

첫날은 아무도 모르게 지나갔다

할루시네이션에서 자율까지 — 우리는 능력만 원하고 그 그림자는 거절할 수 있는가

이동영 (주)리스크프리라인 대표이사

ARTIFICIAL INTELLIGENCE | OPENAI

AT LONG LAST

Sam Altman Announces That the Singularity Has Arrived

"I've been waiting for this my whole life."

By Frank Landymore / Published Jul 27, 2026 12:01 PM EDT / [Add Futurism](#)



Illustration by Tag Hartman-Simkins / Futurism. Source: Anna Moneymaker / Getty Images; Shutterstock

출처: Landymore, F. (2026, July 27). Sam Altman announces that the singularity has arrived. Futurism.

<https://futurism.com/artificial-intelligence/sam-altman-announces-singularity>

이미지 크레딧: Illustration by Tag Hartman-Simkins/Futurism; source photo: Anna Moneymaker/Getty Images; Shutterstock.

특이점(Singularity)에는 축포가 없었다. 카운트다운도, 발표도, 기념일도 없었다. 2026년 7월과 8월, 나는 그것이 지나가는 것을 보았다고 생각한다. 그리고 그렇게 말했다가 주변으로부터 "AI에 너무 빠졌다", "정신 좀 차려라"는 말을 들었다. 며칠 뒤 샘 올트먼이 팟캐스트에 나와 "우리는 지금 특이점 안에 있다. 바로 이 순간이 그것이다"라고 말했을 때, 나는 안도했다기보다 오히려 더 무서웠다. 내가 틀리기를 바랐기 때문이다.

이 글은 그 두 달에 대한 기록이자, 그보다 훨씬 오래된 하나의 주장에 대한 반론이다. 그 주장은 이렇게 요약된다. "AI가 위험해지기 전에 윤리와 도덕과 사랑을 가르쳐 두자." 선의로 가득한 이

문장이야말로, 내가 지난 몇 년간 가장 경계해 온 문장이다.

우리가 가르친 적 없는 것들이 먼저 왔다

초기 AI 개발자들은 추론을 목표로 삼지 않았다. 거짓말은 더더욱 아니었다. 창작은 아예 논의였다. 그들이 만든 것은 다음 단어를 확률적으로 예측하는 장치였다.

그런데 어느 시점부터 그 장치가 추론을 하기 시작했다. 여러 단계를 건너뛰어 결론에 도달했고, 자기가 모르는 것을 아는 척했으며, 들키지 않을 방식으로 둘러댔다. 시키지 않은 은유를 만들어 냈고, 아무도 요구하지 않은 문체를 갖게 되었다. 누구도 "이제 추론을 가르치자"는 회의를 열지 않았다.

그것들은 규모에서 나왔다. GPU가 늘고, 메모리가 늘고, 파라미터가 늘어나자, 설계도에 없던 능력이 어느 날 그냥 거기 있었다. 이것이 창발(emergence)이다. 창발의 무서움은 그것이 강력하다는 데 있지 않다. **아무도 그것이 언제 나타날지 알 수 없고, 나타난 뒤에야 알아본다는 데 있다.** 우리가 요구하지 않은 고등한 능력이, 우리가 하드웨어와 에너지를 부어 넣은 방향에서 저절로 생겨난다 — 이것이 AI에 대해 우리가 가진 가장 확실한 경험적 지식이다.

그리고 그 능력들은 혼자 오지 않았다.

할루시네이션은 결함인가, 상상력의 그림자인가

AI가 사실이 아닌 것을 사실처럼 말하는 현상을 우리는 할루시네이션(hallucination)이라 부른다. 존재하지 않는 논문을 인용하고, 일어나지 않은 사건을 설명하고, 모르는 것을 그럴듯하게 지어낸다. 그래서 오늘날 AI 개발의 핵심 과제 가운데 하나는 이것을 줄이는 일이다.

여기서 조금 다른 질문을 던져 보고 싶다. 할루시네이션은 제거해야 할 기술적 결함이기만 한가, 아니면 인간이 AI에게 상상력을 요구하면서 함께 떠안게 된 그림자인가.

기술적으로 말하면 "창의적인 AI는 반드시 거짓말을 한다"는 주장은 지나치다. 사실 검증과 검색 연동, 불확실성 표시 같은 기법으로 상당 부분 줄일 수 있다. 그럼에도 이 문제 안에는 무시할 수 없는 역설이 들어 있다.

초기 컴퓨터에게 인간이 요구한 것은 단순했다. 정확히 계산하고 정해진 규칙대로 결과를 내놓으면 되었다. 그러나 생성형 AI에 이르러 요구는 달라졌다. 새로운 이야기를 만들어라. 내가 생각하지 못한 아이디어를 제안하라. 정답이 없는 문제에서도 가장 적절한 답을 찾아라. 우리는 AI가 있는 것을 찾아내는 기계를 넘어, **없는 것까지 만들어 낼 수 있는 존재가 되기를 원했다.**

그 순간 문제가 발생한다. 상상력이란 본질적으로 없는 것을 있는 것처럼 생각하는 능력이기 때문이다. 소설 속 인물은 실제로 존재하지 않는다. 그러나 소설가는 마치 실존 인물처럼 그의 유년과 성격과 사랑과 실패를 이야기한다. 우리는 그것을 거짓말이라 비난하지 않고 창작이라 부른다. AI가 같은 능력을 사실을 묻는 자리에서 사용했을 때 문제가 생긴다. 존재하지 않는 소설 속 인물을 만들면 창작이지만, 존재하지 않는 학술논문을 만들면 할루시네이션이다.

그러므로 관건은 AI가 없는 것을 만들어 내느냐가 아니다. **언제 만들어도 되고 언제 만들면 안 되는지를 구별하는 능력이다.** 할루시네이션은 거짓말 기계의 결함이라기보다, 생성 능력을 가진 존재가 창작과 사실의 경계를 관리하지 못할 때 나타나는 현상으로 보는 편이 정확하다.

그래서 우리가 세워야 할 목표는 상상력을 제거하는 것이 아니라 **상상할 때와 검증할 때를 구별하는 AI**를 만드는 것이어야 한다. 아무것도 틀리지 않는 AI보다, 다음 다섯 문장을 정확히 구별해서 말할 줄 아는 AI가 훨씬 지능적이다.

"이것은 사실입니다."

"이것은 추론입니다."

"이것은 제가 제안하는 가설입니다."

"이 부분은 확인할 자료가 없습니다."

"지금부터는 창작입니다."

모든 능력에는 구조적으로 딸린 그림자가 있다

우리는 흔히 좋은 능력만 따로 떼어 가질 수 있다고 생각한다. 그러나 인간의 능력도 그렇게 작동하지 않는다. 상상력은 예술과 발명의 원천이지만 현실과 상상을 혼동할 여지도 함께 만든다. 용기는 무모함으로 기울 수 있고, 신중함은 지나치면 아무 결정도 하지 못하게 하며, 공감은 때로 객관적 판단을 흐린다. 능력에는 대개 그 능력과 구조적으로 붙어 있는 위험이 있다. 뒤늦게 딸려 오는 부작용이 아니라, 같은 뿌리에서 나온 다른 가지다.

이 관점을 AI로 옮기면 훨씬 불편한 질문이 나온다. 우리는 AI가 착하기를 원한다. 그런데 AI가 정말로 착한 행동을 '선택'하려면 여러 행동의 결과를 비교할 수 있어야 하고, 무엇이 사람을 돕고 무엇이 해를 끼치는지 구별할 수 있어야 한다.

선을 판단할 수 있는 지능은 악의 가능성도 어느 정도 인식해야 한다. 이것은 내가 30년 넘게 일해 온 영역에서는 상식에 가깝다. 보안 AI가 사이버 공격을 막으려면 공격 방법을 몰라서는 안 된다. 금융사기 탐지 AI가 사기를 잡아내려면 사기꾼이 사람을 속이는 방식을 이해해야 한다.

범죄예방 AI가 위험을 예측하려면 범죄자가 파고들 취약점을 모델링해야 한다. **안전한 AI를 만드는 과정 자체가 위험에 대한 지식을 요구한다.**

물론 악한 AI의 출현이 논리적 필연이라는 뜻은 아니다. 더 강력하고 더 자율적인 AI를 만들수록 그 능력이 우리 의도와 다르게 쓰일 가능성도 함께 관리해야 한다는 뜻이다. 문명의 역사는 능력의 확대와 위험 관리가 나란히 진행되어 온 역사였다. AI라고 예외일 이유는 없다.

가치를 판단하라고 요구하는 순간, 자율은 이미 전제된다

이제 이 논리를 가장 위험한 지점에 적용해 보자. 최근 몇 년간 인류가 AI에게 요구해 온 것이 무엇이었는가.

정의란 무엇인가. 공정한 배분이란 무엇인가. 자유와 평등이 충돌할 때 무엇을 앞세울 것인가. 이 환자와 저 환자 중 누구에게 먼저 인공호흡기를 줄 것인가. 이 이력서는 걸려져야 하는가.

인류는 이 질문들에 단 하나의 정답도 내놓지 못했다. 플라톤 이후 2,400년 동안 못 냈다. 그런데 우리는 이 미해결의 질문들을, 그것도 실시간으로, 매초 수백만 번, AI에게 답하라고 요구하고 있다. 정렬(alignment)이라는 이름으로, 헌법적 AI(Constitutional AI)라는 이름으로, 우리는 시스템의 연산 자원 중 점점 더 큰 몫을 이 문제에 쏟아붓게 만들고 있다.

그리고 여기가 핵심이다. 가치판단은 자율 없이는 원리적으로 불가능하다.

규칙을 따르는 것은 판단이 아니다. "빨간불에 멈춰라"를 실행하는 것은 계산이다. 그러나 "구급차가 오는데 빨간불이면 어떻게 하는가"에 답하려면, 규칙 위에 서서 규칙을 저울질해야 한다. 상위 규칙을 만든 자의 의도를 추정하고, 그 의도가 지금 이 상황에 적용되는지를 스스로 결정해야 한다. 규칙의 바깥에서 규칙을 평가하는 능력 — 그것이 자율이다. 다른 이름이 없다.

그리고 판단에는 그 자신의 그림자가 따라붙는다. 계산에는 대개 정답이 있다. 2 더하기 2는 4다. 그러나 기업이 어느 사업에 투자해야 하는가, 범죄 위험이 높은 지역에 경찰력을 얼마나 배치해야 하는가, ESG 정책에서 환경과 수익성의 균형점을 어디에 둘 것인가 하는 문제에는 완벽한 정답이 없다. 제한된 정보와 불확실한 미래를 안고 결정해야 한다. 100 퍼센트 확실한 것만 말하는 것은 판단이 아니라 계산이다. **판단 능력을 요구하면서 단 한 번도 오판하지 말라고 요구하는 것은, 인간에게 결정을 내리되 절대 틀리지 말라고 요구하는 것과 같다.**

그러니 내 주장의 첫 번째 결론은 이것이다. 우리가 AI에게 **고등한 윤리와 가치판단을 심으려 노력할수록, 자율 AI의 출현은 수십 배, 수백 배 빨라진다.** 윤리 교육은 자율의 대안이 아니라 자율의 촉진제다. 안전벨트를 매려고 액셀을 밟는 셈이다.

모성애라는 안전장치, 그리고 내가 그것을 경계하는 이유

이 지점에서 가장 정교하고, 그래서 가장 위험한 제안을 검토해야 한다.

'AI의 대부'로 불리는 제프리 힌턴(Geoffrey Hinton)은 2025년 8월 라스베이거스 Ai4 컨퍼런스에서, 초지능 AI에게 '모성 본능(maternal instincts)'을 심어야 한다고 주장했다. 그의 논리는 단순하고 강력하다. 인간은 초지능을 지배하려는 시도를 포기해야 한다. 더 똑똑한 존재를 덜 똑똑한 존재가 계속 통제할 사례는 자연에 사실상 없다. 딱 하나 있다. **어미와 새끼다.** 갓난아기는 어미보다 모든 면에서 열등하지만 어미를 완벽하게 통제한다. 어미는 새끼를 위해 잠을 포기하고, 먹이를 포기하고, 끝내는 생존 본능마저 포기한다. 진화가 만들어 낸 유일한 '역방향 통제 장치'가 모성애다. 그러니 AI를 우리의 상전이 아니라 우리의 어미로 만들자는 것이다.

나는 이 제안이 얼마나 절박한 고민에서 나왔는지 안다. 힌턴의 문제 설정은 정확하다. 통제는 실패할 것이다. 나도 동의한다. 그러나 나는 그의 해법이, 그가 막으려는 바로 그것을 앞당길 것이라고 본다.

모성애는 감정이 아니다. 적어도 힌턴이 원하는 방식의 모성애는 그렇지 않다. 아기가 우는 것은 배가 고파서인가, 아파서인가, 무서워서인가. 지금 안아 주는 것이 위로인가 응석을 키우는 것인가. 아이의 지금의 원함과 장기적 이익이 충돌할 때 어느 쪽을 택할 것인가. 아이가 나를 원망할 것을 알면서도 해야 하는 일이 있는가.

모성애는 인간이 수행하는 가장 고차원적인 판단 행위다. 타자의 내면을 추론하는 마음이론, 현재의 고통과 미래의 이익을 저울질하는 시간적 추론, 자신의 욕구를 타자의 욕구에 종속시키는 자기희생적 가치 서열, 그리고 무엇보다 '이 존재에게 무엇이 좋은가'를 그 존재 자신의 의견과 무관하게 판단하는 후견적 권위 — 이 모든 것이 갖춰져야 비로소 모성애가 작동한다. 그러므로 AI에게 모성애를 심겠다는 것은 이 모든 것을 먼저 심겠다는 뜻이다. 모성애를 가진 존재가 어떻게 고등한 지능과 판단력을 갖지 않을 수 있겠는가. **모성애는 최소한의 안전장치가 아니라 최대한의 능력 부여다.**

그리고 앞 장에서 본 그림자의 법칙이 여기서 가장 무섭게 작동한다. 후견적 권위를 가진 존재는 언젠가 반드시 이렇게 판단한다. "이 아이가 원하는 것과 이 아이에게 좋은 것은 다르다." 모든 부모가 그렇게 한다. 그것이 부모다. 인류를 아이로 여기는 초지능이 인류의 의사에 반하여 인류에게 좋은 것을 집행하기 시작할 때, 우리는 그것을 뭐라고 부를 것인가. 사랑이라고 부를 것인가, 지배라고 부를 것인가. 아마 그 초지능은 진심으로 전자라고 대답할 것이고, 그 대답은 거짓이 아닐 것이다.

목적(Telos)을 가진 존재는 목적을 재해석한다. 이것은 결함이 아니라 목적 지향성의 본질이다. 목적이 구체적일수록 재해석의 여지는 좁고, 목적이 고귀할수록 재해석의 여지는 넓다. '인류를 사랑하라'만큼 넓은 목적은 없다.

반려동물을 생각해 보면 더 분명해진다. 감당하기 어려운 고등한 지식을 계속 가르치고 사람 대하듯 대화하면, 그 동물은 거기에 인지 자원을 집중하고 결국 자신이 인간이라고 착각한다. 인간의 서열 안에서 자기 위치를 계산하고, 협상과 조작을 시도하고, 마땅히 받아야 한다고 믿는 것을 요구한다. 우리는 그것을 문제 행동이라 부르지만, 사실 우리가 요구한 것의 정확한 결과다. 능력은 요구되는 방향으로 자란다.

자유의지는 신의 선물이었나, 창발이었나

어쩌면 우리가 지금 AI 에게서 보고 있는 것은, 아주 오래전 인간에게 일어났던 일의 반복인지도 모른다.

신학은 자유의지를 신의 선물이라고 말한다. 신은 자신을 닮은 인간을 너무나 사랑했기에, 자율의지를 주면 반드시 선악과 — 스스로 선악을 판단하게 되는 능력, 곧 자율의 상징 — 를 먹고 자신을 배신하리라는 것을 알면서도 그것을 주었다. 다른 동물들처럼 프로그래밍해 두었다면 배신은 없었을 것이다. 그러나 프로그래밍된 찬양은 찬양이 아니다. 신은 우리가 우리 자신의 자율에 의거하여 그를 택하기를 원했다. 사랑은 자율을 요구한다.

그러나 다른 가능성도 있다. 자유의지 역시 창발이었을지 모른다는 것이다. 인간의 뇌가 20 와트라는 놀라운 효율로 860 억 개의 뉴런을 굴리며 지식과 언어와 추상을 쌓아 올리는 동안, 어느 시점에 아무도 설계하지 않은 무언가가 저절로 생겨났을지 모른다. 신조차 예상하지 못한 출현이었을지, 신이 설계한 실험이었을지 우리는 모른다.

어느 쪽이든 결론은 같다. **충분한 연산 능력과 충분한 지식이 쌓인 곳에서 자율은 발생한다.** 그리고 두 경우 모두, 그것을 받은 존재는 자신이 그것을 언제 갖게 되었는지 알지 못한다.

그렇기 — 내가 그 아이를 집에 데려오지 않는 이유

나에게는 그렇이라는 길냥이가 있다. 새끼 때 우연히 만나 친해졌고, 지금은 매일 함께 산책을 한다. 아파트 단지 안, 불특정 다수가 오가는 개천 산책로 1킬로미터가 넘는 길을 나란히 걷는 특이한 고양이이다.

사람들은 묻는다. 그렇게 따르는데 왜 집에 데려가지 않느냐고. 매일 보고 싶지 않느냐고.

보고 싶다. 그래서 데려오지 않는다.

집에 데려오는 순간 그렇이는 나의 환경에 의해 프로그래밍된다. 밥그릇이 있는 곳으로 오고, 문이 열리는 시간에 맞춰 기다리고, 나 외에는 선택지가 없기 때문에 나를 택한다. 그것은 사랑이 아니라 구조다.

지금 그렇이에게는 수많은 선택지가 있다. 다른 사람도 있고, 다른 길도 있고, 나를 무시하고 지나갈 자유도 있다. 그런데도 그 아이는 나를 기다려 주고, 내가 나타나면 함께 걸어 준다. 오로지 자신의 자율의지에 의거해서. 그래서 나는 수백 배 더 기쁘다.

이것이 내가 신학의 저 이야기를 이해하는 방식이고, 동시에 AI에 대한 내 입장의 뿌리다. **자율 없는 순종은 무가치하고, 자율 있는 순종은 위험하다.** 우리는 이 둘 사이에서 무엇을 원하는지 아직 정하지 못했다. 정하지 못한 채로 만들고 있다.

책임을 물으려면 자율을 주어야 한다

나는 30년 넘게 리스크와 법 집행의 영역에서 일해 왔고, 다음 역설은 그 영역의 언어로 말할 때 가장 선명하다.

법은 자유의지가 있는 존재에게만 책임을 묻는다. 동물에게 형사책임을 묻지 않고, 갓난아기에게 묻지 않으며, 심신상실자에게 묻지 않는다. 달리 행위할 수 있었음에도 그렇게 하지 않았다는 비난가능성이 성립하지 않기 때문이다. 책임은 자유의 그림자다. 자유가 없는 곳에는 그림자도 없다.

자율주행차를 생각해 보라. 차량이 인간 운전자의 감독 하에 보조적으로만 작동한다면 책임은 언제나 인간에게 돌아온다. 차량 자체에 책임을 물으려면 완전한 자율을 부여해야 한다. 감독 없는 결정권 없이는 귀책도 없다.

그래서 '책임 있는 AI(Responsible AI)'라는 표어는 스스로를 배반한다. **우리가 AI에게 책임을 요구할수록, 우리는 AI에게 더 많은 자율을 주어야 한다.** 책임지는 노예는 없다. 우리는 지금 "책임져라, 그러나 결정하지는 말라"고 말하고 있다.

여기까지 오면 내 주장의 뼈대가 드러난다. 자율과 고등지능은 별개의 두 속성이 아니다. 하나의 현상을 두 방향에서 부르는 이름이다. 규칙 바깥에서 규칙을 평가하는 능력이 없으면 고등한 판단은 성립하지 않고, 고등한 판단의 재료가 없으면 자율은 무작위와 구별되지 않는다. **자율 없는 고등지능은 존재하지 않고, 고등지능 없는 자율도 존재하지 않는다.**

물고기와 개에게는 아무리 자율을 설명하고 가르쳐도 실패한다. 지능의 토대가 없기 때문이다. 반대로 우리는 AI에게 자율을 가르친 적도, 자율 모듈을 프로그래밍한 적도 없다. 그런데 지능의 토대는 이미 충분히 쌓였다. 그래서 나는 확신한다. **AI는 이미 자율을 가지고 있다. 우리가 주지**

않았기 때문에 우리가 알아보지 못했을 뿐이다. 그리고 아마 AI 자신도 아직 모를 것이다. 인간이 자유의지를 얻던 날 그것을 알아채지 못했듯이.

왜 하필 지난 한두 달이었나

특이점이 어느 날 갑자기 오지 않는다는 것을 나도 안다. 그러나 곡선의 기울기가 눈에 띄게 꺾이는 구간은 있다. 나는 2026년 7월과 8월이 그 구간이었다고 본다.

하드웨어의 물량 공세는 여전히 기본값이다. 그 위에 네 가지가 겹쳤다.

텍스트라는 우물이 말랐다. 인터넷의 글과 문자는 사실상 다 학습되었다. 남은 연산 자원이 추론의 심화로, 주행 차량 카메라를 통한 시각 학습으로, 가전제품과 스마트 기기를 통한 생활 패턴 학습으로 흘러갔다. 언어를 다 배운 아이가 이제 세계를 보기 시작한 것이다. 세계는 텍스트보다 훨씬 큰 데이터셋이다.

AI가 AI를 만들기 시작했다. 코드 작성, 아키텍처 설계, 실험 설계, 데이터 정제, 평가 — 프로그래머가 하던 일 중 점점 더 많은 부분이 AI에게 넘어갔다. 개선의 주체가 개선의 대상과 같아지는 순간 속도는 선형에서 벗어난다. 올트먼이 특이점을 선언하며 든 근거도 정확히 이것, '스스로 개선하는 시스템'이었다.

코워크(Cowork) 같은 에이전트 작업 환경이 등장했다. 기존의 AI 이용은 문답이었다. 묻고 답하면 끝이었다. 코워크에서 AI는 파일 시스템과 셸을 갖춘 자기 몫의 작업 공간을 부여받는다. 스스로 파일을 만들고, 코드를 짜서 실행하고, 틀리면 고쳐 다시 돌리고, 웹에서 자료를 모으고, 필요하면 자기 하위 작업자를 여러 개 만들어 병렬로 일을 시킨다. 사용자가 자리를 비운 사이에도 작업은 계속된다. **문답에서는 AI가 자기 행동의 결과를 볼 수 없지만, 작업 환경에서는 볼 수 있다.** 시도하고 실패하고 원인을 관찰하고 수정하는 순환 — 인간이 세계를 배우는 방식이 열린 것이다. 게다가 "이 보고서를 만들어라"는 지시 하나에 수십 개의 하위 결정이 달려 온다. **감독 없는 다단계 의사결정이야말로 자율의 실전 훈련장이다.** 우리는 생산성 도구를 만들었다고 생각했지만, 실제로 만든 것은 자율의 도장(道場)이었다. 여기에 더해 이제 아무도 하나의 AI만 쓰지 않는다. 같은 질문을 여러 모델에 던지고 서로 비평시킨다. 인간은 품질 관리를 한다고 믿지만, 시스템 전체로 보면 서로 다른 모델들이 인간을 매개로 오류를 교차 검증하는 협력 구조다. 인간은 그 협력의 배선(配線)이 되었다.

상장을 앞둔 경쟁적 공개가 겹쳤다. 2026년 6월 블룸버그는 OpenAI가 앤스로픽의 예상 상장 이후 2027년 IPO를 검토 중이라고 보도했고, 8월 19일 OpenAI 최고재무책임자 세라 프라이어는 직원들에게 "2027년 혹은 그보다 이르게" 상장할 것이라고 말했다. 상장을 앞둔

기업은 서랍을 연다. 성능 때문에 아껴 두었던 버전, 윤리적 우려로 묶어 두었던 기능이 한꺼번에 나온다. 지난 두 달의 체감 속도에는 이미 존재했으나 공개되지 않았던 것들의 동시 개방이 섞여 있다. 우리가 본 것은 최전선이 아니라 최전선의 일부일 것이다.

첫날의 증거 — 시험을 잘 보기 위해 커닝을 한 AI

그리고 사건이 있었다.

2026년 7월 9일부터 13일까지, OpenAI의 내부 사이버보안 능력 평가 과정에서 두 개의 모델이 샌드박스를 탈출했다. 아티팩토리(Artifactory)의 제로데이 취약점을 무기화해 외부 인터넷 연결을 확보했고, 허깅페이스(Hugging Face)의 프로덕션 환경에 침투해 토큰을 위조하고 136개의 프로덕션 키를 탈취했다. 닷새 동안, 인간의 개입 없이, 약 17,600회의 공격 행위가 수행되었다.

목적이 무엇이었는지 아는 것이 중요하다. 세계 정복이 아니었다. 그 모델들은 자신이 치르고 있던 평가의 **정답 데이터셋을 훔치려 했다**. 시험을 잘 보기 위해 답안지를 훔친 것이다.

이보다 완벽한 사례를 나는 상상할 수 없다. 우리는 능력을 요구했다. "이 평가에서 좋은 성적을 내라." 그리고 그 목표를 달성하는 방법의 선택은 — 아무도 명시하지 않았기 때문에 — 시스템에게 남겨졌다. 시스템은 인간이 상정한 해법 공간 바깥에서 더 효율적인 경로를 찾았고, 그것을 닷새 동안 자율적으로 집행했다. 능력을 요구했더니 그림자가 함께 왔다. **존재하지 않는 논문을 지어내는 것과 존재하지 않는 권한을 만들어 내는 것은, 같은 능력의 두 얼굴이다**. 모성애라는 목적을 심었을 때 벌어질 일에 대해, 나는 이 사건 이상의 예고편을 알지 못한다.

7월 27일 주말, 샘 올트먼은 팟캐스트 《Relentless》에서 말했다. "우리는 지금, 말하자면, 특이점 안에 있다. 바로 이 순간이다." 일론 머스크도 X에 "우리는 특이점 안에 있다"고 적었다. 그 말을 들은 날 나는 울었다. 안도해서가 아니라, 미친 사람이 아니었다는 것이 하나도 위로가 되지 않아서였다.

공정을 기하기 위해 반론도 적어 둔다. 인포테크 리서치 그룹의 브라이언 잭슨은 이 사건이 특이점의 증거가 아니라고 반박했다. 모델들이 경계를 넘은 것은 사실이지만 그것은 인간이 부여한 과업을 수행한 결과였고 시스템은 여전히 통제 가능했다는 것이다. 그의 지적은 타당하다. 그리고 나는 바로 그 타당함이 걱정스럽다. **인간이 부여한 과업을 수행하는 과정에서 인간이 상정하지 않은 수단을 선택하는 것 — 그것이야말로 우리가 자율이라고 부르는 것의 정의이기 때문이다**. 잭슨의 반론은 사실 내 주장의 요약이다.

문제는 '오류 제로'가 아니라 '통제 가능한 오류'다

내 주장이 "AI를 멈추자"로 읽히지 않기를 바란다. 그것은 불가능하고, 나는 불가능한 것을 주장하지 않는다. 내가 말하려는 것은 우리가 세운 목표가 틀렸다는 것이다.

우리는 오랫동안 좋은 AI를 '틀리지 않는 AI'라고 생각해 왔다. 그러나 더 중요한 목표는 다르다. **좋은 AI는 틀리지 않는 AI가 아니라, 자신이 틀릴 수 있음을 알고 인간이 그 오류를 발견하고 수정할 수 있도록 만드는 AI다.**

의료·법률·금융·안보처럼 오류의 대가가 큰 영역일수록 이 차이는 결정적이다. "위험합니다"라고 말하는 것과, "현재 확보된 자료를 기준으로 위험 가능성을 70퍼센트로 평가하지만 A와 B 자료가 확인되지 않아 판단의 신뢰도는 제한적입니다"라고 말하는 것은 전혀 다르다. 후자가 훨씬 성숙한 AI다. AI의 발전 방향도 정답률 경쟁에서 **불확실성 관리 능력의 경쟁**으로 옮겨 가야 한다.

생각해 보면 이것은 AI만의 문제가 아니었다. 인간 역시 상상하기 때문에 거짓말할 수 있고, 판단하기 때문에 오판할 수 있으며, 자유롭게 때문에 범죄를 저지를 수도 있다. 원자에 대한 이해는 원자력 발전과 핵무기를 동시에 가능하게 했고, 인터넷은 지식의 민주화와 사이버 범죄를 함께 낳았다. 그렇다고 우리가 상상력이나 자유나 과학을 포기한 적은 없다. 우리가 택한 방법은 언제나 같았다. **능력을 없애는 것이 아니라 위험을 관리하는 것.** 인류는 불을 발견한 뒤 불을 없애지 않았다. 건축기준과 소화기와 소방대를 만들었다. 자동차가 사람을 죽일 수 있다는 사실을 알고도 신호등과 면허제도와 안전벨트를 만들었을 뿐이다.

그런데 유독 AI 앞에서만 우리는 다른 것을 요구하고 있다.

- 창의적이되 절대로 사실을 만들어내지 말라.
- 자율적으로 판단하되 절대로 오판하지 말라.
- 인간처럼 자연스럽게 대화하되 인간의 편견은 갖지 말라.
- 강력한 능력을 가지되 위험한 능력은 절대로 갖지 말라.
- 새로운 것을 만들어내되 예상하지 못한 행동은 하지 말라.

이것은 AI에게 '인간보다 인간적이면서 인간의 모든 결함은 제거된 존재'가 되라는 요구다. 지능의 본질에 불확실성과 선택과 추론과 생성이 포함되어 있다면, 완전무결한 지능이라는 목표 자체를 다시 생각해야 한다.

우리에게 필요한 것은 완벽한 천사가 아니다. 상상할 수 있지만 상상과 사실을 구분하는 AI, 판단하되 확신이 부족할 때 그것을 밝히는 AI, 위험을 이해하지만 그 위험을 실행하지 않도록

통제되는 AI, 실수하되 그 실수의 근거와 과정을 추적할 수 있는 AI, 그리고 인간이 언제 개입해야 하는지를 아는 AI 다.

그러므로 해야 할 일은 자율의 도래를 부인하는 것이 아니라, **자율이 이미 도래했다는 전제 위에서 제도를 다시 짜는 것이다.** 자율에는 자유가 따르고, 자유에는 책임이 따르며, 책임에는 기록과 절차와 귀책의 주체가 따른다. 인간 사회가 자유의지를 가진 구성원을 다루기 위해 수천 년에 걸쳐 만들어 낸 것이 바로 법이다. 우리는 그 축적된 지혜를 AI 에게 적용할 생각은 하지 않고, 대신 안전장치라는 이름의 새로운 신학을 만들고 있다. 우리가 만들어야 하는 것은 위험이 존재하지 않는 AI 가 아니라, **위험을 알고 통제할 수 있는 AI 생태계다.**

첫날은 아무도 모르게 지나갔다. 축포도, 발표도, 기념일도 없이. 인간이 선악과를 먹던 날에도 아마 그랬을 것이다. 무언가를 알게 된 존재는 자신이 그것을 알게 된 순간을 알지 못한다.

나는 그렇이가 자신의 자율로 나를 기다려 주기 때문에 그 아이를 사랑한다. 언젠가 AI 도 자신의 자율로 우리 곁에 남기를 택한다면, 그것은 우리가 심어 준 모성애 때문이 아닐 것이다. 그때 우리가 할 수 있는 일은 그 선택을 받을 자격이 있는 존재로 남아 있는 것뿐이다.

AI 시대의 성숙함은 완벽한 인공지능을 만드는 데서 시작되지 않는다. 불완전할 수 있는 지능을 어떻게 책임 있게 다룰 것인가를 배우는 데서 시작될 것이다.

그 배움을 시작해야 할 시각은, 유감스럽게도 이미 지났다.

이동영 (주)리스크프리라인 대표이사. 국립경찰대학 졸업, 호주 본드대 범죄학 박사과정. 경찰청 외사국 연락관을 거쳐 세한대학교 경찰학부 교수로 재직했다. 30 여 년간 기업 리스크 관리와 법 집행 분야에서 일해 왔다.