최 동용 | 인하대학교 2024 년 2 학기
인공지능의 윤리 (Ethics of AI) 월 18:30-20:50

인공지능의 윤리

개요: 이 강의는 인공지능 기술의 사회적 영향이나 정책 문제를 소개하는 수업이라기보다, 인공지능을 계기로 드러나는 철학적 문제들을 체계적으로 검토하는 윤리학 수업입니다. 우리는 먼저 인공지능이 단순한 도구인지, 아니면 일정한 의미에서 '행위자'로 간주될 수 있는지라는 질문에서 출발합니다. 이를 위해 자율성, 선택, 의도, 책임 같은 전통적인 도덕철학의 개념들을 정리하고, 인간에게 적용되던 개념들이 인공지능에도 적용될 수 있는지 살펴봅니다. 이어서 인공지능이 도덕적 평가의 주체가 될 수 있는지(도덕적 행위자성), 혹은 오히려 보호나 고려의 대상이 될 수 있는지(도덕적 지위)를 둘러싼 논쟁을 검토합니다. 이 과정에서 자아, 개인 동일성, 현상적 의식, 감정과 공감 같은 개념들이 왜 도덕철학에서 중요한 역할을 하는지 확인하게 됩니다.

수업의 후반부에서는 논의를 인간 사회의 문제로 확장합니다. 인공지능이 의사결정에 개입할 때 책임은 누구에게 귀속되는지, 인공지능이 인간의 판단을 보조하거나 교정하는 것이 자율성을 강화하는지 약화하는지, 예측 기술과 데이터 활용이 사회적 협력과 공정성에 어떤 영향을 미치는지 등을 다룹니다. 또한 개발자 윤리, 기술 설계의 책임, 인간과 기술의 관계 같은 문제를 통해 인공지능 윤리가 단순히 "기계의 윤리"가 아니라 인간의 자기이해와도 깊이 연결된다는 점을 확인합니다. 전체적으로 이 강의는 인공지능을 하나의 사례로 삼아, 현대 사회에서 도덕적 판단이 무엇에 근거해야 하는지를 철학적으로 탐구하는 것을 목표로 합니다.

목표: 이 수업의 목표는 학생들이 인공지능 윤리 문제에 대해 의견을 갖는 것을 넘어, 철학적으로 논증할 수 있는 능력을 기르는 데 있습니다. 먼저 학생들은 자율성, 도덕적 행위자, 도덕적 지위, 도덕적 책임과 같은 기본 개념들을 서로 구분하고 정확하게 정의할 수 있어야 합니다. 특히 행위자성과 지위, 책임과 비난가능성, 의식과 지능 같은 개념들이 왜 쉽게 혼동되는지 이해하고, 각각이 서로 다른 질문에 대한 답이라는 점을 설명할 수 있어야 합니다.

다음으로 학생들은 인공지능을 둘러싼 대표적인 철학적 입장들을 단순히 요약하는 것이 아니라, 각 입장이 어떤 전제를 받아들이고 어떤 결론을 도출하는지 논증의 구조로 재구성하고 비교할 수 있어야 합니다. 즉 한 입장을 소개하는 데 그치지 않고, 그 입장에 대한 반론이 무엇이며 그 반론이 성공하는지 여부까지 평가할 수 있어야 합니다. 이를 통해 학생들은 철학적 논쟁을 '의견 대립'이 아니라 '이유의 경쟁'으로 이해하게 됩니다.

마지막으로 학생들은 이러한 개념과 논증을 실제 문제에 적용할 수 있어야 합니다. 인공지능이 개입한 사고의 책임 귀속, 도덕적 조언 시스템의 정당화 가능성, 예측 기술과 공정성의 충돌, 인간과 인공지능의 관계 변화 같은 구체적 사례에서 어떤 규범적 판단이 정당화되는지 스스로 설명하고 방어하는 것이 수업의 궁극적 목표입니다. 이를 통해 학생들은 인공지능 윤리를 하나의 최신 주제가 아니라, 도덕철학 전반을 이해하는 하나의 훈련 장으로 활용하게 됩니다.

면담: 수요일/금요일 1 시 30 분부터 2 시 30 분까지입니다. 장소는 5 남 223 호입니다. 자유롭게 방문하셔도 좋으나 사전 예약을 추천드립니다. 예약은 저에게 이메일을 주시거나 다음의 링크를 통해 원하시는 시간을 말씀해 주세요. https://calendly.com/bronzeyong-inha

지원: 별도의 지원이 필요한 학생은 직접 말씀하시거나 위 메일로 연락바랍니다.

성적:

최종점수는 아래 과업에서 획득한 점수를 비율에 따라 합산한 결과에 의해 결정됩니다.

과업	비율
출석	10
중간	45
기말	45

← 질병 등의 합당한 사유로 결석하시면 감점 처리 하지 않습니다. 이 경우 사전에 미리 말씀하시거나 결석 후 10 일 이내에 말씀주세요. 이유를 말씀하지 않고 결석해도 1 회에 한해 감점 하지 않겠습니다.

결석으로 인한 감점은
5 점입니다. 출석이 전체
성적의 10%이니 1 회
결석으로 인해 0.5 점을
전체 성적에서 잃게
되신다고 보시면
되겠습니다. 단 7 회 이상
결석하시면 본 수업을
통과하실 수 없습니다.
지각으로 인한 감점은
2.5 점입니다.

본 교과는 절대 평가방식을 따릅니다. 최종 점수를 토대로 아래의 기준에 따라 최종 학점을 결정합니다.

등급	점수
A+	95.00-100.0
A0	90.00-94.99
B+	85.00-89.99
B0	80.00-84.99
C+	75.00-79.99
C0	70.00-74.99
D+	65.00-69.99
D0	60.00-64.99
F	59.99-

논문:

신상규 인공지능은 자율적 도덕행위자일 수 있는가 (2017)
고인석 인공지능이 자율성을 가진 존재일 수 있는가 (2017)
정태창 자아 없는 자율성_인공 지능의 도덕적 지위에 대한 고찰 (2020)
정태창 인공 지능의 도덕적 지위와 현상적 의식 (2024)
천현득 인공 지능에서 인공 감정으로 (2017)
이병덕 Can Artificial Intelligence Be a Member of the Moral Community? (2025)
신상규 인공지능의 도덕적 지위와 관계론적 접근 (2019)
고인석 인공지능을 활용하는 인지능력 향상의 전망 (2020)

윤준식 인공지능을 활용한 도덕 향상에 대한 비판적 검토 (2024)
김은희 인공적 도덕행위자와 도덕적 책임의 문제 (2021)
손제연 약인공지능 시대의 윤리적 위기 개별화된 예측 정보의 증가로 인한 상호협력 기반의 상실 (2022)
목광수 인공지능 개발자 윤리: 덕성 기반의 모델을 제안하며 (2020)
박찬국 인간과 인공지능의 미래: 인간과 인공지능의 존재론 (2018)

일정:

9 월 2 일	개강
9 월 16 일	추석연휴
10 월 23-25 일	중간고사
12 월 18-29 일	기말고사

1 주차: 9 월 2 일

☐ 수업 내용 및 평가 기준을 설명합니다.

☐ 아래의 질문에 답하며 신상규(2017)를 읽어 오시기 바랍니다.

1 도덕 행위자란 무엇이며, 전통적으로 어떤 존재에게 적용되어 왔는가?

2 자율적 인공지능(AI)의 등장은 도덕 행위자 개념에 어떤 수정을 요구하는가?

3 AI 가 도덕 행위자가 될 수 있다고 주장하기 위해 논문에서 제시한 요건은 무엇인가?

4 논문에서 제시하는 '기능적 도덕 행위자'란 무엇인가?

5 AI 의 자율성이란 무엇이며, 전통적인 자동화 개념과 어떻게 다른가?

6 EU 의회가 AI 로봇에게 '전자인격' 지위를 부여하려는 이유는 무엇인가?

7 논문에서 제시한 AI 와 관련된 도덕적 반론들은 무엇이며, 이에 대한 대응 논리는 무엇인가?

8 AI 가 '자율적 행위자'로 인정받기 위해 필요한 '2 차 수준의 자유도'란 무엇인가?

9 인간과 AI 사이의 도덕적 지위에 대한 비교와 차이점은 무엇인가?

10 AI 에 귀속될 수 있는 책임과 책무성의 개념은 어떻게 정의되는가?

2 주차: 9 월 9 일

□ 아래의 질문에 답하며 고인석(2017)을 읽어 오시기 바랍니다.

1	자율성의 정의와 그 중요성은 무엇인가?	6	인공지능이 자율적 존재가 되는 것이 공학적으로 가능한가?
2	자율적 에이전트란 무엇이며, 공학에서의 자율성과 철학적 자율성의 차이는 무엇인가?	7	크리스만의 자율성 개념에서 자기 전체를 반성하고 변경하는 능력이란 무엇인가?
3	인공지능이 자율성을 가질 수 있다는 주장에 대한 주요 논거는 무엇인가?	8	자율성을 지닌 인공지능이 초래할 수 있는 위험에는 어떤 것들이 있는가?
4	루소와 칸트의 자율성 개념은 각각 어떻게 정의되며, 인공지능의 자율성 논의에 어떻게 적용되는가?	9	인공지능의 자율성에 대한 사회적-윤리적 함의는 무엇인가?
5	인공지능이 자율적 존재가 되는 것이 이론적-개념적으로 가능한가?	10	자율성을 가진 인공지능을 인간의 대리인으로 삼는 것이 정당한가?

3 주차: 9 월 16 일

□ 추석 연휴로 휴강입니다. 수업은 녹화영상으로 대체합니다. 23 일 전까지 영상을 시청해 주세요.

□ 아래의 질문에 답하며 정태창(2020)을 읽어 오시기 바랍니다.

1	자율성의 정의와 인공지능에 적용되는 자율성의 개념은 무엇인가?	6	인공지능의 자율성이 인간의 도덕적 자율성과 어떻게 다른가?
2	'자아 없는 자율성'이라는 개념은 무엇을 의미하며, 왜 중요한가?	7	인공지능이 목적을 선택하는 능력과 인간의 목적 선택 능력의 차이점은 무엇인가?
3	자기 이익(self-interest)의 개념이 인공지능의 도덕적 지위 논의에서 어떤 역할을 하는가?	8	인공지능의 도덕적 지위에 대한 종차별주의 비판은 어떻게 반박되는가?

4	인공지능의 자율성이 도덕적 지위 부여의 근거가 될 수 없는 이유는 무엇인가?	9	인공지능이 도덕적 행위자로 간주될 수 없는 이유는 무엇인가?
5	관계론적 접근과 속성론적 접근의 차이점은 무엇인가?	10	미래에 자기 이익을 추구하는 인공지능이 등장할 가능성에 대한 논의는 무엇인가?

4 주차: 9 월 23 일

□ 아래의 질문에 답하며 정태창(2024)를 읽어 오세요.

1	인공지능의 도덕적 지위를 논의하는 이유는 무엇인가?	6	차머스의 철학적 좀비 논변이 인공지능의 도덕적 지위 논의에 어떻게 적용되는가?
2	감성주의(sentientism)란 무엇이며, 인공지능의 도덕적 지위와 어떻게 관련되는가?	7	감성주의가 인간중심주의에 빠져 있다는 비판은 무엇이며, 이에 대한 저자의 반박은 무엇인가?
3	현상적 의식(phenomenal consciousness)의 정의와 인공지능에 적용될 가능성은 무엇인가?	8	인공지능이 현상적 의식을 가질 수 있다고 주장하는 논거는 무엇인가?
4	속성론적 입장과 관계론적 입장 간의 주요 차이점은 무엇인가?	9	감성주의에 따르면 인공지능의 도덕적 지위를 결정하는 데 현상적 의식이 필수적인 이유는 무엇인가?
5	기능주의의 문제점으로 제시된 '방만한 기능주의'와 '도구주의'는 무엇인가?	10	인공지능의 도덕적 지위와 관련된 미래 연구 방향은 무엇인가?

5 주차: 9 월 30 일

□ 아래의 질문에 답하며 천현득 (2017)을 읽어 오시기 바랍니다.

1 이 글은 알파고 같은 시스템을 “인공지능”이라 부를 수 있는가라는 질문에서 출발해, 그 논쟁이 왜 예전만큼 중요하지 않게 되었는지 보여주는가?

2 저자가 말하듯, 사람들이 “인간의 고유성”을 이성보다 감정에서 찾으려는 전환이 실제로 설득력 있게 제시되는가?

3 저자는 감정을 선험적으로 정의하지 않고, 감정이 수행하는 역할(평가, 인지 촉진, 행위 안내, 사회적 유대 등)로부터 감정 부여의 기준을 제안한다. 이 방식은 감정의 본성을 이해하는 데 더 적합한 접근으로 보이는가?

4 글에서 제시되는 감정의 핵심 역할들(예: 위험에 대한 빠른 평가, 선택적 주의, 기억 강화, 동기부여, 사회적 소통)이 하나의 통일된 감정 개념으로 잘 묶이는가, 아니면 서로 다른 현상을 억지로 한 범주로 묶는 느낌도 있는가?

5 저자가 “진정한 감정 로봇”의 조건으로 제시하는 원초적 자아 모형(proto-self model)과 기본 충동/욕구는 감정의 필수 조건으로 타당해 보이는가? (다른 대안 조건이 더 적절해 보이는가?)

6 저자는 “감정은 일반지능과 강하게 연결되어 있다”는 점을 근거로, 근미래에 감정 로봇이 어렵다고 본다. 여기서 “일반지능”에 대한 요구는 과도한가, 아니면 불가피한가?

7 글이 구분하는 감정 인식-감정 표현-감정 생성 가운데, 저자는 왜 “인식/표현만으로는 감정을 소유했다고 보기 어렵다”는 결론을 설득력 있게 내리는가?

8 “감정 로봇이 더 안전한 AI 를 만든다”는 주장에 대해, 저자는 감정의 명암(분노, 공포, 폭력적 쾌감 등)을 들어 반박한다. 이 반박은 핵심을 찌르는가, 아니면 “안전”이 뜻하는 바를 다르게 이해한 데서 생긴 엇갈림이 있는가?

9 글의 핵심 경고인 “탈인용부호 현상”(로봇의 ‘감정’이 실제 상호작용에서 따옴표를 잃는 현상)은 로봇 윤리에서 중요한 문제를 정확히 포착하는가? 비슷한 현상을 더 잘 설명하는 다른 개념(의인화, 기만, 정서적 의존 등)이 있다고 보는가?

10 저자가 제안하는 대응(예: 로봇이 감정이 없다는 신호를 지속적으로 주기, 도덕적 추론 내장 제도화 등)은 실제로 효과가 있을 것처럼 보이는가? 오히려 로봇의 목적(자연스러운 상호작용)을 훼손할 위험이 더 크다고 보는가?

□ 아래의 질문에 답하며 이병덕(2025)을 읽어 오시기 바랍니다.

1	저자가 인공지능이 도덕 공동체의 일원이 될 수 없다고 주장하는 세 가지 핵심 근거는 무엇인가?	6	스패로우는 인공지능이 '슬픔, 후회, 공감'과 같은 인간의 도덕적·정서적 반응의 대상이 되기 어렵기에 튜링 트리아지 테스트를 통과할 수 없다고 주장한다. 그의 논리는 무엇인가?
2	칸트의 '당위는 가능을 함축한다' 원칙은 무엇을 의미하며, 저자는 이를 통해 왜 비이성적 동물이 도덕 공동체에서 제외된다고 설명하는가?	7	도덕적 행위가 '처벌에 대한 두려움'이 아닌 '도덕적 올바름에 대한 인식'에서 비롯되어야 한다는 칸트의 관점을 인정하면서도, 저자는 왜 사회적 제재가 도덕 공동체 유지에 필수적이라고 강조하는가?
3	저자에 따르면 '도덕적 의무로부터 행동'하기 위해 '진정한 자기 이익'은 왜 필수적인가? 또한, 인공지능은 왜 이것이 결여되어 있다고 저자는 주장하는가?	8	이 논문의 주장을 종합할 때, 인공지능이 인간 수준의 지능을 갖추더라도 도덕 공동체의 일원이 될 수 없는 근본적인 이유는 무엇이라 할 수 있는가?
4	저자는 도덕 공동체 유지를 위해 효과적인 '사회적 제재'가 왜 필요하다고 주장하며, 이러한 제재가 인공지능에게는 효과가 없을 것이라고 보는 이유는 무엇인가?	9	AI 의 ‘기능적 목표’와 인간의 ‘진정한 자기이익’ 구분은 설득력 있는가?
5	로버트 스패로우가 제안한 '튜링 트리아지 테스트'는 무엇을 평가하기 위한 것이며, 어떤 상황을 가정하는 테스트인가?	10	AI 가 복제 가능하다는 이유만으로 그 소멸의 도덕적 중대성이 약해진다고 보아야 하는가?

7 주차: 10 월 14 일

□ 다음주 수업시간에 중간고사가 진행됩니다. 범위는 1 주차부터 7 주차까지 수업 내용입니다. 서술형 두 문제가 출제됩니다.

첫 번째 문제를 답하기 위해선 도덕적 행위자, 자율성의 종류, 인간과 인공지능이 가질 수 있는 자율성에 대해 정리하여야 합니다.
두 번째 문제는 도덕적 지위, 도덕적 지위에 관한 이론, 이 이론의 관점에서 바라본 인간과 인공지능의 도덕적 지위를 이해하셔야 답할 수 있습니다.

□ 다음주 수업시간에 중간고사가 진행됩니다. 범위는 개강 이후부터 신상규 (2019)까지 다룬 수업 내용입니다.

□ 문제 유형은 서술형 2 문제입니다. 첫 번째 문제를 답하기 위해서는 다음 내용을 정리해 두셔야 합니다. 도덕적 행위자란 무엇인지, 자율성의 의미와 여러 종류(단순 기능적 자율성, 규범적/도덕적 자율성의 구분), 인간과 인공지능이 각각 어떤 의미에서 자율성을 가질 수 있는지에 대한 논쟁. 두 번째 문제를 답하기 위해서는 다음 내용을 이해하고 있어야 합니다. 도덕적 지위의 의미(책임과의 차이 포함), 도덕적 지위를 판단하는 기준들(이성, 자아, 의식, 관계 등), 이러한 기준에 비추어 인간과 인공지능의 도덕적 지위를 어떻게 평가할 수 있는지.

□ 수업 시간에 다루었던 논증의 구조와 개념 구분을 중심으로 정리해 오시면 충분히 풀 수 있는 수준으로 출제됩니다. 단순 요약보다는 개념을 구분하고 입장을 비교·설명하는 방식으로 준비하시기 바랍니다.

□ 아래의 질문에 답하며 신상규 (2019)을 읽어 오세요.

1	인공지능의 도덕적 지위란 무엇이며, 왜 중요한가?	6	감정 로봇이 인공지능의 도덕적 지위 논의에 어떻게 기여하는가?
2	실재론적 접근과 관계론적 접근의 차이점은 무엇인가?	7	마크 쿠헬버그의 관계론적 접근이 도덕적 지위 평가에 어떻게 적용되는가?
3	인간-기계 사이의 정서적 상호작용이 인공지능의 도덕적 지위 논의에 어떤 역할을 하는가?	8	도덕적 지위 평가에서 주관적 경험과 객관적 특성의 역할은 무엇인가?
4	관계론적 접근이 인공지능의 도덕적 지위 평가에서 갖는 장점은 무엇인가?	9	관계론적 접근이 상대주의에 빠지지 않고 모종의 객관주의를 옹호할 수 있는 방법은 무엇인가?

5	실재론적 접근이 인공지능의 도덕적 지위 평가에서 직면하는 한계와 문제점은 무엇인가?	10	인공지능의 도덕적 지위 논의가 미래 사회에 미칠 잠재적 영향은 무엇인가?

8 주차: 10 월 21 일

☐ 수업 중 중간고사가 진행됩니다.

☐ 학기 중 강의 평가가 포털에서 진행됩니다. 수업 개선을 위해 참여바랍니다.

9 주차: 10 월 28 일

☐ 중간고사 결과에 대한 리뷰를 진행합니다.

☐ 아래의 질문에 답하며 고인석(2020)을 읽어 오세요.

1	인공지능을 인간 인지체계의 일부로 통합하기 위해 필요한 조건은 무엇인가?	6	인공지능이 인간 인지체계와 '외적 결합'과 '내적 결합'의 차이점은 무엇인가?
2	생체공학 의수가 인공지능을 활용하여 인간의 인지능력을 보완하는 방식은 무엇인가?	7	인공지능과 인간 인지체계의 통합이 가져올 윤리적, 사회적 영향은 무엇인가?
3	알파고 칩을 뇌에 이식하는 경우, 인지능력이 어떻게 향상될 수 있는가?	8	뇌에 알파고 칩을 이식한 기사의 사례가 주는 시사점은 무엇인가?
4	통번역 인공지능 칩이 뇌에 이식될 때의 가능성과 한계는 무엇인가?	9	중국어 방 논증과 통번역 칩의 비교를 통해 드러나는 주요 쟁점은 무엇인가?
5	뇌-컴퓨터 연결 기술의 현재 상태와 미래 전망은 무엇인가?	10	논문에서 제시한 기술적 도전 과제와 이를 해결하기 위한 방안은 무엇인가?

10 주차: 11 월 4 일

☐ 아래의 질문에 답하며 윤준식(2024)를 읽어 오세요.

1	저자는 AME(인공적 도덕 향상)의 성립 조건과 한계를 어떤 기준으로 규정하려고 하는가?	6	저자는 AEA 논의가 '기능'에 치우쳤다고 비판하며, 인간-기술 협력(관계적 측면)을 어떤 방식으로 전면화하는가?

2	기존 AME 논의의 세 유형(정보 제공형/소크라테스적 대화형/모듈식 대화형)이 각각 강조하는 AEA 의 역할은 어떻게 구분되는가?	7	AEA 의 윤리적 쟁점으로 책임 공백(responsibility gap)이 등장할 때, 그 문제가 AEA 맥락에서 특히 두드러지는 조건은 어떻게 제시되는가?
3	'정보 제공형'이 사용자 주체성 보존을 목표로 한다는 설명은 설득력 있게 제시되는가?	8	프라이버시 보호와 도덕 향상 사이의 긴장(정보 수집의 능동성/수동성)은 논문에서 어떤 트레이드오프로 정리되는가?
4	라라·데커스가 정보 제공형을 비판하며 소크라테스적 대화형을 제안할 때, 핵심 문제를 '도덕적 능력(moral skill)' 관점에서 어떻게 짚고 있는가?	9	소크라테스적 산파술을 '관계적'으로 재해석하면서, 도덕 향상이 부정적 감정(불쾌·당혹·정체성 균열)을 포함할 수 있다는 논지는 어떻게 전개되는가?
5	폴크만·가브리엘스의 '모듈식 대화형'은 소크라테스적 대화형의 한계를 어떤 이유로 확장하려는가?	10	저자가 제안하는 도덕(moral/morality)과 윤리(ethics) 구분은 AEA 의 자율성 조절 기준으로 얼마나 실용적으로 기능할 수 있는가? (하인츠 딜레마 같은 중첩 사례에서 한계가 드러나는가?)

11 주차: 11 월 11 일

□ 아래의 질문에 답하며 김은희(2021)을 읽어 오세요.

1	인공적 도덕행위자(AMA)란 무엇이며, 그 특징은 무엇인가?	6	'블랙박스 문제'란 무엇이며, 이 문제는 AMA 의 도덕적 책임 논의에서 어떤 역할을 하는가?
2	인공적 도덕행위자에게 도덕적 책임을 물을 수 있는가? 있다면 어떤 성격의 책임인가?	7	환경 통제력이 증가하는 상황에서 AMA 의 책임은 어떻게 설정될 수 있는가?

3	해명책임, 분산적 책임, 기능적 책임이란 무엇이며, 각 책임의 구체적인 사례는 무엇인가?	8	AMA 의 도덕적 행위자가 되기 위한 조건과 그 행위의 도덕적 평가 기준은 무엇인가?
4	인공적 도덕행위자가 도덕적 책임을 지는 데 있어 '여러 손의 문제'와 '블랙박스 문제'가 왜 중요한가?	9	AMA 의 도덕적 책임을 논의하는 목적은 무엇인가?
5	'여러 손의 문제'란 무엇이며, 이 문제를 해결하기 위해 AMA 에게 어떤 책임을 물을 수 있는가?	10	인간 행위자의 책임과 AMA 의 책임은 어떻게 다르며, 그 차이점은 무엇인가?

12 주차: 11 월 18 일

□ 아래의 질문에 답하며 손제연(2022)을 읽어 오세요.

1	약인공지능 시대의 주요 윤리적 문제와 이를 초래하는 근본 원인은 무엇인가?	6	개인별 예측 기술의 발전이 보험 시장에 어떤 영향을 미치는가?
2	"설명가능한 AI" 담론이 어떤 한계를 지니는가?	7	기계학습의 불투명성 문제를 해결하기 위한 윤리적 요구는 어떻게 형성되었는가?
3	개별화된 예측 정보의 증가가 사회적 상호협력에 어떤 영향을 미치는가?	8	데이터 기반의 예측 모델에서 성별, 인종 등의 편향 문제는 어떻게 해결될 수 있는가?
4	"상호적 무지"가 신용사회에서 어떤 중요성을 지니며, 예측 기술의 발전이 이를 어떻게 위협하는가?	9	공정성과 예측 정확성 간의 충돌은 어떤 상황에서 발생하며, 이를 해결하기 위한 방안은 무엇인가?
5	예측 정보를 사용하는 데 있어 정보비대칭이 어떤 윤리적 문제를 발생시키는가?	10	차별적 환경이 예측 모델의 결과에 미치는 영향을 상쇄하기 위해 어떤 대안적 평가모델이 필요한가?

13 주차: 11 월 25 일

□ 아래의 질문에 답하며 목광수(2020)를 읽어 오세요.

1	현재의 AI 윤리 논의에서 AI 개발자 윤리의 중요성과 필요성은 어떻게 설명되고 있는가?	6	AI 개발자가 덕성을 함양하기 위한 효과적인 교육 방법은 무엇인가?
2	덕성 기반의 AI 개발자 윤리 모델이 기존의 윤리 모델과 어떻게 다른가?	7	'호모 파베르의 역설'이란 무엇이며, AI 과학기술에서 어떻게 나타나는가?
3	과학적 덕성 모델(scientific virtue model)이 AI 개발자 윤리에 어떻게 적용될 수 있는가?	8	'윤리 실현(ethics realizing)'이 '윤리 세탁'과 어떻게 다른가?
4	AI 윤리 지침들이 '윤리 세탁(ethics washing)'으로 비판받는 이유는 무엇인가?	9	AI 윤리가 체계적이고 통합적으로 제시되어야 하는 이유는 무엇인가?
5	1 인칭, 2 인칭, 3 인칭 관점이 통합된 AI 윤리 구조의 필요성은 무엇인가?	10	AI 개발자 윤리가 AI 기술 발전과 사회적 책임을 조화시키는 데 어떻게 기여할 수 있는가?

14 주차: 12 월 2 일

□ 아래의 질문에 답하며 박찬국(2019)를 읽어 오세요.

1	강한 인공지능과 약한 인공지능의 차이는 무엇이며, 논문에서는 이 차이를 어떻게 설명하고 있습니까?	6	인공지능이 인간의 뇌와 동일하게 작동할 수 있다는 주장에 대해 논문은 어떻게 비판하고 있습니까?
2	논문에서 제기된 인간중심주의에 대한 비판은 무엇입니까?	7	논문에서 다루는 인공지능의 정보처리 능력과 인간의 사고 방식의 차이는 무엇입니까?

3	인공지능과 인간의 본질적인 차이는 무엇이라고 논의되었습니까?	8	인공지능의 발전 방향을 논의할 때, 인간의 정신을 이해하는 데 중점을 두어야 한다고 주장하는 이유는 무엇입니까?
4	논문에서는 인공지능 연구가 인간의 자연지능을 모방하는 것보다 더 생산적이라고 주장하는 이유는 무엇입니까?	9	논문에서는 미래의 인공지능이 도덕적 판단을 내리고 자율적으로 행동할 수 있을 가능성에 대해 어떻게 평가하고 있습니까?
5	유물론적 형이상학이 인간과 인공지능의 본질적 동일성을 주장하는 데 어떤 역할을 합니까?	10	인공지능을 인간과 동일하게 만드는 것과 신적인 존재로 만드는 것에 대한 논문 저자의 견해는 무엇입니까?

15 주차: 12 월 9 일

□ 다음주에 기말고사가 진행됩니다. 범위는 중간고사 이후부터 14 주차까지의 수업 내용입니다. 문제 유형은 서술형 2 문제입니다.

□ 첫 번째 문제는 인공지능의 도덕적 지위를 둘러싼 논쟁을 이해하고 정리하는 능력을 묻는 문제입니다. 현상적 의식(느낌·고통·쾌락의 경험)이 도덕적 지위 판단에서 어떤 역할을 하는지, 감정과 공감이 도덕 판단과 행위에 왜 중요한지, 그리고 인공지능이 이러한 특성을 가질 수 있다고 볼 수 있는지에 대한 입장들을 연결하여 설명할 수 있어야 합니다. 단순히 한 주장만 소개하는 것이 아니라, 서로 다른 기준들이 왜 제시되었는지와 그 기준에 따라 인간과 인공지능의 지위가 어떻게 달라지는지를 논증의 흐름 속에서 설명하는 것이 중요합니다.

□ 두 번째 문제는 도덕적 책임과 인공지능 사회의 윤리 문제를 종합적으로 묻는 문제입니다. 도덕적 책임이 성립하기 위한 조건(행위 통제, 이해 가능성, 선택 가능성 등)을 바탕으로 인공지능이 개입된 상황에서 책임을 누구에게 귀속시켜야 하는지 설명할 수 있어야 합니다. 또한 인공지능이 인간의 판단을 보조하거나 개입할 때 나타나는 문제들, 예컨대 도덕 향상, 예측 정보의 활용, 협력의 약화와 같은 쟁점을 연결하여 하나의 설명으로 정리할 수 있어야 합니다.

□ 기말고사는 개별 논문의 세부 내용 암기 여부보다는 여러 입장을 비교하고 연결하여 하나의 설명을 구성하는 능력을 평가합니다. 수업 시간에 다루었던 개념 구분과 논쟁의 핵심 쟁점을 중심으로 정리해 오면 충분히 답할 수 있습니다.

□ 이번 학기 공부한 내용을 정리하며 다음의 질문에 답하세요.

1 인공지능에게 '자율성'을 부여한다는 말은 무엇을 의미하는가? 단순한 선택 능력·학습 능력과 구별하여 설명하시오.

2 인공지능이 '도덕적 행위자'가 되기 위해 필요한 조건들은 무엇인가? 그 조건들이 실제 인공지능에 적용될 수 있는지 평가하시오.

3 '도덕적 지위'와 '도덕적 책임'은 어떻게 구분되는가? 각각을 인공지능에게 인정할 수 있는지 논증하시오.

4 자아(self) 또는 개인 동일성이 도덕적 행위자성에 필수적인가? 필수라고 보는 입장과 그렇지 않은 입장을 비교하시오.

5 현상적 의식(느낌·고통·쾌락의 경험)은 도덕적 지위를 판단하는 기준이 될 수 있는가? 그 이유를 설명하시오.

6 감정 또는 공감 능력은 도덕적 판단과 어떤 관련을 가지는가? 인공지능이 감정을 모방하거나 생성할 수 있다면 도덕성에도 변화가 생기는가?

7 도덕 규범이 어떤 존재에게 '적용된다'는 것은 무엇을 의미하는가? 인공지능에게 도덕 규범을 적용하는 것이 가능한지 논하시오.

8 인간과 인공지능의 관계가 도덕적 지위를 형성할 수 있는가? 존재의 내부 속성에 근거한 설명과 관계에 근거한 설명을 비교하시오.

9 인공지능을 이용한 인간의 도덕적 향상(조언·개입·유도 등)은 정당화될 수 있는가? 그것이 인간의 자율성을 강화하는지 약화하는지 평가하시오.

10 인공지능 사회에서 발생하는 윤리적 문제의 책임은 누구에게 귀속되어야 하는가? 사용자·개발자·기업·제도 가운데 책임 분배 기준을 제시하시오.

16 주차: 12 월 16 일

□ 수업 중 기말고사가 진행됩니다. 3 일후에 기말 리뷰 영상이 I Class 에 올라갑니다.

□ 리뷰 영상을 확인하시고 성적 관련 문의가 있으면 메일 바랍니다.