

챗봇 결과 보고서

(1조) 20180032 남기동

20200046 박민하

20210065 신지원

20220072 서호영

목차

I. 개발 취지

II. 챗봇의 페르소나

III. 챗봇의 구조

IV. 중간 발표 피드백 반영 – 출처 문제

V. 대표 시나리오 소개

V-1 상호작용이 가능한 시나리오

V-2 같은 주제, 서로 다른 의견 비교

V-3 퀴즈를 활용한 시나리오

VI. 부가적인 기능

VI-1. 철학자 명언

VI-2. 설문지를 활용한 사용자 의견 추가

VII. 참고자료

I. 개발 취지

최근 인공지능이 발전함에 따라 많은 윤리적 문제들이 뒤따르고 있다. 국내 기업 스캐터랩에서 출시한 챗봇 이루다의 차별, 혐오 발언과 오픈AI의 챗봇 ChatGPT의 저작권 문제들이 대표적인 인공지능의 윤리적 이슈이다. 이렇게 윤리 문제라고 하면, 인종이나 성별, 소득 간의 차별과 혐오 발언, 저작권, 초상권 문제 같은 단기적인 문제를 떠올리기 쉬운데, 저희 조는 표면적으로 드러나는 윤리적 문제와 더불어, 인공지능 시대 속에서 인간의 삶, 인간의 역할, 그리고 인간의 존엄과 가치와 관련하여 보다 고차원적인 철학적 논의가 필요하다고 생각했다. 인간이 AI를 활용하는 것인지 AI가 인간의 영역을 침범하는 것인지 모호한 지금, 가장 필요한 것은 인간에 대한 성찰이기 때문이다.

그런데 어디서 이러한 철학적인 논의를 할 수 있을까? 기업은 의도적으로 윤리적인 이슈에 대한 언급을 피하고, 인공지능 철학은 신생 학문으로서 정보가 부족한 것이 사실이다. 이에 저희는 챗봇을 철학적 공간으로서 활용하고자 하였다. 철학적 공간을 조성하여 인공지능과 관련한 윤리적, 철학적 토론을 챗봇 내에서 가능하게 만드는 것이 1조의 챗봇 개발 목표이다.

II. 챗봇의 페르소나

우리 조의 챗봇 페르소나는 '로올라 정령'이다. 로올라 정령이라는 이름의 앞부분인 '로올라'는 서강대학교 도서관의 이름을 따온 것이며 이는 가톨릭의 수도자이자 사제이며 예수회의 창설자인 성인 이냐시오 데 로올라에서 따온 것이다. 정령인 이유는 오랜 시간동안 서강대학교의 도서관에서 지내면서 다양한 서적을 읽고 학습하여 챗봇을 통해 철학적 질문을 묻고자 하는 이용자에게 멘토의 역할을 잘 수행할 수 있는 이미지를 갖고 있기 때문이다.

로올라 정령이라는 페르소나는 단순히 가상적인 캐릭터로만 존재하는 것이 아니라 챗봇 사용자가 대화를 하는 중간 중간 상징하는 이미지를 통해 모습을 드러낸다. 로올라 정령에 사용한 이미지는 '서강프렌즈' 중 하나인 '알로스'이다. 서강프렌즈는 서강대학교 재학생을 대상으로 진행한 '이모티콘 공모전'에서 수상한 작품들로 학생들이 공감할 수 있는 캠퍼스 생활, 학사 제도 등을 재미있게 표현한 이모티콘이다. 서강프렌즈는 총 3개의 캐릭터가 있는데 서강이, 알로스, 그리고 소강르타이다. 그 중 알로스는 서강대학교를 대표하는 상징물인 알바트로스 새와 차별화된 캐릭터이며 교색이 주홍색 후드티를 입고 있다.



페르소나로 정한 로올라 정령은 도서관에 살고 있는 정령이라는 이미지와 챗봇을 이용하는 사람에게 교수나 권위 있는 학자와 같은 존재가 아닌, 친숙하게 설명하고 이야기해주는 멘토 이미지를 갖는다. 서강프렌즈 중 하나인 알로스는 서강대학교 재학생에게 매우 친숙한 복장을 하고 있으며 특히나 시험기간에 공부할 때 도서관을 찾는 학생들이 편하게 입는 후드티를 착용하고 있기 때문에 로올라 정령이 추구했던 이미지와 가장 잘 부합했다. 또한 알바트로스 새의 권위 있고 엄중한 이미지와 달리 친근한 이미지를 가진 모습을 하고 있기 때문에 역시 멘토의 이미지를 추구했던 페르소나에 가장 적합하다고 판단되었다.

로올라 정령은 챗봇 내에서 두 가지의 역할을 한다. 첫 번째로 로올라 정령은 인공지능 주제와 관련한 철학적 논의를 주도하는 역할을 한다. 기존에는 인공지능 주제와 관련한 철학적 논의를 매개하는 역할을 하고 챗봇 사용자가 원하는 토론 주제를 선택하면, 로올라 정령이 해당 주제에 맞는 철학자를 소환하는 구조로 챗봇을 설계하였다. 그러나 챗봇이 주로 다루는 주제인 인공지능에 관련하여 철학적 논의를 진행했던 학자들은 대부분 살아 있는 학자들이 많으며, 이 경우에 출처 문제가 복잡해지기 때문에 이를 수정해야만 했다. 따라서 로올라 정령이 기존의 매개자

의 역할을 넘어 인공지능 주제와 관련한 철학적 논의를 주도적으로 이끌어가는 역할을 하도록 설정을 변경하였다.

두 번째로 로올라 정령은 개발자의 철학적인 의견을 제시하는 역할을 한다. 결국 이번 프로젝트를 통해 개발하는 챗봇의 역할은 단순히 인공지능 주제와 관련한 철학적 논의의 정보만을 전달하는 것이 아니다. 오히려 더 중요한 부분은 챗봇을 통해 사용자와 인공지능 주제에 관한 철학적 논의를 진행하고 발전시키는 것이며, 이를 위해서는 분명 챗봇을 통해 개발자의 철학적 견해가 제공되어야 한다. 물론 개발자의 철학적 견해를 제공하는 것은 그러한 특정 견해가 정답이라거나 그러한 특정 견해를 사용자에게 관철시키기 위함이 아니다. 이 챗봇은 철학적 토론이 이루어지는 공간인 만큼, 단순한 정보전달을 넘어 개발자 역시 철학적인 의견을 제시하여 사용자도 스스로의 견해와 입장을 정립하고 발전시킬 수 있도록 돕는 것을 목표로 한다.

Ⅲ. 챗봇의 구조

앞서 설명되었던 페르소나를 바탕으로, 챗봇의 기능적 구조는 기본적으로 페르소나인 '로올라 정령'에 의해 사용자가 원하는 요청과 질문에 대한 답변을 로올라 정령이 이야기하는 구조이다. 하지만 페르소나 설명 부분에서도 언급하였듯, 로올라 정령은 단순히 답을 알려주는 기계가 아니라 철학적 질문과 논의를 하고 싶어하는 사용자에게 도움이 될만한 정보와 개인적인 의견도 간략히 이야기해줄 수 있는 '멘토'와 같은 역할이다. 그렇기 때문에 철학적 질문에 대한 답을 확고히 내리거나 단언하기 보다는 중립적인 정보를 제공하고 이와 관련하여 생각해볼 수 있는 철학적 사고를 소개하는 역할을 하고 있다. 이는 기존의 카카오톡 챗봇이 가지고 있는 이미지와 구분된다. 기존의 카카오톡 챗봇은 친밀한 친구처럼 대화를 나눌 수 있는 것을 목표로 하거나 궁금한 사항에 대해 명확한 답을 내려주는 것을 목표로 하기 때문이다.

우리 조의 챗봇, 즉 로올라 정령이 멘토의 역할을 하기 때문에, 이러한 역할에 부합하는 이미지를 구축하기 위하여 '서가 이동'이라는 구조를 채택하고 있다. '서가 이동'은 로올라 정령이 대화를 나눌 수 있는 10가지의 철학적 주제를 보여주고 사용자에게 관심있는 주제를 물어볼 수 있도록 유도하며 특정 주제를 원한다는 요청에 따라 해당 주제에 관련된 서가로 이동하는 것을 의미한다. '서가 이동'이 가지고 있는 장점은 크게 2가지가 있다.

첫 번째로 '서가 이동'을 도입함으로써 서로 다른, 그리고 다양한 철학적 주제에 대해 독립적인 대화를 나눌 수 있는 환경을 만들 수 있다. 철학적 논의는 서로 관련 없는 이야기들을 끊임없이 질문하는 것이 아니다. 하나의 질문을 통해 이루어진 철학적 논의를 토대로 관련된 내용에 대해 궁금증을 가지고 내용을 심화시키는 것이 필요하다. 이러한 목적에서는 현재 어떤 주제에 대해 이야기를 나누고 있는지 사용자에게 명확하게 인지시켜야 한다. 따라서 10가지 주제 중 원하는 주제를 고르고 관련 서가로 이동했다는 설정은 사용자로 하여금 해당 주제에 대한 철학적 논의가 시작되었고 다른 주제와는 독립성 있는 대화와 논의를 할 수 있다는 인상을 줄 수 있다.

두 번째로 '서가 이동'을 도입하는 것은 한 주제에 대해서도 서로 다른 입장을 가진 철학자의 이야기를 종합하는데 용이하다. 뒤의 대표 시나리오에서 자세히 설명되겠지만, 챗봇이 가진 특징 중 하나는 관련된 주제의 철학자의 사상을 일부 소개하는 것도 포함하고 있다. 하지만 특정 주제에 대해서는 두 명의 철학자가 서로 다른 입장을 견지하는 경우도 있다. 이런 경우 동일한 주제에 대해서 이야기하지만 이야기가 통일되지 않게 이해될 수 있다. 하지만 '서가 이동'의 구조를 채택함으로써 한 주제에 서로 다른 철학자의 주장도 결국에는 한 주제에 관한 것이라는 통일성을 줄 수 있다. 대화의 통일성은 사용자가 철학적 논의에 적극적으로 몰두할 수 있도록 하는 중요한 요소이므로 '서가 이동'을 통해 한 주제에 대한 통일성을 확보하는 것이 큰 도움이 된다.

IV. 중간 발표 피드백 반영 – 출처 문제

중간 발표에서 우리 조는 '철학자와 사용자 간의 직접적인 대화 구조'를 챗봇의 가장 큰 특징으로 삼았고, 챗봇은 그 두 엔티티를 매개하는 방식으로 철학적 논의를 진행하는 것으로 하였다. 그러나 중간 피드백 이후, 우리는 기존 챗봇 방식에 두 가지의 큰 문제점이 있음을 인지하였다.

첫째, 앞서 언급되었듯이 챗봇에 직접 등장하는 철학자들이 대부분 현재 살아있다는 점이다. 챗봇이 최근에서야 활발히 논의되고 있는 '인공지능'을 핵심 주제로 하고 있기 때문에, 학자의 개인적인 의견을 챗봇에 반영하는 과정에서 '학자의 직접적인 발화'를 하나하나 인용하는 방식을 사용하거나, 학자가 저술한 도서나 학술 문서를 참고하여 발화를 재서술하는 방식 등이 있다. 그러나 전자의 경우, 학자의 본래 발화 의도와 맥락을 무시한 일부 인용으로 내용이 왜곡될 수 있다는 문제점이 있다. (그리고 챗봇 속 철학자 모든 발화에 출처를 다는 것 역시 매우 번거로운 작업이며 사실상 불가능하다.) 후자의 경우에는 철학자 본인의 동의 없이 자의적으로 철학자를 대변하는 것이므로 이는 윤리적으로 옳지 못하다고 판단하였다. 따라서 우리는 철학자와 사용자 간의 직접적인 대화 방식을 폐기하는 대신, 서가 이동 안내자로서의 챗봇의 역할을 보다 심화시키는 방식을 택했다.

둘째, 우리는 중간 발표와 피드백을 통해, 철학자의 의견을 챗봇에 사용할 때 '출처 문제'를 어떻게 해결할 것인지 그 방법을 고안해야 했다. 단순히 하나의 말풍선 속 챗봇의 발화가 끝날 때마다 출처도 함께 전달하는 방식은 사용자가 출처를 바로바로 확인할 수 있다는 장점이 있지만 챗봇과의 철학적 논의 흐름 및 몰입도를 깰 수 있다는 명확한 단점이 존재한다. 따라서 우리는 사용자가 쉽고 빠르게 출처를 확인할 수 있게 하면서도 철학적 담화가 이루어지는 상황에서는 숨겨놓는 방식을 도입하였다.

'자율 주행 자동차' 시나리오의 예시를 통해 챗봇의 새로운 '출처' 기능을 살펴볼 수 있다. 자율 주행 자동차 논의의 여러 소주제들 가운데 '자율 주행 시스템의 책임 문제'에서는 최근 '책임 소재'를 명확하게 하기 위해서 차량 내 감시 장치를 두어야 한다는 새로운 주장과 자율 주행의 레벨 3 자동차의 경우에는 조종권이 바뀌는 순간에서의 책임 문제가 발생한다는 내용을 포함한다.

이 부분을 읽은 사용자들은 이와 관련된 최신 기사들 및 관련 학술 문서를 확인하고 싶을 것이다. 따라서 '자율 주행 자동차의 책임 문제'를 다룬 시나리오가 끝난 뒤, 맨 아래에 "출처" 노란 버튼을 삽입하였고, 사용자로 하여금 쉽고 빠르게 관련 문서들의 링크를 확인할 수 있게 하였다. 그 외의 모든 시나리오들도 하나의 소주제 논의가 끝날 때마다 하단에 노란색 '출처' 버튼을 두어, 사용자가 읽었던 내용의 원본 문서를 다시 읽어보며 추가 심화 학습을 할 수 있게 하였다. 나아가, 원본 사진의 출처를 확인할 수 있게 하여 기존 중간 발표에서 제기되었던 글과 이미지의 출처 문제를 해결할 수 있도록 하였다.

V. 대표 시나리오 소개


기본적으로 챗봇의 모든 시나리오는 단순 정보전달을 넘어 토론, 즉 개발자-사용자 간의 상호작용이 가능하다. 이 10개의 시나리오 중에서도 챗봇의 특색이 두드러지는 3가지 시나리오를 대표 시나리오로 선정하여 자세히 설명하고자 한다.

V-1 상호작용이 가능한 시나리오

오전 10:03

자율 주행 자동차의 의미

인공지능캐릭터디자인



"자율주행자동차"란 운전자 또는 승객의 조작 없이 자동차 스스로 운행이 가능한 자동차를 의미해.

(자동차관리법 제2조 제1호의3)

오전 10:03

뭔가 확 와닿지는 않는데?

오전 10:08

인공지능캐릭터디자인

하하, 나도 처음에는 "자율 주행 자동차"라는 말만 들었을 때 무슨 자동차인지 상상이 잘 안 가더라고.

그럼 내가 자율 주행 자동차에 "자율"이라는 개념을 조금 더 자세하게 설명해 줄게!

오전 10:08

오전 10:08

"자율"의 의미 보충 설명

인공지능캐릭터디자인

사실 "자율"이라는 개념은 철학이나 윤리학에서 주로 논의되는 개념이야. 철학에서 "자율"이란 외부의 강제 없이 주체가 스스로 자신의 결정을 내릴 수 있다는 것을 의미해. 이는 곧 '이성'을 가지고 스스로 자신이 따를 법칙들을 스스로 정립한다는 의미지.

근데 이 의미를 자동차에 그대로 적용하는 게 가능할까? 아니지. 절대 불가능하지! 자동차는 이성도 없고 자동차가 다르게 될 특정 법칙들은 인간에 의해 만들어진 것들이니까.

그래서 "자율 주행 자동차"에서 "자율"은 조금 다른 의미로, 자동차가 단순히 '인간의 직접적인 개입 없이' 비교적 오랜 시간동안 운행할 능력이 있다는 의미에서 사용되고 있어!

이제 "자율 주행 자동차"가 어떤 건지 잘 알겠지? 그럼 본격적으로 자율 주행 자동차 관련 윤리 문제를 다루기 전에 잠시 살펴볼 내용이 있어.

자율 주행에도 수준별로 여러 단계가 있다는 거 아니?


잠시 간단하게 살펴보고 윤리 이야기로 넘어가자!

오전 10:08

챗봇을 부른 뒤, '자율 주행 자동차' 주제를 선택하면, 이 '자율 주행 자동차'라는 하나의 주제로 여러 철학적 논의들을 이어나갈 수 있다. 따라서 사용자들로 하여금 '자율 주행 자동차'의 의미, 특히 여기서 '자율'이라는 의미가 철학적 논의에서의 '자율'과 어떤 차이점이 있는지, 개념적 차원에서 명확하게 할 필요가 있다. 따라서 자율 주행 자동차의 의미를 단순히 한 줄로 명시하고 넘어가는 것이 아니라, 여러 단계의 발화를 통해 사용자가 '자율 주행 자동차'의 개념을 정확하게 짚고 넘어갈 수 있게끔 하였다.

또한, 이 '자율 주행 자동차' 시나리오에서는 챗봇과 사용자가 다양한 방식으로 상호작용할 수 있게끔 구조화되어 있다.

(1) 챗봇-사용자의 인터랙션 (읽음 확인 리액션)



인공지능캡스톤디자인


미국 자율 주행 자동차 엔지니어협회 (the Society of Automotive Engineers, SAE)와 자율 주행 자동차 산업 독일 연합 (the German Association of the Automotive Industry, VDA)은 자율 주행의 수준을 분류하기 위한 체계를 제시했어.

아래에 이 분류 체계를 정리해봤는데 한번 읽어볼래?

다 읽었으면 아래의 노란 버튼을 눌러줘!

오전 10:16

다 읽어봤어! 멋진데?



인공지능캡스톤디자인

그치? 특히 레벨 5의 자동차는 "모든 상황"에서 모든 운전 기능을 수행할 수 있다고 하잖아.

그러면 사진에서 보이는 것처럼 자율 주행 자동차가 운전의 모든 것을 담당하고, 사람 운전자는 편하게 책을 읽거나 잠을 자는 등 휴식을 취할 수 있는 거지! 레벨 5의 자율 주행 자동차가 상용화된다면, 운전이라는 개념이 아예 사라질 수도 있겠네!

레벨 3 : 조건적 자율 주행
(운전자는 항상 시스템을 감시할 필요가 없다. 어떤 상황을 하에서는 시스템이 자동적으로 작동할 수 있다. 시스템이 조종을 다시 돌려주기까지 시간(가령 10초 정도)을 운전자에게 준다.)

레벨 4 : 고도 자율 주행
(자동차는 표준 상황을 하에서 모든 운전 기능을 수행할 수 있다. 운전자는 표준적 상황 하에서 조종하도록 요구받지 않는다. 비 표준적 상황은 굵은 낱씨를 포함할 것이다.)

레벨 5 : 완전 자율 주행
(자동차는 모든 상황에서 모든 운전 기능을 수행할 수 있다.)

오전 10:11

다 읽어봤어! 멋진데?

자율 주행 자동차의 책임 문제를 다루기 위해서는 자율 주행 자동차 협회에서 정의된 '자율 주행의 수준 분류 체계'의 레벨 3(조건적 자율 주행)에 대해 사전에 인지하고 있어야 한다. 따라서 이 내용을 자율 주행 자동차의 개념 다음 부분에 자연스럽게 배치하였다. 그러나, 자율 주행의 수준별 구분 내용에 해당하는 텍스트의 양이 많아, 사용자가 다 읽지 않고 넘어갈 가능성이 매우 높다. 따라서 챗봇과의 상호작용을 삽입하여 사용자가 분류 체계의 내용을 읽게끔 유도하였다. "다 읽었으면 아래의 노란 버튼을 눌러줘!"라는 챗봇의 발화와 내용 다음에 위치한 버튼("다 읽어봤어! 멋진데?")을 통해 보다 효과적으로 내용을 전달할 수 있게끔 했다. 나아가, 사용자가 실제로 챗봇과 대화하는 느낌을 부여할 수 있다는 효과도 있을 것이다.

(2) 챗봇-사용자의 인터렉션 (상황 가정 - 사용자의 선택)

자율 주행 자동차 책임 문제

오전 10:28

인공지능캡스톤디자인

자율 주행 자동차 운행에 있어서 가장 우려되는 문제 중 하나는 바로 "사고 발생의 책임이 누구에게 귀속되어야 하는가"야.

자율 주행 자동차라 하더라도, 인간이 자동차의 운행을 책임지고 있었을 때는 당연히 해당 사고의 책임이 인간에게 귀속되어야 할 거야. 그 반대의 경우로 '자율 주행 시스템'이 운행을 맡고 있었을 때에는 그 책임이 자율 주행 자동차를 제조한 회사와 해당 시스템을 개발하고 작동 중인 회사에게 귀속되어야 겠지.

근데 여기서 문제가 있어.

오전 10:28

인공지능캡스톤디자인

그런데 여기서 몇 가지 애매한 지점들이 있어.

우선 사고가 발생했을 때, 인간과 자율 주행 시스템 중 누가 운행을 책임지고 있었는지 규명하는 것이 쉽지는 않을 거야. 그래서 최근에는 책임 소재를 분명하게 하기 위해서, 자율 주행 자동차 내에 감시 장치를 장착하는 것을 필수적으로 요구해야 한다는 주장이 제기되고 있어.

또 한 가지로는, <자율 주행의 수준별 구분>에서 '레벨 3'의 자율 주행 자동차의 경우에는, 기본적으로는 자율 주행 시스템이 자동차를 운행하지만, 특정 상황에서는 인간 운전자의 조종이 필요해.

그래서 그 조종권을 자율 주행 시스템에서 인간에게 돌려주는 데에는 약 10초 정도의 시간이 걸린다고 하는데, 만약 이 10초의 시간 사이에 사고가 발생한다면 사고 발생의 책임은 누구에게 귀속되어야 할까?

시스템 잘못이니 회사 책임

이건 운전자 책임이야!

흐음... 너무 애매한데?

오전 10:29

다음은 '자율 주행 자동차의 책임 문제'를 다루는 시나리오이다. 우선, 자율 주행 시스템이 근본적으로 가지고 있는 책임 귀속 문제에 대해서 사용자에게 간단하게 설명하고 있다. 총 2가지의 케이스를 예시로 들고 있는데, 먼저, 사고가 발생했을 때 누가 자동차의 운행을 책임지고 있었는지 명확하지 못한 경우가 있다는 점을 설명한다. 그 다음으로 앞서 제시했던 "자율 주행의 수준별 구분"에서 레벨 3 자동차의 경우, 조종권이 인간에서 자율 주행 시스템으로, 혹은 자율 주행 시스템에서 인간으로 넘어가는 그 시점의 책임이 모호하다는 것을 사용자에게 제시한다. 그 다음, 이 경우에 사고 발생의 책임이 누구에게 귀속되어야 할지 사용자에게 3가지 선택지를 제시하여 고르게 만든다. ('시스템 잘못이니 회사 책임', '이건 운전자 책임이야!', '흐음...너무 애매한데?')

시스템 잘못이니 회사 책임

오후 11:03

로올라 별관

글쎄...

자동차의 조종권이 자율 주행 시스템에서 인간 운전자로, (혹은 그 반대로 인간 운전자에게서 자율 주행 시스템으로) 옮겨가던 그 찰나의 순간에 사고가 난 거잖아? 그러면 그 사고의 원인을 자율 주행 시스템과 인간 둘 중에 한 명으로 꼽히는 건 매우 어려운 일이 아닐까?

오후 11:04

호음... 너무 애매한데?

로올라 별관

맞아, 나도 이 문제는 정말 애매하다고 생각해.

자동차의 조종권이 자율 주행 시스템에서 인간 운전자로, (혹은 그 반대로 인간 운전자에게서 자율 주행 시스템으로) 옮겨가던 그 찰나의 순간에 사고가 난 거잖아? 그러면 그 사고의 원인을 자율 주행 시스템과 인간 둘 중에 한 명으로 꼽히는 건 매우 어렵겠지.

이처럼 자율 주행 자동차가 사고를 발생시켰을 때 그 책임을 시스템(회사)과 인간 운전자, 둘 중에 누구에게 귀속시켜야 할지가 정말 애매해.

이건 앞으로 자율 주행 자동차를 개발하는 회사들과 자율 주행 시스템을 이용하고자 하는 사회 구성원들이 해결해야 할 영역으로 남아있지!

그럼 다음 주제로 넘어가볼까?

이제 철학적 토론 할래!

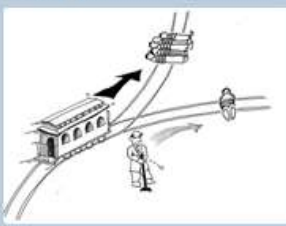
오후 11:04

자율 주행의 윤리적 이점 우려되는 시스템 보안 문제

답변은 총 2가지로 제시된다. "시스템 잘못이니 회사 책임" 또는 "이건 운전자 책임이야!"를 골랐을 경우, 가정된 상황을 다시 차근차근 설명하며, 사고의 원인이 둘 중 하나로 명확하게 좁혀지는 것은 어려움을 강조한다. 그리고 자율 주행 자동차의 책임 문제는 여전히 해결해야 할 과제로 남겨져 있음을 언급하며 해당 주제(자율 주행 자동차의 책임 문제)를 종결한다. 만약, "흐음... 너무 애매한데?"라는 고민의 답지를 고르게 된다면 챗봇은 사용자의 의견에 동의하며 위와 동일한 내용으로 해당 주제를 마무리한다.

이렇게 사용자에게 특정 가정된 상황과 함께 3가지 선택지가 주어지게 된다면, 사용자는 철학적 챗봇과의 대화를 이어 나가기 위해서 답변들 중 하나를 무조건 선택해야 한다. 그러면서 사용자는 자연스럽게 앞선 챗봇의 이야기를 꼼꼼하게 읽어보게 되고, 가정된 상황에서 '어떤 답변이 바람직할지' 고민해보게 된다. 이를 통해 사용자는 철학적 챗봇과의 이야기에 보다 집중하고, 해당 주제에 대해서 깊이 고민해볼 수 있게 된다

로올라 별관



자율 주행 자동차가 본격적으로 도입되었을 때, 가장 대표적으로 우려되는 상황은 윤리학에서 자주 논의되는 "트롤리 문제(trolley problem)"이야.

트롤리 문제(trolley problem)가 무엇인지 간단한 예시를 들어볼게.

만약 선로를 달리고 있는 어떤 기차가 있는데 그 기차의 브레이크가 고장이 난 거야. 기차를 당장 정지시킬 수 없는 기차의 운전자가 있고, 기차의 바로 옆에는 4명의 사람, 다른 옆 방향에는 1명의 사람이 있다고 가정해보자. 기차를 그대로 두면 사람 4명이 죽게 될 것이 분명하다면, 기차의 핸들을 돌려서 다른 한 명의 사람을 죽게 하는 것이 옳은 선택일까?

'자율 주행 자동차가 트롤리 딜레마를 마주한다면'이라는 주제에서도 사용자와 챗봇 사이의 인터랙션을 확인할 수 있다. 자율 주행 자동차가 야기할 수 있는 대표적인 윤리적 문제는 트롤리 딜레마로, '과연 인공지능이 인간 생명의 경중을 판단할 수 있는가'와 관련된 문제이다. 우선 챗봇은 트롤리 딜레마에 대해서 잘 알지 못하는 사용자를 위해 간단한 개념 설명을 한 뒤, 4명(다수)의 생명을 살리기 위해 한 명(소수)의 생명을 희생시킬 수 있는지에 대한 질문을 물어본다. 그 다음 사용자는 '당연히 핸들을 돌려야지!', '핸들을 어떻게 돌려ππ', '음 선택이 쉽지 않은걸?'이라는 세 가지 답변 중 하나를 선택해야 한다.

만약 선로를 달리고 있는 어떤 기차가 있는데 그 기차의 브레이크가 고장이 난 거야. 기차를 당장 정지시킬 수 없는 기차의 운전자가 있고, 기차의 바로 옆에는 4명의 사람, 다른 옆 방향에는 1명의 사람이 있다고 가정해보자. 기차를 그대로 두면 사람 4명이 죽게 될 것이 분명하다면, 기차의 핸들을 돌려서 다른 한 명의 사람을 죽게 하는 것이 옳은 선택일까?

너라면 어떻게 할래?

당연히 핸들을 돌려야지!

핸들을 어떻게 돌려 ㅠㅠ

음 선택이 쉽지 않은걸?

오후 10:58

오후 11:00

당연히 핸들을 돌려야지!

로콜라 별관

사람의 생명이 달린 문제라서 선택이 정말 쉽지 않았지? 우리 인간들도 이렇게 고민하는 문제인데, 만약 자율 주행 자동차가 이 트롤리 딜레마 상황을 갑자기 마주하게 된다면 어떻게 될까? 과연 인간보다 더 현명한 답을 내릴 수 있을까?

그리고 그 답이 현명하다고 판단할 수 있는 기준은 무엇일까? 4명을 살리기 위해서 한 명을 죽게 만든 자율 주행 자동차는 그 한 생명을 어떻게 간주하였을까? 혹은 한 명을 살리기 위해 4명을 죽게 만들었다면 그 자율 주행 자동차에 대해서 어떤 가치 판단을 내릴 수 있을까?

또 다른 트롤리 문제

오후 11:00

오후 11:00

음 선택이 쉽지 않은걸?

로콜라 별관

맞아, 사람의 생명이 달린 문제라 쉽게 줄 중에 하나를 고른다는 것은 정말 어려운 것 같아. 이렇게 우리 인간들도 이렇게 고민하는 문제인데, 만약 자율 주행 자동차가 이 트롤리 딜레마 상황을 갑자기 마주하게 된다면 어떻게 될까? 과연 인간보다 더 현명한 답을 내릴 수 있을까?

그리고 그 답이 현명하다고 판단할 수 있는 기준은 무엇일까? 4명을 살리기 위해서 한 명을 죽게 만든 자율 주행 자동차는 그 한 생명을 어떻게 간주하였을까? 혹은 한 명을 살리기 위해 4명을 죽게 만들었다면 그 자율 주행 자동차에 대해서 어떤 가치 판단을 내릴 수 있을까?

또 다른 트롤리 문제

오후 11:00

핸들을 돌릴지에 대한 질문에 명확한 답변(답변 1, 답변 2)을 내린 사용자들에게는 답변 자체에 대한 이야기보다는 이 딜레마적 상황에서 '과연 자동차가 현명하게 판단할 수 있을까', 추가적인 질문을 제시한다. (사람의 생명이 달린 문제라서 선택이 정말 쉽지 않았지? 우리 인간들도 이렇게 고민하는 문제인데, 만약 자율 주행 자동차가 이 트롤리 딜레마 상황을 갑자기 마주하게 된다면 어떻게 될까? 과연 인간보다 더 현명한 답을 내릴 수 있을까? (이하 생략)) 중립을 선택한 사용자에게도 동일한 내용으로, 자율 주행 자동차에게 트롤리 딜레마적 상황이 어떤 의미를 가지는지에 대해 여러 추가 질문들을 제시한다.

V-2 같은 주제, 서로 다른 의견 비교

가상현실 주제와 관련한 두 가지 시나리오가 있는데, 경험기계와 실재기계 시나리오이다. 사용자가 가상현실 서가를 선택하면, 가상현실에서는 두 가지 사고실험에 관한 간략한 소개와 시나리오로 넘어가는 버튼이 있다. 여기서 기본적인 가상현실 정의를 선택할 수도 있고, 바로 경험기계, 혹은 실재기계 시나리오로 넘어갈 수도 있다. 가상현실의 두 가지 시나리오는 다음과 같은 특색을 지닌다.

첫째, 재미있는 사고실험을 소개해 사용자의 대화 몰입도를 높인다.

차머스는 사고 실험에서 이렇게 가정해. 2095년, 핵 전쟁과 기후 변화로 지구의 표면이 완전히 박살이 난 상황이야. 너는 두 가지 선택을 할 수 있어. 하나는 약탈자와 지뢰를 피해가면서 어렵게 삶을 연명하는 것이고, 또 다른 하나는 네 신체를 안전한 방공호에 가두고 가상 현실에 접속하여 남은 삶을 사는 것이지.

이런 상황에서 너라면 어떤 선택을 할래?

당연히 들어가야지!

난 싫어!

오후 6:40

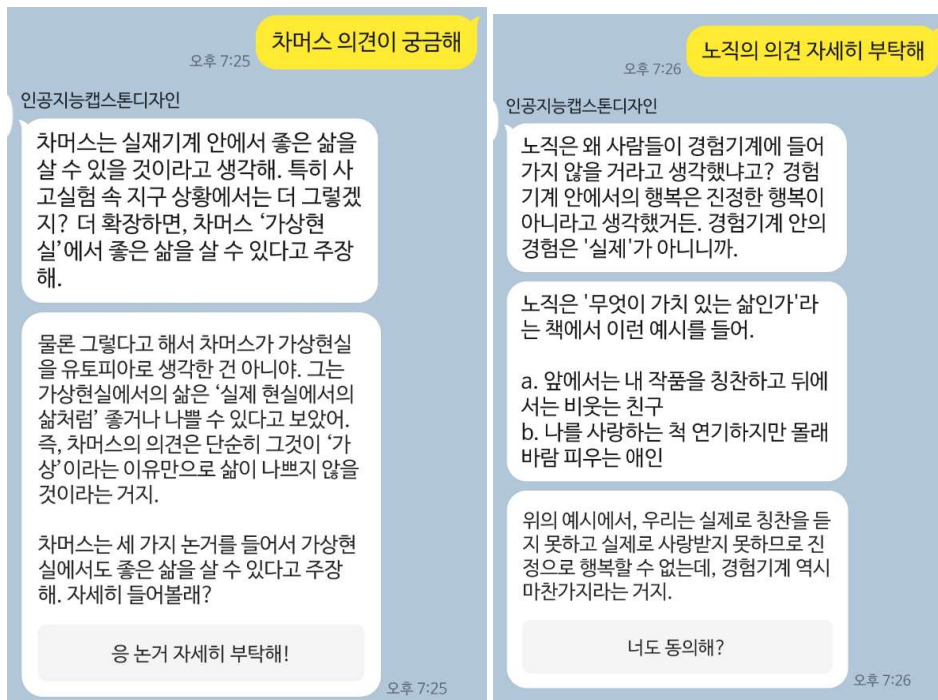
음... 잠깐 고민 좀 할게...

사실 차머스가 가정한 그런 상황, 즉 지구에서 지금처럼 살 수 없을 때는 다른 선택지가 없을 것 같긴 해. 그럼에도 내가 고민하는 건... 지구가 심각하게 파괴되어 일상적인 삶을 살 수 없는 상황에서 만약 실재기계라는 선택지가 없다면, 우리는 어떤 선택을 할 수 있을까?

그냥 포기하고 맞춰서 같이 무법자가 되거나, 그들을 피해서 숨어 살거나, 아니면 그냥 죽음을 선택할 수도 있겠지. 그렇지만... 누군가는 그럼에도 불구하고 이 지구를 바꿔보려고 노력하지 않을까...? 어쩔 수 없이 이 환경에서 살아야 한다면, 그 환경을 바꿔보려는 시도를 해보겠지. 그런데 가상현실이라는 현실의 도피처가 있다면, 아마 아무도 바꿔보려는 노력 없이 지구를 포기할 거야.

노직과 차머스의 철학을 이미 알고 있는 사용자라면 대화를 따라가기 수월하겠지만, 모든 사용자가 사전 지식을 갖고 있는 것은 아니다. 게다가 어렵고 진지한 철학이론으로 논의를 시작하면 사용자가 금방 흥미를 잃을 것으로 판단하였다. 그래서 초반에 재미 요소로서 노직과 차머스의 사고실험을 소개하면서 논의를 시작한다. 왼쪽의 사진은 실재기계를 소개하는 대화이다. 특히 디스토피아 영화의 배경처럼 흥미로운 사고실험의 가정(완전히 망가진 2095년의 지구 등)들이 사용자의 관심을 유도할 수 있을 것이다. 또 이 기계에 들어가고 싶은지를 사용자에게 물어보고 답하는 과정에서 사용자가 대화에 더욱 몰입할 수 있도록 하였다. 또 역으로 사용자가 개발자에게 가상기계로 들어갈 것인지 물어볼 수도 있게 만들었다. 오른쪽의 사진은 실재기계로 들어갈 것이냐는 사용자의 물음에 '들어가지 않겠다'고 답한 후 그 이유를 설명하는 대화이다.

둘째, 두 철학자의 상반된 의견 제시하여 대화를 풍부하게 하고 편향을 방지하였다.



두 철학자는 가상현실에서의 삶에 대해 서로 상반되는 의견을 갖는다. 노직은 경험기계에서의 행복은 진정한 행복이 아니므로 아무도 경험기계로 들어가지 않을 거라고 생각했다. 뿐만 아니라 가상현실에서 좋은 삶을 살 수 없을 것이라고 보았다. 실제로 노직은 생전에 VR, 즉 가상현실기계 자체를 부정적으로 보았다. 반면 차머스는 노직과는 반대로 가상현실에서도 충분히 좋은 삶을 살 수 있다고 주장하고, 이를 실재기계 사고실험에서 설명한다. 이렇게 관련 철학자의 두 가지 의견을 제시하여 사용자가 챗봇과의 대화를 통해 한 가지 의견으로 생각이 편향되지 않도록 하였다.

또 차머스가 자신의 저서에서 경험기계를 언급한 부분을 시나리오에 녹여 풍부한 논의가 이루어지는 동시에 마치 철학자 두 명의 토론을 보는 것과 같은 느낌을 주어 흥미를 유발하는 효과도 기대한다. 오른쪽의 사진이 그 대화의 일부이다. 차머스가 지적한 부분은 노직의 사고실험에서 등장하는 경험기계는 일반의 가상현실과는 차이가 있기 때문에, 경험기계는 가상현실에서 좋은 삶을 살 수 없다는 근거로 제시하는 것은 부적절하다는 것이다. 이렇게 사용자에게 노직의 경험기계는 일반의 가상현실과 다른 점을 설명해줌으로써 더 풍부한 논의가 이루어지게 했다.

둘째, 경험기계는 모든 경험을 사전 프로그래밍하고 나서 들어가기 때문에 자율성이나 주체성, 자유의지가 제한되지만, 실재기계는 그렇지 않아. 물리적 현실에서 두뇌로 내리는 결정 프로세스와 가상현실에서 두뇌로 내리는 결정 프로세스가 차이가 없다면, 가상현실에서 우리의 자유의지와 자율성이 물리적 현실보다 부족하다고 할 수 없다는 게 차머스의 논리야.

셋째, 경험기계는 나 혼자만이 실제 사람이고 나머지는 프로그래밍된 인물이지만, 실재기계에서는 친구와 가족끼리 가상현실을 공유할 수 있어. 즉 동시 접속이 가능하고 상호작용이 가능해.

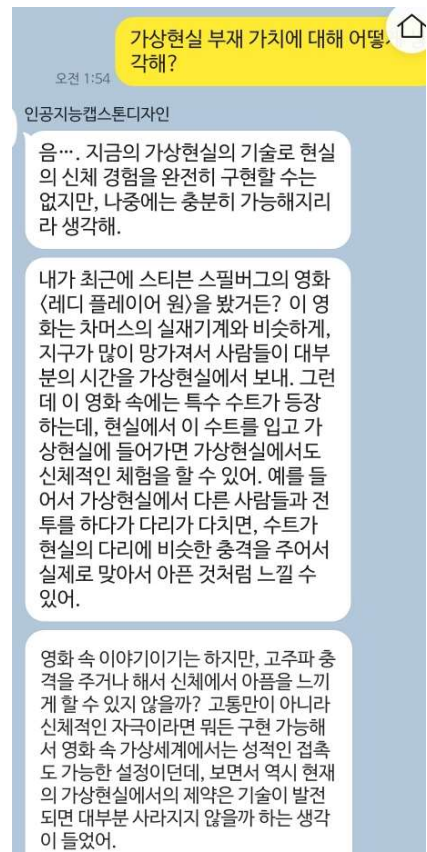
이런 이유로 차머스는 경험기계는 일반적인 가상현실과는 다르기 때문에, 경험기계 논변을 가상현실에서의 삶을 반대하는 논리로서 사용하기에는 적절하지 않다고 보았어.

셋째, 사고실험을 통해 철학적 질문을 이끌어내 사용자의 철학적 사고를 유도한다.



사고실험을 소개해 사용자를 대화에 집중시키고, 이후 본격적으로 가상현실과 관련한 철학적 질문을 제시한다. 노직의 경험기계와 차머스의 실재기계는 '가상현실이 좋다, 나쁘다' 정도의 논의를 넘어, '가상현실에서 좋은 삶을 살 수 있는가?', '좋은 삶이란 무엇인가?', '우리는 삶의 어떤 가치를 추구하고, 추구해야 하는가?'와 같이 보다 본질적인 철학적 질문을 끌어낸다. 챗봇의 페르소나 '로올라 정령'과의 대화로 사용자에게 왼쪽 사진과 같은 철학적 질문들이 유도된다. 그리고 이는 오른쪽 사진처럼 사용자의 의견을 물어보는 대화로 이어진다. 사용자의 의견을 받을 수 있는 구글 폼을 만들어서 사용자 역시 익명으로 자유롭게 의견을 남길 수 있다. 자신의 의견을 제시하고 싶으면 사용자는 '응, 얘기해볼래!' 버튼을 누르고 '정령'이 알려주는 구글 폼 링크를 통해서 의견을 남길 수 있다. 다른 사용자들의 의견을 사용자가 열람할 수 있도록 하여 개발자-사용자 간의 토론을 넘어 사용자-사용자 간의 토론도 가능하도록 했다. (구글 폼에 대한 세부적인 설명은 보고서의 V-2에서 확인할 수 있다)

그런데 사용자 입장에서는 질문에 답하는 것에 어려움을 느끼거나 부담을 느낄 수 있다. 따라서 아직 자신의 생각을 얘기하기 힘든 사용자들은 '정령 네 의견이 궁금해!' 버튼을 눌러 개발자의 의견을 확인하고 넘어갈 수도 있다. 예를 들어 차머스는 저서에서 가상현실에서는 신체적 경험이 구현되기 힘들다고 설명하는데, 시각과 청각은 구현 가능하지만 미각과 후각, 특히 촉각과 관련된 경험을 할 수 없기 때문이다. 이 '가상현실에서 신체적 경험의 부재'에 대해 개발자의 의견을 들려주



는 블록을 만들었다. 아래의 사진이 그 대화의 일부이다. 요지는 기술이 발전하면 신체적인 경험도 가능하다는 것이다. 이를 얘기하면서 스티븐 스피버그 영화 <레디 플레이어 원>을 언급했다. 영화에서 가상현실에서도 신체적 경험을 느낄 수 있는 수트가 나오기 때문이다. 인공지능철학에서는 특히 철학적 상상력이 중요하다고 생각한다. 가상현실의 경우만 봐도 현실의 기술 발전은 아직 두 철학자의 사고실험에 등장하는 기계에 미치지 못한다. 그럼에도 우리는 상상력을 동원해 미래의 기술 발전이 사회, 그리고 개인의 삶에 어떤 영향을 줄 수 있을지 논의할 필요가 있다. 이렇게 SF영화를 소개하여 사용자가 가상현실과 관련한 철학적 상상력의 물꼬를 틀 수 있도록 하였다.

V-3 퀴즈를 활용한 시나리오

우리 챗봇은 철학을 전공하지 않기에 관련 지식이 없거나 부족한 서강대학교 학부생을 기본적인 사용자 타겟층으로 설정하였다. 챗봇을 이끄는 페르소나의 이름이 '로올라 정령'인 까닭도 재학생에게 익숙한 소재를 사용하여 친근감을 느끼게 하는 것에 있다. 챗봇 사용자 대부분은 철학 비전공자일 것이 예상되며, 챗봇의 상용화 시 수익성, 화제성 등을 고려한다면 챗봇 사용자에서 철학 전공자가 차지하는 비율이 적을 것을 상정하고 챗봇을 개발하는 것이 더욱 바람직하다고 볼 수 있다.

그런데 시나리오의 논리 전개 과정이 지나치게 복잡하거나, 최근 들어 학계에서 논의되기 시작하여 아직 완전히 해명되지 않은 소재가 사용된다면, 페르소나가 친근한 말투일지라도 이용자는 심리적 거리감을 느낄 수 있다. 논의를 따라가기 버겁기에 시나리오를 전부 이해하는 과정에서 따분하고 지루한 감정을 느끼거나, 분명 시나리오를 전부 읽었음에도 무슨 말을 하고 있는 건지 헷갈리는 혼란함을 느낄 수 있다.

1. 논리 전개의 복잡성. 2. 소재의 내재적 모호함.

위 두 가지를 챗봇과 사용자의 원활한 상호작용을 방해하는 대표적 요소로 꼽을 수 있다.

위 문제를 인식하여 우리는 챗봇에 작성된 시나리오 전반을 검토하였다. '인간 향상-하버마스-자율성 논변'을 소재로 작성한 시나리오에는 '김건우.(2014).생명공학적 인간향상의 시대에 인간은 얼마나 자유로울 수 있는가- 하버마스의 자율성 논변의 분석 -.법학논집,18(3),447-472.' 전반을 참고한 것으로 인용한 자료가 학술 논문인만큼 내용의 깊이와 논리 전개의 복잡성이 매우 높은 부분으로 평가되었다. 또한 '철학적 좀비' 개념을 서술한 시나리오에는 감각질, 정신과 육체의 관계, 신경과학 등에 대한 기본적인 이해를 요구한다. 인지 과학자 데이비드 차머스는 '의식을 어떻게 소개할까요?'라는 주제의 TED 강연에서 의식, 특히 감각질과 관련된 질문에 설명하는 것이 오늘날 학문의 발달 수준에서 얼마나 어려운 일인지를 토로하기도 하였다. 해당 소재의 내재적 모호함이 사용자에게 혼란을 줄 것이 우려되었다.

따라서 '인공지능과 윤리실습'이라는 과목에 걸맞은 주제의 시나리오를 챗봇에 온전히 담되, 사

용자에게도 잘 융화되며, 우리 챗봇에 오랜 시간 머물며 이탈을 방지하기 위한 전달 방식을 고안하였다. '퀴즈'를 제작하는 것이다. 퀴즈를 챗봇에 도입하였을 때, '첫째, 정답을 맞췄을 때 뿌듯한 감정이 생기며 지적 탐구를 계속하고 싶은 욕구가 생긴다. 둘째, 정답을 맞추지 못하였을 때 연결 버튼을 통해 다시 내용을 점검할 기회를 제공한다.'는 두 가지 효과를 기대하였다. 그렇다면 퀴즈의 난이도는 목적에 따라 달리 설정되어야 할 것이다. 첫째 효과가 실현되기 위해서는 어렵지 않게 정답을 맞출 수 있을 만큼의 난이도가 요구된다. 둘째 효과는 이해하기 쉽지 않은 소재를 퀴즈로 만들었을 때를 가정한다. 그만큼 퀴즈 해결 난이도가 높을 것이다.

'인간 향상-하버마스-자율성 논변' 시나리오에는 하버마스가 자율성 논변의 근거로 제시한 두 가지 논거, '출생의 우연성 논거'와 '보편적 합의 가능성 논거'가 제시된다. 두 개의 논거 모두 4개의 전제와 2개의 결론으로 구성되어 있다. 최종 결론까지의 논리가 복잡하므로 가독성을 위해 3개의 응답으로 나뉘어 제시된다.

인공지능캡스톤디자인

<하버마스의 출생의 우연성 논거>

전제1. 인간은 진정으로 자율적인 (그리고 평등한) 개인이어야 한다.

전제2. 인간이 진정으로 자율적인 (그리고 평등한) 개인이기 위해서는 그 출생이 우연적이어야만 한다.

전제3. 생명공학적인 개입은 출생의 우연성을 제거한다.

-> 결론1. 생명공학적인 개입은 후대의 인간의 자율성을 훼손한다.

전제4. 인간에 대한 개입은 그것이 그 인간으로 하여금 진정으로 자유로운 (그리고 평등한) 개인으로서 살아갈 수 있게 할 때에만 도덕적으로 허용될 수 있다.

-> 결론2. (따라서) 생명공학적인 개입은 도덕적으로 허용될 수 없다.

결론1은 전제1, 2, 3을 모두 함축해.

또한 결론2는 결론1과 전제4를 함축해.

따라서 전제1, 2, 3, 4 모두 참이라면 결론2 또한 참이란다.

인공지능캡스톤디자인

<하버마스의 보편적 합의 가능성 논거>

전제1. 인간에 대한 개입은 그것이 그 인간으로 하여금 진정으로 자유로운 (그리고 평등한) 개인으로서 살아갈 수 있게 할 때에만 도덕적으로 허용될 수 있다.

전제2. 담화(談話)에 참여하는 인간이 진정으로 자율적인 경우, 그리고 오직 그 경우에만 그 참가자들 간에 보편적 합의가 가능하다.

-> 결론1. 생명공학적인 개입은 보편적으로 합의될 만한 것이어야 한다.

전제3. 치료를 목적으로 한 생명공학적인 개입은 후대에 의해 합의될 수 있다.

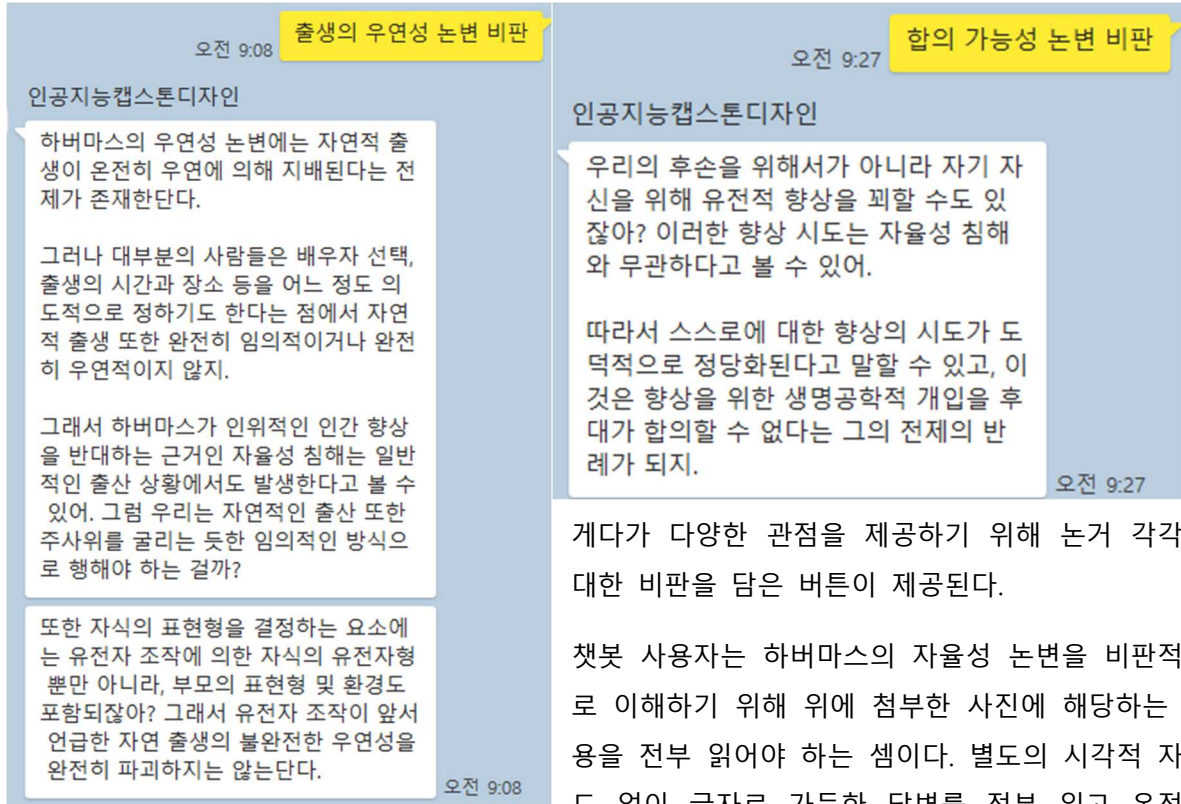
전제4. 향상을 목적으로 한 생명공학적인 개입은 후대에 의해 합의될 수 없다.

-> 결론2. 생명공학적인 개입은 그것이 치료를 목적으로 하는 경우에 한해 도덕적으로 허용될 수 있으며, 향상을 목적으로 하는 경우에는 도덕적으로 허용될 수 없다.

결론1은 전제1, 2를 모두 함축해.

또한 결론2는 결론1과 전제3, 4를 함축해.

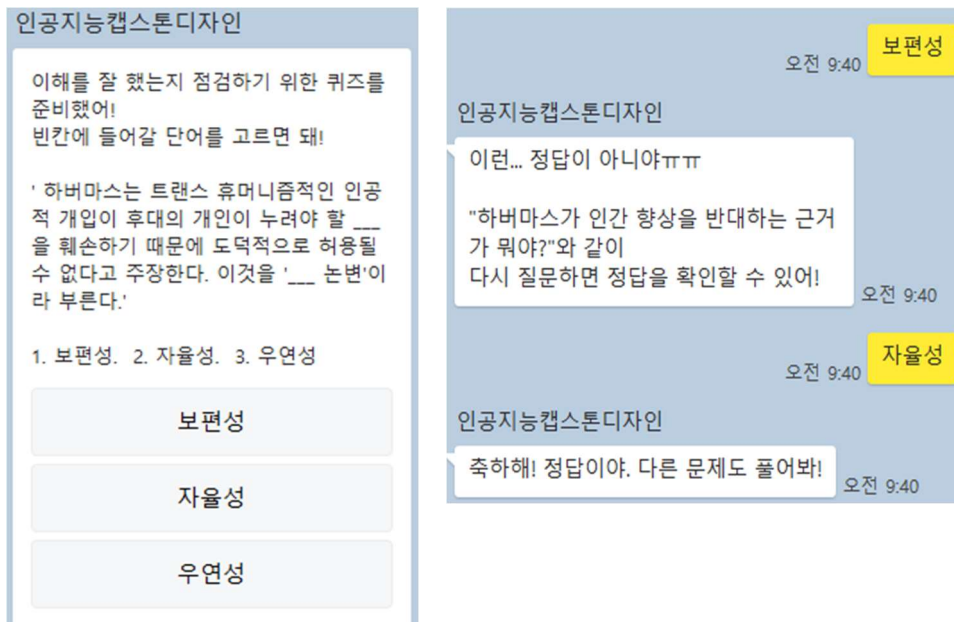
따라서 전제1, 2, 3, 4가 모두 참이라면 결론2 또한 참이란다.



게다가 다양한 관점을 제공하기 위해 논거 각각에 대한 비판을 담은 버튼이 제공된다.

챗봇 사용자는 하버마스의 자율성 논변을 비판적으로 이해하기 위해 위에 첨부한 사진에 해당하는 내용을 전부 읽어야 하는 셈이다. 별도의 시각적 자료도 없이 글자로 가득한 답변을 전부 읽고 온전히

이해하는 것은 쉽지 않아 보인다. 따라서 낮은 난이도와 높은 난이도인 퀴즈 각 1개씩을 바로 연결 버튼으로 제공하였다. 두 개의 퀴즈를 의도적으로 '쉬움->어려움'의 순서로 배열하여 중도 이탈을 방지하였다.



- 낮은 난이도의 퀴즈를 우선 제시하였다.

이해를 잘 했는지 점검하기 위한 퀴즈를 준비했어!
정답으로 옳은 번호를 선택하면 돼!

'하버마스의 출생의 우연성' 논거는 세 개의 전제를 통해 소결론이 도출되고 나머지 하나의 전제를 통해 최종 결론이 도출되는 구조를 보입니다. 이때 출생 우연성 논거를 비판하는 주장이 지적하고 있는 전제와 가장 거리가 먼 것은?

1. 인간이 진정으로 자율적인 개인이기 위해서는 그 출생이 우연적이어야만 한다.
2. 생명공학적인 개입은 출생의 우연성을 제거한다.
3. 인간에 대한 개입은 그 인간으로 하여금 진정으로 자유로운 개인으로서 살아갈 수 있게 할 때에만 도덕적으로 허용될 수 있다.

1

2

3

오전 9:50

2

인공지능캡스톤디자인

이런... 정답이 아니야ㅠㅠ

자식의 표현형은 자식의 유전자형 뿐만 아니라, 부모의 표현형 및 환경도 포함되기 때문에 유전자 조작이 자연 출생의 불완전한 우연성을 완전히 파괴하지는 않는다고 볼 수 있어. 그러므로 '생명공학적인 개입은 출생의 우연성을 제거한다.'는 전제는 부분적으로 한계를 보인다고 볼 수 있지.

더 자세히 알고 싶다면 '하버마스의 출생의 우연성' 논변에 대한 비판에 대해 알려줘! 라고 말해봐!

오전 9:50

오전 9:50

3

인공지능캡스톤디자인

정답이야!

출생의 우연성 논거를 비판하는 주장을 소개하며 '인간에 대한 개입은 그 인간으로 하여금 진정으로 자유로운 개인으로서 살아갈 수 있게 할 때에만 도덕적으로 허용될 수 있다.'는 전제를 직접 반박하지는 않았어.

혹시...너 이 전제의 성립도 받아들이지 못하겠니? 맞아 이 전제 또한 한계가 있을 수 있어. 한 번 스스로 비판적으로 생각해봐!

오전 9:50

응답자가 '보편성'과 같은 오답을 선택하였을 때는 관련된 블록을 호출할 수 있는 패턴 발화를 입력하면 재학습이 가능하며 격려하였다. 두번째 문제는 '틀린 것을 고르시오' 형식이므로 정답을 제외한 나머지 선지는 올바른 내용으로 이루어진 셈이다. 따라서 오답을 선택하였을 땐, 이 선지가 왜 올바른 것인지 이유를 구체적으로 해설해주었다.

다음으로 '철학적 쯔뵤'와 관련된 블록들은 '서

가 이동->철학적 쯔뵤-> 감각질'의 순서로 제시된다. 철학적 쯔뵤를 잘 이해하기 위해서는 감각질에 대한 이해가 수반되어야 한다. 따라서 퀴즈는 감각질에 관해 잘 이해했는지를 확인하기 위해 제작하였다. 서가 이동 때는 흥미 유발의 목적으로 사용한 철학적 쯔뵤 관련 사진을 퀴즈에 다시 사용하는 수미상관 구성을 보인다.

인공지능캡스톤디자인

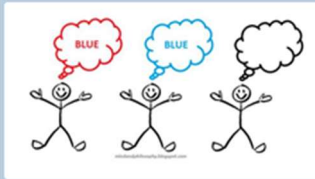


사진 출처: <mindandphilosophy.blogspot.com>



Mind and philosophy
Regularly updated blog about ...
mindandphilosophy.blogspot.com

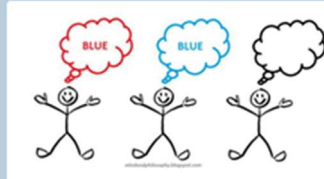
안녕! 이곳은 '철학적 좀비'에 대해 알아볼 수 있는 서가야.

위에 있는 사진 속 같은 모습을 한 세 사람이 보이니? 생김새는 모두 같지만 각자 다른 생각을 하는 것 같지 않니? 철학적 좀비가 무엇인지 알아본다면 이 그림을 더 쉽게 이해할 수 있을 거야. 같이 알아보자.

철학적 좀비의 정의

오전 9

인공지능캡스톤디자인



감각질에 관한 개념을 잘 이해했는지 점검하기 위해 퀴즈를 준비했어! 올바른 선지를 선택하면 돼!

'위 사진은 철학적 좀비 서가에 등장했을 때 제시했던 사진과 같은 것입니다. 이 그림에서 가운데 사람을 파란 하늘을 보고 있는 '나'라고 볼 때, 왼쪽과 오른쪽에 있는 사람을 각각 어떤 상태에 놓인 것이라 볼 수 있을까요? 순서에 맞게 나열한 것을 고르시오. 단, 철학적 좀비와 역전 감각질 상태가 모두 상상 가능하며 논리적으로도 가능하다고 가정합니다.'

1. 철학적 좀비, 역전 감각질
2. 역전 감각질, 철학적 좀비
3. 이원론자, 일원론자

1

2

3

이 문제의 3번 선지는 '이원론자, 일원론자'인데 앞서 제시된 블록에서 등장하지 않는 개념으로 퀴즈를 통해 새로운 개념을 소개하기 위해 의도적으로 제공한 터무니없는 오답이다. 3번 선지를 선택하면 '일원론과 이원론' 개념을 소개하는 블록으로 연결하는 버튼이 제공된다. 물론 해당 블록은 3번 선지를 선택하지 않고 '감각질-철학적 좀비' 담론의 의의-일원론과 이원론' 순서에 따라 확인할 수 있으며, 별도의 패턴 발화가 입력되었을 때도 제공된다.

인공지능캡스톤디자인

틀렸어...

그러나 이 선지를 선택하다니 의외인 걸?
이 문제는 기본적으로 일원론적인 입장에서 만들어진 문제였어. 혹시 일원론과 이원론의 특징과 철학적 좀비의 관련성을 알고 있니? 궁금하다면 아래 버튼을 통해 확인해 봐.

일원론과 이원론

오전 10:05

오전 10:05

3

인공지능캡스톤디자인

의식적 경험이 부재하나, 외형은 인간과 다를 바가 없는 존재인 '철학적 좀비'의 존재 가능성에 대해 이원론과 일원론의 견해가 서로 다르다.

데카르트로 대표되는 이원론과 흄스로 대표되는 일원론이 있어. 내가 간략하게 두 입장을 정리해줄게.

데카르트의 이원론: 정신과 물질은 서로 다른 것이며, 따로 존재할 수 있다. (실체 이원론)
흄스의 일원론: 인간의 정신 활동이라고 불리는 것들 또한 전부 물질적 운동이다. (기계론적 일원론)

그렇다면 이원론자는 철학적 좀비의 존재 가능성을 긍정할까? 부정할까?
마찬가지로 일원론자는 철학적 좀비에 대해 어떻게 생각할까?

간단히 생각한 뒤 아래 버튼을 통해 확인해보자!

이원론자

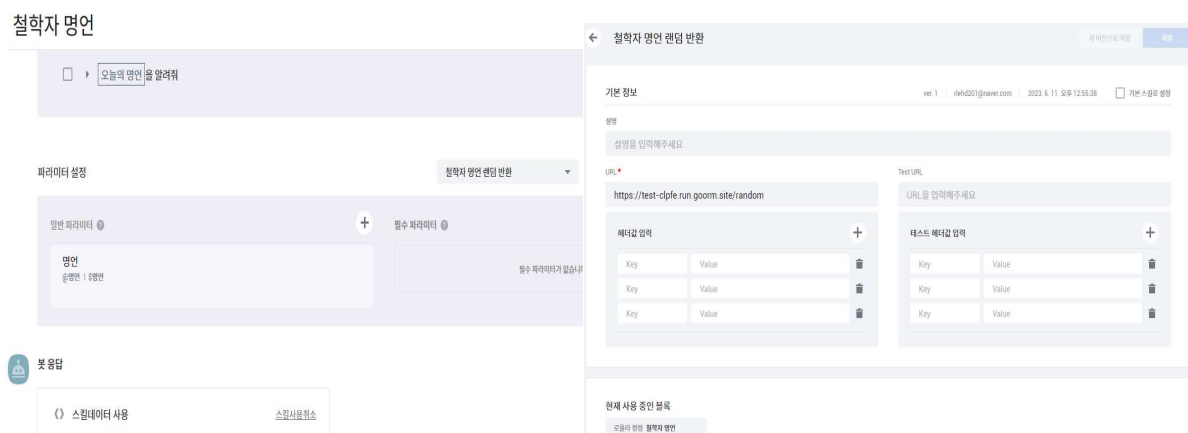
일원론자

오전

VI. 부가적인 기능

비록 이번 프로젝트를 통해 개발하고자 하는 것이 철학적 챗봇이며, 철학적 논의를 전개하는 것이 주된 목적이라 할지라도 사용자와 대화를 나누는 챗봇이라는 특성상 가볍게 사용자들이 요구할 수 있을 기능을 추가하는 것이 필요하다고 생각했다. 이에 따라 철학자의 명언을 랜덤하게 이야기해줄 수 있는 기능을 추가하였으며, 개발자-사용자 간의 상호작용을 넘어 사용자-사용자 간의 토론도 가능하게 하고자 설문지를 활용해 사용자 의견을 추가할 수 있는 기능을 추가하였다.

VI-1. 철학자 명언



기존 시나리오에서 "철학자 명언을 추천해줘", "철학자 글귀를 들려줘", "명언 알려줘" 등과 같은 입력이 주어질 때 철학자 명언을 반환하는 시나리오가 인식되기 원활하도록 '명언'을 엔티티로 설정하였다. 그리고 이러한 입력이 주어질 때 파라미터 설정 부분의 우측에 있는 곳에 기존에 설계한 스킬 데이터를 할당하고 봇 응답을 '스킬 데이터 사용'으로 설정하면 설정한 스킬 데이터를 이용한 답변이 가능해진다.

스킬 데이터에서는 기존에 프로그래밍한 코드를 바탕으로 작동하는 서버의 URL을 연결시키는 방식으로 설계하였다. 이를 통해 챗봇에 사용자가 관련된 입력을 하게 된다면 등록된 URL을 타고 이동하여 구축되어 있는 서버에서 해당 입력을 인식하고 기존에 설정한 12개의 명언 중 랜덤하게 1개의 명언을 챗봇이 반환할 수 있게 된다. 12개의 명언에는 다음과 같은 항목들이 포함되어 있다.

1. 칸트: 네 의지의 준칙이 언제나 보편적 입법의 원리가 될 수 있도록 행위하라
2. 스피노자: 내일 지구가 멸망하더라도 오늘 한 그루 사과나무를 심겠다
3. 소크라테스: 너 자신을 알라
4. 니체: 이 인생을 기뻐하라. 즐겁게 살아가라

5. 사르트르: 실존은 본질에 앞선다
6. 하이데거: 언어는 존재의 집이다
7. 마르크스: 하나의 유령이 유럽을 떠돌고 있다, 공산주의라는 유령이
8. 헤겔: 미네르바의 부엉이는 황혼이 질 때만 날개를 편다
9. 공자: 군자는 의(義)를 밝히고 소인은 이(利)를 밝힌다
10. 노자: 도를 도라고 하면 이미 도가 아니다
11. 비트겐슈타인: 말할 수 있는 것에 대해 분명히 말하고 말할 수 없는 것에 대해 침묵하라
12. 아리스토텔레스: 아는 것에 의해서가 아니라 아는 것을 실천할 때 비로소 지혜로운 사람이 될 수 있다

```
1 from flask import Flask, jsonify
2 import sys
3 import random
4 application = Flask(__name__)
5
6
7 @application.route("/random", methods=["POST"])
8 def random_function():
9     rand = random.randint(1, 12)
10     if rand == 1:
11         message = "네 의지의 준칙이 언제나 동시에 보편적 입법의 원리가 될 수 있도록 행위하라 - 칸트"
12     elif rand == 2:
13         message = "내일 지구가 멸망하더라도 오늘 한 그루 사과나무를 심겠다 - 스피노자"
14     elif rand == 3:
15         message = "너 자신을 알라 - 소크라테스"
16     elif rand == 4:
17         message = "이 인생을 기뻐하라. 즐겁게 살아가라 - 니체"
18     elif rand == 5:
19         message = "실존은 본질에 앞선다 - 사르트르"
20     elif rand == 6:
21         message = "언어는 존재의 집이다 - 하이데거"
22     elif rand == 7:
23         message = "하나의 유령이 유럽을 떠돌고 있다, 공산주의라는 유령이 - 마르크스"
24     elif rand == 8:
25         message = "미네르바의 부엉이는 황혼이 질 때만 날개를 편다 - 헤겔"
26     elif rand == 9:
27         message = "군자는 의(義)를 밝히고 소인은 이(利)를 밝힌다 - 공자"
28     elif rand == 10:
29         message = "도를 도라고 하면 이미 도가 아니다 - 노자"
30     elif rand == 11:
31         message = "말할 수 있는 것에 대해 분명히 말하고 말할 수 없는 것에 대해 침묵하라 - 비트겐슈타인"
32     elif rand == 12:
33         message = "아는 것에 의해서가 아니라 아는 것을 실천할 때 비로소 지혜로운 사람이 될 수 있다 - 아리스토텔레스"
34     else:
35         message = "오늘은 추천해드릴 명언이 떠오르지 않네요"
```

서버를 구현하는 코드는 Flask를 이용하여 작성되었다. Flask는 Python 웹 프레임워크로, 웹 애플리케이션과 API를 개발하는 데 사용됩니다. 간결하고 유연한 설계를 가지고 있는 것이 특징이며, URL 라우팅을 지원하여 요청된 URL에 대해 특정한 함수를 실행할 수 있도록 한다. 이를 통해 요청과 응답을 처리하고 필요한 데이터를 보여줄 수 있다. URL 라우팅이 되는 것이 장점이기 때문에 이번 카카오톡 챗봇의 스킬 데이터를 사용하는 부분에서 URL을 통한 라우팅이 기본적인 작동 구조이기 때문에 이에 부합하는 Flask를 사용하였다.

```
디버그  터미널  🔁  검색  리소스 모니터  린트
root@goorm:/workspace/test# python application.py 80
* Serving Flask app "application" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Running on http://0.0.0.0:80/ (Press CTRL+C to quit)
* Restarting with stat
* Debugger is active!
* Debugger PIN: 573-959-735 172.17.0.1 - - [11/Jun/2023 03:58:52] "POST /random HTTP/1.1" 200 -
172.17.0.1 - - [11/Jun/2023 03:59:00] "POST /random HTTP/1.1" 200 -
172.17.0.1 - - [11/Jun/2023 03:59:04] "POST /random HTTP/1.1" 200 -
172.17.0.1 - - [11/Jun/2023 03:59:17] "POST /random HTTP/1.1" 200 -
```



사진과 같이 서버를 작동시켜서 입력을 받는 중에는 챗봇에서 "철학자 명언"과 관련된 키워드가 입력되었을 때 기존에 미리 설정한 12개의 명언 중 랜덤하게 명언을 대답해주고 있는 것을 확인할 수 있다.

하지만 이러한 구조의 문제점은 서버를 구축하고 서버가 작동되고 있는 동안만 챗봇에 주어진 입력을 서버가 받아들이고 적절한 결과를 다시 챗봇에 보내줄 수 있다는 것이다. 이는 즉 서버가 작동하지 않는다면 명언에 관련된 부분은 제대로 작동하지 않는다는 것을 의미한다. 이번 프로젝트에서 서버를 구축하는데 사용한 사이트는 GoormIDE로 추가적인 결제를 통

해 서버를 항상 켜두는 기능을 사용할 수 있어 실제로 챗봇을 배포하고 이 기능을 유지하는 것에 는 문제가 없을 것이다. 하지만 실제로 챗봇을 상용화하지 않은 상황에서는 수동으로 서버를 작동시키고 있는 상황에서만 해당 기능이 작동한다.

VI-2. 설문지를 활용한 사용자 의견 추가

1조의 챗봇 '로올라 별관'의 궁극적인 목적은 철학적 공간 마련에 있다. 즉 철학적인 토론이 가능한 온라인 공간을 만들고자 하는 것이다. 현재 단계에서는 개발자-사용자 간의 상호작용이 이루어지고 있는데, 나아가 사용자-사용자 간의 토론도 가능하게 하고자 한다.

카카오 챗봇의 기능으로는 사용자-사용자 간의 직접적인 대화는 불가능하다. 따라서 간접적으로 서로의 의견을 확인할 수 있도록 구글 폼으로 '로올라 별관에 사용자 의견 추가하기' 설문지를 만들어서 배포할 예정이다.

구글 폼 '로올라 별관에 사용자 의견 추가하기' 양식은 다음과 같다.

1. 설문지 설명

로올라 별관에 사용자 의견 추가하기

안녕하세요. 챗봇 '로올라 별관' 개발자입니다. '로올라 별관'에서 정령과의 대화가 더 풍성해질 수 있도록 도와주시겠어요? 구글 폼에 원하는 주제에 대한 의견을 제시하면, 이를 챗봇에 추가하여 다른 사용자가 귀하의 의견을 열람할 수 있도록 업로드하고자 합니다.

도움을 주셔서 감사합니다.

- * 작성한 의견은 추후에 삭제는 가능하지만 수정할 수는 없으니 신중하게 작성해주세요.
- * 삭제 문의: rachel01_study@naver.com
- * 사실과 다르거나 근거가 없는 내용은 업로드되지 않습니다.
- * 타인에 대한 비방, 차별적 표현, 혐오 표현이 포함된 경우 업로드되지 않습니다.

설문지 설명에는 설문지의 의도와 목적, 그리고 참고 사항으로 구성하였다. 특히 개인정보보호법에 따라서 개인정보는 개인이 요청할 때 삭제가 가능해야 하므로 삭제가 가능하다는 것을 명시, 그 방법을 설명하였다. 또 의견이 사실과 다르거나 근거가 없는 경우, 그리고 비방, 차별, 혐오 표현이 포함된 경우 업로드되지 않는다는 점을 명시하였다.

2. 설문지 질문

1. 의견을 제시하고 싶은 주제를 선택해주세요.

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> 인공지능의 사고가능성 | <input type="radio"/> 철학적 좀비 |
| <input type="radio"/> 인간 향상 (센델) | <input type="radio"/> 자율주행 자동차 |
| <input type="radio"/> 인간 향상 (하버마스) | <input type="radio"/> 정치와 인공지능 |
| <input type="radio"/> 가상현실 (노직) | <input type="radio"/> 인공지능과 신 |
| <input type="radio"/> 가상현실 (차머스) | <input type="radio"/> 빅데이터와 인공지능 |

2. 선택한 주제에 대해 자유롭게 의견을 말씀해주세요.

- * 타인의 의견을 인용할 때는 출처를 남겨주세요.
- * 사실과 다르거나 근거가 없는 내용은 업로드되지 않습니다.
- * 타인에 대한 비방, 차별적 표현, 혐오 표현이 포함된 경우 업로드되지 않습니다.

3. 귀하의 이메일을 입력해주세요. 추후 삭제 요청을 할 경우 설문을 작성한 본인이 맞는지 확인하기 위한 목적으로 수집합니다.

- * 반드시 답변한 메일 주소로 삭제를 요청해주세요.

4. 귀하의 닉네임을 정해주세요. 챗봇에는 귀하의 닉네임이 표시됩니다.

* 닉네임을 정하지 않으면 '익명의 의견'으로 소개됩니다.

5. 개인정보 수집 및 이용에 동의하십니까? 동의한 사람에 한하여 챗봇에 의견이 업로드됩니다.

1. 수집 목적: 챗봇 '로올라 별관' 내 업로드
2. 수집 항목: 이메일 주소, 닉네임, 개인 의견
3. 보유 및 이용 기간: 입력일로부터 3년까지

☐ 동의

☐ 비동의

2번 질문에서는 앞의 설문지 설명에서 포함한 참고 사항에 더하여, 타인의 의견을 인용할 때는 그 출처를 밝혀야 한다는 점을 명시하여 저작권 문제에 대한 이슈를 최소화하고자 했다.

5번 질문에서는 설문지에서 수집한 이메일 주소, 닉네임, 개인 의견을 개발자가 이용하는 것에 대한 동의를 받고자 한다. 개인정보보호법, 그리고 개인정보보호위원회의 개인정보 처리 동의 안내서를 참고하여 작성하였다. 보유 및 이용 기간은 3년으로 설정하였다.

설문지 링크: <https://forms.gle/trV8Ck5pQ3fvBtAw5>

VII. 참고자료

개인정보 보호법(시행 2020. 8. 5. 법률 제16930호, 2020. 2. 4. 일부개정) 제3조

페르소나 캐릭터 이미지: https://www.sogang.ac.kr/intro/symbol/s_emot.html

박경남, 2023년도 1학기 <인공지능과 윤리> 수업 자료