

저작권 이슈 브리프



COPYRIGHT ISSUE BRIEF

Weekly Report
2026. 10-1



한국저작권위원회
KOREA COPYRIGHT COMMISSION

본 보고서는 EC21R&C(컨설팅사)에서 작성하였고, 국내외 저작권 기술·산업 동향을 조사한 자료로 한국저작권위원회 의견이 반영되어 있지 않습니다.



저작권 이슈 브리프

SUMMARY

산업/기업

기술

기업 | 수노 v6 사례로 본 라이선스 음원 전면 재학습과 AI 생성 음원 관리

▶ AI 음악 생성 서비스의 학습 데이터 출처가 쟁점이 되면서, 수노는 2025년 11월 라이선스 기반 모델을 예고한 뒤 워너뮤직그룹, BMG, 빌리브와 계약을 맺어 학습 음원을 확보하고 있다. 수노는 2026년 9월 9일 v6 모델 3종을 공개하고 기존 모델을 폐기해, 기존 모델을 보완하는 대신 라이선스 데이터로 처음부터 다시 학습시키는 방식을 택했다. 구독 등급별 다운로드 제한과 파생 창작 기능도 도입했으나, 워터마킹과 핑거프린팅은 아직 계획 단계에 머물러 있다. 이에 따라 학습 범위는 계약 진행에 따라 달라지고, 산출물이 플랫폼 밖으로 유통되면 AI 생성 음원인지 확인하기 어렵다. 향후에는 식별 신호를 실제로 삽입하고 유통 단계에서 이를 확인할 수 있는 경로를 마련할 필요가 있다.

산업 | AI 봇 일괄 차단에 한계 및 클라우드플레이어의 용도별 트래픽 관리

▶ AI 봇이 웹사이트 콘텐츠를 수집해 학습이나 응답 생성에 활용하면서도 원문 웹사이트로 이용자를 유입시키지 않자, 웹사이트 소유자가 AI 봇을 일괄 차단하는 수단이 먼저 등장했다. 그러나 일괄 차단은 검색 노출과 학습 허용을 구분하지 못해, 소규모 웹사이트는 학습을 허용하거나 검색 노출을 포기해야 하는 양자택일에 놓였다. 이에 클라우드플레이어는 봇의 이용 목적을 검색, 에이전트, 학습으로 구분해 관리하도록 하고, 2026년 9월 15일부터 광고 페이지에서는 학습과 에이전트 접근을 차단하는 기본값을 적용했다. 다만 검색과 학습을 겸하는 봇에는 가장 제한적인 규칙이 적용되고 robots.txt도 강제력이 없는 만큼, 봇 운영자가 이용 목적을 구분해 밝히고 이를 실제로 준수하는지 확인할 수단이 필요하다.

기업 | 팬그램의 AI 생성 텍스트 탐지 기술과 신뢰성의 한계

▶ 2026년 출판 분야에서 AI 생성 의혹이 잇따르며 출판사 아셰트가 공포소설 '사이겔'의 출간을 취소하는 등 기존 출간 결정까지 반복되고 있다. 그러나 오픈AI는 워터마크 기술을 공개하지 않았고 출판사도 검수 방식을 거의 공개하지 않아, AI 생성 여부를 판별할 수단은 제한적이다. 이에 팬그램은 AI 사용 정도를 백분을 탐지 점수로 제시하며 서브스택 통합과 문학 에이전시의 원고 분석으로 활용 범위를 넓혔다. 팬그램은 오탐률을 0.0041%까지 낮췄다고 밝혔지만, 가벼운 AI 편집은 과하게 판정하고 휴머니저를 거친 AI 생성 텍스트는 대부분 놓치는 한계가 있다. 따라서 탐지 점수는 작업 이력 등 다른 근거와 함께 판단하는 보조 수단으로 활용할 필요가 있다.



저작권 이슈 브리프

SUMMARY

산업/기업

기술

기업 유니파이드의 패키징 단계 C2PA 메타데이터 삽입과 라이브 영상 출처 증명 확대

▶ 영상이 편집, 재패키징, AI 생성을 거쳐 여러 플랫폼으로 유통되면서 출처 기록과 편집 기록이 영상과 함께 전달되지 않아, 특히 라이브 채널과 속보 영상에서 출처를 확인하기 어려워지고 있다. 이에 유니파이드 스트리밍은 2026년 8월 배포 직전의 최종 패키징 단계에서 C2PA 메타데이터를 삽입하는 트러스티드 미디어를 공개했다. 이 도구는 발행자 신원, 출처 표기, 편집 맥락, AI 사용 공개 정보를 CDN 전송 직전에 삽입하고 영상 재생기에서 실시간으로 검증할 수 있게 해, 배포 경로를 새로 구축하지 않고도 기존 워크플로에 출처 증명을 가능케 했다. 또한 정적 이미지를 주로 다뤄 온 기존 C2PA 적용과 달리 스트림마다 동적으로 서명을 삽입해, 라이브 피드와 VOD, 광고 등이 섞인 혼합 선형 스트림까지 적용 범위를 넓혔다.

산업 시스템별 로그 조사의 한계와 보호 시스템 간 데이터 연계 탐지로의 전환

▶ AI로 콘텐츠 불법 유통이 자동화되면서 사후 삭제는 확산 속도를 따라가지 못하고, 개별 로그 조사로는 여러 보호 시스템에 걸친 불법 유통 패턴을 파악하기 어려워지고 있다. 이에 오스트리아 미디어 기술 기업 빅블루마블은 2026년 9월 IBC2026에서 콘텐츠 보호 스위트의 AI 지원 탐지 및 추적 기능을 공개했다. 이 기능은 DRM, 동시 접속 보호, 포렌식 워터마킹의 행위 데이터를 종합해 계정 공유와 무단 재유통 패턴을 식별하고, 불법 스트림 녹화 영상의 워터마크를 분석해 유출 가입자나 세션을 특정한다. 이러한 결과를 각 보호 시스템의 집행 기능과 연계해 탐지부터 조치까지 하나의 절차로 연결한다. 다만 시연 단계인 만큼 실제 성과는 확인이 필요하며, 자동 대응이 확대될수록 운영자가 AI의 판단 근거를 확인하는 절차도 중요해진다.

기술 주간 기술 동향

▶ 생성형 AI로 이미지 일부를 바꾸거나 합성하는 일이 쉬워지면서, 특정 영역에만 워터마크를 삽입하는 로컬 워터마킹이 객체 단위의 출처 확인 수단으로 주목받고 있다. 그러나 기존 연구는 기술마다 평가 조건이 달라 강건성을 서로 비교하기 어려웠다. 이에 워터마킹 기술 5종에 55개 이미지 변형을 동일하게 적용해 페이로드 복원, 워터마크 영역의 위치 추정, 비가시성을 비교한 벤치마크가 제시되었다. 평가 결과 MaskWM이 가장 우수했으나 화질 저하도 가장 컸으며, 모든 기술이 정렬 변화와 생성형 편집에 취약했다. 특히 인페인팅은 워터마크 영역을 건드리지 않아도 이미지 전체를 다시 생성해 페이로드 복원을 실패하게 만들었다. 즉, 현재 로컬 워터마킹 기술은 실제 배포에 충분한 수준이 아니다.



저작권 이슈 브리프

SUMMARY

산업/기업

기술

수노 v6 사례로 본 라이선스 음원 전면 재학습과 AI 생성 음원 관리

AI 음악 생성 서비스의 학습 데이터 출처 대응과 모델 전면 교체의 등장

• 여러 권리자의 침해 주장에 따른 라이선스 음원 확보 확대

- AI 음악 생성 서비스 수노(Suno)는 저작권이 있는 음원을 모델 학습에 사용했다는 이유로 소니뮤직엔터테인먼트(Sony Music Entertainment), 유니버설뮤직그룹(Universal Music Group), 워너뮤직그룹(Warner Music Group) 등 음반사로부터 잇따라 소송을 당함¹⁾
- 이에 수노는 2025년 11월 워너뮤직그룹과의 소송을 합의로 종결하고 2026년 중 라이선스 음원 기반의 재학습 모델을 출시하겠다고 예고하며²⁾, 권리자와의 계약을 통해 학습 데이터를 확보하는 방향으로 대응함
- 이후 수노는 음반사 BMG와 라이선스 계약을 체결해 BMG의 음원을 모델에 추가하기 시작함. 또한 음악 유통사 빌리브(Believe)와는 참여 의사를 밝힌 창작자와 음반사의 음원만 학습에 사용하기로 합의해 라이선스 음원의 확보 범위를 넓힘

• 기존 모델 보완에서 전면 재학습으로의 전환

- 이를 바탕으로 수노는 2026년 9월 9일 학습 데이터 출처가 문제된 기존 모델을 모두 단종하고, 계약을 맺은 권리자와 함께 개발한 v6 모델 3종을 출시해 모델 전체를 교체함
- 수노의 공동창업자 겸 최고경영자 마이키 쉘먼(Mikey Shulman)은 v6가 창작의 가능성을 넓히고 음악 산업 전반에 수익 기회를 제공할 수 있도록 설계되었다고 설명함

학습 단계의 데이터 교체에서 유통 단계의 관리까지 확장

• 라이선스 음원과 선호 데이터로 전면 재학습

- v6 모델 3종은 구독자용 기본 모델 v6, 실험적 모델인 v6-와일드(v6-wild), 처리 속도를 높인 v6-미니(v6-mini)로 구성됨. v6와 v6-와일드는 프로(Pro)와 프리미어(Premier) 구독자를 대상으로 하며, v6-미니는 누구나 이용 가능함
- v6는 기존 모델과 다른 학습 데이터로 전면 재학습됐으며, 학습에 쓰인 음원은 계약을 맺은 권리자가 직접 선정해 제공함. 반면 수노와 소송 중인 유니버설뮤직그룹과 소니뮤직엔터테인먼트의 음원은 학습 데이터에서 제외됨

1) Ivan Mehta, "Suno replaces its AI models with a new one trained on licensed music as copyright suits pile up", TechCrunch, 2026.09.09.,

<https://techcrunch.com/2026/09/09/suno-replaces-its-ai-models-with-a-new-one-trained-on-licensed-music-as-copyright-suits-pile-up/>

2) Murray Stassen, "Suno v6 AI music models launch in partnership with Warner Music, BMG, and Believe", Music Business Worldwide, 2026.09.09.,

<https://www.musicbusinessworldwide.com/suno-v6-ai-music-models-launch-in-partnership-with-wmg-bmg-and-believe/>

- 수노 최고제품책임자 잭 브로디(Jack Brody)는 모델 품질을 좌우하는 요소가 학습 데이터가 아니라 모델 구조와 선호 기반 조정을 포함한 연구개발이며 파트너십의 목적도 데이터 확보가 아니라고 설명해, 제외된 음원을 성능 저하 요인으로 보지 않음을 밝힘

• 창작자의 선택 참여를 전제로 한 편집 기능과 파생 제품 확장

- v6에는 프롬프트나 가사의 특정 단어로 곡 일부를 편집하는 기능, 텍스트, 이미지, 영상을 참고해 곡을 만드는 기능, 샘플에서 악기를 분리해 새 비트를 만드는 기능이 적용돼 이용자가 곡을 수정하거나 참조 자료를 활용할 수 있는 범위가 넓어짐
- 수노는 기존 곡을 리믹스하는 기능도 추가할 계획이나, 음반사와 마련 중인 프로그램에 해당 창작자가 선택 참여(opt-in)한 경우에만 허용할 방침이어서 기존 곡을 활용한 파생 제품에는 별도의 참여 조건이 적용됨
- 술먼은 다음 단계로 개별 창작자를 중심에 둔 선택 참여 제품을 준비한다고 밝힘. 참여 창작자가 보상을 받고 팬이 새로운 방식으로 소통할 수 있도록 해, 파생 제품에서 발생하는 수익을 창작자와 나누겠다는 방향을 제시함

• AI 생성 음원 식별과 다운로드 상한을 통한 유통 관리 병행

- 한편 수노는 AI 생성 음원에 식별 신호를 삽입하는 워터마킹(watermarking)*과 음원의 특징을 대조하는 핑거프린팅(fingerprinting)** 도입을 약속해, AI 생성 음원이 유통된 뒤에도 출처를 확인할 수 있는 식별 수단을 마련하려 함
- 다만 수노는 경쟁사인 유디오(Udio)가 유니버설뮤직그룹 주도의 폐쇄형 생태계(walled garden) 모델에 따라 AI 생성 음원의 외부 반출을 막는 것과 달리, 개방형 스튜디오를 표방하며 구독 등급별 월간 다운로드 상한을 적용함

* 워터마킹(watermarking): 음원에 식별 신호를 삽입해 생성 출처를 확인할 수 있게 하는 기술

** 핑거프린팅(fingerprinting): 음원의 고유한 특징을 추출해 같은 음원인지 대조하고 식별하는 기술

재학습 중심 대응의 적용 범위와 유통 단계 연계 과제

• 개별 계약에 따라 달라지는 학습 데이터 범위의 공백

- 학습 데이터의 범위가 권리자와의 개별 계약에 따라 정해지는 만큼 계약하지 않은 권리자의 음원은 제외되며, 수노가 추가 계약을 추진하는 과정에서 학습 데이터의 범위도 달라질 수 있음
- 또한 학습 데이터보다 연구개발이 품질을 좌우한다는 수노의 설명을 평가하려면, 일부 음원이 제외된 v6가 실제로 기존 모델에 견줄 만한 품질을 내는지 확인할 필요가 있음

• 식별 수단과 유통 단계를 잇는 확인 경로 마련의 필요성

- 다운로드 상한은 내려받는 곡의 수를 제한하지만, 이용자가 AI 생성 음원을 다른 플랫폼에 올린 이후의 출처 확인까지 보장하지는 않음. 워터마킹과 핑거프린팅 역시 도입을 약속한 단계여서 외부로 유통된 AI 생성 음원의 식별 방식은 아직 구체화되지 않음
- 따라서 AI 생성 음원의 출처를 유통 단계까지 확인하려면 수노가 식별 신호를 AI 생성 음원에 실제로 삽입하고, 유통사와 플랫폼도 그 신호를 확인할 수 있는 경로를 마련해야 함



참고문헌

- Murray Stassen, "Suno v6 AI music models launch in partnership with Warner Music, BMG, and Believe", Music Business Worldwide, 2026.09.09., <https://www.musicbusinessworldwide.com/suno-v6-ai-music-models-launch-in-partnership-with-wmg-bmg-and-believe/>
- Ivan Mehta, "Suno replaces its AI models with a new one trained on licensed music as copyright suits pile up", TechCrunch, 2026.09.09., <https://techcrunch.com/2026/09/09/suno-replaces-its-ai-models-with-a-new-one-trained-on-licensed-music-as-copyright-suits-pile-up/>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

AI 봇 일괄 차단 기능의 한계 및 클라우드플레어의 용도별 트래픽 관리

AI 트래픽 일괄 차단 방식의 한계와 용도 구분 요구

• 크롤링과 유입 트래픽의 교환 관계 붕괴와 일괄 차단 기능의 도입

- 그동안 검색엔진은 웹사이트를 크롤링(crawling)*해 색인한 뒤 검색 결과에 원문 링크를 보여줌. 웹사이트 소유자는 크롤링을 허용하는 대가로 웹사이트 방문과 광고 및 구독 수익을 얻었고, 이러한 교환 관계는 30년간 유지되어 옴
- 그러나 생성형 AI는 웹페이지에서 정보를 추출해 답변을 직접 제시하므로 웹사이트를 방문할 필요가 줄어들음. AI가 콘텐츠를 가져가면서도 그 대가인 유입 트래픽은 발생하지 않음
- 이에 미국 클라우드 서비스 기업 클라우드플레어(Cloudflare)는 2025년 원클릭 AI 봇 차단 기능(Block AI Bots)과 페이 퍼 크롤(Pay-Per-Crawl)** 마켓플레이스를 출시해¹⁾, 웹사이트 소유자가 콘텐츠 이용 여부를 직접 정하도록 함

* 크롤링(crawling): 자동화된 프로그램이 인터넷 웹페이지를 돌아다니며 필요한 정보를 수집하고 분류하는 과정

** 페이 퍼 크롤(Pay-Per-Crawl): AI 크롤러가 웹사이트 콘텐츠를 가져갈 때마다 요청 단위로 요금을 부과할 수 있게 하는 과금 모델

• 검색과 학습을 겸하는 봇 운영에서 비롯된 용도 구분의 필요성

- 다만 봇을 한꺼번에 막는 일괄 차단이 모든 웹사이트에 적합한 것은 아님. 예를 들어 소규모 웹사이트는 검색에 노출됨으로써 AI가 학습하도록 내버려두거나, 검색 결과에 나타나지 않을 위험을 감수하는 것 사이에서 하나를 택해야 함
- 이러한 양자택일은 검색과 학습에 동일한 봇을 사용하는 기존 검색 기업에게 유리하게 작용함. 또한 후발 검색 기업들은 기존 검색 기업과의 격차를 좁히기 위해 봇의 용도를 숨길 유인을 갖게 됨
- 더욱이 검색이 결과 화면에서 원문 링크를 늘어놓는 대신 직접 답하는 방향으로 이동하고 있어, 봇이 웹사이트에서 무엇을 하고 무엇을 저장하며 콘텐츠를 어떻게 재공유하는지를 확인할 수 있는 구분이 필요해짐

봇의 용도 분류에서 콘텐츠 이용 수준 설정으로의 확대

• 봇의 행위를 기준으로 삼은 세 용도 분류와 기본값의 전환

- 이 문제에 대응해 클라우드플레어는 2026년 7월 봇의 행위를 기준으로 용도를 검색(Search), 에이전트(Agent), 학습(Training)의 세 가지로 구분함. 웹사이트 소유자는 용도별로 '전면 허용', '전면 차단', '광고 페이지에서 차단' 중에서 선택할 수 있음¹⁾

1) Jin-Hee Lee 외 1명, "Your site, your rules: new AI traffic options for all customers", Cloudflare Blog, 2026.07.01., <https://blog.cloudflare.com/content-independence-day-ai-options/>

- 이 분류에 따라 2026년 9월 15일부터 신규 도메인에는 광고 페이지에서 학습 및 에이전트 용도를 차단하고 검색 용도는 허용하는 기본값이 적용됨. 광고 페이지를 사람이 보도록 만든 것으로 보고, 그 페이지에서는 검색만 허용하는 것임
- 이 기본값은 9월 15일까지 설정을 바꾸지 않은 기존에 무료 요금제를 이용하던 웹사이트 소유자에게도 적용되며, 웹사이트 소유자는 클라우드플레어 대시보드를 통해 설정을 변경할 수 있음
- 아울러 여러 용도를 겸하는 봇에는 모든 행위를 기준으로 가장 제한적인 규칙이 적용됨. 따라서 학습을 금지한 웹사이트에서는 검색과 학습을 함께 수행하는 구글봇(Googlebot), 애플봇(Applebot), 빙봇(BingBot)이 차단될 수 있음

[표 1] 클라우드플레어의 AI 트래픽 용도 분류와 기본값-

구분	봇의 행위	신규 도메인 기본값 (2026년 9월 15일부터)
검색 (Search)	이후 질문에 답하기 위해 웹사이트의 콘텐츠를 수집하고 색인함	허용
에이전트 (Agent)	사람을 대신해 실시간으로 페이지를 방문해 작업을 처리함	광고 페이지에서 차단
학습 (Training)	모델을 학습하거나 미세조정하기 위해 콘텐츠를 수집함	광고 페이지에서 차단

출처: Jin-Hee Lee 외 1명, "Your site, your rules: new AI traffic options for all customers", Cloudflare Blog, 2026.07.01., <https://blog.cloudflare.com/content-independence-day-ai-options/>

• 콘텐츠 이용 수준 설정과 robots.txt 신호로의 확장

- 용도 분류만으로는 봇이 크롤링한 뒤 콘텐츠를 어떻게 보관하고 재공유하는지까지 다루지 못함. 이에 클라우드플레어는 그 범위를 ‘즉시(immediate)’, ‘참조(reference)’, ‘전체(full)’ 세 수준으로 나눠 선택하거나 차단하는 기능을 구축하고 있음
- 세 수준은 봇이 콘텐츠를 재공유할 수 있는 범위 순으로 나뉨. ‘즉시’는 상호작용만 하고 아무것도 저장하거나 재공유하지 않으며, 기본값인 ‘참조’는 색인과 발췌, 원문 링크를 허용하고, ‘전체’는 요약과 재현(reproduce)까지 허용함
- 이렇게 나뉜 이용 수준을 용도 분류와 결합하여, 검색과 광고 노출 확인 용도의 봇은 참조 수준까지만 허용하는 식으로 규칙을 정할 수 있음. 이에 따라 웹사이트 소유자는 봇을 하나씩 관리하는 대신 비슷한 용도끼리 묶어 판단 가능함
- 이와 함께 클라우드플레어는 콘텐츠 시그널(Content Signals)*에 이용 수준을 적는 ‘use’ 항목을 더해 관리형 robots.txt**에 참조 수준을 보여줌. 다만 이 값은 봇을 차단하는 것이 아니라 웹사이트 소유자가 바라는 이용 범위를 알리는 데 그침

* 콘텐츠 시그널(Content Signals): 검색과 AI 학습 등 콘텐츠 이용을 어디까지 허용할지 웹사이트 소유자가 robots.txt에 적어 두는 규칙

** robots.txt: 봇의 접근과 콘텐츠 이용을 어디까지 허용할지 웹사이트 소유자가 적어 두는 파일

용도 기반 관리의 실효성 확보와 기술 과제

• 봇의 용도 분리와 이용 수준 준수에 좌우되는 실효성

- 이번 용도별 관리 도입은 그 실효성이 봇 운영자가 용도마다 봇을 따로 두는지에 달려 있음을 보여줌. 검색과 학습을 하나의 봇으로 함께 수행하면 가장 제한적인 규칙이 적용돼, 웹사이트 소유자가 학습만 막으려 해도 검색 노출까지 함께 잃게 됨
- 이에 따라 봇이 스스로를 어떻게 밝히느냐에 따라 같은 설정에서도 허용 범위가 달라지는 만큼, 용도마다 봇을 나눠 운영하고 그 용도를 정확히 밝히도록 유도하며, 선택한 이용 수준을 지키는지 확인할 수단을 함께 갖추는 일이 과제로 남음

참고문헌

- Jin-Hee Lee 외 1명, "Your site, your rules: new AI traffic options for all customers", Cloudflare Blog, 2026.07.01., <https://blog.cloudflare.com/content-independence-day-ai-options/>
- Nathaniel Lacsina, "The rules of the web change today as Cloudflare cracks down on AI bots", Gulf News, 2026.09.15., <https://gulfnews.com/technology/the-rules-of-the-web-change-today-as-cloudflare-cracks-down-on-ai-bots-1.500674972>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

팬그램의 AI 생성 텍스트 탐지 기술과 신뢰성의 한계

출판 분야의 AI 생성 의혹 확산과 판별 수단의 공백

• 잇따른 AI 생성 의혹에 따른 독자와 출판계 간 신뢰의 균열

- 2026년 들어 출판 분야에서 작가가 AI로 작품을 썼다는 의혹이 잇따르면서 미국 출판사 아세트 북그룹(Hachette Book Group, 이하 아세트)이 공포소설 '샤이걸(Shy Girl)'의 출간을 취소하는 등 이미 내려진 출간 결정까지 뒤집히는 사례가 나옴
- AI 사용 여부가 독자의 작품 수용과 직결되는 만큼 이러한 의혹은 파장이 큼. 실제로 2025년 여론조사기관 유고브(YouGov) 조사에서 미국 독자 가운데 61%는 읽은 작품이 AI로 쓰였음을 알게 되면 만족감이 낮아질 것이라고 답함¹⁾

• 생성 단계 식별 기술의 미공개와 출판사 검수 기준의 불명확성

- 이 가운데 생성 단계에서 식별 정보를 넣는 워터마크(watermark)가 해법으로 거론되나, AI 생성 텍스트에 워터마크를 넣을지를 두고 AI 기업의 대응이 엇갈림
- 생성형 AI 개발사 오픈AI(OpenAI)는 AI 챗봇 챗GPT(ChatGPT)의 생성 텍스트를 99.9% 정확도로 탐지하는 워터마크 기술을 개발했으나 기술의 적용 범위와 신뢰성에 대한 내부 논쟁이 진행 중인데다 사용자 감소를 우려해 공개하지 않았다고 2024년 보도됨
- 반면 경쟁사인 앤트로픽(Anthropic)은 2026년 8월 자사 AI 챗봇 클로드(Claude)의 모든 AI 생성물에 사용자가 선택적으로 해제(opt out)할 수 없는 워터마크를 삽입할 것이라고 밝힘
- 한편 출판사의 검수 기준도 불명확함. 아세트는 작가에게 작품이 본인이 작성한 것임을 서약하도록 요구하나, 출판 전문매체 북라이엇(Book Riot)은 검수 방식을 공개한 출판사가 거의 없다고 지적함

팬그램 탐지 점수의 확산과 판정 신뢰성의 한계

• 합성 미러링 학습과 보수적 판정 설계에 따른 낮은 오탐률

- 판별 수단에 공백이 있는 상황에서 미국 AI 텍스트 탐지 스타트업 팬그램(Pangram)은 텍스트 작성에 AI를 사용한 정도를 백분율 탐지 점수로 제시해, 출판 분야의 대표 탐지 도구로 떠오름

1) Cameron Pugh, "Caught in the backlash over AI, book publishers face mounting questions", The Christian Science Monitor, 2026.08.12., <https://www.csmonitor.com/Arts-Culture/Books/2026/0812/ai-publishing-commonwealth-prize-shy-girl>

- 학습 단계에서는 합성 미러링(synthetic mirroring)*으로 AI의 글쓰기 방식을 익히고, 하드 네거티브 마이닝(hard negative mining)**으로 오탐 사례를 다시 학습해, 사람이 작성한 텍스트를 AI 생성 텍스트로 판정하는 경우를 줄임
- 팬그램의 신규 모델인 팬그램 4의 오탐률(false positive rate)***이 0.0041%로 이전에 보고된 0.01%보다 개선됨²⁾

* 합성 미러링(synthetic mirroring): 사람이 작성한 텍스트를 바탕으로 AI 모델이 비슷한 글을 생성하게 해 AI의 글쓰기 방식을 학습하는 방법

** 하드 네거티브 마이닝(hard negative mining): 탐지 도구의 오탐 사례를 찾아 합성 미러링을 거쳐 학습 자료에 추가하는 방법

*** 오탐률(false positive rate): 사람이 작성한 텍스트를 AI 생성 텍스트로 잘못 판정하는 비율

● 플랫폼 통합과 출판 현장 도입에 따른 판정 영향력의 확대

- 온라인에서 사이겔이 AI로 쓰였다는 의혹이 일자, 팬그램의 공동창업자 겸 CEO 맥스 스페로(Max Spero)는 이 작품을 78% AI 생성으로 판정한 결과를 소셜 미디어 X에 공개했고³⁾, 이후 아셰트는 이 작품의 출간을 취소함
- 실제로 뉴스레터 플랫폼 서브스택(Substack)은 7월 말 팬그램 서비스를 제공해 독자가 AI 사용 가능성을 바로 확인하게 했고, 미국 뉴욕의 문학 에이전시 에이비터스(Aevitas)는 원고와 출간 제안서 분석에 팬그램을 활용함

● 짧은 텍스트와 회피 도구 앞에서 드러나는 탐지 정확도의 한계

- 그러나 탐지 정확도는 텍스트 길이와 문맥에 따라 달라짐. 팬그램은 100단어 미만 텍스트에서 성능이 낮아지고 같은 구절도 단독으로 넣을 때와 긴 문맥 안에 넣을 때 판정이 달라질 수 있음을 인정함
- 더욱이 편집 수준에 따른 판정도 불안정해, 미국 노트르담대학교(University of Notre Dame) 연구 초안에서 팬그램 3.2 모델은 가벼운 AI 편집만 거친 학술 초록을 64~80%의 비율로 AI 생성 텍스트인 것으로 판정함³⁾
- 반대로 AI 생성 텍스트를 휴머니저(humanizer)*로 고치면 팬그램의 탐지율이 4% 미만으로 떨어져, 탐지 점수가 과잉 판정과 미탐지 양쪽으로 모두 흔들릴 수 있음을 보여 줌³⁾
- 이러한 한계는 출판 분야에서 부담이 특히 큼. 미국 미시간대학교(University of Michigan) 박사과정생 타라 라드반드(Tara Radvand)는 의혹 하나로 경력이 무너질 수 있는 출판 분야에서는 1% 오탐률도 높다고 지적함⁴⁾

* 휴머니저(humanizer): AI 생성 텍스트를 사람이 작성한 텍스트처럼 보이도록 고쳐 주는 도구

탐지 점수의 보조적 활용과 생성 단계 식별의 병행 과제

● 탐지 점수를 단독 판정 근거로 삼는 데 따른 오인 가능성

- 팬그램의 판정은 가벼운 AI 편집에는 과하게, 휴머니저로 고친 AI 생성 텍스트에는 약하게 반응해, 특정 탐지 점수만으로 작가의 AI 사용 여부와 정도를 단정하기 어려움

2) Progressive Robot, "Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?", Progressive Robot, 2026.09.02., <https://www.progressiverobot.com/2026/09/02/pangram-ai-detection-gold-standard-trust/>

3) Lexi Pandell, "Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?", WIRED, 2026.09.02., <https://www.wired.com/story/pangram-has-emerged-as-the-gold-standard-of-ai-detection/>

4) Cameron Pugh, "Caught in the backlash over AI, book publishers face mounting questions", The Christian Science Monitor, 2026.08.12., <https://www.csmonitor.com/Arts-Culture/Books/2026/0812/ai-publishing-commonwealth-prize-shy-girl>

- 특히 서브스택처럼 독자가 직접 탐지 점수를 확인하는 곳에서는, 짧은 텍스트라 탐지 정확도가 떨어지는 경우 작가가 AI를 사용한 것으로 오인될 수 있음. 1% 오탐률도 높다고 평가되는 출판 분야에서 이러한 오인은 작가 경력에 직접 타격을 줄 수 있음
- 따라서 탐지 점수는 작업 이력 등 다른 근거와 함께 판단하는 보조 수단으로 활용할 필요가 있음

• 워터마크와 사후 탐지를 결합한 판별 체계의 필요성

- 나아가 사후 탐지는 완성된 텍스트만 보고 AI 사용 정도를 추정하므로 휴머니저로 고친 AI 생성 텍스트나 짧은 텍스트는 가려내기 어려움. AI 사업자가 생성 단계에 삽입하는 워터마크가 이를 보완하는 수단이 될 수 있음
- 다만 생성 단계 식별 역시 오픈AI처럼 기술을 공개하지 않는 기업이 있으면 공백이 생기므로, 워터마크와 사후 탐지 도구를 함께 운영하고 탐지 도구의 성능 수치를 공개하는 방안이 과제로 남음

참고문헌

- Cameron Pugh, "Caught in the backlash over AI, book publishers face mounting questions", The Christian Science Monitor, 2026.08.12., <https://www.csmonitor.com/Arts-Culture/Books/2026/0812/ai-publishing-commonwealth-prize-shy-girl>
- Progressive Robot, "Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?", Progressive Robot, 2026.09.02., <https://www.progressiverobot.com/2026/09/02/pangram-ai-detection-gold-standard-trust/>
- Lexi Pandell, "Pangram Has Emerged as the Gold Standard of AI Detection. Should You Trust It?", WIRED, 2026.09.02., <https://www.wired.com/story/pangram-has-emerged-as-the-gold-standard-of-ai-detection/>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

유니파이드의 패키징 단계 C2PA 메타데이터 삽입과 라이브 영상 출처 증명 확대

유통 과정에서 끊기는 영상 출처 기록과 C2PA 적용 논의의 확대

• 편집과 재가공 과정에서 발생하는 영상 출처 기록의 유실

- 영상이 편집, 클립 추출, 재패키징(repackaging)*, AI 생성 등을 거쳐 여러 플랫폼으로 유통되면서 기존의 출처 및 편집 기록이 영상 파일과 함께 이동하지 않는 경우가 많음. 이로 인해 시청자와 뉴스룸이 원본의 출처와 처리 과정을 확인하기 어려워짐
- 더욱이 이러한 어려움은 라이브 채널과 속보 영상에서 특히 크게 두드러져, 영상이 빠르게 송출되는 환경에서도 출처를 확인할 수 있도록 유통 단계에서 검증 신호를 삽입할 수단이 필요해짐

* 재패키징(repackaging): 이미 배포용으로 처리된 영상을 다른 플랫폼이나 배포 환경에 맞게 다시 처리하는 과정

• 방송 업계의 C2PA 기반 출처 증명 논의 확대

- 이러한 가운데 2026년 9월 암스테르담에서 열린 국제방송박람회(International Broadcasting Convention, 이하 IBC)에서는 C2PA* 참여 기관의 교류 모임이 두 차례 열리는 등 출처 증명(provenance)** 논의가 이어짐
- 실제로 노르웨이 미디어 산업 클러스터인 미디어 클러스터 노르웨이(Media Cluster Norway)가 주관한 조찬 행사에서는 CBC/라디오캐나다(CBC/Radio-Canada) 등 5개 기관이 발표에 나서, 출처 증명이 미디어 플랫폼에 대한 시청자 신뢰를 높일 수 있음을 보여줌¹⁾
- 아울러 뉴스 미디어 국제 표준 기구 IPTC(International Press Telecommunications Council) 등이 함께 쓴 기술 논문은 참여 기관 간 협력을 돕도록 미디어 워크플로(media workflow)***를 정리한 C2PA 출처 증명 프레임워크(framework)를 제시함

* C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 진위 확인을 위한 연합체로, 영상에 삽입해 출처를 검증할 수 있는 메타데이터 규격을 제시함

** 출처 증명(provenance): 영상이 어디에서 만들어졌고 어떤 처리를 거쳤는지를 검증 가능한 정보로 밝히는 것

*** 미디어 워크플로(media workflow): 영상이 제작, 편집, 패키징, 배포를 거쳐 시청자에게 재생되기까지 여러 주체가 이어서 처리하는 작업 흐름

1) Brendan Quinn, "Trust, Story Objects and Agentic Production: IPTC at IBC 2026", IPTC, 2026.09.14., <https://iptc.org/news/trust-story-objects-and-agentic-production-iptc-at-ibc-2026/>

유니파이드의 패키징 단계 C2PA 출처 증명 통합

• 배포 직전 최종 패키징 단계의 C2PA 메타데이터 삽입

- 이처럼 출처 기록이 끊기는 문제에 대응해, 스트리밍 오리진 및 패키징 기술 기업 유니파이드 스트리밍(Unified Streaming, 이하 유니파이드)은 2026년 8월 스트림에 C2PA 메타데이터(metadata)*를 삽입하는 도구 트러스티드 미디어(Trusted Media)를 공개함
- 이 도구는 발행자 신원, 출처 표기, 편집 맥락, AI 사용 공개 정보를 담은 C2PA 메타데이터를 최종 패키징(packaging)** 단계에서 삽입해, 콘텐츠 전송 네트워크(Content Delivery Network, 이하 CDN)로 전송하기 직전에 영상에 검증 신호를 삽입함²⁾
- 이는 유니파이드의 오리진 및 패키징 기술에 내장되어, 방송사, 스트리밍 플랫폼, 권리자는 배포 경로를 새로 구축하거나 아카이브 영상을 다시 처리하지 않고도 기존 워크플로에 출처 증명을 더할 수 있음
- 유니파이드의 최고기술책임자 아르헨 바헤나르(Arjen Wagenaar)도 검증 신호는 영상이 배포용으로 준비되는 지점에서 삽입해야 한다면, 자사 기술이 전 세계 방송사와 플랫폼의 배포 과정에 이미 쓰이고 있어 이 기능을 넣기 적합하다고 설명함

* 메타데이터(metadata): 영상에 관한 정보를 담아 영상과 함께 전달되는 데이터로, C2PA 메타데이터에는 발행자 신원, 출처 표기, 편집 맥락, AI 사용 공개 정보 등이 담김

** 패키징(packaging): 영상을 CDN으로 보내기 전에 배포에 맞게 준비하는 처리 과정

• 혼합 선형 스트림을 대상으로 한 스트림 단위 동적 서명 삽입

- 트러스티드 미디어는 라이브 피드(live feeds), 사전 인코딩된(pre-encoded) 주문형 비디오(Video on Demand, 이하 VOD), 편성 프로그램, 지역별 편성본, 광고와 그래픽이 섞인 혼합 선형 스트림(linear stream)*을 지원함
- 특히 유니파이드는 정적 이미지를 주로 다뤄 온 기존 C2PA 적용과 달리 스트림마다 동적으로 서명을 삽입하는 방식을 택해, 이 방식이 라이브와 가상 채널 워크플로에 적합하다고 설명함

* 선형 스트림(linear stream): 라이브 피드, VOD, 편성 프로그램 등 여러 영상을 하나로 이어 송출하는 스트림

• CDN 전달 과정의 C2PA 메타데이터 유지와 영상 재생기의 실시간 검증

- 이렇게 삽입된 C2PA 메타데이터는 영상과 함께 CDN으로 전달되고 영상 재생기에서 실시간으로 검증될 수 있어, 유통 과정에서 출처 기록이 파일과 분리되던 문제를 배포 경로 안에서 보완함
- 실제로 독일 방송연합체 WDR/ARD의 스트리밍 설계 담당자 마르틴 그로메(Martin Grohme)는 C2PA를 검증 가능한 출처 증명의 유망한 기술 기반으로 보며, 트러스티드 미디어가 영상의 출처와 처리 과정을 보여줘 시청자 신뢰를 높인다고 평가함²⁾

2) TVNewsCheck, "Unified Streaming Launches Trusted Media To Add Verifiable Provenance To Live & On-Demand Streams", 2026.08.11., <https://tvnewscheck.com/tech/article/unified-streaming-launches-trusted-media-to-add-verifiable-provenance-to-live-on-demand-streams/>

배포 경로 중심 출처 증명의 의미와 워크플로 표준화 과제

• 배포 경로에 내재화된 출처 증명의 의미

- 이번 사례는 출처 증명을 별도 절차로 두지 않고 배포 직전의 패키징 과정에 넣어, 기존 설비를 유지한 채 라이브와 VOD 영상까지 C2PA 메타데이터를 적용할 수 있음을 보여줌
- 특히 출처 기록이 끊기기 쉬운 라이브 채널과 속보 영상도 CDN 전송 직전에 서명되고 영상 재생기에서 검증되는 만큼, 시청자가 재생 단계에서 영상의 출처를 검증할 근거를 얻을 수 있음

• 여러 주체가 참여하는 영상 워크플로의 공통 프레임워크 필요성

- 다만 영상은 편집과 재패키징을 거쳐 여러 플랫폼으로 유통되므로, 한 기업의 배포 경로에서 삽입한 서명이 끝까지 효과를 내려면 워크플로에 참여하는 각 주체가 C2PA 메타데이터를 같은 기준으로 다뤄야 함
- IBC에서 참여 기관 간 협력을 돕는 워크플로 기술 모델이 제시된 것도 이러한 조율의 필요성을 보여주며, 향후에는 개별 도구 도입과 함께 공통 프레임워크를 정립하는 작업이 과제로 논의될 전망이다
- 출처 증명이 시청자 신뢰를 높일 수 있다는 발표가 나온 만큼, 제작부터 재생까지 C2PA 메타데이터를 일관되게 전달하는 체계가 갖춰지면 출처 증명이 방송 영상 유통의 기본 관리 수단으로 자리 잡을 가능성이 있음

참고문헌

- TVNewsCheck, "Unified Streaming Launches Trusted Media To Add Verifiable Provenance To Live & On-Demand Streams", 2026.08.11., <https://tvnewscheck.com/tech/article/unified-streaming-launches-trusted-media-to-add-verifiable-provenance-to-live-on-demand-streams/>
- Brendan Quinn, "Trust, Story Objects and Agentic Production: IPTC at IBC 2026", IPTC, 2026.09.14., <https://iptc.org/news/trust-story-objects-and-agentic-production-iptc-at-ibc-2026/>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

시스템별 로그 조사의 한계와 보호 시스템 간 데이터 연계 탐지로의 전환

AI로 가속되는 불법 유통과 사후 대응 중심 보호 체계의 한계

• AI 도입에 따른 불법 유통의 자동화와 규모 확대

- 시청각 불법 유통 대응 연합(Audiovisual Anti-Piracy Alliance) 사무총장 미루나 헤로바누(Miruna Herovanu)는 불법 유통이 조직적인 상업 활동으로 자리 잡았으며, AI가 이러한 활동의 속도와 규모를 확대하는 동시에 운영 비용을 낮추고 있다고 진단함¹⁾
- 실제로 스포츠 중계의 불법 유통에서는 실시간 스트림을 찾아 수집하거나, 차단을 피하기 위해 도메인과 인프라를 바꾸거나, 정식 서비스와 유사한 이용자 화면을 만드는 작업 등이 AI로 자동화되고 있음
- 여기에 에이전틱 AI(agentic AI)*가 활용되면서 스트림 삭제를 감지한 뒤 대체 스트림을 게시하고 인프라를 조정하는 과정까지 사람의 개입이 최소화될 수 있음. 이에 개별 스트림을 삭제하는 조치만으로는 불법 서비스의 운영을 중단시키기 어려워질 전망이다

* 에이전틱 AI(agentic AI): 정해진 범위 안에서 상황을 판단하고 필요한 조치를 자율적으로 수행하는 AI

• 사후 대응의 지연과 시스템별 분석의 한계

- 이처럼 불법 유통이 자동화되면서, 콘텐츠 보안 기업 나그라비전(Nagravision) 부사장 팀 피어슨(Tim Pearson)은 사업자가 대응에 나설 때는 이미 수백만 명이 콘텐츠를 시청해 수익을 잃은 뒤일 수 있다고 지적함¹⁾
- 아울러 DRM(Digital Rights Management)*, 동시 접속 보호(Concurrent Access Protection)**, 포렌식 워터마킹(forensic watermarking)***은 각각 별도의 로그를 남기므로, 이를 개별적으로 조사할 경우 계정 공유나 무단 재유통처럼 여러 시스템에 걸친 정교한 불법 유통을 찾아내기 어려움
- 이에 피어슨은 자동화된 공격에 사람이 하나씩 대응하는 방식으로는 속도 차이를 좁힐 수 없는 만큼, 위협을 평가해 정해진 범위 안에서 스스로 조치까지 수행하는 방어 시스템이 필요하다고 강조함

* 디지털 권리 관리(Digital Rights Management, DRM): 콘텐츠의 무단 복제와 재생을 막기 위해 접근 권한을 관리하는 기술

** 동시 접속 보호(Concurrent Access Protection): 하나의 계정으로 동시에 이용할 수 있는 접속 수를 제한하는 기능

*** 포렌식 워터마킹(forensic watermarking): 콘텐츠에 식별 신호를 삽입해 유출된 콘텐츠가 어느 가입자나 세션에서 나왔는지 추적하는 기술

1) Luca Berton, "IBC 2026: The AI Arms Race Against Content Piracy", lucaberton.com, 2026.09.13., <https://lucaberton.com/blog/ibc-2026-piracy-security/>

빅블루마블, 보호 시스템 로그를 연계한 AI 탐지 기능 공개

• 보호 시스템 간 행위 데이터 연계를 통한 불법 유통 패턴 탐지

- 오스트리아 미디어 기술 기업 빅블루마블(Big Blue Marble)은 DRM, 동시 접속 보호, 포렌식 워터마킹을 통합해 콘텐츠의 불법 복제와 재유통을 방지하는 '콘텐츠 보호 스위트(Content Protection Suite)'를 제공해 왔음
- 2026년 9월 IBC2026에서는 별도 보호 시스템을 구축하지 않고 기존 세 체계의 데이터를 AI로 통합 분석해 불법 이용을 탐지하고 추적하는 기능을 공개함²⁾
- 이 가운데 AI 부정 이용 탐지(AI-fraud detection) 기능은 여러 보호 시스템의 행위 데이터를 함께 분석해, 개별 로그로는 발견하기 어려웠던 계정 공유와 무단 재유통 패턴까지 탐지함
- 이 과정에서 AI는 여러 로그에서 흩어진 증거를 연결해 의심 행위를 식별하고 판단 근거를 설명으로 제시하기 때문에, 운영자는 로그 파일을 직접 대조하지 않고 곧바로 조사에 착수할 수 있음

• 불법 스트림 녹화 영상의 워터마크 분석을 통한 유출 원점 추적

- 탐지 이후에는 AI 지원 포렌식 워터마크 절차를 통해 불법 스트림으로 의심되는 녹화 영상을 분석하고, 콘텐츠가 유출된 가입자나 세션을 추적함
- 이에 따라 유통 중인 스트림뿐 아니라 유출에 사용된 정식 서비스 계정에도 조치를 취할 수 있으며, 삭제 후 대체 스트림이 게시되더라도 유출 원점을 근거로 대응을 이어갈 수 있음
- 빅블루마블 영업 이사 프란시스카 두란 브루스(Francisca Duran Bruce)는 불법 스트림을 탐지해 가능한 경우 콘텐츠를 삭제하고, 삭제가 어려운 상황을 고려해 포렌식 워터마크로 유출 경로를 원점까지 추적해야 한다고 설명함

[표 1] 빅블루마블 콘텐츠 보호 스위트의 AI 지원 기능 구성

단계	기능	분석 대상	결과
탐지	AI 부정 이용 탐지	DRM, 동시 접속 보호, 포렌식 워터마크의 행위 데이터	계정 공유와 무단 재유통 등 불법 유통 패턴 식별 및 판단 근거 설명
추적	AI 지원 포렌식 워터마크	불법 스트림으로 의심되는 녹화 영상	유출된 콘텐츠의 가입자 또는 세션 특정
집행	보호 시스템 및 콘텐츠 전송 단계와 연계된 집행 기능	탐지 및 추적 결과	탐지부터 집행 조치까지 걸리는 시간 단축

출처: IBC, "Big Blue Marble to unveil AI-assisted piracy detection at IBC2026", IBC, 2026.08.10., <https://www.ibc.org/piracy-security/news/big-blue-marble-to-unveil-ai-assisted-piracy-detection-at-ibc/22794>

• 탐지 및 추적 결과의 집행 연계와 기존 운영 환경에 적용

- 탐지 및 추적 결과는 각 보호 시스템과 콘텐츠 전송 단계의 집행 기능으로 연결되므로, 방송사는 불법 유통을 탐지한 뒤 곧바로 대응 조치에 나설 수 있음
- 아울러 빅블루마블 그룹 CEO 미하엘 바젠호퍼(Michael Wagenhofer)는 고객이 이미 갖춘 운영 환경에 맞는 서비스를 선보이겠다고 밝힘²⁾. 새 기능이 기존 보호 시스템에 더해지는 설계인 만큼, 사업자는 운영을 중단하지 않고 탐지 방식만 개선할 수 있음

2) IBC, "Big Blue Marble to unveil AI-assisted piracy detection at IBC2026", IBC, 2026.08.10., <https://www.ibc.org/piracy-security/news/big-blue-marble-to-unveil-ai-assisted-piracy-detection-at-ibc/22794>

AI 지원 탐지의 확산 의미와 운영 관리 과제

• 개별 시스템의 탐지에서 보호 시스템 간 연계 대응으로 확대

- 빅블루마블 사례는 불법 유통 대응의 초점이 새로운 탐지 도구를 추가하는 데서 기존 보호 시스템의 데이터를 연계해 분석하는 데로 옮겨가고 있음을 보여줌
- 여러 시스템에 걸쳐 드러나는 패턴은 개별 도구의 성능을 높이는 것만으로 포착하기 어려우므로, 시스템 간 데이터 연계가 탐지 성능을 좌우하는 요소로 부상함
- 또한 탐지 결과를 추적과 집행으로 즉시 연결하는 구성은 대응 시간이 손실 규모를 좌우하는 실시간 콘텐츠에서 사후 삭제의 공백을 보완할 수 있음

• 실제 운영 환경에서의 성과 검증과 관리 책임 확보

- 다만 이번 기능은 IBC2026에서 시연으로 공개된 만큼, 대량의 실시간 스트림을 처리하는 실제 운영 환경에서의 성과는 향후 사업자 적용 사례를 통해 확인할 필요가 있음
- 두란 브루스는 업계가 AI 실험을 넘어 실제 현장에 배치하고 관리 및 감독하는 데 책임을 져야 하는 단계에 들어섰다고 진단함. 이는 AI 지원 탐지의 평가 기준도 도입 자체에서 안정적인 운영 여부로 옮겨갈 것임을 시사함
- 탐지부터 집행까지 자동으로 이어지는 범위가 넓어질수록 AI의 판단이 실제 조치에 미치는 영향도 커짐. 따라서 대응 절차 안에서 운영자가 AI의 판단 근거를 확인하는 단계를 유지할 필요가 있음

참고문헌

- IBC, "Big Blue Marble to unveil AI-assisted piracy detection at IBC2026", IBC, 2026.08.10., <https://www.ibc.org/piracy-security/news/big-blue-marble-to-unveil-ai-assisted-piracy-detection-at-ibc/22794>
- Newline Report, "Big Blue Marble: cloud, IA y protección de contenidos marcan la agenda de IBC2026", Newline Report, 2026.09.13., <https://www.newslinereport.com/tecnologia/nota/big-blue-marble-cloud-ia-y-proteccion-de-contenidos-marcan-la-agenda-de-ibc2026>
- Luca Berton, "IBC 2026: The AI Arms Race Against Content Piracy", lucaberton.com, 2026.09.13., <https://lucaberton.com/blog/ibc-2026-piracy-security/>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

주간 기술 동향

로컬 워터마킹 기술 5종의 강건성 비교 평가

• AI 생성 이미지 판별의 한계와 영역 단위 출처 추적을 위한 로컬 워터마킹

생성형 AI의 확산으로 누구나 텍스트 명령만으로 실제 사진과 구별하기 어려운 이미지를 손쉽게 만들 수 있게 되었다. 과거에는 이미지 조작에 전문 편집 도구와 기술이 필요했지만, 이제 그러한 진입 장벽은 사실상 사라졌다. 최근 9.11 테러 당시 최소 12명을 구조한 것으로 알려진 웰스 크라우더(Welles Crowther)의 구조 장면이라는 이미지가 온라인에서 확산되었다. 그러나 당시 크라우더의 활동을 담은 사진은 알려진 바 없으며, 해당 이미지는 AI 생성 이미지로 밝혀졌다.

문제는 주요 AI 서비스들이 이 이미지의 진위를 제대로 판별하지 못했다는 점이다. 챗GPT(ChatGPT)는 이 이미지를 9.11 당시 실제로 촬영된 사진으로 판별했고, 클로드(Claude)는 AI 생성 이미지가 아닌 자연 사진으로 판별했으며, 그록(Grok)은 실제 사진이라고 답했다가 재차 질문하자 가짜일 수 있다고 인정했다. 구글의 AI 역시 이 이미지를 1999년 TV 미니시리즈의 한 장면으로 잘못 설명하는 등, AI의 답변만으로는 이미지의 진위를 판별하기 어려운 상황이다.

그 대안으로 주목받아 온 이미지 워터마킹은 워터마크를 삽입하여 출처를 확인하는 방식으로, 이미지 전체에 워터마크를 삽입하는 글로벌 워터마킹(global watermarking)이 주로 쓰였다. 그러나 생성형 AI로 객체를 추가하거나 지우는 일이 쉬워지면서, 여러 출처가 합성된 이미지에서 객체 단위의 출처 확인이 필요해졌다. 이에 특정 영역에만 워터마크를 삽입해, 이미지가 잘리거나 합성되어도 해당 영역에서 워터마크를 검출하는 로컬 워터마킹(local watermarking)이 등장했다. 하지만 기존 연구는 자체적으로 제안한 기법만 서로 상이한 조건으로 평가해, 압축, 회전 등 이미지 변형 이후에도 워터마크가 유지되는 강건성을 비교하기 어렵다.

이러한 배경에서 로컬 워터마킹 기술의 강건성을 동일한 조건에서 평가하는 벤치마크가 제시되었다. 평가 대상은 로컬 워터마킹 기술인 MaskWM, WAM(Watermark Anything Model)과, 워터마크 삽입 범위를 특정 영역으로 제한한 글로벌 워터마킹 기술 3종이다. 4개 데이터셋의 이미지에 55개 이미지 변형을 적용해 워터마크에 담긴 메시지인 페이로드의 복원, 워터마크 영역의 위치 추정, 비가시성을 비교하며, 이미지 일부를 AI로 다시 생성하는 인페인팅 등 생성형 편집도 포함한다. 이를 통해 기술별로 취약한 이미지 변형을 확인하고, 실제 배포를 위한 과제를 제시한다.

[사례] 55개 변형으로 로컬 워터마킹의 취약점을 평가한 강건성 벤치마크

• 기존 강건성 평가의 한계와 벤치마크의 등장 배경

- 생성형 AI로 편집되거나 여러 출처가 합성된 이미지가 늘면서, 이미지 전체에 워터마크를 삽입하는 글로벌 워터마킹과 달리 특정 영역에만 워터마크를 삽입하고 페이로드(payload)* 복원과 워터마크 영역의 위치 추정을 함께 수행하는 로컬 워터마킹이 등장함
- 기존의 강건성 평가는 대부분 각 기술을 제안한 연구진이 자체적으로 수행해 왔음. WAM(Watermark Anything Model) 연구는 로컬 워터마킹 이미지에 이미지 변형을 적용하지 않았고, MaskWM 연구는 이미지 변형 13종만 평가해 인페인팅(inpainting)** 등 생성형 편집을 다루지 않았음
- 본 벤치마크는 로컬 워터마킹의 강건성을 체계적으로 평가한 벤치마크로, 특정 기술에 맞추지 않은 비적응형(non-adaptive)*** 이미지 변형 55개를 4개 데이터셋의 이미지에 적용해 워터마킹 기술 5종을 비교함

* 페이로드(payload): 워터마크에 담기는 메시지로, 0과 1로 이루어진 비트열. 검출 시 이 비트열을 정확히 읽어내는 것을 페이로드 복원이라 함

** 인페인팅(inpainting): 이미지의 일부 영역을 주변 내용에 맞춰 AI로 다시 생성하는 편집 기술

*** 비적응형(non-adaptive): 특정 워터마킹 기술에 맞춰 설정값을 조정하지 않고 모든 기술에 동일하게 적용하는 방식

• 평가 대상 기술과 평가 지표

- 평가 대상은 로컬 워터마킹 기술인 MaskWM, WAM과, 글로벌 워터마킹 기술인 옴니가드(OmniGuard), 트러스트마크(TrustMark), 픽셀실(PixelSeal)의 워터마크 삽입 범위를 특정 영역으로 제한한 3종임
- 페이로드 복원은 두 가지 지표로 측정하며, 올바르게 읽어낸 비트의 비율인 비트 정확도(bit accuracy)와 페이로드 길이를 반영한 증거 점수(evidence score)*를 사용함
- 증거 점수를 함께 사용하는 이유는 기술별 페이로드 길이가 32~256비트로 달라, 비트 정확도만으로는 기술 간 공정한 비교가 어렵기 때문임
- 위치 추정은 예측한 워터마크 영역과 실제 워터마크 영역의 일치도인 IoU(Intersection over Union)**로 측정하며, 픽셀 단위로 워터마크 영역을 출력하는 MaskWM과 WAM만 비교 대상임
- 비가시성은 이미지 변형을 적용하지 않은 상태에서 원본 이미지와 워터마크 삽입 이미지의 화질 차이를 PSNR, SSIM, LPIPS*** 3가지 지표로 비교해 측정함

* 증거 점수(evidence score): 무작위로 비트를 추측했을 때 같은 수 이상의 비트를 맞힐 확률을 로그 척도로 나타낸 값. 값이 클수록 페이로드 복원의 근거가 강하며, 3 이상이면 무작위 추측으로 나올 확률이 0.1% 이하임

** IoU(Intersection over Union): 두 영역의 교집합 넓이를 합집합 넓이로 나눈 값으로, 1에 가까울수록 예측 영역이 실제 영역과 일치함

*** PSNR(Peak Signal-to-Noise Ratio), SSIM(Structural Similarity Index Measure), LPIPS(Learned Perceptual Image Patch Similarity): 각각 화소 단위 왜곡, 구조적 유사도, 딥러닝 특징 기반의 지각적 유사도를 측정하는 화질 지표. PSNR과 SSIM은 높을수록, LPIPS는 낮을수록 원본에 가까움

• 55개 이미지 변형의 4가지 분류 체계

- 신호 왜곡은 화소 위치는 유지한 채 밝기 조정, 흐림 처리, 노이즈 추가, JPEG 압축 등으로 화소값이나 주파수 성분을 바꾸는 이미지 변형임
- 정렬 변화는 잘라내기, 회전, 뒤집기 등으로 화소의 위치를 바꿔 워터마크 검출기가 예상하는 위치와 실제 워터마크 위치를 어긋나게 하는 이미지 변형으로, 이미지 바깥 영역을 새로 생성하는 확장 캔버스 아웃페인팅(outpainting)도 포함함

- 정렬 변화의 영향을 확인하기 위해, 이미지 전체에 동기화 신호를 삽입해 바뀐 화소 위치를 추정하고 검출 전에 원래대로 되돌리는 보조 기술인 싱크실(SyncSeal)의 적용 여부를 나누어 평가함
- 간접 로컬 편집은 워터마크 영역과 무관하게 정한 편집 영역에 블랙아웃(blackout)*, 스플라이싱(splicing)**, 인페인팅을 적용하는 이미지 변형임. 인페인팅은 편집 영역 밖 화소도 바꿀 수 있어, 생성 후 편집 영역 밖을 원래 화소로 되돌린 대조 실험도 수행함
- 직접 워터마크 편집은 워터마크 영역 자체를 편집 영역으로 삼아 인페인팅하거나 스플라이싱하는 이미지 변형으로, 페이로드를 담은 화소가 교체되거나 재생성된 뒤에도 워터마크가 남는지 확인함

* 블랙아웃(blackout): 이미지의 일정 영역을 검은색 화소로 채우는 이미지 변형

** 스플라이싱(splicing): 이미지의 일정 영역을 같은 데이터셋의 다른 이미지 내용으로 교체하는 이미지 변형

[표 1] 5개 워터마킹 기술별 싱크실 적용 여부에 따른 화질 비교

Method	PSNR (dB)		SSIM		LPIPS	
	SyncSeal off	SyncSeal on	SyncSeal off	SyncSeal on	SyncSeal off	SyncSeal on
MaskWM	49.481 $\pm$ 1.082	42.309 $\pm$ 0.617	0.998 $\pm$ 0.001	0.990 $\pm$ 0.004	0.006 $\pm$ 0.001	0.009 $\pm$ 0.002
WAM	62.549 $\pm$ 1.185	43.562 $\pm$ 0.824	1.000 $\pm$ 0.000	0.992 $\pm$ 0.003	0.000 $\pm$ 0.000	0.003 $\pm$ 0.001
PixelSeal	56.800 $\pm$ 1.243	43.353 $\pm$ 0.811	0.999 $\pm$ 0.000	0.992 $\pm$ 0.003	0.000 $\pm$ 0.000	0.003 $\pm$ 0.001
TrustMark	54.092 $\pm$ 2.075	42.978 $\pm$ 0.857	0.999 $\pm$ 0.001	0.991 $\pm$ 0.003	0.000 $\pm$ 0.000	0.003 $\pm$ 0.000
OmniGuard	51.361 $\pm$ 0.837	42.730 $\pm$ 0.841	0.998 $\pm$ 0.001	0.991 $\pm$ 0.003	0.000 $\pm$ 0.000	0.003 $\pm$ 0.001

출처: Kai Yao 외 6인, "What Breaks Local Watermarks? A Robustness Benchmark for Local Invisible Image Watermarking", arXiv, 2026.09.15., <https://arxiv.org/pdf/2609.16832>

• 이미지 변형 범주별 성능 평가 결과

- 이미지 변형이 없는 조건에서 MaskWM은 페이로드 복원과 위치 추정 모두 5종 중 가장 우수했으나, 워터마크 삽입에 따른 화질 저하도 가장 큰 것으로 확인됨
- 신호 왜곡에서는 MaskWM만 모든 세부 유형에서 증거 점수 기준값을 넘었으며, 픽셀실은 노이즈 추가를 제외한 유형에서만 기준값을 넘었고 나머지 3종은 모든 유형에서 기준값에 미치지 못함
- 정렬 변화에서 MaskWM은 상하 뒤집기, 대각선 뒤집기, 확장 캔버스 아웃페인팅에서 페이로드를 거의 복원하지 못했고, 싱크실은 대각선 뒤집기, 회전 등에서 복원을 크게 높였으나 상하 뒤집기에는 효과가 없고 모든 기술의 화질을 낮춤
- 간접 로컬 편집의 인페인팅은 워터마크 영역과 겹치지 않아도 모든 기술을 기준값 아래로 떨어뜨림. 편집 영역 밖 화소를 원래대로 되돌리면 페이로드 복원 성능이 회복되어, 이미지 전체가 재생성되는 과정이 원인인 것으로 확인됨
- 직접 워터마크 편집에서는 모든 기술이 무작위 추측 수준에 그쳤으나, MaskWM과 WAM의 IoU는 상당 수준 유지되어 위치 추정은 페이로드 복원과 별개로 작동함

결론 및 시사점

• 기술적 한계 및 향후 과제

- 현재 로컬 워터마킹 기술은 이미지 변형이 없는 조건에서도 워터마크 영역이 작을수록 성능이 낮아, 영역 크기에 따라 페이로드 길이를 조절하거나 여러 해상도에 걸쳐 워터마크를 삽입하는 방식의 개선이 필요함
- 페이로드 복원 성능과 화질 사이에 상충 관계가 있어 전문 사진이나 의료 영상처럼 화질이 중요한 분야에는 적용이 어려우며, 기술별로 워터마크 강도를 조절해 적정 수준을 찾는 연구가 필요함
- 인페인팅 등 생성형 편집은 워터마크 영역을 겨냥하지 않아도 워터마크를 무력화할 수 있어, 향후 벤치마크와 로컬 워터마킹 기술 모두 생성형 편집에 대한 강건성을 반영해야 함
- 비적응형 이미지 변형만으로도 워터마크 제거가 가능했으므로, 기술 구조나 검출 결과를 활용하는 공격자는 원리상 더 효과적으로 워터마크를 제거할 수 있어 이에 대한 평가가 필요함. 평가 범위 안에서 현재 로컬 워터마킹 기술은 실제 배포에 충분한 수준에 이르지 못함

참고문헌

- Kai Yao 외 6인, "What Breaks Local Watermarks? A Robustness Benchmark for Local Invisible Image Watermarking", arXiv, 2026.09.15., <https://arxiv.org/pdf/2609.16832>
- Matt Novak, "Never Forget, AI Slop Is Polluting 9/11 History on the Internet as We Speak", Gizmodo, 2026.09.11., <https://gizmodo.com/never-forget-ai-slop-is-polluting-9-11-history-on-the-internet-as-we-speak-2000810667>