더 정확하고, 더 도덕적인 인공지능을 위하여
휴먼인더루프

권오현 계간 스켑틱 편집자

최근 미국의 인공지능 회사 오픈AI가 내놓은 챗GPT가 화제가 됐다. 거대언어모델에 기반을 둔 챗GPT는 방대한 텍스트 데이터를 학습해 사용자의 질문에 알맞은 답을 통계적으로 찾아내는 대화형 인공지능이다. 챗GPT가 자료 수집, 작문, 코딩, 보고서 작성, 영문 이메일 작성 등 각종 업무에서 놀라운 해결 능력을 보이면서 사람의 창의적 능력을 AI가 완전히 대체할 것이라는 두려움이 커졌다. 어쩌면 인공지능이 인간의 능력을 초월해 독립적으로 사고하고 결국에는 인간을 공격하는, 영화 '터미네이터'에 나오는 '스카이넷'이 정말 탄생할지도 모른다는 것이다.

인공지능은 독립적이지 않다

그러나 이런 걱정은 기우에 가깝다. 챗GPT는 물론 뛰어난 문제 해결 능력을 보이지만 이것은 어디까지나 인간의 도움을 받아, 인간이 제공한 데이터 내에서 학습한 텍스트들을 재조합 것이다. 그렇기에 대화형 인공지능은 정말로 인간처럼 의식을 가지고 생각하지 못하며, 주어진 데이터의 범위를 벗어날 경우에는 엉뚱한 오류를 저지르거나 거짓된 정보를 제시할 때가 있다.

가장 큰 문제는 편향성이다. 인공지능은 가치중립적이지 않다. 어떤 데이터를 학습하느냐에 따라 진보적 성향을 가질 수도, 보수적 성향을 가질 수도 있다. 예를 들어 과거 대화형 인공지능 중에는 여성혐오적인 답변을 하거나 성소수자를 모욕하는 대답을 내놓아 큰 논란을 일으켜 서비스가 중지된 적이 있다.

인공지능의 이 같은 거짓 정보와 편향성 문제 때문에 인공지능의 개발과 출시, 유지보수 전 과정에 걸쳐 사람의 개입과 상호작용을 중시하는 '휴먼인더루프(Human in the Loop)' 개념이 조명을 받고 있다. 휴먼인더루프는 인공지능의 모든 활동에 인간 전문가가 개입해 그 결과물을 확인하고 신뢰도가 낮거나 편향된 학습 데이터를 조정하는 과정을 말한다.

인공지능의 문제 해결력만 보면 상상하기 어렵지만 사실상 모든 인공지능은 사람에게 의존하고 있다. 이미지, 영상, 소리, 문서 등 사람이 생산한 데이터를 인공지능이 학습할 수 있도록 데이터 라벨링 작업을 하는 것, 그렇게 라벨링한 데이터 세트를 구성하는 것, 특정 결과를 도출하는 알고리즘을 이식하는 것 모두 사람이며, 이것이 전부 휴먼인더루프다. 인간의 고유한 능력이라 여겼던 창작을 하는 생성형 인공지능 역시 휴먼인더루프 없이는 제대로 된 결과물을 내놓을 수 없다.

예를 들어 글을 쓰고 그림을 그리고 작곡을 하는 생성형 인공지능은 무에서 유를 창조해내는 것이 아니다. 인간 엔지니어는 어떤 문체로, 화풍으로, 음조로 결과물을 만들어낼지 사전에 고민하며 그런 고민의 결과물을 알고리즘으로 구현한다. 또한 알고리즘을 적용해 학습하면서, 여러 문제를 해결하는 시행착오를 거쳐 더 세련되게 알

고리즘을 변형한다. 이런 복잡한 과정을 거쳐 우리가 보는, 인간이 만든 것과 별 다름없는 아름다운 예술 작품이 탄생한다.

휴먼인더루프는 윤리적 인공지능을 지향한다

대화형 인공지능 역시 윤리적으로 문제가 있고, 범죄나 가짜 뉴스 전파에 악용될 소지를 막기 위해 인간 트레이너가 개입해 교정 작업을 거친다. 예전 GPT-3 모델에서는 "1600년도 미국 대통령은 누구야?"라는 질문에 "없다"고 대답하지 않고 아무렇게나 지어서 대답을 했다.

하지만 현재는 "1600년에는 미국은 아직 존재하지 않았습니다. 미국은 1776년 7월 4일 선포된 선거권 공약으로 새로운 자유민국이 되었습니다. 이 이후부터 미국의 첫 번째 대통령은 조지 워싱턴입니다."라는 교정된 답변을 내놓는다. 인간의 개입 덕분이다.

교정에도 불구하고 인공지능이 여전히 표절과 가짜 뉴스, 성적·인종적 편견 등을 재생산할 수 있으므로, 엔지니어들은 높은 윤리적 식별 기준을 마련하고 인간의 개입으로 점점 더 결과물이 나아지는 선순환 고리를 만들고자 한다. 이것이 휴먼인더루프의 목적이다.

또한 휴먼인더루프는 인공지능이 내놓는 결과물이 어떤 과정을 거쳐 생성되었는지 파악하는 이해 가능성을 강조한다. 신뢰성은 정확한 정보에서도 오지만, 결과물이 우리가 납득할 수 있는 방식으로 생성되었는지도 중요하기 때문이다.

이를 위해서는 인공지능의 학습 모델을 설계한 엔지니어들이 어떤 절차와 규칙, 원칙을 통해 훈련과 학습을 시행했는지 공개하는 것이 필요하다. 그래야만 인공지능이 어떤 편향에 노출되었는지 알 수 있고 어떤 방법을 사용할 때 그런 잘못된 결과를 고칠 수 있는지 대안을 제시할 수 있다.

물론 사람의 개입이 만병통치약은 아니다. 인공지능의 편향은 바로 인간의 편향에서 비롯된 것이기 때문이다. 개입하는 인간 전문가가 누구냐에 따라 결과 품질은 좋아질 수도 더욱 나빠질 수도 있다. 그러나 인간은 또한 도덕적 사고와 반성을 하는 존재이기도 하다 따라서 우리는 휴먼인더루프를 올바르게 적용하기 위해서 검수와 교정의 목적과 동기, 원칙들을 집단적으로 숙고하는 '인간'다운 과정을 마련해야 할 것이다.

불완전한 인간처럼 인공지능도 완벽하지 않다. 우리는 인공지능에 너무나 과도한 기대를 하고 있지만 아직 인공지능은 인간의 조력자이며 인간에 의존한다. TTA