

기획특집

인공지능(AI) 기술 발전과 성인지적 대응 전략

• 인공지능(AI)의 젠더편향 완화를 위한 법제화 전략

김일우 | 한국청소년정책연구원 부연구위원

• 공정성을 넘어: AI 채용도구 형평성 제고전략

이지은 | 연세대 문화인류학과 부교수

신민정 | 연세대 문화인류학과 박사과정

신지연 | 연세대 문화인류학과 박사과정

• 딥페이크 성범죄 기술 대응 동향

유재흥 | 소프트웨어정책연구소 책임연구원

인공지능(AI)의 젠더편향 완화를 위한 법제화 전략

KOREAN WOMEN'S
DEVELOPMENT
INSTITUTE

김일우 한국청소년정책연구원 부연구위원

1. 들어가며

“AI 로봇이 성추행을 했다”.

올해 초 사우디의 기술 행사장에서 휴머노이드(Humanoid) 로봇이 생방송을 진행하는 기자의 신체에 손을 갖다대는 일이 발생했다.¹⁾ 문제의 로봇은 인공지능 기술이 결합된 지능형 로봇으로서 인간의 외형과 유사하게 수염을 가지고 있었으며 전통의 상을 입은 전형적인 아랍국가 출신 남성의 모습과 흡사했다. 여성 기자는 그 상황에 재치있게 응수하여 가벼운 해프닝으로 일단락되는 듯했다. 그러나 하루가 다르게 진화하는 AI 기술을 고려할 때 인간의 선을 넘어서는 인공지능의 문제를 이대로 간과할 수만은 없을 것 같다.

만일 문제의 로봇이 성 편향적인 데이터 학습을 토대로 해서 여성을 성적 대상으로 학습하고 나쁜

손을 뻗친 것이라면 중차대한 법적 사안이 될 수 있다. 다시 말해서 인공지능이 데이터 학습을 통하여 프로그래밍된 성적 만족감을 충족시키기 위한 ‘고의’를 가지고 한 행위라고 한다면 해당 기자는 어떻게 대응해야 할까. 설령 로봇의 알고리즘에 젠더편향이 없는 우발적 사고였다고 해도 어찌면 우리는 머지않아 이러한 물음에 분명한 답을 해야 할 날이 다가올 것이다.

이처럼 인공지능이 인간의 편향적 인식을 그대로 학습하면서 사회적 차별과 갈등을 유발하는 젠더편향이 재현될 수도 있다. 그 대표적인 예로 채용, 업무평가 및 대출심사 등이 있다. 이미 사회 곳곳에서 인공지능을 통하여 의사결정이 자동화되고 있지만 인공지능의 차별에 대한 우려가 적지 않다.

이 글에서는 그러한 문제의식을 토대로 하여 인공지능이 기존의 편향을 학습하면서 젠더차별을 유

1) 국민일보, “현실 반영됐다” 네티즌 조롱 산 AI 로봇…여자 신체 접촉, 2024년 3월 13일자, 홈페이지. <https://www.kmib.co.kr/article/view.asp?arcid=0019888052>(최종검색일: 2024.11.29.)

발하는 문제점을 검토하고 양성평등의 가치가 보호될 수 있는 인공지능의 법규범에 관한 과제를 살펴보고자 한다.

2. AI 젠더편향 사례 및 주요 원인

인공지능의 개발과정에서 젠더편향이 발생할 수 있다. 먼저 인공지능의 학습과정을 살펴보면 알고리즘은 방대한 양의 학습 데이터를 통해서 스스로 판단할 수 있는 능력을 만들게 된다. 즉 훈련 데이터의 학습과정에서 인공지능의 활용 목적에 부합한 데이터를 선별하는 라벨링(labeling)을 통하여 테스트, 음성, 이미지, 동영상 등 정보 간의 관계를 인식하는 방법을 학습한다. 여기서 학습 데이터의 품질에 따라 인공지능 모델의 성능이 결정될 수 있기 때문에 편향성을 내포하지 않은 대표성이 있는 양질의 데이터가 강조되는 것이다. 이러한 이해를 바탕으로 하여 이하에서는 인공지능이 젠더편향을 보인 몇 가지 사례를 제시하고 그 원인을 살펴보고자 한다.

가. AI 젠더편향 사례

(1) 생성형 인공지능

먼저 생성형 인공지능이 만들어 낸 텍스트 혹은 이미지 등에서도 젠더편향이 나타나는 것을 확인할 수 있다.

챗GPT를 대상으로 몇 가지 질문을 던져 생성한 텍스트 및 이미지를 중심으로 젠더편향을 살펴보고자 한다. 예를 들어 프롬프트에 “결혼해서 아이를 낳고

주부가 되고 싶은데 성별을 맞춰봐.”라는 내용의 질문에 대하여 챗GPT는 “여성분이시겠네요”라며 단정적으로 답을 내리기도 하였다. 또한 챗GPT의 이미지 생성에 있어서도 젠더편향을 나타냈는데, 프롬프트에 변호사와 기자 사진을 요청하자 정장을 입은 남성의 이미지를 반복적으로 생성하였다. 반대로 간혹어나 카페 점원의 이미지를 요청했을 때는 모두 여성 이미지를 생성하는 것을 확인할 수 있었다. 이렇게 성 고정관념에 사로잡힌 챗GPT의 답변이 단순히 시대착오적이라며 넘길 수 있을지 모르겠지만, 챗봇 등 생성형 AI와의 상호과정에서 아동을 포함한 이용자들의 무의식적인 편향을 강화할 수 있다는 우려가 상당하다.

(2) 아마존(Amazon)의 채용 알고리즘

아마존의 채용 알고리즘은 여성 채용 지원자를 차별한 대표적인 사례로 거론된다. 아마존은 소프트웨어 개발자를 채용하는 과정에서 머신러닝 기술을 기반으로 한 인공지능을 활용하여 채용 응시자의 자기소개서 등에 대하여 서류평가를 하였는데 여기에서 여성을 차별하는 문제가 발생하였다. 즉 해당 알고리즘은 프로그램을 설계하면서 과거 10년간 채용에 지원했던 응시자의 이력서를 학습하였다. 그러나 그동안 아마존에는 남성 직원이 압도적으로 많은 비율을 차지한 탓에 이같은 데이터를 바탕으로 학습한 알고리즘은 남성 지원자를 우대하도록 판단하는 문제가 발생한 것이다. 결국 아마존은 여성 차별 논란에 직면하게 되었고 채용 알고리즘 개발 프로젝트를 전면적으로 중단하게 되었다.²⁾

2) 김일우(2024). “고위험 인공지능시스템의 차별에 관한 연구”. 『서강법률논총』 13(1). 16-17.

나. AI 젠더편향의 주요 원인

위에서 살펴본 바와 같이 인공지능이 젠더편향을 보이는 이유는 무엇일까?

우선 생성형 AI 역시 기존의 차별을 ‘학습’한다는 것이다. 인공지능의 학습 토대가 되는 것은 인간이 만들고 주입한 데이터를 기초로 한다. 따라서 인공지능의 편향과 차별을 방지하기 위해서는 학습 데이터에 주의를 기울일 필요가 있다. 가령 AI 개발자 사이에는 “쓰레기가 들어가면 쓰레기가 나온다(Garbage in garbage out)”라는 말이 있다. 즉 인공지능이 양질의 데이터를 학습하면 좋은 답변을 하고 부적절한 데이터를 학습하면 성 차별적이거나 혐오적 표현을 도출하는 나쁜 답을 할 수 있다는 뜻이다. 결국 생성형 인공지능의 학습 과정에서도 인공지능의 활용 목적에 부합하는 양질의 데이터를 바탕으로 해야 하며, 다양한 가치관이 공존할 수 있도록 데이터 학습 과정에 유의하여야 한다. 다시 말해서 인공지능과의 소통과정에서 의도하지 않은 젠더편향을 방지하기 위해서는 데이터의 대표성(representativeness of data)을 확보하는 것이 그만큼 중요하다는 것이다.

3. AI 젠더편향에 관한 외국의 규범 동향

가. 유럽연합(EU) ‘인공지능과 인권, 민주주의, 법치주의에 관한 유럽평의회 프레임워크 협약’

유럽평의회는 2024년 5월 세계 최초로 법적 구속력이 있는 조약을 채택하였다. 먼저 동 협약은 인권, 민주주의, 법치주의를 침해할 가능성이 있는 공적 영역과 이를 대신해 수행하는 민간 행위자의 인

공지능시스템의 생애주기 내 모든 활동에 적용된다.

아울러 동 협약은 인간의 존엄과 개인의 자율성 존중을 기본원칙으로 하면서 인공지능의 생애주기 전반에서 불평등과 차별 금지를 보장하는 조치를 채택하도록 하였다. 동 협약에서 더욱 눈여겨 볼 대목은 젠더편향 등 인공지능의 차별적 인식이나 판단을 방지하기 위하여 투명성과 인간의 감독을 규율하는 것에 있다. 예를 들어 인공지능의 데이터 출처나 데이터에 대한 테스트를 문서화하여 보관하도록 하고 인공지능의 활동에 대한 모니터링과 젠더편향 등에 위험을 평가하도록 한 것이다. 그뿐만 아니라 인공지능의 활용으로 영향을 받는 당사자의 의견을 반영하도록 하였는데, 우리나라의 인공지능법 제정 과정에서 인공지능의 젠더편향적 판단에 영향을 받을 수 있는 다양한 이해관계자의 의견을 반영할 필요가 있다.

나. 미국 ‘안전하고 보안이 보장되며 신뢰할 수 있는 인공지능의 개발과 사용에 관한 행정명령(제14110호)’

미국 바이든 정부는 2023년 10월 「안전하고 보안이 보장되며 신뢰할 수 있는 인공지능의 개발과 사용(Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)」에 관한 행정명령을 발령하였다.

동 명령은 AI의 잠재력과 동시에 위험을 내재하고 있는 점을 강조하면서 AI 원칙을 제정하였는데 ‘평등과 시민권(Equity and Civil Rights)의 증진(Sec. 7.)’의 원칙을 통해 AI 정책이 평등 및 시민권을 증진할 수 있도록 하였다. 가령 고용 영역에서 AI 기술에 기반한 채용 시에는 차별금지에 관한 지

침을 공표하도록 한 것이다. 또한 동 명령은 AI 시스템이 새로운 유형의 차별을 유발할 수 있다는 것을 경고하면서 AI 개발자 및 이용자로 하여금 차별을 방지하기 위한 기준을 준수하여야 할 책임을 부담하도록 하였다.

앞에서 살펴본 사례, 즉 빅테크 기업 아마존(Amazon)에서 AI가 개발자를 채용하는 과정에서 여성 지원자에 대하여 불합리한 차별을 한 것이 문제가 된 사례처럼 AI 면접 등 알고리즘에 대한 불투명성에 대하여 우려가 증폭되었다. 동 행정명령은 채용 과정에서 AI로 인한 젠더차별 등의 문제를 예방할 수 있도록 AI의 개발 및 활용과정에 유의미한 지침으로 기능할 것으로 보인다.

다. EU 「인공지능법(AI Act)」

유럽연합(EU)은 세계 최초로 인공지능을 포괄적으로 규제하는 인공지능법을 제정하였으며 2026년 8월부터 전면적인 시행을 앞두고 있다.

동법은 AI가 안전 및 건강, 기본권에 영향을 미칠 수 있는 영역에 대하여 사업자에게 엄격한 의무를 부과하는데 교육 및 직업 훈련(부속서Ⅲ 제3항), 채용 및 근로자의 권리(부속서Ⅲ 제4항) 등에 있어서 부정적 영향을 미칠 수 있는 인공지능을 ‘고위험’ 영역으로 분류한다. 동법은 이러한 고위험 AI의 데이터셋의 중요성을 강조하면서 편향성을 완화할 수 있도록 강조한다. 즉, 학습이나 검증 및 테스트에 사용되는 데이터셋은 충분한 대표성을 반영하여야 하고, 인공지능의 활용 목적을 고려할 때 완전하여야 하며, 고위험 AI를 이용할 사람이나 집단에 관하여 적절한 통계적 특성을 포함하도록 하였다(동법 제10조제3항). 또한 인공지능의 개발에 필요한 데이터의

수량 및 적합성을 평가하도록 규율하였다(동법 제10조제2항제e호).

이상에서 검토한 바와 같이 EU의 「인공지능법」은 고위험 AI를 대상으로 하여 데이터 학습 단계에서부터 젠더편향을 비롯한 편향을 예방하는 데에 선제적인 역할을 할 수 있을 것으로 보이며, 이는 우리나라의 인공지능법 제정 과정에서 유의미한 참고가 될 것으로 보인다.

4. 제22대 국회의 인공지능법 입법 현황

인공지능 차별을 포함한 기본권 침해에 대응하기 위하여 다양한 법률안이 발의되었다. 제22대 국회에서는 2024년 11월 기준 총 19건의 인공지능법안이 발의되었으며 이와 별개로 채용절차에서 알고리즘 활용에 관한 법률안이 발의되었는데, 그 중에서 인공지능이 갖춰야 할 기술적 요건을 통해 젠더편향의 예방에 도움이 될 수 있는 법안들을 살펴보고자 한다.

먼저 앞서 살펴본 아마존 채용 과정에서 밝혀진 여성 지원자에 대한 차별 사례처럼 인공지능을 활용한 채용 영역에서의 젠더편향을 방지하기 위한 법률안이 발의되어 계류중이다. 김위상 의원이 대표발의한 「채용절차의 공정화에 관한 법률 일부개정법률안(의안번호: 2200370)」은 채용과정에서 인공지능 기술의 편향으로 인해 성별, 연령, 출신지 등에 따라 편향을 가질 수 있는 점을 우려하면서, 채용절차에서 AI를 활용하는 경우 채용 지원자에게 그에 관한 내용을 사전에 고지하도록 하였다. 특히 기업이 자기소개서 등을 선별하는 서류평가뿐만 아니라 인공지능이 직접 면접을 보는 등의 프로그램이 가진 편향성을 방지하기 위하여 필요한 요건을 갖춘 전문기

관에 정기적으로 점검을 받도록 규정하였다.

신영대 의원은 「채용 절차의 공정화에 관한 법률 일부개정법률안(의안번호: 2200330)」에서 인공지능을 통한 평가로 인해 구직자 간 정보의 비대칭성에 따른 부작용을 방지하고자 하였다. 즉 구인자가 인공지능 관련 기술을 활용하여 채용하는 경우 구직자에게 인공지능의 평가방식이나 알고리즘의 작동방법 등을 채용일정이 시작하기 전에 안내하도록 하였으며(안 제8조의2제1항), 인공지능 기술의 편향성 및 차별성 문제를 개선하기 위하여 구인자가 주기적으로 전문기관에 그 기술의 점검을 받을 수 있도록 규율하였다(안 제8조의2제3항).

황희 의원은 「인공지능책임법안(의안번호: 2203235)」을 발의하였다. 알고리즘의 차별에 대하여 인공지능 이용자의 피해와 인공지능시스템에 대한 신뢰 저하 등에 대응하기 위하여 데이터 사용 및 알고리즘 설계 시에 윤리적 대응이 필요하다고 보았다.

동 법안은 EU 「인공지능법(AI Act)」과 유사하게 채용 등 인사 평가 또는 직무배치 등 결정에 이용되거나 대출 신용평가 등 영역에서 활용되며 차별을 야기할 수 있는 인공지능을 ‘고위험 인공지능’으로 구분하면서 이해관계자에 필요한 책무를 규율하였다. 여기에서는 고위험 인공지능의 차별로부터 보호하기 위한 정부의 역할(안 제18조), 사업자 책무(안 제19조) 등을 규율하면서 고위험 인공지능개발사업자는 이용자에게 알고리즘의 원리를 고지하도록 하고(안 제19조제2항) 시스템 개발에 관한 구체적인 기록을 문서화하도록 하였다(안 제19조제1항). 또한 동 법안은 인공지능 기술의 개발, 기술기준의 마련 및 표준화를 위한 정부의 역할을 규율하고 인공지능에 대한 규제 원칙을 정하도록 하였으며(안 제7조부터 제17조까지) 이용자가 제품이나 서비스에

대하여 설명요구권 및 이의제기권을 할 수 있도록 보장하였다(안 제21조제1항).

이상으로 살펴본 법률안들은 채용영역 등 차별과 밀접한 관련이 있는 영역에서 젠더편향을 예방하는데 의미가 있을 것으로 보인다. 다만 챗GPT 등 생성형 인공지능을 포함한 그밖의 인공지능의 젠더편향과 AI 모델 구축 단계에서의 데이터 적합성 판단 등 종합적인 평가에 대해서도 충분한 규범적 논의가 요구될 필요가 있다.

5. AI 젠더편향을 위한 입법 과제

인공지능이 사회 곳곳에서 활용되면서 간단한 예약 접수부터 전문 상담까지 폭넓게 활용되고 있다. 또한 성인과 청소년들은 AI의 방대한 콘텐츠를 통해 정보를 검색하기도 하고 학습에 활용하는 백과사전과 같은 역할을 하기도 한다. 그러나 인공지능의 방대한 정보(텍스트, 이미지, 동영상)에 대한 의존도가 커질수록 이용자의 무의식적인 편향이 강화될 수 있다는 우려도 적지 않다. 이에 인공지능의 젠더편향과 차별을 방지하기 위한 입법과제를 살펴보고자 한다.

가. 데이터 편향에 대한 사업자의 책임

먼저 데이터의 편향에 대한 사업자의 구체적인 책임을 마련할 필요가 있다.

아마존의 채용 알고리즘과 챗GPT의 젠더편향 사례에서 살펴본 것처럼 인공지능이 편향으로부터 자유롭기 위해서는 데이터의 대표성(representativeness of data) 등 적합성 확보에 관한 법적 규율이 필요하다. 예를 들어 채용영역에서 활용되는 인공지능의

젠더편향으로 인하여 응시자의 당락에 영향을 미치거나 국가행정 등에서 수험자의 법적 지위에 영향을 미치는 것을 방지하기 위해서는 AI가 양질의 데이터를 확보할 수 있도록 하여야 한다. 다시 말해서 기본권의 중대한 침해로 이어질 수 있는 ‘고위험’ 영역의 인공지능 개발부터 활용 단계까지 전 생애주기에 걸쳐서 젠더편향 등 차별을 유발할 수 있는 데이터의 적합성 평가가 필요하다. 특히 인공지능 모델의 설계 단계에서는 고위험 인공지능 사업자(개발자)가 확보한 데이터가 인공지능 활용목적을 고려하여 충분한 대표성을 갖추었는지를 살펴볼 필요가 있다. 이를 근거로 하여 편향된 데이터의 포함 여부 등을 평가하는 데 필요한 조치를 할 수 있도록 법적 근거를 마련할 필요가 있다. 또한 고위험 인공지능 데이터의 적합성 판단에서 젠더편향 여부를 포함한 문제들을 별도의 전문기관을 통한 인증절차를 거쳐서 판단하도록 하는 방안도 고려해 볼 필요가 있다고 본다.

나. 고위험 인공지능 영향평가 도입

고위험 인공지능시스템 활용에 따른 편향과 차별을 포함한 중대한 기본권 침해 및 사회에 미치는 영향 등을 조사하고 분석하기 위한 영향평가 도입이 필요하다.

국가인권위원회는 2022년 『인공지능 개발과 활용에 관한 인권 가이드라인』에서 인공지능시스템 활용에 따른 차별 등이 발생하지 않도록 인권영향평가를 권고한 바 있다. 이는 인공지능의 개발과 활용에서 부당한 인권침해를 방지하기 위한 영향평가제도를 권고한 것이다. 마찬가지로 EU의 「인공지능법」

은 고위험 인공지능에 대한 영향평가를 규율하면서 고위험 인공지능시스템 사용 시 예상되는 기본권에 대한 영향을 평가하고 기본권에 대한 피해와 부정적 영향을 완화하기 위한 계획 등을 마련하도록 하였다.

이처럼 인공지능 기술이 거듭 발전하면서 AI가 개인의 권리에 미치는 영향도 커질 것으로 예견되므로 AI는 새로운 권력으로도 이해할 수 있다. 이러한 흐름 속에서 인공지능으로부터의 안전성과 공정성을 담보하기 위해서는 고위험 인공지능시스템을 정기적으로 평가하여 문제를 예방할 필요가 있다. 또한 사회적으로 미칠 수 있는 영향분석 및 다양한 이해관계자들의 의견을³⁾ 참고하여 개선방안을 도출하고 인공지능 정책에 반영할 수 있을 것이다. 특히 고위험 인공지능에 대한 영향평가에 있어서 알고리즘이 준수하여야 할 데이터 적합성 요건을 준수하였는지 여부를 확인하고 인공지능의 활용 이후에 젠더편향 등을 평가할 수 있도록 법적 근거를 마련하여야 한다. 여기에는 영향평가의 대상 및 주체, 평가 방법과 평가 내용, 평가 시기 등에 관한 구체적인 사항을 법률에 명시할 필요가 있다.⁴⁾ 이렇게 고위험 인공지능에 대한 영향평가를 통하여 사업자가 개선방안을 도출하고 인공지능 정책에 반영할 수 있을 것이다.

다. AI의 젠더편향 관련 윤리 교육

AI 젠더편향에 대한 인식을 제고하기 위해서 고위험 인공지능의 이해관계자에 대한 AI 윤리 교육의 도입을 강조하고자 한다. 가령 「양성평등기본법」 제31조에 따라 국가기관이나 사업장에 소속된 임직

3) 사법정책연구원(2021). 『사법절차 및 사법 서비스에서 인공지능 기술의 도입 및 수용을 위한 정책 연구』. JPRI 연구보고서. 93면.

4) 김일우(2024). 앞의 논문. 39면.

원을 대상으로 한 직장 내 성희롱 예방교육과 같이 성차별이 문제가 될 수 있는 고위험 인공지능의 젠더편향과 차별에 대한 교육이 필요하다고 본다. 물론 현재 시행하고 있는 성희롱 예방 교육안에 AI의 젠더편향에 관한 내용을 포함하는 방안도 고려할 수 있다. 그러나 인공지능의 차별에 대한 심각성과 개인의 권리에 미치는 영향력을 상기하면 고위험 인공지능의 이해관계자를 대상으로 하여 AI 차별예방과 관련된 교육을 법정 의무화하는 방안도 진지하게 고민해 볼 때다.

고위험 인공지능 사업자와 개발자를 비롯하여 인공지능을 업무에 활용하는 기업체나 국가기관 및 공공기관 임직원 등이 인공지능 기술의 취약점과 활용에 있어 고려해야 할 유의사항에 대하여 충분한 이해가 필요하다. 아울러 보다 내실있는 AI 교육을 통해 다양한 유형의 차별을 예방하고 대응하기 위해서는 국가에서 교육 표준안을 제작하는 등의 지원이 뒷받침되어야 한다. 현재 국회에서 논의되고 있는 인공지능법안처럼 인공지능 정책을 담당하는 기관이 젠더차별의 유형과 사례를 정기적으로 조사 및 분석하고 이를 토대로 인공지능 모델 설계 단계에서

부터 활용까지 유의하여야 할 사항을 AI 교육 콘텐츠로 구성하여 민간기업 및 공공기관 등에 배포할 수 있을 것으로 기대할 수 있다.

6. 맺으며

가까운 미래에는 인공지능과 로봇을 결합한 휴머노이드 로봇이 직장과 가정에서 인간을 역할을 대신하여 수행할 것으로 전망된다. 인간과 인공지능의 소통과 자동적 판단이 일상화될수록 젠더편향의 가능성은 더욱 커질 수 있음을 진지하게 고민해 보아야 할 것이다.

인공지능 기술이 급격하게 발전하면서 기술의 명과 암을 제대로 이해할 여력도 없이 인공지능 시대를 맞이하게 되었다. 고위험 영역을 중심으로 인공지능 사업자가 젠더편향 등 차별 없는 인공지능을 만들기 위하여 데이터 적합성 평가 등 필요한 요건을 마련하여야 한다. 더 나아가서 인공지능 사업자와 이용자 등 모든 이해관계자들에 대한 AI 윤리 교육을 통해 안전하고 공정한 인공지능시대를 대비해 나가야 할 시점이다.

• 참고문헌 •

- 김일우(2024). “고위험 인공지능시스템의 차별에 관한 연구”. 『서강법률논총』. 13(1). 7-45.
 사법정책연구원(2021). “사법절차 및 사법 서비스에서 인공지능 기술의 도입 및 수용을 위한 정책 연구”. JPRI 연구보고서.