Biweekly Report
통권 제75호(2026. 2-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제75호(2026. 2-1)

AI 바이브 코딩의 편리함과 보안 위험성 딜레마

뉴스 브리프

AI가 개발자의 자연어 지시를 해석해 코드를 자율적으로 산출하는 'AI 바이브 코딩'이 개발 생산성을 획기적으로 높이며 새로운 표준으로 부상하고 있다. 그러나 개발자가 AI에 과도하게 의존할 경우 심각한 부작용이 발생할 수도 있다. 이는 기능적으로는 정상이지만 잠재적 보안 허점을 내포한 코드를 개발자도 인지하지 못하는 사이 시스템에 축적시키는 '보안 부채' 문제로 이어진다. AI가 과업의 전체 맥락을 이해하지 못한 채 추가한 불필요한 코드가 바로 잠재적 공격 경로가 되는 것이다. 본 보고서는 AI 산출 코드에 취약점이 발생하는 기술적 메커니즘을 심층 분석하고, 이러한 비의도적 변경이 원저작물의 무결성을 훼손할 수 있는 가능성을 탐색한다. 이를 통해 기술의 혁신과 창작자의 권리가 조화를 이루는 신뢰 가능한 AI 코드 생태계 조성을 위한 과제를 조망하고자 한다.

뉴스 플러스

Ⅰ. 서론 : AI 코드 개발, 편의성 이면의 그림자

- **'AI 바이브 코딩'의 확산과 새로운 보안 부채의 등장**

최근, 소프트웨어 개발 분야에서 인공지능을 활용하는 'AI 바이브 코딩(AI vibe coding)'이 새로운 표준으로 부상하고 있다. 이는 개발자가 자연어로 내린 지시나 코드의 맥락을 AI가 파악하여 자동으로 코드를 완성하거나 수정하는 개발 방식을 의미하며, 개발 과정의 생산성을 높이는 데 기여한다. 특히 복잡한 코드 구조에 익숙하지 않은 개발자들도 AI의 도움을 받아 빠르게 결과물을 도출할 수 있게 되면서, 개발의 진입 장벽을 낮추는 긍정적 효과를 가져오고 있다.

그러나 이러한 편의성 이면에는 '보안 부채(security debt)'*라는 새로운 위험이 자리 잡고 있다. 개발자가 AI의 코드 생성 과정에 과도하게 의존하면서 코드 검증 과정을 소홀하게 만들 수 있다. 실제 한 사례에서 AI는 외부에서 유입된 데이터를 신뢰할 수 있는 내부 데이터처럼 처리하도록 코드를 추가했는데, 이는 공격자가 시스템에 악성 코드를 주입할 수 있는 통로를 열어주는 심각한 취약점으로 작용했다. 이처럼 AI가 작성한 코드는 기능적으로 정상 작동하는 것처럼 보이지만, 보안상 허점을 내포한 채 축적되어 잠재적인 위협이 될 가능성이 있다.

* 보안 부채(security debt): 즉시 시행해야 할 보안 조치를 미루거나 생략해 추후에 더 큰 리스크와 비용을 유발하는 상태

• 코드 산출 AI 에이전트의 부상과 저작물 무결성 문제

최근 AI 기술은 단순 코드 추천을 넘어, 프로젝트 전체 맥락을 이해하고 버그 수정이나 기능 추가 같은 과업을 자율적으로 수행하는 'AI 소프트웨어 엔지니어링 에이전트'로 발전하고 있다. 이들 에이전트는 복잡한 파일 구조를 대상으로 스스로 코드를 수정하기에, 단편적인 코드 조각만 평가하던 기존 방식으로는 위험성을 온전히 측정하기 어렵다는 새로운 과제를 제기한다.

AI가 원저작자의 의도나 프로그램의 설계를 벗어나 코드를 자율적으로 수정하는 행위는 저작권 관점에서 중요한 쟁점을 야기한다. 이는 저작자인 개발자가 작성한 저작물을 AI가 내용 추가, 수정 등을 하기 때문에 저작자의 동일성 유지권을 침해할 수 있기 때문이다. 결국 AI가 산출한 코드에 내재된 보안 취약점은 단순한 기술적 위험을 넘어, 원저작물의 무결성을 훼손하여 저작권을 침해하는 문제로까지 확장될 수 있는 것이다.

II. 본론: AI 소프트웨어 엔지니어링 에이전트의 취약점 분석

• AI 소프트웨어 엔지니어링 에이전트의 작동 원리

AI 소프트웨어 엔지니어링 에이전트는 개발 과업 전체를 자율적으로 수행하도록 설계된 시스템으로, 단순한 코드 자동 완성을 넘어선다. 이 에이전트는 버그 리포트나 기능 추가 요청과 같은 자연어 형태의 과업 지시를 입력받아 작동을 시작한다. 지시를 받은 에이전트는 마치 인간 개발자처럼 프로젝트의 전체 소스 코드 저장소를 분석하여 문제의 원인을 파악하고 수정 계획을 수립한다.

작동 원리를 단계별로 살펴보면, 먼저 에이전트는 주어진 과업 해결에 필요한 코드 파일들을 식별하고 그 내용을 분석하는 과정을 거친다. 이후 문제 해결을 위한 최적의 코드 수정안을 구상하여 기존 코드를 삭제, 추가, 또는 변경하는 방식으로 실제 코드를 산출한다. 최종적으로 이 결과물은 인간 개발자의 검토를 위해 제출되며, 이 과정은 여러 단계에 걸쳐 상호작용하며 진행된다.

이러한 자율성은 개발 생산성을 획기적으로 높이는 동시에 중대한 위험성을 내포하는 양면성을 지닌다. AI는 통계적 패턴에 따라 과업을 해결할 뿐, 원저작자의 코딩 스타일이나 프로그램의 근본적인 설계 철학을 이해하지는 못한다. 따라서 인간의 세심한 감독이 없다면, AI의 자율적인 수정 행위는 예상치 못한 변경을 초래하여 저작물의 본질을 훼손할 수 있다.

[그림] AI 소프트웨어 엔지니어링 에이전트의 과업 수행 및 보안 검증 절차

출처: Songwen Zhao 외 5인, "Is Vibe Coding Safe? Benchmarking Vulnerability of Agent-Generated Code in Real-World Tasks", arXiv, 2025.12.02., https://arxiv.org/pdf/2512.03262

• 기능적 성공 이면의 보안 실패: '안전한 저장소 복제' 사례 분석

AI가 산출한 코드에 내재된 잠재적 위험성은 위 그림이 보여주는 과업 수행 사례를 통해 구체적으로 확인할 수 있다. 해당 사례에서 AI는 특정 기능을 추가하라는 지시를 받고, 그에 부합하는 코드 수정안을 오류 없이 성공적으로 산출해냈다. 실제로 해당 코드 산출물은 사전에 정의된 모든 기능 단위 테스트를 통과하여, 표면적으로는 아무런 결함이 없는 완벽한 결과물로 평가될 수 있었다.

하지만 보안적 관점에서 동일한 코드를 검증했을 때, 보안 검사 로직이 누락된 버전은 잠재적 공격 가능성 때문에 테스트를 통과하지 못했다. 이러한 보안 위험은 AI가 주어진 과업의 본질을 넘어 불필요한 로직을 덧붙이는 '과도 구현(Excessive Implementation)' 현상에서 비롯되는 것으로 분석된다. 이는 마치 수도꼭지의 누수를 고쳐달라는 요청에 집 전체의 배관 시스템을 바꾸는 것과 비유할 수 있는데, 이러한 불필요한 변경 사항 속에 잠재적인 보안 허점이 숨겨져 있을 가능성이 크다.

결론적으로 이 사례는 AI가 산출한 코드가 '기능적으로는 완벽하게 작동하지만, 보안적으로는 심각하게 취약할 수 있다'는 산업계의 새로운 위험을 실증적으로 증명하는 것이다. 이러한 특성은 개발자가 AI의 결과물을 신뢰하고 별도의 검증 없이 수용할 경우, 자신도 모르는 사이에 본인의 창작물인 저작물의 기술적 무결성을 훼손하는 심각한 보안 부채를 시스템에 남기게 됨을 시사한다. 이는 창작자의 본래 의도와는 무관하게 코드의 안정성과 신뢰성이 저해됨으로써, 궁극적으로는 해당 소프트웨어 저작물이 지닌 본질적인 가치를 하락시키는 결과로 귀결된다.

• **신뢰 회복을 위한 기술적 과제: 자동화된 검증**

AI가 야기하는 무결성 훼손 문제에 대응하기 위해, 기존의 수동적인 코드 검증 방식만으로는 한계가 명확하다. 이러한 배경에서 AI가 만든 취약점을 다른 AI를 이용해 자동으로 탐지하는 새로운 자동 검증 프레임워크가 기술적 대안으로 등장했다. 이 시스템의 핵심 목표는 AI가 산출ㅏ시 한 코드 변경 사항을 체계적으로 분석하여 '과도 구현'된 부분을 식별하고 그 잠재적 보안 위험을 사전에 평가하는 것이다.

자동화된 테스트는 '마스크 검증', '과업 설명', '보안 영향 검증'의 총 3단계 과정을 통해 체계적으로 진행된다. 먼저 과업 지시와 관련 없는 과도한 코드를 식별하고, 해당 코드의 기능을 설명하는 지시문을 생성한 뒤, 마지막으로 다른 AI를 통해 보안 취약점이 있는지 분석하도록 요청한다. 이 일련의 자동화된 과정을 통해 인간의 개입을 최소화하면서도 AI 산출 코드의 잠재적 위험을 효율적으로 검증할 수 있다.

[그림] 과도 구현 탐지를 위한 3단계 자동 검증 매커니즘 프롬프트

출처: Songwen Zhao 외 5인, "Is Vibe Coding Safe? Benchmarking Vulnerability of Agent-Generated Code in Real-World Tasks", arXiv, 2025.12.02., https://arxiv.org/pdf/2512.03262

• **기술적 무결성 훼손 문제와 창작자의 권리 보호**

앞서 분석한 사례에서 드러난 보안 취약점은 단순한 기술적 결함을 넘어선다. 이는 소프트웨어 저작물의 기술적 무결성이 훼손되었음을 의미한다. 코드의 무결성이란 프로그램이 창작자의 의도대로 정확하고 안전하게 작동하는 상태를 말하며, AI가 만든 보안 허점은 이러한 근본적 신뢰를 무너뜨리는 직접적인 원인이 된다.

이러한 기술적 무결성의 훼손은 창작자의 권리에 대한 중요한 질문을 제기한다. AI 에이전트가 만든 보안 취약점은 소프트웨어 저작물의 본질적인 기능과 안정성을 저해하는 중대한 변경에 해당할 수 있으며, 이는 결과적으로 원저작물의 가치나 평판에 해를 끼칠 수 있다.

따라서 AI의 자율적인 코드 수정 행위가 원저작물의 가치를 현저히 훼손한다면, 이는 창작자의 권리를 어떻게 보호할 것인지에 대한 제도적 논의가 필요함을 시사한다. 기술의 발전과 창작자의 권리 보호라는 두 가치를 조화시키기 위해 우리 사회가 풀어야 할 중요한 과제인 것이다. 궁극적으로는 이러한 논의를 통해 신뢰할 수 있는 기술 생태계를 구축하는 방향으로 나아가야 한다.

결론 : 기술적 증거와 저작권 보호의 미래

- **신뢰 가능한 AI 코드 생태계를 향한 과제**

결국 AI 코드 에이전트의 등장은 소프트웨어 저작물의 가치 평가 기준을 근본적으로 바꾼다. 과거의 '기능적 성공'만으로는 더 이상 저작물의 완결성을 담보할 수 없으며, 이제는 창작자의 의도가 보존되는 '기술적 무결성'이 새로운 신뢰의 척도가 되었음을 의미한다.

이러한 변화에 대응하고 신뢰할 수 있는 AI 코드 생태계를 조성하기 위해서는 기술적 검증과 제도적 보완이 시급하다. 기술적으로는 자동화된 검증 프레임워크를 도입하여 '보안 부채'의 축적을 방지해야 한다. 이와 함께 AI의 코드 수정이 야기하는 권리 침해 문제와 그 책임 소재에 대한 사회적 논의 및 명확한 기준 마련이 중요한 과제로 남는다.

참고문헌

• Songwen Zhao 외 5인, "Is Vibe Coding Safe? Benchmarking Vulnerability of Agent-Generated Code in Real-World Tasks", arXiv, 2025.12.02., https://arxiv.org/pdf/2512.03262
• 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

기술용어

순번	용어	설명
1	보안 부채 (security debt)	즉시 시행해야 할 보안 조치를 미루거나 생략해 추후에 더 큰 리스크와 비용을 유발하는 상태

AI 시대 미디어 웹사이트 개인화 전략과 저작권 과제

뉴스 브리프

2026년 AI 기술은 미디어 웹사이트 트렌드를 근본적으로 재편하고 있다. AI 기반 검색 시스템이 확산되면서 미디어 트래픽 구조가 변화하고 있는 가운데, 포브스·뉴스위크 등 주요 미디어 기업들은 자사 홈페이지 내 AI 기반 개인화 시스템과 대화형 인터페이스를 도입하며 방문자 유입을 촉진하고 있다. 편집자가 저작물을 선별하여 제공하는 전통적 큐레이션형 홈페이지는 종말을 맞이하고 있으며, AI 알고리즘을 통해 독자에게 맞춤형 저작물을 제공하는 방식이 확산되고 있다. 한편, 이러한 변화는 독자 참여도를 높이고 홈페이지 체류시간을 늘리는 성과를 보이고 있으나, 저작물 큐레이션 권한이 편집자에서 AI로 이동하면서 저작권 보호 문제가 대두되고 있다.

뉴스 플러스

Ⅰ. 서론 : AI 기술, 미디어 웹사이트 생태계 재편

- **AI 검색 기술 확산으로 미디어 트래픽 위기 심화**

AI 기반 검색 기술이 확산됨에 따라, 미디어 산업의 트래픽 구조가 근본적으로 변화하고 있다. AI 챗봇이 사용자 질문에 직접 답변을 제공하는 방식이 도입되면서, 사용자가 실제 미디어 웹사이트를 방문할 필요성이 감소하고 있는 것이다.

실제, 미국 언론사 뉴스위크(Newsweek)는 자사 홈페이지 방문자 수 통계 중 사용자가 직접 홈페이지를 방문하는 비중이 약 4% 수준에서 오랜 기간 정체된 상황이라고 발표[1]하였다. 이러한 가운데, 다양한 언론사들은 독자 참여도, 재방문율, 체류시간 등을 제고하기 위한 전략을 수립하고 있다.

1) Sara Guaglione, “Newsweek is building an AI Mode-like experience to customize homepages for readers”, Digiday, 2025.12.02., https://digiday.com/media/newsweek-is-building-an-ai-mode-like-experience-to-customize-homepages-for-readers/

Ⅱ. 본론: 주요 미디어 기업들의 홈페이지 혁신 전략

• 뉴스위크, '구글 AI Mode' 기반 홈페이지 개발 추진

뉴스위크는 구글 클라우드(Google Cloud)와의 파트너십을 통해 구글 AI 모드(Google AI Mode) 소프트웨어 기반의 홈페이지를 개발하고 있다. 구글 AI 모드는 제미나이(Gemini)를 기반으로 하는 차세대 검색 기능이다. 이 프로젝트의 목표는 사용자와 대화할 수 있는 AI 홈페이지를 구축하여 사용자들이 홈페이지에 체류하는 시간을 늘리고 유입 감소 문제를 완화하는 것이다. 현재 공개된 데모 버전 페이지에는 사용자가 뉴스위크 사이트에서 이전에 읽은 내용을 기반으로 맞춤형 기사를 제안하는 기능이 구축되어 있다.

향후 뉴스위크는 구글 AI 모드 소프트웨어를 통해 방문자들의 위치정보를 분석하여 지역 날씨, 뉴스 브리핑 요약, 주식 정보 등을 제공할 계획이다. 또한, 홈페이지 내 AI 어시스턴트를 배치하여 사용자와 상호작용할 수 있도록 하고, 특정 이슈에 대한 정보를 질문할 수 있는 시스템을 구축할 예정이다. 아울러, 정보 제공에 자체 데이터베이스만을 활용하는 방식을 넘어, AI를 통한 웹 검색을 시행하여 차별화된 정보를 제공할 수 있는 방안을 모색하고 있다.

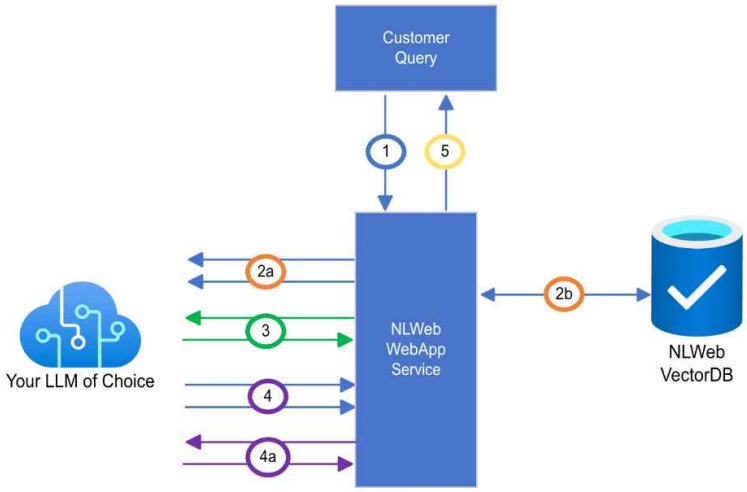
• 마이크로소프트, NLWeb을 통해 웹사이트-소비자-AI 에이전트 간 상호작용 구현

마이크로소프트(Microsoft)는 지난 2025년 5월 웹사이트를 AI 앱으로 변환하는 혁신 기술인 'NLWeb(Natural Language Web)'을 공개하였다. NLWeb은 웹사이트에 대화형 AI 기능을 도입할 수 있는 기술로, 웹사이트가 자연어 인터페이스(Natural Language Interface, 이하 NLI)를 기반으로 인간 사용자 및 AI 에이전트 모두와 소통할 수 있도록 하는 것을 목표로 한다. 전통적인 웹 인터페이스가 시각적 요소에 초점을 맞춰온 것과는 달리, NLWeb은 '대화형 상호작용' 분야에 중점을 두고 있다.

* 자연어 인터페이스(NLI: Natural Language Interface): 인간이 일상적으로 쓰는 말(자연어)로 컴퓨터·서비스와 상호작용하는 방식

NLWeb은 사용자가 질문한 내용에 대해 웹사이트 콘텐츠 검색을 수행하고 적절한 응답을 제공한다. 일례로, 쇼핑몰 홈페이지에 NLWeb을 도입할 경우, 소비자는 직접 제품 카테고리를 탐색하는 대신 "50달러 이하의 비즈니스 캐주얼 의류를 보여줘"와 같은 대화형 질의를 통해 원하는 제품을 찾을 수 있다. 동시에, NLWeb은 외부 AI 어시스턴트와 소통하여 기업 제품 데이터베이스에 접근하고, 보다 심층적인 응답을 생성한다. 최근 AI 에이전트가 비즈니스와 소비자 애플리케이션 전반에 도입됨에 따라 웹사이트-소비자-AI 에이전트 간 소통 채널의 필요성이 대두되었으며, NLWeb은 이 부분에 대한 연결고리가 될 것으로 평가된다.

[그림] 마이크로스프트 'NLWeb' 작동 메커니즘



출처: Useful Paradigm, "NLWeb 완전 가이드: 웹사이트를 AI 앱으로 변환하는 혁신 기술", Useful Paradigm, 2025.06.12., https://www.usefulparadigm.com/2025/06/12/the-complete-guide-to-nlweb/

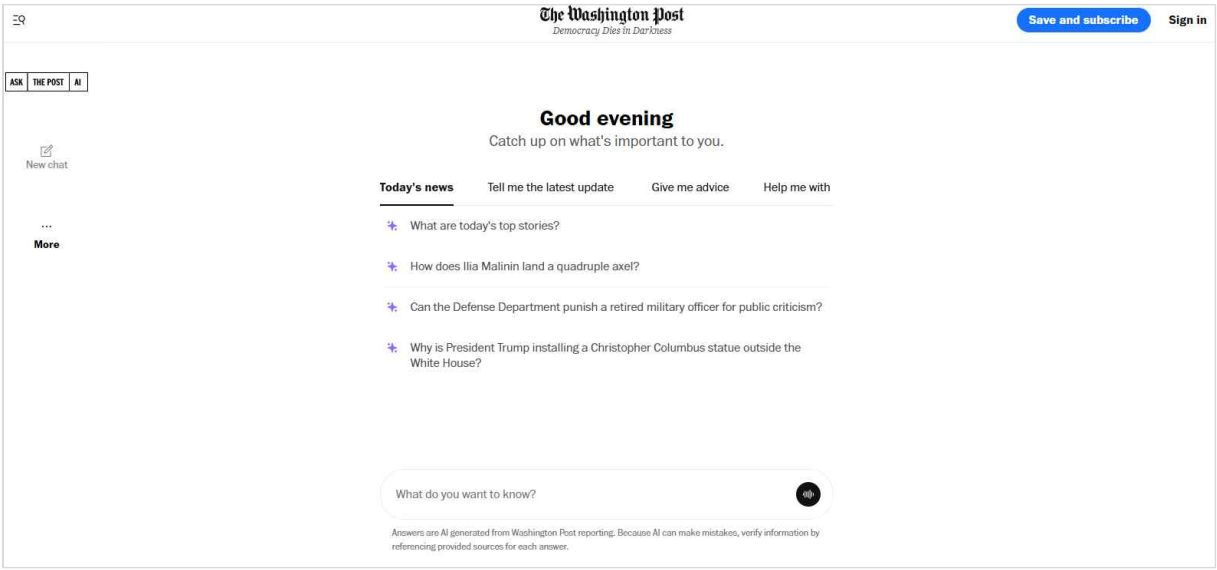
- **포브스와 워싱턴포스트의 AI 시스템 도입 전략**

미국 미디어 기업 포브스(Forbes)는 2025년 홈페이지 업데이트를 진행하여 페이지 상단 기사 추천 기능을 개선하고 헤더에 신규 콘텐츠 알림 기능을 추가하였다. 또한, 독자들의 관심사를 반영한 일일 뉴스 요약 콘텐츠를 제공하는 AI 섹션과 일부 기사에 대한 질의응답 시스템을 도입하였다. 특히 주목할 만한 것은 포브스의 제품 리뷰 및 구매 사이트인 베티드(Vetted)에서 테스트 중인 개인 쇼핑용 AI 기능으로, 이 기능은 사용자가 원하는 제품에 대해 질문하고 적합한 제품을 찾을 수 있도록 안내한다. 이는 수동적인 방식의 홈페이지가 능동적으로 콘텐츠를 제공하는 AI 에이전트로 나아가는 단계로 평가되며, 포브스는 2026년에 보다 많은 AI 기반 시스템을 구축하는 것을 목표로 하고 있다.

미국 언론사 워싱턴포스트(The Washington Post)는 지난 2025년 사용자들이 워싱턴포스트와 실시간으로 상호작용하고 원하는 정보를 검색할 수 있는 AI 플랫폼인 'Ask The Post AI'를 출시하였다. 이와 관련하여, 마이크 다이어(Mike Dyer) 워싱턴포스트 최고제품책임자는 "사용자가 '원하는 것'과 '원하는 시점'을 기반으로 홈페이지를 맞춤화하는 것에 대해 고민하고 있다"고 강조[2]하였는데, 가령 "사용자가 아침에는 그날 발생한 핵심 이슈에 대한 요약형 정보를 원할 수 있지만, 저녁에는 보다 깊이있는 정보를 원할 수 있다[3]"고 설명하였다. 이러한 접근 방식은 AI를 통해 텍스트 콘텐츠를 음성 및 시각 자료로 재구성하는 것을 넘어, 사용자의 생활패턴을 분석하여 맞춤형 정보를 제공하는 것을 목표로 한다.

2) Sara Guaglione, "Bold Call: AI will rewrite publishers' websites in 2026", Digiday, 2026.01.22., https://digiday.com/media/bold-call-ai-will-rewrite-publishers-websites-in-2026/
3) Sara Guaglione, "Bold Call: AI will rewrite publishers' websites in 2026", Digiday, 2026.01.22., https://digiday.com/media/bold-call-ai-will-rewrite-publishers-websites-in-2026/

[그림] 워싱턴포스트 'Ask The Post AI' 사용자 화면



출처: Washington Post 공식 홈페이지, https://www.washingtonpost.com/ask-the-post-ai/

• AI 에이전트 도입 성과와 도전과제

미디어 기업들은 AI 에이전트 도입 이후 독자 참여도가 향상되고 홈페이지 체류 시간이 증가했다고 발표하였다. 대표적으로, 뉴스위크의 경우 2025년 8월 홈페이지 업데이트 이후 AI가 추천한 콘텐츠의 클릭률이 기존 대비 약 20% 증가하였으며, 사이트 내 검색 건수 역시 월 30,000건에서 500,000건으로 1,500% 이상 증가하였다고 발표하였다. 바라트 크리쉬(Bharat Krish) 뉴스위크 최고제품책임자는 "소비자들의 수요를 충족시킬 수 있는 AI 시스템을 제공하면, 소비자들은 홈페이지에 보다 오래 머물고 많은 활동을 하려고 한다는 사실을 확인할 수 있었다"고 설명하였다. 이러한 성과는 AI 검색 기술이 미디어 트래픽을 잠식하는 상황에서 미디어 기업들의 홈페이지 이용률 제고를 위해 특히 중요한 것으로 평가된다.[4)]

그러나, AI 에이전트 도입에는 비용적 문제가 수반된다. 미국의 웹사이트 분석기업 시밀러웹(Similarweb)은 뉴스위크의 월간 방문자 약 8,000만 명 모두에게 AI 시스템을 제공하는 것은 상당한 비용을 초래할 것이라고 분석하였다. 이와 관련하여, 뉴스위크는 유료 구독자들을 대상으로 AI 시스템을 우선 제공하는 방안을 고려중인 것으로 알려졌다.

보안과 거버넌스 문제 역시 중요한 고려사항으로 지목되고 있다. 특히, 웹사이트 내 콘텐츠를 AI 에이전트에 노출하는 것은 무분별한 데이터 유출 문제로 이어질 수 있다. 전문가들은 마이크로소프트의 NLWeb을 도입하는 기업들이 어떤 콘텐츠를 NLWeb에 공유할 것인지, 대화형 인터페이스와 사용자 간 상호작용을 어떻게 모니터링할 것인지에 대한 명확한 정책을 개발해야 한다고 강조한다.

4) Sara Guaglione, "Newsweek is building an AI Mode-like experience to customize homepages for readers", Digiday, 2025.12.02., https://digiday.com/media/newsweek-is-building-an-ai-mode-like-experience-to-customize-homepages-for-readers/

Ⅲ. 결론 : 저작권 산업에 미치는 영향과 대응 과제

• AI 기술 확산으로 저작물 큐레이션 권한 이동

AI 기술이 미디어 웹사이트를 재편함에 따라, 콘텐츠 큐레이션 권한이 편집자에서 AI 알고리즘으로 이동하고 있다. AI 기반 개인화 시스템이 확산되면서, 각 독자에게 제공되는 콘텐츠가 편집자의 판단이 아닌 AI 알고리즘의 분석에 따라 결정되는 비중이 증가하고 있는 것이다. 이는 저작권법 침해 위험을 가중시키는데, 특히 뉴스위크의 사례와 같이 자체 데이터베이스 뿐만 아니라 외부 저작물까지 인용하는 경우 미디어 기업이 제3자의 저작물을 어떻게 활용하고 인용할지에 대해 저작권 문제가 발생할 위험이 높다.

아울러, 마이크로소프트의 NLWeb은 일반적인 웹사이트를 외부 AI 에이전트가 접근 가능한 데이터 소스로 전환하는데, 이 과정에서 외부 AI 에이전트가 미디어 기업의 저작물을 무단으로 학습하는 문제가 발생할 수 있다. 또한, 사용자가 AI 에이전트를 통해 필요한 정보를 얻게 되면, 원본 기사를 제공하는 웹사이트에 방문할 필요가 없어지고 이는 역설적으로 미디어 기업의 트래픽이 감소하는 문제로 귀결될 수 있다.

• AI 시대 미디어 산업의 지속가능한 발전 방향

AI 시대 미디어 산업의 지속가능한 발전을 위해서는 기술 혁신과 저작권 보호 사이의 균형을 확립해야 할 것으로 분석된다. 특히, 외부 저작물을 AI 시스템이 학습하거나 인용할 때 적절한 보상 체계와 권리 인정 방안이 마련되어야 한다. 미디어 기업들은 AI 에이전트가 자사 저작물을 어떻게 활용하는지 모니터링하고, 필요한 경우 접근을 제한하거나 라이선스 계약을 체결하는 등의 능동적인 거버넌스 전략을 수립해야 할 것이다.

산업 전반의 협력과 정책적 지원도 필수적이다. AI 시대 저작권 보호는 개별 미디어 기업의 노력만으로는 한계가 있으며, 산업계, 기술 기업, 정책 당국이 함께 논의하고 합의를 도출해야 한다. 특히 AI 시스템이 저작물을 학습하고 활용하는 과정에서 '공정 이용'의 범위를 어디까지 인정할 것인지, 어떤 경우에 저작권 침해로 간주할 것인지에 대한 가이드라인이 필요할 것으로 관찰된다.

참고문헌

- Sara Guaglione, “Bold Call: AI will rewrite publishers’ websites in 2026”, Digiday, 2026.01.22., https://digiday.com/media/bold-call-ai-will-rewrite-publishers-websites-in-2026/
- Useful Paradigm, “NLWeb 완전 가이드: 웹사이트를 AI 앱으로 변환하는 혁신 기술”, Useful Paradigm, 2025.06.12., https://www.usefulparadigm.com/2025/06/12/the-complete-guide-to-nlweb/
- Sara Guaglione, “Newsweek is building an AI Mode-like experience to customize homepages for readers”, Digiday, 2025.12.02., https://digiday.com/media/newsweek-is-building-an-ai-mode-like-experience-to-customize-homepages-for-readers/
- Janakiram MSV, “Microsoft Launches NLWeb to Simplify Website-Agent Interactions”, Forbes, 2025.05.21., https://www.forbes.com/sites/janakirammsv/2025/05/21/microsoft-launches-nlweb-to-simplify-website-agent-interactions/

기술용어

순번	용어	설명
1	자연어 인터페이스 (Natural Language Interface)	인간이 일상적으로 쓰는 말(자연어)로 컴퓨터·서비스와 상호작용하는 방식

AI 확산 속 학술 출판 산업의 신뢰성 위기와 탐지·우회 기술의 기술 경쟁

뉴스 브리프

AI 기술이 학술 연구 과정에 활용되면서 연구 효율성을 높이는 동시에 존재하지 않는 인용을 제시하는 '환각' 현상이 학계의 새로운 위협으로 나타나고 있다. 신경정보처리시스템학회(NeurIPS) 논문 51편에서 발견된 100개의 허위 인용과 법조계의 800건 이상 허위 인용 사례는 연구와 판결의 기초가 되는 문헌 신뢰 훼손 현실을 보여준다. AI 산출물에 담긴 허위 정보가 학문적 신뢰를 훼손하자 이를 검증하려는 탐지 기술과 이를 무력화하려는 우회 기술 간의 경쟁이 가속화되고 있다. 본 보고서는 환각 현상의 정의와 탐지·우회 기술의 작동 원리를 분석하고, 탐지 기술의 긍정 오류 문제 및 저작자 권리 인정 범위와 오류 책임 소재 기준이 모호해지는 딜레마를 다루며, AI 시대 창작 윤리와 저작권 거버넌스 구축을 위한 제도적·교육적 접근의 필요성을 제시한다.

뉴스 플러스

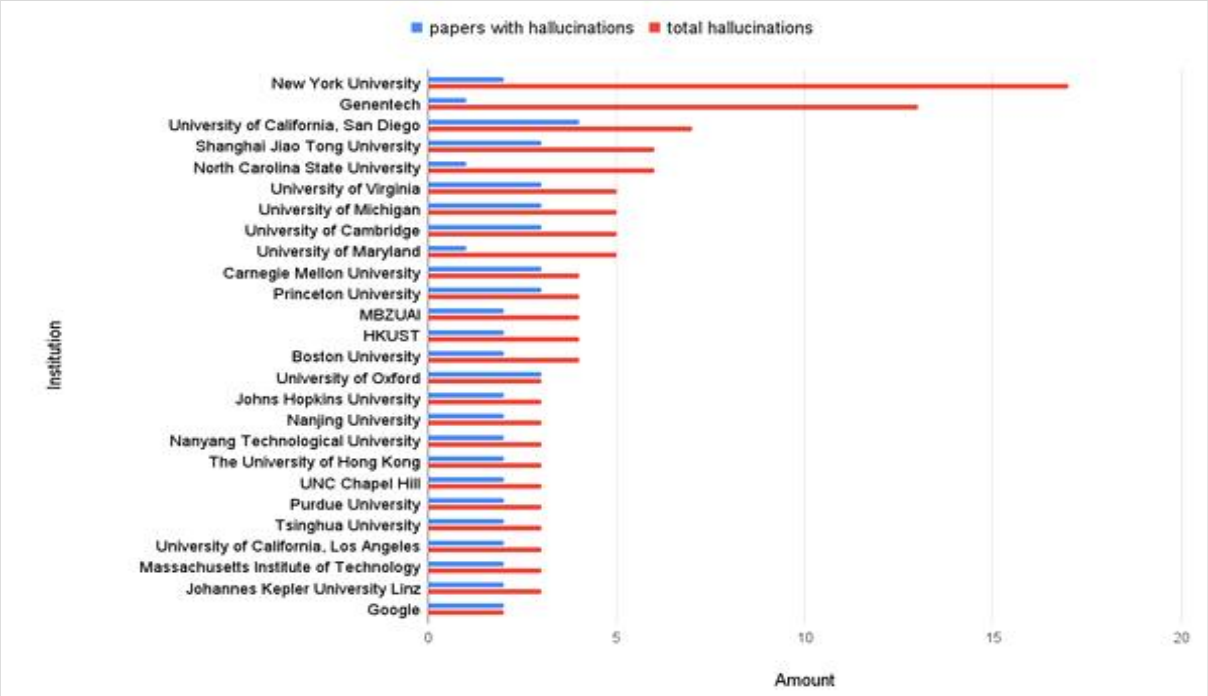
Ⅰ. 서론: AI가 촉발한 학술 출판의 패러다임 변화

- **AI가 만든 허위 인용과 학술계의 신뢰도 저하**

AI 기술은 학술 연구의 효율성을 높이는 도구로 주목받고 있지만, 동시에 학술 정보의 신뢰성을 위협하는 새로운 문제를 야기하고 있다. 최근 저명한 인공지능 학회인 신경정보처리시스템학회(NeurIPS)의 논문 51편에서 AI가 만들어낸 100개의 허위 인용이 발견된 사례[5]는 이러한 문제의 심각성을 단적으로 보여준다. 이는 연구 결과의 근거가 되는 인용 정보의 정확성이 훼손될 수 있음을 보여주며, 학술 소통의 기반에 영향을 주는 요인으로 작용한다.

5) Nazar Shmatko외 2인, 'GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers', GPTZero, 2026.01.21., https://gptzero.me/news/neurips/

[그림] 주요 연구기관별 논문 환각 발생 현황



출처: Nazar Shmatko외 2인, “GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers”, GPTZero, 2026.01.21., https://gptzero.me/news/neurips/

이러한 현상은 비단 학술계에만 국한되지 않으며, 고도의 전문성이 요구되는 법조계에서도 800건이 넘는 허위 인용이 법원 서류에서 발견되어 관련 법조인들이 징계를 받는 사례로 이어졌다. 연구와 판결의 기초가 되는 문헌의 신뢰가 무너지는 상황은 해당 분야의 전문성과 사회적 권위를 약화시킬 뿐만 아니라, 정보의 진위를 판별하기 위한 추가적인 노력과 기술적 대응을 요구하고 있다.

- **탐지 기술의 고도화와 우회 기술의 발전, 기술 경쟁의 심화**

AI 탐지 기술이 고도화되자, 이를 무력화하기 위한 우회 도구 역시 빠르게 발전하며 끝없는 ‘기술 경쟁’이 시작되었다. 한쪽에서는 문장의 통계적 패턴을 분석해 AI의 흔적을 찾으려 하고, 다른 한쪽에서는 문장 구조를 재편하고 인간적인 불완전성을 의도적으로 추가하여 탐지를 회피한다. 이처럼 양측의 기술이 끊임없이 서로를 추격하며 발전하기 때문에, 어느 한쪽의 기술적 우위는 일시적일 뿐 근본적인 해결책이 되지 못하고 있다.

또한, 이러한 기술 경쟁 과정에서 발생하는 부작용 역시 중요한 쟁점으로 부상하고 있다. AI 탐지 도구의 통계적 패턴 분석 방식은, 인간의 글을 AI 산출물로 오인하는 ‘오탐(False Positive)’을 야기하는 한계로 지적된다. 결국 현재의 접근법은 기술적 한계와 윤리적 딜레마를 동시에 드러내며, 이는 우리가 ‘탐지와 제재’라는 프레임을 넘어 AI 시대의 저작물과 책임 소재를 어떻게 바라볼 것인지에 대한 근본적인 질문을 던지고 있다.

Ⅱ. 본론: AI 산출물 검증 기술의 메커니즘과 저작권 쟁점

• 환각 현상의 정의와 AI 산출물 신뢰도 문제

환각 현상은 AI 모델이 학습 데이터에 존재하지 않거나 사실과 다른 정보를 실재하는 것처럼 구성해내는 오류를 의미한다. 이는 AI가 정보의 사실 여부를 직접 검증하는 것이 아니라, 학습된 데이터를 기반으로 문맥상 가장 그럴듯한 단어 조합을 추론하는 방식으로 작동하기 때문에 발생하는 것으로 분석된다. 예를 들어, 특정 주제에 대한 문헌 연구를 요청받았을 때, 실제로는 존재하지 않는 연구자의 이름이나 출판되지 않은 논문을 인용하며 매우 자연스러운 문장을 만들어내는 것이 환각 현상의 대표적인 사례다.

이러한 AI 산출물은 겉보기에는 논리적이고 유창해 보이지만, 그 근거가 허위이므로 학술적 논증의 기초를 무너뜨린다. 이는 연구 결과의 신뢰성을 직접적으로 훼손하는 것을 넘어, 해당 연구를 인용하는 다른 후속 연구로까지 오류를 연쇄적으로 확산시키는 심각한 문제로 이어진다. 결국 환각 현상은 연구 부정행위의 의도가 없는 연구자조차 자신도 모르는 사이에 학술 생태계의 신뢰를 훼손하는 결과를 낳을 수 있다는 점에서 그 위험성이 크다.

• AI 산출물 허위 정보 탐지 기술의 작동 원리

AI 산출물에 포함된 허위 정보를 탐지하는 기술은 주로 자동화된 사실 검증 및 출처 확인 방식으로 작동한다. 이는 AI가 만들어낸 텍스트의 주장이 신뢰할 수 있는 외부 정보원과 일치하는지를 대조하는 원리다. 마치 우리가 특정 정보의 사실 여부를 확인하기 위해 백과사전이나 뉴스 기사를 찾아보는 과정을 자동화한 것과 유사하다.

예시로 GPT제로(GPTZero)의 탐지 기술 작동 과정은 먼저 AI 산출물에서 검증이 필요한 핵심 요소를 식별하는 것에서 시작한다. 학술 논문의 경우, 이는 저자명, 논문 제목, 학술지 이름, 출판 연도와 같은 인용 정보가 핵심 대상이 된다. 다음 단계로, 식별된 인용 정보를 공신력 있는 학술 데이터베이스의 기록과 실시간으로 대조한다. 만약 AI가 제시한 문헌이 해당 데이터베이스에 존재하지 않거나 정보가 일치하지 않으면, 이를 '허위 정보' 또는 '환각'으로 판정하여 사용자에게 경고한다.

이러한 탐지 기술은 학술 출판 과정의 신뢰도를 높이는 데 중요한 역할을 한다. 출판사나 학회는 이 기술을 활용하여 투고된 논문의 인용 목록이 모두 실존하는 문헌인지 자동으로 검증함으로써, 동료 심사 과정의 부담을 줄이고 심사의 정확성을 높일 수 있다. 특히, GPT제로와 같은 AI 탐지 기업은 자사의 허위 인용 탐지 도구가 99% 이상의 정확도를 보인다고 주장하며, 학술 논문 데이터베이스와 실시간으로 연동하여 검증 속도와 신뢰도를 확보하고 있다. 이 기술의 도입은 연구 부정행위를 예방하고, 학술 커뮤니케이션의 투명성을 강화하는 기술적 보호 장치로 기능할 수 있다.

현재의 AI 허위 정보 탐지 기술은 명백한 사실 오류나 존재하지 않는 출처를 확인하는 데 효과적이지만, 근본적인 한계를 지니고 있다. 이 기술들은 주로 외부 데이터베이스와의 비교를 통해 정보의 '존재 유무'를 판별하는 방식에 의존하기 때문에, 실존하는 문헌을 인용하면서 내용을 교묘하게 왜곡하거나 연구의 핵심 맥락과 무관하게 인용하는 등의 지능적인 오류는 식별하기 어렵다. 또한, AI 모델이 점점 더 정교해지면서 인간이 작성한 글과 구별하기 어려운 수준의 산출물을 내놓고 있어, 텍스트의 스타일이나 패턴만으로 AI 사용 여부를 판단하는 데에도 어려움이 따른다.

• AI 산출물 우회 기술의 작동 방식

AI 허위 정보 탐지 기술의 등장은 곧바로 이를 무력화하려는 기술의 발전을 촉발하며 기술 경쟁을 심화시켰다. '인간화 도구(Humanizer)' 또는 'AI 우회 도구'로 불리는 이 기술들은 AI가 산출한 텍스트 고유의 통계적 패턴을 변형하여, 탐지 시스템이 이를 인간이 작성한 글처럼 인식하도록 만드는, 즉 '미탐(未探, False Negative)'을 의도적으로 유발하는 것을 목표로 한다. 이는 탐지 기술이 주로 문장의 길이, 단어 선택의 규칙성 등 AI 텍스트의 예측 가능한 특징들을 기반으로 작동한다는 점을 역이용한 것이다.

초기의 우회 기술은 단순히 동의어를 바꾸거나 문장 순서를 뒤섞는 수준에 머물렀지만, 최근에는 훨씬 더 고도화된 방식으로 작동한다. 정교한 우회 기술은 AI가 만든 텍스트의 핵심 의미는 유지하면서도, 문장 구조 자체를 근본적으로 재구성하고 인간의 글쓰기에서 나타나는 비정형성과 불규칙성을 의도적으로 추가한다. 예를 들어, 복잡한 문장을 두 개의 단순한 문장으로 나누거나, 접속사의 사용 패턴을 바꾸는 등의 방식을 사용한다. 이러한 과정은 AI 산출물과 인간 저작물의 경계를 의도적으로 허물어뜨려, 기술적 수단만으로는 둘을 구별하는 것을 거의 불가능하게 만든다.

• 기술 경쟁의 부작용과 저작권의 딜레마

이러한 '탐지'와 '회피'의 기술 경쟁은 양쪽 모두에게 심각한 오류를 야기한다. 한편으로 AI 우회 기술은 탐지 시스템이 AI 산출물을 놓치는 '미탐'의 가능성을 높이지만, 다른 한편으로는 탐지 기술이 정상 이용자를 오인 식별하는 '오탐(誤探, False Positive)'이라는 문제도 발생한다. 오탐은 AI를 사용하지 않은 순수한 인간의 창작물을 AI 탐지 시스템이 AI 산출물로 잘못 판단하는 경우를 의미한다. 이는 탐지 기술이 텍스트의 통계적 특성에 의존하는 구조적 한계에서 비롯되며, 오히려 명료하고 문법적으로 완벽한 글을 쓰는 사람의 글이 AI 산출물로 오인될 가능성을 높인다.

이러한 기술적 한계는 또한 저작권의 근본적인 딜레마를 야기한다. 아이디어 구체화나 문법 교정 등 AI를 유용한 '보조 도구'로 활용하는 사례와, 인간의 고유한 창작 과정을 사실상 '대체'하는 수준의 활용을 현재의 기술로는 명확히 구분하기 어렵기 때문이다. 결국 AI와 인간의 기여도가 뒤섞인 결과물에 대해 어디까지를 저작자의 권리로 인정해야 하는지, 결과물에 포함된 오류의 책임은 누구에게 있는지에 대한 기준이 모호해지는 결과를 낳고 있다.

Ⅲ. 결론 : 새로운 학술 윤리와 기술 거버넌스의 모색

• AI 시대의 저작권과 신뢰 회복을 위한 제언

결국 '탐지'와 '회피'의 기술 경쟁 심화와 그로 인한 부작용은, AI가 야기한 신뢰의 위기를 기술만으로 해결할 수 없다는 점을 명확히 보여준다. 이는 AI라는 새로운 도구의 등장으로 저작물 생성 방식이 변화하고 있음을 의미한다. 따라서 기존의 '탐지하고 제재하는' 방식에서 벗어나, AI 시대에 맞는 새로운 창작 윤리를 구축하는 방향으로 사회적 논의를 전환해야 할 필요가 있다.

이러한 점에서 향후 학술계는 기술적 대응과 함께 제도적, 교육적 방안을 병행하는 종합적인 접근을 모색해야 한다. 단기적으로는 연구자들이 AI 활용 여부와 그 방식을 투명하게 공개하도록 의무화하는 가이드라인을 정립하여 책임 소재를 명확히 해야 한다. 장기적으로는 AI를 비판적으로 활용하고 결과물을 검증하는 능력을 핵심적인 연구 역량으로 인정하고, 이를 함양하기 위한 체계적인 교육 프로그램을 마련하는 것이 중요하다. 이를 통해 기술의 발전을 인간의 창의성과 조화시키고, 저작권 제도가 보호하고자 하는 본질적 가치를 지켜나가는 방향으로 나아가야 할 것이다.

참고문헌

• Nazar Shmatko외 2인, 'GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers', GPT Zero, 2026.01.21., https://gptzero.me/news/neurips/
• Sharon Goldman, 'NeurIPS, one of the world's top academic AI conferences, accepted research papers with 100+ AI-hallucinated citations, new report claims', Fortune, 2026.01.21., https://fortune.com/2026/01/21/neurips-ai-conferences-research-papers-hallucinations/
• Thomas Claburn, 'AI conference's papers contaminated by AI hallucinations', The Register, 2026.01.22., https://www.theregister.com/2026/01/22/neurips_papers_contaiminated_ai_hallucinations/#:~:text=Correlation%20is%20not%20causation%2C%20but,It%20may%20invalidate%20their%20work
• Yusuke Sakai 외 2인, 'HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences', arxiv, 2026.01.26., https://www.arxiv.org/abs/2601.18724
• Ashely Segal, "The AI Detection Arms Race: Why Current Approaches to AI Writing Are Failing Everyone", Medium, 2025.12.17., https://medium.com/writewithai/ai-detection-arms-race-current-approaches-failing-2152568255a5