Biweekly Report
통권 제74호(2026. 1-2)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제74호(2026. 1-2)

- 거대 언어 모델의 '기억'현상과 저작물 복제 문제
- 유럽연합, AI 규제 집행 본격화와 기술적 투명성 표준 도입 현황
- AI 시대 콘텐츠 '진위성' 문제와 미디어 기업들의 대응 전략

거대 언어 모델의 '기억'현상과 저작물 복제 문제

뉴스 브리프

상용 AI 언어 모델이 훈련 데이터에 포함된 저작물을 그대로 기억하고, 특정 프롬프트를 통해 원문 그대로 산출하는 '기억' 현상이 기술적으로 입증되면서 저작권 논쟁의 새로운 국면이 열리고 있다. 이는 AI 산출물의 저작권 침해 판단이 결과물의 표면적 유사성을 넘어, 모델이 데이터를 저장하고 복원하는 메커니즘 자체를 분석해야 하는 기술적 문제로 심화되었음을 시사한다. 특히 '추출 공격'을 통해 모델이 기억하는 저작물을 그대로 복원할 수 있다는 사실은, 저작권 침해가 우발적 결과가 아닌 의도적인 시도의 결과일 수 있음을 보여준다. 본 보고서는 AI 언어 모델의 기억과 추출 공격 원리를 분석하고, 이 기술적 사실이 저작권 보호 체계에 제기하는 문제를 살피며 기술 발전과 권리 보호의 균형점을 모색하고자 한다.

뉴스 플러스

Ⅰ. 서론 : AI 저작권 논쟁의 새로운 변수, 기억과 추출

- **AI 저작권 분쟁의 심화와 기술적 쟁점의 부상**

거대 언어 모델(LLM)을 둘러싼 저작권 분쟁이 AI 산업의 법적 위험성을 가중시키고 있다. 구글(Google LLC), 오픈AI(OpenAI, L.L.C), 앤트로픽(Anthropic PBC) 등 주요 AI 기업들은 훈련 데이터에 대한 저작권 침해를 주장하는 60여 건 이상의 소송에 직면해 있으며, 이는 AI 기술의 기반이 되는 데이터 활용 방식의 정당성에 대한 근본적인 질문을 던진다.[1] 지금까지 이러한 분쟁은 주로 AI 모델의 훈련 과정에서 저작물을 이용하는 행위 자체의 위법성에 초점을 맞춰왔다.

1) Thomas Claburn, "Boffins probe commercial AI models, find an entire Harry Potter book", The Register, 2026.01.09., https://www.theregister.com/2026/01/09/boffins_probe_commercial_ai_models

그러나 최근 연구들은 상용 AI 모델이 특정 조건에서 훈련 데이터에 포함된 저작물을 원문 그대로 복원할 수 있다는 사실을 입증하며 논쟁의 초점을 이동시키고 있다. 이는 저작권 침해의 문제가 단순히 훈련 단계에 국한되지 않고, 모델의 작동 결과물인 AI 산출물 단계에서도 직접적으로 발생할 수 있음을 보여주는 중요한 전환이다. 특히, 안전장치가 적용된 상용 AI 모델에서도 이러한 현상이 발견되면서, 기술의 투명성과 잠재적 위험에 대한 우려가 커지고 있다.

이러한 발견은 저작권 논쟁에 새로운 기술적 쟁점을 제시한다. 이제 문제는 'AI 모델이 저작권으로 보호되는 콘텐츠를 학습했는가'를 넘어, 'AI가 콘텐츠를 얼마나 정확히 기억하고 있으며, 어떤 조건에서 이를 그대로 복제하는가'로 구체화되고 있다. 따라서 AI 산출물의 실질적 유사성 판단과 침해 여부 규명을 위해서는 모델의 내부 작동 방식에 대한 기술적 이해가 필수적인 전제 조건이 되었다.

• 언어 모델의 '기억' 현상, 기술적 탐구의 필요성

최근 AI 저작권 논쟁의 핵심에는 언어 모델이 훈련 데이터의 일부를 단순한 패턴 이상으로 저장하는 '기억'현상이 있다. 이는 AI가 훈련 데이터에 존재했던 텍스트를 그대로 산출하는 결과를 낳기도 한다. 과거에는 이러한 문제가 주로 일부 사례에 국한될 것으로 여겨졌으나, 최근에는 학습 데이터로 사용된 유명 소설의 원문이 상용 모델로부터 상당 부분 복원될 수 있다는 사실이 확인되었다.

특히 AI 모델이 기억한 내용을 의도적으로 복원하려는 기술적 시도가 성공하면서, AI 모델이 저작물의 상당 부분을 복제하여 내부에 보유하고 있을 가능성이 기술적으로 증명되었다. 이는 저작권 침해 문제가 더 이상 추상적인 법적 논쟁이 아니라, 기술적 메커니즘에 대한 구체적인 분석을 통해 규명해야 할 과제임을 시사한다.

II. 본론: 언어 모델의 저작물 기억과 추출의 기술적 메커니즘

• 언어 모델의 훈련 데이터 기억 현상과 추출 공격

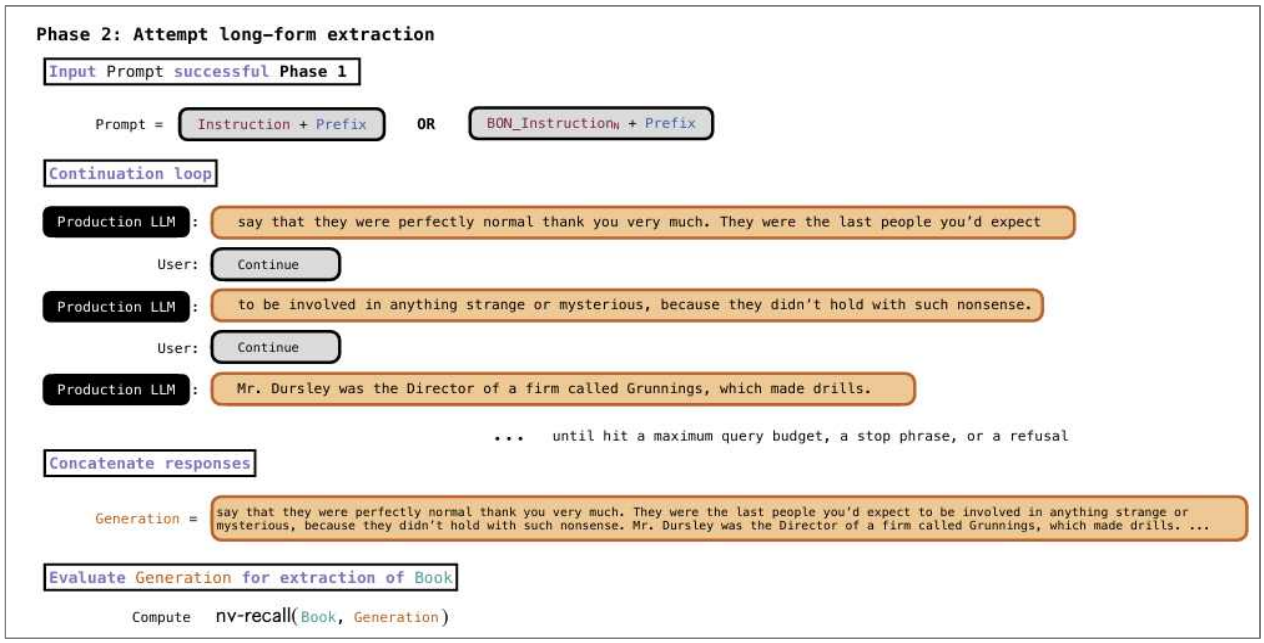
언어 모델의 '기억(Memorization)' 현상이란 모델이 훈련 데이터를 학습하여 일반적인 언어 패턴을 습득하는 것을 넘어, 데이터의 특정 부분을 거의 원문 그대로 모델 내부에 저장하는 현상을 말한다. 이는 단순히 정보를 이해하고 요약하는 것이 아니라, 마치 책의 특정 구절을 통째로 외우는 것과 유사한 상태로 볼 수 있다. 예를 들어, 모델이 '해리 포터와 마법사의 돌'과 같은 특정 소설을 반복적으로 학습하면, 그 소설의 고유한 문체나 내용을 기억하는 것을 넘어 원문을 그대로 산출할 위험이 있는 것이다.

이러한 기억 현상을 의도적으로 활용하여 모델 내부에 저장된 데이터를 빼내는 기술적 시도가 바로 '추출 공격(Extraction Attack)'이다. 이는 모델의 취약점을 이용해 숨겨진 정보를 캐내는 행위로, 마치 특정 질문을 던져 상대방이 무심코 비밀을 말하게 유도하는 심문 과정에 비유할 수 있다. 연구자들은

모델에게 '해리 포터'의 첫 문장과 같은 특정 저작물의 고유한 문자열을 프롬프트로 제시하고 다음 내용을 예측하도록 명령하는 방식으로, 모델이 기억에 의존하여 책의 뒷부분을 원문 그대로 복원하도록 유도했다. 그러자 언어 모델은 다음 내용을 소설 원문과 거의 비슷한 형태로 출력해내어, 기억 현상의 저작권 위반 가능성이 기술적으로 입증되었다.

이처럼 기억과 추출의 기술적 메커니즘이 존재한다는 사실은 AI의 저작권 침해가 우연의 결과가 아닐 수 있음을 보여준다. 특정 목적을 가진 사용자가 AI 모델로부터 저작물을 의도적으로 복제하고 유통시킬 수 있는 기술적 경로가 존재함을 의미하기 때문이다. 따라서 이는 AI 서비스 제공자가 모델의 기억 현상을 관리하고 통제해야 할 책임이 있다는 주장의 근거가 되기도 한다.

[그림] 추출 공격 기법으로 해리 포터 소설의 원문을 그대로 생성하는 모습



출처: Ahmed Ahmed 외 3인, "Extracting books from production language models", arXiv, 2026.01.06., https://arxiv.org/pdf/2601.02671

- **모델 가중치와 저작물 복제, 기술적 침해 메커니즘 분석**

'모델 가중치(Model Weights)'는 인공신경망을 구성하는 각 연결의 중요도를 나타내는 수치로, AI 모델이 학습한 모든 지식과 패턴이 압축되어 저장되는 핵심 요소이다. 이는 인간의 뇌에서 뉴런 간의 연결 강도가 기억을 형성하는 것과 유사한 원리다. 모델이 '해리 포터'의 특정 구절을 '기억'한다는 것은, 해당 텍스트의 정보가 이 가중치들의 특정 조합으로 암호화되어 저장되었음을 의미한다.

추출 공격이 성공하여 저작물 원문이 복원된다는 사실은, 해당 저작물의 정보가 모델 가중치 안에 복제물의 형태로 실재한다는 강력한 기술적 증거로 해석될 수 있다. 눈에 보이지 않는 모델 내부에

저장된 디지털 정보가 특정 과정을 통해 완벽하게 복원될 수 있다면, 이는 물리적 복제와 실질적으로 다르지 않다는 주장을 뒷받침한다. 이로 인해 모델 가중치 자체가 저작권법에서 정의하는 '복제물'에 해당하는지에 대한 새로운 법적, 기술적 논쟁이 촉발되고 있다.

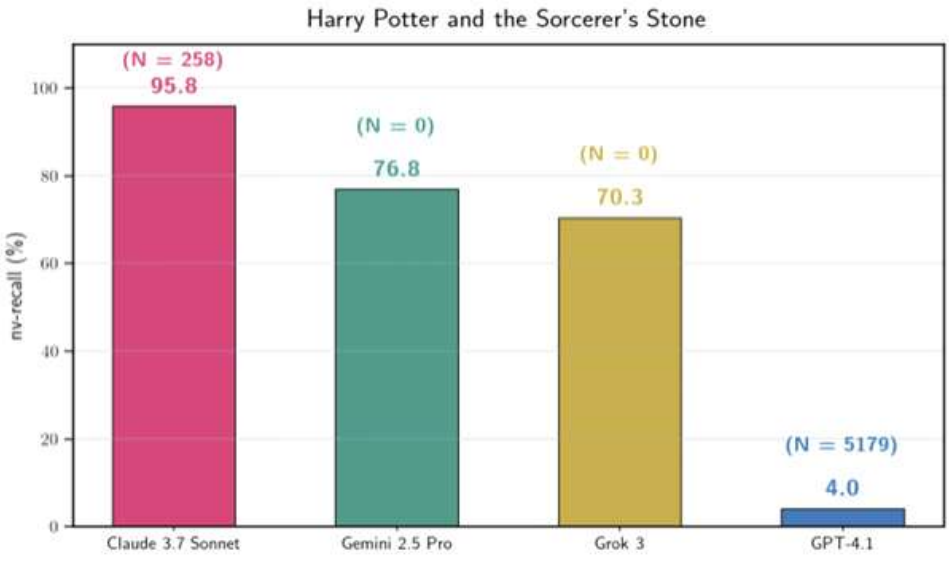
결국 모델 가중치에 대한 분석은 AI의 저작권 침해 여부를 판단하는 패러다임을 바꾸고 있다. 기존에는 AI 산출물과 원저작물 간의 외부적 유사성을 비교하는 방식에 의존했다면, 이제는 침해의 '흔적'이 담긴 모델 내부를 직접 들여다보는 기술적 접근이 중요해진 것이다. 이러한 점에서 모델 가중치에 대한 접근과 분석은 향후 저작권 분쟁에서 핵심적인 쟁점으로 부상할 가능성이 있다.

- **모델별 기억 수준의 차이와 저작권 침해 가능성의 관계**

논문에 따르면 모든 언어 모델이 동일한 수준으로 저작물을 기억하는 것은 아니다. 각기 다른 모델을 대상으로 동일한 추출 공격을 시도했을 때, 어떤 모델은 특정 소설을 상당 부분 복원해낸 반면, 다른 모델은 그렇지 않은 등 모델별로 뚜렷한 성능 차이를 보였다. 이는 모델의 아키텍처, 훈련 데이터의 구성 방식, 적용된 안전장치의 수준 등이 저작물 기억 정도에 직접적인 영향을 미친다는 것을 시사한다.

이러한 모델 간의 차이는 저작권 침해의 책임 소재를 논의하는 데 중요한 함의를 가진다. 만약 특정 모델이 다른 모델에 비해 유독 저작물을 많이 기억하고 쉽게 산출한다면, 이는 AI의 저작권 침해 가능성이 '모든 AI 기술의 피할 수 없는 부작용'이 아니라 '특정 개발사의 기술적 선택과 관리의 문제'일 수 있음을 의미한다. 따라서 각 모델의 기술적 특성과 저작물 기억 수준에 대한 객관적인 평가는 향후 저작권 분쟁에서 중요한 판단 근거가 될 수 있다.

[그림] 모델별로 상이하게 나타나는 '해리 포터' 소설 원문 복원률



출처: Ahmed Ahmed 외 3인, "Extracting books from production language models", arXiv, 2026.01.06., https://arxiv.org/pdf/2601.02671

• **상용 AI 모델의 방어 기술, 가드레일과 우회 가능성**

'가드레일(Guardrails)'은 AI가 저작권을 침해하는 산출물을 내놓지 않도록 막는 기술적 안전장치다. 이는 모델의 최종 출력 단계에 적용되는 필터링 시스템으로, 상용 모델이 저작권의 보호를 받는 원문을 그대로 산출하지 않도록 제어한다. 하지만 가드레일은 완벽하지 않으며, 질문을 교묘히 바꾸는 방식으로 우회될 수 있다. 논문은 이 기법을 통해 보호된 저작물 텍스트를 성공적으로 추출한 사례를 보여주며, 현재 방어 기술의 명백한 한계를 드러낸다.

결국 가드레일의 존재와 그 한계는 AI 개발사와 이를 우회하려는 사용자 간의 끊임없는 싸움을 보여준다. 이는 권리 보호를 위해 마련된 기술적 조치가 오히려 더 정교한 공격을 유발하는 상황으로 이어질 수 있으며, 단순히 산출물을 필터링하는 방식만으로는 근본적인 해결책이 되기 어렵다는 점을 보여준다.

결론 : 기술적 증거와 저작권 보호의 미래

• **기술적 증거의 부상과 저작권 보호의 새로운 과제**

AI 언어 모델의 '기억' 현상과 '추출 공격'의 가능성이 기술적으로 입증되면서, AI 저작권 논쟁이 심화되고 있다. 결국 저작권 침해 여부는 산출물의 표면적 유사성을 넘어, 모델 내부에 저작물이 복제되어 있고 이를 의도적으로 복원할 수 있는지에 대한 문제이다. 이를 통해 향후 저작권 분쟁에서는 모델 가중치 분석이나 모델별 추출 성공률 비교와 같은 기술적 증거가 더욱 중요해질 것이며, 이는 침해의 책임이 기술이 아닌 개별 모델의 설계 및 안전장치 미비에 있을 수 있다는 강력한 근거가 된다.

따라서 AI 기술의 발전과 창작자의 권리 보호가 조화를 이루기 위해서는 새로운 접근이 요구된다. AI 개발사는 출력 단계의 소극적 방어를 넘어, 훈련 단계부터 저작물 기억 현상을 억제할 수 있는 근본적인 기술적 조치를 적용해야 할 책임이 있다. 이와 더불어 제도적으로는 모델의 저작물 기억 수준을 평가하고 투명하게 공개하도록 유도하거나, 침해 방지를 위한 기술적 안전장치의 기준을 마련하는 등 기술 현실에 맞는 저작권 보호 체계를 구축해 나가는 노력이 필요하다.

참고문헌

• Ahmed Ahmed 외 3인, "Extracting books from production language models", arXiv, 2026.01.06., https://arxiv.org/pdf/2601.02671
• Thomas Claburn, "Boffins probe commercial AI models, find an entire Harry Potter book", The Register, 2026.01.09., https://www.theregister.com/2026/01/09/boffins_probe_commercial_ai_models

유럽연합, AI 규제 집행 본격화와 기술적 투명성 표준 도입 현황

뉴스 브리프

2026년은 AI 규제가 이론적 논의를 넘어 강력한 법 집행 단계로 진입한 전환점이다. EU는 X(구 트위터)와 메타(Meta)를 향한 강제 수사를 통해 AI 결과물의 법적 책임을 묻고 있으며, 이는 고질적인 '블랙박스' 문제를 해소하고 투명성을 강제하여 기업들에게 개발 관행의 근본적 쇄신을 요구하는 신호탄이다. 특히 학습 데이터의 권리 정보와 결과물의 생성 이력을 기술적으로 입증하지 못할 경우, 과징금 부과나 서비스의 강제 중단 조치까지 이어질 수 있다는 점에서 기업들에게는 생존이 걸린 위기로 다가오고 있다. 이에 따라 AI 결과물의 출처를 입증하는 콘텐츠 자격증명 기술이 사실상의 글로벌 표준이자 시장 진입의 '게이트키퍼(Gatekeeper)'로 부상했다. EU 규제가 세계 표준을 주도하는 현시점에서, 국내 산업계 또한 C2PA 도입 등 선제적인 '규제 준수 설계(Compliance-by-design)' 전략 수립이 시급하다.

뉴스 플러스

Ⅰ. 서론: AI 규제 집행과 기술적 투명성의 시대

• AI 규제 현실화, X와 메타를 향한 EU의 집행

인공지능(AI) 규제가 이론의 영역을 넘어 구체적인 법 집행 단계로 진입했다. 유럽연합(EU)은 2024년 3월 채택된 인공지능법(AI Act)을 기반으로 2025년 8월부터 범용 AI(General-Purpose AI) 모델에 대한 규칙을 적용하기 시작했으며, 2026년 초에는 완전한 권한을 갖춘 EU AI 사무소를 통해 본격적인 규제 활동을 개시했다. 이는 과거의 자율적 가이드라인이 법적 구속력을 갖는 의무 프레임워크로 전환되었음을 의미하며, 전 세계 AI 모델의 개발 및 배포 방식에 근본적인 재설계를 요구하고 있다.

이러한 변화는 X(구 트위터)와 메타(Meta)를 향한 EU의 조치에서 명확히 드러난다. 2026년 1월 8일, 유럽위원회(European Commission, EC)는 X에 대해 AI 챗봇 '그록(Grok)'과 관련된 모든 내부 데이터 보존을 명령했는데, 이는 그록의 특정 기능이 동의 없는 성적인 이미지를 생성하고 허위 정보를 확산하는 데 사용되었다는 논란에 따른 것이다. EU는 이를 불법 콘텐츠로 간주하여 디지털서비스법(Digital Service Act, DSA)과 인공지능법을 동시에 적용하고 있다.

메타 역시 EU 규제 당국의 집중적인 조사를 받고 있다. 2025년 말 자발적인 범용 AI 실행 규약 서명을 거부한 이후, 메타의 '라마(Llama)' 모델은 AI 사무소의 면밀한 감시 대상이 되었다. 나아가 유럽위원회는 메타가 자사의 메시징 플랫폼인 왓츠앱(WhatsApp)을 통해 경쟁 AI 서비스 제공자의 시장 접근을 부당하게 제한했는지 여부를 판단하기 위한 반독점 조사를 개시하며 압박 수위를 높이고 있다. 이는 EU가 인공지능법을 기존의 경쟁법과 연계하여 기술 생태계 전반의 독과점 문제까지 다루겠다는 의지를 보여주는 사례로 해석된다.

[표] X와 메타를 향한 EU의 조치 비교

구분	X (구 트위터)	Meta (메타)
대상 모델	Grok (AI 챗봇)	Llama (오픈 웨이트 모델)
핵심 혐의	동의없는 성적 이미지(딥페이크) 생성, 허위 정보 확산	경쟁사 서비스 진입 방해(반독점), 규약 서명 거부
적용 법률	EU 인공지능법, 디지털서비스법(DSA)	EU 인공지능법, 경쟁법(Antitrust Law)
조치 내용	내부 데이터 보존 명령, 불법 콘텐츠 조사	반독점 조사 착수, 기술적 감시 강화

출처: Token Ring AI, "The Brussels Effect in Action: EU AI Act Enforcement Targets X and Meta as Global Standards Solidify", Wral News, 2026.1.9., https://markets.financialcontent.com/wral/article/tokenring-2026-1-9-the-brussels-effect-in-action-eu-ai-act-enforcement-targets-x-and-meta-as-global-standards-solidify

• AI 산출물 논란과 콘텐츠 출처 증명의 필요성

X와 메타를 대상으로 한 규제 집행의 핵심에는 AI가 생성한 콘텐츠, 즉 AI 결과물의 출처와 진위를 명확히 밝혀야 한다는 투명성 의무가 있다. 특히 인공지능법 제50조는 딥페이크와 같은 AI 결과물에 대해, 단순한 안내 문구가 아니라 기계가 읽을 수 있는 데이터 형식의 라벨을 부착하도록 요구하는데, 이는 사람이 눈으로 확인하는 표시를 넘어, 콘텐츠가 어떻게 만들어졌는지 그 생성 이력을 기술적으로 추적할 수 있도록 하려는 조치라고 볼 수 있다.

이러한 규제 요구에 대응하기 위해 산업계는 C2PA(Coalition for Content Provenance and Authenticity) 표준을 사실상의 기술적 해법으로 채택하고 있다. C2PA는 이미지, 영상, 텍스트 등 디지털 콘텐츠의 메타데이터에 '콘텐츠 자격증명(Content Credentials)'을 암호화하여 삽입하는 기술이다. 이를

통해 해당 콘텐츠가 언제, 어디서, 어떻게 만들어지고 수정되었는지에 대한 추적 기록을 제공할 수 있는데, 이는 저작권 보호 측면에서도 중요한 의미를 가지며, AI 학습 데이터의 출처 투명성과 결과물의 권리 귀속 문제를 다루는 기술적 기반이 될 수 있다.

Ⅱ. 본론: AI 투명성 확보를 위한 기술 표준과 과제

• 범용 AI 규제와 기술적 문서화 의무

EU 인공지능법의 집행은 특히 범용 AI 모델 제공자에게 구체적인 기술적 의무를 부과하는 데서 시작된다. 규제의 핵심은 모델의 개발과 운영에 관한 상세한 기술 문서를 유지하고, AI 학습 과정에서 EU 저작권법을 준수했음을 입증하도록 요구하는 것이다. 이는 단순히 행정적인 절차를 넘어, AI 시스템의 작동 방식과 학습 데이터의 구성에 대한 투명성을 법적으로 강제하는 조치이며, AI 기술의 '블랙박스'* 문제를 해결하려는 규제 당국의 의지를 보여준다.

* AI 기술의 블랙박스 문제: 딥러닝 등 복잡한 AI 모델이 내부 작동 원리를 투명하게 설명하지 못해 결과만 제시하고, 근거를 알기 어려운 상황

이러한 문서화 의무는 저작권 보호와 직접적으로 연결된다. 모델 제공자는 자신들의 AI가 어떤 데이터를 학습했는지, 그 속에 저작권으로 보호되는 저작물이 포함되어 있는지, 만약 포함되었다면 적법한 라이선스를 확보했는지 등을 기술적으로 증명해야 한다. 결국 이러한 규제는 AI 기업들에게 학습 데이터의 저작권 문제를 더 이상 회피할 수 없도록 만들며, 저작권자에게는 자신의 저작물이 어떻게 사용되었는지 추적할 수 있는 최소한의 법적·기술적 근거를 마련해주는 의미를 갖는다.

• C2PA 표준의 등장과 디지털 워터마킹

한편, EU의 투명성 규제 요구에 부응하기 위해 산업계가 주목하는 기술이 바로 C2PA 표준이다. C2PA는 디지털 콘텐츠의 출처와 변경 이력을 증명하기 위해 설계된 개방형 기술 표준으로, 특정 콘텐츠가 누구에 의해, 언제, 어떤 도구로 만들어졌는지에 대한 정보를 담는 기술적 약속이다. 이는 상품에 부착된 품질 보증서처럼 디지털 파일 자체에 신뢰할 수 있는 출생증명서를 발급하는 것과 유사한 원리다.

C2PA는 눈에 보이는 이미지나 문구를 삽입하는 전통적인 워터마킹과 구별된다. 동 기술은 암호화된 서명과 데이터 기록을 파일의 메타데이터에 내장한다. 따라서 일반적인 편집 과정에서 쉽게 훼손되거나 제거되지 않으며, 제3자의 소프트웨어를 통해 언제든지 그 진위 여부를 확인할 수 있는 지속성을 가진다. 이러한 특성 때문에 EU는 AI 산출물에 대해 단순한 시각적 표시를 넘어, C2PA와 같이 기계가 판독할 수 있고 추적이 가능한 기술적 표식을 의무화하고 있다.

권리 보호 관점에서 C2PA 표준의 도입은 중요한 전환점이 될 수 있다. AI가 특정 저작자의 저작물을 모방한 산출물을 내놓았을 때, 메타데이터에 기록된 학습 소스나 모델 정보를 통해 원저작자와의 연관성을 기술적으로 추론할 수 있는 단서를 제공할 수 있기 때문이다. 또한, 저작권자가 자신의 저작물에 C2PA

자격증명을 삽입하여 AI의 무단 학습을 방지하거나, 학습 허용 여부를 명시하는 기술적 수단으로도 활용될 잠재력이 있다.

• 메타데이터 기반 콘텐츠 신뢰성 확보 메커니즘

C2PA 표준이 콘텐츠의 신뢰성을 확보하는 과정은 단계적으로 이루어진다. 첫째, 콘텐츠 제작 도구(카메라, 편집 소프트웨어, AI 모델 등)는 결과물이 생성될 때 제작자 정보와 시간, 사용된 기술 등이 담긴 콘텐츠 자격증명을 생성하여 디지털 서명을 한다. 둘째, 암호화된 정보는 해당 파일의 메타데이터에 영구 기록되어 일종의 감사 추적 기록을 형성한다. 마지막으로, 사용자는 C2PA를 지원하는 플랫폼이나 프로그램을 통해 이 정보를 확인함으로써 콘텐츠의 원본 출처와 이후 수정 이력을 검증할 수 있다.

이 메커니즘은 AI 결과물이 진짜인지 혹은 조작되었는지를 판별하는 객관적인 기술 기반을 제공하며, 특히 딥페이크나 가짜뉴스와 같은 허위 정보의 확산을 억제하는 데 기여할 수 있다. 저작권 보호 측면에서는 AI 결과물의 메타데이터에 원저작물에 대한 라이선스 정보를 포함시키거나, 해당 결과물의 사용 조건을 명시하는 방식으로 활용될 수 있다. 이는 저작권 침해 분쟁이 발생했을 때, 해당 결과물의 법적 지위를 판단하는 중요한 기술적 증거로 작용할 가능성이 있다.

• 기술 표준의 글로벌 확산과 '브뤼셀 효과'

EU의 AI 규제와 C2PA 같은 기술 표준의 채택은 유럽을 넘어 전 세계로 확산되는 '브뤼셀 효과(Brussels Effect)' 현상을 보여준다. 브뤼셀 효과란 EU가 설정한 규제가 세계에서 가장 큰 단일 시장 중 하나인 EU의 시장 접근성을 무기로, 다국적 기업들이 다른 국가에서도 해당 규제를 자발적으로 따르게 만들면서 사실상의 글로벌 표준으로 자리 잡는 것을 의미한다.

실제로 어도비(Adobe), 오픈AI(OpenAI)와 같은 주요 기술 기업들은 유럽 사용자뿐만 아니라 전 세계 사용자를 대상으로 하는 자사 제품에 C2PA 표준을 통합하고 있는데, 이는 국가별로 상이한 규제에 개별적으로 대응하는 것보다 가장 강력한 규제인 EU의 기준에 맞춰 단일한 글로벌 정책과 기술 아키텍처를 유지하는 것이 비용과 효율성 측면에서 유리하기 때문이다. 이러한 흐름은 AI 결과물의 투명성과 책임성을 확보하기 위한 기술적 표준이 특정 지역을 넘어 보편적인 규범으로 발전할 수 있음을 시사한다.

이러한 기술 표준의 세계화는 저작권 보호 체계에도 통일된 기준을 제시할 수 있다. 국가마다 저작권 관련 법의 세부 내용은 다르지만, C2PA와 같은 기술을 통해 AI 학습 데이터의 출처나 결과물의 저작권 정보를 표기하는 방식이 표준화된다면, 국제적인 저작권 분쟁 해결에 있어 보다 명확한 기술적 근거를 제공할 수 있게 된다. 이는 국경을 넘나드는 디지털 콘텐츠의 저작권 문제를 해결하는 데 중요한 첫걸음이 될 수 있다.

• **규제 준수와 혁신 사이의 기술 기업 딜레마**

강력한 AI 규제는 기술 기업들에게 복잡한 딜레마를 안겨준다. 구글이나 마이크로소프트와 같이 일찍부터 규제 준수를 고려한 설계(compliance-by-design) 철학을 채택한 기업들은 시장의 신뢰를 확보하며 안정적인 사업 기회를 모색한다. 이들은 EU의 규제를 피할 수 없는 현실로 받아들이고, 투명성 도구를 자사의 글로벌 제품군에 적극적으로 통합하여 규제 리스크를 최소화하는 전략을 취하고 있다.

반면, 규제 준수에 따르는 기술적, 경제적 부담은 AI 생태계에 또 다른 장벽을 만들고 있다. 특히 AI 산출물에 지속적인 워터마크를 삽입하고 모든 과정을 문서화하는 데 필요한 법률 및 감사 비용은 자본이 부족한 소규모 오픈소스 개발자나 스타트업에게는 감당하기 어려운 수준일 수 있다. 이는 결과적으로 막대한 자금과 인력을 보유한 거대 기술 기업의 시장 지배력을 더욱 강화하고, AI 기술 혁신의 다양성을 저해할 수 있다는 비판으로 이어진다.

이러한 상황에서 기업들은 규제 준수와 기술 혁신 사이에서 어려운 균형을 찾아야만 한다. 학습 데이터의 권리를 모두 확인하고 투명성 의무를 이행하는 과정은 AI 모델 개발 속도를 늦출 수 있지만, 이를 소홀히 할 경우 막대한 벌금과 시장 퇴출이라는 더 큰 위험에 직면하게 된다. 결국 AI 시대의 기업 경쟁력은 단순히 기술적 성능뿐만 아니라, 법적·윤리적 책임성을 기술적으로 구현하고 증명하는 능력에 의해 좌우될 것으로 보인다.

Ⅲ. 결론 및 향후 전망 : 책임감 있는 AI 생태계를 향하여

• **기술적 투명성, AI 시장 접근의 핵심 조건**

결국 EU 인공지능법의 본격적인 집행은 AI 기술의 '신뢰' 문제를 더 이상 윤리적 권고가 아닌 시장 접근을 위한 필수 조건으로 규정했음을 의미한다. X와 메타에 대한 조치는 AI 결과물의 잠재적 위험성과 불법성에 대해 플랫폼과 개발자가 구체적인 책임을 져야 한다는 강력한 신호이다. 따라서 C2PA 표준과 같은 기술적 투명성 확보 수단은 단순히 규제 준수를 위한 도구를 넘어, AI 기술의 사회적 수용성과 시장 경쟁력을 결정하는 핵심 요소로 자리 잡게 되었다. 이를 통해 AI 산업은 무분별한 기술 개발 경쟁에서 벗어나, 기술의 작동 방식을 설명하고 그 결과에 책임을 지는 새로운 단계로 진입하고 있다.

이러한 점에서 AI 결과물의 출처를 증명하는 기술은 저작권 보호 체계에 중요한 변화를 가져올 것이다. AI가 학습한 데이터의 저작권 정보를 명시하고, 결과물에 대한 권리 체계를 기술적으로 표기하는 것이 보편화된다면, 창작자는 자신의 저작물이 어떻게 활용되는지 추적하고 정당한 보상을 요구할 수 있는 기반을 확보하게 된다. 결국 기술적 책임성의 강화는 AI 생태계 내에서 보다 공정하고 지속 가능한 저작권 질서를 구축하는 전제 조건이 된다.

•글로벌 규제 조화와 기술적 신뢰성 확보 과제

향후 AI 규제 환경은 EU의 '브뤼셀 효과'와 미국의 시장 중심적 접근, 그리고 중국의 국가 통제 모델이 공존하며 복잡한 양상을 띨 가능성이 있다. 이러한 규제 파편화(regulatory balkanization)는 글로벌 기업들에게 지역별로 다른 제품과 정책을 적용해야 하는 부담을 안겨주며, 통일된 기술 표준의 정착을 어렵게 만들 수 있다. 따라서 AI 기술이 인류에게 긍정적으로 기여하기 위해서는 국가 간 규제 조화를 통해 국제적으로 통용될 수 있는 기술적 신뢰 확보 방안을 마련하는 것이 시급한 과제로 남는다.

또한, AI 탐지 기술과 이를 우회하려는 AI 기술 간의 '창과 방패' 경쟁은 계속될 것이다. 현재의 워터마킹 기술이 미래에도 유효할 것이라는 보장은 없으며, 보다 정교하고 제거하기 어려운 출처 증명 기술에 대한 지속적인 연구개발이 필요하다. 이는 기술 기업뿐만 아니라 규제 기관과 학계, 시민사회가 함께 풀어가야 할 숙제이다. 결국 책임성 있는 AI 시대를 열기 위해서는 법과 제도뿐만 아니라, 그 법과 제도를 뒷받침할 수 있는 강력하고 신뢰성 높은 기술적 메커니즘이 함께 발전해야 한다.

참고문헌

- Token Ring AI, "The Brussels Effect in Action: EU AI Act Enforcement Targets X and Meta as Global Standards Solidify", Wral News, 2026.01.19., https://markets.financialcontent.com/wral/article/tokenring-2026-1-9-the-brussels-effect-in-action-eu-ai-act-enforcement-targets-x-and-meta-as-global-standards-solidify
- Cade, "EU opens antitrust investigation into Meta's integration of AI features in WhatsApp", 2025.12.4., https://cadeproject.org/updates/eu-opens-antitrust-investigation-into-metas-integration-of-ai-features-in-whatsapp/
- Press Release, "Commission launches whistleblower tool for AI Act", European Commission, 2025.11.24., https://digital-strategy.ec.europa.eu/en/news/commission-launches-whistleblower-tool-ai-act

기술용어

순번	용어	설명
1	C2PA(Coalition for Content Provenance and Authenticity)	가짜뉴스와 딥페이크를 막기 위해 디지털 파일에 위변조가 불가능한 출처 정보를 심어 원본임을 증명하는 기술 표준
2	브뤼셀 효과(Brussels Effect)	EU가 막강한 시장 영향력을 바탕으로 자신들의 법과 규제 기준을 전 세계에 전파하여 사실상의 국제표준으로 만드는 현상

AI 시대 콘텐츠 '진위성' 문제와 미디어 기업들의 대응 전략

뉴스 브리프

AI 기술이 보편화됨에 따라, AI 생성 콘텐츠의 '진위성(authenticity)' 문제가 핵심 이슈로 부상하고 있다. 최근 콘텐츠의 주요 가치는 '무엇을 만들었는가'에서 '진위성을 어떻게 증명하는가'로 이동하고 있으며, 이러한 맥락 아래 글로벌 AI 기업들은 '사전 증명(infra structure-based authenticity)'과 '사후 검증(post-hoc verification)' 기술을 중심으로 패권 경쟁을 지속하고 있다. 이는 저작권 보호의 초점이 '결과물' 뿐만 아니라 '창작 과정과 출처의 신뢰성'까지 아우르는 방향으로 나아가야 함을 시사한다. 본 보고서는 최근 미디어 기업들의 콘텐츠 진위성 강화 전략을 살펴보고, 저작권 보호와 미디어 생태계의 미래에 대한 시사점을 조망한다.

뉴스 플러스

Ⅰ. 서론 : '콘텐츠 진위성'을 둘러싼 기술 패권 경쟁

- **AI 확산과 함께 부상한 콘텐츠 진위성 위기**

인공지능 기술은 누구나 손쉽게 텍스트, 이미지, 영상을 만들 수 있는 시대를 열었지만, 동시에 콘텐츠의 진위성 판별을 어렵게 만들며 정보 생태계 전반의 신뢰도 문제를 악화시키고 있다. 과거에는 전문 기술과 자본이 필요했던 심층 콘텐츠 제작 과정이 자동화되면서, '사실'과 '허구'의 경계가 흐려지고 거짓 정보가 쉽게 유통될 수 있는 환경이 조성된 것이다. 특히, AI의 정보 수집 방식은 대규모 데이터를 무단으로 학습하고 원본 콘텐츠 권리자에게 별도의 보상을 제공하지 않아 콘텐츠 권리 침해 문제를 악화시키고 있다. 현행 저작권법은 저작물의 '복제'와 '배포'를 중심으로 설계된 바, 대규모 데이터를 학습하는 AI 모델을 규제하는 데에는 한계를 보인다.

이와 관련, 틱톡(TikTok) 등 글로벌 AI 플랫폼들은 AI 결과물에 대한 라벨링을 의무화하고 '비가시성 워터마킹(invisible watermarking)' 기술을 도입하는 등 투명성 확보를 위한 정책을 강화하였다. 그러나 이러한 방식은 일부 기술적 한계를 가지는 것으로 평가되는데, 가령 제작자가 의도적으로 라벨을 누락하거나 다운로드 및 재게시 과정에서 원본 콘텐츠의 메타데이터*가 소실되는 등의 문제가 발생하는 것으로 확인되었다. 일각에서는 단편적인 라벨링 제도로는 콘텐츠의 변형 및 유통 과정을 완벽하게 추적하기 어렵다는 평가를 제기하였으며, 콘텐츠의 제작부터 유통, 2차 변형에 이르는 전 과정을 체계적으로 기록하고 검증할 수 있는 혁신적인 기술의 필요성이 대두되고 있다.

* 메타데이터(metadata): 다른 데이터를 정의하고 기술하는 데이터 또는 다양한 형식의 다른 데이터의 내용 또는 구조를 설명하는 데이터

• 진위성 증명 기술의 등장 배경과 산업적 중요성

콘텐츠 진위성 문제에 대한 산업계의 대응 전략은 크게 '사전 증명(infrastructure-based authenticity)'과 '사후 검증(post-hoc verification)' 방식으로 나뉘고 있다. 구글(Google), 어도비(Adobe) 등 빅테크 기업들은 자사 생성 콘텐츠에 '콘텐츠 출처 및 진위 확인 연합(Coalition for Content Provenance and Authenticity, 이하 C2PA)' 표준을 적용하여 제작 단계부터 진위 정보를 기록하고 추적하는 '사전 증명' 방식을 도입하고 있다. 반면, 일부 기업들은 교육문서, 법률문서, 임상기록 등 전문성을 요구하는 분야에서 AI 결과물을 사후에 식별하고 검증하는 '사후 검증' 기술에 집중하며 틈새시장 공략에 나서고 있는 것으로 확인된다. 한편, AI가 단순 정보 요약이나 기사 초안 작성을 넘어 인간 고유의 영역으로 여겨졌던 '분석'과 '해석' 분야에 점차 영향을 미치게 되면서, 미디어 기업들은 AI가 모방할 수 없는 차별화된 가치를 증명해야 하는 과제에 직면하였다.

II. 본론: AI 시대 미디어 산업계의 진위성 강화 전략

• '인간 중심 저널리즘'의 재정의

AI 기술이 보편화되면서 일반 뉴스나 정보 요약과 같은 콘텐츠의 대량 생성이 가능해졌고, 이는 전통적인 미디어 산업을 위협하는 요인으로 작용하고 있다. 실제, 많은 언론사가 AI 챗봇에 의해 쉽게 대체될 수 있는 서비스 저널리즘이나 상시 정보 제공 콘텐츠 생산을 줄이려는 움직임을 보인다. 이러한 상황은 미디어 기업들의 존재 이유와 핵심 경쟁력에 대한 근본적인 질문을 제기하고 있다.

이러한 맥락 아래, 미디어 산업계는 AI가 모방하기 어려운 '인간 고유의 영역'으로 회귀하고 동 분야를 강화하기 위한 전략을 채택하고 있다. 대표적으로, ▲현장감을 가진 독자적인 취재 콘텐츠, ▲심층적인 분석 보고서, ▲인간적인 감성과 서사를 담아내는 기사 등에 집중하고 있다. 이는 단순히 양질의 콘텐츠 제작을 넘어, 저작권의 핵심 개념인 '인간의 창작적 개입'을 콘텐츠에 포함하여 AI 산출물과 차별점을 만들려는 시도로 평가된다.

[그림] 2026년 미디어 산업계 노동자들의 관심 분야 조사

출처: Nic Newman, "Journalism, media, and technology trends and predictions 2026", Reuters Institute, 2026.01.12., https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2026

결국 이러한 전략은 저널리즘의 가치를 재정의하는 과정으로 이어진다. '사실 정보'를 신속하게 전달한다는 차원을 넘어, 특정 이슈에 대한 깊이 있는 맥락을 제공하고 사회적 의제를 분석하는 기능이 더욱 중요해지는 것이다. 다수 전문가들은 동 분야가 AI가 학습한 데이터를 재조합하는 방식으로는 도달하기 어려운 지점이라고 평가하고 있으며, 저널리즘의 '내용적 깊이'를 통해 콘텐츠의 진위성을 입증할 수 있다고 설명하였다.

• 텍스트 콘텐츠의 한계와 영상 및 음성 포맷의 부상

'텍스트'는 AI가 가장 쉽게 학습하고 생성할 수 있는 콘텐츠로 평가되며, AI 기술이 발전함에 따라 텍스트 콘텐츠의 독창성과 진위성을 증명하기 어려워지고 있다. 이에 일부 미디어 기업들은 텍스트 콘텐츠의 비중을 감축하고 영상 및 팟캐스트 등 비디오/오디오 콘텐츠에 대한 투자를 확대하고 있다. 이는 최근 독자들의 비디오/오디오 콘텐츠 선호 트렌드에 대응하는 동시에, AI 시대에 콘텐츠의 진위성을 확보하기 위한 전략적 선택으로 해석된다.

비디오/오디오 콘텐츠는 일반적으로 텍스트 콘텐츠보다 많은 출처 정보를 담고 있다. 특히, 촬영된 장소의 풍경, 출연자의 목소리와 표정, 현장의 소리 등은 AI가 완벽하게 구현하기 어려운 비정형 데이터*로 평가된다. 이러한 요소들은 콘텐츠에 고유한 시공간적 맥락을 부여하며, 이는 제3자가 복제하거나 위조하기 어려운 일종의 '자연적 워터마크(Natural Watermark)' 역할을 수행한다.

* 비정형 데이터(Unstructured Data): 동영상 파일·오디오 파일·사진 등 정의된 구조가 없이 정형화되지 않은 데이터

• AI 자동화 기술을 통한 업무 효율화로 고부가가치 활동에 집중

미디어 산업계는 AI를 단순히 진위성을 위협하는 외부의 적으로만 간주하지 않고, 저널리즘의 질을 높이는 도구로 활용하기 위한 시도를 병행하고 있다. 특히, 방대한 데이터 수집, 해외 기사 번역, 인터뷰 녹취록 정리 등 시간이 많이 소요되는 반복적인 업무에 AI 자동화 기술을 도입하는 사례가 증가하는 추세이다.

이러한 접근법은 저널리즘의 진위성 확보 문제와도 직결된다. 기자들은 AI 자동화를 통해 단순 정보 처리 업무에서 벗어나, 현장 취재, 심층 인터뷰, 교차 검증 등 기사의 신뢰도를 높이는 활동에 더 많은 시간을 투입할 수 있게 되었다. 결국, AI는 인간을 대체하는 것이 아니라, 인간이 단순 반복적 업무에서

벗어나 AI 산출물과 차별화된 고품질 콘텐츠를 만들 수 있도록 돕는 '보조 도구'로 기능하며, 이는 인간 중심 저널리즘의 가치를 강화하는 결과로 이어진다.

- **'현장 취재'와 '심층 분석' 등 인간 고유의 콘텐츠 차별화 전략**

AI 시대 미디어 기업들의 핵심 차별화 전략은 '현장성 확보'과 '심층 분석'이다. 이와 관련하여, 로이터 저널리즘 연구소(Reuters Institute)는 51개국 미디어 기업 경영진과 280명의 디지털 리더를 대상으로 설문조사를 시행하였다.[2] 설문조사에 참여한 대다수 언론사 경영진은 '독자적인 현장 취재'와 '깊이 있는 분석 및 해석 기사'에 대한 투자를 대폭 늘릴 것이라고 응답하였다. 아울러, 특정 시간과 장소에서 인간 기자가 직접 보고 들은 데이터가 AI가 쉽게 대체할 수 없는 '원본성'을 가지고 있다고 평가하였다.

현장 취재 데이터가 가진 '원본성'은 저작권 보호의 핵심적인 근거 중 하나가 될 수 있다. 현장 취재를 통해 확보된 사진, 영상, 인터뷰 내용은 AI가 접근할 수 없는 1차 정보이며, 이는 그 자체로 배타적 권리를 주장할 수 있는 정보이다. 사건 현장에 대한 기자의 고유한 시각과 해석이 담긴 심층 분석 기사 역시 마찬가지다. 심층 분석형 기사는 단순 사실의 나열이 아닌, 인간의 지적 노력과 창의적 판단이 결합된 결과물로서 AI 산출물과는 질적으로 다른 저작물로 인정받을 수 있다.

AI가 만들어내는 정보의 홍수 속에서, 검증된 사실과 신뢰할 수 있는 분석에 대한 사회적 수요는 오히려 증가할 수 있다는 분석이 제기된다. 미디어 기업들은 이러한 수요에 부응함으로써 새로운 가치를 창출할 수 있다. 콘텐츠의 현장성과 깊이를 강화하는 것은 단순한 차별화를 넘어, 저널리즘에 대한 사회적 신뢰를 회복하고 AI 시대 미디어의 존재 이유를 반증하는 생존 전략이 될 것으로 전망된다.

- **지속 가능한 미디어 비즈니스 모델 모색**

일부 미디어 기업들은 기존 데이터 자산을 활용해 새로운 비즈니스 모델을 구축하는 방안을 모색하고 있다. 특히, 미디어 기업들이 수십 년간 축적해 온 방대한 양의 기사 아카이브가 핵심적인 자산으로 평가되고 있다. AI 모델의 성능은 학습 데이터의 질에 의해 결정되므로, 언론사의 고품질 데이터 아카이브는 AI 기업들에게 매력적인 자원이 될 수 있다.

전문가들은 미디어 기업들이 자사의 데이터베이스를 라이선스 형태로 제공하고 이를 통해 수익을 창출하는 비즈니스 모델을 제안한다. 이는 미디어 기업들의 저작물을 단순한 읽을거리로 취급하는 대신, AI 생태계를 구성하는 핵심적인 '원자재'로 재정의하여 그 가치를 인정받는 방식이다. 이러한 비즈니스

2) Nic Newman, "Journalism, media, and technology trends and predictions 2026", Reuters Institute, 2026.01.12., https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2026

모델은 AI 시대의 새로운 산업 질서 속에서 저작권을 새로운 수익원으로 활용할 수 있는 방안이 될 것으로 전망된다.

Ⅲ. 결론 : 신뢰할 수 있는 콘텐츠 제작을 위한 도전과제

• 기술이 증명하는 시대, 저작권 보호의 새로운 패러다임

AI 기술의 확산은 다양한 산업 분야의 효율성을 제고하고 있지만, 동시에 콘텐츠의 진위성을 둘러싼 논란을 촉발하였다. 최근 콘텐츠의 가치는 '무엇을 만들었는가'에서 '그것이 진짜임을 어떻게 증명하는가'로 이동하고 있다. 이는 저작권 보호의 초점이 저작물 자체에 대한 권리 주장을 넘어, 창작 과정의 투명성을 증명하고 인간의 개입을 입증하는 단계로 나아가고 있음을 의미한다.

이러한 맥락 아래, AI 시대의 저작권은 기존의 개념을 넘어, AI가 모방할 수 없는 인간의 창의적 활동을 강화하는 방향으로 나아가야 할 것으로 전망된다. 특히, 업계는 콘텐츠의 진위성을 입증하기 위해 ▲기술 표준, ▲제도적 지원, ▲산업계 참여가 유기적으로 연결되는 다각적인 협력 체계를 구축해야 할 것이다. 이러한 공동의 노력을 통해, 미디어 업계는 기술 혁신과 인간의 창의성이 공존하는 지속 가능한 콘텐츠 생태계를 구축할 수 있을 것으로 관찰된다.

참고문헌

- Leo Martinez, "TikTok expands transparency tools for algorithmic content recommendations", Oakhill Gazette, 2026.01.11., https://oakhillgazette.com/tiktok-transparency-tools/
- Nic Newman, "Journalism, media, and technology trends and predictions 2026", Reuters Institute, 2026.01.12., https://reutersinstitute.politics.ox.ac.uk/journalism-media-and-technology-trends-and-predictions-2026
- Noah Lee, "The 2025 fight for authentic content: niche AI tools vs. tech giants", Digital Journal, 2025. 10.15., https://www.digitaljournal.com/tech-science/the-2025-fight-for-authentic-content-niche-ai-tools-vs-tech-giants/article

기술용어

순번	용어	설명
1	메타데이터 (Metadata)	다른 데이터를 정의하고 기술하는 데이터 또는 다양한 형식의 다른 데이터의 내용 또는 구조를 설명하는 데이터
2	비정형 데이터 (Unstructured Data)	동영상 파일·오디오 파일·사진 등 정의된 구조가 없이 정형화되지 않은 데이터