저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 AI 생성 이미지를 탐지하고 근거까지 설명하는 RealOrRender 공개

▶ 생성형 AI 기술이 고도화되면서 실제 촬영물과 구별하기 어려운 이미지가 늘어나고, 이러한 이미지가 출판물이나 광고, 플랫폼 콘텐츠에 삽입되는 사례가 확산되고 있다. 배포된 이후에 진위를 소급 확인하던 기존의 사후 검증 방식은 어떤 근거로 진위를 판단했는지 설명하기 어려워, 콘텐츠의 신뢰성을 확보하는 데 구조적 한계를 보이고 있다. 이에 따라 이미지의 진위를 단순히 탐지하는 데 그치지 않고, 왜 그렇게 판정했는지 근거를 함께 제시하며 검증 과정을 투명하게 드러내는 방향으로 무게 중심이 이동하고 있다. 이러한 변화는 일부 탐지 도구의 개선에 머무르지 않고, 콘텐츠의 진위 검증과 출처 확인, 나아가 저작권 콘텐츠 검수 전반의 신뢰 기반을 다시 설계하는 흐름으로 확산될 가능성이 있다.

산업 베트남 언론 AI 활용 규제로 본 AI 콘텐츠 제작 과정의 책임 내재화

▶ 베트남은 언론의 AI 활용이 확대되면서 AI 콘텐츠로 인한 진위 혼동과 신뢰 저하 우려가 커지자, 시행령 제237호와 베트남기자협회의 AI 활용 10대 규칙을 통해 사전적 관리 체계를 마련하였다. 시행령은 AI로 생성하거나 편집한 콘텐츠에 라벨링을 적용하고, 진위와 적법성 검증, 담당자별 책임, 활용 기록 보관 의무를 언론사에 부과하는 한편 허위정보와 권익 침해 콘텐츠의 생성·유포를 금지한다. 이와 함께 자율 규범은 AI를 기자의 판단을 지원하는 보조 도구로 한정하고, 기자의 최종 책임과 발행 전 검토, 다중 검증, 데이터 보안, 저작권 보호를 요구한다. 이는 제작 과정의 투명성과 책임을 강화하는 접근으로서 의미가 있으며, 제도의 실효성을 높이기 위해서는 내부 심의 절차와 감독 도구의 실제 작동 여부를 지속적으로 점검해야 한다.

산업 리얼리티체크, 실시간 오디오 딥페이크 탐지를 통한 사후 분석 한계의 보완

▶ 오디오 딥페이크는 기업 임원과 정치인, 가족을 사칭한 전화 사기뿐 아니라 허위 콘텐츠 제작과 콜센터 사기, 음성 인증 우회 등으로 악용 범위가 확대되고 있다. 그러나 기존 탐지는 녹음이 끝난 파일을 사후 분석하는 방식이어서 피해 발생 전에 대응하기 어렵다. 하이더웨이 디지털의 리얼리티체크는 통화와 방송 중 전달되는 음성을 실시간으로 분석하고, 주파수 변화와 운율, 압축 과정의 왜곡, 생성 모델 고유 패턴을 여러 AI 모델로 교차 탐지한다. 판정 결과는 기업의 통화와 방송, 인증 서비스에 연동되어 위험 표시와 이용 차단, 추가 인증 등에 활용된다. 다만 짧은 음성만으로 정확도를 유지하고 새로운 합성 기법에 대응해야 하므로, 다중 인증과 거래 이상 탐지 등 다른 보안 수단과 결합해 활용할 필요가 있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 AI 학습 거부 의사표시 방식의 콘텐츠 내재화와 공통 표준 마련 과제

▶ AI 학습용 데이터 수집이 늘면서 저작권자의 학습 거부 의사를 기계가 읽을 수 있게 표시하는 방식이 저작권 관리의 과제 중 하나로 다뤄지고 있다. robots.txt 등 기존의 서버 단위 표시는 거부 신호를 웹사이트나 서버에 저장해, 콘텐츠가 복제되어 옮겨지면 표시와 콘텐츠가 분리되는 한계가 있다. 이에 최근에는 거부 정보를 콘텐츠의 출처 증명 메타데이터에 함께 기록해 파일이 이동해도 표시가 따라가도록 하는 방식이 제시되며, 학습 허용·불허·조건부처럼 세부적인 의사표시도 가능하다. 다만 이 표시 역시 제거될 수 있어 워터마크 등으로 보완하나 완전하지는 않다. 유럽연합이 공통 인정 목록을 정하는 논의를 진행하며 거부 표시 준수가 규제 이행 요건으로 다뤄지지만, 여러 방식이 병존해 공통 표준 논의의 실효성이 관건이 된다.

산업 대형 스포츠 이벤트 불법 스트리밍 대응의 다층화와 차단 한계

▶ 대형 스포츠 생중계는 월드컵과 프리미어리그 등을 중심으로 다수의 도메인과 미러 사이트를 통해 불법 스트리밍이 대규모로 유통되나, 경기 종료 후 이루어지는 사후 개별 차단은 실시간 중계의 짧은 상업적 가치를 보호하기 어렵고 대체 도메인의 재생성으로 실효성에 한계가 있다. 이에 대응 수단은 도메인 압수와 국제 공동 단속, 법원의 동적 금지명령과 ISP 기반 차단, 광고 수익 차단 등 여러 층으로 나뉘고 있다. 다만 미러 사이트, 대체 인프라가 지속적으로 재생성되고 침해 인프라가 여러 국가에 분산되어 있어 집중 단속을 넘어선 상시적, 다층적 집행 체계와 저작권, 소비자 보호 및 사이버보안의 연계 필요성이 제기된다.

기술 주간 기술 동향

▶ AI 음악 산출이 빠르게 확산되면서 음악의 출처를 검증하는 워터마킹의 중요성이 커지고 있다. 그러나 기존 오디오 워터마킹은 대부분 음악을 산출한 뒤 신호를 심는 사후 방식이라, 의미적 내용만 남기고 나머지를 버리는 신경망 코덱 재합성 앞에서 워터마크가 쉽게 사라진다. 뮤직마크는 이 한계를 넘어, 음악을 산출하면서 동시에 의미 잠재 공간에 워터마크를 직접 심는 최초의 음악용 생성형 워터마킹 기술이다. 워터마크 어댑터와 분리형 교차 어텐션으로 워터마크를 음악의 의미 표현 안에 결합하고, 기반 모델은 고정한 채 어댑터와 검출기만 학습해 음질을 해치지 않는다. 그 결과 코덱 재합성이나 커버송 변환 같은 공격에도 사후 방식보다 안정적으로 워터마크를 지켜내, AI 음악의 출처와 저작권 귀속을 검증하는 보조적 잠재 수단으로 주목받는다.

AI 생성 이미지를 탐지하고 근거까지 설명하는 RealOrRender 공개

AI 생성 이미지 확산에 따른 사후 진위 검증 방식의 한계

- **생성형 AI 고도화에 따른 이미지 진위 판별의 어려움 확대**

- 최근 생성형 AI가 사람 얼굴이나 사물, 장면 전체를 실제처럼 만들어내는 수준에 이르면서, 원본과 구별하기 어려운 조작 이미지가 인터넷에 빠르게 확산되고 있음
- 이렇게 만들어진 이미지는 창작 분야에서 새로운 활용 가능성을 열어주는 한편, 허위 정보가 퍼질 가능성을 함께 키워 증거를 왜곡하거나 사회적 신뢰를 떨어뜨릴 수 있다는 우려가 제기됨
- 조작된 이미지가 정교해질수록 사람의 눈으로 진위를 가려내기 어려워, AI 생성 여부를 안정적으로 판별하는 기술적 장치의 필요성이 콘텐츠 검증 측면에서 핵심 과제로 인식됨

- **탐지 결과만 제시하는 검증 방식의 신뢰 확보 한계**

- 기존 딥페이크 탐지 기술은 배포된 이미지가 조작인지 여부만 판정하는 방식이 일반적이어서, 어떤 근거로 판단했는지는 이용자가 확인하기 어려운 구조였음
- 판정 근거가 드러나지 않으면 탐지 결과를 신뢰하거나 재검증하기 어렵고, 보안 기관이나 언론사, 플랫폼이 실제 업무에 활용할 때 판단의 책임 소재를 확인하기 어렵다는 제약이 있었음
- 이미지 생성 모델의 성능이 계속 높아지면서 기존 탐지 방식의 정확도만으로는 대응이 충분하지 않아, 탐지 성능과 판정 근거의 설명 가능성을 함께 확보하는 방향으로 기술 개선의 필요성이 제기됨

탐지와 판정 근거 제시를 결합한 기술 구조의 작동 방식

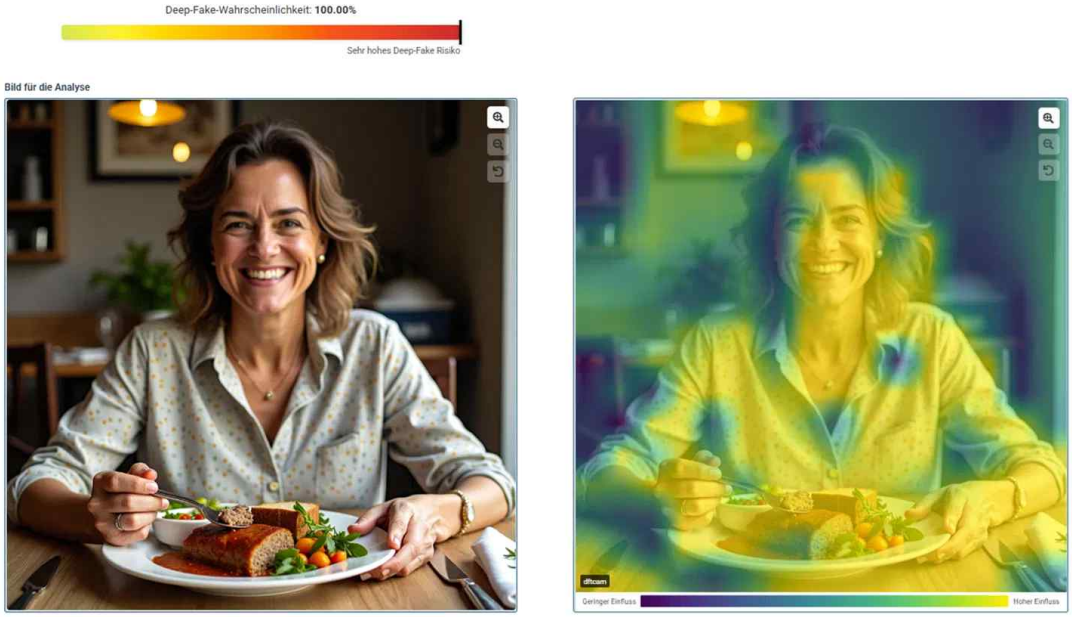
- **재구성 오차와 딥러닝 분류를 결합한 하이브리드 탐지 방식**

- 독일의 응용연구기관인 프라운호퍼 IOSB(Fraunhofer Institute of Optronics, System Technologies and Image Exploitation)가 독일 연방정보보안청(Bundesamt für Sicherheit in der Informationstechnik, 이하 BSI)의 지원을 받아 수행한 프로젝트를 통해 AI 생성 이미지를 실제 이미지와 구별하는 RealOrRender를 개발함
- 이 기술은 딥러닝(deeplearning) 기반의 전통적 분류 방식과, 생성 모델이 해당 이미지를 얼마나 잘 재구성하는지 확인하는 방식을 함께 쓰는 하이브리드 구조를 채택해 탐지 정확도를 높인 것이 특징임
- 작동 과정은 먼저 AI 이미지 생성기로 대상 이미지를 재구성하고, 이어 AI 모델이 재구성 오차를 산출해 딥페이크 여부를 분류하고, 최종적으로 인식 정확도를 백분율 추정값으로 제시하는 순서로 이루어짐
- 분류 결과는 AI 생성 또는 실제 이미지로 분류한 신뢰도 추정값으로 제시하는 방식으로 표시되어, 이용자가 해당 이미지의 AI 생성 여부를 수치로 확인할 수 있게 함

• XAI를 활용한 판정 근거 시각화 구조

- RealOrRender는 조작 여부를 탐지하는 데 그치지 않고 왜 그렇게 판정했는지를 함께 보여주는 것을 핵심 목표로 삼아, 설명가능 인공지능(explainable AI, 이하 XAI) 방식을 적용해 판정에 기여한 이미지 영역과 구조를 드러냄
- RealOrRender는 이러한 XAI 접근을 통해 특정 질감이나 특유의 주파수 패턴처럼 이미지가 AI로 생성됐음을 시사하는 특징이 드러난 영역을 표시해, 이용자가 판정 근거를 직접 확인할 수 있게 함
- 판정 근거를 제시하는 방식은 설명 결과의 표현 형식에 따라 히트맵 기반과 세그먼트 기반으로 나뉘며, 두 방식은 근거를 서로 다른 형태로 시각화함
- 히트맵 기반 방식은 영향도가 큰 위치를 시각적으로 나타내 판단에 크게 기여한 영역을 빠르게 확인하는 데 유용하고, 세그먼트 기반 방식은 이어진 이미지 구획을 의미 단위로 분석해 어떤 영역이 판단에 기여했는지 파악하는 데 유용함

[그림 1] RealOrRender의 딥페이크 탐지 결과(왼쪽)와 XAI 판정 근거 시각화(오른쪽)



출처: Fraunhofer-Gesellschaft, "Reliably Detecting and Clearly Explaining Deepfake Images", 2026.07.01., https://www.fraunhofer.de/en/press/research-news/2026/july-2026/reliably-detecting-and-clearly-explaining-deepfake-images.html

• 대규모 데이터셋 검증으로 확인된 성능과 대응 견고성

- 프라운호퍼 IOSB는 대규모 이미지 데이터셋을 대상으로 시험한 결과, 탐지 성능과 판정 근거의 설명 가능성을 결합한 이번 방식이 기존의 탐지 방식보다 나은 성능을 보였다는 점을 확인함
- 하이브리드 판정을 통해 측정한 전체 탐지율은 프라운호퍼가 사용한 데이터셋과 평가 조건에서 81~91%의 탐지 성능을 기록한 것으로 수준으로 나타났으며, 이미지의 특성에 따라 개별 사례에서는 이보다 더 높은 정확도를 보이기도 함
- 탐지 방식과 근거 설명 방식을 함께 쓰는 구조는 딥페이크 탐지의 견고성을 높여, 성능이 점점 높아지는 이미지 생성 모델에 대해서도 대응력을 유지할 수 있는 것으로 나타남
- 이러한 검증 결과는 탐지 정확도와 판정 근거의 설명 가능성을 동시에 확보하려는 접근이 실제 대규모 환경에서도 작동할 수 있음을 보여주며, 콘텐츠 검증 도구로서의 활용 기반을 뒷받침함

시사점: 콘텐츠 진위 검증 기술의 저작권 검수 활용 가능성과 과제

- **저작권 콘텐츠 검수로의 확산 가능성과 남은 과제**

- RealOrRender는 개발 단계에서부터 보안이나 법 집행 기관뿐 아니라 소셜 미디어 플랫폼 운영사, 미디어 기업, 사진 통신사 등 콘텐츠를 다루는 여러 주체가 활용할 수 있는 기술로 제시됨
- AI 생성 이미지가 출판물, 광고물, 플랫폼 게시물 등에 삽입되는 사례가 늘면서, 탐지 결과와 판정 근거를 함께 확인할 수 있는 방식은 콘텐츠의 진위 검증과 출처 확인, 나아가 저작권 검수 단계로 연결될 여지가 있음
- 다만 이미지의 AI 생성 여부를 기술적으로 탐지하는 것과 해당 이미지의 저작권 침해 여부를 법적으로 판단하는 것은 별개의 절차이므로, 기술적 탐지 결과가 곧바로 권리 판단을 대체하지는 않는다는 점을 전제로 삼아야 함

참고문헌

- Fraunhofer-Gesellschaft, "Reliably Detecting and Clearly Explaining Deepfake Images", 2026.07.01., https://www.fraunhofer.de/en/press/research-news/2026/july-2026/reliably-detecting-and-clearly-explaining-deepfake-images.html
- IDW, “Reliably Detecting and Clearly Explaining Deepfake Images”, 2026.07.01., https://idw-online.de/en/news873688

베트남 언론 AI 활용 규제로 본 AI 콘텐츠 제작 과정의 책임 내재화

언론의 AI 활용 확산에 따른 규제 필요성 증대

• 취재와 제작 전반의 AI 활용 확대에 따른 콘텐츠 진위 혼동 우려

- 베트남 정부는 AI 활용을 언론 활동의 효율성 제고 수단으로 명시하고, 언론사가 AI를 연구하고 기사의 취재, 제작, 분석, 유통 전 과정에 적용하도록 장려함
- 그러나 소셜미디어와 생성형 AI가 정보의 생산과 확산 방식을 빠르게 바꾸면서 사실과 허위 정보의 구분은 갈수록 어려워지고 있음
- 특히 AI로 생성하거나 편집한 콘텐츠가 실제 사건이나 인물에 대한 이용자의 판단을 흐릴 수 있다는 우려도 커지고 있음

• AI 콘텐츠의 진위와 적법성 검증을 위한 규제 요구 대두

- 이러한 문제의식은 2026년 6월 열린 제3회 전국언론포럼(National Press Forum)에서도 공유되었으며, 가짜뉴스 자체보다 언론에 대한 신뢰의 약화가 더 근본적인 위험이라는 인식 아래 신뢰를 유지하기 위한 대응이 논의됨[1)]
- 이러한 흐름 속에서 베트남 정부는 2025년 언론법의 이행 기준을 구체화한 시행령 제237호(Decree 237/2026/ND-CP)*를 마련해 2026년 7월 1일부터 시행함[2)3)]
- 베트남기자협회(Vietnam Journalists Association) 또한 2026년 6월 20일 언론 활동에서의 AI 활용에 관한 10대 규칙을 발표하면서, 정부의 법적 규제와 언론계의 자율 규범을 병행하는 관리 체계가 마련됨[1)]

* 시행령 제237호(Decree 237/2026/ND-CP): 2026년 6월 26일 공포되었으며, 언론사가 AI를 이용할 때 지켜야 할 책임은 제24조에 규정됨

생성 단계부터 작동하는 언론 AI 관리 체계

• AI 활용 라벨링과 책임 소재 명확화를 위한 법적 규제

- 시행령 제237호는 취재, 제작, 유통 과정에서의 AI 활용을 장려하는 한편, AI로 생성하거나 편집한 콘텐츠의 진위와 적법성을 검증할 책임을 언론사에 부과함으로써 기술 활용과 그에 따른 책임을 함께 규정함

1) Hoàng Khôi, "Announcing 10 rules for using artificial intelligence in journalistic activities", Lao Dong, 2026.06.20., https://news.laodong.vn/thoi-su/cong-bo-10-quy-tac-su-dung-tri-tue-nhan-tao-trong-hoat-dong-bao-chi-1722215.ldo
2) Bao Anh 외 1인, "Vietnam requires media to label potentially misleading AI-generated content from July", Tuoi Tre News, 2026.06.27., https://news.tuoitre.vn/vietnam-requires-media-to-label-potentially-misleading-ai-generated-content-from-july-103260627111100181.htm
3) Vietnam Law and Legal Forum, "AI-generated press content must be labelled under new rules", 2026.07.09., https://vietnamlawmagazine.vn/ai-generated-press-content-must-be-labelled-under-new-rules-79961.html

- 핵심 의무는 실제 사건이나 인물의 진위에 대해 이용자를 오인시킬 수 있는 AI로 생성하거나 편집한 텍스트, 이미지, 음성, 영상 콘텐츠에 대해, 이용자가 쉽게 인식할 수 있는 위치에 AI 이용 사실을 표시하도록 한 이른바 라벨링(labelling) 규정에 해당함
- 이와 함께 AI를 이용해 허위, 왜곡, 오도성 정보를 생성하거나 유포하는 행위와 국가 안보, 사회 질서와 안전, 조직과 개인의 명예와 권익을 침해하는 콘텐츠를 만드는 행위를 명시적으로 금지함
- 언론사는 AI 활용에 관한 심의, 편집, 위험관리 절차를 마련하고 업무 담당자별 책임을 구체화해야 하며, AI 활용 기록과 관련 기술 문서를 보관해 당국의 점검 요청에 따라 제출해야 함
- 이에 더해 문화체육관광부는 사이버공간에서 이루어지는 언론 활동을 실시간으로 관찰하고 분석하는 디지털 도구를 개발하고 운영할 예정이지만, 해당 도구를 통해 언론 콘텐츠에 직접 개입하지는 않겠다는 방침을 명확히 함

• 자율 규범과 다중 검증을 통한 제작 과정 관리

- 베트남기자협회가 2026년 6월 발표한 10대 규칙은 법적 강제력을 지닌 시행령과 달리 언론계가 자율적으로 준수하는 실무 및 윤리 기준으로, AI를 기자의 판단을 돕는 보조 도구로 한정하고 최종 책임은 기자가 부담하도록 함
- 10대 규칙은 AI 활용 사실을 표시하는 데 그치지 않고, AI를 취재원으로 인정하지 않는 원칙과 발행 전 검토와 승인 절차, 민감한 주제에 대한 다중 검증 절차를 포함해 콘텐츠 제작 전반에 적용됨
- 또한 기밀 정보 및 내부 자료와 민감한 개인정보를 AI에 입력하거나 타인의 저작물을 무단으로 복제하거나 재작성하는 행위를 금지함으로써, AI에 투입되는 데이터까지 함께 관리하도록 함

[표1] 베트남기자협회가 발표한 AI 활용 10대 규칙

번호	담당 영역	규칙 내용
1	**인간 최종 책임**	AI는 기자를 대체하는 수단이 아닌 보조 도구로 활용, 모든 산출물에 대한 최종 책임은 기자가 부담
2	**정보 검증**	AI를 취재원으로 인정하지 않으며, AI가 제시한 사실은 신뢰할 수 있는 출처와 대조 및 검증
3	**발행 전 편집 통제**	AI 콘텐츠는 발행 전 편집, 수정, 맥락 보완, 승인 절차를 거쳐 공개
4	**허위정보 방지**	AI를 이용한 가짜뉴스, 허위 데이터, 조작된 발언, 가공 인물의 생성 및 유포 금지
5	**딥페이크 방지**	당사자의 동의 없이 음성, 이미지, 얼굴을 복제하거나 이를 이용해 대중을 기만하는 행위 금지
6	**데이터 보안**	기밀 문서, 미공개 원고, 원본 데이터, 내부 정보, 민감 개인정보의 AI 업로드 금지
7	**저작권 보호**	AI로 타인의 저작물을 복제, 재작성, 부당 활용해 저작권 의무를 우회하는 행위 금지
8	**표시·투명성**	AI를 상당 부분 활용한 콘텐츠는 적절한 표시와 고지를 통해 투명성 확보
9	**민감 주제 검증**	민감한 주제에 AI를 활용할 경우 복수의 출처와 절차를 통한 다층 검증 수행
10	**법령·윤리 준수**	언론, 데이터 보호, 디지털기술산업, 지식재산권 관련 법령과 언론 직업윤리 준수

출처: Vietnam Law and Legal Forum, "AI-generated press content must be labelled under new rules", 2026.07.09., https://vietnamlawmagazine.vn/ai-generated-press-content-must-be-labelled-under-new-rules-79961.html

사전적 언론 AI 관리에 대한 함의

• 법적 규제와 자율 규범을 결합한 사전적 언론 AI 관리

- 베트남의 이번 규제는 AI 콘텐츠의 진위 검증과 책임 소재 판단을 생성과 편집 단계부터 적용한다는 점에서 사전적 언론 AI 관리의 한 방향을 제시함
- 특히 법적 강제력을 지닌 시행령과 언론계의 자율 윤리 규범을 함께 운영함으로써, 편집 실무부터 직업윤리 영역까지 폭넓게 포괄했다는 점에서 의의가 있음
- 이러한 규제의 초점은 위반 행위에 대한 사후 처벌보다 AI 활용 과정의 투명성과 언론에 대한 공적 신뢰 유지에 맞춰져 있어, 유사한 문제를 겪는 다른 국가와 언론사에도 참고가 될 수 있음
- 다만 표시와 검증 절차가 형식적 준수에 그치지 않으려면 담당자별 책임 체계와 내부 심의 절차가 정착될 필요가 있으며, 라벨링이 형식적 절차에 머무는지, 감독 도구가 실제로 가동되는지 등을 중심으로 제도의 실효성을 지속적으로 점검할 필요가 있음

참고문헌

- Hoàng Khôi, "Announcing 10 rules for using artificial intelligence in journalistic activities", Lao Dong, 2026.06.20., https://news.laodong.vn/thoi-su/cong-bo-10-quy-tac-su-dung-tri-tue-nhan-tao-trong-hoat-dong-bao-chi-1722215.ldo
- Bao Anh 외 1인, "Vietnam requires media to label potentially misleading AI-generated content from July", Tuoi Tre News, 2026.06.27., https://news.tuoitre.vn/vietnam-requires-media-to-label-potentially-misleading-ai-generated-content-from-july-103260627111100181.htm
- Vietnam Law and Legal Forum, "AI-generated press content must be labelled under new rules", 2026.07.09., https://vietnamlawmagazine.vn/ai-generated-press-content-must-be-labelled-under-new-rules-79961.html

리얼리티체크, 실시간 오디오 딥페이크 탐지를 통한 사후 분석 한계의 보완

오디오 딥페이크의 확산에 따른 사후 탐지의 한계

• 실시간 통화와 방송에서 악용되는 오디오 딥페이크

- AI가 특정 인물의 음성을 모방한 오디오 딥페이크(audio deepfake)*가 새로운 디지털 사기의 수법으로 부상하면서, 기업 임원과 정치인, 가족 등을 사칭한 실시간 전화 사기 사례가 증가함
- 이러한 오디오 딥페이크는 전화 사기에 그치지 않고, 대중을 겨냥한 허위 콘텐츠 제작, 콜센터 상담원을 사칭한 사기 등으로 악용 범위가 확대되고 있음
- 나아가 목소리로 본인 여부를 확인하는 음성 인증을 속여 금융과 정부 시스템에 타인의 신원으로 접근하는 신원 도용 공격에도 활용되면서, 개인뿐 아니라 자동화된 인증 체계까지 주요 표적이 됨

* 오디오 딥페이크(audio deepfake): 인공지능을 활용해 특정 개인의 목소리를 똑같이 모방하거나 새로운 음성을 합성하는 기술

• 사후 분석에 머문 기존 탐지 방식의 한계

- 그러나 기존 오디오 딥페이크 탐지는 대부분 녹음이 완료된 파일을 분석하는 방식에 머물러 있음
- 이에 따라 판정 결과가 통화나 방송 등 상황이 끝난 뒤에야 제공되므로, 실제 음성이 전달되는 동안에는 오디오 딥페이크를 탐지하거나 피해 확산을 즉시 막기 어려움
- 이러한 한계를 보완하기 위해 녹음이 완료된 파일을 사후에 분석하는 방식에서 벗어나, 전송 중인 음성을 곧바로 분석해 피해가 발생하기 전에 개입할 수 있는 기술에 대한 요구가 커짐

실시간 오디오 딥페이크 탐지의 작동 원리와 적용 방식

• 전송 중인 음성을 분석하는 실시간 포렌식 탐지

- 캐나다의 AI 인프라 및 사이버 보안 기업 하이더웨이 디지털(Hydaway Digital)은 2026년 7월 7일 자사 보안 플랫폼 리얼리티체크(RealityChek)에 실시간 오디오 딥페이크 탐지 기술인 스트리밍 오디오 디텍션(Streaming Audio Detection)을 기업 고객용으로 추가했다고 발표함[1)]
- 이 기술은 통화나 방송 중 오가는 음성에 대해 라이브 스트림(live stream)을 실시간으로 포렌식 분석*해, 음성이 수신되는 동안 합성 여부 판정 결과를 제공한다고 설명함

* 포렌식 분석: 디지털 데이터의 특성을 정밀하게 분석해 위조나 조작 여부를 탐지하는 기법

1) IRW-Press, "Hydaway Digital's RealityChek Launches Real-Time Streaming Audio Deepfake Detection", IRW-Press, 2026.07.07., https://irw-press.com/en/hydaway-digitals-realitychek-launches-real-time-streaming-audio-deepfake-detection/

• 다중 음향 단서와 특화 AI 모델을 결합하여 판정

- 실시간 판정을 위해 단일 지표에 의존하지 않고, 오디오 딥페이크에서 나타나는 스펙트럼 패턴(spectral pattern)*과 운율, 압축 아티팩트(compression artifact)**, 생성 모델 시그니처(generative model signature)*** 등 여러 음향적 특징과 기술적 특징을 종합적으로 분석함
- 이를 위해 각 단서의 분석에 특화된 여러 AI 모델을 결합하여, 최신 음성 합성 기술이 남기는 흔적을 서로 다른 관점에서 교차 탐지하도록 설계됨
- 구체적으로 각 AI 모델이 비정상적인 주파수 변화와 음성의 리듬, 높낮이, 강세에서 나타나는 부자연스러움을 포착하고, 오디오의 생성, 녹음, 전송 과정에서 발생한 미세한 왜곡과 AI 모델 특유의 반복적 흔적을 탐지함

* 스펙트럼 패턴(spectral pattern): 음성을 시간과 주파수 영역으로 변환했을 때 나타나는 주파수별 에너지 분포와 변화 양상

** 압축 아티팩트(compression artifact): 음성을 코덱으로 압축하거나 다른 형식으로 재인코딩하는 과정에서 원래 신호에 생기는 왜곡이나 정보 손실

*** 생성 모델 시그니처(generative model signature): 음성 합성 모델의 구조와 처리 방식이 생성된 음성 파형에 남기는 모델별 고유 패턴

[표1] 오디오 딥페이크 탐지에 활용되는 주요 신호

탐지 신호	원어	포착 대상
스펙트럼 패턴	spectral patterns	주파수 변화의 비정상성
운율	prosody	말의 리듬과 높낮이, 강세의 부자연스러움
압축 아티팩트	compression artifacts	녹음과 전송, 조작 과정에서 생긴 미세한 왜곡
생성 모델 시그니처	generative model signatures	AI 음성 생성 도구가 남기는 흔적

출처: Abhishek Jadhav, "Hydaway introduces real-time enterprise audio deepfake detection", Biometric Update, 2026.07.09., https://www.biometricupdate.com/202607/hydaway-introduces-real-time-enterprise-audio-deepfake-detection

• 판정 결과의 서비스 연동과 기업 보안 활용

- 판정 결과는 리얼리티체크의 API와 기업용 관리 화면을 통해 제공되며, 기업의 통화와 방송, 인증 서비스에 연동되어 실시간에 가깝게 전달됨
- 기업은 이를 바탕으로 오디오 딥페이크로 의심되는 음성을 위험하다고 표시하거나, 위험 수준에 따라 서비스 이용을 차단하거나 추가 인증을 요구해 실제 피해가 발생하기 전에 대응할 수 있음
- 탐지 결과는 기존 업무 시스템과 직접 연결할 수 있어, 콜센터 사기 방지와 실시간 미디어 모니터링, 음성 인증 시스템 보호 등 음성 기반 통신 및 인증 환경 전반에 활용 가능함

실시간 탐지로의 전환이 갖는 의미와 기술적 과제

• 탐지 시점이 사후 판별에서 실시간 개입으로 이동

- 이번 사례는 오디오 딥페이크의 탐지 시점을 콘텐츠 생성 이후에서 통화나 방송이 진행되는 도중으로 앞당겨, 사기가 실제 피해로 이어지기 전에 개입할 수 있는 시간을 확보했다는 점에서 의미가 있음
- 이러한 수요에 대응해 미국 AI 음성 기술 기업 옴니스피치(OmniSpeech)는 화상회의에 오디오 딥페이크 탐지 기능을 접목하고 있으며, 미국 정보보안 기업 핀드롭(Pindrop)은 영국 통신사 BT와 협력해 콜센터 음성 사기 대응을 강화하는 등 업계 전반에서 실시간 탐지 기술의 도입이 이어지고 있음[2)]
- 리얼리티체크도 기존의 영상 딥페이크 판별 기능에 이번 실시간 오디오 딥페이크 탐지를 추가하면서, 영상과 음성 등 여러 유형의 콘텐츠 진위를 함께 확인하는 통합 탐지 체계를 구축해 나가고 있음

• 실시간 탐지 기술의 한계와 보완 과제

- 다만 실시간 탐지는 발화가 끝나기 전의 짧은 음성만으로 판단해야 하므로, 처리 지연을 줄이면서 정확도를 유지하는 일이 핵심 과제로 남음. 이 과정에서 실제 음성을 합성 음성으로 분류하는 오탐과 합성 음성을 실제 음성으로 분류하는 미탐이 발생할 수 있음
- 현재의 탐지 모델은 주요 음성 합성 기술에서 나타나는 특징을 학습한 것이므로, 기존 탐지 방식을 우회하는 새로운 생성 기법에 대응하려면 학습 데이터와 모델을 지속적으로 갱신해야 함
- 아울러 오디오 딥페이크 탐지만으로 사기 위험을 완전히 차단하기는 어려운 만큼, 다중 인증과 거래 이상 탐지, 영상 진위 확인 등 다른 보안 기술과 결합해 위험 판단을 보완하는 수단으로 활용할 필요가 있음

참고문헌

- Abhishek Jadhav, "Hydaway introduces real-time enterprise audio deepfake detection", Biometric Update, 2026.07.09., https://www.biometricupdate.com/202607/hydaway-introduces-real-time-enterprise-audio-deepfake-detection
- IRW-Press, "Hydaway Digital's RealityChek Launches Real-Time Streaming Audio Deepfake Detection", IRW-Press, 2026.07.07., https://irw-press.com/en/hydaway-digitals-realitychek-launches-real-time-streaming-audio-deepfake-detection/
- Ali Nassar-Smith, "RealityChek Adds Real-Time Detection for Audio Deepfakes", IDTechWire, 2026.07.07., https://idtechwire.com/realitychek-adds-real-time-detection-for-audio-deepfakes/

2) Ali Nassar-Smith, "RealityChek Adds Real-Time Detection for Audio Deepfakes", IDTechWire, 2026.07.07., https://idtechwire.com/realitychek-adds-real-time-detection-for-audio-deepfakes/

AI 학습 거부 의사표시 방식의 콘텐츠 내재화와 공통 표준 마련 과제

웹사이트 단위 학습 거부 표시 방식의 도달 한계

• 기계 판독 가능한 거부 의사표시에 대한 규제 요구

- 유럽연합(EU)의 디지털 단일시장 저작권 지침(Directive on Copyright in the Digital Single Market)*은 인터넷에 공개된 저작물의 경우, 창작자가 이를 AI 학습에서 제외하려면 사람이 아니라 기계가 읽을 수 있는 형태로 거부 의사를 밝히도록 규정하고 있음

- 또한 유럽연합 인공지능법(EU AI Act)은 이러한 학습 거부 표시에 대한 확인 및 준수 의무를 AI 기업에게 부과했으며, 위반 시 제재는 2026년 8월부터 적용될 예정임

* 디지털 단일시장 저작권 지침(Directive on Copyright in the Digital Single Market): 2019년 유럽연합이 채택한 저작권 규정으로, 인터넷에 공개된 저작물을 AI 학습 등 텍스트·데이터 마이닝에 활용하는 것을 저작권자가 사전에 거부(opt-out)할 수 있도록 정하고 있음

• 서버 단위 표시 방식의 작동 구조와 적용 한계

- 그동안 학습 거부에 주로 쓰인 방식은 로봇 배제 표준(Robots Exclusion Protocol)*에 따라 웹사이트에 특정 파일(robots.txt)을 두는 것으로, 수집 로봇을 종류별로 각각 지정해야 하고 이를 준수할지는 수집 주체의 자율에 맡겨져 있었음

- 그러나 이 방식은 저작권자가 직접 관리하지 않는 다른 웹사이트에 자신의 저작물이 게시된 경우 거부 표시를 부착할 수 없어, 정작 보호가 필요한 상황에서 활용되기 어렵다는 한계가 지적됨

- 독일 함부르크 고등법원(Higher Regional Court of Hamburg)이 2025년 12월 관련 소송에서 이를 다뤘으나[1], 유효한 기계 판독 가능 거부 표시의 구체적 기준과 이용약관의 일반 문구가 이에 해당하는지는 여전히 쟁점으로 남아 있음

* 로봇 배제 표준(Robots Exclusion Protocol): 웹사이트 운영자가 검색·수집 로봇에게 어떤 페이지의 접근을 허용할지 알리기 위해 만든 규약으로, robots.txt 파일을 통해 적용되며 강제력 없이 로봇의 자율적 준수에 의존함

거부 표시의 콘텐츠 내재화를 통한 신호 지속성

• 사이트 및 서버 단위 신호의 작동 방식과 콘텐츠 이동 시 유실

- robots.txt의 한계를 보완하려는 방식으로는 웹 표준화 기구인 월드 와이드 웹 컨소시엄(World Wide Web Consortium)이 제시한 TDM 권리 유보 프로토콜(TDM Reservation Protocol, 이하 TDMRep)*이 있으며, 이는 거부 여부와 함께 이용 문의 경로까지 표시할 수 있도록 한 점이 특징임

1) Hilma Mäkitalo-Saarinen, "Text and Data Mining for AI Training", Hannes Snellman, 2026.01.27., https://www.hannessnellman.com/news-and-views/blog/text-and-data-mining-for-ai-training/

- 다만 해당 방식은 거부 정보를 콘텐츠가 아니라 콘텐츠가 놓인 서버나 웹페이지에 저장하므로, 콘텐츠 파일 자체에는 표시가 남지 않고 원래 게시 위치에만 신호가 남는 구조임
- 이 때문에 사진과 영상이 복제되어 다른 곳으로 옮겨지면 거부 표시는 원래 서버에 남고 콘텐츠만 이동해 표시와 콘텐츠가 서로 분리되는 한계가 존재함

* TDM 권리 유보 프로토콜(TDM Reservation Protocol, TDMRep): 월드 와이드 웹 컨소시엄이 마련한 규약으로, 저작권자가 자신의 콘텐츠에 대한 텍스트·데이터 마이닝 거부 의사와 이용 문의 경로를 기계가 읽을 수 있게 표시하도록 한 방식

• 콘텐츠 파일 내 거부 의사 기록 방식의 작동 구조

- 최근에는 거부 정보를 콘텐츠의 출처 증명 메타데이터*에 함께 기록해, 파일이 어디로 이동하든 거부 표시가 콘텐츠와 함께 따라가도록 하는 방식이 제시됨
- 이 방식은 허용, 불허, 조건부 허용처럼 어떤 활용을 거부하는지 세부적으로 구분해 표시할 수 있어, 단순 차단보다 정교한 의사표시가 가능함
- 이처럼 콘텐츠에 거부 표시를 담는 방식은 이미지에서 시작되었으나, 문서와 전자책 등 다른 형식의 파일에도 거부 정보를 담을 수 있도록 확대되면서 그 적용 범위가 넓어지고 있음

* 출처 증명 메타데이터: 콘텐츠의 제작·편집 이력과 출처를 파일 안에 함께 기록하는 정보로, 대표적으로 C2PA 콘텐츠 자격증명(Content Credentials) 방식이 활용됨

• 내재형 표시의 잔존 한계와 소프트 바인딩

- 그러나 콘텐츠 파일에 포함된 거부 표시 또한 규격상 제거가 허용되거나 중계 서버에서 처리, 전송되는 과정에서 함께 삭제될 가능성이 존재함
- 이를 보완하기 위한 방법으로는 소프트 바인딩(soft binding)*을 적용해 거부 표시가 사라진 경우에도 별도 저장소에서 관련 정보를 확인할 수 있음
- 다만 소프트 바인딩 역시 오인식 가능성이 있으며, 공개 규격을 따르는 워터마크는 구조가 이미 알려져 있어 의도적인 제거를 완전히 방지하기 어렵다는 한계가 남음

* 소프트 바인딩(soft binding): 콘텐츠의 특징을 별도로 계산해 저장하거나 눈에 보이지 않는 워터마크를 삽입하는 방식으로, 메타데이터가 사라져도 콘텐츠를 다시 식별할 수 있게 하는 보조 수단

[표1] AI 학습 거부 신호와 보완적 식별 기술의 특성 비교

구분	신호 저장 위치	콘텐츠 동시 이동	거부 의사 표현	주요 한계
서버 단위(robots.txt)	웹사이트 서버	불가	수집 허용, 거부 중심	로봇별 개별 지정, 자율 준수 의존
서버 단위(TDMRep)	서버, 웹페이지	불가	거부 및 문의 경로 제공	콘텐츠 이동 시 신호와 분리
콘텐츠 내재형 (출처 증명 메타데이터)	콘텐츠 파일	원칙적 가능	허용, 불허, 조건부 표현 가능	처리, 변환, 전송 중 제거 가능
소프트 바인딩	콘텐츠 + 저장소	가능(통계적)	거부 의사 표현 기능 부재	오인식 및 역이용 가능

출처: 참고문헌 종합하여 재구성

시사점: 기계 판독 가능한 거부 표시의 표준화 논의

- **거부 표시 준수의 규제 이행 요건화**

- 표시 기술의 변화는 표시 준수가 수집 주체의 자율적 선택에서 규제상 이행 사항으로 옮겨가는 흐름을 보여주며, 거부 표시를 무시하는 행위가 규제의 미이행으로 비춰질 수 있음을 의미함
- 유럽연합이 여러 표시 방식 중 공통으로 인정되는 방식을 정하는 가운데, 목록에 오른 표시 방식을 따르는 것이 비교적 안전한 이행 경로가 될 것으로 전망됨
- 앞으로는 거부 표시를 확인하고 준수했는지가 AI 개발사의 저작권 준수를 입증하는 근거로 다뤄질 수 있어, 표시 준수 여부를 기록, 관리하는 체계의 필요성이 커짐

- **다중 표시 방식 병존에 따른 상호운용성 과제**

- 현재의 AI 학습 거부 표시는 서버 단위 표시, 콘텐츠 내재형 표시, 워터마크 등 방식이 병존하므로 저작권자는 여러 형식으로 거부 의사를 밝혀야 하고 AI 개발사는 여러 위치를 모두 확인해야 하는 부담이 존재함
- 이러한 형식이 하나로 모이지 않으면 확인 누락이나 해석 차이로 거부 의사가 실제로 지켜지지 않을 수 있어, 공통 표준을 마련하려는 논의의 실효성이 관건이 되고 있음

참고문헌

- Hilma Mäkitalo-Saarinen, "Text and Data Mining for AI Training", Hannes Snellman, 2026.01.27., https://www.hannessnellman.com/news-and-views/blog/text-and-data-mining-for-ai-training/
- IPTC Media Provenance Advocacy Working Group, "IPTC Frequently Asked Questions on C2PA Content Credentials", IPTC, 2026.07.08., https://iptc.org/news/iptc-c2pa-content-credentials-faqs-released
- European Commission, "Stakeholder consultation on AI and copyright compliance", Shaping Europe's digital future, 2025.12., https://digital-strategy.ec.europa.eu/en/faqs/stakeholder-consultation-ai-and-copyright-compliance
- IPTC, "IPTC Generative AI Opt-Out Best Practice Recommendations (Version 2.0)", IPTC, 2026.03., https://iptc.org/std/guidelines/data-mining-opt-out/IPTC-Generative-AI-Opt-Out-Best-Practices.pdf

대형 스포츠 이벤트 불법 스트리밍 대응의 다층화와 차단 한계

스포츠 불법 스트리밍의 확산과 사후 차단 방식의 한계

• 대형 스포츠 중계의 불법 유통 확대

- 월드컵과 잉글리시 프리미어리그(English Premier League, EPL) 등 주요 경기가 열리는 시기에 불법 스트리밍 사이트가 크게 늘어나며, 다수의 도메인과 미러 사이트(mirror site)*를 통해 대규모로 유통됨
- 대표적 불법 스트리밍 사이트인 로하디렉타(RojaDirecta)의 변형 사이트는 2025년 12월 기준 월간 2,440만 회의 방문을 기록한 것으로 나타나는 등 이용 규모가 상당한 수준으로 확인됨[1)]
- 이러한 불법 유통은 독점 중계권을 확보한 공식 사업자의 상업적 이익과 직접 충돌하며, 경기 시작과 동시에 여러 경로로 확산되어 사전에 전면 차단하기 어려운 특성을 지님

* 미러 사이트(mirror site): 원본과 같은 콘텐츠를 다른 도메인 주소에서 제공하도록 복제한 사이트

• 사후 개별 차단의 시간적, 관할적 한계

- 경기가 끝난 뒤에야 이루어지는 사후 차단으로는 실시간 중계 특유의 한시적인 상업적 가치를 보호하기 어렵다는 한계가 존재함
- 침해에 활용되는 특정 도메인을 차단하더라도, 동일한 콘텐츠를 제공하는 대체 도메인이 신속하게 재생성되면서 개별 사이트 차단만으로는 유통 자체를 중단시키기 어려운 것으로 분석됨
- 특히 일부 운영자는 미국 등 집행기관의 접근이 어려운 해외 최상위 도메인(Top-Level Domain, 이하 TLD)*과 대체 인프라로 이동하며 단속을 회피하는 것으로 나타남

* 최상위도메인(Top-Level Domain, TLD): 도메인 주소에서 .com, .ir와 같이 가장 뒤에 위치하는 최상위 구분자

도메인, 법원, 수익원을 포괄하는 대응 수단의 다층화

• 대형 스포츠 이벤트를 계기로 한 국제 공동 단속의 확대

- 미국은 2026년 북중미 월드컵(FIFA World Cup) 2026 기간 중 세 차례에 걸친 단속을 통해 1,000개 이상의 불법 스트리밍 도메인을 압수했으며, 이는 2022년 카타르 월드컵 당시의 5배를 넘는 규모임[2)]
- 중남미에서도 콜롬비아가 주도한 단속에 9개국이 참여해 총 1,970개의 도메인을 차단하고 콜롬비아 내에서 관련 인원 15명을 체포함

1) The Online Citizen, "47 Piracy Websites to Be Blocked in Singapore Under Premier League Court Order", The Online Citizen, 2026.02.05., https://theonlinecitizen.com/2026/02/05/47-piracy-websites-to-be-blocked-in-singapore-under-premier-league-court-order
2) Ernesto Van der Sar, "FIFA World Cup Triggers a Global Anti-Piracy Crackdown (Updated)", TorrentFreak, 2026.07.19., https://torrentfreak.com/fifa-world-cup-triggers-a-global-anti-piracy-crackdown/

- 이러한 단속은 국제 지식재산 및 사이버범죄 대응 프로그램(International Computer Hacking and Intellectual Property)* 네트워크를 통한 정보 공유를 바탕으로 진행된 것으로 분석됨

  * 국제 지식재산 및 사이버범죄 대응 프로그램(International Computer Hacking and Intellectual Property): 미국 법무부가 해외 국가의 사이버범죄와 지식재산권 침해 대응 역량을 지원하기 위해 파견하는 법률 전문가 프로그램

• 동적 금지명령과 ISP 협력을 통한 신속 차단 확대

- 인도 델리고등법원(Delhi High Court)은 2026년 6월 3일 월드컵 불법 스트리밍에 대한 동적 금지명령(dynamic injunction)*을 내려, 이후 확인되는 신규 사이트를 별도의 소송 없이 추가로 차단할 수 있도록 조치함

- 싱가포르 고등법원(Singapore High Court) 또한 2026년 1월 저작권법에 근거해 프리미어리그 경기를 무단 중계한 47개 사이트의 차단을 명령했으며, 인터넷 서비스 제공사업자(Internet Service Provider, 이하 ISP)**가 이를 이행함

- 동남아시아에서는 프리미어리그 측의 요청 등을 통해 인도네시아 2만 1,000개 이상, 베트남 8,000개, 말레이시아 400개 이상의 관련 도메인이 누적 차단된 것으로 나타남[3)]

- 이는 법원의 최초 차단명령에 그치지 않고 신규, 우회 도메인을 지속적으로 차단 대상에 반영하는 한편, ISP의 기술적 이행을 통해 불법 스트리밍 접근을 제한하는 체계가 확대되고 있음을 보여줌

  * 동적 금지명령(dynamic injunction): 사전에 특정된 사이트뿐 아니라 이후 확인되는 미러·신규 사이트까지 범위에 포함해, 반복 소송 없이 추가 차단할 수 있도록 한 명령

  ** 인터넷 서비스 제공사업자(Internet Service Provider, ISP): 이용자에게 인터넷 접속과 데이터 전송 서비스를 제공하는 사업자로, 유선·무선 인터넷망 운영, IP 주소 제공, 회선 관리 등의 업무를 수행함

• 광고 수익과 온라인 중개자를 겨냥한 집행 확대

- 최근 민간 광고업계 단체인 신뢰가능광고그룹(Trustworthy Accountability Group, 이하 TAG)은 불법 도메인 배제 목록을 통해 스포츠 불법 중계 등 월드컵 관련 콘텐츠를 제공한 1,376개 사이트의 광고 수익을 차단함[4)]

- 이러한 조치는 도메인 차단이라는 단일 수단을 넘어, 불법 사이트의 수익 기반과 유통 경로까지 겨냥하는 방향으로 집행 지점이 확대됨을 시사함

[표1] 스포츠 불법 스트리밍 대응수단별 비교

사례	주요 수단	집행 대상	특징	한계
미국 법무부	도메인 압수	불법 스트리밍 도메인	국제 공조, 단계적 압수	해외 TLD, 대체 도메인 이동
국제 합동(중남미)	사이트 차단, 체포	도메인, URL, 운영조직	미주 다국가 공동 단속	변종, 복제 사이트 지속
인도 델리고등법원	동적 금지명령	미러, 신규 사이트	반복 소송 없이 추가 차단	권리자 신고, ISP 집행 의존
싱가포르 고등법원	ISP 차단명령	EPL 불법 스트리밍 사이트	기존 차단의 누적, 확대	미러 도메인 재생성
TAG	광고 수익 차단	불법 사이트 수익원	도메인 외 경제적 기반 겨냥	타 수익모델로 이동 가능

출처: 참고문헌 종합하여 재구성

3) The Online Citizen, "47 Piracy Websites to Be Blocked in Singapore Under Premier League Court Order", The Online Citizen, 2026.02.05., https://theonlinecitizen.com/2026/02/05/47-piracy-websites-to-be-blocked-in-singapore-under-premier-league-court-order

4) Ernesto Van der Sar, "FIFA World Cup Triggers a Global Anti-Piracy Crackdown (Updated)", TorrentFreak, 2026.07.19., https://torrentfreak.com/fifa-world-cup-triggers-a-global-anti-piracy-crackdown/

시사점: 지속적 우회에 대응하는 집행 과제

- **대체 인프라 재생성에 따른 차단 한계**

- 일부 사례와 같이, 특정 도메인을 차단하더라도 운영자가 대체 도메인과 해외 인프라로 이동하는 만큼 개별 차단의 효과는 일시적이며 단속 성과를 지속적으로 유지하기 어려운 것으로 분석됨
- 또한 불법 스트리밍 인프라가 여러 국가에 분산되어 있어 한 국가의 집행만으로는 관할권 밖 영역을 포괄하기 어려우며, 국제 공조의 실효성도 참여국의 범위에 좌우됨
- 이에 따라 대형 이벤트 시기의 집중 단속을 넘어 상시적 대응 체계를 갖추어야 한다는 논의가 제기되며, 아시아, 태평양 지역에서도 사이트 차단의 제도화가 정책 의제로 다뤄지는 양상임

- **저작권, 사이버보안, 소비자 보호의 연계 필요성**

- 불법 스트리밍은 저작권 침해에 그치지 않고 사기, 악성코드 및 개인정보 침해 등 이용자 피해와 결합될 수 있어, 저작권 보호를 소비자 보호 관점과 함께 다룰 필요성이 제기됨
- 이러한 위험에 대응하기 위해서는 권리자와 ISP, 광고업계, 사이버보안 기관이 함께 참여하는 협력 구조가 요구되며, 단일 주체의 대응만으로는 한계가 있음
- 아울러 단속 강화와 병행하여 불법 이용 수요를 지속시키는 합법 서비스의 가격 및 접근성 문제에 대한 점검도 함께 이루어질 필요가 있는 것으로 분석됨

참고문헌

- Ernesto Van der Sar, "FIFA World Cup Triggers a Global Anti-Piracy Crackdown (Updated)", TorrentFreak, 2026.07.19., https://torrentfreak.com/fifa-world-cup-triggers-a-global-anti-piracy-crackdown/
- AVIA, "State of Piracy Roundtable @ APOS", 2026.06.16., https://avia.org/all_events/piracy-roundtable-apos/
- Legal Maestros, "Delhi High Court Issues Injunction Against Rogue Websites Streaming FIFA World Cup 2026", Legal Maestros, 2026.06.05., https://legalmaestros.com/current-legal-update/inside-the-supreme-courts-new-draft-ai-framework-for-indian-courts/
- The Online Citizen, "47 Piracy Websites to Be Blocked in Singapore Under Premier League Court Order", The Online Citizen, 2026.02.05., https://theonlinecitizen.com/2026/02/05/47-piracy-websites-to-be-blocked-in-singapore-under-premier-league-court-order

주간 기술 동향

생성 과정에 워터마크를 심는 음악 생성용 워터마킹 기술, MusicMark

• AI 음악 생성의 확산과 출처 검증을 위한 생성형 워터마킹 기술의 부상

오픈소스 음악 생성 모델과 상업용 AI 음악 플랫폼이 빠르게 발전하면서, 이용자는 텍스트 프롬프트와 가사만으로 고품질 음악을 산출하고 배포할 수 있게 됐다. 이렇게 AI를 활용하여 산출된 음악이 대량으로 확산되면서, 이러한 음악이 어디에서 어떤 모델로 산출됐는지 확인하는 출처 검증의 중요성 또한 커지고 있다. 음악의 기원과 산출 모델을 추적할 수 없다면 창작물의 귀속과 책임 소재를 가리기 어렵기 때문이다. 이에 AI 산출 음악에 식별 정보를 삽입하는 워터마킹 기술이 대응 수단으로 주목받고 있다. 워터마킹은 음악에 식별용 신호를 심어 두었다가 이후 그 신호를 검출해 출처를 검증하는 기술이다.

그러나 지금까지 오디오 워터마킹 연구는 대부분 음성을 중심으로 이뤄져 왔으며, 이를 음악에 그대로 적용하기는 기술적인 어려움이 있었다. 일반적으로 음악은 음성보다 넓은 주파수 대역에 걸쳐있고, 여러 악기와 목소리가 멜로디, 화성, 리듬으로 얽힌 복잡한 구조를 지니기 때문이다. 더욱이 사람의 귀는 음악의 미세한 불협화음이나 선율의 어긋남에도 민감해, 워터마크로 인한 작은 왜곡조차 음질 저하로 이어질 수 있다. 결국 음악 워터마킹은 음악의 섬세한 음질을 그대로 유지하면서도, 압축이나 변환 같은 여러 처리를 견뎌내야 하는 과제를 안고 있다.

기존 오디오 워터마킹은 대부분 음악을 산출한 뒤 별도의 단계에서 들리지 않는 신호를 심는 사후 삽입 방식이다. 사후 삽입 방식에서 워터마크는 음악 내용과 결합되지 않은 외부 신호로 남아, 시간축 변형이나 압축 같은 변환에 쉽게 손상된다. 이 약점은 신경망 코덱 재합성 앞에서 특히 두드러진다. 신경망 오디오 코덱의 재합성 과정은 지각적으로 중요한 정보를 중심으로 복원하므로 미세한 후처리 신호가 손실될 수 있음. 또한, 음악을 산출하는 단계와 워터마크를 심는 단계가 분리돼 있어, 워터마킹을 생략하거나 우회하기 쉽다는 점도 출처 검증을 약화시킨다.

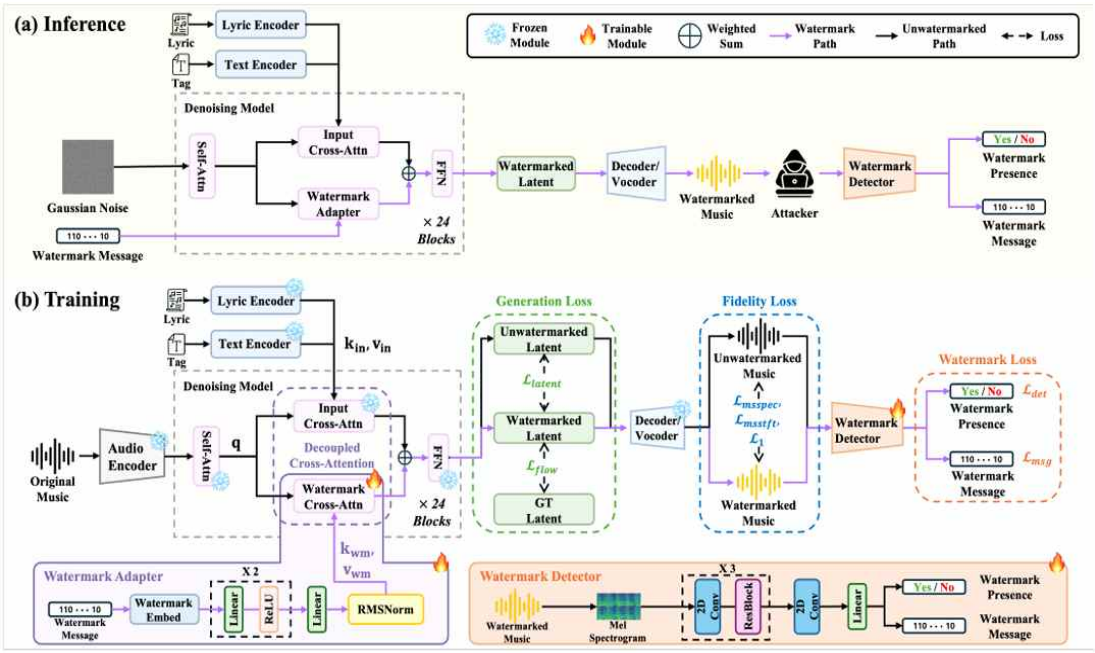
이러한 한계를 극복하기 위해 등장한 것이 음악을 위한 생성형 워터마킹 기술인 뮤직마크(MusicMark)다. 뮤직마크는 워터마크를 사후에 삽입하는 대신, 음악 산출 과정 중에 의미 잠재 공간에 메시지를 직접 심는다. 워터마크가 음악 내용의 일부로 통합되기 때문에, 신경망 코덱 재합성을 비롯한 다양한 공격에도 견디는 강인성을 확보하도록 설계됐다. 본 보고서에서는 이 기술이 어떤 원리로 워터마크를 삽입하고, 기존 사후 삽입 방식의 취약점을 어떻게 해결했는지 분석한다.

[사례] 음악 산출 과정에 워터마크를 심는 생성형 워터마킹, 뮤직마크

• 기존 사후 워터마킹의 한계와 뮤직마크의 등장 배경

- 기존 오디오 워터마킹은 대부분 음악을 산출한 뒤 들리지 않는 신호를 심는 사후 삽입 방식으로, 워터마크가 음악 내용과 결합되지 못한 채 음악 위에 외부 신호로만 남음
- 사후 삽입 방식은 시간축 변형, 주파수 변환, 압축 등 외부 공격에 의해 삽입 신호가 손상되기 쉬우며, 특히 의미적 내용만 남기고 잔여 신호를 버리는 신경망 코덱 재합성에서 워터마크가 함께 사라짐
- 음악 산출과 워터마킹이 서로 다른 단계로 나뉘어 있어, 워터마킹 단계를 생략하거나 우회할 경우 워터마크 자체가 존재하지 않게 되어 음악의 출처를 검증하기 어려움
- 뮤직마크는 이러한 사후 삽입 방식의 한계를 극복하기 위해 등장한 음악용 생성형 워터마킹 기술로, 워터마크를 음악 산출 과정에 통합하는 접근법을 사용함

[그림 1] 뮤직마크의 추론 과정(a)과 학습 과정(b) 개요



출처: Seohwan Yun 외 4인, "MusicMark: A Robust Generative Watermarking Framework for Music Generation", arXiv, 2026.07.13., https://arxiv.org/pdf/2607.11117

• 의미 잠재 공간에 워터마크를 심는 원리와 분리형 교차 어텐션 구조

- 뮤직마크는 워터마크로 넣을 메시지와 가사나 태그 같은 조건을 함께 입력받아, 음악을 산출하는 도중에 그 메시지를 음악 속에 심고, 완성된 음악에서 다시 메시지를 읽어내는 방식으로 작동함
- 이 과정에서 모델이 실제로 다루는 것은 완성된 음악 파일이 아니라, 음악의 의미와 특징을 압축해 담은 중간 데이터인 잠재 표현(latent representation)*이며, 뮤직마크는 바로 이 단계에 워터마크를 심음
- 잠재 표현 단계에 워터마크를 심으면 워터마크가 나중에 더해진 신호가 아니라 음악의 의미 잠재 표현 안에 녹아들기 때문에, 압축이나 코덱 변환 같은 공격에도 쉽게 지워지지 않도록 설계됨
- 워터마크를 심는 역할은 워터마크 어댑터가 맡으며, 어댑터는 여러 자리의 이진 메시지를 자리마다 학습된 벡터로 바꾼 뒤 이를 다듬어 하나의 워터마크 표현으로 만듦

- 워터마크는 교차 어텐션(cross-attention)**이라는 구조를 통해 잠재 표현에 주입되는데, 뮤직마크는 가사나 태그를 처리하는 기존 경로와 별개로 워터마크 전용 경로를 하나 더 두는 분리형 방식을 씀
- 경로를 둘로 나눈 덕분에 기존 음악 산출 품질은 그대로 두고 워터마크만 따로 주입할 수 있으며, 두 경로의 출력은 학습으로 조절되는 가중치에 따라 합쳐져 최종 잠재 표현에 반영됨

* 잠재 표현(latent representation): 신경망이 원본 데이터를 직접 다루는 대신, 그 데이터의 핵심 특징만 추려 압축해 담아둔 내부 표현
** 교차 어텐션(cross-attention): 신경망이 어떤 데이터를 처리할 때 그와 별개인 다른 정보를 함께 참조하도록 연결하는 구조

• 결합 학습 전략과 손실 함수 설계, 워터마크 검출 방식

- 학습 시 기반 음악 산출 모델은 그대로 고정한 채 워터마크 어댑터와 검출기만 함께 학습하며, 잠재 생성 손실, 충실도 손실(fidelity loss)*, 워터마크 손실 세 가지를 함께 써서 음질 유지와 견고한 워터마크 삽입을 동시에 달성함
- 잠재 생성 손실에 속하는 잠재 일관성 손실은 워터마크가 삽입된 잠재 표현을 같은 조건에서 워터마크 없이 만든 잠재 표현에 가깝게 유지하여, 워터마크 삽입에 따른 음질 저하를 억제함
- 충실도 손실은 워터마크가 삽입된 음악과 삽입되지 않은 음악의 파형이나 스펙트로그램의 차이를 비교해 청각적 품질을 보존하고, 워터마크 손실은 검출기가 워터마크 존재 여부와 메시지를 정확히 복원하도록 이끎
- 검출기는 완성된 음악을 멜 스펙트로그램(mel spectrogram)**으로 바꾼 뒤 잠재 표현과 시간축이 맞춰진 특징을 뽑아내, 각 시점마다 워터마크가 있는지와 어떤 메시지가 심겼는지를 예측함

* 충실도 손실(fidelity loss): 산출 결과가 기준이 되는 원본에 얼마나 가까운지를 수치로 나타내, 그 차이를 줄이도록 학습을 이끄는 기준 값
** 멜 스펙트로그램(mel spectrogram): 소리를 시간, 주파수, 세기로 나타내되, 주파수 축을 사람의 귀가 느끼는 방식에 맞춰 변환한 시각적 표현

• 실험 설계 및 성능 평가 결과

- 뮤직마크는 오픈소스 음악 산출 모델을 토대로 구현됐으며, 잡음 추가나 압축, 코덱 변환, 가수 목소리를 바꾸는 커버송 변환 등 20종의 공격 환경에서 성능을 평가함
- 공격이 없을 때는 워터마크 존재 여부와 담긴 메시지를 거의 완벽하게 찾아냈고, 다양한 공격이 가해진 뒤에도, 워터마크가 가장 쉽게 지워지는 신경망 코덱 재합성 상황에서 검출 성능이 안정적으로 유지됨
- 가수 목소리를 바꾸는 커버송 변환처럼 음악에 특화된 공격에도 워터마크를 안정적으로 지켜냈으며, 워터마크를 심어도 원래 음악과 음질 차이가 거의 없어 산출 품질이 우수하게 유지됨

결론 및 시사점

• 기술적 의의 및 향후 과제

- 뮤직마크는 워터마크를 음악의 의미 표현 안에 심음으로써, 사후 방식이 가장 취약했던 신경망 코덱 재합성이나 커버송 변환 같은 공격에도 워터마크가 함께 살아남게 한 점에서 의의가 큼
- 기반 산출 모델은 고정한 채 어댑터와 검출기만 학습하는 구조라 기존 모델의 성능을 해치지 않으면서 워터마킹을 더할 수 있어, AI가 산출한 음악의 출처와 저작권 귀속을 검증하는 AI 생성 음악의 출처 확인을 보조할 잠재적 수단으로 기대됨
- 다만 현재는 담을 수 있는 메시지 용량이 16비트로 제한적이고 검증도 텍스트 조건 기반 음악에 집중돼 있어, 더 큰 용량과 다양한 입력 조건으로의 확장이 향후 과제로 남음

참고문헌

• Seohwan Yun 외 4인, "MusicMark: A Robust Generative Watermarking Framework for Music Generation", arXiv, 2026.07.13., https://arxiv.org/pdf/2607.11117