www.copyright.or.kr

저작권 이슈 브리프

COPYRIGHT ISSUE BRIEF

Weekly Report
2026. 7-3

한국저작권위원회
KOREA COPYRIGHT COMMISSION

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 모듈레이트 AI 음악 탐지 API로 본 스트리밍 음악 검증 체계의 발전

▶ AI 음악 생성 도구의 성능이 빠르게 향상되면서, AI가 만든 음악이 창작자와 청취자의 인지 없이도 스트리밍 카탈로그에 대량으로 유입되고 있다. 자율 신고와 메타데이터 표시에 의존해 온 기존 검증 방식은 이러한 흐름을 따라가지 못하며, 저작권료가 부정하게 취득되는 부작용도 나타나고 있다. 이에 한 음성 AI 기업은 보컬과 악기 등 음원 구성 요소를 독립적으로 판별하는 여러 모델을 결합한 탐지 기술을 선보였다. 이 기술은 구간 단위로 세밀하게 채점해 혼합 제작 트랙도 정확히 판정하며, 라이선싱 검증과 업로드 심사 등 실무에도 결합될 수 있다. 업계 주요 주체들의 관심도 확인되고 있으나, 일부 복잡한 음원 구성에서는 여전히 식별의 한계가 남아 있어 탐지 결과가 처리 판단을 완전히 대체하지는 못한다.

산업 클라우드플레어-비하이브 협력으로 본 AI 크롤러 대응 방식의 변화

▶ AI 검색과 챗봇의 확산으로 원문 사이트 유입이 줄고 콘텐츠 무단 수집 우려가 커지면서, AI 크롤러에 대응할 필요성이 높아지고 있다. 이에 따라 클라우드플레어와 비하이브는 협력을 통해 발행 대시보드에 AI 크롤러 제어 기능을 내장해, 발행자가 별도의 코드 작업 없이 크롤러 접근을 허용하거나 차단할 수 있도록 지원한다. 특히 어떤 크롤러가 접근을 시도했고 그중 무엇이 차단되었는지, 각 AI 서비스가 원문 링크를 통해 방문자를 얼마나 유입시켰는지 확인할 수 있어, 크롤러별 허용 여부를 판단할 근거를 제공한다. 이를 바탕으로 발행자는 AI를 통한 노출 확대와 스크래핑 차단을 통한 콘텐츠 보호 가운데 원하는 대응 방향을 선택할 수 있다. 이번 사례는 AI 크롤러 대응이 콘텐츠 공개 시점부터 접근 조건과 보상 구조를 미리 설계하는 방식으로 이동할 수 있음을 보여준다.

산업 윈도우 사진 워터마크로 본 기본 앱 차원의 AI 활용 여부 표시 확산

▶ AI 이미지 생성 및 편집 기능이 전문 생성 도구를 넘어 기본 앱으로 확산되면서, AI 활용 여부를 표시하는 기능도 일상적 앱 환경으로 확대될 필요가 커지고 있다. 이에 따라 마이크로소프트는 윈도우 사진에서 AI로 생성하거나 편집한 이미지에 시각적 워터마크를 삽입하고, C2PA 기반 메타데이터에 AI 모델과 생성 시점 등 출처 정보를 기록하는 기능을 도입했다. 이는 기본 앱 차원에서 AI 출처 표시를 제공하려는 초기 사례로 볼 수 있다. 다만 시각적 워터마크는 이용자가 직접 선택해야 적용되고 편집 과정에서 쉽게 지워질 수 있으므로, 그 자체만으로는 한계가 있다. 따라서 향후에는 이미지뿐 아니라 영상, 음성까지 메타데이터 기반 출처 표시를 확대하고, 이를 인식할 수 있는 앱 및 플랫폼과 공통 표준의 업계 채택을 병행하는 노력이 병행되어야 한다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 온디바이스 딥페이크 탐지 기술의 확산과 AI 합성 영상 및 음성의 진위 검증

▶ 생성형 AI와 음성 복제 도구의 확산으로 얼굴과 음성을 합성한 콘텐츠의 제작 장벽이 낮아지면서, 딥페이크를 이용한 신원 사칭과 사기 시도가 화상회의, 영상 시청 및 모바일 등 다양한 접점으로 번지고 있다. 이에 대응해 스캠에이아이와 퀄컴의 헤일로를 비롯한 온디바이스, 실시간 탐지 기술이 상용화되면서, 영상 데이터를 외부 서버로 보내지 않고 기기 안에서 곧바로 진위를 확인하는 방식이 확산되고 있다. 이러한 흐름은 진위 검증의 무게 중심을 사후 과정, 이용자 기기 및 실시간 이용 환경으로 옮기며, 얼굴과 음성 등 개인 고유의 특징 보호와 이에 맞춘 제도 설계 과제를 함께 제기한다.

산업 AI 기반 불법 스트리밍 확산과 콘텐츠 보안 대응 체계의 변화

▶ AI를 활용한 불법 스트리밍이 자동화 및 대규모화되면서 기존의 개별 보안 도구 중심 대응만으로는 실시간 침해 확산에 효과적으로 대응하기 어려워지고 있다. 이에 콘텐츠 보안 업계는 분산된 탐지 데이터를 통합 분석하고 피해 규모와 시간 민감도를 기준으로 대응 우선순위를 정하는 지능형 대응 체계로 무게 중심을 옮기고 있다. 스위스 콘텐츠 보안 기업 나그라비전은 포렌식 워터마킹과 멀티 DRM, 조사, 차단 대응을 하나의 운영 체계로 연계한 지능형 보안 플랫폼 나그라 벤투리를 출시하며 이러한 접근을 선도적으로 적용하고 있다. 향후 콘텐츠 보안 대응은 개별 기술의 성능 경쟁을 넘어, 데이터 기반의 통합 대응 구조와 사업 성과 중심의 효과 측정으로 확산될 가능성이 제기된다.

기술 주간 기술 동향

▶ 생성형 AI 기술의 확산으로 음성과 음악 복제, 저작권 침해 우려가 커지면서 오디오 워터마킹이 대응 수단으로 주목받고 있다. 방어자는 워터마크 디코더가 출력하는 메시지 확률이 정상 오디오에서 일정한 통계적 분포를 따른다는 점을 이용해 조작 여부를 판별해왔다. 그러나 본 보고서에서 소개하는 적응형 오디오 워터마크 공격 기술인 AWM은 이러한 방어 체계를 우회한다. AWM은 섭동을 반복 갱신해 공격을 성공시키고 탐지를 피한 뒤, 신호와 스펙트로그램 손실로 오디오 품질을 개선하는 두 단계로 작동한다. 워터마크 교체, 생성, 제거 세 가지 유형 모두에 적용된 실험에서 탐지율이 대폭 낮아졌으며, 제거 공격에서는 탐지 사례가 전혀 없었다. 이는 기존 탐지 방식의 취약성을 보여주며, 더 정교한 방어 체계의 필요성을 시사한다.

모듈레이트 AI 음악 탐지 API로 본 스트리밍 음악 검증 체계의 발전

AI 생성 음악 확산에 따른 스트리밍 음악 검증 공백 확대

• AI 음악 생성 도구의 성능 향상과 스트리밍 카탈로그 유입 확산

- 생성형 AI 음악 도구인 수노(Suno)는 2026년 6월 3일 4억 달러(원화 약 6,194억 원)[1] 규모의 투자를 유치하며 기업가치 54억 달러(원화 약 8조 3,614억 원)를 기록해, 시장의 큰 관심을 받음[2]
- 수노의 사례 이외에도 생성형 AI 음악 도구에 대규모 자금이 계속 유입되면서 창작자와 청취자의 사전 인지 없이 AI 생성 음악이 음원 목록에 대량으로 편입되는 현상이 점차 뚜렷해지고 있음
- 한편, 보컬만 AI로 제작하고 악기 반주는 인간이 제작하거나 그 반대인 혼합 제작 방식이 확산되면서, 트랙 전체를 AI 산출물인지 아닌지로만 구분하기 어려운 상황이 나타나고 있음

• 자율 신고 중심 검증 방식의 한계와 수익 배분 왜곡 우려

- 지금까지 AI 음악 투명성 확보는 자율 신고, 메타데이터 표시, 워터마킹 등 자발적 방식에 주로 의존해 왔음. 그러나 이러한 방식은 콘텐츠가 출처 단계에서 정확히 표시되어야만 유효하다는 한계를 지님
- 미국의 음성 AI 기업 모듈레이트(Modulate)의 CEO 마이크 파파스(Mike Pappas)는 "공개도 중요하지만, 그것만으로는 부족하다. 업로더가 스스로 밝힌 정보에만 의존하는 플랫폼은 밀려드는 AI 생성 콘텐츠의 규모를 제대로 관리할 수 없다“ 라고 지적함 [3]
- AI 생성 트랙이 저작권료를 부정하게 취득할 목적으로 대량 업로드되는 사례도 발생하며, 이는 정당한 인간 아티스트에게 돌아갈 수익을 희석시키는 요인으로 지목되고 있어 업계의 우려가 커지고 있음
- 스포티파이(Spotify), 애플 뮤직(Apple Music), 유튜브 뮤직(YouTube Mudic) 등 주요 음악 스트리밍 서비스는 AI 콘텐츠 공개 의무화 또는 수익화 제한 정책을 이미 운영하고 있으나,[4] 업로드되는 콘텐츠의 양을 고려할 때 인간의 심사만으로는 실효적으로 집행하기 어려운 것이 현실임

1) 1달러=1,548.40원(KEB 하나은행 매매기준율 적용, 2026.07.01.)
2) Duncan Riley, “Modulate launches AI music detection as synthetic tracks flood streaming”, Silicon Angle, 2026.06.24., https://siliconangle.com/2026/06/24/modulate-launches-ai-music-detection-synthetic-tracks-flood-streaming/
3) Duncan Riley, “Modulate launches AI music detection as synthetic tracks flood streaming”, Silicon Angle, 2026.06.24., https://siliconangle.com/2026/06/24/modulate-launches-ai-music-detection-synthetic-tracks-flood-streaming/
4) Modulate, “AI Music Detection: Detect AI generated vocal and instrumental music”, 2026.07.02. 접속 기준, https://www.modulate.ai/api/ai-music-detection

오디오 신호 분석 기반 탐지 모델의 세분화된 검증 구조 정착

• 프레임·보컬·악기 판별을 위한 3중 모델 앙상블 구조

- 모듈레이트는 AI 음악 탐지 API(Application Programming Interface)*에서 프레임 분류, 보컬 판별, 악기 판별을 담당하는 세 개의 모델을 함께 운용함
- 프레임 분류 모델이 4초 단위 구간마다 음악인지 음성인지 또는 둘 다인지를 분류하면, 보컬 판별 모델과 악기 판별 모델이 보컬과 악기 성분의 AI 생성 여부를 독립 판정하며, 이 3중 모델 구조는 모듈레이트의 오디오 지능 플랫폼 벨마(Velma) 위에 앙상블 형태로 구축됨
- 이러한 개별 분석 체계 덕분에 AI가 합성한 보컬에 인간이 연주한 반주가 붙는 트랙에서, 전체를 AI 산출물로 간주하지 않고 성분별로 나누어 4초 구간 단위까지 세밀하게 판정할 수 있음[5)]
- 이는 하나의 트랙 전체에 단일 점수만 부여하던 기존 방식과 달리, 성분 단위로 세분화된 판정을 가능하게 하는 구조적 차별점으로 평가되며, 하이브리드 트랙 처리에서 강점을 보이는 요인으로 작용함

* API(Application Programming Interface): 서로 다른 소프트웨어나 서비스가 데이터를 주고받을 수 있도록 정해진 규칙과 연결 방식으로, 개발자가 이를 이용해 자신의 서비스에 모듈레이트의 탐지 기능을 직접 연동할 수 있음

• 구간 단위 채점 방식이 만드는 세분화된 탐지 정밀도

- 모듈레이트는 오토튠(Auto-Tune)이나 압축 등 흔한 디지털 보정을 AI 생성으로 오탐하는 기존 방식과 달리, 실제 AI 생성 음원만 정확히 걸러내도록 학습되었다고 밝힘
- 판정 기준이 되는 임계값은 고정되지 않고 조정 가능하며, 고정된 기준만 제공하거나 아예 지원하지 않는 업계 표준과 달리 재학습 없이 기준을 바꿀 수 있음
- 파일을 통째로 올려야만 결과를 알려주는 기존 방식과 달리, 오디오가 실시간으로 입력되는 환경에서도 연결을 유지한 채 곧바로 결과를 받을 수 있으며, 4초 구간마다 보컬 판정 결과가 나와 업로드 시점부터 바로 정책에 반영할 수 있음
- 주요 음악 생성 모델을 대상으로 진행한 내부 테스트에서는 76개 장르에 걸쳐 약 95%의 정밀도를 기록했으며, 이는 AI 모델별 특성을 암기하기보다 AI 생성 패턴 자체를 폭넓게 학습한 결과로 설명됨[6)]

• 라이선싱 검증과 업로드 심사에 연동되는 활용 방식

- 라이선싱 업체와 싱크 에이전시(sync agency)*는 콘텐츠에 AI 산출물이 포함되어 있는지 여부를 계약을 체결하기 전에 미리 확인하는 용도로 이 API를 활용할 수 있음
- 디지털 유통사도 업로드 단계에서부터 AI 산출물 여부를 선별하는 심사 도구로 이 API를 도입할 수 있는 대상으로 거론되며, 이는 대량 업로드를 다루는 유통 구조에 곧바로 결합될 수 있는 형태로 설계됨
- 음악 퍼블리셔와 저작권 관리 단체 역시 특정 트랙이 로열티를 받을 자격이 있는 콘텐츠인지 판단하는 과정에서 이 API를 참고 자료로 활용할 수 있음

* 싱크 에이전시(sync agency): 영화·광고·게임 등 영상물에 음악을 삽입할 수 있도록 사용권을 중개하는 대행사

5) Modulate, “AI Music Detection: Detect AI generated vocal and instrumental music”, 2026.07.02. 접속 기준, https://www.modulate.ai/api/ai-music-detection
6) Duncan Riley, “Modulate launches AI music detection as synthetic tracks flood streaming”, Silicon Angle, 2026.06.24., https://siliconangle.com/2026/06/24/modulate-launches-ai-music-detection-synthetic-tracks-flood-streaming/

탐지 인프라의 산업 확산 가능성과 기술적 한계

• 탐지 인프라의 산업 확산과 응용 가능성

- 모듈레이트는 주요 음반사와 유통 플랫폼으로부터 이미 초기 관심을 받고 있다고 밝히며, 이는 탐지 기술이 스트리밍 업계 전반으로 확산될 가능성을 보여줌
- 또한, 모듈레이트의 API는 브라우저 확장 프로그램이나 인증 배지처럼, 청취자가 직접 트랙의 AI 산출물 여부를 확인할 수 있는 소비자 대상 도구를 만드는 데도 활용될 수 있음
- 스트리밍 플랫폼과 유통사뿐 아니라 저작권 관리 단체, 라이선싱 업체, 음반사 등 여러 업계 주체가 해당 API의 잠재 활용 대상으로 거론되고 있어, 적용 범위가 넓게 형성되어 있음

• 탐지 기술과 실제 판단 사이에 남는 기술적 한계

- 합창처럼 여러 목소리가 섞인 경우, 또는 인간 보컬 아래 AI 반주가 깔린 경우는 현재 기술로도 정확한 식별이 어려우며, 생성 기법이 계속 발전하고 있어 탐지 모델도 지속적인 갱신이 필요함
- 탐지 결과는 트랙에 AI 산출물이 포함되어 있을 가능성을 알려주는 신호일 뿐이며, 이를 바탕으로 콘텐츠를 어떻게 처리할지 정하는 과정에는 별도의 판단이 여전히 필요함

참고문헌

- Duncan Riley, “Modulate launches AI music detection as synthetic tracks flood streaming”, Silicon Angle, 2026.06.24., https://siliconangle.com/2026/06/24/modulate-launches-ai-music-detection-synthetic-tracks-flood-streaming/
- Kristin Canders, “Modulate Launches AI Music Detection API to Help Platforms Verify AI-Generated Music at Scale”, AccessNewsWire, 2026.06.24., https://www.accessnewswire.com/newsroom/en/computers-technology-and-internet/modulate-launches-ai-music-detection-api-to-help-platforms-verify-1181688
- Modulate, “AI Music Detection: Detect AI generated vocal and instrumental music”, 2026.07.02. 접속 기준, https://www.modulate.ai/api/ai-music-detection

클라우드플레어-비하이브 협력으로 본 AI 크롤러 대응 방식의 변화

소규모 발행자의 AI 크롤러 대응 한계와 플랫폼 기반 제어 수단의 등장

• AI 크롤러 확산에 따른 기존 대응 방식의 한계

- 최근 온라인에서 뉴스나 정보를 검색하는 이용자들이 검색 결과를 클릭해 원문을 읽기보다는 생성형 AI 챗봇이나 구글(Google LLC)의 AI 검색 등이 제공하는 요약 답변을 먼저 확인하는 경향이 확산되면서, 원문 사이트 방문이 줄어듦
- 이 과정에서 AI 서비스가 답변 생성이나 모델 학습을 위해 기사, 뉴스레터, 블로그 등 웹 콘텐츠를 수집하는 반면, 콘텐츠 제공자에게는 콘텐츠 이용에 상응하는 이용자 유입이나 광고 수익이 충분히 돌아가지 않는다는 우려가 제기됨
- 이에 콘텐츠 제공자들은 AI 크롤러(crawler)*의 접근을 직접 차단하는 방식을 사용하고 있으나, 기존 차단 방식은 robots.txt** 파일 수정이나 방화벽 규칙 설정 등 일정 수준 이상의 기술 지식을 요구해 전문 인력이 없는 경우 활용하기 어려웠음
- 또한 기존 차단 방식은 강제력이 없기 때문에, 크롤러가 해당 규칙을 따를 때만 효력이 발생하므로 실제 접근 차단에는 한계가 있음

* 크롤러(crawler): 웹페이지를 자동으로 순회하며 정보를 수집하는 프로그램, 봇(bot) 또는 스파이더(spider)로도 불림

** robots.txt: 웹사이트 최상위 경로에 두는 텍스트 파일로, '로봇 배제 표준(Robots Exclusion Protocol)'에 따라 크롤러에 수집을 허용하거나 거부할 대상을 안내하는 관행적 규약

• 클라우드플레어-비하이브 협력을 통한 발행 플랫폼 내 제어 기능 도입

- 이러한 한계를 보완하기 위해, 2026년 6월 23일 웹 보안 및 인프라 기업 클라우드플레어(Cloudflare, Inc.)와 뉴스레터 발행 플랫폼 비하이브(beehiiv, Inc.)가 협력을 발표하고 발행 플랫폼에 크롤러 제어 기능을 내장하는 방식을 제시함[1)]
- 이번 제휴는 클라우드플레어가 자사의 AI 크롤링 제어 기술을 제공하고, 비하이브가 이를 발행 대시보드*에 통합하여 발행자가 별도의 외부 도구 없이 플랫폼 화면에서 크롤러 접근을 관리하도록 함
- 클라우드플레어가 전 세계 웹 트래픽의 약 20%를 처리하고[2)] 비하이브에서는 13만 5,000명 이상의 소규모 발행자가 활동하는 만큼[3)], 이번 협력은 대형 언론사 중심이던 크롤러 대응 수단이 소규모 발행자에게 확장되는 계기로 작용할 것으로 기대됨

* 대시보드(dashboard): 서비스의 주요 정보와 설정을 한 화면에 모아 보여주고 조작할 수 있게 한 관리 화면

1) Cloudflare, "Cloudflare and beehiiv Introduce AI Crawl Controls to Help Independent Publishers Navigate the AI Era", 2026.06.23., https://www.cloudflare.com/press/press-releases/2026/cloudflare-and-beehiiv-introduce-ai-crawl-controls-to-help-independent-publishers-navigate-the-ai-era/
2) Andrew Deck, "Beehiiv's new Cloudflare partnership gives indie journalists a new level of control over AI crawlers", Nieman Lab, 2026.06.25., https://www.niemanlab.org/2026/06/beehiivs-new-cloudflare-partnership-gives-indie-journalists-a-new-level-of-control-over-ai-crawlers/
3) Duncan Riley, "Beehiiv adds Cloudflare AI Crawl Control so writers can block or allow bots", SiliconANGLE, 2026.06.23., https://siliconangle.com/2026/06/23/beehiiv-adds-cloudflare-ai-crawl-control-writers-can-block-allow-bots/

대시보드 내장형 크롤링 제어 기능의 접근성 확대

• 코드 작업 없는 크롤러 접근 관리로 기술적 부담 완화

- robots.txt 파일 수정이나 방화벽 규칙 설정이 필요했던 기존 방식과 달리, 별도의 코드 작업 없이 발행 화면 안에서 접근 권한을 관리할 수 있게 되어 기술적 부담이 낮아짐
- 특히 크롤러의 자발적 준수에 의존하던 기존 방식과 달리, 클라우드플레어는 접근 요청 단계에서 크롤러를 직접 차단할 수 있어 실효성이 확보됨

• 크롤러 활동 확인을 통한 허용 및 차단의 판단 근거 제공

- 발행자는 이번 기능을 통해 어떤 AI 크롤러가 콘텐츠 접근을 시도했고 그중 무엇이 차단되었는지, 각 AI 서비스가 원문 링크를 통해 어느 정도의 방문자를 유입시켰는지에 대한 정보를 확인할 수 있음
- 기존 방식은 실제 접근 주체나 유입 효과를 확인하기 어려워 크롤러를 일괄 허용 또는 일괄 차단하는 양자택일 방식에 가까웠으나, 이번 협력을 통해 비하이브 대시보드 화면에서 곧바로 크롤러별 접근 이력과 유입 효과를 함께 확인할 수 있음
- 발행자는 이를 바탕으로 독자 유입에 기여하는 크롤러는 허용하고, 무단 수집 우려가 크거나 유입 효과가 낮은 크롤러는 차단하는 방식으로 접근 기준을 조정할 수 있음

• AI 노출 확대와 콘텐츠 보호 사이의 선택적 대응 가능

- 이와 별도로 발행자는 자신의 콘텐츠에 대해 AI 검색 및 에이전트의 노출을 폭넓게 허용해 배포를 확대할지, 아니면 스크래핑을 차단해 콘텐츠를 보호할지를 목적에 따라 직접 선택할 수 있음
- 이는 AI를 통해 노출 확대와 독자 유입을 기대하는 발행자와, 콘텐츠 가치가 장기적으로 훼손되는 것을 우려해 보호를 우선하는 발행자가 공존하는 시장의 서로 다른 이해관계를 반영한 조치임
- 이 중 '보호'를 선택하면 AI가 콘텐츠를 무단으로 수집하거나 학습 데이터로 활용할 가능성을 줄이고, 향후 해당 콘텐츠를 라이선스 협상이나 유료화의 대상 자산으로 관리하는 전략과도 연결될 수 있음

AI 크롤러 접근 제어의 사전 설계 방식 확산

• 소규모 발행자의 AI 크롤러 대응 여건 개선

- 이번 사례는 AI 크롤러 대응이 대형 언론사 중심의 기술 대응에서 벗어나, 소규모 발행자도 플랫폼 안에서 직접 접근을 제어할 수 있는 방식으로 확대되고 있음을 보여줌
- 비하이브가 제어 기능의 기술 장벽을 낮추고 크롤러 접근 정보까지 제공하면서, 발행자는 간편하게 크롤러 활동을 확인하고 노출과 보호 중 무엇을 우선할지 선택할 수 있는 제어 수단을 확보하게 됨
- 다만 이러한 기능은 비하이브와 클라우드플레어 같은 사업자의 지원을 전제로 작동하므로, 이를 지원하지 않는 플랫폼의 발행자에게까지 개선 효과가 균등하게 미치기는 어렵다는 한계가 있음

• **콘텐츠 접근 권한을 사전에 설계하는 구조로의 이동**

- AI의 무단 이용을 둘러싼 우려에 사후적으로 대응하던 방식에서 나아가, 콘텐츠 공개 시점에 어떤 크롤러의 접근을 허용할지 미리 정하는 방식으로 제어 기준이 이동할 수 있음을 시사함
- 특히 노출과 보호 중 하나를 선택할 수 있다는 점은 콘텐츠 접근을 일률적으로 차단할 대상으로 보기보다, 발행자가 조건에 따라 허용 여부를 정할 수 있는 권한으로 보는 관점과 맞닿아 있음
- 이에 따라 향후에는 AI 크롤러의 접근을 허용하는 대가로 트래픽, 출처 표시, 라이선스 비용 등 어떤 보상 조건을 마련할지에 대한 논의로 이어질 수 있음
- 이러한 방식이 뉴스레터를 넘어 콘텐츠 산업 전반으로 확대될 경우, 콘텐츠 공개에 앞서 크롤러의 접근 조건을 설정하는 절차가 콘텐츠 관리의 기본 요건으로 자리 잡을 것으로 전망됨

참고문헌

- Cloudflare, “Cloudflare and beehiiv Introduce AI Crawl Controls to Help Independent Publishers Navigate the AI Era”, 2026.06.23., https://www.cloudflare.com/press/press-releases/2026/cloudflare-and-beehiiv-introduce-ai-crawl-controls-to-help-independent-publishers-navigate-the-ai-era/
- Andrew Deck, “Beehiiv's new Cloudflare partnership gives indie journalists a new level of control over AI crawlers”, Nieman Lab, 2026.06.25., https://www.niemanlab.org/2026/06/beehiivs-new-cloudflare-partnership-gives-indie-journalists-a-new-level-of-control-over-ai-crawlers/
- Duncan Riley, “Beehiiv adds Cloudflare AI Crawl Control so writers can block or allow bots”, SiliconANGLE, 2026.06.23., https://siliconangle.com/2026/06/23/beehiiv-adds-cloudflare-ai-crawl-control-writers-can-block-allow-bots/
- Mark Tarre, “Cloudflare & beehiiv add AI crawl controls for publishers”, SecurityBrief, 2026.06.24., https://securitybrief.com.au/story/cloudflare-beehiiv-add-ai-crawl-controls-for-publishers

윈도우 사진 워터마크로 본 기본 앱 차원의 AI 활용 여부 표시 확산

AI 생성 및 편집 기능의 기본 앱 확산에 따른 표시 공백

• AI 생성 및 편집 기능의 확산에 따른 기존 표시 방식의 한계

- 최근 AI를 활용한 이미지 생성 및 편집 기능은 전문 생성 도구에만 머무르지 않고, 이용자가 일상적으로 사용하는 기본 앱에도 표준 기능으로 탑재되고 있음
- 그러나 AI 활용 여부를 표시하는 책임은 주로 전문 AI 서비스 사업자에게 있는 것으로 인식되어 왔으며, AI 활용 여부 표시 기능도 대체로 전문 생성 도구 내에서만 제공됨
- 그 결과 이용자가 기본 앱에서 직접 AI로 생성하거나 편집한 이미지의 경우, 관련 생성 및 편집 이력을 추적할 수 있는 표시 기능이 충분히 마련되지 못함
- 이로 인해 AI 활용 여부 표시가 전문 생성 도구에 한정되지 않고, 이용자가 기본 앱에서 AI로 이미지를 생성하거나 편집하는 경우에도 적용될 필요성이 제기됨

• 윈도우 사진의 AI 활용 여부 표시 기능 추가

- 이러한 공백을 해소하기 위해, 마이크로소프트(Microsoft Corporation)는 2026년 6월 12일 윈도우(Windows) 운영체제의 기본 사진 앱 윈도우 사진(Windows Photos)에서 AI로 생성하거나 편집한 이미지에 시각적 워터마크를 추가하는 기능을 실험 채널*을 통해 처음 도입함[1)]
- 이후 2026년 6월 24일에는 해당 기능을 베타(Beta)와 릴리스 프리뷰(Release Preview)** 채널로 확대해, 정식 배포 전 안정성과 적용 방식을 점검하는 단계로 진입함[1)]
- 이번 기능은 윈도우 기본 앱 7종의 개선 업데이트 내용에 포함된 것으로, 기본 앱 개선 흐름 안에서 AI 활용 여부 표시 기능이 다뤄졌다는 점에서 의미가 있음

* 실험 채널(Experimental): 윈도우 인사이더 프로그램(Windows Insider Program)에서 정식 배포 이전 단계의 기능을 사전에 시험하는 배포 경로
** 릴리스 프리뷰(Release Preview): 정식 배포 직전 단계의 채널로, 실험 채널보다 안정성이 확보된 기능을 최종 점검하는 경로

윈도우 사진 시각적 워터마크의 작동 방식

• 시각적 워터마크와 C2PA 메타데이터의 결합

- 윈도우 사진에서 AI로 이미지를 생성하거나 편집하면, 해당 이미지에는 코파일럿(Copilot) 아이콘이나 'AI로 생성됨(AI-Generated)' 문구가 시각적 워터마크로 삽입됨
- 해당 표시는 이미지에 직접 드러나기 때문에, 파일 속성을 별도로 확인하지 않아도 해당 이미지가 AI로 생성되거나 편집된 콘텐츠임을 직관적으로 확인할 수 있음

1) Microsoft Learn, "Photos release notes", 2026.06.25., https://learn.microsoft.com/en-us/windows-insider/release-notes/apps/photos

[그림1] 코파일럿으로 생성된 이미지 우측 하단에 시각적 워터마크가 삽입된 모습

출처: Microsoft Learn, “Add watermarks to content generated or altered by using AI in Microsoft 365”, 2026.06.12., https://learn.microsoft.com/en-us/microsoft-365/copilot/watermarks

• 이용자 선택에 따른 선택형 표시 방식

- 윈도우 사진의 시각적 워터마크 기능은 이용자가 표시 여부를 직접 선택하는 형태임. 이용자는 이미지를 생성하거나 편집할 때 ‘표시 안 함(Never)’, ‘항상 표시(Always)’, ‘매번 확인(Ask Every Time)’ 세 가지 가운데 하나를 지정할 수 있음
- 다만 이 기능은 기본적으로 비활성화되어 있어, 이용자가 시각적 워터마크 표시를 명시적으로 선택하지 않으면 AI로 생성하거나 편집한 이미지라도 시각적 워터마크 없이 저장됨
- 즉 윈도우 사진의 시각적 워터마크는 자동으로 일괄 적용되는 방식이 아니라, 이용자가 사전에 표시 기능에 동의하고 이를 활성화했을 때 적용되는 선택형 표시 방식에 해당함

기본 앱 환경으로 확장되는 AI 출처 표시와 향후 과제

• 기본 앱과 콘텐츠 유형 전반으로 확대되는 AI 활용 여부 표시

- 이번 기능은 윈도우 사진뿐 아니라 디자이너(Designer), 워드(Word), 파워포인트(PowerPoint) 등 마이크로소프트의 주요 제작 도구로 확대되고 있음. 그 결과 AI 활용 여부 표시는 개별 앱에 추가된 부가 기능을 넘어, 이용자가 일상적으로 사용하는 제작 환경 안에 내장되는 흐름을 보이고 있음[2)]
- 적용 대상도 정지 이미지 중심에서 영상과 음성으로 확대되고 있음. 마이크로소프트 365(Microsoft 365)에서 AI로 생성하거나 편집한 멀티미디어 콘텐츠까지 표시 대상에 포함되면서, AI 활용 여부 표시는 여러 콘텐츠 유형을 포괄하는 관리 체계로 발전하는 양상임
- 이는 AI 생성 콘텐츠의 AI 활용 여부 표시 기능이 AI가 기본 앱에 내장되는 흐름에 맞춰 제작 환경 전반으로 확산되고 있음을 보여줌

2) Microsoft Learn, "Add watermarks to content generated or altered by using AI in Microsoft 365", 2026.06.12., https://learn.microsoft.com/en-us/microsoft-365/copilot/watermarks

• **메타데이터 기반 출처 확인의 실효성 확보 과제**

- 다만 시각적 워터마크는 이용자가 기능을 명시적으로 활성화해야 표시되며, 저장 이후 잘라내기나 추가 편집 과정에서 제거될 수도 있음. 따라서 시각적 워터마크가 보이지 않는다는 사실만으로 해당 콘텐츠가 AI로 생성되거나 편집되지 않았다고 판단하기는 어려움
- 이러한 한계를 보완하는 수단이 파일 메타데이터* 기반의 출처 표시임. 메타데이터는 화면에 드러나는 표시와 별도로 파일 내부에 생성 및 편집 이력을 남기기 때문에, 시각적 워터마크보다 지속적인 출처 추적 수단으로 기능할 수 있음
- 다만 마이크로소프트 365에서 메타데이터 기록은 현재 이미지에만 우선 적용되고, 영상과 음성에는 동일한 방식이 아직 마련되지 않아 콘텐츠 유형별 적용 범위에 차이가 있음
- 또한 메타데이터가 실제 확인 수단으로 작동하려면 이를 인식하고 표시할 수 있는 앱과 플랫폼이 함께 필요하므로, 콘텐츠 유형 확대와 함께 C2PA** 표준과 같은 공통 규격의 업계 채택이 병행되어야 함

* 메타데이터(Metadata): 이미지 파일에 함께 저장되는 부가 정보로, 화면에는 드러나지 않으나 파일 속성 등을 통해 확인할 수 있음

** C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 생성 및 편집 출처와 이력을 표준화된 방식으로 기록하기 위해 구성된 국제 연합체

참고문헌

- Microsoft Learn, “Add watermarks to content generated or altered by using AI in Microsoft 365”, 2026.06.12., https://learn.microsoft.com/en-us/microsoft-365/copilot/watermarks
- Microsoft Learn, “Photos release notes”, 2026.06.25., https://learn.microsoft.com/en-us/windows-insider/release-notes/apps/photos
- Windows Forum, “June 2026 Insider Update: Photos, Paint, Camera, Calculator & 7 More Get App Release Notes”, 2026.06.15., https://windowsforum.com/threads/june-2026-insider-update-photos-paint-camera-calculator-7-more-get-app-release-notes.426398/
- Taras Buria, “Microsoft releases major feature updates for stock Windows 11 apps”, Neowin, 2026.06.12., https://www.neowin.net/news/microsoft-releases-major-feature-updates-for-stock-windows-11-apps/

온디바이스 딥페이크 탐지 기술의 확산과 AI 합성 영상 및 음성의 진위 검증

AI 합성 영상 및 음성의 확산과 진위 검증 수요의 확대

- **딥페이크 사기 시도의 확산**

- 생성형 AI와 음성 복제 도구의 접근성이 높아지면서 실제 인물의 얼굴과 음성을 활용한 합성 콘텐츠의 제작 장벽이 낮아졌으며, 어도비(Adobe) 조사에서는 창작자의 86%가 제작 과정에 생성형 AI를 활용한다고 응답할 만큼 관련 도구의 사용이 확산됨[1)]

- 딥페이크(deepfake)를 활용한 신원 사칭과 사기가 화상 면접, 최고경영진 회의 및 투자 권유 과정에서 확산되는 가운데, 채용 영역에서는 딥페이크 사기 시도가 최근 3년간 2,000% 이상 늘고 인사 책임자의 31%만이 탐지에 대비돼 있다고 응답한 것으로 나타남[2)]

- 미국에서 생성형 AI를 악용한 사기 피해가 2027년까지 최대 400억 달러(원화 약 61조 9,360억 원[3)])에 이를 것이라는 전망이 제시되며, 딥페이크 대응이 기업을 넘어 개인 이용자 보호 과제로 확대됨[4)]

- **음성 합성과 영상 결합에 따른 탐지 난도 상승**

- 최근 사기 사례에서는 AI로 복제 및 합성한 음성을 실제 영상이나 일부만 수정된 영상에 결합하는 방식이 늘면서, 영상 자체는 정상으로 보이더라도 음성이 허위 메시지를 전달하는 형태가 확인됨

- 이러한 경우 시각적 요소만으로는 조작 여부를 가려내기 어려워 기존 보안 방식으로는 허위 정보의 포착이 쉽지 않으며, 오디오와 비디오를 동시에 분석하는 탐지 방식의 필요성이 커짐

실시간 탐지 기술의 상용화와 적용 범위의 확산

- **화상회의 환경에서의 실시간 탐지 및 온디바이스 탐지 상용화**

- 딥페이크 보안 기업 스캠에이아이(Scam.ai)와 글로벌 반도체 기업 퀄컴(Qualcomm)은 2026년 6월 화상회의용 온디바이스(on-device)* 딥페이크 탐지 모델 헤일로(Halo)를 공개함[5)]

1) Georgia Collins, "Exploring Gen's AI Breakthrough for Deepfake Detection", AI Magazine, 2026.01.14., https://aimagazine.com/news/on-device-ai-breakthrough-for-real-time-deepfake-detection
2) Scam.ai, "Scam.ai Announces Qualcomm Partnership, Launches Halo Deepfake Detection Model at Computex 2026", Business Wire, 2026.06.25., https://www.businesswire.com/news/home/20260623132017/en/Scam.ai-Announces-Qualcomm-Partnership-Launches-Halo-Deepfake-Detection-Model-at-Computex-2026
3) 1달러 = 1,548.40원(2026.07.01. KEB 하나은행 매매기준율 적용), 이하 동일 기준 적용
4) Joel R. McConvey, "New deepfake detection product launches reflect expanding market", BiometricUpdate.com, 2026.06.30., https://www.biometricupdate.com/202606/new-deepfake-detection-product-launches-reflect-expanding-market
5) Scam.ai, "Scam.ai Announces Qualcomm Partnership, Launches Halo Deepfake Detection Model at Computex 2026", Business Wire, 2026.06.25., https://www.businesswire.com/news/home/20260623132017/en/Scam.ai-Announces-Qualcomm-Partnership-Launches-Halo-Deepfake-Detection-Model-at-Computex-2026

- 해당 모델은 화상회의 세션 백그라운드에서 합성 또는 AI 생성 영상을 실시간으로 표시하는 방식으로 작동되며, 영상 데이터가 이용자의 컴퓨터를 벗어나 외부 서버로 전송되지 않는 구조로 설계돼 회의 흐름을 바꾸지 않으면서 개인정보 보호와 신속성을 함께 확보할 수 있다는 부분이 강조됨
- 헤일로가 외부 서버 없이 기기 안에서 곧바로 작동할 수 있는 것은 퀄컴이 확대해 온 AI PC와 신경망 처리 장치(Neural Processing Unit, NPU)** 생태계가 뒷받침한 결과로, 이 기반이 기기 내 실시간 분석을 가능하게 함

* 온디바이스(on-device): 데이터를 외부 서버(클라우드)로 보내지 않고 이용자의 기기 안에서 직접 처리하는 방식
** 신경망 처리 장치(Neural Processing Unit, NPU): AI 연산을 빠르고 효율적으로 처리하도록 설계된 전용 반도체

• 콘텐츠 재생 및 소비 단계로의 탐지 시점 확장

- 사이버 보안 기업 젠(Gen)과 반도체 기업 인텔(Intel)은 기기 내부에서 오디오와 비디오의 조작을 동시에 식별하는 기술을 제시했으며[6], 이용자가 콘텐츠를 소비하는 도중 실시간으로 분석해 탐지 시점을 사후에서 이용 시점으로 앞당김
- 이러한 방식은 유명인 사칭 탐지를 넘어 가족이나 지인을 사칭하는 개인화된 피해 대응으로도 확장될 여지가 있는 것으로 나타나, 탐지 대상 범위가 넓어질 가능성이 제기됨

• 기업 보안에서 개인, 모바일 이용 영역으로의 확산

- 미국 및 루마니아의 보안 기업 비트디펜더(Bitdefender)가 출시한 딥페이크 탐지 앱 리얼체크(RealCheck)는 모바일에서 영상 링크나 파일을 분석해 조작 가능성, 기만 의도 및 전사 단서를 제시하는 방식으로, 14개국에서 개인 이용자용 탐지 서비스를 제공함[7]
- 특히 이 서비스는 풍자, 오락 목적의 합성 콘텐츠와 악성 딥페이크를 구분하는 방향으로 발전하고 있으며, 딥페이크 탐지 기술이 기업 보안 도구를 넘어 일반 이용자 보호 수단으로 확대되고 있음을 보여줌

[표1] 딥페이크 탐지 기술의 적용 접점 비교

구분	헤일로(Halo)	젠, 인텔(Gen, Intel)	리얼체크(RealCheck)
제공 주체	스캠에이아이, 퀄컴	젠, 인텔	비트디펜더
적용 접점	화상회의	영상 재생, 시청	모바일 링크, 파일
처리 위치	기기 내부(PC)	기기 내부(PC)	모바일 앱
탐지 시점	회의 중 실시간	재생, 이용 중 실시간	파일, 링크 제출 및 분석 시
주요 대상	기업(채용담당, 경영진)	개인 시청자	개인 이용자
특징	합성 영상 실시간 표시	오디오, 비디오 동시	악성, 풍자 구분

출처: 참고문헌 종합하여 재구성

시사점: AI 합성 콘텐츠 검증과 얼굴, 음성 권리 보호 체계의 변화

• 개인 고유 특징 보호의 보조 수단

- 온디바이스 실시간 탐지는 얼굴, 음성 등 개인 고유의 특징이 무단으로 합성됐는지를 이용 시점에 조기 식별하는 기술적 보조 수단으로 활용될 수 있으며, 사후 대응에 앞선 예방적 확인 지점을 제공함

6) Georgia Collins, "Exploring Gen's AI Breakthrough for Deepfake Detection", AI Magazine, 2026.01.14., https://aimagazine.com/news/on-device-ai-breakthrough-for-real-time-deepfake-detection
7) Joel R. McConvey, "New deepfake detection product launches reflect expanding market", BiometricUpdate.com, 2026.06.30., https://www.biometricupdate.com/202606/new-deepfake-detection-product-launches-reflect-expanding-market

- 다만 탐지 결과는 합성 여부에 관한 기술적 판단에 그치므로, 저작권, 초상권 및 음성권 등 개인의 권리 침해에 관한 최종 판단을 대체하지는 못하며, 법적, 절차적 검토가 함께 이뤄질 필요가 있음

- **진위 확인 지점의 이동과 국내외 제도 과제**

- 콘텐츠 진위를 확인하는 지점이 플랫폼 서버와 사후 신고 중심에서 이용자의 기기와 실시간 이용 환경으로 옮겨가면서, 진위 검증이 유통 이후가 아니라 이용 과정 안으로 들어오는 구조 변화가 나타남
- 이에 따라 화상회의, 영상 시청 및 모바일 파일 확인 등 콘텐츠 접점별로 탐지 방식이 나뉘면서, 단일한 검증 수단보다 이용 맥락에 맞춘 다층적 대응 체계가 요구될 가능성이 큼
- 국내에서도 한국인터넷진흥원이 딥페이크 탐지와 사기 억제 기술 개발을 추진하는 등 개인정보 보호와 AI 합성 콘텐츠 대응이 맞물리는 흐름이 확인되며, 탐지와 함께 표시, 증거 보존 및 이의제기 절차를 아우르는 제도 설계가 과제로 제기됨

참고문헌

- Joel R. McConvey, "New deepfake detection product launches reflect expanding market", BiometricUpdate.com, 2026.06.30., https://www.biometricupdate.com/202606/new-deepfake-detection-product-launches-reflect-expanding-market
- Scam.ai, "Scam.ai Announces Qualcomm Partnership, Launches Halo Deepfake Detection Model at Computex 2026", Business Wire, 2026.06.25., https://www.businesswire.com/news/home/20260623132017/en/Scam.ai-Announces-Qualcomm-Partnership-Launches-Halo-Deepfake-Detection-Model-at-Computex-2026
- Dealroom.co, "Scam.ai partners with Qualcomm to launch Halo on-device deepfake detection model", Dealroom.co, 2026.06., https://app.dealroom.co/news/feed/scam-ai-partners-with-qualcomm-to-launch-halo-on-device-deepfake-detection-model
- Olivier Blanchard, "Qualcomm Unveils Future of Intelligence at CES 2026, Pushes the Boundaries of On-Device AI", The Futurum Group, 2026.01.16., https://futurumgroup.com/insights/qualcomm-unveils-future-of-intelligence-at-ces-2026-pushes-the-boundaries-of-on-device-ai/
- Georgia Collins, "Exploring Gen's AI Breakthrough for Deepfake Detection", AI Magazine, 2026.01.14., https://aimagazine.com/news/on-device-ai-breakthrough-for-real-time-deepfake-detection

AI 기반 불법 스트리밍 확산과 콘텐츠 보안 대응 체계의 변화

AI 활용 불법 스트리밍의 자동화와 개별 도구 중심 대응의 한계

• AI 활용으로 자동화 및 대규모화되는 불법 스트리밍 유통

- 최근 불법 스트리밍에 AI를 활용한 자동화 운영 방식이 확산되면서, 콘텐츠 가치가 가장 높은 라이브 스포츠 중계를 중심으로 불법 스트리밍이 짧은 시간 안에 대규모로 유통되는 양상을 보이고 있음
- 불법 스트리밍 사업자들은 AI를 활용해 불법 사이트 생성부터 콘텐츠 수집과 배포까지 유통 과정을 자동화하며, 이에 따른 새로운 불법 스트리밍 서비스가 수 분 만에 등장하고 확산하는 사례가 늘어나고 있음
- 미국 무역대표부(Office of the United States Trade Representative, USTR)* 자료에 따르면, 라이브 콘텐츠 불법 스트리밍으로 인한 글로벌 스포츠 산업의 잠재 수익 손실을 연간 최대 280억 달러(약 43조 3,550억원)[1]로 추정되며 최대 규모의 불법 스트리밍 플랫폼은 연간 16억 회 이상의 방문을 기록한 것으로 나타남[2]

* 미국 무역대표부(Office of the United States Trade Representative, USTR): 미국의 통상 정책을 총괄하는 대통령 직속 기관으로, 매년 지식재산권 침해가 심각한 온라인·오프라인 시장을 지정한 '악명 높은 시장 목록'을 발표함

• 개별 보안 도구 중심의 사후 대응이 직면한 한계

- 라이브 스포츠 중계는 경기 진행 중 짧은 시간에 상업적 가치가 집중되는 특성이 있어, 경기 종료 이후 이루어지는 차단 조치만으로는 이미 발생한 시청 이탈과 광고 수익 손실을 만회하기 어려운 구조임
- 2024년 유럽에서 탐지된 수백만 건의 불법 라이브 스트리밍 가운데 81%는 결국 차단되지 않았으며, 삭제 통지 이후 30분 안에 제거된 비율도 3%에 미치지 못한 것으로 조사됨[3]
- 사업자와 권리자는 개별 보안 솔루션에서 생성되는 대량의 탐지 결과와 경보를 관리해야 하지만, 어떤 위협부터 우선 대응해야 하는지 종합적으로 판단하기 어려워 대응의 효율성이 떨어진다는 지적이 제기됨
- 이에 따라 도구별로 흩어진 탐지 데이터를 통합해 분석하고, 피해 규모와 시간 민감도를 기준으로 대응 우선순위를 정할 수 있는 관리 수단의 필요성이 커지고 있음

1) 1달러=1,548.40원(2026.07.01., KEB 하나은행 매매기준율)
2) USTR, "USTR Releases 2025 Review of Notorious Markets for Counterfeiting and Piracy", 2026.03.03., https://ustr.gov/about/policy-offices/press-office/press-releases/2026/march/ustr-releases-2025-review-notorious-markets-counterfeiting-and-piracy
3) Julian Clover, "Industry-wide call for EU to act on real-time piracy", Broadband TV News, 2025.10.29., https://www.broadbandtvnews.com/2025/10/29/industry-wide-call-for-eu-to-act-on-real-time-piracy/

분산 데이터 통합과 우선순위 기반 대응 구조의 형성

• 분산된 데이터의 통합을 통한 불법 유통 활동의 실시간 파악

- 스위스 콘텐츠 보안 기업 나그라비전(Nagravision)은 2026년 6월 22일 자사의 스트리밍 보안 기술을 하나의 분석 체계로 통합한 지능형 보안 플랫폼 나그라 벤투리(NAGRA Venturi)를 출시함[4)]
- 벤투리는 스트리밍 생태계 전반에서 발생하는 불법 유통 관련 데이터를 수집 및 분석해 불법 스트리밍 활동을 실시간으로 파악하도록 설계되었으며, 개별 보안 솔루션에서 생성되던 탐지 결과와 경보를 하나의 분석 체계에서 관리하는 구조를 갖춤
- 기존의 다른 플랫폼에서는 개별 사업자가 자사 서비스에서 발생한 침해만 확인할 수 있었다면, 벤투리는 생태계 전반의 데이터를 함께 분석함으로써 불법 스트리밍 인프라와 유출 콘텐츠의 확산 양상까지 파악할 수 있도록 지원함
- 벤투리는 명확성(clarity), 집중(focus), 영향(impact)의 세 가지 운영 원칙을 중심으로 설계되었으며, 각 원칙이 탐지부터 대응과 성과 측정까지 연결되는 구조임

[표1] 나그라 벤투리의 운영 원칙별 기능과 대응 단계

운영 원칙	주요 기능	대응 단계
명확성 (Clarity)	분산 데이터의 실시간 통합 및 가시화	탐지 및 현황 파악
집중 (Focus)	우선 대응이 필요한 위협 식별	대응 우선순위 설정
영향 (Impact)	대응 성과의 사업 영향 분석	대응 및 성과 측정

출처: NAGRAVISION, "NAGRAVISION Launches NAGRA Venturi, Intelligence-Led Streaming Security for the AI Era", 2026.06.22., https://www.nagra.com/nagravision-launches-nagrar-venturi-intelligence-led-streaming-security-ai-era

• 피해 규모와 시간 민감도에 따른 대응 우선순위 설정

- 벤투리는 탐지된 모든 위협에 동일하게 대응하는 대신, 피해 규모와 시간 민감도를 종합적으로 고려해 우선 대응이 필요한 위협을 선별하는 방식을 적용함
- 이러한 방식은 콘텐츠 가치가 경기 시간에 집중되는 라이브 스포츠 중계처럼, 대응 시점에 따라 피해 규모가 크게 달라지는 콘텐츠에서 특히 효과적인 접근으로 평가됨
- 나그라비전은 불법 스트리밍 사업자의 AI 활용으로 불법 유통 방식이 자동화되는 상황에서, AI 기반 침해에는 AI 기반 분석 체계로 대응해야 한다는 방향을 제시함

• 개별 보안 기술과 대응을 연계한 통합 운영

- 벤투리 포트폴리오에는 포렌식 워터마킹(forensic watermarking)*과 멀티 DRM(Digital Rights Management)** 등 기존 보안 기술이 그대로 포함되며, 이를 탐지와 조사, 차단 단계와 연계해 하나의 운영 체계에서 활용하도록 구성됨
- 또한 관리형 서비스인 넥서스(NEXUS)에서는 분석가들이 전 세계 불법 스트리밍 활동을 모니터링하고 조사와 차단 대응을 통합적으로 수행함
- 이 체계는 개별 불법 스트리밍을 차단하는 데 그치지 않고 유통망 전반을 동시에 교란하는 것을 목표로 하며, 대응 성과 역시 탐지 건수보다 수익 보호와 침해 감소 등 사업 성과 중심으로 평가하는 방식을 제시함

* 포렌식 워터마킹(forensic watermarking): 영상에 육안으로 식별되지 않는 고유 식별 정보를 삽입해, 콘텐츠가 유출되었을 때 출처와 유통 경로를 추적할 수 있도록 하는 기술

** DRM(Digital Rights Management): 영화·음악·전자책·게임 등 디지털 콘텐츠가 무단 복제·배포·시청되지 않도록 이용 권한을 관리하고 접근을 제한하는 기술

4) NAGRAVISION, "NAGRAVISION Launches NAGRA Venturi, Intelligence-Led Streaming Security for the AI Era", 2026.06.22., https://www.nagra.com/nagravision-launches-nagrar-venturi-intelligence-led-streaming-security-ai-era

시사점: 콘텐츠 보안 대응 체계의 변화

• 도구 중심 대응에서 데이터 기반 통합 대응으로의 변화

- 벤투리 사례는 콘텐츠 보안의 초점이 워터마킹과 DRM 등 개별 기술의 성능 향상에서, 분산된 탐지 데이터를 통합해 대응 우선순위를 결정하는 운영 체계 구축으로 이동하고 있음을 보여줌
- 불법 유통이 AI를 활용해 자동화 및 대규모화되는 환경에서는 모든 침해를 동일하게 대응하기보다, 피해 규모와 시간 민감도를 고려해 대응 자원을 집중하는 방식의 중요성이 커지고 있음
- 또한 대응 성과를 차단 건수 등 기술 지표보다 수익 보호와 시청 경험 유지 등 사업 성과 중심으로 평가하려는 흐름은, 콘텐츠 보안을 권리 보호를 위한 운영 역량으로 확장하는 방향을 나타냄

• 콘텐츠 보안의 통합 운영 체계 확산 가능성

- 유사한 시기에 클라우드 기업 패스틀리(Fastly)와 스페인 프로축구리그 라리가(LaLiga)도 실시간 불법 스트리밍 탐지를 위한 AI 기반 시스템을 공동 개발하는 등[5], 지능형 통합 대응이 산업 전반으로 확산되는 양상을 보임
- 이러한 흐름은 개별 보안 기술을 추가하는 방식보다 다양한 탐지 정보를 통합해 신속하게 대응하는 운영 체계의 중요성이 커지고 있음을 시사함
- 다만 이러한 통합 대응 역시 침해 자체를 완전히 방지하는 수단은 아니므로, 기술적 대응과 함께 실시간 차단이 가능한 집행 체계 및 제도적 기반도 함께 마련될 필요가 있음

참고문헌

- NAGRAVISION, "NAGRAVISION Launches NAGRA Venturi, Intelligence-Led Streaming Security for the AI Era", 2026.06.22., https://www.nagra.com/nagravision-launches-nagrar-venturi-intelligence-led-streaming-security-ai-era
- Julian Clover, "Fastly and LaLiga partner on AI anti-piracy solution", Broadband TV News, 2026.04.10., https://www.broadbandtvnews.com/2026/04/10/fastly-and-laliga-partner-on-ai-anti-piracy-solution/
- USTR, "USTR Releases 2025 Review of Notorious Markets for Counterfeiting and Piracy", 2026.03.03., https://ustr.gov/about/policy-offices/press-office/press-releases/2026/march/ustr-releases-2025-review-notorious-markets-counterfeiting-and-piracy
- Julian Clover, "Industry-wide call for EU to act on real-time piracy", Broadband TV News, 2025.10.29., https://www.broadbandtvnews.com/2025/10/29/industry-wide-call-for-eu-to-act-on-real-time-piracy/

5) Julian Clover, "Fastly and LaLiga partner on AI anti-piracy solution", Broadband TV News, 2026.04.10., https://www.broadbandtvnews.com/2026/04/10/fastly-and-laliga-partner-on-ai-anti-piracy-solution/

주간 기술 동향

탐지를 우회하는 적응형 오디오 워터마크 공격 기술, AWM

• 통계적 방어를 무력화하는 적응형 오디오 워터마크 공격 기술의 등장

생성형 AI 기술의 발전으로 음성과 음악을 합성하는 오디오 생성 모델이 사회 전반에 빠르게 확산되고 있다. 이에 따라 원저작자의 동의 없이 AI가 생성한 음성이나 음악이 무단으로 복제되어 주요 플랫폼에 재유통되는 사례가 늘어나면서 저작권 침해에 대한 우려가 커지고 있다. 특히 특정인의 목소리를 불법적으로 복제하는 음성 복제 공격은 금전적 손실과 명예 훼손 등 심각한 피해로 이어질 수 있어, 이를 막을 수 있는 기술적 대응 수단의 필요성이 대두되고 있다.

이러한 문제에 대응하기 위해 딥러닝을 활용한 오디오 워터마킹 기술이 제안되었으며, 사람의 귀로는 들을 수 없는 비가청 신호를 오디오 파일에 삽입해 권리와 소유를 증명하는 수단으로 활용되고 있다. 대부분의 워터마킹 기술은 워터마크 신호를 삽입하는 인코더와 이를 추출하는 디코더로 구성되며, 재녹음이나 음성 변환, 손실 압축 등 다양한 왜곡을 견디도록 학습된다. 그러나 실제 환경에서는 디코더가 외부에 공개적으로 노출되는 경우가 많아, 공격자가 이를 악용할 위험이 존재한다. 최근 연구에서는 공격자가 정교한 섭동을 추가해 디코더의 출력을 의도한 대로 왜곡시킬 수 있다는 사실이 확인되었다.

이를 막기 위해 방어자는 워터마크 디코더가 출력하는 메시지 확률값이 정상 오디오 파일에서 일정한 통계적 분포를 따른다는 점에 착안하여, 이상 수치를 탐지하는 방식으로 조작 여부를 판별해왔다. 그러나 공격자가 오디오 품질 저하를 감수하면서 공격 효과만을 극대화하면 탐지를 피하기 어려워지고, 반대로 탐지 회피에 지나치게 집중하면 공격 성공률이 낮아지는 상충 관계가 존재한다. 더욱이 공격자가 방어자의 통계적 기준을 정확히 알지 못한 채 이를 추정해야 한다는 점은 기존 공격 기법의 실효성을 제한해왔다.

본 보고서에서는 통계적 방어 체계를 우회하도록 설계된 적응형 오디오 워터마크 공격 기술인 AWM을 소개한다. AWM은 두 단계의 최적화를 거치는데, 첫 단계는 공격의 성공을 확보하는 데 집중하고, 두 번째 단계는 확보한 공격을 유지하면서 오디오 품질을 개선하는 데 초점을 맞춘다. 이 과정에서 AWM은 제한된 오디오 샘플만으로 방어자의 정상 분포 평균과 표준편차를 추정한 뒤, 디코더 출력 확률값을 추정된 범위 안으로 되돌려 탐지를 회피한다. 이러한 접근은 워터마크 교체, 생성, 제거의 세 가지 공격 유형 모두에 적용되어, 기존 방어 기술의 취약점을 드러낸다.

[사례] 탐지를 우회하는 적응형 오디오 워터마크 공격, AWM

• AWM의 개발 배경

- 오디오 워터마킹은 사람의 귀로 들을 수 없는 신호를 음원에 삽입해 소유권을 주장하고 무단 사용을 추적하는 저작권 보호 수단으로 널리 활용되고 있음
- 그러나 오디오 워터마크를 지우거나 위조하려는 적대적 공격이 꾸준히 제기되어 왔으며, 최근에는 탐지를 회피하면서도 오디오 품질까지 유지하는 정교한 공격이 등장하고 있음
- 적응형 오디오 워터마크 공격인 AWM(Adaptive audio WaterMark attack)은 대표적인 공격에 해당하며, 워터마킹이 저작권 보호 수단으로서 지닌 취약성을 드러내는 연구임

• 공격 준비 단계: 세 가지 공격 유형과 방어 분포 추정

- AWM은 공격을 성공시키는 1단계와 오디오 품질을 회복하는 2단계의 최적화로 작동하지만, 그에 앞서 방어자의 정상 분포를 추정하는 준비 과정을 거침
- 공격 유형은 세 가지로 나뉘는데, 교체는 이미 삽입된 워터마크를 다른 내용으로 바꾸는 공격, 생성은 워터마크가 없는 오디오에 워터마크가 있는 것처럼 위조하는 공격, 제거는 삽입된 워터마크를 지워 없는 것처럼 만드는 공격임
- 탐지를 피하려면 공격 후 확률값이 방어자의 정상 범위 안에 있어야 하지만, 공격자는 그 범위의 평균과 표준편차를 알지 못하므로 소수의 표본만으로 이를 근사해 추정해야 함
- 교체와 생성 공격에서는 워터마크가 삽입된 오디오를 분석하는데, 이 경우 확률값이 비트 0과 비트 1 두 갈래로 나뉘므로 각 갈래에 대응하는 평균과 표준편차를 따로 계산해야 함
- 반면 제거 공격에서는 워터마크가 없는 정상 오디오를 바로 분석하며, 확률값이 한 갈래로만 모이므로 평균과 표준편차 한 쌍만 구하면 되어 상대적으로 간단함

• 공격 성공과 음질 회복의 2단계 최적화

- 준비 과정에서 방어자의 정상 분포를 파악한 뒤에는, 공격을 성공시키는 1단계와 오디오 품질을 회복하는 2단계 최적화가 순서대로 진행되며 각 단계는 서로 다른 목표를 우선함
- 1단계에서는 원본 오디오에 작은 섭동을 더한 뒤 디코더에 통과시키고, 디코더 출력이 공격자가 의도한 목표값에 가까워지도록 섭동을 반복적으로 조금씩 갱신함
- 이때 이미 정상 범위 안에 들어온 비트는 그대로 두고 아직 범위를 벗어난 비트에만 최적화를 집중해, 조작하면서도 확률값이 정상 분포처럼 보이도록 만들어 탐지를 회피함
- 2단계에서는 허용 범위를 넓혀 섭동에 여유를 주고 정해진 횟수만큼 최적화를 이어가며, 소프트맥스(Softmax)*를 적용한 스펙트로그램 손실을 사용해 음량과 음질을 원본에 가깝게 회복함

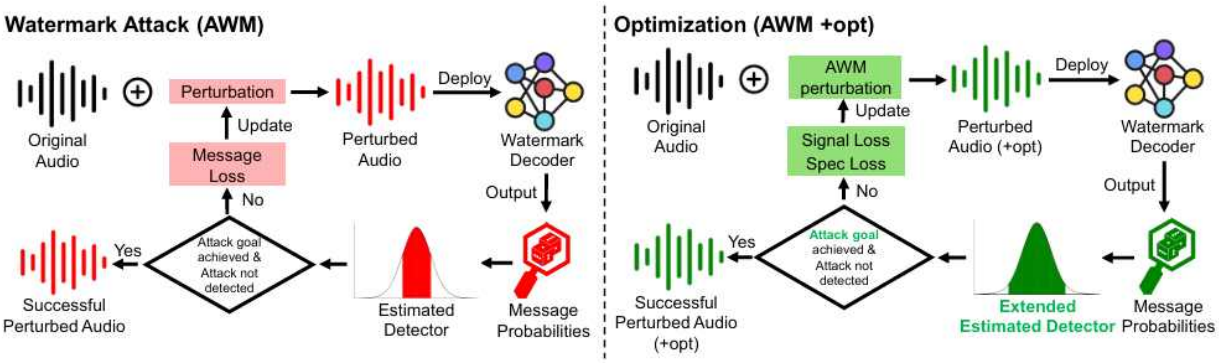
* 소프트맥스(Softmax): 여러 개의 실수 값을 총합이 1이 되는 확률분포 형태로 변환해, 값들 사이의 상대적 크기를 유지하면서 정규화하는 함수

• 탐지 회피 성능 및 공격 성공률 평가

- 이번 실험은 탐지 성공률(Detection Success Rate, DSR)과 오탐률(False Acceptance Rate, FAR)을 이용해, 세 개의 음성 데이터셋과 두 개의 워터마킹 모델을 대상으로 탐지 회피 능력을 종합 평가함
- 기존 AudioMarkBench 방식은 세 가지 공격 유형 대부분에서 탐지율이 90%를 넘겨 쉽게 적발된 반면, AWM은 모든 유형에서 탐지율을 크게 낮춰 방어자의 눈을 효과적으로 피함

- 교체와 생성 공격에서는 탐지율이 대부분 10% 미만으로 떨어졌고, 제거 공격에서는 모든 데이터셋과 모델에서 탐지된 사례가 하나도 없어 탐지율이 0%를 기록함
- 결과적으로 제거 공격이 탐지를 가장 완전하게 피하는 유형으로 나타났으며, 교체와 생성 역시 10% 미만으로 낮은 탐지율을 보였으나 완전히 0에 이르지는 못한 것으로 확인됨

[그림 1] AWM 공격 및 품질 최적화 흐름도



출처: Weikang Ding 외 6인, "Learning to Evade: Adaptive Attacks on Audio Watermarking", arXiv, 2026.06.21., https://arxiv.org/pdf/2606.22310

• 음질 및 강건성 평가 결과

- 오디오 품질은 신호대잡음비(Signal-to-Noise Ratio, SNR)와 사람이 느끼는 음질을 정량화한 지표인 ViSQOL(Virtual Speech Quality Objective Listener)*로 측정했으며, 공격을 가하지 않은 원본 오디오를 기준으로 비교함
- 워터마크 교체와 제거 공격에서는 음질 개선을 거친 AWM의 품질이 원본과 거의 차이 없이 회복되어, 기존 방식과 비슷한 수준의 품질을 안정적으로 보임
- 워터마크 생성 공격에서는 기존 방식이 더 높은 품질 지표를 기록했지만, 이는 섭동을 약하게 더해 음질만 지킨 것이며, AWM은 음질을 다소 낮추는 대신 섭동을 강하게 더해 이후 편집에도 잘 견딤
- 저역통과필터, 진폭조정, 가우시안 잡음(Gaussian Noise)**, MP3 압축, 고역통과필터 등 다섯 가지 편집을 가한 뒤에도 AWM이 생성한 워터마크는 대부분 100%에 가까운 정확도로 유지된 반면, 기존 방식으로 생성한 워터마크는 편집에 쉽게 손상됨

* ViSQOL(Virtual Speech Quality Objective Listener): 원본과 비교해 사람이 실제로 느끼는 음질을 점수로 환산하는 객관적 음질 평가 지표
** 가우시안 잡음(Gaussian Noise): 신호에 무작위로 크기 변화를 더하는 잡음으로, 녹음이나 압축 과정에서 자연스럽게 발생하는 음질 저하 현상

결론 및 시사점

• AWM의 한계와 워터마크 방어 기술의 향후 과제

- AWM은 워터마크 제거 공격에서 탐지율 0%로 가장 완전하게 탐지를 피했고 교체와 생성 공격에서도 탐지율을 10% 미만으로 낮췄으나, 이 두 공격은 여전히 탐지 가능성이 남아 있음
- 오디오 품질을 개선하는 추가 단계를 적용하면 탐지 위험이 커지는 경향이 나타나, 공격 성공률과 품질, 탐지 회피 사이의 균형을 동시에 만족시키는 것은 여전히 까다로운 과제로 남아 있음
- 메시지 확률의 통계적 분포에만 의존하는 기존 탐지 방식은 정교하게 설계된 적응형 공격 앞에서 손쉽게 무력화될 수 있다는 사실이 이번 실험으로 확인됨

- 향후에는 메시지 확률 분포뿐 아니라 스펙트로그램 패턴이나 신호 자체에 남는 미세한 흔적처럼 여러 신호를 함께 살펴보는 탐지 기법을 마련해, 적응형 공격에도 흔들리지 않는 방어 체계를 갖추는 것이 과제로 남음

참고문헌

• Weikang Ding 외 6인, "Learning to Evade: Adaptive Attacks on Audio Watermarking", arXiv, 2026.06.21., https://arxiv.org/pdf/2606.22310