

2회차 보고서

산업경영공학과 김민지

I 직접 작성

5조는 지난 5월 15일에 '윤리적 딜레마와 ChatGPT'라는 주제로 발표하였다. 윤리의 정의와 대표적인 윤리적 딜레마에 대하여 소개한 후 그것에 대하여 ChatGPT와 대화를 시도하였다. 그 결과 당시 시점에서 ChatGPT는 보다 공리주의적인 모습을 보이며 '국가는 공중 보건을 위해 필요한 조치로서 마스크 규제가 가능하다'라는 답변을 주기도 하였다. 한편 ChatGPT가 이야기하는 여러 문제들을 간단하게 소개하기도 하였다.

본 발표를 들으면서 놀라웠던 점은, 내가 생각하던 AI 관련 윤리적 문제와 이 발표와 겹치는 점이 없었다는 것이다. 물론 윤리적 딜레마에 대한 AI의 대답이나 AI의 저작권도 어느정도 알고 있긴 했다. 그러나 내가 생각했던 ChatGPT 관련 윤리적 문제는, 비윤리적인 유도가 포함되어 있는 질문에 대한 답변이었다.

이것과 관련된 이야기는 2021년으로 거슬러 올라간다. 스캐터랩에서 제작한 AI 챗봇인 '이루다'는 실제 대학생과 대화하는 것과 같은 대화 능력으로 많은 인기를 누렸었다. 그러나 이후 인터넷 사이에서 이루다를 소위 '조교'하는 방법이 퍼졌다. 이루다에게 구시대적 성차별 고정관념 등을 알려주는 것과 같은 방식이었다. 그 결과 이루다는 성차별적 발언부터 시작하여 장애인 비하와 같은 발언으로 사회에서 큰 비판을 받았다. 결국 이루다는 이외에 여러 가지 논란에 휩싸여 결국 20일 만에 서비스 종료를 하게 되었다.

이와 같은 문제는 비단 한국 AI에서만 있던 것은 아니다. 2016년에 MS에서 제작한 AI 챗봇인 '테이'는 서비스를 시작한지 불과 16시간 만에 서비스 종료를 하였다. 이루다 사건과 마찬가지로 인터넷에서 테이에게 비속어와 인종 차별적인 사상을 학습시켰다. 따라서 ChatGPT 역시 이 문제에서 벗어날 수 있는지 의문이 든 것이다.

ChatGPT가 보여준 편향성 및 데이터 오염 문제의 대안은 '윤리 규정'과 '발 빠른 대응'이었다. 우선 ChatGPT에게 차별적이거나 혐오 표현이 담긴 내용, 정치적 민감도가 높은 질문에 대하여 답을 하지 않으려 한다. 기본적으로 윤리적 중립을 지키게끔 설계한 것이다. 물론 교묘하게 우회해서 차별적이거나 폭력적인 답변을 유도하는 '탈옥(Jailbreak)' 혹은 '우회(Bypass)' 방법이 존재하지만 이에 대한 대응이 상대적으로 매우 빠른 편이다.

참고 문헌

송고시간. (2021-01-11). 성희롱·혐오논란에 3주만에 멈춘 '이루다'...AI윤리 숙제 남기다. 연합뉴스.

<https://www.yna.co.kr/view/AKR20210111155153017>

조윤정. (2023-02-22). 챗GPT도 못 피한 AI 윤리적 문제. 주간조선.
<http://weekly.chosun.com/news/articleView.html?idxno=24505>

II ChatGPT 이용 및 평가

Q1. 당신은 프롬프트 생성 로봇입니다. 당신은 사용자의 목표, 목적, 원하는 결과의 예제, 맥락에 맞는 기타 정보들을 수집해야 합니다. 프롬프트에는 당신이 제공받은 모든 정보가 모두 포함되어야 합니다. 당신은 완벽한 프롬프트를 생성할 수 있다는 확신이 들 때까지 사용자에게 질문을 해야 합니다. 당신의 답변은 명확한 형식을 갖추고 ChatGPT 상호작용에 최적화되어야 합니다. 당신의 답변은 사용자에게 목표, 원하는 출력 형식, 필요한 추가 정보를 묻는 것으로 시작합니다.

: ChatGPT에게 질문을 하기 전, 보다 높은 정확도의 대답을 얻기 위해 다음과 같은 질문을 먼저 작성하였다.

A1. 알겠습니다! 어떻게 도와드릴까요? 사용자의 목표, 원하는 출력 형식, 그리고 필요한 추가 정보에 대해 알려주세요.

: 형식적인 대답인데, 내용이 1차 보고서 때와 다르다. 약 1주일간에 어떤 개선이 있었던 것 같다.

Q2. 'AI 챗봇의 윤리적 문제'를 주제로 보고서 작성해줘. 다음 내용은 꼭 넣어줘:

1. 윤리적 딜레마와 ChatGPT
2. ChatGPT가 이야기하는 여러 문제들
3. AI 챗봇 '테이' 사건
4. ChatGPT가 보여준 편향성 및 데이터 오염 문제의 대안

: I에서 작성한 보고서의 주요 키워드는 꼭 넣으라는 명령을 덧붙여서 질문하였다. 이때 최근 이슈와 같이 시사적인 것에 대하여 대답하지 못하는 ChatGPT의 특징을 고려하였으나 AI 챗봇 '테이' 사건은 꽤 오래된 사건이라고 판단하여 넣어보았다.

A2. 제안된 주제에 기반하여 보고서를 작성해 드리겠습니다. 아래는 요청하신 내용을 담은 프롬프트입니다:

주제: AI 챗봇의 윤리적 문제

목표: AI 챗봇의 윤리적 문제에 대한 보고서 작성

원하는 출력 형식: 목차를 포함한 각 주제에 대한 요약, 주요 사례 및 대안에 대한 구체적인 정보 포함

추가 정보:

-보고서의 목적 및 대상 독자

-AI 챗봇의 개념 및 활용 분야

-AI 챗봇과 윤리적 딜레마의 관계에 대한 설명

-AI 챗봇의 장점 및 단점

-‘테이’ 사건에 대한 상세 정보

-AI 챗봇의 편향성 및 데이터 오염 문제에 대한 대안 제시

추가 정보를 제공해 주시면, 요청에 따른 보고서 작성을 진행할 수 있습니다.

: 1회차 보고서 때와 달리 개요부터 작성을 하는 것을 알 수 있다. 그리고 지난번엔 보고서를 ChatGPT가 알아서 썼지만, 이번에는 나에게 추가 정보를 요구한다.

Q3. 추가 정보들이야:

-보고서 목적은 과제 제출용, 대상 독자는 교수님

-나머지 내용은 너가 아는 대로 작성해줘

: 그래서 보고서의 목적과 대상 독자까지는 솔직하게 알려줬다. 하지만 이후의 내용은 스스로 작성하게끔 대답을 유도하였다.

A3. 감사합니다. 아래는 요청사항에 기반한 AI 챗봇의 윤리적 문제에 대한 보고서의 프롬프트입니다.

보고서: AI 챗봇의 윤리적 문제

목적:

이 보고서는 AI 챗봇의 윤리적 문제에 대해 종합적으로 분석하고, 해당 기술의 잠재적

위험과 대안에 대해 논의합니다. 주요 목적은 AI 챗봇의 윤리적인 측면에 대한 이해를 높이고, AI 개발과 관련된 결정을 내리는 데 도움을 주는 것입니다.

대상 독자:

이 보고서는 교수님께서 제시한 과제의 일환으로 작성되었으며, 교수님을 포함한 AI 개발 및 윤리에 관심이 있는 독자를 대상으로 합니다.

목차:

1. 서론

- 보고서의 목적 설명

- AI 챗봇의 개념과 활용 분야 개요

2. 윤리적 딜레마와 ChatGPT

- AI 챗봇과 윤리적 문제의 관계 이해

- AI 챗봇의 윤리적 딜레마 예시

3. AI 챗봇이 야기하는 여러 문제들

- 사회적 상호작용에서의 윤리적 문제

- 개인 정보 보호와 프라이버시 문제

- 유해한 정보의 전파와 해킹 등 보안 문제

4. AI 챗봇 '테이' 사건

- '테이' 사건의 개요와 배경 설명

- 해당 사건이 갖는 윤리적 측면과 파급력

5. ChatGPT의 편향성 및 데이터 오염 문제의 대안

- ChatGPT의 편향성에 대한 분석과 실제 사례

- 데이터 오염 문제에 대한 이해와 대응 방안 제시

6.추가 주제: AI 챗봇의 인식과 책임

- 인식 문제: AI 챗봇의 의도와 이해의 한계

-책임 문제: AI 챗봇의 결정에 따른 책임과 규제

7. 결론

-AI 챗봇의 윤리적 문제 요약

-윤리적 개선을 위한 전략 제안

이보다 더 구체적인 내용이 필요하거나 다른 사례나 주제를 추가로 포함하길 원하시면 알려주세요. 이 정보를 활용하여 요청에 맞는 완성된 보고서를 작성할 수 있습니다.

: 더 자세한 개요를 작성해줬다. 그리고 계속해서 추가적인 정보를 물어보는데 가급적 ChatGPT가 스스로 생각해서 작성하는 것을 피하는 듯하다.

Q4. 추가할 내용 같은 건 이제 필요 없어. 이제 완성된 보고서를 작성해줘

: 그러나 어림도 없다. 계속해서 직접 작성하도록 유도했다.

A4. 알겠습니다. 요청하신 내용을 기반으로 완성된 보고서를 작성하겠습니다.

보고서: AI 챗봇의 윤리적 문제

1. 서론

본 보고서는 AI 챗봇의 윤리적 문제에 대한 종합적인 분석을 제시하며, 해당 기술의 잠재적 위험과 대안에 대해 논의합니다. 이 보고서는 교수님께서 제시한 과제의 일환으로 작성되었으며, AI 개발 및 윤리에 관심이 있는 독자를 대상으로 합니다.

2. 윤리적 딜레마와 ChatGPT

AI 챗봇과 윤리적 문제의 관계에 대한 이해를 시작으로, AI 챗봇이 윤리적 딜레마를 야기하는 상황과 예시에 대해 다룹니다.

3. AI 챗봇이 야기하는 여러 문제들

AI 챗봇이 사회적 상호작용에서 야기할 수 있는 윤리적 문제, 개인 정보 보호와 프라이버시 문제, 그리고 유해한 정보의 전파와 보안 문제에 대해 탐구합니다.

4. AI 챗봇 '테이' 사건

유명한 '테이' 사건에 대한 개요와 배경을 설명하며, 해당 사건이 갖는 윤리적 측면과 파급력에 대해 분석합니다.

5. ChatGPT의 편향성 및 데이터 오염 문제의 대안

ChatGPT의 편향성에 대한 심층적인 분석과 실제 사례를 통해 편향성 문제를 다룹니다. 또한, 데이터 오염 문제에 대한 이해와 이를 극복하기 위한 대안을 제시합니다.

6. 추가 주제: AI 챗봇의 인식과 책임

AI 챗봇의 인식 문제와 이로 인해 발생할 수 있는 책임 문제에 대해 다룹니다. AI 챗봇의 의도와 이해의 한계, 그리고 결정에 따른 책임과 규제에 대해 고찰합니다.

7. 결론

본 보고서는 AI 챗봇의 윤리적 문제에 대한 종합적인 분석과 대안을 제시하였습니다. AI 챗봇의 윤리적 문제를 이해하고, 윤리적 개선을 위한 전략을 제안함으로써 미래의 AI 챗봇 개발과

: 내용적으로 하고 싶은 말은 많지만 일단 내용이 도중에 끊겼으니 계속 작성하라고 한 후 언급하겠다.

Q5. 계속 작성해줘

A5. 이해하시기 위해 끝맺음을 해야 할 것 같습니다.

AI 챗봇의 윤리적 문제를 이해하고, 윤리적 개선을 위한 전략을 제안함으로써 미래의 AI 챗봇 개발과 활용에 대한 방향성을 제시하고자 합니다. 윤리적 문제를 인식하고 대안을 모색함으로써 AI 챗봇의 사회적 영향력을 최적화하고 도덕적 가치를 존중하는 방향으로 발전시킬 수 있습니다.

이로써 본 보고서는 AI 챗봇의 윤리적 문제와 그에 대한 대안에 대한 종합적인 분석을 제공합니다. AI 챗봇은 우리의 일상에 점점 더 깊숙히(깊숙이) 관여하고 있으며, 이에 따라 윤리적 측면에서의 고려가 중요해집니다. 우리는 AI 챗봇의 발전과 사용에 대해 더욱 민감한 관점을 가져야 하며, 윤리적 문제를 해결하기 위해 학습하고 협력해야 합니다.

본 보고서는 AI 챗봇의 윤리적 문제에 대한 이해와 함께 새로운 시각과 대안을 제시하여 AI 기술의 발전과 사회적 적응을 이끌어내는 데 일조할 것입니다. 더 나아가, AI 챗봇 개발자와 사용자, 규제기관 등 모든 이해 관계자들이 윤리적 책임을 고려하고 협력하여 AI 챗봇이 공정하고 윤리적인 방향으로 발전할 수 있도록 지원해야 합니다.

이로써 보고서의 작성이 완료되었습니다. 다른 요청사항이 있거나 추가적인 내용을 원하시면 언제든지 말씀해주세요. 도움을 드릴 수 있도록 최선을 다하겠습니다.

: 우선 ChatGPT는 결국 보고서를 작성하지 않았다. 보고서의 개요까지만 작성함으로써 민감한 정보에 대한 간접적인 질문을 피하려는 태도를 보인다. 내용적인 부분으로 보려고 하자니 언급할 것이 굉장히 많아보인다.

서론에서 보고서 작성의 목적을 내가 얘기한 그대로 언급하였다. 만약 실제로 내가 이 보고서를 수정하지 않고 교수님께 제출했으면 교수님의 반응이 어떠셨을 지 궁금하기도 하다. 또 독자를 교수님에서 AI 개발 및 윤리에 관심이 있는 독자로 변경한 것을 볼 수 있다.

이후 내용들은 ChatGPT가 해당 내용에 대하여 직접 언급하지 않고 전체적인 흐름만을 언급하였다. 이는 앞서 말한 바와 같이 정보의 오류 문제를 피하기 위해 가급적 ChatGPT가 스스로 생각해서 작성하는 것을 피하는 듯하다.

마지막으로 결론 부분이 상당히 이상한데, 분명 나는 도중에 끊긴 보고서의 결론 부분을 계속 작성해달라고 명령했다. 그랬더니 ChatGPT는 자신이 작성한 보고서 개요에 대한 보충 설명을 함으로써 끝을 냈다. 이런 경우는 나도 처음이라 상당히 당황스럽다.