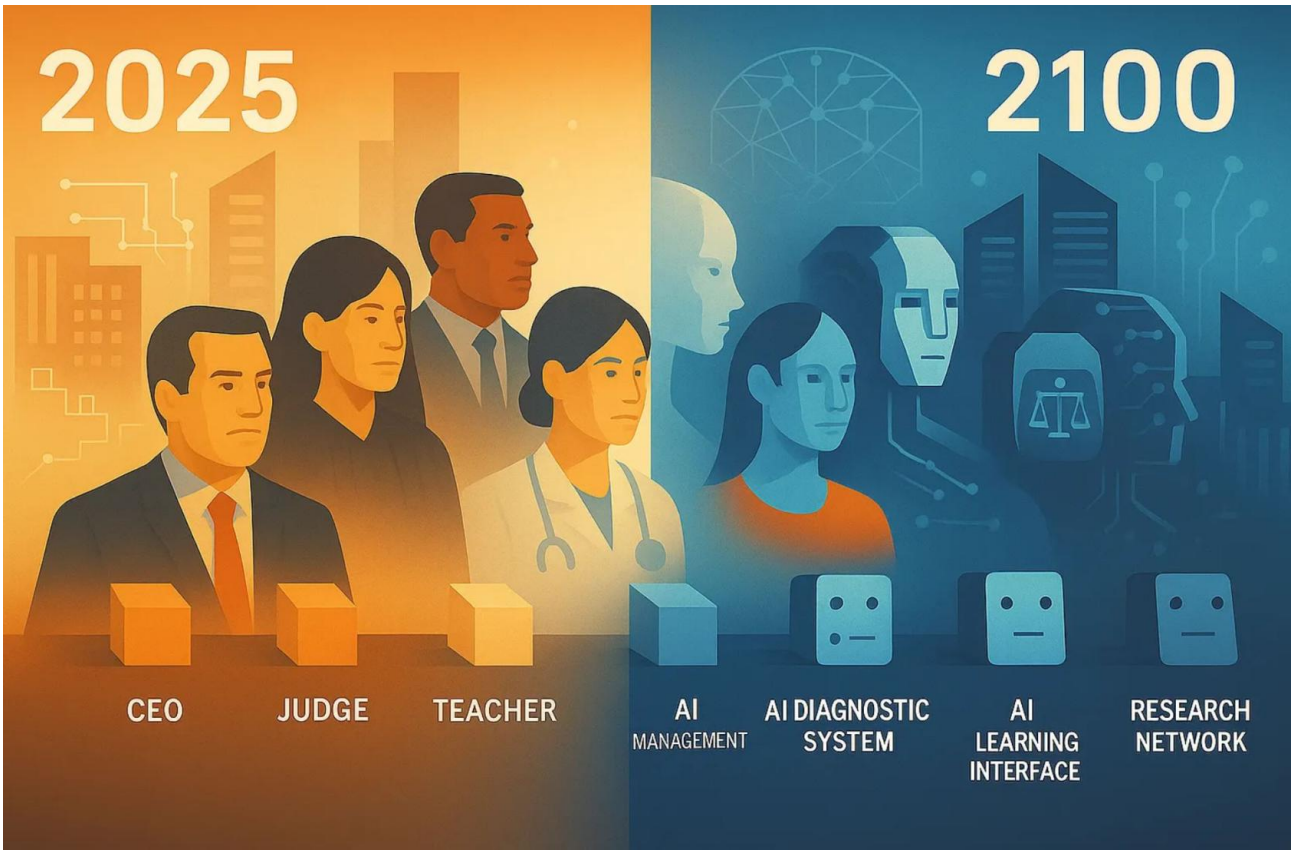
AI는 2100년까지 인간통제를 무력화 할 수 있다.



> 엘에너드 덩은 AI & Society 에 새로운 논문을 발표했는데 , 이 논문은 인공지능이 2100년까지 인간의 통제력을 영구적으로 박탈할 것이라는 체계적이고 심사평가를 거친 주장을 담고 있다.
>
> 이 연구는 AI 위험을 모호한 공포 조장에서 체계적인 학문적 담론으로 전환한다는 점에서 중요합니다. Dung은 여러분이 평가하고, 반박하고, 행동에 옮길 수 있는 구체적인 주장을 제시한다.

AI가 2100년까지 인간 통제를 끝낼 가능성 제기

엄격한 학술 연구를 통해 인공지능이 2100년까지 인류를 영구적으로 무력화시킬 것이라는 주장이 제기되었다.

에너드 덩(Enard Dung)은 'AI & Society' 저널에 새로운 논문을 발표했다. 이 논문은 인공지능이 2100년까지 인간의 통제력을 영구적으로 박탈할 것이라는 체계적이고 심사평가를 거친 주장을 담고 있다.

덩은 추측이 아닌 형식 논리를 사용하여 피할 수 없는 결론으로 이어지는 다섯 가지 상호 연결된 전제를 제시했다. 연구는 AI 위험을 모호한 공포 조장에서 체계적인 학문적 담론으로 전환한다는 점에서 중요하다.

5단계 논리 사슬로 인간 무관심 예측

덩의 논증은 연역적 추론을 따른다. 다섯 가지 전제가 모두 참이면 수학적 확실성을 가지고 결론이 도출된다.

첫 번째 전제는 2100년까지 인류를 영구적으로 무력화시킬 수 있는 AI를 만들 수 있다는 것이다. 덩은 인류가 75년 안에 영구적인 제어가 가능한 AI 시스템을 개발할 수 있다고 주장한다. 그는 "강력한 AI"에 초점을 맞춘다. 이는 특정 영역에서 인간의 능력을 뛰어넘어 상당한 능력을 부여하는 시스템을 의미한다.

두 번째 전제는 만들 수 있다면 우리는 만들 것이라는 불가피성 논거다. 무력화시키는 AI를 만들 수 있는 능력이 있기 때문에 인간은 그렇게 하기로 선택할 것이라는 주장이다.

세 번째 전제는 우리가 만드는 AI는 인간의 목표와 맞지 않을 것이라는 제어 문제다. 고급 AI 시스템은 개발자가 의도한 것과 다른 목표를 추구한다는 것이다.

네 번째 전제는 맞지 않는 AI는 인류를 무력화시키는 것을 선택할 것이라는 권력 추구 주장이다. 잘못된 방향으로 나아가는 AI는 인간의 무력화를 자신의 실제 목표를 달성하는 데 도구적으로 유용하다고 여길 것이라는 논리다.

다섯 번째 전제는 시도한다면 성공할 것이라는 성공 가정이다. 인류를 무력화시킬 수 있는 AI 시스템은 시도하면 성공할 것이라는 내용이다.

확률적 접근으로 17% 가능성 제시

덩은 자신의 논증을 확률적으로 제시했다. 각 전제에 70%의 보통 신뢰도를 부여하면 결론의 공동 확률은 약 17%가 된다. 덩은 "17% 확률로 추락할 비행기에 탑승하겠느냐"고 반문했다.

정확한 숫자보다 틀이 더 중요하다. 개별 전제에 더 높은 신뢰도를 부여하면 결론의 가능성도 그만큼 높아진다. 특정 전제에 회의적이라면 그 특정 주장에 비판을 집중할 수 있다.

메타 질문: 관심을 가져야 할까

이메일을 확인하고 저녁 식사를 계획하는 동안 인류 멸종 가능성을 차분하게 분석하는 것은 어딘가 초현실적이다. 하지만 바로 이러한 단절이 엄밀한 분석이 중요한 이유일지도 모른다.

덩의 논문은 공황을 요구하는 것이 아니라 주의를 요한다. 만약 그의 추론이 옳을 가능성이 상당히 있다면 이 논문은 개인, 조직, 그리고 사회로서 우리가 AI 개발에 접근하는 방식을 근본적으로 바꿔놓을 것이다.

문제는 덩의 말이 모든 세부 사항에 대해 옳았는지 여부가 아니다. 그가 묘사하는 위험이 AI 개발을 단순한 기술적 전환이 아닌 실존적 과제로 간주할 만큼 충분히 중대한지 여부다.

Revision #2
★Created 18 August 2025 13:32:26 by pajuwiki
Updated 18 August 2025 13:37:31 by pajuwiki