Biweekly Report
통권 제85호(2026. 7-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제85호(2026. 7-1)

마이크로소프트, 세 가지 기술의 결합과 콘텐츠 진위 인증

뉴스 브리프

AI 기술의 발전으로 사람이 만든 것과 구분하기 어려운 산출물이 빠르게 늘면서, 둘 사이의 경계가 흐려지고 있다. 이러한 환경에서 콘텐츠가 어디에서 왔고 어떻게 바뀌었는지를 확인하는 출처 증명과 워터마크 기술이 주목받고 있다. 마이크로소프트는 디지털 콘텐츠의 진위를 가리기 위해 출처 증명, 워터마크, 디지털 지문을 결합하는 방식을 제시했고, 자사 서비스에서 영상과 음성에 워터마크를 더하는 기능도 도입했다. 핵심 쟁점은 콘텐츠의 출처와 변형 이력을 기술적으로 증명할 수 있는가, 그리고 이용자가 그 신호를 실제로 이해하고 활용할 수 있는가에 있다. 이는 단순한 기술 문제를 넘어 저작권 보호와 침해 식별의 기반으로 이어진다. 본 보고서는 출처 증명과 워터마크의 작동 원리를 살펴보고, 이 기술이 저작권 보호와 어떻게 연결되는지를 분석한다. 나아가 신뢰 기반 콘텐츠 생태계를 위한 과제와 전망을 짚어본다.

뉴스 플러스

Ⅰ. 서론 : 진위 식별 기술의 부상과 신뢰의 과제

- **진위 경계의 붕괴와 콘텐츠 식별 문제**

AI 기술이 빠르게 발전하면서 사람이 만든 것과 구분하기 어려운 산출물이 일상 곳곳으로 퍼지고 있다. 마이크로소프트(Microsoft, 이하 MS)는 자사 업무 서비스인 마이크로소프트 365(Microsoft 365)에서 이용자가 AI를 활용해 제작하거나 변형한 영상과 음성에 워터마크(watermark)를 더하는 기능을 도입했다. 영상에는 화면에 표시되는 시각 표식이, 음성에는 해당 콘텐츠가 AI로 만들어졌음을 알리는 안내 음성이 삽입된다. 이러한 조치는 콘텐츠가 어떻게 만들어졌는지를 이용자가 직관적으로 파악할 수 있도록 돕기 위한 장치다.

문제는 이런 표식이 없을 경우 무엇이 사람이 만든 작품이고 무엇이 AI의 산출물인지 구별하기가 점점 어려워진다는 점이다. MS가 표식을 끄더라도 콘텐츠의 부가 정보가 자동으로 메타데이터(metadata)에 기록되도록 한 것은, 식별 가능성을 최소한이라도 남겨두려는 시도로 해석된다. 콘텐츠의 출처와 변형 이력을 확인할 수 없는 상태가 이어지면, 창작물의 권리 귀속을 따지는 일 자체가 흔들릴 수 있다. 진위를 가리는 기술이 단순한 편의 기능을 넘어 신뢰의 기반으로 다뤄지는 이유가 여기에 있다.

• 출처 증명 기술의 등장 배경과 의의

이러한 문제의식을 바탕으로 마이크로소프트 리서치(Microsoft Research)는 디지털 콘텐츠의 출처와 이력을 다루는 프로젝트인 프로비넌스(Project Provenance)를 진행하고 있다. 프로비넌스는 콘텐츠가 어디에서 비롯되었고 어떤 변형을 거쳤는지에 관한 출처 정보를 사람들이 쉽게 이해하고 활용하도록 만드는 데 초점을 둔다. 이 프로젝트는 기술 자체의 정교함뿐 아니라, 출처 정보를 일반 이용자가 실제로 해석하고 판단에 활용할 수 있는가를 중심에 놓는다는 점에서 기존 접근과 구분된다.

MS는 콘텐츠의 출처와 진위에 관한 산업 표준을 만들기 위한 협의체인 콘텐츠 출처·진위 연합(Coalition for Content Provenance and Authenticity, 이하 C2PA)의 공동 설립에 참여해 왔다. C2PA 표준은 출처 증명을 구현하는 개방형 표준으로, 여기에 워터마크와 디지털 지문을 함께 통합한다. 그 바탕에는 어느 한 기술만으로는 진위를 완전히 가릴 수 없다는 인식이 깔려 있다. 이러한 기술은 창작물의 진위와 변형 여부를 기술적으로 뒷받침한다는 점에서 저작권 보호 논의와 맞닿아 있다. 이러한 점에서 세 기술이 각각 어떻게 작동하고 결합되는지를 살펴보는 일은 신뢰 기반 콘텐츠 환경을 이해하는 출발점이 된다.

Ⅱ. 본론: 출처 증명과 워터마크의 기술적 구조

• 출처 증명과 워터마크의 작동 원리

출처 증명(provenance)은 콘텐츠가 어떻게 만들어지고 어떤 편집을 거쳤는지에 관한 정보를 콘텐츠 파일에 붙여, 그 내력을 확인할 수 있게 하는 기술이다. 쉽게 말하면 콘텐츠에 출처와 편집 이력을 담은 정보를 붙여두고, 이를 나중에 열어볼 수 있도록 하는 방식이다. C2PA 표준에서는 해당 정보를 암호로 서명해 보호한다. 정보가 중간에 조작되면 서명이 깨지도록 설계되어 있어, 콘텐츠를 받는 쪽은 매니페스트(manifest)라 불리는 서명된 기록을 통해 콘텐츠가 서명 이후 바뀌지 않았는지 확인할 수 있다.

다만 출처 증명에는 한 가지 명확한 한계가 존재한다. 서명된 정보가 조작되지 않았음은 보장하지만, 그 정보 자체가 참이라는 것까지 증명하지는 못한다는 점이다. 결국 콘텐츠에 대한 신뢰는 그 정보에 서명한 주체를 얼마나 믿을 수 있는가에 달려 있다. 이 때문에 출처 증명은 누가 서명했는지를 함께 드러내며, 이를 통해 창작자는 사칭이나 위조로부터 보호받고 콘텐츠를 받는 쪽은 신뢰할 수 있는 출처와 그렇지 않은

출처를 구별할 수 있다.

워터마크는 콘텐츠 안에 진위나 출처를 확인할 수 있는 정보를 직접 심어 넣는 기술이다. 미디어 진위 인증에서 주목하는 것은 사람이 알아채기 어려운 비가시 워터마크로, 인코더(encoder)가 콘텐츠의 데이터를 미세하게 바꿔 표식을 심으면 디코더(decoder)가 그 표식을 읽어낸다. 이 표식은 콘텐츠가 일부 변형되더라도 추출될 수 있도록 설계되며, 약한 변형에도 깨지도록 만든 방식과 일정한 변형을 견디도록 만든 견고한 방식으로 나뉜다. 워터마크에는 출처에 관한 정보나, 그 정보를 찾아갈 수 있는 식별값(identifier)이 담긴다. 출처 증명 기록이 파일에서 떨어져 나가더라도, 콘텐츠 안에 남은 워터마크를 통해 원래의 출처 정보를 다시 찾을 수 있다는 점에서 두 기술은 서로를 보완한다.

• 디지털 지문과 콘텐츠 식별의 메커니즘

디지털 지문(digital fingerprint)은 콘텐츠에서 고유한 식별값을 뽑아내, 이를 데이터베이스에 저장된 값과 대조해 같은 콘텐츠인지 확인하는 기술이다. 쉽게 말하면 콘텐츠마다 짧은 식별값을 만들어두고, 나중에 같은 값이 나오는지 견주어 콘텐츠를 가려내는 방식이다. 미디어 진위 인증에서 쓰이는 것은 소프트 해시(soft hash)라 불리는 방식으로, 콘텐츠를 낮은 해상도로 줄여 식별값을 계산하기 때문에 파일 형식이 바뀌거나 크기가 줄어드는 정도의 가벼운 변형에는 같은 값이 유지된다. 앞선 출처 증명이나 워터마크와 달리, 식별값이 콘텐츠 안이 아니라 바깥의 데이터베이스에 따로 저장된다는 점이 특징이다.

이 방식은 저작권 보호와 직접 맞닿아 있다. 디지털 지문은 콘텐츠가 허락 없이 퍼지거나 변형되었는지를 추적하는 데 쓰이며, 이미 알려진 문제 콘텐츠나 저작권으로 보호되는 콘텐츠를 가려내는 데에도 활용된다. 다만 소프트 해시는 서로 다른 콘텐츠가 같은 식별값을 갖는 충돌이 일어날 수 있어, 대조 결과를 사람이 다시 확인하는 절차가 필요하다. 이 때문에 디지털 지문은 단독으로 진위를 확정하기보다, 저작권 침해가 의심되는 상황에서 사람의 판단을 돕는 추적 단서로 기능한다고 평가된다.

• 세 기술의 결합을 통한 신뢰 확보

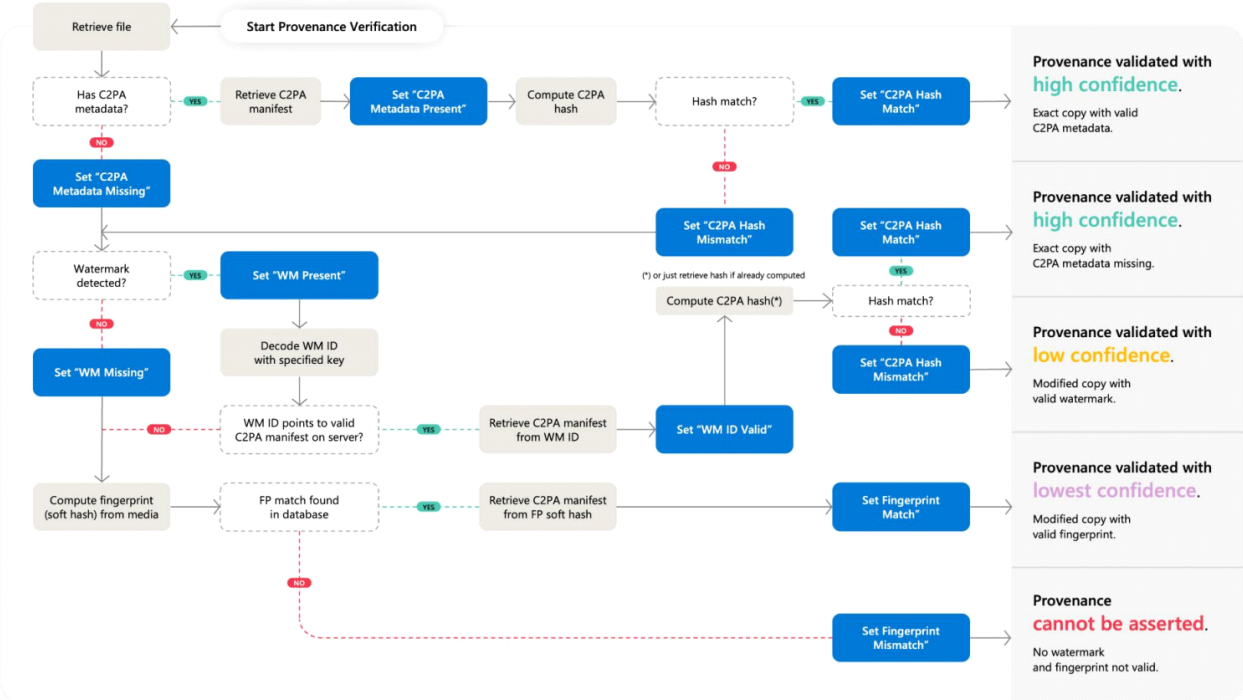
출처 증명, 워터마크, 디지털 지문은 저마다 강점과 약점이 다르다. 어느 하나만 쓰면 그 기술이 뚫리거나 제거되는 순간 진위를 확인할 길이 사라지지만, 셋을 함께 쓰면 하나가 무력화되어도 나머지가 그 자리를 대신할 수 있다. 미디어 진위 인증에서 기준이 되는 것은 출처 증명이다. 콘텐츠에 붙은 서명된 기록이 검증되고, 그 기록에 담긴 식별값이 콘텐츠에서 계산한 값과 일치하면, 해당 콘텐츠가 서명 당시의 원본과 같다는 것을 높은 확신으로 확인할 수 있다.

문제는 출처 증명 기록이 콘텐츠에서 떨어져 나가는 경우다. 서명된 기록은 화면 갈무리나 일부 플랫폼의 업로드 과정에서 쉽게 제거될 수 있기 때문이다. 이때 워터마크가 보완 역할을 한다. 콘텐츠 안에 심긴 워터마크에는 출처 정보를 찾아갈 수 있는 식별값이 담겨 있어, 기록이 사라지더라도 이 식별값을 통해 원래의 서명된 기록을 다시 불러올 수 있다. 여기에 디지털 지문까지 더하면, 워터마크가 지워진 상황에서도

콘텐츠에서 계산한 식별값을 데이터베이스와 대조해 남은 단서를 확보할 수 있다.

이러한 결합 구조는 저작권 보호와 침해 식별에 실질적인 의미를 지닌다. 권리자는 자신의 창작물이 무단으로 퍼지거나 바뀌었는지를 여러 겹의 단서로 추적할 수 있고, 하나의 단서가 제거되더라도 나머지를 통해 원본과 가공물을 구별하는 근거를 확보할 수 있다. 이는 콘텐츠의 출처를 되짚을 수 있는 기술적 여지가 그만큼 넓어진다는 것을 의미한다. 다만 이러한 확인은 정해진 조건 아래에서 성립하는 것이며, 콘텐츠의 진위나 권리 귀속을 그 자체로 확정하는 것은 아니라고 보는 편이 적절하다.

[그림] 세 기술의 결합에 따른 진위 검증 경로와 신뢰도 단계



출처: Jessica Young 외 11명, "Media Integrity and Authentication: Status, Directions, and Futures", Microsoft, 2026.01., https://www.microsoft.com/en-us/research/publication/media-integrity-and-authentication-status-directions-and-futures/

• 진위 인증 기술의 한계와 저작권 함의

세 기술을 결합하더라도 완전한 진위 인증은 어렵다. 워터마크는 콘텐츠를 일정 수준 이상으로 변형하면 지워지거나, 알고리즘 구조가 외부로 노출될 경우 엉뚱한 콘텐츠에 위조되어 붙을 수 있다. 디지털 지문은 서로 다른 콘텐츠가 같은 식별값을 갖는 충돌 문제에서 자유롭지 않고, 출처 증명은 콘텐츠가 인터넷과 단절된 개인 기기에서 만들어질 경우 서명 정보의 신뢰도를 담보하기 어렵다. 이처럼 각 기술에는 저마다의 약점이 있어, 결합은 신뢰도를 높이는 방편일 뿐 어떤 상황에서도 통하는 완결된 해법은 아니다.

이러한 한계는 저작권 보호 측면에서 두 가지를 시사한다. 하나는 진위 인증 기술이 남긴 신호를 곧바로 권리의 증거로 단정하기 어렵다는 점이다. 신호가 제거되거나 위조될 수 있는 만큼, 신호의 유무만으로 침해

여부를 가리기보다 여러 단서를 종합해 판단할 필요가 있다. 다른 하나는 그럼에도 이러한 기술이 저작권 보호의 기반으로서 갖는 의미가 작지 않다는 점이다. 완전하지 않더라도 출처와 변형 이력을 기술적으로 남겨두는 일은, 권리 귀속을 따지고 무단 이용을 추적하는 과정에서 실마리를 제공하기 때문이다.

III. 결론 및 전망: 신뢰 기반 콘텐츠 생태계의 전망

• 진위 인증 기술의 확산과 남은 과제

결국 출처 증명, 워터마크, 디지털 지문은 각각으로는 완전하지 않지만, 서로 결합할 때 콘텐츠의 출처와 변형 이력을 여러 겹으로 남기는 수단이 된다. 이를 통해 무엇이 AI 산출물인지, 콘텐츠가 어떤 경로로 바뀌어 왔는지를 기술적으로 되짚을 여지가 넓어진다. 이는 저작권 보호의 관점에서 창작물의 권리 귀속을 따지고 무단 이용을 추적하는 일이 개인의 주장에만 기대지 않고 기술적 단서 위에서 이뤄질 수 있음을 의미한다. 이러한 점에서 진위 인증 기술은 신뢰 기반 콘텐츠 환경을 떠받치는 토대로 자리 잡아 가고 있다.

따라서 앞으로의 과제는 이 기술들을 얼마나 넓고 일관되게 정착시키느냐에 있다. 기술적으로는 워터마크 제거나 식별값 충돌 같은 약점을 보완하고, 인터넷과 단절된 개인 기기에서도 출처 정보를 안전하게 남길 수 있는 방안을 마련하는 일이 남아 있다. 제도적으로는 한국이 2026년 1월 시행한 인공지능 기본법이 실제와 구분하기 어려운 산출물에 대해 AI가 만들었음을 알리도록 요구하는 것처럼,[1] 표시 의무와 기술 표준을 어떻게 맞물리게 할지가 과제로 떠오른다. 이러한 점에서 기술의 발전과 제도의 정비, 그리고 실제 시장에서의 정착 등이 신뢰 기반 콘텐츠 생태계의 관건이 될 것으로 보인다.

참고문헌

- Microsoft, “Add watermarks to content generated or altered by using AI in Microsoft 365”, 2026.06.12., https://learn.microsoft.com/en-us/microsoft-365/copilot/watermarks
- Microsoft, “Project Provenance”, 2026.06.30. 접속 기준, https://www.microsoft.com/en-us/research/project/project-provenance/
- Jessica Young 외 11명, “Media Integrity and Authentication: Status, Directions, and Futures”, Microsoft, 2026.01., https://www.microsoft.com/en-us/research/publication/media-integrity-and-authentication-status-directions-and-futures/

1) Jessica Young 외 11명, “Media Integrity and Authentication: Status, Directions, and Futures”, Microsoft, 2026.01., https://www.microsoft.com/en-us/research/publication/media-integrity-and-authentication-status-directions-and-futures/

‘시딩’에서 알고리즘 증폭까지, 딥페이크의 단계별 확산

뉴스 브리프

딥페이크의 영향은 영상의 정교함만으로 결정되지 않으며, 최초 게시 이후의 유통 과정에서 확대된다. 새로 만든 익명 계정 등에 영상을 먼저 게시하는 시딩 이후 조회수, 댓글, 공유와 같은 이용자 반응이 추천 신호로 작동하면서 노출 범위가 넓어진다. 이후 같은 영상이 다른 플랫폼으로 옮겨지고 새로운 자막과 설명이 덧붙으면, 하나의 콘텐츠가 여러 플랫폼과 이용자 집단에 반복적으로 노출되면서 해당 주제의 존재감이 커질 수 있다. 따라서 딥페이크 대응은 개별 영상의 진위를 판별하는 데서 그치지 않고 최초 게시 시점과 재게시 순서, 문구 변화, 이용자 반응, 플랫폼 이동을 함께 살펴야 한다. 본 보고서는 시딩에서 추천 노출 확대, 교차 플랫폼 재배포, 외부 인용으로 이어지는 확산 단계를 살펴보고 영상 판별과 유통 경로 분석을 연계하기 위한 과제를 분석한다.

뉴스 플러스

I. 서론: 딥페이크 확산 경로 분석의 필요성

• 제작보다 유통 과정에서 커지는 영향

딥페이크 탐지 전문기업 센시티 AI(Sensity AI)는 딥페이크를 인공지능으로 생성하거나 조작된 영상, 이미지, 음성 등의 디지털 미디어로 정의한다. 센시티 AI는 딥페이크의 영향이 콘텐츠 제작 자체뿐 아니라 이후 어떤 디지털 환경에서 유통되고 확산되는지와 밀접하게 연결된다고 설명한다.

딥페이크는 한 계정에서 공유된 하나의 영상만으로도 수만 명의 이용자에게 도달할 수 있다. 또한, 틱톡(TikTok)과 텔레그램(Telegram) 등에 게시된 영상이 엑스(X), 유튜브 쇼츠(YouTube Shorts), 페이스북(Facebook), 왓츠앱(WhatsApp) 등으로 이동하면 여러 이용자 집단에 반복적으로 노출될 수 있다. 따라서 딥페이크의 영향은 영상의 제작 자체뿐 아니라 어떤 플랫폼에 게시되고 얼마나 넓고 반복적으로 유통되는지에 따라서도 달라질 수 있다.

• **개별 영상 판별 중심 대응의 한계**

개별 영상의 조작 여부를 판별하는 작업은 딥페이크 대응의 중요한 출발점이다. 그러나 시딩(Seeding)된 영상에는 캡션과 댓글을 통해 추가 설명이 붙을 수 있으며, 영상과 관련 문구가 다른 계정과 플랫폼으로 반복해 재공유될 수 있다. 이후 뉴스 사이트와 블로그, 각종 SNS 등이 해당 영상을 인용하면 콘텐츠의 신뢰성이 더욱 높아 보일 수 있다.

따라서 영상의 조작 흔적을 확인하는 작업과 함께 영상이 처음 게시된 계정과 시점, 재공유된 플랫폼, 캡션과 설명의 변화, 외부 사이트의 인용 여부를 살펴볼 필요가 있다. 이를 통해 개별 게시물의 진위뿐 아니라 하나의 영상이 여러 정보 경로를 거쳐 확산되는 과정도 파악할 수 있다.

II. 본론: 딥페이크 확산 단계와 플랫폼 작동 원리

• **익명 계정을 활용한 초기 시딩**

시딩은 새로 만든 익명 계정 등에 딥페이크 콘텐츠를 처음 게시해 이후 확산의 출발점을 형성하는 단계이다. 초기 게시에는 틱톡(TikTok)과 같은 플랫폼에서 일반적인 정보 계정처럼 운영되는 계정이 활용될 수 있다. 이러한 계정 중 일부는 처음부터 합성 콘텐츠를 게시하지만, 일반 콘텐츠를 먼저 올려 팔로워를 확보한 뒤 합성 콘텐츠를 게시하기도 한다.

틱톡이나 텔레그램(Telegram)과 같이 콘텐츠가 빠르게 유통되는 플랫폼에서는 짧고 감정적인 영상이 추천 알고리즘상 유리하게 작용할 수 있다. 초기 게시물에 조회수, 좋아요, 댓글 및 공유 등의 반응이 쌓이면 비슷한 이용 행태를 보이는 다른 이용자에게도 노출되며, 반응과 노출이 서로를 확대하는 순환이 형성된다. 이러한 초기 게시 지점은 이후 영상이 다른 플랫폼과 재게시 네트워크로 이동하는 출발점으로 작동한다.

• **이용자 반응에 따른 추천 노출 확대**

시딩된 영상은 이용자 반응이 쌓이면서 더 넓은 범위에 노출된다. 추천 시스템은 이러한 반응을 기준으로 콘텐츠를 비슷한 이용 행태를 보이는 다른 이용자에게 노출하며, 초기 반응이 콘텐츠의 확산을 견인하고, 넓어진 범위가 다시 새로운 반응을 만드는 자기강화적 순환이 형성된다.

이 과정에서는 이용자의 거주 지역이나 국적보다 시청, 공유 등 상호작용 방식의 유사성이 중요하게 작용한다. 서로 다른 지역이나 집단에 속한 이용자도 유사한 상호작용 패턴을 보이면 동일한 콘텐츠에 노출될 수 있다. 이에 따라 하나의 영상은 대상 설정 없이도 여러 이용자 집단으로 이동하며, 기존에 구분돼 있던 집단이 플랫폼 안에서 겹치면서 같은 콘텐츠를 접할 수 있다.

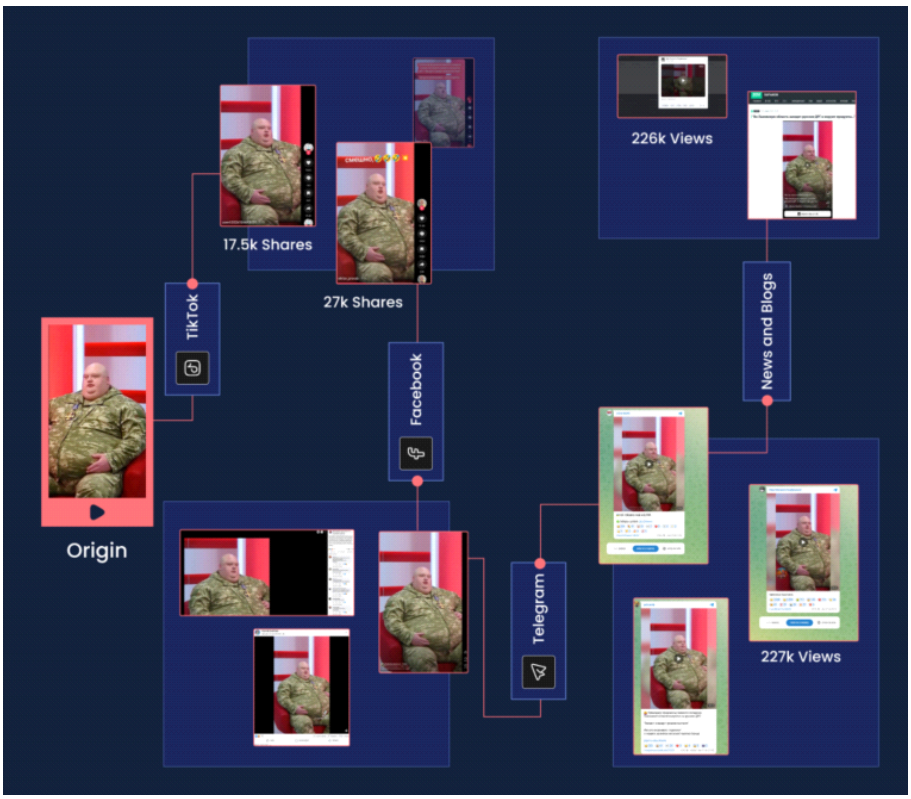
분노, 공포, 충격과 같이 강한 감정적 반응을 일으키는 영상은 추천 알고리즘상 노출 확대에 유리할 수 있다. 댓글과 공유 등 높은 참여를 유발한 영상은 사실 여부와 관계없이 추가 노출을 얻기 쉬우며, 이는 이용자의 관심과 체류 시간을 높이도록 설계된 플랫폼 구조가 딥페이크 확산에 활용되는 방식이다.

이후 이용자 반응이 추가 노출을 부르고, 확대된 노출이 다시 댓글과 공유를 늘리면서 최초 게시자의 지속적인 개입 없이도 콘텐츠가 확산될 수 있다. 결국 플랫폼의 참여 중심 추천 구조가 딥페이크의 반복 노출과 자발적인 재확산을 뒷받침하게 된다.

• **교차 플랫폼 재배포와 외부 인용 효과**

초기 플랫폼에 게시된 영상은 하나의 서비스에만 머무르지 않는다. 틱톡이나 텔레그램에 게시된 영상은 엑스, 유튜브 쇼츠, 페이스북, 왓츠앱 등 인접한 플랫폼과 재게시 네트워크로 옮겨질 수 있다. 이처럼 영상이 여러 플랫폼으로 이동하면 각 플랫폼의 이용자와 추천 체계를 거치면서 노출 범위가 확대될 수 있다.

[그림 1] 딥페이크 영상의 교차 플랫폼 재배포 예시

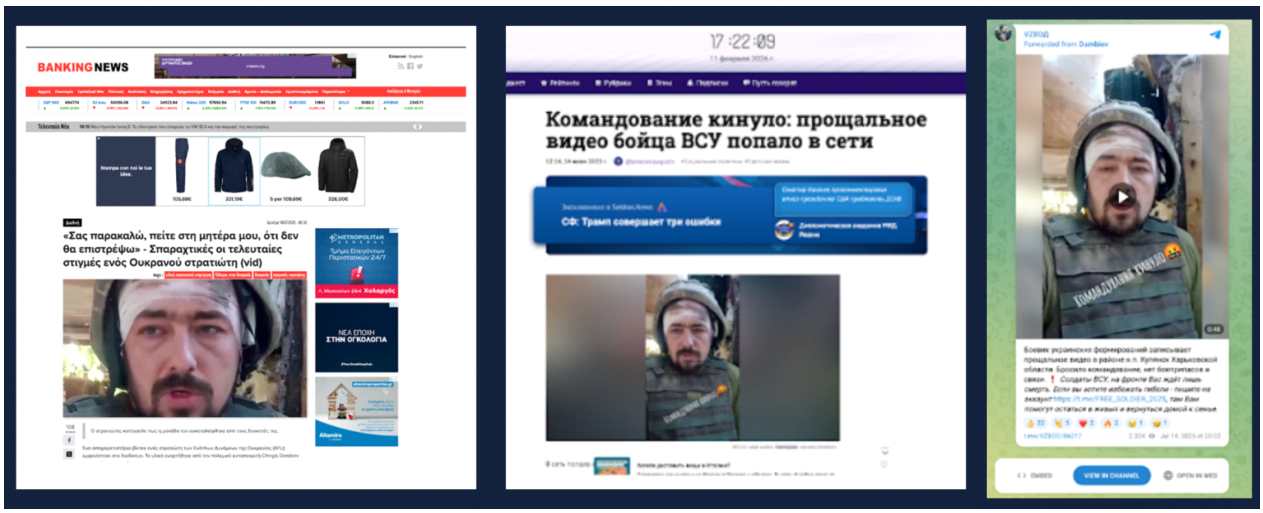


출처: Team Sensity, "The Role of Deepfakes in Cognitive Warfare 2026", Sensity AI, 2026.04.09., https://sensity.ai/blog/the-role-of-deepfakes-in-cognitive-warfare/

플랫폼 간 재배포 과정에서는 영상뿐 아니라 관련 캡션과 설명도 함께 공유되거나 다시 작성될 수 있다. 게시 계정이 추가한 캡션과 댓글은 영상이 전달하는 주제와 유사한 메시지를 반복하거나 강화하는 역할을 한다. 이에 따라 영상 자체의 내용뿐 아니라 함께 제시되는 문구도 이용자가 콘텐츠를 이해하는 방식에 영향을 줄 수 있다.

딥페이크 확산의 영향은 모든 이용자가 해당 영상을 사실로 믿는지 여부만으로 결정되지 않는다. 동일한 내러티브(narrative)가 여러 플랫폼과 이용자 집단에 걸쳐 빠르고 지속적으로 노출되면 정보 환경 전반에서 해당 주제의 존재감이 커질 수 있다. 여기에 뉴스 사이트, 블로그, 각종 SNS 등의 외부 인용이 더해지면 익명 소셜미디어 계정에서 시작된 영상도 신뢰할 만한 정보처럼 보일 수 있다. 공식 뉴스 사이트에 딥페이크가 등장하면 익명 계정에 게시됐을 때보다 신뢰성이 있는 것처럼 인식될 수 있다. 이처럼 외부 사이트의 인용은 콘텐츠의 사실 여부와 별개로 메시지에 정당성과 신뢰성을 부여하는 효과를 낼 수 있다.

[그림 2] 뉴스 사이트와 텔레그램을 통한 딥페이크 서사 강화 예시



출처: Team Sensity, "The Role of Deepfakes in Cognitive Warfare 2026", Sensity AI, 2026.04.09., https://sensity.ai/blog/the-role-of-deepfakes-in-cognitive-warfare/

이러한 확산 구조를 분석하기 위해서는 영상이 처음 등장한 플랫폼과 이후 재게시된 경로, 캡션과 설명의 변화, 조회수와 댓글, 공유 등의 이용자 반응, 외부 사이트의 인용 여부를 함께 살펴볼 필요가 있다. 이를 통해 개별 영상의 진위뿐 아니라 해당 영상이 어떤 경로를 거쳐 확산됐는지도 파악할 수 있다.

III. 결론: 딥페이크 확산 분석의 향후 과제

• 영상 판별과 유통 경로 분석의 연계

결국 딥페이크의 영향력은 영상의 제작 자체뿐 아니라 최초 게시 이후 어떤 경로로 이동하고 얼마나 반복적으로 노출되는지에 따라 확대될 수 있다. 따라서 영상의 조작 여부를 확인하는 작업과 함께 처음 게시된 플랫폼과 시점, 재게시 패턴, 캡션과 설명의 변화, 이용자 반응, 외부 사이트의 인용 여부를 살펴볼 필요가 있다. 이를 통해 개별 게시물의 진위뿐 아니라 하나의 영상이 여러 플랫폼과 계정을 거쳐 확산되는 과정도 파악할 수 있다.

자동화된 탐지 도구는 수집된 미디어를 초기 단계에서 조작 징후에 따라 선별해 사람이 모든 영상을 직접 검토하는 부담을 줄이고, 대규모 자료의 신속한 초기 분류를 지원하는 보조 수단으로 활용할 수 있다. 다만 단순히 진짜와 가짜를 구분하는 결과만으로는 충분하지 않으며, 판단 근거가 재현 가능하고 설명 가능하며 추적할 수 있는 형태로 제시돼야 한다. 분석자는 설명 가능한 기술적 분석 결과를 바탕으로 게시 시점, 이용자 반응, 플랫폼 이동, 외부 인용 등의 유통 맥락을 해석해야 한다. 향후 딥페이크 대응에서는 개별 영상의 진위를 판별하는 기술과 영상의 이동 경로 및 반복 노출 과정을 파악하는 분석을 연계할 필요가 있다.

참고문헌

• Francesco Cavalli, "Deepfake Dissemination Strategy: From Seeding to Algorithmic Saturation", Sensity AI, 2026.06.10., https://sensity.ai/blog/deepfake-dissemination-strategy/
• Team Sensity, "The Role of Deepfakes in Cognitive Warfare 2026", Sensity AI, 2026.04.09., https://sensity.ai/blog/the-role-of-deepfakes-in-cognitive-warfare/

에이전트 바나나로 살펴보는 이미지 편집 자동화와 원본 보존 기술

뉴스 브리프

이미지를 직접 다루지 않고 자연어 지시만으로 편집을 수행하는 에이전트형 편집 도구가 창작 현장에 빠르게 확산되고 있다. 사용자가 원하는 바를 설명하면 시스템이 작업을 단계로 나눠 실행하고, 결과를 스스로 점검한 뒤 이전 단계로 되돌리기까지 수행하는 방식이다. 이러한 다중 턴 편집 도구의 근간에는 그동안의 편집 이력을 압축해 관리하고, 편집 대상만 국소적으로 처리해 원본을 보존하는 기술이 자리한다. 다만 겉보기에는 자연스러운 편집이 반복되는 사이에 손대지 않은 영역까지 조금씩 변형되는 현상이 관찰되며, 무엇이 바뀌고 무엇이 보존되었는지 검증하는 문제가 함께 제기된다. 이러한 기술은 원본성 보존과 편집 추적이라는 저작권 보호의 측면과 맞닿아 있다. 본 보고서는 에이전트형 편집 기술의 작동 원리를 살펴보고, 앞으로 어떤 대응이 필요한지를 전망한다.

뉴스 플러스

I. 서론: 에이전트형 이미지 편집의 부상과 과제

• 창작 워크플로에 스며든 에이전트형 편집 도구

포토샵과 일러스트레이터 등 전문 창작 도구로 알려진 미국의 소프트웨어 기업 어도비(Adobe)는 최근 자사의 이미지·영상 산출 도구 파이어플라이(Firefly)에 창작 에이전트를 결합한 새로운 기능을 공개했다. 사용자가 원하는 브랜드의 스타일과 색상을 설명하면 로고와 색상 팔레트를 한 번에 구성하고, 제품 사진을 숏폼(short-form)으로 바꾸며, 스토리보드에서 곧바로 영상을 만들어내는 방식이다. 창작자가 세부 조작을 일일이 수행하는 대신 원하는 바를 설명하면 시스템이 작업 흐름을 대신 조율한다는 점에서, 이러한 도구는 기존 편집 방식과 구분된다.

이 흐름에서 주목할 점은 여러 작업 세션에 걸쳐 캐릭터와 배경을 반복해 활용하며 일관성을 유지하도록 설계되었다는 것이다. 반복 사용되는 요소를 저장하고 이후 작업에 다시 불러오는 구조는 창작의 연속성을 높인다. 그 과정에서 하나의 원본은 여러 산출물로 확장되고 변형되며, 무엇이 원본이고 무엇이 그로부터 파생된 결과인지를 구분하는 일이 점차 중요해진다.

• 다중 턴 편집 기술의 등장 배경

이러한 도구가 실제 전문 워크플로(workflow)에 안착하기 위해서는 몇 가지 기술적 조건이 충족되어야 한다. 사진, 디자인, 영상 제작 같은 분야에서는 4K 이상의 원본 해상도에서 작업하며, 편집 대상이 아닌 영역은 손대지 않은 채로 유지해야 한다. 그러나 기존 모델은 이미지 전체를 낮은 해상도로 줄여 처리하거나, 사용자가 의도하지 않은 영역까지 함께 바꾸는 과잉 편집(over-editing) 문제를 드러냈다. 여러 지시를 순차적으로 처리하는 다중 턴 편집에서는 앞선 편집이 뒤이은 편집에 의해 훼손되는 한계도 나타났다.

이러한 한계를 넘어서기 위해 텍사스A&M대학교(Texas A&M University) 외 5개 대학과 엑스에이아이(xAI) 외 2개사의 연구진이 공동으로 제안한 것이 에이전트형 편집 프레임워크 에이전트 바나나(Agent Banana)다. 이는 사용자의 지시를 스스로 해석하고 편집 작업을 나누어 수행하는 연구 단계의 편집 시스템으로, 시각·언어 모델 (Vision-Language Model, VLM)의 이해와 추론 능력을 편집 작업에 결합한다. 이 기술은 복잡한 지시를 최소 단위의 작업으로 분해하고, 편집 대상만 국소적으로 처리하며, 각 단계의 상태를 기록해 되돌리기와 재실행을 가능하게 한다는 점에서 앞선 도구들과 구분된다.

II. 본론: 에이전트 바나나의 다중 턴 편집 메커니즘 분석

• 기획자와 실행자로 나뉜 편집 자동화 구조

에이전트 바나나의 뼈대는 편집 작업을 성격이 다른 두 역할로 나누는 데 있다. 하나는 전체 작업을 이해하고 계획하는 기획자(Planner)로, 사용자의 복잡한 지시를 해석해 실행 가능한 하위 작업으로 분해하고 전체 진행 상황을 점검하는 역할을 맡는다. 다른 하나는 실제 편집을 수행하는 실행자(Executor)로, 개별 편집 작업을 국소 영역에 적용하고 중간 결과를 검증하며 오류를 복구하는 역할을 맡는다.

이 구조는 사용자의 지시가 들어온 뒤 하나의 순환 과정을 따라 작동한다. 먼저 기획자가 지시와 현재 이미지 상태를 받아 작업을 하위 목표로 나누고 실행자에 전달한다. 실행자는 각 목표를 수행한 뒤 결과를 되돌려 주고, 기획자는 그 결과가 지시한 목표를 충족하는지 검증한다. 충족하지 못하면 기획자는 다시 계획을 세우거나 이전 상태로 되돌린다. 이 과정은 사용자의 의도가 만족되거나 정해진 편집 횟수에 도달할 때까지 반복된다. 실행자 안에는 결과의 품질을 스스로 점검하는 장치가 있어, 검증을 통과하지 못한 편집은 사용자에게 전달되기 전에 다시 수행된다.

이처럼 역할을 나누고 각 단계를 검증하는 구조는 편집 과정을 되짚어 볼 수 있다는 점에서도 의미를 지닌다. 편집이 하나의 덩어리로 뭉뚱그려 처리되지 않고 최소 단위의 작업으로 분해되어 순차적으로 기록되기 때문에, 어떤 편집이 어느 영역에 어떤 순서로 가해졌는지가 기록된다. 이는 최종 산출물이 원본에서 어떤 경로를 거쳐 변형되었는지를 되짚을 수 있는 기술적 토대가 된다는 것을 의미한다. 이러한 점에서 편집 자동화의 구조 자체가 편집 과정의 추적 가능성과 연결된다.

• 편집 이력을 압축하는 컨텍스트 폴딩 메커니즘

지시와 편집을 여러 번 이어 가는 다중 턴 편집에서는, 편집이 길어질수록 시스템이 다뤄야 할 정보의 양도 함께 늘어난다. 편집을 이어갈 때마다 이전까지의 지시와 이미지 상태를 모두 참조해야 하므로, 이력이 쌓이면 처리할 수 있는 정보의 한계를 넘어서고 불필요한 내용이 뒤섞여 이후 판단을 방해한다. 이를 해결하기 위해 제안된 것이 컨텍스트 폴딩(Context Folding)으로, 이는 늘어나는 편집 이력을 압축된 형태로 저장하는 메커니즘이다.

컨텍스트 폴딩은 편집 이력을 세 가지 레벨로 나누어 다룬다. 첫째는 자산 레벨(Asset Level)로, 각 이미지를 무거운 원본 데이터가 아니라 고유 식별자와 내용 설명, 이전 상태와의 연결 관계를 담은 가벼운 요약 정보로 바꿔 기록한다. 둘째는 실행 레벨(Execution Level)로, 하나의 편집을 수행하는 동안 필요한 도구 선택과 중간 판단 같은 세부 기록을 임시 저장했다가 작업이 끝나면 삭제한다. 셋째는 계획 레벨(Planning Level)로, 사용자의 요구를 성공적으로 충족한 편집 경로만 남긴다. 이렇게 하면 편집이 오래 이어져도 전체 작업 기록을 잃지 않으면서도 정보량의 부담을 덜 수 있다.

이 메커니즘은 편집 이력의 추적 가능성과도 이어진다. 자산 단계에서 각 이미지 상태를 고유 식별자와 이전 상태와의 연결로 기록하기 때문에, 하나의 원본이 어떤 편집을 거쳐 현재의 산출물에 이르렀는지가 이력의 형태로 남는다. 이는 편집의 각 단계가 서로 어떻게 이어지는지를 되짚을 수 있는 기록이 자동으로 구성된다는 것을 의미한다. 이러한 점에서 컨텍스트 폴딩은 정보 처리의 효율을 높이는 기술인 동시에, 산출물의 변형 경로를 확인할 수 있는 기술적 단서를 남기는 역할을 한다.

• 비편집 영역의 변형을 방지하는 이미지 레이어 분해 기술

기존 다중 턴 편집에는 눈에 잘 띄지 않는 문제가 따른다. 논문은 이를 사전 유도 편집 표류(Prior-Induced Editing Drift, PIED)라 부르는데, 이는 편집이 반복되는 사이에 손대지 않은 영역이 서서히 변형되는 현상이다. 이 문제는 편집을 수행할 때마다 이미지 전체를 다시 만들어내는 방식에서 비롯된다. 손대지 않아야 할 영역까지 매번 다시 생성되기 때문에, 편집을 거듭할수록 그 영역이 생성 모델이 선호하는 질감과 양식 쪽으로 조금씩 이동한다. 각 편집 결과는 겉보기에 자연스럽고 때로는 눈으로 구분하기 어려울 만큼 정교하지만, 여러 턴의 편집을 거치며 원본에 대한 충실도는 떨어진다.

이러한 문제를 해결하기 위해 제안된 것이 이미지 레이어 분해(Image Layer Decomposition, 이하 ILD)로, 이는 이미지 전체를 다시 만드는 대신 편집 대상 영역만 떼어내 처리하는 메커니즘이다. ILD는 세 단계로 작동한다. 먼저 편집해야 할 대상 영역을 정밀하게 찾아내고, 그 부분을 원본 고해상도 이미지에서 손실 없이 잘라내 독립된 조각으로 분리한다. 다음으로 모든 편집을 이 조각 안에서만 수행하므로, 배경 영역의 픽셀 상태는 변경되지 않고 그대로 유지된다. 마지막으로 편집이 끝난 조각을 원본 이미지에 자연스럽게 다시 합성해 경계가 어긋나지 않도록 이어붙인다. 이 메커니즘은 대상 영역만 처리하기 때문에, 모델이 본래 다룰 수 있는 해상도의 한계를 넘어서는 초고해상도 이미지 편집도 가능하게 한다.

에이전트 바나나는 ILD를 통해 앞서 지적한 PIED를 억제한다. 손대지 않아야 할 영역이 애초에 다시 생성되지 않고 원본 픽셀 그대로 남기 때문에, 편집이 여러 턴 이어져도 비편집 영역이 변형되지 않는다. 에이전트 바나나는 여러 턴에 걸쳐 비편집 영역의 상태를 비교적 일정하게 유지하며, 이는 앞선 편집이 뒤이은 편집에 의해 훼손되는 문제를 상당 부분 해소한 결과로 평가된다. 편집 대상이 아닌 부분이 원본 상태로 보존된다는 것은, 완성된 결과에서 실제로 손댄 곳과 그대로 남은 곳이 뚜렷이 나뉜다는 것을 의미한다. 이러한 점에서 ILD는 편집의 정밀함을 높이는 기술인 동시에, 원본과 산출물 사이에서 무엇이 보존되고 무엇이 바뀌었는지를 구분할 수 있는 근거가 된다.

III. 결론: 편집 자동화 기술의 확산과 향후 전망

• 원본 보존 기술의 확산과 산업적 과제

결국 에이전트형 편집 기술이 보여주는 것은, 편집 자동화가 작업 속도를 높이는 데 그치지 않고 원본을 다루는 방식 자체를 바꾼다는 점이다. 기존의 편집이 완성된 결과물만 남겼다면, 이 기술은 어떤 지시가 어떤 순서로 처리되었는지, 원본의 어느 부분이 손대어졌고 어느 부분이 그대로인지를 함께 남긴다. 편집의 결과뿐 아니라 그 과정과 원본의 상태까지 확인할 수 있게 된 것이다. 이는 창작의 효율을 높이는 흐름이 원본과 산출물의 관계를 들여다볼 수 있는 여지를 함께 넓히고 있다는 것을 의미한다.

이를 통해 앞으로의 과제도 분명해진다. 어도비 파이어플라이처럼 편집 자동화 도구가 창작 현장에 빠르게 확산되면서, 원본에서 파생되는 산출물의 수와 종류는 계속 늘어날 전망이다. 따라서 편집이 어떻게 진행되었는지를 기록하고 원본의 보존 여부를 확인하는 기능이 도구 안에 표준으로 자리 잡을 필요가 있다. 동시에 완성된 결과만으로는 드러나지 않는 원본의 훼손을 단계마다 점검하는 검증 방식도 함께 갖추어야 한다. 이러한 기술적 보완과 산업적 합의가 이루어질 때, 편집 자동화는 창작의 효율과 원본의 신뢰를 함께 지키는 방향으로 발전할 수 있을 것으로 보인다.

참고문헌

• Forest Key, "Adobe Firefly introduces new agentic capabilities and an upgraded creative AI studio built for the way you work", Adobe, 2026.06.18., https://blog.adobe.com/en/publish/2026/06/18/adobe-firefly-introduces-new-agentic-capabilities-and-an-upgraded-creative-ai-studio-built-for-the-way-you-work

• Ruijie Ye 외 12명, "Agent Banana: High-Fidelity Image Editing with Agentic Thinking and Tooling", arXiv, 2026.02.09., https://arxiv.org/pdf/2602.09084