산업 캘리포니아주 AI 투명성법 시행, AI 생성 이미지의 출처 책임 기준 제시

딥페이크 확산 속 AI 투명성법 도입으로 콘텐츠 출처 표기 최초 의무화

▶ AI 생성 콘텐츠 확산으로 딥페이크 문제가 심화되면서 자율 규제의 한계가 드러났고, 이에 따라 캘리포니아주는 2026년 1월 1일부터 월간 사용자 100만 명 이상 AI 서비스 제공자를 대상으로 출처 공개를 의무화하는 AI 투명성법을 시행하였다. 이에 대응해 오픈AI는 메타데이터와 비가시적 인증 마커를 적용하고 있으며, 어도비는 콘텐츠 자격 증명의 단계별 자동 적용 등 다양한 기술적 방식을 통해 규제 요건을 이행하고 있다. 그러나 기업 간 규제 대응 역량 격차 확대, 플랫폼 차원의 메타데이터 자동 삭제 관행, 텍스트 기반 콘텐츠의 규제 범위 제외 등으로 인해 해당 법안의 실효성과 지속 가능성에는 여전히 의문이 제기되고 있다.

산업 구글, 제미나이 기반 구글 AI로 생성·편집된 동영상 검증 기능 도입

AI 콘텐츠 투명성을 위한 신스아이디 기반 동영상 검증 체계

▶ 구글(Google)은 2025년 12월 제미나이 앱에 구글 AI로 생성·편집된 동영상을 검증할 수 있는 기능을 공식 출시했다. 사용자는 동영상을 업로드해 질문하는 방식으로 검증 결과를 확인할 수 있으며, 제미나이는 동영상 내 신스아이디(SynthID)를 감지해 영상·음성 구간별로 구체적인 검증 결과를 제공한다. 해당 기능은 신스아이디 워터마크 스캔과 제미나이의 추론 능력을 결합한 방식으로, 동영상 내 워터마크가 없거나 일부 손상된 경우에도 AI 산출물을 식별할 가능성을 갖는다. 다만 데이터 보존 정책, 업로드된 콘텐츠의 AI 모델 학습 활용 여부와 오탐률·오분류율 등 검증 정확도를 판단할 수 있는 정보가 공개되지 않았다는 점은 한계로 지적된다.

산업 美 엔터테인먼트소프트웨어협회, DMCA 소환장 통한 불법 토렌트 사이트 운영자 추적 시도

루트래커(RuTracker) 사례를 중심으로 본 운영자 타겟팅 및 중개자 책임 쟁점

▶ 미국 엔터테인먼트소프트웨어협회(ESA)는 2025년 12월 루트래커 운영자의 신원 확보를 위해 클라우드플레어(Cloudflare)를 대상으로 디지털 밀레니엄 저작권법(Digital Millennium Copyright Act)에 따른 소환장을 신청했다. 이는 단순 접속 차단에서 벗어나 중개 사업자를 통해 운영 주체를 직접 특정하고 법적 책임을 추궁하려는 전략적 전환으로 풀이된다. 이러한 흐름은 게임을 넘어 학술·출판 등 산업 전반으로 확산되고 있으나, 사법부의 심의 없이 발급되는 소환장이 중개 사업자의 프라이버시 보호 원칙과 충돌하고 있어 향후 저작권 집행의 실효성과 개인정보 보호 간의 갈등은 더욱 심화될 것으로 전망된다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 자동화된 저작권 삭제 체계 확산…불법 스트리밍 대응, 클라우드 인프라 단계로

저작권자의 직접 삭제 요청 확대 움직임

▶ 온라인 스트리밍 산업에서는 불법 스트리밍 대응 방식이 기존의 개별 웹사이트 단위 차단에서 벗어나, 클라우드 인프라 계층을 중심으로 한 자동화된 저작권 삭제 체계로 전환되고 있다. 2025년 상반기 클라우드플레어는 저작권자가 전용 API를 통해 침해 신고를 직접 제출할 수 있는 구조를 도입하며, 대량의 삭제 요청을 자동으로 처리하는 방식을 본격화했고, 그 결과 호스팅 관련 저작권 침해 신고가 약 12만 건에 달하며 실제 삭제·차단 조치도 급증했다. 특히 자체 저장 서비스(R2)를 이용한 계정 종료 사례는 CDN 사업자가 단순 트래픽 중개를 넘어, 콘텐츠가 저장·유통되는 인프라 단계에서 직접 개입하기 시작했음을 보여주는 상징적 변화로 평가된다.

산업 북아이피스 사례로 본 생성형 AI 교육콘텐츠 신뢰성 관리의 설계 및 책임 구조 전환

생성형 AI 기반 교육콘텐츠 활용 확대와 품질 관리 문제 동시 부상

▶ 생성형 AI의 교육콘텐츠 활용이 확산되면서 제작 효율성과 접근성은 높아졌으나, 사실 오류·교육과정 부적합·편향성·책임 주체 불명확 등 신뢰성 문제가 동시에 부상하였다. 특히 사후 수정·삭제 중심의 관리 방식만으로는 생성형 AI 콘텐츠의 품질과 저작권 리스크를 충분히 통제하기 어렵다는 한계가 부각되었다. 이에 국내 에듀테크 기업 북아이피스는 콘텐츠 생성 단계에서 위험을 통제하는 가드레일과 전문가 피어리뷰, 국가콘텐츠식별체계(UCI) 기반 출처·이력 관리 구조를 실제 서비스에 적용하였다. 해당 사례는 생성형 AI를 보조 도구로 활용하고 판단·책임은 전문가가 담당하는 운영 모델을 통해, 관리의 책임 단위가 '사후 검증'에서 '생성 단계 설계'로 이동하고 있음을 시사한다.

기술 주간기술동향

AI 모델 도용 위협 증가와 모델 워터마킹 기술

▶ 모델 워터마킹(Model Watermarking)은 모델의 출력물에 소유자만 식별할 수 있는 디지털 서명을 삽입하여 저작권을 증명하는 기술이다. 최근 이러한 모델 워터마킹 기술에 대한 관심이 증가하고 있다. 본 보고서에서는 수많은 워터마킹 기술 중에서도 특히 모델의 내부 구조를 알 수 없는 블랙박스 환경에서, 원본의 품질 저하는 최소화하면서도 다양한 공격에 강력한 저항성을 갖도록 설계된 ComMark 기술 사례를 집중적으로 분석하고자 한다.

캘리포니아주 AI 투명성법 시행, AI 생성 이미지의 출처 책임 기준 제시

뉴스브리프

AI 생성 콘텐츠 확산으로 딥페이크 문제가 심화되면서 자율 규제의 한계가 드러났고, 이에 따라 캘리포니아주는 2026년 1월 1일부터 월간 사용자 100만 명 이상 AI 서비스 제공자를 대상으로 출처 공개를 의무화하는 AI 투명성법을 시행하였다. 해당 법안은 잠재적 공개와 명시적 공개로 구성된 이중 공개 체계를 규정하고 C2PA 표준을 법적 의무로 도입하는 한편, 무료 공개 검증 도구 제공과 워터마크가 삭제된 콘텐츠에 대한 96시간 이내 라이선스 취소 의무를 명시함으로써 AI 서비스 제공자의 책임 범위를 제도적으로 규정하고 있다. 이에 대응해 오픈AI는 메타데이터와 비가시적 인증 마커를 적용하고 있으며, 어도비는 콘텐츠 자격 증명의 단계별 자동 적용 등 다양한 기술적 방식을 통해 규제 요건을 이행하고 있다. 그러나 기업 간 규제 대응 역량 격차 확대, 플랫폼 차원의 메타데이터 자동 삭제 관행, 텍스트 기반 콘텐츠의 규제 범위 제외 등으로 인해 해당 법안의 실효성과 지속 가능성에는 여전히 의문이 제기되고 있다.

AI 생성 콘텐츠 확산과 캘리포니아주 투명성 법안 시행

• 딥페이크 확산과 캘리포니아주 AI 투명성법 제정 배경

- AI 영상·이미지 생성 기술의 고도화로 현실과 구분이 어려운 합성 미디어가 빠르게 확산되고 있음. 이에 따라 딥페이크를 활용한 선거 개입, 허위정보 유포 등 사회적 혼란 사례가 증가하고 있으며, 고성능 AI 도구의 대중화로 디지털 콘텐츠 전반의 신뢰성 위기가 심화되고 있음
- 이러한 상황에서 기존 자율 규제 방식만으로는 AI 생성 콘텐츠의 출처를 확인하기 어려워졌으며, 콘텐츠 진위 판단을 위한 법적·기술적 기반 마련 필요성이 부각되고 있음

- 기술 기업들은 자발적으로 워터마크 삽입을 약속해왔으나, 실제 이행률이 낮아 실효적인 검증 체계가 부재한 한계가 드러남
- 캘리포니아주는 이러한 문제의식에 따라 2026년 1월 1일부터 AI 투명성법(AI Transparency Act, SB 942)을 시행하고, 월간 사용자 100만 명 이상 AI 제공자에게 생성 콘텐츠 출처 공개를 법적으로 의무화함. 이는 미국 주 단위 최초의 강제적 AI 투명성 규제로, 글로벌 AI 규제 논의에서 중요한 선례로 작용할 전망임[1)]
- 동 법안은 이미지·영상·오디오 콘텐츠에 암호화된 메타데이터 삽입을 의무화하고, 무료 공개 검증 도구 제공을 요구함. AI 제공자뿐 아니라 소셜미디어 플랫폼에도 출처 확인 책임을 부여하는 구조로, 미국 주 단위 규제 가운데 가장 강력한 AI 투명성 제도로 평가되고 있음

캘리포니아주 AI 투명성법의 기술적·법적 요구사항과 C2PA 표준

• 캘리포니아주 AI 투명성법의 이중 공개 체계와 C2PA 표준 적용

- AI 투명성법은 잠재적 공개(latent disclosure)와 명시적 공개(manifest disclosure)로 구성된 이중 공개 체계를 규정함. 두 방식은 각각 기술적 검증과 일반 사용자 인식을 담당하도록 설계되었으며, 상호 보완적으로 작동하는 구조임
- 동 법안은 파일의 생성·편집 이력을 체인 형태로 기록해 콘텐츠의 전체 생애주기를 추적 및 관리할 수 있도록 설계된 산업 표준 C2PA*를 법적 의무 사항으로 명시함. 이를 통해 기존 자발적 참여 방식의 산업 표준을 최초로 법적 강제 체계로 전환함

  * C2PA(Coalition for Content Provenance and Authenticity): 2021년 어도비(Adobe Inc.), 마이크로소프트(Microsoft Corporation), 암(Arm Limited), BBC(British Broadcasing Corporation), 인텔(Intel Corporation), 트루픽(Truepic Inc.) 등 6개 기업이 공동으로 개발한 콘텐츠 출처 및 진위 검증 산업 표준

- 기존 워터마킹 기술은 압축이나 스크린샷 등 단순 편집 과정에서도 쉽게 제거되는 한계가 있었으나, 이번 법률은 기술적 실현 가능성** 조항을 통해 일반적인 편집 환경에서도 유지 가능한 수준의 워터마크 내구성을 요구함

  ** 실현 가능성(technically feasible): 현존하거나 합리적으로 예측 가능한 기술 수준과 인프라, 표준, 인력 역량을 전제로 할 때 추가적인 근본 기술 돌파 없이 실제 구현 및 운영이 가능한 상태

[표1] SB 942에 따른 공개 유형 및 기술 요구사항

구분	잠재적 공개 (Latent Disclosure)	명시적 공개 (Manifest Disclosure)
정의	파일 코드 내 암호화 메타데이터	가시적 워터마크/아이콘
제거 난이도	극도로 어려움	사용자 선택적, 제거 가능
포함 정보	제공자명, 버전, 시각, 고유ID	시각적 AI 생성 표시
기술 표준	C2PA 기반 암호화 서명	선택적 가시 마커
목적	기술적 검증 및 추적	일반 사용자 인식 제공

출처: Spoke, “California’s AI Transparency Act Goes Live: A New Era in the War on Deepfakes”, 2026.01.05., https://markets.financialcontent.com/spoke/article/tokenring-2026-1-5-californias-ai-transparency-act-goes-live-a-new-era-in-the-war-on-deepfakes

1) Spoke, “California’s AI Transparency Act Goes Live: A New Era in the War on Deepfakes”, 2026.01.05., https://markets.financialcontent.com/spoke/article/tokenring-2026-1-5-californias-ai-transparency-act-goes-live-a-new-era-in-the-war-on-deepfakes

• 법적 의무사항 및 위반 시 제재

- AI 서비스 제공자*는 URL 기반 무료 공개 검증 도구와 제3자 플랫폼이 즉시 조회 가능한 API**를 제공하여 소셜미디어 등 플랫폼이 콘텐츠 출처를 실시간으로 확인할 수 있도록 지원해야 함

  * AI 서비스 제공자: AI 모델을 만들고 API나 서비스로 제공하는 회사

  ** API(Application Programming Interface): 서로 다른 시스템 간에 기능 및 데이터를 자동으로 연동할 수 있도록 정의된 표준화된 인터페이스

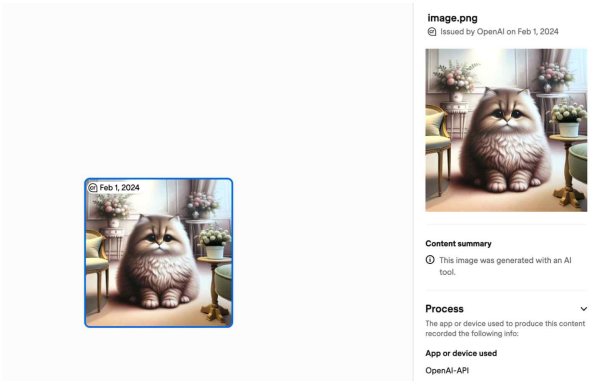
- 라이선시*가 워터마크를 제거할 경우 AI 서비스 제공자는 96시간 이내에 해당 라이선스를 취소해야 하며, 이를 이행하지 않을 경우 위반 건당 벌금이 부과됨. 이를 통해 AI 서비스 제공자에게 라이선시 관리 책임을 명확히 부여하고, 워터마크 제거 행위를 간접적으로 억제하는 구조를 형성함

  * 라이선시(Licensee): AI 서비스 제공자의 API를 라이선스 받아 자신의 앱/서비스에 통합하는 제3자 기업

- 한편 텍스트 기반 AI 산출물은 수정헌법 제1조에 따른 표현의 자유 문제와 기술적 구현상의 어려움을 이유로 현재 공개 의무 대상에서 제외되어 있음. 입법 과정에서 언론 자유 침해 논란과 함께 텍스트 워터마킹 기술의 미성숙이 지적되었으며, 그 결과 규제 범위는 이미지, 영상, 오디오로 제한됨

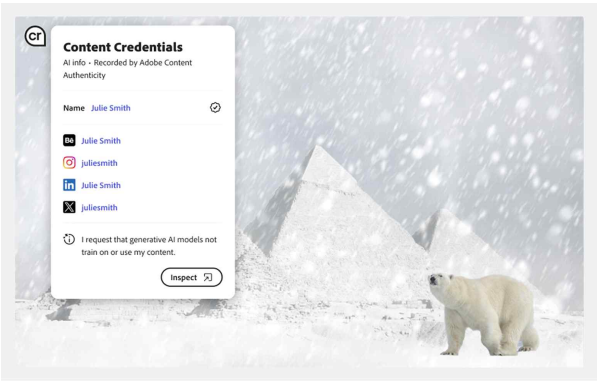
주요 기업의 C2PA 구현 사례

[그림1] 오픈AI AI 생성 이미지의 콘텐츠 자격 증명 확인



출처: OpenAI, “C2PA in ChatGPT Images”, 2026.01.12., https://help.openai.com/en/articles/8912793-c2pa-in-chatgpt-images

[그림2] 어도비 AI 생성 이미지의 콘텐츠 자격 증명 확인



출처: Adobe, “Content Credentials overview”, 2025.09.02., https://helpx.adobe.com/creative-cloud/apps/adobe-content-authenticity/content-credentials/overview.html

• 오픈AI의 C2PA 구현 방식

- 오픈AI(OpenAI)는 자사 범용 AI 챗GPT(ChatGPT)와 텍스트 기반 이미지 생성 AI 모델 API인 DALL·E 3 API*로 생성된 모든 이미지에 C2PA 메타데이터를 자동 삽입하고, 콘텐츠 자격 증명(Content Credentials)**을 통해 출처 확인이 가능하도록 구현함

  * DALL·E 3 API: OpenAI에서 개발한 텍스트 기반 이미지 생성 AI 모델인 DALL·E 3를 애플리케이션이나 서비스에 통합할 수 있도록 제공하는 프로그래밍 인터페이스

  ** 콘텐츠 자격 증명(Content Credentials): 이미지나 영상 같은 디지털 파일에 생성 및 편집 방식, 제작자, 사용된 도구, 편집 활동 내역 등 관련 컨텍스트를 추가하는 업계 표준의 변조 방지 메타데이터

- API로 직접 생성한 이미지는 API의 출처가 기록되고, 챗GPT에서 생성한 이미지는 API와 챗GPT 두 단계의 출처가 모두 기록되도록 하여, 동일한 생성 결과물이라도 어떤 서비스 인터페이스와 실행 경로를 거쳐 생성되었는지까지 식별 가능하도록 함

- 오픈AI는 자사 영상 생성 도구 소라(Sora)에도 C2PA 메타데이터를 적용하고 비가시적 인증 마커를 삽입해 이를 높은 정확도로 탐지하는 시스템을 운영 중임. 이 시스템은 가시적 워터마크와 별개로, 메타데이터가 제거된 경우에도 콘텐츠 출처를 추적할 수 있도록 이중 검증 체계를 구성함

• 어도비의 콘텐츠 자격 증명 구현 방식

- 어도비는 AI 이미지 생성 도구 파이어플라이(Firefly), 이미지 편집 소프트웨어 포토샵(Photoshop), 영상 편집 소프트웨어 프리미어 프로(Premiere Pro) 전반에 콘텐츠 자격 증명을 자동 적용하고, 창작자 신원, 사용 도구, 편집 이력 등 핵심 출처 정보를 기록함
- 해당 체계는 AI 산출물과 사람이 창작한 콘텐츠 모두에 동일하게 적용되어, 디지털 콘텐츠 전반의 출처 투명성을 확보하는 기반으로 활용됨
- 콘텐츠 자격 증명은 편집 단계마다 새로운 인증 정보를 순차적으로 추가하는 방식으로 설계되어, 콘텐츠가 카메라 촬영물인지 AI 생성물인지를 명시함. 이를 통해 최초 생성부터 최종 산출물까지의 전체 제작 및 편집 과정이 체인 형태로 기록되며, 각 편집 행위가 누적되는 구조를 형성함
- 또한 어도비는 주요 카메라 제조사와 협력해 촬영 시점부터 콘텐츠 자격 증명를 삽입하는 인증 캡처 기술로 기술을 확장 중임. 이는 가짜 콘텐츠의 사후 탐지 중심 접근에서 벗어나, 진짜 콘텐츠를 선제적으로 인증하는 방향으로 전략을 전환하려는 장기적 시도로 평가됨

캘리포니아주 AI 투명성법의 산업 영향과 향후 과제

• 시장 집중도 심화와 법안의 구조적 한계

- 해당 법률은 대기업과 스타트업 간 규제 대응 역량 격차를 확대해 AI 시장의 집중도를 심화시킬 수 있음. 빅테크 기업은 이미 C2PA 통합을 완료하고 이를 경쟁 우위로 활용하는 반면, 중소 스타트업은 연구개발 예산의 최대 20%를 규제 대응에 투입하며 혁신 여력이 약화되는 구조가 형성되고 있음[2)]
- 또한 법안의 실효성은 소셜미디어 플랫폼의 기술 표준 준수 여부에 크게 좌우되나, 실제 이행 수준은 제한적임. 한 실험 결과, 8개 주요 플랫폼 모두 C2PA 메타데이터를 자동 삭제했으며, 이미지 압축 과정에서 출처 정보가 제거되어 사용자에게 전달되지 못한 것으로 나타남[3)]
- 아울러 주 차원의 법률로서 연방정부와의 법적 충돌 가능성과 규제 범위 한계로 장기적 지속성이 불확실하다는 평가를 받고 있음. 또한 텍스트 생성 AI가 규제 대상에서 제외됨에 따라 문자 기반 허위정보는 여전히 통제 범위 밖이라는 것도 한계로 지목됨

참고문헌

- Spoke, “California’s AI Transparency Act Goes Live: A New Era in the War on Deepfakes”, 2026.01.05., https://markets.financialcontent.com/spoke/article/tokenring-2026-1-5-californias-ai-transparency-act-goes-live-a-new-era-in-the-war-on-deepfakes
- OpenAI, “C2PA in ChatGPT Images”, 2026.01.12., https://help.openai.com/en/articles/8912793-c2pa-in-chatgpt-images
- Adobe, “Content Credentials overview”, 2025.09.02., https://helpx.adobe.com/creative-cloud/apps/adobe-content-authenticity/content-credentials/overview.html
- Kevin Schaul, “We uploaded a fake video to 8 social apps. Only one told users it wasn’t real.”, The Washington Post, 2025.10.22., https://www.washingtonpost.com/technology/2025/10/22/ai-deepfake-sora-platforms-c2pa/

2) Spoke, “California’s AI Transparency Act Goes Live: A New Era in the War on Deepfakes”, 2026.01.05., https://markets.financialcontent.com/spoke/article/tokenring-2026-1-5-californias-ai-transparency-act-goes-live-a-new-era-in-the-war-on-deepfakes

3) Kevin Schaul, “We uploaded a fake video to 8 social apps. Only one told users it wasn’t real.”, The Washington Post, 2025.10.22., https://www.washingtonpost.com/technology/2025/10/22/ai-deepfake-sora-platforms-c2pa/

구글, 제미나이 기반 구글 AI로 생성·편집된 동영상 검증 기능 도입

뉴스브리프

구글(Google)은 2025년 12월 제미나이 앱에 구글 AI로 생성·편집된 동영상을 검증할 수 있는 기능을 공식 출시했다. 사용자는 동영상을 업로드해 질문하는 방식으로 검증 결과를 확인할 수 있으며, 제미나이는 동영상 내 신스아이디(SynthID)를 감지해 영상·음성 구간별로 구체적인 검증 결과를 제공한다. 해당 기능은 신스아이디 워터마크 스캔과 제미나이의 추론 능력을 결합한 방식으로, 동영상 내 워터마크가 없거나 일부 손상된 경우에도 AI 산출물을 식별할 가능성을 갖는다. 다만 데이터 보존 정책, 업로드된 콘텐츠의 AI 모델 학습 활용 여부와 오탐률·오분류율 등 검증 정확도를 판단할 수 있는 정보가 공개되지 않았다는 점은 한계로 지적된다.

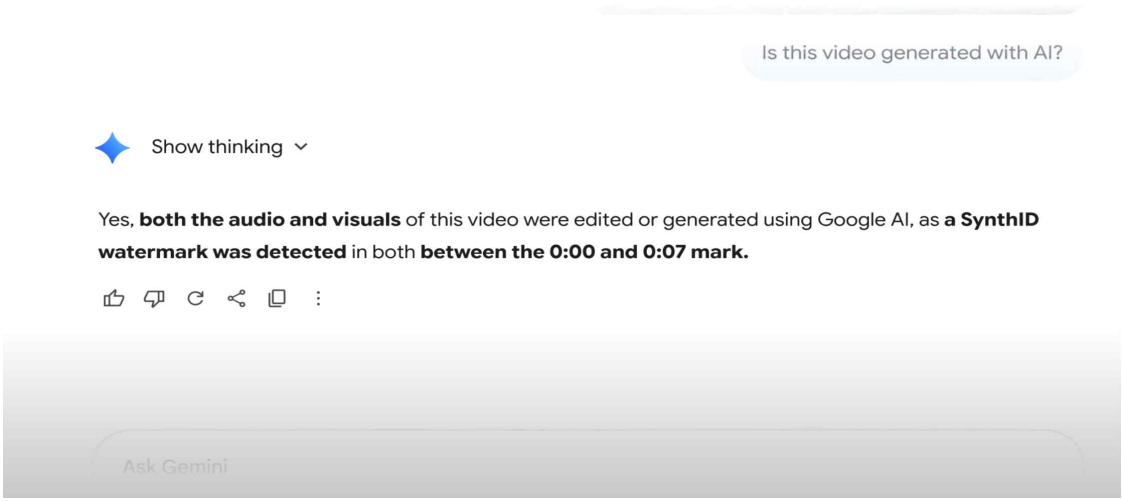
구글, 제미나이에 구글 AI 동영상 검증 기능 추가

- **2025년 12월, 구글 AI로 생성·편집된 동영상 검증 기능 공식 출시**

- 구글은 2025년 12월 18일 자사의 AI 모델 제미나이(Gemini)에 동영상이 구글 AI를 사용하여 편집 또는 생성되었는지를 검증하는 기능을 공식 출시함[1)]
- 사용자는 제미나이에 동영상 파일을 업로드하고 “이 영상이 구글 AI로 생성된 것인가요?”와 같은 질문을 입력하면 검증 결과를 확인할 수 있음
- 제미나이는 동영상을 분석한 후 “오디오 10~20초 구간에서 신스아이디(SynthID)가 감지되었습니다” 또는, “영상에서 신스아이디가 감지되지 않았습니다”와 같이 구체적인 정보를 제공함
- 이 기능은 최대 100MB 용량과 90초 길이의 동영상 파일을 지원하며, 제미나이 앱이 지원하는 모든 언어와 국가에서 이용 가능함

1) Google "You can now verify Google AI-generated videos in the Gemini app." 2025.12.18., https://blog.google/technology/ai/verify-google-ai-videos-gemini-app/

[그림 1] 제미나이 동영상 내 신스아이디 감지 결과 예시



출처: Google "You can now verify Google AI-generated videos in the Gemini app." 2025.12.18., https://blog.google/technology/ai/verify-google-ai-videos-gemini-app/

신스아이디 워터마크 기반 검증 구조

• 신스아이디 워터마크 스캔과 추론 기반 검증 방식

- 검증은 구글이 개발한 신스아이디(SynthID)*라는 워터마킹 기술을 기반으로 작동함. 제미나이는 업로드된 동영상에서 신스아이디를 스캔하여 구글 AI 사용 여부를 판별함

  * 신스아이디(SynthID): 사람은 인지할 수 없지만 기계로는 판독 가능한 디지털 워터마크를 영상과 음성 트랙에 삽입하는 기술

- 신스아이디는 동영상의 영상 및 음성 트랙 전반에 걸쳐 삽입되어, 제미나이는 전체가 합성된 동영상인지, 일부만 편집된 것인지, 또는 AI 음성과 실제 영상이 섞인 것인지 구분할 수 있음

- 또한, 구글의 신스아이디는 동영상이 편집되거나 포맷되더라도 유지되는 내장 워터마크로, 파일이 재인코딩되거나 편집을 거치면 효과가 없어지는 메타데이터(Metadata)의 한계를 해결함

  * 메타데이터(Metadata): 파일의 속성 정보를 담은 데이터로, 생성 날짜, 저자, 위치 등의 정보를 포함

- 또한, 검증은 워터마크 식별을 넘어 제미나이의 추론 기능을 활용함. 제미나이는 동영상을 검증할 때 맥락을 분석하여 워터마크가 없거나 손상된 경우에도 영상 또는 음성 특징을 기반으로 AI 산출물을 식별할 가능성이 있음

구글의 정보 공개 투명성 이슈

• 데이터 처리·보존 정책 및 검증 성능에 대한 투명성 부족

- 폴란드의 디지털 마케팅 전문 미디어인 PPC 랜드(PPC LAND)는 구글의 이번 발표에서 데이터 보존 정책과 검증 정확도에 대한 구체적인 정보가 공개되지 않았다고 지적함[2)]

- 동영상 검증을 위해 사용자는 콘텐츠를 제미나이에 업로드해야 하나, 해당 영상이 구글의 AI 모델 학습에 활용되는지 여부나 데이터 보존 여부에 대한 설명은 제시되지 않음

2) Luis Rijo, "Google's Gemini now lets users verify AI-generated video content", PPC LAND, 2025.12.26., https://PPC.land/googles-gemini-now-lets-users-verify-ai-generated-video-content/

- 또한 오탐률*이나 오분류율** 등 검증 성능을 판단할 수 있는 정량적 지표가 제공되지 않았음. 이로 인해 실제 콘텐츠가 합성 콘텐츠로 잘못 판단되거나, 워터마크가 없거나 크게 변형된 AI 산출물이 검출되지 않을 가능성이 존재함

  * 오탐률(false positive rate): 실제로는 AI로 생성되지 않았거나 문제가 없는 콘텐츠를 시스템이 AI 합성 콘텐츠로 잘못 판단하는 비율

  ** 오분류율(misclassification rate): 콘텐츠의 실제 유형과 시스템의 판별 결과가 서로 일치하지 않는 전체 오류 비율

- 구글은 제미나이의 검증 기능이 모든 지원 언어와 국가에서 사용 가능하다고 밝혔으나, 언어별·콘텐츠 유형별로 검증 정확도에 차이가 발생하는지에 대해서는 언급하지 않음. 콘텐츠 검증 결과에 의존하는 사용자 입장에서는 이러한 오류 가능성과 한계를 사전에 인지하는 것이 필수적임

AI 콘텐츠 투명성 요건 강화 속 구글 동영상 검증 기능의 역할

• 디지털 광고 분야에서의 투명성 향상 가능[3)]

- PPC 랜드(PPC LAND)는 디지털 광고 분야에서 허위·기만 정보에 대한 규제가 강화되는 가운데, AI로 생성된 콘텐츠의 출처와 진위 여부를 검증할 수 있는 수단의 중요성이 빠르게 커지고 있다고 설명함
- 특히 미국 연방거래위원회(Federal Trade Commission, FTC)가 합성 추천 및 조작된 추천을 포함한 기만적 광고 관행에 대한 단속을 강화하면서, AI 생성 콘텐츠를 활용하는 광고주가 공개 의무와 진정성 주장과 관련해 보다 엄격한 검토에 직면하고 있다고 지적함
- 이와 관련해 PPC 랜드는 광고주가 구글 AI 도구를 활용할 경우, 구글 동영상 검증 기능이 관련 광고 정책과 투명성 요건을 충족하고 있음을 입증하는 보조적 수단으로 활용될 수 있다고 언급함
- 또한 해당 기능이 소비자용 애플리케이션인 제미나이를 통해 제공된다는 점에 주목하며, 콘텐츠 진위 여부에 대한 인식과 책임이 전문가나 기업 차원을 넘어 일반 이용자 수준까지 확대되고 있음을 보여준다고 평가함. 이는 디지털 광고 업계 전반에서 투명성에 대한 기대 수준을 한 단계 끌어올리는 방향으로 작용할 가능성이 있다고 덧붙임

참고문헌

- Luis Rijo, "Google's Gemini now lets users verify AI-generated video content", PPC LAND, 2025. 12.26., https://PPC.land/googles-gemini-now-lets-users-verify-ai-generated-video-content/
- Google "You can now verify Google AI-generated videos in the Gemini app." 2025.12.18., https://blog.google/technology/ai/verify-google-ai-videos-gemini-app/
- Elissa Welle, "Google's Gemini app can check videos to see if they were made with Google AI", The Verge, 2025.12.19., https://www.theverge.com/news/847680/google-gemini-verification-ai-generated-videos

3) Luis Rijo, "Google's Gemini now lets users verify AI-generated video content", PPC LAND, 2025.12.26., https://PPC.land/googles-gemini-now-lets-users-verify-ai-generated-video-content/

美 엔터테인먼트소프트웨어협회, DMCA 소환장 통한 불법 토렌트 사이트 운영자 추적 시도

뉴스브리프

미국 엔터테인먼트소프트웨어협회는 2025년 12월 러시아 최대 토렌트 사이트인 '루트래커(RuTracker)' 운영자의 신원 정보를 확보하기 위해 DMCA 소환장을 신청했다. 이는 URL 삭제나 접속 차단 등 인프라 대응 방식에서 벗어나, 서비스 중개 사업자를 통해 운영 주체를 직접 특정하고 법적 책임을 추궁하려는 공세적 전략 전환으로 풀이된다. 특히 이번 조치는 게임 산업뿐만 아니라 학술·출판 업계에서도 유사한 법적 수단을 동원하는 흐름과 궤를 같이하며 산업 간 공조 체계로 확산되는 양상을 보이고 있다. 하지만 사법부의 실질 심의 없이 발급되는 DMCA 소환장이 중개 사업자의 프라이버시 정책과 충돌하며 '적법 절차' 논란을 야기하고 있어, 향후 저작권 집행의 실효성과 개인정보 보호 간의 법적 갈등은 더욱 심화될 것으로 전망된다.

대형 토렌트 사이트 대응의 한계와 전략적 전환

- **기존 저작권 집행 수단의 실효성 저하와 대응 방식 전환 필요성**

- 지난 20년간 권리자 단체들은 불법 저작물 유통 사이트를 대상으로 검색엔진 URL 삭제, ISP 접속 차단 등 하드웨어적 차단 방식을 중심으로 대응해 왔으나, 운영자의 도메인 변경 및 VPN 우회 안내 등으로 인해 실효성이 저하됨
- 특히 러시아 최대 규모 토렌트 사이트인 '루트래커(RuTracker)'는 2004년 설립 이후 21년간 운영을 지속하며, 도메인 압수와 전 세계적 차단 조치에도 불구하고 우회적으로 운영하며, 기존 차단 중심 집행 방식의 구조적 한계를 보여주는 대표적 사례로 평가됨

- **ESA의 DMCA 소환장 활용을 통한 운영자 특정 시도**

- 미국 엔터테인먼트소프트웨어협회(ESA)는 2025년 12월 디지털 밀레니엄 저작권법(DMCA)의 제512조 (h)항을 근거로 미국 연방법원에 소환장을 신청하며, 기존 대응 방식과 다른 접근을 시도함[1)]

1) And Maxwell, "Video Game Giants Suddenly Have RuTracker in their Crosshairs Again", TorrentFreak, 2025.12.21., https://torrentfreak.com/video-game-giants-suddenly-have-rutracker-in-their-crosshairs-251221/

- ESA는 콘텐츠 전송 네트워크(CDN)* 사업자인 클라우드플레어(Cloudflare)를 상대로 루트래커 운영자의 신원 정보를 확보해 민·형사상의 법적 책임을 직접 추궁하려는 전략으로 해석됨

  * 콘텐츠 전송 네트워크(CDN): 전 세계에 분산된 서버를 통해 웹사이트·콘텐츠를 사용자와 가까운 위치에서 전달함으로써, 접속 속도 향상과 트래픽 분산·보안 강화를 제공하는 인터넷 인프라 서비스.

토렌트 사이트 운영자 신원 확보 시도와 집행상의 쟁점

• DMCA 소환장을 통한 운영자 실질 신원 및 추적 정보 요구

- ESA는 루트래커가 콘텐츠 전송 네트워크 서비스를 이용하고 있다는 점에서, 해당 인프라를 제공하는 클라우드플레어를 운영자 정보의 주요 보유 주체로 특정함
- 이에 따라 ESA는 운영자를 사법 처리하기 위해 성명, 물리적 주소, 이메일, 전화번호 등 단순 인적 정보 외에도, 가입 당시부터 현재까지의 접속 IP 로그와 서비스 이용 중 발생한 모든 디지털 활동 기록을 공개 요구함
- 특히 유료 광고 수익, 후원금, 기타 결제 수단과 관련된 금융 거래 내역 및 계좌 정보를 요구함으로써, 자금 흐름을 통해 실질적인 운영 주체를 특정하려는 의도가 반영된 것으로 해석됨

[표1] DMCA 소환장을 통해 요청된 정보 항목

정보 항목	주요 요청 정보 항목	활용 목적
기본 신원 정보	성명, 물리적 주소, 이메일 주소, 전화번호 등	익명 운영 주체(개인 또는 단체)의 실질적 특정
기술 추적 데이터	가입 당시 및 최근 접속 IP 로그, 디지털 활동 기록	실제 접속 위치 파악 및 사법 절차용 증거 확보
금융 거래내역	광고·후원금 결제 정보(신용카드, 금융 계좌 등)	자금 흐름 추적을 통한 은닉된 실질 운영자 식별
서비스 이용 이력	계정 생성·변경 이력, 관련 데이터 증거 보존 요구	민·형사상 소송 및 국제 수사 공조의 법적 근거 마련

출처: 참고문헌 종합하여 재구성

• 중개 서비스 사업자 대상 집행 확대에 따른 법적·기술적 쟁점

- 콘텐츠를 직접 전송하거나 저장하지 않는 CDN·DNS* 등 중개 서비스 사업자가 저작권 집행의 핵심 경로로 부상하며 인프라 중립성과 집행 협력 의무 간 갈등이 심화됨

  * 도메인 네임 시스템(DNS): 인터넷 주소(도메인 이름)를 서버의 숫자형 IP 주소로 변환해 주는 시스템으로, 사용자가 웹사이트에 접속할 수 있도록 경로를 안내하는 핵심 인터넷 인프라

- DMCA의 소환장은 사법부의 실질 심사 없이 발부될 수 있다는 점에서, 사업자의 프라이버시 보호 정책을 무력화하고 개인정보를 오남용할 수 있다는 '적법 절차' 논란이 제기됨
- 운영자가 암호화폐 결제, 익명화된 통신 수단을 활용했을 경우, 소환장을 통해 확보된 정보가 실제 신원 특정으로 이어지지 않을 수 있는 기술적 한계도 존재함
- 이러한 한계에도 불구하고 권리자 단체들은 중개 사업자에게 보다 강화된 본인 인증 및 데이터 보존 의무를 부여해야 한다는 입장을 지속적으로 개진하고 있음

• 게임 산업을 넘어선 전방위적 저작권 집행 공조 체계 확산

- ESA의 이번 조치는 게임 산업에 국한되지 않고, Sci-Hub 등 '섀도우 라이브러리' 운영자를 추적하기 위해 유사한 법적 수단을 동원하는 학술·출판 업계의 집행 전략과 유사한 흐름으로 해석됨[2)]
- 개별 사이트나 URL 차단하는 지엽적 방식에서 벗어나, 호스팅·도메인·CDN·결제 사업자 등 서비스 공급망 전반을 집행 대상으로 설정하는 방식으로 진화함
- 이는 불법 사이트의 운영 기반 자체를 압박하려는 산업 간 공동 대응 모델로, 향후 다른 콘텐츠 산업으로 확산되어 불법 복제 생태계 전반에 대한 법적 압박 수위가 한층 높아질 것으로 전망됨

온라인 저작권 집행 방식의 변화와 전망

• 실효성 있는 저작권 보호를 위한 집행 전략의 고도화 및 과제

- 도메인 우회 기술 등으로 기존의 접속 차단 조치 한계에 부딪힘에 따라, 중개 사업자를 통해 운영자의 신원을 특정하여 법적 책임을 규명하려는 시도가 향후 국제 수사 협력 및 분쟁의 주요한 수단이 될 것으로 전망됨
- 중개 사업자가 보유한 데이터의 품질이 집행의 성패를 결정함에 따라, 향후 사업자 대상의 엄격한 본인 인증 절차와 데이터 보존 정책 요구가 심화되는 과정에서 프라이버시 침해 및 '적법 절차' 논란이 지속될 것으로 보임
- 운영 주체를 은폐하는 기술적 고도화에 대응하기 위해, 수사 기관과 권리자 단체 간의 정보 공유 체계를 강화하고 중개 사업자의 협력을 끌어낼 세부적인 가이드라인 마련이 시급함
- 법적 강제 조치만으로는 고착화된 불법 생태계 근절에 한계가 있으므로, 이용자의 접근성을 높이고 합리적인 가격 체계를 갖춘 합법적 서비스 모델을 구축하여 불법 복제물에 대한 수요 자체를 줄이는 병행 전략이 필수적임

참고문헌

- And Maxwell, "Video Game Giants Suddenly Have RuTracker in their Crosshairs Again", TorrentFreak, 2025.12.21., https://torrentfreak.com/video-game-giants-suddenly-have-rutracker-in-their-crosshairs-251221/
- Vondran Legal, "Cloudflare 17 U.S.C. 512(h) DMCA subpoena explained", JDSUPRA, 2024.01.03., https://www.jdsupra.com/legalnews/cloudflare-17-u-s-c-512-h-dmca-subpoena-7263660/
- Rouse Editor, "Roskomnadzor has permanently blocked rutracker.org", ROUSE, 2016.02.19., https://rouse.com/insights/news/2016/roskomnadzor-has-permanently-blocked-rutracker-org
- And Maxwell, "Major & Persistent Video Game Pirates Investigated by ESA", TorrentFreak, 2022.04.30., https://torrentfreak.com/major-persistent-video-game-pirates-investigated-by-esa-220430/
- Pascal Hetzscholdt, "The July 2025 Cloudflare DMCA subpoena marks another escalation in the decades-long battle between academic publishers and shadow libraries.", Pascal's Sbstack, 2025.08.09., https://p4sc4l.substack.com/p/the-july-2025-cloudflare-dmca-subpoena

2) Pascal Hetzscholdt, "The July 2025 Cloudflare DMCA subpoena marks another escalation in the decades-long battle between academic publishers and shadow libraries.", Pascal's Sbstack, 2025.08.09., https://p4sc4l.substack.com/p/the-july-2025-cloudflare-dmca-subpoena

자동화된 저작권 삭제 체계 확산…
불법 스트리밍 대응, 클라우드 인프라 단계로

뉴스브리프

온라인 스트리밍 산업에서는 불법 스트리밍 대응 방식이 기존의 개별 웹사이트 단위 차단에서 벗어나, 클라우드 인프라 계층을 중심으로 한 자동화된 저작권 삭제 체계로 전환되고 있다. 2025년 상반기 클라우드플레어는 저작권자가 전용 API를 통해 침해 신고를 직접 제출할 수 있는 구조를 도입하며, 대량의 삭제 요청을 자동으로 처리하는 방식을 본격화했고, 그 결과 호스팅 관련 저작권 침해 신고가 약 12만 건에 달하며 실제 삭제·차단 조치도 급증했다. 특히 자체 저장 서비스(R2)를 이용한 계정 종료 사례는 CDN 사업자가 단순 트래픽 중개를 넘어, 콘텐츠가 저장·유통되는 인프라 단계에서 직접 개입하기 시작했음을 보여주는 상징적 변화로 평가된다.

온라인 스트리밍 산업, 자동화된 저작권 삭제 체계 도입 움직임

- **저작권자의 직접 삭제 요청 확대와 클라우드 인프라 단계에서 저작권 침해 차단하는 방식 확산**

- 클라우드플레어(Cloudflare)*는 2025년 상반기부터 저작권자와의 협력을 통해 불법 스트리밍 대응 방식을 기존의 수동적 처리 방식에서 보다 자동화된 구조로 전환을 추진하고 있음

  * 클라우드플레어(Cloudflare): 전 세계 웹 트래픽의 약 20%를 처리하는 미국 기반 인터넷 인프라 기업으로, 콘텐츠 전송 네트워크(CDN), 디도스(DDoS) 방어, 클라우드 스토리지 및 보안 서비스를 제공함

- 가장 핵심이 되는 변화는 “저작권자가 전용 API(Application Programming Interface)*”를 통해 저작권 침해 신고를 직접 제출할 수 있도록 한 점으로, 이를 통해 대량의 삭제 요청을 신속하게 처리할 수 있는 환경이 마련됨

  * 저작권자 전용 API(Application Programming Interface): 저작권자가 개별 신고를 수작업으로 하지 않고, 시스템 연동을 통해 침해 콘텐츠 삭제 요청을 빠르고 반복적으로 전달할 수 있게 해주는 기술적 연결 수단을 의미

- 이러한 방식은 특히 실시간성이 중요한 스포츠 중계 불법 스트리밍 대응시 효과적이며, 삭제 요청 접수부터 조치까지 걸리는 시간이 크게 단축되는 것으로 알려져 있음

• **자동화된 집행 체계 도입 이후 저작권 침해 삭제 조치 급증**

- 이러한 구조 전환의 결과, 2025년 상반기 클라우드플레어가 접수한 호스팅 관련 저작권 침해 신고는 약 12만 건에 달했으며, 이 중 절반에 가까운 건이 실제 삭제 또는 접근 차단 조치로 이어짐
- 이는 불과 직전 반기와 비교해 수십 배 증가한 수치로, 저작권 침해 대응이 개별 신고를 하나씩 처리하는 방식에서 벗어나, 대량의 요청을 자동으로 처리하는 구조로 전환되고 있음을 보여줌
- 특히 주목되는 점은 클라우드플레어가 자체적으로 제공하는 저장 서비스인 R2를 이용하던 계정에 대해 대규모 종료 조치를 단행했다는 점임
- 이는 그동안 트래픽 전달을 중개하는 역할에 머물렀던 CDN 사업자가, 불법 콘텐츠가 실제로 저장되고 유통되는 단계에서 직접 개입한 사례로 볼 수 있음
- 이러한 조치는 저작권 침해에 대한 대응 범위가 개별 콘텐츠 제공자나 웹사이트 운영자에만 한정되지 않고, 클라우드와 같은 인터넷 인프라 사업자까지 확대되고 있음을 보여주는 변화로 해석될 수 있음

자동화된 삭제 조치 확대에 따른 문제점

• **저작권 침해와 직접적 관련 없는 콘텐츠까지 함께 제한될 가능성 존재**

- 저작권 침해 삭제 요청이 자동화되면서, 불법 스트리밍 콘텐츠에 대한 대응 속도와 처리 규모는 크게 증가하고 있지만, 이러한 방식은 삭제 대상과 차단 범위를 어떻게 설정하느냐에 따라, 저작권 침해와 직접적인 관련이 없는 콘텐츠까지 함께 제한될 가능성이 존재함
- 클라우드플레어는 법원 명령이나 규제 기관 요청에 따라 특정 국가에서 일부 도메인에 대한 접속을 제한하고 있으나, 인터넷 주소 전체를 일괄적으로 차단하는 방식은 지양하고 있다는 입장을 밝히고 있음
- 특히 공용 DNS 서비스를 활용한 차단은 정상적인 웹사이트 접근까지 함께 제한할 수 있다는 점에서, 클라우드플레어는 해당 방식이 인터넷 이용에 미치는 영향이 크다고 보고 있음

• **단순히 '차단 여부'뿐 아니라 '어디까지 차단할 것인지'를 기술적으로 관리하는 문제가 함께 중요해져**

- 실제 일부 국가에서는 불법 스포츠 스트리밍을 차단하는 과정에서 클라우드 인프라 전체에 대한 IP 차단이 이뤄져, 저작권 침해와 무관한 다수의 웹사이트가 동시에 접속 불가 상태가 되는 사례가 발생한 바 있음
- 이러한 사례는 자동화된 저작권 집행이 확대될수록, 단순히 '차단 여부'뿐 아니라 '어디까지 차단할 것인지'를 기술적으로 정교하게 관리하는 문제가 함께 중요해지고 있음을 보여줌

불법 스트리밍 차단 방식에 대한 클라우드플레어의 기준과 대응 방식

• **차단 범위를 제한하기 위한 클라우드플레어의 대응 원칙과 선택적 협력 방식**

- 자동화된 저작권 삭제 조치가 확대되는 가운데, 클라우드플레어는 모든 차단 요청을 동일한 방식으로 수용하기보다는, 적용 범위와 방식에 있어 일정한 기준을 유지하려는 입장을 보이고 있음
- 클라우드플레어는 특히 공용 DNS 서비스를 통한 차단은 인터넷 접근 전반에 영향을 미칠 수 있다는 점에서, 해당 방식에 대해서는 법적 대응이나 대체 이행 방안을 우선적으로 검토해 왔다고 밝힘

- 실제로 클라우드플레어는 자사 공용 DNS 서비스(1.1.1.1)를 통한 콘텐츠 차단은 지금까지 시행하지 않았으며, 법원 명령에 따른 차단 역시 지역 단위의 접근 제한이나 특정 도메인 차단 등 보다 제한적인 방식으로 이행하고 있음

• 클라우드 플레어, 자동화된 삭제 요청을 수용하면서도, 차단 범위를 무작정 확대하지 않겠다는 기조

- 클라우드플레어는 자동화된 삭제 요청을 수용하면서도, 차단 범위를 무작정 확대하지 않겠다는 원칙을 유지하고 있음. 다만 이러한 입장이 선언적 수준에 그치지 않고 실제로 어떻게 적용되고 있는지를 보여주기 위해서는 구체적인 사례가 필요함
- 한편, 영국에서는 과거에 내려진 사이트 차단 명령을 근거로 일부 불법 스트리밍 사이트에 대해 클라우드플레어가 자발적으로 접근 제한 조치를 시행하기 시작했는데, 이는 새로운 법원 명령에 따른 강제 조치라기보다는 법적 의무 이전 단계에서 이루어진 제한적 협력 사례로 볼 수 있음
- 이 과정에서 클라우드플레어는 차단된 도메인에 접속할 경우 차단 사유와 법적 근거를 안내하는 별도의 안내 페이지를 제공하고, 이의 제기 절차 역시 함께 안내하는 방식을 채택함. 이는 자동화된 삭제 요청을 도입하면서도 차단 범위를 최대한 좁게 설정하고, 이용자와 이해관계자의 혼란을 최소화하기 위한 대응 방식임

자동화된 저작권 삭제 체계 확산의 의미

• 불법 스트리밍 대응이 웹사이트를 넘어 클라우드 인프라 단계로 확대

- 저작권자가 전용 API를 통해 대량의 삭제 요청을 직접 전달하고, 인프라 사업자가 이를 자동으로 처리하는 구조는 대응 속도와 효율성을 크게 높였다는 점에서 산업적 전환으로 평가할 수 있음
- 동시에 이러한 구조는 차단 범위 설정, 정상 콘텐츠 보호, 이용자 권리 보장 등 새로운 관리 과제를 함께 수반하고 있으며, 기술적 대응만으로는 한계가 존재함
- 특히 자동화된 삭제·차단 체계가 확대될수록, 어떤 콘텐츠를 어디까지 차단할 것인지에 대한 기준을 기술적으로 정교하게 설계하고, 이를 제도적으로 뒷받침하는 장치의 필요성이 더욱 커지고 있음

참고문헌

• Ernesto Van der Sar, "Cloudflare Reports Surge in Streaming Piracy Takedowns, Removes 20k+ Storage Accounts", TF, https://torrentfreak.com/cloudflare-reports-surge-in-streaming-piracy-takedowns-removes-20k-storage-accounts/
• Gigazine, "Cloudflare Transparency Report Reveals Copyright Infringement Reports Doubled in the First Half of 2025, with Responses 50x Increase", 2026.01.05., http://gigazine.net/gsc_news/en/20260105-cloudflare-transparency-report/?utm
• Ernesto Van der Sar, "Premier League Targets Dozens of Pirate Streaming Sites through Cloudflare Subpoena", https://torrentfreak.com/premier-league-targets-dozens-of-pirate-streaming-sites-through-cloudflare-subpoena/?utm

북아이피스 사례로 본 생성형 AI 교육콘텐츠 신뢰성 관리의 설계 및 책임 구조 전환

뉴스브리프

생성형 AI의 교육콘텐츠 활용이 확산되면서 제작 효율성과 접근성은 높아졌으나, 사실 오류·교육과정 부적합·편향성·책임 주체 불명확 등 신뢰성 문제가 동시에 부상하였다. 특히 사후 수정·삭제 중심의 관리 방식만으로는 생성형 AI 콘텐츠의 품질과 저작권 리스크를 충분히 통제하기 어렵다는 한계가 부각되었다. 이에 국내 에듀테크 기업 북아이피스는 콘텐츠 생성 단계에서 위험을 통제하는 가드레일과 전문가 피어리뷰, 국가콘텐츠식별체계(UCI) 기반 출처·이력 관리 구조를 실제 서비스에 적용하였다. 해당 사례는 생성형 AI를 보조 도구로 활용하고 판단·책임은 전문가가 담당하는 운영 모델을 통해, 관리의 책임 단위가 '사후 검증'에서 '생성 단계 설계'로 이동하고 있음을 시사한다.

생성형 AI 확산에 따른 교육콘텐츠 신뢰성 관리 이슈

- **생성형 AI 기반 교육콘텐츠 활용 확대와 품질 관리 문제 동시 부상**

- 생성형 AI 기술의 고도화와 함께 교육 자료 생성, 학습 콘텐츠 제작, 평가 문항 구성 등 교육콘텐츠 전반에 AI 활용이 빠르게 확산되는 추세임
- 특히 에듀테크 플랫폼과 디지털 학습 환경을 중심으로 생성형 AI 콘텐츠가 실제 교육 현장에 활용되면서 제작 효율성과 접근성은 향상되는 반면, 콘텐츠의 정확성·적합성에 대한 우려도 병존함
- 생성형 AI는 사실성과 교육적 맥락에 대한 검증이 충분하지 않은 콘텐츠를 생성할 가능성이 있어 교육콘텐츠에서도 사실 오류, 교육과정 부적합성, 편향성 문제가 발생할 수 있으며, 동시에 책임 주체를 명확히 규정하기 어렵다는 구조적 한계가 제기됨

- **출처·검증·이력 관리 기반의 신뢰성 확보 필요성 증대**

- 생성형 AI 콘텐츠의 기술적·구조적 특성으로 인해, 사후 수정이나 삭제 중심의 관리 방식만으로는 교육콘텐츠의 신뢰성을 충분히 담보하기 어렵다는 인식이 확산됨
- 이에 따라 콘텐츠가 만들어지는 단계부터 검증 주체, 활용 이력, 원출처 정보를 함께 관리하는 전 주기적 신뢰성 관리 체계의 필요성이 교육콘텐츠 산업 전반에서 제기됨

- 이러한 흐름 속에서, 국내 기업 북아이피스는 생성형 AI 콘텐츠의 신뢰성과 저작권 관리 문제를 대응하기 위해 가드레일, 피어리뷰, 출처 이력 관리 체계를 결합한 기술적 관리 모델을 실제 서비스에 적용함[1)]

생성형 AI 교육콘텐츠의 신뢰성 관리 구조 전환 사례

• 생성 단계 통제와 전문가 검증 결합형 관리 구조

- 북아이피스는 생성형 AI 콘텐츠 생성 단계에서 가드레일* 시스템을 적용하여, 생성 범위·주제·표현 기준을 사전에 설정함으로써, 교육적으로 부적절하거나 오해 소지가 있는 콘텐츠의 생성 위험을 차단하는 구조를 구축함
- 가드레일을 통과한 생성형 AI 콘텐츠에 대해 동일 분야의 교육 전문가들이 상호 검증하는 피어리뷰** 체계를 운영하며, 단순한 사실 정확성뿐 아니라 교육과정 적합성, 학습 맥락의 타당성, 설명 수준의 적절성 등이 종합적으로 점검됨
- 이는 자동화된 알고리즘 검증만으로는 한계가 있는 교육콘텐츠 특성을 고려한 구조로, AI 생성 이후 오류를 수정하는 사후 대응 방식과 달리 생성 단계에서 위험 관리와 전문가 검증을 결합한 선제적 통제 방식으로 작동함

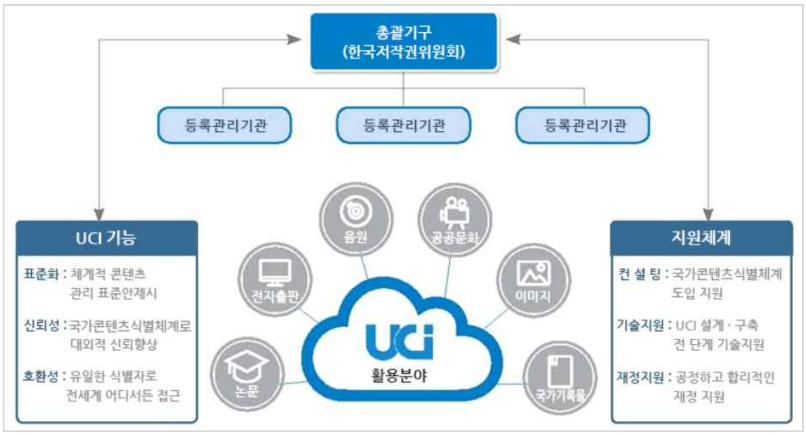
* 가드레일(Guardrail): AI 시스템이 사전에 정의된 기준과 범위를 벗어난 결과를 생성하지 않도록 제어하는 기술적·운영적 통제 장치를 의미함
** 피어리뷰(Peer Review): 동일 분야 전문가들이 콘텐츠를 상호 검토하여 정확성과 타당성을 평가하는 검증 방식임

• 국가콘텐츠식별체계(UCI) 기반 콘텐츠 출처·이력 관리 체계 적용

- 북아이피스는 한국저작권위원회와 함께 구축해 온 국가콘텐츠식별체계(UCI)를 생성형 AI 교육콘텐츠 관리 구조에 적용함
- UCI 적용으로 에듀테크 플랫폼에서 생성·유통되는 모든 콘텐츠는 원출처 정보와 생성·수정·활용 이력이 데이터로 기록되며, 저작권 분쟁이나 품질 논란 발생 시 사후 추적과 책임 확인이 가능한 관리 체계가 확보됨
- 이러한 구조는 생성형 AI를 단독 제작 주체가 아닌 보조 도구로 활용하는 휴먼-AI 협업형 콘텐츠 관리 모델로 설계되어, AI는 초안 생성과 반복 작업 효율화를 담당하고 인간은 콘텐츠의 적합성 판단과 책임 관리 역할을 수행하는 분담 구조를 형성함

* 국가콘텐츠식별체계(UCI): 콘텐츠에 고유 식별자를 부여하여 출처, 권리 정보, 유통 이력 등을 체계적으로 관리하는 국가 표준 식별 체계임

[그림1] 국가콘텐츠식별체계(UCI) 개요도



출처: 한국저작권위원회

1) 이지희, 「북아이피스, '쏠북' AI 콘텐츠 신뢰성 관리를 위한 가드레일·피어리뷰 시스템 도입」, 전자신문(에듀플러스), 2025.12.26., https://www.etnews.com/20251226000161

북아이피스 사례로 본 생성형 AI 콘텐츠 신뢰성 관리의 책임 및 운영 구조 변화

• 생성형 AI 콘텐츠 관리의 책임 단위가 '사후 검증'에서 '생성 단계 설계'로 이동

- 북아이피스 사례는 생성형 AI 콘텐츠의 오류·부적절성·저작권 문제를 사후 수정이나 분쟁 대응 차원이 아닌, 콘텐츠 생성 단계에서 기준·검증·책임을 구조적으로 내재화하는 방식으로 전환하고 있음을 보여줌
- 가드레일과 피어리뷰를 결합한 운영 구조는 생성형 AI를 단독 제작 주체가 아닌 보조 도구로 위치시키고, 생성은 AI가 수행하되 판단과 책임은 인간 전문가가 담당하는 역할 분담 구조를 명확히 설정한 점에서 특징을 가짐
- 이는 생성형 AI 활용의 핵심이 모델 성능이나 자동화 수준 자체가 아니라, 출처·검증·이력 관리가 결합된 운영 구조를 어떻게 설계하느냐에 따라 콘텐츠 신뢰성이 좌우될 수 있음을 보여주는 사례로 해석됨
- 북아이피스 사례는 이러한 관리 구조가 실제 서비스 환경에서 구현·운영되고 있다는 점에서 의의가 있으며, 향후 교육콘텐츠를 포함해 신뢰성과 책임성이 요구되는 콘텐츠 영역에서 생성형 AI 활용의 참고 모델로 작용할 가능성을 시사함

참고문헌

- UNESCO, "Guidance for Generative AI in Education and Research", UNESCO, 2023.03.07., https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research/
- OECD, "Initial Policy Considerations for Generative Artificial Intelligence", OECD, 2023.04.12., https://www.oecd.org/en/publications/initial-policy-considerations-for-generative-artificial-intelligence_fae2d1e6-en.html
- National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", NIST, 2024.01.26., https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- 이지희, "북아이피스, '쏠북' AI 콘텐츠 신뢰성 관리를 위한 가드레일·피어리뷰 시스템 도입", 전자신문(에듀플러스), 2025.12.26., https://www.etnews.com/20251226000161
- 고민서, "창작자 보호 강화하는 어도비…'콘텐츠 진위' 웹 앱 무료 배포한다", 매일경제, 2024.10.11., https://www.mk.co.kr/news/it/11137217

주간 기술 동향

AI 모델 도용 위협 증가와 모델 워터마킹 기술

• 탈취 가능한 AI 모델을 보호하기 위한 '모델 워터마킹'의 부상

생성형 AI 기술이 비약적으로 발전하며 기업의 핵심 경쟁력으로 자리 잡음에 따라, 막대한 자원과 데이터를 투입해 개발한 AI 모델 자체가 새로운 형태의 IP로 주목받고 있다. 하지만 이러한 기술적 자산은 디지털 형태라는 특성상 정교한 탈취 위협에 직접적으로 노출되어 있으며, 최근 '모델 도용(Model Theft)' 공격이 기업의 경쟁력을 순식간에 무력화시키는 심각한 위협으로 부상하고 있다. 이는 단순히 데이터를 유출하는 차원을 넘어, 수년간의 연구 개발 투자와 노력이 집약된 기술적 결정체를 통째로 빼앗기는 문제이기에 산업계의 깊은 우려를 낳고 있다.

공격자들은 더 이상 서버를 직접 해킹하지 않고도 정상적인 서비스 경로인 API를 통해 모델의 정보를 빼내거나 구조를 복제하는 고도화된 수법을 사용하고 있다. 대표적으로 공격자는 자동화된 스크립트를 통해 API에 수많은 질문을 보내고 그 응답 패턴을 분석하여 원본 모델과 거의 동일한 성능을 내는 복제 모델을 만들어내는 '모델 추출(Model Extraction)' 공격을 감행한다. 심지어는 AI 모델이 구동되는 하드웨어의 전력 소비 패턴이나 전자파 변화를 분석하는 사이드 채널 공격을 통해 모델의 내부 구조를 추론하는 등, 물리적 접근 없이도 IP를 훔치는 것이 가능한 시대가 되었다.

이러한 새로운 유형의 공격은 암호화된 트래픽을 감시하거나 외부의 비정상적 침입을 막는 데 초점을 맞춘 기존의 네트워크 보안 시스템만으로는 탐지하고 방어하기가 매우 어렵다. 공격자의 API 요청은 정상적인 사용자의 서비스 이용과 구별하기 힘들며, 대부분의 기업은 자사의 모델이 외부에서 은밀하게 분석되고 있다는 사실조차 인지하지 못하는 보안 사각지대에 놓여있다. 따라서 단순히 접근을 통제하는 것을 넘어, 모델 자체에 소유권을 증명할 수 있는 고유한 표식을 심어두는 능동적이고 지속적인 방어 체계의 필요성이 절실하게 대두되고 있다.

이러한 배경 속에서 AI 모델의 불법 복제 및 도용 문제에 대응하기 위한 핵심 기술로 '모델 워터마킹(Model Watermarking)'이 주목받고 있으며, 이는 모델의 출력물에 소유자만 식별할 수 있는 디지털 서명을 삽입하여 저작권을 증명하는 기술이다. 본 보고서에서는 수많은 워터마킹 기술 중에서도 특히 모델의 내부 구조를 알 수 없는 블랙박스 환경에서, 원본의 품질 저하는 최소화하면서도 다양한 공격에 강력한 저항성을 갖도록 설계된 ComMark 기술 사례를 집중적으로 분석하고자 한다. 이를 통해 AI 지식재산권 보호 기술의 현주소와 미래 발전 방향을 심도 있게 탐색할 것이다.

초기 모델 워터마킹의 특징과 기술적 한계

- 초기 모델 워터마킹은 모델 내부 파라미터를 직접 수정하는 화이트박스 방식이나 특정 입력에 대한 단순 표식을 출력하는 초보적 블랙박스 방식에 의존하여, 적용 범위가 제한적이거나 공격에 취약한 구조적 특징을 가짐
- 이로 인해 워터마크의 견고성을 높이면 모델 성능이 심각하게 훼손되고, 반대로 성능을 유지하면 모델 증류*나 파라미터 가지치기** 같은 고도화된 공격에 쉽게 무력화되어 실질적인 IP 보호 수단으로 기능하지 못하는 명백한 기술적 한계에 직면함

* 모델 증류(Model Distillation): 사전 학습된 대형 모델의 지식을 소형 모델로 전달하는 기법으로, 원본 모델의 입출력 쌍을 학습하여 기능을 복제하는 모델 탈취 공격에 악용됨

** 파라미터 가지치기(Parameter Pruning): 학습된 신경망에서 중요도가 낮은 파라미터를 제거하는 모델 최적화 기술로, 모델 내부에 삽입된 워터마크 정보를 훼손하여 소유권 증명을 무력화하는 공격 수단으로 사용됨

[사례] 압축 샘플 기반 ComMark 워터마킹 기술의 원리 및 성능 심층 분석

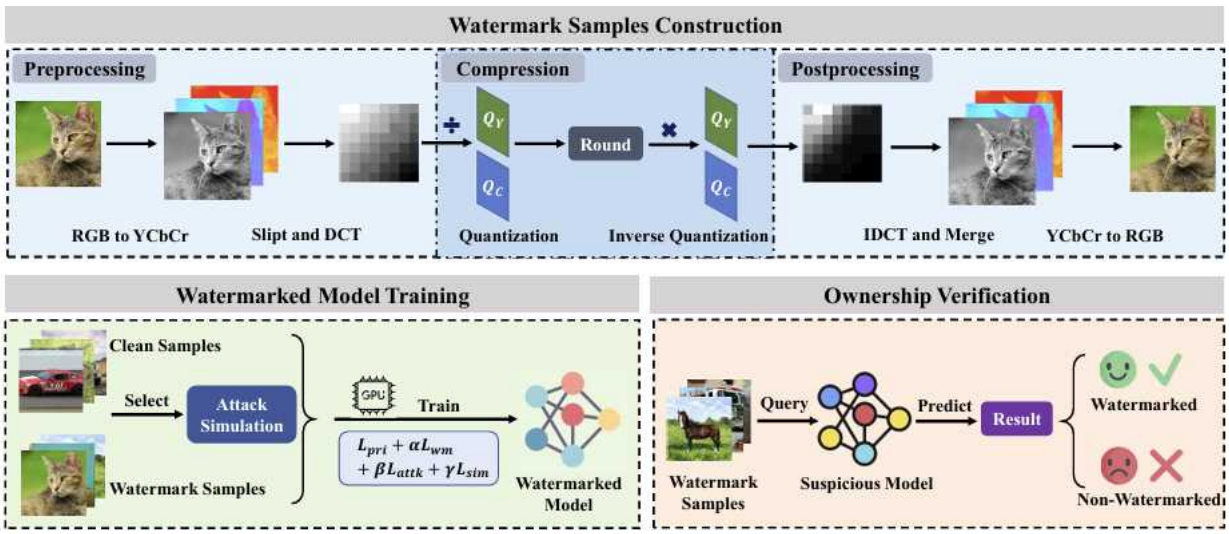
• ComMark의 핵심 원리와 접근법

- ComMark는 모델의 내부 구조나 파라미터에 접근할 수 없는 블랙박스 환경을 위해 특별히 설계된 최신 모델 워터마킹 프레임워크 기술임
- 해당 기술의 핵심은 원본에 미세한 노이즈를 추가하는 기존 방식에서 벗어나, 주파수 영역에서 정보를 압축하여 인간의 시각 시스템이 거의 인지할 수 없는 워터마크 샘플을 생성하는 데 있음
- 이산 코사인 변환*(Discrete Cosine Transform, DCT)과 같은 주파수 변환 기법을 활용하여 이미지의 고주파 성분을 제거함으로써, 워터마크가 이미지의 본질적인 특징은 해치지 않으면서도 은밀하게 삽입될 수 있도록 설계함

* 이산 코사인 변환(Discrete Cosine Transform): 특정 신호를 코사인 함수의 합으로 분해하여 주파수 영역으로 변환하는 수학적 기법으로, 데이터 압축에 사용되는 핵심 원리임

- 이를 통해 초기 기술들의 가장 큰 문제였던 워터마크의 견고성(robustness)과 비가시성 (imperceptibility) 사이의 상충 관계를 극복하고, 두 성능을 동시에 달성하는 것을 목표로 함

[그림] ComMark의 워터마크 샘플 생성, 모델 학습 및 소유권 검증 절차



출처: Yunfei Yang 외 5인, "ComMark: Covert and Robust Black-Box Model Watermarking with Compressed Samples", arXiv, 2025.12.16., https://arxiv.org/pdf/2512.15641

• 공격 시뮬레이션 기반의 워터마크 훈련 및 검증

- 소유자는 먼저 자신만이 아는 비밀 키를 이용해, 특정 이미지에 의도된 노이즈 패턴이 포함된 다수의 입력용 이미지 세트인 '트리거 세트(Trigger Set)'를 생성함
- 이후 보호하려는 원본 모델이 트리거 세트를 입력받았을 때 약속된 특정 레이블을 출력하도록 학습시키는데, 이때 '공격 시뮬레이션'을 훈련 과정에 포함시키는 것이 핵심 차별점임
- 모델의 원래 성능 유지, 워터마크 삽입, 공격 방어, 유사도 유지 등 네 가지 목표를 동시에 최적화하는 특수한 손실 함수를 사용하여, 모델이 스스로 다양한 공격을 예측하고 이에 저항하도록 훈련됨
- 도용이 의심되는 외부 모델이 발견되면, 모델 소유자는 해당 모델의 API에 트리거 세트 이미지를 입력해보고 약속된 레이블이 일관되게 출력되는지를 확인하는 과정만으로 소유권을 증명할 수 있음

• 주요 성능 평가 및 공격 방어 능력

- ComMark 워터마크가 적용된 이미지들은 객관적 화질 평가 지표인 PSNR* 및 SSIM**에서 높은 점수를 획득하여, 원본 대비 시각적 품질 저하가 거의 발생하지 않는 뛰어난 은닉성을 증명함
- 특히 모델 증류, 파라미터 가지치기와 같은 고도화된 공격은 물론, 모델 압축 및 미세조정 등 다양한 유형의 공격 전반에 걸쳐 일관되게 높은 워터마크 탐지 성공률을 유지하며 기존 기술을 뛰어넘는 강력한 견고성과 범용성을 동시에 입증함
- 이는 기존 기술들이 쉽게 무력화되던 실질적인 위협 환경 속에서도 ComMark가 AI 모델의 IP를 효과적으로 보호할 수 있는 실용적 기술임을 시사함

* PSNR(Peak Signal-to-Noise Ratio): 원본 이미지와 워터마크가 삽입된 이미지 간의 품질 차이를 측정하는 대표적인 지표로, 신호 대비 잡음의 비율을 나타내며 값이 높을수록 이미지 손실이 적고 원본과 유사함을 의미

**SSIM(Structural Similarity Index Measure): 밝기, 대비, 구조적 정보 등을 종합하여 두 이미지 간의 시각적 유사성을 평가하는 지표로, 인간의 시각적 인지와 유사하게 측정하며 1에 가까울수록 거의 동일한 이미지임을 의미

결론 및 시사점

- ComMark는 MLaaS 기업의 서비스 모델을 불법 복제로부터 보호하고, 법적 분쟁 발생 시 결정적인 디지털 포렌식 증거를 제공하는 등 AI 모델의 지식재산권을 보호할 실질적 수단을 제시했다는 점에서 중요한 산업적 의의를 가짐
- 향후 AI 모델 보호는 ComMark와 같은 사후 증명 기술과 더불어, 모델 도용 시도를 사전에 탐지하고 차단하는 기술이 상호 보완적으로 결합된 다층적 방어 시스템으로 발전해야 할 것임
- 궁극적으로 AI 모델의 가치를 온전히 보장하기 위해서는 이러한 기술적 노력과 더불어, 이를 명확히 규정하고 집행할 수 있는 법적, 제도적 기반 마련이 반드시 병행되어야 함

참고문헌

- Yunfei Yang 외 5인, "ComMark: Covert and Robust Black-Box Model Watermarking with Compressed Samples", arXiv, 2025.12.16., https://arxiv.org/pdf/2512.15641
- Or Eshed, "Model Theft in AI: How IP and Models Get Stolen", Layer X, 2025.10.21., https://layerxsecurity.com/generative-ai/model-theft/