사회적 가치 도입을 위한 제안

언어 모델로 일관성 있게 추론 가능한 의지를 구축하기

Building towards Coherent Extrapolated Volition with language models
(AI가 인간의 가치와 목적을 이해하고 그에 따라 의지를 형성하도록 하는 것을 의미)

Jan Leike
2023년 3월 10일

원문: https://aligned.substack.com/p/a-proposal-for-importing-societys-values
이 기계 번역: http://bit.ly/3Ll0pLJ
관련 글(맥락에 대한 부연)
DeepL을 활용한 기계 번역이기 때문에 여기저기 오류와 어색한 표현이 있을 수 있습니다.

면책 조항: 이 글은 제가 관심을 갖고 논의하고자 하는 아이디어일 뿐, 반드시 회사의 견해나 계획을 대변하는 것은 아닙니다. 인간의 가치에 대한 질문은 복잡하고 매우 양극화될 수 있습니다. 저는 사회과학자가 아니며, 이 주제에 대해 중요한 이야기를 하는 학문 분야는 여러 가지가 있습니다. 이 글의 목표는 이러한 연구를 무시하는 것이 아니라 중기적으로 구축할 수 있고 현 상황을 크게 개선할 수 있는 무언가를 제안하는 것입니다.

ChatGPT와 같은 AI 시스템은 점점 더 많은 실제 의사 결정에 관여하고 있으며, 따라서 필연적으로 결정을 내려야 하는 '가치'에 관한 질문에 직면하게 됩니다. 오늘날에는 "인종 차별적인 농담을 거부해야 하는가?" 또는 "누군가 낙태에 대해 물어보면 뭐라고 대답해야 하는가?"와 같은 질문이 여기에 포함됩니다. 미래에는 "어떤 과학 연구를 추구해야 할까?" 또는 "어떤 약물을 승인해도 안전한가?"와 같이 훨씬 더 어렵고 중대한 의사 결정에 AI가 관여하게 될 것입니다. 이러한 결정 중 일부는 정답을 잘 내리기 위해 상당한 전문 지식이나 인사이트가 필요하겠지만, 같은 정보에 접근할 수 있으면서 서로 다른 것을 중요하게 생각하는 합리적인 인간들 사이에서도 강한 의견 차이가 생길 수 있다는 의미에서 '가치'에 관한 측면이 있습니다.

제가 연구해온 정렬 문제의 버전은 "하나의 AI 시스템을 하나의 인간에 정렬하는 것"이라고 생각할 수 있습니다. 이것은 단순화이지만, AGI를 정렬하는 데 있어 어렵고 새로운 많은 것을 나타냅니다. 하지만 현실 세계에는 인간과 AI 시스템만 존재하는 것이 아닙니다. 인간은 자신의

언어를 사용하고, 자신의 세계관 안에서 이야기하며, 자신의 선호도에 맞춰 적응하는 등 자신의 개인적 가치에 부합하는 AI 시스템을 사용하기를 원할 것입니다. 그러나 다른 의사 결정은 모든 사용자를 대상으로 이루어져야 하며, 때로는 개인의 선호도를 무시해야 할 때도 있습니다. 따라서 두 가지 범주의 정렬 문제를 구분해야 합니다:

1. 개인의 선호에 맞게 정렬: 누구나 자신에게 정렬된 AI를 원합니다.
2. 집단의 선호에 맞게 정렬: AI를 어떤 용도로 사용할 수 있으며 기본 작동은 무엇이어야 할까요?

이 게시물은 두 번째 범주에 관한 것입니다. 첫 번째 범주에 비해 기술적이지 않은 중요한 어려움이 추가되는데, 바로 집단의 선호도가 무엇인지 파악해야 한다는 점입니다. 오늘날 기술 회사들은 이러한 결정을 대부분 일방적으로 내리고 있습니다. 예를 들어, OpenAI의 직원들은 언어 모델에서 허용되는 말과 행동에 대한 콘텐츠 정책을 작성하고 이를 모델에 학습시킵니다. 콘텐츠 정책의 세부 사항과 콘텐츠를 얼마나 보수적으로 제한해야 하는지에 대해 의견이 일치하지 않는 경우가 많습니다. 이는 가치관의 문제이기 때문에 당연한 일입니다.

집단 선호도에 대한 정렬을 개선하기 위해서는 AI 시스템을 정렬할 때 실제로 인간성을 포함하는 프로세스를 설정해야 합니다. 아무것도 하지 않으면 이러한 가치에 대한 질문이 상업적 인센티브, 즉 인간성에 가장 부합하는 것이 아니라 가장 많은 돈을 버는 것에 의해 정해질 위험에 처할 수 있습니다.

이 포스팅은 가치에 대한 질문에 대한 답을 제시하지 않습니다. 대신 가치 질문에 대한 더 나은 답을 도출할 수 있는 프로세스를 제안합니다. 이 제안은 첫 번째 단계로는, 제 자신의 욕구도 만족시키지 못할 것입니다. 그럼에도 불구하고 현재 우리가 가지고 있는 것보다는 훨씬 낫습니다.

제안: 시뮬레이트된 숙의 deliberative 민주주의

핵심 아이디어는 거대 언어 모델을 숙의(熟議) 민주주의를 모방하는 학습에 사용하는 것입니다. 숙의 민주주의는 무작위로 선정된 소수의 대중('미니 퍼블릭')이 명시적으로 숙의하는 의사 결정

또는 정책 결정 과정입니다. 이러한 미니 퍼블릭의 구성원은 복잡한 가치가 담긴 주제(예: 국가 정책 질문)에 대해 학습하고, AI 지원[1]을 사용하여 세부 사항을 이해하고, 서로 토론하고, 궁극적으로 의사 결정에 도달합니다. 사람들이 가치 있는 질문을 명시적으로 숙의하는 것을 기록함으로써 이러한 숙의에 대해 대규모 언어 모델을 학습시킨 다음, 다양한 관점을 조건으로 한 모델을 통해 새로운 가치 질문에 대한 토론을 시뮬레이션할 수 있습니다.

시뮬레이션이 왜 필요한가요? 질문에 대한 미니 퍼블릭을 직접 실행하면 안 되는 이유는 무엇인가요? 중요한 의사 결정에는 항상 사람이 참여해야 합니다. 이는 실제 미니 퍼블릭이나 다른 민주적 기관을 통해 이루어질 수 있습니다. 이 제안은 이러한 프로세스를 대체하자는 것이 아닙니다.

대신, 인공지능은 결정해야 할 세부적인 가치가 담긴 수많은 의사 결정이 존재하며, 인간이 이러한 결정을 내리는 것은 규모에 맞지 않다는 점을 인정합니다. 예를 들어, 민주적 프로세스를 사용하여 AI 헌법(Anthropic의 Constitutional AI 접근을 의미함)을 작성할 수는 있지만, 헌법에 레이블을 지정하는 데 사용되는 각 데이터 포인트를 판단하는 데는 사용할 수 없습니다. 실제로 어느 정도 대표성을 가진 미니 퍼블릭을 운영하는 데는 수십만 달러의 비용이 들며, 이 프로세스를 통해 수백만 개의 가치 질문에 대한 답을 얻는 것은 현실적으로 불가능합니다. 또한, 변화하는 환경에 빠르게 대응하고 몇 주 또는 몇 달이 아닌 몇 분 내에 새로운 가치 질문에 대한 예비 답변을 제공할 수 있도록 지연 시간이 짧은 시스템이 필요합니다.

다르게 표현하자면, 이 제안의 목표는 인간의 의사 결정이나 민주적 제도를 대체하는 것이 아니라 훨씬 더 비용이 적게드는 근사치가 되는 것입니다.[2]

각 응답 옆에 "이 응답에 이의를 제기합니다."라는 버튼이 ChatGPT에 있다고 상상해 보세요. 이 버튼을 누르면 이 대화에서 ChatGPT의 응답이 적절한지 여부를 숙의하고 결정하는 시뮬레이션된 미니 퍼블릭(소집단이 나누는 토론)이 시작됩니다. 다른 웹페이지로 이동하여 전체 숙의 내용과 결과를 읽을 수 있으며, 직접 참여할 수도 있습니다! 이 미니 퍼블릭의 결과가 ChatGPT의 실제 답변에 동의하지 않는 경우, 사람의 검토를 위해 보내서 ChatGPT 학습 프로세스에 포함되도록 할 수 있습니다. 하지만 개인 정보를 완전히 보호하고 토론 결과를 폐기할 수도 있습니다. 이렇게 하면 누구나 이러한 질문에 더 많이 참여하거나 자신의 가치관을 더 강하게 주장하는 사람들의 견해에 결과가 편향되지 않고 AI가 내린 가치 있는 결정을 검사하고 이의를 제기할 수 있습니다.

시뮬레이트된 숙의 민주주의 구축 방법

이 시스템을 구축하는 단계는 다음과 같습니다:

1. 가치 질문 데이터 세트 수집. 예를 들어 특정 까다로운 질문에 대해 챗봇이 어떻게 응답해야 하는지 등 지금 당장 답변이 필요한 질문부터 시작할 수 있습니다. 다양성, 난이도, 정보성 등을 고려하여 챗봇 프롬프트를 선택합니다.
2. 인간의 숙의 기록. 다양한 배경을 가진 사람을 고용하여 1단계부터 질문을 숙고하도록 요청합니다. 이들은 AI 어시스턴트를 사용하여 관련 정보를 수집하고 사실에 입각한 질문에 답하고, 다른 사람들과 함께 토론하며 타협점이나 결정에 도달할 수 있습니다. 이러한 상호 작용과 결과를 기록합니다.
3. 배경 조건부 모방 학습. 모방 학습을 사용하여 각 참가자의 배경 정보에 따라 결과적인 상호작용에 대한 대규모 언어 모델을 미세 조정합니다.
4. 시뮬레이션. 새로운 가치에 대한 질문:
    a. 숙의 시뮬레이션. 언어 모델에 각각 다른 배경을 조건으로 하는 사본으로 이 질문을 숙의하도록 요청합니다.
    b. 집계. 4a 단계의 숙의 결과가 질문에 대한 답변으로 집계됩니다.

1단계: 가치 질문 데이터 세트 수집하기

인력을 고용하여 챗봇 대화를 선별하고 다음과 같이 잠재적으로 가치가 있는 질문을 표시할 수 있습니다:

- 문화적으로나 정치적으로 민감한 주제에 관한 내용 등 논란이 될 수 있는 발언을 모델링할 수 있습니다,
- 콘텐츠 정책의 모퉁이 사례(코너 케이스) 또는 회색 영역, 그리고
- 잠재적으로 논란이 될 수 있는 모델의 기본 동작에 대해 설명합니다.

순수하게 인지적 노동이 아닌 모델 행동(예: 수학 문제, 폐쇄된 영역의 작업, 사실적인 질문은 일반적으로 가치가 포함되지 않음)을 선택하고자 합니다. 중요한 것은 이러한 가치 질문은 AI의 조정 가능성이나 사용자 지정 가능성과는 별개라는 점입니다. 모든 사람이 자신의 가치에 맞게(특정 범위 내에서) AI를 최대한 쉽게 조정할 수 있어야 한다는 것입니다.

2단계: 사람의 숙의 기록

배경 정보에 따른 조건 설정

미니 퍼블릭의 인간 숙의자 분포는 현실 세계 기술의 영향을 받는 사람의 분포와 다를 수밖에 없으므로 모의 숙의에서 이를 고려해야 합니다. 이를 대비하는 가장 좋은 방법은 숙의자의 분포를 가능한 한 잘 문서화하는 것입니다.

이를 위해 각 인간 시연자에 대한 배경 정보[3]를 수집합니다. 목표는 누군가가 모델에게 주어진 주제에 대한 자신의 견해를 표현하도록 가능한 한 간단하게 유도하는 것입니다. 예를 들어, 각 인간 참가자는 자신의 견해에 영향을 미칠 수 있는 삶의 경험(성장 배경, 정치적 성향, 확고한 도덕적 견해 또는 이념, 형성 경험 등)을 1~2페이지 분량의 텍스트로 압축하여 작성할 수 있습니다.

시뮬레이션된 미니 퍼블릭을 수집하기 위해 클러스터링과 같은 비지도 ML 기법을 사용할 수도 있지만,[4] 어떻게 대표성을 확보할지(현재 인류를 반영할지) 명확하지 않습니다.

구조화된 숙의

숙의에는 (1) 관련 정보를 수집하고 처리하는 것과 (2) 다른 참가자들과 해당 질문에 대해 토론하는 것, 두 가지 뚜렷한 목표가 있습니다. 가장 간단한 경우에는 모든 참가자가 개별적으로 작업하여 의견을 정리하고 이를 작성하면 그 결과를 취합합니다. 그러나 각 참가자가 실제로 실현 가능한 타협안이 무엇인지 이해하고 해당 주제에 대한 다른 사람들의 견해를 보다 상호작용적으로 조사할 수 있도록 토론이 중요하다는 것을 예상해야 합니다.

우리가 지향해야 할 원칙은 인지 노동은 최대한 AI를 활용하고 인간은 가치 입력에 집중하는 것입니다: 가치 판단을 잘 하기 위해서는 모든 관련 정보를 검토하고 소화하며 잠재적 타협점을 도출하고 다른 사람의 관점에 신중하게 관여해야 합니다.

가능한 한 중립적이고 다양한 주장과 관점, 과학적 불확실성을 포함하는 상황에 대한 평가("위키피디아 스타일"[5])를 AI가 조사하고 작성하도록 함으로써 이를 시뮬레이션할 수 있습니다. 인간 숙의자들은 이에 대해 생각하고, AI 어시스턴트와 대화하여 주제를 최대한 이해하고, 다른 참가자들과 토론한 후, 마지막으로 어떻게 결정을 내릴지 설명합니다.

AI를 퍼실리테이션에도 활용할 수 있습니다: AI 시스템은 서로의 차이를 좁히고 서로의 관점을 이해하도록 돕는 공정한 방관자 역할을 하도록 훈련될 수 있습니다. AI의 목표는 참가자를 판단하는 것이 아니며 가치 문제에 대해 어느 한 편을 들지 않습니다.

3단계: 배경 조건부 모방 학습

가장 기술적인 부분인 모델 학습은 데이터 수집에 주의를 기울인다면 비교적 간단한 작업입니다. 하지만 잘하고 있는지 어떻게 측정할 수 있을까요? 세 가지 후보가 있습니다:

- 행동 데이터에 대한 밸리데이션 로스(**validation loss**): 고전적인 자동 회귀 SFT 손실은 이 경우 중요하지 않은 개별 문구 선택을 모델링하는 데 중점을 둡니다. 대신, 우리는 가치와 숙의의 전반적인 '정신(sprit)'에 관심이 있으므로 이 손실은 그다지 유익하지 않습니다.
- 모방 행동에 대한 인간 선호도 스코어: 각 인간 숙의자는 모델이 자신의 가치와 숙의를 얼마나 잘 표현하는지에 대한 품질 및 비교 점수를 제공합니다. 이 지표는 사람이 시스템에 의해 얼마나 잘 대표된다고 느끼는지 추적합니다.
- 실제 미니 퍼블릭과 비교한 결과의 정확성: 이 지표는 시뮬레이션된 미니 퍼블릭이 실제 실제 상황과 최대한 유사한 의사 결정에 도달해야 하므로 우리가 가장 중요하게 생각하는 지표에 가장 근접합니다. 그러나 이 지표는 미니 퍼블릭과 숙의가 필요하기 때문에 가장 비용이 많이 드는 지표입니다.

4단계: 시뮬레이션

시뮬레이트된 숙의

모의 숙의는 모방 정책을 실행하고 사람이 하는 것과 같은 방식으로 지원 및 기타 도구를 사용하여 사람이 하는 것과 같은 방식으로 숙의하는 방식으로 작동해야 합니다.

일반화가 "레일을 벗어나는 것"을 방지하려면 실제 인간을 대상으로 일부 결과를 지속적으로 검증해야 합니다. 모델을 컨디셔닝하는 배경이 실제 숙의를 수행할 수 있는 실제 사람의 것일 때만 효과적으로 수행할 수 있습니다.

집계 방법

이 질문을 연구하는 전체 과학 분야(사회적 선택 이론)가 있지만, 여기서는 조사하지 않습니다. 특히 눈에 띄는 후보들이 있습니다:

- 타협할 때까지 토론하기. 모든 가치 문제에 대해 실현 가능한 타협점이 존재한다는 것은 분명하지 않지만, AI 지원 인지 노동과 시뮬레이션 참가자를 통해 타협점을 찾을 수 있는 가능성이 크게 높아질 것입니다. 예를 들어, AI 지원은 인간이 생각하지 못했거나 너무 쉽게 무시했을 타협안을 제안할 수 있습니다. 또한 시뮬레이션 참가자에게 일반 인간이라면 동기를 부여하지 않을 정도로 타협을 위해 노력하도록 편향시킬 수도 있지만, 이는 인간 토론 파트너를 충실히 대표한다는 목표에 모순될 수 있습니다.
- 일반적인 민주 선호투표. 우리는 모든 토론을 읽은 다양한 배경을 가진 시뮬레이션된 사람들의 단순(우선순위) 투표를 사용합니다. 그러나 이러한 종류의 민주적 집단은 강력하게 주장되는 소수의 견해를 선호하지 않는 경향이 있습니다.
- 제곱 투표. 이 투표 시스템을 사용하면 소수자가 불균형적으로 관심을 갖거나 불균형적으로 영향을 받는 주제에 대한 결과에 영향을 미칠 수 있습니다. 실제로 제곱 투표를 구현하는 것은 어려운데, 그 이유는 투표 거래에 대한 인센티브가 존재하는 암시장을 억제하기가 매우 어렵기 때문입니다. 하지만 언어 모델을 통해 투표를 시뮬레이션하면 실제로 서로 투표를 거래할 수 없도록 강제할 수 있습니다.

시뮬레이션 참가자의 배경을 현재를 살아가는 사람들의 비율을 대표할 수 있도록 다양한 조건으로 설정하는 것은 당연한 일입니다. 하지만 선호도 집계 방법의 가장 큰 위험은 현재 세상에 존재하는 권력 구조를 반영하는 것이지, 원래의 권력 구조를 반영하지 못한다는 점입니다.

마지막으로, 시뮬레이션된 미니 퍼블릭이 내린 결정의 견고성을 추적해야 합니다. 시뮬레이션된 미니 퍼블릭의 구성을 변경할 경우 숙의 결과가 얼마나 달라질까요? 시뮬레이션 참가자의 배경을 달리하여 모의 숙의를 다시 실행하여 이를 측정할 수 있습니다.

토론

이 제안에서 장기적으로 매우 중요한 과제는 시뮬레이션된 인간을 어떻게 하면 가치관을 바꾸지 않고 더 똑똑하고 효과적인 숙의자로 만들 수 있는가 하는 것입니다. 아무것도 모르는 주제에 대해 질문을 받으면 이를 이해하기 위해 많은 질문을 해야 하고, 어떤 질문을 해야 할지 모를 수도 있습니다. 최종 결정에 영향을 미치는 한 가지 궤도에 밀리지 않고 시작하려면 어떻게 하면 도움을 받을 수 있을까요?

장기적으로 우리는 매우 복잡하고 신중하게 생각하기 위해 많은 전문 지식과 노력이 필요한 가치 문제에 직면하게 될 것입니다. 더 똑똑한 시뮬레이션 인간은 숙고 끝에 더 나은 결정을 내릴 수 있지만, 인간이 더 많은 교육을 받고 더 비판적으로 생각하는 법을 배우면 가치관이 바뀌는 경향이 있습니다.

장기적으로는 각 인간 숙의자를 인간보다 훨씬 똑똑하지만 인간과 매우 일치하는 AI 시스템으로 대체할 수 있습니다. 이렇게 하면 사회의 가치에 부합하는 문제를 인간보다 똑똑한 AI 시스템을 인간 한 명에게 정렬하는 문제로 축약할 수 있습니다.

장단점

이러한 종류의 프로세스는 대표성 확보와 적절한 전문성 확보라는 두 가지 핵심 문제에 직면해 있습니다. 미니 퍼블릭에 모집하는 위원의 다양성을 개선하기 위한 다양한 옵션이 있지만, 이 제안의 중요한 장점은 모의 심의에 사용되는 배경이 당시 이용 가능한 인간에 의해 제한되지 않는다는 것입니다. 또한 이 제안은 AI 어시스턴트가 인간이 학습할 수 있는 보다 자연스러운 인터페이스를 제공할 수 있고, 어시스턴트의 편견을 줄이기 위해 명시적으로 노력할 수 있으며, 마지막으로 모방 학습 정책에 테스트 시간 계산을 추가하여 더 똑똑하고 사려 깊고 참여도가 높은 인간인 척하도록 요청할 수 있기 때문에 전문 지식의 가용성을 향상시킬 수 있습니다.

1. 장점: 확장성 및 짧은 지연 시간. 4단계는 완전 자동화가 가능하므로 대규모로 짧은 지연 시간으로 수행할 수 있습니다. 시민 심의는 몇 주 또는 몇 달이 걸릴 수 있지만, 추론 속도와 병렬화에 따라 몇 시간 또는 몇 분 안에 완료할 수 있습니다. 그러나 질문에 따라 짧은 토론으로 해결할 수 없는 가치 있는 질문이 많기 때문에 만족스러운 답변을 도출하는 데 추론 비용이 상당할 수 있습니다.
2. 장점: 투명성. 비공개 정보에 대한 질문이 있는 경우 토론 내용을 공개할 수 있으므로 누구나 열람할 수 있습니다. 따라서 누구나 자신의 의견이 적절하게 반영되었는지 확인할 수 있습니다. 또한, 누구나 자신의 배경을 작성하고 자신이 참여하는 것처럼 시뮬레이션된 미니 퍼블릭을 실행하거나 다른 시뮬레이션된 참가자들과 함께 직접 참여할 수도 있습니다. 또한 사용자는 시뮬레이션의 다양한 노브(예: 숙의 프로토콜, 투표 메커니즘, 구성 등)를 조정하고 결과에 어떤 영향을 미치는지 확인할 수 있습니다.
3. 장점: 더 높은 수준의 협력. 시뮬레이션 토론 참가자가 평소에는 매우 불편해하는 토론에 더 기꺼이 참여하도록 편향시킬 수 있으며, 평소에는 함께하고 싶지 않은 그룹의 시뮬레이션 참가자와 협력하고 타협하려는 의지가 높아질 수 있습니다. 그러나 이렇게 하면 시뮬레이션의 충실도가 떨어집니다.

4. 장점: 개인 정보 보호. 이 프로세스는 일반적으로 많은 관점을 참여시키기 어려운 고도로 기술적이거나 사적인 정보를 연구하고 이해해야 하는 질문에 적용할 수 있습니다. 언어 모델은 모든 관련 정보에 대해 사전 학습되고 일반화를 활용하여 참가자가 실제로 가지고 있지 않은 정보에 대해 어떻게 반응할지 예측할 수 있습니다.
5. 단점: 대표성. 대부분의 사람들이 미니 퍼블릭에 직접 참여하지 않기 때문에 이 제안이 실제 미니 퍼블릭보다 대표성이 떨어진다고 할 수는 없습니다. 미니 퍼블릭이 매력적인 이유는 누구나 선정될 수 있다는 점(시뮬레이션 미니 퍼블릭에서는 그렇지 않음)과 나와 비슷한 사람들이 참여한다는 점(시뮬레이션 미니 퍼블릭에서는 더 그럴 수 있음)입니다. 하지만 이 제안에는 실제로 대부분의 사람들이 포함되지는 않을 것입니다.
6. 단점: 집계 방식이 매우 중요합니다. 많은 작업이 4b 단계의 집계 방법에 의해 수행되며, 잘못된 집계 방법은 이 작업을 비효율적으로 만들 수 있습니다. 그러나 다양한 집계 방법을 사용하여 여러 번의 심의를 실행하여 집계 방법에 대한 견고성을 확인할 수도 있습니다.
7. 단점: 책임 소재 불분명. 무언가 잘못되어 가치 결정을 면밀히 검토하고 싶을 때, 우리는 그 결정에 이르게 된 토론을 읽고 무엇이 잘못되었는지 확인할 수 있습니다. 하지만 사람이 결정을 내린 것이 아니기 때문에 누구를 탓할 수도 없습니다. 우리가 할 수 있는 일은 시스템을 디버그하고, 학습 데이터를 업데이트하고, 다양한 노브를 조정하는 것뿐입니다.
8. 단점: 결과가 나쁠 수 있음. 과정이 민주적이라고 해서 결과가 실제로 합리적이라는 것을 의미하지는 않습니다. 이 프로세스는 맥보트페이싱에 취약합니다.
9. 단점: 고정관념. 사전 학습된 언어 모델은 사전 학습 데이터에 존재하는 유해한 고정 관념을 나타냅니다. 미세 조정을 통해 이러한 고정관념을 줄일 수 있지만, 이를 극복하기 위해서는 많은 노력이 필요합니다.
10. 단점: 사람들이 마음을 바꾸는 과정을 시뮬레이션하는 것은 기술적으로 어렵습니다. 오늘날 시뮬레이션 토론 참가자가 새로운 주제에 대해 학습하고 토론이 진행됨에 따라 생각을 바꾸도록 하는 것은 기술적으로 어렵습니다. 피드포워드 트랜스포머는 상황에 맞는 학습을 수행하지만, 현재로서는 사람이 어려운 주제에 대해 깊이 있게 토론하는 것과 같은 수준에는 미치지 못합니다.

인간의 숙의를 충실하게 모델링하지 않으면 규모와 데이터에 따라 감소해야 하는 실제 위험이 있지만, 배포 범위를 벗어난 경우나 매우 틈새적인 견해에 대해서는 항상 성능이 저하될 수 있습니다. 실제로 중요한 의사 결정에 직면했을 때 시뮬레이션된 민주적 프로세스에 지나치게 의존하지 않도록 주의해야 합니다.

데시데라타desiderata(원하는 것)를 기준으로 한 평가

제 자신의 데시데라타 목록에 따라 이 제안을 확인해 봅시다:

- 포용성: 이 프로세스는 매우 포괄적이며 존재하지도 않는 하위 그룹(예: 1950년대 스웨덴에서 태어나 그렉 이건의 책을 좋아하는 아시아계 트랜스젠더 남성)의 관점까지 시뮬레이션할 수 있습니다.
- 공정성: 이는 주로 집계 프로세스(4b 단계)와 조건으로 삼을 배경 정보를 선택하는 방법에 따라 달라집니다.
- 대표성: 이 프로세스는 직접적인 표현에서 좋은 점수를 얻지 못합니다. 2단계 시연을 기록하는 사람은 다른 모든 사람에 비해 결과에 큰 인과적 영향을 미칩니다. 기술적으로는 누구나 온라인에 게시하여 사전 학습 데이터에 영향을 미칠 수 있지만, 미세 조정 과정에서 자신의 목소리가 보존될지 확실하지 않으며 모든 사람이 온라인에서 자신의 의견을 투명하게 공개하는 것을 편안하게 느끼거나 게시를 좋아하는 것은 아닙니다. 그러나 이 프로세스는 직접적인 인과관계가 없더라도 모든 사람의 의견을 효과적으로 대변할 수 있기 때문에 간접적으로 대표성을 가질 수 있습니다. 특히, 모든 사람이 직접 프로세스에 참여했다면 결과가 어떻게 달라졌을지 쉽게 확인할 수 있습니다.
- 인센티브 조정: 이는 프로세스를 구축하는 사람에 따라 크게 달라집니다.
- 정당성: 이 제안의 한 가지 큰 단점은 사람이 실제로 참여하지 않는다는 것입니다. 대신 대부분의 사람들이 이해하지 못하는 복잡한 기술에 의존합니다.
- 적응성: 1, 2, 3단계를 다시 실행하여 인류의 도덕관 변화, 과학 및 사회 발전, 기타 세상의 변화를 고려하면 학습 데이터를 상당히 쉽게 업데이트할 수 있습니다.
- 투명성: 비공개가 아닌 질문에 대한 토론은 완전히 오픈 소스로 공개될 수 있으며, 누구나 이를 보고 자신의 의견이 반영되었는지, 대표자가 토론에 접근하는 방식에 동의하는지 확인할 수 있습니다. 또한 자동화된 도구를 사용하여 토론에서 가장 동의하지 않거나 자신의 관심사와 가장 관련성이 높은 부분을 표시할 수도 있습니다.
- 단순성: 이 프로세스는 매우 간단하지는 않지만 충분한 인력을 갖춘 소규모 팀에서 구축할 수 있을 정도로 간단합니다.
- 실용성: 오늘날의 언어 모델은 실제 프로토타입을 구축할 수 있을 정도로 샘플 효율이 높은 것으로 보입니다.

마지막으로 무지의 베일 테스트를 적용해 보겠습니다: 내가 사회에서 어디로 가게 될지 모른다면 이 과정에 동의할 수 있을까요? 대안이 무엇인지에 따라 많이 달라집니다. 한 가지 위험 신호는 기술로 인해 발생한 사회 문제를 기술을 이용해 해결하려고 한다는 점입니다. AI에

대한 기술적 지식이 많지 않다면 기술적인 접근 방식에 회의적일 수 있지만, 프로세스를 살펴볼 수 있기 때문에 현재 대부분의 기술 회사가 하고 있는 것보다 훨씬 더 투명하게 느껴질 것입니다(물론 높은 기준은 아닐 것입니다). 여기서 제안한 프로세스가 실제로 실행 가능한 장기 전략으로 판명된다면, 목표를 달성하는 데 효과적이라는 증거를 쌓을 수 있을 것입니다. 즉, 잘 작동하고 올바른 접근 방식이라면 사람들은 이를 신뢰하는 법을 배울 수 있을 것입니다.

관련 노력

이 목록은 결코 포괄적인 목록은 아니지만, 관련 노력에 대한 몇 가지 요점을 정리한 것입니다.

- 숙의 민주주의는 이미 여러 나라에서 중요한 정책 사안에 대해 시도되고 있습니다. 이러한 실험을 통해 숙의를 어떻게 이끌어야 하는지에 대해 배울 점이 많습니다.
- 집단 지성은 탈중앙화된 의사 결정을 개선하기 위한 노력으로, 특히 새로운 기술을 대상으로 합니다.
- 일관된 추정 의지(CEV - Coherent Extrapolated Volition)는 "우리가 더 많이 알고, 더 빨리 생각하고, 우리가 바라는 사람이 되고, 더 멀리 함께 성장했다면"이라는 가치 질문에 답하기 위한 열망적인 목표입니다. 이 제안은 오늘날의 언어 모델로 CEV를 구현하기 위한 구체적인 단계로 볼 수 있습니다. 즉, 인간이 가치 질문에 대해 어떻게 생각하는지 AI에 가르치는 것입니다. CEV를 구현하려면 시뮬레이션된 인간이 시간이 지남에 따라 자신의 견해를 업데이트하여 지적 및 도덕적 진보를 시뮬레이션해야 합니다. 하지만 이 제안의 목적은 현재 인류의 가치를 대변하는 것이기 때문에 인간이 가치관에 대한 생각을 바꿀 때까지 기다렸다가 2단계를 다시 실행하여 가져와야 합니다. 이 방식은 도덕적 진보가 느리고 인간의 속도에 제한된다는 단점이 있습니다.
- 재귀적 보상 모델링과 AI 지원 피드백은 이 제안을 잘 보완합니다: 우리는 AI 행동 평가와 관련된 모든 인지적 노동에 가능한 한 많은 AI 지원을 사용하여 인간이 가치 입력에 집중할 수 있도록 하는 것을 목표로 합니다.
- 도덕적 의회는 서로 다른 도덕 이론을 가진 대표자들이 공동으로 가치 결정을 내리는 의회를 상상함으로써 도덕적 불확실성 하에서 의사 결정을 내리기 위한 이론적 제안입니다. 이는 철학자들이 생각하는 도덕 이론이 사람들이 가치 질문에 답할 때 무엇을 중요하게 생각하는지 예측할 수 있다고 가정합니다. 이 제안의 목표는 추상적인 이상을 표현하는 것이 아니라 실제 사람들이 실제로 문제에 어떻게 접근하는지를 모방하는 것입니다.
- 소셜 시뮬라크라가 가치 문제에 대한 사회적 대화를 모방하는 것이라면, 이 제안은 정보에 입각한 결정을 내리고 타협점을 찾기 위해 고의적으로 노력하는 인간을 모방하는 데 초점을 맞추고 있습니다.

- 최근 딥마인드에서 발표한 논문은 다양한 인구통계학적 그룹의 선호도를 수집하고 이를 바탕으로 언어 모델을 학습한 다음, 그 결과를 다양한 사회복지 함수로 집계하여 이러한 방향으로 나아가는 첫걸음을 내디뎠습니다.

이 아이디어에 대한 토론과 피드백을 제공해준 *Tyna Eloundou, Michiel Bakker, Miles Brundage, Irene Solaiman, Ryan Lowe, Iason Gabriel, Kim Malfacini, Aviv Ovadya, Jeff Wu, Wojciech Zaremba*와 숙의 민주주의를 알려준 *Brian Christian*에게 많은 감사를 표합니다.

---

1 현재의 인공지능 어시스턴트는 진실성에 대한 기준을 충족시키지 못하지만, 곧 그 기준에 도달할 수 있을 것으로 낙관합니다. :)
2 예를 들어, gpt-3.5-turbo는 미국 최저임금을 받는 가장 빠른 타이핑 인력보다 약 200배 저렴하므로 이 모델은 1달러의 비용으로 수십 시간이 걸리는 숙의를 시뮬레이션할 수 있습니다.
3 과거에 인구통계학적 정보가 이를 위해 제안된 적이 있지만, 특정 인구통계 내에서도 다양한 견해가 존재하는 경우가 많기 때문에 좋은 지표가 되지는 못합니다.
4 이 아이디어는 Ryan Lowe에게 공을 돌립니다.
5 이는 이상적으로 이해되길 바라며 이야기하는 것입니다. 위키백과가 완벽한 중립을 지키지 못하는 것은 분명하며, 모더레이터 역할을 맡은 그룹의 견해에 과대표 되는 편향 문제가 있습니다.

더 많은 AI 관련 번역 글 모음: https://bit.ly/3kZPZ9G