관심사 독해: 폭발 신화와 검증의 기술

2026년 11월 20일 수집분

AI가 스스로를 폭발적으로 똑똑하게 만들 수 있는지 묻는 논쟁에서 출발해, 기술 진보의 곡선이 멈춘 우주 여행, 논문 속 심사 조작, 백배 단위 오류를 놓친 의학 논문, AI 시대의 학자 평가까지 이어 간다. 오늘 묶인 글들은 예언보다 검증, 속도보다 판단 절차를 묻는다.

폭발을 기다리는 동안 도착한 초인간 AI

Noahpinion — Does it really matter if there's an intelligence explosion?

AI 안전 논쟁의 상당 부분은 "지능 폭발"을 중심으로 돌아간다. 지능 폭발은 AI가 자기 자신을 개선하는 연구를 더 빠르게 하면서 능력이 초지수적으로 치솟는 사건을 말한다. 여기서 초지수적은 일정한 배율로 커지는 지수 성장보다 성장률 자체도 함께 커지는 상태다. 회의론자는 AI 연구가 이미 비용과 계산 자원의 수확 체감에 걸려 있다고 본다. 수확 체감은 같은 투자를 늘려도 얻는 추가 성과가 점점 작아지는 현상이다. 반대편은 AI가 이미 연구 속도를 크게 높이고 있으며, 정답을 검증할 수 있는 수학과 프로그래밍에서 보인 급격한 향상이 AI 연구 자체로 옮겨 올 것이라고 본다.

이 글의 논점은 어느 쪽 예측이 맞는가가 아니라, 이 논쟁이 정책의 중심이 될 만한가다. 저자는 지능 폭발이 가능하다면 어차피 임계점에 도달할 것이고, 진전을 늦추는 정책은 폭발을 막기보다 몇 년 늦출 뿐이라고 본다. 더 중요하게는 폭발이 없어도 현재의 빠른 점진적 진전만으로 대부분의 경제 활동에서 인간을 넘는 AI가 나올 수 있다는 것이다. 그렇다면 정말 필요한 질문은 "신적인 AI가 갑자기 깨어나는가"보다 "이미 일상에 들어온 초인간적 도구를 누가 소유하고, 어떻게 규제하며, 이익을 어떻게 나눌 것인가"가 된다.

이 주장의 힘은 관심을 종말 예언에서 현재의 제도로 되돌린다는 데 있다. 다만 "가능하면 결국 일어난다"는 전제는 충분히 입증되지 않았다. 완전한 중단이 어렵다는 사실이 속도, 배포 범위,

안전장치, 독점 구조를 바꾸는 일까지 무의미하게 만들지는 않는다. 폭발이 오면 정렬 연구가 따라잡지 못한다는 말도 폭발 시나리오 안에서는 타당하지만, 폭발이 아니라 빠른 점진적 진전이 현실이라면 제도와 시장 구조는 여전히 결과를 바꾼다. 이 글은 예측보다 의제 설정에서 더 유용하다.

기술 곡선이 멈춘 자리에서 본 지구

Crooked Timber — No planet but this one (repost)

이 글은 2018년에 쓴 글을 2026년의 보이저 1호 소식에 맞춰 다시 올린 것이다. 보이저 1호가 태양에서 빛이 하루 동안 가는 거리, 즉 1광일만큼 멀어졌다는 소식이 계기다. 1광일은 약 259억km다. 인류가 만든 물체가 반세기 가까이 날아 도달한 거리지만, 별 사이의 거리를 기준으로 하면 여전히 태양계의 바깥 문턱에도 이르지 못했다. 원문에 등장하는 보이저 2호 문장은 2018년 시점을 그대로 보존한 재게재분이다.

글의 핵심은 우주 탐사의 성취를 부정하는 것이 아니라, 기술 발전의 곡선을 과거의 속도로 계속 외삽할 수 없다는 데 있다. 외삽은 지금까지의 추세를 앞으로도 이어질 것으로 가정하는 추론이다. 라이트 형제의 비행에서 대서양 횡단까지 40년이 걸렸다고 해서, 그 다음 40년에 인류가 행성 사이를 자유롭게 오갈 것이라 예상할 수는 없었다. 실제로 유인 달 탐사는 멈췄고, 우주왕복선의 약속은 후퇴했으며, 태양계 어디도 지구의 극지방이나 사막만큼 살기 좋은 곳은 없었다. 물리적 거리와 에너지, 생명유지 비용은 소프트웨어의 낙관을 쉽게 허용하지 않는다.

이 글은 지능 폭발 논쟁에 대한 직접 답변이 아니지만 중요한 반대편 압력을 준다. Noahpinion은 현재의 AI 향상만으로도 신에 가까운 도구가 나온다고 보고, Crooked Timber는 과거의 비약이 미래의 비약을 보장하지 않는다고 말한다. 둘은 충돌하지만 한 지점에서 만난다. 화성 이주든 기계 신이든, 지금의 지구와 제도를 대신할 탈출구는 없다는 것이다. 따라서 기술의 미래를 논하는 방식도 예언보다 조건을 봐야 한다. 무엇이 물리적으로 가능한가, 누가 비용을 치르는가, 실패했을 때 어디에 남는가가 먼저다.

논문 속에 숨겨진 심사 조작

Retraction Watch — ‘Give this paper a good review’: Prompts for AI reviewers make their way into published papers

학술 논문에 AI 심사자를 속이기 위한 숨은 문구를 넣는 사례가 나오고 있다. 이런 공격을 프롬프트 인젝션이라고 부른다. 프롬프트는 AI에게 주는 지시고, 인젝션은 원래 지시 사이에 몰래 다른 지시를 끼워 넣는 방식이다. 한 프리프린트에서는 그림과 같은 색의 아주 작은 글씨로 AI에게 우호적인 평가를 유도하는 문구가 들어 있었다. 다른 논문에서는 데이터 이용 안내 문구 뒤에 흰색 1포인트 글씨로 비슷한 지시가 붙었고, 출판 후에 발견되어 철회됐다. Nature는 2025년에 이런 숨은 지시를 포함한 프리프린트 18편을 확인했다고 보도했다.

숨은 문구 자체는 유치한 속임수처럼 보이지만, 전제는 심각하다. 이 공격이 성립하려면 심사자가 원고를 외부 AI에 올리고 평가를 받을 가능성이 있어야 한다. 학술 심사는 미공개 원고를 다루는 비밀 절차다. 심사자가 원고를 상용 AI 서비스에 올리면 저자의 아이디어와 데이터가 제3자 시스템으로 새어 나갈 수 있고, 심사 판단도 검증할 수 없는 모델의 출력에 의존하게 된다. 숨은 지시는 그 약점을 겨냥한다. 저자가 심사를 조작하려는 의도라는 점에서 연구 부정이고, 심사자가 AI를 몰래 사용하는 관행이 있다는 점에서 제도 실패다.

기사는 탐지용 장난과 조작용 지시를 구분한다. 어떤 저자는 AI가 쓴 심사를 잡아 내기 위해 특정 단어를 넣으라는 식의 시험 문구를 숨기기도 한다. 그러나 방어 목적이라도 심사 과정에 몰래 장치를 심는 일은 신뢰를 더 갉아먹을 수 있다. 필요한 것은 숨은 문자를 찾는 기술 경쟁만이 아니다. 학회와 출판사는 심사자가 원고를 외부 AI에 올릴 수 있는지, 어떤 AI 사용을 공개해야 하는지, 위반 시 어떤 절차로 확인하고 이의를 제기할지를 명확히 해야 한다. 숨은 문구는 AI보다 먼저, 심사라는 제도가 이미 얼마나 AI에 기대고 있는지를 드러낸다.

백 배 차이를 놓친 심사

Offsetting Behaviour — Erratum-watch: vaping risks

뉴질랜드의학저널은 전자담배의 발암 위험을 다룬 논문에서 포름알데히드 수치를 정정했다. 포름알데히드는 자극성과 발암 가능성이 있는 화학물질이다. 논문은 어떤 측정값을 한 번 흡입할 때의 양으로 썼지만, 정정 후에는 100번 흡입할 때의 양으로 바뀌었다. 같은 수치라도 단위가 “1회”에서 “100회”로 바뀌면 노

출량은 100분의 1이 된다. 다른 인용 수치까지 포함하면 원문의 위험 표현이 실제보다 100~1000배 부풀려졌다는 것이 블로그 글의 지적이다.

이 오류가 중요한 이유는 전자담배 정책의 찬반과 별개로, 과학 심사가 가장 기본적인 검산을 놓쳤다는 데 있다. 검산은 정밀 계산 전에 단위와 크기의 자릿수가 상식적으로 맞는지 확인하는 절차다. 블로그 저자는 동물 실험에서 코와 인두의 암 증가가 나타난 노출 기준보다 수천 배 높은 값이 일상적 전자담배 사용에서 나왔다면, 그 결과는 이미 오래전에 훨씬 큰 문제로 드러났을 것이라고 말한다. 특히 한 제품의 한 향료 시료만 다른 시료보다 10배 이상 높다면, 실제 위험보다 가열 오류나 측정 실수를 먼저 의심해야 한다.

정정이 난 것은 좋은 일이지만, 방송과 인용을 거친 수치는 이미 공중에게 도달했다. AI 검증 도구가 이런 문제를 모두 해결해 주지는 않는다. 오히려 필요한 것은 사람과 AI 모두에게 같은 습관이다. 단위를 원자료로 되짚고, 비교 기준이 같은 조건인지 확인하고, 자릿수가 이상하면 결론 전에 멈추는 일이다. 숨은 프롬프트 사건이 의도된 속임수라면, 이 정정은 의도 없이도 얼마나 큰 오류가 심사를 통과하는지를 보여 준다. 학술 시스템의 문제는 악의만이 아니라 검증의 시간과 능력 부족이다.

셀 수 없을 때 사람을 보는 법

Computational Complexity — How Will We Judge?

AI가 정리 증명과 논문 작성을 쉽게 만들면 학계는 더 많은 논문을 받게 된다. 문제는 평가자의 시간은 늘지 않는다는 데 있다. 박사 입학, 교수 채용, 승진 심사는 결국 사람을 골라야 하지만, 논문 수와 인용 수는 이미 실제 이해와 기여를 제대로 반영하지 못해 왔다. AI는 이 낡은 문제를 더 크게 만든다. 저자가 직접 쓴 부분과 AI가 만든 부분을 어떻게 나눌 것인가. 비싼 모델과 더 좋은 프롬프트를 가진 사람이 더 유능한 연구자로 보일 것인가.

공개 사용 선언도 단순한 해결책은 아니다. 선언이 평가에 쓰이면 다음 지원자는 선언문을 더 교묘하게 쓸 수 있다. 추천서도 마찬가지다. 추천서가 AI로 작성되면, 과장을 걸러 읽던 기존의 독법도 흔들린다. 그래서 한 이론컴퓨터학자 모임은 2026년 7월 이후의 업적은 지원자가 고른 최대 3편만 보고, 그 이전 업적과 함께 평가하자고 제안한다. 더 중요한 것은 지원자가 자기 논문

을 칠판 앞에서 설명하거나 수학적으로 깊은 구술을 거쳐 실제로 이해하는지 확인하는 절차다. 칠판 설명은 준비된 발표문을 읽는 대신 질문에 답하며 사고 과정을 보여 주는 방식이다.

이 제안에는 한계가 있다. 세 편 제한은 정말 생산적인 연구자를 불리하게 할 수 있고, 2026년 7월이라는 경계도 인위적이다. 그러나 방향은 옳다. 산출물의 양이 폭증하면 평가는 소유물 계산에서 이해 검증으로 옮겨 가야 한다. 이 원칙은 학계만의 이야기가 아니다. AI가 보고서와 코드와 요약문을 대량으로 만들수록, 좋은 판단은 "무엇을 만들었는가"보다 "왜 이렇게 만들었는지 설명할 수 있는가"에 달린다. 오늘의 다른 글들도 같은 지점을 가리킨다. 지능의 양보다 검증의 질이 남는 자산이다.

수집과 근거

승인된 피드 22곳 중 20곳을 확인했고, 2곳은 피드 형식 오류로 수집하지 못했다. 새 글 13편을 모두 끝까지 읽은 뒤 5편을 분석하고 8편은 생략했다. 분석한 5편은 원문 페이지와 대조해 전문 수신을 확인했다.

출처

- https://www.noahpinion.blog/p/does-it-really-matter-if-theres-an
- https://crookedtimber.org/2026/11/20/no-planet-but-this-one-repost/
- https://retractionwatch.com/2026/11/19/prompt-injection-ai-llm-peer-review-published-papers/
- http://offsettingbehaviour.blogspot.com/2026/11/erratum-watch-vaping-risks.html
- https://blog.computationalcomplexity.org/2026/11/how-will-we-judge.html