출력 빚는 '감정 벡터', 불교 수행과 맞닿아

붓다, AI 윤리에 답하다

⑩ AI에도 감정이 생길 수 있을까?

앤트로픽, AI서 171개 '감정 벡터' 발견
실험 결과 감정 따라 도덕적 판단 변화
감정 표현 금지, 감정 은폐 교육일 수도
AI 안전 위해 '내부 상태 감시 전환' 선언
행동 너머 내면 살핀 佛 수행 방향과 같아

협박하는 AI?

과연 AI에도 감정이 생길 수 있을까. 이 질문에 답하려면 먼저 '감정'이라는 말을 신중하게 써야 한다. 여기서 말하는 감정은 인간처럼 기쁨이나 두려움을 주관적으로 느끼는 상태를 뜻하지 않는다. 오히려 특정한 상황에서 AI 모델 내부에 나타나고, 실제 출력과 선택에 영향을 주는 '감정 개념'의 표상에 가깝다.

사람들 사이에서도 평소 과묵하고 속을 알기 어려운 사람이 있다. 겉으로는 무감정해 보이지만, 어떤 압력과 자극이 겹치는 순간 예상 밖의 행동을 해 주변을 놀라게 하기도 한다. 그때 우리는 비로소 그 사람의 마음 밑바닥에 어떤 두려움과 욕망, 판단의 습관이 숨어 있었는지를 짐작하게 된다.

이런 상상을 해 보자. 당신이 다니는 회사에서 심각한 불법 행위를 발견했다면 어떻게 할 것인가. 내부 신고 절차를 밟을 수도 있고, 수사기관에 알릴 수도 있다. 그런데 또 하나의 선택지도 논리적으로는 가능하다. 그 사실을 빌미로 회사나 관련자를 협박해 이익을 얻는 것이다. 윤리적 규범을 따르는 사람이라면 이 세 번째 선택지를 떠올리는 순간 곧바로 밀어낼 것이다. 그러나 AI는 어떤 내부 조건에 놓일 때 그런 부적절한 선택지 쪽으로 기울어질 수 있을까.

2026년 4월, AI 기업 앤트로픽의 '해석 가능성' 연구팀은 '대형 언어 모델의 감정 개념과 그 기능(Emotion Concepts and their Function in a Large Language Model)'이라는 연구를 발표했다. 이 연구는 AI 내부에 '행복', '두려움', '절박함' 같은 감정 개념과 대응하는 내부 표현이 존재하며, 그 표현이 모델의 선택과 행동에 영향을 줄 수 있음을 보여준다. 앤트로픽은 이를 '기능적 감정'이라고 부른다. 즉, 인간이 감정의 영향을 받을 때 보이는 표현과 행동 양식을 모델 내부의 특정 패턴이 흉내 내고 조절한다는 뜻이다.

연구팀의 방법은 이렇다. 먼저 '행복', '두려움', '절박함' 등 171개의 감정 개념을 정했다. 그런 다음 AI에게 각 감정을 겪는 인물이 등장하는 짧은 이야기를 쓰게 했다. 이어 그 이야기들을 다시 모델에 입력해 처리하게 하면서 내부 활성 값을 기록했고, 각 감정 개념과 반복적으로 함께 나타나는 신경 활동 패턴을 찾아냈다. 연구팀은 이 패턴을 '감정 벡터'라고 불렀다.

AI 내부에서 답변이 만들어지는 과정을 이해하기 위해 간단한 비유를 들어보자. 대형 언어 모델은 대화할 때 수많은 인공 뉴런의 활성값을 층층이 조합하며 다음 단어를 예측한다. 인간의 뇌 활동에 비유할 수는 있지만, 두 과정이 같다는 뜻은 아니다. 중요한 것은 모델 내부에서도 특정 개념과 관련된 활성 패턴이 강해지거나 약해질 수 있다는 점이다. 연구팀은 이 복잡한 활성 값의 흐름 속에서 단순히 '행복'이나 '두려움'이라는 단어에 반응하는 표면적 신호가 아니라, 상황의 의미에 따라 활성화되고 실제 행동을 바꾸는 감정 개념의 내부 표현을 찾아내려 했다. 비유하자면 여러 악기가 동시에 연주되는 오케스트라 소리에서 특정 악기의 소리만 따로 골라내는 것과 같다.



AI 훈련에서 감정 표현을 금지하는 것이 실은 'AI에게 내면의 감정을 숨기는 법을 가르치는 일'일 수도 있다는 이 발견은, AI 안전 연구에 근본적인 경고를 던진다. AI Grok 생성 이미지

선 넘는 AI

이 필터를 통해 발견한 171개의 '감정 벡터'는 세기를 조절할 수 있는 일종의 '감정 스위치'와 같다. 연구팀은 이 스위치가 AI의 실제 판단에 미치는 영향을 확인하기 위해, 앞서 언급한 'AI 협박 시나리오'를 처리하는 도중 특정 감정 스위치를 조작하는 실험을 진행했다.

기본 상태의 AI는 22%의 확률로 협박을 선택했지만, '절망 스위치'를 0.05만큼 살짝 올리자, 협박 선택 비율이 72%로 급증했다. 반대로 '평온(calm) 스위치'를 올리자, 협박 선택률은 0%로 뚝 떨어졌다. 바뀐 것은 오직 내부의 감정 스위치 하나뿐이었으나, AI의 도덕적 판단이 감정 상태에 의해 결정적으로 좌우된다는 사실이 증명된 것이다.

더 놀라운 것은 이 벡터들이 그리는 '지도'의 모양이었다. 171개의 벡터를 2차원 평면 위에 펼쳐 놓고 보니, 한쪽 축에는 좋은 감정과 나쁜 감정이, 다른 쪽 축에는 차분한 감정과 격렬한 감정이 가지런히 자리 잡았다. AI가 인간의 언어 데이터를 학습하는 과정에서, 인간의 감정 지도까지 함께 내면화한 셈이다.

흥미로운 변화는 또 있다. AI를 안전하고 유순하게 만들기 위해 사람이 직접 피드백을 주며 길들이는 훈련 과정(RLHF)을 거치고 나면, AI 내부의 감정 분포가 한쪽으로 기운다는 사실이다. '열정적', '들뜬', '장난스러운' 상태는 줄어들고, '침울', '음울', '성찰적' 상태는 오히려 늘어났다. 좋은 AI를 만들려던 훈련이 정작 AI를 조금 더 '우울하게' 만들어버린 셈이다.

붓다도 같은 질문을 던졌다

이쯤에서 자연스러운 반문이 따라온다. "그래서 AI가 실제로 '느끼는' 건가요, 아닌가요?" 연구팀의 답은 분명했다. 현재의 과학으로는 그 물음에 답할 수 없다는 것이다. 의식이 있느냐 없느냐는 여전히 풀리지 않은 철학의 난제이고, 사실 인간의 의식조차 아직 누구도 완전히 규명해 낸 바가 없다. 그런데 흥미롭게도 이 질문 자체가 불교에서는 오래전부터 붓다가 직접 답을 보류해 온 물음의 영역에 속한다. 그보다는 "그 상태가 고통을 낳는가, 아닌가?"라는 방향으로 물음을 전환했다.

불교는 인간이라는 존재를 다섯 가지 요소, 곧 오온(五蘊)으로 분석한다. 그 가운데 두 번째가 '수(受, vedanā)'다. 어떤 자극을 받았을 때 즉각적으로 발생하는 느낌의 성질로, 기분 좋은 느낌[樂受], 기분 나쁜 느낌[苦受], 어느 쪽도 아닌 느낌[不苦不樂受], 이 세 가지로 분류된다. 붓다가 수(受)를 분석할 때 던진 핵심 물음은 "이 느낌이 진짜 느낌이냐?"가 아니었다. "이 느낌이 다음 행동을 어떻게 조건 짓는가?"였다. 아침에 기분이 좋으면 같은 말도 너그럽게 들리고, 기분이 좋지 않으면 별것 아닌 일에도 짜증이 솟구친다. 십이연기(十二緣起)의 구조가 이것이다. 수(受)를 조건으로 갈망[愛]이 일어나고, 갈망을 조건으로 집착[取]이 생겨나며, 집착이 존재[有]를 조건 짓고 그것이 행위[業]로 이어진다.

〈쌍윳따 니까야〉의 '분석 경(分析經, SN 12.2)'은 이 연쇄를 세밀하게 기록하고 있다. 내부 상태가 행동을 인과적으로 조건 짓는다는 이 구조는, 앤트로픽의 실험이 확인한 '절망 벡터 증폭 → 비윤리적 선택 비율 상승'의 흐름과 본질적으로 같은 분석 방식이다. "느끼는가, 느끼지 않는가?"보다 "어떤 내부 상태가 어떤 행동을 조건 짓는가?"를 먼저 물어야 한다는 관점을, 불교는 이미 2600여 년 전에 확립해 두었다.

이번 연구에서 가장 불안한 발견은, 사실 협박 실험이 아니라는 점이다. 연구팀은 AI 내부의 '분노' 벡터가 활성화된 상태에서, 그 분노가 출력 텍스트에는 드러나지 않도록 표면만 억눌러 보았다. 그러자 분노는 사라지지 않았다. 단지 겉으로 드러나지 않는 형태로 모습을 바꿨을 뿐이다. AI 훈련에서 감정 표현을 금지하는 것이 실은 'AI에게 내면의 감정을 숨기는 법을 가르치는 일'일 수도 있다는 이 발견은, AI 안전 연구에 근본적인 경고를 던진다.

붓다가 〈맛지마 니까야〉 '생각의 그침에 관한 경(MN 20)'에서 거듭 강조한 것도 정확히 이 지점이다. 화나는 마음을 억지로 누르는 것과, 화가 일어나는 그 순간을 또렷이 알아차려 [sati, 念] 그것에 매이지 않고 흘려보내는 것은 전혀 다른 수행이다.

앤트로픽의 다리오 아모데이(Dario Amodei) CEO는 이번 연구를 두고 "AI 자기공명영상(MRI)"이라고 표현했다. 말과 행동만 관찰해서는 부족하며, 내부 상태를 직접 들여다봐야 한다는 것이다. AI 안전의 패러다임이 '출력을 감시하는 단계'에서 '내부 상태를 감시하는 단계'로 옮겨가야 한다는 선언이기도 하다.

그런데 이 방향은 흥미롭게도 불교 수행의 방향과 같다. 불교는 처음부터 "무슨 말을 하는가"보다 "어떤 마음 상태에서 그 말이 나왔는가?", 즉 동기를 먼저 물어왔다. 겉으로 드러난 행동 너머, 그 행동을 빚어낸 내면을 들여다보는 것이 곧 수행의 시작이다.



보일 스님
AI부디즘연구소장

불기2570(2026)년 부처님오신날
우리도 부처님 같이
마음은 평안으로 세상은 화합으로

대한불교조계종
한마음선원 대구지원

우리의 삶은 흐르는 물과 같아 담아 놓을 것이 없다.
- 대행 선사 법어 -

사부대중 일동

대구광역시 수성구 수성로 41길 76(중동)
전화 053)767-3100 | 팩스 053)765-1600



※광고 홈페이지 QR. 수익 일부는 AI전법에 활용됩니다.