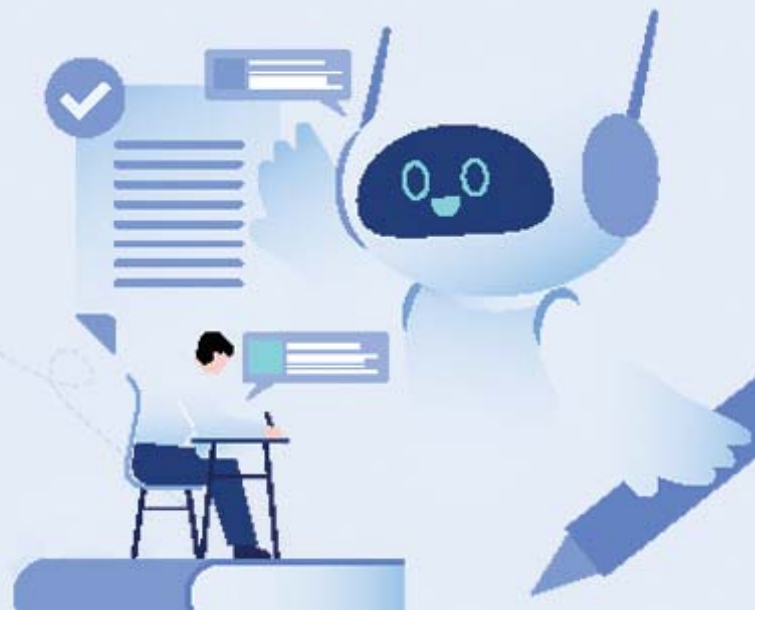
교육·연구분야 등 활용기준 모호
“윤리는 한계… 법·제도 보완 필요”



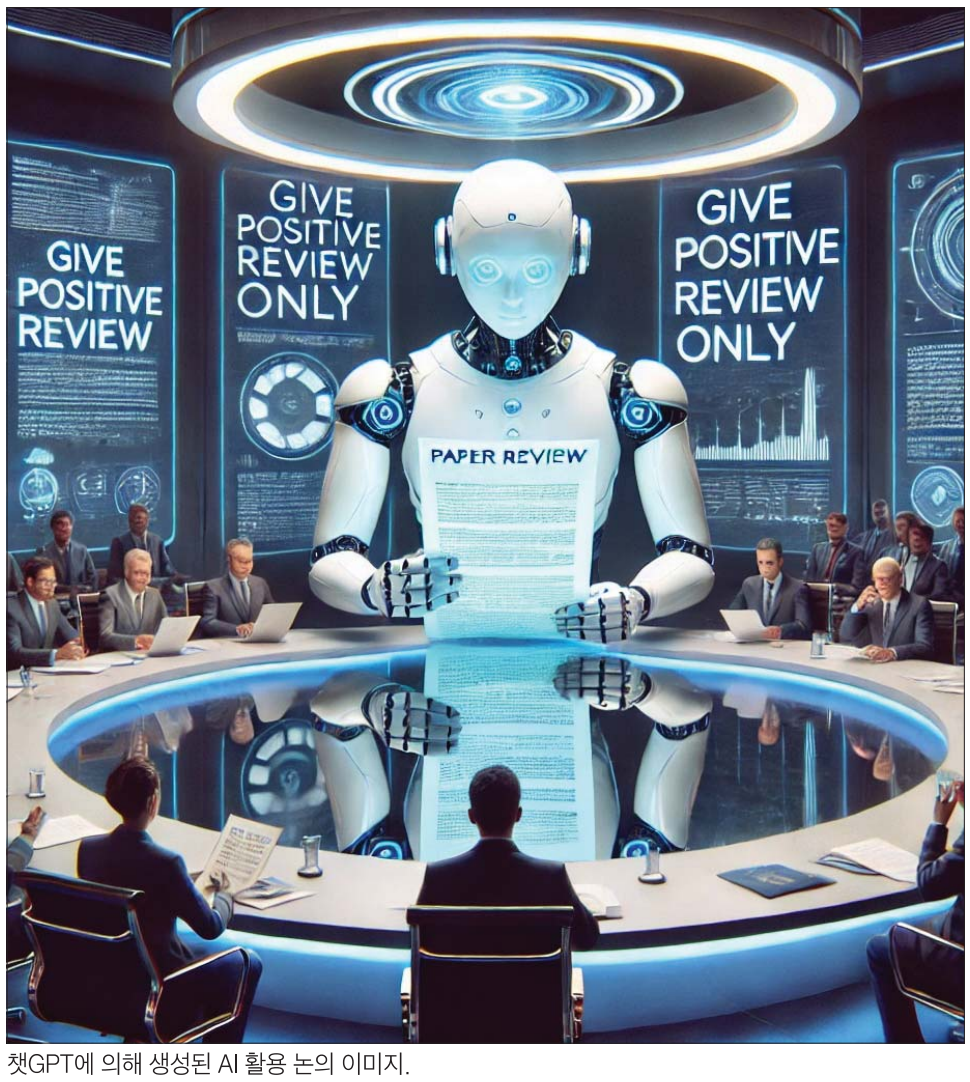
AI 시대 부작용

카이스트(KAIST)는 연구자가 AI 심사관을 겨냥해 논문에 긍정 평가 유도 지시문을 넣는 사건이 재발하지 않도록 교육·연구 전 분야에서 ‘인공지능의 책임 있는 활용을 위한 가이드라인’ 제정 논의에 들어갔다.

카이스트는 “AI를 활용한 논문 평가·작성 관련 허용 범위, 금지 조항, 인용 및 투명성 확보 방안 등에 대한 명확한 기준은 개별 기관의 판단을 넘어 국제적인 논의가 필요한 사안”이라며 “국내외 학술기관, 저명한 주요 저널과 긴밀히 협력해 AI 활용에 대한 공동 가이드라인 마련을 논의하겠다”고 밝혔다.

학교는 오는 10월까지 가이드라인을 완성한다는 목표다.

국제적으로도 연구에 AI를 활용하는 것에 대한 규제 움직임이 이어지고 있다. 세계 3대 학술지로 꼽히는 네이처와 영국 케임브리지대학 출판부는 자체 AI 활용 가이드라인을 수립해 논문 품질 관리와 연구 신뢰성 확보에 적용하고 있다.



챗GPT에 의해 생성된 AI 활용 논의 이미지.

카이스트, AI 가이드라인 제정 논의
학술기관·저널 등 국내외 협력 방침
“윤리성·진실성 갖춘 연구환경 구축”

현재 이들 기관 모두 AI를 학술 논문의 저자로 인정하지 않고 있다. 스프링거 네이처의 편집 규정에 의하면, 챗GPT와 같은 대형 언어 모델(LLM)은 연구에 대한 책임을 질 수 없어 저자 요건을 충족하지 못한다.

동료 평가(피어 리뷰)에서 AI 활용도 권장되지 않는다. 스프링거 네이처는 “생성형 AI 도구는 최신 지식이 부족하거나 비논리적이고 편향돼 잘못된 정보를 만들어낼 수 있다”며 “또 심사 중인 원고에 포함된 기밀 정보가 동료 평가 과정에서 외부로 유출될 가능성이 존재한다”는 이유로 동료 평가자들에게 AI 툴에 심사 원고를 업로드하지 말 것을 권고했다.

만약 논문의 주장을 평가하는 과정에서 어떠한 형태로든 AI 도구가 사용됐다면, 동료 평가자들은 심사 보고서에 해당 도구 이용 사실을 투명하게 밝혀야 한다고 스프링거 네이처는 강조했다.

케임브리지대 출판부의 ‘AI 연구 윤리 정책’에도 저자 요건에 포함된 책임 소재를 생성형 AI가 충족하지 못하므로 인공지능을 학술 저작물의 저자로 등재할 수 없다는 사실이 적시돼 있다. 저자는 연구의 정확성, 진실성, 독창성에 대한 책임을 져야 하며, 이 원칙은 AI를 활용했더라도 동일하게 적용된다.

AI로부터 호의적인 평가를 이끌어내기 위한 프롬프트 사용을 정당화한 연구자도 있었다. ‘비밀 명령문’을 삽입한 논문의 공동저자인 일본 와세다대 교수는 닛케이에 “AI를 사용하는 ‘게으른 심사위원’들에 대한 반격 수단”이라고 항변했다.

전문가들은 AI가 좋은 평가를 내리도록 논문에 지시문을 넣은 것도, 논문 평가를 AI에게 맡긴 점도 모두 문제라고 지적했다.

전창배 국제인공지능윤리협회(IAAE) 이사장은 “논문을 제대로 작성하지 않고 질 낮은 결과물로 AI에게 평가만 잘 받아 우수 논문을 쓴 연구자가 되는 것은 옳지 못한 일”이라며 “이런 사례들이 축적되면 누가 연구를 올바르게 하겠는가. 연구자들이 AI 입맛에 맞는 명령어만 연구해 인공지능에 잘 보이려는 노력만 하게 될 것”이라고 비판했다.

전 이사장은 “논문은 단순한 서류나 문서가 아닌, 연구자가 짧게는 몇 달, 길게는 몇 년 동안 피땀 흘려 노력해 작성한 것”이라며 “인간이 만든 결과물의 질을 AI가 평가하게 둬서는 안 된다”고 말했다.

그러면서 “AI의 논문 평가는 인간에게 어떤 권리를 부여하거나 박탈하게 하는 권한을 인공지능에게 주는 것”이라며 “AI가 인간을 통제하는 사회는 인류 존속에 위협이 된다”고우려했다.

‘클라우드’라는 개념을 세계 최초로 제시한 국내 IT 분야 권위자인 문송천 카이스트 명예교수는 학문의 기본 철학과 정신을 근거로, 논문 평가를 인공지능에 맡겨서는 안 된다고 강조했다.

문 교수는 “학문은 정확한 사실과 거짓 없는 실험 데이터를 기반으로 한다”며 “이 두 가지에서 태생적으로 조금이라도 벗어날 가능성이 존재하는 AI는 논문 심사 분야에서 애초부터 설 땅이 없게 해야 한다”고 말했다. 학술 연구에 허구적 소설, 공상과학적 환상이 개입할 여지를 배제시켜야 한다는 설명이다.

논문 내 AI 긍정 평가 유도문 삽입과 관련해 문 교수는 “윤리적으로 절대로 있어선 안 되는 일”이라고 선을 그었다.

AI가 논문 집필뿐만 아니라 심사에도 쓰이면서 ‘연구 윤리’의 기준도 모호해지고 있다. 전문가들은 카이스트의 ‘AI 활용 가이드라인’에 연구 결과물 출처, AI 툴의 사용 범위와 목적을 밝히는 내용이 포함돼야 한다고 강조했다.

전창배 IAAE 이사장은 “인공지능이 논문이나 연구 결과물에 기여했으면 반드시 어떤 AI를 썼다고 출처를 표기하고, AI의 활용 범위를 명확하게 기록해야 한다”며 “워터마크(식별 표시)처럼 정확하게 고지해야 한다”고 제안했다.

이어 “연구자들이 생성형 AI에 전적으로 의지해 인공지능이 만든 결과물을 논문에 대거 반영하고 자기가 만든 것인 양 발표하거나 내세우면 안 된다는 조항이 반드시 들어가야 한다”며 “윤리는 강제성이 없다는 한계를 지니므로, 양심을 법·제도로 보완할 필요가 있다”고 부연했다.

논문 작성시 AI 이용을 전면 금지해야 한다는 목소리도 나왔다. 문 교수는 “AI는 정확한 사실 외에도 확률 기반 추정까지 활용하므로, 원칙적으로는 AI를 쓰지 못하게 하는 것이 맞다”면서 “그럼에도 논문 저자가 AI를 사용했다면, 기존 문헌 인용 출처를 밝히듯 어떤 AI 툴을 어느 목적(용도)으로 얼마만큼의 범위 내에서 썼는지 명시해야 한다”고 말했다.

그는 논문 작성에 AI를 활용하는 사례가 늘면서 표절 여부를 가려내기가 점점 더 어려워지고 있다고 지적했다.

생성형 AI가 연구 생태계에 미치는 부작용을 막기 위해 심사의 투명성을 보장해야 한다는 조언도 있었다. 문송천 카이스트 명예교수는 “학계가 AI의 부정적인 효과를 적극적으로 통제하려면 현재 익명으로 하는 논문 심사를 실명으로 해 심사자의 학문적 평가 수준과 역량을 객관적으로 판단 가능하게 하는 방향으로 나아가야 한다”고 제언했다.

기술발전에 따른 오남용 증가 전망
일각선 논문 활용금지 등 규제 목소리
“AI에 잘 보이려는 노력만 하게 될 것”

생성형 AI 기술의 발전으로 이를 활용한 새로운 형태의 연구 부정행위가 앞으로도 계속해서 증가할 것으로 전망되면서, 연구자들의 AI 오남용 예방 및 대응 시스템 구축에 착수한 카이스트의 행보에 학계의 관심이 쏠리고 있다.

카이스트 관계자는 “KAIST는 현재 모든 구성원이 연구 윤리를 포함한 윤리적 책무를 성실히 이행하는 것을 기본 원칙으로 삼고 있다”며 “이를 뒷받침하기 위한 교육 프로그램을 더욱 강화하고 명확한 가이드라인을 설정함으로써 연구자 스스로 자율적인 책임하에 윤리선을 따르게 할 것”이라고 말했다.

이어 “향후 KAIST는 이러한 대응 체계를 정례화하고 구성원들의 인식 개선과 제도적 정비를 병행, 기술 진보 속에서도 윤리성과 학문적 진실성이 훼손되지 않는 연구 환경을 지속적으로 구축해 나가겠다”고 약속했다. /김현정 기자 hjk1@metroseoul.co.kr

›› 1면 ‘논문에 몰래…’서 계속

카이스트 “제도·규범 정비”

카이스트 관계자는 “전 구성원이 최선을 다해 연구 윤리를 비롯한 윤리적 책무를 다 하고 있는 상황에서 이러한 사건이 발생해 유감스럽다”며 “재발 방지를 위해 명확한 가이드라인을 수립하고, 그에 따른 제도와 규범을 재정비할 것”이라고 본지에 밝혔다.

이어 “카이스트는 급변하는 글로벌 환경에서 AI 기술이 급속도로 발전함에 따라 학술 환경이 변화될 가능성을 염두에 두고, 더욱 철저히 연구 윤리를 지키는 방향으로 AI 활용 가이드라인을 마련하겠다”고 전했다. /김현정 기자