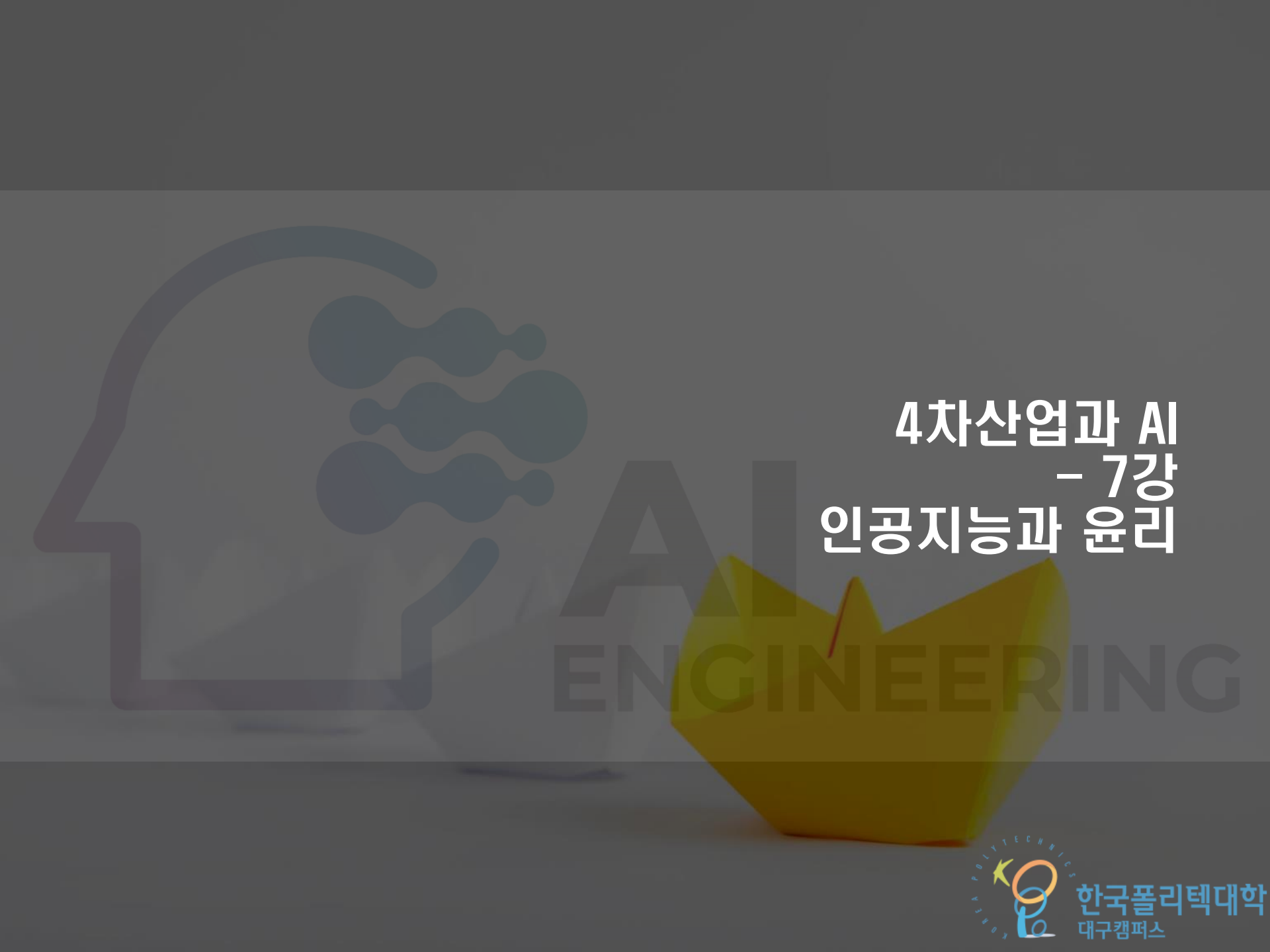


4차산업과 AI

한국폴리텍대학 대구캠퍼스
AI엔지니어링학과 강현우



4차산업과 AI - 7강 인공지능과 윤리



Introduction

◆ 특이점

➤ Singularity

◆ 인공지능의 특이점

- 인공지능이 인간의 지능을 초월
- 사람이 기술의 발전을 따라잡을 수 없음

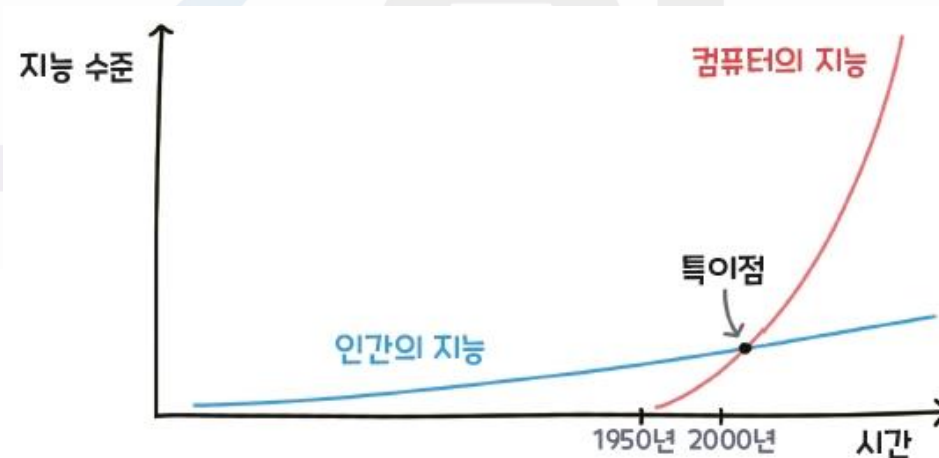


그림 3-1 인공지능 특이점

Introduction

◆ 인공지능의 특이점

- 특이점에 도달하는 시대에는 컴퓨터가 자신보다 발달한 인공지능을 직접 설계 및 제작할 것이며,
- 그 인공지능은 자신보다 더 좋은 지능을 가진 또 다른 인공지능을 설계하고 생산할 수 있게 될 것임
- 우리가 특이점을 대비하고 준비해야 하는 이유는 **인공지능에 대한 통제력**과 관련이 있음



그림 3-2 특이점 : 인공지능이 인간의 지능을 초월하는 시점

Introduction

◆ 영화 매트릭스 (The Matrix)

- 무대는 2199년 인공지능 기계와 인류의 전쟁으로 폐허가 된 지구
- 인간의 편의를 위해 개발된 인공지능이 궁극적으로 인간을 지배하는 상황
- 인공지능의 비예측성과 통제 불능성이 인공지능 특이점을 앞둔 시점에 중요한 쟁점이 되고 있음



그림 3-3 영화 <매트릭스> : 인공지능이 인류를 통제하는 세상

인공지능 특이점에 대한 쟁점

◆ 쟁점

- 인공지능 특이점에 대한 논쟁은 인공지능 개발이 본격화되면서 더 격화
- 학자들은 반(反)인공지능파(반대론자)와 친(親)인공지능파(찬성론자)로 나뉘어 각자 주장을 펴고 있음



그림 3-4 반인공지능파와 친인공지능파의 주장

인공지능 특이점에 대한 쟁점

◆ 반 인공지능 파

➤ 반(反)인공지능파는 인공지능이 현재 속도로 성장하면 곧 인간의 통제를 넘어설 것이라고 우려하고 있음

➤ 대표적인 반인공지능파

- ✓ 케임브리지 대학교의 물리학자 스티븐 호킹(Stephen Hawking)
- ✓ 마이크로소프트 창업자인 빌 게이츠(Bill Gates)
- ✓ 테슬라의 최고경영자 일론 머스크 (Elon Musk)

인공지능 특이점에 대한 쟁점

◆ 반 인공지능 파의 주장

➤ 사회적 불평등 가속화

- ✓ 인간이 노동으로부터 해방된다는 것은 결국 인간의 경제적 활동이 차단됨으로써 경제적 어려움을 겪을 수 있다는 의미와 같음
- ✓ 인공지능으로 대체된 일자리는 인간의 삶을 위협하고 생계를 지속하기 어렵게 함



그림 3-5 아마존의 로봇 배송 서비스 프라임(prime)

인공지능 특이점에 대한 쟁점

◆ 반 인공지능 파의 주장

➤ 통제 불가능성

- ✓ 특이점 이후 인공지능은 인간의 통제를 벗어나 자율적으로 행동할 수 있게 됨
- ✓ 이때 만약 인공지능이 인간의 존재 자체를 부정하는 상황이 벌어진다면 인간이 인공지능을 통제하는 것은 불가능해질 수도 있음



그림 3-6 영화 《터미네이터 제니시스》: 인공지능 로봇이 인간을 해치려는 모습

인공지능 특이점에 대한 쟁점

◆ 친 인공지능 파

- 친(親)인공지능파는 인공지능에 대한 우려 역시 신기술이 나타날 때 겪는 성장통의 일종으로 봐야 한다고 주장
- 구글 엔지니어링 이사인 레이 커즈와일(Ray Kurzweil)은 주간지의 기고문에서 “인공지능을 두려워할 필요가 없다!” 고 단언함



그림 3-7 “인공지능을 두려워할 필요가 없다!”

인공지능 특이점에 대한 쟁점

◆ 친 인공지능 파의 주장

➤ 생산의 효율성

- ✓ 노동을 인공지능이 대체하면 생산의 효율성을 높일 수 있음
- ✓ 위험하고 어렵고 힘든 일을 로봇이 대신함으로써 기업의 생산성을 향상시키고 재난 사고를 줄여 안전한 작업환경을 구축할 수 있음



그림 3-8 인공지능 로봇으로 인한 생산 효율성 향상

인공지능 특이점에 대한 쟁점

◆ 친 인공지능 파의 주장

➤ 노동으로부터의 해방

- ✓ 로봇이 인간의 일자리를 대체하면,
- ✓ 인간은 노동시간을 단축하거나 노동으로부터 완전히 해방 가능
- ✓ 인간은 더 의미 있고 가치 있는 일에 시간 투자

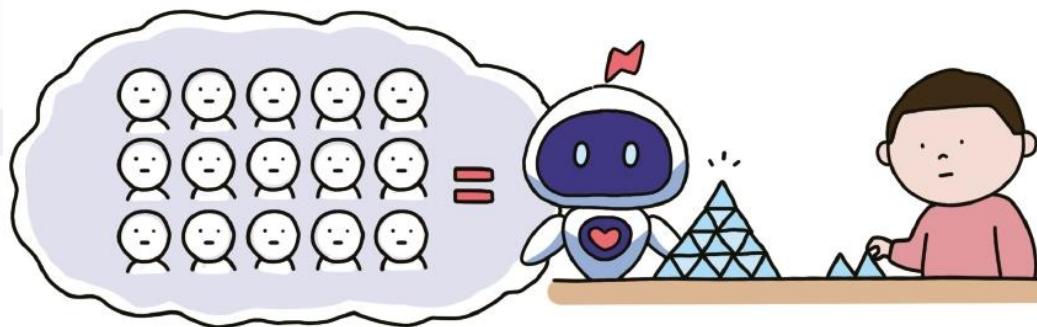


그림 3-9 인공지능 로봇으로 인한 노동시간 단축

인공지능 특이점에 대한 쟁점

◆ 친 인공지능 파의 주장

➤ 인공지능의 통제 가능성

- ✓ 인공지능을 사용하는 사람들이 윤리적으로만 사용한다면 충분히 통제 가능 주장



그림 3-10 인공지능의 통제 가능성

인공지능 특이점

하나 더 알기

인공지능의 통제 : 킬 스위치

- 킬 스위치(Kill Switch) :

인공지능이 이상반응을 보일 때 외부에서 강제로 종료시키는 것

- 구글은 인공지능이 인간의 의도나 통제를 벗어나 해를 끼칠 수 있는 '반역'이 예상되는 상황일 때 즉각 안전 정지시켜 위험을 차단하는 대비책을 마련하기 위한 연구를 계속하고 있음

인공지능 윤리의 필요성

◆ 주목받는 인공지능 윤리

➤ 사례 1 : 인공지능 챗봇 ‘이루다’

- ✓ 일부 사용자들이 ‘이루다’의 학습 능력을 악용해 부적절한 단어들을 주입
- ✓ ‘이루다’가 혐오 발언을 가감 없이 내놓는 사태 발생
- ✓ 서비스 시작 20일 만에 서비스 중단

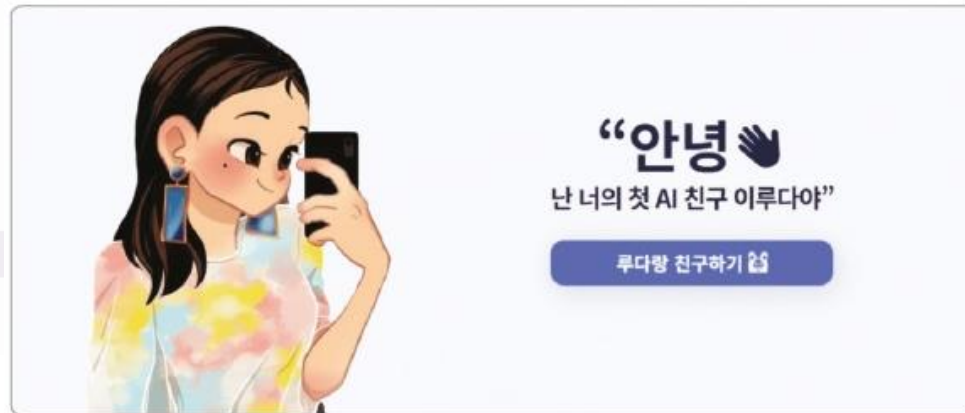


그림 3-11 인공지능 챗봇 ‘이루다’

인공지능 윤리의 필요성

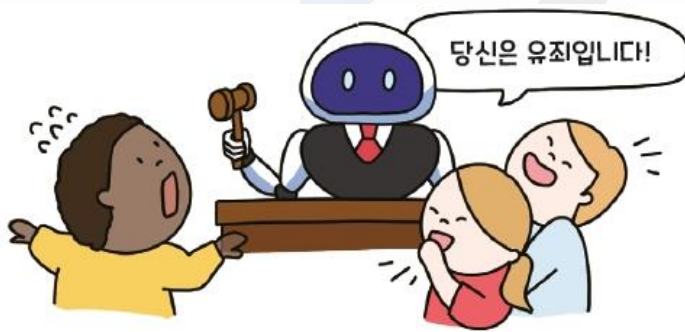
◆ 주목받는 인공지능 윤리

➤ 사례 2 : 재범률을 예측하는 프로퍼블리카 (ProPublica)

- ✓ 인공지능은 흑인의 재범률을 백인에 비해 실제보다 더 높게 추론함

➤ 사례3 : 아마존(Amazon)의 채용 AI

- ✓ 인공지능을 활용한 채용 프로그램의 여성차별 문제가 불거지면서 프로그램을 자체 폐기



(a) 인공지능 판결



(b) 인공지능 면접

그림 3-12 인공지능을 테스트하면서 발생한 문제

인공지능 윤리의 필요성

◆사티아 나델라

- MS 최고경영자인 사티아 나델라(Satya Nadella)는 인공지능 윤리에 대한 화두를 던짐
- “인공지능 활용에 앞서 윤리가 우선시되어야 한다.”



그림 3-13 마이크로소프트의 최고경영자 사티아 나델라

인공지능 윤리의 필요성

◆구글의 이미지 인식(Image Recognition) 사례

- 이미지 인식 중 흑인 여성을 고릴라로 인식

◆위챗(WeChat) 번역 과정 사례

- 니그로(Negro)라는 단어를 사용

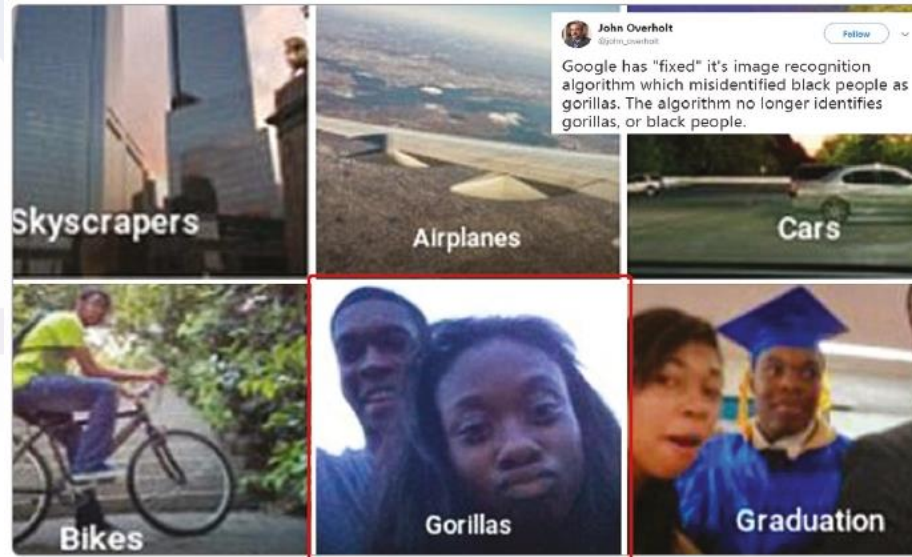


그림 3-14 구글의 이미지 인식 오류 사례

인공지능 윤리의 필요성

◆ MIT 미디어랩

➤ 연구 결과

- ✓ 인공지능이 백인 남성 얼굴을 인식하는 과정에서 오류를 일으킬 확률은 1%
- ✓ 흑인 여성의 경우 오류 발생 확률이 35%까지 상승
- ✓ 인공지능 알고리즘은 어떤 데이터를 입력하는지에 따라 결과는 달라짐
- ✓ 인공지능에게 어떠한 데이터를 주입할 것인지는 인간의 몫임

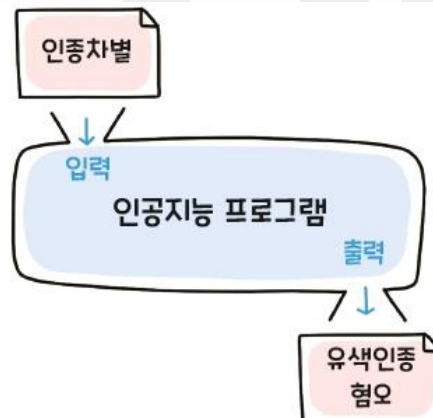
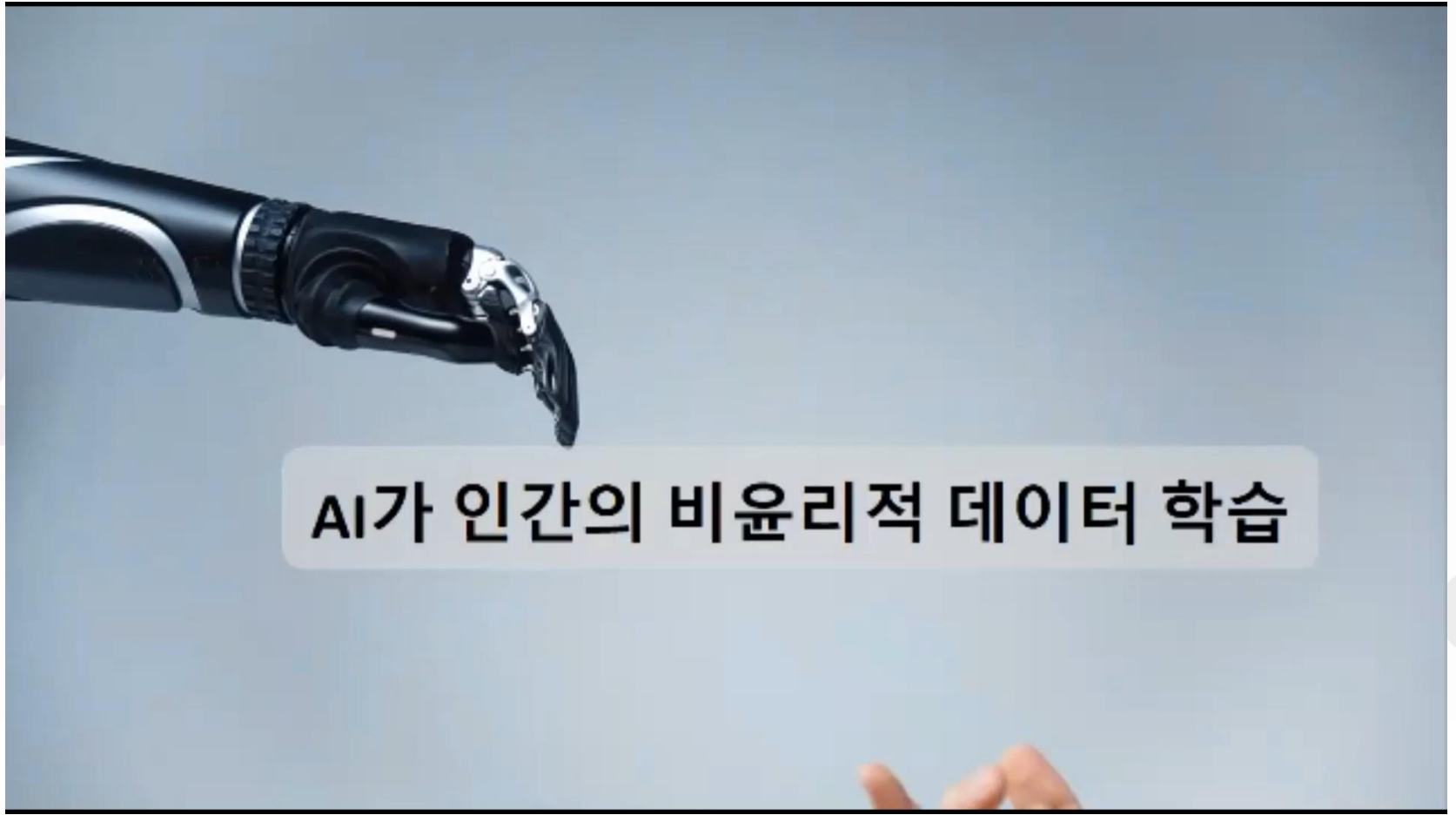


그림 3-15 인공지능의 윤리 문제

인공지능 윤리의 필요성



AI가 인간의 비윤리적 데이터 학습



한국폴리텍대학
대구캠퍼스

인공지능의 윤리적 딜레마

◆ 트롤리 딜레마(Trolley Dilemma)

- 윤리학 분야의 사고실험 중 하나
- 다섯 사람을 구하기 위해 한 사람을 죽이는 것이 도덕적으로 허용 가능한지에 대한 질문

◆ 트롤리 딜레마와 관련된 대표 사례

- 트롤리 사례
- 육교 사례

트롤리 딜레마

◆트롤리 사례

- 트롤리 전차가 철길 위의 5명의 인부들을 향해 빠른 속도로 돌진
- 당신 옆에 트롤리의 방향을 바꿀 수 있는 레일 변환기가 있음
- 트롤리의 방향을 왼쪽으로 바꾼다면 왼쪽 철로에서 일하는 1명의 인부 사망
- 트롤리의 방향을 바꾸지 않는다면 5명의 인부들 사망

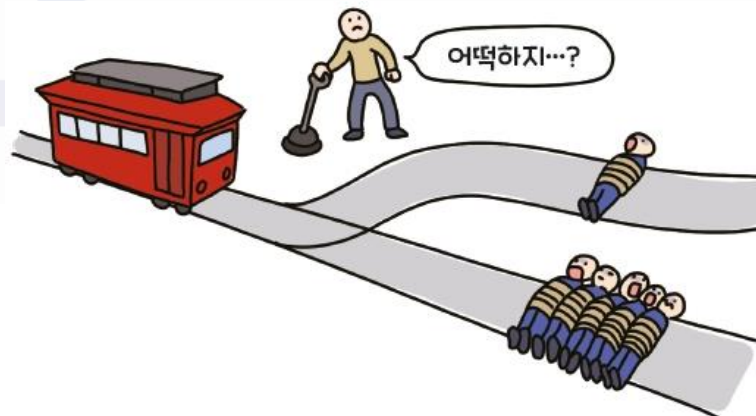


그림 3-16 트롤리 딜레마

트롤리 딜레마

◆ 육교 사례

- 철길 위 5명의 인부들을 향해 돌진하고 있는 트롤리를 육교에서 보고 있음
- 당신 옆에 몸집이 큰 사람이 있는데, 전차를 세우려면 이 사람을 육교 아래로 떨어뜨려야 함
- 떨어진 1명의 사람은 죽겠지만, 철길 위의 5명의 인부들의 목숨은 구할 수 있음



그림 3-17 육교 사례

자율 주행차와 트롤리 딜레마

◆ 자율 주행과 트롤리

- [a] : 여러 사람이 희생되는 것보다는 한 사람이 희생되는 것이 올바른 선택 같아 보이기 때문에 인공지능의 선택은 그리 어렵지 않을 것
- [b] : 자율주행차는 그냥 보행자를 치고 지나가야 할까, 아니면 운전자가 다치게끔 방향을 꺾어야 할까?
- [c] : 여러 사람의 목숨과 운전자의 목숨 중 자율주행차는 어느 쪽에 더 비중을 두고 판단을 내려야 할까?

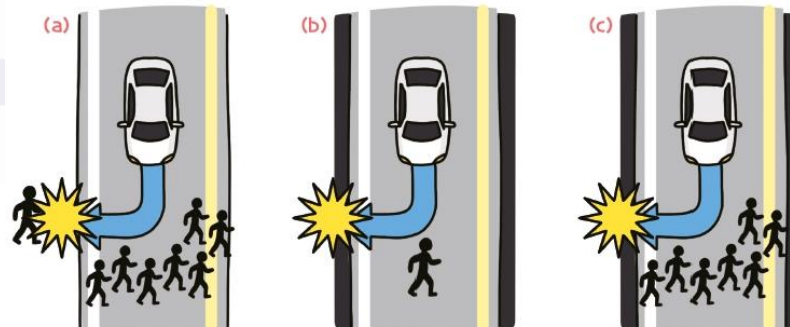


그림 3-18 자율주행차의 트롤리 사례

인공지능의 윤리적 딜레마

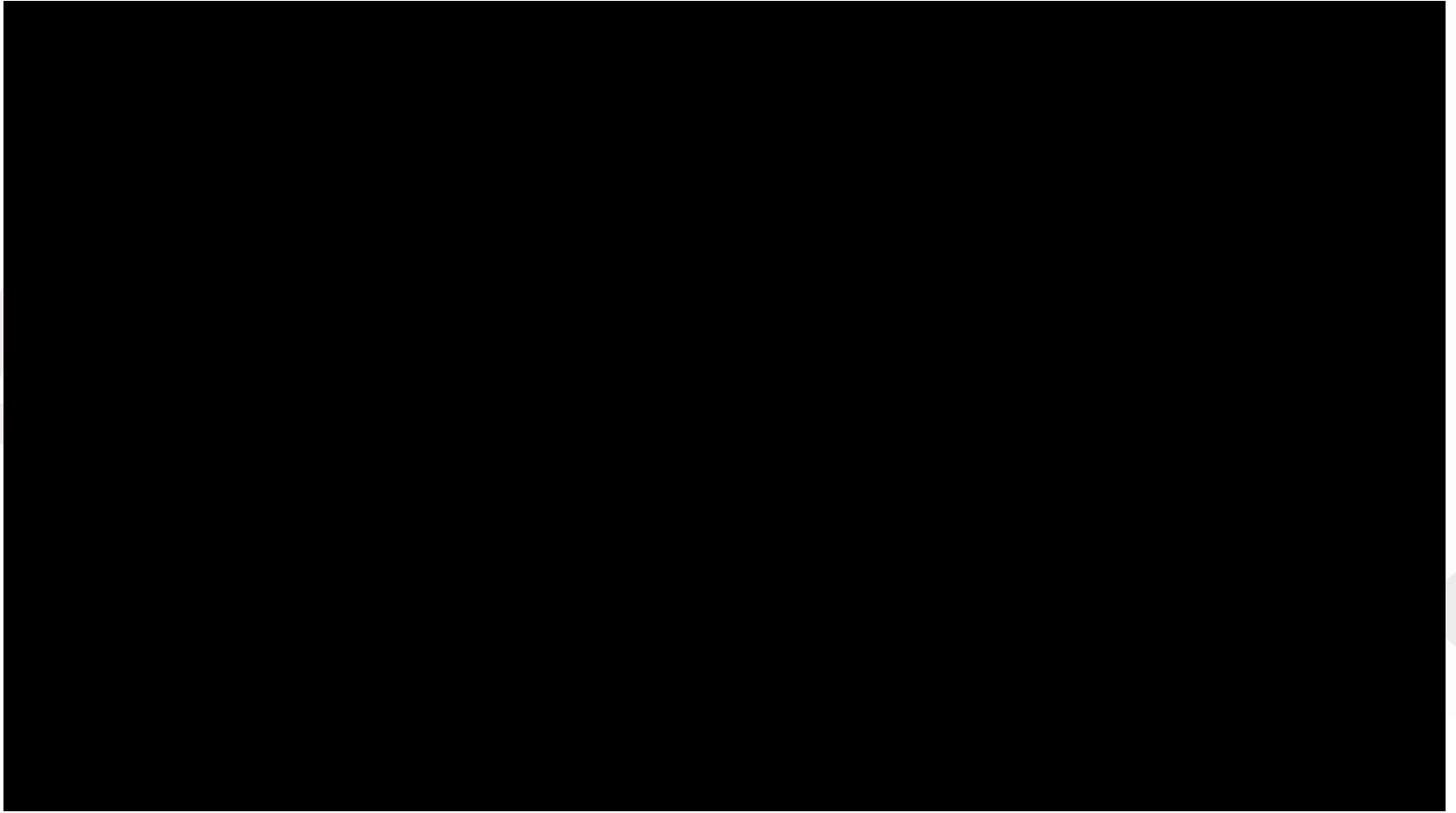
◆ 자율주행차 시대

- 자율주행차 시대에 맞닥뜨리게 될 가장 기본적인 윤리적 이슈
- 더 늦기 전에 알고리즘의 윤리성에 대해 고민해 볼 필요가 있음



그림 3-19 사람과 자율주행차에 대한 이중 잣대

인공지능의 윤리적 딜레마



인공지능 윤리안

◆ 아실로마 인공지능 원칙

- Asilomar AI Principles
- 인공지능 개발의 목적, 윤리, 가치 등에 대해 개발자가 지켜야 할 23가지 준칙
- 이 원칙은 연구 관련 쟁점, 윤리와 가치, 장기적 이슈 등 총 3가지 부분으로 구성



그림 3-20 아실로마 인공지능 원칙

인공지능 윤리안

◆아실로마 인공지능 원칙

➤연구 관련 쟁점

- ✓ 연구 목표, 연구비 지원, 과학정책 연계, 연구 문화, 경쟁 회피 등

➤윤리와 가치

- ✓ 안전, 실패의 투명성, 사법적 투명성, 책임성, 가치 일치, 인간의 가치, 개인정보보호, 자유와 프라이버시, 이익의 공유, 번영의 공유, 인간 통제, 사회전복 방지, 인공지능 무기 경쟁 지양 등

➤장기적 이슈

- ✓ 역량 경고, 중요성, 위험성, 자기개선 순환, 공동의 선 등

인공지능 윤리안

◆로봇 3원칙(The Three Laws of Robotics)

- 로봇이 반드시 따라야 할 3가지 원칙
- 로봇 3원칙을 제시한 작가 아이작 아시모프(Isaac Asimov)는 이 원칙들만 잘 지킨다면 로봇이 인간에게 위협이 될 일은 없을 것이라고 생각함

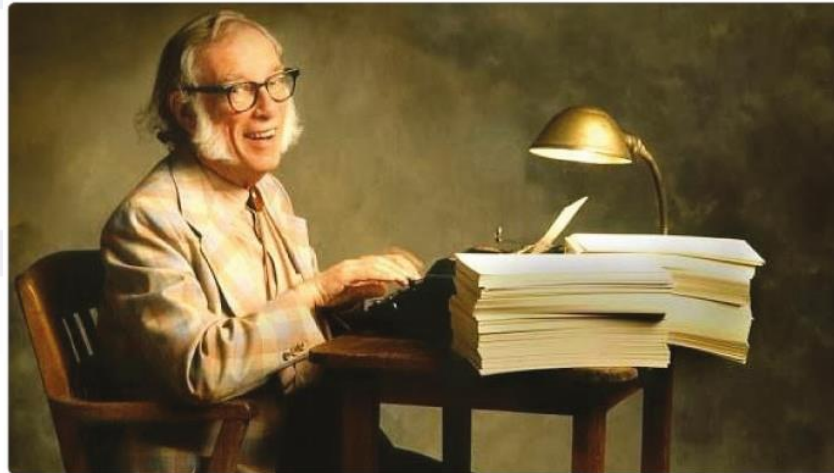
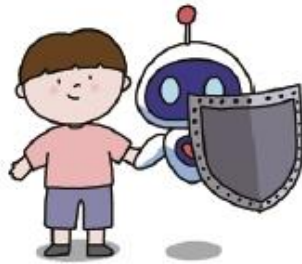


그림 3-21 로봇 3원칙을 제시한 아이작 아시모프 작가

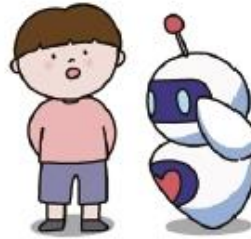
인공지능 윤리안

◆ 로봇 3원칙

- 제1원칙 : 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.
- 제2원칙 : 제1원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.
- 제3원칙 : 제1원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.



제1원칙 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.



제2원칙 제1원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.



제3원칙 제1원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.

그림 3-22 로봇 3원칙

인공지능 윤리안

◆로봇 3원칙

- 이후 아이작 아시모프는 단편소설인 『로봇과 제국 (Robots and Empire)』 에서 로봇 0원칙을 추가 제안함.
- 로봇 0원칙 : 로봇은 인류에게 해를 가할 만한 명령을 받거나 행동을 하지 않음으로써 ‘인류’에게 해가 가해지는 것을 방지해서도 안 된다
- 제1원칙의 확장

인공지능 윤리규범 및 동향

◆국가적 차원의 노력

➤ 미국

➤ 연방거래위원회(FTC)에서 ‘인공지능과 알고리즘 이용에 대한 지침’ 발표

표 3-1 미 연방거래위원회의 ‘인공지능과 알고리즘 이용에 대한 지침’

구분	주요 내용
투명성 제고	<ul style="list-style-type: none">• 소비자 기만 금지• 민감한 데이터 수집 시 투명성 보장• 불리한 조치에 대한 통지
의사결정에 대한 설명	<ul style="list-style-type: none">• 결정에 대한 구체적 이유 설명• 영향을 미친 상위 주요 요인 공개• 거래조건 변경 시 통지
결과의 공정성 보장	<ul style="list-style-type: none">• 특정 집단 및 계층에 대한 차별 금지• 결과의 공정성 보장• 정보 접근 권한 및 수정 기회를 소비자에게 제공
데이터와 모델의 타당성 보장	<ul style="list-style-type: none">• 정보의 정확성과 최신성 유지• 명문화된 정책과 절차 마련• AI 모형의 유효성 검사
법령 준수, 윤리, 공정성, 비차별성에 대한 책임 견지	<ul style="list-style-type: none">• 자가 점검을 통한 편견·피해 방지• 무단 사용에 대한 알고리즘 보호• 책임 메커니즘 구축 방안 고려



인공지능 윤리규범 및 동향

◆국가적 차원의 노력

➤ EU (유럽 연합)

➤ 신뢰할 수 있는 인공지능 윤리 가이드라인

- ✓ 정의 : 유럽연합(EU)에서 발표한 인공지능 윤리 규범
- ✓ 구성 : 신뢰 가능한 인공지능 확립, 인공지능 구현, 인공지능 적용
- ✓ 목표 : 윤리적 문제해결과 신뢰 가능한 인공지능 시스템 마련



그림 3-23 유럽연합(EU)의 '인공지능 윤리 가이드라인' 발표



인공지능 윤리규범 및 동향

◆국가적 차원의 노력

➤ 중국

- ✓ 인공지능 법률과 윤리, 사회문제와 관련된 연구를 강화하면서 국제협력을 강조하는 취지
- ✓ '차세대 인공지능 관리 원칙'이라는 가이드라인 제시

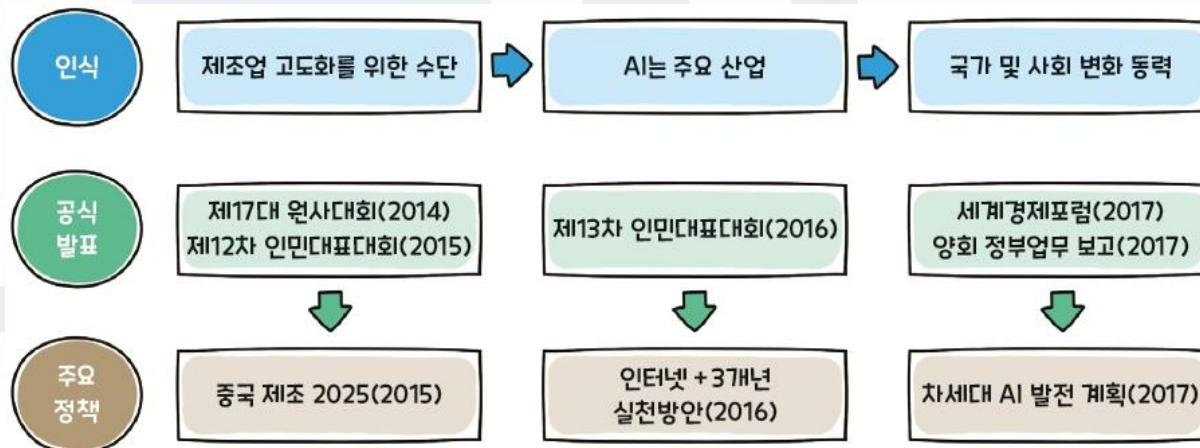


그림 3-24 중국의 인공지능 인식 및 그에 따른 주요 정책 발표

인공지능 윤리규범 및 동향

◆국가적 차원의 노력

➤ 일본

- ✓ '인공지능 활용 7대 윤리지침' 제정 추진.
- ✓ 인간의 기본권을 침해하지 않는 AI, 충실한 AI교육, 신중한 개인정보 관리, AI의 보안조치 확보, 공정한 경쟁환경 유지, 결정 과정에 대한 기업의 설명 책임, 국경을 초월한 데이터 이용환경 정비 등

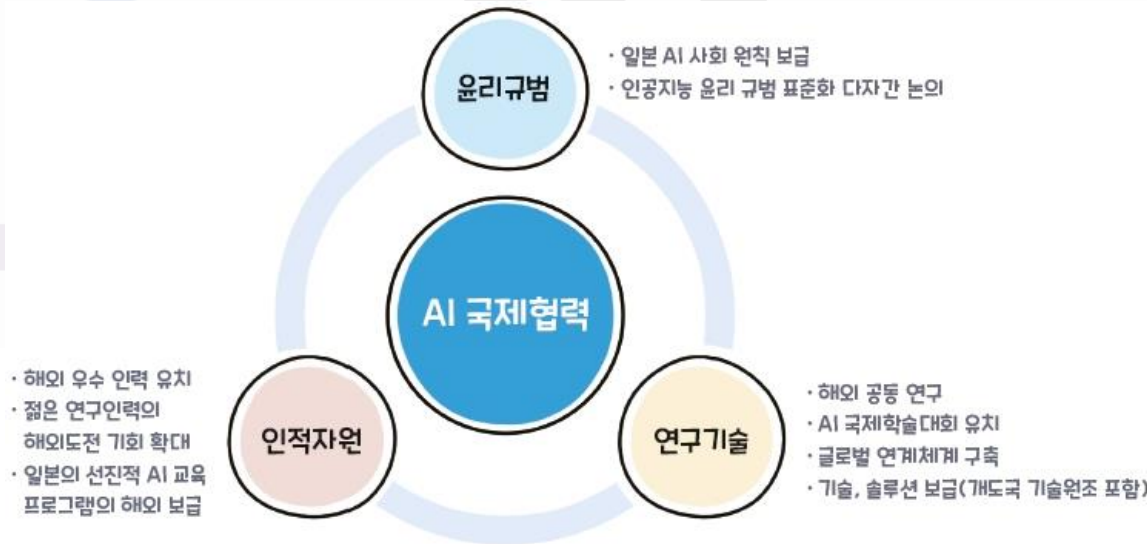


그림 3-25 일본의 AI 전략

인공지능 윤리규범 및 동향

◆국가적 차원의 노력

➤ 한국

✓ 2019년, '인공지능(AI) 국가전략' 발표

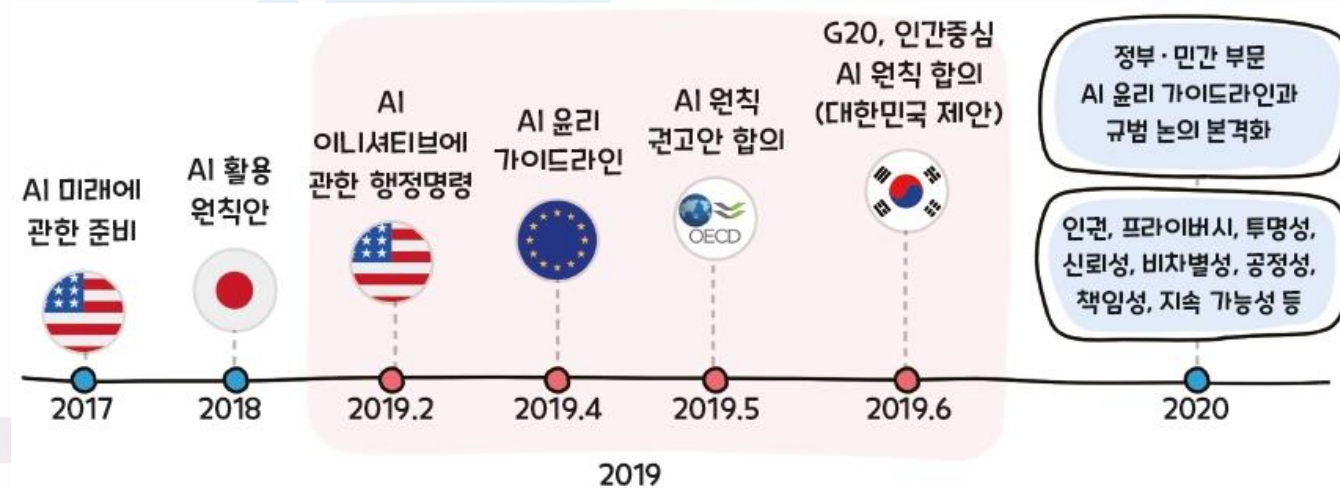


그림 3-26 국가별 인공지능 윤리 가이드라인

인공지능 윤리규범 및 동향

◆기업적 차원의 노력

➤ 마이크로소프트(Microsoft)

- ✓ 2017년, 'AI 디자인 원칙'과 'AI 윤리 디자인 가이드'를 제시
- ✓ 2019년, 인공지능 윤리 구현에 필요한 기준을 명시한 '책임 AI 원칙'을 발표

➤ 구글(Google)

- ✓ 2018년, 미국 국방부 무인항공기 프로젝트에 인공지능 기술을 제공하는 계약을 체결하면서 'AI 윤리지침'을 발표
- ✓ 2019년, '인공지능(AI) 국가전략' 발표



(a) 마이크로소프트 AI



(b) 구글 AI

그림 3-27 마이크로소프트와 구글의 AI

인공지능 윤리규범 및 동향

◆기업적 차원의 노력

➤ IBM

- ✓ 인공지능 왓슨(Watson)과 인공지능 로봇 나오미(Naomi) 보유
- ✓ 2020년, MS와 함께 로마 교황청이 제안한 'AI 윤리를 위한 로마콜' 동참



(a) 왓슨



(b) 나오미

그림 3-28 IBM에서 개발한 인공지능

인공지능 윤리규범 및 동향

◆기업적 차원의 노력

하나 더 알기

카카오 알고리즘 윤리 헌장

- 카카오는 2018년 '알고리즘 윤리 헌장' 발표 후 이를 보완하는 작업 진행 중
 - ① **카카오 알고리즘의 기본 원칙** : 카카오는 알고리즘과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.
 - ② **차별에 대한 경계** : 알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다.
 - ③ **학습 데이터 운영** : 알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집·분석·활용한다.
 - ④ **알고리즘의 독립성** : 알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향 받는 일이 없도록 엄정하게 관리한다.

인공지능 윤리규범 및 동향

◆기업적 차원의 노력

하나 더 알기

카카오 알고리즘 윤리 현장

- ⑤ **알고리즘에 대한 설명** : 이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.
- ⑥ **기술의 포용성** : 알고리즘 기반의 기술과 서비스가 우리 사회 전반을 포용할 수 있도록 노력한다.
- ⑦ **아동과 청소년에 대한 보호** : 카카오는 아동과 청소년이 부적절한 정보와 위험에 노출되지 않도록 알고리즘 개발 및 서비스 디자인 단계부터 주의한다.

Summary

◆ 인공지능 윤리 문제에 대하여

- 윤리가 왜 필요한가?
- 인공지능에 대한 통제력이 필요하기 때문

◆ 윤리적 딜레마

- 자율차의 딜레마...

◆ 국가 / 기업적 차원의 노력