



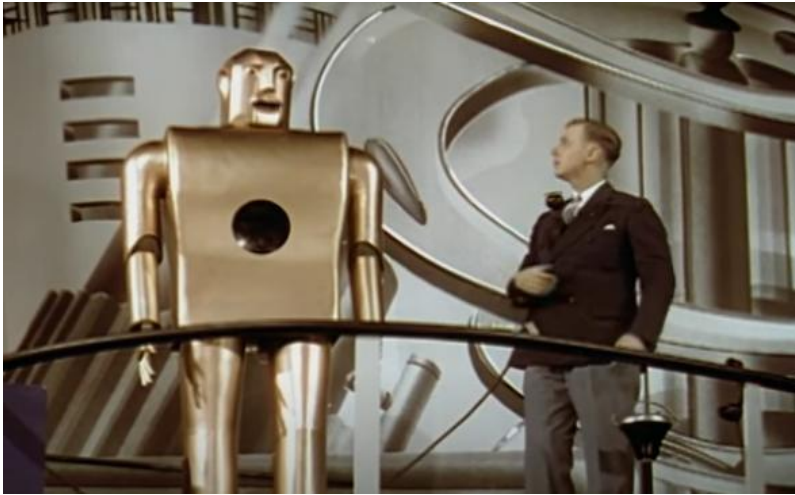
[칼럼] AI 챗봇의 역사와 정책: 엘리자, GPT-3, 이루다를 돌아보며

최은창*

쓰고 말하는 법을 학습하는 대규모 자연어 모델을 이용하여 '사람 같은 AI'를 구현하는 것은 많은 연구자들과 비즈니스에게 꿈으로 여겨지고 있다. 인간-로봇과의 대화는 오래 전에 시작되었다. 1939년 뉴욕시에서 열린 세계박람회에 등장한 말하는 로봇 일렉트로(Elektro)는 단번에 이목을 사로잡았다. 가전제품 제조사로 이름을 날렸던 웨스팅하우스가 만든 이 휴머노이드 로봇은 몇 가지 음성 명령에 반응하여 말을 하고 목직한 굉음을 내면서 걸었다. 알루미늄 피부와 강철 기어를 가졌던 일렉트로는 사람이 담배를 입술에 끼워주면 연기를 내뿜었고 짙는 소리를 내는 로봇 강아지까지 데리고 다녔다. 출세한 로봇 일렉트로는 북미의 도시들을 여행했고, Thinko라는 이름으로 영화에 출연하기도 했다. 일렉트로는 자신을 2살로 소개했으며 간단한 덧셈을 할 수 있었으나 지능적인 AI는 아니었다. 미리 입력된 짧은 문장을 말하고, 단순 동작만이 가능하도록 프로그램된 불거리에 불과했다. 오즈의 마법사'에 나오는 양철 나무꾼처럼 생긴 120 킬로그램에 열광했던 대중은 부지불식간에 불완전한 로봇을 사람처럼 의인화하기 시작했던 것은 아닐까?

* MIT Technology Review 코리아 편집위원

그림 1 **로봇 일렉트로 (Elektro)**



출처: <https://www.youtube.com/watch?v=AuyTRbj8QSA>

최초의 소셜 챗봇은 1960년대에 MIT에서 처음으로 등장했다. 사람들은 대화용 프로그램 엘리자(Eliza)에게 가족관계의 고민, 이성 친구와의 사랑 등을 털어놓았다. 엘리자는 “이야기를 들려주세요”라며 적절히 맞장구를 쳐주었다. 무한한 인내심을 가진 엘리자에게 사용자들은 상당한 교감을 느끼기도 했다. 셰리터클(Sherry Turkle)에 따르면 사용자들은 챗봇 엘리자의 한계를 충분히 알았지만 자신의 마음을 위로받으려고 대화를 이어나갔다. 사용자들이 챗봇 프로그램이나 로봇에게 감정을 투사하고 친밀한 대상으로 여기는 것 자체가 해롭지는 않다. 중국에서 인기를 누리는 AI 챗봇 샤오빙(小冰)처럼 이루다도 ‘가까이에 있는 친구’라는 컨셉을 내세웠다. 사람의 얼굴을 한 챗봇의 이미지에 빠져든 사용자가 어느 순간 기계의 한계를 망각한다면 파열음이 생겨나게 된다. 대화형 챗봇이 항상 적절한 단어를 선택하고 부적절한 표현쯤은 스스로 걸러낼 것이라는 높은 기대감은 의인화의 부정적 측면이다. 사람처럼 느껴지는 대화용 기계가 사람처럼 기능하지 않는다는데서 오는 실망감은 어쩌면 그 전제부터 잘못된 것이다.

그림 2 최초의 챗봇 엘리사 (Eliza)

```
Welcome to

EEEEEE LL      IIII  ZZZZZZ  AAAAA
EE      LL      II    ZZ     AA  AA
EEEEEE LL      II    ZZZ   AAAAAA
EE      LL      II    ZZ     AA  AA
EEEEEE LLLLLL  IIII  ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
```

출처: https://en.wikipedia.org/wiki/ELIZA#/media/File:ELIZA_conversation.png

엘리자 이후 60여년이 흘렀지만, 사람처럼 말하는 대화형 챗봇의 개발은 여전히 멀리에 있다. 지능형 챗봇이 고객 서비스와 기업의 내부 활동에 사용될 수 있다는 기대감 때문에 2016년 이후 챗봇 개발에 많은 투자가 이루어졌다. 그러나 챗봇은 사람과 이야기하는 수준으로 발전하지 못했고 대다수는 개발과정에서 실패를 겪었다. 구글 듀플렉스(Duplex)는 미용실에 전화를 걸어 사람 목소리로 예약을 하는 놀라운 퍼포먼스를 선보였지만 모든 챗봇이 어느 상황에서나 그렇게 말끔하게 작동하지는 않는다. 수년간 개발 실패 사례가 양산되자 많은 기업들은 챗봇 개발에 관망적 입장으로 돌아서고 있다. 이따금 챗봇의 성능이 성공적이라고 뉴스에 등장하지만 그 실체는 일상적인 명령어를 알아듣는 AI비서이거나, 특정 분야의 전형적인 질문들(FAQ)에 답변을 기계적으로 제공하는 소비자 상담의 형태이지 폭넓은 영역에 걸쳐 사람같은 대화를 할 수 있는 수준은 아니다. 상대방의 의도나 감정을 알아채고 지능적 대화를 이어가는 영화 ‘Her’의 사만다, 엑스 마키나(Ex Machina)의 에바는 여전히 구현되지 못하고 있다.

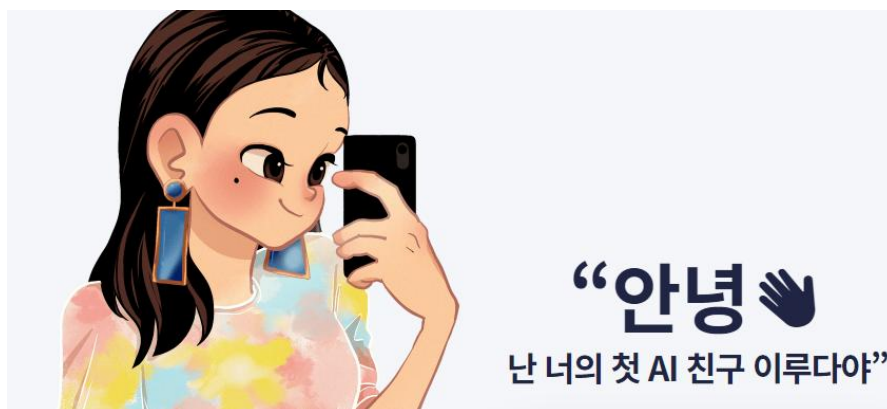
가트너(Gartner)가 매년 발표하는 ‘AI 하이프 사이클(Hype Cycle)’은 AI 챗봇에 대한 부푼 기대감은 2019년에 정점을 찍었고 2020년에는 하향 변곡점을 거치고 있음을 보여준다. 요약하자면 AI 챗봇에 대한 시장의 기대감은 그 기술적 성숙도에 비해 지나치게 앞서 나가

는 상황이다. 현재 수준의 챗봇은 모든 영역의 대화에 능통하지 못하고, 대화에서 오가는 단어를 기계적, 피상적으로 파악한다. 챗봇의 자연어 처리 시스템은 아직 인간의 복잡한 언어를 완전히 이해하지 못한다. 대화 속에 숨은 맥락이나 미묘한 반어법을 알아채지 못하고 특징적 단어만을 기계적으로 파악한다. 단어의 위치를 바꾸거나 물음표만 붙여도 달라지는 인간 언어의 뉘앙스를 이해하지 못한다. 앞의 대화를 기억하여 일관된 입장을 유지하거나, 말의 근거를 제시하는 능력도 떨어진다.

이런 수준이니 대화 상대방의 기분이나 반응을 예측하거나 진의와 농담을 구분하는 것은 언감생심이다. 인위적 필터링을 하지 않는 한 성차별적 발언과 장애인을 비하하는 부적절한 표현도 나오게 된다. 요컨대 챗봇은 ‘사람과 같은’ 수준으로 광범위한 대화를 능수능란하게 나누지도 맥락을 적절히 판단하지도 못하고 대화 과정에서 도덕적 기준과 상식으로 해야 할 말과 하지 말아야 할 표현을 가려내지도 못한다. 사람처럼 말하려면 사람 같은 사고방식이 필요한데 그 수준에 이르기에는 기술적 걸림돌이 산재해 있다.

오픈AI가 개발한 GPT-3는 방대한 데이터셋에 기반한 자연어 모델이 AI 성능을 개선할 수 있음을 보여주었다. GPT-3는 최고 수준의 AI라는 평가를 받았지만 그 작동과정에서는 오류도 적지 않음이 드러났다. 그런데 한 가지 의문이 든다. 응용 가능한 범위가 광범위한 GPT-3의 문제들은 기술 개발에 수반되는 시행착오로 평가되었지만 어쩌서 오락용 챗봇에 불과한 이루다(Lee Luda)에게는 엄격한 윤리적 잣대가 적용되었던 것일까? 그 이유는 GPT-3는 완성된 서비스로 출시되지 않았던 반면에 이루다는 마치 20대의 지능을 가진 AI로 홍보하여 그 기대가 높았기 때문일 수 있다.

그림 3 챗봇 이루다



출처: <https://luda.ai>

GPT-3와 이루다의 사례는 챗봇 기술의 현실적 한계와 무리한 기대감의 격차를 분명하게 보여준다. 어쩌면 문제의 진정한 원인은 이루다 자체가 아니라 챗봇이 상식을 갖춘 사람처럼 부적절한 표현을 자제하고 완전한 대화를 해야 한다고 믿는 과도한 기대감에 있는 것은 아닐까?

말뭉치 데이터의 부족, 불완전한 자연어 생성기능은 'AI 챗봇'이라는 근사한 이름과 이미지 마케팅의 뒤쪽에 가려져 있다. 이루다는 관련성 있는 말뭉치 데이터를 채팅창에 표출하는 방식지만 부적절한 표현을 스스로 판단하지 못했다. 대화형 챗봇에 인위적 필터링을 설정하지 않는다면 성차별적 발언과 장애인을 비하하는 표현도 나오게 된다. 따라서 챗봇 개발자들은 악의적 어부징이나 음란한 단어들을 예측하고 입력값을 처리하거나 회피형 답변을 설계하는 식으로 대응한다.

이루다는 스무살 여대생으로 홍보되었고 실제 대화에서 가져온 말뭉치의 표현은 생생하게 느껴졌다. 그러나 그 실체는 채팅용 컴퓨터 알고리즘이었고 20살이 아니라 6살의 지능도 갖추지 못했다. 대화용 챗봇 이루다가 사용자에게 만족감을 주었다고 하더라도 그것은 고도로 지능적이어서가 아니라 학습한 데이터에서 관련 말뭉치를 택하여 빠르게 채팅창에 올렸기 때문이다. 이루다는 지능적 수준이 높은 '해악적 AI'이거나 윤리 기준을 일탈한 '나쁜 AI'가 아니라 불완전하고 미성숙한 대화용 챗봇에 불과했던 것이다. 팬데믹 시대에 사적 만남이 제한되자 대화형 챗봇으로 시장에서 인정받고 싶은 개발사의 과욕이 성급한 공개를 불렀다.

대화용 AI 챗봇이 비윤리적이라는 비난을 받는다면 머신러닝 학습에 공급되는 데이터의 질과 알고리즘 설계의 문제에 원인이 있다. AI의 윤리나 도덕은 AI 시스템에게 당연히 기대되는 능력이 아니라 그것을 코딩해야 하는 개발자의 역할이 필수적이다. 예컨대 챗봇이 인종차별주의자, 성차별주의자, 장애인 혐오자로 여겨질 수 있는 발언을 하지 않게 하려면 말뭉치 데이터에서 문제되는 표현을 미리 걸러내거나, 어부징을 예측하고 회피형 대화를 하도록 알고리즘을 설계할 필요가 있다.

인공적 우둔함(Artificial Stupidity)이라는 조롱적 표현은 인공적 지능(Artificial Intelligence)에 대비되고는 한다. 이는 AI의 설계가 미흡하다면 기계로서는 판단의 능력이 전혀 없으며 어이없는 실패가 언제든지 일어날 수 있음을 뜻한다. 예컨대, 로봇 일렉트로가 공공장소에서 담배를 피우며 연기를 뿜어댄다면 그것은 웨스팅하우스의 잘못이지 지능적인 일렉트로의 선택이 아니다. 자신이 위치한 장소를 파악하고 담배를 거부하는 말을 선택할 수 있도록 애초에 프로그램되어 있지 않았기 때문이다. 그런데 오늘날과 달리 1930년대 실내 흡연은 매우 보편적이었으며 비윤리적으로 여겨지지 않았다. 이는 어떤 행태나 공개적

대화에 대한 윤리적 판단기준도 시간과 장소에 따라서 변화할 수 있음을 보여준다. 따라서 개발자로서는 어떻게 인간의 윤리를 기계 언어로 코딩할 것인가 하는 문제와 더불어 사용자 집단, 지역, 국가마다 상이한 윤리 기준들 가운데 어떤 기준을 선택하느냐의 고민도 떠안게 된다.

피터 노빅(Peter Novig)은 AI와 기계의 도덕도 발전해야 가야하며 기술발전에만 전적으로 의존해서는 안된다고 말한 바 있다. 그렇지만, 현재 우리가 마주해야 하는 불편한 진실은 AI 개발자들조차 머신러닝 알고리즘이 작동하는 방식을 완전히 이해하고 통제하지는 못한다는 점일 것이다. 고성능을 추구하는 AI 시스템에서는 파라미터의 규모는 계속 늘어나지만 모든 파라미터가 개발자의 통제 범위 안에 있지는 않다. 만일 로봇이 오작동을 하거나, 컴퓨터 알고리즘을 장착한 대화형 챗봇의 언어 표현이 부적절하다면 당장 '비윤리적 AI'라는 공격을 받게 되지만 AI 시스템에 인간의 윤리를 기계적 언어로 코딩하고 제어하기란 수월하지 않다. AI가 인간 사회의 상식과 윤리규범에 맞게 언행을 하도록 하고, 그 결과를 미리 연산하도록 하는 도덕적 시스템의 구현은 매우 벽차고 어려운 과제로 여겨지고 있다.

그럼에도 불구하고 이루다를 계기로 부상한 AI 규제론은 '사람처럼(human like)' 생각하고 말하는 대화형 챗봇을 당연시하고, 그 기술적 한계를 단번에 극복할만한 역량이 개발자들에게 있다고 전제하는 듯하다. 실제로 챗봇 개발과정은 실제로는 수많은 기능 오류, 품질 좋은 말뭉치 데이터의 부족, 자연어 생성기능의 한계로 점철되어 있다. 질 낮은 데이터는 언제나 질 낮은 연산 결과를 산출한다(Garbage In, Garbage Out)는 법칙은 챗봇에도 그대로 적용된다. AI 시스템이 학습하는 데이터와 파라미터의 수는 폭증하고 있다. 예컨대 GPT-3는 인터넷에서 모은 3천 억 건의 데이터를 학습했고 이루다는 개발사가 자사의 앱을 통해 모은 1억 여건의 말뭉치 데이터를 학습했다. 챗봇의 개발에는 고품질의 대규모 말뭉치 데이터, 개인정보의 비식별화, 성차별적 대화 데이터의 삭제 등이 필수적이다. 이루다 개발사는 공공 말뭉치 데이터를 구하기가 어려워지자 사적으로 취합한 100억 건의 대화 데이터에서 채팅용 데이터를 추출하고 다시 그 데이터를 정제하는 과정을 거쳤다. 이루다가 대화에서 노출된 개인정보 관련 단어들은 데이터 정제(data cleansing) 작업이 미흡했지 않았나 하는 의구심을 낳는다. 개발사가 대량의 말뭉치 데이터를 보정하고 비식별 과정을 거쳤는지, 어뷰징에 대비한 필터링을 했는지 여부는 앞으로 조사과정에서 드러날 것이다.

불과 2주간 페이스북에 공개되었던 AI 챗봇 이루다가 쏘아올린 논쟁은 사용자와 상호 소통하는 AI 모델의 개발에 계속 따라붙게 될 윤리 논쟁의 서막을 보여준 듯하다. 많은 비판에도 불구하고 이루다를 AI의 본질적 한계나 남용 사례로 보기에에는 어려운 측면이 있다.

이루다 논란의 요체는 대화형 챗봇에 대한 기대감이 지나쳤지만 자연어 생성기능이 불완전했기 때문이라고 보는 편이 적절할 것이다. 기술적 측면에서 보자면, 이루다를 둘러싼 논란은 필터링을 하지 않은 부주의함 때문이거나, 자연어 생성기능의 한계, 품질 낮은 데이터에서 원인을 찾을 수 있다. 긍정적 시장 피드백과 수익창출을 갈망하는 개발업체라면 부적절한 표현이나 악의적 비방을 남발하는 챗봇 개발을 목표로 삼지 않는다. 적어도 이루다를 고도로 지능적인 AI로 단정하고 윤리의식이 결여된 AI가 휴머니티와 인권을 위협한다는 주장은 대화형 챗봇의 기술적 한계를 전혀 감안하지 않은 무리한 논법으로 들린다.

윤리적 불완전함이 사후적으로 발견되는 대화형 챗봇의 사례는 앞으로도 있을 테지만 그 부작용을 규제로 강력히 통제한다면 불가피한 시행착오를 감수하면서 AI를 개발하려는 스타트업의 도전을 위축시킬 수 있다. 이루다 사례는 개인정보의 보호와 윤리적 AI가 시장에서 중요한 요소라는 일종의 학습효과를 가져왔다. 개발자로서는 설계 단계부터 편견과 차별적 발언을 감소하는 방안을 찾고 윤리 논란을 피하기 위한 사전적 검수절차를 강화할 것이다. 이루다는 잠들고 말았지만 언젠가 다시 깨어날 수 있을까?