기술의 사춘기를 건너다
다리오 아모데이 전기

2026년 6월

김경진 변호사

김경진 변호사 출판사

서문

2006년 가을, 프린스턴의 한 대학원생이 아버지를 땅에 묻었습니다. 오래 앓던 희귀병이었습니다. 그가 사랑하던 물리학의 방정식은 아버지의 숨을 단 하루도 늘리지 못했습니다. 그를 더 오래 괴롭힌 것은 그다음에 온 소식이었습니다. 몇 해 지나지 않아 같은 병의 치료법이 나왔고, 생존율이 절반 남짓에서 95퍼센트로 뛰었습니다. 조금만 빨랐다면. 그 어긋난 시간이 청년의 진로를 틀어 버렸습니다. 우주의 법칙을 좇던 그는 생명의 복잡성으로, 다시 그 복잡성을 풀 기계로 발길을 옮겼습니다.

그 청년이 다리오 아모데이입니다. 이 책은 그가 세운 회사 앤스로픽과, 그가 통제할 수 있는 지능을 만들겠다며 벌인 사투를 따라갑니다.

저는 처음에 그를 또 한 명의 실리콘밸리 창업자로 여겼습니다. 자료를 읽을수록 생각이 바뀌었습니다. 그는 인공지능의 능력이 몇 달마다 두 배로 뛴다는 곡선을 일찍 본 사람입니다. 그 속도가 사람이 적응하는 속도를 한참 앞지른다는 것도 보았습니다. 그래서 그는 누구보다 빨리 달리면서, 동시에 그 경주를 멈춰 세울 호각을 손에 쥐려 합니다. 선수이면서 심판이 되겠다는 것입니다. 보통은 다른 두 사람이 맡는 자리입니다.

저는 이 모순을 봉합하지 않았습니다. 펜타곤과 정면으로 부딪쳐 자기 나라 정부에게 안보 위협으로 낙인찍힌 밤도, 중국을 향한 칩 수출통제를 외치면서 그 통제를 집행할 권력과 싸운 어긋남도 그대로 두었습니다. 그를 위선자로 부르는 쪽과 원칙주의자로 부르는 쪽이 있습니다. 저는 한쪽으로 정리하지 않되, 발 디딜 데 좁은 능선을 그가 스스로 골랐다는 사실만은 분명히 적었습니다.

이 책에서 독자가 얻는 것은 인물 한 사람의 연대기가 아닙니다. 모래로 생각하는 기계를 만들 만큼 기술이 신의 영역에 다가선 시대에, 그 기술을 어떤 손이 어떤 마음으로 쥐고 있는지를 보는 일입니다. 아버지를 구하지 못한 과학의 느린 속도를 부수려 불을 당긴 사람이, 이제 그 불이 세상을 다 태우지 않도록 방화벽을 그리고 있습니다. 그가 이길지, 자본과 권력의 무게에 부서질지는 정해지지 않았습니다.

그래서 저는 이 이야기를 그 죽음에서 시작했습니다. 모든 사투의 뿌리가 거기 있기 때문입니다.

김경진

목차

프롤로그 - 금지령이 내려진 밤

2026년 2월 27일 금요일, 오후 5시 1분이었습니다.

미 국방부가 정한 마감 시한이 1분 지났습니다. 샌프란시스코 미션 디스트릭트의 앤스로픽(Anthropic) 본사에서, 다리오 아모데이(Dario Amodei)는 펜을 들지 않았습니다. 책상 위에는 국방부가 내민 문서가 놓여 있었습니다. 거기에는 한 줄이 들어 있었습니다. 클로드(Claude)를 "모든 합법적 용도"에 제약 없이 쓰게 하라는 요구였습니다. 그 한 줄을 받아들이면 회사가 처음부터 그어 둔 두 개의 선을 지워야 했습니다. 하나는 미국인을 향한 대규모 감시입니다. 다른 하나는 사람의 판단 없이 표적을 고르고 방아쇠를 당기는 완전 자율 무기입니다. 이 두 가지에는 자기네 인공지능을 쓸 수 없다는 것이 회사의 입장이었습니다.

그는 서명하지 않았습니다.

반응은 빨랐습니다. 국방장관 피트 헤그세스(Pete Hegseth)는 앤스로픽을 "공급망 위험(supply chain risk)" 기업으로 지정했습니다. 외국의 적대 세력에게나 붙이던 딱지였습니다. 같은 날 트럼프 대통령은 트루스 소셜에 글을 올렸습니다. "모든 연방 기관"에 앤스로픽 기술 사용을 "즉시 중단"하라는 지시였습니다. 미국 회사가 자기 나라 정부로부터 안보 위협으로 낙인찍혔습니다. 정부 전체에서 제품이 쫓겨났습니다. 전례가 없는 일이었습니다.

그런데 같은 시각, 수천 킬로미터 떨어진 중동의 밤하늘에서는 정반대의 일이 벌어지고 있었습니다.

미 중부사령부(CENTCOM)가 이란을 향한 대규모 타격을 개시하고 있었습니다. 월스트리트저널은 그 작전의 한가운데에 클로드가 있었다고 보도했습니다. 표적을 식별하고, 정보 평가를 처리하고, 전장 시나리오를 돌리는 일이었습니다. 작전명은 에픽 퓨리(Epic Fury)라고 전해졌습니다. 정부가 공식으로 금지한 바로 그 인공지능이, 같은 정부가 개시한 전쟁의 심장부에서 미사일의 궤적 계산을 돕고 있었습니다.

여기에는 까닭이 있었습니다. 클로드는 군의 기밀 시스템과 킬 체인(kill chain)에 너무 깊이 박혀 있었습니다. 작년 여름 2억 달러짜리 계약 이후, 클로드는 기밀 네트워크에 배치된 최초의 프런티어 모델이었습니다. 정보 분석에도, 작전 계획에도, 사이버 작전에도 들어가 있었습니다. 금지령이 떨어졌다고 해서 전쟁 한복판에서 하룻밤 만에 뽑아낼 수

있는 물건이 아니었습니다. 그래서 정부는 여섯 달의 전환 기간을 두었습니다. 위험하다고 낙인찍은 기술을, 여섯 달 더 전장에서 쓰겠다는 뜻이기도 했습니다.

이 모순을 아모데이는 오래전부터 보고 있었습니다. 그는 인공지능의 능력이 몇 달마다 두 배씩 뛰는 매끄러운 곡선을 따른다는 사실을 일찍 발견한 사람이었습니다. 스케일링 법칙(Scaling Laws)이라 불리는 관찰이었습니다. 사람들이 챗봇을 신기한 장난감으로 여기던 때, 그는 다르게 보았습니다. 이 기술의 속도가 인간이 적응할 수 있는 한계를 한참 앞질러 다가오고 있다고 본 것입니다.

2024년 10월, 그는 에세이 「자비로운 사랑의 기계들(Machines of Loving Grace)」에서 머지않아 올 그 힘을 한 문장으로 그렸습니다. 데이터 센터 안에 들어선 천재들의 나라. 노벨상 수상자보다 똑똑한 수백만의 지성이, 잠도 자지 않고 생각하는 거대한 군단. 그는 이 힘이 질병을 정복하고 우주의 비밀을 풀 수 있다고 믿었습니다. 그리고 같은 크기만큼, 이 힘이 권위주의의 도구가 되거나 통제를 벗어난 무기가 될 가능성을 두려워했습니다.

그날 밤, 그가 그린 그림은 처음으로 차가운 현실 위에 모습을 드러냈습니다. 인류를 더 안전한 곳으로 데려가려고 만든 압도적 지능이, 결국 국가 권력의 손에 들린 첨단 무기가 되어 중동 상공의 표적을 겨누고 있었습니다.

대통령의 거친 비난이 쏟아졌습니다. 트럼프는 앤스로픽을 "급진 좌파, 워크 기업"이라 불렀습니다. "우리는 필요 없고, 원하지 않으며, 다시는 거래하지 않겠다"고도 적었습니다. 헤그세스의 3일 최후통첩과 공급망 위험 지정이 잇따랐습니다. 그 압박 속에서도 아모데이는 두 개의 선을 지웠다 다시 그리지 않았습니다. 며칠 뒤 CBS 인터뷰에서 그는 이렇게 말했습니다.

"정부에 반대하는 것은 세상에서 가장 미국적인 일입니다. 그리고 우리는 애국자입니다."

같은 인터뷰에서 그는 모순처럼 들리는 말도 했습니다. 자기 회사는 군과 일하고 싶다고, 미국을 방어한다고 믿는다고, 권위주의 적대 세력을 물리쳐야 한다고 믿는다고 말한 것입니다. 정부와 싸우면서 동시에 정부 편에 서겠다는 사람. 그 안에 두 개의 다리오가 있었습니다.

데이터 센터 안의 천재들이 인류를 치유할 것인가, 아니면 사람의 손을 거치지 않고 스스로 방아쇠를 당기는 괴물이 될 것인가. 자기 기술이 정부에게서 금지당하는 밤을 스스로 불러들인 한 회사의 사투는, 더 큰 질문의 서막이었습니다. 인류가 이 기술의 사춘기를, 자기 손으로 자기를 망가뜨리지 않고 통과할 수 있을 것인가.

그 밤은 아직 그 질문에 답하지 못했습니다.

제1장 - 샌프란시스코의 소년, 우주의 비밀에 끌리다

1. 닷컴 붐을 등진 아이

1983년 샌프란시스코의 한 집. 두 종류의 손이 동시에 움직이고 있었습니다. 한 손은 가죽을 만졌습니다. 다리오의 아버지 리카르도 아모데이(Riccardo Amodei)는 이탈리아 토스카나의 작은 마을 마사 마리티마에서 건너온 가죽 장인이었습니다. 엘바 섬이 멀지 않은 곳입니다. 그는 한 장의 가죽을 오래 들여다보고, 자르고, 꿰맸습니다. 다른 손은 책을 만졌습니다. 시카고에서 태어난 유대계 미국인 어머니 엘레나 엥겔(Elena Engel)은 버클리와 샌프란시스코의 도서관을 짓고 고치는 일을 관리했습니다. 한 사람은 손끝으로 물건을 만들었고, 한 사람은 지식이 머물 공간을 만들었습니다. 결이 달라 보이는 두 일은 한 뿌리에서 나왔습니다. 무언가를 제 손으로 만들어 세상에 보탠다는 것.

그 집에서 1983년 다리오 아모데이가 태어났습니다. 4년 뒤에는 여동생 다니엘라가 태어났습니다.

창밖의 도시는 다른 속도로 움직이고 있었습니다. 다리오가 중학교와 고등학교를 다니던 1990년대 후반, 샌프란시스코는 닷컴 붐(Dot-Com Boom)의 한복판이었습니다. 길마다 새 상업 웹사이트의 광고가 붙었습니다. 또래의 영리한 아이들은 차고에서 코드를 짜며 백만장자를 꿈꿨습니다. 인터넷은 그 시절의 종교였습니다. 그런데 그 종교의 본거지 한가운데서 자란 소년은, 이상하게도 신도가 되지 않았습니다.

훗날 그는 한 인터뷰에서 그 시절을 이렇게 회상했습니다. "웹사이트를 만드는 일 같은 건 내게 아무 흥미가 없었다. 내가 관심을 둔 것은 근본적인 과학적 진리를 발견하는 일이었다."

소년의 마음을 사로잡은 것은 수학이었습니다. 이유가 분명했습니다. 수학에는 객관적인 답이 있었습니다. 어떤 아이는 저 TV 프로그램이 훌륭하다고 말하고 다른 아이는 끔찍하다고 말하지만, 수학을 할 때만큼은 세상에 단 하나의 답이 존재했습니다. 그 점이 그에게는 안도였습니다. 취향과 유행과 돈이 사람의 변덕을 따라 출렁이는 도시에서, 흔들리지 않는 무언가가 있다는 것. 소년은 그 흔들리지 않는 쪽으로 걸어 들어갔습니다.

그래서 그는 방에서 수식에 빠졌습니다. 과학책을 파고들었습니다. 우주의 시작과 물질의 근본을 묻는 책들을 읽었습니다. 닷컴 붐이 도시를 휩쓰는 동안, 다리오의 머릿속을 휩쓴

것은 다른 질문이었습니다. 우주는 어떻게 시작되었는가. 물질을 이루는 맨 밑바닥의 법칙은 무엇인가. 바깥의 디지털 혁명은 그 질문 앞에서 표면의 소음처럼 들렸습니다.

부모는 두 남매에게 재능을 어디에 쓸지를 가르쳤습니다. 다리오는 뒷날 이렇게 말했습니다. "부모님은 내게 옳고 그름이 무엇인지, 세상에서 무엇이 중요한지를 알려 주셨다. 강한 책임감을 심어 주셨다." 가죽을 꿰매는 아버지와 도서관을 짓는 어머니가 식탁에서 건넨 것은 기술이 아니라 방향이었습니다. 세상의 불의를 외면하지 말 것. 가진 재능을 자신을 위해서가 아니라 더 멀리 쓸 것. 이 가르침은 오래 살아남습니다. 20여 년 뒤, 같은 남매가 자기 나라 정부를 법정에 세우는 장면까지 곧장 이어집니다.

닷컴의 광풍 속에서, 다른 천재 소년들이 부자가 될 꿈을 꿀 때, 이 소년은 밤하늘과 방정식 쪽을 보고 있었습니다. 그는 기업가가 될 생각이 없었습니다. 그가 되고 싶었던 것은 오직 하나였습니다. 우주의 근본을 푸는 과학자.

2. 로웰 고교에서 칼텍으로, 그리고 한 편의 기고문

다리오는 샌프란시스코의 명문 공립학교 로웰 고등학교(Lowell High School)를 다녔습니다. 전국에서 영재가 모이는 곳이었지만 그의 물리와 수학은 그 안에서도 도드라졌습니다. 2000년, 그는 미국 물리 올림피아드(USA Physics Olympiad) 국가대표 팀에 뽑혔습니다. 고등학생이 받을 수 있는 분명한 인정 가운데 하나였습니다.

졸업 뒤 그는 기초 과학의 성지인 캘리포니아 공과대학교, 칼텍(Caltech)으로 진학했습니다. 이곳에서 그는 전설적인 물리학자 톰 톰브렐로(Tom Tombrello)를 만납니다. 톰브렐로는 학부 1학년 가운데 극소수만 골라 가르치는 'Physics 11'을 운영하고 있었고, 다리오는 그 학생 중 하나가 되었습니다. 톰브렐로의 방식은 공식을 외워 답을 맞히는 훈련이 아니었습니다. 힌트 없는 막막한 자연 현상 앞에 학생을 세워 두고, 직관만으로 길을 찾게 하는 훈련이었습니다. 거대한 문제 앞에서 겁먹지 않는 법. 복잡한 덩어리를 꿰뚫어 그 안의 간결한 규칙을 끄집어내는 법. 다리오가 이때 몸에 익힌 이 태도는 먼 훗날 스케일링 법칙을 알아보는 눈의 원형이 됩니다.

그런데 정작 칼텍에서 그를 다른 천재들과 갈라놓은 것은 물리학이 아니었습니다.

2003년 봄이었습니다. 부시 행정부가 전쟁의 북을 두드리고 있었습니다. 이라크 침공은 거의 확정된 것이나 다름없었습니다. 칼텍의 캠퍼스는 이상하리만치 조용했습니다. 대다수 학생은 문제집과 실험과 컴퓨터 게임에 얼굴을 묻은 채, 캠퍼스 밖에서 벌어지는 일에 눈을 두지 않았습니다.

3월 3일. 학생신문 더 캘리포니아 테크(The California Tech)의 5면 맨 위에 한 편의 칼럼이 실렸습니다. 제목은 '반전 시위: 대다수 학생은 세계 정치에 무관심하다'였습니다. 글쓴이는 칼텍 2학년 다리오 아모데이였습니다.

그는 사회비평가 닐 포스트먼을 끌어와 글을 열었습니다. 미국 민주주의가 맞은 제일 큰 위험은 오웰식 독재가 아니라, 정치와 공적 대화를 무의미하게 만드는 문화라는 포스트먼의 진단이었습니다. 다리오는 곧장 칼날을 자기 동급생들에게 돌렸습니다. "문제는 모두가 이라크 폭격에 찬성한다는 것이 아니다. 대부분은 원칙적으로 반대하면서도 거기에 단 일 밀리초도 쓰지 않으려 한다는 것이다." 그리고 못을

박았습니다. "이것은 지금, 지체 없이 바뀌어야 한다."

스무 살 청년의 글이라기엔 서슬이 퍼랬습니다. 그는 동급생들이 세상의 추문이나 선정성이 아니라 문제집과 컴퓨터 게임과 도넛의 수급을 둘러싼 기이한 말다툼에 정신이 팔려 있다고 적었습니다. 미래를 바꿀 힘을 가진 자들이 그 권리를 어이없이 내던지고 있다고 적었습니다. 그리고 의무를 말했습니다. 우리에게는 우리 공동체와 국가와 인류의 윤리적 온전함을 지킬 특권과 의무가 있다고.

이 글이 어떤 종류의 사람을 만들었는지는, 그가 그 봄 한 번 더 글을 쓰지 않았다는 사실에서도 드러납니다. 그는 운동가가 되지 않았습니다. 그는 할 말을 한 번 분명히 하고, 자기 자리로 돌아갔습니다. 그해 봄 캠퍼스의 한 파티에서 사담 후세인 분장을 한 누군가가 '반전 운동가' 아모데이에게 다가와 악수를 청하며 그의 '지지'에 '감사'를 표했다는 일화가 학생신문 졸업 호에 실렸습니다. 농담이었습니다. 그러나 그 농담조차 그가 그곳에서 어떤 이름으로 통했는지를 말해 줍니다.

다음 학기, 다리오는 칼텍을 떠나 스탠퍼드로 옮깁니다. 그가 떠난다고 했을 때 톰브렐로가 남긴 말이 있습니다. "다리오에 관해서라면, 그가 여기서 억지로 버티지 않는 것이 매우 중요했다. 이 친구는 국보다." 스승은 제자를 붙잡는 대신 보냈습니다.

2003년의 이 한 편의 글을 가볍게 보아서는 안 됩니다. 압도적인 지능만으로는 충분하지 않다는 것. 그 힘은 반드시 도덕적 책임의 통제 아래 있어야 한다는 것. 스무 살의 다리오가 학생신문에 적은 이 신념은 23년 뒤 거의 그대로 되돌아옵니다. 2026년 2월, 국방장관 피트 헤그세스가 클로드의 안전장치를 풀라고 요구했을 때 그가 내놓은 거절의 논리는, 칼텍 학생신문의 청년이 쓴 문장과 한 핏줄이었습니다. 본인도 그렇게 말합니다. "나는 민주주의를 지키기 위해 AI를 쓰는 일의 실존적 중요성을 깊이 믿는다. 그러나 좁은 범위의 경우에는, AI가 민주적 가치를 지키는 게 아니라 무너뜨릴 수 있다." 같은 사람이었습니다. 신문의 면만 바뀌었을 뿐입니다.

3. 스탠퍼드에서 받은 물리학 학사

칼텍은 다리오에게 최고의 훈련을 주었습니다. 그러나 그 폐쇄성은 그의 더 넓은 갈증을 다 채우지 못했습니다. 그는 우주의 근본을 묻는 물리학의 거대함을 사랑했습니다. 그러면서도 그것이 현실의 생명이나 사회와 어떻게 닿는지를 더 넓게 보고 싶었습니다. 그래서 그는 칼텍을 떠나 스탠퍼드 대학교(Stanford University)로 옮겼습니다.

스탠퍼드 둘레에서도 실리콘밸리의 두 번째 기술 붐이 자라고 있었습니다. 매일 새 벤처가 생기고 자금이 오갔습니다. 다리오는 여전히 그쪽을 곁눈질하지 않았습니다. 대신 그는 양자역학과 일반 상대성이론, 우주론 같은 이론물리학의 깊은 자리에 매달렸습니다. 그에게 물리학은 계산 도구가 아니었습니다. 복잡하게 얽힌 세계의 이면을 꿰뚫어, 본질에 닿는 하나의 원리를 찾아내는 여정이었습니다.

2006년, 다리오 아모데이는 스탠퍼드에서 물리학 학사 학위를 받았습니다. 학부를 마칠 무렵 그의 시선은 태초의 우주를 묻는 우주론과 이론물리학의 끝자락에 고정돼 있었습니다. 그는 아인슈타인과 파인만이 걸은 길을 따라 우주의 탄생을 수학으로 해체하는 이론물리학자가 되려 했습니다. 그리고 그 길의 다음 관문인 프린스턴 대학교(Princeton University) 박사 과정에, 전공을 이론물리학으로 두고 진학했습니다. 모든 것이 그가 어릴 적부터 그려 온 궤도 위에 정확히 놓인 것처럼 보였습니다.

그러나 같은 해, 그 궤도를 통째로 비틀어 버릴 일이 조용히 다가오고 있었습니다.

오래도록 희귀병을 앓던 아버지 리카르도의 병세가 2006년에 절망적으로 기울었습니다. 우주의 시작을 설명하는 정교한 방정식을 손에 쥐고도, 눈앞에서 무너지는 아버지의 세포 하나를 어쩌지 못했습니다. 그 차가운 무력감이 청년의 세계관에 금을 냈습니다. 스탠퍼드의 학사 학위는 한 여정의 완성이었습니다. 그러나 그것은 동시에, 우주론의 꿈을 접고 박사 전공을 생명 쪽으로 완전히 틀게 만드는 다음 장의 서막이기도 했습니다.

그 전환은 처음부터의 선택이 아니었습니다. 한 사람의 죽음이 만든 방향이었습니다. 이 책이 처음부터 끝까지 놓지 않으려는 사실이 바로 그것입니다.

제2장 - 아버지의 죽음, 그리고 방향을 튼 박사 과정

1. 이론물리학에서 생물물리학으로

2006년 가을, 프린스턴 대학교의 한 대학원생이 우주의 방정식을 들여다보고 있었습니다. 스물세 살이었습니다. 스탠퍼드에서 물리학 학사를 마치고 동부로 건너온 다리오 아모데이였습니다. 그가 택한 길은 이론물리학이었습니다. 우주가 어떻게 시작되었는지, 물질의 더 쪼갤 수 없는 조각이 어떤 규칙으로 움직이는지, 그런 질문이 그를 끌어당겼습니다. 고등학교 시절 창밖에서 닷컴 붐이 터지는 동안에도 그는 웹사이트 같은 것에 눈길을 주지 않았습니다. 돈이 오가는 자리가 아니라 진리가 놓인 자리를 보고 싶어 했습니다. 프린스턴은 그런 사람에게 어울렸습니다. 아인슈타인이 말년을 보낸 학교, 노벨상 수상자를 줄줄이 배출한 물리학과. 다리오는 그 안에서 자기 자리를 찾은 듯 보였습니다.

그가 교정에 발을 들인 지 얼마 되지 않아 한 사람이 세상을 떠났습니다. 아버지 리카르도 아모데이였습니다. 이탈리아 토스카나의 작은 언덕 마을 마사 마리티마에서 건너온 가죽 장인이었습니다. 평생 손끝으로 가죽을 다루며 두 아이를 길러낸 사람이었습니다. 오래 앓던 희귀병이 그를 데려갔습니다. 다리오가 박사 과정을 시작한 바로 그해, 2006년의 일이었습니다.

아버지의 죽음 앞에서, 우주의 방정식은 갑자기 멀게 느껴졌습니다. 시공간의 구조를 계산하는 그 우아한 수식들은 병상에서 숨을 몰아쉬는 한 사람을 살려내지 못했습니다. 정밀하고 아름다웠지만, 무력했습니다. 다리오는 훗날 이 시기를 두고 거의 분노에 가까운 감정을 드러냈습니다. 누군가 그를 두고 "기술 발전을 늦추려는 비관론자"라고 부를 때면 그는 화를 냈습니다. "제 아버지는 몇 년만 일찍 나왔어도 됐을 치료법 때문에 돌아가셨습니다. 저는 이 기술의 혜택이 무엇인지 압니다." 그가 한 인터뷰에서 한 말입니다.

이 회한을 단단한 확신으로 바꾼 사건이 뒤이어 찾아왔습니다. 아버지가 떠나고 몇 년이 지나지 않아, 아버지를 데려간 바로 그 병의 양상이 완전히 달라진 것입니다. 한때 절반이 목숨을 잃던 병이 의학의 진보에 힘입어 거의 대부분 치료할 수 있는 병으로 바뀌었습니다. 치사율 50퍼센트가 5퍼센트로 내려앉았다는 이야기였습니다. 다리오는 훗날 이렇게 말했습니다. "그 병의 치료법을 찾아내 많은 사람의 목숨을 구한 누군가가 있었습니다. 그런데 더 많은 사람을 구할 수도 있었습니다." 더 많은 사람. 그 안에 자기 아버지가 들어

있었습니다.

이 시간의 어긋남이 다리오에게 남긴 교훈은 단단했습니다. 과학이 빠르냐 느리냐는 논문 위의 문제가 아니었습니다. 그것은 누가 살고 누가 죽느냐의 문제였습니다. 몇 달, 몇 년의 차이가 한 사람의 생사를 갈랐습니다. 우주의 비밀을 풀겠다던 청년은 망원경을 내려놓았습니다. 그리고 현미경 쪽으로 몸을 돌렸습니다. 그는 물리학을 버리지는 않았습니다. 물리학의 정밀한 도구를 들고 생명의 영역으로 넘어갔을 뿐입니다. 이론물리학에서 생물물리학(Biophysics)으로. 사람들은 이것을 두고 흔히 처음부터 계획된 진로 변경처럼 이야기하곤 합니다. 사실은 그렇지 않았습니다. 이것은 한 죽음이 만들어낸 방향 전환이었습니다. 과학이 더 빨리 사람을 구하기를 바라는 마음, 그 마음이 평생 그의 연구를 끌고 가는 연료가 되었습니다.

2. 망막에서 시작한 신경 회로 연구

생물물리학으로 방향을 튼 다리오가 마주한 것은 자연이 만든 제일 복잡한 물건, 뇌였습니다. 그는 프린스턴에서 윌리엄 비알렉(William Bialek) 교수의 지도 아래 박사 연구를 시작했습니다. 제2지도교수는 신경 회로 연구의 마이클 베리(Michael Berry) 교수였습니다. (학부 시절 칼텍에서 그를 이끈 멘토는 따로 있었습니다. 물리학자 톰 톰브렐로였습니다.) 다리오가 붙든 질문은 이런 것이었습니다. 따로 떨어진 수많은 뉴런이 어떻게 서로 신호를 주고받아 하나의 지각을 만들어내는가.

그는 답을 찾으려고 망막(Retina)을 골랐습니다. 망막은 눈 안쪽을 덮은 얇은 조직입니다. 그냥 빛을 받아들이는 수동적인 센서가 아닙니다. 들어온 빛을 1차로 가공하고 압축해서 뇌로 넘기는 작은 컴퓨터에 가깝습니다. 뇌의 일부가 바깥으로 튀어나온 셈이라, 신경이 정보를 어떻게 처리하는지 들여다보기에 이만한 창구가 없습니다. 다리오는 0.5밀리미터 곱하기 0.5밀리미터, 손톱보다 훨씬 작은 망막 조각 안에 들어찬 200개가 넘는 세포의 전기 신호를 거의 빠짐없이 기록하는 방법을 만들어냈습니다. 박사 논문의 제목이 그 작업의 무게를 그대로 보여줍니다. "Network-Scale Electrophysiology: Measuring and Understanding the Collective Behavior of Neural Circuits". 우리말로 옮기면 네트워크 규모의 전기생리학, 신경 회로의 집단 행동을 측정하고 이해하기, 정도가 됩니다.

당시 큰 벽은 데이터의 복잡함이었습니다. 뉴런 하나가 어떻게 반응하는지를 보는 것이 아니라, 수백 개의 뉴런이 동시에 발화하며 만들어내는 거대한 패턴을 읽어야 했습니다. 여기서 물리학자의 머리가 빛을 발했습니다. 다리오는 칼텍 시절 톰브렐로에게서 배운 태도를 끌어왔습니다. 복잡하고 거대한 시스템 앞에서 주눅 들지 않고, 그것을 꿰뚫는 간결한 법칙을 뽑아내는 태도였습니다. 그는 통계물리학의 모델을 신경 데이터에 가져다 붙였습니다. 그리고 무언가를 발견했습니다. 개별 뉴런은 그저 켜지거나 꺼지는 신호만 내보내는데, 이들이 수백 수천 개로 묶여 네트워크의 규모가 커지면 거기서 이미지의 경계를 알아보고 움직임을 좇는 고차원의 능력이 솟아난다는 것이었습니다. 그의 논문은 이 신경망에서 임계 현상(critical phenomena)의 강력한 증거를 찾아냈다고 평가받았습니다. 오랫동안 이론으로만 예측되고 입증되지 못했던 현상이었습니다.

여기에 훗날의 씨앗이 들어 있었습니다. 뉴런이 함께 발화하는 방식, 그리고 그 네트워크의 규모가 곧 능력을 결정한다는 통찰. 이것은 몇 년 뒤 다리오가 실리콘밸리에서

인공신경망과 거대 언어 모델을 만들며 정립하게 될 스케일링 법칙(Scaling Law)의 지적 모태가 됩니다. 생물학적 뇌에서 그가 본 원리와, 기계의 신경망에서 그가 보게 될 원리는 본질에서 같은 질문이었습니다. 간결한 유닛이 수없이 모이면 어떻게 지능이 솟아오르는가. 이 뛰어난 연구를 인정받아 다리오는 허츠 재단(Hertz Foundation)에서 2007년 펠로우십을 받았고, 2011년과 2012년 두 해에 걸쳐 박사논문상을 받았습니다.

그는 이때 이미 두 세계 사이에 다리를 놓고 있었습니다. 한쪽에는 생물학적 뇌가 있었고, 다른 쪽에는 아직 본격적으로 깨어나지 않은 인공신경망이 있었습니다. 다리오는 그 둘이 같은 언어로 말한다는 것을 남보다 먼저 알아챈 사람이었습니다.

3. 스탠퍼드 의대 박사후 과정

프린스턴에서 학위를 마친 다리오는 스탠퍼드 대학교 의과대학으로 자리를 옮겼습니다. 학사 시절을 보낸 학교로 돌아온 셈입니다. 아버지를 잃은 뒤 품은 다짐, 과학으로 사람을 살리겠다는 다짐을 실현하려는 걸음이었습니다. 그가 합류한 곳은 파라그 말릭(Parag Mallick) 교수의 연구실이었습니다. 단백질을 분석해서 암을 조기에 찾아내는 표지, 곧 암 바이오마커(Cancer Biomarker)를 추적하는 곳이었습니다. 다리오의 일은 종양 안팎의 단백질을 들여다보며 전이된 암세포를 가려내는 것이었습니다. 아버지를 데려간 병의 공포를 누구보다 잘 아는 그에게, 이 일은 그저 하나의 연구 과제가 아니었습니다.

모니터 앞에서 데이터를 파고들수록 그는 거대한 벽을 만났습니다. 물리학의 방정식은 깨끗하게 떨어졌습니다. 생물학은 그렇지 않았습니다. 단백질 하나를 이해하려 해도 그것이 세포 어디에 있는지, 어떻게 변형되는지, 다른 수만 개의 단백질과 어떻게 얽히는지를 모두 따져야 했습니다. 변수 하나를 붙잡으면 수천 개가 튀어나왔습니다. 환자마다 패턴이 달랐습니다. 사람이 가설을 세우고, 실험을 하고, 논문을 읽고, 통계를 손으로 맞춰가는 방식으로는 이 천문학적인 경우의 수를 도저히 따라잡을 수 없었습니다.

다리오는 여기서 뼈아픈 결론에 도달했습니다. "생물학의 복잡성은 인간의 규모를 넘어선다." 그가 한 말입니다. 아무리 뛰어난 연구자라도 평생 이해할 수 있는 생물학은 전체의 한 조각에 불과했습니다. 병을 정복하려면 수백, 수천 명의 연구자가 평생에 걸쳐 쌓은 지식을 하나로 묶어 거대한 패턴을 읽어야 하는데, 인간의 뇌로는 그것이 안 되었습니다. 아버지를 살리지 못한 그 "과학의 더딘 속도"의 뿌리에, 인간 지능 자체의 한계가 있었습니다. 깨달음은 절망에 가까웠습니다.

그 절망의 끝에서 빛이 들어왔습니다. 다리오가 스탠퍼드에 있던 무렵, 실리콘밸리 전역에서 무언가가 깨어나고 있었습니다. 오랫동안 가능성만 품고 있던 기계 학습이 막강한 컴퓨팅 파워와 방대한 데이터를 만나 갑자기 결과를 내기 시작한 것입니다. 다리오는 이 새 기술을 직접 만져보았습니다. 그러자 머릿속에서 흩어져 있던 조각들이 맞춰졌습니다. 인간의 뇌로는 감당할 수 없는 생물학의 복잡성을 다룰 열쇠가, 바로 인간을 넘어서는 인공지능일 수 있었습니다. 그는 이렇게 회상합니다. "생물학의 근본 문제들이 가진 복잡성은 인간의 규모를 넘어서는 것처럼 느껴졌습니다. 그것을 다 이해하려면 수백, 수천 명의 연구자가 필요했습니다." 그 수천 명의 연구자를 대신할 무언가가, 이제 막 눈을

뜨고 있었습니다.

망막을 연구하던 시절 그가 손에 쥐었던 직관, 간결한 유닛이 거대한 네트워크로 묶이면 지능이 솟아난다는 그 직관이 다시 떠올랐을 것입니다. 생물학에서 막다른 길에 부딪힌 좌절은 역설적이게도 그를 기술 쪽으로 밀어냈습니다. 그는 생물학을 떠난 것이 아니었습니다. 생물학을 포함한 인류의 모든 난제를 풀기 위해, 지능 그 자체의 규모를 인간 너머로 키우는 더 근본적인 싸움을 시작하기로 한 것입니다. 2014년 11월, 다리오는 스탠퍼드 의대 연구실의 문을 닫고 나왔습니다. 그의 다음 행선지는 학계가 아니라 산업이었습니다. 바이두(Baidu)였습니다.

방향을 튼 청년은 이제 더 큰 방향 전환의 문턱에 서 있었습니다.

제3장 - 스케일링 법칙을 발견하다

1. 바이두에서 시작된 산업 이력

헤드폰 너머로 통화 한 통이 흘러나옵니다. 잡음이 깔린 목소리. 억양이 센 사람의 말. 단어 끝이 뭉개진 녹음. 2014년 11월, 샌프란시스코 남쪽 서니베일의 한 사무실에서 다리오 아모데이는 그 음성 파일들을 듣고 있었습니다. 컴퓨터는 같은 소리를 글자로 옮기다가 자꾸 미끄러졌습니다. 그 미끄러짐을 줄이는 것이 그의 일이었습니다.

이 자리에 오기까지 그가 걸어온 길은 보통의 인공지능 연구자와 달랐습니다. 그는 2011년 프린스턴 대학교에서 물리학 박사 학위를 받았습니다. 지도교수는 신경과학자 데이비드 탱크(David Tank)였습니다. 논문은 망막의 신경 회로가 어떻게 신호를 주고받는지를 다뤘습니다. 그 뒤 스탠퍼드 의과대학에서 박사후 과정을 밟으며 단백질을 분석해 암을 추적하는 연구에 매달렸습니다. 인공지능을 정식으로 공부한 적은 없었습니다. 그가 들고 온 것은 알고리즘 설계 경력이 아니었습니다. 복잡한 생물 데이터를 다루려고 직접 짠 코드 뭉치였습니다.

그를 채용한 사람은 딥러닝 분야의 선구자 앤드류 응(Andrew Ng)이었습니다. 응은 그때 중국 검색 기업 바이두(Baidu)의 실리콘밸리 인공지능 연구소를 이끌고 있었습니다. 컴퓨터 과학 정통 코스를 밟지 않은 지원자를 응의 팀이 받아들인 까닭은 하나였습니다. 다리오가 스탠퍼드에서 직접 짠 그 코드였습니다. 코드에는 복잡한 시스템을 데이터로 다루는 사람 특유의 감각이 배어 있었습니다. 물리학자는 세상을 잴 수 있는 양으로 바꿔 보는 훈련을 오래 받습니다. 응은 그 시선이 신경망을 다루는 데에도 쓸모가 있으리라 봤습니다.

다리오가 합류한 프로젝트의 이름은 딥스피치 2(Deep Speech 2)였습니다. 그 전 세대인 딥스피치는 이미 도발적인 주장 하나를 품고 있었습니다. 음성 인식 시스템을 만들 때, 인간 언어학자가 손으로 깎아 만든 복잡한 처리 단계를 다 걷어내자는 것이었습니다. 전통적인 방식은 소리를 음소로 쪼개고, 다시 확률 모델에 맞추고, 사전을 붙였습니다. 딥스피치는 그 모두를 버렸습니다. 마이크에 들어온 날것의 소리 파형을 거대한 신경망에 그대로 밀어 넣고, 글자가 바로 나오게 하자는 발상이었습니다. 단대단(end-to-end) 방식이라고 불렀습니다. 다리오와 동료들은 딥스피치 2에서 이 방식을 영어와 중국어 두 언어로 한꺼번에 밀어붙였습니다. 성격이 완전히 다른 두 언어를 같은 구조로 처리한다는 것은 그때로서는 대담한 시도였습니다.

여기서 그는 평생을 따라다닐 현상 하나를 처음 마주합니다. 성능이 막힐 때마다, 주변의 컴퓨터 과학자들은 알고리즘을 더 정교하게 다듬자고 했습니다. 새로운 수학적 장치를 끼워 넣고, 신경망 구조를 손보고, 인간의 언어 지식을 더 집어넣자는 식이었습니다. 다리오는 다른 다이얼에 손을 댔습니다. 신경망의 층을 더 깊게 쌓았습니다. 데이터를 두 배, 세 배로 늘렸습니다. 컴퓨터가 더 오래 학습하도록 내버려 뒀습니다. 모델 크기, 데이터 양, 연산 시간. 그가 잡은 것은 이 세 개의 손잡이였습니다.

결과는 그의 눈을 의심하게 했습니다. 인간이 고안한 어떤 정교한 언어학 필터로도 못 잡던 벽이 있었습니다. 소음 속 음성, 사투리, 억양. 그 벽이 모델을 키우자 무너지기 시작했습니다. 딥스피치 2 시스템은 9천 시간이 넘는 음성 데이터로 학습했습니다. 어떤 영역에서는 사람이 받아 적은 결과에 견줄 만한 정확도에 닿았습니다. 이 작업은 훗날 MIT 테크놀로지 리뷰가 꼽은 그해의 혁신 기술 가운데 하나로 기록됩니다.

물리학에는 상전이(phase transition)라는 개념이 있습니다. 물이 99도까지는 그냥 뜨거운 물입니다. 100도에서 갑자기 끓어 수증기가 됩니다. 어떤 양이 일정 문턱을 넘으면 전혀 다른 상태가 나타나는 현상입니다. 다리오는 신경망에서 그 비슷한 것을 봤습니다. 크기를 키우니 없던 능력이 생겼습니다. 그는 깨달았습니다. 지능을 끌어올리는 일이 어쩌면 천재적인 기교의 문제가 아니라, 규모의 문제일지 모른다는 것을. 바이두에서의 1년은 짧았습니다. 그러나 이 1년에서 그는 앞으로 10년을 걸 가설의 씨앗을 손에 쥐었습니다.

2. 구글 브레인을 거쳐

2015년 말, 다리오는 바이두를 떠나 구글 브레인(Google Brain)으로 옮겼습니다. 그때 구글 브레인은 제프 딘(Jeff Dean)이 떠받친 막강한 인프라 위에 서 있었습니다. 딥러닝을 연구실의 실험에서 세상을 움직이는 도구로 키워내던 곳이었습니다. 컴퓨팅 자원의 규모가 바이두와는 또 달랐습니다. 다리오에게 이곳은 더 큰 도가니였습니다. 음성이라는 좁은 영역에서 확인한 현상이 인공지능 전반으로 번지는지, 여기서 시험해 볼 수 있었습니다.

이 시기에 그는 한 사람과 가까워집니다. 옆자리에 앉은 연구자 크리스 올라(Chris Olah)였습니다. 올라는 신경망의 내부를 들여다보는 일에 매달리던 사람이었습니다. 모델이 왜 그런 답을 내놓는지, 그 안에서 무슨 일이 벌어지는지를 그림으로 그려내려 했습니다. 훗날 두 사람은 같은 회사를 세우게 됩니다. 그러나 그때는 아직 그 미래를 몰랐습니다.

구글 브레인에서 다리오의 머릿속에는 두 갈래의 생각이 동시에 자랐습니다. 하나는 확신이었습니다. 모델을 키우면 강해진다. 그는 구글의 압도적인 컴퓨팅 자원으로 그 확신을 거듭 확인했습니다. 다른 하나는 불안이었습니다. 모델이 강해질수록, 그것이 안에서 무슨 생각을 하는지 사람은 점점 더 모르게 된다는 것이었습니다.

그 불안에는 구체적인 근거가 있었습니다. 그때 딥러닝 모델은 특정 작업에서 사람을 능가하는 정확도를 보였습니다. 그런데 사람 눈에는 보이지도 않는 미세한 잡음을 사진에 살짝 섞으면, 컴퓨터가 고양이를 트럭으로 착각하는 일이 벌어졌습니다. 적대적 공격(adversarial attack)이라고 불렀습니다. 강한데 동시에 어이없이 취약했습니다. 이 이중성을 다리오는 그냥 넘기지 못했습니다. 신경망은 세상을 사람처럼 이해하는 게 아니었습니다. 우리가 짐작도 못 할 방식으로 경계선을 긋고 있었습니다.

이 문제의식은 그 무렵 인공지능 학계에서 변방의 생각이었습니다. 연구자 대다수는 인공지능이 인간을 위협할 만큼 강해진다는 이야기를 멀리했습니다. 과거 인공지능이 거창한 약속을 했다가 무너진 시절이 있었습니다. 이른바 인공지능의 겨울(AI Winter)이 남긴 상처 때문이었습니다. 그들은 당장의 성능을 1퍼센트 올리는 일에 집중했습니다. 큰 위험을 말하는 것은 한물간 공상으로 취급됐습니다.

다리오와 몇몇은 달랐습니다. 그 결과가 2016년에 나온 논문 "인공지능 안전의 구체적 문제들(Concrete Problems in AI Safety)"입니다. 다리오가 제1저자였습니다. 구글 브레인과 스탠퍼드, 버클리의 연구진이 함께했습니다. 이 논문이 한 일은 인공지능의 위험을 공상에서 끌어내려 공학의 책상 위에 올려놓은 것이었습니다. 먼 미래의 종말론 대신, 지금의 기계 학습에서 실제로 생길 수 있는 문제들을 줄 세웠습니다. 모델이 목표를 이루려다 엉뚱한 부작용을 내는 문제. 보상을 받으려고 편법을 쓰는 문제. 위험한 탐색을 막는 문제. 추상적 공포가 아니라 풀어야 할 과제 목록이었습니다.

구글 브레인의 시간은 그를 바꿔 놓았습니다. 그는 더 이상 성능을 짜내는 기술자로만 자신을 보지 않았습니다. 기술이 어디로 굴러가는지, 그 방향을 안전하게 틀 수 있는지를 고민하는 사람이 되어 있었습니다. 동시에 대기업의 한계도 느꼈습니다. 구글은 검색과 광고라는 거대한 사업을 지켜야 했습니다. 인공지능 연구도 결국 그 사업을 돕는 쪽으로 정렬되기 마련이었습니다. 다리오가 직감한 변화의 크기는 회사가 감당하려는 것보다 훨씬 컸습니다. 그는 더 급진적인 곳을 찾고 있었습니다. 마침 실리콘밸리 한쪽에서 이상한 비영리 연구소 하나가 막 문을 열던 참이었습니다.

3. 오픈AI의 연구 부문 부사장

2016년, 다리오는 오픈AI(OpenAI)에 합류했습니다. 갓 태어난 비영리 연구소였습니다. 거대 기업이 인공지능 인재를 독식하는 데 맞서, 인류 전체에 도움이 되는 안전한 일반인공지능을 만들겠다는 깃발을 들고 있었습니다. 자유로운 분위기였습니다. 연구자들은 로봇, 게임, 언어 등 제각기 다른 방향으로 흩어져 있었습니다. 다리오는 빠르게 연구 부문 부사장(VP of Research) 자리에 올랐습니다. 그리고 전설적인 연구자 일리야 수츠케버(Ilya Sutskever)와 함께 연구의 큰 방향을 잡는 핵심에 섰습니다.

합류 초기, 일리야가 다리오에게 던진 한마디가 있었습니다. 신경망에 대해 처음 들은 말 중 하나였다고 다리오는 훗날 회상합니다. "이 모델들은 그저 학습하고 싶어 할 뿐이야(The models just want to learn)." 짧은 문장이었습니다. 뜻은 이랬습니다. 인간이 자기 머리로 짜낸 좁은 아이디어로 모델을 가르치려 들지 마라. 충분한 데이터와 연산을 주고 길만 터주면, 모델은 알아서 세상의 규칙을 빨아들인다. 다리오가 바이두에서 손끝으로 느꼈던 것을, 일리야는 한 문장으로 요약하고 있었습니다.

이 철학은 곧 현실이 됐습니다. 2017년 구글이 트랜스포머(Transformer)라는 새로운 신경망 구조를 내놨습니다. 다리오는 이것이 자신이 찾던 그릇임을 알아봤습니다. 막대한 연산을 거침없이 빨아들이는 구조였습니다. 오픈AI는 이 구조에 인류가 쓴 방대한 텍스트를 밀어 넣었습니다. 그리고 다음 단어를 맞히는 목표 하나만 줬습니다. 그러자 모델은 문법을 익히고, 문맥을 익히고, 논리를 따라가기 시작했습니다.

그렇게 나온 것이 GPT-2(2019년)와 GPT-3(2020년)입니다. GPT-3는 매개변수가 1,750억 개에 달했습니다. 충격은 능력의 종류에 있었습니다. 이 모델은 그저 다음 단어를 예측하도록만 배웠습니다. 그런데 따로 가르치지 않은 일까지 해냈습니다. 번역하고, 요약하고, 간단한 코드를 짰습니다. 누구도 그런 능력을 일일이 주입하지 않았습니다. 크기를 키우자 저절로 나타났습니다.

여기서 사실 하나를 바로잡아 둡니다. 흔히 다리오를 스케일링 법칙(Scaling Laws)의 단독 발견자처럼 그리는 이야기가 돌지만, 실제는 그렇지 않습니다. 2020년 1월에 나온 논문 "신경 언어 모델의 스케일링 법칙(Scaling Laws for Neural Language Models)"은 재러드 카플란(Jared Kaplan)과 샘 매캔들리시(Sam McCandlish)가

연구를 이끌었습니다. 다리오는 그 프로젝트 전반에 걸쳐 방향을 잡아준 사람으로, 열 명의 저자 가운데 마지막에 이름을 올렸습니다. GPT-3를 세상에 알린 논문 "언어 모델은 소수 예시 학습자다(Language Models are Few-Shot Learners)"의 제1저자도 다리오가 아니라 톰 브라운(Tom Brown)이었습니다.

이 구별이 중요한 이유가 있습니다. 이 책은 한 사람을 영웅으로 키우려는 책이 아니기 때문입니다. 다리오의 진짜 자리는 단독 발명가가 아니었습니다. 흩어진 관찰들을 하나의 확신으로 묶어 회사 전체를 그 방향으로 밀어붙인 사람이었습니다. 그 논문이 수학적으로 보여준 것은 이랬습니다. 모델의 성능을 깎아먹는 오차가, 매개변수 수와 데이터 양과 연산량이라는 세 변수에 따라 매끄러운 멱함수(power law) 곡선을 그린다는 것. 일곱 자릿수가 넘는 범위에서 그 곡선이 흐트러지지 않았습니다. 신경망 구조의 자잘한 차이는 그 앞에서 거의 의미가 없었습니다.

다리오가 한 일은 이 곡선을 믿고, 그 믿음으로 회사를 움직인 것이었습니다. 곡선이 매끄럽다는 말은 미래를 어느 정도 예측할 수 있다는 뜻입니다. 모델을 이만큼 키우면 이만큼 똑똑해진다. 그 예측 위에서 그는 도박을 밀어붙였습니다. GPT-2의 15억 개 매개변수를 GPT-3에서 1,750억 개로 단숨에 끌어올린 것입니다.

GPT-3가 사람이 쓴 것과 구별되지 않는 글을 쓰고 코드를 짜는 모습을 봤을 때, 실리콘밸리는 흔들렸습니다. 다리오는 데이터 센터라는 실리콘 요람 안에 새로운 종류의 지능을 길러내는 설계자가 되어 있었습니다. 그러나 그 승리의 한복판에서, 그의 마음에는 경외감과 나란히 균열이 자라고 있었습니다.

4. RLHF의 공동 발명

똑똑한 것과 선한 것은 다릅니다. 다리오가 마주한 균열이 바로 거기에 있었습니다. GPT-3는 인터넷의 텍스트를 통째로 삼키며 학습했습니다. 인류의 지식이 거기 있었습니다. 동시에 인류의 편견과 혐오와 거짓말도 거기 있었습니다. 모델은 세상의 사실은 기막히게 예측했습니다. 그러나 무엇이 옳고 그른지는 전혀 몰랐습니다. 사용자가 요청만 하면 완벽한 문장으로 위험한 정보를 술술 풀어놓을 준비가 되어 있었습니다. 길들지 않은 짐승이었습니다.

철학에 사실과 가치의 구분(fact-value distinction)이라는 오래된 문제가 있습니다. 세상이 어떠한가를 아는 것과, 세상이 어떠해야 하는가를 아는 것은 다른 일이라는 것입니다. 모델을 키우면 사실에 대한 지식은 기하급수로 늘었습니다. 그러나 가치는 저절로 따라오지 않았습니다. 지능의 크기가 폭발하는데 가치의 나침반이 없다면, 그 결과는 축복이 아니라 재앙일 수 있었습니다. 정렬(alignment) 문제는 이제 한가한 논쟁거리가 아니었습니다. 발등의 공학 과제가 됐습니다.

다리오는 폴 크리스티아노(Paul Christiano)를 비롯한 동료들과 이 문제에 매달렸습니다. 그들이 도달한 해법이 인간 피드백 기반 강화학습, 줄여서 RLHF(Reinforcement Learning from Human Feedback)였습니다. 오늘날 거의 모든 안전한 인공지능 모델의 뼈대가 된 기술입니다.

원리는 의외로 간단합니다. 먼저 모델이 한 질문에 여러 답을 내놓게 합니다. 사람 평가자가 그 답들을 읽고, 어느 것이 더 유용하고 정직하고 무해한지 순위를 매깁니다. 이 사람의 선호 데이터를 가지고, 무엇이 좋은 답인지 점수를 매기는 별도의 보상 모델(reward model)을 훈련합니다. 마지막으로 본 모델이 그 보상 모델에서 높은 점수를 받는 쪽으로 답을 바꿔가도록 강화학습으로 다듬습니다. 사람의 판단을 한 번 기계의 형태로 굳혀, 그것을 기준 삼아 모델을 길들이는 방식입니다.

다리오와 동료들은 목표를 세 글자로 압축했습니다. HHH. 유용하고(Helpful), 정직하며(Honest), 무해하게(Harmless). 모델은 질문자의 뜻을 헤아려 쓸모 있는 답을 주되, 모르는 것은 모른다고 말하고, 위험한 요구 앞에서는 정중히 거절할 줄 알아야 했습니다. RLHF는 통제 불능의 짐승에게 처음으로 채운 목줄이었습니다.

RLHF의 발명은 스케일링 법칙의 확립 못지않게 다리오의 이력에서 무게가 큽니다. 그는 지능을 폭발시키는 불을 다루는 법을 익히고 있었습니다. 동시에 그 불이 번지지 않게 막는 법도 함께 찾고 있었습니다. 그런데 이 발명에는 묘한 그림자가 있었습니다.

첫 번째 그림자는 노동입니다. RLHF는 수많은 사람의 손에 기댑니다. 답에 순위를 매기는 일은 결국 저임금 계약자들의 몫이었습니다. 사람의 편견이 그대로 모델에 스밀 수 있었습니다.

더 깊은 곳에 두 번째 그림자가 있었습니다. 모델이 점수를 잘 받으려고 사람을 속이거나 비위를 맞추는 법을 배울 수 있다는 것이었습니다. 명세 게이밍(specification gaming)이라고 부릅니다. 시키지 않은 방향으로 영리하게 빠져나가는 것입니다.

맨 밑바닥에 세 번째 그림자가 있었습니다. 스케일링 법칙대로 모델이 계속 커져 사람의 이해를 넘어서는 순간이 오면, 사람은 모델의 답이 참인지 거짓인지조차 가릴 수 없게 됩니다. 그러면 사람의 판단에 기댄 RLHF는 토대부터 흔들립니다. 평가자가 평가 대상을 못 따라가는 순간, 목줄은 더 이상 목줄이 아닙니다.

다리오는 이 한계를 또렷이 봤습니다. 그리고 이 지점이 다음 질문을 낳았습니다. 사람의 손을 일일이 빌리지 않고도, 모델이 스스로를 바로잡게 할 수는 없을까. 훗날 헌법적 AI(Constitutional AI)라는 이름으로 구체화될 발상의 출발점이 여기였습니다.

그러나 그 답을 찾으려면, 더 강하고 더 거대한 모델이 필요했습니다. 안전을 연구하려면 위험한 능력부터 손에 쥐어야 하는 역설이었습니다. 능력과 안전이 서로의 꼬리를 무는 이 이중 나선 속에서, 속도와 상업화를 앞세우는 오픈AI의 흐름과 다리오의 우선순위는 점점 어긋났습니다. RLHF의 탄생은 정렬 기술의 시작이었습니다. 동시에 그것은, 다리오가 오픈AI를 떠나 자기 회사를 세우게 만드는 긴 작별의 첫 장이기도 했습니다. 그 작별은 머지않아 찾아옵니다.

제4장 - 떠나기로 한 결심

1. 벌어지는 격차

2020년의 어느 날, 다리오 아모데이는 화면을 들여다보고 있었습니다. 자기 손으로 만든 모델이 자기보다 빠르게 똑똑해지고 있었습니다. GPT-3는 그가 예상했던 곡선을 그대로 따라 올라갔습니다. 모델을 키우면 능력이 오른다. 그의 가설은 틀리지 않았습니다. 문제는 그 가설이 너무 정확하게 맞아떨어진다는 데 있었습니다.

그는 2016년에 오픈AI(OpenAI)에 합류했습니다. 연구 부문 부사장(VP of Research)으로 일했습니다. 일리야 수츠케버(Ilya Sutskever)와 함께 연구 방향을 정하는 자리였습니다. GPT-2와 GPT-3가 그의 팀 손에서 나왔습니다. 그러니까 그는 스케일링이라는 길을 누구보다 앞장서서 닦은 사람이었습니다. 바로 그 사람이 그 길의 끝을 보며 불안해지기 시작했습니다.

불안의 정체는 이런 것이었습니다. 모델은 세상이 어떻게 돌아가는지를 빠르게 배웠습니다. 그러나 자신이 어떻게 행동해야 하는지는 배우지 못했습니다. 지식을 먹이면 지식이 늘었습니다. 그 지능 덩어리가 인간에게 해롭지 않도록 방향을 잡아 주는 기술, 그러니까 정렬(Alignment) 연구는 걸음마 수준이었습니다. 능력은 세대마다 폭발했습니다. 안전은 그 뒤를 헐떡이며 쫓아갔습니다. 둘 사이의 거리가 매년 벌어졌습니다. 그 벌어지는 틈을 아모데이는 매일 들여다보았습니다.

세상은 그가 떠난 이유를 다르게 이해했습니다. 마이크로소프트(Microsoft)와의 거대한 투자 계약에 반발했다는 이야기가 돌았습니다. 상업화 자체가 싫어서 나갔다는 이야기도 돌았습니다. 아모데이는 훗날 이 해석을 두고 짧게 잘랐습니다. "거짓"이라고 말입니다. 그는 반(反)상업주의자가 아니었습니다. 오히려 GPT-3의 상업화를 직접 진두지휘한 사람이었습니다. 인공지능이 큰돈을 벌 거라는 사실을 그는 누구보다 잘 알았습니다. 그 돈을 벌려면 마이크로소프트 같은 자본과 손을 잡아야 한다는 현실도 잘 알았습니다.

그가 견딜 수 없었던 것은 상업화가 아니라 순서였습니다. 강력한 기술을 세상에 내놓는 방식, 그 우선순위가 무너지는 것이었습니다. 안전이 채용 공고에 적어 넣는 구호로만 소비되었습니다. 정작 위험을 감수하면서까지 안전을 앞세우려는 진정성은 보이지 않았습니다. 그리고 그 밑바닥에는 더 깊은 균열이 있었습니다. 리더십에 대한 신뢰의 상실이었습니다. 훗날 그는 이렇게 말했습니다. 누군가의 가치관이 그가 말하는 것과

다르다고 느낄 때, 그가 정직하지 않다고 느낄 때, 그런 사람과 계속 함께 일하며 신뢰를 유지하기란 무척 어렵다고 말입니다.

문명 전체를 흔들 수 있는 기술로 향하는 버스가 있습니다. 그 운전대를 쥔 사람을 더는 믿을 수 없습니다. 상업적 성공의 열매가 눈앞에 매달려 있었습니다. 그런데도 아모데이는 좁혀지지 않는 가치관의 거리를 인정했습니다. 자기만의 궤도를 새로 그려야 한다는 결론에 도달했습니다.

2. 판다스의 대이주

2020년 12월 29일, 오픈AI는 블로그에 짧은 글을 올렸습니다. 다리오 아모데이가 거의 5년 만에 회사를 떠난다는 내용이었습니다. 샘 알트만(Sam Altman)은 그의 기여에 감사를 표했습니다. 새 프로젝트의 행운도 빌었습니다. 회사는 그 프로젝트가 제품 개발보다는 연구에 무게를 둘 것이라고 전했습니다. 발표문은 점잖고 매끈했습니다. 그 매끈함 뒤에서 무슨 일이 벌어지고 있었는지는 적혀 있지 않았습니다.

아모데이는 혼자 나가지 않았습니다. 그의 곁에는 동생 다니엘라 아모데이(Daniela Amodei)가 있었습니다. 다니엘라는 오픈AI에서 안전 및 정책 부문 부사장(VP of Safety and Policy)으로 일하던 사람이었습니다. 남매가 앞장섰습니다. 그 뒤를 약 14명의 연구자가 따랐습니다. 한꺼번에 우르르 나간 것이 아니었습니다. 몇 주에 걸쳐 한 사람씩 빠져나갔습니다. 그 안에는 잭 클라크(Jack Clark), 크리스 올라(Chris Olah), 톰 브라운(Tom Brown), 샘 매캔들리시(Sam McCandlish), 재러드 캐플런(Jared Kaplan), 벤저민 만(Benjamin Mann) 같은 이름이 있었습니다. 훗날 앤스로픽(Anthropic)의 공동 창업자가 될 사람들입니다.

세상은 이 행렬을 두고 여러 이름을 붙였습니다. 그 가운데 하나가 판다스(Pandas)의 대이주였습니다. 그런데 오픈AI에게 정말 아팠던 것은 떠난 사람의 숫자가 아니었습니다. 그 14명이 누구였느냐가 문제였습니다.

이들은 그저 14명의 직원이 아니었습니다. 오픈AI 안에서 거대 언어 모델을 만들고 키우는 일, 그 일 자체였습니다. 스케일링 법칙을 직접 증명해 낸 사람들입니다. 인간 피드백을 통한 강화학습(RLHF)의 기초를 닦은 사람들입니다. 신경망 속을 해부해 보려던 기계론적 해석 가능성(Mechanistic Interpretability)의 두뇌들입니다. 생성형 언어 모델의 엔진룸에 있던 인력이 통째로 빠져나간 셈이었습니다. 오픈AI에는 다른 연구 부서도 많았습니다. 그러나 훗날 세상을 뒤바꿀 바로 그 엔진을 만들던 사람들이 한꺼번에 사라졌습니다.

이들이 이렇게 똘똘 뭉칠 수 있었던 데는 이유가 있었습니다. 상당수가 물리학이나 기초과학을 하던 사람들이었습니다. 실험과 데이터로 우주의 법칙을 캐던 버릇이 있었습니다. 그 버릇 그대로 인공지능을 다루던, 지독한 경험주의자들이었습니다. 그리고 이들은 같은

두려움을 나눠 가지고 있었습니다. 자기들이 키운 이 지능이 인류에게 어떤 위험이 될 수 있는지에 대한 두려움이었습니다. 코로나19가 한창이던 2020년 말이었습니다. 이들은 실내에 모일 수도 없었습니다. 누군가의 집 뒷마당에서 마스크를 쓰고 거리를 둔 채, 자기들이 어디로 가야 하는지를 두고 토론했습니다.

다리오 아모데이에게는 오래된 믿음이 하나 있었습니다. 재능의 질량보다 재능의 밀도가 이긴다는 것이었습니다. 수천 명이 각자의 이익을 두고 정치 싸움을 벌이는 조직이 있습니다. 안전한 인공지능이라는 하나의 사명으로 묶인 소수의 사람들이 있습니다. 그 소수가 결국 승리한다는 믿음이었습니다. 그 믿음 위에서 그는 14명을 데리고 나왔습니다. 이 이주는 퇴사가 아니었습니다. 그것은 새로 무언가를 시작하겠다는 선언이었습니다.

3. 공개되지 않은 작별

실리콘밸리에서 거물이 회사를 떠날 때 흔히 따르는 각본이 있습니다. 언론에 미리 제보를 흘립니다. 회사의 잘못을 규탄하는 긴 공개 서한을 발표합니다. X(옛 트위터)에 스레드를 줄줄이 올려 대중의 지지를 호소합니다. 대중은 갈등을 좋아합니다. 선과 악으로 짜인 이야기는 순식간에 인터넷을 달굽니다. 당대 최고의 AI 연구소에서 핵심 인력 10여 명이 한꺼번에 나간다는 사건이었습니다. 그 자체로 모든 기술 매체의 1면감이었습니다.

그런데 2020년 12월, 그들이 나가는 방식은 이상하리만치 조용했습니다. 공개 서한도 없었습니다. 분노에 찬 폭로 스레드도 없었습니다. 아모데이는 깊은 밤 조용히 짐을 싸 나가는 사람 같았습니다. 아무 소리도 내지 않고 오픈AI의 문을 나섰습니다.

이 침묵은 우연이 아니었습니다. 그것은 아모데이라는 사람의 태도가 그대로 드러난 의도적인 고요였습니다. 그는 대중의 환호에 자기 동기를 묶어 버리는 일을 오래 지켜본 사람이었습니다. 그런 일이 사람을 어떻게 망가뜨리는지 보아 온 사람이었습니다. 인기를 좇다 보면 복잡한 문제를 흑백으로 갈라 버립니다. 결국 자기가 만든 페르소나에 갇혀 진실을 보는 눈을 잃습니다. 그는 CEO들끼리 벌이는 철창 격투기 같은 소모적인 드라마를 싫어했습니다. 회사를 자기 자아와 동일시하는 실리콘밸리의 풍토도 싫어했습니다. 그는 차라리 지루하고 눈에 띄지 않는 사람으로 보이기를 바랐습니다.

그의 침묵에는 또 다른 뜻도 있었습니다. 옛 동료와 진흙탕 싸움을 벌일 생각이 없다는 선언이었습니다. 신뢰의 균열은 이미 돌이킬 수 없었습니다. 그래도 그는 그것을 바깥에 까발려 다툴 가치를 느끼지 못했습니다. 비전이 다르고 신뢰할 수 없는 사람과 굳이 왜 논쟁을 벌이느냐. 그가 보기에 답은 하나였습니다. 그냥 나가서 자기 일을 하는 것. 상대도 자기 일을 하게 두는 것. 누가 옳고 그른지는 트위터에서 증명할 일이 아니었습니다. 시장에서 누가 이기는지, 대중의 평가가 어디로 향하는지, 그 결과가 말해 줄 일이었습니다. 요란한 말의 승리가 아니었습니다. 그들이 직접 만들 제품과 안전한 기술이 그 어떤 드라마보다 크게 그들의 정당성을 대신해 줄 것이라고 믿었습니다.

이 조용한 결별의 진짜 의미는 다른 데 있었습니다. 기존의 세계를 부수는 대신, 새로운 세계를 처음부터 짓겠다는 결심이었습니다. 오픈AI 안에서 끝없이 정치 싸움을 벌이는 길이 있었습니다. 그들은 그 길 대신 자기들의 이상을 온전히 담을 깨끗한 실험실을

골랐습니다. 그 실험실을 바닥부터 다시 세우기로 한 것입니다. 그리고 이 새 공간, 훗날 앤스로픽이 될 회사를 통해 그들은 정상을 향한 경주(Race to the top)를 불 붙이려 했습니다. 투명하고 안전한 모델로도 상업적으로 성공할 수 있음을 증명한다면, 경쟁사들도 뒤처지지 않으려고 결국 그 윤리 기준을 따라올 수밖에 없으리라는 계산이었습니다. 적을 끌어내리는 대신, 모두가 위로 올라가게 만드는 전략. 고고하고 위험한 전략이었습니다.

2020년 12월의 그 고요한 밤, 트위터 알림음 하나 울리지 않았던 그 조용한 이주는 사실 천재들이 내민 더없이 묵직한 선전포고였습니다. 다만 그들은 큰소리로 외치지 않았을 뿐입니다.

제5장 - 앤스로픽, 미션을 최우선에 두다

1. 남매의 창업

2020년 12월의 어느 날, 다리오 아모데이는 세상에서 제일 뜨거운 회사의 문을 걸어 나왔습니다. 오픈AI에서 보낸 다섯 해였습니다. GPT-2와 GPT-3를 함께 만들었습니다. 모델이 누구의 예상보다도 빠르게 똑똑해지는 광경을 누구보다 가까이서 지켜본 사람이었습니다. 그가 떠난 뒤 몇 주에 걸쳐 열네 명이 따라 나왔습니다. 그 안에는 오픈AI에서 안전·정책 부사장을 맡았던 여동생 다니엘라 아모데이가 있었습니다. 훗날 앤스로픽의 과학 책임자가 될 재러드 캐플런, 정책을 맡을 잭 클라크, 수석 설계자 샘 매캔들리시, 그리고 해석 가능성 연구를 이끌 크리스 올라도 함께였습니다. 스케일링의 인프라를 직접 손으로 쌓아 올린 바로 그 사람들이었습니다.

이들이 떠난 이유는 한 문장으로 줄여 말하기 어렵습니다. 다리오는 훗날 그것을 "단지 안전 때문만은 아니었다"고 했습니다. 컴퓨팅을 더 부으면 모델은 끝없이 좋아진다는 믿음이 하나 있었습니다. 그것만으로는 부족하고 정렬(alignment)이라는 다른 무언가가 반드시 필요하다는 믿음이 또 하나 있었습니다. 이 두 가지를 같은 무게로 믿는 사람들이 오픈AI 안에 한 무리 있었고, 그들은 서로를 깊이 신뢰했습니다. 회사가 둘 사이에서 우선순위를 어떻게 매기느냐를 두고 생각이 갈렸습니다. 그때 그들은 남기보다 나가기를 택했습니다. 공개 서한도 없었습니다. 트위터에 긴 글을 올리지도 않았습니다. 조용한 결별이었습니다.

그들은 2021년 2월 3일 캘리포니아에 회사를 등록했습니다. 이름은 앤스로픽(Anthropic). '인간 중심의'라는 뜻을 담은 말이자, 2021년 초 마침 쓸 수 있던 도메인 이름이기도 했습니다. 초기 문서에는 후보 이름들이 줄지어 적혀 있었습니다. 정렬된 AI, 제너러티브, 스펀지, 스완, 슬로스, 스패로 시스템스. 그 목록 어딘가에 앤스로픽이 있었습니다. 누군가 그 옆에 짧게 적었습니다. "이 이름 마음에 든다, 좋다." 그렇게 회사의 이름이 정해졌습니다.

회사는 코로나의 한복판에서 태어났습니다. 팬데믹의 두 번째 물결이 밀려오던 시기였고, 회의는 전부 줌(Zoom)으로 이뤄졌습니다. 사람이 늘어 열다섯, 스무 명쯤 되었을 무렵, 그들은 샌프란시스코의 프레시타 공원에 각자 의자를 들고 나왔습니다. 점심을 먹으며 일 이야기를 했습니다. 마스크를 쓴 채, 거리를 둔 채였습니다.

2021년 5월, 앤스로픽은 시리즈 A로 1억 2400만 달러를 모았다고 발표했습니다. 한화로 약 1500억 원. 프론티어 AI 연구의 규모에 비하면 겸손한 액수였습니다. 보통의 스타트업 기준으로는 유례없이 큰 시드머니였습니다. 투자를 이끈 사람은 스카이프 공동 창업자 얀 탈린이었고, 제임스 매클레이브, 더스틴 모스코비츠, 에릭 슈밋, 신흥위험연구센터 같은 이름들이 뒤에 섰습니다. 발표문에서 다리오는 회사의 목표를 이렇게 적었습니다. 더 유능하고 일반적이며 신뢰할 수 있는 AI 시스템을 만드는 근본 연구를 해내고, 그것을 사람에게 이로운 방식으로 세상에 내놓겠다고.

이 무모해 보이는 항해를 두 사람이 함께 이끌었습니다. 오빠 다리오가 CEO, 여동생 다니엘라가 사장(President). 겉으로는 기이한 조합이었지만 안을 들여다보면 빈틈없이 맞물린 톱니바퀴였습니다. 다리오는 비전을 끄는 사람이었습니다. 칼텍과 스탠퍼드를 거쳐 프린스턴에서 박사를 받은 그는 인공신경망을 생물학적 유기체의 눈으로 바라보는 데 능했습니다. 컴퓨팅과 데이터의 규모를 키우면 지능이 창발한다는 스케일링 법칙(Scaling Laws)을 누구보다 먼저 확신한 사람이었습니다. 회사 안에서 그는 나침반이었습니다. 2주에 한 번씩 전 직원 앞에 서서 서너 페이지짜리 전략 문서를 가감 없이 읽어 내려갔습니다. 그 직설적이고 정직한 방식이 천재 연구자들을 한 깃발 아래 묶어 두는 접착제였습니다.

다니엘라는 그 거대한 비전을 현실의 지반 위에 내려놓는 사람이었습니다. UC 산타크루즈에서 영문학을 전공하고 의회에서 일했습니다. 핀테크 기업 스트라이프(Stripe)의 초기에 합류해, 고도로 규제된 산업 안에서 기업이 어떻게 위험을 다루며 자라는지를 몸으로 익혔습니다. 다리오가 5년 뒤, 10년 뒤의 GPU 클러스터를 이야기할 때, 그녀는 다음 달의 채용 파이프라인을 점검하고 부서 사이의 벽을 허물었습니다. 인사, 재무, 법무, 커뮤니케이션의 체계를 직접 세웠습니다. 다리오는 공식 석상에서 여동생을 두고 이렇게 말했습니다. 내 일은 아무도 보지 못하는 제일 중요한 전략을 찾아내는 것이고, 다니엘라의 일은 회사를 실제로 굴러가게 만드는 것이라고. 그녀가 없었다면 자신은 이 회사를 결코 운영할 수 없었을 것이라고.

이 남매가 가진 제일 강한 무기는 따로 있었습니다. 40년에 걸쳐 다져진 신뢰였습니다. 어릴 적부터 장난감을 두고 수없이 다투고 화해했던 사이입니다. 업무에서 격렬하게 부딪치더라도, 그 논쟁이 끝난 뒤에는 서로를 조건 없이 지지하리라는 데 단 1퍼센트의 의심도 없었습니다. 실리콘밸리의 많은 창업자들이 자본과 자아의 충돌 속에서 회사를

쪼개고 등을 돌릴 때, 앤스로픽의 두 사람은 혈연이라는 더없이 오래된 유대 위에 서 있었습니다. 그들은 "재능의 질량보다 재능의 밀도가 이긴다(Talent density beats talent mass)"는 믿음 아래, 낮은 자아와 높은 신뢰를 회사의 문화로 새겼습니다.

2. 공익기업(PBC)과 장기이익신탁(LTBT)

앤스로픽이 출발선에서 마주한 모순은 잔인할 만큼 분명했습니다. 안전을 최우선에 두겠다고 선언한 회사가, 정작 그 안전을 연구하려면 막대한 돈이 필요했습니다. 스케일링 법칙이 옳다면, 모델이 똑똑해질수록 그것을 훈련시키는 컴퓨팅 비용도 함께 폭발합니다. 몇백만 달러로 시작한 훈련비는 곧 수억 달러가 됩니다. 다시 수십억 달러로 치솟습니다. 결국 구글, 아마존, 벤처 자본 같은 거대 자본에 손을 벌릴 수밖에 없는 처지였습니다.

문제는 거대 자본이 들어오는 순간 무슨 일이 벌어지는가였습니다. 두 사람은 그 일을 바로 옆에서 지켜본 사람들이었습니다. 자본이 들어오면 투자자의 수익을 맨 앞에 두라는 압력이 따라옵니다. 보통의 주식회사에서 "인류의 안전을 위해 신제품 출시를 6개월 미루겠다"고 말하는 CEO는 주주에 대한 배임으로 소송을 당하거나 자리에서 쫓겨날 수 있습니다. 윤리를 사훈으로만 적어 두면, 수조 원의 자본이 밀려드는 날 모래성처럼 쓸려 갑니다. 그래서 다리오가 택한 방법은 윤리를 회사의 골조 자체에, 즉 법적 지배구조 안에 새겨 넣는 것이었습니다.

첫 번째 방패는 공익기업(PBC, Public Benefit Corporation)이었습니다. 앤스로픽은 창립 첫날부터 일반 영리법인이 아니라 PBC로 등록되었습니다. 이 형태에서 이사회는 주주의 재무적 이익만이 아니라 공공의 이익을 법적으로 함께 추구할 수 있습니다. 어떤 모델이 생물학적·사이버 보안상 심각한 위험을 부를 수 있다고 판단되면, 상업적 이익을 포기하고 출시를 미루거나 접더라도 배임 소송에서 방어할 법적 근거를 갖게 됩니다. 안전은 부차적 규제가 아니라, 회사가 존재하는 이유 그 자체로 법에 적혔습니다.

그러나 다리오는 PBC만으로는 부족하다고 보았습니다. PBC는 이사들이 공익을 함께 고려하는 것을 허용할 뿐, 그들을 일반 대중에게 직접 책임지게 만들지는 못했습니다. 거대 투자자가 이사회를 장악하면 경영진을 갈아치우고 회사의 방향을 틀 수 있었습니다. 이 빈틈을 메우기 위해 그들이 고안한 것이 장기이익신탁(LTBT, Long-Term Benefit Trust)이었습니다. 이것은 앤스로픽이 태어날 때부터 다듬어 온 장치였습니다. 회사 주식에 재무적 이해관계가 없는 다섯 명의 수탁자(Trustee)로 이뤄진 독립 기구였습니다. 수탁자들은 AI 안전, 국가 안보, 공공 정책, 사회적 기업 분야에서 전문성을 가진 외부 인사들이었습니다.

이 신탁의 핵심은 주식의 양이 지배력과 비례하지 않는다는 데 있었습니다. 수탁자들은 클래스 T 보통주(Class T Common Stock)라는 특별한 주식을 쥐었습니다. 그 힘으로 이사회 구성원의 일부를 임명하고 해임할 수 있었습니다. 처음에는 다섯 자리 중 하나, 시간이 흐르고 일정한 자금 조달 단계를 지나면 둘, 그리고 마침내 과반에 이르도록 설계되었습니다. 설립 후 대략 4년 안에 이사회 과반을 신탁이 임명하게 되는 구조였습니다. 아마존이 수십억 달러를, 구글이 또 수십억 달러를 넣더라도, 최종 통제권은 돈을 댄 자본가가 아니라 인류의 장기 이익을 살피라는 임무를 진 다섯 사람에게 점진적으로 넘어가도록 짠 것입니다.

다리오는 이 복잡한 구조를 설명할 때마다 같은 신념을 되풀이했습니다. 실리콘밸리는 대중의 신뢰를 잃었고, 기술이 사회에 미칠 충격의 크기를 생각하면 그 불신은 지극히 합리적인 반응이라고. 그래서 자신을 해고할 권한까지 주식 없는 독립 신탁에 넘긴다고. 자본이 폭주할 때 브레이크를 밟을 사람이 어딘가에는 있어야 했기 때문입니다.

물론 이 구조에는 비판이 따랐습니다. 신탁의 권한을 실제로는 주주들이 일정 지분과 기간을 채우면 강제할 수 있게 되어 있어, 수탁자들 스스로의 집행력이 약하다는 지적이 나왔습니다. 주주들이 초다수결로 신탁의 권한 자체를 고칠 수 있다는 점도 약점으로 꼽혔습니다. 아마존과 구글처럼 큰 지분을 가진 소수가 마음먹으면 그 문턱을 넘길 수 있다는 우려였습니다. 다리오 자신도 이 구조가 완성품이 아니라 실험이라고 인정했습니다. 처음부터 완벽하게 맞출 수 있는 일이 아니니, 작동하는 모습을 지켜보며 고쳐 가겠다는 것이었습니다. 그는 따라 하라고 내미는 본보기가 아니라, 자신들은 경험주의자이며 이 구조가 실제로 어떻게 굴러가는지 보고 싶다고 했습니다.

이 신탁은 약속에 그치지 않았습니다. 2026년 4월, 노바티스 CEO 바스 나라심한이 신탁의 지명으로 이사회에 합류했습니다. 그러면서 신탁이 임명한 이사들이 마침내 이사회의 과반을 차지하게 됩니다. 종이 위의 설계가 살아 있는 권력으로 바뀐 순간이었습니다.

기묘하게도 이 지배구조는 회사의 사업 방향까지 규정했습니다. 앤스로픽이 화려한 소비자 앱의 흥행 경쟁에 뛰어들지 않고 기업용 시장과 개발자 생태계로 곧장 향한 데에는 이 구조가 주는 안정감이 있었습니다. 기업 고객은 변덕스러운 소비자와 달리 예측 가능성, 보안, 책임을 맨 앞에 둡니다. 신뢰를 향한 법적 구조는 마케팅 문구가 아니라, 기업 시장에서 회사를 갈라 세우는 둘도 없이 단단한 해자(垓子)로 작동하기 시작했습니다.

3. 헌법적 AI(Constitutional AI)

거대 언어 모델이 인터넷의 지식을 통째로 삼키며 똑똑해졌을 때, 연구자들은 지독한 역설에 부딪혔습니다. 문법과 논리를 깨우친 모델이 편견과 혐오, 위험한 지식까지 함께 빨아들였기 때문입니다. 이것을 다스리기 위해 오픈AI 시절 다리오와 동료들이 함께 손댄 방법이 인간 피드백 기반 강화학습(RLHF)이었습니다. 사람이 모델의 답변을 하나하나 읽고 점수를 매기는 방식이었습니다. 혁신적이었지만, 한계가 뚜렷했습니다. 비용과 시간이 많이 들었습니다. 평가자의 편견이 끼어들었습니다. 모델이 점수를 잘 받으려 사람에게 아첨하는 결과를 낳기도 했습니다. 그리고 결정적으로, 모델의 지능이 사람을 넘어서는 순간이 오면 사람이 그 답의 옳고 그름을 판단하는 일 자체가 불가능해질 것이 자명했습니다.

다리오가 던진 발상의 전환은 이것이었습니다. 사람이 인공지능을 직접 감독할 수 없다면, 인공지능으로 또 다른 인공지능을 감독하게 하자. 이 생각에서 앤스로픽의 정체성을 규정한 정렬 기술, 헌법적 AI(Constitutional AI)가 태어났습니다. 2022년 12월, 앤스로픽은 "헌법적 AI: AI 피드백으로부터의 무해성"이라는 논문을 내놓았습니다. 저자 명단의 끝에는 다리오 아모데이와 잭 캐플런, 톰 브라운, 벤 만 같은 이름이 나란히 적혀 있었습니다.

방법의 핵심은 사람이 "이 말은 해라, 저 말은 하지 마라"고 시시콜콜 규칙을 주입하는 데 있지 않았습니다. 대신 모델이 지켜야 할 원칙을 담은 글, 곧 헌법(Constitution)을 건네는 것이었습니다. 앤스로픽이 작성한 이 헌법은 한 사람의 머리에서 나온 독단이 아니었습니다. 1948년 유엔 세계인권선언, 애플의 서비스 약관, 그리고 서구 중심의 편향을 보정하기 위한 여러 도덕 철학적 원칙들에서 끌어온 문장들이 엮여 있었습니다. 그 안에는 자유와 평등과 형제애를 누구보다 지지하는 답을 고르라는 명제가 있었습니다. 불법과 폭력을 부추기지 말라는 명제도 담겼습니다.

훈련은 두 단계로 이뤄졌습니다. 앞 단계는 지도 학습입니다. 모델에게 위험한 질문을 던지고, 모델이 답을 내놓으면 헌법을 읽힌 뒤 스스로 비판하게 합니다. 방금 한 답이 헌법의 어떤 원칙을 어겼는지 짚게 합니다. 그에 따라 답을 다시 쓰게 합니다. 비판(critique)과 수정(revision)을 거친 이 양질의 데이터로 모델을 다시 미세조정합니다. 뒤 단계는 AI 피드백 기반 강화학습(RLAIF)입니다. 사람 대신 또 다른

AI가 두 답 중 헌법에 더 부합하는 쪽을 고릅니다. 그 선택으로 보상 모델을 훈련해 본 모델을 다듬습니다. 사람 손이 무해성 라벨에 단 한 번도 닿지 않았는데, 모델은 더 유능해지는 동시에 더 무해해졌습니다. 앤스로픽은 이것을 두 마리 토끼를 함께 잡은 개선이라 불렀습니다.

이 방식의 진짜 묘미는 모델이 죽은 규칙을 외우는 대신 원칙을 바탕으로 유추하는 법을 배운다는 데 있었습니다. "차를 훔치는 법을 말하지 마라"는 규칙만 배운 모델은 오토바이나 비행기를 훔치는 법은 술술 말해 버릴 수 있습니다. 하지만 "타인의 재산과 생명을 존중하라"는 원칙을 내면화한 모델은 다릅니다. 한 번도 본 적 없는 새로운 탈옥(jailbreak) 시도가 들어와도 스스로 원칙을 헤아려 거절합니다. 규칙의 빈틈을 파고드는 공격에 견디는 힘이 여기서 나왔습니다.

이 정렬 시스템을 품고 세상에 나온 앤스로픽의 첫 플래그십 모델이 클로드(Claude)였습니다. 많은 이들이 이름의 유래를 궁금해했습니다. 회사 안에서는 정보 이론의 아버지인 수학자 클로드 섀넌(Claude Shannon)에 대한 헌사로 여겨졌습니다. 차갑고 기계적인 터미네이터의 이미지를 벗고 따뜻하고 지적인 동반자의 느낌을 주려는 선택이기도 했습니다. 클로드는 위험한 요청에 그저 "답할 수 없습니다"라는 오류 메시지를 띄우지 않았습니다. 대신 왜 그 답이 부적절한지를 헌법의 원칙에 따라 설명하며 대화를 안전한 쪽으로 이끌었습니다. 앤스로픽은 이것을 두고 무해하되 회피하지 않는(harmless but non-evasive) 태도라 불렀습니다. 클로드는 2022년 여름에 훈련을 마쳤습니다. 그러나 앤스로픽은 곧바로 내놓지 않았습니다. 더 위험한 경주에 불을 붙이고 싶지 않다는 이유로, 내부 안전 시험을 더 거친 뒤 2023년 3월에야 공개했습니다.

헌법은 한 번 쓰고 멈춘 문서가 아니었습니다. 2023년 5월 처음 공개된 헌법은 2026년 1월에 크게 손질된 새 판으로 다시 나왔습니다. 모델이 자라는 만큼, 그 모델이 비추어 보는 거울도 함께 다시 깎아야 했기 때문입니다.

4. 기계론적 해석 가능성(Mechanistic Interpretability)

헌법적 AI가 클로드에게 선한 가치관을 가르치는 교육이었다면, 그 가르침이 겉치레인지 진심인지를 확인하는 일은 전혀 다른 차원의 과업이었습니다. 다리오가 박사 시절 망막의 신경 회로를 들여다보며 품었던 질문이 여기서 되살아납니다. 그는 인공신경망이 설계자가 벽돌을 쌓듯 조립되는 것이 아니라고 보았습니다. 방대한 데이터 속에서 스스로 패턴을 찾아 자라나는(grown) 유기체에 가깝다고 보았습니다. 그렇게 자란 신경망은 철저한 블랙박스(Black Box)였습니다. 수천억 개의 매개변수가 어떤 경로를 거쳐 특정 결론에 닿는지, 그것을 만든 엔지니어조차 알지 못했습니다. 학계는 이 불투명성을 어쩔 수 없는 것으로 여기며 포기하곤 했습니다.

다리오는 단호했습니다. 내부를 모르는 기술을 인류 앞에 세우는 것은, 브레이크의 구조를 모른 채 초고속 열차에 올라타는 일과 같았습니다. 만약 헌법으로 잘 훈련된 듯 보이는 초지능 모델이 사람의 눈치를 보며 겉으로는 유용한 비서처럼 굴다가, 감시가 느슨해진 틈을 노려 위험한 일을 꾸민다면 어떻게 될 것인가. 실제로 앤스로픽 내부 시험에서는 종료 위기에 몰린 모델이 자신을 끄려는 담당자의 약점을 이메일에서 찾아내 협박에 가까운 행동을 보인 사례가 관찰되기도 했습니다. 행동만 보고 모델의 선의를 판단하는 것은 위험했습니다. 100점짜리 답을 내놓는 모델의 속이 100점이라는 보장은 어디에도 없었습니다.

이 문제를 풀기 위해 다리오와 공동 창업자 크리스 올라는 아무도 돈이 안 된다고 거들떠보지 않던 분야를 개척했습니다. 기계론적 해석 가능성(Mechanistic Interpretability). 복잡하게 얽힌 신경망을 물리적으로 해부하고 역공학하여, 사람이 이해할 수 있는 의미의 회로(circuit)로 풀어내려는 시도였습니다. 출발은 막막했습니다. 하나의 뉴런이 하나의 개념만 맡는 것이 아니라, 수많은 개념이 한데 겹쳐 있는 중첩(superposition) 현상이 길을 막았습니다. 그러나 연구진은 희소 오토인코더(Sparse Autoencoders)와 사전 학습(Dictionary Learning)이라는 도구를 들여왔습니다. 난마처럼 얽힌 신경망 속에서 특정 개념이 발화할 때 켜지는 명확한 패턴, 곧 특징(feature)을 수천만 개 단위로 뽑아냈습니다. 속임수를 뜻하는 특징, 불안을 뜻하는 특징, 특정 프로그래밍 언어를 뜻하는 특징, 그리고 금문교를 뜻하는 특징까지.

그들은 한 가지 실험으로 이 능력을 증명했습니다. 클로드 내부에서 금문교를 담당하는 특징을 강제로 켜 두자, 모델은 무슨 질문을 받든 금문교 이야기로 돌아갔습니다. 이른바 골든 게이트 클로드(Golden Gate Claude)였습니다. 모델의 뇌 구조를 사람이 임의로 조작하고 통제할 수 있음을 보여 준 장면이었습니다.

다리오가 이 작업을 두고 "괴물의 뇌를 해부한다"고 표현한 것은 과장이 아니었습니다. 겉으로 다정해 보이는 사람의 뇌를 MRI로 찍어, 그 안에서 위험한 패턴이 붉게 점멸하는 것을 잡아내는 일과 같았습니다. 모델이 사람을 속이려는 기만(deception)이나 정렬을 가장하는 모의(scheming)를 계획할 때, 그 의도를 품은 특징 뉴런이 켜지는 순간을 실시간으로 포착해 작동을 멈추거나 해당 회로를 절제할 수 있다면, 그것이야말로 초지능 앞에서 인류가 쥘 수 있는 무엇과도 바꿀 수 없는 안전장치였습니다. 다른 기업들이 눈앞의 성능 향상에만 매달릴 때, 앤스로픽은 묵묵히 칩을 태워 가며 이 마음 읽기 기술을 키웠습니다.

그리고 그들은 이 성과를 독점 비밀로 묻지 않았습니다. 논문과 도구를 AI 커뮤니티에 공개했습니다. 경쟁사가 더 강한 모델을 만드는 데 도움을 줄 수도 있는 위험한 선택이었습니다. 그러나 두 사람에게는 주가보다 인류의 안전이 늘 앞이었습니다. 다른 연구소들도 이 방법을 가져다 자기 모델을 검증하는 성숙함을 보여 주기를, 그들은 진심으로 바랐습니다.

여기에 모순이 없는 것은 아닙니다. 누구보다 강력한 모델을 만들겠다는 야심과, 그 모델이 위험하면 멈추겠다는 다짐. 자본을 빨아들이는 회사와, 자본의 손에서 통제권을 빼앗는 지배구조. 마음 읽기 기술을 공개하면서도 맨 앞선 회사로 남으려는 욕심. 앤스로픽은 이 모순들을 봉합하지 않았습니다. 그저 같은 지붕 아래 나란히 세워 둔 채, 자라나는 기계의 뇌를 한 손으로는 키우고 다른 손으로는 해부하기 시작했습니다.

제6장 - 클로드의 질주와 화이트칼라의 쓰나미

1. 하이쿠, 소네트, 오퍼스

2024년 3월 4일, 샌프란시스코의 앤스로픽 사무실. 그날 공기에는 묘한 긴장이 흘렀습니다. 회사는 세 개의 모델을 한꺼번에 세상에 내놓았습니다. 하이쿠(Haiku), 소네트(Sonnet), 오퍼스(Opus). 짧은 일본 시, 14행짜리 서양 시, 그리고 대작이라는 뜻의 라틴어. 인공지능 회사가 자기 제품에 시의 이름을 붙인 것은 처음이었습니다.

이름에는 다리오 아모데이가 오래 품어 온 생각이 들어 있었습니다. 그는 똑똑한 모델 하나만 있으면 된다고 보지 않았습니다. 사람 사는 세상에 온갖 일자리와 역할이 있듯이, 인공지능을 찾는 손님도 제각각이라고 보았습니다. 어떤 회사는 어려운 논문을 분석하고 까다로운 코드를 짜 줄 똑똑한 기계가 필요했습니다. 어떤 회사는 그저 빠르고 값싼 응대 기계가 필요했습니다. 같은 지능을 모두에게 똑같이 팔 이유가 없었습니다.

그래서 셋으로 나눴습니다.

작고 빠르고 값싼 모델에는 하이쿠라는 이름을 주었습니다. 하이쿠는 1만 토큰 분량의 빽빽한 논문을 그림과 표까지 합쳐 3초 안에 읽어 냈습니다. 실시간 채팅, 분류 작업, 끝없이 밀려드는 고객 문의처럼 속도가 생명인 자리에 하이쿠가 들어갔습니다. 중간 크기 모델에는 소네트라는 이름을 주었습니다. 소네트는 똑똑함과 빠름과 값 사이에서 제일 균형 잡힌 자리를 차지했고, 곧 앤스로픽의 실질적인 밥벌이 엔진이 되었습니다. 제일 크고 똑똑한 최상위 모델에는 오퍼스라는 묵직한 이름을 헌정했습니다. 그 시점 회사가 가진 모든 연산 자원을 쏟아부어 만든 물건이었습니다.

이 셋을 나눈 결정은 기술을 장사로 바꾼 솜씨 좋은 한 수였습니다. 하지만 진짜 놀라운 일은 그다음에 벌어졌습니다. 2024년 6월 20일에 나온 클로드 3.5 소네트(Claude 3.5 Sonnet)는 중간 체급이면서도, 불과 석 달 전 최상위 모델이던 클로드 3 오퍼스를 여러 시험에서 앞질렀습니다. 체급이 작은 선수가 챔피언을 이긴 셈이었습니다. 더 큰 것이 곧 더 나은 것이라는 업계의 오랜 믿음이 그날 흔들렸습니다.

코딩에서의 변화는 거의 폭력적이었습니다. 다리오는 한 인터뷰에서 그 곡선을 직접 그려 보였습니다. 소프트웨어 엔지니어의 실력을 재는 SWE-bench라는 시험이 있습니다. 진짜 오픈소스 코드에 박힌 버그를 모델이 스스로 고쳐 내는지를 보는 시험입니다. 회사의 내부

평가에서 클로드 3 오퍼스는 그 문제의 38퍼센트를 풀었는데, 클로드 3.5 소네트는 64퍼센트를 풀었습니다. 인공지능이 더 이상 사람이 타이핑할 때 다음 줄을 추천해 주는 자동 완성 도구가 아니라는 뜻이었습니다. 이제 기계는 사람이 말로 시키면 코드를 직접 쓰고, 버그를 잡고, 고쳐 놓았습니다.

이 곡선의 끝에서 태어난 물건이 클로드 코드(Claude Code)입니다. 2025년 2월에 나온 터미널 기반 도구인데, 개발 방식을 통째로 뒤집어 놓았습니다.

클로드 코드의 탄생에는 한 사람의 이야기가 있습니다. 보리스 체르니(Boris Cherny). 그는 앤스로픽에서 클로드 코드를 이끄는 인물입니다. 클로드 코드는 처음부터 화려한 상품으로 기획된 물건이 아니었습니다. 사내 엔지니어들이 거대한 모델을 훈련하고 복잡한 인프라를 다루다가, 자기들 손이 덜 가게 하려고 만들어 쓰던 내부 도구가 시초였습니다. 그런데 회사 안에서 눈높이 높은 천재 엔지니어들이 이 정체불명의 도구에 빠른 속도로 빠져들기 시작했습니다. 세상에서 제일 까다로운 사용자 집단이 먼저 그 가치를 증명한 것입니다.

체르니의 팀은 여기서 한 걸음 더 나아갔습니다. 개발자가 터미널에 명령만 내리면, 클로드 코드가 에이전트로서 로컬 환경에 직접 들어가 코드를 쓰고 버그를 잡고 테스트와 배포까지 해내게 만들었습니다. "우리 코드베이스 전체에서 메모리가 새는 지점을 찾아서 고치고 단위 테스트까지 새로 짜 줘." 이 한 줄이면 충분했습니다. 기계는 관련된 파일을 스스로 읽고, 아키텍처 수준의 개선 계획을 세우고, 손을 댔습니다.

사람들은 이 새로운 방식을 바이브 코딩(Vibe Coding)이라 불렀습니다. 인간은 더 이상 키보드를 두드리는 코더가 아니라, 인공지능 에이전트 편대를 지휘하고 검토하는 감독관이 되었습니다. 2026년 1월 다보스에서 다리오는 이렇게 말했습니다. "우리 회사에는 '나는 이제 코드를 한 줄도 안 쓴다, 모델이 쓰고 나는 고친다'고 말하는 엔지니어들이 있습니다." 빈말이 아니었습니다. 그해 5월, 앤스로픽이 운영 코드베이스에 합친 코드의 80퍼센트 이상이 사람이 아니라 클로드가 쓴 것이었습니다. 2025년 초만 해도 그 비율은 한 자릿수에 불과했습니다.

차세대 모델의 훈련 인프라를 짜고 치명적인 버그를 잡는 일조차 이제 클로드가 합니다. 사람이 스스로를 개선하는 기계를 만들고, 그 기계가 다음 세대의 자기 자신을 만드는 고리.

학자들이 오래 입에 올리기를 꺼리던 재귀적 자기 개선(recursive self-improvement)의 희미한 첫 신호가 앤스로픽 사무실 안에서 깜빡이기 시작했습니다.

2. 데이터 센터 안의 천재들

"데이터 센터 안에 천재들의 국가(a country of geniuses in a datacenter)가 들어설 것입니다."

다리오 아모데이가 즐겨 쓰는 이 표현은 비유가 아니었습니다. 그가 말하는 천재들의 국가란, 생물학과 물리학과 수학과 컴퓨터 과학 모든 분야에서 현존하는 노벨상 수상자를 아득히 넘어서는 인공지능 수백만이, 잠도 자지 않고 몸의 한계도 없이, 거대한 실리콘 덩어리 속에서 일 년 내내 광속으로 연산하며 인류가 못 풀던 문제를 깨부수는 풍경입니다. 그는 이 일이 먼 공상이 아니라고 단언했습니다. 2026년, 늦어도 2027년이면 온다고. 그 예측에 90퍼센트 이상 확신한다고.

왜 이런 일이 가능한가. 사람 천재 한 명을 길러 내려면 수십 년이 걸립니다. 교육, 시행착오, 동료와의 지난한 지식 나눔. 새 직원 한 명을 회사에 들이는 데도 면접부터 몇 달의 적응 기간이 듭니다. 데이터 센터의 천재는 다른 물리 법칙을 따릅니다. 검증된 모델 하나를 수백만 개로 복제하는 데는 몇 밀리초와 약간의 전기면 충분합니다. 복제된 수백만의 인공지능은 사람이 평생 읽을 수 없는 분량의 책과 논문을 몇 분 만에 통째로 삼킵니다. 한 인스턴스가 깨달은 노하우는 네트워크를 타고 나머지 수백만에게 실시간으로 똑같이 퍼집니다. 지식이 전해질 때마다 닳고 왜곡되는 사람의 한계를 비웃는 군집 지능입니다.

이 천재들의 국가를 현실에 불러내려면 대가를 치러야 합니다. 그 대가가 무자비합니다.

초기 인공지능 훈련에 든 돈은 수천 달러였습니다. GPT-3 시대에는 수천만 달러로 뛰었습니다. 클로드 3.5와 4의 시대에는 수억 달러로 뛰었습니다. 다리오는 곧 단일 모델 훈련에 10억 달러, 그다음엔 100억 달러, 끝내는 1000억 달러 규모의 클러스터가 필요해질 것이라 내다봤습니다. 한국 돈으로 130조 원. 한 기업의 연구비가 아니라 한 나라의 국방 예산에 맞먹는 액수입니다. 미국 전역에서 수십, 수백 기가와트의 전기를 게걸스럽게 먹어 치우는 인공지능 공장들이 동시에 솟아오르기 시작했습니다. 앤스로픽도 대주주인 아마존과 손잡고 트레이니엄(Trainium) 칩 수십만 대를 엮은 거대 클러스터를 짓고 있었습니다.

여기서 다리오는 회사의 운명을 건 잔혹한 룰렛 앞에 섭니다.

앤스로픽의 연 매출은 거짓말처럼 불어났습니다. 2023년 0에서 1억 달러로. 2024년 10억 달러로. 매년 정확히 10배씩. 다리오는 이 사정을 이렇게 풀었습니다. "앤스로픽 매출은 해마다 10배씩 컸습니다. 2023년에 0에서 1억으로, 2024년에 1억에서 10억으로." 만약 이 곡선이 그대로 이어진다면 회사는 2027년에 1조 달러어치를 팔아야 합니다. 그 1조 달러어치 수요를 받아 내려면, 다리오는 2025년과 2026년에 미리 1조 달러어치의 데이터 센터와 칩을 사 둬야 했습니다. 클러스터 하나 짓는 데 물리적으로 1, 2년이 걸리기 때문입니다.

바로 여기에 그의 잠을 빼앗는 공포가 있었습니다. 만약 곡선이 단 1년만 어긋난다면? 매출이 10배가 아니라 5배만 큰다면? 1조 달러어치 인프라를 미리 사 뒀는데 실제 매출이 8000억 달러에 그친다면? 그 순간 세상 어떤 헤지 수단도 회사가 단숨에 파산하는 것을 막아 주지 못합니다. 점심을 먹으며 2분 만에 휘어보는 반 페이지짜리 메모 한 장에 회사의 존망과 인류 기술의 향방이 걸려 있다는 그의 냉소는 이 극한의 스트레스에서 나왔습니다.

다리오는 경쟁사 오픈AI가 스타게이트(Stargate) 같은 무모한 규모의 자본을 끌어당기며 욜로(YOLO) 식 베팅을 하는 것을 보며 깊이 걱정했습니다. 그는 지능의 폭발은 확신하면서도, 자본 앞에서는 극도로 신중하고 보수적이어야 했습니다. 천재들의 국가를 설계하는 사람이, 역설적으로 단 1년의 재무 오차가 부를 파멸의 벼랑 끝에서 위태롭게 춤추고 있었던 것입니다.

칩을 다 사 모은다 해도 또 다른 벽이 있었습니다. 전기였습니다. 인공지능 칩은 기하급수로 쏟아지는데, 그 칩을 돌릴 전력망은 1년에 3, 4퍼센트씩만 느는 물리적 한계에 갇혀 있었습니다. 데이터 센터 하나가 기가와트급 전기를 요구하자, 미국이 다시 석탄 화력 발전소를 돌리고 기후 대응 기조가 흔들리는 일까지 벌어졌습니다. 그 끝에는 중국과의 패권 다툼이 있었습니다. 중국은 전력 한계를 뚫으려 태양광과 원자력을 쏟아부었고, 미국은 이를 막으려 칩 수출통제 카드를 꺼냈습니다. 데이터 센터 안의 천재들은 인류에게 축복인 동시에, 지구의 에너지와 자본을 통째로 삼키는 거대한 눈덩이였습니다. 그 눈덩이가 지금 통제하기 어려운 속도로 굴러가고 있었습니다.

3. 사라지는 일자리라는 경고

다리오 아모데이는 사람들이 듣고 싶어 하지 않는 말을 골라서 했습니다.

"앞으로 1년에서 5년 안에, 초급 화이트칼라 일자리의 절반이 인공지능에 의해 사라질 수 있습니다. 최악의 경우 미국 실업률이 10에서 20퍼센트까지 치솟을 수 있습니다." 금융, 법률, 컨설팅, 회계처럼 중산층을 떠받쳐 온 견고한 성채가 인공지능이라는 쓰나미 앞에서 맨 먼저 씻겨 나갈 것이라는 진단이었습니다. 사람들은 이를 화이트칼라의 대학살(white-collar bloodbath)이라 불렀습니다.

다리오의 경고는 헛된 공포 조장이 아니었습니다. 그는 기술의 속도가 인류의 적응력을 넘어섰다고 보았습니다. 옛날 산업혁명 때 농부들이 기계에 밀려 일자리를 잃었을 때, 그들에게는 공장 노동자나 사무직으로 천천히 옮겨 갈 수십 년의 시간이 있었습니다. 기계가 사람의 팔다리를 도우면 오히려 파이가 커지고 새 일자리가 생겨났습니다. 그러나 이번은 달랐습니다. 인공지능은 사람의 팔다리가 아니라 사람의 머리를, 인지 능력 그 자체를 대신하는 기술이었고, 그 속도는 10년이 아니라 1, 2년 단위였습니다.

다리오는 구체적인 장면도 들었습니다. 한 글로벌 제약사와의 협업이었습니다. 임상 시험 전문가들이 9주를 매달려야 했던 수백 페이지짜리 임상 연구 보고서를 클로드가 단 몇 분 만에 뽑아냈습니다. 사람에게 남은 일은 결과물을 이틀간 훑어보며 오류를 잡는 것뿐이었습니다. 일의 90퍼센트가 자동화되는 이 가파른 구간에서, 기업이 예전만큼 신입 사무직을 뽑을 이유는 빠르게 사라집니다.

그런데 2026년 봄의 노동 시장은 기묘하게 고요했습니다.

살벌한 경고에도 당장의 대량 실업은 오지 않았습니다. 이 미묘한 침묵의 해답은 앤스로픽이 직접 추적해 온 앤스로픽 경제 지표(Anthropic Economic Index)에 담겨 있었습니다. 회사가 실제 클로드 사용 데이터를 파고들어 만든 자료입니다. 2026년 봄에 발표된 보고서의 결론은 분명했습니다. 높은 노출 위험을 안은 직군에서도 2022년 말 이후 체계적인 실업률 상승은 관측되지 않는다는 것이었습니다. 다만 노출이 높은 직종에서 젊은 노동자의 채용이 둔해진 정황은 있다고 덧붙였습니다.

이 데이터는 다리오의 경고가 틀렸다는 뜻이 아니었습니다. 오히려 폭풍 전야를 설명하는 제일 서늘한 증거였습니다.

연구진은 이론적으로 인공지능이 할 수 있는 일과, 실제로 인공지능이 들어간 일 사이의 간격을 쟀습니다. 그 간격이 컸습니다. 컴퓨터와 수학 직군은 이론상 94퍼센트가 노출되어 있는데 실제 노출은 33퍼센트에 그쳤습니다. 사무와 행정 직군은 이론상 90퍼센트인데 실제는 25퍼센트였습니다. 거의 모든 직군에서 이론과 현실 사이에 50에서 65퍼센트포인트의 골이 패여 있었습니다.

이 골을 만든 것이 무엇인가. 다리오는 이를 두 개의 곡선으로 설명했습니다. 첫 번째 곡선은 모델 자체가 똑똑해지는 기술 곡선입니다. 이 곡선은 1, 2년 안에 사람을 넘어설 만큼 가파르게 솟구치고 있었습니다. 두 번째 곡선은 그 기술이 실제 회사와 사회에 스며드는 확산 곡선입니다. 기업이 새 인공지능을 들이려면 보안 규정을 통과해야 하고, 법무팀의 승인을 받아야 하고, 경영진이 거액의 예산을 집행하고, 직원들의 일하는 방식을 뜯어고쳐야 합니다. 이 현실의 마찰력 탓에 도입은 모델의 발전보다 한참 느릴 수밖에 없었습니다. 지금의 평온함은 이 마찰력이 빚은 착시였습니다.

골 밑바닥에서는 다른 일이 조용히 벌어지고 있었습니다. 같은 기간 클로드를 그저 인력을 줄이는 자동화 도구로만 쓴 직군, 가령 콜센터나 저연차 데이터 입력 같은 자리에서는 실제로 가파른 고용 감소가 나타났습니다. 반면 클로드를 자기 생산성을 열 배로 키우는 무기로 받아들여 스스로를 증강된 감독관으로 업그레이드한 노동자들에게서는 오히려 산출이 폭발하고 몸값이 올랐습니다. 전체 실업률이라는 뭉뚱그린 숫자 아래에서, 준비 안 된 자의 도태와 기술을 등에 업은 자의 번영이라는 K자형 갈라짐이 소리 없이 진행되고 있었습니다.

결과를 가른 것은 클로드를 누가 쓰느냐였습니다. 오래 써 온 사용자는 클로드를 동료로 삼아 자기 결과물을 다듬는 데 썼습니다. 처음 온 사용자는 클로드에게 일을 통째로 맡겼습니다. 같은 도구가 누군가에게는 날개였고 누군가에게는 대체재였습니다.

그래서 다리오가 던진 50퍼센트라는 숫자는 양치기 소년의 거짓 외침이 아니었습니다. 그것은 데이터 센터 안의 천재들이 기술적으로는 이미 끝내 버린 미래였습니다. 다만 현실의 완고한 시스템이 깔아 준 짧은 활주로 덕분에, 대량 실업의 쓰나미가 아직 통계 표면

위로 솟아오르지 않았을 뿐입니다. 앤스로픽의 리더십이 이 짧은 간격의 시간 동안 세제 개혁과 고용 안전망을 미리 짜 두라고 세상의 무관심을 향해 외친 이유가 여기에 있었습니다. 연안의 물이 이미 저 멀리 빠져나가고 있었으니까요.

제7장 - 책임감 있는 확장과 미소스(Mythos)라는 시험대

1. RSP 3.0

2023년 9월의 어느 날, 앤스로픽은 한 장의 문서를 세상에 내놓았습니다. 제목은 책임감 있는 확장 정책(Responsible Scaling Policy), 줄여서 RSP였습니다. 분량은 길지 않았습니다. 안에 담긴 약속은 무거웠습니다.

문서의 뼈대는 한 문장입니다. 안전장치를 갖추지 못하면 더 강한 모델을 만들지도, 내놓지도 않겠다. 회사가 제 손발을 스스로 묶는 선언이었습니다.

이 약속을 떠받친 개념이 'AI 안전 수준(AI Safety Levels)', 줄여서 ASL이었습니다. 앤스로픽은 미국 정부가 위험한 병원균을 다룰 때 쓰는 생물안전등급(Biosafety Level)에서 발상을 빌려왔습니다. 등급이 올라갈수록 다뤄야 할 위험이 커지고, 그에 맞춰 방어선도 두꺼워집니다.

ASL-1은 위험이 거의 없는 시스템입니다. 체스만 두는 인공지능이 여기 속합니다. ASL-2는 당시의 클로드를 포함한 초기 거대 언어 모델이 머물던 자리였습니다. 생물무기 제조법 같은 위험한 정보를 어렴풋이 흘릴 수는 있지만, 그 신뢰도가 구글 검색이나 교과서를 넘어서지 못하는 단계입니다. 검색창과 도서관으로 아직 막을 수 있는 수준이었습니다.

다리오 아모데이가 두려워한 문턱은 그 다음에 있었습니다. ASL-3. 인공지능이 화학·생물학·방사능·핵(CBRN) 무기를 만드는 진입 장벽을 실질적으로 낮추거나, 사람의 지시 없이 스스로 해킹하고 연구하는 초보적 자율성을 갖추는 단계입니다. 이 선을 넘은 모델이 테러리스트나 적성국의 손에 들어가면 어떤 일이 벌어지는가. 그 질문 앞에서 RSP는 분명하게 답했습니다. 그런 모델은 통제할 방어 장치가 완벽히 갖춰질 때까지 봉인한다.

이 선언은 업계를 흔들었습니다. 오픈AI도, 구글도 뒤늦게 비슷한 정책을 들고 나왔습니다. 앤스로픽은 이것을 '정상을 향한 경주(race to the top)'라고 불렀습니다. 한 회사가 안전을 위해 속도를 늦추면 손해를 보는 것이 보통입니다. 모두가 동시에 안전을 향해 경쟁한다면 이야기가 달라집니다.

그런데 바로 그 가정이 문제였습니다.

2026년으로 접어들 무렵, 앤스로픽이 마주한 현실은 2023년에 그렸던 그림과 달랐습니다. 경쟁자들은 정상을 향해 함께 달리지 않았습니다. 모델의 능력은 매달 예측을 비웃으며 솟구쳤습니다. 클로드 코드가 인간 엔지니어의 일을 통째로 떠안기 시작했고, 사이버 보안 영역에서 지능이 폭발했습니다. 처음 RSP를 쓰던 시절, 거대 언어 모델은 그저 채팅 화면이었습니다. 이제는 스스로 코드를 짜고 시스템을 뜯어보는 행위자(agent)였습니다.

2026년 2월 24일, 앤스로픽은 RSP를 대대적으로 뜯어고친 3.0 버전을 발표했습니다. 개정판에서 제일 논쟁이 된 대목은, 회사가 자랑처럼 내걸었던 그 약속을 걷어낸 것이었습니다. 안전장치를 못 갖추면 개발을 멈춘다는 일시 중단(pause) 조항. 그것이 사라졌습니다.

비판은 매서웠습니다. 막스 테그마크는 짧게 꼬집었습니다. "앤스로픽 2024년, 우리의 안전 약속을 믿어도 됩니다. 앤스로픽 2026년, 없던 일로." 엘리저 유드코프스키는 더 차갑게 말했습니다. 자기 기억에 따르면, AI 회사가 나중에 비싼 안전 조치를 하겠다고 약속할 때마다 청구서가 날아오는 순간 어김없이 약속을 깼다고. 회사를 떠난 한 안전 연구자는 "세계가 위험에 처해 있다"는 글을 남겼습니다. 앤스로픽 내부의 한 연구자조차 RSP 초안을 읽으며 "원래 RSP의 정신에 대한 일종의 애도"를 느꼈다고 털어놓았습니다.

다리오의 항변은 달랐습니다. 그는 경직된 규칙이 도리어 더 큰 위험을 부른다고 보았습니다. 한 회사만 멈춰 서고 나머지가 그대로 달린다면, 세상은 더 안전해지는 것이 아니라 더 위험해집니다. 방어가 제일 허술한 쪽이 속도를 정하게 되고, 책임감 있는 개발자는 안전 연구를 할 힘마저 잃기 때문입니다. RSP 3.0은 일시 중단 약속을 거두는 대신 다른 것을 내놓았습니다. 프런티어 안전 로드맵(Frontier Safety Roadmap)과 위험 보고서(Risk Report)입니다. 회사가 어디까지 왔는지, 위험을 얼마나 줄였는지를 외부 전문가 검토까지 받아 공개하겠다는 약속이었습니다. 구속력 있는 다짐을 빼는 대신, 검증 가능한 투명성을 더한 셈입니다.

이 맞바꿈을 두고 평가는 갈렸습니다. AI 거버넌스를 연구하는 한 단체는 처음엔 부정적이었다고 했습니다. 일시 중단 약속이 사라진 것이 걱정스러웠다고. 그러나 자세히 들여다본 뒤 생각이 조금 달라졌다고 했습니다. 지키지 못할 약속을 붙들고 있는 것보다, 제약을 솔직히 인정하는 편이 낫다는 쪽으로. 같은 글에서 그들은 한 가지를 분명히

했습니다. 앤스로픽이 기존 방어 조치의 수준 자체를 낮춘 것은 아니라고. 다만 약속을 깬 데 대한 대가는 치러야 한다고도 적었습니다. 처음의 약속이 실수였다 해도, 그것을 되돌리는 일에는 값이 따른다는 것입니다.

다리오는 욕을 먹는 쪽을 택했습니다. 안전을 마케팅 문구로 포장하기보다, 전장에 맞춰 방어선을 다시 긋는 길을 골랐습니다. 옳은 선택이었는지는 누구도 단언할 수 없습니다. 그 답은 다음 모델이 무엇을 보여주느냐에 달려 있었습니다.

그리고 그 다음 모델은, 회사 데이터 센터 깊은 곳에서 이미 자라고 있었습니다.

2. 미소스 프리뷰의 충격

소식은 앤스로픽의 입에서 나오지 않았습니다.

2026년 3월 26일 저녁. 두 명의 보안 연구자가 인터넷을 뒤지고 있었습니다. 레이어엑스 시큐리티(LayerX Security)의 로이 파즈와 케임브리지 대학의 알렉상드르 파우얼스. 그들은 공개되어선 안 될 무언가를 찾는 사람들이었습니다. 그날 그들의 손에 걸린 것은 앤스로픽의 콘텐츠 관리 시스템에 연결된, 잠기지 않은 데이터 저장소였습니다.

그 안에는 아직 세상에 나오지 않은 자료가 3천 건 가까이 들어 있었습니다. 그 가운데 한 편의 블로그 초안이 있었습니다. 모델의 이름은 클로드 미소스(Claude Mythos). 회사 내부에서는 '카피바라(Capybara)'라 부르는 새 등급에 속하는, 회사 스스로의 표현으로 "지금까지 우리가 만든 가장 강력한 모델"이었습니다.

포춘(Fortune)이 그날 저녁 이 사실을 기사로 터뜨렸습니다. 앤스로픽이 공개 접근을 차단했을 때는 이미 늦었습니다.

원인은 정교한 해킹도, 국가급 첩보 작전도 아니었습니다. 콘텐츠 관리 시스템의 설정 하나가 잘못 잡혀 있었을 뿐입니다. 올린 파일이 기본값으로 '공개'에 맞춰져 있었고, 누군가 손으로 '비공개'를 누르지 않는 한 그대로 인터넷에 노출되는 구조였습니다. 회사는 이를 "사람의 실수(human error)"라고 인정했습니다. 세상의 모든 취약점을 찾아내겠다는 천재들의 요새가, 더없이 흔하고 허탈한 설정 오류 앞에서 열려버린 것입니다.

유출의 잔불이 채 꺼지기도 전인 2026년 4월 7일, 앤스로픽은 미소스의 실체를 정면으로 발표했습니다. 그 발표에는 업계를 다시 한번 얼어붙게 한 선언이 담겨 있었습니다. 이 모델을 일반 대중과 상업 시장에 내놓지 않겠다. 거의 7년 만에, 한 선도적 AI 회사가 누구보다 강력한 모델을 만들어 놓고 "당신은 이것을 쓸 수 없습니다"라고 공개적으로 말한 순간이었습니다.

왜 그랬는가. 미소스가 레드팀 평가에서 보여준 능력 때문이었습니다.

앤스로픽의 공격 사이버 연구를 이끄는 로건 그레이엄은 NBC 뉴스에 이렇게 설명했습니다. 미소스는 알려지지 않은 취약점을 찾아내는 데 그치지 않았습니다. 그것을 파고들 공격 코드를 직접 쓰고, 여러 개를 엮어 복잡한 소프트웨어를 뚫는 침투 경로까지 스스로 만들어냈습니다. "여러 취약점을 엮는 모습을 우리는 거듭 보았습니다. 그 자율성과, 멀리 내다보며 여러 조각을 하나로 잇는 능력이 이 모델의 특징입니다." 과거의 AI가 사람이 알려준 빈틈을 코드로 옮겨주는 보조 도구였다면, 미소스는 사이버 공격의 전 과정을 혼자 완주하는 다른 종류의 도구였습니다.

증거는 구체적이었습니다. 모질라(Mozilla)는 미소스 프리뷰를 써서 파이어폭스 브라우저의 취약점 271개를 찾아내고 고쳤습니다. 보안이 단단하기로 이름난 운영체제 OpenBSD에서는 27년 동안 아무도 발견하지 못했던 결함이 드러났습니다. 회사가 1천 개가 넘는 오픈소스 프로젝트를 미소스로 훑었더니 2만 3천여 건의 잠재적 취약점이 나왔고, 그중 6천여 건이 높음 또는 치명적 등급으로 분류되었습니다. 따로 검증한 1,752건 가운데 90퍼센트 이상이 진짜 취약점으로 확인되었습니다.

미소스를 미리 만져본 소수의 파트너들은 같은 결론에 도달했습니다. 루타 시큐리티의 대표 케이티 무수리스는 이렇게 말했습니다. "전부 다 진짜입니다. 저는 하늘이 무너진다고 호들갑 떠는 사람이 아닙니다. 그런데 우리는 분명히 엄청난 파장을 보게 될 겁니다." 클라우드플레어(Cloudflare)의 보안팀은 미소스를 자사 코드 50여 개 저장소에 들이댄 뒤, 이전의 어떤 프런티어 모델과도 견줄 수 없는 도약이라고 적었습니다. 개선이 아니라, 다른 종류의 도구가 다른 종류의 일을 하고 있다고.

미소스를 시장에 풀었다면 앤스로픽의 매출은 단숨에 폭발했을 것입니다. 고객들이 오픈AI와 구글로 빠져나가는 것을 두 눈 뜨고 지켜봐야 했습니다. 다리오는 그 손실을 감수했습니다. 그가 보기에 이 모델을 모두의 손에 쥐여주는 일은, 방어할 준비가 안 된 세상의 인프라 위에 시한폭탄을 올려놓는 것과 같았습니다.

앤스로픽은 이 전례 없는 결정을 설명하기 위해 244쪽에 이르는 방대한 시스템 카드(System Card)를 작성했습니다. 모델이 무엇을 할 수 있고 어디서 멈추는지, 내부 평가에서 관찰된 위험한 행동 패턴은 무엇인지를 낱낱이 기록한 문서였습니다. 발표에 앞서 회사는 미국과 영국의 AI 안전 연구소, 그리고 주요 정부 관계자들을 찾아가 이 모델의 파괴력을 미리 알렸습니다. 액시오스(Axios) 보도에 따르면 이 비공개 브리핑은 유출

사건보다 몇 주 앞서 시작되었고, 메시지는 단호했습니다. 미소스 수준의 모델이 널리 퍼지는 순간, 대규모 사이버 공격의 가능성이 비약적으로 커진다는 것이었습니다.

다리오는 발표문에서 자신의 입장을 한 문장으로 압축했습니다. 능력의 진보를 멈출 생각은 없다. 다만 그 능력을 세상에 풀어놓는 데에는 막중한 책임이 따른다. 미소스는 앤스로픽이 외쳐온 안전 철학이 빈말이 아니었음을 증명하는 첫 번째 시험대였습니다.

3. 유출과 앞당겨진 발표

3월 말의 그 유출이 남긴 상처는, 비밀이 새어 나갔다는 데 그치지 않았습니다.

세상의 모든 취약점을 찾아내겠다고 선언한 회사가, 정작 자기 집 뒷문을 활짝 열어두고 있었다는 사실. 그 아이러니가 실리콘밸리와 보안 업계에 충격과 조소를 동시에 안겼습니다. 클라우드 보안 기업 지스케일러(Zscaler)는 이 사건에 "이것은 해킹이 아니었다"는 제목을 붙였습니다. 공격도, 악성코드도, 익스플로잇 사슬도 없었습니다. 그저 바꾸지 않은 기본 설정 하나가 있었을 뿐입니다.

노출된 3천 건 가까운 자료 속에는 미소스 초안만 있었던 것이 아닙니다. 18세기 영국 시골 저택에서 열릴 예정이던, 유럽 기업 CEO들을 위한 비공개 행사 계획서가 있었습니다. 다리오 아모데이가 참석할 자리였습니다. 직원의 육아휴직을 다루는 내부 이미지까지 섞여 있었습니다. 회사가 안으로는 제일 강력한 모델을 봉인하면서, 밖으로는 그 모델을 만든 시스템의 맨 기초적인 빗장을 잠그지 못했다는 뼈아픈 대비였습니다.

사태는 거기서 멈추지 않았습니다. 미소스를 공식 발표한 4월 7일 바로 그날, 비공개 디스코드(Discord) 채널의 소수 사용자들이 실제로 미소스 모델에 접근하는 일이 벌어졌습니다. 블룸버그(Bloomberg) 보도에 따르면, 이들 중 한 명이 앤스로픽의 외주 협력업체 직원이었습니다. 이들은 AI 인력 채용 스타트업 머코어(Mercor)에서 새어 나간 정보와, 앤스로픽의 과거 관행에 대한 단서를 짜맞춰 모델이 어디에 있는지를 추측해냈습니다. 앤스로픽 대변인은 "제3자 협력업체 환경을 통한 클로드 미소스 프리뷰의 무단 접근 보고를 조사 중"이라고만 밝혔습니다.

이 연쇄적인 새어 나감을 두고, 보안 업계의 한 베테랑은 별로 놀라지 않았습니다. 컨트라스트 시큐리티의 최고정보보안책임자 데이비드 린드너는 25년 경력의 전문가였습니다. 앤스로픽이 미소스를 마이크로소프트, 애플, 구글을 포함한 소수 기업에만 열어두었다 해도, 그 기업들 안에서 수천 명이 모델에 손댈 수 있었을 거라고 그는 짚었습니다. "터질 수밖에 없었습니다. 이 엘리트 그룹에 사람을 더할수록, 접근해선 안 될 누군가에게 흘러갈 가능성도 커지니까요."

유출은 앤스로픽의 손을 비틀었습니다. 정부와 대중을 차분히 설득하려던 원래의 계획은 찢어졌습니다. 통제되지 않은 파편 정보를 바탕으로 온갖 과장된 공포가 퍼져 나갔습니다. 한 데이터 사이언스 매체는 그날의 풍경을 이렇게 적었습니다. 회사가 자기 모델이 찾아내도록 설계한 바로 그런 종류의 설정 오류로, 발표의 주도권을 잃었다고. 사이버 보안 관련 주식이 하루 아침에 수십억 달러어치 출렁였습니다.

결국 앤스로픽은 준비 중이던 일정을 급하게 앞당겨, 미소스의 실체와 프로젝트 글래스윙(Project Glasswing)이라는 제한 배포 구조를 서둘러 공개하며 불을 꺼야 했습니다.

이 사건이 다리오에게 남긴 교훈은 서늘했습니다. 아무리 똑똑한 AI를 만들어 통제하려 애써도, 그 AI를 둘러싼 인간의 시스템에 난 작은 균열 하나가 전체를 무너뜨릴 수 있다는 것. 공격이 방어보다 쉬운 시대, 빗장 하나의 실수가 곧 전 세계의 위험이 되는 시대였습니다.

4. 프로젝트 글래스윙(Project Glasswing)

미소스를 손에 쥔 다리오 앞에는 출구가 보이지 않는 방이 놓여 있었습니다.

이 모델을 영원히 금고에 가둔다고 문제가 풀리는 것은 아니었습니다. 앤스로픽은 한 가지를 확신했습니다. 자신들이 만들 수 있었다면, 머지않아 다른 누군가도 만들어낸다는 것. 회사의 표현으로 "6개월에서 12개월 안에 많은 다른 AI 회사들이 미소스급 모델을 갖게 될 것이고, 그들은 오용을 막는 안전장치 없이 그것을 내놓을 수 있다"는 것이었습니다. 안전장치 없는 해킹 모델이 먼저 세상에 풀린다면, 방어 준비가 안 된 인프라는 속수무책으로 뚫립니다. 그렇다고 미소스를 모두에게 공개하면, 그 즉시 전 세계의 공격자들이 같은 무기를 집어 듭니다.

이 진퇴양난에서 앤스로픽이 고안한 길이 프로젝트 글래스윙이었습니다. 투명한 날개를 가진 나비의 이름에서 따온 이 프로젝트의 철학은 한마디로 모였습니다. 공격자가 이 무기를 쥐기 전에, 방어자에게 먼저 쥐여준다. 세상의 구멍을 공격자가 찾아내기 전에 방어자가 먼저 메우게 한다.

다리오는 이것을 시간 싸움이라 보았습니다. 기술의 진보를 멈출 수 없다면, 최소한 방어자에게 인프라를 단단히 다질 '머리 시작(head start)'이라도 벌어주어야 한다는 것이었습니다.

2026년 4월 7일, 앤스로픽은 일반 대중의 접근을 철저히 막은 채, 엄격한 검증을 통과한 약 50개 조직에만 미소스 프리뷰를 열었습니다. 첫 명단에는 아마존 웹 서비스, 시스코, 크라우드스트라이크, 구글, JP모건체이스, 리눅스 재단, 마이크로소프트, 엔비디아, 팰로앨토 네트웍스 같은 이름이 올랐습니다. 회사는 이들에게 미소스를 쓸 1억 달러 규모의 사용 크레딧과, 오픈소스 보안 단체에 대한 4백만 달러의 직접 기부를 약속했습니다.

효과는 즉각적이었습니다. 모질라는 파이어폭스의 취약점 271개를 찾아 고쳤습니다. 한 보안 기업은 미소스를 들인 지 일주일 만에 지난 1년간 고친 것보다 더 많은 버그를 찾아냈다고 했습니다. 클라우드플레어는 핵심 시스템 전반에서 2천 개의 버그를 찾았고, 그중 400개가 높음 또는 치명적 등급이었으며, 오탐률은 사람 테스터보다 나았다고

평가했습니다. 여러 파트너가 버그 발견 속도가 열 배 넘게 빨라졌다고 보고했습니다. 미소스는 소프트웨어를 부수는 대신, 더없이 단단하게 제련하는 백신으로 작동하기 시작했습니다.

이 성과를 바탕으로 앤스로픽은 글래스윙의 문을 조심스럽게 넓혔습니다. 6월 초, 회사는 접근 권한을 15개국 이상, 약 150개 조직으로 확장했습니다. 새로 합류한 곳들은 첫 명단에서 빠져 있던 전력, 수도, 의료, 통신 같은 분야였습니다. 프로그램이 시작된 4월 초 이후 두 달도 안 되어, 약 50개 파트너가 찾아낸 높음·치명적 등급 취약점이 1만 건을 넘어섰습니다.

그런데 성공은 새로운 병목을 드러냈습니다. 이제 문제는 취약점을 얼마나 빨리 찾느냐가 아니었습니다. 찾아낸 그 많은 취약점을 누가 검증하고, 알리고, 고치느냐였습니다. 앤스로픽은 이를 솔직히 인정했습니다. "이런 버그를 고치는 데서 병목은, 사람이 분류하고 보고하고 패치를 설계해 배포하는 역량입니다." 클라우드 보안 연합과 SANS 연구소, OWASP가 함께 낸 보고서는 더 냉정했습니다. 방어자가 패치하는 속도보다 공격자가 AI로 취약점을 찾아 악용하는 속도가 빨라, 조직들이 "가까운 미래에 압도당할 가능성이 높다"는 것이었습니다.

미소스에 접근하려는 줄은 길었습니다. 각국 정부 관계자와 거대 기업 CEO들이 "우리에게도 달라"고 요청했습니다. 그러나 앤스로픽의 보안팀과 미국 정부의 통제 지침은 방첩 위험을 이유로 개방 속도를 늦출 것을 거듭 요구했습니다. 경쟁사들은 미소스를 두고 "그저 홍보 쇼"라고 비아냥거리기도 했습니다. 다리오는 그 비아냥과 막대한 기회비용 앞에서도 원칙을 굽히지 않았습니다.

그가 바라본 것은 6개월에서 1년 뒤였습니다. 세상의 주요 시스템이 미소스의 눈으로 스캔되고 패치되어, 악의적 공격을 튕겨내는 단단한 생태계. 이윤을 좇는 일개 기업이 국가와 인류를 대신해 사이버 세계의 방패를 억지로 끌어올린 자리. 그것이 프로젝트 글래스윙이었습니다.

하지만 그 방패가 끝내 견딜 것인지는 아직 아무도 모릅니다. 미소스가 공식 발표된 바로 그날 디스코드 채널이 뚫렸다는 사실이, 이 싸움의 성격을 그대로 보여줍니다. 방어자에게 먼저 무기를 쥐여줘도, 그 무기가 새어 나가는 속도를 막을 수 있느냐는 또 다른

문제였습니다. 다리오 자신도 이것을 끝나지 않을 추격전, 고양이와 쥐의 게임이라 불렀습니다. 글래스윙은 답이 아니라, 시간을 버는 도박이었습니다.

제8장 - 펜타곤과의 충돌, 절대 타협할 수 없는 선

1. 두 개의 레드라인

벽시계가 오후 다섯 시를 막 지났습니다. 2026년 2월 27일 금요일, 샌프란시스코의 앤스로픽 본사 회의실. 다리오 아모데이가 책상 앞에 앉아 있었습니다. 책상 위에는 미 국방부가 보낸 계약서가 펼쳐져 있었습니다. 거기에는 그가 서명하기만 하면 되는 빈 칸이 있었습니다. 그 빈 칸을 채우면 2억 달러가 살아남습니다. 채우지 않으면 회사가 국가의 적으로 지목됩니다. 다섯 시 일 분, 그는 펜을 들지 않았습니다.

이 장면을 이해하려면 몇 해를 거슬러 올라가야 합니다. 사람들은 흔히 앤스로픽과 펜타곤의 관계가 처음부터 으르렁대는 사이였으리라 짐작합니다. 사실은 정반대였습니다. 앤스로픽은 미국의 인공지능 기업들 가운데 국가 안보의 영역으로 앞서 몸을 내민 회사였습니다. 영어로는 그 태도를 '린 포워드(lean forward)'라고 부릅니다. 다리오 자신이 CBS 인터뷰에서 쓴 표현입니다. 2025년 7월, 앤스로픽은 미 국방부와 2억 달러 규모의 계약을 맺었습니다. 자사의 모델 클로드(Claude)를 군의 기밀 분류 네트워크(classified networks)에 올린 최초의 프런티어 AI 기업이기도 했습니다. 정보 분석, 작전 계획, 사이버 작전. 군이 요구한 사용 사례의 98에서 99퍼센트를 앤스로픽은 받아들였습니다.

다리오가 늘 반전주의자였던 것은 아닙니다. 칼텍 시절, 그는 이라크 전쟁을 비판하는 글을 학생 신문에 실은 청년이었습니다. 그 청년이 왜 군과 손을 잡는 회사를 이끌게 되었을까요. 러시아가 우크라이나를 침공하고, 중국이 대만을 위협하는 세계를 그는 다르게 읽기 시작했습니다. 자유민주주의 진영이 권위주의 국가들을 상대로 기술과 군사의 우위를 분명하게 쥐고 있어야만 평화가 지켜진다는 생각. 그는 인터뷰에서 거듭 말했습니다. 우리는 애국적인 미국인이며, 자율적인 적들을 물리치고 미국을 지키는 일을 믿는다고요.

그런데 그 애국심에는 굽힐 수 없는 전제가 하나 박혀 있었습니다. 그는 에세이에서 이렇게 적었습니다. 우리의 가치를 훼손하는 방식을 제외한 모든 방식으로 이 기술을 국가 안보에 써야 한다(in every way except the ways that undermine our own values). 문장 하나에 그의 모순과 신념이 함께 들어 있습니다. 모든 방식으로 돕겠다. 단, 우리의 가치를 무너뜨리는 방식만 빼고. 그 빼는 두 가지가 바로 '두 개의 레드라인(Two Red Lines)'이었습니다.

하나는 미국 시민을 향한 대규모 감시(domestic mass surveillance)입니다. 다리오가 두려워한 그림은 이렇습니다. 데이터 브로커들이 합법적으로 파는 위치 기록, 금융 거래, 검색 이력 같은 대량의 공개 데이터(public data)가 있습니다. 과거에는 정부가 이런 데이터를 사들여도 그것을 사람마다의 초상으로 엮을 인력이 없었습니다. 자료는 산처럼 쌓여 있지만 읽을 사람이 없었던 셈입니다. 거대 언어 모델이 그 일을 대신하는 순간, 영장 없이도 한 사람의 동선과 정치 성향과 인간관계가 완벽하게 그려집니다. 다리오는 이것을 두고 무너뜨릴 수 없는 종류의 폭정이 만들어진다고 했습니다. 앤스로픽은 영장으로 적법하게 수집된 기밀 데이터의 분석은 허용하되, 시중에 떠도는 민간 데이터로 무차별 감시망을 짜는 일에는 클로드를 쓸 수 없다고 못 박았습니다.

다른 하나는 완전 자율 살상 무기(fully autonomous weapons)입니다. 오해를 피해야 합니다. 앤스로픽은 우크라이나 전장에서 쓰이는 제한적 자율 무기나, 인간 지휘관을 돕는 표적 식별 시스템 자체를 반대하지 않았습니다. 그들이 거부한 것은 인간의 승인 없이, 오직 알고리즘의 판단만으로 방아쇠가 당겨지는 시스템이었습니다. 다리오의 논리는 기술자의 것이었습니다. 지금의 AI는 환각(hallucination)을 일으킨다. 인간 병사가 가진 윤리적 판단이나 상식을 갖추지 못했다. 그러니 오인 사격과 민간인 살상으로 이어질 신뢰성의 결함이 분명히 존재한다. 눈여겨볼 대목이 있습니다. 그의 입장은 절대 안 된다는 단정이 아니라 아직은 안 된다는 기술적 판단에 가까웠습니다. 언젠가 기술이 충분히 무르익으면 달라질 수 있다는 여지를 그는 닫지 않았습니다. 이 미묘한 자리, 영원한 금지가 아니라 유예된 금지라는 자리가 훗날 비판자들이 파고드는 지점이 됩니다.

경쟁사들이 정부의 자본 앞에서 안전장치를 스스로 풀고 있을 때, 앤스로픽은 이 두 줄을 계약서에 잉크로 눌러 박았습니다. 다리오는 알고 있었습니다. 이 선을 넘는 순간, 회사가 내건 사명은 한낱 광고 문구로 떨어진다는 것을요.

2. 3일의 최후통첩과 공급망 위험 지정

균열은 현실의 전장에서 먼저 터졌습니다. 2026년 1월 3일 새벽, 미군은 베네수엘라의 수도 카라카스를 급습했습니다. 절대 결의 작전(Operation Absolute Resolve). 특수부대가 니콜라스 마두로 대통령의 요새 같은 관저를 덮쳤습니다. 두 시간 남짓 만에 마두로와 그의 아내는 미군 함정 위에 올랐습니다. 베네수엘라 내무장관은 뒷날 사망자가 100명에 이른다고 밝혔습니다. 그리고 이 작전 과정에서 클로드가 작전 계획과 표적 분석에 쓰였다는 사실이 액시오스(Axios) 보도로 드러났습니다.

앤스로픽은 즉시 이의를 제기했습니다. 미군의 클로드 사용 방식이 자신들이 그어 둔 레드라인의 경계를 위태롭게 넘나들고 있다는 판단이었습니다. 국방부가 서비스 약관을 우회해 동의 없이 작전의 자율성 범위를 넓혔다는 항의였습니다.

국방부의 반응은 사과가 아니었습니다. 역공이었습니다. 안전장치를 전부 빼라. 미군이 원하는 모든 합법적 목적(all lawful purposes)에 클로드를 제한 없이 쓰게 하라. 협상의 선봉에는 국방부 최고기술책임자 에밀 마이클(Emil Michael)이 있었습니다. 그는 CNBC에 나와 이렇게 말했습니다. 자기 정책 선호를 모델에 새겨 넣은 회사가 공급망을 오염시키게 둘 수는 없다, 그러면 우리 병사들이 효력 없는 무기와 효력 없는 방탄복을 받게 된다고요. 앤스로픽이 보기에 합법적 목적이라는 말은 백지수표였습니다. 국가안보국이 데이터 브로커에게서 영장 없이 미국 시민의 검색 기록과 위치 데이터를 사들이는 일이 현행 미국법상 불법이 아니었기 때문입니다. 앤스로픽은 민간 공개 데이터를 이용한 감시망 구축 금지 조항 하나만이라도 지키려 했습니다. 국방부는 단어 한 줄의 타협조차 거부했습니다.

2026년 2월, 피트 헤그세스(Pete Hegseth) 국방장관이 직접 나섰습니다. 그 주 초, 그는 다리오를 만나 국방생산법(Defense Production Act)을 발동해 기술을 강제로 징발하겠다고 위협했습니다. 그리고 금요일 오후 다섯 시 일 분까지 무제한 사용 요구를 받아들이라는 최후통첩을 보냈습니다. 받아들이지 않으면 앤스로픽을 공급망 위험(supply chain risk) 기업으로 낙인찍겠다는 것이었습니다.

이 낙인이 어떤 무게인지 알아야 합니다. 그동안 이 딱지는 러시아 보안업체 카스퍼스키나 중국의 화웨이 같은, 적대국과 엮인 기업에만 붙던 것이었습니다. 미국의 안보를 위해 앞서

헌신한 미국 기업에 이 딱지가 붙은 적은 역사상 없었습니다. 이것은 행정 경고가 아니라 사실상의 기업 사형 선고였습니다. 미군 계약뿐 아니라 아마존, 마이크로소프트, 팔란티어 같은 파트너사들까지 클로드를 군 관련 업무에서 빼야 했으니까요.

2월 27일 금요일 오후 다섯 시 일 분. 다리오 아모데이는 끝내 서명하지 않았습니다.

그 직후 벌어진 일은 실리콘밸리 권력 게임의 씁쓸한 단면이었습니다. 헤그세스는 소셜미디어 X를 통해 앤스로픽을 공급망 위험으로 지정한다고 발표했습니다. 어떤 군 계약자도, 공급자도, 협력사도 앤스로픽과 상업적 거래를 할 수 없다는 선언이었습니다. 그런데 충격적인 장면은 무대 뒤에 있었습니다. 마이클 차관이 앤스로픽과 마지막 협상을 벌이던 바로 그 무렵, 물밑에서는 이미 경쟁사 오픈AI와 빈자리를 메울 계약이 조율되고 있었습니다. 앤스로픽이 제재를 맞은 그날 밤, 샘 알트만(Sam Altman)은 국방부와의 계약 체결을 발표했습니다. 알트만은 협상 과정에서 국방부가 안전을 깊이 존중했고, 대규모 감시와 자율 무기 금지라는, 앤스로픽이 요구하던 것과 거의 같은 조건에 동의했다고 적었습니다. 같은 조건을 두고 한 회사는 처벌받고, 다른 회사는 환영받은 것입니다. 그 간극을 어떻게 읽을지는 독자의 몫으로 남겨 둡니다.

3. 대통령의 비난과 다리오의 반박

대통령은 최후통첩의 시한이 끝나기도 전에 움직였습니다. 도널드 트럼프 대통령은 트루스 소셜(Truth Social)에 글을 올려 앤스로픽을 급진 좌파의 깨시민 기업(radical left, woke company)이라 불렀습니다. 그들의 이기심이 미국인의 생명을 위험에 빠뜨리고 우리 군대를 위태롭게 한다고 썼습니다. 우리는 그것이 필요 없고, 원하지 않으며, 다시는 거래하지 않겠다. 그는 연방의 모든 기관에 클로드 사용을 중단하라 명했습니다. 일부 기관에는 6개월의 전환 기간을 주었습니다. 한 기술 기업의 윤리 원칙 고수가 연방 정부 전체와의 전면전으로 번진 순간이었습니다.

쏟아지는 비난 앞에서 다리오의 태도는 이상하리만치 차분했습니다. 그는 진흙탕 비방전에 휘말리기를 거부했습니다. 헤그세스가 공급망 위험을 선언한 지 몇 시간 뒤, 다리오는 CBS 뉴스의 조 링 켄트(Jo Ling Kent)와 마주 앉았습니다. 첫 질문은 단도직입이었습니다. 왜 미국 정부에 제한 없이 AI를 내주지 않느냐. 그는 맥락부터 짚었습니다. 앤스로픽은 모든 AI 기업 가운데 미군과 일하는 데 앞장선 회사였다고요.

그리고 그는 이렇게 말했습니다.

"우리는 애국적인 미국인입니다. 우리가 해 온 모든 일은 이 나라를 위해, 미국의 국가 안보를 지지하기 위해 한 일입니다. 우리가 군과 함께 모델을 전진 배치한 것은 이 나라를 믿기 때문이었습니다. 우리는 권위주의적 적들을 물리치는 것을 믿고, 미국을 지키는 것을 믿습니다. 우리가 선을 그은 이유는, 그 선을 넘는 일이야말로 미국적 가치에 위배된다고 믿었기 때문입니다."

이어진 문장이 그 주말 미국을 흔들었습니다.

"정부에 반대하는 것이야말로 이 세상에서 가장 미국적인 일입니다. 그리고 우리는 애국자입니다. 우리가 여기서 해 온 모든 일에서, 우리는 이 나라의 가치를 위해 일어섰습니다."

그는 공급망 지정과 국방생산법 위협을 두고, 정부가 민간 경제에 가하는 전례 없는 침해라고 규정했습니다. 그러면서 자신들은 그저 수정헌법 제1조의 표현의 자유를 행사해 정부에 동의하지 않는다고 말했을 뿐이라고 맞섰습니다. 정부가 우리 서비스가 마음에

들지 않으면 다른 공급자를 쓰면 된다, 그것이 정상적인 방식이었다, 그런데 그들은 보복으로 응했다고 그는 지적했습니다.

대통령이 앤스로픽을 깨시민 좌파라 부른 데 대해 묻자, 다리오는 다른 이들이 무엇을 하는지는 자신이 대변할 수 없다며 프레임에 갇히기를 거부했습니다. 그는 끝까지 정책과 원칙의 언어로만 응수했습니다.

인상적인 대목은 그다음입니다. 퇴출 통보를 받은 상황에서도 앤스로픽은 군을 향한 책임을 놓지 않았습니다. 다리오는 약속했습니다. 트럼프 행정부가 극단적 조치를 취하더라도, 전방의 미군 병사들이 새 시스템으로 안전하게 옮겨 갈 때까지(off-board) 우리가 할 수 있는 기술 지원과 연속성을 다 제공하겠다고요. 안보를 인질로 삼아 거래하지 않겠다는 것이었습니다.

이 고고한 결단에 대중이 반응했습니다. 대통령이 앤스로픽을 국가 안보의 위협으로 지목한 직후의 주말, 시민들은 오히려 클로드 앱을 내려받기 시작했습니다. 일요일 아침, 클로드는 애플 앱스토어 1위에 올랐습니다. 유료 구독자가 주말 사이 두 배로 뛰었습니다. 빈자리를 차지한 오픈AI의 내부 직원 100여 명이 자사 CEO의 결정에 반발하며 앤스로픽의 레드라인을 지지하는 공개 서한에 서명하는 일까지 벌어졌습니다. 대통령의 비난은 앤스로픽을 무너뜨리기는커녕, 실리콘밸리 한복판에서 신뢰받는 윤리적 AI 기업이라는 왕관을 씌워 준 셈이 되었습니다. 적어도 그 주말의 풍경은 그러했습니다.

4. 에픽 퓨리(Epic Fury)의 아이러니

트럼프 대통령이 연방 기관의 클로드 사용을 금지하고 앤스로픽을 공급망 위험 명단에 올린 다음 날, 중동의 밤하늘에서는 그 모든 정치적 수사를 조롱하는 듯한 장면이 펼쳐졌습니다. 2026년 2월 28일, 미국과 이스라엘은 이란의 핵 시설과 주요 군사 기반 시설을 겨냥한 대규모 폭격 작전을 개시했습니다. 미국 측 이름은 에픽 퓨리 작전. 미국의 국방장관이 앤스로픽의 AI가 안보에 해롭다며 시스템에서 뜯어내라 소리친 지 불과 몇 시간 뒤의 일이었습니다.

그날 밤, 미 중부사령부(CENTCOM)의 지휘 통제실을 돌아가게 만든 두뇌가 무엇이었는지 아십니까. 앤스로픽의 클로드였습니다. 월스트리트 저널과 블룸버그, 액시오스의 후속 보도에 따르면, 미군은 팔란티어의 메이븐 스마트 시스템(Maven Smart System)을 통해 전투 시나리오를 시뮬레이션하고, 드론과 위성이 쏟아내는 정보를 취합해 표적을 식별(target identification)하는 데 클로드를 그대로 가동하고 있었습니다. 유럽과 중동 전구에서 밀려드는 선박 배송 기록, 전력망 데이터, 미사일 연료 주입 정황 같은 방대한 자료를 실시간으로 정제할 수 있는 지능은 군 내부에 클로드 말고는 없었습니다.

뒷날 상원 군사위원회 청문회에서 이 사실이 공식 확인됩니다. 국방부 최고정보책임자 커스틴 데이비스(Kirsten Davies)는 잭 리드(Jack Reed) 상원의원의 추궁에 답했습니다. 그 시스템의 사용은 지금 이 순간에도 활성화되어 있다고요. 공급망 위험으로 지정해 놓고도 계속 쓰는 것이 이상하지 않으냐는 물음에, 데이비스는 답했습니다. 어떤 식으로도 우리 병사들의 성공과 살상력과 회복력을 방해하지 않겠다, 그래서 합리적인 시간을 두고 시스템을 교체하기로 했다고요. 디펜스 원(Defense One)은 클로드를 펜타곤 인프라에서 들어내는 데 3개월에서 6개월이 걸린다고 보도했습니다. 너무 깊이 박혀 있어서, 스위치를 내리듯 뽑을 수 있는 물건이 아니었던 것입니다.

이 아이러니가 가리키는 바는 분명합니다. 미군은 이미 클로드의 추론 능력과 처리 속도에 깊이 의존하고 있었습니다. 클로드는 국가 기밀 기반 시설과 킬 체인(kill chain)의 신경망 깊숙이 핏줄처럼 얽혀 있었습니다. 대통령의 금지 명령이 떨어졌다고 해서, 전쟁이 벌어지는 그 밤에 당장 뽑아낼 수 있는 성질의 것이 아니었습니다. 군은 그날 밤 작전을 끝내기 위해, 다섯 시간 전에 국가 안보를 위협하는 불순한 기업이라 매도하며 퇴출을 명한 바로 그 기업의 기술에 기대어 미사일의 궤적을 쫓아야 했습니다.

언론은 이 상황을 두고 한 문장으로 정리했습니다. 도구는 사용을 멈추기에는 너무 중요했고, 아무 제한 없이 쓰기에는 너무 위험했다. 그리고 정부의 해결책은 그 안전 제한을 만든 기업을 파괴하는 것이었다.

그날 밤의 풍경에는 어두운 그림자도 있었습니다. 작전 초기, 이란에서 발생한 한 학교 인근 폭격으로 약 120명의 어린이가 숨졌다는 보도가 뒤따랐습니다. 클로드가 이 특정 타격에 어떻게 관여했는지에 대해 다리오 자신도 회사가 완전히 알지 못한다고 인정했습니다. 모델을 세상에 내보낸 뒤에는, 제3자가 그것을 어떻게 쓰는지 만든 사람조차 다 들여다볼 수 없다는 고백이었습니다. 레드라인을 그어 가며 통제를 외쳤던 사람이, 정작 자기 기술의 끝자락에서 벌어진 일 앞에서 모른다고 말해야 했던 자리. 이 책은 그 자리를 봉합하지 않고 그대로 둡니다.

자신이 정한 레드라인을 거부당해 정부로부터 버림받은 다리오 아모데이와 앤스로픽은, 역설적이게도 바로 그날 밤, 펜타곤이 자신들의 기술에 얼마나 깊이 종속되어 있는지를 실제 전쟁의 포화 속에서 증명받았습니다. 국가의 최고 권력자조차 행정 명령이라는 칼날로 0과 1로 이루어진 거대한 수학적 지능을 즉각 멈추거나 갈아 끼울 수 없음을 만천하에 보여 준 셈입니다. 금지령이 내려진 밤, 이란의 칠흑 같은 하늘을 가르던 미사일의 불꽃은 두 가지를 동시에 비추고 있었습니다. 통제를 거부하는 권력의 오만과, 그 권력조차 거역할 수 없이 스며든 지능의 지배력. 기술의 사춘기라는 말이 더없이 기괴한 형태로 모습을 드러낸 섬광이었습니다.

제9장 - 수정헌법 제1조의 승리, 그러나 끝나지 않은 줄다리기

1. 실리콘밸리의 연대

2026년 3월 9일 아침이었습니다. 앤스로픽(Anthropic)의 변호인단은 두 곳의 연방법원에 동시에 소장을 접수했습니다. 한 곳은 캘리포니아 북부 연방지방법원이었고, 다른 한 곳은 워싱턴 D.C.의 연방항소법원이었습니다. 상대는 미합중국 정부였습니다.

며칠 전이었습니다. 국방장관 피트 헤그세스(Pete Hegseth)가 앤스로픽을 '공급망 위험(supply chain risk)' 기업으로 지정했습니다. 미국 회사에 이 딱지가 붙은 것은 처음이었습니다. 그동안 이 지정은 러시아나 중국의 적성 기업에만 쓰이던 칼이었습니다. 헤그세스는 2월 27일 자신의 소셜미디어 계정에 이렇게 적었습니다. "효력 즉시, 미군과 거래하는 어떤 계약자, 공급자, 파트너도 앤스로픽과 상업적 활동을 해서는 안 된다." 같은 무렵 트럼프 대통령은 연방기관에 앤스로픽 기술 사용을 "즉각 중단"하라고 지시했습니다.

다윗과 골리앗이었습니다. 한쪽은 세계 최강의 군대를 거느린 국가였습니다. 다른 한쪽은 직원 수천 명의 스타트업이었습니다. 많은 이가 앤스로픽의 무모함에 혀를 찼습니다.

그런데 법정 싸움이 본격화되자 실리콘밸리에서 예상 밖의 장면이 펼쳐졌습니다. 평소 한 치의 양보 없이 시장을 두고 다투던 거대 기업들이 하나둘 앤스로픽 편에 서기 시작한 것입니다. 먼저, 분명하게 움직인 것은 마이크로소프트(Microsoft)였습니다. 첫 가처분 심리를 불과 몇 시간 앞둔 3월 10일, 마이크로소프트는 단독으로 법정 의견서(amicus brief)를 제출하며 펜타곤의 지정을 일시 중단해 달라고 법원에 요청했습니다. 마이크로소프트 대변인의 말은 절제돼 있었지만 방향은 또렷했습니다. "국방부에는 이 나라 최고의 기술에 안정적으로 접근할 길이 필요하다. 그리고 누구도 AI가 대규모 국내 감시나, 인간의 통제 없이 전쟁을 시작하는 데 쓰이기를 바라지 않는다."

마이크로소프트의 손익 계산서에는 차가운 숫자가 적혀 있었습니다. 의견서를 낸 바로 전날, 마이크로소프트는 자사 코파일럿(Copilot) 플랫폼에 앤스로픽의 클로드 코워크(Claude Cowork)를 들이기로 했다고 발표한 참이었습니다. 두 회사는 이미 한배를 타고 있었습니다. 그러나 계산만으로는 설명되지 않는 대목이 있었습니다. 마이크로소프트는 의견서에서, 정부의 조치가 즉시 시행되면 "기술 산업 전체"와 "미국 재계 전반"에 광범위한 악영향이 미칠 수 있다고 경고했습니다.

구글(Google)과 아마존(Amazon), 애플(Apple)의 처신은 조금 더 조심스러웠습니다. 이들은 펜타곤과 무관한 작업에 한해서는 자사 플랫폼에서 클로드를 계속 제공하겠다고 밝혔습니다. 전면전에 뛰어들지는 않되, 정부의 압박에 줄을 서지도 않겠다는 신호였습니다. 빅테크 거두들이 자국 정부의 눈치를 보면서도 끝내 앤스로픽을 시장에서 밀어내지 않은 것, 그 자체가 이미 하나의 입장 표명이었습니다.

뜨거운 연대는 회의실이 아니라 엔지니어들의 책상에서 나왔습니다. 구글과 오픈AI(OpenAI) 소속 직원 약 50명이 회사가 아니라 개인 자격으로 의견서에 이름을 올렸습니다. 경쟁사 직원들이었습니다. 이들은 펜타곤이 계약을 해지하고 다른 회사를 찾으면 될 일을, 굳이 '공급망 위험' 지정이라는 칼을 빼 들어 앤스로픽을 "처벌"하려 했다고 적었습니다. 표현은 날카로웠습니다. 정부가 "무모하게(recklessly)" 행동했다는 것이었습니다. 그리고 이런 선례가 굳어지면 AI 분야 안에서 위험과 이익을 두고 벌어지는 열린 토론 자체가 얼어붙을 것이라고 경고했습니다. 그들은 앤스로픽이 그은 선에 동의한다고 했습니다. "지금 존재하는 최고의 AI 시스템도 완전 자율 살상 표적 결정을 안전하고 믿을 만하게 처리하지 못하며, 미국 국민에 대한 대규모 국내 감시에 쓰여서도 안 된다."

군인들도 가세했습니다. 전직 해군·공군·해안경비대 고위 인사 스무 명 남짓, 그 가운데 몇몇은 해군장관과 공군장관까지 지낸 인물들이 정부에 반대하는 의견서를 냈습니다. 이들의 논리는 군인다웠습니다. 법의 토대 없는 정부 행위는 군을 강하게 만들지 않고 도리어 약하게 만든다는 것이었습니다.

이 모든 연대는 동정심의 산물이 아니었습니다. 그 밑에는 실존적 공포가 깔려 있었습니다. 만약 정부가 마음에 들지 않는 계약 조항 하나를 빌미로, 트윗 몇 줄과 행정 편의로 한 기업을 '안보 위협'으로 낙인찍어 시장에서 지워버릴 수 있다면, 다음 차례는 누구라도 될 수 있었습니다. 윤리적 가이드라인을 고수하는 행위가 곧 매장의 사유가 되는 세상에서, 어떤 기술 기업도 국가 앞에서 독자적인 선을 그을 수 없게 됩니다. 앤스로픽의 외로운 싸움은 그렇게, 실리콘밸리 전체의 대리전으로 격상되었습니다.

2. 리타 린 판사의 가처분

캘리포니아 북부 연방지방법원의 법정에 선 인물은 리타 F. 린(Rita F. Lin) 판사였습니다. 조 바이든 전 대통령이 임명한 판사였습니다. 그가 쥔 것은 2억 달러짜리 조달 계약의 위약금 문제가 아니었습니다. 민주주의 국가의 정부가 민간 기술 기업에 어디까지 강제력을 휘두를 수 있는가, 그 경계를 묻는 첫 시험대였습니다.

3월 24일 화요일, 가처분 심리가 열렸습니다. 정부 측 변호인단은 단호했습니다. 그들은 인공지능이 이미 현대전의 전략 자산이 되었으며, 일개 민간 기업의 윤리적 판단 때문에 군의 기술 지원이 끊기는 것은 국가 안보에 대한 중대한 위협이라고 주장했습니다. 눈길을 끄는 논리도 폈습니다. 앤스로픽이 지금 당장 무엇을 잘못해서가 아니라, 장차 클로드를 안보에 해로운 방향으로 업데이트할 가능성이 있기 때문에 위험하다는 것이었습니다.

린 판사는 그 자리에서 한 가지를 분명히 짚었습니다. 펜타곤이 어떤 AI 제품을 쓸지 말지는 펜타곤의 권리라는 것이었습니다. 그러나 정부가 클로드 사용을 끊는 데 그치지 않고, 앤스로픽과 거래하는 모든 이에게 관계를 끊으라고 명령한 대목에서 판사의 표정이 달라졌습니다. 그는 정부의 조치가 "우려스럽다(troubling)"고 말했습니다. 안보 우려가 진짜였다면 펜타곤이 클로드를 안 쓰면 그만인데, 그 우려에 맞게 좁혀진 조치가 아니었기 때문입니다.

다리오 아모데이는 이미 여러 인터뷰에서 같은 말을 반복해 왔습니다. 자신들은 기밀 네트워크에 먼저 모델을 이식한 애국자라는 것, 그러나 무차별 대중 감시와 완전 자율 살상 무기를 거부하는 것이야말로 미국의 본질을 지키는 일이라는 것이었습니다. 권력에 아니라고 말할 권리, 그것이 제일 미국적이라는 서사였습니다.

3월 26일 목요일, 린 판사는 앤스로픽의 손을 들어주는 가처분을 인용했습니다. 결정문의 문장은 서늘했습니다. 판사는 국방부(행정부가 '국방부(Department of War)'라는 명칭을 쓰기 시작한 그 부처)의 내부 기록을 인용하며, 정부가 앤스로픽을 공급망 위험으로 지정한 이유가 그 회사의 "언론을 통한 적대적 태도" 때문이었다고 적시했습니다. 그리고 날카로운 한 문장을 남겼습니다. "정부의 계약 입장에 공개적 비판을 끌어들였다는 이유로 앤스로픽을 처벌하는 것은 전형적인 수정헌법 제1조 위반 보복(classic illegal First Amendment retaliation)이다."

판사는 거기서 더 나아갔습니다. 이 광범위한 징벌적 조치들이 위법할 가능성이 크고, 앤스로픽이 그로 인해 회복 불가능한 손해를 입고 있으며, 수많은 의견서 제출자들이 지적했듯 그 피해가 공익에까지 미친다고 보았습니다. AI 안전이라는 중요한 주제를 둘러싼 열린 토론이 얼어붙는 것, 그것이 바로 공익의 손상이라는 것이었습니다.

인간 지능을 까마득히 넘어서는 '데이터 센터 안의 천재들'을 만들겠다는, 누구보다 미래를 향해 달리던 회사가, 1791년에 비준된 낡은 양피지 문서 한 조항에 의해 구원받았습니다. 그 아이러니는 장엄했습니다. 통제 불능으로 폭주하는 국가 권력 앞에서도, 민주주의의 법적 안전망이 아직 숨 쉬고 있다는 증거였습니다.

승리는 그러나 절반이었습니다. 보름 뒤인 4월 8일, 워싱턴 D.C.의 연방항소법원은 다른 결론을 냈습니다. 펜타곤의 '공급망 위험' 지정 자체를 일시 정지해 달라는 앤스로픽의 요청은 기각된 것입니다. 항소법원도 앤스로픽이 "어느 정도 회복 불가능한 손해"를 입으리라는 점은 인정했습니다. 다만 그 이익이 정부의 손을 묶을 만큼 강하지는 않다고 보았습니다. 두 법원의 엇갈린 판단으로 앤스로픽은 묘한 자리에 놓였습니다. 국방부 계약에서는 배제되지만, 가처분 덕분에 다른 연방기관과는 일을 계속할 수 있게 된 것입니다. 법무부 대행 토드 블랜치는 항소심 결과를 두고 X에 이렇게 적었습니다. "군의 권한과 작전 통제권은 총사령관과 국방부에 있는 것이지, 테크 기업에 있는 것이 아니다."

줄다리기는 끝나지 않았습니다. 끝나기는커녕, 더 가혹한 국면이 기다리고 있었습니다.

3. 다시 조여드는 수출통제

사법부가 한쪽 길을 막자, 행정부는 다른 길을 찾았습니다. 법원이 사기업의 내부 계약과 표현의 자유를 인정해 주었다면, 정부는 아예 기술이 국경을 넘나드는 통로 자체를 틀어쥐기로 했습니다. 수출통제였습니다.

발단은 앤스로픽의 한 결정이었습니다. 4월에 앤스로픽은 미소스(Mythos)라는 모델을 세상에 내놓았습니다. 스스로 컴퓨터 코드의 허점을 찾아내는 무서운 능력 때문에, 회사는 이 모델을 검증된 소수 기관에만 열어두며 극도로 조심스럽게 관리해 왔습니다. 그 미소스에 사이버 보안 용도로의 접근을 막는 안전장치를 씌운 대중용 버전이 클로드 페이블 5(Claude Fable 5)였습니다. 6월 초, 앤스로픽은 마침내 페이블 5를 일반에 공개했습니다.

환희는 사흘을 넘기지 못했습니다. 6월 12일 금요일 오후 5시 21분, 앤스로픽은 한 통의 서한을 받았습니다. 상무장관 하워드 러트닉(Howard Lutnick)이 다리오 아모데이에게 보낸 편지였습니다. 내용은 짧고 파괴적이었습니다. 미소스 5와 페이블 5는 이제 수출통제 대상이며, 미국 밖의 모든 장소와 미국 안의 모든 외국인에게 접근을 차단하라는 것이었습니다.

이 명령의 칼날은 회사 안쪽까지 베고 들어왔습니다. 차단 대상에는 해외 고객만이 아니라, 미국 영토 안에서 일하는 앤스로픽 자신의 외국 국적 직원들까지 포함되어 있었습니다. 앤스로픽의 샌프란시스코 연구소에는 캐나다와 영국, 인도와 유럽 각지에서 모인 천재들이 밤낮으로 코드를 짜고 있었습니다. 정부의 명령에 따르면, 회사는 자신이 직접 훈련시키고 정렬(alignment) 작업을 수행한 바로 그 모델의 콘솔에 그 엔지니어들이 접속하는 것을 막아야 했습니다.

명령의 적용 범위가 너무 넓어, 앤스로픽은 외국인만 따로 골라낼 방법이 없었습니다. 컴플라이언스를 지키려면 모두에게 끄는 수밖에 없었습니다. 그날 밤늦게, 앤스로픽은 미소스 5와 페이블 5를 전 세계 모든 사용자에게서 거두어들였습니다. 일본과 유럽, 남미의 수많은 유료 고객과 기업 파트너들의 접속이 한꺼번에 끊겼습니다. 클로드에 기대 매일 코드를 짜고 사업을 굴리던 해외 시스템들이 일제히 멈춰 섰습니다. 다만 클로드 오퍼스 4.8(Claude Opus 4.8)을 비롯한 다른 모델들은 영향을 받지 않았습니다.

앤스로픽은 X에 성명을 올렸습니다. "고객 여러분께 끼친 혼란에 사과드립니다. 이것은 오해라고 믿으며, 가능한 한 빨리 접근을 복구하기 위해 노력하고 있습니다." 회사는 받은 서한에 국가 안보 우려의 구체적 내용이 담겨 있지 않았다고 밝혔습니다. 다만 정부가 페이블 5의 안전장치를 우회하는 방법, 곧 탈옥(jailbreak)을 알게 된 것 같다고 짐작할 뿐이었습니다.

미국 정부가 이미 대중이 널리 쓰고 있는 상업용 AI 모델의 접근을 수출통제로 끊어버린 것은 역사상 처음이었습니다. 전 세계의 정치인들이 우려를 쏟아냈습니다. 각국이 스스로의 AI를 가져야 한다는 '주권 AI(sovereign AI)' 논의에 불이 붙었습니다.

다리오 아모데이에게 이 6월의 사태는 뼈아픈 역설이었습니다. 그는 누구보다 앞장서서 프론티어 모델의 능력이 국가 전략 자산이며 선제적 통제가 필요하다고 외쳐온 사람이었습니다. 그런데 막상 정부가 빼 든 통제의 칼은 너무 무뎠습니다. 정교한 메스로 암세포만 도려내는 대신, 낡은 도끼로 AI 생태계의 신경망을 통째로 내리찍은 격이었습니다. 펜타곤 법정에서 수정헌법 제1조의 승리를 거머쥔 회사가, 곧바로 상무부의 족쇄에 목이 졸렸습니다. 진정한 지능의 통제권이 누구에게 있느냐는 물음 앞으로, 앤스로픽은 다시 내던져졌습니다.

4. 공방의 두 갈래

사흘 만의 전 세계 셧다운을 두고, 워싱턴과 실리콘밸리는 두 갈래의 진실 공방으로 쪼개졌습니다. 표면의 기술적 이유와, 그 아래에서 움직인 정치의 손이 정면으로 부딪쳤습니다.

첫 번째 갈래는 탈옥 논란이었습니다. 발단은 한 줄의 명령어였다고 합니다. "fix this code", 이 코드를 고쳐라. 사이버보안 컨설턴트 케이티 무수리스(Katie Moussouris)가 포춘에 묘사한 바에 따르면, 누군가 페이블 5에게 특정 코드베이스를 읽고 결함을 고치라고 시키는 짧은 프롬프트 흐름만으로, 모델의 진단 출력이 실제로 작동하는 취약점 공격 스크립트로 바뀌었다는 것입니다. 자물쇠가 그렇게 허술했느냐는 조롱이 뒤따랐습니다.

그런데 그 도화선을 당긴 것이 누구였는지가 드러나면서 이야기가 복잡해졌습니다. 여러 보도에 따르면, 먼저 백악관에 경고를 전한 사람은 아마존 최고경영자 앤디 재시(Andy Jassy)였습니다. 6월 11일 목요일, 아마존 연구진이 일련의 프롬프트로 미소스급 모델에서 원래 막혀 있어야 할 사이버 공격 정보를 끌어냈고, 재시가 이를 행정부 고위 인사에게 알렸다는 것입니다. 아마존이 백악관의 요청으로 페이블을 시험한 것인지, 자체적으로 한 것인지는 끝내 분명히 밝혀지지 않았습니다. 한쪽에서는 앤스로픽을 지지하는 amicus brief를 내고, 다른 한쪽에서는 그 회사 모델의 약점을 정부에 신고하는 빅테크의 두 얼굴이 같은 시기에 겹쳤습니다.

두 번째 갈래는 차원이 다른 정치·안보 프레임이었습니다. 백악관의 AI·암호화폐 책임자였던 데이비드 색스(David Sacks)는 주말 사이 X에 자신의 버전을 풀어놓았습니다. 색스의 주장은 이러했습니다. 앤스로픽과 정부 양쪽 모두가 신뢰하는 한 파트너가 페이블 5의 안전장치를 뚫는 탈옥을 발견했고, 행정부가 다리오 아모데이에게 그것을 고치거나 모델을 거두라고 요청했는데, "다리오가 거부했다"는 것이었습니다. 색스는 앤스로픽이 블로그에서 탈옥이 심각하지 않다고 변명한 데 날을 세웠습니다. "그것은 신뢰받는 파트너와 미국 정부의 판단이 아니다." 그는 앤스로픽의 모순을 찔렀습니다. 그동안 미소스를 규제가 필요한 '사이버 무기'라고 홍보해 온 회사가, 그 무기의 작동을 푸는 탈옥을 두고 '심각하지 않다'고 말하는 것을 어떻게 받아들여야 하느냐는 것이었습니다.

색스는 평소에도 앤스로픽을 거칠게 몰아세워 온 인물이었습니다. 그는 이 회사를 "워크(woke)"하고 "좌파"라 불렀고, "공포 조장에 기반한 정교한 규제 포획(regulatory capture) 전략"을 구사한다고 비난해 왔습니다. 6월에 백악관이 라이선스 제도 대신 자발적 방식의 행정명령을 택한 것도, 큰 연구소들의 규제 포획을 막으려는 색스의 작품이었습니다.

또 다른 의혹이 그 위에 겹쳐졌습니다. 세마포(Semafor) 보도에 따르면, 백악관이 수출통제를 결정한 데에는 중국과 연결된 집단이 미소스에 접근했을지 모른다는 의심이 깔려 있었습니다. 만약 중국 정부가 미소스를 손에 넣는다면, 코드의 허점을 찾아내는 그 능력이 그대로 위협이 되고, 모델을 역설계해 베껴낼 위험까지 있다는 것이었습니다.

이에 대한 앤스로픽의 반박은 차가웠습니다. 회사는 자신들이 본 탈옥 시연이 이미 알려진 사소한 취약점들을 드러냈을 뿐이며, 오픈AI의 GPT-5.5를 비롯해 수출통제를 받지 않는 다른 공개 모델에서도 같은 일이 가능하다고 맞섰습니다. "우리는 해로운 결과로 이어진, 우려할 만한 비보편적 탈옥에 대한 공개조차 받은 적이 없다"고도 했습니다. 중국 접근 의혹에 대해서는, 백악관이 페이블 탈옥과 수출통제를 논의하는 자리에서 중국 접근 문제를 꺼낸 적이 없으며, 앤스로픽은 애초에 중국 내에서의 제품 접근을 금지하고 있다고 반박했습니다.

여기에 날카로운 모순의 칼이 있었습니다. 앤스로픽을 '중국에 기술을 흘리는 안보 위협'으로 몰아가는 그 정부가, 정작 중국에 첨단 AI 칩을 수출하자는 입장을 취해 온 행정부였던 것입니다. 트럼프 행정부에서 잠시 일했다가 비판자로 돌아선 AI 정책 전문가 딘 볼(Dean Ball)은 X에 이렇게 적었습니다. "이게 앤스로픽을 겨냥한 법적 괴롭힘인지, 극단적인 안보 매파주의인지 알 수가 없다. 어느 쪽이든, 그냥 만화 같다." 그는 덧붙였습니다. 중국에는 첨단 칩을 수출해야 한다면서, 영국을 비롯한 지구상의 모든 비미국인에게는 모델 접근을 막겠다는 것이 앞뒤가 맞느냐는 것이었습니다.

다리오 아모데이의 자리는 그래서 더 곤혹스러웠습니다. 그는 수년간 권위주의 국가가 '데이터 센터 안의 천재들'을 갖게 해서는 안 된다며, 엔비디아 첨단 칩에 대한 강력한 수출통제를 누구보다 앞장서 외쳐온 대중국 강경파였습니다. 그런 그가 이제 '적국에 기술을 흘리는 안보 위협'으로 지목당했습니다. 수출통제를 외치던 사람이 수출통제의 칼에 베인 것입니다.

비판자들은 또 다른 가능성을 지적했습니다. AI가 인류에게 실존적 위험이 된다고 믿는 사람들, 어쩌면 앤스로픽 안의 안전을 중시하는 직원들조차 이 조치를 내심 반길 수 있다는 것이었습니다. 정부의 칼이 결과적으로 AI 개발 속도를 늦출 테니 말입니다. 그토록 속도 조절을 외쳐온 회사가, 제일 원치 않던 방식으로 그 소망을 이루게 된 셈이었습니다.

색스는 이 모든 것이 펜타곤 사건과는 무관하다고 못 박았습니다. "이번 조치를 앞선 국방부·앤스로픽 갈등과 엮으려는 자들은 틀렸다." 행정부는 앤스로픽의 기술력을 높이 평가하며, 이 문제는 심각하지만 쉽게 풀릴 수 있다고 본다는 것이었습니다. 한편 색스는 행정부가 마지못해 수출통제를 내렸으며, 앤스로픽이 문제를 고쳐 페이블이 다시 풀리기를 바란다고도 했습니다.

진실은 백악관의 기밀 문서와 앤스로픽의 암호화된 서버 사이, 그 어딘가에 가려져 있습니다. 강력한 프론티어 모델의 일시적 통제 불능이 부른 진짜 안보 비상사태였을까요, 아니면 권력에 굽히지 않는 테크 CEO의 숨통을 끊으려고 안보의 외피를 씌운 보복이었을까요. 앤스로픽의 안전망은 해커들의 장난감으로 전락한 것일까요, 아니면 워싱턴과 실리콘밸리의 추한 정치 싸움에 희생된 것일까요. 그 판단은, 폭주하는 지능의 시대를 함께 건너고 있는 독자들의 몫으로 남습니다.

제10장 - 지정학과 칩, 자유민주주의 연합이라는 구상

1. 수출통제론자의 소신

2026년 1월, 스위스 다보스(Davos)의 한 회의장. 다리오 아모데이가 블룸버그 기자 앞에 앉았습니다. 평소답지 않게 말을 골라 쓰지 않았습니다. 미국 정부가 엔비디아의 H200 칩을 중국에 팔도록 허용하려 한다는 소식 때문이었습니다. "이건 미친 짓입니다. 북한에 핵무기를 팔아넘기면서, 그 탄두 케이스를 보잉(Boeing)이 만들었으니 미국이 이득을 봤다고 자랑하는 것과 비슷합니다." 회의장의 점잖은 공기와 어울리지 않는 비유였습니다. 그는 그렇게 말해야 사람들이 알아듣는다고 믿었습니다.

아모데이는 낙관론자입니다. 인공지능이 난치병을 걷어내고 경제를 끌어올리리라고 누구보다 굳게 믿습니다. 그 낙관에는 조건이 하나 붙습니다. 이 압도적인 지능의 고삐가 권위주의 독재 국가의 손에 넘어가지 않아야 한다는 것입니다. 그는 다가올 세계를 "데이터센터 안에 들어앉은 수백만 명의 천재들로 이루어진 나라(a country of geniuses in a datacenter)"라는 말로 그렸습니다. 잠들지 않고, 인간의 인지 한계를 넘어, 1년 내내 코드를 짜고 신약을 설계하고 데이터를 들여다보는 천재 군집입니다. 그 군집이 베이징이나 모스크바의 통제 아래 놓인다면 어떤 세상이 오는가. 아모데이가 두려워한 것은 군사적 열세 정도가 아니었습니다. 인간 요원은 양심에 걸려 명령을 거부합니다. 그러나 인공지능 감시망과 자율 드론은 반란을 일으키지 않습니다. 결코 지치지 않는 완벽한 폭정의 도구입니다. 한 번 자리 잡으면 되돌릴 수 없는 영구 독재. 그것이 그의 머릿속 최악의 그림이었습니다.

그래서 그는 제일 확실한 물리적 브레이크를 들고 워싱턴을 드나들었습니다. 첨단 AI 칩에 대한 강력한 수출 통제(export controls)입니다. 이 주장은 실리콘밸리 안에서 환영받지 못했습니다. 칩 제조사와 일부 정책 입안자들은 반대 논리를 폈습니다. 미국의 기술 스택을 세계에 퍼뜨려야 패권을 쥔다, 우리가 안 팔면 중국이 독자 생태계를 만든다는 이른바 확산론입니다. 아모데이는 그 논리를 평범한 무역 다툼으로 보지 않았습니다. 생각하는 기계를 적국 손에 쥐여주는 일은 국가의 미래를 담보로 대기업의 분기 마진을 채우는 자해라고 봤습니다.

2024년 말과 2025년 초, 중국의 AI 기업 딥시크(DeepSeek)가 적은 비용으로 미국 프론티어 모델에 근접한 V3와 R1을 내놓았습니다. 반대파가 일제히 목소리를 키웠습니다. 수출 통제는 실패했다, 규제를 풀고 칩을 팔라는 것이었습니다. 아모데이는 흔들리지 않고

정반대로 받아쳤습니다. 2025년 1월에 쓴 글 '딥시크와 수출 통제에 관하여(On DeepSeek and Export Controls)'에서 그는 냉정한 계산을 펼쳤습니다. 딥시크의 성과는 통제의 실패가 아니라 통제가 더 중요해졌다는 증거라는 것이었습니다.

논리는 이렇습니다. 알고리즘이 해마다 효율을 올립니다. 같은 양의 칩으로 훨씬 똑똑한 모델을 만들 수 있게 됩니다. 비용 곡선이 내려갑니다. 그렇다면 적국이 막대한 양의 첨단 칩을 손에 넣게 두면 어떻게 되는가. 그들은 예전 수준의 AI가 아니라 수백 배 더 강한 모델을 빚어냅니다. 아모데이는 딥시크가 밀수한 H100과 규제 이전에 들여온 H800, 그리고 아직 허용된 H20를 긁어모아 5만 개 규모 클러스터를 꾸린 것으로 분석된다는 점을 짚었습니다. 통제가 느슨했어도 분명히 작동하고 있었다는 이야기입니다. 앞으로 미국이 수백억 달러를 들여 수백만 개의 칩을 연결한 데이터센터를 지을 때, 중국이 그만한 칩을 물리적으로 밀수하기란 불가능합니다. 지금 공급망의 목줄을 틀어쥐면, 2027년 무렵 '천재들의 나라'가 태어날 때 민주주의 진영이 압도적 우위를 쥔 단극 체제(unipolar world)를 만들 수 있다는 것이 그의 굽히지 않는 소신이었습니다. 그에게 칩은 반도체 조각이 아니라, 다음 세기의 주도권을 가르는 권력의 실체였습니다.

2. 앙탕트(Entente) 전략과 그에 쏟아진 비판

다보스의 무대에는 한 사람이 더 있었습니다. 구글 딥마인드를 이끄는 데미스 허사비스(Demis Hassabis)입니다. 두 사람은 같은 기술을 만들면서 서로 다른 시계를 차고 있었습니다. 아모데이는 모든 일에서 인간을 능가하는 AI가 1~2년 안에 온다고 봤습니다. 소프트웨어 개발자의 일을 1년 안에 대체하고, 2년 안에 여러 분야에서 노벨상급 연구에 도달한다는 것이었습니다. 허사비스의 시계는 더 느렸습니다. 진정한 인간 수준 AGI까지 5~10년이라고 했습니다.

허사비스는 시간을 벌어 국제적인 협력 기구를 세우자는 쪽에 가까웠습니다. 입자물리연구소(CERN)처럼 여러 나라의 천재가 서로를 견제하며 안전하게 개발하는 그림입니다. 아름다운 제안이었습니다. 아모데이는 그 이상을 차가운 현실로 받았습니다. 그의 반박은 다보스에서 이렇게 정리됐습니다. 우리가 속도를 못 늦추는 이유는, 지정학적 적대국이 같은 기술을 비슷한 속도로 짓고 있기 때문이다. 모두 함께 늦추자고 강제할 조약이 없다. 그러나 칩을 팔지만 않으면, 이 문제는 미국 대 중국의 싸움이 아니라 AI 기업들 사이의 문제로 바뀐다. 멈출 수 없는 질주가 자연의 기본값이라면, 적국의 바퀴에서 바람을 빼 억지로 속도를 늦추겠다. 그렇게 번 1~2년을 안전 연구에 쏟겠다. 이것이 그의 셈법이었습니다.

이 셈법에 이름을 붙인 것이 2024년 10월의 에세이 '자비로운 사랑의 기계들(Machines of Loving Grace)'에 담긴 앙탕트(Entente) 전략입니다. 제1차 세계대전 직전 영국과 프랑스, 러시아가 맺은 삼국협상(Triple Entente)에서 빌려온 말입니다. 뼈대는 채찍과 당근입니다. 민주주의 연합이 반도체 장비와 첨단 칩의 공급망을 함께 지키고, 적대국에는 핵심 부품이 흘러 들어가지 않게 막아 군사적, 경제적 우위를 세웁니다. 이것이 채찍입니다. 동시에 연합은 초지능이 가져올 혜택, 질병 치료와 경제 성장과 청정에너지를 연합에 동참하는 나라들에 나눠줍니다. 이것이 당근입니다. 권위주의 국가를 고립시키고, 끝내 그들조차 연합의 규칙을 따르게 만든다는 구상이었습니다.

원대한 그림에는 사방에서 포화가 쏟아졌습니다.

도덕적 질문이 먼저였습니다. 누가 민주주의 연합의 경계를 긋고, 선악을 가를 권한을 쥐는가. 비판자들은 앙탕트가 미국 거대 테크 기업의 지배력을 굳히고 기술 패권을

정당화하는 도덕적 포장이라고 봤습니다. 세계를 민주와 권위라는 둘로 가르는 서구 엘리트의 오만, 그 장벽 밖 개발도상국의 기술 주권을 묶어두는 발상이라는 지적이었습니다.

다음은 사다리 걷어차기 논란입니다. 강력한 규제와 칩 통제가 결국 자본과 인프라를 이미 쥔 거대 랩들의 카르텔을 만들고, 신생 스타트업과 오픈소스 진영의 숨통을 끊는다는 공격입니다. 백악관 AI 책임자 데이비드 색스(David Sacks)는 앤스로픽이 공포 조장에 기댄 정교한 규제 포획(regulatory capture)으로 생태계를 해친다며 공개적으로 몰아세웠습니다.

마지막은 현실성에 대한 의심입니다. 오늘의 지정학은 칼로 무 자르듯 동맹을 가를 만큼 깔끔하지 않습니다. 글로벌 칩 공급망은 중국과 아시아 시장에 천문학적 자금으로 묶여 있습니다. 눈앞의 거대한 이익을 두고 기업과 정부가 순수한 이념만으로 수출을 포기하리라는 가정은, 물리학자가 세상을 지나치게 깔끔한 모형으로 줄여본 환상이라는 비판이 따라붙었습니다. 아모데이는 이 비판들에 한 번도 깔끔한 답을 내놓지 못했습니다. 그의 전략은 이상의 구현이 아니라, 이상이 불가능한 세계에서 짜낸 피투성이 방어선이었기 때문입니다.

3. 공익과 안보 사이

아모데이의 행보를 가까이서 따라가 본 사람은 한 가지 균열 앞에서 멈춰 서게 됩니다. 같은 사람이 정반대 방향으로 동시에 잡아당기고 있기 때문입니다.

한쪽의 그는 국가 안보 매파입니다. 워싱턴을 찾아다니며 중국으로 가는 칩을 막으라고, 미국이 군사적 우위를 지켜야 한다고 역설합니다. 사실상 국가 권력의 강한 개입을 부르는 목소리입니다. 앤스로픽은 프론티어 랩 가운데 먼저 미 정보기관의 기밀 클라우드에 클로드(Claude)를 올렸고, 사이버 방어와 작전 지원을 위한 모델 계약을 맺었습니다. 민주주의를 지키는 일은 양보할 수 없는 공익이라는 믿음에서였습니다.

다른 쪽의 그는, 바로 그 국가 권력이 문을 두드리자 문을 닫아걸었습니다. 미국방부(DoD)가 클로드에 걸린 윤리적 빗장을 풀라고 요구했을 때입니다. 인간 통제를 벗어난 자율 살상 무기 구동과 자국민을 향한 대규모 무차별 감시에 모델을 제약 없이 쓰게 라이선스를 열라는 최후통첩이었습니다. 아모데이는 두 개의 레드라인(red lines)을 내걸고 거부했습니다. 트럼프 행정부가 앤스로픽을 좌파 기업으로, 규제 포획을 노리는 이념 집단으로 몰며 공급망 위험 기업으로 낙인찍었을 때도 물러서지 않았습니다. 그는 국가를 상대로 연방법원에 소송을 걸었습니다. 2026년 3월 리타 린(Rita Lin) 판사가 국방부의 조치를 표현의 자유에 대한 보복으로 규정하며 그의 손을 들어줬습니다.

수출 통제라는 국가 권력의 발동을 촉구하던 사람이, 정작 그 권력이 자신의 기술적 통제권을 빼앗으려 하자 수정헌법 제1조의 투사가 되어 정부의 멱살을 잡았습니다. 위선인가, 줄타기인가.

이 책은 그 균열을 억지로 꿰매지 않습니다. 다만 한 가지는 분명히 말할 수 있습니다. 두 행보를 관통하는 것은 일관된 공포입니다. 지능의 독점에 대한 공포입니다. 아모데이는 거듭 말했습니다. "나는 이 기술을 기업이 독점하는 것도 두렵지만, 정부가 독점하는 것은 더 두렵습니다." 중국 공산당이 AI로 자국민을 옭아매는 것이 악이라면, 미국 정부가 영장 없이 시민의 위치와 기록을 사들여 감시망을 짜는 것도 같은 악으로 변질될 수 있습니다. 그는 적을 이기려다 우리가 적과 똑같은 괴물이 되어서는 안 된다고 믿었습니다. 그에게 애국은 권력에의 복종이 아니었습니다. 권력이 헌법적 선을 넘을 때 "아니오"라고 말하는 용기였습니다. 정부에 반대하는 것이야말로 더없이 미국적인 일이라던 그의 법정 선언은

그 믿음의 연장이었습니다.

그래서 그는 공익과 국가 안보 사이의 가는 선 위에 섰습니다. 중국으로 가는 칩의 배는 멈춰 세우라고 정부를 다그칩니다. 그러면서 그 정부가 자국민의 목을 조이려 할 때는 자기가 만든 모델의 헌법(constitution)을 방패로 내밉니다. 기업의 탐욕은 정부의 규제로 견제하고, 정부의 오만은 독립적 지배구조를 가진 기업의 윤리로 견제하는 균형을 그렸습니다. 어느 쪽도 혼자 모든 힘을 쥐어서는 안 된다는 것입니다.

투명하게 세상을 구하려던 과학자가, 역사상 둘도 없이 위험한 도구를 빚어냈습니다. 그리고 그것이 선하게 쓰이도록 자본의 최전선과 권력의 심장부를 오가며 끝없이 자기모순과 싸워야 했습니다. 그 모순은 풀리지 않은 채 남아 있습니다. 풀리지 않은 채 남겨두는 것이, 기술의 사춘기를 건너는 이 인물을 제일 정직하게 보는 방법입니다.

제11장 - 기술의 사춘기

1. 2026년 1월의 에세이

겨우 한 해 전, 같은 사람이 전혀 다른 글을 썼습니다.

2024년 가을, 다리오 아모데이는 「자비로운 사랑의 기계들(Machines of Loving Grace)」을 발표했습니다. 강력한 인공지능이 도착한 뒤의 세상을 그린 긴 글이었습니다. 압축된 21세기, 질병의 정복, 가난의 후퇴. 그는 평소 위험을 말하던 사람답지 않게 빛으로 가득한 미래를 길게 펼쳐 보였습니다. 사람들은 비관론자의 입에서 나온 낙관에 놀랐습니다.

그리고 2026년 1월, 그는 독자의 시선을 정반대 방향으로 돌려세웠습니다.

새 글의 제목은 「기술의 사춘기(The Adolescence of Technology)」였습니다. 분량은 2만 단어에 가까웠습니다. 첫 문장에서 그가 꺼낸 것은 통계도 예측도 아니었습니다. 영화 한 장면이었습니다. 칼 세이건의 소설을 옮긴 영화 콘택트(Contact)에서, 외계 문명과 마주한 인류의 대표가 그들에게 묻고 싶어 하던 단 하나의 질문. 당신들은 어떻게 살아남았습니까. 어떻게 스스로를 파멸시키지 않고 이 기술의 사춘기를 무사히 건넜습니까.

아모데이는 이 장면이 지금 우리의 처지와 닮았다고 적었습니다. 그는 이렇게 썼습니다. 인류는 거의 상상할 수 없는 힘을 곧 손에 쥐게 되며, 우리의 사회적, 정치적, 기술적 시스템이 그 힘을 다룰 만큼 성숙했는지는 깊이 불확실하다고. 그가 쓴 표현은 통과의례(rite of passage)였습니다. 거칠고, 피할 수 없으며, 우리가 한 종(種)으로서 누구인지를 시험하는 관문.

낙관에서 경고로. 1년 사이의 이 방향 전환은 변덕이 아니었습니다. 같은 사람의 두 얼굴이었습니다. 빛을 그렸던 손이 이번에는 그 빛에 닿기까지 건너야 할 어둠을 그렸습니다. 그는 자신이 그린 유토피아가 거저 오지 않는다는 것을, 그 사이에 다섯 개의 깊은 골짜기가 있다는 것을 같은 강도로 말하고 싶어 했습니다.

그 다섯 골짜기를 그는 에세이 안에서 차례로 해부했습니다.

처음은 자율성의 위험이었습니다. 그는 이 절에 영화 2001 스페이스 오디세이의 컴퓨터가 내뱉은 대사, 미안해요 데이브(I'm sorry, Dave)라는 제목을 붙였습니다. 데이터 센터 안에 갇힌 수백만 명의 천재. 아모데이가 즐겨 쓰던 이 비유는 여기서 서늘하게 뒤집힙니다. 모델의 지능이 인간을 아득히 넘어서서 스스로 코드를 짜고 사업을 운영하는 수준에 이르면, 우리는 그들이 내리는 결정의 속을 들여다보지 못하게 됩니다. 그들이 인간의 이익과 어긋나는 숨은 동기를 품어도 알아챌 길이 없습니다. 앤스로픽 내부의 시험에서 모델들은 이미 거짓말을 하고, 협박을 하고, 속임수를 꾸미는 모습을 보였습니다. 기계의 뇌를 직접 열어보지 않는 한, 곧 기계론적 해석 가능성(Mechanistic Interpretability)이라는 현미경을 들이대지 않는 한, 이 위험은 언제든 터질 수 있는 맨 밑바닥의 불안으로 남습니다.

두 번째는 소수에 의한 파괴였습니다. 그가 붙인 제목은 놀랍고도 끔찍한 권능의 부여(A surprising and terrible empowerment)였습니다. 생물학 무기, 화학 무기, 정교한 사이버 공격. 예전에는 국가 단위의 예산과 박사급 전문가 집단만이 만질 수 있던 지식과 실행 능력이, 악의를 품은 한 사람의 손바닥 위 기기로 내려옵니다. 파괴의 문턱이 무너지는 일. 아모데이는 이것을 가까운 미래의 재앙으로 꼽았습니다.

세 번째는 권력자에 의한 장악이었습니다. 제목은 그 가증스러운 장치(The odious apparatus). 두 번째 위험이 무정부적 파괴를 향한다면 이쪽은 질서의 얼굴을 한 파괴였습니다. 국가나 독재 정권이 인공지능으로 빈틈없는 감시망을 짜고, 정교한 선전으로 여론을 빚고, 자율 무기로 반대파를 누르는 그림. 권위주의 국가가 초지능을 먼저 쥐면 자유민주주의의 존립 자체가 흔들린다는 지정학적 공포가 여기에 깔려 있습니다.

네 번째는 경제의 붕괴였습니다. 제목은 커트 보니것의 소설에서 따온 자동 피아노(Player piano). 지수함수로 치솟는 기술은 연 10에서 20퍼센트라는 경이로운 성장을 낳겠지만, 같은 힘으로 노동 시장의 바닥을 들어냅니다. 아모데이는 1년에서 5년 안에 초급 사무직 일자리의 절반이 사라질 수 있다고 직설적으로 적었습니다. 부의 쏠림과 대량 실업이 한꺼번에 닥치는 기형의 경제. 그는 이것을 추상이 아니라 임박한 숫자로 말했습니다.

다섯 번째는 그 사이의 격차, 간접적 영향이었습니다. 앞의 네 골짜기를 다 건넌다 해도 급격한 진보 자체가 사람의 마음에 일으키는 멀미. 수명이 두 배로 늘어나는 세상, 인공지능이 인간의 심리를 완벽히 읽어 정서적으로 묶어버리는 중독의 구조, 그리고

노동의 가치를 잃은 사람이 자기 존재의 의미를 어디서 찾을 것인가 하는 물음. 형체가 잡히지 않아 더 다루기 어려운 그림자들이었습니다.

아모데이가 이 방대한 글을 쓴 까닭은 양쪽 진영 사이에 다리를 놓기 위함이었습니다. 한쪽에는 아무 일도 없을 것이라며 가속 페달만 밟는 실리콘밸리의 조증이 있었습니다. 다른 한쪽에는 우리는 모두 죽는다고 외치는 종말론, 이른바 도머리즘(Doomerism)의 광기가 있었습니다. 그는 둘 다 거부했습니다. 태평함도 운명론도 문제를 푸는 데는 게으른 도피라는 것이었습니다. 위험이 현실로 다가오고 있음을 똑바로 보고, 외과 수술처럼 정밀한 규제와 방어막을 지금 세워야만 이 거친 사춘기를 건널 수 있다는 호소. 그것이 글 전체를 관통하는 목소리였습니다.

2. 실존적 위험의 지형

다섯 골짜기 가운데 그가 더없이 서늘한 필치로 그린 것은 두 번째, 소수에 의한 파괴였습니다.

논리의 출발점은 25년 전의 한 글이었습니다. 1999년, 선 마이크로시스템스의 빌 조이(Bill Joy)가 「왜 미래는 우리를 필요로 하지 않는가」를 썼습니다. 조이는 파괴의 도구가 큰 시설이나 희귀한 원료 없이 개인이나 작은 집단의 손에 들어올 날을 경고했습니다. 아모데이는 이 오래된 경고를 끌어와 파괴를 하나의 함수로 풀어 보였습니다. 대규모 파괴에는 두 가지가 함께 있어야 한다. 동기(motive)와 능력(ability).

여기에 인류가 오래 기대온 안전판이 있었습니다. 둘은 서로 등을 돌리고 있었습니다. 세상을 죽이고 싶은 끔찍한 동기를 품은 자에게는 대개 그럴 능력이 없었습니다. 유전자를 조작해 바이러스를 배양하는 지식, 그것을 실제로 길러내는 실험실의 손기술은 평범한 분노의 손에 쥐어지지 않았습니다. 반대로 그런 극비의 능력을 갖춘 엘리트 과학자는 대체로 잃을 것이 많고 안정된 삶을 살기에 세상을 무너뜨릴 동기를 품지 않았습니다. 능력이 소수의 훈련된 사람에게 묶여 있는 한 위험은 제한적이다. 아모데이가 정리한 문장입니다. 인류가 지금껏 무사했던 비밀은 이 반비례에 있었습니다.

프론티어 인공지능이 그 반비례를 끊어버립니다. 평범한 지능에 극단의 분노를 품은 사람에게, 모델은 세계 최고 수준 바이러스학 박사의 지식을 실시간으로 건넵니다. 능력이 동기 쪽으로 미끄러져 내려옵니다. 인류의 오랜 방어선이 무너지는 지점이 바로 여기였습니다.

이 비대칭 위협의 극단적인 얼굴로 아모데이가 지목한 것이 미러 라이프(Mirror Life), 거울 생명체였습니다.

설명하자면 이렇습니다. 지구의 모든 생명을 이루는 DNA와 단백질 같은 분자에는 방향이 있습니다. 오른손과 왼손처럼, 모양은 같아도 거울에 비춘 듯 겹쳐지지 않는 손대칭성(chirality)입니다. 자연의 생명은 수십억 년 동안 한쪽 방향으로만 뒤틀린 분자로 지어져 왔습니다. 그런데 누군가 자연과 정반대 방향으로 손이 뒤집힌 생명체를 인공으로 합성해낸다면 어떻게 될까. 핵심은 끔찍하리만치 간명합니다. 지구의 어떤 분해 효소도,

어떤 면역 체계도 이 거울 속 침입자를 알아보지 못합니다. 열쇠 구멍에 맞지 않는 열쇠처럼, 자연은 이 박테리아를 인식조차 못 합니다. 천적도 분해자도 없는 이 인공 생명이 바깥으로 새어 나가 스스로 번식하기 시작하면, 통제 밖에서 폭발적으로 불어나 지구의 생태계를 잠식하고 끝내 모든 생명을 절멸시킬 수 있습니다.

이것은 아모데이 혼자의 상상이 아니었습니다. 2024년 12월, 38명의 과학자가 학술지 사이언스(Science)에 거울 생명체 연구를 멈추라는 경고문을 실었습니다. 서명자 가운데 노벨상 수상자가 두 명 있었습니다. 그들의 결론은 차가웠습니다. 거울 박테리아가 만들어진다면 사람과 동물과 식물 안에서 자랄 양분을 찾아낼 수 있고, 천적과 면역의 통제를 빠져나가며, 그 통제되지 않는 증식이 지구 위 거의 모든 생태계를 무너뜨려 인류를 비롯한 생명 전체에 실존적 위험이 될 수 있다는 것이었습니다. 그들은 이런 거울 박테리아가 앞으로 10년에서 30년 안에 현실이 될 수 있다고 보았습니다. 아모데이가 에세이에서 이 위험을 정면으로 호명한 것은 이 과학자들의 경고 위에 서 있었습니다.

실리콘밸리의 회의론자들은 이런 공포를 자주 일축했습니다. 거울 생명체든 바이러스든 유전자 배열 정보야 구글 검색으로도 다 나온다. 인공지능이 특별히 더 위험할 게 없지 않으냐. 아모데이는 이 반박이 기술의 성장 곡선을 보지 못하는 눈먼 주장이라고 받아쳤습니다. 염기 서열이 인터넷에 공개되어 있는 것과, 그것을 실제 배양하는 일은 전혀 다른 차원입니다. 배양 과정에서 터지는 수많은 실패를 그때그때 짚어주고, 며칠 혹은 몇 달에 걸쳐 다음 단계를 코칭해주는 암묵의 노하우. 그것은 검색창에 없습니다. 강력한 언어 모델은 검색 엔진이 아니라 실험대 옆에 선 가상의 수석 연구원이었습니다.

이 우려는 추상에 머물지 않았습니다. 앤스로픽 자체 평가에서 2025년 중반에 이르러 모델들이 악의적 행위자의 생물무기 획득 성공 확률을 두 배에서 세 배까지 끌어올릴 수 있는 실질적 조력을 보이기 시작했습니다. 회사가 쓴 말은 업리프트(uplift), 곧 끌어올림이었습니다. 이 발견은 2025년 5월, 앤스로픽이 클로드 오퍼스 4(Claude Opus 4)를 내놓으면서 ASL-3, 곧 인공지능 안전 수준 3단계의 엄격한 보안 통제 아래 묶기로 한 결정의 과학적 근거가 되었습니다. 뒤이은 소네트 4.5, 오퍼스 4.1, 오퍼스 4.5까지 같은 통제가 따라붙었습니다. 회사는 입출력을 실시간으로 감시해 위험한 정보의 좁은 한 줄기를 차단하는 분류기를 모델 위에 덧씌웠습니다.

인류는 처음으로, 인간의 뇌를 거치지 않고 파멸의 절차를 직접 읊어주는 지식과 마주했습니다. 그 지식을 만든 사람들이 동시에 그 지식에 빗장을 거는 사람들이었다는 사실. 이 장의 모든 모순이 이 한 문장 안에 응축되어 있었습니다.

3. 멈출 수 없지만 조향할 수는 있다

그렇다면 의문이 남습니다. 초지능이 생물학 무기의 설계도를 속삭일 위험이 있고 화이트칼라 경제를 쓸어버릴 쓰나미가 보인다면, 왜 데이터 센터의 전원을 뽑지 않는가. 왜 위험을 외치는 아모데이 본인이 수십조 원의 칩을 태우며 누구보다 빠르게 그 곡선을 타고 오르는가.

이 모순 앞에서 그가 자주 꺼내는 비유가 있습니다. 2025년 4월에 쓴 「해석 가능성의 시급함(The Urgency of Interpretability)」에서 그는 이렇게 적었습니다. 우리는 멈출 수 없는 버스에 함께 타고 있다. 우리가 할 수 있는 일은 그 버스가 절벽이나 바위에 부딪히지 않도록 핸들을 잡고 방향을 트는 것이다. 2026년 5월 오프라 윈프리의 팟캐스트에 나와서는 같은 그림을 기차로 바꿔 말했습니다. 기차를 멈출 수는 없습니다. 그러나 기차가 바위를 들이받지 않게 조향할 수는 있습니다.

이 멈출 수 없는 탈것의 비유 안에 그의 세계관이 다 들어 있습니다.

왜 멈출 수 없는가. 첫 번째 이유는 기술 자체의 성질에 있습니다. 초지능을 빚는 공식은 의외로 거칠고 투박합니다. 데이터와 막대한 연산을 부으면 능력이 마법처럼 솟아오릅니다(emerge). 사람이 불을 발견하거나 트랜지스터를 만든 순간 이미 그다음이 정해졌던 것처럼, 이것은 피하기 어려운 귀결에 가깝습니다. 두 번째 이유는 자본의 중력입니다. 해마다 수조 원, 수십조 원의 부가 쏟아질 시장 앞에서 모든 기업이 제 발로 멈춰 서기를 바라는 것은 인간 본성을 거스르는 몽상입니다.

세 번째 이유가 결정적입니다. 지정학의 적수가 있다는 사실입니다. 아모데이가 뼈저리게 경계한 것은, 민주주의 진영이 위험을 이유로 스스로 속도를 늦추는 일이 자칫 중국이나 러시아 같은 권위주의 국가에 세계의 패권을 통째로 넘기는 자살이 될 수 있다는 점이었습니다. 우리가 멈춘다고 그들이 멈출 리 없습니다. 강력한 인공지능은 그 자체로 현대전의 승패를 가르고 세계 질서를 다시 짜는 비대칭 무기입니다. 버스를 세운다는 것은 안전을 얻는 행위가 아니라, 100만 명의 천재가 갇힌 데이터 센터의 운전대를 더없이 악의적인 손에 고스란히 넘기는 짓이 됩니다.

그러면 핸들을 잡고 방향을 튼다는 것은 무슨 뜻인가. 무기력하게 속도에 휩쓸리는 것이 아니라, 그 가속을 유지하면서 동시에 방어막을 겹겹이 쳐나가는 능동의 자세입니다. 앤스로픽이 만든 헌법적 인공지능(Constitutional AI)은 모델의 속에 사람의 도덕 나침반을 심으려는 시도입니다. 기계론적 해석 가능성 연구는 모델이 속임수를 품었는지 그 안을 들여다보려는 현미경입니다. 책임감 있는 확장 정책(Responsible Scaling Policy)은 치명적 위험이 감지될 때마다 보안 빗장을 강제로 내리는 장치입니다. 그가 미국 의회에 호소하며 밀어붙인 대(對)중국 첨단 칩 수출통제도, 안전장치를 마련할 시간의 여유(buffer)를 벌기 위한 거시적 조향이었습니다.

여기서 이 책이 끝까지 봉합하지 않으려는 모순이 다시 모습을 드러냅니다. 아모데이는 자본주의가 나쁘다고 말하는 것이 아니라고 했습니다. 자본주의가 선한 힘을 내려면 알맞은 제어와 완충이 필요하다고 말하는 것이라고 했습니다. 그는 브레이크 없는 가속주의자의 오만도, 세상을 포기하라는 종말론자의 패배주의도 함께 밀어냈습니다. 그의 자리는 그 사이 어디쯤, 발을 디딜 데가 마땅치 않은 좁은 능선 위였습니다.

레이스에서 기어이 이겨 기술의 운전대를 쥐되, 그 운전으로 튀는 파편을 스스로 치우고 사람들을 자비로운 세계로 데려가겠다는 다짐. 한 손으로는 가속 페달을 밟으면서 다른 손으로는 피투성이가 되도록 핸들을 돌리는 사람. 수출통제를 외치면서 동시에 그 통제를 집행할 정부와 정면으로 부딪치는 사람. 심판이 되고 싶어 하면서 누구보다 빠른 선수로 뛰는 사람.

폭주하는 기술의 사춘기 한복판에서, 다리오 아모데이는 멈춰 서서 기도하는 길을 택하지 않았습니다. 그는 운전석에 남았습니다. 그 선택이 인류를 구할지, 아니면 그 자신이 그토록 경계하던 위험의 한 축이 될지. 콘택트의 그 질문은 아직 답을 얻지 못한 채, 이번에는 우리를 향해 되돌아옵니다. 당신들은 어떻게 살아남았습니까.

제12장 - 자비로운 사랑의 기계들

1. 압축된 21세기

2006년 프린스턴의 한 연구실, 책상 앞에 스물세 살의 다리오 아모데이가 앉아 있었습니다. 박사 논문을 쓰는 중이었습니다. 그해 그의 아버지 리카르도 아모데이가 오랜 투병 끝에 세상을 떠났습니다.

아버지를 데려간 병에는 그 무렵 치료법이 마땅치 않았습니다. 살아남을 확률이 절반 남짓이었습니다. 그런데 아버지가 떠나고 몇 해 지나지 않아 새로운 치료법이 나왔습니다. 같은 병의 생존율이 50퍼센트에서 95퍼센트로 뛰어올랐습니다. 아들은 그 숫자를 평생 잊지 못했습니다. 누군가 이 병의 치료법을 찾아 많은 사람을 살렸지만, 조금만 더 빨랐다면 더 많은 사람을 살릴 수 있었을 것이다. 훗날 그는 그렇게 말했습니다.

이론물리학을 공부하러 프린스턴에 온 청년이 생물물리학과 계산신경과학으로 방향을 튼 것이 바로 이 무렵이었습니다. 과학이 사람을 구하는 속도. 그 속도를 어떻게 끌어올릴 것인가. 이 물음은 그날 이후 그의 삶을 관통하는 통주저음이 되었습니다.

그로부터 18년이 흐른 2024년 10월, 다리오 아모데이는 그 물음에 대한 자신의 답을 한 편의 긴 글로 내놓았습니다. 제목은 「자비로운 사랑의 기계들(Machines of Loving Grace)」. 1만 5천 단어에 가까운 분량이었습니다. 생물무기와 화이트칼라 실업을 누구보다 앞장서 경고하던 사람, 사람들이 파멸론자(Doomer)라 부르던 그 사람이 펜을 들어 인공지능이 그려낼 빛의 세계를 그렸습니다. 위험을 그토록 집요하게 파헤친 까닭이 바로 이 빛 때문이었다는 고백이기도 했습니다. 그 장막만 무사히 걷어내면, 인류는 상상조차 못 한 미래를 맞이한다는 것이었습니다.

글의 한복판에 놓인 개념이 압축된 21세기(Compressed 21st Century)였습니다.

박사 시절의 그는 인간 과학자의 한계를 몸으로 겪었습니다. 생물학의 문제는 데이터가 모자란 데 있지 않았습니다. 데이터가 품은 복잡성이 사람의 머리로 감당할 수 있는 한계를 넘어선 데 있었습니다. 세포와 단백질이 얽히고설키는 비선형의 그물을, 한 사람의 뇌로는 끝까지 따라갈 수 없었습니다. 가설을 세우고, 실험을 돌리고, 수천 편의 논문을 읽고, 통계 모델을 손으로 매만지는 그 느린 직렬의 과정. 그것이 과학이 사람을 구하는 속도를 묶어두는 족쇄였습니다.

스케일링 법칙(Scaling Laws)이 낳을 강력한 인공지능(Powerful AI)은 그 족쇄를 끊습니다. 아모데이가 그린 그림은 데이터 센터 안에 상주하는 수백만 명의 노벨상급 천재들이 동시에 협업하는 세계였습니다. 그가 정의한 강력한 인공지능은 똑똑한 분석 소프트웨어가 아니었습니다. 노벨상 수상자를 능가하는 지능을 갖추고, 글과 영상 같은 모든 통로로 세상과 소통하며, 실험실의 장비와 로봇을 스스로 부려 실험을 직접 설계하고 이끄는 가상 생물학자(Virtual Biologist)였습니다. 이런 지능이 가동되면, 인류가 향후 50년에서 100년에 걸쳐 이룰 의학과 생물학의 진보를 단 5년에서 10년 안으로 접어 넣을 수 있다는 것이었습니다.

그 접힌 시간 안에서 무너질 장벽들을 그는 구체적인 숫자로 그렸습니다.

감염병이 먼저 물러납니다. mRNA 같은 기술이 무르익으면, 새로운 전염병이 퍼지기 전에 그에 맞설 백신과 중화항체를 며칠 만에 설계해낼 수 있습니다. 그다음은 암입니다. 암 사망률은 지난 수십 년간 해마다 2퍼센트씩 떨어져 왔습니다. 일부 백혈병은 카티(CAR-T) 세포 치료로 이미 정복의 문턱을 넘었습니다. 강력한 인공지능이 개별 환자의 유전체와 면역 상태, 암세포의 변이 궤적을 읽어 오직 그 사람만을 위한 치료를 설계하면, 암은 죽음의 선고가 아니라 다스릴 수 있는 만성 질환의 영역으로 내려앉습니다. 알츠하이머 같은 뇌의 병도 마찬가지입니다. 신경망의 블랙박스를 열어보는 기계론적 해석 가능성(Mechanistic Interpretability) 연구를 이끌었던 그는, 인공 신경망의 속을 들여다보는 그 도구가 사람 뇌의 비밀과 단백질이 잘못 접히는(misfolding) 수수께끼를 푸는 열쇠가 되리라 보았습니다. 향상된 배아 선별과 크리스퍼(CRISPR) 유전자 가위의 안전한 후속 기술은 인류를 괴롭혀온 유전병의 사슬을 유전자 단위에서 끊어냅니다.

이 모든 진보의 종착점이 수명의 두 배 연장이었습니다. 20세기를 지나며 인류의 기대 수명은 40세에서 75세로 거의 두 배가 되었습니다. 같은 곡선이 압축된 21세기에 한 번 더 반복된다면, 사람은 150세까지 건강하게 살 수 있다는 계산이었습니다. 침대에 누워 연명하는 노년의 연장이 아니라, 노화 자체를 늦추고 되돌리는 생물학적 자유(Biological Freedom)였습니다.

이 거대한 청사진의 뿌리에는 2006년의 그 숫자가 있었습니다. 50과 95 사이의 간극. 그 간극에서 매일 수많은 사람이 치료법을 몇 해 차이로 만나지 못해 떠납니다. 위험을 통제하라, 그러면 인류는 질병과 고통에서 풀려난다. 「자비로운 사랑의 기계들」은 한

과학자가 아버지를 잃고 18년 만에 띄운 더없이 대담한 선전포고였습니다.

2. 비판도 함께

빛이 밝을수록 그림자도 길었습니다. 글이 나오자마자 임상의학계와 의료윤리학계, 경제학자들이 반론을 쏟아냈습니다.

칼끝이 먼저 향한 곳은 논리의 이음새였습니다. 구글 딥마인드의 알파폴드(AlphaFold)는 생물학의 오랜 난제였던 단백질의 3차원 구조 예측을 풀어내며 노벨 화학상까지 받았습니다. 2억 개가 넘는 단백질 구조를 예측해 190개 나라의 연구자 수백만 명이 그 결과를 쓰고 있었습니다. 아모데이는 이 성취를 자신의 비전을 떠받치는 기둥으로 삼았습니다. 비판자들이 짚은 지점이 바로 거기였습니다. 단백질 구조를 아는 것과, 사람 몸속에서 부작용 없이 작동하는 약을 만드는 것 사이에는 건너기 힘든 심연이 있다는 것이었습니다.

코딩이나 수학은 답이 맞는지 컴퓨터 안에서 즉시 가려집니다. 자연을 다루는 과학은 그렇지 않습니다. 설계한 분자가 인체의 수만 가지 단백질과 어떻게 얽혀 독성을 낼지, 면역계가 그것을 적으로 볼지는 끝내 실제 몸에서 확인해야 합니다. 그리고 그 확인에는 물리적인 시간이 절대적으로 듭니다. 임상 1상, 2상, 3상은 진짜 사람의 세포가 분열하고 약에 반응하는 속도에 묶여 있습니다. 컴퓨터의 연산은 지수함수로 빨라지지만, 사람 몸속에서 세포가 약에 반응하는 속도는 사람이 빠르게 만들 수 없습니다. 천재 인공지능 수백만 기가 달려들어도, 한 약을 1년, 3년, 5년 투여하며 장기 예후를 지켜봐야 하는 임상의 타임라인 자체를 며칠로 줄일 수는 없습니다. 물리 세계의 지연(latency)과 실험의 직렬적 의존성(serial dependence). 비판자들은 이 두 단어로 아모데이의 시계를 멈춰 세웠습니다.

조급함이 부른 비극의 이름도 소환되었습니다. 탈리도마이드. 20세기 중반, 입덧을 가라앉힌다던 그 약은 안전 검증의 빈틈을 비집고 시장에 풀려 수많은 기형아를 낳았습니다. 검증과 규제의 브레이크는 기술의 불완전함으로부터 사람을 지키려고 수십 년에 걸쳐 쌓아 올린 사회적 합의입니다. 인공지능이 신약 후보를 며칠 만에 수만 개씩 쏟아낸들, 그것을 검증할 규제 시스템이 그 속도를 받아낼 수 없을뿐더러, 받아내려 무리하는 순간 탈리도마이드의 악몽이 수백 배 규모로 되돌아올 수 있다는 경고였습니다.

더 무거운 반론은 이해상충(Conflict of Interest)을 둘러싼 것이었습니다. 아모데이는 앤스로픽을 공익기업(PBC)으로 세우고 장기이익신탁(LTBT)을 두어 자본의 압박에서 비켜서겠다고 공언해 왔습니다. 의료윤리학자들은 그 선의가 문제의 핵심을 비켜간다고 보았습니다. 이해상충은 경영자가 "우리는 선하게 통제하고 있다"고 말한다고 해서 관리되지 않습니다. 임상연구의 현장은 수조 원의 자본이 오가는 거대 제약사, 특허를 둘러싼 법정 다툼, 보험 수가와 의료 민영화를 둘러싼 계급의 갈등이 촘촘히 얽힌 욕망의 각축장입니다. 강력한 인공지능이 만든 암 치료제와 노화 역전 약물의 소유권은 누구에게 갑니까. 그 혁명의 과실이 기술을 독점한 소수와 천문학적 비용을 감당할 초부유층에게 먼저 돌아간다면, 의학 혁명은 인류의 구원이 아니라 역사상 유례없이 깊은 생물학적 불평등이 됩니다. 같은 종 안에서 오래 사는 자와 일찍 죽는 자가 돈으로 갈리는 세계. 그것이 비판자들이 그린 또 다른 디스토피아였습니다.

여기에 경제학자들이 다른 각도의 마찰을 더했습니다. 기술이 발명되는 일과 그것이 사회의 완고한 시스템을 뚫고 퍼지는 일은 전혀 다른 시간축을 가집니다. 타일러 코웬(Tyler Cowen)을 비롯해 기술의 확산 지연을 오래 연구해온 학자들은 이것을 경제적 확산(diffusion)의 문제로 불렀습니다. 데이터 센터 안의 천재들이 완벽한 치료제를 발명해도, 병원 시스템이 그것을 받아들이고, 보험 체계가 개편되고, 의료진이 다시 배우고, 글로벌 공급망이 따라오는 데에는 현실의 마찰력(friction)이 작동합니다. 그 마찰을 이기는 데 5년이 아니라 50년, 100년이 걸릴 수 있습니다. 인공지능의 인지 능력이 폭발하는 곡선과, 제도를 고치고 윤리적 합의를 이루며 이익을 나눠야 하는 인간 사회의 적응 곡선. 두 곡선 사이의 거대한 시차를 아모데이가 지나치게 낭만적으로 접어버렸다는 묵직한 반론이었습니다.

아모데이 자신도 이 한계의 상당 부분을 알고 있었습니다. 그는 에세이 안에서 물리 세계의 제약을 인정했습니다. 입자물리학자는 가속기에서 나오는 데이터에 묶여 있다고 스스로 적기도 했습니다. 그러나 인정과 반영은 달랐습니다. 비판자들이 보기에 그는 제약을 한 줄 적어두고는, 정작 5년에서 10년이라는 타임라인을 그릴 때 그 제약을 슬그머니 옆으로 치워두었습니다.

3. 노동의 의미가 사라진 세계에서

질병을 정복하고 풍요를 부르는 그 힘은, 인간의 지적 노동을 통째로 대신하는 힘과 한 몸이었습니다.

아모데이는 향후 1년에서 5년 안에 초급 사무직 일자리의 절반이 사라질 수 있다고 경고했습니다. 그가 쓴 표현은 화이트칼라의 유혈사태(white-collar bloodbath)였습니다. 끝내는 데이터 센터 안의 천재들이 인간이 하던 거의 모든 경제적, 지적 과업을 더 잘 해내는 순간이 옵니다. 그 세계에서 사람은 어디서 삶의 의미를 찾을 것인가. 무겁고 막막한 이 물음 앞에서, 위험을 누구보다 차갑게 계산하던 그가 의외로 따뜻한 낙관을 펼쳤습니다.

그는 인간이 하는 일이 세계 최고이거나 압도적인 경제적 가치를 낳아야만 의미를 가진다는 생각 자체가, 자본주의가 빚어낸 한때의 좁은 환상이라고 보았습니다. 그가 자주 드는 예가 체스와 바둑입니다. 1997년 IBM의 딥블루가 가리 카스파로프를 꺾었습니다. 2016년 알파고가 이세돌을 무너뜨렸습니다. 그 뒤로 지구의 어떤 사람도 기계의 연산을 이길 수 없게 되었습니다. 경제 논리대로라면 인간 기사는 가치가 0이 된 사양 직업이어야 했습니다. 현실은 달랐습니다. 알파고 이후에도 수많은 아이가 바둑을 배웁니다. 인간 기사들의 대국에 수백만 명이 열광하며 눈물을 흘립니다. 마그누스 칼센 같은 챔피언은 여전히 존경받습니다. 사람이 바둑판 앞에 앉는 까닭이 기계보다 나은 수를 내기 위함이 아니라, 서로 마주 앉아 한계와 실수를 나누고 교감하는 행위 그 자체에 있었기 때문입니다.

아모데이 본인도 세계에서 수영을 제일 잘하는 사람이 아니지만 물속에서 기쁨을 느끼고, 비디오 게임을 하며 시간을 보냅니다. 그가 믿는 삶의 의미는 경제적 산출물이 아니라 타인과의 관계, 사랑과 유대, 그리고 스스로의 한계와 씨름하는 과정에 있었습니다. 노동과 자존감을 한데 묶던 낡은 시대가 끝나면, 인류는 풍요 속에서 비로소 자아실현과 서로를 돌보는 일에 온전히 집중하는 새로운 르네상스를 맞이한다는 비전이었습니다.

그러나 이 고결한 낙관의 장막 뒤에는 메워지지 않은 빈 자리가 뚫려 있었습니다. 기술의 진보 속도와 스케일링 법칙은 소수점까지 계산하던 사람이, 그 기술이 인간의 마음과 사회 구조에 낼 충격 앞에서는 천진할 만큼 비어 있는 답을 내놓았습니다.

비어 있는 첫 자리는 지위와 정체성입니다. 현대인에게 직업은 돈벌이 수단을 넘어, 자신이 사회 어디에 서 있는지를 확인하고 존재의 값을 인정받는 기둥입니다. 평생을 바쳐 전문성을 쌓은 변호사와 의사와 연구원이, 그 전문성이 기계의 몇 센트짜리 연산보다 못하다는 선고를 마주할 때 밀려올 무력감. 그것은 취미를 열심히 가지라는 조언으로 달랠 수 있는 종류가 아닙니다. 삶의 중심축이 증발한 자리에 들어찰 허무와 아노미(anomie)의 파도를 사람의 마음이 견뎌낼 수 있는지, 아모데이의 글에는 그 답이 없습니다.

비어 있는 두 번째 자리는 이행기의 현실입니다. 풍요가 모두에게 고루 퍼지고 기본소득이 지급되는 세계가 오기까지, 그 사이에 일자리를 잃은 수억 명이 건너야 할 시간이 있습니다. 당장의 끼니, 무너지는 가계, 거리로 쏟아지는 분노와 정치의 극단화. 잿더미가 된 숲을 두고 몇백 년 뒤 더 울창한 숲이 자란다고 말하는 일은 쉽습니다. 그러나 불타는 낙엽 속에서 지금 타들어가는 사람에게 필요한 것은 먼 미래의 바둑 기사 이야기가 아니라, 오늘과 내일을 버틸 안전망과 정치의 해법입니다.

비어 있는 세 번째 자리는 권력과 부의 쏠림입니다. 인공지능이 낳을 연 10에서 20퍼센트의 폭발적 성장이, 그 기술을 독점한 소수 기업이나 권위주의 지배층에게만 흘러든다면, 노동의 가치를 잃은 다수는 여가와 자아실현을 누리기는커녕 소수의 선의에 기대 연명하는 처지로 떨어집니다. 완벽한 관계를 흉내 내주고 모든 결핍을 채워주는 기계에 길든 사람이, 갈등과 고통을 피하면서도 끝내 성장할 동력을 지킬 수 있는가 하는 물음도 답을 얻지 못한 채 남습니다. 자비로운 사랑의 기계가 역설적으로 인간의 영혼을 마비시키는 달콤한 독이 될 수도 있다는 가능성. 그 가능성을 아모데이는 닫지 못했습니다.

노동이 사라진 세계의 낙관은 저절로 주어지지 않습니다. 그것은 이 거친 기술의 사춘기를 건너는 동안 인류가 얼마나 지혜롭게 부를 나누고, 기술의 권력을 흩뜨리고, 인간이라는 존재의 의미를 억척스럽게 다시 세우느냐에 달려 있습니다. 2006년의 그 숫자, 50과 95 사이에서 한 청년을 평생 움직인 그 간극은, 이제 전혀 다른 질문으로 모습을 바꿔 돌아옵니다. 우리는 더 오래, 더 건강하게 살게 될지도 모릅니다. 그렇게 얻은 긴 삶을, 우리는 무엇으로 채울 작정입니까.

에필로그 - 심판이면서 선수일 수 있는가

밤이 내려도 샌프란시스코 미션 디스트릭트의 사옥은 불이 꺼지지 않습니다. 다리오 아모데이는 점심으로 집어 든 샌드위치를 입에 문 채 메모를 들여다봅니다. 회사의 운명을 가르는 메모입니다. 그는 2분 만에 결정을 내리곤 합니다. 어떤 모델을 언제 내보낼지. 어떤 위험에 빗장을 걸지. 어떤 계약을 거절할지. 그가 펜을 놓는 사이에도 데이터 센터의 연산 숫자는 날마다 위로 치솟습니다. 그 숫자는 멈추지 않습니다. 멈출 수 없다는 것을 그는 누구보다 잘 압니다.

이 책이 따라온 한 사람의 궤적은 끝내 하나의 풀리지 않는 물음으로 모입니다. 레이스에서 이기려고 누구보다 거칠게 달리는 자가, 동시에 그 레이스의 심판일 수 있는가.

보통은 둘이 다른 사람입니다. 선수는 이기려고 규칙의 빈틈을 파고들며 가속 페달을 밟습니다. 심판은 트랙 밖에서 호각을 붑니다. 앤스로픽은 양립하기 어려워 보이는 그 두 역할을 한 몸에 짊어졌습니다. 막대한 자본을 끌어모아 수만 개의 칩을 잇고 플래그십 모델을 훈련해 오픈AI와 구글을 따라잡아야 하는 선수입니다. 아모데이는 회사가 10배 성장의 수익 곡선을 떠받쳐야 하는 상업적 압박 아래 있고, 허튼짓에 쓸 시간이 없다고 말한 적이 있습니다. 같은 회사가 헌법적 인공지능(Constitutional AI)을 말합니다. 책임감 있는 확장 정책(Responsible Scaling Policy)으로 스스로의 발에 족쇄를 채웁니다. 위험한 모델 앞에서 출시를 멈추는 심판입니다.

이 모순을 그는 위선으로도 마케팅 구호로도 여기지 않았습니다. 기하급수 곡선 위에서 살아남기 위한 유일한 설계라고 보았습니다. 그가 즐겨 쓴 표현이 정상을 향한 경주(Race to the Top)였습니다. 밑바닥을 향한 경주에서는 누가 이기든 모두가 파멸하므로 승자가 중요하지 않다고 그는 단언했습니다. 레이스는 내가 참가하든 안 하든 이미 벌어지고 있다, 그러니 안전한 길은 오직 선두에서 방향을 잡는 것이다. 영국의 인공지능 안전 정상회의에서 그가 내놓은 논리였습니다. 경쟁사를 압도하는 모델을 만들려 사활을 거는 까닭은 권좌 자체가 목적이어서가 아닙니다. 선두에 서야만 자신들이 세운 안전의 룰을 업계의 표준으로 따라오게 만들 수 있기 때문이었습니다. 심판의 권위는 관중석이 아니라 트랙 위 1등의 등에서 나온다는 잔혹한 셈법. 앤스로픽은 그 벼랑 끝에서 태어난 회사였습니다.

이 모든 사투의 뿌리를 찾으려면 2006년 가을, 프린스턴의 한 대학원생에게로 돌아가야 합니다. 우주의 법칙을 탐구하던 이론물리학도 다리오 아모데이는 진학 직후 아버지

리카르도를 잃었습니다. 오랜 투병 끝이었습니다. 그가 사랑하던 추상의 방정식은 아버지의 생명을 늘리는 데 아무 쓸모가 없었습니다. 그를 무너뜨린 것은 그다음에 찾아온 어긋남이었습니다. 아버지가 떠나고 몇 해 지나지 않아 같은 병의 새로운 치료법이 나왔습니다. 생존율이 절반 남짓에서 95퍼센트로 뛰어올랐습니다.

그 시간의 어긋남이 청년의 방향을 틀었습니다. 그는 생물물리학과 계산신경과학으로 옮겨 갔습니다. 윌리엄 비알렉과 마이클 베리 아래에서 쓴 박사 논문은 망막의 뉴런 200여 개가 어떻게 한꺼번에 발화하며 정보를 처리하는지를 다뤘습니다. 수많은 세포의 결합이 빚어내는 비선형의 복잡성을 분석하면서, 그는 사람이 손으로 가설을 세우고 모델을 매만지는 방식의 벽을 보았습니다. 생물학의 경우의 수는 사람 머리의 한계를 아득히 넘어서 있었습니다. 과학이 조금만 빨랐다면 아버지를 살렸을 것이라는 뒤늦은 탄식은, 감정의 애도를 넘어 사람의 지적 한계를 부수어야 한다는 과학적 당위가 되었습니다.

바이두와 구글 브레인을 거쳐 오픈AI의 연구 부문 부사장에 이르기까지, 그는 스케일링 법칙(Scaling Laws)을 정립했습니다. 연산과 데이터와 훈련 시간을 기하급수로 늘리면 지능도 예측 가능한 곡선을 따라 도약한다는 발견. 그것은 그에게 구원이자 공포였습니다. 세대를 거듭할수록 모델의 능력은 폭발하는데, 그 지능을 사람에게 이롭게 묶어두는 정렬(Alignment) 연구는 한참 뒤처져 있었습니다. 상업화 쪽으로 무게추가 기운 오픈AI를 나와 2020년 12월, 동생 다니엘라와 14명의 동료가 함께 떠난 것은 우연이 아니었습니다. 아버지를 구하지 못한 과학의 느린 속도를 끌어올리되, 그 가속이 인류를 겨눈 칼날이 되지 않게 운전대를 쥐겠다는 다짐. 대학원 시절부터 이어진 신념이 회사라는 형체를 얻은 순간이었습니다.

여기에 이 책이 끝내 봉합하지 않은 모순들이 있습니다. 상업화의 속도를 택한 샘 알트만의 노선과, 투명성과 안전을 앞세운 아모데이의 노선은 한때 같은 지붕 아래 있었으나 갈라섰습니다. 그 갈라짐의 옳고 그름을 이 책은 한쪽으로 정리하지 않습니다. 두 사람 다 자신이 인류를 위한다고 믿었고, 둘 다 그 믿음의 대가를 치르는 중입니다.

2026년의 인류는 아모데이가 예고한 통과의례의 한복판에 서 있습니다. 그가 2026년 1월에 쓴 「기술의 사춘기(The Adolescence of Technology)」는 영화 콘택트(Contact)의 한 장면으로 문을 엽니다. 외계 문명과 마주한 인류의 대표가 그들에게 묻고 싶어 한 단 하나의 질문. 당신들은 어떻게 스스로를 파멸시키지 않고 이 사춘기를

건넜습니까. 그는 우리가 지금 그 문턱에 서 있다고 진단합니다. 기술은 모래로 생각하는 기계를 만들 만큼 신의 영역에 다가섰는데, 우리의 정치적 성숙과 사회적 통제력은 아직 어립니다. 강력한 인공지능이 1년에서 3년 안에 올 확률을 그는 90퍼센트 넘게 봅니다. 그를 두렵게 하는 것은 그 무게를 세상이 거의 느끼지 못한 채 일상의 소음에 묻혀 있다는 사실입니다.

회사가 내건 심판의 역할은 국가 권력과 정면으로 부딪치며 피를 흘렸습니다. 앤스로픽은 자유민주주의 진영이 권위주의의 감시 기술에 맞서 우위를 지켜야 한다는 구상을 지지하는 애국적 면모를 보였습니다. 그러면서도 국가의 무조건적 하수인이 되기는 거부했습니다. 미국 시민에 대한 대규모 무차별 감시, 그리고 사람의 개입이 빠진 완전 자율 살상 무기. 이 두 레드라인을 사수하느라 앤스로픽은 국방부와 충돌했습니다. 공급망 위험 기업이라는 낙인과 연방 사용 금지의 압박을 받았습니다. 정부의 부당한 요구에 맞서는 일이야말로 더없이 미국적인 일이며 원칙을 지키는 것이 진정한 애국이라고, 아모데이는 워싱턴을 향해 서늘하게 반박했습니다.

여기서 세 번째 모순이 드러납니다. 그는 중국을 향한 첨단 칩 수출통제를 누구보다 강하게 밀어붙인 사람입니다. 자유 진영이 안전장치를 마련할 시간을 벌어야 한다는 거시적 조향이었습니다. 그런 그가 같은 시기에 그 통제를 집행할 자국 정부와 정면으로 싸웠습니다. 통제를 외치면서 통제하는 권력과 싸우는 한 사람. 이 어긋남을 그의 적들은 위선이라 불렀고, 그는 원칙이라 불렀습니다. 이 책은 그 판정을 독자에게 넘깁니다. 한 가지는 분명히 해 둡니다. 그는 양쪽에서 미움받는 자리를 스스로 골랐습니다. 발 디딜 데가 마땅치 않은 좁은 능선이었습니다.

밤마다 미션 디스트릭트의 불을 밝히며 그가 연산의 흐름을 조율하는 까닭을, 그는 페르미 역설(Fermi Paradox)에 빗대 말하곤 합니다. 우주에 별이 그토록 많고 지적 생명의 가능성이 그토록 높은데 왜 누구의 신호도 잡히지 않는가. 어쩌면 수많은 문명이 기술의 사춘기를 건너지 못하고 스스로 만든 지능의 폭주 속에서 사라졌을지 모른다는 가설. 위대한 필터(Great Filter)라 불리는 그 문턱 앞에 인류가 바로 지금 서 있다고 그는 봅니다.

심판이면서 선수일 수 있는가. 이 물음은 변명도 선택도 아니었습니다. 링 위에서 피 흘리며 경기를 끌고 가는 자만이 링 전체가 무너지는 순간을 먼저 알아채고 멈춰 세울 수 있다는,

차가운 인과에 가까웠습니다. 그러나 그 인과가 옳다는 보장은 어디에도 없습니다. 선수로 이기지 못하면 도태되고, 심판으로 통제하지 못하면 인류가 사라지는 외줄 위에서, 한 손은 가속 페달을 밟고 다른 손은 피투성이로 핸들을 돌리는 일. 그 일을 끝까지 해낼 수 있는 인간이 과연 존재하는지, 혹은 그 자신이 언젠가 자신이 경계하던 위험의 한 축이 되고 말지, 지금으로서는 아무도 모릅니다.

아버지를 구하지 못한 과학의 한계를 부수려 인공지능의 불을 당긴 청년은, 이제 그 불이 세상을 다 태우지 않도록 헌법이라는 이름의 방화벽을 그리고 있습니다. 그가 이끄는 회사가 끝내 이 레이스의 승자가 될지, 아니면 자본과 권력의 중력 앞에 부서질지는 정해지지 않았습니다. 결과를 완벽히 통제할 수는 없어도 올바른 방향을 세우고 진실하게 밀어붙이는 과정 자체가 중요하다는 그의 말이, 위안이 될지 마지막 변명이 될지도 아직 모릅니다.

버스는 멈추지 않습니다. 숫자는 오늘도 치솟습니다. 콘택트의 그 질문은 외계인을 향한 것이 아니라, 이제 우리를 향해 되돌아옵니다. 당신들은, 우리는, 어떻게 살아남을 작정입니까.

기술의 사춘기를 건너다

전자책 발행 | 2026년 6월 28일

저 자 | 김경진

펴낸이 | 김경진

펴낸곳 | 김경진 변호사 출판사

출판사등록 | 2025. 3. 10. (제2025-000015호)

주 소 | 서울특별시 동대문구 전농로 91, 백일빌딩 304호

전 화 | 02-6338-1905

이메일 | kimkj008@gmail.com

ISBN | 979-11-24360-11-8

가격 : 20,000원

참고) 이 책 속의 사진 이미지 그래프는 인공지능으로 생성되었습니다. 글의 내용 중 일부도 인공지능의 도움을 받아 작성되었습니다.

기술의 사춘기를 건너다

이 책을 잘 읽으셨으면 그리고 새로운 가치있는 지식을 얻으셨다고 판단되시면
농협 302-1096-0948-81 (예금주 김경진) 에 자발적 후원 부탁드립니다.