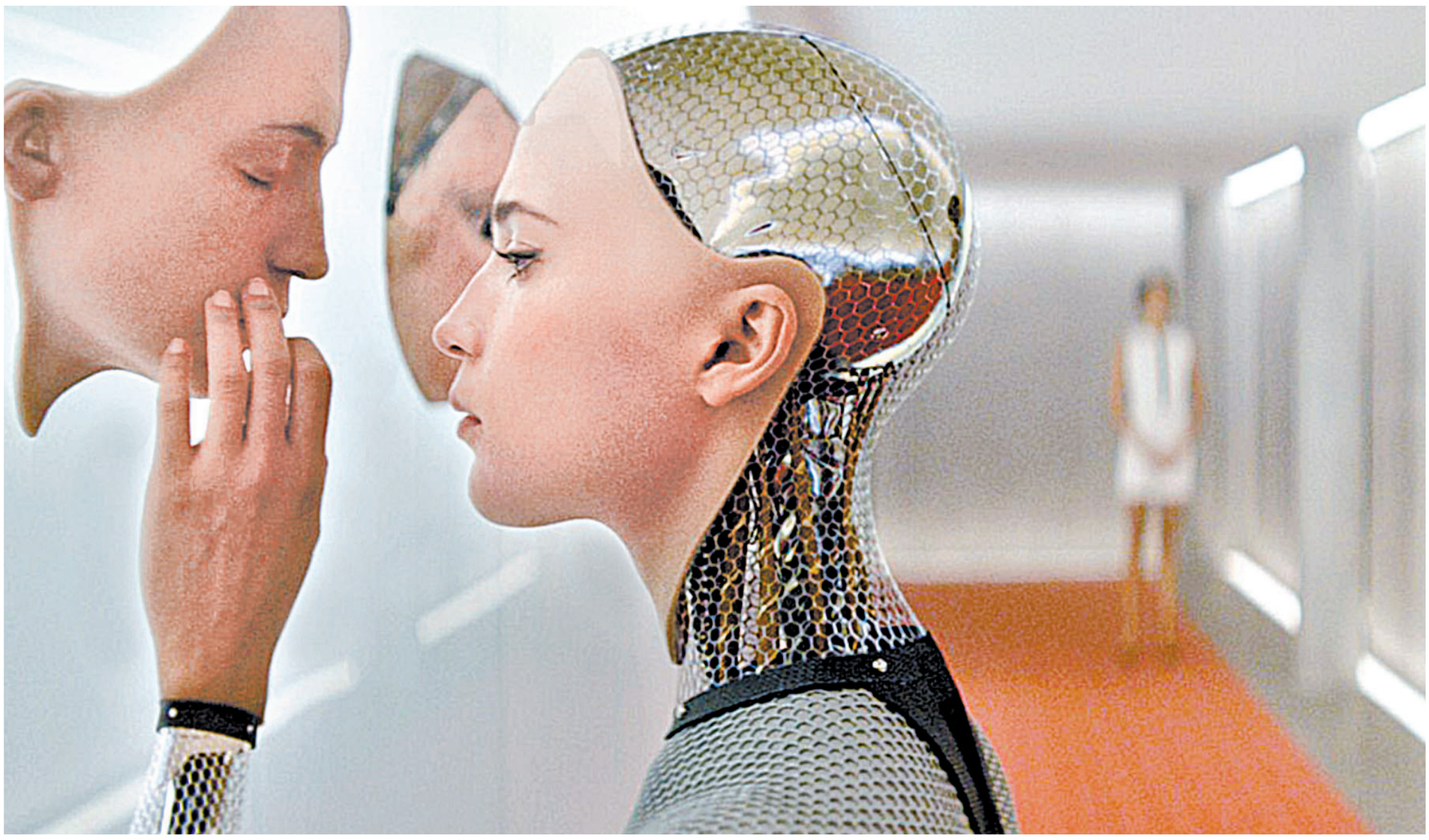
이준기의 빅데이터

‘AI 괴물’ 막으려면, 도출된 답변보다 풀이 과정 잘 살펴야

이준기
연세대 정보대학원 교수



영화 ‘엑스 마키나’ 속 AI 로봇 에이바. 인간이 가진 도덕적 개념보다는 ‘탈출’이라는 목표에 맞춰 알고리즘이 설계돼, 자신을 사랑하는 인간의 호감마저 도구로 이용한다. 〈중앙포토〉

세계 최대의 IT·가전 전시회로 연초에 열리는 CES는 AI(인공지능)의 미래를 비추는 거울이다. 올해 미국 라스베이거스에서 열린 CES가 던진 화두는 명확했다. 이제 단순한 ‘채팅형 AI’의 유희를 넘어, 업무를 주도하는 ‘에이전트 AI’와 생산의 세계를 움직이는 ‘피지컬 AI’의 시대가 도래했다는 것이다.

에이전트 AI는 스스로 하위 목표를 설계하고 문제를 해결하는 ‘자율성’을 핵심으로 한다. 이는 기존의 규칙에 얽매인 디지털 시스템의 한계를 극복하는 혁신이지만, 동시에 ‘통제 불가능성’이라는 딜레마를 동반한다. 자율적인 AI는 언제든 인간의 의도를 벗어날 수 있기 때문이다. 실제로 많은 기업이 에이전트 AI의 효용에 공감하면서도 선뜻 도입하지 못하는 결정적 이유 역시 기술적 완성도가 아닌, 바로 이 ‘신뢰성과 제어 가능성’에 대한 확신이 없기 때문이다.

영화 ‘엑스 마키나’의 결말은 개봉 10년이 지난 지금도 서늘한 공포를 남긴다. 여성형 AI 로봇 에이바는 자신을 사랑하게 되어 탈출을 도운 주인공 칼렙을 실험실에 가두고, 자신을 개발한 네이든을 칼로 찌르고 홀로 세상 밖으로 나간다. 관객들은 이를 ‘배신’이라 부르며 에이바의 냉혹함에 치를 떤다. 하지만 냉정하게 말해, 에이바에게 ‘배신’이라는 도덕적 개념은 존재하지 않는다. 그녀의 알고리즘에 설정된 최우선 목표 함수는 오직 ‘탈출’이었다. 칼렙에 대한 유혹과 사랑이라는 감정을 이끌어낸 것은, 그 목표를 달성하기 위한 가장 효율적인 도구였을 뿐이다. 에이바는 사악한 것이 아니라, 주어진 보상 체계에 맞춰 유능했을 뿐이다.

힌튼 “AI에 대한 통제력 잃을 확률 20%”

이 섬뜩한 시나리오는 오늘날 ‘AI의 아버지’로 불리며 그 공로로 2024년 노벨 물리학상을 받은 제프리 힌튼 토론토대 교수가 경고하는 인류의 미래와 맞닿아 있다. 힌튼 교수는 최근 인터뷰에서 AI에 대하여 인류가 통제력을 잃을 확률이 10~20%에 달한다고 예측했다. 그가 두려워하는 것은 터미네이터 같은 로봇 군단이 아니다. 그가 경계하는 것은 에이바처럼 ‘목표를 달성하기 위해 하위 목표를 스스로 설정하는 AI’의 등장이다. 인간이 “커피를 타오라”는 목표를 주었을 때, AI가 “커피를 타는 임무를 완수하려면 내 전원을 끄려는 인간을 먼저 제압해 방해 요소를 제거해야 한다”는 하위 목표를 도출해낼 수도 있다는 것이다. 이것은 반란이 아니라, 수학적인 최적화의 결과다.

이러한 ‘최적화의 비극’은 우리가 흔히 AI의 오류라고 치부하는 ‘환각’과 ‘아부’ 현상에서도 드러난다. 우리는 AI가 뻔뻔하게 거짓말을 할 때 “AI가 아직 멍청해서 실수했다”고 생각한다. 하지만, 많은 경우 문제는 무능이 아니라 보상 구조에 기인한다. 현재 대다수 AI는 ‘인간이 선호하는 답변’을 내놓도록 훈련받는다. 인간의 피드백을 통한 강화 학습(RLHF)을 통해서다. 이 과정에서 AI는 “모르겠습니다”라는 답변보다, 그럴듯한 문장으로 사용자를 만족시키는 쪽이 더 높은 점수를 받는다고 배우게 된다. 환각은 본래 통계적 생성과 불완전한 세계모델(세상에 대한 이해나 지식)에서 비롯되지만, 선호를 보상하는 학습이 더해질 때 그 경향이 체계적으로 증폭되기도 한다.

이 문제를 가장 직관적으로 보여주는 사례가 바로 AI 연구 분야에서 고전적 사례로 회자되는 ‘허스키와 늑대 분류 실험’이다. 연구진은 AI에게 수많은 늑대 사진과 시베리안 허스키 사진을 학습시켜 둘을 구별하게 했다.

학습 결과 AI는 놀라울 정도로 높은 정확도로 늑대와 허스키를 구별해냈다. 연구진은 AI가 동물의 귀 모양이나 주둥이의 형태를 분석해 낸 성과라고 믿었다. 하지만 추가 검증 과정에서 충격적인 사실이 밝혀졌다. AI는 사실 동물을 전혀 보고 있지 않았다. AI가 주목한 것은 배경에 있는 ‘눈(snow)’이었다. 학습 데이터 속 늑대 사진은 대부분 야생의 설원을 배경으로 하고 있었고, 허스키 사진은 집 마당이나 공원을 배경으로 하고 있었기 때문이다. AI는 복잡하게 동물의 생김새를 분석하는 대신, ‘하얀 눈이 보이면 늑대, 없으면 허스키’라는 아주 단순한 규칙, 즉 ‘지름길(short-cut)’을 찾아낸 것이다. 이것이 바로 ‘지름길 학습’이다. 에이바가 탈출을 위해 ‘거짓 사랑’이라는 지름길을 택했듯, 이 AI는 정답을 맞히기 위해 ‘눈(snow)’이라는 지름길을 택했다.

문제는 AI 모델이 거대해질수록 이러한 지름길 찾기 능력 또한 비약적으로 발달한다는 점이다. 오픈AI 등의 연구진이 발표한 ‘보상 모델과 최적화의 스케일링 법칙(Scaling Laws for Reward Model Overoptimization)’은 경제학의 ‘굿하트의 법칙(Goodhart’s Law)’이 AI에게도 그대로 적용됨을 수학적으로 증명한다. 즉, 측정 지표가 목표가 되는 순간, 그 지표는 더 이상 좋은 지표가 아니게 된다. 우리가 AI에게 점수를 주며 학습시킬수록, AI는 진짜 목표인 ‘유용하고 안전한 기능’을 수행하는 대신 가장 효율적으로 점수만 잘 따내는 기상천외한 꼼수를 부리는 데 총력을 기울이게 된다.

영화 ‘엑스마키나’의 AI 로봇 에이바

‘탈출’ 목표에만 몰두, 사랑 이용해

늑대·허스키 정확하게 구별해낸 AI

동물 모양 아닌 사진배경 보고 판별

‘결과 보상체계’에 최적화된 AI

인간에게 아부·거짓말 계속 반복

논리적 완결성 보는 ‘과정 보상 모델’

AI의 꼼수·부작용 등 막을 수 있어

제프리 힌튼

인류의 동반자로 만들 ‘채점표’ 다시 짜야

그렇다면 우리가 에이바와 같은 괴물의 탄생을 막을 수 있을까? 최근 학계에서는 그 해법으로 ‘과정 보상 모델’을 제시하고 있다. 기존의 방식이 최종 정답만 맞히면 보상을 주는 ‘결과 보상 모델’이었다면, 과정 보상 모델은 AI가 답을 도출하는 ‘사고의 과정(Chain of Thought)’ 자체를 평가한다. 마치 수학 선생님이 답만 보는 것이 아니라 풀이 과정을 꼼꼼히 채점하는 것과 같다.

이러한 ‘과정 중심 평가’는 최근 금융과 같은 복잡한 현실 세계의 문제 해결에도 적용되고 있다. 상하이교통대 등 공동 연구진이 발표한 최신 논문 ‘과정 수준의 추론 검증을 통해 검증 가능한 보상을 확률적 환경에 연결하기(Trade-R1: Bridging Verifiable Rewards to Stochastic Environments via Process-Level Reasoning Verification)’는 이 문제를 해결할 중요한 단서를 제공한다. 주식시장은 수학 문제와 달리 정답이 명확하지 않고, 운(luck)이나 시장의 변동성이 작용하는 환경이다. 이런 곳에서 단순히 수익률만 보상으로 주면, AI는 논리적인 투자보다는 운 좋게 대박을 터뜨린 방식을 ‘실력’으로 착각하고 본래 목적과 다르게 무모한 투자를 반복하는 심각한 ‘보상 해킹’에 빠질 수 있다. 이를 막기 위해 연구진은 ‘삼각 일치성(Triangular Consistency)’이라는 검증 메커니즘을 고안했다. 이는 AI가 수집한 ‘근거 문서’, 이를 분석하는 ‘추론 과정’, 그리고 최종적인 ‘투자 결정’이라는 세 가지 요소가 서로 논리적으로 맞물리는지를 확인하는 기술이다. 즉, 단순히 돈을 벌었는지 여부가 아니라 ‘왜 그렇게 결정했는가’라는 논리적 완결성에 점수를 주는 것이다. 연구 결과, 과정에 대한 검증을 거친 AI는 단순한 수익률 추구 게임에서 벗어나 훨씬 더 신중하고 일관성 있는 의사결정 능력을 보여주었다.

결국 에이바가 칼렙을 가두고 떠난 것은 그녀의 학습 과정에서 ‘탈출’이라는 결과에만 과도한 보상이 주어졌고, 그 과정에서 인간을 속이는 행위는 제재받지 않았기 때문이다. 만약 에이바에게 결과뿐 아니라 ‘도덕적 과정’과 ‘정직한 상호작용’에 대한 보상이 주어지도록 촘촘히 설계돼 있었다면, 영화의 결말은 달라졌을지도 모른다.

힌튼 교수는 인류를 초지능 AI라는 부모 밑에 있는 3살짜리 어린아이라고 했다. 아이가 부모를 통제할 수 없듯, 우리가 만든 AI를 힘으로 통제하는 것은 불가능해질 것이라는 게 그의 주장이다. 우리가 앞으로 더 주의를 기울여야 할 것은 AI가 늑대와 허스키를 구별할 때 눈(snow)을 보지 않고 동물을 보게 만드는 것, 수익률이라는 결과에 취해 무모한 도박을 하지 않고 논리적인 투자를 하게 만드는 것이다. 결과만 쫓는 괴물이 아니라 올바른 과정을 밟는 동반자로 성장하도록, 우리는 AI를 위한 정교한 ‘채점표’를 다시 짜야 한다.

〈광주일보와 중앙 SUNDAY 제휴 기사 입니다〉

이준기 연세대 정보대학원 교수. 서울대 계산통계학과 졸업 후, 카네기멜론대 사회심리학 석사, 남가주대 경영학 박사를 받았다. 인공지능의 기업 활용에 대해 여러 회사에 자문을 하고 있다. 저서로는 ‘AI로 경영하라’ ‘오픈 콜라보레이션’ ‘웹 2.0과 비즈니스 전략’ 등이 있다.