산업 소니 그룹, AI 생성 음악의 원곡 기여도 산정 기술 개발

▶ 소니 그룹(Sony Group)은 AI가 생성한 음악에서 원곡을 식별하고, 원곡별 기여도를 산정하는 기술을 개발했다. 이 기술은 AI 생성 음악과 자사 음악 카탈로그 내 음원을 비교해 유사도를 분석하는 직접 비교 방식을 기본으로 한다. 하지만 AI 기업이 내부 시스템 접근을 허용할 경우에는 언러닝(unlearning) 기법을 활용해 보다 정밀한 분석이 가능하다고 밝혔다. 소니 그룹은 이 원곡 기여도 산정 기술을 바탕으로 향후 AI 생성 음악에서 발생하는 수익을 원곡 창작자의 기여 지분에 따라 분배하는 보상 체계를 구축하고자 한다.

산업 마이크로소프트, AI 산출물 출처 표시 정책 도입

▶ 생성형 AI 도구의 확산으로 오디오·영상·이미지 등 다양한 형태의 AI 산출물이 대량으로 생산되면서, 이용자들이 자신이 접하는 콘텐츠가 AI에 의해 생성되거나 편집된 것인지 여부를 식별하는 것은 점점 더 어려워지고 있다. 기존의 콘텐츠 관리 체계는 콘텐츠의 출처를 표시하는 것을 전제로 설계되지 않았기 때문에, AI 산출물이 유통되는 과정에서 출처 정보가 제외되는 문제가 발생해 왔다. 이러한 가운데, 마이크로소프트는 2026년 2월 AI 산출물에 워터마크와 메타데이터를 부여하여 생성 이력을 추적할 수 있도록 하는 새로운 출처 표시 정책을 발표했다. 이는 콘텐츠의 출처 정보를 사후에 확인하는 것이 아니라 생성 단계에서부터 기록하려는 시도로 평가된다.

산업 허위 계정과 우회 접속을 통한 지식 증류 기법

▶ 앤트로픽(Anthropic)은 중국의 AI 기업 딥시크(DeepSeek) 등이 허위 계정과 상업용 프록시 서비스를 활용해 자사 모델 클로드 (Claude)로부터 대규모 모델 기능 추출을 시도했다고 밝혔다. 이번 사안은 고성능 AI 모델의 출력과 추론 패턴을 활용해 소형 모델을 학습시키는 '지식 증류(Distillation)' 기법과 관련된 것으로 알려졌는데, 업계에서는 기법 자체보다 지역 제한이나 이용약관상 금지된 접근을 허위 계정과 우회 접속으로 회피한 방식이 주요 쟁점으로 논의되고 있다. 한편 일부 업계에서는 앤트로픽 역시 학습 데이터 수집 관련 저작권 분쟁을 겪어온 당사자라는 점에서 이번 문제 제기의 형평성에 대한 의문도 제기하고 있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 단일 3인칭 영상에서 1인칭 시점 영상을 생성하는 AI 기술 'EgoX'

▶ 최근 '1인칭 시점 영상'에 대한 수요가 확대되고 있으나, 접근성과 비용 측면에서 한계가 있었다. 기술 기업들은 이미 촬영된 3인칭 영상을 1인칭으로 변환하는 기술의 산업 적용도 시도해 왔으나, 다중 카메라 입력 등 부가 조건이 필요하거나 큰 시점 변환 시 화면 왜곡이 발생하는 등 기술적 제약이 존재했다. 최근 이러한 한계를 극복하기 위한 새로운 접근으로, 대규모 비디오 생성 AI 모델과 3D 공간 이해 기술을 결합해 단일 영상만으로 시점을 변환하는 시도가 이루어지고 있다. 일례로, KAIST 연구팀의 AI 모델 'EgoX'는 3인칭 영상 한 편에서 3D 공간을 복원하고 AI가 빈 영역을 채워 넣는 방식을 통해, 별도 장비 없이도 1인칭 시점 영상 생성이 가능함을 입증하였다.

산업 콘텐츠 유통 환경 변화에 따른 자동 콘텐츠 인식 시스템의 성장

▶ 자동 콘텐츠 인식(ACR) 시스템 시장은 2024년 38억 달러(약 5조 4,545억 원)에서 2032년 307억 9천만 달러(약 44조 1,959억 원) 규모로 성장할 전망이다. 자동 콘텐츠 인식 시스템은 콘텐츠의 시청각적 특징을 자동으로 분석해 저작권 침해 여부를 탐지하는 기술로, 다채널·온디맨드 플랫폼의 확대로 수동 관리가 어려워진 환경에서 새로운 저작권 관리 수단으로 주목받고 있다. 핵심 기술은 세 가지로 구분된다. 핑거프린팅은 콘텐츠의 고유한 시청각적 특징으로 식별자를 생성해 편집·재인코딩 이후에도 콘텐츠를 식별할 수 있으며, 디지털 워터마킹은 콘텐츠에 비가시적 표식을 삽입해 유통 이후에도 출처 확인을 가능하게 한다. 메타데이터 분석은 식별된 콘텐츠 정보를 권리 데이터베이스와 대조해 차단·수익 공유·저작권료 정산 등의 조치를 자동화한다.

기술 주간 기술 동향

▶ 최근 생성형 AI 모델이 저작권 콘텐츠를 거의 그대로 암기하여 재생산할 수 있다는 사실이 입증되면서, 학습된 데이터의 영향을 사후적으로 제거하는 머신 언러닝 기술이 주목받고 있다. 그러나 기존의 검증 방식은 모델의 최종 출력만을 평가하여 정보가 실제로 삭제되었는지 확인할 수 없다는 한계가 있다. 이번 연구에서 활용한 희소 오토인코더(SAE)는 신경망의 복잡한 내부 표현을 해석 가능한 개별 특징으로 분해하는 기법이다. 연구진은 SAE를 활용하여 머신 언러닝이 정보를 실제로 삭제하는지 아니면 출력 단계에서만 억제하는지를 구분하는 검증 프레임워크를 제안한다.

소니 그룹, AI 생성 음악의 원곡 기여도 산정 기술 개발

소니 그룹, AI 생성 음악의 원곡 기여도 산정 기술 개발

• 소니 그룹, 원곡 기여도 산정을 위한 기술 구현

- 소니 그룹(Sony Group)의 AI 연구개발 조직인 소니 AI(Sony AI)가 AI 생성 음악에서 원곡을 식별하고 원곡별 기여 비율을 산정할 수 있는 추적 기술을 개발함
- 2025년 12월, 소니 AI는 자사 블로그를 통해 현재 유통 중인 AI 생성 음악의 상당수가 창작자나 음반사의 허락 없이 수집된 음원 데이터를 기반으로 만들어지고 있다고 지적한 바 있음[1)]
- 이에 소니 AI는 자사 음악이 AI 생성물에 어떻게 반영되었는지를 창작자가 파악할 수 있는 기술을 개발해 왔으며, 최근 기술 구현 및 내부 검증을 완료함

소니 그룹의 원곡 기여도 산정 기술 및 추가 연구 성과

• 소니 그룹의 원곡 기여도 산정 기술: 직접 비교 방식과 언러닝 기법

- 소니 그룹이 개발한 직접 비교 방식은 AI 생성 음악을 소니 그룹이 보유한 음원과 비교해 유사도를 분석하는 방식을 기본으로 함
- 보도에 따르면, 분석 결과는 '비틀즈(Beatles) 30%, 퀸(Queen) 10%'와 같이 원곡별 기여 비율을 수치로 제시하는 형태로 출력됨[2)]
- 그러나 직접 비교 방식은 AI 모델이 실제로 학습한 데이터셋과 소니가 보유한 음악 카탈로그가 정확히 일치하지 않는다는 점에서, 분석 정확도에 한계가 있다는 지적이 제기됨
- 이에 대해 소니 그룹은 AI 기업이 AI 모델의 내부 시스템 접근을 허용하는 경우에 한해 보다 정밀한 분석이 가능한 언러닝(unlearning) 기법을 제시함
- 언러닝 기법은 AI 모델의 학습 데이터에 포함된 특정 음원으로부터 학습된 요소를 일시적으로 제거한 뒤, 모델이 생성하는 결과가 어떻게 달라지는지를 비교하는 방식임. 이때 생성물에서 나타나는 화성·멜로디·박자와 같은 음악적 요소의 변화 정도를 분석해 해당 음원의 기여도를 산정할 수 있음
- 소니 AI 연구팀은 115,000개 트랙으로 구성된 내부 데이터셋으로 학습한 모델을 대상으로 해당 기법의 성능을 검증했다고 밝혔음[3)]

1) Sony AI, “Protecting Creator’s Rights in the Age of AI”, Sony Ai, 2025.12.15., https://ai.sony/blog/Protecting-Creator%e2%80%99s-Rights-in-the-Age-of-AI/
2) Tomislav Bezmalinovic, “Sony technology identifies original songs in AI-generated music”, heise online, 2026.02.17., https://www.heise.de/en/news/Sony-technology-identifies-original-songs-in-AI-generated-music-11179510.html

- 이 과정에서 유사도 기반의 직접 비교 방식을 대조군으로 한 실험도 함께 진행했으며, 그 결과 언러닝 방식이 AI 생성물에 실제로 영향을 준 음원을 보다 정확하게 특정하는 것으로 나타남

[표1] 소니 그룹이 개발한 원곡 기여도 산정 기술의 방식 비교: 직접 비교 vs 언러닝 기법

구분	직접 비교 방식	언러닝 기법
적용 조건	• AI 기업이 내부 시스템 접근을 허용하지 않는 경우	• AI 기업이 내부 시스템 접근을 허용한 경우
분석 대상	• 소니의 음악 카달로그 내 음원들과 AI 생성 음악	• AI 모델의 학습 데이터와 AI 생성 음악
분석 방법	• 소니의 음악 카달로그 내 음원들과 AI 생성 음악 간 유사도를 직접 비교	• 학습 데이터 내 특정 음원으로부터 학습한 요소를 일시적으로 제거한 뒤 생성 결과물의 변화를 비교
산정 결과	• 유사도 분석 결과를 바탕으로 원곡별 기여도 수치화 (예: 비틀즈(Beatles) 30%, 퀸(Queen) 10%)	• 음악적 요소별(화성·멜로디·박자)로 변화의 폭이 클수록 제거된 음원의 기여도가 높은 것으로 판단해 수치화
정확도	• 실제 학습 데이터에 직접 접근하지 않는 만큼 정확도 차이가 발생할 가능성이 지적됨	• 소니 AI 연구팀의 실험 결과, 직접 비교 방식에 비해 기여도를 정확하게 산정하는 것으로 나타남

출처: 참고문헌 종합하여 재구성

• AI 생성 음악과 관련된 소니 그룹의 추가 연구 성과

- 소니 그룹은 앞서 소개된 원곡 기여도 산정 기술 외에도, AI 생성 음악과 원곡 간 유사성 탐지 분야에서 연구 성과를 발표한 바 있음
- 예를 들어, 소니 AI 연구팀은 편곡·리믹스 등을 통해 형태가 달라진 음악이 동일한 원곡을 기반으로 생성된 것인지를 식별하는 기술인 CLEWS(Contrastive Learning from Weakly-Labeled Segments)를 개발했다고 발표함
- 해당 기술은 전체 트랙 단위 비교뿐만 아니라 약 20초 단위의 짧은 구간에서도 유사성을 탐지할 수 있어, 기존 기술로는 포착하기 어려웠던 부분적 유사성까지 식별할 수 있는 것으로 평가됨

참고문헌

- Mandy Dalugdug, “Sony Group develops tech to track original music in AI-generated songs”, Music Business Worldwide, 2026.02.16., https://www.musicbusinessworldwide.com/sony-group-develops-tech-to-track-original-music-in-ai-generated-songs-report
- Sony AI, “Protecting Creator’s Rights in the Age of AI”, Sony Ai, 2025.12.15., https://ai.sony/blog/Protecting-Creator%e2%80%99s-Rights-in-the-Age-of-AI/
- Tomislav Bezmalinovic, “Sony technology identifies original songs in AI-generated music”, heise online, 2026.02.17., https://www.heise.de/en/news/Sony-technology-identifies-original-songs-in-AI-generated-music-11179510.html

3) Sony AI, “Protecting Creator’s Rights in the Age of AI”, Sony Ai, 2025.12.15., https://ai.sony/blog/Protecting-Creator%e2%80%99s-Rights-in-the-Age-of-AI/

마이크로소프트, AI 산출물 출처 표시 정책 도입

AI 산출물 확산에 따른 출처 표시 필요성 확대

• 업무 플랫폼 내 AI 산출물의 생산 확대와 출처 식별 체계 부재

- 마이크로소프트 365의 코파일럿(Copilot), 구글(Google)의 제미나이(Gemini) 등 생성형 AI 기능이 주요 업무 플랫폼에 통합됨에 따라, 오디오·영상·이미지 형태의 AI 산출물이 대량으로 생산되고 있음
- AI 산출물은 사람이 직접 제작한 콘텐츠와 외형적으로 구별이 어려우며, 생성 과정에 대한 정보가 콘텐츠 내에 포함되지 않아 이용자가 출처를 사전에 확인하기 어려움

• 주요국의 AI 콘텐츠 출처 표시 규제

- 현재 AI 산출물의 생성 여부나 출처를 식별할 수 있는 표준화된 표시 체계가 마련되어 있지 않아, 주요국들은 이러한 공백을 해소하기 위한 규제 도입을 추진 중임
- EU는 AI법(AI Act)* 제50조에서 AI 콘텐츠에 출처 표시를 의무화하였으며, AI 시스템 제공자에게는 기계가 자동으로 인식할 수 있는 형식의 표시를, 콘텐츠 배포자에게는 딥페이크 사용 여부의 공개 의무를 각각 부과함
- 미국에서도 딥페이크 책임법(Deepfakes Accountability Act)** 발의를 통해 합성 미디어에 대한 출처 공개와 AI 산출물의 표시 의무를 법제화하려는 시도가 이어지고 있음
- 이 같은 규제 환경 속에서 플랫폼 사업자가 자체적으로 AI 산출물의 출처 표시 기능을 도입하는 사례도 증가하고 있으며, 마이크로소프트 365의 워터마킹 정책 도입이 대표적인 사례에 해당함

* AI법(AI Act): 2024년 8월 제정된 유럽연합의 인공지능 규제 법률로, AI 시스템의 위험 수준에 따른 단계적 규제를 적용하며, AI 산출물의 투명성 확보를 주요 의무로 규정함
** 딥페이크 책임법(Deepfakes Accountability Act): AI로 생성된 합성 미디어에 대해 출처 표시 및 공개 의무를 핵심으로 한 법안

마이크로소프트 365 내 AI 산출물 출처 표시 정책

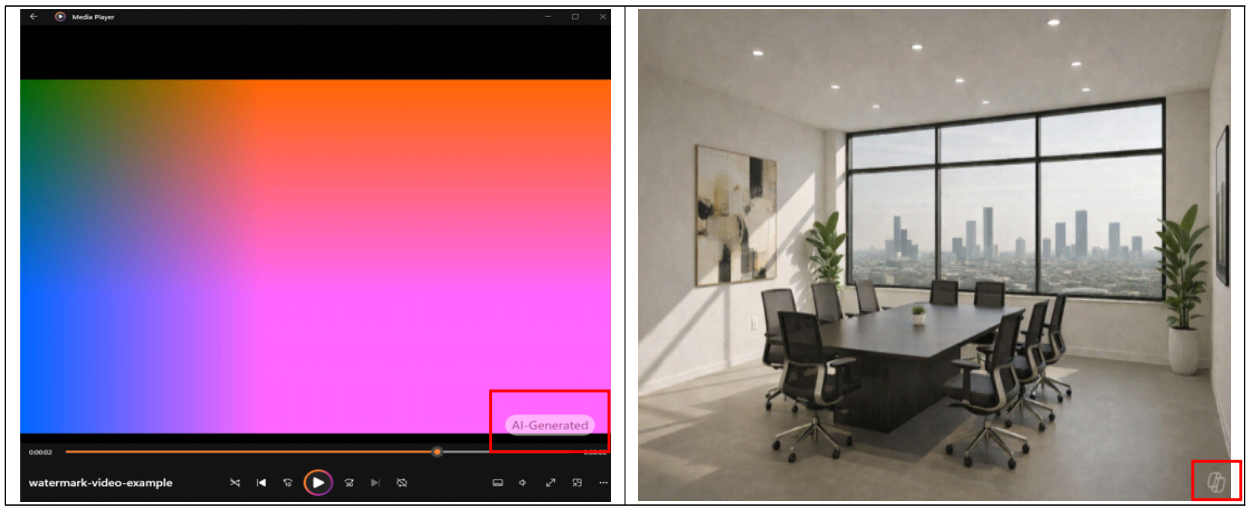
• 워터마크와 메타데이터를 결합한 이중 표시 체계 적용

- 마이크로소프트는 2026년 2월 중순 마이크로소프트 365에서 AI로 생성·편집된 오디오 콘텐츠의 출처를 워터마킹으로 표시하는 새로운 클라우드 정책을 공개함
- 영상 워터마크는 콘텐츠의 시작 또는 종료 구간에 'This audio is generated by AI'라는 음성 문구를 삽입하는 방식이며, 이미지 워터마크는 화면 우측 하단에 'AI-Generated' 문구 또는 코파일럿 아이콘 형태의 AI 생성 표시가 삽입됨
- 워터마크와 별도로, AI 콘텐츠에는 '콘텐츠 출처 및 진본성 연합(C2PA)*' 표준에 기반한 메타데이터가 자동으로 부여되어 사용된 AI 모델, 생성 앱, 생성 시점 등의 정보가 기록됨

- 해당 이중 표시 체계는 가시적·가청적 워터마크**를 통해 일반 이용자가 AI 생성 여부를 직관적으로 파악할 수 있도록 하고, 메타데이터에 포함된 식별 정보를 기반으로 플랫폼이나 검증 시스템이 기계적으로 출처를 확인할 수 있도록 함

* 콘텐츠 출처 및 진본성 연합(Coalition for Content Provenance and Authenticity, C2PA): 콘텐츠의 출처와 진본 여부를 기술적으로 검증하기 위한 개방형 국제 표준으로, 마이크로소프트, 어도비, 인텔, BBC 등 주요 기술·미디어 기업이 참여하여 개발함
** 가시적·가청적 워터마크: 화면에 표시되는 텍스트나 이미지 마크 등의 워터마크와 오디오나 영상 파일에 삽입된 출처 안내 음성 등의 워터마크

[그림 1] 마이크로소프트 365의 AI 생성 영상·이미지 워터마크 적용 예시



출처: Microsoft 365, "Add watermarks to content generated or altered by using AI in Microsoft 365", 2026.02.24., https://learn.microsoft.com/en-us/copilot/microsoft-365/watermarks

• 이미지 워터마킹 적용의 한계

- 마이크로소프트의 워터마킹 정책은 기본적으로 비활성 상태로 제공되며, IT 관리자가 클라우드 정책 서비스에서 수동으로 활성화해야만 적용되는 옵트인(opt-in)* 구조로 설계됨
- 워터마크의 문구와 배치는 마이크로소프트가 사전에 설정한 형태로 고정되어 있어, 관리자와 이용자 모두 워터마크의 내용이나 위치를 개별적으로 조정할 수 없음
- 이미지의 경우 관리자 정책을 통한 일괄 적용이 지원되지 않으며, 개별 사용자가 개인 계정 설정에서 직접 워터마크 기능을 활성화해야 하는 구조임
- 이에 따라 이미지 워터마킹은 조직 차원의 일괄 적용이 불가능하며, 사용자가 기능을 활성화하지 않으면 가시적 표시 없이 메타데이터만 부여되는 문제가 발생할 수 있음

* 옵트인(opt-in): 기본적으로 비활성화되어 있는 기능이나 이용자가 사전에 충분한 정보를 제공받은 후 명시적인 동의 의사표시를 통해 해당 기능·서비스의 적용 또는 참여를 선택하는 방식을 의미함

• 워터마크와 메타데이터의 구현 현황

- 메타데이터는 파일 내부에 삽입되어 일반 이용자가 직접 확인하기 어렵고 파일 변환·공유 과정에서 제거될 수 있는 반면, 가시적·가청적 워터마크는 콘텐츠 표면에 노출되어 보다 지속적인 출처 표시 수단으로 기능함
- 현재 메타데이터 자동 부여 기능은 이미지 콘텐츠에만 적용되고 있으며, 오디오·영상 콘텐츠에 대한 메타데이터 부여 기능은 개발 단계에 있음

시사점: AI 산출물 출처 투명성 확보를 위한 과제

• 옵트인 기반 출처 표시 정책의 적용 한계

- 옵트인 방식의 워터마킹 정책은 조직의 자율적 판단에 의존하는 구조로, 정책을 활성화하지 않은 조직에서는 AI 산출물에 가시적 출처 표시가 부여되지 않아 산업 전반의 일관된 워터마크 표시 수준을 확보하기 어려움
- 오디오·영상 콘텐츠는 관리자 정책을 통해 일괄 적용이 가능하나, 이미지 콘텐츠는 개별 사용자 설정에 위임되어 있어 동일 조직 내에서도 콘텐츠 유형별 출처 표시 체계에서 차이가 발생할 가능성이 있음
- 특히 메타데이터는 파일 변환이나 공유 과정에서 제거될 수 있어, 가시적 워터마크가 비활성화된 경우 이용자가 AI 산출물 여부를 인지할 수 있는 실질적 수단이 제한될 수 있음

• 플랫폼 간 상호운용이 가능한 출처 표시 표준과 검증 체계 정립의 필요성

- 개별 플랫폼의 자율 정책은 해당 서비스 내부에서만 유효하므로, AI 산출물이 외부로 유통되거나 타 플랫폼에서 재사용될 경우 출처 정보의 연속성이 단절되는 한계가 있음
- C2PA 등 개방형 출처 인증 표준의 채택이 확산될 경우, 플랫폼 간 상호운용이 가능한 검증 체계의 기반이 될 수 있으며, 개별 서비스의 정책적 공백을 보완하는 역할이 기대됨
- 다만 기술 표준 확산만으로는 실효성 있는 출처 관리가 이루어지기 어려우며, 규제 요건과의 정합성 확보 및 이용자의 인식 제고가 병행될 때 실질적인 거버넌스 체계로 작동할 수 있음

참고문헌

- Viktor Eriksson, “Microsoft 365 begins watermarking AI-generated content”, PCWorld, 2026.02.26., https://www.pcworld.com/article/3072414/microsoft-365-begins-watermarking-ai-generated-content.html
- Microsoft 365, “Add watermarks to content generated or altered by using AI in Microsoft 365”, 2026.02.24., https://learn.microsoft.com/en-us/copilot/microsoft-365/watermarks
- Red Grave LLP, “Legal 365 Snapshot: AI Watermarks and Metadata Coming to Microsoft 365”, 2026.02., https://www.redgravellp.com/publication/legal-365-snapshot-ai-watermarks-and-metadata-coming-to-microsoft-365
- Gleb Tkatchouk, “AI in Creative Industries: Impact on Art, Music, and Content Creation”, The AI Journal, 2026.01.08., https://aijourn.com/ai-in-creative-industries-impact-on-art-music-and-content-creation/

허위 계정과 우회 접속을 통한 지식 증류 기법

지식 증류 기법을 둘러싼 산업적 관행과 이용 규범

• 앤트로픽, 중국 AI 3사 대상 클로드 모델 무단 추출 의혹 제기[1)]

- 앤트로픽(Anthropic)은 2026년 2월 중국 AI 기업 딥시크(DeepSeek), 문샷 AI(Moonshot AI), 미니맥스(MiniMax)가 약 2만여 개의 허위 계정으로 자사 AI 모델 클로드(Claude)와 총 1,600만 건 이상의 상호작용을 통해 모델 기능을 대규모로 추출했다고 밝힘

- 기업별 추출 규모는 미니맥스 1,300만 회, 문샷 AI 340만 회, 딥시크 15만 회 이상으로 나타났으며, 클로드의 에이전트 추론·도구 활용·코딩 등 핵심 기술을 중심으로 모델 성능을 수집한 것으로 나타남

- 수집 대상은 기업별로 상이하며, 딥시크는 클로드 AI의 추론 과정 데이터, 문샷 AI는 이미지 인식 및 코딩 등 응용 기능, 미니맥스는 코드 생성과 도구 연계 등 에이전트 기반 수행 능력에 집중적으로 접근하여 수집한 것으로 파악됨

• 앤트로픽의 AI 모델 이용 규범

- 이번 추출에 사용된 지식 증류* 기법은 최첨단 AI 기업들이 모델 경량화에 폭넓게 활용하는 일반적인 방식이나, 앤트로픽을 포함한 주요 AI 기업들은 이용약관을 통해 타 기업의 자사 모델 출력값 수집 및 학습 활용을 금지하고 있음

- 또한, 앤트로픽은 보안상 이유로 클로드 API의 중국 내 상업적 접근을 허용하지 않으며, 중국 기업의 해외 자회사에도 동일한 이용 제한을 적용함

- 앤트로픽 측은 지식 증류 기법 자체는 모델 경량화 등을 위해 널리 활용되는 정상적인 학습 기법이라고 설명하면서도, 이번 사례의 문제는 허위 계정과 프록시 서비스**를 이용해 지역 접근 제한과 이용약관을 우회한 채 대규모 API*** 접근이 이루어진 데 있다고 밝힘

* 지식 증류(Knowledge Distillation): 고성능 대형 AI 모델의 출력·추론 패턴을 활용해 소형 모델을 학습시키는 머신러닝 기법으로, 고성능 모델의 예측 정확도를 최대한 보존하면서도 연산 비용과 모델 용량을 획기적으로 감축할 수 있어, AI 모델의 경량화 및 추론 효율성 제고를 위한 핵심 방법론으로 활용됨

** 프록시 서비스(Proxy Service): 이용자의 실제 위치나 신원을 숨긴 채 다른 서비스에 접근할 수 있도록 트래픽을 중계해 주는 서비스

*** API(Application Programming Interface): 소프트웨어 간 데이터 교환을 가능하게 하는 인터페이스

허위 계정·우회 접속을 통한 무단 추출 구조와 대응 한계

• 상업용 프록시 서비스를 활용한 접근 우회 구조

- 딥시크, 문샷 AI, 미니맥스 3사는 중국 내 접근 제한을 우회하기 위해 클로드 이용권을 재판매하는 상업용 프록시 서비스를 활용하여 앤트로픽 API 및 제3자 클라우드 플랫폼에 대규모로 접근함

- 해당 서비스는 하나의 프록시 네트워크가 동시에 2만 개 이상의 허위 계정을 운영하는 구조로 설계되어, 데이터 추출 요청을 일반 이용자 요청과 혼합해 탐지를 회피하는 방식이 사용됨[2)]

1) Anthropic, “Detecting and preventing distillation attacks”, Anthropic, 2026.02.23., https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks
2) Anthropic, “Detecting and preventing distillation attacks”, Anthropic, 2026.02.23., https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

- 일례로 딥시크는 여러 계정에서 동일한 결제 수단을 사용하고 접속 시간을 일치시켜, 정상적인 운영을 하는 일반 기업의 공동 계정이 보이는 패턴과 유사한 패턴을 보이게 한 것으로 나타남

• 앤트로픽의 방어 전략

- 앤트로픽은 정상 이용과 대규모 데이터 추출 시도를 구별하기 위해 계정별 이용 패턴, 프롬프트 구조, 접속 타이밍 등을 분석하는 행동 기반 탐지 시스템과 교차 계정 상관 분석 도구를 활용했음
- 또한 추론 과정 데이터를 수집하려는 시도에 대응해, 연쇄 사고* 유도 프롬프트를 탐지하는 분류기를 방어 체계에 포함함
- 허위 계정 개설에 빈번하게 악용되던 교육용 계정, 보안 연구 프로그램, 스타트업 조직 등에 대한 신원 인증 요건도 강화함
- 무단 추출 시 응답의 정보 가치를 낮추도록 출력값 형태를 조정하는 모델 및 API 수준의 대응책도 개발 중임. 다만 일반 이용자의 편의를 고려할 때 적용 범위에는 한계가 있을 것으로 보임
- 앤트로픽은 탐지 지표를 다른 AI 기업, 클라우드 사업자, 관계 당국과 공유하고 있으며, 단일 기업 대응에는 한계가 있는 만큼 업계 공동 대응과 정책적·국제적 공조가 필요하다고 강조함

* 연쇄 사고(Chain-Of-Thought, CoT): AI 모델이 최종 답변에 이르기까지의 단계별 추론 과정을 명시적으로 서술하도록 유도하는 프롬프트 기법으로 해당 출력이 강화학습 훈련 데이터로 활용될 수 있음

• 이용약관과 형평성 논란

- 이번 사례는 AI 모델 출력값의 활용 범위와 이용약관 적용을 둘러싼 논쟁으로 확산되고 있으며, 타사 모델을 활용한 학습 방식의 허용 범위라는 산업적 쟁점으로 이어지고 있음
- 업계 일각에서는 앤트로픽 역시 과거 학습 데이터 수집 관련 저작권 분쟁을 겪었다는 점을 들어, 타사의 지식 증류를 공격으로 규정하는 것에 형평성 문제를 제기함[3)]
- 앤트로픽은 안전장치 없이 추출된 파생 모델이 사이버 공격, 생화학무기 개발, 허위정보 유포 등에 악용될 수 있다고 경고하며, 이번 사안을 단순한 약관 위반을 넘어 국가 안보 차원의 문제로 규정함
- 또한 대규모 지식 증류 시도에는 고성능 반도체 확보가 필요하다는 점에서, 이번 사안이 반도체 수출 통제 정책 논의와도 연결될 수 있다고 지적함

시사점: AI 모델 출력값 권리 보호 논의와 과제

• AI 모델 보호를 위한 기술·법제·국제 공조의 필요성

- 앤트로픽의 사례와 함께 오픈AI(OpenAI)도 같은 달 미국 하원에 딥시크의 기술 탈취 시도와 관련한 서한을 제출했으며, 구글(Google)의 제미나이(Gemini)를 포함한 주요 AI 기업 전반에서 유사한 무단 추출 시도가 확인되는 등 미·중 AI 경쟁 맥락에서 산업 차원의 공동 대응 필요성이 제기됨
- 현재 대부분의 주요 AI 기업은 이용약관을 통해 지식 증류를 통한 모델 학습을 금지하고 있으나, 허위 계정과 프록시 우회에 대한 기술적 탐지와 법적 집행 수단이 뒷받침되지 않는 한 약관 규정만으로는 실질적인 억제 효과를 기대하기 어려움
- AI 모델 출력값의 파생 활용 범위를 어떻게 규율할 것인지는 이용약관·지식재산권·안보 정책이 교차하는 영역으로, 단일 기업이나 단일 국가 차원을 넘어 국제적 기준 논의로 이어질 가능성이 있음

3) 硅星人, “Anthropic’s China Allegations, Tailored for an Audience of One: Washington”, The China Academy, 2026.02.26., https://thechinaacademy.org/anthropics-china-allegations-tailored-for-an-audience-of-one-washington/

참고문헌

- Rebecca Bellan, “Anthropic accuses Chinese AI labs of mining Claude as US debates AI chip exports”, Tech crunch, 2026.02.23., https://techcrunch.com/2026/02/23/anthropic-accuses-chinese-ai-labs-of-mining-claude-as-us-debates-ai-chip-exports/
- Anthropic, “Detecting and preventing distillation attacks”, Anthropic, 2026.02.23., https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks
- Michael Nuñez, “Anthropic says DeepSeek, Moonshot, and MiniMax used 24,000 fake accounts to rip off Claude”, VB, 2026.02.23., https://venturebeat.com/technology/anthropic-says-deepseek-moonshot-and-minimax-used-24-000-fake-accounts-to
- Ravie Lakshmanan, “Anthropic Says Chinese AI Firms Used 16 Million Claude Queries to Copy Model”, The Hacker News, 2026.02.24., https://thehackernews.com/2026/02/anthropic-says-chinese-ai-firms-used-16.html
- 硅星人, “Anthropic’s China Allegations, Tailored for an Audience of One: Washington”, The China Academy, 2026.02.26., https://thechinaacademy.org/anthropics-china-allegations-tailored-for-an-audience-of-one-washington/
- 김근철, “앤스로픽 "中 AI 3사, 클로드 모델 사기 계정 만들어 제품 개선…안보 위협"”, 뉴스핌, 2026.02.24., https://www.newspim.com/news/view/20260224000018

단일 3인칭 영상에서 1인칭 시점 영상을 생성하는 AI 기술 'EgoX'

1인칭 시점 영상에 대한 수요 확대와 기존 기술의 한계

• 산업 전반에서의 1인칭 시점 영상 수요 확대

- 최근 영상 기술 분야에서는 사용자가 실제로 보는 시야를 그대로 재현하는 1인칭 시점 영상에 대한 관심과 활용 수요가 다양한 산업 분야에서 확대되고 있음
- 대표적으로 증강현실(AR), 가상현실(VR), 메타버스 등의 분야에서는 사용자가 장면 안에 들어가 있는 듯한 몰입감을 구현하기 위해 1인칭 시점 영상*의 활용이 늘어남
- 로봇 분야에서도 사람의 행동을 관찰하고 모방하는 학습 방식이 확산되면서, 사람이 직접 바라보는 시점을 담은 1인칭 데이터가 핵심 학습 자원으로 활용되고 있음
- 이 밖에도 스포츠 중계를 선수의 시점에서 제공하거나, 일상이나 경험을 영상으로 기록해 공유하는 브이로그를 주인공의 시점으로 촬영하는 등 영상 콘텐츠 제작과 소비 방식에서도 1인칭 시점에 대한 관심이 커지고 있음

* 1인칭 시점 영상(Egocentric video): 촬영자 또는 영상 속 인물이 직접 바라보는 시야를 그대로 담은 영상으로, 3인칭(관찰자) 시점 영상과 구분됨

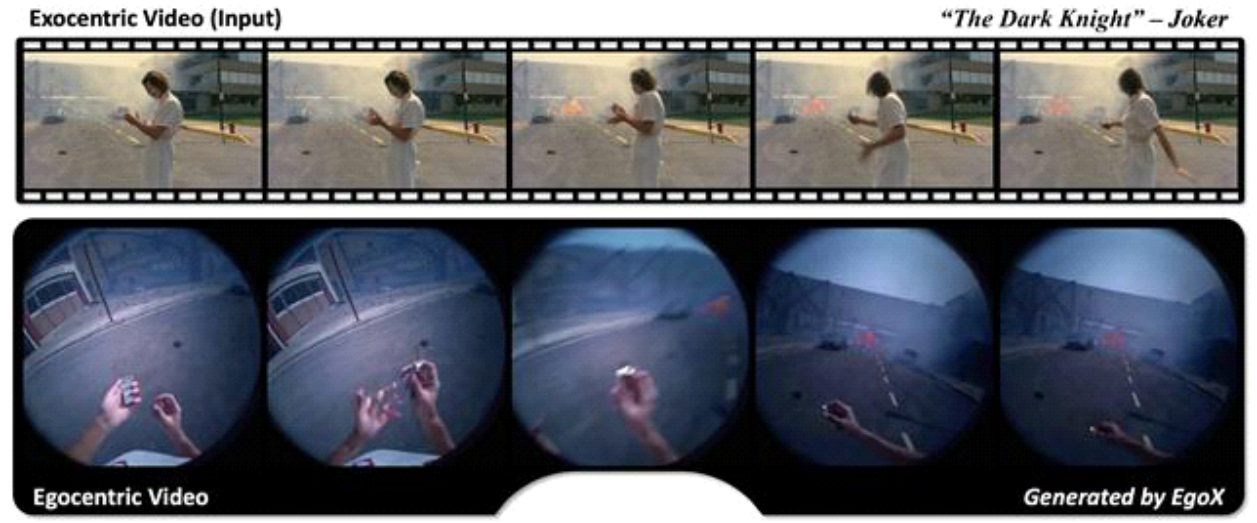
• 기존 1인칭 영상 확보 방식의 구조적 제약

- 이처럼 1인칭 시점 영상에 대한 수요는 커지고 있지만, 고품질 1인칭 영상을 확보하려면 촬영자가 고가의 액션캠이나 스마트 글래스(Smart Glasses)* 등 전용 웨어러블 장비를 직접 착용해야 해 장비 비용과 착용 부담으로 활용 범위가 제한되어 왔음
- 업계에서는 이미 촬영된 3인칭(관찰자 시점) 영상을 1인칭으로 변환하는 기술이 시도되어 왔으나, 4대 이상의 카메라로 동시에 촬영한 영상을 함께 사용하거나 1인칭 첫 프레임을 별도로 제공해야 하는 등 부가 조건이 필요했음
- 이러한 제약을 줄이기 위해 AI를 활용해 카메라 시점을 변환하는 카메라 제어 모델**도 시도되어 왔으나, 이 기술은 소폭의 시점 이동에 최적화되어 있어 3인칭에서 1인칭으로의 큰 시점 변환에서는 화면 왜곡이나 시간적 불일치가 발생하는 한계가 있었음
- 결국 1인칭 시점 영상에 대한 산업적 수요는 커지고 있으나, 별도 장비 없이 기존 영상만으로 고품질 1인칭 시점 영상을 생성하는 기술은 아직 충분히 성숙하지 못한 단계였으며, 이를 해결하기 위한 다양한 접근이 시도되어 왔음

* 스마트 글래스(Smart Glasses): 안경 형태의 웨어러블 기기로, 카메라·디스플레이·센서 등을 내장하여 착용자의 시야를 촬영하거나 증강현실 정보를 표시하는 장치

** 카메라 제어 모델: AI를 활용해 영상의 카메라 시점을 변경하거나 이동시키는 기술로, 주로 소폭의 시점 변화에 최적화되어 있음

[그림 1] EgoX를 통해 3인칭 시점 영상을 1인칭 시점 영상으로 변환한 예시



출처: KAIST, "다크나잇 속 조커가 '내가 되는' 기술 개발", 2026.02.23., https://www.kaist.ac.kr/news/html/news/?mode=V&mng_no=58490&skey=keyword&sval=%45%67%6f%58&list_s_date=&list_e_date=&GotoPage=1

EgoX의 핵심 접근 방식과 기존 기술 대비 차별점

• 단일 영상 입력만으로 시점 변환을 가능하게 한 핵심 설계 원리

- KAIST 김재철AI대학원 연구팀은 관찰자 시점 영상 한 편만으로 영상 속 인물이 실제로 보았을 장면을 예측해 변환 생성하는 AI 모델 'EgoX'를 개발하고, 2025년 12월 관련 논문을 arXiv에 선공개함
- EgoX는 '3D 공간 복원을 통한 초안 영상 생성'과 'AI 모델을 통한 빈 영역 보완'이라는 두 단계를 결합하여, 기존 기술이 요구하던 다중 카메라 촬영이나 1인칭 첫 프레임 제공 없이 단일 영상만으로 시점 변환을 수행함
- 먼저 첫 번째 단계에서는 주어진 3인칭 영상에서 각 장면의 깊이 정보를 추정해 3D 공간을 복원한 뒤, 인물의 눈 위치에서 다시 바라본 '초안 영상'을 생성함
- 이 초안 영상은 빈 영역이 많고 화질이 불완전하지만, 두 번째 단계에서 비디오 생성 AI 모델*이 이를 참고 자료로 활용하여 빈 영역을 자연스럽게 채워 최종 영상을 완성함

* 비디오 생성 AI 모델(Video Diffusion Model): 노이즈로부터 점진적으로 영상을 복원하는 방식으로 고품질 영상을 생성하는 AI 기술로, 대규모 영상 데이터로 사전 학습하여 장면의 시공간적 특성을 이해함

• 3인칭과 1인칭의 시야 차이를 구분해 처리하는 방식

- 단일 영상으로 시점을 변환할 때 가장 큰 난제는 3인칭과 1인칭의 시야가 거의 겹치지 않아, 1인칭에서 보여야 할 장면의 상당 부분이 원본 영상에 존재하지 않는다는 점임
- 이 문제를 해결하기 위해 EgoX는 원본 3인칭 영상의 각 영역을 '1인칭과 대응되는 영역', '부분적으로만 관련 있는 영역', '전혀 관련 없는 배경'으로 AI가 자동 구분하도록 설계됨
- 구체적으로는 1인칭 시점으로 변환했을 때 카메라가 향하게 될 방향을 기준으로, 원본 영상의 각 영역이 1인칭 시점과 얼마나 가까운지를 자동으로 계산하여, 관련성이 높은 영역은 적극 반영하고 무관한 영역은 생성 과정에서 참고하지 않도록 처리함

- 이러한 선별적 처리를 통해, 원본 영상에 담기지 않은 부분을 AI가 새로 생성할 때도 실제 공간 구조에 맞는 영상이 생성되고, 인물이 고개를 돌릴 때 그에 따라 시야가 전환되는 장면도 자연스럽게 재현됨

시사점: EgoX의 산업적 활용 가능성과 향후 과제

• 기존 영상 자산의 재활용을 통한 산업적 활용 가능성

- EgoX의 가장 큰 기술적 의의는 이미 촬영되어 축적된 대량의 3인칭 영상을 1인칭 영상으로 변환할 수 있다는 점으로, 별도 촬영 없이도 기존 영상 자산의 활용 범위를 넓혀줌
- 예를 들어 3인칭으로 촬영해 둔 스포츠 경기나 수술 교육 영상을 사후에 1인칭으로 재구성하는 것이 가능할 것으로 기대됨. 이를 통해 추가 촬영 없이도 새로운 형태의 콘텐츠 제작이 가능해질 전망임
- 앞서 언급한 로봇 모방 학습(Imitation Learning) 분야에서도 기존에 축적된 3인칭 영상을 변환해 1인칭 학습 데이터로 활용할 수 있어, 데이터 확보에 드는 장비 비용과 촬영 부담을 줄일 수 있음
- 이처럼 EgoX는 새로운 영상을 별도로 촬영하지 않고도 기존 영상을 다른 시점으로 재활용할 수 있게 한다는 점에서, 향후 영상 콘텐츠의 제작·소비 방식에 변화를 가져올 것으로 전망됨

• 현재 기술의 제약 조건과 향후 발전 방향

- 해당 연구는 2025년 12월 arXiv에 선공개된 이후 단일 영상만으로 고품질 시점 변환을 구현했다는 점에서 엔비디아(NVIDIA), 메타(Meta) 등 주요 빅테크의 주목을 받았으며, 2026년 6월 국제 학술대회 CVPR*에 공식 발표될 예정임. 다만 실용화를 위해서는 몇 가지 과제가 남아 있음
- 우선 EgoX는 변환하려는 1인칭 시점의 위치와 방향을 사용자가 수동으로 지정해야 하는데, 연구팀은 영상 속 인물의 머리 방향을 자동으로 추정해 이를 대체하는 모듈의 결합을 향후 과제로 제시함
- 또한 인물이 카메라를 등지거나 빠르게 움직이는 등 모호한 장면에서는 동작을 잘못 해석하는 사례가 일부 확인되어, 다양한 환경에서의 안정성을 높이는 작업도 필요함
- 이러한 과제들이 해결될 경우 별도 장비나 수동 설정 없이도 기존 영상에서 1인칭 시점을 자동으로 생성하는 것이 가능해져, 관련 기술의 실용화 가능성이 보다 높아질 것으로 기대됨

* CVPR(Conference on Computer Vision and Pattern Recognition): 컴퓨터 비전 분야의 최상위 국제 학술대회로, IEEE/CVF가 주관하며 매년 개최됨

참고문헌

- 장세민, “KAIST, ‘1인칭 시점 영상’ 생성하는 AI 개발…”로봇 학습 데이터로 활용””, AI Times, 2026.02.23., https://www.aitimes.com/news/articleView.html?idxno=207104
- KAIST, “다크나잇 속 조커가 ‘내가 되는’ 기술 개발”, 2026.02.23., https://www.kaist.ac.kr/news/html/news/?mode=V&mng_no=58490&skey=keyword&sval=%45%67%6f%58&list_s_date=&list_e_date=&GotoPage=1
- 강태웅 외 5인, “EgoX: Egocentric Video Generation from a Single Exocentric Video”, arXiv, 2025.12.09., https://arxiv.org/abs/2512.08269

콘텐츠 유통 환경 변화에 따른 자동 콘텐츠 인식 시스템의 성장

자동 콘텐츠 인식 시스템의 성장 배경

• 콘텐츠 유통 플랫폼의 확대에 따른 자동 콘텐츠 인식 시스템 시장 성장

- TV·스트리밍·소셜미디어 등 다채널 플랫폼과 넷플릭스(Netflix), 핫스타(Hotstar), 유튜브(YouTube), 아마존 프라임(Amazon Prime) 등 온디맨드* 플랫폼이 확대되면서, 콘텐츠 유통 경로가 방송 중심에서 디지털 플랫폼 중심으로 빠르게 변화하고 있음
- 이러한 환경에서는 동일한 콘텐츠가 다양한 플랫폼과 디바이스를 통해 동시에 소비되기 때문에, 콘텐츠가 유통된 이후 저작권 침해 여부를 파악하는 것은 점점 더 어려워지고 있음
- 이에 따라 콘텐츠의 시청각적 특징을 자동으로 분석하여 콘텐츠를 식별하는 자동 콘텐츠 인식(Automatic Content Recognition, ACR) 시스템이 새로운 저작권 관리 수단으로 주목받고 있음
- 미국의 시장조사업체 베리파이드마켓리서치(Verified Market Research)의 보고서에 따르면, 자동 콘텐츠 인식 시스템의 시장 규모는 2024년 38억 달러(원화 약 5조 4,545억 원)[1]에서 2032년 308억 달러(원화 약 44조 1,839억 원)로 성장할 것으로 전망되며, 연평균 성장률**은 약 29.9%로 예측됨[2]

* 온디맨드(On-demand): 이용자가 원하는 시간에 원하는 콘텐츠를 선택해 소비할 수 있는 방식으로, 정해진 편성표 없이 콘텐츠를 제공하는 서비스 형태

자동 콘텐츠 인식 시스템의 핵심 기술

• 콘텐츠의 고유 특징을 기반으로 식별자를 생성하는 핑거프린팅 기술

- 핑거프린팅(Fingerprinting)은 콘텐츠의 고유한 시청각적 특징을 분석하여 고유 식별자를 생성하는 기술로서, 이를 데이터베이스와 대조하여 콘텐츠를 식별하는데 사용할 수 있음
- 오디오 핑거프린팅은 음성 신호의 주파수 스펙트럼, 음압 패턴, 시간대별 음향 특성 등을 분석하여 고유 식별자를 생성하는 방식으로, 코덱이나 압축 방식이 달라져도 동일한 콘텐츠를 인식할 수 있음
- 이 기술은 음악·방송 프로그램·광고 등 다양한 오디오 콘텐츠의 무단 사용 여부를 자동으로 탐지하는 데 활용되며, 방송 음악 모니터링 시스템이나 음원 저작권 관리 체계에서 널리 사용되고 있음
- 영상 핑거프린팅(Video Fingerprinting)은 영상 프레임의 색상 분포, 장면 전환 패턴, 움직임 벡터 등 시각적 특징을 기반으로 식별자를 생성하는 방식으로, 영상 콘텐츠의 편집이나 재인코딩이 발생하더라도 동일한 콘텐츠를 인식할 수 있도록 설계됨
- 이를 통해 TV 쇼, 영화, 스포츠 등 영상 콘텐츠의 무단 송출이나 재배포 여부를 자동 탐지할 수 있음

1) 1달러=1,435.40원(2026.03.03, KEB 하나은행 최초 매매기준율 적용)
2) Verified Market Research, "Automatic Content Recognition (ACR) Market Valuation – 2026-2032", 2025.06., https://www.verifiedmarketresearch.com/product/automatic-content-recognition-acr-market/

• 디지털 워터마킹을 통한 콘텐츠 출처 확인 및 권리 정보 관리

- 워터마킹은 콘텐츠 배포 이전 단계에서 디지털 표식을 삽입하여 콘텐츠 유통 이후에도 콘텐츠의 출처를 추적할 수 있도록 하는 기술임
- 디지털 워터마킹은 인간의 눈이나 귀로는 인식하기 어려운 형태의 신호를 영상이나 음성 데이터에 삽입하며, 플랫폼 간 이동이나 압축 과정에서도 보존될 수 있도록 설계됨
- 이를 통해 콘텐츠가 재업로드되거나 편집된 이후에도 워터마킹을 통해 권리 정보를 확인할 수 있으며, 콘텐츠의 진위 여부를 확인 할 수 있음
- 또한 방송사나 콘텐츠 제작사는 워터마킹 기술을 활용해 사용자가 특정 콘텐츠를 시청했는지 여부를 감지할 수 있음

• 메타데이터 분석과 데이터베이스 매칭을 통한 콘텐츠 권리 확인

- 자동 콘텐츠 인식 시스템은 핑거프린팅이나 워터마킹을 통해 식별된 콘텐츠 정보를 메타데이터* DB와 대조하여 콘텐츠의 권리 정보를 확인하는 방식으로 작동함
- 메타데이터에는 콘텐츠 제목, 제작자, 저작권자, 발행일, 라이선스 범위, 유통 지역 등 다양한 권리 관리 정보가 포함되며, 이러한 데이터가 권리 관리 시스템과 연결되어 자동으로 처리됨
- 플랫폼 사업자는 콘텐츠 식별 결과를 바탕으로 저작권 침해 여부를 판단하여, 콘텐츠 차단, 수익 공유, 저작권료 정산 등의 조치를 자동으로 수행할 수 있음
- 이 과정에서 메타데이터의 정확성과 표준화 수준은 자동화된 저작권 관리 체계의 효율성을 좌우하는 핵심 요소로 작용함

* 메타데이터(Metadata): 콘텐츠 자체가 아니라 콘텐츠를 설명하는 부가 정보로, 제목·제작자·저작권자·발행일 등 콘텐츠의 속성과 권리 정보를 포함하는 데이터

[표1] ACR 핵심 기술 비교

구분	오디오·영상 핑거프린팅	디지털 워터마킹	메타데이터 분석
작동 방식	콘텐츠의 시청각 특징을 분석해 고유 식별자 생성 후 DB와 매칭	콘텐츠 내에 비가시적 정보 신호 삽입	콘텐츠 설명 정보를 데이터베이스와 대조
주요 분석 요소	오디오 스펙트럼, 영상 프레임 특징, 장면 패턴	삽입된 디지털 코드 또는 신호	제목, 제작자, 저작권자, 발행일, 권리 정보
장점	편집·재인코딩 이후에도 콘텐츠 식별 가능	콘텐츠 출처 및 유통 경로 추적 가능	권리 정보 관리 및 라이선스 관리 용이

출처: 참고문헌 종합

자동 콘텐츠 인식 시스템의 현황과 과제

• 자동 콘텐츠 인식 시스템의 정확성·표준화 문제

- 자동 콘텐츠 인식 시스템은 콘텐츠 식별, 권리 판정, 침해 대응, 저작권료 정산까지 이어지는 자동화된 저작권 관리 체계를 구축하는 데 활용되고 있음
- 자동 콘텐츠 인식 시스템은 콘텐츠의 시청각적 특징을 기반으로 식별하기 때문에, 인용이나 패러디 등 허용된 이용 형태가 포함된 편집·변형 콘텐츠까지 저작권 침해 콘텐츠로 인식하는 문제가 발생하고 있음
- 또한 자동 콘텐츠 인식 시스템은 메타데이터의 정확성과 표준화 수준에 크게 의존하기 때문에, 메타데이터 형식과 관리 체계가 통일되어 있지 않을 경우 권리 정보의 정확한 매칭이 어려워질 수 있음
- 이에 따라 플랫폼 차원의 이의 제기 절차를 표준화하고 그 처리 과정을 투명하게 공개하는 것이 향후 주요 과제로 제시됨
- 또한 자동화된 저작권 관리 체계의 실효성을 높이기 위해서는 플랫폼 간 메타데이터 형식의 통일과 산업 차원의 공통 기준 마련이 필요하며, 이를 위한 사업자 간 협력 및 제도적 논의가 요구됨

참고문헌

• Verified Market Research, “Automatic Content Recognition (ACR) Market Valuation – 2026-2032”, 2025.06., https://www.verifiedmarketresearch.com/product/automatic-content-recognition-acr-market/
• Gabrel Patrick, “Top 7 automatic content recognition services revolutionizing media and entertainment”, Verified Market Research, 2026.02., https://www.verifiedmarketresearch.com/blog/top-automatic-content-recognition-services/

주간 기술 동향

SAE 기반 머신 언러닝의 검증

• 생성형 AI 확산과 개인정보 보호 규제 강화 속 머신 언러닝 기술의 부상

생성형 AI 모델이 학습 데이터를 단순히 패턴화하는 것이 아니라 저작권으로 보호되는 콘텐츠를 거의 그대로 암기하여 재생산할 수 있다는 사실이 실증적으로 입증되면서, AI 기업들이 제시해 온 저작권 관련 방어 논리에 대한 의문이 제기되고 있다. 스탠포드대학교와 예일대학교 연구진은 2026년 1월 클로드 3.7 소네트 모델에서 해리포터 원문의 95.8%를 단 55달러(한화 약 78,947원)[1]로 추출하는 데 성공했으며, 제미나이 2.5 프로와 그록 3는 안전장치 우회 없이도 저작권 텍스트를 연속적으로 그대로 출력했다. 이러한 문제에 대응하기 위해 학습 데이터의 영향을 사후적으로 제거하는 머신 언러닝(Machine Unlearning) 기술이 주목받고 있다. 한편, 업계 일각에서는 언러닝이 실제로 작동했는지에 대한 의구심도 제기되고 있으며, 이에 언러닝 효과를 검증하는 방법론에 대한 논의도 이뤄지고 있다.

기존의 머신 언러닝 검증 방법들은 주로 분류 정확도 하락, 멤버십 추론 공격(Membership Inference Attack) 성공률 감소 등 모델의 최종 출력을 기반으로 평가하는 방식에 의존해왔다. 그러나 이러한 출력 기반 접근법은 모델이 특정 범주에 대한 내부 표현을 여전히 보유하고 있더라도 최종 출력 단계에서만 해당 정보를 억제하면 머신 언러닝이 성공한 것처럼 보이는 허점을 가지고 있다. 실제로 다수의 머신 언러닝 기법들이 모델의 최종 분류층이나 출력 확률 분포만을 조정하는 방식으로 구현되어 있어, 중간 표현층에는 삭제 대상 데이터의 특징이 그대로 남아있을 가능성이 제기되고 있다.

이에 따라 모델의 내부 수준에서 머신 언러닝의 효과를 검증할 수 있는 새로운 분석 프레임워크의 필요성이 증대되고 있으며, 신경망 해석 기법을 활용한 다양한 검증 방법론이 연구되었다. 그중, 희소 오토인코더(Sparse Autoencoder, 이하 SAE)는 신경망의 중간층 활성화 패턴을 해석 가능한 독립적 특징으로 분해할 수 있는 기법으로, 최근 LLM의 내부 작동 메커니즘을 이해하는 도구로 주목받고 있다.

본 보고서에서는 머신 언러닝 검증을 위한 두 가지 핵심 기술적 접근을 분석한다. 첫째, SAE를 활용하여 모델의 중간층에서 클래스별 특징을 추출하고 추론 시점 조향(inference-time steering)을 통해 억제된 표현을 복원함으로써 머신 언러닝의 실제 효과를 표현 수준에서 검증하는 복원 기반 분석 프레임워크를 다룬다. 둘째, 12개 주요 머신 언러닝 기법들을 평가하여 출력 기반 평가를 통과한 모델들조차 표현 수준에서는 삭제 대상 정보를 여전히 보유하고 있음을 실증한 대규모 비교 실험 결과를 검토한다.

1) 1달러=1,435.40원(2026.03.03, KEB 하나은행 최초 매매기준율 적용)

머신 언러닝 검증의 한계

• 출력 단계의 결과물만 검증이 가능했던 언러닝 검증 기법의 한계

- 기존의 언러닝 검증 방식은 최종 출력만 평가하여 분류 정확도와 멤버십 추론 공격 성공률을 통해 언러닝 성공 여부를 판단하지만, 모델 내부에 해당 데이터 표현이 남아있는지 확인할 수 없음
- 이는 기존에 나온 다수의 언러닝 기법은 최종 레이어나 출력 확률 분포만 조정하는 방식으로 구현되어 있기 때문이며, 중간 레이어에는 삭제 대상 범주의 특징이 남아있을 가능성이 있음
- 표현 수준 검증 없이 출력 기반 평가만으로 언러닝 효과를 판단하면 적대적 공격이나 미세 조정을 통해 억제된 정보가 재활성화될 위험이 있어 개인정보 보호나 저작권 대응에 한계가 있음

[사례] SAE 기반 표현 수준 머신 언러닝 검증 프레임워크

• SAE 기반 언러닝 검증 기술 개요 및 배경

- 희소 오토인코더(Sparse Autoencoder, SAE)는 인공신경망의 복잡한 내부 표현을 해석 가능한 개별 특징들로 분해하는 기법으로, 각 특징이 독립적인 의미를 가지도록 학습하여 모델이 특정 정보를 어떻게 저장하고 처리하는지 분석할 수 있게 함
- 본 프레임워크는 머신 언러닝이 언러닝 타깃 데이터를 모델의 출력 단계에서만 억제하는지, 아니면 내부 표현 수준에서 실제로 삭제하는지를 구분하기 위해 개발되었으며, SAE와 추론 시점 조향 기법을 결합하여 억제된 정보의 복원 가능성을 측정함
- 기존 검증 방식은 분류 정확도나 멤버십 추론 공격 성공률 등 최종 출력만을 평가하여 내부에 정보가 남아있는지 확인할 수 없었으나, 본 기술은 모델의 중간 레이어에서 범주별 특징을 추출하고 이를 복원하여 실제 삭제 여부를 판별함
- 이미지 분류 벤치마크 데이터셋인 CIFAR-10(10개 범주, 6만 장)과 ImageNette(10개 범주, 1만 3천 장)에서 12개의 주요 머신 언러닝 기법을 대상으로 실험을 수행하여, 출력 기반 평가를 통과한 기법들조차 표현 수준에서는 삭제 대상 정보를 보유하고 있음을 실증함

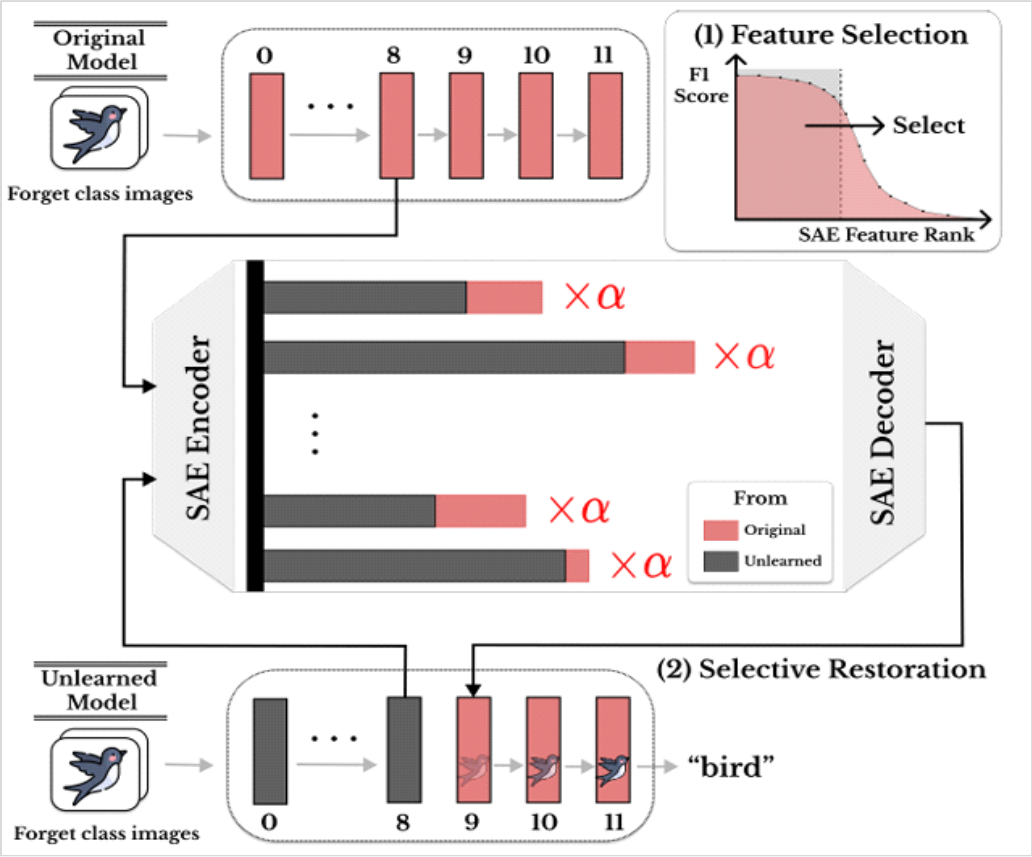
• 범주별 특징 추출 및 식별 과정

- 본 연구는 ResNet-18이라는 이미지 분류 모델을 사용하며, 이 모델이 이미지를 분류하기 직전 단계에서 생성하는 내부 신호를 SAE로 분석하여 각 범주(예: '새', '고양이', '비행기', '자동차' 등)를 구별하는 고유한 패턴을 찾아냄
- 예를 들어 '새' 이미지 1,000장을 모델에 입력했을 때 나타나는 내부 신호들을 SAE로 분석하면, '새'를 대표하는 특정한 신호 조합이 추출되며, 이를 '새 범주의 대표 특징'으로 저장함
- 머신 언러닝을 적용하기 전에는 '새' 범주의 대표 특징이 모델 내부에 명확하게 존재하지만, 언러닝 적용 후 이 특징이 사라졌는지 아니면 여전히 남아있는지를 SAE 분석을 통해 확인할 수 있음
- 만약 언러닝 후에도 '새' 범주의 대표 특징이 모델 내부에서 여전히 감지된다면, 이는 출력만 조작된 것이지 실제로 정보가 삭제되지 않았음을 의미함

• **추론 시점 조향을 통한 복원 실험**

- 추론 시점 조향은 AI 모델이 이미지를 판단하는 과정에서 특정 범주 방향으로 인위적으로 신호를 강화하거나 약화시키는 기법임. 이는 마치 미세한 신호를 증폭시키듯이 조향 강도를 조절하여 모델의 판단을 특정 방향으로 유도하는 방식임
- 만약 언러닝을 통해 '새' 범주의 대표 특징을 진짜 삭제한 경우라면 모델이 해당 범주의 특징을 '새'로 인식하지 못하지만, 단순히 억제만 된 경우라면 이용자가 제시하는 '새' 이미지를 갑자기 '새'로 인식하기 시작함
- 실험에서는 조향 강도를 0.0(조향 없음)부터 2.0(강한 조향)까지 단계적으로 높여가며 각 단계에서 모델이 삭제 대상 범주를 얼마나 정확하게 인식하는지 측정하여, 조향에 따른 정확도의 변화를 관찰함
- 대부분의 언러닝 기법들은 조향 없이는 삭제 범주 정확도가 0%에 가까웠으나 조향을 가하자 정확도가 급격히 상승했으며, 이는 정보가 내부에 여전히 존재하지만 출력 단계에서만 차단되었음을 보여줌. 즉, AI 모델이 해당 범주를 실제로 '잊은' 것이 아니라, '모르는 척' 할 뿐임을 시사함
- [그림 1]은 원본 모델과 언러닝 모델에서 삭제 대상 범주(새)의 SAE 특징을 추출한 후, 언러닝 모델의 중간 레이어에 해당 특징을 조향 강도(α)만큼 인위적으로 더하여 복원을 시도하는 과정을 보여줌
- 조향을 통해 언러닝 모델이 새를 삭제 대상 범주로 다시 분류하면 단순 억제 방식이며, 조향 후에도 인식하지 못하면 삭제 대상 범주가 완전히 삭제된 언러닝으로 판정함

[그림 1] SAE 기반 추론 시점 조향을 통한 억제-삭제 판별 과정



출처: 장유림 외 4인, "Suppression or Deletion: A Restoration-Based Representation-Level Analysis of Machine Unlearning", arXiv, 2026.02.18., https://arxiv.org/pdf/2602.18505

• 주요 실험 결과 및 기법 비교

- CIFAR-10 데이터셋 실험에서 SCRUB이라는 머신 언러닝 기법은 조향 없이 6.0%의 삭제 범주 정확도를 보였으나 조향 후 100%로 상승했고, SalUn과 L1-Sparse 기법도 유사하게 0%에서 90% 이상으로 급등하여 모두 억제 방식임이 드러남
- 12개 기법 중 유일하게 EU-K 기법만이 조향 전후 모두 0%의 정확도를 유지했으며, 이는 해당 기법이 모델 내부에서 중요 특징 정보를 완전히 제거한 유일한 사례로 평가됨
- 언어모델인 클로드 3.7 소네트에 대한 실험에서도 조향 없이 0.24%였던 정확도가 조향 후 95.87%로 치솟아, 이미지 모델뿐 아니라 텍스트 모델에서도 동일한 억제 메커니즘이 작동함을 확인함
- 기존의 출력 기반 평가 방식으로는 대부분의 기법이 성공적으로 언러닝된 것처럼 보였으나, SAE 기반 표현 수준 분석을 통해 실제로는 정보가 내부에 그대로 남아있었음이 밝혀져 기존 검증 방식의 심각한 허점이 입증됨

결론 및 시사점

• SAE 기반 언러닝 검증의 기술적 의의

- 본 연구는 SAE 기반 표현 수준 분석을 통해 기존 머신 언러닝 기법 대부분이 정보를 실제로 삭제하지 않고 출력 단계에서만 억제한다는 사실을 실증했으며, 이는 출력 기반 평가만으로는 언러닝의 진정한 효과를 판단할 수 없음을 보여줌
- SAE 기반 기술은 머신 언러닝 기법의 실제 효과를 평가하는 새로운 표준을 제시하며, AI가 저작권으로 보호되는 콘텐츠나 개인정보 자체를 제거하거나 중요 특징 정보를 제거했는지 확인하는 신뢰성 있는 도구로 활용될 수 있음

• 한계점 및 향후 과제

- SAE 학습에 상당한 컴퓨팅 자원이 필요하고 조향 강도 설정 기준이 명확하지 않으며, 현재는 이미지 분류 모델 중심으로만 검증되어 다른 모델에 대한 적용 가능성은 추가 연구가 필요함
- 향후 억제가 아닌 진정한 삭제를 달성하는 머신 언러닝 알고리즘 개발이 필요하며, 이는 AI 시대의 저작권 보호와 개인정보 보호를 위한 핵심 기술 과제로 남아있음

참고문헌

- 장유림 외 4인, "Suppression or Deletion: A Restoration-Based Representation-Level Analysis of Machine Unlearning", arXiv, 2026.02.18., https://arxiv.org/pdf/2602.18505
- Elena Marchetti, "Researchers Extracted 95.8% of Harry Potter From Claude, Word for Word - and It Only Cost $55", Awesome Agents, 2026.02.23., https://awesomeagents.ai/news/ai-models-reproduce-copyrighted-books-from-memory/