저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

기업 유튜브 뮤직, 파트너 신고와 자체 탐지를 결합한 AI 생성 음원 라벨 도입

▶ 생성형 AI를 활용한 음악 제작이 늘면서 AI 생성 음원이 빠르게 증가했고, 청취자들은 듣는 음원에 AI가 사용됐는지 알려 달라고 요구해 왔다. 이에 유튜브 뮤직은 10월 14일 주로 생성형 AI로 제작한 음원의 재생 화면에 라벨을 표시하기 시작했다. 라벨은 파트너가 완전 생성형 AI, 부분 생성형 AI, 생성형 AI 미사용 세 항목으로 신고한 내용을 토대로 하되, 자체 탐지 결과로 신고를 보완하고 미신고 음원도 판별한다. 유튜브는 AI 사용만을 이유로 수익화를 제한하지 않아, 추천이나 저작권료 배분을 제한하는 타 음악 스트리밍 서비스와 달리 정보 공개에 중점을 둔다. 다만 항목의 해석이 파트너에게 맡겨져 있고 미신고 음원의 판별이 탐지 성능에 좌우되는 만큼, 라벨과 추천, 저작권료 지급의 관계를 설명하는 일이 과제로 남는다.

기업 디텍티프AI의 AI 생성 음성 탐지 모델 개방과 판별 결과의 활용 조건

▶ 짧은 녹음만으로 목소리를 복제할 수 있게 되면서, 화자 검증과 별도로 음성의 AI 생성 여부를 확인할 필요가 커졌다. 이에 디텍티프AI는 2026년 10월 AI 생성 음성 탐지 모델 알레테이아와 화자 검증 모델 에이도스를 API와 MCP로 개발자에게 개방했으며, 두 모델은 하나의 키로 함께 쓸 수 있다. 알레테이아는 전화 음성을 기준으로 생성 흔적을 찾아 판정과 구간별 점수를 반환하며, 포도노스 벤치마크에서 정확도 94.47%를 기록했다. 다만 실제 음성의 8.4%를 AI 생성으로 판정했고, 영어 음성에 맞춰 조정돼 전화 음질에서는 신뢰도가 낮아진다. 따라서 탐지 결과는 다른 검증과 결합한 보조 신호로 쓰고, 성능 수치는 자체 환경에서 다시 검증해야 한다.

산업 우크라이나 공영방송의 C2PA 도입과 보도 출처 증명 체계 구축

▶ 전쟁 중인 우크라이나에서 AI로 생성한 허위 콘텐츠가 대규모로 퍼지면서, 언론이 보도의 신뢰를 지키기 어려워지고 있다. 이에 우크라이나 공영방송 수스필네는 허위 콘텐츠를 모두 가려내는 대신 자사 보도임을 증명하는 쪽으로 대응 방향을 바꾸고 C2PA 표준을 도입했다. 수스필네는 소수 인력으로 6개월 만에 서명 도구를 구축해 C2PA 신뢰 목록에 올랐으며, 서명된 보도 사진은 누구나 허가 없이 검증할 수 있다. 다만 서명 대상은 아직 선별된 보도 사진에 한정되어 있다. 수스필네가 강조했듯 출처 증명은 업계가 함께 구축할 때 비로소 작동하므로, 앞으로는 서명 대상을 넓히는 동시에 업계 전반의 참여를 확산할 필요가 있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

기업 하쿠호도 사례로 본 AI 생성 이미지의 기존 저작물 유사성 확인 자동화

▶ 생성형 AI로 광고 제작물을 만드는 일이 늘면서 AI 생성 이미지가 기존 저작물과 비슷한지 확인하는 작업이 요구되고 있지만, 사람이 직접 확인하면 많은 시간이 소요되고 확인 담당자마다 판단이 달라지는 문제가 있다. 이에 하쿠호도는 그룹 내부에서 쓰던 이미지 유사도 체커를 2026년 10월 19일부터 고객사에 제공하기 시작했다. 이 도구는 생성형 AI로 만든 키워드로 웹에서 유사 이미지를 찾고, 의미와 구조 등 여러 관점의 유사도 점수를 목록으로 보여주며, 확인 과정을 리포트와 감사 로그로 남긴다. 다만 하쿠호도는 이 도구가 저작권 침해 여부를 판정하지 않으며 최종 권리 판단은 이용 기업이 내린다는 전제를 밝혀, 유사도 점수와 확인 이력을 기업의 제작 절차와 연결해 최종 판단까지 이어지도록 운영하는 일이 과제로 남는다.

산업 일본 방송업계의 C2PA 도입과 촬영부터 시청까지 이어지는 출처 증명의 확대

▶ 일본에서 AI로 생성한 허위 광고와 재해 관련 허위 이미지가 늘면서, 방송사들은 제보 영상의 진위 확인 부담과 보도 신뢰 하락이라는 문제에 직면하고 있다. 2026년 10월 도쿄에서 열린 미디어 프로버넌스 서밋에서는 카메라 렌즈부터 시청 화면까지 출처 정보를 유지하는 글래스 투 글래스가 강조되었다. NHK와 후지TV는 제작 과정에 C2PA를 적용했고, 후지TV는 광고 송출에도 이를 도입했으며, NTT도코모와 액세스는 출처 정보의 복원과 시청 화면 검증 기술을 선보였다. 다만 복원 기술은 실증 단계이고 브라우저는 시제품 단계이며, 'cr' 아이콘은 일본 시청자에게 큰 의미를 전달하지 못하고 있다. 이에 관련 기술의 상용화와 함께 시청자가 출처 정보의 의미를 쉽게 이해할 수 있는 표시 방식을 마련할 필요가 있다.

기술 주간 기술 동향

▶ 유럽연합이 AI법의 투명성 요건을 적용하기 시작하면서 AI 생성 텍스트의 출처를 확인하는 기술이 주목받고 있다. 그러나 기존 텍스트 워터마킹은 편집 이후 식별 정보를 복원하는 성능과 생성 품질을 동시에 확보하기 어렵고, 같은 앞 토큰이 반복될 때 생성 결과가 치우치거나 워터마크를 삽입할 기회가 줄어드는 한계가 있다. ClockMark는 후보 토큰들이 카운트다운 경주를 벌여 다음 토큰을 선택하는 방식으로, 이전 선택의 진행 상태를 저장하는 적응형 삽입과 현재 확률을 따르는 확률 의존 삽입을 결합한다. 그 결과 AI 모델의 원래 생성 확률을 유지하면서도, 토큰 삭제, 단어 교체, 문장 전체 바꿔 쓰기 등의 편집 이후에도 기존 기술보다 높은 식별 정보 복원 정확도를 기록한다.

유튜브 뮤직, 파트너 신고와 자체 탐지를 결합한 AI 생성 음원 라벨 도입

AI 생성 음원의 증가와 유튜브 뮤직의 라벨 도입

• AI 생성 음원의 증가에 따른 청취자의 정보 공개 요구 확대

- 생성형 AI(generative AI)*를 활용한 음악 제작 방식이 다양해지면서 스트리밍 서비스에서 유통되는 AI 생성 음원(AI-generated music)**이 빠르게 늘고 있음
- 이에 따라 청취자들은 자신이 듣는 음원에 AI가 사용됐는지 알려 달라고 요구해 왔으며, 구글의 동영상 플랫폼 유튜브(YouTube)도 청취자 의견을 통해 이러한 정보 공개의 필요성을 확인함[1)]

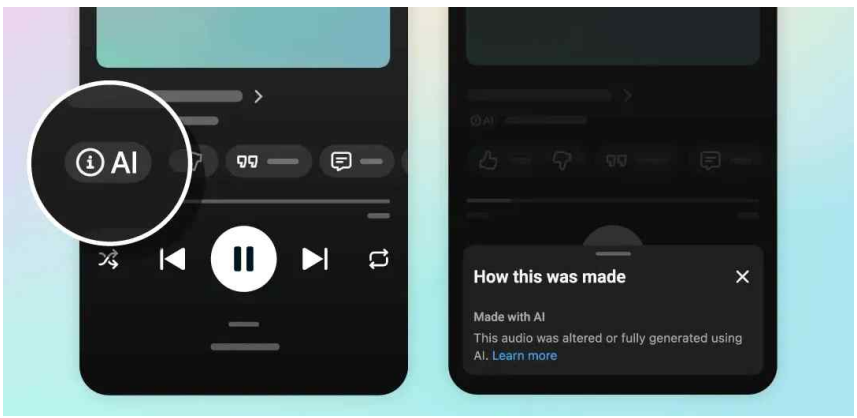
* 생성형 AI(generative AI): 학습한 데이터를 바탕으로 음원 등 새로운 콘텐츠를 만들어내는 AI

** AI 생성 음원(AI-generated music): 생성형 AI로 만든 음원

• 재생 화면에 AI 생성 여부를 표시하는 라벨 도입

- 이러한 요구에 대응해 구글의 음악 스트리밍 서비스 유튜브 뮤직(YouTube Music)은 2026년 10월 14일 작곡과 보컬을 비롯해 곡웁 핵심 요소를 생성형 AI로 제작한 음원에 라벨(label)을 붙이겠다고 발표함[1)]
- 라벨은 재생 화면의 곡명 아래에 'AI'로 표시되며, 이를 누르면 AI로 변형하거나 생성한 음원이라는 설명이 나타나 청취 중에도 AI 사용 여부를 확인할 수 있음

[그림 1] 유튜브 뮤직 재생 화면에 표시되는 AI 라벨과 안내 문구



출처: Abner Li, "YouTube Music rolling out AI labels in the Now Playing screen", 9to5Google, 2026.10.14., https://9to5google.com/2026/10/14/youtube-music-ai-label/

1) Murray Stassen, "YouTube Music starts labeling AI-generated songs, using partner disclosures and its own detection tools", Music Business Worldwide, 2026.10.14., https://www.musicbusinessworldwide.com/youtube-music-starts-labeling-ai-generated-songs-using-partner-disclosures-and-its-own-detection-tools/

파트너 신고와 자체 탐지를 결합한 라벨 부착 절차와 적용 범위

• 콘텐츠 전달 단계에서 이뤄지는 파트너의 생성형 AI 사용 범위 신고

- 라벨 부착의 기본 자료는 음반사나 유통사 등 유튜브에 음악 콘텐츠를 공급하는 파트너의 신고임. 파트너는 디지털 데이터 표준화(Digital Data Exchange, 이하 DDEX)*나 CSV(Comma-Separated Values)** 양식으로 유튜브에 콘텐츠를 전달할 때 생성형 AI 사용 여부를 함께 신고함
- 신고 항목은 AI 사용 범위에 따라 '완전 생성형 AI(Fully Gen AI)', '부분 생성형 AI(Partly Gen AI)', '생성형 AI 미사용(No Gen AI)'으로 나뉘며, 오토튠(Auto-Tune)***을 통한 음정 보정은 생성형 AI 미사용으로 분류됨
- 유튜브는 각 항목의 예시를 제시하되 DDEX가 정한 정의의 해석은 파트너에게 맡기고 있어, 개별 음원을 어느 항목으로 신고할지는 파트너가 판단함

* 디지털 데이터 표준화(Digital Data Exchange, DDEX): 음악 콘텐츠와 관련 정보를 주고받기 위한 표준

** CSV(Comma-Separated Values): 쉼표로 값을 구분해 데이터를 저장하는 텍스트 파일 형식

*** 오토튠(Auto-Tune): 음정 보정 소프트웨어

[표 1] 유튜브 뮤직 파트너 신고의 생성형 AI 사용 항목

항목	정의	유튜브 제시 예시
완전 생성형 AI (Fully Gen AI)	음원 전체를 AI로 생성	텍스트 프롬프트를 입력한 뒤 AI 음악 생성기에서 내려받은 음원
부분 생성형 AI (Partly Gen AI)	AI 생성 요소와 사람이 직접 제작한 요소를 결합	AI로 만든 베이스라인과 현악 파트에 직접 녹음한 보컬과 연주를 추가한 음원, 사람이 연주한 음원에 AI 모델로 제작한 뮤직비디오를 결합한 콘텐츠
생성형 AI 미사용 (No Gen AI)	생성형 AI를 사용하지 않음	오토튠 등 음정 보정, 자동 AI 마스터링

출처: Lauren Forristal, "YouTube Music launches new labels as AI-generated music floods streaming platforms", TechCrunch, 2026.10.14., https://techcrunch.com/2026/10/14/youtube-music-launches-new-labels-as-ai-generated-music-floods-streaming-platforms/

• 자체 탐지를 통한 파트너 신고의 보완과 미신고 음원의 판별

- 유튜브는 파트너의 신고 내용과 자체 탐지 도구의 분석 결과를 함께 검토해 라벨 부착 여부를 결정함
- 파트너가 생성형 AI 사용 여부를 신고하지 않은 경우에는 자체 탐지 결과를 바탕으로 완전 생성형 AI 또는 부분 생성형 AI로 분류해, 미신고 음원도 판별 대상에 포함함
- 이는 유튜브가 2026년 5월 영상 부분에 선제 도입했던 라벨 업데이트에서 창작자가 AI 사용을 밝히지 않아도 시스템이 사실적인 AI 사용을 상당 부분 감지하면 라벨을 자동으로 붙이도록 한 원칙을 음원으로 확대한 것임[2)]

2) Murray Stassen, "YouTube Music starts labeling AI-generated songs, using partner disclosures and its own detection tools", Music Business Worldwide, 2026.10.14., https://www.musicbusinessworldwide.com/youtube-music-starts-labeling-ai-generated-songs-using-partner-disclosures-and-its-own-detection-tools/

• 수익화 제한 없이 정보 공개에 한정된 라벨의 적용

- 유튜브는 AI 사용만을 이유로 음원의 수익화를 제한하지 않으며, 영상 라벨에 대해서도 라벨 자체는 추천 방식이나 수익 창출 자격에 영향을 주지 않는다고 밝힌 바 있음
- 이는 AI 생성 음원을 알고리즘 추천과 추천 플레이리스트에서 제외하는 프랑스 음악 스트리밍 플랫폼 디저(Deezer), 저작권료 배분 대상에서 제외하는 음악 스트리밍 플랫폼 타이달(TIDAL)과 달리 음원 제작에 AI가 사용됐음을 알리는 데 중점을 둔 방식임[3)]
- 다만 2026년 10월 14일 발표와 후속 공식 안내에서는 AI 음원 라벨 부착에 따른 추천이나 저작권료 지급 기준 및 생성형 AI를 부분적으로 활용한 음원의 라벨 표시 여부가 명시되지 않음

신고 해석과 탐지 성능에 좌우되는 라벨 방식의 과제

• 파트너의 해석 차이와 탐지 누락에 따른 판별의 일관성을 확보하기 위한 과제

- 파트너가 신고 기준을 해석하는 만큼 AI를 같은 정도로 사용한 음원도 서로 다른 항목으로 신고될 수 있어, 일관된 판별을 위해서는 해석 차이를 줄일 세부 기준이 필요함
- 미신고 음원은 자체 탐지에 의존하므로 AI 사용을 놓치면 라벨 없이 재생될 수 있으며, 잘못된 판별을 줄이는 탐지 성능과 오류를 바로잡는 절차도 중요함
- 부분 생성형 AI 음원의 라벨 표시 여부도 명확하지 않아, 청취자는 라벨이 없다는 사실만으로 해당 음원에 생성형 AI가 사용되지 않았다고 판단하기 어려움

• 플랫폼별 대응 차이에 따른 AI 생성 음원 노출과 수익의 분화

- 유튜브 뮤직은 정보 공개에 중점을 두는 반면 타 음악 스트리밍 서비스는 추천이나 저작권료 배분을 제한하는 경우도 있어, 같은 AI 생성 음원도 유통 플랫폼에 따라 노출 기회와 수익이 달라질 수 있음
- 유튜브 뮤직 역시 정보 공개 중심의 원칙이 실제 운영에 어떻게 적용되는지 확인할 수 있도록, 라벨 부착과 추천, 저작권료 지급의 관계를 구체적으로 설명할 필요가 있음

참고문헌

- Lauren Forristal, "YouTube Music launches new labels as AI-generated music floods streaming platforms", TechCrunch, 2026.10.14., https://techcrunch.com/2026/10/14/youtube-music-launches-new-labels-as-ai-generated-music-floods-streaming-platforms/
- Murray Stassen, "YouTube Music starts labeling AI-generated songs, using partner disclosures and its own detection tools", Music Business Worldwide, 2026.10.14., https://www.musicbusinessworldwide.com/youtube-music-starts-labeling-ai-generated-songs-using-partner-disclosures-and-its-own-detection-tools/
- Abner Li, "YouTube Music rolling out AI labels in the Now Playing screen", 9to5Google, 2026.10.14., https://9to5google.com/2026/10/14/youtube-music-ai-label/

3) Lauren Forristal, "YouTube Music launches new labels as AI-generated music floods streaming platforms", TechCrunch, 2026.10.14., https://techcrunch.com/2026/10/14/youtube-music-launches-new-labels-as-ai-generated-music-floods-streaming-platforms/

디텍티프AI의 AI 생성 음성 탐지 모델 개방과 판별 결과의 활용 조건

음성 복제 위협의 확대와 탐지 모델의 개발자 개방

• 짧은 녹음만으로 가능해진 음성 복제와 화자 검증의 한계

- AI로 음성을 생성하거나 복제하는 기술을 사용해 짧은 녹음만으로도 익숙한 목소리를 만들 수 있게 되면서, 은행과 상담 센터는 계좌 변경, 송금, 신원 확인 시 고객의 음성이 실제 음성인지 확인해야 하는 상황에 놓임
- 실제로 복제 음성은 상담 센터에 본인을 사칭해 전화하여 등록된 전화번호를 변경하거나, 임원이나 가족을 사칭해 급한 송금을 요구하는 데 악용될 수 있음. 비대면 계좌 개설 때 본인 확인을 통과하는 데도 쓰일 수 있어 피해 범위가 넓음
- 더욱이 정교한 복제 음성은 등록된 성문(voiceprint)*과 일치할 수 있어, 송금이나 인증 정보 재설정, 연락처 변경처럼 민감한 정보 요청 전에는 화자 검증(speaker verification)**과 별도로 음성의 AI 생성 여부를 확인해야 함

* 성문(voiceprint): 화자 검증을 위해 미리 등록해 둔 특정인의 목소리 기준 정보

** 화자 검증(speaker verification): 새로 입력된 음성이 미리 등록한 특정인의 목소리와 일치하는지 확인하는 기술

• 디텍티프AI의 AI 생성 음성 탐지 모델 개발자 개방

- 이에 AI 생성 음성 탐지 기업 디텍티프AI(DetectifAI)는 2026년 10월 19일 자사 탐지 모델을 API(Application Programming Interface)*와 MCP(Model Context Protocol)** 서버를 통해 개발자에게 개방함[1)]
- 개방 대상은 AI 생성 음성을 판별하는 알레테이아(Aletheia)와 화자 검증 모델 에이도스(Eidos)로, 두 모델을 하나의 API 키로 이용할 수 있어 개발자는 판별과 화자 검증을 함께 구성할 수 있음[2)]
- 아울러 개발자는 대시보드에서 API 키를 직접 발급받을 수 있어, API 키 발급, 음성 파일 전송, 판별 결과 확인의 세 단계를 거쳐 판별 결과를 받을 수 있음

* API(Application Programming Interface): 프로그램끼리 데이터와 기능을 주고받도록 정한 접속 규격

** MCP(Model Context Protocol): AI 어시스턴트가 외부 도구에 연결해 기능을 호출하도록 정한 통신 규격

1) Ryan Merket, "DetectifAI opens its voice deepfake detector to developers", RuntimeWire, 2026.10.19., https://runtimewire.com/article/detectifai-opens-voice-deepfake-detection-api
2) Detectif.ai, "Aletheia", 2026.10.22. 조회, https://platform.detectif.ai/docs/aletheia/

음성 입력부터 판정 반환까지 단계별로 이뤄지는 AI 생성 음성 판별

• API와 MCP를 통한 음성 입력 경로의 확대

- 디텍티프AI의 API는 알레테이아와 에이도스에 공통으로 WAV, MP3 등 6개 형식의 음성을 최대 100MB까지 받으며, 재시도할 때 동일한 요청 키를 붙이면 저장된 결과를 돌려줘 같은 음성을 중복 판별하지 않음[3)]
- 이에 더해 MCP 서버에서도 같은 API 키와 권한으로 동일한 기능을 쓸 수 있어, 개발자는 AI 어시스턴트에 파일의 AI 생성 여부를 자연어로 묻는 것만으로 판별을 실행할 수 있음
- 다만 MCP에서는 음성을 텍스트 형식(base64*)으로 변환해 보내야 해 상한이 25MB로 낮아지고, 녹음 길이도 4초 초과 120초 이하만 허용돼 긴 통화는 판별할 구간을 잘라 보내야 함[3)]

* base64: 이진 데이터를 64개의 ASCII 문자로 구성된 텍스트 형식으로 변환하는 인코딩 방식

• 생성 흔적 분석을 통한 AI 생성 음성 판별

- 디텍티프AI는 초당 8,000샘플로 전달되고 300~3,400Hz 대역만 남는 전화 음성을 설계 기준으로 삼아[4)], 고음질 녹음으로 학습한 탐지기가 의존하는 세부 정보가 사라진 신호에서도 판별하도록 함
- 판별 과정에서는 음성을 짧게 겹치는 구간으로 나눠 음높이와 숨소리를 함께 보는 스펙트럼 표현으로 바꾼 뒤, 지나치게 깨끗한 배음이나 숨소리 부재처럼 사람이 듣지 못하는 규칙성을 찾아 점수를 매김
- 다만 개발자 문서는 판별이 영어 음성에 맞춰 조정돼 있고 전화 음질에서는 깨끗한 녹음보다 신뢰도가 낮다고 안내해, 언어와 음질에 따라 성능이 달라질 수 있음

• 판정과 구간별 점수로 구성된 탐지 결과의 반환

- 판별 결과는 실제(real), AI 생성 가능성이 높은 가짜(fake)의 2단계 판정, 그리고 AI 생성 가능성을 0~1 사이로 나타낸 점수로 반환됨[3)]
- 아울러 겹치는 구간별 점수가 함께 반환돼 전체 녹음 중 검토가 필요한 구간을 가려낼 수 있고, 판정이 맞았는지 회신하는 기능으로 결과의 정오를 기록할 수도 있음
- 이러한 판별의 정확도는 음성 탐지 벤치마크 운영사 포도노스(Podonos)의 평가에서 23개 시스템 중 9위로 중간 수준이었으나[5)], 회사가 보고한 지연 시간은 이를 공개한 상용 시스템 중 가장 짧았음

[표 1] 포도노스 벤치마크에서 나타난 디텍티프AI의 성능 지표

구분	내용
평가 시점	2026년 9월
평가 규모	23개 시스템, 음성 파일 4,524개(디텍티프AI는 전체 파일 판정)
정확도	94.47%(23개 시스템 중 9위, 포도노스 채점)
탐지율	AI 생성 음성 파일의 97.3% 탐지
오탐률	실제 음성 파일의 8.4%를 AI 생성으로 판정
지연 시간	파일당 24ms(자체 보고, 시간 정보를 공개한 상용 11개 시스템 중 1위)

출처: Detectif.ai, "Detectif.ai: Audio Deepfake Detection and Speaker Verification", 2026.10.22. 조회, https://detectif.ai/

3) Detectif.ai, "Aletheia", 2026.10.22. 조회, https://platform.detectif.ai/docs/aletheia/
4) Detectif.ai, "Detectif.ai: Audio Deepfake Detection and Speaker Verification", 2026.10.22. 조회, https://detectif.ai/
5) Ryan Merket, "DetectifAI opens its voice deepfake detector to developers", RuntimeWire, 2026.10.19., https://runtimewire.com/article/detectifai-opens-voice-deepfake-detection-api

탐지 결과의 활용 범위와 성능 수치 해석의 과제

• 탐지 결과를 보조 신호로 활용하는 검증 절차 결합의 필요성

- 실제 음성의 일부를 AI 생성으로 잘못 판정하는 오탐이 있는 만큼 판정만으로 요청을 막으면 정상 이용자도 걸러질 수 있어, 디텍티프AI도 점수를 확정이 아닌 증거로 보고 중요한 결정에는 사람의 검토를 거치라고 안내함
- 더욱이 복제 음성은 화자 검증을 통과할 수 있고 판별 성능은 언어와 음질에 따라 달라지므로, 판별과 화자 검증, 사람의 검토를 결합해 각 수단의 공백을 서로 보완하는 절차가 필요함
- 이에 따라 개발자는 지연 시간이 짧은 탐지 도구를 자사 서비스에 연동하기 쉬워졌지만, 빠른 판별과 오탐에 따른 불필요한 경보 사이에서 균형을 따져야 하는 과제를 안게 됨

• 지표별로 달리 나타나는 성능 수치의 해석 과제

- AI 생성 음성만 대상으로 한 탐지율은 실제 음성의 판정 결과까지 반영한 정확도보다 높게 나타나, 실제 음성을 AI로 잘못 판정하는 경우를 함께 보지 않으면 탐지 성능을 과대평가할 수 있음
- 아울러 정확도는 포도노스가 비공개 정답으로 채점한 수치인 반면 지연 시간은 회사가 자체 보고한 수치로 하드웨어와 배포 환경에 따라 달라지므로, 수치마다 측정 주체와 조건을 구분해 읽어야 함
- 따라서 공개 벤치마크 수치는 비교 기준으로 활용하되, 언어와 음질처럼 성능을 좌우하는 조건은 도입하려는 서비스의 음성으로 다시 검증하는 과정이 필요함

참고문헌

- Ryan Merket, "DetectifAI opens its voice deepfake detector to developers", RuntimeWire, 2026.10.19., https://runtimewire.com/article/detectifai-opens-voice-deepfake-detection-api
- Detectif.ai, "Detectif.ai: Audio Deepfake Detection and Speaker Verification", 2026.10.22. 조회, https://detectif.ai/
- Detectif.ai, "Aletheia", 2026.10.22. 조회, https://platform.detectif.ai/docs/aletheia/
- Detectif.ai, "Detectif.ai Platform: audio authentication through one API", 2026.10.22. 조회, https://platform.detectif.ai/

우크라이나 공영방송의 C2PA 도입과 보도 출처 증명 체계 구축

전쟁 속 AI 생성 허위 콘텐츠의 확산과 언론 신뢰의 위기

• 전쟁과 함께 확산되는 AI 생성 허위 콘텐츠의 위협

- 우크라이나 공영방송 수스필네(Suspilne)에 따르면 군인, 가족, 징집 대상 연령 남성 등 평범한 우크라이나인을 가장해 절망감과 병역 회피에 대해 이야기하는 AI 생성 영상이 100만 회 넘게 조회되는 등, 허위 콘텐츠가 대규모로 확산되고 있음[1)]
- 더욱이 모병사무소 제복을 입은 남성이 외국인을 강제 징집하는 것처럼 조작한 AI 생성 사진 한 장은 다음 날까지 약 300건의 게시물로 퍼져나가[1)], 수스필네는 이를 물리적 전쟁과 함께 벌어지는 인지전(cognitive war)으로 규정함
- 아울러 우크라이나인의 72%가 텔레그램(Telegram)을 통해 뉴스를 접할 만큼 이용자가 언론사 웹사이트를 떠나는 가운데, 전 세계 뉴스 신뢰도도 37%까지 떨어져[1)] 언론이 보도의 신뢰를 지키기 어려워지고 있음

• 허위 콘텐츠 탐지에서 보도의 진본 증명으로의 대응 전환

- 이에 따라 수스필네는 모든 허위 콘텐츠를 잡아내는 대신 자사 보도의 진본성을 입증하는 쪽으로 대응 방향을 전환했으며, 허위 콘텐츠가 즉시 만들어지는 만큼 검증도 지체 없이 이뤄져야 한다는 입장임
- 이를 위해 C2PA(Coalition for Content Provenance and Authenticity)* 표준을 도입한 수스필네는 2026년 10월 13일 도쿄에서 열린 미디어 프로버넌스 서밋(Media Provenance Summit)에서 도입 경험을 공유하며, 출처 증명(provenance)**은 업계가 함께 구축할 때 작동한다고 강조함[2)]

* C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 기록하고 서명으로 이를 검증할 수 있게 하는 기술 표준

** 출처 증명(provenance): 데이터나 디지털 콘텐츠(텍스트, 이미지, 영상 등)가 어디서 시작되었고, 어떤 과정을 거쳐 수정·유통되었는지 전체 이력을 기록하고 검증하는 개념

1) Brendan Quinn, "From Kyiv to Tokyo: an “infrastructure for truth” moves from pilot to production at the Media Provenance Summit Tokyo", IPTC, 2026.10.14., https://iptc.org/news/media-provenance-summit-tokyo-2026-report/
2) Suspilne Ukraine, “Suspilne Ukraine team presents its experience implementing C2PA content provenance technology in Japan”, 2026.10.20., https://corp.suspilne.media/en/news/73783-suspilne-ukraine-team-presents-its-experience-implementing-c2pa-content-provenance-technology-in-japan/

구축부터 검증까지 이어지는 수스필네의 C2PA 서명 체계

• 소수 인력과 외부 협력으로 이룬 서명 도구의 구축

- 수스필네는 카메라 제조사 소니(Sony)의 C2PA 지원 카메라 2대와 엔지니어 2명, 제품 관리자 1명으로 6개월 만에 서명 도구를 구축해[1], 출처 증명 도입에 큰 투자가 필요하지 않음을 보여줌
- 아울러 영국 공영방송사 BBC의 국제 개발 자선 단체인 BBC 미디어 액션(BBC Media Action)의 보조금, 글로벌 인증기관 SSL.com의 인증서, 서명 콘텐츠 검증 플랫폼 픽셀스트림(Pixelstream)의 엔지니어링 지원이 더해져, 외부 협력이 소규모 팀의 구축을 지원함

• 공식 인증 절차를 거친 서명 신뢰성의 확보

- 이 도구가 C2PA 적합성 프로그램(C2PA Conformance Program)*을 통과하면서, 수스필네는 동유럽 공영방송사 최초로 C2PA 신뢰 목록(C2PA Trust List)**에 올라 공식 서명 주체로 인정받음
- 또한 서명 키(signing key)***는 하드웨어 보안 모듈(hardware security module)****에서 생성되고 외부로 반출되지 않아, 수스필네가 발행한 보도라는 서명의 신뢰성을 뒷받침함

* C2PA 적합성 프로그램(C2PA Conformance Program): C2PA가 서명 도구의 표준 요구사항 충족 여부를 확인하는 절차

** C2PA 신뢰 목록(C2PA Trust List): 적합성 프로그램을 통과한 서명 주체를 C2PA가 공식 등재하는 목록

*** 서명 키(signing key): 콘텐츠에 서명할 때 사용하는 암호 키로, 이 키로 만든 서명이 C2PA 신뢰 목록과 연결돼 발행 주체를 증명함

**** 하드웨어 보안 모듈(hardware security module): 서명 키를 생성하고 외부로 반출되지 않도록 보관하는 보안 장치

• 누구나 확인할 수 있는 보도 사진의 출처 검증

- 이처럼 서명된 보도 사진에는 수스필네가 발행했다는 신원 정보, 파일의 출처와 편집 이력, C2PA 신뢰 목록과 연결된 서명이 함께 담겨, 누구나 수스필네의 허가 없이 해당 사진을 검증할 수 있음
- 이 서명은 사진에 보이지 않는 디지털 형태로 붙어 있어 사람뿐 아니라 플랫폼과 자동화 시스템도 이를 검증할 수 있으며, 이용자에게는 'cr' 표시 창으로 그 내용이 나타남
- 'cr' 표시 창에 표시되는 정보를 콘텐츠 자격증명(Content Credentials)*이라 하며, 이용자가 이를 통해 출처를 확인할 수 있어 수스필네의 원본 사진은 이미 AI 생성 이미지와 구분되고 있음
- 다만 현재 서명 대상은 포격 피해 현장 사진 등 선별된 보도 사진에 한정되며, 수스필네는 스포츠, 문화, 교육, 아동 콘텐츠, 아카이브와 음성, 영상 등으로 적용 대상을 넓힐 계획임

* 콘텐츠 자격증명(Content Credentials): C2PA 표준에 따라 디지털 파일에 덧붙는 출처 정보로, 콘텐츠를 누가 발행했고 어떻게 만들어졌으며 어떤 변경을 거쳤는지를 담음

소규모 조직의 출처 증명 도입 가능성과 확산 과제

• 소규모 조직까지 넓어진 출처 증명 도입의 가능성

- 수스필네가 소수 인력과 외부 협력으로 6개월 만에 C2PA 신뢰 목록 등재까지 마친 만큼, 대규모 기술 조직을 갖추지 않은 언론사도 출처 증명을 도입할 수 있는 여건이 마련되고 있음
- 더욱이 수스필네가 강조했듯 출처 증명은 여러 언론사가 함께 구축해야 작동하므로, 개별 언론사의 도입을 넘어 업계 전반의 참여가 확산의 조건이 될 전망임

• **서명 대상 확대와 출처 정보 인지 확산의 과제**

- 다만 현재 서명은 선별된 보도 사진에 머물러 있어, 수스필네 보도 전반에서 진본을 증명하려면 음성과 영상, 아카이브까지 서명 대상을 넓히는 작업이 뒤따라야 함
- 아울러 출처 정보는 이용자가 'cr' 표시 창을 열어야 확인할 수 있으므로, 이용자가 이 표시의 의미를 알고 활용하도록 하는 일이 출처 증명의 효과를 좌우할 전망임

참고문헌

• Brendan Quinn, "From Kyiv to Tokyo: an "infrastructure for truth" moves from pilot to production at the Media Provenance Summit Tokyo", IPTC, 2026.10.14., https://iptc.org/news/media-provenance-summit-tokyo-2026-report/
• Suspilne Ukraine, "Suspilne Ukraine team presents its experience implementing C2PA content provenance technology in Japan", 2026.10.20., https://corp.suspilne.media/en/news/73783-suspilne-ukraine-team-presents-its-experience-implementing-c2pa-content-provenance-technology-in-japan/

하쿠호도 사례로 본 AI 생성 이미지의 기존 저작물 유사성 확인 자동화

생성형 AI 활용 확대와 기존 저작물 유사성 확인의 부담

• 생성형 AI 활용 광고 제작의 확대에 따른 유사성 확인 부담의 증가

- 생성형 AI를 활용한 광고 제작물과 콘텐츠 제작이 늘면서, AI 생성 이미지가 기존 저작물과 유사한지 검토하는 등 생성형 AI를 안전하게 쓰기 위한 확인 체계가 기업에 요구되고 있음
- 그러나 사람이 직접 유사성을 확인하면 상당한 작업 시간이 소요되고 확인 담당자마다 판단이 달라질 수 있어, 확인 절차를 효율화하고 일정한 기준으로 운영할 수단을 마련하는 일이 과제로 제기됨

• 그룹 내부 활용을 거친 하쿠호도의 이미지 유사도 체커 제공 확대

- 이러한 확인 부담을 줄이기 위해, 일본 광고 대행사 하쿠호도(Hakuhodo)는 2026년 10월 19일 AI 생성 이미지의 저작권 침해 위험을 자동 판정하는 이미지 유사도 체커(画像類似度チェッカー)를 고객사에 제공하기 시작함[1)]
- 이 도구는 모기업 하쿠호도 DY 그룹(Hakuhodo DY Group) 내부에서 생성형 AI로 광고 제작물을 만들 때 쓰던 확인 도구로, 고객사 측에서도 동일한 검수 작업의 효율화 요구가 잇따르면서 외부에 제공하게 됨

검색, 판정, 기록으로 이어지는 이미지 유사도 체커의 확인 절차

• 생성형 AI 기반 키워드 생성을 통한 유사 이미지 검색의 정확도 향상

- 이미지 유사도 체커는 확인 대상 이미지를 기준으로 웹에서 비슷한 이미지를 검색하고 유사성이 높은 이미지를 자동으로 찾아 보여줘, 확인 담당자가 직접 비교 대상을 찾는 작업을 줄여줌
- 더욱이 이미지의 특징을 생성형 AI로 분석해 관련 키워드를 자동으로 만들고 이를 검색에 반영해, 생성형 AI를 확인 대상이 아닌 확인 수단으로도 쓰면서 유사 이미지 검색의 정확도를 높임
- 아울러 하쿠호도는 고객사 제공에 맞춰 싱글 사인온(Single Sign-On)*을 포함한 인증 기능과 이용 이력을 기록하는 감사 로그 기능을 새로 강화해, 이용 기업이 도구에 접속하는 사람과 이용 내역을 직접 관리할 수 있도록 함

* 싱글 사인온(Single Sign-On): 한 번의 로그인으로 여러 서비스에 접속할 수 있도록 하는 인증 방식

1) 吹越友一, "博報堂、著作物との類似性を自動判定する「画像類似度チェッカー」の企業提供を開始", AI Watch, 2026.10.20., https://ai.watch.impress.co.jp/docs/news/2149474.html

• 복수 관점의 유사도 점수 표시를 통한 확인 결과의 비교 기준 마련

- 검색된 이미지는 의미와 구조 등 여러 관점에서 매긴 유사도 점수와 함께 표시되며, 확인 담당자는 유사도 점수를 목록 형태로 일괄 확인할 수 있어 여러 이미지의 확인 결과를 빠르게 비교할 수 있음
- 아울러 유사도는 복수의 알고리즘으로 판정돼, 확인 담당자마다 판단이 달라지기 쉬운 수작업 확인과 달리 같은 점수 기준을 토대로 확인 결과를 비교할 수 있는 근거를 제공함

• 리포트 출력과 감사 로그 저장을 통한 확인 과정의 기록

- 검색 결과와 판정 내용은 리포트 형식으로 출력할 수 있어, 확인 작업을 효율화하는 동시에 어떤 결과를 근거로 판단했는지를 문서로 보여주고 기록으로 남길 수 있음
- 또한 감사 로그에는 누가, 언제, 어떤 이미지를 확인했고 어떤 판정 결과가 나왔는지가 저장돼, 확인 결과뿐 아니라 확인 작업 자체의 이력도 추적할 수 있음
- 하쿠호도는 이러한 기록 기능이 기업의 AI 거버넌스(AI governance)*와 감사 대응을 지원한다고 보고, 이 도구를 단순한 유사 이미지 검색 도구가 아닌 기업의 생성형 AI 활용을 관리하는 기반으로 규정함

* AI 거버넌스(AI governance): 기업이 AI를 안전하게 활용하도록 관리 기준과 절차를 운영하는 체계

유사성 확인 자동화의 의미와 남은 과제

• 검색부터 기록까지 하나로 묶은 유사성 확인 절차의 표준화 가능성

- 이번 사례는 유사 이미지 검색과 유사도 판정, 기록을 하나의 도구에 결합해, 확인 담당자마다 달라지던 수작업 확인을 동일한 기준의 점수와 남겨진 이력으로 관리할 수 있음을 보여줌
- 특히 리포트와 감사 로그로 판단 과정이 남는 만큼, 유사성 확인을 확인 담당자 개인의 판단에 맡기지 않고 사후에 다시 점검할 수 있는 기업 내부의 절차로 운영할 수 있음

• 유사도 판정과 권리 판단의 구분에 따른 기업 확인 절차와의 연계 과제

- 다만 하쿠호도는 이 도구가 기존 이미지와의 유사성 확인을 돕는 수단으로 저작권 침해 여부를 직접 판정하지 않으며, 최종적인 권리 검토와 판단은 이용 기업의 책임이라는 점을 명확히 함[2)]
- 따라서 유사도 점수는 침해 여부의 결론이 아니라 기업이 추가로 확인할 이미지를 추리는 정보로 사용해야 하며, 점수를 바탕으로 최종 판단을 내리는 기업 내부의 확인 절차가 함께 운영돼야 함
- 하쿠호도도 향후 생성형 AI 이용 지침 마련과 운영 흐름 정비 지원까지 검토하고 있어, 자동화된 확인의 효과는 점수와 기록을 기업의 제작 절차에 어떻게 연결하는지에 따라 달라질 전망임

참고문헌

- 吹越友一, "博報堂、著作物との類似性を自動判定する「画像類似度チェッカー」の企業提供を開始", AI Watch, 2026.10.20., https://ai.watch.impress.co.jp/docs/news/2149474.html
- 博報堂, "博報堂、生成AI時代のクリエイティブガバナンスを支援する「画像類似度チェッカー」の企業提供を開始", 2026.10.19., https://www.hakuhodo.co.jp/news/newsrelease/267420/

2) 博報堂, "博報堂、生成AI時代のクリエイティブガバナンスを支援する「画像類似度チェッカー」の企業提供を開始", 2026.10.19., https://www.hakuhodo.co.jp/news/newsrelease/267420/

일본 방송업계의 C2PA 도입과 촬영부터 시청까지 이어지는 출처 증명의 확대

일본의 허위 콘텐츠 확산과 출처 증명 요구의 확대

• 허위 광고와 재해 허위 이미지 확산에 따른 기존 확인 방식의 한계

- 일본에서는 유명 기업과 공인을 사칭한 허위 광고가 소셜 미디어(social media)에서 늘고 있으며 그 상당수가 AI로 생성됨.
- 그러나 HTTPS(HyperText Transfer Protocol Secure)*는 웹 서버의 신원을 확인하고 통신을 보호하지만, 광고를 만든 광고주는 인증하지 못해 기존 보안 수단만으로는 대응하기 어려움
- 아울러 재해가 잦은 일본에서는 재해 때 허위 이미지가 늘어나며, 방송사는 시청자 제보 영상에 대해서도 AI 생성과 변조 여부를 확인해야 하는 부담을 안고 있음
- 더욱이 방송사들은 자사 보도를 변조한 허위 영상과 언론사를 사칭한 소셜 미디어 게시물의 확산을 우려하고 있어, 언론이 허위 콘텐츠에 속지 않는 것만으로는 보도의 신뢰를 지키기 어려움

* HTTPS(HyperText Transfer Protocol Secure): 콘텐츠를 전달하는 웹 서버를 인증하는 보안 통신 규격

• 촬영부터 시청까지 출처 정보를 잇는 C2PA 적용의 확대

- 이러한 가운데 2026년 10월 도쿄에서 열린 미디어 프로버넌스 서밋(Media Provenance Summit)에서는 방송사와 기술 기업 관계자 수백 명이 C2PA(Coalition for Content Provenance and Authenticity)* 기반 출처 증명(provenance)** 기술의 실제 적용 사례를 공유함[1)]
- 출처 정보가 처음 기록되는 촬영 단계에서는 소니(Sony), 캐논(Canon), 니콘(Nikon) 등 카메라 제조사가 이미 C2PA 지원 카메라를 출시해, 출처 증명을 적용할 수 있는 기반이 마련돼 있음
- 특히 서밋의 여러 발표에서 카메라 렌즈부터 스마트폰 등의 디스플레이까지 출처 정보를 일관되게 유지 및 전달하는 글래스 투 글래스(Glass to Glass)가 강조돼, 출처 증명의 적용 범위가 촬영 단계를 넘어 화면 표시 단계까지 넓어지고 있음[2)]

* C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 기록하고 서명으로 이를 검증할 수 있게 하는 기술 표준

** 출처 증명(provenance): 데이터나 디지털 콘텐츠(텍스트, 이미지, 영상 등)가 어디서 시작되었고, 어떤 과정을 거쳐 수정·유통되었는지 전체 이력을 기록하고 검증하는 개념

1) Brendan Quinn, "From Kyiv to Tokyo: an “infrastructure for truth” moves from pilot to production at the Media Provenance Summit Tokyo", IPTC, 2026.10.14., https://iptc.org/news/media-provenance-summit-tokyo-2026-report/
2) 小山安博, "フェイク情報に対抗する技術「C2PA」の現在地", デジカメ Watch, 2026.10.21., https://dc.watch.impress.co.jp/docs/column/2149298.html

제작부터 시청까지 단계별로 넓어지는 일본의 C2PA 적용

• 제보 영상 확인 부담을 덜기 위한 방송 제작 과정의 C2PA 적용

- 제보 영상 확인 부담을 덜기 위해 일본 공영방송 NHK는 프로그램 제작 과정에서 출처 정보를 확인하고 부여할 수 있도록 C2PA를 적용했으며, 이 개발 성과를 오픈소스로 공개할 계획임
- 일본 상업 방송사 후지TV(Fuji Television)도 같은 이유로 영상 라이브러리에서 C2PA 출처 정보를 표시하도록 해, 시청자 제보 영상의 출처를 제작 과정에서 바로 확인할 수 있게 함

• 어드레서블 광고 송출 과정에 결합된 후지TV의 출처 정보 삽입과 검증

- 아울러 후지TV는 지역별로 서로 다른 광고를 내보내는 어드레서블 광고(addressable advertising)의 수작업 심의가 어려워지자, 2025년 6월부터 광고 송출에 C2PA를 적용함[3)]
- 이에 따라 광고 영상이 광고주에서 광고 심의 기관으로 넘어갈 때 대표적인 동영상 포맷 및 스트리밍 규격인 MP4와 MPEG-DASH* 영상에 출처 정보가 삽입되고, TV 앱의 광고 재생 기능이 이를 추출 및 검증해 시청자 화면에 결과를 표시함
- 실제로 이 시스템은 2025년 6월 이후 120개 실제 TV 프로그램 방송에서 방송당 수십만 건의 광고 노출(ad impressions)**을 처리해[3)], 대규모 방송 환경에서도 작동함을 보여줌

* MPEG-DASH: 영상을 조각 단위로 나눠 인터넷으로 전송하는 스트리밍 형식

** 광고 노출(ad impressions): 디지털 광고와 마케팅에서 광고가 사용자의 화면에 표시된 횟수

• 유통과 시청 단계에서 출처 정보를 전달하기 위한 복원과 검증 기술의 개발

- 한편 소셜 미디어는 게시 과정에서 메타데이터(metadata)*를 삭제하는 경우가 많아, 일본 통신 기업 NTT도코모(NTT Docomo) 등은 기지국 위치 정보와 시각 정보를 워터마크(watermark)**와 결합해 지워진 출처 정보를 복원하는 기술을 실증하고 있음[4)]
- 시청 단계에서는 일본 임베디드 소프트웨어 기업 액세스(ACCESS)가 출처 증명 기능을 내장한 TV용 넷프론트 브라우저 BE(NetFront Browser BE)의 시제품을 공개해[3)], 시청자가 별도 검증 사이트를 찾지 않아도 출처 정보를 볼 수 있게 함
- 이 시제품은 출처 정보의 조회부터 검증과 화면 표시까지 브라우저 엔진 안에서 처리해, JPEG, PNG 이미지와 MPEG-DASH, HLS(HTTP Live Streaming)*** 스트리밍 영상을 검증하고 영상 재생 중에도 실시간으로 검증 결과를 아이콘으로 표시함
- 다만 후지TV는 콘텐츠 자격증명(Content Credentials)****을 나타내는 'cr' 아이콘이 일본 시청자 다수에게 큰 의미를 전달하지 못한다고 판단하고, HTTPS 자물쇠와 비상구 표시처럼 대중적인 인지도를 확보한 기호로 자리 잡은 사례를 참고하여 검토 중임

* 메타데이터(metadata): 콘텐츠의 촬영 일시, 기기, 위치, 편집 이력 등 데이터를 설명하기 위해 파일에 함께 저장되는 정보

** 워터마크(watermark): 콘텐츠에 식별 신호를 삽입하는 기술로, 메타데이터가 삭제됐을 때 출처 정보를 복원하는 기준으로 활용됨

*** HLS(HTTP Live Streaming): MPEG-DASH와 마찬가지로 영상을 작은 구간 단위로 나눠 인터넷으로 전송하는 스트리밍 형식

**** 콘텐츠 자격증명(Content Credentials): C2PA 표준에 따라 콘텐츠에 덧붙는 출처 정보로, 누가 만들고 어떤 기기로 촬영했으며 어떻게 편집했는지를 담음

3) Brendan Quinn, "From Kyiv to Tokyo: an “infrastructure for truth” moves from pilot to production at the Media Provenance Summit Tokyo", IPTC, 2026.10.14., https://iptc.org/news/media-provenance-summit-tokyo-2026-report/
4) 小山安博, "フェイク情報に対抗する技術「C2PA」の現在地", デジカメ Watch, 2026.10.21., https://dc.watch.impress.co.jp/docs/column/2149298.html

출처 증명 체계의 완성을 위한 과제

• 촬영부터 시청까지 이어지는 출처 증명 기술의 상용화 과제

- 카메라 제조사의 지원부터 방송 제작, 광고 송출, 시청 화면까지 단계마다 적용 사례가 나오면서, 일본에서는 글래스 투 글래스를 구현할 기술 요소가 하나씩 갖춰지고 있음
- 다만 NTT도코모의 복원 기술은 실증 단계에 있고 액세스의 브라우저도 아직 시제품 단계에 있어, 소셜 미디어를 거쳐 시청 화면까지 출처 정보가 이어지려면 이들 기술의 상용화가 뒤따라야 함

• 시청자가 이해할 수 있는 출처 표시 방식의 마련 필요성

- 출처 정보가 화면에 표시되더라도 시청자가 'cr' 아이콘의 뜻을 모르면 검증으로 이어지지 않으므로, 브라우저와 TV에 검증 기능을 내장하는 것과 함께 직관적인 표시 방식이 마련돼야 함
- 아울러 C2PA는 허위 콘텐츠 탐지 도구가 아니라 출처를 기록하는 수단인 만큼, 출처 정보의 검증이 콘텐츠에 담긴 사실관계의 진실성까지 보장하지 않는다는 점을 시청자가 이해하도록 돕는 일도 과제로 남음

참고문헌

- Brendan Quinn, "From Kyiv to Tokyo: an “infrastructure for truth” moves from pilot to production at the Media Provenance Summit Tokyo", IPTC, 2026.10.14., https://iptc.org/news/media-provenance-summit-tokyo-2026-report/
- 小山安博, "フェイク情報に対抗する技術「C2PA」の現在地", デジカメ Watch, 2026.10.21., https://dc.watch.impress.co.jp/docs/column/2149298.html

주간 기술 동향

편집에 강하고 생성 품질을 유지하는 텍스트 워터마킹 기술, ClockMark

• AI 생성 콘텐츠 투명성 요건 시행과 편집에 강한 텍스트 워터마킹 기술의 필요성

AI 생성 텍스트는 사람이 쓴 글과 구별하기가 점점 어려워지고 있다. 이에 따라 어떤 텍스트가 AI 모델에서 생성되었는지, 나아가 어떤 모델이나 사용자가 생성했는지 확인할 수 있는 수단에 대한 요구도 커지고 있다. 유럽연합은 2026년 8월 2일부터 AI법의 새로운 투명성 요건을 적용하기 시작했다. 이 요건은 AI가 생성하거나 변형한 콘텐츠를 탐지하기 쉽도록 기계가 판독할 수 있는 형태로 알리도록 요구한다.

같은 날부터 유럽연합 집행위원회는 범용 AI 모델 제공자가 AI법상 의무를 지키지 않을 경우 과징금을 부과할 수 있는 권한을 갖게 되었다. 범용 AI 모델 제공자는 기술 정보를 문서화하고 유럽연합 저작권법을 준수해야 하며, 투명성 의무 위반 역시 과징금이 부과될 수 있는 사안에 포함된다. 이에, 텍스트 워터마킹은 AI 모델이 텍스트를 생성하는 과정에서 워터마크를 삽입해 생성 여부를 확인하는 방법으로 주목받고 있다. AI 모델은 텍스트를 토큰이라는 작은 단위로 하나씩 이어 생성하므로, 워터마크 역시 토큰을 하나씩 선택하는 과정에서 삽입된다.

초기 텍스트 워터마킹 기술은 텍스트에 워터마크가 삽입되었는지 여부만 판별할 수 있어 구체적인 출처 추적에는 한계가 있었다. 이후 등장한 멀티비트 워터마킹은 모델이나 사용자를 구분하는 식별 정보를 삽입하고 나중에 복원할 수 있게 하였다. 그러나 텍스트 일부가 삭제되거나 다른 단어로 바뀌는 편집을 거친 뒤에도 식별 정보를 복원할 수 있게 하면서, 동시에 생성 품질을 유지하는 것은 여전히 어려운 과제이다. 특히 앞 토큰을 기준으로 워터마크를 삽입하는 방식은 동일한 직전 토큰이 반복해서 등장할 때 생성 결과가 한쪽으로 치우치거나, 이를 피하려 건너뛰면 워터마크를 삽입할 기회가 줄어드는 문제를 안고 있다.

이러한 배경에서 토큰 생성 과정을 고려한 텍스트 워터마킹 기술인 ClockMark가 제안되었다. ClockMark는 이전 토큰 선택의 진행 상태를 저장해 다음 선택에 활용하고, AI 모델이 현재 각 토큰에 부여한 확률에 따라 토큰 선택과 상태 갱신을 조절한다. 또한 워터마크가 삽입된 모든 토큰에 식별 정보 전체를 반영하여, 편집 이후 남은 부분만으로도 식별 정보를 복원할 수 있도록 설계되었다. 이 기술은 AI 모델의 원래 생성 품질을 유지하면서도 텍스트가 편집되었을 때 식별 정보를 복원하는 성능을 높여, AI 생성 텍스트의 출처 추적을 위한 실용적 가능성을 제시한다.

[사례] 이전 선택의 진행 상태를 활용하는 텍스트 워터마킹 기술, ClockMark

• 기존 텍스트 워터마킹의 한계와 ClockMark의 등장 배경

- 멀티비트 워터마킹(multi-bit watermarking)*의 일부 방식은 토큰(token)**마다 식별 정보의 특정 부분만 삽입하므로, 편집으로 해당 토큰이 사라지면 그 부분의 식별 정보를 복원하기 어려움
- AI 모델은 다음 토큰 후보마다 선택될 확률을 매긴 뒤 하나를 고르는데, 워터마크를 강하게 삽입하려고 이 확률을 직접 바꾸면 생성 품질이 떨어질 수 있음
- 앞 토큰을 기준으로 워터마크를 삽입하는 방식은 토큰 하나만 편집되어도 뒤 토큰의 워터마크까지 손상될 수 있음. 기준을 앞 토큰 하나로 좁히면 손상은 줄지만, 같은 앞 토큰이 자주 반복되어 생성 결과가 치우치거나 워터마크를 삽입할 기회가 줄어드는 문제가 생김
- ClockMark는 토큰을 고를 때마다 이전 선택의 진행 상태를 저장해 이어서 활용하고, AI 모델이 매긴 현재 확률을 그대로 따라 토큰을 고르는 방식으로 이러한 한계를 해결하고자 함

* 멀티비트 워터마킹(multi-bit watermarking): 워터마크의 삽입 여부만 판별하는 방식과 달리, 모델이나 사용자를 구분하는 식별 정보를 워터마크로 삽입하여 나중에 복원할 수 있게 하는 기술

** 토큰(token): AI 모델이 텍스트를 처리하고 생성하는 기본 단위로, 단어 또는 단어의 일부에 해당함. AI 모델은 토큰을 하나씩 선택해 이어 붙이는 방식으로 텍스트를 생성함

• 식별 정보 분산과 카운트다운 경주 방식

- ClockMark는 식별 정보 전체를 하나로 묶어 각 후보 토큰의 점수에 반영하므로, 토큰마다 식별 정보의 특정 부분을 따로 배정할 필요가 없음
- 각 후보 토큰에는 비밀 키(secret key), 바로 앞 토큰, 후보 토큰 자체를 바탕으로 무작위 값이 주어지며, 이 값은 같은 비밀 키를 가진 쪽만 똑같이 다시 만들 수 있음
- 후보 토큰의 점수는 이 무작위 값이 식별 정보와 얼마나 같은 방향을 가리키는지로 계산되며, 점수가 높을수록 해당 후보 토큰에 짧은 카운트다운(countdown)이 주어짐
- 이후 모든 후보 토큰의 카운트다운이 동시에 줄어드는 일종의 경주가 진행되며, 카운트다운이 가장 먼저 0에 도달한 후보 토큰이 다음 토큰으로 선택됨
- 카운트다운이 줄어드는 속도는 AI 모델이 각 후보 토큰에 매긴 확률에 따라 정해짐. 따라서 식별 정보와 방향이 맞는 후보 토큰이 선택되기 쉬워지며, AI 모델의 원래 확률이 토큰 선택에 그대로 반영됨

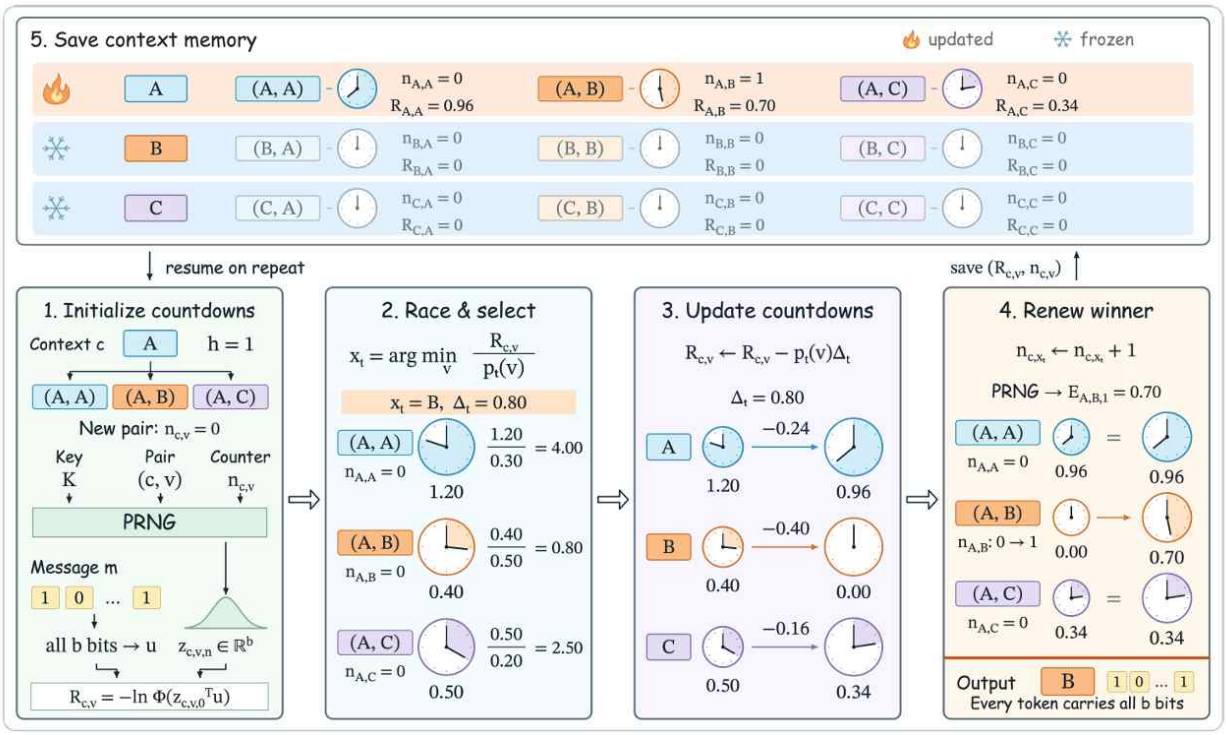
• 진행 상태 저장과 확률에 따른 토큰 선택

- 경주가 끝나면 모든 후보 토큰의 카운트다운이 같은 시간 동안 각자의 속도로 줄어듦. 이후 새 카운트다운은 선택된 토큰에만 주어지고 선택되지 않은 후보 토큰은 처음부터 다시 시작하지 않고 남은 카운트다운을 다음 경주에 이어서 사용함
- 남은 카운트다운은 바로 앞 토큰별로 진행 상태로 저장되며, 동일한 토큰이 다시 등장하면 처음부터 시작하지 않고 저장된 상태에서 경주를 이어가는데, 이를 적응형 삽입(adaptive embedding)이라 함
- 카운트다운이 줄어드는 속도는 매 경주마다 AI 모델의 현재 확률로 다시 정해지며, 현재 확률이 토큰 선택과 진행 상태 갱신을 함께 결정하는 방식을 확률 의존 삽입(host-dependent embedding)이라 함
- 두 방식을 결합해 바로 앞 토큰 하나만 기준으로 삼으면서도 반복 구간에서 워터마크를 계속 삽입할 수 있어, 편집으로 인한 워터마크 손상 범위가 줄어듦. 또한 AI 모델의 원래 생성 확률이 유지되는 것으로 수학적으로 증명됨

• 워터마크 복원과 탐지 방식

- 워터마크 복원 과정에서는 AI 모델 없이 비밀 키와 전달받은 텍스트만 사용하며, 텍스트에서 서로 이웃한 두 토큰으로 이루어진 토큰 쌍을 모두 추출한 뒤 각 토큰 쌍의 무작위 값을 다시 만듦
- 같은 토큰 쌍은 텍스트 어디에 등장하든 같은 무작위 값을 가지므로, 편집으로 일부가 사라져도 다른 위치에 남아 있는 같은 토큰 쌍으로 무작위 값을 다시 만들 수 있음
- 모든 토큰 쌍의 무작위 값을 하나로 더한 뒤, 그 결과가 가리키는 방향에서 식별 정보를 복원하고 같은 결과의 크기로 워터마크가 삽입되었는지 탐지함
- 탐지 기준은 텍스트 길이와 관계없이, 워터마크가 없는 텍스트를 워터마크가 삽입된 텍스트로 잘못 탐지할 가능성을 미리 정한 수준 이하로 통제하도록 설계됨

[그림 1] ClockMark의 워터마크 삽입 과정



출처: Yupeng Qiu 외 2인, "ClockMark: Adaptive and Host-Dependent Watermarking for Large Language Models", arXiv, 2026.10.19., https://arxiv.org/pdf/2610.27079

• 실험 설계 및 성능 결과

- 연구진은 서로 다른 두 종류의 공개 AI 모델로 긴 답변형 질의응답 텍스트와 뉴스형 웹 텍스트를 생성하고, ClockMark를 기존 멀티비트 워터마킹 기술 12종과 비교함
- 편집 상황은 토큰 일부 삭제, 일부 단어를 비슷한 단어로 교체, 다른 AI 모델을 활용한 문장 전체 바꿔 쓰기의 세 가지 방식으로 구성됨
- 편집 전 텍스트에서 ClockMark는 비교 기술 중 가장 높은 식별 정보 복원 정확도와 최상위 수준의 탐지 성능을 기록했으며, 식별 정보가 길어질수록 기존 기술과의 격차가 커지는 것으로 확인됨. 생성 품질은 워터마크가 없는 텍스트와 비슷한 수준으로 유지됨
- 편집 후에도 ClockMark는 논문에서 평가한 편집 조건과 비교 기술 범위에서 가장 높은 복원 정확도를 기록했으며, 특히 문장 전체 바꿔 쓰기에서 기존 기술과의 차이가 두드러짐. 다만 텍스트의 절반을 비슷한 단어로 교체하는 강한 편집에서는 모든 기술의 탐지 성능이 크게 낮아짐

결론 및 시사점

• 기술적 의의 및 산업적 활용 가능성

- ClockMark는 이전 선택의 진행 상태를 저장하는 적응형 삽입과 현재 확률을 따르는 확률 의존 삽입을 결합하여, 생성 품질을 유지하면서도 편집 이후 식별 정보를 복원하는 성능을 높일 수 있음을 확인함
- 생성 품질 유지는 상용 AI 서비스에 워터마킹을 적용할 때 필수 조건이며, 일부 상용 AI 모델도 원래 생성 확률을 유지하는 워터마킹 방식을 사용하고 있어 ClockMark의 설계 방향은 실제 서비스 환경과 부합함
- 식별 정보 복원과 워터마크 탐지에는 AI 모델 없이 비밀 키와 텍스트만 필요하므로, 텍스트를 생성한 AI 모델에 접근할 수 없는 상황에서도 출처를 확인할 수 있음
- 다만 텍스트 상당 부분이 비슷한 단어로 교체되는 강한 편집에서는 탐지 성능이 크게 낮아지는 것으로 확인되어, 강한 편집 상황에서의 탐지는 여전히 해결해야 할 과제로 남아 있음

참고문헌

• Yupeng Qiu 외 2인, “ClockMark: Adaptive and Host-Dependent Watermarking for Large Language Models”, arXiv, 2026.10.19., https://arxiv.org/pdf/2610.27079
• Forklog, “OpenAI to Introduce Text Watermarking in the EU”, 2026.10.06., https://forklog.com/en/openai-to-introduce-text-watermarking-in-the-eu/