 제50조는 AI 생성 콘텐츠가 기계 판독 가능(Machine-readable)한 형태로 출처 표기를 의무화하고 AI 챗봇과의 상호작용 및 딥페이크 여부 등에 대한 이용자 고지 의무를 규정한 조항임
- 이번 가이드라인 초안은 그동안 사업자가 자율적으로 해석해 온 적용 대상과 표시 방식을 조항 시행 전에 정리한 첫 공식 해석 문서에 해당함
- 다만, 기존 생성형 AI 서비스의 워터마킹 의무는 관련 기술 표준이 미비한 점을 감안하여 2026년 12월 2일까지 이행이 유예됨

* AI 법(AI Act): 유럽연합(EU)이 제정한 인공지능 규제법으로, AI 생성 콘텐츠의 표시 및 탐지 의무 등을 포함하며 2024년 8월 발효됨. AI 산출물의 워터마킹 의무는 2026년 12월 적용이 논의 중임

• C2PA 메타데이터의 보존 한계와 AI 생성 콘텐츠 표시 기능 도입의 지연

- AI 생성 콘텐츠 표시의 주요 기법 중 하나인 C2PA(Coalition for Content Provenance and Authenticity)* 메타데이터는 콘텐츠 생성 시점의 출처를 암호학적으로 기록함. 다만 플랫폼 업로드 과정의 압축이나 재인코딩으로 정보가 훼손될 수 있어, 유통 단계까지 안정적으로 보존되기 어렵다는 한계가 지적됨
- 실제 산업 조사에서도 워터마킹을 적정 수준으로 구현한 도구는 38%, 가시적 라벨링까지 적용한 도구는 18%에 그쳐, 시행 임박 시점에도 AI 생성 콘텐츠 표시 기능을 갖춘 도구가 제한적인 수준에 머무른 것으로 나타남[1)]
- 이러한 한계는 매체별 취약점 차이에서도 드러남. 텍스트 콘텐츠에 활용되는 통계적 워터마킹은 의미는 유지한 채 어휘와 문장구조만 변경하는 패러프레이즈 공격(Paraphrase Attack)에 의해 탐지력이 저하됨
- 또한, 이미지 콘텐츠의 비가시적 워터마크는 이미지를 자르거나 재인코딩하는 것만으로도 훼손될 수 있는 등, 단일 기법만으로는 모든 매체의 AI 생성 여부를 일관되게 표시하기 어려움

* C2PA(Coalition for Content Provenance and Authenticity): 디지털 콘텐츠의 출처와 진본성을 기록·검증하기 위한 기술 표준을 개발하는 국제 협의체로, 캐논(Canon) 등 주요 카메라 제조사와 뉴스 기관이 참여하고 있음

1) AI CERTS, "EU Compliance Countdown: AI Watermarking Rules", 2026.05.08., https://www.aicerts.ai/news/eu-compliance-countdown-ai-watermarking-rules/

복수 표시 기법 결합과 단계별 책임 구조의 형성

• 표시 요건: 단일 표시 방식에서 복합 표시 방식으로의 변화

- 이번 초안은 어떠한 단일 표시 방식도 효과성, 상호운용성, 견고성 및 신뢰성 등 네 가지 요건을 동시에 충족하기 어렵다고 명시함. 이는 단일 기술 중심 표시 방식의 한계를 정책 차원에서 공식화한 첫 사례임
- 유럽위원회는 사업자가 제50조 의무 이행을 입증할 수 있도록 자율적 행동강령인 Code of Practice*를 마련함. 해당 강령은 C2PA 메타데이터를 비롯한 여러 표시 방식을 결합해 적용하는 방식과, 이미지·영상·텍스트 등 매체별 라벨 표시 방안을 구체적으로 제시함
- 특히 영상 매체에 대해서는 긴 영상의 경우 시작 지점과 일정 간격마다 라벨을 반복 표시하고, 짧은 영상의 경우 시작 시점부터 재생 전반에 걸쳐 라벨을 지속 표시하도록 세부 기준을 제시함
- 가이드라인은 표시 의무의 적용 범위를 개별 사업자의 사정이 아니라 업계 전반의 기술 수준에 따라 판단하는 '기술적 실행 가능성(technical feasibility)' 기준을 제시함. 이에 따라 비용이나 인력 부족과 관계없이 기업 규모를 막론하고 동등한 수준의 식별 표시 의무를 이행해야 함을 명시함

* Code of Practice: AI 사업자가 제50조(2)·(4) 의무 준수를 입증하기 위해 활용할 수 있도록 유럽위원회가 마련 중인 자율적 행동강령

• 책임 배분: 콘텐츠 생성·공개·유통 단계별 의무 구분

- 가이드라인 초안은 콘텐츠 생성부터 이용자 노출까지의 과정을 단계별로 구분하고, 생성·공개·유통 단계별 행위자에게 부여되는 표시·고지 책임을 명시함
- AI 서비스 제공자는 생성 단계에서 AI 산출물에 워터마크나 메타데이터처럼 기계가 인식할 수 있는 표시를 삽입해야 함
- AI 생성 콘텐츠를 공개하는 배포자는 유통 단계에서 AI 시스템이 생성·조작한 이미지·음성·영상 콘텐츠가 딥페이크에 해당하는 경우, 또는 공익적 사안에 관해 대중에게 정보를 제공할 목적으로 공개되는 AI 생성·조작 텍스트인 경우 해당 콘텐츠가 AI로 생성·조작됐다는 사실을 표시해야 함
- 제3자가 생성한 콘텐츠를 단순히 전달하는 온라인 플랫폼은 AI 서비스를 직접 운영하거나 통제하지 않기 때문에 배포자에 해당하지 않음. 다만 생성 단계에서 적용된 표시와 라벨이 유통 과정에서 훼손되거나 사라지지 않도록 적절한 조치를 취할 것이 별도로 권고됨

[표1] AI 생성 콘텐츠 표시 의무의 주체별 구분

주체	단계	주요 의무	적용 조항
제공자	생성 단계	상호작용 AI 고지, 산출물에 기계 판독 표시	50(1), 50(2)
배포자	노출 단계	감정인식·생체분류 고지, 딥페이크·공익 텍스트 라벨링	50(3), 50(4)
단순 전송자 (온라인 플랫폼)	유통 단계	의무 아님 (생성 단계에서 적용된 표시 보존 등 적절한 조치 권고)	권고

출처: Jadzia Pierce 외 2인, "10 Takeaways: European Commission Draft Guidelines on AI Transparency under the EU AI Act", Inside Global Tech, 2026.05.12., https://www.insideglobaltech.com/2026/05/12/10-takeaways-european-commission-draft-guidelines-on-ai-transparency-under-the-eu-ai-act/

• 이행 강제: Code of Practice 기반 자율 준수와 차등 감독

- Code of Practice에 서명한 사업자는 유럽 위원회와 관할 시장감시당국 (competent market surveillance authorities)이 강령 이행 여부를 감독함.
- 반면 서명하지 않은 사업자는 자체적으로 마련한 조치가 Code of Practice와 어떻게 다른지 별도의 차이 분석 자료를 제출해 의무 이행 수준을 입증해야 함
- 이에 따라 Code of Practice는 형식상 자율적 도구이나, 미서명 사업자에게 추가 입증 부담을 부과함으로써 사실상 가입 유인을 제공하는 실질적 준수 장치로 기능함

표시 기술 표준 경쟁과 사업자 대응 과제

• 표시 기술 표준 경쟁과 사업자별 대응 격차 확대

- EU AI 법 제50조 시행은 AI 생성 콘텐츠 표시 의무를 단순 고지 수준을 넘어 기술 표준 경쟁의 문제로 확장시킴. 이에 따라 C2PA 기반 메타데이터, 워터마킹, 가시적 라벨 등 복수의 표시 방식을 어떤 조합으로 적용할지가 향후 준수 체계의 핵심 쟁점으로 부상함
- 다만 결합 적용 방식이 아직 확정되지 않은 만큼, 국제표준화기구(International Organization for Standardization, ISO)의 후속 표준 논의가 향후 가이드라인 보완 방향을 좌우할 주요 변수로 거론됨
- 사업자 입장에서는 복수의 표시 방식을 도입하는 데 그치지 않고, 각 표시 방식이 콘텐츠 생성 이후 유통 과정에서도 유지되는지 검증해야 하는 부담이 확대됨
- 특히 표시 기능을 갖춘 도구가 제한적인 상황에서 결합 적용 요건이 시행될 경우, 사업자는 표시 기술 도입, 유지 여부 검증, 이행 자료 작성, 감독 대응을 동시에 부담해야 하며, 이러한 부담은 대형 사업자보다 중소 사업자에게 더 크게 작용할 가능성이 큼
- 결국 EU AI 법의 표시 의무 시행은 단순한 라벨 부착을 넘어 기술 도입·검증·운영 체계 정비를 요구하는 과제로 전환되고 있음. 향후 사업자 간 대응 역량 차이는 규제 준수 비용뿐 아니라 콘텐츠 신뢰성 확보와 시장 경쟁력에도 영향을 미칠 전망임

참고문헌

- European Commission, "Code of Practice on marking and labelling of AI-generated content", 2026.05.22., https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content
- European Commission, "Consultation on the draft guidelines on transparency obligations under the AI Act", 2026.05.08., https://digital-strategy.ec.europa.eu/en/consultations/consultation-draft-guidelines-transparency-obligations-under-ai-act
- Jadzia Pierce 외 2인, "10 Takeaways: European Commission Draft Guidelines on AI Transparency under the EU AI Act", Inside Global Tech, 2026.05.12., https://www.insideglobaltech.com/2026/05/12/10-takeaways-european-commission-draft-guidelines-on-ai-transparency-under-the-eu-ai-act/
- Dan Whitehead 외 4인, "EU legislators agree to delay for high-risk AI rules", Hogan Lovells, 2026.05.07., https://www.hoganlovells.com/en/publications/eu-legislators-agree-to-delay-for-highrisk-ai-rules
- AI CERTS, "EU Compliance Countdown: AI Watermarking Rules", 2026.05.08., https://www.aicerts.ai/news/eu-compliance-countdown-ai-watermarking-rules/

AI 학습 데이터 저작권 관리를 위한 블록체인 기반 컴플라이언스 체계 구축 논의

AI 학습 데이터 저작권 관리 체계에 블록체인 도입 요구 확대

- **AI 학습 데이터 관리 체계 변화 필요성 제기**

- 최근 AI 기업에 대한 투자 및 인수합병이 확대되는 과정에서 AI 모델이 어떤 데이터를 학습했으며 해당 데이터에 대한 이용 권한을 확보했는지가 주요 쟁점으로 나타남
- 이는 라이선스 미확보 데이터나 권리 관계가 불명확한 콘텐츠가 학습에 포함되어 향후 저작권 분쟁이나 손해배상 책임으로 이어지는 것을 방지하기 위한 목적임
- 따라서 AI 모델의 결과물을 사후 분석해 저작권 침해 여부를 판단하는 방식만으로는 한계가 있으며, 학습 단계에서부터 이용 권한, 사용 이력을 기록·검증할 수 있는 관리 체계의 필요성이 커지고 있음
- 이러한 배경에서 블록체인(Blockchain)은 AI 학습데이터의 권리관계를 추적하고 입증하는 지식재산 감사 및 컴플라이언스 인프라*로 주목받고 있음

* 컴플라이언스 인프라(compliance infrastructure): 기업이 관련 법령·계약상 의무·라이선스 조건 등을 준수하고 있음을 체계적으로 기록·관리·입증할 수 있도록 설계된 운영 체계

지식재산 보호를 위한 블록체인 기반 IP 감사 인프라의 적용

- **학습 데이터 출처·라이선스 이력의 기록 및 추적 수단으로서의 블록체인 활용**

- AI 기업의 투자와 인수합병 과정에서 지식재산 실사(IP diligence)*는 어떤 데이터가 모델 학습에 사용되었는지, 해당 데이터에 대한 적법한 권리를 확보했는지를 핵심 항목으로 검토함
- 이 과정에서 블록체인 기반의 학습 데이터 이력 관리는 데이터의 출처, 수집 시점, 이용권한 등을 이력이 남는 방식으로 기록하고 보존하는 인프라로서 데이터 사용의 적법성을 입증할 수 있는 근거로 활용 가능함
- 이는 블록체인이 데이터 항목별로 출처, 수집 시점, 이용 권한 등을 시간순으로 기록할 수 있어 학습 데이터의 권리 관계를 추적하는 데 용이하기 때문임
- 또한 기록된 데이터를 사후에 임의로 변경하거나 삭제하기 어려운 구조적 특성을 지니고 있어, 학습 데이터의 출처와 사용 이력에 대해 기록 이후의 변경 가능성을 낮출 수 있음
- 이에 따라 블록체인 기반 학습 데이터 이력 관리는 컴플라이언스 인프라로서, 지식재산 실사 과정에서 필요한 확인 절차를 간소화하고 계약 체결 이후 발생할 수 있는 손해배상 청구 위험을 줄이는 기능을 할 수 있다는 평가가 제기됨

* 지식재산 실사(IP diligence): 기업 투자·인수합병 등 주요 계약 전에 상대 기업이 보유한 기술·콘텐츠·데이터의 권리 귀속 관계와 잠재적 법적 리스크를 검토하는 절차

블록체인 기반 학습 데이터 컴플라이언스 체계의 활용 가능성

• 강화되는 학습 데이터 출처 검증 요구와 블록체인 활용 논의

- AI 기업에 대한 투자와 인수합병 과정에서 일부 인수 계약 실무에서는 AI 모델 개발자가 자사 모델에 포함된 지식재산에 대해 기존보다 엄격한 수준의 진술 및 보증을 요구받는 사례가 나타나고 있음[1)]
- 이는 기존에 요구되던 단순히 '학습 데이터셋이 제3자의 지식재산을 침해하지 않는다'는 수준의 보증에서 한층 강화된 개념임
- 최근에는 이러한 계약과 진술 과정에서, 저작권 및 라이선스 준수 여부를 검증할 수 있는 감사 가능한 컴플라이언스 체계가 존재하는지에 대한 검토가 함께 이루어지는 추세임
- 이러한 흐름 속에서 블록체인을 학습 데이터 컴플라이언스 인프라로 활용하자는 논의가 제기되고 있으며, 향후 실무에서 활용되기 위해서는 감사 가능한 데이터 추적 구조의 신뢰성, 라이선스 정보의 정합성 등이 후속 과제로 제기됨

[표1] 블록체인 기반 학습 데이터 저작권 관리 논의 요약

구분	내용
배경	• 생성형 AI 확산에 따른 대규모 학습 데이터 활용 일반화 • 학습 데이터 저작권·라이선스 적법성 문제의 주요 쟁점화 • AI 기업 투자·인수합병 과정에서 데이터 출처·이용 권한 확보 여부의 핵심 실사 항목화
블록체인 활용 논의	• 사후 탐지 중심 저작권 관리 방식의 한계 제기 • 학습 단계에서의 데이터 출처·이용 권한·사용 이력 기록·검증 필요성 부각 • IP 감사 및 컴플라이언스 인프라로서 블록체인 활용 논의 확산
블록체인 특징	• 데이터 항목별 출처, 수집 시점, 라이선스 조건 및 사용 이력의 시간순 기록 • 사후 변경·삭제가 어려운 구조적 특성으로 위·변조 가능성 차단 • 사전 구축된 기록 인프라를 통한 데이터 적법성의 외부 입증 근거 확보
시사점	• 학습 데이터의 합법적 출처·권리 관리 역량을 기업 가치 평가 요소로 반영하려는 흐름의 확산 • 일부 계약에서 감사 가능한 통제 체계의 존재 여부가 거래 조건에 반영 • 블록체인의 학습 데이터 컴플라이언스 인프라 활용 논의 확산

출처: 참고문헌 종합하여 재구성

참고문헌

• Baker Botts, “Blockchain, AI Training Data, and Protecting Intellectual Property in the Next Deal”, 2026.05.06., https://www.bakerbotts.com/thought-leadership/publications/2026/may/blockchain-ai-training-data-and-protecting-intellectual-property-in-the-next-deal

1) Baker Bootts, “Blockchain, AI Training Data, and Protecting Intellectual Property in the Next Deal”, 2026.05.06., https://www.bakerbotts.com/thought-leadership/publications/2026/may/blockchain-ai-training-data-and-protecting-intellectual-property-in-the-next-deal

비가시적 워터마킹 확산과 이미지 저작권·진본성 검증의 변화

기존 이미지 저작권 관리 방식의 한계

• 메타데이터 기반 출처 관리의 한계

- 디지털 이미지는 포맷 변환·재인코딩(re-encoding)* 등 일상적인 가공 과정에서 파일에 포함된 메타데이터(metadata)**가 삭제될 수 있으며, 이에 따라 원본 출처와 저작권 정보가 유통 과정에서 소실될 가능성이 있음
- 이미지 유통 사업자의 경우 취급하는 이미지가 수억 건에 달해 이미지를 유통하는 환경에서는 개별 파일 단위로 출처와 저작권 정보를 추적하는 방식에 현실적 한계가 존재함
- 더불어 AI 기술의 발전으로 진본 이미지와 AI 생성 이미지를 육안으로 구분하기 어려워지면서, 이미지 유통 이후에도 출처와 진본성을 검증할 수 있는 별도 관리 수단의 필요성이 확대되고 있음

* 재인코딩(re-encoding): 영상·음성을 다시 압축·변환하는 작업으로, 화질 조정이나 포맷 변경 등에 활용됨

** 메타데이터(metadata): 이미지 파일에 내장된 촬영 일시, 저작권자, 위치 정보 등의 부가 정보로, 포맷 변환이나 편집 과정에서 삭제되거나 변경될 수 있음

• 에이전시·언론사의 저작권 집행 및 진본성 검증

- 온라인 뉴스 등 이미지가 유통되는 현장에서는 저작권 위반 탐지와 법적 분쟁 대응을 위해 이미지별 식별 정보를 확보해야 할 필요성이 커지고 있으며, 사후 모니터링 중심의 기존 방식만으로는 이러한 요구를 충족하기 어려움
- 보도 사진의 진위 검증 요구가 확대되면서, 보도 이후의 사후 검증을 넘어 편집·배포 등 이미지 유통 단계에서 검증 수단을 내재화하는 관리 방식에 대한 논의가 확대되고 있음
- C2PA(Coalition for Content Provenance and Authenticity)* 등 국제 표준을 중심으로 디지털 콘텐츠의 출처 검증 체계에 관한 논의가 확산되며, 이미지 출처 관리가 기존의 기업별 자체 운영 방식에서 벗어나 공통 표준 기반의 검증 체계로 이동하는 흐름이 나타남

* C2PA(Coalition for Content Provenance and Authenticity): 디지털 콘텐츠의 출처와 진본성을 기록·검증하기 위한 기술 표준을 개발하는 국제 협의체로, 캐논(Canon) 등 주요 카메라 제조사와 뉴스 기관이 참여하고 있음

비가시적 워터마킹의 구조와 계층적 확산

• 비가시적 워터마킹과 출처 기록 방식의 차이

- 비가시적 워터마킹(invisible watermarking)은 이미지의 픽셀 단위에 육안으로 식별되지 않는 정보를 삽입하는 기술로, 사이즈 변환·자르기·포맷 변환·재인코딩 이후에도 식별자가 유지되어 유통 경로 전반에 걸친 추적이 가능함

- C2PA의 콘텐츠 크리덴셜(Content Credentials)*이 촬영 일시, 장비 정보, 편집 이력 등 이미지의 출처 정보를 메타데이터 형태로 기록·서명하는 방식이라면, 비가시적 워터마킹은 메타데이터가 제거된 이후에도 이미지 자체에 식별 정보를 유지하는 방식임
- 이에 따라 콘텐츠 크리덴셜은 이미지의 출처와 편집 이력을 설명하는 역할을 하고, 비가시적 워터마킹은 이미지 파일이 변형되거나 메타데이터가 삭제된 이후에도 추적 가능성을 유지하는 보완적 역할을 수행할 수 있음

* 콘텐츠 크리덴셜(Content Credentials): 이미지의 촬영 정보, 제작자, 편집 이력 등 출처 관련 정보를 기록해 콘텐츠의 생성·수정 과정을 확인할 수 있도록 하는 디지털 표시 체계

• 이미지 유통 단계의 적용 사례

- 독일 대표 사진 에이전시인 디피에이 픽처 얼라이언스(dpa Picture Alliance)는 2026년 5월 이미지 보호 기술 기업인 아이마태그(IMATAG)와의 계약을 통해 이미지 카탈로그 전반에 걸쳐에 비가시적 워터마킹 기술을 단계적으로 적용하기 시작함
- 해당 계약은 약 3억 건의 이미지와 약 350 개 파트너 에이전시 네트워크를 대상으로 하며, 각 이미지에 라이선스와 연결되는 식별 정보를 삽입하여 저작권 집행, 무단 사용 모니터링 및 진본성 검증을 지원하는 방식으로 설계됨[1)]
- 이는 비가시적 워터마킹이 선택적 보호 수단을 넘어, 전문 이미지 유통 단계에서 출처 추적과 저작권 관리를 뒷받침하는 기본 관리 계층으로 편입되는 흐름을 보여주는 사례로 평가됨

• 촬영 및 유통 단계의 관리 방식 결합

- 앞선 사례가 이미지 유통 단계에서 비가시적 워터마킹을 적용한 경우라면, 캐논(Canon)의 진본성 이미징 시스템(Authenticity Imaging System, 이하 AIS)*는 이미지가 촬영되는 시점부터 출처 정보를 기록·관리함
- AIS는 촬영 시점에 이미지의 출처와 처리 이력을 확인할 수 있는 C2PA 기반 정보를 생성해, 이후 편집·배포·게시 단계에서도 출처를 검증할 수 있도록 설계됨
- 이에 따라 촬영 단계에서는 C2PA 기반 출처 기록을 남기고, 유통 단계에서는 비가시적 워터마킹을 통해 이미지 자체의 추적성을 보완하는 구조가 가능해짐
- 이러한 결합은 이미지 저작권·진본성 관리가 단일 기술에 의존하는 방식에서 벗어나, 촬영, 유통 및 검증 단계별 보호 수단을 함께 적용하는 계층형 인프라로 확장될 수 있음을 보여줌

* 진본성 이미징 시스템(Authenticity Imaging System, AIS): 캐논이 C2PA 표준에 기반해 개발한 출처 관리 솔루션으로, 촬영 시점부터 게시 단계까지 이미지의 출처 이력을 검증 가능한 형태로 유지하는 기능을 제공함

시사점: 이미지 저작권 관리 인프라의 변화

• 저작권 집행 수단으로서의 워터마킹

- 비가시적 워터마킹은 이미지의 변형 이후에도 식별 정보를 유지할 수 있다는 점에서, 저작권 분쟁 발생 시 저작권 귀속을 입증하는 수단으로 활용될 가능성이 커지고 있음
- EU AI 법(AI Act)*에 따른 AI 산출물에 대한 워터마킹 의무 적용이 2026년 12월로 잠정 합의되면서, 워터마킹 기술은 저작권 보호 수단을 넘어 AI 산출물의 출처 검증과 규제 준수를 지원하는 핵심 인프라로 역할이 확장되는 흐름을 보임

1) IMATAG, "dpa Picture Alliance Selects IMATAG as Preferred Invisible Watermarking Technology", IMATAG, 2026.05.06., https://www.imatag.com/blog/dpa-picture-alliance-selects-imatag-as-preferred-invisible-watermarking-technology

- 이에 따라 이미지 유통 사업자 입장에서는 워터마킹 기술 도입이 선택적 보호 수단이 아니라, 규제 대응과 출처 관리 체계를 위한 운영 요건의 일부로 자리 잡을 가능성이 있음

  * AI 법(AI Act): 유럽연합(EU)이 제정한 인공지능 규제법으로, AI 생성 콘텐츠의 표시 및 탐지 의무 등을 포함하며 2024년 8월 발효됨. AI 산출물의 워터마킹 의무는 2026년 12월 적용이 논의 중임

• 촬영·유통·검증 단계의 기술 연계 필요성

- C2PA 기반 출처 기록과 비가시적 워터마킹은 각각 촬영 단계와 유통 단계에서 서로 다른 역할을 수행하며, 단일 기술만으로는 이미지의 저작권·진본성을 전 과정에 걸쳐 보장하기 어려움
- 그러나 카메라 제조사, 이미지 유통 사업자 및 검증 단계의 언론사·플랫폼이 각자의 단계에서 호환 가능한 출처 검증 체계를 운영하고 단계 간 정보 연동이 이루어지는 경우, 이미지 유통 전 과정에서 저작권 관계를 일관되게 추적할 수 있는 기반이 구축될 것으로 분석됨
- 향후 이미지 저작권 관리의 실효성은 개별 기술의 성능보다 촬영·유통·검증 단계의 관리 체계를 얼마나 유기적으로 연결하는지에 달려 있으며, 업계에서는 공통 표준을 중심으로 한 협력 구조의 필요성이 증대되고 있음

참고문헌

- IMATAG, “dpa Picture Alliance Selects IMATAG as Preferred Invisible Watermarking Technology”, IMATAG, 2026.05.06., https://www.imatag.com/blog/dpa-picture-alliance-selects-imatag-as-preferred-invisible-watermarking-technology
- Canon Inc., “Canon Introduces C2PA—Compliant Authenticity Imaging System for News Organizations”, Canon Global, 2026.05.11., https://global.canon/en/news/2026/20260511.html
- Foo Yun Chee, “EU countries, lawmakers clinch provisional deal on watered-down AI rules”, Reuters, 2026.05.07., https://www.reuters.com/world/eu-countries-lawmakers-strike-provisional-deal-watered-down-ai-rules-2026-05-07

스테그닷에이아이(Steg.AI), 비가시적 워터마킹으로 콘텐츠 유출 추적 기술 확대

기업용 비가시적 워터마킹의 등장과 적용 확산

• 비가시 식별 정보를 결합한 워터마킹 솔루션의 공개

- 미국 워터마킹 솔루션 기업 스테그닷에이아이(Steg.AI)는 2026년 5월 4일 뉴욕대 창업 인터뷰 시리즈에서 자사 비가시적 워터마킹 기술과 기업 적용 사례를 공개함
- 스테그닷에이아이는 첫 기업형 고객인 미국 오디오 기기 기업 소노스(Sonos)를 통해 마케팅 자산 보호 분야에 비가시적 워터마킹 기술을 적용했으며, 이후 라이선스 기반 완구 제조사 펑코(Funko), 영화·TV·게임 스튜디오, 소비자 가전 기업 등으로 고객군을 확대함

• 기업 보안 환경에 맞춘 세 가지 제공 방식의 마련

- 스테그닷에이아이 솔루션은 웹에서 바로 사용하는 방식, 기업 내부 시스템과 연동하는 API 방식, 자체 서버에 설치해 외부 클라우드 업로드 없이 운용할 수 있는 온프레미스(on-premise) SDK 방식으로 제공됨
- 이러한 제공 구조는 출시 전 마케팅 자료, 라이선싱 이미지, 영상 콘텐츠처럼 외부 공유가 잦고 유출 위험이 큰 디지털 자산을 관리하는 기업 환경에 부합함

비가시적 워터마킹의 작동 방식과 C2PA 연계 구조

• 콘텐츠 내부에 보이지 않는 식별 정보의 삽입 메커니즘

- 스테그닷에이아이는 자체 개발한 딥러닝 모델을 활용해 이미지·영상에는 사람이 알아보기 어려운 수준으로 픽셀값을 변형하고, 문서·오디오 등 다른 파일 형식에도 비가시적 워터마킹을 적용할 수 있다고 설명함
- 동일 콘텐츠라도 공유 대상에 따라 서로 다른 식별 정보를 삽입할 수 있어, 같은 홍보 영상을 여러 파트너사에 전달할 때 각 파일에 서로 다른 워터마크가 들어감
- 워터마크에는 수신자, 콘텐츠 사용 권한, 출처 등 사용자가 지정한 정보가 포함될 수 있으며, 콘텐츠의 외형을 바꾸지 않으면서 파일 내부에 추적 단서를 남기는 구조임

• 편집 이후에도 남아있는 워터마크 설계 구조

- 스테그닷에이아이의 자체 설명에 따르면, 비가시적 워터마크는 콘텐츠 내부에 삽입되는 방식으로 설계됨. 이에 따라 자르기, 크기 조정, 재인코딩, 스크린샷 및 색상 변경 등 일반적인 편집 이후에도 워터마크가 유지될수 있다고 설명함
- 실제 유출 콘텐츠는 원본 파일 그대로 유통되기보다 캡처, 재저장, 압축 및 편집을 거친 형태로 퍼지는 경우가 많아, 파일에 내장된 고유 식별 정보를 판독하는 방식이 출처 추적 수단으로 제시됨
- 또한 스크린샷 1장이나 짧은 클립만 확보된 경우에도 출처 식별을 시도할 수 있도록, 영상의 모든 프레임에 고유 식별 정보를 삽입하는 단일 프레임 워터마킹(single-frame watermarking)* 방식이 함께 제시됨

* 단일 프레임 워터마킹(single-frame watermarking): 영상의 각 프레임에 식별 정보를 넣어, 1프레임이나 스크린샷만으로도 추적할 수 있도록 하는 방식

[그림 1] 스테그닷에이아이의 비가시적 워터마킹 적용 예시: 왼쪽 원본, 오른쪽 워터마크 삽입 이미지

출처: Steg,ai, "Content Verification and Deepfake Detection", 2026.05.18. 접속 기준, https://steg.ai/products/content-authentication/

• C2PA 자격증명과 비가시적 워터마크의 보완 구조

- C2PA 자격증명은 콘텐츠의 생성 정보와 편집 이력을 메타데이터 형태로 기록·검증할 수 있도록 설계된 산업 표준 체계임. 다만 플랫폼 업로드나 파일 변환 과정에서 관련 정보가 제거될 수 있다는 한계가 있음
- 스테그닷에이아이는 워터마크 내부에 연결된 원격 매니페스트(manifest)* 정보를 통해 C2PA 자격증명 손실 상황에서도 출처 정보를 다시 확인할 수 있다고 설명함
- 이는 C2PA 자격증명만으로 출처 정보를 확인하기 어려운 상황에서, 콘텐츠 내부의 비가시적 워터마크를 함께 활용하는 보완 구조로 볼 수 있음
- 다만 스테그닷에이아이가 제시한 검증 시나리오에서는 콘텐츠가 크게 편집되어 C2PA 자격증명과 워터마크가 모두 확인되지 않는 경우, 해당 콘텐츠가 검증되지 않은 것으로 처리됨

* 매니페스트(manifest): 콘텐츠의 출처, 생성자, 편집 이력 등을 담은 정보 묶음

콘텐츠 보안 체계의 사후 추적 기능 강화

• 기존 보안 도구의 한계와 워터마킹 보완 필요성

- 기존 콘텐츠 보안 체계는 암호화, 접근 통제, 데이터 손실 방지처럼 유출을 사전에 차단하는 방식에 초점을 두었음
- 그러나 실제 유출 발생 시 해당 콘텐츠가 유출된 정확한 유출 주체 및 경로를 파악하기 어려운 구조적 한계가 존재함
- 스테그닷에이아이의 콘텐츠 유출 방지 솔루션 설명에 따르면, 콘텐츠 유출은 출시 동력 약화, 매출 손실, 복구 비용 증가로 이어질 수 있으며, 유출 사고 1건당 평균 약 1,000만 달러 규모의 손실과 넷플릭스(Netflix) '아케인(Arcane)' 시즌 2 유출 사고가 관련 사례로 제시됨
- 스테그닷에이아이는 이러한 한계를 보완하기 위해 유출 이후 추적 기능을 제공하며, 콘텐츠가 보안 시스템을 벗어난 뒤에도 파일 내부에 남은 식별 정보를 판독해 어떤 사본이 외부에 노출됐는지 확인하는 방식임
- 이에 따라 기업은 워터마킹을 기존 보안 도구의 대체 수단이 아니라, 유출 이후 출처 확인을 위한 보완 계층으로 설계할 필요가 있음
- 출시 전 마케팅 자료, 영상 파일, 디자인 시안, 내부 발표 자료처럼 외부 공유가 잦고 유출 위험이 높은 콘텐츠부터 우선 적용하는 방식의 검토가 요구됨

참고문헌

- Abhi Das, "Beyond the Bench: Steg.AI's Forensic Watermarking Secures the Provenance of Digital Content", NYU Entrepreneurship, 2026.05.04., https://entrepreneur.nyu.edu/blog/2026/05/04/beyond-the-bench-steg-ais-forensic-watermarking-secures-the-provenance-of-digital-content/
- Steg,ai, "Stop content leaks. Defend your IP.", 2026.05.18. 접속 기준, https://steg.ai/products/leak-protection/
- Steg,ai, "Content Verification and Deepfake Detection", 2026.05.18. 접속 기준, https://steg.ai/products/content-authentication/

AI 클린룸 재구현과 오픈소스 라이선스 쟁점

오픈소스 라이선스와 클린룸 설계

• 오픈소스 라이선스의 작동 원리

- 저작권법은 아이디어 자체보다 이를 구체적으로 표현한 결과물을 보호 대상으로 삼으며, 소프트웨어 분야에서는 특정 기능 자체보다 해당 기능을 구현한 코드가 보호 대상이 됨
- 오픈소스 라이선스는 코드의 자유로운 이용, 수정 및 배포를 허용하되, 라이선스 유형에 따라 저작자 표시, 동일조건변경허락(Share-Alike)*, 소스코드 공개 등의 의무를 부과하는 방식으로 운영되어 왔음
- 이러한 라이선스 체계는 코드의 자유로운 활용을 보장하는 동시에, 수정·개선 결과가 다시 공유되도록 함으로써 오픈소스 생태계가 유지되는 기반으로 작동해 왔음

* 동일조건변경허락(Share-Alike): 저작물을 수정·변형해 배포할 경우, 원저작물과 동일한 라이선스 조건으로 공개하도록 요구하는 라이선스 조항

• 클린룸 설계 방식의 등장 배경

- 클린룸 설계(Clean Room Design)*는 IBM PC와 호환되는 시스템을 개발하려 했던 경쟁사들이 고안한 역설계(reverse engineering)** 방식으로, 원본 코드를 분석해 사양을 정리하는 팀과 이를 바탕으로 새 코드를 구현하는 팀을 분리하는 구조를 취함
- 이 방식은 원본 코드를 직접 복제하지 않고 별도의 구현 과정을 거치면 저작권 침해 위험을 낮출 수 있다는 논리에 기반하며, 이후 소프트웨어 역설계 분야에서 주요한 법적·실무적 방식으로 활용됨
- 다만 기존 클린룸 방식은 사양 분석과 코드 구현 과정을 분리해야 해 별도 인력과 장기간 작업이 필요했으며, 높은 비용 부담으로 인해 일반 개발 환경에서는 제한적인 방식으로 인식됨

* 클린룸 설계(Clean Room Design): 원본 코드를 직접 본 사람과 새 코드를 작성하는 사람을 분리해, 기존 소프트웨어와 같은 기능을 하면서도 원본 코드를 복제하지 않도록 개발하는 방식

** 역설계(reverse engineering): 완성된 제품이나 소프트웨어를 분석해 작동 방식이나 구조를 파악하는 과정

AI 클린룸 재구현의 등장과 쟁점

• 말러스의 작동 구조와 라이선스 우회 주장

- 말러스(Malus.sh)*는 AI를 활용한 오픈소스 소프트웨어를 클린룸 방식으로 재구현하는 서비스로, 기존 클린룸 설계에서 개발자가 수행하던 사양 분석과 코드 구현 과정을 각각 사양 분석 봇과 코드 생성 봇으로 대체하는 구조를 취하고 있음
- 기존 클린룸 설계가 원본 코드 분석과 구현 과정을 조직적으로 분리하는 방식이었다면, 말러스는 이러한 절차를 단일 서비스 안에서 자동 수행할 수 있다는 점에서 차이를 보임

- 말러스는 카피레프트(Copyleft)** 의무와 저작자 표시 없이도 새 코드를 사용할 수 있다고 주장하고 있으나, 코드 생성에 사용된 대규모 언어모델(Large Language Model, LLM)***이 어떤 자료를 학습했는지 외부에서 확인하기 어려운 구조임
- 이에 따라 해당 서비스가 원본 소프트웨어와 실제로 분리된 상태에서 작동하는지 검증하기 어렵다는 지적이 제기되며, 기존 클린룸 방식의 핵심 전제였던 원본 코드와의 실질적 분리가 AI 환경에서도 유지되는지에 대한 논란이 이어짐
- 또한 말러스는 오픈소스 라이선스 체계의 맹점을 지적하는 성격의 프로젝트라고 설명하면서도 실제 유료 상용 서비스 형태로 운영 중이며, 보도를 통해 클로드 코드(Claude Code)****를 활용해 기존 오픈소스 라이브러리를 재구현한 사례도 알려지면서[1] 원저작자 표시 의무와 AI 기반 클린룸 재구현의 정당성을 둘러싼 논란이 이어지고 있음

* 말러스(Malus.sh): AI 봇을 이용해 기존 오픈소스 소프트웨어와 같은 기능을 수행하는 코드를 새로 생성한다고 주장하는 서비스로, '오픈소스 라이선스 의무로부터의 해방'을 표방함

** 카피레프트(Copyleft): 공개된 코드를 활용해 새 프로그램을 만들 경우, 그 결과물도 다시 공개·공유하도록 요구하는 오픈소스 라이선스 원칙

*** 대규모 언어모델(Large Language Model, LLM): 대규모 텍스트 데이터로 학습된 AI 언어 모델로, 코드 생성·번역·요약 등 다양한 언어 처리 작업에 활용됨

**** 클로드 코드(Claude Code): 앤트로픽(Anthropic)이 제공하는 AI 코딩 도구로, 사용자의 지시에 따라 코드를 작성·수정하거나 기존 코드의 구조를 분석하는 데 활용됨

• AI 생성 코드의 독립성과 권리 문제

- AI 클린룸 방식으로 만들어진 코드가 원본과 독립된 새로운 저작물인지에 대해서는 법적 해석이 엇갈리고 있으며, LLM의 학습 데이터에 기존 오픈소스 코드가 포함됐을 가능성이 있다는 점에서 산출물의 독립성을 둘러싼 논란이 이어짐
- 미국 저작권법 해석상 인간의 창작적 기여가 충분히 확인되지 않는 경우, AI가 생성한 코드는 저작권 보호 대상이 되기 어려우며, 이 경우 AI를 통해 생성된 코드를 이용자가 법적으로 온전히 소유하거나 독점적으로 보호받기 어려울 수 있음
- 나아가 원본 코드를 직접 복제하지 않았더라도 생성된 코드가 기존 특허가 보호하는 기술적 방법이나 처리 절차를 구현할 경우 별도의 지식재산권 문제가 발생할 수 있어, AI 클린룸 방식의 법적 리스크는 저작권 문제에만 한정되지 않는 것으로 해석됨

• 저작권법 해석의 불확실성과 산업 영향

- 2021년 미국 연방대법원은 구글(Google)이 오라클(Oracle)의 API를 활용한 사례를 공정이용(Fair Use)*으로 인정한 바 있으나[2], 이러한 판단을 AI 기반 소프트웨어 재구현 사례에 동일하게 적용할 수 있는지는 아직 명확하지 않음
- 또한 말러스와 유사한 서비스를 이용했다는 사실 자체가 기존 소프트웨어를 대체할 목적으로 AI를 활용했다는 정황으로 해석될 가능성도 제기됨
- 특히 AI 기반 재구현 방식은 비교적 낮은 비용만으로도 기존 서비스의 핵심 기능을 손쉽게 구현할 수 있는 환경을 조성하며, 서비스형 소프트웨어(Software-as-a-Service)** 업계에서는 기능 모방과 경쟁 심화에 대한 우려가 확대되고 있음

* 공정이용(Fair Use): 저작권자의 허락 없이도 비평, 연구, 교육, 보도 등 일정한 목적과 조건 아래 저작물을 제한적으로 이용할 수 있도록 허용하는 법적 원칙

** 서비스형 소프트웨어(Software-as-a-Service): 프로그램을 직접 설치하지 않고 인터넷을 통해 구독형으로 이용하는 소프트웨어 서비스

1) Victor Tangermann, "Devious New AI Tool 'Clones' Software So That the Original Creator Doesn't Hold a Copyright Over the New Version", Futurism, 2026.04.26., https://futurism.com/artificial-intelligence/malus-clones-software-copyright
2) Jonathan Bailey, "Cleanroom as a Service: AI-Washing Copyright", Plagiarism Today, 2026.03.24., https://www.plagiarismtoday.com/2026/03/24/cleanroom-as-a-service-ai-washing-copyright/

시사점: AI 클린룸 재구현의 대응 과제

• AI 클린룸 방식의 법적 불확실성

- 기존 클린룸 방식은 원본 코드를 분석하는 팀과 새 코드를 작성하는 팀을 분리해 진행했기 때문에, 새 코드가 원본을 직접 복제하지 않았는지 확인할 수 있는 절차가 비교적 명확했음
- 그러나 AI가 이 과정을 대신할 경우, 코드 생성에 사용된 AI 모델이 원본 소프트웨어나 유사한 오픈소스 코드를 학습했는지 외부에서 확인하기 어려움
- 이에 따라 인간 개발자의 역할 분리를 전제로 형성된 기존 클린룸 법리를 AI 자동화 방식에도 동일하게 적용할 수 있는지는 아직 명확하지 않으며, 관련 법적 불확실성이 당분간 이어질 가능성이 있음

• AI 생성 코드의 출처·라이선스 확인 필요성

- AI를 이용해 기존 소프트웨어와 같은 기능의 코드를 새로 만드는 방식이 확산되면서, 새 코드가 기존 오픈소스 라이선스 조건을 준수하는지 확인할 수 있는 검증 절차가 중요해지고 있음
- 특히 AI가 저작자 표시 없이 기존 오픈소스 코드와 유사한 기능을 수행하는 코드를 쉽게 만들어낼 경우, 원저작자의 기여와 라이선스 조건이 제대로 반영되지 않을 수 있다는 우려도 제기됨
- 이에 따라 법 개정 논의와 별개로, 업계에서는 AI 생성 코드의 출처와 라이선스 준수 여부를 확인할 수 있는 실무 관리 체계의 필요성이 확대되고 있음
- 다만 저작권법의 기본 원칙인 아이디어와 표현의 구분 자체를 바꾸는 방식은 소프트웨어뿐 아니라 다른 저작물 전반에도 영향을 줄 수 있어, 관련 제도 논의에는 신중한 접근이 필요하다는 지적도 함께 나타남

참고문헌

- Emanuel Maiberg, “This AI Tool Rips Off Open Source Software Without Violating Copyright”, 404 Media, 2026.04.21., https://www.404media.co/this-ai-tool-rips-off-open-source-software-without-violating-copyright
- Victor Tangermann, “Devious New AI Tool ‘Clones’ Software So That the Original Creator Doesn’t Hold a Copyright Over the New Version”, Futurism, 2026.04.26., https://futurism.com/artificial-intelligence/malus-clones-software-copyright
- BeauHD, “AI Tool Rips Off Open Source Software Without Violating Copyright”, Slashdot, 2026.04.22., https://news.slashdot.org/story/26/04/22/1631212/ai-tool-rips-off-open-source-software-without-violating-copyright
- Jonathan Bailey, “Cleanroom as a Service: AI-Washing Copyright”, Plagiarism Today, 2026.03.24., https://www.plagiarismtoday.com/2026/03/24/cleanroom-as-a-service-ai-washing-copyright/

주간 기술 동향

워터마크 제거 및 위조에 대응하는 사후 가이던스 방어 기술, PGID

• AI 생성 이미지의 신뢰성 문제와 워터마크 공격에 대한 방어 필요성 대두

인공지능 기술의 발전으로 고품질 이미지 생성이 보편화되면서 저작권 침해나 허위 정보 확산과 같은 부작용에 대한 우려도 함께 증가하고 있다. 이러한 문제를 완화하고 AI 생성 콘텐츠의 출처를 명확히 하기 위한 기술적 장치로서 디지털 워터마킹의 중요성이 부각되고 있다. 이에 따라 생성형 AI 모델 자체에 저작권 정보를 내재화하여 이미지 품질 저하 없이 저작권을 보호하는 워터마킹 기술 개발이 산업계의 주요 과제로 논의되고 있다.

최근에는 확산 모델의 잠재 공간 내 노이즈(noise)에 정보를 주입하는 의미론적 워터마킹 기법은 초기 노이즈 단계부터 워터마크를 내장하여 이미지를 생성하므로 인위적인 제거가 어렵고 시각적 품질 저하가 적다는 장점을 갖는다. 하지만 이러한 견고한 워터마킹 시스템에도 불구하고, 확산 모델의 역방향 프로세스를 이용하는 새로운 형태의 공격들이 등장하며 기술적 한계를 드러내고 있다. 이러한 공격들은 워터마크 신호를 교란시켜 저작권 정보를 무력화하거나 허위 정보를 삽입하여 진위를 조작하는 문제를 일으킨다.

최근 보고된 워터마크 제거 및 위조 공격은 현재 의미론적 워터마킹이 가진 기술적 취약점을 이용하는 사례로, 워터마크가 삽입된 잠재 벡터를 워터마크가 없는 영역으로 유도하는 방식으로 탐지 과정을 회피한다. 기존의 단일 단계 역방향 탐지 방식은 교란된 신호를 복원하거나 위조 여부를 판별하지 못하여 공격에 효과적으로 대응하기 어렵다는 한계를 보인다. 이러한 문제로 워터마크의 신뢰성이 저하되어 저작권 보호 체계의 실효성에 대한 우려가 커지고 있다.

이러한 새로운 위협에 대응하기 위해 워터마크 탐지 과정의 강건성을 높이는 방어 프레임워크의 필요성이 꾸준히 제기되고 있으며, 본 보고서는 해당 공격을 대응하기 위해 고안된 사후 식별 가이던스(Progressive Guided Inversion and Denoising, 이하 PGID) 기술을 분석하고자 한다. PGID는 별도의 모델 학습 없이 적용 가능한 플러그 앤 플레이(plug and play) 방식의 방어 프레임워크로, 공격으로 변형된 잠재 벡터를 원본 상태로 복원하는 접근법을 사용한다. 이 기술은 제거된 워터마크 신호를 복구하고 위조된 이미지를 식별함으로써 AI 콘텐츠 저작권 보호 기술의 발전 방향을 제시한다는 점에서 의의가 있다.

[사례] 워터마크 공격 방어를 위한 PGID 기술

• 현 워터마킹 기술의 특징과 한계

- 현재 생성형 AI에 적용되는 워터마킹 기술은 사용자에게 보이지 않는 비가시성과 이미지 압축, 변형 등에도 정보가 유지되는 강인성 확보를 핵심 목표로 함
- 그러나 현재까지의 기술은 확산 모델의 생성 원리를 역이용하여 워터마크를 무력화하는 제거 공격(removal attack)이나, 워터마크가 없는 이미지에 허위 정보를 삽입하여 탐지 시스템을 속이는 위조 공격(forgery attack)에 취약한 한계를 보임
- 현재 출시된 대부분의 워터마킹 시스템은 공격을 받은 이후 이를 탐지하거나 훼손된 정보를 능동적으로 복원하는 방어 메커니즘이 부재하며, 이로 인해 고도화된 공격에 효과적으로 대응하지 못하고 워터마크의 신뢰성이 저하되는 문제를 안고 있음

• PGID 기술 개요 및 작동 원리

- PGID는 기존 워터마킹 시스템의 구조, 내용, 내부 데이터 등을 변경하지 않고 적용 가능한 사후 방어 기술로, 제거 및 위조 공격을 '데이터가 비정상 영역으로 이동한 상태'로 정의하고 확산 모델의 역방향 과정에 개입하여 이를 원래 상태로 되돌림
- 이러한 복원 과정은 공격받은 이미지 데이터에 의도적으로 노이즈를 추가했다가 다시 제거하는 점진적 유도 복원 방식을 통해 이루어지며, 비정상 영역으로 이동한 잠재 벡터(latent vector)*를 원본 워터마크가 존재하는 올바른 영역으로 점진적으로 유도함
- 최종적으로 이 과정을 통해 손상된 워터마크 신호를 복원하거나, 원본이 아닌 이미지에 삽입된 위조 신호를 식별해낼 수 있음

* 잠재 벡터(latent vector): 인공지능과 머신러닝에서 고차원의 원본 데이터(이미지, 텍스트 등)를 압축하여 핵심 특징만 숫자로 표현한 벡터

• 핵심 기능 1 – 지워진 신호를 복원해 제거 공격을 방어하는 PGID-R

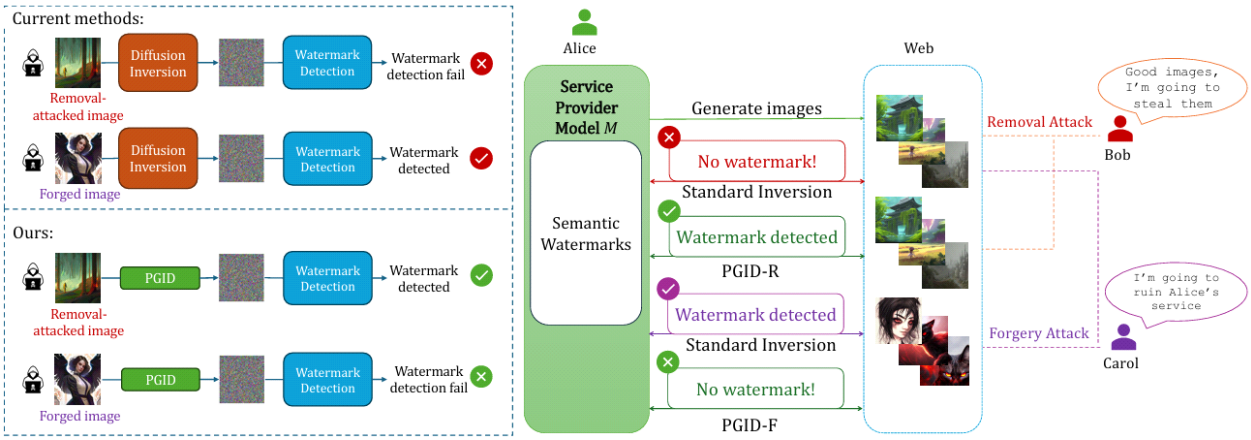
- 제거 공격을 방어하는 PGID-R 기능은 공격으로 훼손된 이미지의 내부 데이터를 순수한 노이즈 상태에 가깝게 되돌린 뒤, 이 노이즈를 단계적으로 제거하며 숨겨진 워터마크 정보를 다시 살려내는 방식임
- 이 복원 과정의 정밀도는 내부 설정값을 통해 조절할 수 있으며, 특히 이미지 데이터를 되돌리는 복원 단계를 충분히 수행할수록 워터마크가 안정적으로 복원되는 특징을 보임
- 실제로 특정 공격 기법으로 훼손된 스테이블 디퓨전(Stable Diffusion)로 생성한 이미지에 이 기술을 적용했을 때, 원래 워터마크 정보의 약 93%를 성공적으로 되살리는 결과를 보여주었음
- 이 모든 과정은 AI 모델을 재학습시킬 필요 없이 기존 시스템에 추가 기능처럼 덧붙여 방어 능력을 부여하는 방식이므로 실용성이 높음

• 핵심 기능 2 – 가짜 신호를 판별해 위조 공격을 방어하는 PGID-F

- 위조 공격을 방어하는 PGID-F 기능은 워터마크가 없는 데도 있는 것처럼 조작된 이미지의 미세한 불일치성을 증폭시켜 진위를 판별하는 메커니즘을 사용함
- 원본 이미지와 위조 이미지를 동일한 PGID 노이즈 제거 과정에 통과시켰을 때 나타나는 최종 워터마크 신호 강도의 뚜렷한 차이를 이용하여 둘을 구분하는 원리임
- PGID-F는 대부분의 공격 시나리오를 적용한 결과 AUC 0.96점 이상의 점수를 기록했으며, 이는 소수의 오류를 제외하고는 매우 높은 신뢰도로 위조 이미지를 판별해냄을 시사함

- 이는 단순히 워터마크의 존재 유무를 넘어, 해당 정보가 정당한 생성 과정을 거쳤는지 직접 검증하는 능동적 방어 체계라는 점에서 기존 방식과 뚜렷한 차별점을 가짐

[그림 1] PGID 기술의 작동 개요



출처: Minh Quoc Duong 외 2인, "PGID: Progressive Guided Inversion and Denoising for Robust Watermark Detection", arXiv, 2026.05.10., https://arxiv.org/html/2605.09319v1

• 종합 성능 및 적용성

- PGID는 기존 AI 모델을 수정하거나 재학습할 필요 없이 간편하게 추가할 수 있는 솔루션으로, 다양한 시스템에 쉽게 적용할 수 있는 높은 범용성을 지님
- 실험 결과, 이 기술은 널리 사용되는 스테이블 디퓨전 모델뿐만 아니라 다른 구조의 최신 이미지 생성 모델에서도 안정적인 방어 성능을 유지하는 것으로 확인되었음
- 가이던스 강도와 같은 내부 설정값을 유연하게 조절할 수 있어, 다양한 공격 유형과 시스템 환경에 맞춰 최적의 방어 수준을 설정할 수 있는 특징을 가짐

결론 및 시사점

• 기술적 한계와 향후 과제

- PGID 기술은 기존 워터마킹이 공격에 수동적으로 버티는 방식에서 벗어나, 훼손된 정보를 능동적으로 복원하고 위조 여부를 판별하는 새로운 방어 패러다임을 제시했다는 점에서 중요한 의의를 가짐
- 다만, 고도로 최적화된 공격 시나리오에서는 방어 및 탐지 성능이 일부 저하될 수 있으며, 새로운 생성 모델에 적용 시에는 모델별로 최적의 하이퍼파라미터*를 다시 설정해야 하는 실용적 한계를 지님
- 따라서 향후에는 여러 가지 공격 유형에 대한 방어 강건성을 높이고, 다양한 모델 구조에 자동으로 최적화되는 범용적인 방어 프레임워크로 발전시키는 방향의 후속 연구가 핵심 과제로 요구됨

* 하이퍼파라미터(hyperparameter): 인공지능 모델이 학습을 시작하기 전, 학습 과정이나 모델 구조를 제어할 목적으로 사용자가 직접 설정하는 외부 구성 값

참고문헌

• Minh Quoc Duong 외 2인, "PGID: Progressive Guided Inversion and Denoising for Robust Watermark Detection", arXiv, 2026.05.10., https://arxiv.org/html/2605.09319v1