기업 앤트로픽, 클로드 생성 파일의 출처를 확인하는 무료 검증 도구 공개

▶ 앤트로픽은 EU 인공지능법의 기계 판독 가능 표시 의무화에 대응해 클로드가 생성하거나 처리한 파일에 C2PA 기반 콘텐츠 자격증명을 삽입하고, 이를 누구나 확인할 수 있는 무료 도구인 콘텐츠 체커를 공개했다. 콘텐츠 체커는 이용자 브라우저에서 자격증명 유무만 판독하며, 파일 내용은 분석하지 않고 외부로 전송하지도 않는다. 그러나 파일 변환이나 메타데이터 삭제 과정에서 자격증명이 사라질 수 있어, 표시가 없다는 결과만으로 AI를 사용하지 않았다고 단정하기는 어렵다. 이에 따라 향후에는 유통 과정에서 표시가 유지되도록 하고, 사업자별 확인 결과를 한곳에서 교차 대조할 수 있는 체계를 마련할 필요가 있다.

산업 벨기에, 불법 웹사이트 차단의 한계에 따른 도메인 등록 정보의 확보

▶ 벨기에는 지난 2년간 불법 웹사이트의 도메인 1,500개 이상을 차단했으나, 불법 웹사이트가 주소를 빠르게 바꿔 도메인 차단만으로는 집행 효과를 거두기 어려웠다. 이에 BAPO는 2026년 8월 사이트 접속을 막는 대신 도메인 등록을 매개한 중개자가 운영자 정보를 제출하도록 하는 결정 5건을 발부했다. 등록대행사 대상 4건은 이름과 연락처, 결제 수단, 접속 로그 등 7개 범주를 요구했다. 반면 등록소 대상 1건은 등록자 정보와 네임서버 변경 이력만 요구해, 확보되는 정보의 범위가 중개자 유형에 따라 갈렸다. 다만 확인 대상이 중개자가 보관해 온 기록에 그치고 운영자도 신원을 여러 단계 은폐한 것으로 추정되는 만큼, 추적의 실효성은 중개자별 데이터 보관 항목과 기간에 따라 달라질 것으로 보인다.

기업 TPN의 픽스 잇 오어 리스크 잇 공개로 본 콘텐츠 보안의 실행 격차

▶ 콘텐츠 보안 사고는 새로운 공격 기법보다 이미 알려진 취약점을 방치하는 데서 기인한다. TPN이 2026년 처음 발간한 STAR 보고서는 평가가 주기적으로 이어졌는데도 미준수율이 3년간 개선되지 않았고 확인된 문제의 70%가 해결되지 않은 채 남았음을 확인했다. 이에 TPN은 2026년 9월 IBC2026에서 비영리 캠페인 픽스 잇 오어 리스크 잇을 공개해 평가 중심 대응을 이행 지원으로 넓혔다. 픽스잇 툴킷은 실무 지침과 정책 템플릿, 교육, 격차 분석을 무료로 제공하고 각 조치를 노력 수준에 따라 정리해, 비용이 들지 않는 조치부터 처리하도록 안내한다. 다만 이행 여부는 조직이 자체적으로 정하고 툴킷에는 적용한 보안 통제가 이후에도 작동하는지 확인할 수단이 담겨 있지 않다. 상시 모니터링과 자동화, 책임 소재의 명확화를 함께 갖추는 일이 남은 과제다.

산업 팬그램의 서브스택 적용 사례로 본 AI 개입 정도 판별 방식의 변화

▶ AI가 생성한 글과 이미지가 채용 지원서와 보험 청구에까지 흘러들고 사람이 쓴 뒤 AI가 다듬은 혼합 콘텐츠가 늘면서, AI와 사람 둘 중 하나로 나누는 판별만으로는 대응하기 어려워졌다. 이에 팬그램은 2026년 7월 개입 비율을 재는 텍스트 판별 모델과 이미지 판별 모델을 공개했다. 사람이 쓴 문서에 어조를 맞춘 합성 대조본을 대조하는 방식이어서 사전 표식이 없는 콘텐츠에도 적용된다. 서브스택은 이 모델을 연동해 독자가 게시물의 사람과 AI 작성 비율을 확인하도록 했다. 다만 사람이 다시 쓴 문장을 AI로 표시하는 오판이 확인돼, 판별 결과 표시와 정정 절차를 함께 갖출 필요가 있다.

산업 프렌드MTS와 애자일TV의 사전 통합으로 본 서버 측 워터마킹 도입의 간소화

▶ 라이브 스포츠의 불법 유통에 대응하는 서버 측 워터마킹은 사업자마다 워터마킹 백엔드와 영상 전송 설비를 잇는 개별 통합 작업이 필요해 도입이 늦어져 왔다. 이에 프렌드MTS와 애자일TV는 2026년 9월 3일 그 연결 작업을 미리 끝낸 첫 통합을 마쳤다고 발표했다. 프렌드MTS가 계정별 식별 신호를 배정하고, 애자일TV의 CDN이 세션 토큰을 바탕으로 가입자마다 다른 스트림을 만들어 전달하는 방식이다. 이로써 불법 스트림에서 식별 신호를 해독해 유출 계정을 특정하고 실시간으로 차단할 수 있게 되었다. 다만 그 효과는 CDN 구간과 선별된 프리미엄 콘텐츠에 한정되고 헤드엔드와 비즈니스 로직 연동은 사업자별 과제로 남아, 이 방식이 자리 잡으려면 공급업체 간 조합 확대와 남은 구간의 부담 완화가 함께 필요하다.

기술 주간 기술 동향

▶ 텍스트 지시문 하나로 사진을 편집하는 도구가 확산되면서, 원본 대부분이 남는 편집본이 진짜 사진으로 오인되는 문제가 커지고 있다. 촬영 단계에서 진위 정보를 기록하는 방식은 촬영 시점과 기기만 알려줄 뿐이고, 기존 워터마킹은 미리 정해둔 식별 번호만 삽입해 편집 내용을 설명하지 못하며 별도 모듈로 삽입되어 제거되기도 쉽다. 앤젤핑거프린트는 편집 프롬프트의 임베딩을 워터마크로 삽입하는 기술이다. 저순위 적응을 통해 확산 모델의 가중치에 직접 삽입하므로 제거할 별도 모듈이 없다. 원본 모델의 예측값을 기준으로 삼아 편집 품질을 유지하고, 중간 주파수 대역에만 워터마크를 삽입해 비가시성과 압축 강건성을 확보한다. 후보 200개 중 실제 사용된 편집 프롬프트를 86% 정확도로 찾아냈다.

앤트로픽, 클로드 생성 파일의 출처를 확인하는 무료 검증 도구 공개

AI 생성 콘텐츠 표시 의무화와 사업자별 확인 도구 제공 확대

• 기계 판독 가능 표시 의무화에 따른 AI 생성 콘텐츠 표시 도입

- EU 인공지능법(EU AI Act) 제50조 제2항이 2026년 8월 2일부터 AI 생성 콘텐츠에 기계 판독 가능 표시(machine-readable labeling)*를 의무화하면서, AI 사업자는 콘텐츠 생성 단계에서 파일에 식별 정보를 삽입하는 기능을 갖추게 됨[1)]
- 이러한 흐름에 맞춰 AI 개발사 앤트로픽(Anthropic)은 2026년 7월 약 190개 기업이 서명한 'AI 콘텐츠 투명성 관련 EU 자율 행동강령(EU Code of Conduct for Transparency of AI Content)'에 참여해 표시 의무 이행에 나섬
- 이에 따라 앤트로픽은 자사 AI 챗봇 클로드(Claude)의 신규 모델 페이블 5.1(Fable 5.1)과 미토스 5.1(Mythos 5.1)부터 표시 기능을 적용했고, 해당 표시는 웹 애플리케이션, API, 클라우드 파트너 등 모든 접근 경로에서 이용자의 별도 설정 없이 자동으로 삽입됨
- 아울러 앤트로픽에 따르면 이미 제공 중인 구 모델에도 표시 기능을 순차적으로 추가하고 있어, 신규 모델에서 시작한 표시 적용 범위를 클로드가 생성, 처리하는 파일 전반으로 넓혀 가는 중임

* 기계 판독 가능 표시(machine-readable labeling): 사람의 눈이 아닌 소프트웨어가 읽어 AI 생성 여부를 식별할 수 있도록 콘텐츠에 삽입하는 표시

• 앤트로픽의 무료 도구 '콘텐츠 체커' 공개

- 앤트로픽은 표시 기능 도입에 이어 2026년 9월 파일에 클로드의 표시가 포함돼 있는지 누구나 확인할 수 있는 무료 검증 도구인 '콘텐츠 체커(Content Checker)'를 자사 웹사이트에 공개해, 표시를 삽입하는 데 그치지 않고 이용자가 이를 직접 확인할 수 있는 수단까지 마련함[2)]
- 콘텐츠 체커는 JPG, PNG, SVG 등 이미지 10종과 MP4 등 영상 3종, MP3 등 오디오 4종의 파일을 100MB 이하 크기까지 확인할 수 있음. 계정 등록이나 소속 증명을 요구하지 않아 특정 기관이나 이용자에 한정하지 않고 누구나 이용하도록 개방함

1) Brian Beckmann, "Anthropic launches free checker tool for Claude watermarks", Beckmann, 2026.09.04., https://beckmann.ai/en/security/2026-09/anthropic-claude-watermark-checker
2) Isaiah Richard, "Anthropic's Claude Content Checker Tool Is Now Available—Here's How to Use the Detector", iTech Post, 2026.09.02., https://www.itechpost.com/articles/237212/20260902/anthropics-claude-content-checker-tool-now-availableheres-how-use-detector.htm

[표 1] 앤트로픽 콘텐츠 체커의 주요 내용

구분	내용
공개 시점	2026년 9월
이용 경로	claude.com/check-files, 계정 등록 또는 소속 증명 불필요
지원 형식	- 이미지 10종(JPG, PNG, GIF, WEBP, TIFF, HEIC, AVIF, SVG, DNG, JXL) - 영상 3종(MP4, MOV, AVI) - 오디오 4종(WAV, MP3, M4A, FLAC)
확인 가능 파일 크기	파일당 100MB 이하
처리 위치	이용자 브라우저 내 처리
판독 대상	C2PA 콘텐츠 자격증명 유무

출처: Anthropic, "Check if a file was made with Claude", claude.com, 2026.09.10. 조회, https://claude.com/check-files

콘텐츠 체커의 작동 방식과 확인 범위의 제한

• 개방 표준 C2PA를 따르는 콘텐츠 자격증명의 삽입

- 클로드는 차트 등 지원 형식의 파일을 이용자가 내려받을 때 해당 파일을 자신이 만들거나 처리했을 수 있다는 정보를 암호학적으로 서명해 메타데이터(metadata)*에 첨부함. 콘텐츠 체커는 이렇게 삽입된 콘텐츠 자격증명(Content Credential)**을 판독함

- 이 자격증명은 개방형 산업 표준 C2PA(Coalition for Content Provenance and Authenticity)***를 따름. 이에 따라 앤트로픽이 제공하는 콘텐츠 체커뿐 아니라 C2PA를 지원하는 다른 도구에서도 해당 정보를 확인할 수 있음

* 메타데이터(metadata): 파일의 내용과 별도로 파일에 관한 정보를 기록한 데이터

** 콘텐츠 자격증명(Content Credential): 파일의 출처와 처리 이력을 기록해 파일 메타데이터에 결합하는 암호학적으로 서명된 정보

*** C2PA(Coalition for Content Provenance and Authenticity): 콘텐츠의 출처와 편집 이력을 기록하는 개방형 표준을 마련한 산업 연합체

• 자격증명 유무만 읽는 브라우저 내 판독과 확인 대상의 제한

- 앤트로픽에 따르면 이용자가 파일을 올리면 콘텐츠 체커는 브라우저 안에서 콘텐츠 자격증명만 확인하며, 파일의 실제 내용은 분석하지 않음[3]. 이에 따라 파일이 이용자의 기기 외부로 전송되거나 저장되지 않으며 다른 용도로 활용되지도 않음

- 다만 파일 내용 자체를 분석하지 않기 때문에, 파일의 실제 내용을 분석해 AI 생성 여부를 추정하는 일반적인 AI 탐지 도구와 달리 확인할 수 있는 대상과 범위가 제한적임

- 또한 텍스트에는 자격증명 대신 글 자체에 워터마크(watermark)를 삽입하며, 이를 확인하는 워터마크 탐지 API (Watermark Detection API)는 EU 규제기관, 언론, 연구기관 등에만 비공개 프리뷰로 제공됨. 이에 따라 누구나 확인할 수 있는 대상은 현재 콘텐츠 체커가 지원하는 파일 형식에 한정됨

• 양성과 음성 결과 모두에서 나타나는 AI 생성 여부 판정의 한계

- 자격증명이 식별되어 양성으로 판정된 경우. 클로드가 파일을 만들거나 처리하는 데 관여했을 가능성을 보여줌. 그러나 직접 만든 파일을 편집이나 변환 목적으로 올린 경우에도 자격증명이 붙을 수 있어, 결과만으로 클로드가 어느 정도 관여했는지는 판단할 수 없음

3) Anthropic, "Check if a file was made with Claude", claude.com, 2026.09.10. 조회, https://claude.com/check-files

- 반대로 메타데이터가 제거되거나 파일이 변환 및 수정되면 클로드가 만든 파일에서도 자격증명이 사라질 수 있음. 앤트로픽도 결과가 음성이라고 해서 클로드를 사용하지 않았음을 입증할 수는 없다고 밝힘[4)]
- 이에 따라 콘텐츠 체커는 의심스러운 보도 사진이나 AI 생성 그래픽의 출처를 확인하는 1차 참고 수단으로 활용할 수 있으나, 콘텐츠 생성 경위와 조작 여부를 종합적으로 판단하는 포렌식 분석(forensic analysis)을 대신하기는 어려움

개방 표준 채택의 의미와 사업자별 확인 도구 병립의 과제

• 개방 표준 채택에 따른 사업자 간 호환 가능성

- 앤트로픽은 자체 확인 도구를 개발하면서도 판독 대상에는 독자 형식이 아닌 개방 표준을 채택함. 사업자별로 확인 도구가 달라도 표시 형식을 통일하면, 콘텐츠를 생성한 사업자와 관계없이 출처 정보를 판독할 수 있는 기반이 마련됨
- 다만 표시 적용이 아직 모든 모델로 확대되지 않았고 메타데이터 편집 과정에서도 표시가 사라질 수 있음. 사업자 간 호환성을 높이는 표준화와 함께 콘텐츠가 유통되는 과정에서도 표시가 유지되도록 하는 조치가 필요함

• 확인 결과의 해석 한계와 사업자 간 교차 확인 체계의 부재

- 이용자가 콘텐츠 체커의 음성 결과를 사람이 만든 파일이라는 근거로 해석하면 오인이 발생할 수 있음. 따라서 표시가 없다는 결과가 AI 생성이 아님을 의미하지 않는다는 점과 도구가 확인할 수 있는 범위를 함께 안내할 필요가 있음
- 아울러 사업자마다 별도의 확인 도구를 제공하는 현재 구조에서는 도구가 늘어날수록 이용자가 사업자별 서비스를 일일이 확인해야 하는 불편이 커짐
- 어느 도구에서도 출처를 확인하지 못하는 파일이 남을 수 있는 만큼, 표시 형식의 표준화를 넘어 여러 사업자의 확인 결과를 한곳에서 교차 대조할 수 있는 체계 마련이 과제임

참고문헌

• Anthropic, "Check if a file was made with Claude", claude.com, 2026.09.10. 조회, https://claude.com/check-files
• Brian Beckmann, "Anthropic launches free checker tool for Claude watermarks", Beckmann, 2026.09.04., https://beckmann.ai/en/security/2026-09/anthropic-claude-watermark-checker
• Isaiah Richard, "Anthropic's Claude Content Checker Tool Is Now Available—Here's How to Use the Detector", iTech Post, 2026.09.02., https://www.itechpost.com/articles/237212/20260902/anthropics-claude-content-checker-tool-now-availableheres-how-use-detector.htm

4) Brian Beckmann, "Anthropic launches free checker tool for Claude watermarks", Beckmann, 2026.09.04., https://beckmann.ai/en/security/2026-09/anthropic-claude-watermark-checker

벨기에, 불법 웹사이트 차단의 한계에 따른 도메인 등록 정보 확보

도메인 차단의 한계와 운영자 식별을 위한 대응

• 반복되는 도메인 변경으로 제한된 차단 효과

- 벨기에 경제부 산하 온라인 침해 대응 부서(Department for Combating Online Infringement, 이하 BAPO)는 브뤼셀 상사법원(Business Court of Brussels)의 명령을 근거로 불법 웹사이트의 도메인 차단 결정을 발부해, 지난 2년간 도메인 1,500개 이상을 차단함[1)]
- 그러나 불법 웹사이트 운영자가 도메인 이름(domain name)*을 빠르게 바꾸며 차단을 피해 가면서, 접속 차단만으로는 침해를 막는 데 한계가 나타남
- 법원은 이번 사안에서 대규모 침해가 조직적으로 이루어지고, 운영자가 여러 단계에 걸쳐 신원을 숨기고 있다고 봄. 이에 운영자의 신원 정보와 결제 및 접속 기록 확보가 필요하다고 판단함

* 도메인 이름(domain name): 숫자로 된 복잡한 IP 주소 대신 사람이 쉽게 기억할 수 있도록 문자로 표시한 인터넷 주소

• 도메인 등록 정보를 통한 운영자 식별 추진

- 이러한 판단에 따라 BAPO는 2026년 8월 도메인 등록 단계의 중개자에게 운영자 정보를 제출하도록 요구하는 결정 5건을 발부함[1)]. 이번 결정은 사이트 접속을 차단하는 대신, 도메인 등록 과정에서 남은 기록을 확보하는 데 초점을 맞춤
- 해당 결정 5건 중 4건은 도메인 등록대행자(registrar)*를, 1건은 등록인(registrant)** 기록을 직접 관리하는 등록기관(registry)***를 대상으로 함. 이름이 확인된 등록대행자 3곳은 모두 EU 역내에 소재하지만 벨기에 국외에 있는 것으로 확인됨[1)]

* 등록대행자(registrar): 도메인 등록을 대행하며 등록인의 계정 정보와 결제 정보를 보유하는 사업자

** 등록인(registrant): 도메인을 등록해 사용 권리를 보유한 개인이나 조직으로, 실질적인 도메인 소유자로 인정받는 주체

*** 등록기관(registry): 도메인의 등록인 기록을 직접 보유하고 관리하는 기관

중개자 유형에 따라 달라지는 요구 범위

• 등록대행자에 계정 정보와 결제 및 접속 기록 제출 요구

- 등록대행자를 대상으로 한 결정 4건은 모두 7개 범주의 정보를 요구함. 요구 정보는 이름, 연락처, 청구지 주소와 계좌 정보, 결제 수단, 같은 고객이 등록한 다른 도메인, 계정 생성 당시의 IP 주소와 접속 시각, 최근 12개월간의 접속 기록으로 구성됨[2)]
- 신원 관련 항목에는 계정에 연결된 우편 주소와 이메일 주소, 전화번호뿐 아니라 해당 정보의 수정 이력도 포함되어, 운영자가 과거에 사용한 연락처까지 확인할 수 있도록 함

1) Ernesto Van der Sar, "Belgian Orders Demand Pirate Site Operators' Bank Details, Crypto Wallets and Server Logs", TorrentFreak, 2026.08.28., https://torrentfreak.com/belgian-orders-demand-pirate-site-operators-bank-details-crypto-wallets-and-server-logs/
2) FPS Economy, "Department for Combating Infringements of Copyright and Related Rights Committed Online and the Illegal Exploitation of Online Games of Chance of 19 August 2026 CONCERNING ORDER RR/26/XXX", FPS Economy, 2026.08.19., https://economie.fgov.be/en/themes/intellectual-property/intellectual-property-rights/copyright-and-related-rights/sanctions-and-legal-actions/online-piracy

- 결제 관련 항목에는 전체 IBAN(International Bank Account Number)*과 예금주명, 카드 발급 은행과 발급국, 인터넷 결제 서비스 페이팔(PayPal)의 계정 ID가 포함됨
- 암호자산(crypto-asset)**으로 결제한 경우에는 자산 종류와 지갑 주소(wallet address)***, 해시 ID (hash ID)**** 등도 요구 정보에 해당함

* IBAN(International Bank Account Number): 국가와 은행, 계좌번호를 하나로 표기하는 국제 은행 계좌 식별 번호

** 암호자산(crypto-asset): 블록체인 등 분산원장 기술을 바탕으로 암호화 기술을 적용해 안전하게 생성 및 거래되는 디지털 자산

*** 지갑 주소(wallet address): 블록체인에서 디지털 자산을 송금하고 수신하기 위해 사용하는 고유 식별자

**** 해시 ID(hash ID): 암호화폐 거래가 시작될 때 생성되는 문자열로, 블록체인 상에서 거래를 추적하는 데 사용되는 고유 식별자

• 등록기관에는 도메인 등록 및 관리 정보 중심으로 제출 요구

- 반면 등록기관에는 등록인 정보, 해당 도메인의 등록과 관리를 맡은 등록대행자의 신원, 사용 중인 네임서버(name server)*와 변경 이력을 요구함. 등록대행자와 달리 도메인의 등록 및 관리 기록에 초점을 맞춘 것임
- 이에 따라 계정의 결제 수단과 접속 기록은 요구 정보에서 제외되어, 등록대행자에 요구한 범위보다 좁음. 같은 사안에서도 어떤 중개자를 대상으로 하느냐에 따라 요구 정보가 달라짐

* 네임서버(name server): 도메인 이름에 연결된 서버의 주소 정보를 알려주는 서버

계정 추적의 활용 가능성과 남은 한계

• 도메인 변경에 영향을 받지 않는 추적 경로의 확보

- 이번 결정은 차단할 도메인을 추가하는 대신, 도메인을 등록한 계정의 기록을 확보하도록 함. 이를 통해 운영자가 도메인을 바꾸더라도 계정에 남은 연락처와 결제 수단, 접속 기록을 대조하며 추적을 이어갈 수 있음
- 등록기관에 요구한 등록대행자의 신원과 네임서버 변경 이력도 추적의 단서가 될 수 있음. 한 중개자의 기록만으로 운영자를 특정하기 어려운 경우, 해당 기록에 나타난 다른 중개자를 통해 추가 기록을 확인할 수 있기 때문임

• 기록 보관 범위와 신원 은폐 및 국외 집행의 한계

- 다만 확보할 수 있는 기록은 중개자가 실제로 보관해 온 범위에 한정되며, 접속 기록의 요구 범위도 최근 12개월로 제한됨. 중개자마다 보관하는 기록의 종류와 기간이 다르므로, 같은 정보를 요구하더라도 실제 확보 결과는 달라질 수 있음
- 운영자가 여러 단계에 걸쳐 신원을 숨긴 경우에는 등록 정보와 결제 기록을 확보하더라도 실제 운영 주체를 특정하기 어려울 수 있음. 이름이 확인된 등록대행자 3곳도 모두 벨기에 국외에 있어, 벨기에 법에 근거해 조치하는 BAPO가 이들에게 정보 제출을 강제할 수 있을지는 불확실함

참고문헌

• Ernesto Van der Sar, "Belgian Orders Demand Pirate Site Operators' Bank Details, Crypto Wallets and Server Logs", TorrentFreak, 2026.08.28., https://torrentfreak.com/belgian-orders-demand-pirate-site-operators-bank-details-crypto-wallets-and-server-logs/

• FPS Economy, "Department for Combating Infringements of Copyright and Related Rights Committed Online and the Illegal Exploitation of Online Games of Chance of 19 August 2026 CONCERNING ORDER RR/26/XXX", FPS Economy, 2026.08.19., https://torrentfreak.com/images/260819-BAPO-D-FR-020-EN.pdf

TPN의 픽스 잇 오어 리스크 잇 공개로 본 콘텐츠 보안의 실행 격차

알려진 취약점에서 반복되는 사고와 보안 조치의 지연

• 취약점 발견과 악용 가속에 따른 보안 경보의 급증

- 미국영화협회(Motion Picture Association, 이하 MPA)가 소유하고 운영하는 콘텐츠 보안 기구 트러스티드 파트너 네트워크(Trusted Partner Network, 이하 TPN)는 글로벌 미디어 및 엔터테인먼트 산업을 대상으로 보안 경보를 발행하고 있음
- 최근 보안 경보에서는 자격증명(credential)* 유출, 다중 인증(Multi-Factor Authentication)** 미비, 보안 설정 미흡, 취약점 관리 부실 등이 반복적으로 지적됨. 이들 항목 상당수는 평가 단계에서 미리 확인할 수 있는 보안 문제이지만, 실제 사고에서도 되풀이되고 있음
- 근본적인 취약점 유형에는 큰 변화가 없으나, 공격자의 탐색 및 악용 속도는 가속화되고 있음. 이에 보안 경보 발행 건수는 2026년 1분기에 이미 2025년 연간 건수를 웃돌았으며, 6월 기준으로는 두 배를 넘어섬[1)]

* 자격증명(credential): 시스템이나 웹사이트, 서비스에서 사용자의 신원과 권한을 확인하는 데 사용하는 인증 정보

** 다중 인증(Multi-Factor Authentication): 비밀번호 외에 별도의 인증 수단을 추가로 요구해 계정 접근을 통제하는 방식

• 평가 결과가 실제 조치로 이어지지 않는 실행 격차

- TPN은 이러한 위협에 대응하기 위해 MPA 콘텐츠 보안 모범사례(MPA Content Security Best Practices)를 바탕으로 독립 평가를 운영해 옴
- 그러나 2026년에 처음 발간한 보안 평가 데이터 분석 보고서 'STAR 보고서(TPN STAR Report)'에 따르면, 2023~2025년 평가가 주기적으로 이뤄졌음에도 산업 전체 미준수율(non-compliance)*은 23.45%에 달함[2)]
- 평가에서 확인된 보안 문제 중 해결되지 않은 비중은 70%에 이르며, 조치를 취하기까지 평균 112.34일이 걸림[2)]. 문제를 확인하고도 조치가 제때 이뤄지지 않아, 반복적인 평가가 산업 전체 미준수율 감소로 이어지지 못하고 있음
- 이에 TPN은 평가에서 드러난 보안 문제를 실제로 조치하도록 지원하고자, 2026년 9월 9일 미디어 산업 행사 IBC2026에서 비상업 캠페인 '픽스 잇 오어 리스크 잇(Fix It or Risk It)'을 공개함[3)]

* 미준수율(non-compliance): 평가자가 모범사례의 부분 이행 또는 미이행으로 판정한 항목의 비율

1) Trusted Partner Network, "New Report from the MPA's Content Security Initiative Links Control Failures to Content Security Incidents Across the Entertainment Industry", Trusted Partner Network, 2026.09.09., https://trustedpartnernetwork.org/2026/09/trusted-partner-network-launches-fix-it-or-risk-it/
2) Trusted Partner Network, "TPN STAR Report 2026", Motion Picture Association, 2026.04.28., https://www.ttpn.org/star-access/
3) Steven Hawley, "IBC 2026: MPA's Trusted Partner Network launches Fix It or Risk It security awareness campaign", Piracy Monitor, 2026.09.04., https://piracymonitor.org/ibc-2026-mpas-trusted-partner-network-launches-fix-it-or-risk-it-security-awareness-campaign/

픽스 잇 오어 리스크 잇, 비용 부담을 낮추고 조치 순서를 안내

• 평가나 인증 없이 활용할 수 있는 무료 실무 자료 제공

- 픽스 잇 오어 리스크 잇은 엔터테인먼트 산업의 조직들이 일상적인 보안 문제를 파악하고 작은 취약점이 큰 사고로 이어지기 전에 조치하도록 돕는 비상업 캠페인임. 평가나 인증을 거치지 않고도 활용할 수 있는 자료를 산업 전반에 공개함
- 이를 위해 실무 지침과 정책 템플릿, 교육 자료, 미디어 자료, 격차 분석으로 구성된 온라인 도구 모음 픽스 잇 툴킷(Fix It Toolkit)을 모두 무료로 제공해 조직이 보안 조치를 마련하는 데 드는 비용 부담을 낮춤
- 이러한 비용 부담 완화가 실제 보안 조치로 이어진다는 점은 앞선 사례에서도 확인됨. TPN은 2025년 1월 무료 정책 템플릿 공개 이후 평균 조치 기간이 36% 단축됐다고 보고함[4)]

• 조치 우선순위와 반복 점검을 통한 지속적인 보안 개선 지원

- 툴킷은 각 조치에 필요한 노력의 정도에 따라 실행 순서를 제시함. 비용이 들지 않고 간단하면서도 보안 효과가 큰 조치부터 시작해, 이후 더 많은 조정이나 장기 계획이 필요한 개선으로 범위를 넓히도록 안내함
- 또한 보안 기준이 계속 높아지는 만큼 한 차례 조치로 끝내지 않고 정기적으로 점검하고 개선하도록 권고함. 외부 지원이 필요한 조직에는 TPN 평가 수행 자격을 갖춘 전문가 명부를 제공함
- 반복 점검의 효과는 STAR 보고서에서도 확인됨. TPN에 2년간 참여한 조직 가운데 재평가를 마친 곳은 34.41%에 그쳤으나, 재평가를 거친 조직의 미준수율은 25.46%에서 17.99%로 낮아짐[4)]

평가 중심 대응의 전환과 지속 이행 확인 수단의 마련

• 문제 확인에서 조치 실행으로 옮겨간 대응의 초점

- 이번 사례는 평가로 드러난 보안 문제를 실제 개선으로 연결하는 것이 중요함을 부각함. 문제가 확인되더라도 해결되지 않는 상황에서, 평가를 확대하기보다 조치를 지연시키는 비용과 우선순위 판단의 부담을 줄이려는 접근임
- 무료 정책 템플릿 제공 이후 조치 기간이 36% 단축된 점은 이러한 지원의 효과를 보여줌. 다만 툴킷의 활용 여부와 조치 범위는 각 조직이 결정하므로, 이번 캠페인이 산업 전반의 개선으로 이어지는지는 향후 미준수율의 변화를 통해 확인할 필요가 있음

• 자율적 재점검의 한계를 보완하는 상시 관리의 필요성

- 조치를 실행한 이후에는 해당 기능이 계속 작동하는지도 확인해야 함. 그러나 자발적인 참여에 의존하는 것만으로는 보안 조치의 유지 여부를 파악하는 데 한계가 있음
- 이에 STAR 보고서는 상시 모니터링과 자동화를 확대하고 보안 조치의 성과에 대한 책임 소재를 명확히 할 것을 제시함. 개별 조직의 점검 의지에 의존하는 방식을 보완하려면, 지속적인 확인 수단과 관리 책임을 함께 갖출 필요가 있음

4) Trusted Partner Network, "TPN STAR Report 2026", Motion Picture Association, 2026.04.28., https://www.ttpn.org/star-access/

참고문헌

- Steven Hawley, "IBC 2026: MPA's Trusted Partner Network launches Fix It or Risk It security awareness campaign", Piracy Monitor, 2026.09.04., https://piracymonitor.org/ibc-2026-mpas-trusted-partner-network-launches-fix-it-or-risk-it-security-awareness-campaign/
- Fix it or risk it, "Big security incidents can start with small moments. Let's fix it.", 2026.09.18. 조회, Fix it or risk it, https://www.fixitorriskit.org/
- Trusted Partner Network, "TPN STAR Report 2026", Motion Picture Association, 2026.04.28., https://www.ttpn.org/star-access/
- Trusted Partner Network, "New Report from the MPA's Content Security Initiative Links Control Failures to Content Security Incidents Across the Entertainment Industry", Trusted Partner Network, 2026.09.09., https://trustedpartnernetwork.org/2026/09/trusted-partner-network-launches-fix-it-or-risk-it/

팬그램의 서브스택 적용 사례로 본 AI 개입 정도 판별 방식의 변화

AI 생성 콘텐츠 확산과 이분법적 판별의 한계

• 작성 주체가 불분명해지면서 커진 콘텐츠 신뢰 판단의 곤란

- AI가 생성한 글과 이미지가 소셜 미디어(social media)를 넘어 채용 지원서와 상품 리뷰, 보험 청구 자료에도 사용되면서, 플랫폼과 독자 모두 AI 생성물과 사람의 창작물을 명확히 구별하기 어려워지고 있음
- 이러한 변화는 독자가 콘텐츠를 읽는 태도에도 영향을 준다는 지적이 제기됨. AI가 만들어 낸 오류를 의심하며 읽어야 할지, 사람이 직접 취재한 글로 믿고 읽어도 될지를 판단하는 데 작성 주체가 중요한 기준이 되기 때문임

• AI 사용 여부만으로 구분하기 어려운 사람-AI 혼합 콘텐츠의 증가

- 판별 대상에는 AI가 전부 작성한 글뿐 아니라 사람이 쓴 글을 AI로 다듬은 글도 포함됨. 이러한 사람-AI 혼합 콘텐츠(mixed human-AI content)*는 사람과 AI 중 어느 한쪽이 작성한 것으로 판별하기 어려움
- 이에 AI 사용 여부를 가르는 것보다 개입 정도를 파악하는 일이 더 어렵지만 유용할 수 있다는 견해가 나옴[1]. AI를 사용한 사실을 숨기지 않고 공개한다면 그 사용 자체는 받아들일 수 있다는 시각도 함께 제시됨

* 사람-AI 혼합 콘텐츠(mixed human-AI content): 사람이 직접 쓴 글을 AI로 다듬는 등 사람과 AI의 작성이 함께 들어간 콘텐츠

팬그램의 AI 개입 정도 판별 모델과 서브스택 적용

• 사람의 원문과 AI가 다시 쓴 문서를 비교하는 학습 방식의 적용

- 미국 AI 판별 스타트업 팬그램(Pangram)은 2026년 7월 900만 달러(원화 약 123억 6,240만 원[2])를 조달하고, 차세대 텍스트 판별 모델 팬그램 4(Pangram 4)와 이미지 판별 모델 팬그램 이미지(Pangram Image)를 함께 공개함[3]
- 팬그램 4는 사람이 쓴 문서 수천만 건과 이를 AI가 다시 작성한 합성 대조본(synthetic mirror)*을 짝지어 학습함. 합성 대조본은 최신 대규모 언어 모델(Large Language Model, LLM)을 이용해 원문과 주제, 길이, 어조가 같도록 작성함

1) Rebecca Bellan 외 1명, "We're 'dangerously close' to dead internet theory, says Pangram's CEO", TechCrunch, 2026.09.02., https://techcrunch.com/podcast/were-dangerously-close-to-dead-internet-theory-says-pangrams-ceo/
2) 1달러=1,373.60원(KEB 하나은행 매매기준율 적용, 2026.09.01.)
3) Rebecca Bellan, "As AI content floods the internet, Pangram raises $9M to detect it", TechCrunch, 2026.07.29., https://techcrunch.com/2026/07/29/as-ai-content-floods-the-internet-pangram-raises-9m-to-detect-it/

- 모델은 두 문서를 비교하면서 AI가 반복적으로 사용하는 문체의 특징을 학습하므로, 워터마크(watermark)와 같은 사전 표식이 없는 글에도 적용할 수 있음. 팬그램 측은 AI가 전부 작성한 글과 사람-AI 혼합 콘텐츠를 99%가 넘는 정확도로 판별한다고 밝힘

* 합성 대조본(synthetic mirror): 사람이 쓴 문서와 주제, 길이, 어조를 맞춰 AI로 다시 작성한 문서로, 원문과 비교해 AI 특유의 문체를 학습하는 데 사용

[표 1] 팬그램이 2026년 7월 공개한 판별 모델 2종 비교

구분	팬그램 4	팬그램 이미지
판별 대상	텍스트	이미지
학습 방식	사람 문서와 합성 대조본의 대조	화소 단위 분포 차이 학습
판별 결과	AI 개입 비율	AI 생성 여부
사전 표식 의존	없음	없음
공개 상태	정식 공개	연구용 시험 공개

출처: Rebecca Bellan, "As AI content floods the internet, Pangram raises $9M to detect it", TechCrunch, 2026.07.29., https://techcrunch.com/2026/07/29/as-ai-content-floods-the-internet-pangram-raises-9m-to-detect-it/

• 서브스택에서 독자에게 AI 개입 정도를 보여주는 기능 제공

- 팬그램의 판별 기술은 같은 달 뉴스레터 발행 플랫폼 서브스택(Substack)에 연동됨. 독자는 앱에서 100자를 넘는 게시물과 댓글을 검사해, 사람과 AI가 각각 작성한 비율의 추정치를 확인할 수 있음[4)]
- 서브스택은 이 기능의 목적이 AI 사용을 금지하거나 제재하는 데 있지 않다고 밝힘. 작성자가 글을 작성한 과정을 설명하도록 유도하기 위한 것으로, AI 사용 사실을 직접 알릴 수 있는 표시 기능도 함께 제공함
- 작성자가 판별 결과를 확인하고 정정을 요청할 수 있는 절차도 마련함. 발행 전에 원고를 미리 검사할 수 있으며, 잘못된 것으로 판단한 결과는 신고해 삭제할 수 있음

• 실제 검증에서 드러난 문장 및 이미지 판별의 오류

- 다만 실제 검증에서는 회사가 제시한 정확도와 별개로 문장별 판별 오류가 확인됨
- 팬그램 공동창업자 맥스 스페로(Max Spero)는 사람이 쓴 문서 1만 건 중 오탐 사례가 1건에 불과하다고 밝혔으나, 미국 IT 매체 테크크런치(TechCrunch)의 검증에서는 기자가 AI 문장을 통째로 다시 쓴 경우에도 AI가 작성한 것으로 잘못 판단함[5)]
- AI 개입 비율을 산출하는 과정에서도 판별이 일관되지 않은 사례가 나타남. 기자가 자신의 기사를 AI로 다듬은 뒤 검사한 결과 개입 비율은 13%로 나왔으나, 일부 문장의 작은 단어 변화는 판별하면서 다른 문장의 수정은 놓쳤고 사람이 쓴 문장도 AI가 개입한 것으로 표시함[5)]
- 이미지 판별에서도 오류가 확인됨. 팬그램 이미지는 실제 사진 속에 포함된 AI 이미지를 판별하기도 했지만, AI 이미지를 촬영한 사진을 사람이 직접 만든 콘텐츠로 판별한 사례도 있었음

4) Sarah Perez, "Substack's new tool tells you who's been writing their newsletters with AI", TechCrunch, 2026.07.22., https://techcrunch.com/2026/07/22/substacks-new-tool-tells-you-whos-been-writing-their-newsletters-with-ai/
5) Rebecca Bellan, "As AI content floods the internet, Pangram raises $9M to detect it", TechCrunch, 2026.07.29., https://techcrunch.com/2026/07/29/as-ai-content-floods-the-internet-pangram-raises-9m-to-detect-it/

AI 개입 정도 판별의 의미와 결과 활용의 과제

• 사전 표식 없는 콘텐츠와 판별 범위 확대

- 이번 사례는 콘텐츠 자체의 문체와 이미지 내부의 통계적 특성인 화소 단위 분포(pixel-level distribution)*를 근거로 판별한다는 점에서 의미가 있음
- 생성 단계에서 사전 표식을 넣지 않은 콘텐츠도 유통 이후 판별할 수 있어, 특정 사업자가 사전 표식을 부여한 콘텐츠에 국한되지 않고 판별을 시도할 수 있음
- 텍스트 판별의 범위도 AI 사용 여부를 넘어 개입 정도를 추정하는 데까지 넓어짐. 서브스택 사례는 이러한 결과가 콘텐츠를 제재하는 근거뿐 아니라, 작성자가 AI 사용 사실과 작성 과정을 독자에게 알리도록 유도하는 수단으로 활용될 수 있음을 보여줌

* 화소 단위 분포(pixel-level distribution): 이미지를 이루는 화소의 밝기와 색상값이 퍼져 있는 양상으로, AI가 생성한 이미지는 촬영물과 이 양상이 다르게 나타남

• 판별 오류를 고려한 결과 해석과 정정 절차 마련의 필요성

- 다만 콘텐츠 전체의 AI 개입 비율과 개별 문장의 작성 주체에 대한 판단은 구분할 필요가 있음. 전체 비율을 대체로 맞히더라도 특정 문장을 잘못 판별할 수 있으므로, 문장별 판별 결과만으로 AI 사용 여부를 단정하기는 어려움
- 서브스택이 결과를 정정할 수 있는 절차를 함께 마련한 것도 이러한 오류 가능성을 고려한 설계로 볼 수 있음. 판별 기능을 도입하는 플랫폼은 결과를 표시하는 방식과 함께, 작성자가 잘못된 결과를 정정할 절차도 마련할 필요가 있음
- 아울러 어느 수준의 AI 개입부터 표시할 것인지 기준을 정하는 일도 과제로 남음. 개입 정도가 하나의 비율로 제시되더라도 문장별 판별에 오류가 있을 수 있는 만큼, 해당 수치의 의미와 한계를 고려해 표시 기준을 정할 필요가 있음

참고문헌

• Rebecca Bellan 외 1명, "We're 'dangerously close' to dead internet theory, says Pangram's CEO", TechCrunch, 2026.09.02., https://techcrunch.com/podcast/were-dangerously-close-to-dead-internet-theory-says-pangrams-ceo/
• Rebecca Bellan, "As AI content floods the internet, Pangram raises $9M to detect it", TechCrunch, 2026.07.29., https://techcrunch.com/2026/07/29/as-ai-content-floods-the-internet-pangram-raises-9m-to-detect-it/
• Sarah Perez, "Substack's new tool tells you who's been writing their newsletters with AI", TechCrunch, 2026.07.22., https://techcrunch.com/2026/07/22/substacks-new-tool-tells-you-whos-been-writing-their-newsletters-with-ai/

프렌드MTS와 애자일TV의 사전 통합으로 본 서버 측 워터마킹 도입의 간소화

라이브 스포츠 보호 요구의 확대와 워터마킹 도입의 부담

• 접근 제한을 보완하는 유출 출처 식별의 필요성

- 라이브 스포츠와 같은 고가치 프리미엄 콘텐츠의 불법 유통을 막기 위해 스포츠 권리자들이 더 강력한 보호 수단을 요구하고 있음. 이에 영상 전송 단계에서 식별 신호를 넣는 서버 측 워터마킹(server-side watermarking)*의 중요성이 커지고 있음
- 접근 권한을 관리하는 디지털 권리 보호(Digital Rights Management, 이하 DRM)**은 무단 접근을 막지만, 정상적으로 전달된 영상이 유출되면 그 출처까지 식별하지는 못함
- 워터마킹은 온라인에 유통되는 불법 영상에서 식별 신호를 읽어 유출 출처를 특정함으로써 이러한 한계를 보완함

* 서버 측 워터마킹(server-side watermarking): 영상을 시청자에게 전송하는 단계에서 시청 세션마다 다른 식별 신호를 넣는 기술

** 디지털 권리 보호(Digital Rights Management, DRM): 자격을 갖춘 이용자만 콘텐츠를 재생할 수 있도록 접근 권한을 관리하는 기술

• 개별 통합 작업에 따른 도입 지연과 보호 수준 확보의 부담

- 그러나 워터마킹을 실제 서비스에 적용하려면 사업자마다 워터마킹 백엔드(watermarking backend)*와 영상 전송 설비를 연결하는 작업이 필요함. 이러한 개별 통합은 도입 기간을 지연시킬 뿐 아니라, 시스템 연동 과정에서의 장애 및 운영 리스크를 가중시킴
- 이처럼 통합 부담이 커지면 사업자는 도입 속도와 보호 수준을 함께 확보하기 어려워짐. 빠른 도입을 우선하면 보호 수준을 낮춰야 하고, 충분한 보호 수준을 갖추려면 도입 시점을 늦춰야 하는 부담이 발생함

* 워터마킹 백엔드(watermarking backend): 어떤 계정에 어떤 식별 신호를 배정할지 정하고 그 대응 관계를 보관해, 나중에 신호를 해독할 때 참조하는 서버 설비

프렌드MTS와 애자일TV의 사전 통합과 유출 차단 절차의 연결

• 워터마킹 백엔드와 CDN의 사전 통합을 통한 도입 부담 완화

- 이러한 부담을 줄이기 위해 콘텐츠 보호 기업 프렌드MTS(Friend MTS)와 스페인 영상 전송 솔루션 기업 애자일TV(AgileTV)는 2026년 9월 3일 애자일TV의 콘텐츠 전송 네트워크(Content Delivery Network, 이하 CDN)*와 프렌드MTS의 서버 측 워터마킹을 처음으로 통합한 사례를 공개함[1)]

1) Friend MTS, "Friend MTS and AgileTV Complete CDN Watermarking Integration to Fast-Track Anti-Piracy Protection for Operators", 2026.09.03., https://www.friendmts.com/news/friend-mts-and-agiletv-complete-cdn-watermarking-integration-to-fast-track-anti-piracy-protection-for-operators

- 프렌드MTS는 워터마킹의 핵심 처리를 담당하는 백엔드를 제공하고, 애자일TV의 CDN은 선별된 프리미엄 콘텐츠에 워터마킹이 적용된 스트림을 전달함. 양사는 각각 보호 기능과 영상 전송을 맡아 두 기능이 함께 작동하도록 통합함
- 이에 따라 두 솔루션 중 하나를 이미 사용하는 사업자는 다른 하나를 도입할 때 시스템 통합 부담을 덜 수 있음
- 프렌드MTS 최고경영자 셰인 맥카시(Shane McCarthy)는 사업자가 도입 속도와 보호 수준 중 하나를 포기할 필요 없이, 두 요건을 갖춘 공동 솔루션을 빠르게 도입할 수 있게 됐다고 설명함

* 콘텐츠 전송 네트워크(Content Delivery Network, CDN): 영상 등의 콘텐츠를 여러 지역의 거점 서버에 분산해 두고 시청자와 가까운 곳에서 전달하는 전송망

• 세션 토큰을 활용한 가입자별 워터마킹 스트림 생성

- 가입자를 구분하는 식별 신호는 영상 전송 단계에서 구성됨. 애자일TV의 CDN은 가입자의 세션 토큰(session token)*을 바탕으로 가입자마다 서로 다른 A/B 워터마킹(A/B watermarking)** 스트림을 만들어 전송함
- 같은 콘텐츠라도 가입자마다 식별 신호가 다르므로, 불법 영상이 온라인에 유통되면 해당 신호를 읽어 최초 유출 경로가 된 가입자 계정을 특정할 수 있음. 양사는 이러한 워터마킹이 영상 서비스 품질에 영향을 주지 않는다고 설명함

* 세션 토큰(session token): 이용자가 서비스에 접속할 때 발급되는 식별값으로, 개별 시청 세션을 구분하는 데 쓰임

** A/B 워터마킹(A/B watermarking): 같은 영상을 미세하게 다른 두 형태로 준비한 뒤, 세션마다 다른 순서로 조합해 고유한 식별 신호를 만드는 방식

• 불법 유통 감시와 연계한 유출 계정의 실시간 차단

- 워터마킹은 프렌드MTS의 불법 유통 감시 기능과 연계돼 계정 차단으로 이어짐. 감시 과정에서 유출 계정이 확인되면 사업자가 미리 정한 규칙에 따라 실시간으로 차단함
- 다만 이번 사전 통합은 CDN 구간에 한정되므로, 전체 서비스를 구축하려면 추가 통합 작업이 필요함. 애자일TV 최고경영자 콜도 우나누에(Koldo Unanue)는 헤드엔드(headend)*와 비즈니스 로직(Business Logic)을 포함한 전체 구축 방식은 사업자의 환경과 요구사항에 따라 달라진다고 밝힘[2)]

* 헤드엔드(headend): 영상 인코딩과 전송용 형식 변환 등 송출 전 단계의 처리를 담당하는 설비

사전 통합 방식의 의미와 남은 적용 과제

• 공급업체 간 공통 통합 작업의 선행을 통한 도입 부담 완화

- 이번 사례는 워터마킹을 별도 시스템으로 도입하는 데서 나아가, 영상 전송 과정에 보호 기능을 미리 통합하는 방식을 보여줌. 사업자는 전송 설비와 통합된 보호 기능을 함께 도입해 운영할 수 있음
- 특히 사업자마다 반복하던 통합 작업을 공급업체가 미리 수행했다는 점에서 의미가 있음. 사업자가 직접 처리해야 할 통합 작업이 줄어들면서 도입을 결정하고 추진하는 데 따르는 부담과 위험도 낮아짐

2) AgileTV, "AgileTV and Friend MTS Complete CDN Watermarking Integration to Fast-Track Anti-Piracy Protection for Operators", 2026.09.03., https://agiletv.com/agiletv-and-fiend-mts-complete-cdn-watermarking-integration-to-fast-track-anti-piracy-protection-for-operators/

• 타 솔루션으로의 확산과 사업자별 추가 통합의 과제

- 다만 이러한 효과는 애자일TV의 CDN과 프렌드MTS의 워터마킹을 함께 사용하는 사업자에게 한정됨. 다른 조합을 사용하는 사업자는 여전히 개별 통합이 필요하므로, 도입 부담을 폭넓게 낮추려면 공급업체 간 사전 통합 조합이 늘어나야 함
- 적용 대상 역시 선별된 프리미엄 콘텐츠에 한정돼, 현재의 적용 범위만으로는 사업자가 제공하는 모든 콘텐츠의 유출 출처를 가입자 단위로 식별할 수 없음
- 아울러 사전 통합에 포함되지 않은 헤드엔드와 비즈니스 로직의 통합 부담을 줄일 필요가 있음. 이 작업이 사업자마다 별도로 요구되는 만큼, 전체 도입 기간의 단축 효과는 추가 통합에 걸리는 시간까지 고려해 판단해야 함

참고문헌

• Friend MTS, "Friend MTS and AgileTV Complete CDN Watermarking Integration to Fast-Track Anti-Piracy Protection for Operators", 2026.09.03., https://www.friendmts.com/news/friend-mts-and-agiletv-complete-cdn-watermarking-integration-to-fast-track-anti-piracy-protection-for-operators
• AgileTV, "AgileTV and Friend MTS Complete CDN Watermarking Integration to Fast-Track Anti-Piracy Protection for Operators", 2026.09.03., https://agiletv.com/agiletv-and-fiend-mts-complete-cdn-watermarking-integration-to-fast-track-anti-piracy-protection-for-operators/
• Julian Clover, "Friend MTS and AgileTV integrate CDN watermarking to speed anti-piracy deployments", Broadband TV News, 2026.09.03., https://www.broadbandtvnews.com/2026/09/03/friend-mts-and-agiletv-integrate-cdn-watermarking-to-speed-anti-piracy-deployments/

주간 기술 동향

모델 가중치에 워터마크를 삽입하는 편집 이력 추적 기술 AngelFingerprint

• AI를 활용한 이미지 편집 도구의 발전과 출처 확인 기술의 과제

AI를 활용한 이미지 편집 도구는 짧은 프롬프트 하나만으로 사진의 일부를 편집하는 것이 가능하다. 이러한 편집은 이미지를 처음부터 만들어내는 방식보다 허위정보 측면에서 위험이 더 클 수 있다. 원본 사진의 대부분이 편집된 이미지에 그대로 남아 있기 때문에, 편집된 이미지를 보는 사람은 이를 실제 사진으로 받아들이기 쉽다. 조작된 이미지가 진본과 구분하기 어려운 원인이 바로 여기에 있으며, 이미지가 편집되었는지와 무엇이 바뀌었는지를 확인할 방법론이 요구되고 있다.

이러한 문제를 해결하기 위해 촬영 단계에서 진위 확인 정보를 기록하는 방식이 확산되고 있다. 구글(Google)은 2025년부터 안드로이드 스마트폰 카메라에 C2PA 표준을 따르는 진위 확인 소프트웨어를 넣기 시작했으며, 이 방식은 니콘(Nikon), 소니(Sony), 캐논(Canon) 등의 디지털 카메라에도 적용되고 있다. 애플(Apple)도 촬영한 원본 데이터를 회사 서버로 보내 검증한 뒤 고유 식별 번호를 부여하는 기능을 검토했다. 이 식별 번호를 통해 이용자는 해당 사진이 언제 어떤 기기에서 촬영됐는지 확인할 수 있다.

그러나 이러한 방식에도 한계가 있다. 애플이 검토하는 기능은 기본적으로 꺼진 상태로 제공되어 이용자가 직접 켜야 하고, 이미 촬영해 둔 사진에는 적용되지 않는 것으로 알려졌다. 무엇보다 이 정보는 촬영 시점과 촬영 기기 정보만 알려줄 뿐 이미지에 담긴 내용까지는 보증하지 못한다. 생성 단계에서 워터마크를 삽입하는 방식도 사정은 비슷하다. 대부분 모델 번호나 이용자 번호처럼 미리 정해둔 값을 삽입하는데, 이 값은 편집과 무관하게 정해지기 때문에 어느 모델에서 나온 이미지인지만 알려줄 뿐 무엇이 어떻게 바뀌었는지는 알려주지 못한다. 게다가 이러한 워터마크는 분리 가능한 별도 모듈을 통해 삽입되므로, 확산 모델이 오픈소스로 공개된 환경에서는 공격자가 해당 모듈을 찾아내 제거할 수 있다.

이러한 배경에서 편집 프롬프트 임베딩(editing prompt embedding)을 워터마크로 삽입하는 기술이 제시되고 있다. 편집 프롬프트는 이미지를 어떻게 바꿀지 지시하는 문장이며, 임베딩은 이 문장의 의미를 인공지능이 다룰 수 있도록 일정한 길이의 숫자 배열로 바꾼 값이다. 앤젤핑거프린트(AngelFingerprint)는 이 숫자 배열을 이미지에 워터마크로 삽입하고, 이후 이미지의 픽셀만으로 해당 값을 복원한다. 고정된 번호 대신 편집 내용에서 나온 값이 워터마크로 삽입되므로, 이미지가 어떤 모델에서 나왔는지와 함께 어떤 편집이 이루어졌는지까지 확인할 수 있다.

[사례] 편집 프롬프트 임베딩을 워터마크로 삽입하는 앤젤핑거프린트

• 기존 워터마킹 기술의 한계와 세 가지 요건

- 기존 워터마킹 기술 대부분은 미리 정해둔 식별 번호를 워터마크로 삽입하는 방식으로, 이를 복원해도 식별 번호만 확인될 뿐 원본에서 무엇이 어떻게 바뀌었는지에 관한 정보는 담겨 있지 않음
- 또한 다수의 방식이 외부 인코더(encoder)나 사후 처리 모듈처럼 생성 모델과 분리된 별도 모듈을 통해 워터마크를 삽입하기 때문에, 화이트박스(white-box)* 환경에서는 공격자가 해당 모듈을 찾아내 제거할 수 있음
- 이에 따라 편집된 이미지를 대상으로 한 워터마킹은 추적성, 설명가능성, 화이트박스 은닉성 세 가지를 동시에 충족해야 한다는 요건이 제시됨. 추적성은 이미지의 원본을 추적할 수 있는 성질이고, 설명가능성은 편집 내용과 그에 담긴 의도를 파악할 단서를 제공하는 성질임
- 화이트박스 은닉성은 원본 모델과 동일한 구조와 연산 과정을 유지하여 워터마크를 탐지하거나 분리해내기 어렵게 만드는 성질로, 기존 연구에서는 세 가지 요건을 모두 충족한 사례가 없음

* 화이트박스(white-box): 모델의 구조, 가중치, 연산 과정을 모두 열람할 수 있는 상태를 가리키며, 가중치가 공개된 모델을 내려받아 사용하는 환경이 이에 해당함

• 워터마크에 담는 정보와 삽입 위치

- 앤젤핑거프린트는 텍스트 기반 이미지 편집 모델에 적용하는 워터마킹 기술로, 워터마크를 삽입하는 부분과 이미지에서 워터마크를 복원하는 추출기로 구성됨
- 앤젤핑거프린트는 편집 프롬프트를 대조 언어-이미지 사전학습(Contrastive Language-Image Pre-training, CLIP)* 텍스트 인코더에 통과시켜 얻은 768차원 임베딩을 워터마크로 삽입함
- 기존 방식은 모든 이미지에 동일한 값을 삽입하지만, 앤젤핑거프린트는 편집마다 서로 다른 값을 삽입하므로 워터마크를 복원하면 어떤 편집이 이루어졌는지 확인할 수 있음
- 편집 프롬프트를 문장 그대로 삽입하지 않고 임베딩 형태로 삽입하는 이유는 두 가지임. 첫째, 확산 모델이 애초에 텍스트를 임베딩 형태로만 입력받기 때문에 문장을 그대로 넣으려면 별도 모듈이 추가되어야 하고, 둘째, 문장은 연속적인 값이 아니어서 학습 과정에 사용할 수 없음
- 워터마크는 저순위 적응(Low-Rank Adaptation, LoRA)**을 통해 확산 모델의 가중치에 직접 삽입되며, 구조와 연산 과정, 추론 코드가 원본 모델과 동일하게 유지되므로 공격자가 분리해 제거할 수 있는 독립 모듈이 존재하지 않음

* 대조 언어-이미지 사전학습(Contrastive Language-Image Pre-training, CLIP): 이미지와 텍스트를 같은 기준의 숫자 배열로 변환해 둘 사이의 의미적 유사도를 계산할 수 있도록 학습된 모델

** 저순위 적응(Low-Rank Adaptation, LoRA): 모델 전체를 다시 학습시키지 않고 소수의 가중치만 추가 학습시켜 모델의 기능을 조정하는 기법

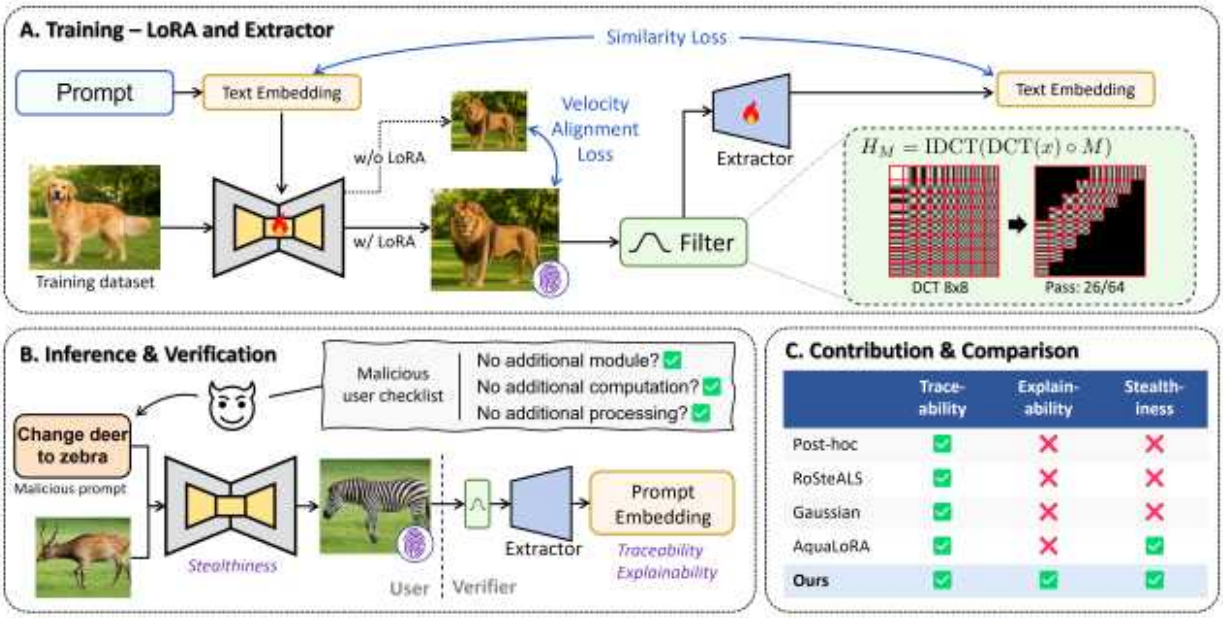
• 편집 품질을 유지하기 위한 학습 방식

- 워터마크가 추출기에 읽히려면 이미지가 바뀌어야 하지만, 변형 폭이 커질수록 워터마크가 눈에 띄고 편집 품질도 저하됨. 두 목표가 서로 충돌하는 구조임
- 복원 손실만으로 학습하면 추출기가 읽어낼 수 있는 변화라면 무엇이든 보상을 받기 때문에, 결과물이 원본 모델의 출력에서 점점 멀어지게 됨
- 앤젤핑거프린트는 워터마크를 삽입하지 않은 원본 모델의 예측값을 기준으로 삼아, 워터마크를 삽입한 모델의 예측값이 여기서 벗어나지 않도록 붙잡는 방식으로 이를 해결함
- 이때 편집 지시문을 넣었을 때의 예측값과 넣지 않았을 때의 예측값을 모두 기준에 맞추는데, 합쳐진 결과값만 맞추면 이용자가 편집 강도를 바꿔 생성할 경우 결과물이 원본 모델과 달라지기 때문임
- 지시문을 넣지 않았을 때의 예측값에 더 큰 가중치를 두어 고정하고, 지시문을 넣었을 때의 예측값은 상대적으로 느슨하게 두어 워터마크를 삽입할 여지를 남김

• 중간 주파수 대역에만 워터마크를 삽입하는 설계

- 앤젤핑거프린트는 생성된 이미지를 주파수 영역으로 변환한 뒤, 중간 대역만 남기고 낮은 대역과 높은 대역을 차단한 신호를 추출기에 전달하는 필터를 둠. 차단된 대역은 학습 과정에도 반영되지 않으므로 워터마크는 중간 대역에만 삽입됨
- 낮은 대역은 이미지의 외형을 좌우하는 영역으로, 이곳이 차단되면 워터마크를 삽입해도 추출기가 읽지 못하므로 결과적으로 워터마크가 낮은 대역에 삽입되지 않아 편집 결과가 육안상 그대로 유지됨
- 또한 낮은 대역에는 이미지의 내용이 담겨 있어, 이를 차단하지 않으면 추출기가 워터마크를 읽는 대신 이미지 내용을 보고 편집 프롬프트를 추측하는 방식으로 학습될 수 있음
- 높은 대역은 압축이나 크기 변경 과정에서 가장 먼저 손실되는 영역이므로 이를 배제하고 상대적으로 안정적인 중간 대역을 사용하며, 이 대역 설정은 이미지 압축 표준인 JPEG 처리 방식을 본떠 64개 구간 중 26개를 남기는 방식으로 정해짐

[그림 1] 앤젤핑거프린트의 워터마크 삽입 및 복원 구조



출처: Bo-Han Kung 외 4인, "AngelFingerprint: A Traceable, Explainable, and White-Box Stealthy Watermark for Text-Guided Image Editing", arXiv, 2026.09.04., https://arxiv.org/pdf/2609.04709

• 편집 프롬프트 복원 성능 검증

- 편집 이미지 데이터셋으로 실험한 결과, MagicBrush 데이터셋의 200개 후보 조건에서 실제 사용된 편집 프롬프트를 정확히 찾아낸 비율이 86%로, 기존 프롬프트 역추론 기법의 20%대를 크게 상회함
- 동일한 방식을 다른 편집 모델에 적용했을 때도 65.5%의 정확도를 기록해 특정 모델에만 적용 가능한 기술이 아님이 확인됨
- 워터마크를 삽입하지 않은 이미지에서는 정확도가 사실상 0에 가깝게 떨어져, 복원 결과가 이미지 내용이 아니라 삽입된 워터마크에서 비롯된 것임이 입증됨

결론 및 시사점

• 기술적 의의 및 산업적 활용 가능성

- 앤젤핑거프린트는 이미지가 어떤 모델에서 생성되었는지를 넘어 어떤 편집이 이루어졌는지까지 확인할 수 있게 함으로써, 편집된 이미지를 이용한 허위정보 대응에 활용될 수 있음
- 워터마크가 모델 가중치에 통합되어 있어 별도 모듈이 필요하지 않으므로, 서비스 제공자는 워터마크를 삽입한 모델을 일반적인 미세조정 모델과 동일한 형태로 공개할 수 있음
- 다만 워터마크가 삽입된 이미지에는 미세한 격자 형태의 무늬가 남음. 일반적인 관찰로는 구별하기 어려운 수준이지만, 워터마크의 존재를 인지한 상태에서 자세히 들여다볼 경우 흔적이 드러날 수 있음
- 또한 검증이 두 종류의 편집 모델과 후보군 200개 규모를 대상으로 이루어졌기 때문에, 더 다양한 편집 모델과 대규모 후보군을 다루는 실제 운영 환경에서의 성능은 추가로 확인되어야 함

참고문헌

• Bo-Han Kung 외 4인, "AngelFingerprint: A Traceable, Explainable, and White-Box Stealthy Watermark for Text-Guided Image Editing", arXiv, 2026.09.04., https://arxiv.org/pdf/2609.04709
• T.J. Thomson, “More phones may soon let us prove photos are real – but will it solve the AI fake image crisis?, The Conversation, 2026.08.23., https://theconversation.com/more-phones-may-soon-let-us-prove-photos-are-real-but-will-it-solve-the-ai-fake-image-crisis-290051