서강법률논총 제13권 제1호 DOI : 10.35505/slj.2024.02.13.1.7

고위험 인공지능시스템의 차별에 관한 연구*

김 일 우**

국문초록

지난 2023년 챗 GPT-4의 출현으로 인공지능은 뜨거운 화두가 되었으며, 사회의 다양한 영역에서 인공지능 기술이 도입되면서, 지능정보사회의 발전을 견인하고 있다.

예를 들어, 사적 영역에서는 채용 및 업무평가, 신용평가 등에서 인공지능 기술의 도입이 확산되고 있으며, 공적 영역에서는 인공지능을 활용한 자동적 행정처분이 도입되고, 법관의 재판 업무를 보조하는 인공지능 기술 등의 도입이 추진되고 있다.

이처럼, 인공지능시스템은 다양한 업무처리의 신속성과 효율성을 높이는 혁신적인 신기술로서 상당한 관심과 기대를 받고 있지만, 인공지능의 데이터와 알고리즘에 내재된 기존의 편향과 차별에 대한 우려도 적지 않다. 이와 같은 문제에 관하여 현재 국회에는 평등권 등 중대한 기본권 침해를 야기할 수 있는 고위험 인공지능의 규제에 관한 법률안이 다수 계류되어 있기는 하지만, 인공지능 차별을 예방하고, 대응하기 위한 실

---

* 이 글은 저자의 박사학위 논문인 “고위험 인공지능시스템의 기본권 침해에 관한 헌법적 연구” 중 제3장 제1절 및 제4장을 ‘고위험 인공지능의 차별’이라는 주제에 맞게 보완 · 발전한 것입니다.

** 서강대학교 법학연구소 선임연구원, 법학박사

효성 있는 법적 근거의 마련이 절실하다. 특히, '고위험' 영역에 활용되는 인공지능시스템의 정의를 분명하게 규정할 필요가 있다.

우선 이 글에서는 인공지능시스템의 잠재적 위험을 고려한 위험 기반 접근방식에 따라 기존의 편향과 차별을 야기하여 평등권을 침해할 수 있는 인공지능을 고위험 영역으로 분류한다. 이러한 고위험 인공지능은 온라인 플랫폼 등을 통해 일상에서 매일 활용되고 있다고 해도 과언이 아닐 정도로 활용도가 높으며, 다양한 의사결정 과정에서 적극적으로 활용되면서 평등권 침해에 밀접한 영향을 미치고 있다. 따라서 이와 같은 고위험 인공지능의 차별문제에 대해 헌법적 해명이 필요한 시점이라고 본다. 이 글에서는 고위험 인공지능시스템의 차별에 관한 구체적인 사례를 바탕으로 평등권 침해 문제점을 살펴보고, 외국의 인공지능 규범을 바탕으로 적절한 개선방안을 도출해 낼 것이다.

무엇보다도, 고위험 인공지능시스템의 활용으로 인한 불합리한 차별을 방지하기 위해서는 공정하고 신뢰할 수 있는 인공지능을 구현하기 위한 규범적 노력이 필요하다. 우선, 고위험 인공지능에 관한 비교법적 경향을 살펴보면 OECD와 UNESCO는 인공지능의 신뢰성과 공정성 원칙을 구현하기 위한 윤리원칙을 통해 인공지능이 준수해 나가야 할 방향성을 제시하였다. 또한, EU 「인공지능법안」과 미국 「알고리즘 책임법안」은 차별 등 기본권 침해를 야기할 우려가 큰 '고위험' 영역을 구체적으로 정하고, 인공지능 사업자 및 사용자 등에 대한 구체적인 의무를 규정하고 있다. 이와 같은 외국의 법제는 인공지능의 편향과 차별로부터 기본권을 보호하는 데 참고할 부분이 적지 않다.

끝으로 고위험 인공지능시스템의 차별을 방지하기 위해서는 인공지능의 신뢰성과 투명성 원칙이 구현될 수 있어야 한다. 이를 위해서는 데이터와 알고리즘이 편향이나 부당한 차별을 내포하지 않도록 사업자와 사용자의 감독의무가 필요하며, 고위험 인공지능시스템의 운용기록에 대한 정보접근권 보장 및 인공지능 영향평가 도입을 진지하게 논의할 필요가 있다.

• 핵심어 고위험 인공지능시스템, 편향과 차별, 인공지능 영향평가, 지능정보사회, 정보기본권

Ⅰ. 들어가며

오늘날 지능정보사회에서 인공지능은 인간의 인지능력과 유사하게, 스스로 학습하고 판단하면서 종래 과학기술과는 차원을 달리[1]하며, 거듭 진화해 나가고 있다.

이러한 인공지능 기술의 발달을 바탕으로, 인공지능은 다양한 영역에서 활용되는데, 사적 영역에서는 채용이나 업무평가 및 업무할당, 신용평가를 비롯하여 자동추천 알고리즘 등을 일상에서 쉽게 접하고 있으며, 공적 영역에서는 자동적 행정처분이 도입되었으며, 재판 업무 지원을 돕는 인공지능이 추진되고 있다.

그러나, 인공지능 활용에 따른 다양한 우려가 발생하고 있다. 즉, 객관적인 데이터를 기반으로 결론을 도출하는 인공지능이 부당한 차별로부터 보호해 줄 것이라는 기대와는 달리 오히려 인공지능이 기존의 차별적 성향이나 판단을 강화할 수 있어 문제가 된다. 예를 들어, 금융회사는 대출 신청자에 대한 신용평가에서 인공지능을 활용할 경우, 대출 신청자의 상환능력과 직접적인 관련이 없는 데이터를 판단 요소로 포함함으로써 차별을 야기할 수 있다. 특히, 데이터에 수집된 성별 · 인종 · 연령 등에 따라 부당한 대우를 하는 것은 불합리한 차별이 될 수 있으며, 이는 곧 평등권 침해가 문제 될 수 있다.

이처럼 불합리한 차별을 포함한 중대한 기본권을 침해할 우려가 큰 고위험 인공지능시스템에 대해서는 강도 높은 법적 규율이 요구되고 있다. 그 밖에도 고위험 영역에 활용되는 인공지능은 인간의 건강과 안전 등 본질적인 기본권을 위협할 수 있으며, 민주주의와 법치주의 등 헌법적 가치를 침해할 우려가 커지고 있다.

이 글은 그와 같은 문제의식에서 출발하여, 아래 제2장(Ⅱ)에서는 고위험 인공지능의 출현에 따른 우려와 고위험 인공지능의 정의를 살펴본다. 다음으로, 제3장(Ⅲ장)에서는 고위험 인공지능의 구체적인 활용 사례를 바탕으로 평등권 침해 문제를 규명하고, 제4장(Ⅳ)에서는 고위험 인공지능을 규율하는 주요 국가의 법규범에 대

1) 김광수, "인공지능 과학기술과 행정법학", 『서강법률논총』 11-1호, 서강대학교 법학연구소, 2022, 49면.

하여 비교법적으로 분석한다. 마지막으로, 제5장(Ⅴ)에서는 인공지능 차별로부터 평등권을 보호하기 위한 실효성 있는 개선방안을 제시해 보고자 한다.

Ⅱ. 고위험 인공지능시스템의 출현과 차별 문제

1. 고위험 인공지능시스템의 출현

인간은 오랜 시간 동안 차별과 불평등에 대항하였다. 인공지능을 포함한 ICT 기술의 발달로 사회 곳곳에서 자동화된 의사결정이 보편화된 오늘날에도 차별과 불평등은 여전히 존재하며, 인공지능이 보다 객관적이고 합리적인 판단을 할 것이라는 기대와는 다르게 기존의 사회적 편향과 차별이 강화되는 문제가 계속해서 나타나고 있다.

또한, 인간의 고정관념이나 편향은 인공지능시스템에도 그대로 반영되어 다양한 생활영역에서 활용되면서 예상하지 못한 부정적인 영향이 야기되고 있다. 특히, 인공지능 기술이 거듭 발전해 나갈수록 인간의 기본권과 법적 지위에 부정적인 영향을 미치는 고위험 인공지능시스템에 대한 우려가 커질 수밖에 없다. 특히, 고위험 인공지능시스템이 야기하는 또 다른 문제점은 사회적 취약계층이나 소수자에 대한 편향이나 차별을 강화할 수 있는 점이다. 무엇보다도, 현대 디지털사회에서 정보의 신뢰성 및 의사결정의 투명성이 강조되고 있음에도, 인공지능의 차별적인 판단에 대한 구체적인 설명을 요구할 수 있는 권리가 마련되어 있지 않다.

따라서 인공지능 차별을 포함한 중대한 기본권 침해 위험이 높은 고위험 인공지능시스템의 대상 및 분류 기준을 정하고, 고위험 인공지능 사업자 및 사용자 등 모든 이해관계자가 준수해야 할 원칙에 관한 법적 규율 방안을 마련할 필요가 있다.

2. 고위험 인공지능시스템의 정의

위와 같이 고위험 인공지능시스템의 출현으로 인한 편향과 차별은 중대한 기본권 침해로 이어지고 있으며, 국가 대 사인 간 혹은 사인 간 법률 분쟁을 야기하고 있다. 고위험 인공지능시스템은 위험도를 구분하여 각 위험 등급별로 강한 규제가 적용될 수 있기 때문에, 사업자와 사용자 등 다양한 인공지능 이해관계자에 따라 고위험에 대한 정의에 관하여 다양한 의견이 있을 수 있다. 따라서 고위험 인공지능시스템의 중대한 기본권 침해에 대응하기 위해서는 '고위험'의 범위를 법률에 규정할 필요가 있다.

이러한 관점에서 유럽연합(EU)이 입법을 추진하고 있는 「인공지능법안」은 인공지능의 위험성에 따라 단계별로 구분하는 위험 기반 접근을 기반으로, 인공지능 활용이 제한되는 금지 영역과 고위험 및 제한된 영역으로 구분하고 있다. 동 법안에 따르면 인간의 안전이나 생명에 명박한 위협이 되거나 인간의 행동을 조작하기 위하여 고안된 알고리즘 등은 허용되지 않는다. 다음으로, 민주주의 및 법치주의, 알 권리[2], 재판에 대한 권리에 중대한 영향을 초래할 수 있는 인공지능 등을 고위험군으로 분류하도록 규정하고 있다.

그 밖에도 고위험 인공지능시스템의 분류에 있어서는 후술할 EU「인공지능법안」과 같이 인공지능시스템의 사용 목적과 사용 대상도 함께 고려할 필요가 있으며, 인공지능 시스템이 건강이나 안전에 대해 미칠 수 있는 위험성이나 기본권 침해 가능성 등도 고려할 필요가 있다.

그 밖에 EU「인공지능법안」뿐만아니라 2022년 미국 상·하원 의회에서 각각 발의된「알고리즘 책임법안」도 참고할 내용이 적지 않다. 동 법안은 인간에게 중대한 영향을 미칠 수 있는 중요한 결정에 관여하는 인공지능을 이른바 '강화된 중요 결정절차'로 정의하면서, 평가나 인증을 포함한 교육 및 직업훈련이나 근로자 관리

2) 유럽연합 의회는 2023년 5월 EU「인공지능법안」에 대한 수정안을 통해 정치 캠페인이나 소셜 미디어의 자동추천시스템을 활용하여 유권자에 영향을 미칠 수 있는 인공지능을 고위험군으로 추가하였다.

또는 주택담보대출이나 여신 등 금융서비스 및 법률서비스, 전기나 수도 등 필수 인프라에 활용되는 인공지능시스템을 고위험으로 분류하고 있다. 동 법안의 고위험 인공지능시스템의 정의도 EU 인공지능법안과 전반적으로 매우 유사한 내용을 포함하고 있다.

이러한 비교법적 경향을 바탕으로, 불합리한 차별로 평등권을 침해하는 인공지능시스템을 고위험 영역으로 분류하고자 한다. 그 밖에도 지능정보사회에서 다양한 온라인 플랫폼이나 알고리즘 활용에 따라 알 권리 및 재판청구권 침해를 야기할 수 있는 인공지능도 고위험 영역에 해당한다고 판단한다.

3. 고위험 인공지능시스템의 차별 유형과 문제점

인공지능시스템에 의한 차별은 알고리즘 개발 및 제조 과정에서 개발자의 차별의도 유무에 따라 의도적 차별과 비의도적 차별로 분류할 수 있다. 여기서 의도적 차별과 비의도적 차별은 직접차별과 간접차별에 각각 연결된다.

먼저 의도적 차별이란 문언 그대로 개발자가 알고리즘을 고안하는 단계에서 차별의도가 내포되어 차별적인 결과를 도출한다는 것을 뜻한다. 예를 들어, 금융회사에서 여신업무에 사용하는 알고리즘이 여성 고객의 여신한도를 남성 고객에 비하여 합리적인 근거 없이 낮게 평가하도록 고안하였다거나 로봇이나 인공지능 비서에게 그 역할에 부합하는 특정 성별을 부여하는 경우가 해당될 수 있다.

이와 반대로, 비의도적 차별이라 함은 개발자가 알고리즘을 고안할 때 직접적으로 차별을 의도하지는 않았지만, 차별을 의도한 것과 다르지 않게 차별이 발생하는 경우를 의미한다. 특히, 이와 같은 간접차별은 인공지능 개발자나 사용자의 차별의도가 없는 경우에도 데이터의 결함 등으로 차별이 발생하는 고위험 인공지능시스템에서 더욱 문제가 되고 있다. 예를 들어, 채용 과정에서 활용된 알고리즘이 여성 지원자를 탈락시키지는 않았지만, 남성 위주로 채용되었던 기존의 데이터를 학습한 알고리즘이 자기소개서를 통하여 채용 응시자를 여성으로 추론하여 부당하게 탈락

시킨 사례가 보고되었다.

이에 관하여 개발자가 의도하지 않은 알고리즘의 차별은 다음과 같은 원인이 있을 수 있다. 첫째, 인공지능의 학습 데이터의 부족으로 소수자에 대한 대표성이 결여되는 것이다. 둘째, 알고리즘 내부의 학습 데이터에 편견이 내포되는 경우이다. 셋째, 알고리즘으로 인하여 어떠한 특정한 현상이 또 다른 현상에 영향을 주는 '스필오버 효과(Spillover effect)[3]'로 인해 뜻하지 않은 차별이 발생하는 경우이다. 예를 들어, 알고리즘 자체는 성적 편향이 반영되지 않았지만 광고시장 등에서 다른 광고와 경쟁하면서 편향적 광고만 노출[4]되는 경우가 해당될 수 있다.

이처럼 현대 사회에서 인공지능 기술의 발달에 따라 알고리즘 개발자의 차별의도가 반영되지 않았더라도 차별 문제가 발생할 수 있다. 특히, 의도하지 않은 알고리즘 차별은 차별 피해자가 차별 원인에 관한 충분한 설명을 받을 수 없고, 인공지능 사업자가 차별에 대한 입증책임을 부담하지 않음으로써 차별로 인한 손해 및 원상회복에 능동적으로 대응하기 어려운 문제가 있다. 이하에서는 고위험 인공지능시스템이 야기한 차별 문제에 관한 구체적인 사례를 중심으로 문제점을 살펴보고자 한다.

Ⅲ. 고위험 인공지능시스템의 평등권 침해 사례

1. 인종 차별

(1) 경찰 안면인식 알고리즘

미국 애틀랜타에서 2022년 한 흑인 남성이 절도 혐의로 체포되었다. 경찰은 루이

3) 스필오버효과(Spillover effect)는 특정 현상이 다른 현상에 파급효과를 끼치는 것을 의미한다. 네이버 지식백과 참조, 홈페이지 https://terms.naver.com/search.naver?query=%EC%8A%A4%ED%95%84% EC%98%A4%EB%B2%84(최종검색일: 2024.01.10.)

4) 한애라, "인공지능과 젠더차별", 『이화젠더법학』 제11권 제3호, 이화여자대학교 젠더법학연구소, 2019, 5면.

지애나주의 뉴올리언스에 위치한 한 상점에서 발생한 가방 절도사건을 수사하면서 안면인식 알고리즘을 활용하여 CCTV에 찍힌 가방 절도 용의자가 흑인 남성의 운전면허증 사진과 동일 인물로 판단하였다. 경찰은 해당 남성에 대한 지명수배를 내렸고, 그를 체포한 후 구치소에 구금하였다. 그러나, 해당 남성은 루이지애나주에 방문한 적조차 없었으며, 결백함을 입증할 수 있는 명백한 알리바이를 경찰에 제시하여 구금 6일 만에 석방될 수 있었다. 그 후 해당 남성은 경찰에 대해 직권남용 및 불법감금 등의 혐의로 조지아주 애틀랜타 연방법원에 소송을 제기하였다.[5)]

루이지애나주 경찰은 2019년 안면 인식 알고리즘을 수사에 활용하기 위하여 알고리즘 개발사 'Clearview AI'와 연간 사용료 2만5천불을 지급하는 계약을 체결하였고, 해당 업체는 소셜 미디어 등 온라인에서 수십억 장의 사진 스크랩을 바탕으로 안면인식 알고리즘을 제조하였다.

이 사건과 관련하여 AP통신은, 최근 몇 년간 미국에서 안면인식 오류 문제로 인하여 법집행 기관을 상대로 소송을 제기한 흑인이 다수 있으며, 경찰이 활용하는 안면인식 기술이 흑인 등 유색인종을 용의자로 오판하는 비율이 더 높다는 사실을 밝혔다.[6)] 이에 미국의 한 인권단체는 경찰의 안면인식 알고리즘이 "인종차별 치안"라고 비판하면서, 경찰의 알고리즘이 흑인에 대한 차별 문제를 더욱 악화시키는 우려를 표명하였다.[7)] 이 사건을 안면인식 통하여 미국 경찰의 안면인식 알고리즘의 오판 문제가 알려지면서 흑인 등 유색인종에 대한 편향과 차별 문제가 본격적으로 제기될 수 있었다.

---

5) 조선일보, "AI가 인종차별? 美 흑인남성 억울한 옥살이한 이유가", 2023년 9월 26일 자, https://www.chosun.com/international/international_general/2023/09/26/D334UXOO6BAH5DPRCSWLXS2B2M/(최종검색일: 2024.01.10.)

6) AP통신, "Facial recognition tool led to mistaken arrest, lawyer says", 2023년 1월 3일 자, https://apnews.com/article/technology-louisiana-baton-rouge-new-orleans-crime-50e1ea591aed6cf14d248096958dccc4(최종검색일: 2024.01.10.)

7) The New York times, "'Thousands of Dollars for Something I Didn't Do'", 2023년 3월 31일 자, https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html(최종검색일: 2024.01.10.)

(2) 페이스북(Facebook) 맞춤형 광고 알고리즘

온라인 플랫폼은 인공지능을 활용하여 고객들의 관심사와 취향에 부합하는 다양한 맞춤형 광고 서비스를 제공하고 있다. 예를 들어, 온라인 맞춤형 타겟광고(Online Behavioral Advertisement, 이하 'OBA')가 소비자의 온라인 활동을 수집하여 머신러닝 분석을 바탕으로, 맞춤형 광고를 제공하는 것이 대표적인 사례이다.

그런데, OBA는 고위험 인공지능으로 분류되는 고용 영역, 즉 기업의 채용 광고에 있어서도 인종과 성차별을 유발할 수 있다. 가령, 페이스북(Facebook) 온라인 광고가 인종과 성별에 따라 편향적으로 노출된다는 실증연구가 밝혀졌다. 즉, 목재 산업 분야의 채용 광고는 약 70%가 백인에게 노출되었고, 마트 계산원 채용광고는 약 85%가 여성에게, 택시 운전사 채용 광고는 약 75%가 흑인에게 노출된 것으로 드러났다.

그뿐만 아니라 주택 광고에서도 인종에 따른 편향이 드러났다. 가령, 주택 매매 광고는 백인에게 더 노출되었으며, 임대 광고는 흑인에게 상대적으로 더 노출된 사실이 밝혀졌다.[8] 결국, 미 주택도시개발부(The US Department of Housing and Urban Development('HUD'))은 페이스북이 『공정주택법(Fair Housing Act)』을 위배하고, 광고 플랫폼을 통하여 인종, 성별 및 종교에 따른 차별을 조장하였다는 이유로 소송을 제기하였다.[9] 그 밖에도 OBA는 소비자의 프라이버시 침해 및 프로파일링(Profiling) 등으로 소비자에 대한 차별위험을 내포하고 있는 것으로 지적되었다.[10]

---

8) Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Rieke, A.(2019), "Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes", 199:4면.

9) Hao, K.(2019. 4. 5.), "Facebook's ad-serving algorithm discriminates by gender and race", MIT Technology Review, https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/(최종검색일: 2024.01.10.)
덧붙여서 페이스북은 맞춤형 광고의 차별행위에 대하여 다음과 개선안을 마련하였다. 즉 광고주는 고용, 주거, 신용 분야에 있어서 성별과 연령 등에 따라 기회를 제한시키는 등의 차별을 하지 않겠다는 대안을 제시하였다. https://www.aclu.org/blog/womens-rights/womens-rights-workplace/facebook-settles-civil-rights-cases-making-sweeping(최종검색일: 2024.01.10.)

10) Wachter, S.(2019.05.15.), "Affinity Profiling and Discrimination by Association in Online

(3) 비자 승인 알고리즘

영국 내무부는 2015년 비자 승인처리 업무처리에 인공지능 활용을 도입하였다. 즉, 해당 인공지능 알고리즘은 비자 신청자가 제공한 개인정보에 국적 정보를 포함하였는데, 국가의 위험성에 따라 각 신청자에게 녹색 · 황색 · 적색 등 색상 코드를 부여한 것으로 확인되었다.

그런데, 영국의 한 시민단체는 해당 인공지능시스템이 백인보다는 비백인 인구가 많은 특정 국가 출신의 비자 신청자들의 심사가 지연되거나, 관계 당국이 합리적 이유 없이 비자 신청을 거절한 문제를 지적하였다. 또한, 동 시민단체는 영국정부의 비자 승인 알고리즘이 국적을 기준으로 사실상 인종 차별을 야기하였으며 평등법을 위반하였다고 지적하면서, 해당 알고리즘이 기존의 편향과 차별을 반영함으로써 차별을 더욱 강화하였다고 주장하였다.[11] 이에, 영국의 관할 당국은 해당 시민단체의 문제 제기에 관하여, 비자신청 처리시스템의 작동 과정을 점검하겠다는 계획을 밝히며, 인공지능 활용을 중단하기로 결정하였다.[12]

2. 성 차별

(1) 아마존(Amazon) 채용 알고리즘

빅테크 기업 아마존(Amazon)은 머신러닝 기술을 기반으로 한 인공지능을 활용하여 채용응시자를 선별하였다.[13] 그런데, 아마존(Amazon)이 사용한 알고리즘이 소프

Behavioural Advertising", Berkeley Technology Law Journal vol.35 no.2, 2020 Forthcoming, 4-13면.

11) 김용철, SBS 뉴스, "[취재파일] AI에 의한 입시부정?…'학력 평가 알고리즘'에 뿔난 영국 학생들", 2020년 8월 27일 자, https://news.sbs.co.kr/news/endPage.do?news_id=N1005950655(최종검색일: 2024.01.10.)

12) BBC, "Home Office drops 'racist' algorithm from visa decisions", 2020년 8월 4일 자, https://www.bbc.com/news/technology-53650758(최종검색일: 2024.01.10.)

13) THE VERGE, "Amazon reportedly scraps internal AI recruiting tool that was biased against women", https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report (최종검색일: 2024.01.10.)

트웨어 개발이나 기술 담당자 채용에 응시한 여성에 대한 차별이 문제가 되었다.

해당 알고리즘은 과거 10년간 채용 응시자의 이력서를 학습하였고, 알고리즘은 채용 응시자들에게 각각 1~5점 사이의 점수를 부여[14]하였다. 그런데, 아마존의 기술 담당자는 남성 직원이 압도적으로 많은 비율을 차지하여, 해당 알고리즘이 남성 지원자를 우대하는 방향으로 학습하는 문제가 발생하였다. 즉, 채용 응시자의 이력서에 '여자 체스클럽 주장'으로 기재되었거나, '여대 졸업'과 같이 '여성'이 기재된 이력서에는 알고리즘이 낮은 점수를 부여한 것으로 밝혀졌다. 결국, 아마존은 채용 과정에서 여성 차별 논란에 직면하게 되었고, 채용 알고리즘 개발 프로젝트를 전면적으로 중단[15]하였다.

(2) 애플카드(Apple Card) 알고리즘

인공지능의 성차별은 금융사의 신용 평가에서도 문제가 되었다. 미국의 골드만삭스(Goldman Sachs)의 '애플카드(Apple Card)'는 신용한도 책정에서 인공지능을 활용한 대표적인 사례이다. 문제가 된 것은 배우자와 자산 및 계좌를 공유하는 고객 중에서, 남성 고객보다 여성 고객이 현저하게 낮은 신용한도를 부여받은 것으로 밝혀졌다. 이에 뉴욕주 금융당국(DFS)은 애플카드가 활용한 인공지능의 성차별에 관한 조사를 착수하였다.

1) 사실의 개요[16]

애플과 골드만삭스는 2019년 8월 신용카드 '애플 카드'를 출시하였다. 그런데, 한 남성 고객은 배우자와 공동 세금 신고서를 제출하는 등 재산을 공동으로 관리하고 있음에도, 배우자보다 약 20배 가까이 높은 신용한도를 제공받은 사실을 지적하였

14) THE VERGE, 위의 기사.
15) 한애라, 앞의 논문, 13면.
16) New York State Department of Financal Services(2021), "Report on Apple Card Investigation", 4-5면.

다. 또한, 애플의 공동창업자 역시 애플카드의 성차별 논란을 비판하면서, 자신의 아내와 재산을 공동으로 관리하지만, 본인의 카드 신용한도가 배우자보다 약 10배 높은 사실을 들며, 카드사의 성차별 문제를 제기하였다. 또한, 카드사 고객들은 해당 알고리즘의 성 차별적 판단을 비판하면서, 해당 알고리즘이 고객의 신용한도에 관한 구체적인 설명을 제공하지 않는 문제를 지적하였다.

2) 대상사건의 문제점

위의 문제에 관하여 뉴욕주 금융당국은 카드 고객의 신용평가에 있어서 남녀차별 없이 동등한 대우를 한 것으로 판단하여 성차별을 인정하지 않았다.

그러나 뉴욕주 금융당국은 그러한 결론을 내리면서도, 카드사의 알고리즘의 구체적인 판단근거와 알고리즘의 안전성을 확인하지 못했다. 또한, 카드사의 신용점수 부여 및 신용평가 방법 등에 관한 상세한 정보를 공개하지 않았다.

이와 같은 문제점과 유사한 내용으로서, 미국 『신용기회평등법(Equal Credit Opportunity Act)』은 모든 신용거래에서 성별 · 종교 · 연령 · 인종 · 혼인 여부 등에 따라 신용 차별을 금지하고 있다. 특히, 동법의 취지에 따라 개발된 FICO 점수[17]는 대출자의 부채와 청구서 납부 기록 등 재정 상태에 관한 변수만을 고려하여 신용평가점수를 산정한다.[18]

그러나 현실에서는 기업들이 위와 같은 공적인 신용평점 외에, 다양한 빅데이터를 확보하여 소비자들의 신용평가에 반영하고 있는 실정이다. 실제로 미국 사기업이 평가하는 신용점수에는 소비자의 재정 상태뿐만 아니라 거주지와 투표 여부, 흡연 여부 등 내밀한 개인정보를 수집하여 신용평가에 반영하고 있다.[19] 이러한 신용평가점수는 공적인 신용평가 도구인 FICO와는 다르게, 편향성을 내포하는 데이터에 대해서 아무런 규제를 받지 않으며, 신용평가에 관한 기업의 정보공개 의무가

17) FICO는 미국의 페어아이잭코프(Fair Isaac Corporation)사가 제공하는 개인신용평가 서비스를 뜻한다.
18) 한애라, 앞의 논문, 15면.
19) 한애라, 위의 논문, 15면.

없기[20]때문에 소비자가 불합리한 차별에 대응하기 어려운 문제가 있다.

해당 사례와 같이 인공지능이 생활 영역에 깊숙이 파고들면서, 성별에 따라 불합리한 차별적 판단을 하는 문제가 계속해서 제기될 것으로 예견된다. 또한, 소비자들에게 알고리즘의 판단 원리에 관한 투명한 정보 접근이 제공되지 않는 한, 데이터를 반복적으로 학습하는 인공지능의 특성에 따라 인공지능의 편향과 차별이 강화될[21] 우려가 있다.

한편, 우리나라의 금융사에서 활용하는 인공지능의 차별문제에 대응하여 정부(금융위원회)는 2021년 7월 「금융분야 AI 운영 가이드라인」을 제정하였는데, 금융회사 등이 개인에 대한 불합리한 차별을 비롯하여 개인의 권익 및 자유에 대한 중대한 위험을 초래할 수 있는 서비스를 고위험 서비스로 분류하였다. 또한, 금융 업무에서 인공지능을 활용할 경우에는 적절한 내부통제 활동 및 승인 절차를 마련하고, 승인 책임자를 지정하도록 규정[22]하였다.

3. 소득에 따른 차별

성적 예측 알고리즘

국제사회는 2020년 코로나19 팬데믹으로 사회적 혼란에 직면하였다. 특히, 영국은 학교 수업을 진행할 수 없는 상황에 직면했고, 모든 국가시험을 취소하는 결정을 내렸다.

---

20) 캐시 오닐, 「대량살상수학무기」, 『흐름출판』, 2016, 237-267면.

21) 한애라, 위의 논문, 17면.

22) 「금융분야 AI 운영 가이드라인」(금융위원회)

2. 거버넌스의 구축

(가~다.) 생략

라. 금융회사 등이 개인에 대한 부당한 차별 등 개인의 권익와 안전, 자유에 대한 중대한 위험을 초래할 수 있는 서비스(이하 '고위험 서비스'라 한다.)에 대해 AI 시스템을 활용하는 경우, 적절한 내부통제 활동 및 승인절차를 마련하고, 승인 책임자를 지정한다, 승인 책임자는 책임있는 업무 수행이 가능한 지위로 하되 최고위험관리책임자, 신용정보보호·관리인, 최고정보보호책임자 등 유사 업무와 겸직할 수 있다.

이에 따라 2020년 3월 예정되었던 대학 입학고사 'A-Level' 시험이 취소되었고, 영국 잉글랜드 · 웨일스 · 북아일랜드 3개 자치정부 지역의 고교 졸업반 30여만 명은 졸업시험에 해당하는 '에이(A)-레벨 시험'을 치르지 못했다.[23] 이에, 영국의 대입 시험 감독기관인 Ofqual(Office of Qualifications and Examinations Regulation)[24]은 시험을 실시하지 않는 대신에 기존 데이터에 근거하여 학생들에게 적합한 학점을 부여하는 '직업센터학업모델(The Direct Centre Perforamance Model)'라는 알고리즘을 이용하여 대입에 반영한다는 계획을 발표하였다.

문제의 성적 예측 알고리즘은 교사들이 예측한 학생들의 성적, 기존 모의고사 성적, 학교 내신과 같은 학생 개인적 요소뿐만 아니라 학생이 소속된 학교와 지역의 기존 입시 성적 등과 같은 외부 요소들을 함께 반영하였다. 그런데, 이 성적들은 영국 자격시험관리국과 교사가 예측한 수치보다 약 40%정도 저하된 성적을 보였다.

특히, 부유층 가정 출신 학생의 입학비율이 높은 사립학교와 공립학교 간의 평가에서도 차별이 드러났다. 즉, 사립학교에서는 성적 예측 알고리즘을 통한 학업평가를 실시하기 전인 2019학년도에 비해 종합학교보다 상위권 성적을 내는 학생의 비율이 증가하였다. 또한, 부유층 가정 출신의 학생이 많은 사립학교 학생의 약 49%가 A 등급 이상을 받은 반면에 공립학교 학생은 약 22%가 같은 등급을 받았다.

이처럼 해당 인공지능시스템은 공립학교, 사회 취약계층, 이민자 등 특정 집단에 속한 학생들의 성적을 낮게 평가하는 차별적 판단이 문제되면서 해당 알고리즘은 모두 폐기되었다.[25]

---

23) 한겨레, "AI가 준 학점, 가난한 학생들 차별했다", 2020년 8월 24일 자., https://www.hani.co.kr/arti/international/europe/959055.html(최종검색일: 2024.01.10.)

24) Ofqual은 영국에서 대입 시험 등을 담당하는 시험 감독청이다.

25) The Guardian, "Who won and who lost: when A-levels meet the algorithm", 2020년 8월 13일자., https://www.theguardian.com/education/2020/aug/13/who-won-and-who-lost-when-a-levels-meet-the-algorithm(최종검색일: 2024.01.10.)

Ⅳ. 인공지능의 평등권 보호에 대한 외국의 법제 동향

이상에서는 인공지능 차별에 관한 주요 사례를 살펴보았다. 현재 국회에 계류 중인 인공지능법안은 EU「인공지능법안」이 규정하는 고위험 인공지능시스템의 정의를 비롯하여 인공지능 사업자 및 사용자 등에 대한 규제 방안을 상당 부분 참고한 것으로 보인다. 한편, EU「인공지능법안」외에도 미국의「알고리즘 책임법안」은 고위험 인공지능시스템의 차별 등을 포함한 부정적 영향을 예방하는 데 필요한 유의미한 내용을 담고 있다.

본 장에서는 고위험 인공지능의 차별을 방지하기 위한 입법에 관하여 국제기구 및 주요 국가에서 논의된 입법례를 중심으로 비교법적 경향을 검토하고자 한다.

다만, 그와 같은 논의에 앞서 현재 국제사회는 인공지능 기술에 대한 주도권 확보를 위해 치열한 경쟁을 벌이며, 인공지능이 국제질서의 지각 변동에 적지 않은 역할을 차지하는 점을 이해할 필요가 있다. 특히, 미국 등 주요 국가는 인공지능을 활용한 안보 강화와 경제성장을 위해 기술과 자원을 결집[26]하고 있다. 이러한 측면에서 EU「인공지능법안」은 차별 등 기본권 침해를 위협하는 고위험 인공지능에 대한 규제뿐만 아니라, 미국의 인공지능 기술 패권을 견제하기 위한 유럽 국가들의 목적이 있음을 숙지할 필요가 있다.

이하에서는 국제기구(OECD · UNESCO)의 대응과 유럽연합(EU)과 미국이 규정하는 인공지능에 활용에 따른 평등권 보호에 관한 내용을 중심으로 고위험 인공지능시스템의 차별에 관한 비교법적 분석을 하고자 한다.

---

26) 지능정보사회진흥원, "디지털 기술 주도권 확보 전략 및 대응 방안-미국 · 중국 · EU 정책 분석을 중심으로-", 2019, 1면.

1. 국제기구

(1) OECD, 「OECD AI 권고안」

1) 제정 배경

본 권고안이 제정되기 전까지 인공지능에 관한 윤리적 접근은 주요 선진국과 글로벌 빅테크 기업을 중심으로 논의되어 왔다. 가령, 유럽연합(EU)은 개별적인 지침을 마련하여 인공지능 윤리를 규정하거나 구글과 마이크로소프트 같은 빅테크 기업에서 내부 지침형식으로 인공지능 윤리 원칙을 마련하였다.

그와 같은 인공지능 윤리 원칙은 강제규범이 아닌 기업 등의 자율적 준수에 의지하고 있으며, 인공지능의 투명성과 공정성, 신뢰성 등 인공지능 윤리 및 기본권 보호를 위한 법률이 지금까지 제정되지 않아, 각 국가와 기업 등 인공지능 사업자 및 사용자가 인공지능 규제에 관한 구체적인 원칙과 책무를 준수하기에 한계가 있었다.

이에 OECD는 2019년 5월 인공지능이 야기하는 다양한 기본권 침해에 대비하고 지속가능하고 신뢰할 수 있는 인공지능 구현을 위한 'OECD AI 권고안(Council Recommendation on AI)'을 OECD 각료이사회에서 채택[27]하였다. 동 권고안은 선언(Declaration)보다 강한 효력을 가지는 권고(Recommendation)의 형식으로[28] 구성되었으며, 국제사회에서 최초로 마련된 인공지능 규범이라는 점에서 의의가 있다. 즉, OECD 소속의 주요 선진국이 참여하여 마련한 인공지능 원칙은 유럽연합과 미국이 인공지능법을 제정하는 데 있어 단초가 되었을 것으로 보인다. 특히, OECD가 2013년 개인정보 보호 지침을 마련한 후, 미국과 유럽에서 「개인정보보호법」을 제정하는 데 영향을 주었던 점을 상기해 보면, 동 권고안은 향후 인공지능법안을 제정하는데 있어 단단한 토대가 될 것으로 예견된다.[29] 한편, 동 권고안은 미국 ·

27) 한국정보화진흥원, "OECD 인공지능 권고안", 『TTA저널』 187호, 2020년, 3면.
28) 홍석한, "미국 "2022 알고리즘 책임법안"에 대한 고찰", 『미국헌법연구』 제34권 제1호, 2023, 100면.
29) 동아일보, 'OECD "인공지능 활용 원칙 권고안 만장일치 채택"', 2019년 5월 27일 자.

영국 · 독일 등 OECD 36개 회원국과 아르헨티나 · 브라질 · 콜롬비아 · 코스타리카 · 루마니아 등 비회원국 42개 국가가 참여하였는데, 동 권고안의 구체적인 내용을 살펴보면 다음과 같다.

2) 평등권 보호를 위한 인공지능 원칙

OECD의 동 권고안은 인공지능시스템의 활용에 따른 부당한 차별을 방지하기 위하여 몇가지 원칙을 이행하도록 권고하였다.

먼저, 인공지능 이해관계자는 사회적 소수집단에 대한 포용력을 진전시키고, 경제적 · 사회적 · 성차별 및 사회적 불평등을 제거하도록 하였다. 또한, 인공지능에 직접 또는 간접적으로 관련이 있거나 인공지능 활용에 의해 영향을 받는 모든 사업자와 사용자 등 이해관계자는 신뢰할 수 있는 인공지능 구현을 위한 관리 감독의 책임이 있다는 점을 선언하였다.

다음으로, 인공지능시스템을 직접 운영하는 조직과 개인 등 인공지능의 활동 주체[30]는 인공지능시스템 개발부터 활용 단계를 포함한 인공지능의 전 수명주기 동안 인권 및 민주적 가치를 준수할 것을 권고하였다. 여기에는 인간의 존엄과 평등 및 차별 방지를 위해 노력하여야 하며, 사회적인 다양성 보호를 위하여 노력할 것을 천명하였다. 동 권고안은 인공지능의 차별 방지뿐만 아니라 인공지능 활용에 따른 소수자 보호를 최초로 언급한 대목에서 주목할 만하다.

---

https://www.donga.com/news/article/all/20190526/95707383/1(최종검색일:2024.01.06.)

30) 동 권고안의 인공지능 '활동 주체'(AI actors)의 정의는 인공지능시스템을 활용하는 조직이나 개인을 포함하며, 인공지능시스템 수명주기 동안 적극적인 역할을 수행하는 이들을 의미한다고 정의하였다. 반면, 인공지능시스템으로부터 영향을 받는 조직 및 개인은 '이해관계자'로 분류된다.

(2) UNESCO, 「인공지능 윤리에 관한 권고」

1) 제정 배경

유네스코(UNESCO)는 2019년 11월 유네스코 총회에서 「인공지능 윤리에 관한 권고(Recommendation on the ethics of artificial intelligence)」를 마련하였다. 특히, 인공지능 윤리에 관한 국제표준을 마련하기 위하여 193개 유네스코 회원국이 인공지능 윤리규범을 마련한 점에서 의의가 있다. 동 권고는 위에서 살펴본 OECD 권고안이 제시한 인공지능 차별에 관한 내용을 보다 구체화하였다.

동 권고문이 제출된 배경으로는 코로나 팬데믹으로 인한 전 세계의 사회·경제적 영향으로 인하여 국가 간 협력이 보다 강조되었고, 국제 사회에서 인공지능 기술의 도입이 급속히 확대되면서 차별 등 부정적인 영향이 보고됨에 따라 인공지능 규제에 대한 목소리가 커진 것이다.

UNESCO 권고안도 OECD 권고안과 유사하게 인공지능 거버넌스 논의에 있어 다양한 사회구성원의 참여를 기초로 하고 있다. 즉, 인공지능 거버넌스에 대한 포용적 접근방향으로 정부와 민간기업, 연구자뿐만아니라 사회 소수자에 대한 참여를 보장하도록 한 점이 주목된다.

생각건대 인공지능의 편향과 차별을 방지하고, 신뢰성과 공정성을 보장하기 위해서는 인공지능 정책수립 과정에 있어서도 다양한 사회구성원의 참여가 강조되어야 할 것으로 보인다. 동 권고안은 인공지능의 차별과 불평등을 방지하도록 주문하면서, OECD의 인공지능 권고안에서 제시한 사회적 다양성과 소수자 보호 등 헌법적 기본가치를 보호하는 데 기여할 수 있었다고 평가한다. 동 권고안의 평등권 보호에 관한 주요 내용을 살펴보면 다음과 같다.

2) 평등권 보호를 위한 인공지능 원칙

동 권고안은 인공지능의 기본가치와 윤리원칙을 명시하여, 공정하고 신뢰할 수

있는 인공지능을 구현할 수 있도록 다음의 내용을 규정하였다.

우선, 동 권고안의 전문은 인공지능시스템이 편향된 정보를 학습함으로써 야기할 수 있는 차별과 불평등 및 정보격차 심화 등을 경계하면서, 인간의 존엄과 양성 평등 및 사회적 다양성 보호의 필요성을 강조하였다. 이를 위하여 인공지능은 사회적 정의를 증진하고 국제법에 따라 차별금지 원칙을 준수하도록 권고하고, 인공지능시스템의 차별적이고 편향된 알고리즘 결정에 대하여 효과적인 구제 방안을 마련하도록 하였다. 또한, 인공지능 서비스 제공에 있어서 문화적 다양성이 반영된 인공지능 서비스를 제공하여야 하며, 인공지능 개발 전 단계에서도 개발자의 다양성을 보장하도록 하였다.

다음으로, 동 권고안은 인권과 인간의 존엄에 대한 보호를 강조하면서, 인공지능시스템의 전 수명주기 전반에서 실현되어야 함을 규정하였다. 또한, 동 권고안은 인간의 존엄의 의미에 대하여 인종과 성별, 종교, 국적, 사회적 배경과 및 장애여부 등에 관계없이 모든 인간이 가지는 본질적으로 동등한 가치를 존중하는 것에서 비롯된다는 것을 확인하였다. 또한, 인공지능의 활용에 있어서도 다양성을 보장하도록 강조하였는데, 인종과 성별, 연령, 종교, 출신배경, 장애 등에 관계없이 모든 인간의 적극적인 참여를 보장함으로써 평등권을 보호하기 위한 기본원칙을 제시하였다.

2. 유럽연합(EU), 「인공지능법안」

(1) 제정 배경

앞에서 살펴본 국제기구(OECD · UNESCO)의 권고안은 인공지능시스템의 투명성과 공정성 구현에 대한 원칙을 확립하고, 인공지능 차별을 방지하기 위한 이정표 역할을 하였지만, 강행규정이 아닌 권고적 효력에 그칠 수밖에 없는 한계가 있었다.

따라서 인공지능의 급속한 성장에 따라 예측하지 못했던 부정적 영향과 기본권 침해를 방지하고, 인공지능이 보다 안전하게 활용되기 위해서는 관련 문제점을 구

체적으로 규율할 수 있는 법 제정을 통하여 인공지능의 신뢰성 원칙을 구현하기 위한 규제가 필요하다.[31] 특히, 'GPT' 등 생성형 인공지능의 출현으로 인하여 편향과 차별, 프라이버시권 침해 등 기본권에 대한 위협이 심화되었다. 유럽연합은 그와 같은 문제에 대한 헌법적 공감을 기초로 하여, 고위험 인공지능을 중심으로 강한 규제를 위한 지속적인 논의를 하였다.

유럽연합 집행위원회는 2018년 학계와 기업, 시민사회 출신의 고위전문가 그룹(High-Level Expert Group on Artificial Intelligence, AI HLEG)을 구성하여 인공지능 기본권 침해와 투명성과 신뢰성 제고 등에 관한 논의를 시작하였고[32], 2019년 신뢰할 만한 인공지능을 위한 윤리 가이드 라인을 발표[33]하였다. 나아가 EU 집행위원회는 2020년 2월 인공지능에 대한 백서를 발표하면서 차별을 비롯한 공정한 재판을 받을 권리, 표현의 자유 등 인공지능의 기본권 침해를 지적하고, EU가 추구하는 기본권적 가치가 인공지능에 반영될 수 있도록[34], 인공지능 법률 제정에 대한 논의가 본격적으로 진행되었다.

마침내 유럽연합은 2021년 4월 인공지능에 관한 유럽 내 일치된 규범을 정립하기 위한 EU 인공지능법안 초안(Proposal for a regulation laying down harmonisesd rules on artificial intelligence, Artificial Intelligence Act, 이하 'EU 인공지능법안')을 마련하였다.

동 법안은 유럽 의회와 유럽연합 이사회 및 집행위원회의 합의[35]를 거쳐 제정될 수 있을 것으로 보이며, 우리나라의 고위험 인공지능시스템의 개발과 활용에 있어서 기본권 침해를 방지하기 위한 참고서가 될 수 있을 것으로 판단한다.

---

31) 홍석한, "유럽연합 '인공지능법안'의 주요 내용과 시사점", 『유럽헌법연구』 제38호, 유럽헌법학회, 2022, 274면.

32) 유럽연합(EU), https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance.(최종검색일: 2024.01.10.)

33) 박혜성 · 김법연 · 권헌영, "인공지능 통제를 위한 규제의 동향과 시사점", 『정보법학』 제25권 제2호, 한국정보법학회, 2021, 7면.

34) 김진우, "유럽연합의 인공지능 백서에 관한 고찰", 『외법논집』 제44권 제4호, 한국외국어대학교 법학연구소, 2020, 152면.

35) 이재훈, "EU의 입법절차", 『법제연구』 제61호, 한국법제연구원, 2018, 52면.

(2) 인공지능시스템의 분류

EU 인공지능법안은 인공지능시스템에 내포된 위험성을 기준으로, 다음과 같이 인공지능의 위험등급을 분류하고 있다.

첫째, '수용할 수 없는 위험(Unacceptable risk)'로서 중대한 기본권 침해를 야기할 수 있는 위험이 있는 인공지능시스템은 사용이 금지된다. 가령 인간의 잠재의식을 조정하는 기술을 통하여 사람의 신체나 정신에 해악을 끼치는 등 행동을 조작하는 기술과 아동이나 신체적 또는 정신적 장애가 있는 사람의 취약성을 이용하여 행동을 조작하는 인공지능기술은 허용되지 않는다. 또한, 국가 또는 공공기관이 사람의 인격적 특성이나 사회적 행동을 바탕으로 개인을 평가하기 위한 인공지능시스템 역시 금지된다. 가령, 중국의 사회신용점수 제도가 대표적인 사례로서 단순히 개인의 신용 상태에 대한 평점 부과를 넘어 교통법규 위반, 상거래 내역[36] 등 일상생활에 관한 정보를 수집하여 점수를 부여하는 방식으로 평가하는 것은 금지된다.

둘째, '고위험(High risk)'으로서 앞에서 살펴본 경찰 안면인식 알고리즘처럼 수사기관의 사실판단이나 형사사법절차에서 재범 예측 등에 활용되는 인공지능시스템이 대표적인 예이다. 이 밖에도 인력 채용이나 직업훈련기관에서의 교육대상자 선발, 실시간 원격 생체인식시스템, 수도 · 전기 · 가스 등 중요 인프라 관리와 운영 등에 관한 인공지능시스템도 고위험으로 분류된다.

셋째, '제한된 위험영역(limited risk)'으로서 사람과 대화 등을 통해 상호작용하도록 고안된 인공지능시스템이나 감정인식시스템 및 생체정보 기반 범주화시스템 등이 포함된다. 뿐만 아니라 사람이나 사물 등이 실존하는 것처럼 기망하는 이미지나 동영상, 음성 등을 생성하거나 조작하는 시스템 이른바 '딥페이크'가 포함된다.

넷째, '최소 위험(Minimal risk)'으로서 위 세 가지 분류에는 들어가지 않는 기타 인공지능시스템이 모두 포함된다. 가령, 광고성 스팸메일을 분류하는 등 사람에게

36) 이승은, "중국 사회신용시스템의 현황 및 전망: '빅브라더'와 빅데이터", 대외경제정책연구원, 2017, 2면.

미치는 위험성이 거의 없는 인공지능시스템이다.

(3) 평등권 보호 관련 내용

우선, EU「인공지능 법안」은 유럽연합이 추구하는 인간의 존엄과 평등권, 표현의 자유 등 기본권 보호 가치를 바탕으로 한다. 특히, 동 법안의 전문에서는 인공지능시스템을 규율하는 공통적인 규범이 유럽연합의 기본권 헌장에 부합하여야 하는 점을 선언하였다. 이를 통하여 인공지능시스템의 개발과 활용에 있어서 인간의 생명과 안전 및 기본권 보장 등 헌법적 가치를 수호하고자 하는 점을 알 수 있다. 또한, 이러한 가치를 보호하기 위하여 표준규범의 제정 필요성을 강조하였다.

동 법안에서 규정된 평등권 보호에 관한 내용을 살펴보면, 동 법안의 전문에서는 인공지능시스템의 발전에 따라 다양한 분야에서 활용되는 동시에 인공지능 기술이 오용될 수 있는 위험이 상존하는 점을 강조하면서, 조작과 착취를 야기하며 사회적 통제 실행을 위한 도구가 될 수 있는 위험성을 지적하였다. 이러한 문제에 대하여 실효성 있는 대응을 위해서는 인공지능의 위험단계에 따른 위험기반접근 방식에 따라 필요한 규범의 내용을 조정하여야 하고, 고위험 인공지능시스템의 요건과 인공지능 운영자의 의무 및 투명성 의무를 명시하여야 한다고 규정하였다. 이러한 원칙을 바탕으로 동 법안에서는 인공지능시스템의 평등권 침해로부터 보호하기 위하여 다음과 같은 규범적 보호방안을 마련하였다.

첫째, 고위험 인공지능시스템에 대한 인간의 감독을 규정하였다. 즉, 고위험 인공지능시스템의 제조과정에 있어서 인간이 감독할 수 있는 방식으로 설계 및 개발할 것을 원칙으로 하였다(동법안 제14조 제1항). 특히, 고위험 인공지능시스템의 작동상태를 모니터링함으로써 미연에 발생할 수 있는 기본권 침해 등에 대비하고, 문제점이 발생할 경우 지체 없이 해결할 수 있어야 한다고 규정하였다. 가령, 인공지능시스템의 '정지' 버튼을 통하여 고위험 인공지능시스템의 작동을 언제든지 중단할 수 있어야 한다는 것이다. 다음으로 고위험 인공지능시스템의 특성을 고려하여 고

위험 인공지능시스템의 산출물을 올바르게 해석할 수 있도록 한다는 점을 규정하였는데(동 법안 제14조 제4항), 이러한 내용은 설령 인공지능시스템에 차별 의도가 반영되지는 않았더라도, 인공지능의 판단에 관한 인간의 편향된 해석으로 미연의 차별을 방지하기 위한 것으로 이해할 수 있다.

둘째, 고위험 인공지능시스템 사업자가 준수해야할 의무를 규정하였다. 여기서 사업자는 인공지능시스템을 개발하거나 서비스를 제공하는 자연인이나 법인 및 기관 등을 의미한다. 고위험 인공지능시스템 사업자가 준수하여야 할 주요 내용을 살펴보면, 고위험 인공지능시스템이 준수하여야 할 품질관리체계 확보 의무를 규정하면서(동 법안 제16조), 고위험 인공지능의 개발과 품질관리에 사용하는 기술과 절차에 관한 지침을 문서로 마련하도록 하였다. 또한, 인공지능 데이터 관리에 필요한 시스템과 절차, 특히 데이터 수집과 데이터 분석 등에 관한 내용이 포함되도록 하였다(동 법안 제17조).

이 밖에도 고위험 인공지능시스템 사업자는 '기술문서' 작성을 통하여 고위험 인공지능시스템이 갖춰야 할 요건을 준수하고 있는지 입증하여야 하고, 해당 요건의 준수여부에 관한 평가정보를 관할 당국과 인증기관에 제공하도록 하였다(동 법안 제18조). 또한, 고위험 인공지능시스템 사업자는 인공지능의 시장 출시 혹은 서비스 개시 전에 인증기관에 의한 품질관리체계 평가와 기술문서에 대한 평가를 바탕으로 진행하는 적합성 평가를 하여야 한다(동 법안 제43조). 적합성 평가내용은 동 법안에 규정된 고위험 인공지능시스템에 적용되는 요건으로서, 고위험 인공지능시스템의 잠재적인 위험성을 분석하고, 기본권과 안전에 미칠 수 있는 영향과 편향을 완화하기 위한 조치여부, 데이터의 적절성 등에 관한 평가가 포함된다.

이러한 적합성평가를 거치고 난 이후에 고위험 인공지능시스템이 필요한 요건을 준수한 것으로 확인될 경우에는 사업자는 유럽연합 적합성 선언서를 작성(동 법안 제48조)하고, 'CE 마크'를 부착하여야 한다(동 법안 제19조). 즉, 이러한 과정은 제품의 품질과 안전을 확인하는 인증마크를 부착하는 것과 유사한 것으로서, 고위험 인공지능시스템에 대한 신뢰성과 안전성을 담보하는 데 유용한 기능을 할 것으로

판단한다.

셋째, 고위험 인공지능시스템 사용자에 대한 의무를 규정하였다. 여기서 사용자는 자체 권한을 바탕으로 인공지능시스템을 사용하는 자연인이나 법인, 공공기관, 기관, 그 밖의 기구를 뜻하고, 비전문적인 사적활동을 수행하는 사용자는 제외된다.

고위험 인공지능시스템 사용자는 인공지능이 기본권을 침해할 수 있는 가능성에 유의하고, 인공지능시스템의 데이터가 본래 목적과 관련성을 가지는지 확인할 의무가 있다. 즉, 사용자는 고위험 인공지능시스템에 데이터를 입력함으로써 특정한 판단이나 결과물을 제공받는데, 이러한 인공지능의 판단이 본래의 사용 목적 범위 내에서 활용되는지를 점검하여야 한다. 또한, 사용자는 동 법안에서 정하는 사용 안내에 따라 고위험 인공지능시스템이 본래의 목적에 따라 운영되고 있는지 감시해야 한다. 따라서 인공지능시스템이 인간의 건강이나 안전 또는 기본권을 침해할 수 있다면, 사용자는 사용을 중지하고, 인공지능 사업자 또는 유통업자에게 그 내용을 고지하여야 한다(동 법안 제29조 제4항). 이 밖에도 사용자는 고위험 인공지능시스템의 운용기록을 보관하여야 하는데, 미연에 발생할 수 있는 인공지능시스템의 기본권 침해에 대한 원인을 사후적으로 진단하는데 유의미할 것으로 판단한다.

3. 미국 「알고리즘 책임법안」[37]

(1) 제정배경

미국의 인공지능 정책은 EU의 강한 규율과는 달리, 그보다 상대적으로 완화된 규율방식을 취하였는데, 인공지능 산업 육성을 위한 국가적 차원의 인공지능 개발 지원 등에 초점이 맞춰져 있었다. 이에 따라 인공지능 활용에 따른 기본권 침해 등을 방지하기 위한 실질적 보호 측면에서는 다소 한계가 존재할 수밖에 없었다.

특히, 미국의 글로벌 기업에서 인공지능이 적극적으로 활용되면서, 그에 따른 차

37) The Algorithmic Accountability Act of 2022.

별 문제가 심화되었고, 인공지능의 공정성과 투명성을 제고하기 위한 윤리규범을 논의하기 시작하였는데, 미국 의회와 정부에서 2019년 인공지능으로부터 기본권 침해를 보호하기 위한 다양한 입법이 추진되었다. 미 하원의회에서는 2019년 「알고리즘 책임법안」[38]이 최초로 발의되었으나 임기만료로 폐기되었고, 그 후 미 상원의회[39]와 하원의회[40]는 2022년 「알고리즘 책임법안」을 동시에 발의하였다. 2022년 「알고리즘 책임법안」은 2019년 법안과 유사한 구조이나 약간의 차이가 있는데, 인공지능을 '자동 결정 시스템(Automated Decision System)'과 '강화된 중요 결정 절차(Augmented Critical Decision Process)'로 구분[41]하는 점이다.

한편, 미국의 「알고리즘 책임법안」의 기본방향은 고위험 인공지능시스템을 보다 구체적으로 한정하여 규제하는 방식으로 접근하며, 구체적인 규율은 연방통상위원회(FTC)에 위임함으로써 EU 인공지능법안보다는 유연한 규제방식으로 볼 수 있다.[42] 한편, 동 법률안은 알고리즘의 투명성과 책임성을 확보를 위하여 강화된 중요 결정 절차에 대한 영향평가를 실시하도록 규정하는 점에서 주목할 만하다.

인공지능 활용에 따른 문제는 다양하나, 데이터의 편향으로 인한 차별적인 판단, 프라이버시 침해 등 중대한 기본권 침해가 주로 문제된다. 이러한 문제에 대하여 영향평가를 통해 인공지능 사업자가 사용자, 이해관계자 및 인공지능 감독기관에 인공지능에 관한 필요한 정보 제공 여부, 인공지능의 결정이나 판단에 따른 다양한 영향에 대한 사전적인 분석과 평가[43] 등을 기대할 수 있다. 이를 통해 인공지능의

---

38) 116th Congress (2019-2020), "H.R.2231.-Algorithmic Accountability Act of 2019", https://www.congress.gov/bill/116th-congress/house-bill/2231/text,(최종검색일: 2024.01.10.)

39) 미 상원의원 Ron Wyden, Cory Booker는 2019년에 발의되었던 알고리즘 책임 법안을 보완하여 재발의하였다. 이 법안에 관한 자세한 내용은 아래에서 참고할 수 있다. 117th Congress (2021-2022), "H.R3572. - Algorithmic Accountability Act of 2022". 홈페이지 https://www.congress.gov/bill/117th-congress/senate-bill/3572/text(최종검색일: 2024.01.10.)

40) 117th Congress (2021-2022), "H.R.6580.-Algorithmic Accountability Act of 2022", https://www.congress.gov/bill/117th-congress/house-bill/6580(최종검색일: 2024.01.10.)

41) 김광수, "인공지능 알고리즘 규율을 위한 법제 동향-미국과 EU 인공지능법의 비교를 중심으로-", 『행정법연구』 제70호, 행정법이론실무학회, 2023, 186면.

42) 김광수, 위의 논문, 190면.

43) 지능정보사회진흥원, "미국, '알고리즘 책임법안(Algorithmic Accountability Act)'발의", 『디지털 법제 Brief』 제22-01호, 2022, 1면.

투명성 원칙을 구현하고, 특히 고위험 인공지능의 오판에 대한 책임소재를 보다 명확하게 확인할 수 있을 것으로 보인다.

(2) 고위험 인공지능시스템의 정의 및 평등권 보호 관련 내용

1) 고위험 인공지능시스템의 정의

동 법률안에서 기본권 보호에 관한 내용을 살펴보기에 앞서, 고위험 영역의 인공지능시스템에 대한 정의를 살펴볼 필요가 있다. 우선, 동 법안은 '고위험'이라는 용어를 직접적으로 사용하는 것 대신에 '자동 결정 시스템(Automated Decision System)'과 '강화된 중요 결정 절차(Augmented Critical Decision Process)'로 고위험 인공지능을 구분하고 있다. 다만, 본 연구에서는 평등권 등 중대한 기본권 침해를 야기하는 인공지능시스템을 '고위험'에 해당하는 것으로 전제하기 때문에, 편의상 고위험 인공지능시스템으로 분류하기로 한다.

먼저, '자동결정시스템'은 결정 또는 판단의 기초가 되는 모든 소프트웨어로서, 머신러닝과 통계학 및 데이터 처리 및 인공지능 기술이 포함된다. 여기에는 인공지능시스템 사용자의 관심사나 취향을 파악하여 맞춤형 정보를 제공하는 자동추천시스템이 포함된다.

다음으로, '강화된 중요 결정 절차'는 인간에게 중대한 영향을 미칠 수 있는 '중요 결정'에 있어서 자동화된 의사결정 시스템을 활용하는 모든 과정과 활동 등이 포함된다. 여기서 '중요 결정'이라 함은 소비자의 생활에서 법적, 사실적[44] 또는 이와 유사한 중대한 영향을 미치는 결정과 판단을 의미하는데, 다음과 같이 구체적으로 열거하고 있다. ① 평가, 인증을 포함한 교육 및 직업 훈련, ② 고용, 노동자 관리 또는 자영업, ③ 전기, 난방, 수도, 인터넷 등 통신망 또는 교통을 포함한 필수공익사업, ④ 입양 서비스 또는 출산 관련 서비스를 포함한 가족계획, ⑤ 주택담보 등 금융회사 회사, 금융 중개업자 또는 대부업자에 의한 여신 등 금융서비스, ⑥ 의료

44) 김광수, 위의 논문, 186면.

서비스, ⑦ 임대 또는 숙박시설 등 주거 관련 서비스, ⑧ 민간 조정 또는 중재를 포함한 법률 서비스, ⑨ 연방통상위원회(FTC)가 소비자의 생활에 법적, 실질적 또는 이와 유사하게 중대한 영향을 미치는 서비스, 프로그램 및 기회에 관한 결정 등이 포함된다.

2) 평등권 보호 관련 내용

동 법안에서 평등권 보호에 관한 내용을 살펴보면, '자동 결정 시스템'이나 '강화된 중요 결정 절차'에서 발생할 수 있는 편향과 차별적 판단을 방지하기 위하여 소비자의 인종 · 성별 · 성별 · 연령 · 종교 · 장애 · 가족형태 · 사회경제적 배경 등에 따른 차별적 판단 여부를 평가하도록 하였다. 또한, 해당 인공지능시스템의 판단과정에서 데이터가 미치는 영향에 관한 정보를 확인할 수 있도록 문서화할 것을 규정하였다.

다음으로 동 법안에 따라 인공지능시스템이 불합리한 차별을 야기하거나 개인정보 침해 등에 관한 영향평가 실시 등을 준수하지 않을 경우에는 관할 당국의 법집행이 가능하다. 가령, 인공지능 시스템으로 인하여 국민이 위협을 받거나 불리한 대우를 받을 경우에 주 검찰총장은 피해자의 법적 구제를 위하여 당사자를 대리하여 민사소송을 제기할 수 있다(SEC. 9(b)(1)). 또한, 주의 검찰총장은 필요한 경우 관련 조사를 수행할 수 있으며, 이해관계인에게 출석 또는 문서 제출 등을 명령할 수 있다(SEC. 9(b)(3)).

생각건대 고위험 인공지능시스템의 기본권 침해에 대해서 당사자들의 법적 대응이 쉽지 않은 현실을 고려할 때, 국가가 인공지능 차별 문제를 지원하는 것은 당사자의 피해회복을 돕고, 미연의 인공지능 차별에 대한 사전 예방적 효과를 거둘 수 있을것으로 판단한다. 또한, 동 조항과 같이 국가가 인공지능으로부터 예상치 못한 손해를 입은 당사자를 대리하는 것은 앞에서 살펴본 EU 「인공지능법안」에서는 찾아볼 수 없는 유의미한 특징이기도 하다.

Ⅴ. 고위험 인공지능시스템의 평등권 침해에 대한 개선방안

인공지능시스템의 개발과 활용에 있어서 공정성과 투명성 원칙의 구현은 무엇보다 중요하다고 할 수 있다. 먼저, 공정성 원칙은 고위험 인공지능시스템이 사회적 편향이나 고정관념을 학습하지 않도록 개발하기 위한 인공지능 윤리 구현의 이정표이자, 차별을 방지하기 위한 인공지능의 주요 원리이다. 또한, 투명성 원칙은 기본권 침해를 비롯한 부정적 영향을 방지하고, 인공지능에 관한 정보제공을 실현하기 위한 핵심 가치로 이해할 수 있다. 이처럼 인공지능의 두 가지 기본원칙의 구현과 관련하여 평등권 보호를 구체화하기 위한 다음과 같은 개선방안을 제언한다.

1. 평등권 보호를 위한 기본전제

우선, 평등권을 위협하는 인공지능시스템은 고위험으로 분류하여 강화된 규제가 필요하다. 오늘날 민주사회에서 기본권 보장의 목적은 차별과 불평등을 제거해 불합리한 차별을 방지하고, 보다 정의로운 사회를 실현하기 위함으로 이해할 수 있다. 특히, 사회적 통합을 위한 전제로서 평등권 보호가 더욱 강조된다.[45]

제4장(IV)에서 살펴본 바와 같이, 유럽과 미국은 인공지능시스템의 잠재적인 위험성을 기준으로 고위험 영역에 대한 규제를 강화하고 있으며, 차별과 불평등을 야기할 수 있는 인공지능시스템을 고위험 영역에 포함하여 헌법상 평등권을 보호하기 위한 강한 법적규율을 추진하고 있다.

예를 들어, EU「인공지능법안」은 인력 채용이나 업무평가 등 고용 영역이나, 교육 · 직업훈련 등에 관한 교육영역, 대출 · 신용평가 등 금융서비스 영역 및 재범 가능성 판단과 범죄예측 등 법 집행 영역을 고위험 영역으로 규정하고 있다. 또한, 동 법안은 새로운 고용형태로 출현한 플랫폼 노동자에 대한 업무할당뿐만 아니라 일반적인 고용관계에서 특정 인종이나 특정 사회적 배경 등을 가진 피고용인에 대

45) 명재진 · 이한태, "평등권 이론에 관한 현대적 전개로서 적극적 평등실현조치",『법학연구』제18권 제2호, 충남대학교 법학연구소, 2007, 19면.

한 차별을 유발할 수 있는 인공지능시스템을 고위험 영역으로 규정하고 인공지능 사업자에게 보다 강한 책임을 부담하도록 한다.

미국「알고리즘 책임법안」도 그와 유사하게 인력 채용이나 근로자 관리 등 고용 영역, 평가나 직업훈련 등 교육영역, 대출 등 금융영역, 전기 · 수도 · 통신 등 중요 인프라 영역에서 인공지능시스템을 통한 의사 결정을 '중요 결정'으로 규정하였다. 즉, 해당 영역에서 활용되는 인공지능시스템을 일종의 고위험 영역으로 분류하는 것으로 볼 수 있으며, 그 위험에 따른 구체적인 요건을 준수하도록 한다.

현재 우리나라는 플랫폼 기업 등 사적영역을 중심으로 자동화된 의사결정이 주로 활용되고 있지만, 공적영역에서도 인공지능 기술을 활용한 자동화된 행정을 도입[46]하였다. 즉,「행정기본법」 제20조[47]는 '자동적 처분'을 도입하면서 행정청이 인공지능 기술을 적용한 완전히 자동화된 시스템으로 행정처분을 할 수 있는 법적 근거를 마련하였다. 또한, 대법원은 '스마트법원 4.0'을 통하여 인공지능을 활용한 판례검색 및 판결문 작성 등을 추진하고 있다.

그런데, 인공지능에 의한 차별과 편향은 이용자의 법적지위에 미치는 영향이 크고, 차별에 따른 원상회복이 쉽지 않아 더욱 문제된다. 따라서 현대 지능정보사회에서 차별이 야기되거나 법적지위에 영향을 미칠 수 있는 영역에 대한 실효성 있는 규제가 필요하고, 앞에서 살펴본 EU「인공지능법안」과 미국「알고리즘 책임법안」에서 평등권을 보호하도록 규정하는 바와 같이, 차별을 유발할 수 있는 인공지능시스템을 고위험으로 영역으로 분류할 필요가 있다. 이를 통하여 고위험 인공지능시스템 사업자와 사용자 및 그 밖의 이해관계자가 인공지능 차별을 분명하게 인식하고, 인공지능 개발과 사용에 있어서 미연의 차별을 방지하는 데 기여할 수 있을 것이라고 판단한다.

---

46) 김광수, "인공지능 과학기술과 행정법학",『서강법률논총』제11권 1호, 서강대학교 법학연구소, 2022, 52면.

47)『행정기본법』제20조(자동적 처분) 행정청은 법률로 정하는 바에 따라 완전히 자동화된 시스템(인공지능 기술을 적용한 시스템을 포함한다)으로 처분을 할 수 있다. 다만, 처분에 재량이 있는 경우는 그러하지 아니하다.

2. 구체적 개선방안

(1) 사업자 및 사용자의 의무

고위험 인공지능시스템의 차별을 방지하기 위해서는 인공지능 사업자[48] 및 사용자[49]에 대해서 다음과 같은 법적 책임과 의무를 부과할 필요가 있다.

첫째, 인공지능시스템의 데이터와 알고리즘이 부당한 편향을 내포하거나 차별을 야기하지 않도록 사업자의 의무를 부과하여야 한다. 즉, EU「인공지능법안」제16조에서 고위험 인공지능시스템의 데이터와 알고리즘이 준수하여야 할 품질보증 요건을 규정하는 내용을 확인하였다. 또한, 동 법안 제17조는 품질관리 체계를 바탕으로, 고위험 인공지능시스템의 개발과정부터 활용까지 잠재적 편향을 방지하기 위한 감독 의무를 규정하고 있는데, 사업자의 책임을 강화하기 위한 유의미한 수단이라고 본다.

다음으로, 고위험 인공지능시스템이 준수하여야 할 요건 등 안정성에 관한 입증책임을 사업자에게 부과할 필요가 있다. 특히, 편향과 차별 여부에 대한 사업자의 감독 의무 및 안전성에 관한 입증책임은 평등권 침해를 방지하는 데 중요한 보호수단이 될 것으로 예상된다. 특히, 인공지능의 차별 문제는 그 차별의도가 분명하지 않은 간접차별에서 더욱 문제되는 점을 고려할 때, 인공지능으로부터 불합리한 차별을 당한 피해자로 하여금 입증의 부담을 완화하는 데 의미가 있다고 본다.

둘째, 인공지능시스템을 활용하는 사용자의 의무도 규정할 필요가 있다. EU「인공지능법안」제29조에서 사용자의 의무를 규정하는 것처럼, 인공지능 사용자가 인공지능시스템의 활용 안내에 따라 그 운영을 감독하고, 편향이나 차별을 유발할 수

48) 인공지능시스템 사업자는 인공지능시스템을 시장에 출시 또는 인공지능을 활용한 서비스를 제공하는 자연인이나 법인, 공공기관 및 그 밖의 기구를 포함한다(Artificial Intelligence Act Article3(2)).

49) 인공지능시스템 사용자의 정의는 EU 인공지능법안 수정안의 '배포자(Deployer)'와 동일한 개념으로(각주 54), 개인적 비전문적 활동 과정에서 인공지능시스템을 사용하는 경우를 제외하고, 인공지능시스템을 사용하는 자연인이나 법인, 공공기관 또는 그 밖의 기관을 뜻한다(Artificial Intelligence Act Article3(4)).

있다고 판단될 경우에는 사업자에게 그 내용을 고지하고, 사용을 중단하도록 해야 할 것이다.

더 나아가, 인공지능시스템의 활용에 따라 제3자가 차별을 받았다고 판단될 경우에는 그 당사자에게도 차별에 관한 내용을 고지하도록 하여야 한다. 특히, 인공지능시스템의 사업자와 사용자가 사실상 동일한 경우가 적지 않은 점을 고려하면 사업자뿐만 아니라 사용자에게도 인공지능 차별을 방지하기 위한 능동적인 감시자로서 윤리적 책무를 다하는 데 도움이 될 것으로 판단한다.

(2) 고위험 인공지능시스템의 운용기록에 대한 정보접근권

인공지능 기술이 거듭 진화할수록 성별이나 출신지, 혼인 여부 등 민감한 개인정보가 반영된 데이터를 기초로 하면서, 고위험 인공지능시스템의 평등권 침해가 더욱 문제가 될 것으로 예견된다.

위에서 언급한 것처럼, 평등권을 위협하는 인공지능시스템에 대하여 이해관계자가 분명한 법적 대응을 하지 못하는 이유는 인공지능시스템의 데이터와 알고리즘의 불투명성에서 기인한다. 이러한 문제는 알 권리 침해와도 관련이 있는데, 인공지능 사업자는 영업비밀을 사유로 구체적인 알고리즘 정보공개를 꺼리고 있는 실정이다.

따라서 인공지능 차별에 대한 법적 대응을 하기 위해서는 데이터 및 특정한 판단을 도출하도록 설계된 알고리즘에 관한 정보를 기록하고, 판단과정을 추적할 수 있도록 시스템 처리과정을 저장하도록 할 필요가 있다. 특히, 딥러닝 기술을 활용하는 인공지능시스템의 경우 상당한 수량의 매개변수를 활용[50]하기 때문에, 인공지능시스템에 관한 정보처리 기록을 보관하지 않는다면, 인공지능 차별에 대한 법적 대응 및 원상회복은 사실상 기대하기 어려운 점을 유념할 필요가 있다. 따라서 이러한 '블랙박스' 문제를 방지하기 위한 사전 예방 수단이 강조된다. EU「인공지능법안」제12조는 고위험 인공지능시스템의 운용과정에 발생하는 모든 판단이 자동으

50) 박도현, "인간 편향성과 인공지능의 교차",『서울대학교 法學』제63권 제1호, 서울대학교 법학연구소, 2022년, 166면.

로 저장되도록 하고, 인공지능시스템의 개발부터 시판 후 활용단계까지의 전체 수명주기 동안, 정보처리를 수행한 이력에 대하여 추적 가능할 수 있도록 문서화할 것을 규정하고 있다.

물론, 동 조항이 차별을 원천적으로 방지할 수는 없으나, 인공지능시스템의 모든 데이터와 알고리즘의 판단과정이 시스템에 기록된다면, 차별에 관한 법적 분쟁이 발생할 경우 당사자가 해당 정보를 열람할 수 있을 것이다. 또한, 인공지능시스템이 도출한 판단 이유에 관하여 이해할 수 있도록 문서화한다면, 해당 시스템을 개발하는 사업자 역시 차별을 방지하기 위한 다양한 조치를 취할 수 있을 것으로 기대된다.

(3) 인공지능 영향평가

고위험 인공지능시스템 활용에 따른 편향과 차별을 감지하기 위한 법제도가 요구되며[51], 중대한 기본권 침해 및 사회적 영향을 조사하고 분석하기 위한 영향평가 제도 도입을 고려할 필요가 있다. 유럽연합(EU) 의회는 2023년 5월 고위험 인공지능시스템을 활용하여 서비스를 제공하는 배포자[52]에게 인공지능 영향평가를 수행하도록 의무를 부과하는 수정안을 제출하였다. 특히, 동 법안은 고위험 인공인공지능시스템의 배포자에게 다음과 같은 의무를 준수하도록 하였는데, ① 고위험 인공지능시스템 사용 시 예상되는 기본권에 대한 영향, ② 취약계층에 발생할 수 있는 위해 가능성, ③ 기본권에 대한 피해와 부정적 영향을 완화하기 위한 세부 계획, ④ 인간에 의한 감독 방안을 포함한 AI 거버넌스 구축 계획 등을 포함하도록 하였다.[53]

---

51) Sandra Wachter 외 2인, "WHY FAIRNESS CANNOT BE AUTOMATED: BRIDGING THE GAP BETWEEN EU NON-DISCRIMINATION LAW AND AI", *Computer Law & Security Review* Volume 41, Elsevier, 2021, 48면.

52) '배포자(Deployer)'는 인공지능법안 초안에서 'User'로 사용된 개념으로, 2023년 5월 수정안을 통해 명칭이 변경되었다. 즉, 배포자는 개인적 비전문적 활동 과정에서 인공지능시스템을 사용하는 경우를 제외하고, 인공지능시스템을 사용하는 자연인이나 법인, 공공기관 또는 그 밖의 기관을 뜻한다(Artificial Intelligence Act Article3(4)).

53) Artificial Intelligence Act Article 29a

이와 유사하게, 국가인권위원회는 2022년 '인공지능 개발과 활용에 관한 인권 가이드라인'을 제시하면서 인공지능시스템 활용에 따른 차별 등을 보호하기 위한 인권영향평가를 권고한 바 있다. 특히, 인권위는 동 가이드라인을 통하여, 인공지능이 인간의 존엄성 및 개인의 자율성과 다양성을 보장해야 하며, 인간의 존엄과 가치 및 행복을 추구할 권리에 부합하는 방향으로 개발·활용되어야 한다는 기본원칙을 천명하였다. 또한, 국가는 인공지능 개발 및 활용과 관련하여 인권침해나 차별을 방지하기 위한 인공지능 인권영향 평가제도를 마련할 것을 권고[54]하였다.

앞에서 살펴본 바와 같이, 국제사회에서도 인공지능시스템이 과연 인간에게 무해한 편익을 제공하는지 구체적으로 점검하고, 인권에 미치는 영향 등에 관한 인공지능 영향평가 제도 도입을 위한 법안 등을 추진하고 있다. 특히, UN 인권이사회는 2021년 10월 '디지털시대 프라이버시권 결의'[55]에서 인공지능시스템이 인권에 미치는 위협을 방지하기 위한 인권 실사를 도입을 결정하였으며, 국가나 기업이 활용하는 인공지능시스템의 전 수명주기 동안 인권실사를 권장한 바 있다.

일상의 다양한 영역에서 인공지능시스템 도입이 확산되면서 인공지능은 지능정보사회의 동반자가 되었다. 이러한 새로운 물결 속에서 인공지능의 신속성과 공정성을 제고하기 위해서는 고위험 인공지능시스템의 위험성을 정기적으로 평가하여 문제점을 예방하고, 사회적으로 미칠 수 있는 영향분석 및 다양한 이해관계자들의 의견을[56] 참고하여 개선방안을 도출하고 인공지능 정책에 반영할 수 있을 것이다.

끝으로 고위험 인공지능에 대한 영향평가가 안정적으로 실행되기 위해서는 법적 근거를 마련하여야 하고, 영향평가의 대상 및 주체, 평가 방법과 평가 내용, 평가 시기 등에 관한 사항을 법률에 규정할 필요가 있다.

---

54) 국가인권위원회 결정, 「인공지능 개발과 활용에 관한 인권 가이드라인」, 2022년 4월 11일, 6면.

55) Resolution adopted by the Human Rights Council on 7 October 2021, 48/4(Right to privacy in the digital age), https://digitallibrary.un.org/record/3985679(최종검색일: 2023.05.10.)

56) 사법정책연구원, "사법절차 및 사법 서비스에서 인공지능 기술의 도입 및 수용을 위한 정책 연구", 2021, 109면.

VI. 맺으며

이상에서 고위험 인공지능시스템의 차별에 관한 구체적인 사례를 살펴보고, 외국의 입법례와 우리의 법 현실을 바탕으로 구체적인 개선방안을 제시해 보았다. 고위험 인공지능은 인공지능 차별을 야기하는 주된 원인으로서, 미연의 편향과 차별로부터 보호하기 위해서는 고위험 인공지능의 데이터와 알고리즘에 대한 몇 가지 원칙이 강조되는데, 위에서 살펴본 논의를 정리하면 다음과 같다.

먼저, 고위험 인공지능시스템의 편향과 차별을 방지하기 위해서는 알고리즘을 개발하거나 활용하는 사업자 및 사용자 등 모든 인공지능 이해관계자가 준수해야 할 구체적인 의무를 마련할 필요가 있다. 특히, 인공지능의 차별을 방지하기 위해서는 해당 프로그램을 직접 제조하거나 활용하는 사업자가 인공지능의 데이터의 적합성과 대표성 등 필요한 요건을 갖추었는지 확인할 수 있는 감독 의무가 강조된다. 또한, 인공지능시스템을 활용하는 사용자 역시 활용 안내에 따라 운영을 감독하고, 법적 지위에 부정적 영향을 미치거나 기본권 침해를 야기할 수 있다고 판단될 경우에는 사업자가 해당 인공지능시스템 사용을 중단할 수 있도록 필요한 내용에 관한 고지의무를 부과하는 방안도 고려해 볼 수 있을 것이다.

다음으로, 고위험 인공지능시스템의 활용에 따른 사회적 소수자와 취약계층의 기본권과 법적 지위를 보호할 수 있도록 각별한 규범적 노력이 필요하다. 위에서 살펴본 바와 같이, 인공지능은 인종, 성별, 소득 수준 등에 따라 차별을 야기하고 있다. 특히, 고위험 인공지능의 차별 문제는 직접 차별을 의도하지 않았음에도, 데이터의 편향이나 대표성 부족 등으로 흑인이나, 저소득 취약계층이 불합리한 처우를 받거나 때로는 범죄 용의자로 지목돼 억울한 누명을 쓸 수 있는 사례가 확인되었다.

이에 대하여, 유럽연합과 미국은 모두 인공지능의 투명성과 책임성 원칙이 바람직한 인공지능사회를 위하여 공통으로 추구해야 할 기본적 가치[57]로 전제하면서, 인공지능 시스템이 기존의 편향을 강화해 불평등과 차별을 야기하고, 유색 인종이

57) Sandra Wachter 외 4인, "Artificial Intelligence and the 'Good Society': t he US, EU, and UK approach", *Science and Engineering Ethics* volume 24, Springer Science+Business Media, 2017, 523면.

나 사회적 취약 계층의 법적 권리를 침해할 수 있는 위험에 대한 구체적인 보호 방안의 필요성을 명확하게 인식하고 있다. 특히, 유럽연합의 「인공지능법안」과 미국의 「알고리즘 책임법안」은 고위험 인공지능의 출현에 따른 기본권 침해로부터 보호하기 위하여 다양한 법적 규율을 마련하고 있다. 유럽연합과 미국은 리스크 기반 접근방식을 통하여 고위험 인공지능을 중심으로 보다 강화된 조치를 취하는 점에서 유사한 측면이 있다.[58]

물론, 유럽과 미국의 인공지능법안은 고위험 인공지능의 대상 및 이해관계자의 의무와 책임 등 구체적인 규제 방식에 있어서 적지 않은 차이를 보인다. 우선, EU 「인공지능법안」은 인공지능시스템의 위험성을 중심으로 4가지 유형의 인공지능으로 분류하고, 고위험 인공지능에 대해서는 개발단계부터 활용 단계까지 보다 강한 규제를 하는 방식을 취한다. 이와 달리, 미국 「알고리즘 책임법안」은 고용과 근로자 관리, 및 직업 훈련, 법률 서비스, 금융 서비스 등에 활용되는 인공지능을 고위험 영역으로 구체적으로 열거하는데, 그 밖에도 다양한 영역에서 활용되며 중대한 기본권 침해를 위협하는 경우를 상기했을 때, 실효성 있는 규제 측면에서 한계가 있는 것으로 판단된다.

끝으로 고위험 인공지능시스템으로부터 평등권을 보호하기 위해서는 인간의 감독이 재차 강조된다. 앞에서 주요 국가의 인공지능법안에서 살펴보았듯이, 인공지능 기술이 거듭 진화해 나가면서, 법원의 재판 업무에서도 생성형 인공지능을 도입하는 사업이 추진되고 있다. 따라서 이와 같은 고위험 인공지능에 대한 적절한 규제를 마련하여 지능정보사회가 직면한 새로운 차별 문제에 더욱 적극적으로 대응해 나갈 때라고 본다. 이를 위해서는, 고위험 인공지능시스템 사업자 및 사용자가 데이터와 알고리즘이 준수해야 할 요건을 감독할 수 있도록 법적 근거를 마련하고, 데이터의 대표성과 편향 여부 등을 사전적으로 점검하기 위한 인공지능 영향평가 도입에 대해서도 진지한 논의가 필요한 때라고 본다.

▶ 논문투고일: 2024. 01. 15. 논문심사일: 2024. 01. 28. 게재확정일: 2024. 02. 05

58) 김광수, 앞의 책, 190면.

▌참고문헌

1. 국내문헌

고학수 외 2인, “인공지능과 차별”, 『저스티스』 제171호, 한국법학원, 2019.

고학수 외 2인, “유럽연합 인공지능법안의 개요 및 대응방안”, 『DAIG Magazine, 서울대학교 인공지능 정책 이니셔티브』 2021년 제2호, 2021.

국가인권위원회, 『2022 국가인권위원회 통계』, 2023.

_____, 『4차 산업혁명 시대에서 정보인권 보호를 위한실태조사』, 2018.

_____,, 『인공지능(AI) 개발과 활용에서의 인권 가이드라인 연구(최종보고서)』, 2021.

_____,, 『인공지능 인권영향평가 도입 방안 연구(최종보고서)』, 2022.

김광수, 『인공지능법 입문』, 내를건너서숲으로, 2021.

_____, “인공지능 과학기술과 행정법학”, 『서강법률논총』 제11권 1호, 서강대학교 법학연구소, 2022.

김진우, “유럽연합 인공지능법안에 따른 고위험 인공지능 시스템 공급자 등의 의무”, 『과학기술과 법』 제12권 제2호, 충북대학교 법학연구소, 2021.

_____, “유럽연합의 인공지능 백서에 관한 고찰”, 『외법논집』 제44권 제4호, 한국외국어대학교 법학연구소, 2020.

남중권, “간접차별과 머신러닝에 의한 특성추론”, 『법철학연구』 제22권 제2호, 한국법철학회, 2019.

명재진 · 이한태, “평등권 이론에 관한 현대적 전개로서 적극적 평등실현조치”, 『법학연구』 제18권 제2호, 충남대학교 법학연구소, 2007.

박노형 외 1인, “자동화된 결정에 관한 개인정보보호법 정부 개정안 신설 규정의 문제점– EU GDPR과의 비교 분석”, 『사법』 제62호 제1권, 사법발전재단, 2022.

박도현, “인간 편향성과 인공지능의 교차”, 『서울대학교 法學』 제63권 제1호, 서울대학교 법학연구소, 2022.

박혜성 · 김법연 · 권헌영, “인공지능 통제를 위한 규제의 동향과 시사점”, 『정보법학』 제25권 제2호, 한국정보법학회, 2021.

선지원, “인공지능 생애주기의 관점에서 본 규제와 거버넌스”, 『연세법학』 제39호, 연세법학회, 2022.

양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안”, 『법조』 제66권 제3호, 법조협회, 2017.

이승은, “중국 사회신용시스템의 현황 및 전망: ‘빅브라더’와 빅데이터”, 대외경제정책연구원,

2017.
이재훈, “EU의 입법절차”, 『법제연구』 제61호, 한국법제연구원, 2018.
지능정보사회진흥원, “디지털 기술 주도권 확보 전략 및 대응 방안-미국 · 중국 · EU 정책 분석을 중심으로-”, 2019.
_____, “미국, ‘알고리즘 책임법안(Algorithmic Accountability Act)’발의”, 『디지털 법제 Brief』 제22-01호, 2022, 1면.
카나리나 츠바이크(유영미 번역), 『무자비한 알고리즘』, 니케북스, 2021.
한애라, “인공지능과 젠더차별”, 『이화젠더법학』 제11권 제3호, 이화여자대학교 젠더법학연구소, 2019.
홍석한, “유럽연합 ‘인공지능법안’의 주요 내용과 시사점”, 『유럽헌법연구』 제38호, 유럽헌법학회, 2022.
_____, “미국 “2022 알고리즘 책임법안”에 대한 고찰”, 『미국헌법연구』 제34권 제1호, 미국헌법학회, 2023.

2. 외국문헌

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Rieke, A.(2019), “Discrimination through optimization: How Facebook’s ad delivery can lead to skewed outcomes”
Cathy O’nell, 『WEAPONS OF MATH DESTRUCTION』, Broadwaybooks, 2016.
European Commission, White Paper on Artificial Intelligence — A European approach to excellence and trust, COM(2020) 65 final, 2020.
_____, High-level Expert Group on Artificial Intelligence, “ETHICS GUIDELINES FOR TRUSTWORTHY AI”, 2019.
HLEG, Ethics Guidelines for Trustworthy AI, 2019.
_____, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2020.
New York State Department of Financal Services(2021), “Report on Apple Card Investigation”, 2021.
Sandra Wachter 외 2인,“WHY FAIRNESS CANNOT BE AUTOMATED: BRIDGING THE GAP BETWEEN EU NON-DISCRIMINATION LAW AND AI”, 『Computer Law & Security Review』Volume 41, Elsevier, 2021.
Sandra Wachter 외 4인, “Artificial Intelligence and the ‘Good Society’: the US, EU, and UK approach”, 『Science and Engineering Ethics』 volume 24, Springer Science+ Business Media, 2017.

Abstract

A Study on the Discrimination by High-Risk Artificial Intelligence Systems

Kim, Il Woo*

In the contemporary intelligent information society, high-risk artificial intelligence(AI) systems are widely utilized across both private and public domains. Specifically, AI technology is actively used in private sectors such as recruitment, business evaluation, credit evaluation. In the public sector, automated administrative measures utilizing AI are being introduced, and discussions about the deployment of AI systems to support judicial work are in progress.

Despite the societal and economic advancements (e.g., the speed and efficiency of business processing) brought about by these high-risk AI systems, there are significant concerns as they cause various consitutional issues such as discrimination and bias.

Accordingly, there is a need for a law that can effectively regulate high-risk artificial intelligence systems that can cause significant infringement of basic rights, including unreasonable discrimination. Such AI corresponds to AI systems that are widely used to the extent that they are utilized daily in everyday life. Furthermore, this study investigates the infringement of basic rights by high-risk AI systems by examining specific cases regarding potential threats or infringements on basic rights caused by the widespread use of high-risk AI systems in various fields.

Above all, strict regulation of AI is needed to prevent the infringement of basic rights by high-risk AI systems. When examining comparative legal trends regarding high-risk AI, UNESCO and the OECD have established ethical principles to implement the principles of reliability and fairness of AI. The EU AI Act and the U.S. Algorithm Accountability Act define high-risk AI systems specifically and regulate duties of AI businesses and users.

Finally, safety and transparency of AI must be implemented to prevent the infringement of basic rights by high-risk AI systems. It is necessary to set the duty of supervision by businesses or users to ensure that data and algorithms do not embody unjust discrimination and bias. There is a need to guarantee access to information for high-risk AI system users

* Senior Researcher, The Institute for Legal studies of Sogang University, Ph.D. in Law

and the right to reject AI usage.

• **Key Words** High-risk artificial intelligence systems, bias and discrimination, AI impact assessment, intelligent information society, informational basic rights