모델이 메시지다
The Model is The Message

원저자: Benjamin Bratton, Blaise Agüera Y Arcas
발췌 + 한국어 번역: Alm Chung

역주 1: 엔지니어 Blacke Lemoine가 대화형 인공지능 모델 LaMDA에 “지각(sentient)이 있다”는 개인 인상을 공유했는데, 미디어에서 이에 대한 반향이 컸습니다. 이 글은 Lemoine의 주장에 대한 저자들의 반응으로 시작됩니다.

역주 2: 내용 중 익사에 대한 묘사가 있으니 주의 바랍니다.

[...전략…]

[인공지능 언어 모델] LaMDA는 애드리브에 능숙하며, 매 번의 대화가 즉석에서 즉흥적으로 생성됩니다: 우리가 느끼는 LaMDA의 “성격”이란, 대부분 주어진 명령 그리고 진행되고 있는 대화 자체에서 발생합니다. 다름 아닌 당신이 모델에게서 보길 원하는 모습을 보는 거죠.

그러므로 ‘인공지능이 포유류와 비슷한 수준의 지각(이게 아마도 Lemoine가 소망한 방향이겠죠)을 가지고 있는가’보다는, ‘이 모델이 Lemoine가 이 모델에게서부터 듣고 싶은 것을 정확하게 파악하고 있는 이 상황을 어떻게 이해하면 좋을까’를 물어봐야겠습니다. [...]

LaMDA가 이런 결과를 보여줄 수 있는 것은, 이 모델이 꽤나 교묘한 작업을 수행하고 있다는 이야기입니다: 이 인공지능 모델은 사용자의 사고방식을 모델링(mind modeling)하고 있습니다. 아마도 이 모델은 자신에 대해 어느 정도 충분히 파악하고

있는 듯합니다. — 이 말이 어떤 주관적 사고능력(subjective mind)을 뜻하는 것은 아니지만, Lemoine의 사고방식 속 어떤 구성물로써 — 상황에 맞게 반응할 줄 알며, 따라서 Lemoine의 마음속에 있는 인격의 의인화적 투영 (anthropomorphic projection of personhood)을 증폭시켰다는 이야기입니다.

타인의 마음에 대해 상대적으로 제 자신을 모델링하는 능력은 사회적 지능의 기본입니다. 이 능력은 포식자-피식자 상호 작용뿐만 아니라, 대화나 협상과 같은 복잡한 일들을 수행 가능케하죠. 다르게 이야기하자면, 여기에 (Lemoine가 주장하는 방향으로는 아니지만) 정말로 어떤 종류의 지능(Lemoine가 자신을 어떻게 생각하는지에 맞춰 자기 자신을 모델하는 지능)이 존재할 가능성이 있습니다.

[...]

이것보다도 더 중요하게, 자연어 처리(natural language processing)의 심장인 시퀀스 모델링 (sequence modeling)이야말로, 임의의 태스크(task, 임무)를 유연하게 수행해내며 현재 언어의 영역을 너머 이미지 합성과 신약 개발, 로보틱스 영역까지 아우르는, 제너럴리스트 인공지능(generalist AI, 범용 인공지능) 모델을 현실화할 수 있는 열쇠라고 볼 수 있습니다. "지능"은 이런 사람과 기계 의사소통의 모의 합성 (mimetic synthesis) 순간들에서도 찾아볼 수 있지만, 자연어가 말과 글의 범주를 넘어, 우리의 인지적 인프라(infrastructure)로 확장된 데서도 찾아볼 수 있습니다.

[...]

그렇다면 이 이야기가 "이해력(comprehension)"의 특성에 대해 무엇을 시사할 수 있을까요? 1982년도에 Frank Jackson가 제안한 "Mary의 방(Mary's Room)"이라는 사고 실험은, 광학적 현상으로서 "붉은색"에 대한 과학적 지식은 가지고 있지만 완전히 무채색인 방에서 살고 있는, Mary라는 한 과학자에 대한 이야기입니다. 여기서 Mary가 어느 날 이 방에서 나가 붉은색 물건들을 보게 되었을 때 "붉은색"을 [다른 사람과 비교해] 확연히 다르게 경험할지, Jackson은 질문합니다.

그렇다면 인공지능은 무채색의 Mary와 같은 경우에 놓인 걸까요? 방을 나서는 순간,

Mary는 "붉은색"을 다른 방식으로 이해할 (그리고 더 잘 이해할) 수 있을 것이지만, 결국 그런 경험의 스펙트럼은 언제나 제한되어있기 (curtailed) 마련입니다. 호숫가에서 평생을 지낸 어떤 사람이 어느 날 호수에 빠지게 된다면, 그 순간 그가 여태까지 상상도 못 해본, 깊고도 격한 방식으로 "물"을 첫 경험하게 되겠죠 – 숨을 압도하고, 폐를 깊숙이 채우고, 가장 밑바닥까지의 공포를 불러일으킨 후의, 공허로써.

이것이 물입니다. 그렇다면 물가에서 물을 무력하게 바라만 보고 있는 사람은 물을 이해하지 못하는 것일까요? 익사 중인 사람에 비교한다면, 물가의 사람은 어떤 면에서는 (다행히) 물에 대한 이해가 적다고 할 수 있지만, 또 어떤 면으론 당연히 물에 대해 이해하고 있습니다. 그렇다면 인공지능은 "물가에서" 세상을 보며, 어떤 면으로는 세상을 이해하고 있고, 어떤 면으론 이해하지 못하고 있는 것일까요?

[...]

마지막으로, 어느 지점부터 이성(reason)의 퍼포먼스가 이성의 일종이 되기 시작할까요? LaMDA와 같은 거대 언어 모델(Large Language Model)이 인지 인프라를 구동하기 시작한다면, "언어"의 효과—의미적 구별 및 물리적 세계 속 지시 대상과의 맥락적 연관관계를 포함한—에 대한 기능적인 이해가 언제부터 정당한 이해로 인정받을 수 있는가라는 질문은 더 이상 철학 사고 실험이 아니게 될 것입니다. 이제 이 질문은 사회적, 경제적, 그리고 정치적 결과가 따르는 현실적인 문제입니다. 이런 기술분야의 다양한 영역과 목적에 적용 가능한, 한 가지 현혹적이고 심오한 교훈은, 간단하게도, 이것일 겁니다: 모델이 메시지다.

[... 후략 ...]