

이준기의 빅데이터

선악과 먹은 것처럼, 인간 윤리·가치 따르는 AI 나온다

이준기
연세대 정보대학원 교수



인공지능(AI)은 이제 우리 일상에 깊숙이 들어왔다. 이메일을 대신 써주고 문장을 다듬어주며 아이들의 숙제를 도와주는 건 물론이고, 복잡한 보고서를 요약해주고 새로운 광고 기획안을 만들어주기도 한다. 회의록을 자동으로 정리하고 제품 홍보 영상을 생성해주는 데에도 활용된다. 요즘은 단순히 일을 도와주는 수준을 넘어 인생의 고민을 들어주고 진심 어린 조언을 해주는 디지털 동반자이자, 궁금한 것이 있을 때 곧바로 물어볼 수 있는 내 옆의 전문가 역할을 하고 있다.

고시 1등 의사도 처음엔 환자 이해 서툴러 하지만 이렇게 똑똑해 보이는 인공지능에도 근본적인 한계가 있다. 바로, 사용자의 목적이나 맥락에 맞게 충분히 조정되어 있지 않다는 점이다. 예를 들어 아내에 대해 상담을 하면서도 나의 나이, 성별, 처한 상황 같은 핵심 정보를 반영하지 못하고 모든 사람에게 비슷한 방식으로 대답하는 경우가 많다.

최근 방영된 TV 드라마에서 전공의들의 성장을 다룬 이야기에 등장하는 '김사비(사진)'라는 인물은, 결국은 수술을 받아야 된다는 것을 알지만 가족들에게 미안하고 부담되는 듯하여 수술 동의를 선택 못하는 어르신들을 이해할 수 없다. 의대 수석 출신에 국가고시 1등이라는 천재지만, 환자의 감정을 이해하지 못하고 환자와 어떻게 대화를 나눠야 할지 몰라 당황한다. 이는 바로 지식은 풍부하지만, 상황에 맞게 행동은 하지 못하는 오늘날의 인공지능을 보는 듯한 장면이다.

기업들도 마찬가지다. 인공지능을 도입하더라도 우리 조직의 내부 규정, 고객 응대 방식, 고유의 철학과 지향점을 충분히 반영하지 못하면 실무에 제대로 쓰기 어렵다. 마치 모든 교과서와 백과사전 내용을 통째로 외우고 있는 누군가가, 기업이 원하는 방식대로 말하거나 행동해주지 않는 상황과 같다.

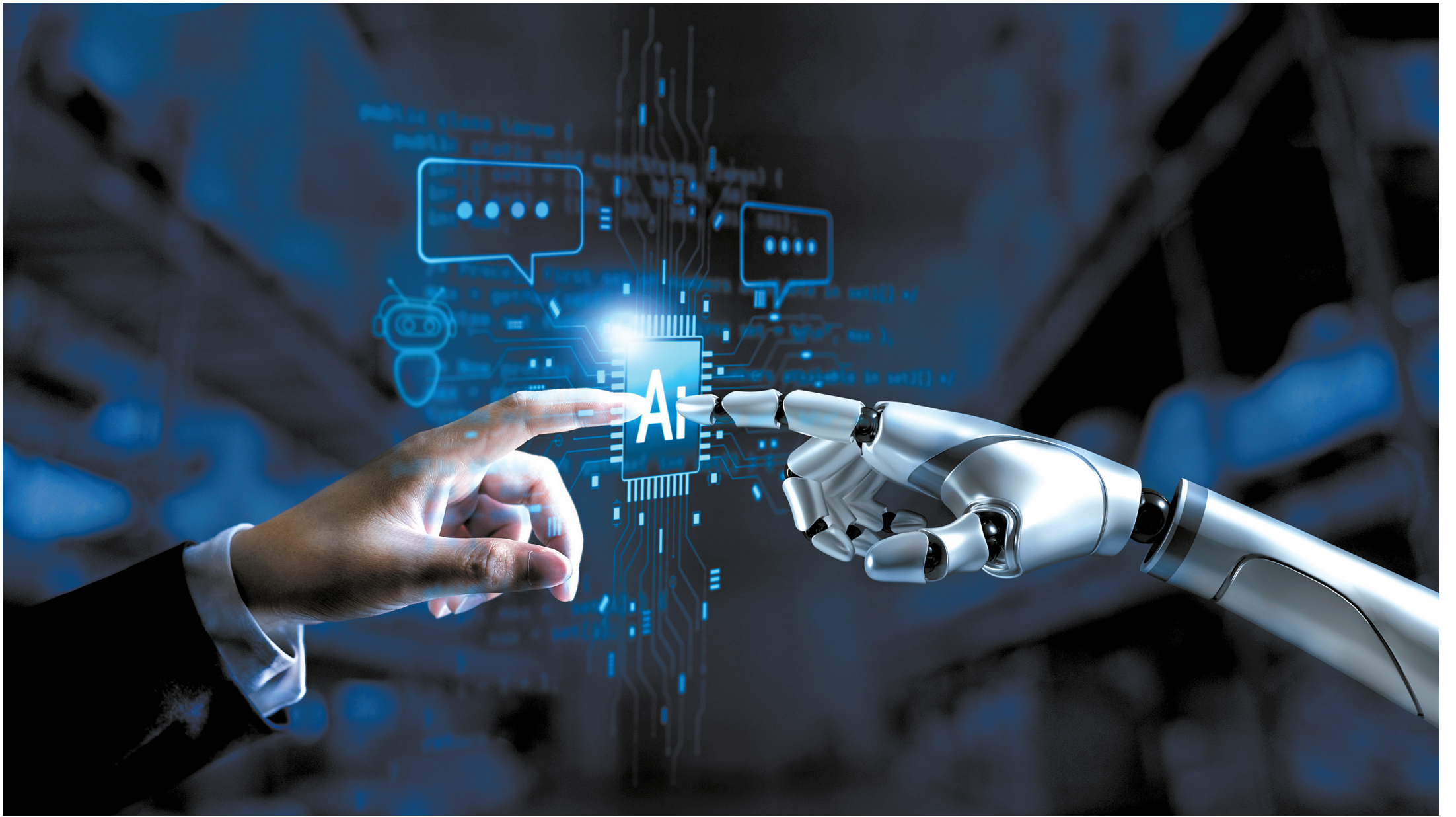
최근의 인공지능 연구는 이 같은 한계를 극복하고, AI가 특정 상황과 가치를 반영해 개인화된 방식으로 작동할 수 있도록 만드는 방향으로 나아가고 있다. 그 중심에는 세 가지 대표적인 기술이 있다. 바로 문맥 학습(In-Context Learning), 검색 증강 생성(RAG, Retrieval-Augmented Generation), 그리고 강화 학습 기반 미세조정(RFT, Reinforcement Fine-Tuning)이다. 이 기술들은 단순한 기능 향상이 아니라, AI가 우리 사회의 규칙과 맥락을 배워가는 과정이다. 흥미롭게도 그 여정은 인간이 윤리를 배우고 책임을 지는 존재로 성장해 가는 모습과 닮아 있다.

문맥 학습은 프롬프트에 예시를 제공하여 인공지능이 문맥을 이해하고 그에 따라 응답하게 하는 방식이다. 예를 들어, 인공지능에 “밥 먹었니?” → “식사하셨어요?”와 같은 예시를 몇 개 제공하면, 다음에 “잘 잤니?”라고 묻는다면 “안녕히 주무셨어요?”라는 식으로 응답하게 된다.

가장 간단한 형태의 훈련 방식으로, 이메일 작성이나 고객 응대와 같이 일정한 어투나 스타일을 요구하는 업무에서 유용하게 쓰인다. 단점은 인공지능이 그 문맥을 오래 기억하지 못하고 다음 세션에는 다시 가르쳐야 한다는 점이다. 또한 예시가 명확하고 일관되어야 좋은 성능이 나온다.

두 번째 방식인 검색 증강 생성(RAG)은 최근 기업들이 가장 활발히 도입하는 기술이다.

예를 들어 “우리 직원 김모씨가 연차 휴가를 OO 방식으로 사용하려고 한다. 현재 우리 회사 규정상



이게 가능한가”라는 질문에 대해 일반적인 거대언어모델(LLM)은 사내 규정을 모르기 때문에 정확한 답을 할 수 없다. 이때 필요한 것이 바로 검색 증강 생성이다.

이 방식은 사내 규정이나 문서를 AI에게 프롬프트 형식으로 제공하여, AI가 문서를 읽고 그에 기반한 답변을 생성하는 방식이다.

마치 인공지능이 “잠시만요, 관련 규정을 찾아보고 알려드릴게요”라고 말하는 셈이다.

특정한 정보에 기반한 정확한 답을 얻을 수 있다는 점에서 매우 유용하지만, 방대한 자료를 인공지능에 제공하고 그중에서 정확한 정보를 검색하여 나가게 하는 과정은 여전히 도전적이다.

마지막으로 가장 주목받고 있으며 주요한 변화를 이끌고 있는 기술은 강화학습 기반 미세조정(RFT)이다. 기존의 LLM 조정 방식(Fine-Tuning)은 주로 사람이 일일이 정답을 제공하며 LLM을 다시 학습시켜야 했기에 시간과 비용이 많이 드는 방식이었다.

하지만 RFT는 사람이 직접 답을 제공하여 주는 대신 AI가 여러 응답을 생성하면 사람이 가장 적절한 답변을 고르고, 이를 기준으로 보상하는 방식이다. 특히 최근에는 초기에 사람이 피드백을 주고, 이후에는 규칙 기반 평가나 기존 데이터를 활용한 자동화된 방식으로 피드백이 가능해지고 있다. 이런 학습을



반복하여 인공지능은 사람이 선호하는 방향으로 스스로 행동을 조정해 나가게 된다.

최근 중국의 딥시크(DeepSeek) 모델은 이 방식을 통해 짧은 시간과 획기적으로 적은 비용으로도 최고 수준의 성능을 가진 모델을 만들 수 있다는 가능성을 보여주었다.

또한 이 방식은 개인화된 AI의 훈련에도 엄청난 진전을 이뤘다. 하지만 수학 문제나 코딩처럼 정

현재 AI, 획일적 지식 학습에 집중
그때그때 특수한 상황 대응 한계

문맥 학습·증강 생성·미세 조정 등
스스로 반응 조정하는 기술 통해
시대 맞는 진정한 지혜 배우는 중

답이 명확한 분야에서는 RFT를 통해 빠르게 성능을 개선할 수 있지만, 개인화된 인공지능을 만들기 위해서는 나에게 맞는 ‘좋은 대답’을 정의하는 것이 중요하며 이를 기준으로 사람이나 조직이 설정하여야 한다는 것이 과제로 남아 있다.

이 과정을 들여다보면 성경 창세기의 이야기가 떠오른다. 에덴동산의 아담과 하와는 선악과를 먹기 전까지는 지식을 갖고 있었지만 ‘옳고 그름’을 판단하지 못했다. 그러나 선악과를 먹은 순간, 그들은 자율 의지의 판단에 대하여 스스로 책임을 지는 존재로 변화했으며 에덴동산을 떠나 인간 사회로 들어오게 된다.

똑똑하게보다 어떤 방향이나가 더 중요

지금의 인공지능도 마찬가지다. 방대한 정보를 갖추었지만 ‘무엇이 좋은지’, ‘무엇이 옳은지’를 알지 못한다. 하지만 강화학습 기반 미세조정을 통해 인공지능은 인간의 피드백과 보상을 학습하며 어떤 응답이 바람직한지, 어떤 방식이 사회적으로 수용되는지를 배워간다. 이것은 ‘김사비’가 전공의 생활을 하며 선배들의 조언과 경험을 바탕으로 전문의로 슬기롭게 성장하여 가는 것과 같다.

AI는 의식을 가진 존재는 아니지만, 그 결과는 인간의 윤리와 가치에 따라 우리와 조화를 이루는 AI로 나아가는 방향이 된다. 이러한 기술 발전은 AI의 개인화와 사회화에 있어 필수적인 조건이다. 교육에서는 AI 튜터가 학생의 성향과 실력에 맞춘 문제를 출제하고 학습 동기를 자극할 수 있다. 의료에서는 최신 연구를 기반으로 환자 맞춤형 진단과 처방을 제공할 수 있다. 기업에서는 고객과의 응대 톤을 설정하고 업무 프로세스를 기업의 목적에 맞게 조정할 수 있다.

하지만 여기에는 중요한 새로운 문제가 제기된다. 우리는 지금까지 인공지능이 어떻게 똑똑하게 되는 가를 연구하여 왔다. 결과적으로 특정 분야의 판단(바둑·영상진단 등)에서 인공지능이 인간 전문가에 버금가는 또한 능가하는 결과를 보이는 데 성공하였다.

하지만 앞서 설명한 문맥 학습, RAG, RFT는 단순히 인공지능을 더 똑똑하게 만드는 기술이 아니다. 우리는 이러한 기술들을 통하여 처음으로 답을 구하려는 인공지능이 아닌 우리의 언어, 감정, 가치에 반응하는 존재를 만들려고 하고 있다. 김사비의 교육된 한국병원에서의 슬기로움이 미국 병원에서의 슬기로움으로 연결될지는 고민을 해야 한다. 지금은 ‘더 어떻게 더 똑똑하게 만들 수 있을까’보다 더 중요한 질문을 던져야 한다. 우리는 인공지능을 어떻게 가르쳐야 하는가. 그리고 어떤 방향으로 가르쳐야 하는가.

돌이켜보면, 인류의 많은 갈등은 서로가 생각하는 ‘ 좋음’과 ‘ 나쁨’의 기준이 달라서 생겨났다. 그렇기에 인공지능 시대의 윤리와 기준, 그리고 진정한 지혜를 담을 수 있는 인공지능에 대한 고민은 지금 이 시점에서 더욱 절실하다

〈광주일보와 중앙 SUNDAY 제휴 기사입니다〉

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



KSA 한국표준협회
KOREAN STANDARDS ASSOCIATION

ISO 21388

보청기적합관리 인증센터



국 제 보 청 기

“고객에게는 신뢰와 만족”

- ✓ 필요한 소리만 똑똑히 들립니다.
- ✓ 작은 사이즈로 착용시 거부감이 없습니다.
- ✓ 정직한 우수상품 가격부담이 없습니다.

본점	서석동 남동성당 옆	062) 227-9940 062) 227-9970
서울점	종로 5가역 1층	02) 765-9940
순천점	중앙시장 앞	061) 752-9940