SNU AI 연구원 2021 봄학기 콜로퀴움 제1강 '인공지능 시대, 인공지능 윤리' - SNU 법학전문대학원 고학수 교수 Note

> 2021-3-11. 서울대학교 AI 연구원 17:00 - 19:00 콜로퀴움 Youtube 온라인 강의를 듣고 정리한 Rough임.
>
> 강의자: 고학수 교수 (서울대학교 법학전문대학원)
> 강의 주제: 인공지능 시대, 인공지능 윤리

Motivation

대한민국에서는 연초에 '이루다 사태'가 터지면서 인공지능 윤리에 대한 사회적 관심이 부상. 하지만 이 '인공지능 윤리'가 학계에서 등장한지는 얼마 되지 않았다.

'인공지능 윤리'에서의 '윤리'란 철학/사회적 의미, 법적 의미, 공학적 의미 등 다양한 의미가 결부된 용어임. 단순한 차원으로 이해해서는 안되는 용어임. (현재 학계의 '인공지능 윤리'에 대한 시각)

이 강의에서는 고학수 교수님 주변의 인공지능 윤리에 관한 일부를 보기로 하자.

도처에서 나타나는 인공지능 윤리와 규범의 문제

학계에서는 2017년 이후로 하여 인공지능과 관련된 '윤리'가 학회에 급부상하였음. 즉, 2017년 이후부터 '인공지능 윤리'가 급부상하였음.

경제학, 윤리학, 철학, 공학적으로 이 '인공지능 윤리'에 관한 다양한 접근을 하고 있는 중임. 특히, 3명의 컴퓨터 공학 박사가 'Fairness and machine learning'이라는 초고를 공동 집필하기도.

EU에서도, 그리고 우리나라 정부에서도 인공지능의 윤리 및 이후 정책을 고민하는 중.

- EU에서는 2018년에 인공지능 전문가 그룹을 만들어, 공식적으로 AI 윤리에 관한 Publication을 냄. -> 인공지능의 윤리의 논의 이전에 '윤리'에 대한 엄밀성 논의 필요를 말함.
- 우리나라에서는 최근 과학기술부 주관으로 인공지능 윤리 기준을 발표.
- CVPR (비전 영역에 있어 잘 알려진 인공지능 학술 대회), 그리고 IEEE(국제 공학 표준 기구)에서도 윤리와 관련된 문제를 최근 화두로 제시하고 있음. IEEE는 P7000이라는 AI 윤리에 관한 표준안을 이미 발표하였음.

AI 윤리는 현재 추상적 차원이 아닌, 보다 명확하고 객관적인 차원에서 논의되고 있음.

'이루다'를 어떻게 바라보아야 하는가?

- 이루다 사건은 인공지능과 생활이 연계되는 경우, 우리가 윤리적 문제 등 사회적 규범적 차원에서 생각해야 할 여러가지 문제를 제시하였음.
- 국립국어원에서는 이 사건의 여파로 '모두의 말뭉치'의 다운로드를 막아버림.
- 그러면 이제 이러한 '말뭉치'를 어떻게 정제 작업을 해야 하는가?

자연어처리(NLP)에서의 윤리

- 자연어처리에서의 편향에 관한 논문이 쏟아져나오고 있으며, 이들 논문을 정리한 문헌정리(survey)가 나오고 있는 추세이기도.

검색 및 추천 알고리즘: 아마존 - 의 biased 추천 알고리즘

- 아마존에서도 약 1년 전에 검색 알고리즘에서 자사 제품을 보다 더 추천하도록 하는 편향된 알고리즘을 사용했다가 발견되어 월 스트리트 저널 등에 대서특필됨.
- 아마존 추천 알고리즘 이외에도, 유튜브, 카카오택시, 네이버쇼핑 '네이버페이', 구글 자사 쇼핑 서비스 검색 서비스 노출 등과 관련된 AI 알고리즘의 편향성에 대한 문제 제기가 계속 이어지고 있음.

Adaptive clothing과 편향

- Adaptive clothing: 장애인용 패션 용품을 기존 SNS의 이미지 알고리즘이 '장애인용 용품'으로 구분할지, '패션'으로 구분할지 헷갈려서 Flagging을 계속 해버림. -> 계속 수동적 관리를 해 주어야 해서, 회사 입장에서 Resource가 추가 소모됨.

인공지능을 활용한 의사 결정 사례와 이슈

1. 신용 평가의 경우와 인공지능

- 원래 개인 신용의 조회는 금융 영역의 데이터를 모아 조회한 다음, 이를 이용해 평가하는 방식이었는데, 현재는 금융 영역 이외에도 비금융 영역에서도 정보가 추가되어 평가하고 있음.
- 또 이러한 평가 정보는 금융 영역 외에도 활용되고 있음.
- 과거는 중간 정보 수집 기관이 금융 기관에 한정되었으나, 데이터 기업의 추가 발달 등으로 인하여 금융 영역 이외의 사업자가 데이터 수집에 가세.
- 미국에서는 최근 신용 평가에, 금융 외 정보 (렌트 이력, 전화 및 Utility 이용력, 고용력, 부동산 소유 이력, SNS 이력, 웹 페이지 조회 이력 등)가 이용되기 시작 *나의 우려: 개인 프라이버시 침해 아닌가, 이 정도면?*
- 관련된 하나의 에피소드
    - 애플에서 1년 ~ 2년 전 즈음에 신용카드를 출시했음. 골드만 삭스와 제휴를 해서. Apple Card가 나온 다음에, 유력 인사 한 명이 트위터에 자기 아내와 자신이 신용 카드를 같이 발급 받았는데, 두 사람이 경제 활동의 대부분을 같이 했음에도 불구하고, 아내에 비하여 자신이 20배 넘는 한도를 판단받았다고, 뭔가 차별적 요소가 있지 않느냐고 이의 제기
    - 골드만 삭스 측에서는 데이터 훈련에 성적, 인종, 연령, 성적 지향 등의 정보를 사용하지는 않는다고 하기는 하던데.
- 중국: 몇년 전 부터 '사회적 신용(Social Credit)' 시스템이 적용됨.
    - 사람들의 사회 생활에 관한 점수 부여라고 생각하면 되는데 *나의 생각: 진짜 망할 개인정보 침해인 것 같은데? 디스토피아 아니야?*
    - 선한 행동을 하면 사회적 신용 점수가 오르고, 악한 행동을 하면 사회적 신용 점수가 낮아짐. 주변 사람들의 점수가 좋은 경우 자신의 점수도 올라가는(?) *나의 생각: 이거 괜찮은 거 맞냐?*
    - 이 제도에 대한 평이 엇갈리고 있음. 중국은 신용 평가 시스템을 겨우 만들어냈기 때문에 호평이 있지만, 이러한 것을 통해 사회적 통제 및 감시, 끼리끼리 모이게 만드는 것 아니냐(인적 네트워크의 island화)는 비판이 있음 (엇갈리는 평가)
- 우리나라의 개인 신용 평점에 관해 추가적으로:
    - 학력, 신용조회 이력정보는 신용 평점에 반영이 안 된다고 하는데...
    - 학력을 안 본다는 것은 그러면 학력과 대출을 갚지 않을 확률 등이 아예 통계적으로 관련이 없다는 것을 지시하는가...?
    - 근데 과거에는 고려가 되었다. 그러면 왜 지금은 고려하지 않는가? 다른 방식이 존재하기 때문!
        - 자신의 신용 정보를 잘 찾아보는 사람의 경우는, 자신의 신용 등급이 불안한 요소가 있거나, 조만간 거래할 것으로 추정되기 때문에, 이것을 부정적 정보로 고려하기도 했음.
        - 반론: 신용 정보의 잘못된 정보를 찾아보기 어렵다 (찾아보는 것이 부정 평가라면)...

    - **인공 지능의 학습 데이터는 뭐를 넣어야 하는가? 라는 질문으로 이어지는 것.** (알고리즘 이외에도, 학습 데이터에 관한 윤리적 문제가 제기되는 것)

2. 인공지능 면접

- 1970 ~ 1980년대 미국에서 오케스트라에 블라인드 면접을 실시하여, 음악 소리만 듣고 선발하도록 했더니 기존 5%에 달하던 여성 단원 비율이 크게 증가함.
    - 오케스트라에서 '소리만' 듣고 선발하면, '다른 단원들과의 협업'과 같은 요소를 평가하기 힘들다는 입장이 있음.
- 회사에서의 '채용 및 선발'은 '능력' 이외에도 '커뮤니케이션, 공존 역량, 리더십' 등 다양한 요소를 고려해야 한다는 점에서 상당히 어려운 부분 (따라서 인공지능이 면접에 투입되기에는 다소 어려움...)
- 몇 년 전 아마존에서 '인공지능 고용' 인공지능 팀을 해체함. 학습 데이터와 관련된 문제가 제일 어려움. 회사 내 채용 여성 비율이 너무 낮았다는 문제 등으로 인하여, 알고리즘도 성적 편향으로 이어질 것 같아서...
- 우리나라에서도 인공지능 면접이 발달하고 있는데, 과연 주장대로 '편견과 차별이 없는지' 고민해야 할 것.
- 인공지능의 업무 성과 예측 등 다양한 분야에서는, 편향되었을 수도 있는 과거 데이터를 기반으로 알고리즘이 작성되므로, 편향 부분에 대해 고민해야 할 것.

알고리즘의 공정성과 사회적 문제

- 알고리즘의 분류 문제(classification)과 차별(discrimination)
- 인공지능의 활용 과정에서, 그룹별 분류에 따라 차등적인 결과가 나타난다면, 사회적 차별에 관한 논쟁이 될 것.
- '직접 차별'의 경우는 학습 데이터의 구축 문제, 그리고 모델링의 통제 과정에 관한 문제가 될 것이며, '간접 차별'의 경우는 결과값을 보고 이것의 해석에 있어서에 관한 문제가 될 것이다.
- 그렇다고 학습 데이터를 규율하는 경우, 분석 데이터가 줄어들기 때문에 분석 결과의 정확도, 신뢰도가 하락해버리고, 세분류 그룹별로 나누어 보기 시작하는 경우, 그룹별로 파악하지 못한 특징으로 이익/불이익이 달리 나타날 수 있음.
- 결과값 규율: 어떤 결과값에 대하여 '문제' 상황으로 판단할지, 그리고 그 경우 문제의 '해결책'을 어떻게 찾을 것인지 문제이다... (결과 해석에 관한 애매함, 그리고 결과가 던지는 문제의 해결책에 관한 애매함)
- 인공지능에 의한 차별은 사람에 의한 차별보다는 파악하기 쉬운 편이 많음.
    - 사람의 경우는 사람의 마음을 알 수가 없으니... 입증도 어렵고 내면 심리 파악도 어려움.
    - 인공지능은 그 알고리즘, 인풋을 알 수 있으므로 비교적 차별에 관하여 파악하기는 쉬움.
- 인공지능 차별의 원인: 각각의 모든 인공지능 모델을 만들어내는 과정에서 차별의 위협이 존재.
    - 알고리즘 작성, 데이터베이스화, 실제 인풋 데이터의 결정, 해석 등의 전반에서...
    - 어떤 데이터를 선택하느냐, 어떤 대체 데이터를 선택하느냐, 그리고 누락된 데이터... 등에 의하여 다양한 biased 요소가 발생 가능함.
- 실제 공학계에서의 인공지능 구현과, 실제 사회 정책 결정자 / 윤리 결정자에서의 Ping-Pong을 막으려면, 인공지능 윤리에 관한 협업이 중요할 것.

인공지능이 윤리적일 수 있을까?

- 국내외적으로 인공지능의 윤리에 관한 다양한 규범이 제시되고 있음.
- 인공지능 윤리에 관해 논할 때, 공정성, 투명성, 책무성, 강건성, 프라이버시 등 주요 제시되는 키워드들이 존재. 이들 키워드의 개념을 구체화하고, 어떻게 적용할 것인가 등의 논의 필요
- 트롤리 딜레마 너머, 보다 근본적인 '윤리적 문제'에 관한 논의가 필요할 것.

- 인공지능의 적용에 관한 다양한 시각의 존재: 인공지능에게 모든 것을 믿고 맡겨도 된다. 사람의 개입이 그래도 필요하다!
    - 센싱 오류, 알고리즘 오류 등 다양한 알고리즘 상의 오류로 인하여 항공 등 교통 사고 등에 인명 사고가 발생하기도 했음 (보잉 737 Max의 경우 알고리즘 동작 오류로 대형 인명 사고)
- 인공지능에 과연 모든 것을 맡겨도 될까? 아니면 그럼에도 불구하고 인간의 개입이 필요한가?