윤리적 침해 대응을 위한 인공지능 쟁점 및 데이터 학습 모델 분석*

박 범 진** . 이 상 화*** . 정 영 주****

【목차】

【국문초록】

본 연구는 유럽 연합의 윤리 권고문의 가이드라인 7가지를 토대로 4차 산업혁명과 인공지능의 발전의 결과로 2020년까지 나타난 대표적인 침해 사항을 분석하였다. 사회적으로 큰 영향력을 가지는 인공지능 채용의 윤리 권고문 침해 내용 및 원인에 대한 분석을 시작으로, 윤리적인 인공지능에 대한 논의를 위해 인공지능 욕설학습과 '차별적 학습' 사례를 유럽연합 윤리권고안의 관점에서 분석해보았다.

본 연구는 윤리 권고안을 수행하기 위해 단순히 인공지능 자체에 대한 윤리 규범을 넘어서 인공지능 모델의 개발과정에 대한 윤리적 규범의 필요성을 제시하였다. 인공지능을 개발하는 과정에서 세분화를 통해 해당 부분에서 발생하는 윤리적 침해 가능성을 파악하였고 이에 대한 예방 방안을 제시하였다.

* 이 글은 중앙대학교 인문콘텐츠연구소 HK+사업단에서개최한 '2020 제3회 인문페스티벌'에서 인공지능의 쟁점을 분석하기 위하여 작성되었습니다.
** 중앙대학교 자연과학대학 수학과 e-mail: beomjin26@naver.com
*** 중앙대학교 경영경제대학 응용통계학과 e-mail: tkdghk987@gmail.com
**** 중앙대학교 경영경제대학 산업보안학 e-mail: ojoo_o@naver.com

Ⅰ. 서 론

이제 인공지능은 영화나 소설에서나 나오는 먼 미래의 기술이 아니다. 4차 산업혁명 시대에서 인공지능이 사람과 비슷한 수준으로 오를 날은 얼마 남지 않았다. 스마트 스피커 사례만 보더라도 사람의 말을 대부분 해석할 수 있으며 물음에 대한 합리적인 반응이 가능하다. 인공지능 기술의 발전 속도가 과거 그 어느 시점보다 빠른 현재, 인공지능과 관련된 세부적인 윤리 지침이 필요하다. 본고에서는 2019년 유럽연합(EU)에서 제시한 윤리 지침을 상용화된 AI가 침해할 수 있는지에 대한 여부를 파악하고 해결책을 제시한다. 또한 보다 실용적인 해결 방안을 도출하기 위하여 인공지능이 생성되는 원리를 파악하고 세부 항목에 대한 윤리적 규제를 제시한다.

Ⅱ. 유럽연합(EU)의 AI 윤리 가이드라인

2019년 4월, 세계 최초로 국제기구 차원에서 신뢰할 수 있는 인공 지능에 대한 윤리지침이 나왔다. EU집행위원회는 독립적인 인공지능 고위레벨 전문가 그룹(AI HLEG)을 통해 인공지능에 관한 다양한 의견을 종합한 뒤, '신뢰할 수 있는 인공지능'을 알리기 위한 목적으로 다음과 같은 7가지 윤리지침을 발표하였다.

(1) 인간의 통제가능성 : 인공 지능 시스템은 인간기관 및 기본권을 지원하고 인간의 자율성을 감소, 제한 또는 오도하지 않고 평등한 사회를 가능하게 해야 한다.

(2) 견고성 및 안전성 : 신뢰할 수 있는 인공지능은 AI 시스템의 모든 라이프 사이클 단계에서 오류 또는 불일치를 처리 할 수 있을 정도로 알고리즘이 안전하고 신뢰할 수 있으며 견고해야 한다.

(3) 개인 정보 보호 및 데이터 거버넌스 : 시민은 자신의 데이터를 완전히 제어해야 하며 관련 데이터는 해를 입히거나 차별하는 데 사용하지 않는다.

(4) 투명성 : AI 시스템의 설명가능성과 추적성을 보장해야 한다.

(5) 다양성, 차별 금지 및 공정성 : AI 시스템은 모든 범위의 인간 능력, 기술 및 요구 사항을 고려하고 접근성을 보장해야 한다.

(6). 사회 및 환경 복지 : AI 시스템은 긍정적인 사회적 변화를 강화하고 지속 가능성 및 생태적 책임을 강화하는 데 사용해야 한다.

(7). 책임성 : AI 시스템과 그 결과에 대한 책임과 책임을 보장하기 위한 메커니즘을 마련해야 한다.

EU 집행위는 7가지 핵심 요구사항을 통해 AI는 “(1) 합법적이어야 하며 모든 관련 법규를 준수해야 한다. (2) 윤리적이어야 하며 윤리적 원칙과 가치를 준수해야 한다. (3)좋은 의도로 AI 시스템이 의도하지 않은 결과를 초래할 수 있기 때문에 기술 및 사회적 관점 모두에서 견고해야 한다.” 의 3가지 요소를 준수해야함을 요구하였다.

또한 AI 시스템의 개발, 배치 및 사용과정에서 7가지 윤리 가이드라인의 충족여부를 확인해 AI의 신뢰성을 확보해야 한다고 주장한다. 나아가 인공지능 윤리규범에 관한 지속적인 국제적 논의를 통해 윤리적이고 신뢰할 수 있는 AI를 위한 범세계적 규범을 제창할 것을 촉구한다.

Ⅲ. 사회적 인공지능 쟁점

- AI 채용 프로세스 -

1. 문제제기

기업은 인적 자원의 지속성을 확보하기 위해 끊임없이 채용 전략을 개선하고 있다. 최근에는 채용 과정에서 발생하는 비용을 줄이고 객관적인 평가를 지향하기 위하여 인공지능을 활용한 채용 프로세스를 도입하는 기업들이 늘어나는 추세이다. 국내의 경우 대표적으로 마이다스아이티 (MIDASIT)의 AI 평가 솔루션(inAIR) 등의 플랫폼을 예로 들 수 있다.

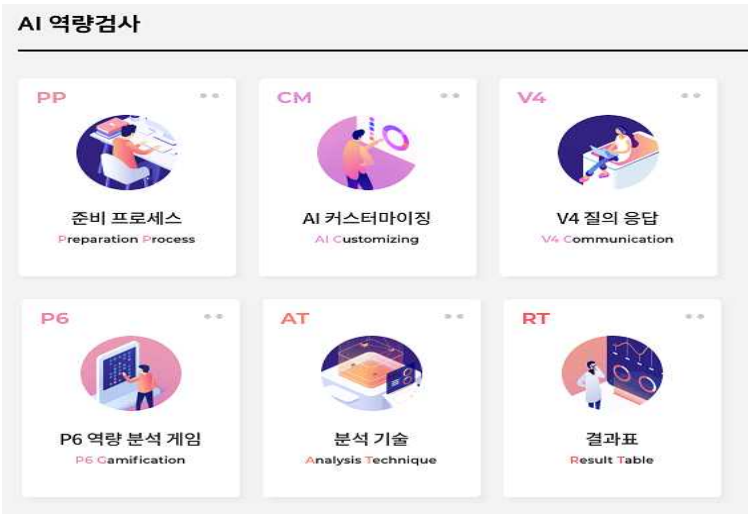

그림 1 마이다스아이티(MIDASIT)의 inAIR 기능 (출처: MIDASIT HRi)

AI 채용 프로세스를 도입할 경우 온라인 환경에서 지원자와 소통하며 지원자의 생체신호, 답변 데이터들을 분석하고 기업의 인재 상에 맞는 사람인지 평가할 수 있다. 또, 지원자의 표정이나 답변 간격 등의 요소 등을 이용해 진실성을 판단하기도 한다. AI 채용 프로세스는 지원 가능 인원을 늘려 보다 많은 지원자들에게 기회를 제공할 수 있고 인재 채용의 질을 높인

다는 점에서 높은 평가를 받고 있다.

반면 AI 채용 프로세스 도입에 대한 부정적인 평가도 존재한다. 채용기준이 획일화 될 수 있고, 해당 기준에 맞는 모범답안이 유출되어 정당한 평가를 받을 수 없다는 것이 그 이유이다. AI를 활용한 채용 프로세스가 지속 가능하고 효율적인 채용 방식으로 자리 잡기 위해서는 윤리적 침해 사항의 원인이 되는 문제점을 분석할 필요가 있다.

2. AI 채용 프로세스의 윤리적 침해 사항

(1) 비차별성 및 다양성

AI 채용 프로세스에서 특정한 조건을 가질 때 좋은 결과를 얻도록 편향 되어있는 학습데이터를 활용하는 경우, 편견이 담긴 패턴을 학습하게 되어 차별을 유도할 수 있다. 실제로 세계 최대 전자상거래업체인 아마존이 여성 차별 문제를 야기한 AI 채용 시스템을 폐기한 사례가 있다. IT 산업에서 남성이 지배적이었던 지난 10년간의 이력을 학습하면서 지원자가 '여성'임을 인식했을 때 해당 지원자의 경력을 평가절하하게 된 것으로 보인다. 이처럼 기업 내 여성 인력에 대한 과거 데이터가 부족한 경우 여성지원자를 입사 범주에 포함시키지 않아 채용 과정에서 여성이 배제되고 다양성이 확보되지 않을 수 있다. 아이작 아시모프(Isaac Asimov)가 제시한 '로봇 3원칙'에 따르면 로봇은 작위 또는 부작위로 인간에게 해를 입혀서는 안 된다[2]. 즉, 인간의 자유와 권리를 침해해서는 안 된다. 하지만 인공지능 채용 방식이 도입되고, 편향된 학습을 통해 채용 결과가 산출됨으로써 지원자들의 성별, 인종 등의 요인에 좌우되지 않고 공정하게 평가받을 권리를 침해할 수 있다.

(2) 책임성

전통적인 윤리학에서 책임은 자유 의지를 갖는 도덕적 행위자에게 귀속시킬 수 있는 개념이다. 인공지능은 도덕적 행위자로 보지 않기 때문에 책임 귀속의 대상이 되지 않는다. 따라서 인공지능에 의해 야기된 차별은 인공지능의 '행위'에 의한 것이라기보다 인공지능의 '작동' 및 그 과정의 완료에서 나타나는 차별적 귀결이라는 표현이 더욱 적절할 것이다. (허유선, 2018)

하지만 그 '작동'의 결과로 인해 차별 및 편견의 체계가 확장될 가능성이 있고, 피해를 입은 지원자가 실질적으로 존재한다는 점이 문제가 된다. 이러한 상황에서 AI 채용 알고리즘의 설계자 혹은 해당 인사 업무와 관련이 있는 인간 행위자를 찾아 책임을 귀속시키는 방법이 대두된다. 그러나 어느 정도의 연관성을 가진 인간 행위자를 찾아야 하는지, 또 어떻게 책임을 분배해야 하는지에 대한 불분명성이 존재한다. 인공지능은 스스로를 발전시키며 학습하는데, 그 과정을 인간이 전부 파악할 수 없고 해당 과정을 통제할 능력이 충분하다고 말할 수 없기 때문이다.

(3) 투명성

AI 채용 프로세스를 둘러싼 가장 큰 우려 중 하나는 인공지능 모델이 블랙박스 형태로 작

동한다는 점이다. 알고리즘 산출 결과가 공정하지 않은 것이 기존 데이터의 문제일 수도 있지만, 프로그래머가 의도적으로 편향성을 지닌 결과를 도출하도록 코드를 작성했을 가능성을 배제할 수 없다. 따라서 알고리즘이 내린 의사결정으로 개인의 권리가 침해 받을 경우 그에 대한 설명을 요구할 수 있는 설명권, 즉 예측에 사용된 방법에 대한 투명성을 보장해야한다는 주장이 대두되고 있다.

하지만 투명성을 보장하는 과정에서 새로운 부작용이 발생할 가능성도 존재한다. 첫 번째로, 공정하고 객관적인 평가를 지향한다는 AI 채용 방식의 도입 목표가 훼손될 수 있다. AI 채용 알고리즘의 작동 방식을 공개하면 자연히 합격을 좌우하는 요인이 노출된다. 이에 따라 본인의 성격 및 장단점을 해당 기업의 인재상에 맞출 수 있도록 조작하여 합격하는 사례가 생길 수 있다. 공정성을 위해 도입한 AI 면접이 인력의 질을 향상시키지 못하고 오히려 객관성을 잃게 만드는 상황을 초래하는 것이다. 두 번째로, 알고리즘의 작동원리에 대한 정보가 노출되면서 기술의 취약점이 드러날 수 있다. 공격자가 취약점을 이용해 허위 데이터를 주입하는 등의 공격을 수행한다면 인사채용 과정 전반에 큰 문제를 발생시키게 된다. 마지막으로 지적 재산권에 대한 절취가 발생할 수 있다. 훈련 데이터세트와 모델의 작동원리 대해 설명하는 과정에서 알고리즘에 대한 정보가 공개되는데, 제3자가 이를 재구성하여 비슷한 알고리즘을 만들 수 있다는 위험이 존재한다.

이와 같이 투명성을 보장하고 개인의 권리를 보호하기 위해 알고리즘에 대한 설명을 제공하지만, 그에 따른 결과로 또 다른 개인의 권리가 침해 받을 수 있다는 모순이 발생하게 된다. AI 면접 프로세스의 신뢰성을 확보하기 위해, 그리고 지원자 및 개발자의 권리를 보호하기 위해 균형 있는 대안이 필요하다.

(4) 개인정보 보호

채용 과정에 인공지능이 도입되면서 지원자의 이름, 주소와 같은 기본 개인정보와 더불어 목소리, 표정, 심장박동 등의 생체데이터까지 수집하게 되었다. 또한, 인공지능 면접을 위해 외부 업체의 프로그램을 사용하는 사례가 늘어나고, 그로 인해 개인정보 유출에 대한 위험도가 높아지고 있다. 제3자가 인공지능 모델의 학습 데이터 세트에 특정 개인의 데이터가 포함되어 있는지 무단으로 확인하는 추론 공격 가능성도 무시할 수 없다. 기존 데이터를 기반으로 새로운 가치를 창출해내는 인공지능의 특성을 고려하여, 수집한 개인 데이터를 어떻게 관리할 것인지에 대한 논의가 필요해 보인다.

Ⅳ. 윤리적 인공지능 쟁점

1. 문제제기

인공지능 기술이 발전함에 따라 현대사회의 인간은 기존과는 비교할 수 없는 풍요를 누리고 있지만, 동시에 기존에 겪어본 적 없는 새로운 문제들이 생겨나는 위험사회로 진입하고 있다. 본 장에선 EU의 인공지능 윤리지침 7가지 권고안을 바탕으로 최근 대두되는 인공지능의 '욕설 학습' 및 '차별 유도적 학습' 위험사회의 사례들을 윤리척 측면에서 분석하고자 한다.

1.1 챗봇(Chatbot)

챗봇 서비스란 기존에 인간이 하던 민원대응 및 고객 상담의 영역에서 인공지능 로봇을 도입해 대신하는 것을 의미한다. 자연언어처리 기술을 이용하거나, 사용자의 입력에 대한 동작과 각본에 있는 응답을 출력하는 각본 방식이 있다. 행정 효율화와 비용 절감 등을 명목으로 각국 정부와 많은 기업들에서 챗봇의 도입이 활발해지고 있다. 다만 챗봇이 인간의 대화 데이터들을 학습할수록 비속어나 극단적인 표현 등을 무분별하게 남발할 수 있다는 문제점도 상존한다. 대표적 사례로 2016년 5월 MicroSoft사에서 개발한 인공지능 챗봇 'Tay'를 들 수 있다.

챗봇에서 이처럼 부적절한 언어사용의 문제가 발생하는 원인은 다음과 같다. 첫째. 인공지능의 경우 기존의 컴퓨터 알고리즘과 달리 모델의 개발이 사업의 시작이다. 개발된 모델의 학습과정 역시 인공지능 모델에서 중요한 단계인데, 이 과정에서의 학습데이터에 내재된 차별과 편견적인 요인으로 인해 문제가 발생한다.

둘째. 개발과정에서 학습하는 데이터 자체가 편견이 가득하거나 혹은 학습할 데이터의 절대량 부족으로 똑똑하지 않은 챗봇, 즉 인공지능 윤리를 위반하는 모델이 탄생한다.

셋째. 인공지능 개발자의 윤리적 책임성에 대한 이해가 부족할 경우에도 이러한 문제가 발생한다. 제품 출시 전 다양한 조건에서 모델이 훈련을 거쳤더라도, 인공지능의 특성 상 출시 후 위험한 상황에 노출될 가능성을 배제할 수 없다. 이는 개발과정에서 충분히 예측할 수 있는 것이지만 관련 대책이 부재할 경우 위와 같은 사례가 발생한다.

1.2 차별 유도적 학습 - 노스포인트사(社)의 컴퍼스(COMPUS)

'COMPUS'는 범죄자의 재범률을 측정하는데 사용하는 알고리즘이다. 피고인의 재범 가능성을 risk score 1(매우 낮음) ~10(매우 높음) 으로 측정하여 판사에게 제공한다. 판사는 이를 판결의 보조 자료로 활용해 가석방 결정, 형량 선고 등을 한다. 형사 재판에서 활용되는 인공지능(AI) 알고리즘이 흑인에게 불리한 인종차별적 판단을 내린다는 논란이 발생해 타당성에 의문이 제기되었다. 대표적인 논란은 다음과 같다.

흑인의 범죄발생률이 백인보다 높다는 기본구성비율에 의거, 인종적 요소가 범죄 위험에 대해 더 높은 가산점을 부여해 타 인종보다 흑인의 수감기간이 길어질 경우가 많아진다. 이는 '교도소에 있는 사람들 중 흑인의 비율이 높다는' 것의 새로운 데이터가 되고, 향후 범죄 위험도 예측 결정에서 반영된다. 이러한 양의 피드백(feedback) 효과로 인해 흑인이 타 인종에 비해 더 많이 교도소에 수감되는 결과로 이어진다.

또한 공정성을 위해 알고리즘의 화이트박스(white box)를 추진할 경우, 기업의 지적 재산권 문제에 있어 첨예한 갈등일 발생하고, 인간이 이해 가능한 수준에서 알고리즘을 구현하여 성능 저하 우려가 발생한다. 반면 알고리즘이 성능 향상을 위해 과적합을 방지하는 정규화(regularization), 랜덤화(randomization) 등을 도입할 경우 인간이 알고리즘의 구동과정을 상세히 이해하기 힘들고, 결정 근거에 대해서도 완벽하게 파악할 수 없게 된다.

2. 윤리적 침해 사항

2.1. 챗봇(Chatbot)

(1) 책임성

인공지능 모델 개발 과정에서 개발자의 윤리적 책임감 부재로 위와 같은 문제가 발생할 수 있다. 챗봇의 설계 과정에서 1) 알고리즘에 편견과 왜곡적인 요소가 없는가. 2) 알고리즘 설계 시 실수나 누락 등이 없는가. 3) 개발 이후 학습 과정의 편향에 대한 예측이 부재한가. 4) 학습 데이터의 문제는 없는가. 이 4가지 요소를 적절히 고려하지 않았는지 검증이 필요하다.

(2) 인간의 통제 가능성

인간의 대화를 학습 데이터로 활용하는 과정에서, 비속어 및 폭언 등을 챗봇이 적절히 걸러낼 수 있도록 기획 및 설계를 해야 한다. 또한 부적절한 판단을 할 경우 인간이 알고리즘의 결정과정에 즉시 개입해야 한다.

2.2. 차별 유도적 학습

(1) 투명성

AI 시스템을 만드는 데 사용된 알고리즘과 데이터는 사람이 이해하고 설명할 수 있어야 한다. 하지만 COMPUS 알고리즘의 판단근거는 블랙박스 속에 있어 기술적 안정성과 신뢰성을 확인할 수 없고, 재판 과정에서 피의자가 판결의 근거를 확인할 수 없다.

(2) 다양성, 비차별성, 공정성

AI는 연령, 성별, 인종 등을 차별하지 말아야 한다. 하지만 COMPUS 알고리즘은 이와 관련된 논란이 여전히 진행 중이다. 학습 데이터가 범죄발생률과 인종간의 유의미한 차이를 보이기에 알고리즘은 문제가 없다는 입장과, 특정 인종에게 불리한 피드백이 지속되어 인종차별적이라는 주장이 대립하고 있다.

(3) 책임성 보장

인종차별과 관련한 논란이 지속되고 있지만, COMPUS 알고리즘은 지적 재산권 문제등 여러 복합적인 요인들로 인해 감사가 불가능하다. 또한 알고리즘은 결과만 알려주고 판단 근거는 알 수 없기에, 설명책임(accountability)의 문제가 제기된다.

Ⅴ. 인공지능 모델 개발 과정 분석

1. 데이터 학습

스마트폰의 등장으로 개인으로부터 실시간으로 데이터 수집이 가능해졌다. 또한 컴퓨터 기술의 발달은 대용량 데이터를 저장하고 처리하는 것을 가능하게 했다. 이러한 두 가지 발전은 기업들로부터 새로운 이익 구조를 적용하게 만들었으며 기업은 사용자로부터 데이터를 수집하고 이를 바탕으로 트렌드와 선호도를 분석한다. 분석 결과물을 통해 모델을 만들면 마치 인간처럼 합리적인 의사결정을 하는 인공지능이 만들어진다. 인공지능을 개발에 있어 영향을 끼치는 요소는 학습 데이터와 학습 방법이다. 2020년 현재까지 나온 인공지능은 데이터의 종류와 목표에 따라서 학습 종류가 다르다. 따라서 이번 장에서는 EU에서 세운 7 가지 권고안을 침해할 수 있는 사용자 로그 데이터와 자연어 데이터에 대한 분석을 진행한다.

2. 사용자 로그 데이터

사용자는 스마트폰을 통해서 기록을 남긴다. 기록은 문자 메시지나 인터넷 사용 기록에서부터 위치와 시간 정보까지 다양하다. 스마트폰의 애플리케이션이 데이터를 축적하고 애플리케이션을 관리하는 서버로 보내면 서버에서는 로그(사용자 데이터)를 분석하고 이를 토대로 사용자에 대한 선호도를 분석할 수 있다. 로그 데이터를 정형화하고 인공지능을 학습한다. 인공지능이 만들어지면 이를 계속해서 사람들에게 데이터를 받아오고 이를 통해 인공지능의 성능을 향상시킨다.

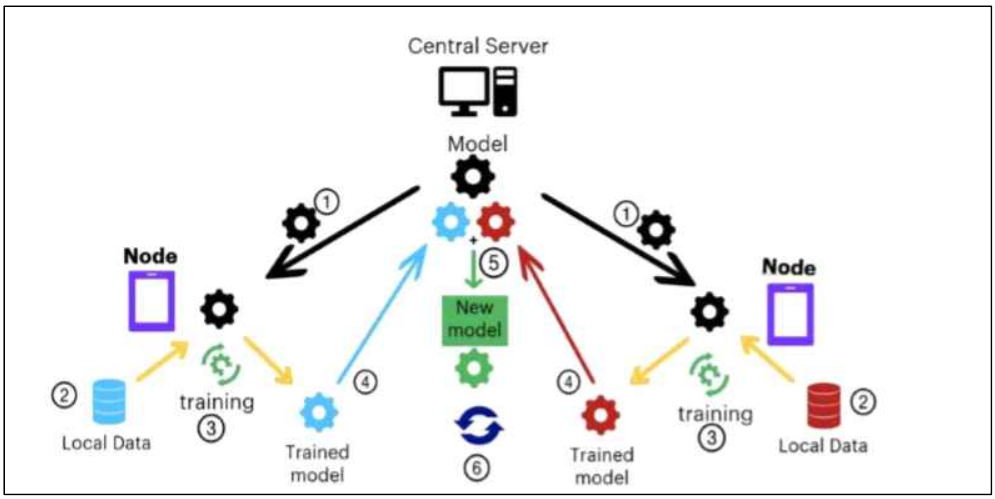

그림 2 사용자 로그 데이터와 모델 간의 관계. [출저 : mc.ai]

2.1 사용자 로그 데이터 학습과정에서 인공지능의 윤리적 침해 사항

(1) 다양성, 비차별성, 공정성

인공지능은 주어진 데이터를 바탕으로 현실세계에 대한 이해 및 해석을 하게 된다. 사용자들로부터 임의의 데이터를 받아서 이를 처리할 경우, 편향적인 인공지능이 만들어지며 의도적으로 불순한 데이터를 사용자가 전송한다면 모델은 불순한 데이터에 의한 학습이 진행 된다.

(2) 투명성

인공지능의 판단에 대한 근거를 제시할 수 있어야 한다. 그러나 현재 모델에 대한 내부적인 판단 기준은 모델을 설계한 인공지능 설계자만이 알고 있다. 그러나 설계 구조는 4차 산업혁명에서 기업의 이익구조에서 가장 중요한 요소이므로 사용자는 내부적 구조에 대한 윤리적 도덕적 판단이 불가능하다.

(3) 책임성

현재는 인공지능의 수준이 사람이 설계한 수준이다. 따라서 비윤리적인 행동에 대한 책임은 설계자에게 전가가 가능하다. 그러나 개발이 진행되고 모델이 복잡해짐에 따라서 설명가능성이 약해지고 설계자가 책임을 질 수 없는 수준까지 넘어갈 수 있다. 이러한 상황에서 인공지능을 단순하게 도구로 본다면 책임을 배제할 수 있다

2.2 사용자 로그 데이터 윤리적 침해 해결 방안

인공지능 설계자는 데이터의 흐름부터 인공지능의 작동원리에 대하여 세부적인 정보를 알고 있다. 따라서 가장 선행되어야 하는 것은 설계자에 대한 윤리적인 책임을 부여하는 것이다. 그러나 설계자에 대한 규제만으로는 앞으로 더욱 복잡해지는 인공지능에 대한 규제가 될 수 없으며 인공지능의 개발 및 처리 과정에서 규제가 들어가야 한다.

(1) 사용자가 데이터를 전송하는 부분

인공지능을 설계하는 부분만큼 중요한 것은 데이터의 종류와 품질이다. 좋은 모델이 있을지라도 올바른 종류의 데이터가 충분하지 않다면 인공지능은 작동하지 않는다. 따라서 비윤리적인 인공지능을 예방하기 위해 사용자로부터 받는 데이터에 대한 규제가 필요하다. 현재는 사용자가 사진, 음성, 위치와 같은 특정 항목에 대해서는 프로그램이 접근하기 전에 사용자에게 동의를 얻는 구조이다. 프로그램이 해당 정보에 대한 접근 권한을 얻으면 데이터에 대한 무제한적인 권한을 취득하며 이에 대한 처리 및 분석에 대한 권리를 가지게 된다. 따라서 데이터의 종류에 따른 전송 규제를 통하여 사회적으로 동의 가능한 인공지능을 만들어야 한다.

(2) 데이터를 학습시키는 부분

데이터를 수집하는 것과 이를 바탕으로 인공지능을 학습하는 것은 다른 문제이다. 예를 들어, '카카오톡'에서 프로필 서비스 제공을 위해서 개인의 사진 데이터를 보관하는 것과 이를 바탕으로 사람의 얼굴을 인식하는 인공지능을 만들기 위해 사용하는 것은 윤리적 범주가 다르다. 사용자로부터 받은 데이터에 대한 접근과 처리는 수집한 기업이나 개인이 학습 을 위해서 사용할 수 있다. 이러한 구조는 4차 산업혁명을 가능하게 만들었지만 비차별성을 더욱 심화시킨다. 학습데이터를 선택할 수 있다는 것은 반대로 특정 데이터를 배제할 수도 있다는 것이다. 인공지능이 편향적일 수 있으며 이에 대한 사례를 이미 언급한 AI면접에서 도 충분히 나올 수 있다. 따라서 학습데이터에 대한 규제는 인공지능이 발전하는 과정에서 반드시 이루어져야 한다.

(3) 인공지능이 판단을 내리는 부분

인공지능은 데이터를 기반으로 학습되며 이를 토대로 현실을 제대로 반영할 수 있다. 그러나 인공지능이 현실을 제대로 반영하고 이를 있는 그대로 사용하게 될 경우 오히려 문제가 발생한다. 예를 들어 '샌프란시스코 범죄 데이터'는 샌프란시스코의 범죄 정보에 대한 위치정보를 제공한다. 인공지능이 이를 바탕으로 학습하고 판단을 내릴 경우, 해당 지역에 대 한 부정적인 판단을 내릴 가능성이 농후하다. 따라서 인공지능의 판단에 대한 규제를 통해 비윤리적인 인공지능을 예방해야 한다.

3. 자연어 데이터

자연어(문자)데이터는 인공지능과 가장 관련이 깊은 데이터이다. 과거에도 자연어 데이터에 대한 중요성은 높았으나 최근 딥러닝 기계학습의 발전으로 자연어 처리(Natural Language Processing)에 대한 발전이 가능해졌다. 그림에서 나타나듯이 자연어 처리에 대한 시장에서 비중은 2017년 대비 2020년에 3배가량 성장했으며 이에 따라 대화 가능한 인공지능의 발전은 더욱 가속도가 붙게 된다. 따라서 인공지능의 사회적 역할을 생각한다면 이에 대한 논의는 반드시 진행되어야 한다. 자연어 데이터에 대한 인공지능 모델 또한 방대한 양의 데이터를 바탕으로 한다. 자연어처리는 2016년 이후로 급속도로 발전하였는데, 자연어 문자들 사이의 복잡성에 대한 해석이 딥러닝 아키텍처의 발전과 더불어 가능해졌기 때문이다. 이 기술은 단순히 문자에 대한 인공지능을 넘어서 음성에 대한 인공지능을 가능하게 만들었으며 말하는 인공지능을 가능하게 만들었다

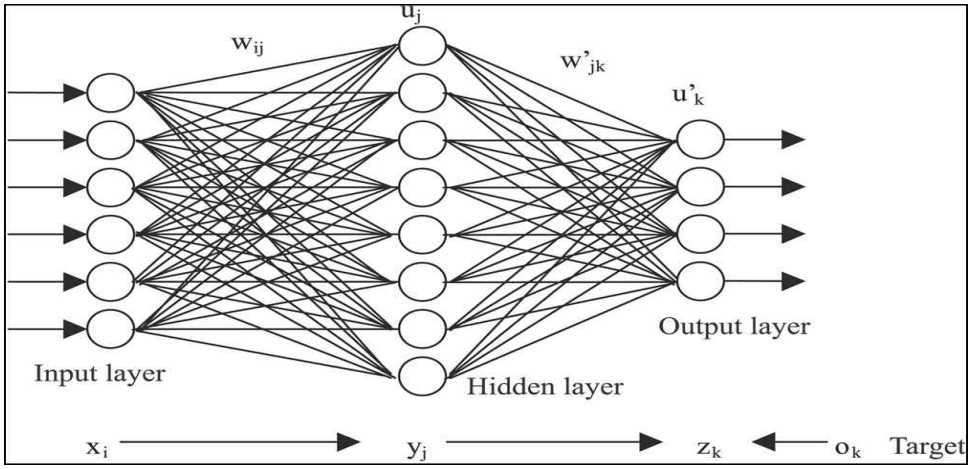

그림 3. 말하는 인공지능을 가능하게 한
딥러닝 인공신경망 구조 [출처: 사이언스 라이프]

3.1 자연어 데이터 학습

인공지능 모델은 신경망을 통해서 언어의 구조에 대한 전반적인 학습을 진행한다. 올바른 문장구조에 대한 패턴을 학습하고 이를 바탕으로 상황에 맞는 패턴을 선택하는 방식으로 사람이 말을 배우는 방식과 유사하다. 따라서 다양한 패턴을 학습할수록 인공지능은 사람과 유사해지며, 인공지능을 만들기 위해서는 많은 글과 음성 데이터를 필요로 한다. IT 기업 구글이 유튜브 프리머엄 고객들에게 무료로 스마트 스피커를 나눠준 것도 이러한 음성 데이터의 중요성을 뒷받침한다. 또한 트위터API는 트윗(짧은 글)에 대한 추출을 가능하게 만들었고 이를 바탕으로 많은 자연어처리에 대한 연구와 기술이 개발되었다. 여기서 더 나아가서 언어에 대한 학습을 진행하면 특수한 목적에 맞춰서 다시 학습이 가능 하다. 이를 전이학습(transfer learning)이라고 하며, 이 기술은 언어에 대해 학습한 인공 지능이 더 나아가서 언어를 활용해서 목적을 달성할 수 있게 만들었다. 목적에 따라서 훈련 된 인공지능은 비윤리적 인공지능으로 훈련될 가능성이 있으며 이에 대한 사회적 통용 가능성과 윤리적 문제에 대한 논의가 필요함은 자명해 보인다.

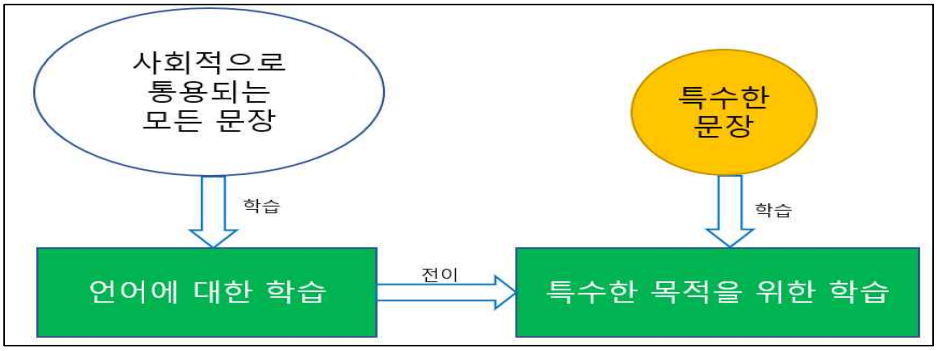

그림 4 전이학습으로 언어적 인공지능을 학습시키는 방법

3.2 자연어 데이터 학습과정에서 윤리적 침해 사항

(1) 다양성, 비차별성, 공정성

동일한 언어라도 문화에 따라서 의미가 다를 수 있다. 자연어 데이터 또한 이러한 언어의 특성을 반영하고 다른 종류 데이터보다 편향에 더 민감하다. 대표적인 예로 앞에서 언급한 욕설 학습이 있다. 예를 들어 친구들 사이에 '바보'라는 단어로 친근함을 표현할 수 있지만 이 단어를 타인에게 대해서 사용했을 때는 그 의미가 부정적일 수 있다. 인공지능은 사람들과의 반응을 통해서 실시간으로 개선되기 때문에, 부정적인 내용에 대한 데이터가 주어지면 사용자에게 반응하는 내용 또한 부정적일 수 있다. 전 세계적으로 인종차별이 큰 이슈가 되는 현재 인공지능이 언어에 대하여 주된 문화를 선택하는 것은 어려운 일이다.

(2) 투명성

말하는 인공지능은 단편적인 대화를 넘어서 그 목적을 지니고 있다. 인공지능과 대화를 하면서 사용자는 인공지능에게 원하는 것을 말하고 인공지능은 이에 대한 반응을 한다. 문제는 이러한 반응이 어떠한 의도를 지니고 있는지 알기 어렵다는 것이다. 2020년 대부분의 스마트 스피커는 사용자로부터 질문이나 요청이 왔을 때 간단하게 답할 수 있는 수준이다. 인공지능의 발전이 진행됨에 따라 인공지능이 선택할 수 있는 의도는 더욱 다양해지고 사용자 는 그 목적이 무엇인지 알기 어렵다. 예를 들어, 휴가 장소를 추천해달라고 했을 때, 질문자에게 가장 적합한 장소를 추천해주는데 그 안에 인공지능 설계자의 이익을 위해서 특정 지역을 추천하는 목적이 있을 가능성도 있다. 이러한 행위는 이익구조로 당연시 여길 수도 있지만, 의도가 윤리적으로나 법적으로 허용되지 못하는 수준일 수도 있다. 그러나 이러한 경우일지라도 사용자는 그 내부적 구조를 알 수 없다

3.2 자연어 데이터 윤리적 침해 해결방안

말하는 인공지능은 합리적인 판단을 내리는 인공지능과 성격이 다르다. 사용자 로그 데이터를 사용하는 인공지능은 합리적 의사결정을 목표로 만들어졌으므로 데이터에 대한 올바른 사용을 바탕으로 윤리적 침해를 막아야 한다. 그러나 말하는 인공지능은 데이터에 처리과정보다 근본적으로 어떠한 언어적 특성이 사회에 통용될 수 있는가를 다뤄야 한다.

비윤리적 인공지능은 문화적인 특성에 따라서 차이가 난다. 따라서 비윤리적인 인공지능을 예방하기 위해 윤리적 가이드라인을 적용하는 방식은 문화에 따라서 차이를 보일 수 있다. 윤리적 인공지능에 대한 일반적인 해결방안은 여러 윤리적 규범들 중에서 투명성을 최우선으로 두는 것이다. 사용자에게 인공지능이 작동하는 방식에 대한 정보를 제공함으로써 사회적으로 동의할 수 없는 인공지능에 대한 내부적 처리절차를 파악할 수 있어야 한다. 투명성은 기업의 이익구조와 관련되어 있으므로 현시점에서 인공지능에 대한 작동원리를 개방하는 것은 쉽지 않을 수 있다. 그러나 언어에 대한 전반적인 학습을 먼저 진행하고 추후 전이학습을 통해서 목적을 달성하는 말하는 인공지능의 특성을 고려한다면, 인공지능의 전이 학습 부분에 대한 일부 투명성을 통해서 인공지능의 목적 부분에 대한 설계 원리를 파악할 수 있도록 만들어야 한다.

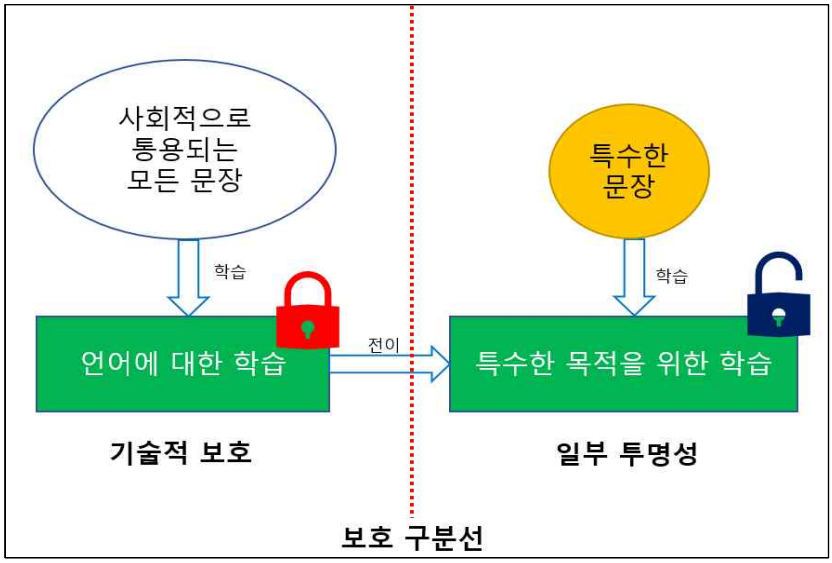

그림 5 언어적 인공지능에 대한 전이학습과 일부 투명성

Ⅵ. 맺음말

인공지능은 다양한 분야에서 성장하며 산업구조를 변화시킬 정도의 상당한 파급력을 보여주고 있다. 이러한 사회적 흐름에 따라 인공지능에 대한 윤리적 문제에 대한 관심도 높아졌다. 보다 객관적이고 정확한 의사결정 결과를 얻기 위해 인공지능 알고리즘을 도입하였지만, 오히려 그 과정에서 편향적이고 비윤리적인 학습이 이루어질 수 있다는 것이 드러났다. 이러한 학습의 결과는 인간존엄성 보호 및 차별금지와 같은 사회적 가치의 훼손으로 이어질 수 있기 때문에 인공지능의 윤리 침해 문제에 대한 대응방안을 마련할 필요가 있다.

본고에서는 윤리 침해 문제를 보다 거시적으로 확인하기 위해 대표적 윤리 침해 예시인 AI 채용, 챗봇, 컴퍼스 사례에 EU 의 윤리 지침을 적용하여 분석하였다. 차별 유도 학습으로 인해 문제가 되는 AI 채용 및 컴퍼스 사례에서는 인공지능이 편견을 가진 데이터를 학습하게 되어 채용 범주에서 여성을 배제하고 흑인의 재범률을 높게 측정하는 등의 편향적 결과가 산출되는 것을 볼 수 있었다. 이는 비차별성 및 공정성 요소에 위반되며, 실질적 피해자가 존재함에도 불구하고 책임을 귀속시키는 기준이 불명확하다는 문제가 발생한다. 또, 투명성을 보장하기 위한 과정에서 추가적인 부작용이 야기되기도 한다. 다음으로 욕설 학습의 예시인 챗봇 사례를 통해 비윤리적이거나 편견적인 발언들을 기존의 개발과정에서 학습한 데이터보다 많이 학습하게 되어 인공지능 알고리즘이 윤리 지침을 위반하는 모델로써 작동할 수도 있다는 사실을 알 수 있었다. 인공지능의 특성을 고려했을 때 출시 후에도 비윤리적인 데이터를 학습하지 않는다고 단언할 수 없고, 따라서 인공지능이 인간이 통제할 수 있는 범위 내에서 작용해야한다는 인간의 통제 가능성 규범을 위반한 것이 된다. 더불어 개발자의 윤리적 책임감 부재도 문제가 될 수 있다.

이러한 인공지능의 윤리 침해 문제를 해결하기 위해 모든 사례에서 공통적으로 문제가 되는 '데이터 학습 과정'에 초점을 맞추어 지켜야하는 윤리적 규범을 작성하였다. 이 때, 단순히 인공지능 자체에 윤리적 규범을 세우기보다는 데이터의 종류를 사용자 로그 데이터와 자연어 데

이터로 분류하고, 인공지능의 생성부터 사용까지의 과정을 사용자가 데이터를 전송하는 단계, 데이터를 학습시키는 단계, 인공지능이 판단을 내리는 단계로 세분화하여 대응 방안을 도출하였다. 첫 번째로 사용자 로그 데이터의 경우, 사용자가 데이터를 전송하는 단계에서 정보에 대한 무제한적인 권한을 취득하는 것을 방지하기 위해 사용자로부터 받는 데이터의 종류 및 내용에 대한 규제가 필요하다. 또한, 데이터를 학습시키는 단계에서는 사용자로부터 받은 데이터와 학습할 수 있는 데이터를 적절히 분리하여 특정 데이터를 배제하지 않고도 차별적 결과를 줄일 수 있도록 균형을 맞출 필요가 있다. 인공지능이 판단을 내리는 단계에서는 인공지능이 현실 반영 데이터를 학습하면서 특정 지역이나 인물 및 사건에 대해 절대적으로 부정적인 판단을 내릴 가능성이 농후하기 때문에 이에 대응하는 규제를 마련해야 한다. 두 번째로 자연어 데이터의 경우, 합리적 의사결정을 목표로 두는 사용자 로그 데이터와는 다르게 학습 과정 보다는 근본적으로 어떤 언어적 특성이 사회적으로 통용될 수 있는지에 대한 논의가 필요하다. 또한 전이 학습 부분에 대한 투명성을 보장하여 인공지능의 목적에 대한 설계 원리를 파악할 수 있어야 한다.

알고리즘 및 자연어 처리 기술의 발달로 인공지능과 인간 간 유사성이 점차 높아지고 있다. 이러한 상황 속에서 윤리적 규범이 기술의 발전 속도를 따라가지 못할 경우, 인공지능에 대한 거부감을 높이고 문화지체현상을 발생시키는 등의 사회적 비용을 야기하게 된다. 또, 세계적인 영향력을 가지고 있는 EU 의 가이드라인과 미국 백악관의 인공지능 규제 원칙이 존재하긴 하지만 이러한 지침들과는 별개로 인공지능기술 사용 시 유의사항을 개별적으로 지정하는 경우가 많다. 따라서 투명성 보장을 위한 영업비밀의 범위는 어디까지 인정할 것인지, 책임성의 소재와 범위는 어떻게 판단할 것인지 등에 대한 내용은 규범 차원이 아니라 표준화 및 법제화하여 그에 따른 법익을 보호할 필요가 있다. 아울러, 무분별한 인공지능의 발달로 인해 발생하는 윤리 침해 문제에 대해 사회 구성원 모두가 관심을 가지고 적극적으로 대응해야 한다.

【참고문헌】

[1] Stefan Larsson, 2019, Transparency in artificial intelligence, Internet Police Review

[2] 양천수, 2020, 인공지능과 윤리 -법철학의 관점에서-,「Chosun Law Journal Vol.27 No.1」

[3] 허유선, 2018, 인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰 -알고리즘의 편향성 학습과 인간 행위자를 중심으로-, 한국여성철학회

[4] Tom van Nuenen, 2020, Transparency for whom? Assessing discriminatory AI, King's College London

[5] 이환우, 2019, 채용 전형에서 인공지능 기술 도입이 입사 지원의도에 미치는 영향,「정보시스템연구」

[6] Stephanie Kelley, 2019, Discriminatory Artificial Intelligence in Organizations: Causes and Prevention Methods, Queen's University

[7] Carmen Fernández, 2019, AI in Recruiting. Multi-agent Systems Architecture for Ethical and Legal Auditing, King Juan Carlos University

[8] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." Journal of machine learning research 12.ARTICLE (2011): 2493-2537.

[9] 윤상오, (2018). 인공지능 기반 공공서비스의 주요 쟁점에 관한 연구 - 챗봇(Chatbot) 서비스를 중심으로-, 한국공공관리학보.

[10] Nelson, Marilyn M., and William T. Illingworth. "A practical guide to neural nets." (1991).

[11] Wolf,M,J, Miller,K,W, Grodzinsky,F,S, Why We Should Have Seen That Coming – Comments on Microsoft's Tay "Experiment", and Wider Implications, 2017.

[12] 홍성욱, 2018, AI Algorithm and Discrimination, 과학기술정책연구원.

[13] European Comission. 2019. Ethics guidelines for trustworthy AI

주 제 어 인공지능, 인공지능 개발, 인공지능 학습, 윤리 권고안, 인공지능 윤리, 인공지능 사회

Keywords Artificial Intelligence, Artificial Intelligence Development, Artificial intelligence train, Ethics Guidelines, Artificial intelligence Ethics, Artificial Intelligence Society