



KOREAN WOMEN'S
DEVELOPMENT
INSTITUTE

인공지능(AI) 활용의 젠더 편향성 실태 및 개선방안¹⁾

문미경 한국여성정책연구원 선임연구위원

1. 서론

현재 인공지능(AI)의 발전 속도는 가늠할 수 없을 정도로 빠르며, 인공지능(AI)은 우리의 일상생활에 광범위하게 영향을 미치면서 깊이 파고들고 있다. 인공지능(AI)은 편향적이고 평가적 특성을 지닌 인간에 의해 만들어지기 때문에 인간만큼이나 편향성이나 오류 가능성은 항상 존재할 수 있다. 그 때문에 인공지능(AI)이 항상 공정하고 객관적이지 않음을 인식하는 것이 중요하다. 인공지능(AI)의 편향성에 의해 야기되는 차별 문제는 차별금지라고 하는 사회적 가치를 훼손할 여지가 있다(양종모, 2017; 변순용, 2020). 따라서 인공지능(AI)이 인간을 대신하며 사회적으로 그 비중이 점점 커지고 있는 시점에서 인공지능(AI) 편향성에 의해 야기되는 차별을 사회적 문

제로 인식하고 이를 심층적으로 고찰할 필요가 있다.

본고는 인공지능(AI) 젠더 편향성(gender bias) 문제에 초점을 맞추어 인공지능(AI) 젠더 편향성 실태²⁾ 및 개선에 필요한 방안들을 살펴보았다.

2. 인공지능 젠더 편향성

인공지능(AI)은 자율주행 자동차, 사람을 가장하는 로봇, 기계학습 등 사람들에게 각기 다른 의미를 부여하는 개념과 기술의 집합체이며 그 응용 프로그램은 어디에서나 볼 수 있다. 최근 인공지능(AI)의 정의는 인간의 지능으로 할 수 있는 문장이해, 영상 인식, 음성인식, 학습 등을 컴퓨터가 실행하도록 하는 방법을 연구하는 컴퓨터 공학 및 정보기술의 한

1) 문미경·김복태 외(2022). 인공지능 딥러닝 활용의 젠더 편향성 실태 및 개선방안. 한국여성정책연구원 연구보고서 내용을 발췌하여 정리함.

2) 인공지능(AI) 젠더 편향성 사례를 인공지능(AI) 기술 구성 단계인 기획 및 설계, 데이터 처리단계, 알고리즘 생성 및 학습 등 모델링 단계로 유형화하여 살펴봄.

분야로서, 컴퓨터가 인간의 지능적인 행동을 모방할 수 있도록 하는 것을 말한다.³⁾

인공지능(AI)을 통해 컴퓨터는 방대한 양의 데이터를 활용하고 학습된 지능을 사용하여 인간에 의해 소요되는 시간보다 훨씬 짧은 시간 안에 원하는 결과물을 만들어 낼 수 있다.

최근에 인공지능(AI) 분야에서 표면적으로 눈에 띄게 드러나는 이슈는 편향성이었고, 특히 인공지능(AI) 편향성의 문제는 인종과 젠더의 문제로 집중되고 있는 양상을 보인다. 이에 사회적인 우려의 시각과 이에 대한 해결방안을 모색하려는 윤리적, 사회적, 법적, 기술적인 노력들이 강조되고 있다. 인공지능(AI) 편향성과 관련하여 개발된 알고리즘의 젠더 편향성을 검증하는 알고리즘도 개발되고 있는 추세이며, 알고리즘에 대한 윤리적 검증시스템의 도입이 앞으로도 필요해질 것으로 예측되고 있다.

인공지능(AI) 젠더 편향성을 살펴보기 위해서는 ‘무엇을 젠더 편향성이라 할 수 있는가?’의 개념 정의가 필요하다. 본고에서는, 젠더 편향성은 성별에 따라 또는 젠더에 미치는 영향이 편향적으로 나타나는 현상을 의미하며, 인공지능(AI) 젠더 편향성은 이러한 편향성이 AI 기술 구축의 초기 단계인 인공지능(AI) 시스템의 기획 및 설계, 데이터 처리(수집·가공·관리 등), 알고리즘 생성 및 학습 등의 모델링 각 기술 단계에서 발생하거나 활용단계에서 나타나는 것으로 정의하였다.

인공지능(AI) 활용의 젠더 편향성은 데이터의 젠더 편향성, 알고리즘의 젠더 편향성, 그리고 AI 활

용단계에서 나타나는 AI의 젠더 편향성, 그리고 AI에 대한 젠더 편향성 외에도 이러한 단계들이 교차하여 나타날 수 있다. 본고에서는 데이터의 젠더 편향성, 알고리즘의 젠더 편향성, 그리고 AI 활용단계에서 나타나는 AI의 젠더 편향성, AI에 대한 젠더 편향성을 중심으로 살펴보았다.⁴⁾

가. 데이터의 젠더 편향성

데이터의 젠더 편향성은 AI의 학습데이터의 생성 과정에서 데이터 자체의 젠더 편향성이 문제가 되는 경우에 발생한다. 인공지능(AI)이 정제되지 않은 학습데이터를 활용할 경우 특정 인종·연령·성별에 대한 현실의 차별이 여과 없이 학습되거나, 특정 언어권의 문화와 가치관만이 학습 데이터로 입력되어 다양성이나 공정성의 문제가 제기될 수 있다. 예를 들어, GPT-3는 방대한 양의 인터넷 웹 자료를 8년간 모은 CommonCrawl⁵⁾을 학습하였는데, CommonCrawl이 수집한 웹 자료의 대부분은 미국이나 영국에 사용되는 영어가 모국어인 20~30대 백인 남성이 작성한 텍스트이다. GPT-3이 학습데이터 활용한 데이터가 영미 계열의 백인 남성들이 발화한 내용들이었다는 것이 알려지자 학습데이터의 인종적, 성별적 편향성에 문제가 제기되기도 하였다. 학습데이터의 잠재적인 젠더 편향성을 학습한 인공지능(AI)이 실제로 성차별적 표현을 사용할 가능성이 있다는 점이다.

3) 네이버 지식백과, <https://terms.naver.com/entry.naver?docId=1136027&cid=40942&categoryId=32845>, 접근일: 2022.4.12.

4) 이하 ‘가. 데이터의 젠더 편향성’~’라. 인공지능(AI)에 대한 젠더 편향성’의 내용은 문미경·김복태 외(2022)의 인공지능 딥러닝 활용의 젠더 편향성 실태 및 개선방안 연구자문회의(2022.10.13.)의 변순용 교수님 발표자료를 기반으로 작성함.

5) 자동화된 방법으로 웹사이트나 하이퍼링크, 데이터, 정보 자료를 수집, 분류, 저장한 크롤링(crawling)한 데이터를 누구나 쉽게 접근하고 분석하도록 공개(open) 저장소에서 유지·관리하는 데이터를 의미함. <http://www.apple-economy.com>, 접근일: 2024.3.27.

나. 알고리즘의 젠더 편향성

젠더 편향성이 제로인 데이터의 존재 자체가 불가능하기 때문에, 알고리즘 학습과정에서 어느 정도의 젠더 편향성을 학습할 수밖에 없을 것이다. 이러한 현실적 전제하에서 알고리즘이 설계되는 과정이나 사용되는 과정, 그리고 그 과정에서 산출되는 결과물의 편향성은 존재할 수밖에 없다.

다. 인공지능(AI)의 젠더 편향성

인공지능(AI) 젠더 편향성은 생산된 인공지능(AI)에서 젠더 편향성이 표출되는 것이다. 예를 들면 인공지능(AI)에서 제공하는 비서의 디폴트 음성이 여성으로 설정된 것 등이다. 우리나라의 챗봇 서비스인 이루다 역시 여자 대학생의 정체성을 가지게 된 것이나 섹스로봇이 여성의 모습으로 먼저 선을 보였다는 것 등에서 AI의 젠더 편향성을 알 수 있다. 인공지능(AI) 비서나 소셜 로봇들이 대체로 여성형으로 설계되고 출시되는 것 등이 AI의 젠더 편향성의 좋은 예가 될 수 있다.

라. 인공지능(AI)에 대한 젠더 편향성

인공지능(AI)에 대한 젠더 편향성은 앞의 인공지능(AI)의 젠더 편향성과 중첩될 수 있는 개념이지만, 사람들이 인공지능(AI)에 대하여 성적으로 편향적인 태도를 취하는 경우를 의미한다. 2016년 마이

크로소프트사의 챗봇인 테이에게 이용자들이 젠더 차별에 대한 학습을 유도하여 테이가 이런 유형의 말을 하도록 한 경우가 좋은 사례이다. 이와 더불어 성적 착취와 수익화의 도구로 인공지능(AI)에 근거한 딥페이크 내지 디지털 휴먼을 불법적으로 이용하는 경우도 여기에 포함될 수 있다. 인공지능(AI)의 젠더 편향성은 인공지능(AI) 자체가 가지고 있는 젠더 편향성이라면, 인공지능(AI)에 대한 젠더 편향성은 인공지능(AI)의 유저들이 인공지능(AI)에 대해 행사하는 젠더 편향성이라고 설명될 수 있다.

3. 인공지능(AI) 젠더 편향성 실태

인공지능(AI) 활용의 젠더 편향성 사례를 인공지능(AI) 기술 구성 단계인 기획 및 설계, 데이터 처리 단계, 알고리즘 생성 및 학습 등 모델링 단계로 구분하고 이에 대해 살펴보았다. 우선, 기획 및 설계단계에서는 인공지능(AI)의 콘셉트, 목적, 내용, 필수조건, 기본전제, 프로토타입 설계 등의 활동 등 ‘인공지능(AI)의 젠더화’, ‘성적 착취 및 수익화 목적과 인공지능(AI)의 결합’, ‘젠더 편향 및 차별을 조장할 수 있는 젠더 타깃 연구’로 유형화할 수 있다. 이 단계는 인공지능(AI) 기술의 사용 목적 및 타깃을 규정하는 단계이기 때문에, 이에 속하는 사례는 역사적·문화적으로 누적된 사회의 젠더 편향 및 성차별이 두드러지게 나타난다.

〈표 1〉 인공지능(AI) 기획 및 설계 단계 주요 사례와 유형 분석

세부 유형	사례	분석
인공지능(AI)의 젠더화	AI 비서, 소셜 로봇 등의 여성 음성, 이미지화, 성적 괴롭힘에 대한 순응적 반응 등	가장 많이 보고되는 사례 유형. 현존하는 부정적 젠더 편향성. 고정관념을 반영하고 이를 재강화.

〈표 1〉 인공지능(AI) 기획 및 설계 단계 주요 사례와 유형 분석(계속)

세부 유형	사례	분석
성적 착취 및 수익화 목적과 인공지능(AI) 기술의 결합	딥페이크 기술 활용 포르노, 여성 신원 추적 목적 알고리즘 개발 등	기존의 성적 학대 및 착취 행위, 디지털 성범죄, 성적 착취를 통한 수익화 목적 등과 인공지능(AI) 기술이 결합. 인공지능(AI) 기술을 통해 과거보다 더 수월하게 목적 성취, 더 큰 피해 및 피해 구제의 더 큰 어려움 야기.
젠더 편향 반영 및 차별을 조장할 수 있는 젠더 타깃 연구	얼굴 이미지만으로 성적 지향 식별 가능한 안면인식 알고리즘 연구 등	인공지능(AI) 기술이라는 것만으로 비과학적인 알고리즘 분류 및 연구조차 과학적인 것으로 간주됨. 부정적 젠더 편향에 따른 정체성 식별, 성차별 강화 목적으로 알고리즘 연구 및 설계가 이용될 수 있음.

둘째, 데이터 처리 단계에서는 ‘데이터 자체 편향’과 ‘데이터 수집 및 가공 기준에 따른 데이터 처리 편향’으로 유형화할 수 있다. 데이터 처리 단계는 데이터의 수집·가공·품질 검사 등 관리, 데이터세트의 특성의 문서화, 데이터세트의 용처 혹은 목적, 형성 방법, 구성, 유지 방법 등의 정보를 포함하는 메타데이터의 문서화 및 지속적 관리 활동 전반을 포함한다. 이 단계의 사례는 데이터 편향은 과소대표

혹은 과대대표하여 대표성을 갖지 못하는 ‘데이터 편향(Data Bias)’이 원인이 된다. 학습용 데이터세트 가공 및 선별만이 아니라, 가공 전 원 데이터(raw data)의 젠더 편향성도 포함한다. 젠더 고정관념을 반영하지 않고 기획된 인공지능(AI)조차 이용자의 젠더에 따라 그 결과물에 커다란 격차를 드러낸다. 궁극적 원인은 데이터에 반영된 역사적·사회구조적 젠더 편향과 그에 대한 문제의식 결여라 할 수 있다.

〈표 2〉 데이터 처리 단계 주요 사례와 유형 분석

세부 유형	사례	분석
데이터 자체 편향	언어처리 영역의 기계번역, 단어 임베딩에서 성별중립 혹은 여성형 단어를 남성형으로 자동 번역 및 연관 짓는 경우, 자동 데이터 라벨링 혹은 자동 이미지 생성 알고리즘의 젠더에 따라 다른 결과 산출 등	가공하지 않은 데이터 자체 편향이 주요 원인. 곧, 사회 내 데이터에 역사적 젠더 편향의 누적 및 반영
데이터 수집 표본 집단 설정에서 젠더 편향	안면인식 프로그램의 백인 남성과 흑인 여성 간에 두드러지는 인식률 격차, 의료데이터 남성 집단 편향성 등	데이터 수집 및 설정의 범위가 시·공간적으로 지나치게 협소하거나 편중됨. 시간에 따른 사회적 변화를 고려하지 않음

마지막으로, 알고리즘 생성 및 학습 등 모델링 단계이다. 이 단계에서 나타나는 사례는 세부적으로 ‘부정적인 젠더 편향에 대한 고려없는 알고리즘 설계’, ‘젠더 편향 알고리즘 기반 기계학습’, ‘알고리즘

의 투명성·설명가능성’의 어려움으로 유형화할 수 있다. 이 단계에는 알고리즘의 생성, 선택, 교정이나 알고리즘 훈련(학습) 및 알고리즘의 의미를 설명하는 일 등이 포함된다. 이 단계에 나타나는 편향성

은 ‘알고리즘 편향성(algorithm bias)’으로 불린다. 그러나 알고리즘 생성 및 훈련은 인공지능 기획, 데이터 처리 등과 개발과정 내 여러 단계와 긴

밀한 연관 속에서 상호작용을 한다. 따라서 이 단계의 사례는 대부분 복합적 원인에 의한 결과로 볼 수 있다.

〈표 3〉 알고리즘 생성 및 학습 등 모델링 단계 주요 사례와 유형 분석

세부 유형	사례	분석
부정적인 젠더 편향 실태 및 그 개선에 대한 고려 없는 알고리즘 설계	‘STEM’(Science, Technology, Engineering, Math)으로 대표되는 과학기술 분야의 경력 개발 광고 노출의 젠더 격차, 전통적 젠더 고정관념을 강화하는 컴퓨터 시각의 모델 훈련 등	알고리즘이 젠더 편향을 고려하여 설계되지 않았기에 발생, 변수 간의 상관관계를 강한 종속 관계로 해석하는 등, AI의 최초 목적 외에 젠더 편향의 가능성 및 부작용을 고려하지 않아 결과적으로 젠더 편향을 더욱 강화하도록 설계.
젠더 편향 데이터 기반 알고리즘 기계학습	전통적으로 남성에게 편중되었던 고임금 일자리 광고 노출의 젠더 격차, AI 채용 과정에서 여성 연관 키워드에 관한 무조건적 감점 등	현존하는 사회 내 젠더 간 직무 차이, 임금 격차 등이 반영된 훈련용 데이터셋에 따라 알고리즘 학습이 원인, 알고리즘의 변수를 설정할 때, 다양성, 젠더 평등에 대한 고려 없이 과거 채용 패턴과 크게 다르지 않은 채용 알고리즘을 선호한 결과.
알고리즘 투명성 혹은 설명가능성의 어려움	남편과 공동자산 소유자인 여성에 대한 더 낮은 신용 한도 책정, 유색 인종 여성 키워드 검색 시 선정적 콘텐츠의 검색 과다 노출(검색 최상단 제시 등)	알고리즘의 추론 과정이 인간의 추론 과정과 동일하지 않기 때문에, 기술의 특성상 알고리즘이 어떻게 지금과 같은 결과를 산출하였는지 설명 및 이해가 어렵다는 알고리즘의 기술적 불투명성/설명성의 어려움, 영업비밀이나 지적 재산권을 근거로 한 알고리즘 접근의 어려움이 주요 원인, 젠더 편향으로 인한 피해 인지가 어렵고, 주요 원인 확인이 어려움. 따라서 문제 감지 및 이의제기, 피해 구제 등 문제 상황의 개선 역시 어려워짐.

4. 인공지능(AI) 활용의 젠더 편향성 개선 방안

인공지능(AI) 젠더 편향성 사례 및 연구가 꾸준히 이어지고 있다. 이는 인공지능(AI) 젠더 편향성에 대한 문제의식이 사회적으로 확산되고 있음을 보여 주지만, 다른 한편 인공지능(AI) 젠더 편향성이라는 문제가 충분히 개선되지 않고 있다는 방증으로 읽을 수도 있다.

인공지능(AI) 활용의 젠더 편향성 문제를 개선하기 위해서는 근본적 원인에 더욱 주목할 필요가 있

다. 인공지능(AI) 활용의 젠더 편향성은 인공지능(AI)의 객관성에 대한 근거 없는 믿음(인간보다 기계가 편향적이지 않을 것이라는 가정)에서 발생하는 경향이 있다. 그리고 이에 대한 대응이 쉽지 않은 까닭은 인공지능(AI)이 산출하는 결과물의 과정과 근거를 인간이 정확하게 이해하기도 어렵고, 인공지능(AI) 기술의 특성(투명성 혹은 설명가능성의 문제)상 일반인이 쉽게 그러한 정보에 접근하기도 어렵기 때문이다. 그러나 이러한 문제는 인공지능(AI) 기술이 사회적 문제와 서로 얽혀 연결되어 있음을 보여 준다. 인공지능(AI) 젠더 편향성의 사례를 볼

때 근본적 원인은 누적된 역사적·구조적 젠더 편향의 영향으로 해석될 수 있다.

따라서 인공지능(AI) 활용의 젠더 편향성을 개선하기 위해서는 세 가지 측면에서 대응할 필요가 있다.

첫 번째는, 기술공학적 차원에서 기술 구성 및 활용 전체 단계의 젠더 편향성 문제를 완화하는 것이다. 앞서 인공지능(AI)의 젠더 편향성 문제는 인공지능(AI) 기술 구성의 각 단계별로 발생할 수 있음을 확인하였다. 그러므로 인공지능(AI) 기획 및 설계, 데이터 처리(수집·가공·관리 등), 알고리즘 생성 및 학습 등의 모델링이라는 인공지능(AI) 기술 구성의 전체에 걸쳐, 각 단계마다 인공지능(AI)의 젠더 편향성 문제를 고려하고 젠더 편향성을 완화할 수 있는 유용한 도구들이 개발될 필요가 있다.

두 번째는, 법적 제도적 차원에서 젠더 편향성 완화를 유도하는 것이다. OECD, 유럽연합(EU), UNESCO 등의 국제기구들은 인공지능(AI) 윤리에 대한 권고안을 마련하여, 각 국가들이 인공지능(AI) 기술을 개발하고 활용함에 있어 신뢰성을 확보하고 인공지능(AI) 윤리를 고려해야 함을 강조하고 있다. 현재 국내 인공지능(AI) 관련 현행 법령 및 정책에서는 인공지능(AI) 기술의 모든 단계에 차별과 편향이 발생하지 않게 할 것을 제시하고 있으나, 대부분의 법률 제정안에서 젠더에 대하여 명시적으로 고려한 조항은 미비한 상황이다. 또한 알고리즘 및 인공지능(AI)의 진흥, 인공지능(AI) 산업 기반 조성 및 육성에 있어서 인간의 기본적 인권과 존엄성을 보호하도록 하고, 이를 위한 국가 및 지방자치단체 책무를 제시하고 있으나, 성차별 방지를 위한 구체적인 내용은 여전히 미비한 상황이다. 해외 주요 국가들의 법제화와 같이 인공지능(AI)이 인종·성차별에 미칠 수 있는 파급효과에 대하여 투명한 방식으로 평

가 및 공개하는 방안이 젠더 편향성을 포함시킬 필요가 있으며, 우리나라 인공지능(AI) 기술 근거를 이루는 지능정보화 기본법 속에 젠더 편향성을 완화할 수 있는 내용을 포함할 필요가 있다. 또한 유네스코 인공지능(AI) 윤리 권고에서 시사점을 얻어 윤리 영향평가를 수행할 전담 기구를 설치하고 운영할 때 젠더 편향성을 완화할 수 있는 내용들을 포함하여 구성할 필요가 있다.

세 번째는, 인공지능(AI) 기술의 젠더 편향성 문제를 사회 전체의 문제로 인지하고, 사회문화적 차원에서 젠더 편향성을 완화하는 것이다. 인공지능(AI) 활용의 젠더 편향성의 궁극적 원인은 사회 내 젠더 고정관념 및 성차별 문제에 있기 때문이다. 사회 내 모든 행위자에 대한 인공지능(AI) 윤리교육, 특히 젠더 편향성과 관련된 커리큘럼, 콘텐츠가 필수적으로 포함된 윤리교육이 요청된다. 인공지능(AI) 제품 및 서비스의 직접적 이용자는 아니지만, 기술 환경이나 직접 이용자와 연관되어 인공지능(AI) 제품 및 서비스에 간접적 영향을 받는 일반 시민 모두가 각자에게 인공지능으로 인해 부정적 영향을 받을 수 있는 만큼, 이해 가능한 어휘로 교육이 필요하다. 아울러 기술과학 등 기존에 여성이 많이 진출하거나 교육받지 않은 분야의 여성 인재 부족 등이 개선되어야 한다. 특정 직종 및 산업 분야, 예를 들면 인공지능(AI) 연구 및 개발자 분야의 여성 인재 부족은 성별에 따라 학문적, 직업적 관심사가 다르게 형성되는 것을 당연하게 생각하고 선택하도록 조장해 온 역사적·구조적 젠더 편향성의 결과이자 다시 그것을 강화하는 요인이다. 따라서 이러한 부분에서 개선이 필요하다.

• 참고문헌 •

변순용(2020). 데이터 윤리에서 인공지능 편향성 문제에 대한 연구. 윤리연구, 1(128), 143-158.

양종모(2017). 인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안. 법조, 66(3), 60-105.

네이버 지식백과 인공지능(<https://terms.naver.com/entry.naver?docId=1136027&cid=40942&categoryId=32845>,
접근일: 2022.4.12.

OECD(2019). 『Artificial Intelligence in Society』, Paris:OECD Publishing.