숭실대학교 Soongsil University

인공지능의 안전성과 윤리

숭실대학교 베어드교양대학

서유화 교수

yhsuh@ssu.ac.kr

지난시간 배운 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티처블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 윤리적 쟁점과 다양한 이슈	1. 자율시스템으로써의 인공지능과 딥러닝 2. 인공지능 안전성과 윤리 3. 인공지능의 윤리적 쟁점 (자율주행자동차, AI로봇, 트랜스휴먼, 프라이버시 문제)	1. 비윤리적 데이터 생성과 수집 - 웹 스크레핑(크롤링)을 이용한 데이터 수집
7		기말고사	

목차

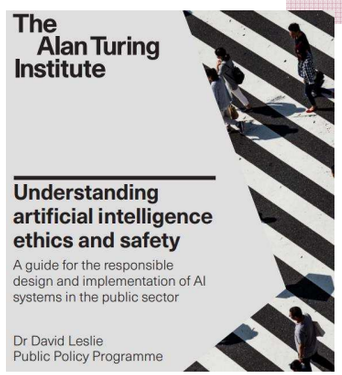
- 인공지능의 위험가능성
- 인공지능 안전성의 개념과 침해 사례
- 인공지능 윤리의 개념과 침해 사례
- 인공지능의 안전성과 윤리의 차이

학습 목표

- 인공지능의 위험성을 설명할 수 있다.
- 인공지능의 안전성 침해 사례를 설명할 수 있다.
- 인공지능 윤리의 개념을 설명할 수 있다.
- 인공지능의 윤리 침해 사례를 설명할 수 있다.
- 인공지능 위험성과 윤리의 차이를 설명할 수 있다.

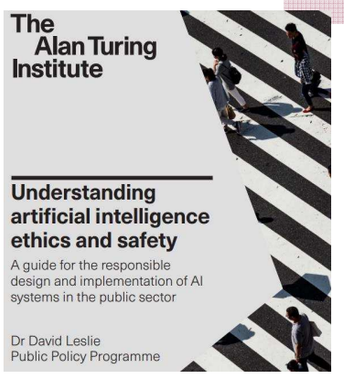
인공지능의 위험가능성

- 영국 앨런튜링 연구소의 '인공지능 위험가능성'
  - 편견과 차별
  - 개인의 자율성, 의지, 권리 거부
  - 불투명, 설명불가능, 정당하지 않은 결과
  - 프라이버시 침해
  - 사회적 관계에서의 고립과 분열
  - 신뢰할 수 없거나 안전하지 않거나, 품질이 낮은 결과



인공지능의 위험가능성

- 영국 앨런튜링 연구소의 '인공지능 위험가능성'
  - 편견과 차별
  - 개인의 자율성, 의지, 권리 거부
  - 불투명, 설명불가능, 정당하지 않은 결과
  - 프라이버시 침해
  - 사회적 관계에서의 고립과 분열
  - 신뢰할 수 없거나 안전하지 않거나, 품질이 낮은 결과



Soongsil University

인공지능의 위험가능성

- **편견과 차별**
  - AI에 사용되는 데이터는 모든 데이터를 대표하지 않는다.
  - 공정하지 못한 데이터와 알고리즘의 사용은 그 결과도 편견과 차별을 가짐
    - 데이터 결함 및 편향을 가진 학습데이터 사용에 따른 편견과 차별 복제
    - 설계자의 선입견과 편견 복제

- **개인의 자율성, 의지, 권리, 거부**
  - AI 시스템의 결정이 개인에게 영향을 미치는 경우 그 결과에 대한 책임을 묻기 어려움
    - 설계, 생산, 구현 등 모든 프로세스가 매우 복잡하고 세밀한 분산
    - 인간에게 부상이나 부정적인 결과를 만든 경우 책임 격차문제 발생
    - 그 결과에 영향을 받는 개인의 권리 침해

Soongsil University

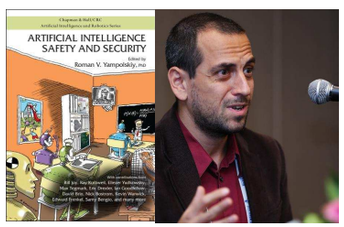
인공지능의 위험가능성

- **불투명성, 설명 불가능, 정당하지 않은 결과**
  - 알고리즘을 통해 결과가 생성, 결과에 대한 근거가 불투명 함
    - AI의 추론 과정이 고차원적 상관관계에서 도출되어 그 과정을 분석하기 어려움
    - 편견이나 불평등, 불공정한 결과에 대한 설명이 불가능
- **프라이버시 침해**
  - AI에 사용되는 데이터는 데이터 주체의 적절한 동의없이 캡쳐나 추출될 수 있음
    - 드론, SNS 데이터 등
- **사회적 관계에서의 고립과 분열**
  - 사회적 관계의 초개인화와 양극화
    - AI기반 초개인화 서비스에 따른 나와 다른 세계관에 대한 노출 제한
- **신뢰할 수 없거나, 안전하지 않거나, 품질이 낮은 결과**
  - 무책임한 데이터 관리, 부주의한 설계 및 생산, 의심스러운 배포
  - 개인의 복지와 공공복지에 피해

Soongsil University

인공지능 안전성(AI Safety)

- 인공지능의 안전성이란?
  - 인공지능 기술 · 구조적 한계 및 특징으로 인해 발생할 수 있는 의도치 않은 각종 위험들에 대비하는 개념 및 분야
    - 인공지능은 인간의 지능을 모사하기 위해 다양한 SW와 HW기술들이 융합되어 구현되는 복합체
    - 개발 과정의 오류, 상용화된 이후의 기능적 오작동 위험성을 내포
    - 예) 자율주행자동차의 물체나 환경 인식 오류, 돌발 상황, 책임규명의 한계



로만 얌폴스키(미국 루이빌대학교 교수)

- AI safety, AI 안전성이라는 용어를 처음 사용
- "특정 인간이 안전한지를 확인하는 문제" 로 축소시켜 볼 수 있으며 이것을 안전한 인간문제로 정의
- AI 시스템에서 완전한 안전은 없고, 확률적으로 안전한 것
- 완전 자율 기계는 절대로 안전할 수 없다고 가정
- 약한 인공지능 -> 강한 인공지능/초지능 발전시 안전성 확보 문제 심화

이미지 출처 : http://m.joongdo.co.kr/view.php?key=20161017000025098

인공지능의 안전성 침해 사례

- 한맥투자증권 자동매매 시스템 알고리즘 오류 (2013.2)
  - 국내 첫 사례, 2분만에 460억원 손실로 파산
  - https://youtu.be/cT8VLwjZVDg
- 러시아 AI 프로모봇 IR77 탈출 사고 (2016.6)
  - 테스트 중 두 번의 탈출 사고 발생
  - 해당로봇은 재발 방지를 위해 해체 결정
- 미국의 스탠포드쇼핑센터 경비로봇 'K5' 어린아이 공격 상해(2016. 7)

이미지 출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

인공지능의 안전성 침해 사례

- 미국 애리조나 주 우버 자율자동차 사고 (2018.3)
  - 애리조나에서 최초 완전 자율 택시 서비스를 시작
  - 자율 주행 중이던 우버 차량이 보행자를 치어 사망
    - https://youtu.be/vTkBPa1gwa4
  - 구글 웨이모는 현재 애리조나 피닉스 이스트 밸리에서 운전자가 탑승하지 않는 완전 자율주행 택시를 정식 운행 중

- 미국 플로리다 주 법무부가 사용하는 리걸테크 (legal tech) 콤파스(COMPAS) 인종차별 (2016)
  - 인터넷 언론 프로퍼블리카지의 탐사보도
  - 흑인과 백인 재범 발생 확률이 비슷함에도 흑인을 백인보다 훨씬 위험하다고 평가

이미지 출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

인공지능 윤리

- 인공지능 윤리
  - 인공지능관련 이해관계자들이 준수해야하는 보편적인 사회규범 및 관련 기술
  - 사회적 합의, 가이드라인, 규제
  - 나쁜 의도를 가진 개발자 또는 관련 제품 및 서비스 판매자, 이용자가 인공지능을 악용하는 행위는 인공지능 윤리의 가치를 위반

인공지능 윤리 침해 사례

- 인공지능 자율살상 무기 개발
  - 2025년 기준으로 인공지능을 탑재한 군사용 로봇과 드론 시장 규모는 민간 시장 규모를 압도할 것으로 예상
    - 2019. 1월 예맨 정부군 공군기지 행사에서 후티 반군 소행의 드론 폭발사고 (정부군 6명 사망, 관료 12명 부상)
    - 2018년 8월 니콜라스 마두로 베네수엘라 대통령의 군 창설 연설 도중 드론 테러 (7명 부상)

  - 구글의 군사용 메이븐 프로젝트 지원(2018. 6)
    - 구글이 인공지능 기술을 미국 공군의 메이븐(Maven) 프로젝트에 제공했다는 사실이 알려짐
    - 메이븐 프로젝트는 인공지능을 이용해 군사용 드론의 이미지를 분석하고 드론의 목표 타격률을 향상하는 미 국방부 프로젝트
    - 구글은 해당 프로젝트에서 손을 떼고 '인공지능 윤리강령 ' 을 만듦

이미지 출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

인공지능 윤리 침해 사례

- 정치조작 및 여론 조작
  - 2016년 페이스북 대선 개입 사실이 알려짐
    - 트럼프 선거 운동, 브렉시트 등에 개입
    - 성격테스트에 참여한 27만 명과 그들의 친구목록에 5천만명의 개인정보를 수집해 당사자 동의없이 케임브리지 애널리티카에 판매
    - 개인 맞춤형 트럼프에 대한 4천개의 서로 다른 온라인 광고가 수백만명의 미국인들에 의해 15억회나 조회됨

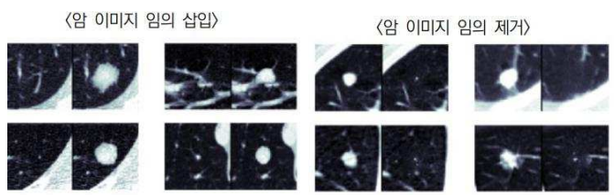
2018년 3월 25일 마크 주커버그가 영국 일간지에 낸 사과 광고

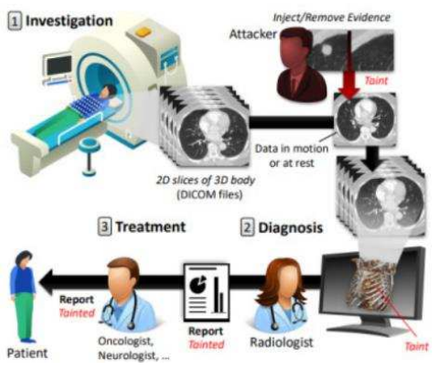
- 허위 왜곡 정보의 범람
  - 의도적인 가짜 이미지, 영상, 뉴스, 음성 생성 배포
    - 딥페이크를 이용한 편집물의 인터넷 유포
    - https://youtu.be/bOQOzX6MpA8

이미지 출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

인공지능 윤리 침해 사례

- 데이터 조작, 해킹 등 사이버 공격에 활용
  - 이스라엘 벤구리온(Ben-Gurion) 대학교 연구진의 딥러닝을 이용한 의료영상 조작 실험(2019)
    - 해킹으로 조작된 이미지를 3명의 방사선과 의사들에게 진의여부를 물어 폐암 이미지가 삽입된 가짜 이미지의 99%를 암환자 이미지로 판단
    - 가짜 이미지의 94%를 건강한 사람 영상으로 착각

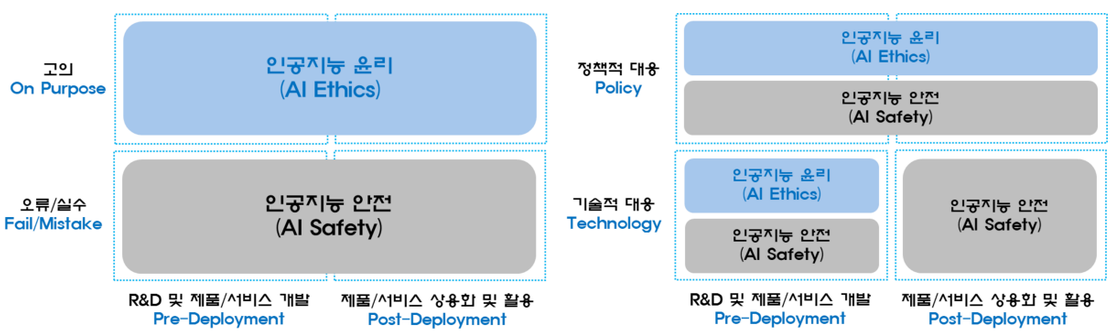




이미지 출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

인공지능의 안전성과 윤리

- 인공지능 안전성과 윤리 개념적 범위



출처 : 과학기술정책연구원 인공지능 기술 전망과 혁신정책 방향(2차년도) 보고서(2019.12)

다음시간에 배울 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티처블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 윤리적 쟁점과 다양한 이슈	1. 자율시스템으로써의 인공지능과 딥러닝 2. 인공지능 안전성과 윤리 3. 인공지능의 윤리적 쟁점 (자율주행자동차, AI로봇, 트랜스휴먼, 프라이버시 문제)	1. 비윤리적 데이터 생성과 수집 - 웹 스크레핑(크롤링)을 이용한 데이터 수집
7		기말고사	