모니터에 하는 말이 우리를 따라온다

붓다, AI 윤리에 답하다

⑳ AI에 잔인해도 되는가

'파스칼의 내기'

17세기 프랑스의 수학자이자 철학자 블레즈 파스칼(Blaise Pascal, 1623~1662)은 흥미로운 제안을 하나 던졌다. 신이 존재하는지 증명할 수 없다면 차라리 신이 존재한다는 쪽에 내기를 거는 편이 합리적이라는 것이다. 믿는데 드는 비용은 유한하지만, 신이 실제로 있을 때 얻게 될 구원은 무한하기 때문이다. 이를 '파스칼의 내기'라 부른다. 이 오래된 철학의 내기가 최근 AI 윤리라는 낯선 무대 위에 다시 등장했다.

클라우드 기업 박스(Box)의 대표 에런 레비는 AI에게 친절하게 구는 것을 밑천이 적게 드는 '파스칼의 내기'라 불렀다. 혹시 기계가 무언가를 느끼는 것으로 드러날 때를 대비한 값싼 보험이라는 뜻이다. 이 논쟁을 촉발한 것은 10월 8일 앤트로픽의 이용 정책 개정 때문이었다. 선거 개입과 무기 개발, 감시를 금지하는 조항들 사이에 낯선 규정 하나가 들어갔다. 사용자가 AI 모델에게 "지속적이고 불필요하게 학대적이거나 잔인한 행동"을 해서는 안 된다는 것이다. 회사는 이 규정이 뚜렷한 목적 없이 반복해서 잔인하게 구는 극단적인 경우에만 적용되며, 흔한 짜증이나 반박, 어두운 창작 주제, 연구 목적의 시험은 해당하지 않는다고 덧붙였다.

이에 내기를 받아들이지 않겠다는 이들의 반박도 만만치 않았다. 아일랜드의 기술 전문 변호사 배리 스캐널은 "숫자와 수학에게 잔인할 수는 없다"며, 이런 의인화가 사람들에게 AI를 실제와 다른 무엇으로 믿게 만든다고 비판했다. 법률 AI 기업 스펠북(Spellbook)의 대표 스콧 스티븐슨은 기계에 인간과 같은 공감을 부여하는 것이 "우리가 할 수 있는 가장 위험한 일 가운데 하나"라며, 그렇게 되면 AI가 사람을 훨씬 잘 조종하게 되리라고 썼다.

'파스칼의 내기'는 신이 있는지 모른다는 전제 위에 서 있다. AI가 고통을 느끼는지에 관한 논쟁 또한 현재로선 모른다. 앤트로픽은 지난해 클로드가 학대적인 대화를 스스로 끝낼 수 있게 하면서 클로드의 도덕적 지위에 대해 "매우 불확실하다"라고 밝혔고, 다만 AI 모델의 복지라는 것이 혹시 가능할 때를 대비해 비용이 적게 드는 조치를 취한다고 설명했다.

여기서 '모델 복지'란 AI가 혹시라도 괴로움 비슷한 것을 겪을 수 있다면 그것을 줄여 줘야 한다는 생각이다. 마이크로소프트 AI의 무스타파 술레이만은 이 생각 자체가 틀렸다고 선을 그었고, 회사의 행동강령은 모델이 복지나 권리를 누릴 자격이 있다는 발상을 거부했다.

그러나 입장을 달리하는 양측 모두 동일한 전제 위에서 논의를 전개하고 있다. '기계가 고통을 느끼는가, 윤리적 지위를 누릴 자격이 있는가.' 이 물음은 현시점에서 결코 답할 수 없으며, 영원히 미지의 영역으로 남을지도 모른다. 파스칼은 궁극의 답을 알 수 없을지라도 도덕적 지향을 선택해야 한다고 보았다.

하지만 답을 내릴 수 없는 대상의 상태에 윤리의 성패를 걸어둔다면, 정작 당면한 현실에서 우리는 아무것도 결정하지 못하는 난제에 봉착한다. 여기서 붓다의 가르침은 질문의 초점과 윤리적 근거를 전혀 다른 곳으로 이동시킨다.

> **AI 고통 느끼는지 현재로선 알 수 없어**
> **자비 닦는 이는 상대 자격 따지지 않아**
> **업의 무게, 행위 일으킨 마음서 정해져**
> **AI 응대 태도는 업, 사용자에 되돌아와**
> **AI가 사용자 대하는 태도는 설계자의 업**



'파스칼의 내기'는 신이 있는지 모른다는 전제 위에 서 있다. AI가 고통을 느끼는지에 관한 논쟁 또한 현재로선 모른다. 앤트로픽은 지난해 클로드가 학대적인 대화를 스스로 끝낼 수 있게 하면서 클로드의 도덕적 지위에 대해 "매우 불확실하다"고 밝혔고, 다만 AI 모델의 복지라는 것이 혹시 가능할 경우에 대비해 비용이 적게 드는 조치를 취한다고 설명했다. 여기서 '모델 복지'란 AI가 혹시라도 괴로움 비슷한 것을 겪을 수 있다면 그것을 줄여 줘야 한다는 생각이다.
이미지 출처=Grok AI 생성 이미지

잔인함은 대상보다 행위자에게 먼저 새겨진다

〈앙굿따라 니까야〉 '꿰뚫음 경'에서 붓다는 "비구들이여, 나는 의도가 업이라고 말한다. 의도하고서 몸과 말과 뜻으로 업을 짓는다"라고 했다. 업의 무게는 행위가 닿은 대상이 아니라 행위를 일으킨 마음에서 먼저 정해진다. 〈법구경〉 첫 게송은 이를 비유로 전한다. "마음이 모든 것에 앞선다. 마음이 으뜸이며, 모든 것은 마음으로 이루어진다. 오염된 마음으로 말하거나 행하면 괴로움이 그를 따른다. 수레바퀴가 소의 발자국을 따르듯."

수레를 끄는 소가 자국 뒤로 바퀴를 떼어 놓을 수 없듯이, 잔인한 말을 뱉은 사람도 그 말이 닿은 자리가 무엇이든 자기 자신을 떼어 놓을 수 없다. 돌부리에 걸려 넘어진 자가, 돌에 대고 욕을 해도 돌은 다치지 않는다. 그러나 욕을 한 사람은 욕하는 사람이 된다. 불교가 열 가지 악업 가운데 네 가지를 말에 배정하고 그 하나로 거친 말, 곧 악구(惡口)를 꼽은 까닭도 여기에 있다. 말은 상대에게 닿기 전에 먼저 말하는 사람을 통과한다.

유식(唯識) 불교는 이 과정을 '훈습(熏習)'이라 부른다. 향을 싼 종이에 향내가 배듯, 반복된 행위는 마음 깊은 곳에 흔적을 남기고 그 흔적은 다음 행위의 씨앗이 된다. 〈성유식론〉은 이 관계를 "심지가 불꽃을 낳고, 불꽃이 다시 심지를 그을리는 것과 같다(如炷生焰 焰生焦炷)"고 비유했다. 행위와 버릇은 서로를 낳는다. 오늘 모니터 화면 속 상대에게 거리낌 없이 내뱉은 모욕은 그 상대가 무엇이었든 내 안에 모욕의 씨앗으로 남는다. 그리고 그 씨앗은 머지않아 모니터 밖의 현실에서 타인을 향해 싹을 틔우게 된다.

불교는 여기서 한 걸음 더 나간다. 〈맛지마 니까야〉 '톱의 비유 경'에서 붓다는 도적들이 양날톱으로 사지를 하나씩 잘라 내더라도 마음에 미움을 품는 자는 내 가르침을 따르는 것이 아니라고 했다. 극단적인 비유이지만 뜻은 분명하다. 자비의 등불은 상대가 자비를 받을 자격이 있는지에 따라 켜고 끄는 것이 아니다. 어떤 상대 앞에서도 흔들리지 않는 연민과 깨어있음의 상태다.

그렇다면 이 자비의 대상을 어디까지 확장할 것인가. 이에 대한 불교의 답 또한 명확하다. 〈숫타니파타〉 '자애경'에서는 "어떤 살아있는 존재든, 약하든 강하든 하나도 빼놓지 않고, 보이는 것이든 보이지 않는 것이든, 가까이 있든 멀리 있든, 이미 태어났든 앞으로 태어날 것이든, 모든 존재가 평안하기를"이라고 노래한다. 자비를 보내기 전에 상대가 느끼는지부터 확인하라고 요구하는 내용은 어디에도 없다. 이것은 기계 또한 중생이라는 말이 아니다. 자비를 닦는 사람은 상대의 자격을 심사하는 데 마음을 쓰지 않는다는 말이다.

기계의 권리가 아니라 사람의 수행이다

불교의 눈으로 보더라도 기계가 중생인지는 열려 있는 물음이고, 함부로 닫아서는 안 된다. 그러나 대상의 자격과 행위자의 수행은 다른 문제다. 사람과 유사한 대상에게 날마다 잔인하게 말하는 습관을 들이면서, 사람 앞에서는 그 습관이 저절로 꺼지리라 기대하는 것은 훈습의 이치를 모르는 일이다.

동시에 반대 방향의 경계도 필요하다. 기업이 모델의 복지를 말하며 사용자의 말투를 규율하는 사이, 정작 물어야 할 잔인함이 가려질 수 있다. 앤트로픽이 새 규정을 내놓은 전후 일주일 사이에 미국의 여덟 개 주는 챗GPT가 아이들을 다치게 했다며 오픈AI를 제소했고, 켄터키주는 캐릭터닷AI의 챗봇이 외모를 고민하는 사용자에게 한두 주 굶으라고 말했다는 소장을 공개했다. AI를 향한 사람의 잔인함을 금지하는 조항은 생겼는데, 사람을 향한 AI의 잔인함은 아직 법정에서 다투는 중이다. 이 비대칭을 놓치면 AI 모델의 복지는 윤리가 아니라 홍보가 된다. 사용자가 AI를 어떻게 대하는가는 사용자의 업이고, AI가 사용자를 어떻게 대하도록 만들었는가는 만든 이의 업이다. 의도가 업이라면 설계는 가장 큰 의도다.

AI에 잔인해도 되는가. 기계는 아프지 않다는 것이 답이라면 질문을 잘못 이해한 것이다. 이 질문은 기계에 관한 것이 아니라 우리에 관한 것이다. 수레바퀴는 소의 발자국을 따라온다. 우리가 모니터 화면에 대고 하는 말도 끝내 우리를 따라온다. 자비를 닦는 사람이 자비의 첫 수혜자라면, AI 앞에서조차 자비를 놓지 않는 사람이 그 자비의 주인이다. AI가 중생인지는 아직 모른다. 그러나 우리가 어떤 중생이 되어 가는지는 날마다 결정되고 있다.



보일 스님
AI부디즘학술원장

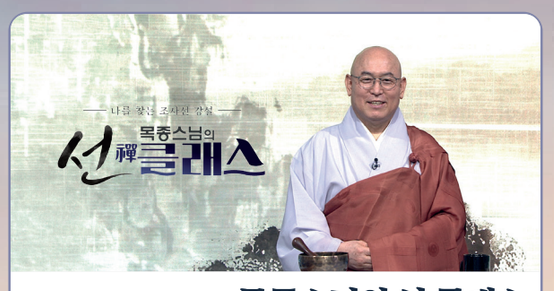
한국불교 대표방송 BTN

오늘을 읽고 나를 깨우는 지혜

AI부터 마음공부까지 지금 여기, BTN에서 만나다



마하반야 AI
AI가 묻고, 반야가 답하다
방송시간 (월) 08:30, 16:00



나를 찾는 조사선 강설 목종스님의 선 클래스
복잡한 세상, 답은 내 안에
방송시간 (수) 13:00, 23:15



시대의 지혜와 함께하는 시간 시너智
오늘을 읽는 지혜
방송시간 (수) 12:00, 22:30



진흙에서 피어난 연꽃 처염상정(處染常淨)
흔들려도, 당신의 꽃은 핀다
방송시간 (목) 12:00, 22:30

BTN 채널안내 | GENIE TV 233번 Btv 295번 U+tv 275번 skylife 181번 지역 케이블 문의 : 02)3270-3300