이슈보고서

인공지능에 대한 인권기반접근

영향 받는 사람들의 인공지능

2025년 11월

사단법인 정보인권연구소

오병일, 장여경, 희우

정보인권연구소 Institute for Digital Rights

HEINRICH BÖLL STIFTUNG 서울 동아시아 | 국제 대화

정보인권연구소 이슈보고서
인공지능에 대한 인권기반접근
: 영향 받는 사람들의 인공지능

발행일 2025년 11월 30일

발행처 사단법인 정보인권연구소, idrsec@proton.me

정보인권연구소 Institute for Digital Rights

저자 오병일 (정보인권연구소 연구위원)
장여경 (정보인권연구소 상임이사)
희우 (디지털정의네트워크 활동가)

후원 하인리히뵐 재단

HEINRICH BÖLL STIFTUNG 서울 동아시아 | 국제 대화

사단법인 정보인권연구소는 시민사회 관점에서 정보인권을 지지하는 대안적 정책을 연구하고 생산하기 위해 활동하고 있습니다.
http://idr.jinbo.net

정보인권연구소 이슈보고서

인공지능에 대한 인권기반접근

: 영향 받는 사람들의 인공지능

목 차

들어가며

인공지능(AI)은 글, 이미지, 오디오, 영상을 생성하는 도구에서부터 자율주행차, 산업용 로봇, 그리고 공공 행정 시스템에 이르기까지, 우리 사회의 모든 영역에 빠르게 확산되고 있습니다. 인공지능은 기존의 자동화 시스템보다 뛰어난 자율성과 적응성으로 인간의 지능이 필요했던 예측, 추천, 결정 등의 결과물을 생성할 수 있습니다. 이러한 기술 혁신이 생산성을 향상시키고 효율성을 증대시킬 것이라는 기대감이 높아지면서, 한국 정부 또한 'AI 강국' 실현을 목표로 강력한 산업 지원 정책을 추진하고 있습니다.

하지만, 인공지능이 활용하는 방대한 데이터의 원천은 결국 '사람'이며, 그 예측과 결정의 대상이 되는 것 또한 '사람'입니다. 따라서 인공지능은 평범한 시민들의 인권에 중대한 영향을 미칠 수밖에 없습니다. 특히, 사법, 고용, 복지 등 사람의 삶과 노동에 필수적이고 중요한 영향을 미치는 영역에 인공지능이 사용될수록, 그 영향은 개인의 인권을 심각하게 침해하거나 사회적 차별을 악화시키는 결과를 낳을 수도 있습니다.

이 보고서는 인공지능과 그 데이터 및 알고리즘이 인권과 사회에 미치는 영향을 살펴보고자 합니다. 더불어 인공지능으로부터 인간의 존엄성과 인권을 보호하기 위해서 국제 인권 규범이 지지하는 인권 기반 접근 Human Rights-Based Approach 을 소개합니다.

특히 공공기관과 민간기업이 인공지능을 개발하고 활용할 때 인권 의무를 준수해야 하고 인공지능으로부터 영향 받는 사람을 구제해야 할 책무가 있음을 강조하였습니다. 마지막으로 인권적 관점에서 보았을 때 한국에서 제정한 인공지능기본법이 어떤 내용과 한계를 가지고 있는지를 검토하였습니다.

2025년 11월

사단법인 정보인권연구소

* 이 자료는 독일 하인리히 뵐 재단의 후원으로 발간되었습니다.

인공지능의 인권 영향

1. 인공지능과 인권의 관계

인공지능의 개념

우리의 삶과 노동의 여러 영역에 인공지능(AI) 기술이 빠른 속도로 확산되고 있습니다. 글, 이미지, 오디오, 영상을 생성해주는 다양한 생성형 AI가 등장하여 널리 쓰이고 있으며, 배달 앱, 가전 제품, 자율주행차처럼 우리 주변의 제품이나 서비스에 인공지능 알고리즘이 빠르게 적용되고 있습니다. 최근에는 직장, 학교, 사회복지 등 우리 삶에 중요하고 필수적인 영역에도 인공지능 알고리즘이 도입되고 있습니다.

인공지능은 인간의 인지 기능을 기계적으로 모방하려는 컴퓨터 과학 분야를 설명하는 데 사용되는 포괄적 용어입니다. 인공지능 기법으로 개발된 컴퓨터 시스템은 특히 학습, 추론, 인식, 문제 해결 등 일반적으로 인간 지능과 관련된 인지 문제를 해결하는 데 탁월하며, 기존의 컴퓨터에 비하여 뛰어난 자율성과 적응성으로 예측, 추천, 결정 등의 결과물을 산출합니다.

「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 '인공지능기본법')」은 인공지능 시스템을 다음과 같이 정의합니다. OECD나 유럽연합 등 세계 인공지능 관련 규범들도 유사하게 인공지능 시스템을 정의하고 있습니다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법

제2조(정의)
“인공지능 시스템”이란 다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 인공지능 기반 시스템을 말한다.

최근의 놀라운 인공지능 혁신은 데이터, 하드웨어, 알고리즘 측면의 발전을 배경으로 합니다. 우선 데이터 처리 기술의 발전으로 과거에는 다루기 어려웠던 이미지, 영상, 음성 등 비정형 데이터를 활용할 수 있게 되었습니다. 대규모 빅데이터를 실시간으로 처리하는 일도 과거보다 확연하게 저렴하고 수월해졌습니다. 더불어 클라우드 컴퓨팅 기술이 발전하고 병렬연산방식의 GPU가 상용화하면서 대규모 데이터를 저장하고 처리하는 효율이 크게 향상되었습니다.

무엇보다 머신러닝 알고리즘 기법의 발전을 빼놓을 수 없습니다. 2012년 이미지 인식 분야에서 딥러닝(심층신경망)이 비약적인 성능 향상을 보였고, 2017년에는 이후 GPT와 같은 대규모언어모델의 기반이 된 트랜스포머 아키텍처가 공개되었습니다. 2022년 챗GPT가 출시된 이후로는 일반 시민들도 텍스트·이미지·음성 등 다양한 형태의 콘텐츠를 직접 생성할 수 있는 생성형 인공지능(Generative AI)을 널리 활용하게 되었습니다.

이러한 머신러닝 알고리즘들은 기본적으로 대규모 데이터에서 자동으로 ‘패턴’이나 ‘상관관계’를 찾아냅니다. 이 과정이 “데이터를

‘학습’한다”고 표현됩니다. 기계의 학습, 즉 ‘머신러닝’은 학습을 통해 무언가를 분류하거나 예측하는 결과물을 산출하는 수학적 모델을 만드는데, 이러한 절차를 ‘알고리즘’이라고 합니다. 기존의 컴퓨터 프로그램은 사람이 정한 규칙에 따라 작동하였지만, 머신러닝은 사람이 일일이 프로그래밍하지 않아도 스스로 규칙을 발견하고 적용하는 자율성을 지닙니다. 새로운 데이터가 주어지면 학습 모델을 갱신하며 환경 변화에 적응할 수도 있습니다.

데이터를 분석하여 추론하는 결과물을 산출하는 인공지능 알고리즘은, 쇼핑 기록이나 시청 기록에 대한 분석을 통해 개인의 취향을 추론하거나, 얼굴 인식을 통해 그 사람이 누구인지를 추론할 수 있습니다. 나아가 사람의 얼굴 표정이나 음성 분석을 통해 그 사람이 화가 났거나 긴장했는지 그 감정 상태를 추론하거나, 특정인에 대한 다양한 개인정보 분석을 통해 질병, 부정행위, 재범 위험을 추론하는 것도 가능합니다.

이처럼 데이터로부터 학습하고 그 결과를 바탕으로 높은 자율성과 적응성을 가지고 작동하는 머신러닝 알고리즘은 기존의 컴퓨터 프로그램보다 더 탁월한 성능을 발휘할 수 있습니다. 복잡한 분석이나 예측을 수행하고 자동화된 결정도 내릴 수 있으며, 글, 이미지, 영상 등 새로운 창작물도 생성할 수 있습니다.

인공지능과 인간 존엄성

디지털기술은 장애인이나 농촌지역주민 등 취약한 사람들의 인권을 증진하고 권리를 행사할 수 있는 새로운 수단이 될 수 있습니다. 다만 이러한 기술이 인권에 미치는 영향을 충분히 고려하지 않고 배

치될 경우 부정적인, 심지어 치명적인 영향을 미칠 수도 있습니다.

특히 인공지능을 개발하고 활용하면서 인간을 존엄한 존재로 대우하지 않을 경우 인권침해로 이어질 수 있습니다. 인공지능이 사람을 데이터·점수·객체로만 취급하는 경우, 자동화된 결정에 대하여 사람이 거부할 수 있는 기회를 주지 않거나 설명하지 않고 그 대상으로만 취급하는 경우, 감시와 통제로 인간의 자율성과 주체성을 무시하는 경우에 인간의 존엄성에 대한 문제가 발생합니다.

2017년 페이스북 프로필에 "좋은 아침"이라는 글을 올린 한 팔레스타인 남성이 이스라엘 경찰에 체포되는 사건이 일어났습니다. 서안 지구의 건설 노동자인 그 남성은 불도저에 기대어 있는 자신의 사진을 "يصبحهم"이라는 문구와 함께 게시했습니다. 이는 아랍어에서 "좋은 아침"을 의미합니다. 하지만 페이스북 자동 번역 인공지능은 이 문구를 "그들을 공격하라"로 번역하였고 누군가 그를 경찰에 신고하였던 것입니다. 당시 체포가 진행되기 전 아랍어를 구사하는 이스라엘 경찰관 중 누구도 게시물을 직접 읽지 않았던 것으로 드러났습니다. 이 남성은 부당하게 체포되어 경찰의 심문을 받아야만 했습니다.[1)]

미국에서도 유사한 사건이 일어났습니다. 얼굴인식 AI의 오류로 무고한 흑인이 범인으로 체포되는 일이 잇따른 것입니다. 뉴저지주에 살고 있는 니지어 파크스씨는 경찰 얼굴인식 AI가 절도 사건 범인으로 지목하는 바람에 10일간 무고하게 구금되었습니다. 경찰 얼굴인식 AI가 엉뚱한 사람을 체포한 사례가 2020년 최소 3건이었으며 피해자는 모두 흑인이었습니다. 파크스 씨는 "구치소에 있는 동안 경찰은 내 지문과 DNA를 확인하는 등 추가 증거를 확보하지 않았

다"고 토로했으며 변호사도 "얼굴인식 AI를 제외한 다른 모든 증거는 파크스 씨가 범인이 아님을 가리켰다"고 비판했습니다.[2)]

이 사건들은 국가 공권력이 사람을 부당하게 체포하고 구금함으로써 인간의 존엄을 훼손하고 인권을 침해한 사건입니다. 여기서 국가는 인간의 판단보다 편향적이고 오류가 있는 인공지능에 일방적으로 의존하였습니다.

유럽국가인권기구네트워크는 특히 인공지능 감시 기술이 인간의 자율성, 인간의 주체성, 자기통치권, 자기결정권을 침식한다고 지적하였습니다. 감정 인식 기술은 개인을 고유한 가치와 존엄성에서 분리된 데이터 포인트로 축소하여 비인간화할 위험이 있습니다.[3)]

유럽연합 기본권청(FRA)은 국가기관이 공개된 장소에서 실시간으로 얼굴이나 동작 등 생체 인식을 통해 사람들을 감시하는 것은 일반적인 인간존엄성의 권리에 관한 문제라고 지적하였습니다. 얼굴 이미지 처리는 다양한 방식으로 인간 존엄성에 영향을 미칠 수 있습니다. 사람들은 얼굴인식 감시 하에 있는 공공장소에 출입하는 것에 불편함을 느낄 수 있습니다. 사회생활을 취소하거나 감시중인 주요장소를 방문하지 않게 되고 기차역을 피하거나 문화사회스포츠행사 참석을 줄이는 등 감시로 인해 행동을 바꿀 수 있습니다. 얼굴인식 기술이 적용되는 정도에 따라 사람들은 삶 속에서 감시 기술을 의식하게 되는데 이 의식은 개인이 존엄한 삶을 영위할 역량에 영향을 미칠 정도로 매우 중대할 수 있습니다.[4)]

인공지능이 인간의 존엄성에 미치는 위험과 관련하여 가장 논쟁적인 사례는 중국의 '사회신용점수' 시스템입니다. 중국의 사회신용점수는 개인의 금융 정보, 범죄 이력 등 국가가 수집한 모든 개인정보와 인공지능, 얼굴인식 등 첨단기술을 활용하여 모든 시민에게 점수를 부여합니다. 사회신용점수는 대출, 교육, 의료, 취업 등 개인의 모든 사회생활에 범용으로 활용되며, 사회신용점수가 낮은 사람은 심지어 고속철도, 항공기와 같은 교통편이나 고급 호텔을 이용할 수 없습니다.[5)]

인공지능의 예측 기능이 인간에 대해 어디까지 예측해도 되는지도 문제가 됩니다. 2012년 미국 미니애폴리스의 '타겟' 매장에서는 십대 여성의 임신 사실을 가족보다 먼저 예측한 알고리즘이 임신 축하 할인 쿠폰을 발송한 사건이 발생했습니다. 당시 대형마트 타겟은 당시 무향 로션, 비타민, 화장솜, 손소독제, 면봉 등 약 25가지 특정 품목을 구매한 고객을 기준으로 임신가능성을 예측하는 알고리즘을 사용하고 있었습니다. 이 알고리즘은 그 십대 여성 고객의 임신 가능성이 높다고 예측하였고 아기 옷 등에 대한 맞춤형 할인 쿠폰이 발송되었습니다. 임신 축하 쿠폰을 발견한 고객의 아버지는 마트 측에 잘못 발송하였다며 항의하였지만, 나중에 해당 여성이 실제로 임신하였다는 사실이 드러났습니다. 이 사건은 알고리즘이 사람에 대해 일방적으로 예측해도 되는지, 알고리즘의 예측이 인간의 존엄성을 어떻게 존중해야 할지에 대한 논란을 불러왔습니다.[6)]

인공지능에 대한 인권 중심 접근

인공지능과 관련한 규범은 인공지능 기술에 대한 전문가들이 실행자 측면에서 '인공지능 윤리'에 대한 논의로 먼저 시작하였습니다. 다만 국제 인권 규범은 윤리적 접근 방식으로 인권 위험을 해결할 수 없다고 지적합니다. 유엔 표현의 자유 특별보고관은 "윤리는 인공지능 분야에서 발생하는 특정한 문제들을 해결하는 데 중요한 개념틀이지만, 모든 국가에서 법률로 구속되는 인권을 대체할 수 없다."고 지적하였습니다.[7] 윤리 기준은 인공지능을 개발하고 활용하는 기업과 같은 실행자 측면에서 '자율적' 준수를 강조하기 때문입니다.

이와는 달리 인권에 기반한 접근 방식의 경우, 헌법과 국제 인권법 등 구속력 있는 규범에 근거를 두고 있습니다. 국가인권위원회는 인공지능의 개발과 활용에 있어서 다음과 같이 인간의 존엄과 가치를 비롯한 인권을 존중해야 한다고 밝혔습니다.

> 국가인권위원회
> 〈인공지능 개발과 활용에 관한 인권 가이드라인〉
>
> 15. 어떠한 활동도 인간의 존엄에서 유래하는 다양한 권리들의 희생을 강요해서는 안 되며, 궁극적으로 모든 활동은 인간의 존엄과 가치를 향상시키는 방향으로 수행되어야 합니다.
>
> 16. 인공지능은 인간으로서의 존엄과 가치 및 행복을 추구할 권리에 부합하는 방향으로 개발 및 활용되어야 하며, 개인의 선택과 판단 및 결정을 강요하거나 자율성을 침해해서는 안 됩니다.

17. 인공지능의 개발과 활용은 개인의 행복과 사회적 공공성의 증진에 위배되어서는 안 되고, 표현의 자유, 집회 및 결사의 자유, 노동권 등 인권을 인공지능의 부정적 영향으로부터 보호해야 합니다.

26.「대한민국헌법」 제10조에서 보장하고 있는 인간의 존엄과 가치는 누구나 누려야 할 불가침의 기본적 인권으로, 모든 권리의 출발점인 동시에 종국적으로 보장되어야 할 인권적 가치입니다.

국가인권위원회가 인공지능 인권 기준으로 제시한 인간의 존엄과 가치, 행복추구권, 인격의 자율성은 우리 헌법에서 기본권으로 보호하는 영역입니다. 또한 인공지능 환경에서 보호하고자 하는 대상은 인간의 존엄에서 유래하는 다양한 권리들, 즉 모든 '인권'입니다. 무엇보다 국가와 기업은 이 인권을 보호하고 존중해야 하는 의무를 준수해야 합니다.

즉, 인공지능에 대한 인권 기반 접근이란, 인공지능 기술을 개발하고 활용하는 국가와 기업 등 의무주체들이, 인공지능 기술의 대상이 되는 권리주체가 국제 인권법 등에서 확인한 인권을 보호하고 존중하며, 인권침해가 발생한 경우 필요한 구제 절차를 제공하도록 법과 제도로서 감독하는 것입니다.

유엔 사무총장은 인공지능과 같은 신기술의 개발과 배치가 견고한 인권 기반에 뿌리를 두는 것이 기술 발전의 혜택을 온전히 누리고 피해 가능성을 최소화하는 길이라고 설명합니다.[8] 국제 인권법은 국가 간에 합의된 내용이고 국가적, 지역적, 국제적 메커니즘이 모니터링

하고 있습니다. 따라서 끊임없이 변화하는 기술 환경 문제에 대응할 때 보편적인 국제 인권법이 핵심적인 지침이 될 수 있습니다.

특히 유엔 사무총장은 “사람을 권리주체 개인으로 대우하여야 한다.”고 강조하면서, 인공지능의 잠재적 위험을 해결하고 그 혜택을 살리기 위해서는 인권 기반 접근을 취해야 한다고 요구하였습니다.

> 유엔 사무총장(2020)
>
> 인공지능에 대한 “인권 기반 접근”이란 사람을 권리주체 개인으로 대우하고, 권리주체가 자신의 권리를 행사하고 인권침해와 유린에 대하여 구제수단을 찾을 수 있도록 개인의 역량을 강화하고 법·제도적 환경을 조성하는 것입니다.

2025년 유엔 인권최고대표는 인공지능에 대하여 우리가 인권 기반 접근을 해야 할 이유가 인간에게 유익하고 책임 있는 인공지능 혁신 생태계를 구축하는 데 필수적이기 때문이라고 말했습니다.[9)]

> 유엔 인권최고대표 (2025)
>
> 인권 보장이 내포되지 않은 인공지능은 우리가 추구하는 결과를 도출하지 못할 뿐 아니라, 오히려 개발을 저해하고 평화와 안보를 훼손할 수 있습니다. 인공지능을 개발하고 배치할 때 국가의 역할과 의무, 그리고 기업의 책임을 명확히 하는 것이야말로 인류에게 유익하고 책임 있는 인공지능의 혁신 생태계를 구축하는 데 필수적입니다.

그래서 인공지능에 대한 규제는 보안이나 안전에 미치는 위험에 대한 중시를 넘어 “사람에게 미치는 영향에 초점을 맞춰야 한다.”는 것입니다. 아마도 이것이 인공지능에 대한 인권 기반 접근의 핵심적인 목표일 것입니다.

2. 데이터의 인권 영향

데이터 기반 기술과 개인정보

오늘날 인공지능 개발의 핵심 방법인 ‘머신러닝’은 대규모 데이터를 기반으로 모델이 학습하고 구축되며, 이렇게 구축된 모델은 활용 과정에서도 지속적으로 데이터를 수집하거나 처리합니다. 이 과정에서 인공지능이 사용하는 데이터에 개인정보가 포함될 수 있습니다.[10] 공개된 장소나 인터넷 등에서 정보주체가 모르는 새 무작위로 수집되었을 수 있고, 거래되었을 수도 있습니다.

인공지능의 개인정보에 대한 수집·분석·추론 기능은 정보주체가 자신의 개인정보 처리에 대해 통제할 수 있는 권리를 약화시킬 수 있으며, 정보주체가 원치 않는 개인정보 학습이 이루어질 수 있습니다. 학습 결과물에서 개인정보가 노출되거나, 편향적인 데이터 품질이 편향적인 결과물을 생성할 위험도 존재합니다. 이처럼 인공지능 기술은 기존의 개인정보 보호 원칙의 이행과 정보주체의 개인정보자기결정권의 행사에 여러 영향을 미치고 있습니다.

인공지능이 자동화된 방식으로 개인에 대하여 추론하고 결정을 내리는 기능과 관련된 개념이 '프로파일링'입니다. 프로파일링이란 개인에 관련한 특정 측면을 평가하기 위하여, 특히 개인의 업무 성과, 경제적 상황, 건강, 개인적 선호, 관심사, 신뢰도, 행태, 위치 또는 이동에 관한 측면을 분석하거나 예측하기 위하여, 개인정보를 사용하여 이루어지는 자동화된 개인정보의 처리를 말합니다.[11]

예를 들어 기업이 특정 소비자에 대하여 프로필을 생성할 때 기존의 개인정보에만 기반하는 것이 아니라 기업 자체적으로 이 소비자의 재정적 취약성에 초점을 맞춰 범주화("외곽거주자로 간신히 끼니를 때움", "도시에 거주하는 재정적으로 힘든 이주민", "젊은 미혼부모")하거나 이들에 '등급'을 매기는 경우가 프로파일링에 해당합니다.[12] 프로파일링은 자동화된 개인정보 처리의 한 방식으로, 개인에 대한 평가를 토대로 정보주체의 권리에 미치는 위험성을 증가시킬 우려가 있습니다. 이러한 이유로 「유럽연합 일반개인정보보호규정(GDPR)」은 프로파일링을 포함한 자동화된 결정에 대하여 정보주체가 행사할 수 있는 권리를 법적으로 규정하고 보호하고 있습니다.

국가인권위원회는 인공지능 개발 및 활용 과정에서도 정보주체의 권리가 보장되어야 한다고 말합니다. 더불어 인공지능을 개발하거나 활용하는 개인정보처리자들이 목적 제한 원칙, 최소화 원칙, 정확성 원칙, 투명성 원칙, 정보주체 참여 원칙 등 개인정보의 처리의 원칙을 준수할 것을 요구하였습니다.

국가인권위원회
〈인공지능 개발과 활용에 관한 인권 가이드라인〉

27. 인공지능과 관련하여 정보주체의 권리는 처리된 개인정보에 관하여 고지를 받을 권리, 개인정보 접근 및 열람권, 개인정보 처리 동의권 및 정정·삭제권, 처리정지요구권 등을 포함하며,

정보주체는 자신의 개인정보가 사용되는 방법을 이해하고 그에 대한 통제권을 가지는 것이 중요합니다. 정보주체는 인공지능 서비스가 언제, 어디서 자신의 개인정보를 수집하고, 어떻게 개인정보를 처리하여 사용, 보관, 삭제하는지에 대해 알고 참여할 권리가 있습니다.

28. 인공지능의 개발과 활용에서 개인정보는 목적에 필요한 최소한의 범위 내에서 처리되어야 하며, 처리목적 달성에 필요한 기간 동안만 보관되어야 합니다. 또한 이러한 개인정보 처리 방침은 정보주체가 확인할 수 있도록 공개되어야 합니다.

29. 인공지능의 개발과 활용에서 민감정보를 처리할 때에는 특별한 주의를 기울여 보호하여야 합니다. 더불어 의사결정의 내용과 관련성이 없거나, 부정확한 데이터에 기반한 의사결정이 이루어지지 않도록 데이터의 정확성, 완전성, 최신성을 보장해야 합니다.

인공지능 챗봇 '이루다' 사례

2021년 4월 28일 개인정보 보호위원회는 인공지능 챗봇 '이루다' 개발사 (주)스캐터랩에 대하여 개인정보보호법 위반 사실을 확인하고 총 1억 330만원의 과징금과 과태료를 부과했습니다.[13)]

이 사건은 우리나라에서 국가 개인정보 보호 감독기구가 인공지능 학습 및 서비스 과정에서 이루어진 개인정보 처리에 대하여 직권으로 조사하고 제재한 첫 사례입니다.

회사는 2013년 카카오톡 대화 감정분석서비스인 '텍스트앳'을 출시하고 2016년에는 카카오톡 대화 연애 심리 검사를 내세운 '연애의 과학'을 출시한 바 있었습니다. 인공지능 챗봇 이루다는 이들 서비스에서 수집한 회원들의 개인정보와 비공개 대화문장을 이용하여 학습이 이루어졌으며 2020년 12월 출시되었습니다. 개인정보 보호위원회 조사 결과 전체 회원은 60만 명이며 그중 약 20만 명이 만 14세 미만 아동이었습니다. 이처럼 회사가 7년 전과 4년 전에 출시한 서비스에서 수집한 회원정보는 회원을 탈퇴하거나 1년 이상 서비스를 이용하지 않았어도 파기되지 않고 이루다 서비스에 이용되었습니다.

회사는 이루다 학습을 위해 '학습 DB'를 구축하였는데 이 데이터세트는 회원정보와 카카오톡 대화문장 94억 건으로 구성되었으며, 특히 대화문장은 아무런 보호조치 없이 원문 그대로 학습에 이용되었습니다. 한편 회사는 학습 DB에 저장된 카카오톡 대화문장 중 20대 여성의 대화문장을 약 1억 건을 추출한 후 '응답 DB'를 구축하여 이루다의 출력에 이용하였습니다. 회사는 이 대화문장에서 실명, 장소명, 숫자/영문, 선정적 표현을 제거하였다고 밝혔으나, 개인정보보호위원회 조사 결과 일부 주소, 휴대전화번호 등 개인정보가 노출된 상태였던 것으로 나타났습니다.[14)]

인공지능 학습이나 서비스에 사용되는 데이터세트가 개인정보를 이용한 경우 개인정보보호법을 준수하여야 합니다. 그러나 회사는 개인정보 동의를 받을 때 인공지능 학습과 서비스에 이용된다는 사

실을 정보주체가 명확하게 인지할 수 있도록 알리고 동의를 받지 않았습니다. 회사가 이루다 이전에 출시했던 서비스는 “로그인함으로써 이용약관 및 개인정보처리방침에 동의합니다”라고만 안내하였으며, 개인정보 처리방침은 수집된 개인정보를 “신규 서비스 개발 및 마케팅·광고에의 활용”하겠다며 추상적으로만 안내하였습니다.

개인정보 보호위원회는 “개인정보처리방침에 ‘신규 서비스 개발’이 명시되어 있다는 이유만으로, 이용자가 ‘이루다’와 같은 기존 서비스와 전혀 다른 신규 서비스의 개발과 서비스 운영에 자신의 개인정보가 이용될 것을 예상하고 이에 동의하였다고 보기 어렵다.”고 지적하였습니다. 특히 만 14세 미만 아동의 개인정보를 수집할 때에는 법정대리인의 동의를 받아야 하지만 회사는 이를 준수하지 않았습니다.

사생활 침해

범용 AI가 개인정보 보호에 초래하는 위험에 대하여 종합적으로 검토한 〈과학보고서〉[15]는 이를 학습 위험, 사용 위험, 악의적 피해 위험으로 구분하였습니다. 학습 위험의 경우 범용 AI가 학습 데이터 중 일부를 암기하여 개인정보를 노출시키거나 민감한 개인정보를 추론하는 위험을 말합니다. 정보주체가 알지 못하거나 동의하지 않은 상태에서 수집된 데이터세트로 학습이 이루어지는 문제도 있습니다. 사용 위험의 경우 인공지능 시스템을 사용하거나 배치하는 과정에서 민감한 개인정보가 유출되거나 의도하지 않은 방식으로 이용되는 문제입니다. 악의적 피해 위험은 악의적인 이용자가 인공지능 시스템을 사이버 공격에 사용하거나, 공개하지 않은 개인의 민감한 속성을

추론하거나, 스토킹 행위를 심화하거나, 딥페이크 허위조작정보를 생성하는 상황 등이 있습니다.

특히 생성형 인공지능이 확산되기 시작하면서 딥페이크 기술로 생성된 허위조작정보가 큰 우려를 사고 있습니다. 인공지능을 이용하여 타인의 얼굴이나 목소리 등 개인정보를 악용하고 사생활을 침해하는 사건들이 급증한 것입니다. 가족이나 지인의 목소리를 흉내 내고 개인의 취약점을 악용하는 보이스피싱이 심각한 피해를 낳기도 하였습니다.[16] 2024년 8월에는 전국의 학교와 대학교 수백 곳에서 딥페이크 기술로 여학생들 사진을 허위로 음란물에 합성하고 유포하는 사건이 발생하여 세계적 충격을 주었습니다.[17]

인공지능을 감시 목적으로 사용할 경우 과거보다 양적으로나 질적으로 더욱 심각하게 사생활의 권리를 침해할 수 있습니다. OECD는 인공지능이 성적 지향, 정치적 선호, 소득, 미래의 범죄가능성 등 개인에게 매우 민감한 추론을 내리는 데 사용될 수 있다고 경고하였습니다. 이는 사생활의 권리 뿐 아니라 자동화된 차별을 양산하며 정치적 반대자를 억압하는 데 남용될 수도 있습니다. 생체 인식 기술의 사용은 이러한 위험을 증폭시킵니다. 얼굴 인식 문제가 가장 널리 알려져 있지만 동작, 보행, 심장 박동과 같은 생체 특성을 통해 식별하는 것도 가능합니다. 인공지능 감시를 통해 직장에서 과거보다 심하게 노동조합 활동을 감시하고, 공공장소에서 표현의 자유와 집회시위의 자유를 제한하거나, 개인이나 집단을 표적으로 삼아 차별하는 감시도 이루어질 수 있습니다.[18]

국제노동기구(ILO) 역시 최근 직장에 도입되는 인공지능 모니터링 시스템들이 전례 없는 정교함, 속도, 규모로 노동자의 생각, 감정,

생리적 상태 역시 추적 및 분석할 수 있으며, 노동자의 특정 행동 또한 예측할 수 있다고 우려하였습니다.[19] 노동자별로 평가할 수 있는 개인 프로파일링을 생성하여 과거보다 은밀하게 비교하는 일도 생겨났습니다. 예를 들어, 노동자의 친분 관계에 대한 개인정보에 기반하여 노동조합 결성 가능성을 예측하거나 노동자의 억양이나 주소지에 대한 개인정보를 업무 신뢰성과 연관 짓는 식입니다.

인공지능 모니터링은 과거의 감시보다 더 방대하게 감시하며, 생체 인식 정보나 특정 행동 등 개인의 더 개인적인 특성, 심지어 개인의 감정이나 인간관계 등 매우 사적인 특성에 대한 침습적인 감시가 이루어질 수 있습니다. 특히 지속적이고 실시간으로 은밀하게 이루어지는 감시 시스템은, 간헐적이고 특정한 표적에 대해 이루어지는 감시에 비하여, 개인에게 더 부정적인 결과를 초래할 수 있습니다.

자동화된 추론과 결정

인공지능은 입력된 데이터를 분석하여 예측, 추천, 결정 등 다양한 결과물을 추론합니다. 이때 입력되고 분석되는 데이터에 다양한 개인정보가 포함될 수 있습니다. 쇼핑 기록이나 시청 기록으로 개인의 취향을 추론하거나, 얼굴 인식으로 그 사람이 누구인지 추론할 수 있으며, 사람의 얼굴 표정이나 음성 분석으로 감정 상태를 추론하거나, 개인정보 분석을 통해 개인의 질병, 부정행위, 재범 위험도 추론할 수 있습니다.

유엔 인권최고대표는 인공지능의 추론과 예측 기능이, 사람들의 자율성과 자신의 정체성에 대한 세부사항을 확립할 권리를 포함하여, 프라이버시권의 향유에 깊은 영향을 미친다고 지적하였습니다.

이는 사상과 의견의 자유에 대한 권리, 표현의 자유, 공정한 재판 관련 권리 등 다른 권리에도 여러 문제를 야기할 수 있습니다.[20] 인공지능 시스템이 성적 지향, 정치적 선호, 소득, 미래의 범죄가능성 등 민감한 추론을 내리는 데 사용되면, 개인이 공개하지 않은 개인정보를 추론하거나 잘못된 추론을 낳을 수 있습니다.[21] 경찰, 의료, 교육, 고용 등 중요한 삶의 영역에서 사람에 대한 결정을 자동화하거나 지원하는 인공지능이, 잘못되거나 불투명한 판단을 내릴 수도 있습니다. 미국에서는 경찰이 사용하는 얼굴인식도구가 개인의 얼굴을 잘못 추론하여 무고한 사람이 구금되는 사건이 발생하기도 하였습니다.[22]

특히 인공지능은 개인에 대한 다양한 데이터 학습이나 생체 인식을 통해, 정보주체가 공개하지 않은 개인의 사회적 배경이나 개인적 속성을 자동적으로 추론하여 사생활을 침해할 수 있습니다.[23] 한 연구에 따르면 인터넷 게시판에서 수집한 데이터를 학습한 인공지능은 개인이 밝히지 않은 위치, 소득, 성별 등 다양한 속성을 추론할 수 있었습니다.[24]

인공지능 개인정보 침해에 대응하는 요구

과학보고서는 범용 AI에서 학습 위험과 사용 위험을 감소시키고 개인정보를 보호하기 위한 기술적 조치를 다양하게 제안했으며, 최소화 원칙, 목적 제한 원칙 등 개인정보 보호 원칙이 여전히 중요하다고 강조하였습니다.

특히 정보주체는 인공지능이 처리한 데이터에 자신의 개인정보가 포함되어 있는지 여부나 자동화된 추론이나 결정에 이용된 자신의

개인정보에 대하여 알 권리를 보장받을 수 있어야 합니다. 개인정보보호법은 자신에 대하여 일반적으로 이루어지는 개인정보처리에 대하여 열람할 수 있는 정보주체의 알 권리와, 자동화된 알고리즘이 자신의 개인정보에 기반하여 내리는 결정에 대하여 설명을 요구할 수 있는 권리로 나누어 보호하고 있습니다. 자신의 개인정보가 어떻게 처리되고 있는지 열람한 후에는 이에 대한 정정·삭제, 처리정지를 요구할 수 있습니다. 인공지능에 기반한 완전히 자동화된 결정 과정에서 처리되는 개인정보에 대해서는 그 처리를 거부하거나 설명을 요구할 수 있습니다.

개인정보보호법 제4조(정보주체의 권리)

정보주체는 자신의 개인정보 처리와 관련하여 다음 각 호의 권리를 가진다.
1. 개인정보의 처리에 관한 정보를 제공받을 권리
2. 개인정보의 처리에 관한 동의 여부, 동의 범위 등을 선택하고 결정할 권리
3. 개인정보의 처리 여부를 확인하고 개인정보에 대한 열람(사본의 발급을 포함한다. 이하 같다) 및 전송을 요구할 권리
4. 개인정보의 처리 정지, 정정·삭제 및 파기를 요구할 권리
5. 개인정보의 처리로 인하여 발생한 피해를 신속하고 공정한 절차에 따라 구제받을 권리
6. 완전히 자동화된 개인정보 처리에 따른 결정을 거부하거나 그에 대한 설명 등을 요구할 권리

하지만 OECD는 생성형 인공지능을 비롯한 인공지능의 개인정보 처리가 기존의 개인정보 보호 원칙과 정보주체의 권리 행사를 보장하지 않을 수 있다고 우려하였습니다.[25] 예를 들어 인공지능이 사용하는 대규모 데이터세트에서 개인이 자신의 개인정보에 대하여 접근하거나 정정 또는 삭제를 요구하는 것이 수월하지 않을 수 있습니다.

그럼에도 플랫폼 노동자들은 개인정보보호법을 활용하여 불투명한 인공지능이 자신의 개인정보를 어떻게 처리하였는지 알고자 분투해 왔습니다. 2021년 3월 11일 네덜란드 암스테르담 지방법원은, 차량공유서비스 '우버'의 운전자 노동자의 요청에 따라 회사는 승객이 부과한 각각의 평점을 익명으로 공개해야 한다고 판결하였습니다. 또다른 차량공유서비스 '올라'의 경우, 완전히 자동화된 방식으로 운전자 벌점 및 소득 시스템을 운영하여 법적 효력과 유사하게 본인에게 중대한 영향을 미치는 결정을 내렸습니다. 이에 법원은 운전자 노동자가 요청한 대로 '부정행위 위험 점수'를 생성하는 데 사용된 개인정보 및 프로파일링, 업무 할당에 영향을 미치는 소득 프로필을 생성하는 데 사용된 개인정보 및 프로파일링, 금전적인 불이익 결정을 내리는데 사용되는 부정행위 경고 시스템과 개인정보 등을 공개하라고 판결하였습니다.

3. 알고리즘의 인권 영향

기계 편향

인간의 판단과 의사결정은 기존의 개인적 친분관계, 편견, 고정관념 등 다양한 요인에 의해 객관적이거나 공평하지 않은 결론을 내릴 수 있고, 자기 스스로도 인지하지 못하는 다양한 주관적, 내재적, 무의식적 원인에 의해 판단과 의사결정이 좌우되는 문제, 즉 인지편향(cognitive bias)이 나타날 수 있습니다. 인지편향에 영향을 받은 인간의 판단과 의사결정이 실제로 차별적 행동으로 이어지기도 합니다. 이러한 이유에서 데이터와 알고리즘에 기반하여 내리는 판단과 결정이 인간이 가지고 있는 편향과 오류보다 객관적이고 정확할 것이라고 기대하는 목소리가 있어 왔습니다.

그런데 최근에는 인공지능에 의한 판단과 결정에서도 인간과 유사한 편향이나 차별이 나타날 수 있다는 사실이 확인되고 있으며, 이러한 경향을 기계 편향(Machine Bias)이라고 부르기도 합니다.[26] 국제인권 규범과 법률이 불합리한 차별을 금지하고 있음에도, 공공과 민간의 주요 영역에서 개발되고 활용되는 인공지능이 인종, 성별, 문화, 연령, 장애, 정치적 견해 등 다양한 인간 정체성과 관련해 편향된 결과를 산출하고 차별적 대우를 낳는 문제가 발생하고 있는 것입니다.

2018년 아마존이 개발 중인 AI채용 시스템이 여성 지원자를 차별하는 것으로 드러나 도입이 취소됐습니다.[27] 아마존의 AI채용 시스템 개발팀은 2014년부터 지원자의 이력서를 검토해 인재를 가려낼

수 있는 기술을 개발해왔습니다. 연구진은 이 시스템을 활용해 채용 완료된 직원의 이력서에 적용해 실제 결과와 일치하는지 확인하는 실험을 했습니다. 그러나 연구진이 2015년까지의 자료를 바탕으로 실험한 결과 이 시스템이 여성을 차별하는 문제점을 안고 있는 것으로 나타났습니다. 특히 소프트웨어 개발자 및 기타 기술 직무 지원자가 성중립적으로 평가되지 않았다는 점이 드러났습니다. 아마존의 기존 재직자 데이터는 남성의 비율이 매우 높았고, 이 학습데이터로 학습한 알고리즘 또한 지원자의 이력서에서 '여성'이라는 단어를 선호하지 않게 된 것입니다.

인공지능의 편향이 나타나는 원인에 대해서는 여러 분석이 이루어지고 있습니다. 과학보고서[28]는 인공지능의 편향이 발생하는 데에는 학습데이터나 알고리즘 설계와 관련된 다양한 원인이 있다고 설명합니다.

가장 많은 원인은 아마존 채용 AI의 사례처럼 학습데이터에 특정 집단이 과소대표되어 있거나 사회적 편견이 반영되어 있을 때입니다. 특히 생성형 AI 모델은 인터넷에서 수집한 대규모 데이터세트로 학습하였기 때문에 우리 사회의 고정관념과 권력 구조를 그대로 재생산할 위험이 매우 높습니다. 구글과 메타의 대규모언어모델의 학습데이터에는 인터넷 게시판, 언론, 공공기관 등 공개된 웹사이트에서 스크랩한 방대한 양의 텍스트가 포함되어 있습니다. 이는 이 데이터로 학습한 모든 생성형 AI 모델이 혐오 발언에서 광고에 이르기까지 모든 것을 포함하는 콘텐츠를 '학습'하였다는 의미이며, 이러한 모델이 생성하게 될 결과물에도 영향을 미칠 수 있다는 것을 의미합니다.

예를 들어, 인종차별적인 콘텐츠가 많이 포함된 인터넷 게시판에서 데이터를 스크랩하는 경우, 해당 데이터에 대해 학습한 모델은 인종차별적 결과물을 다시 생성할 위험이 있습니다. 실제로 이미지 생성형 AI는 여성, 특히 유색인종 여성을 남성보다 훨씬 더 높은 비율로 성적 대상화하는 경향을 보였습니다. '아프리카 노동자'와 같은 프롬프트는 육체 노동자의 사진을 생성하는 경향이 있었고 '유럽 노동자'는 사무직의 사진을 생성하는 경향이 있었습니다. 구글과 메타의 이미지 인식 인공지능은 피부색이 어두운 사람을 고릴라 또는 영장류로 분류해서 비판을 받기도 했습니다. 또다른 언어 모델도 장애인을 더 부정적인 감정 단어와 연결하는 경향을 나타냈습니다.[29] '엔지니어'를 검색하면 남성이, '사회복지사'와 '가사도우미'를 검색하면 유색인종 여성이, 'CEO'를 검색하면 백인남성의 이미지가 생성됐습니다.[30] 편향적인 데이터가 제대로 정제되지 않으면 성착취·인종차별·성별 고정관념 등 편향적인 경향이 AI결과물에 나타날 수 있는 것입니다.

이러한 편향은 인공지능의 모든 수명주기에서 일어날 수 있는 문제입니다. 인공지능이 학습하는 데이터에 이미 사회적 편견이나 차별적 요인이 내재되어 있어 인공지능도 그러한 편견·차별을 학습할 뿐 아니라, 라벨링[31]이나 강화학습에 참여한 사람[32]이나 개발자[33]가 가지고 있는 편견이 인공지능에 반영될 수도 있습니다. 인종과 같은 차별 속성을 의도적으로 직접 고려하지 않더라도 주거 형편 등 차별 속성과 상관관계가 높은 대리변수(proxy)를 사용하는 경우 비의도적이거나 간접적인 차별이 나타날 수 있습니다.

차별

아마존 채용 AI 사례가 심각한 이유는 학습데이터에 반영된 편향(bias)이 인공지능의 자동화된 결정을 거쳐 실제 차별(discrimination)로 이어졌다는 것입니다. 이와 유사한 문제로서, 비장애인 데이터 위주로 학습한 채용 AI는 장애가 있는 채용 지원자의 신체적, 행동적 특성을 불공정하게 평가하는 결과를 낳을 가능성이 높은 것으로 나타났습니다.[34] 장애로 인해 나타난 특성에 대하여, 긴장하고 있다거나 거짓말을 한다고 오해할 수 있는 것입니다.

인공지능이 사용한 데이터에 내포된 사회적이고 역사적인 편향이 그 결과물에 반영되면 인공지능의 결과물에 기반하여 이루어지는 의사결정 또한 사회적 차별을 악화시킬 위험이 있습니다. 인공지능의 편향이 실제 현실로 이어져 고정관념을 강화하거나, 특정 집단이나 관점을 불리하게 대우하는 차별적 결과를 야기할 수 있는 것입니다. 인공지능의 개발과 활용 결과에서 국내·외 인권 규범에서 금지하고 있는 직·간접적 차별이 때로는 은밀하게 실행될 수 있습니다. 실제로 미국 경찰의 경우 얼굴인식도구의 오류로 여러 번 무고한 흑인을 체포해서 소송이 제기되기도 하였습니다. 조사 결과 미국 경찰이 공권력에 사용하는 인공지능 얼굴인식도구가 흑인이나 아시아계를 잘못 인식할 가능성이 백인보다 10배~100배 높은 것으로 나타났습니다. 여성도 잘 식별하지 못했고, 중년보다 노년의 얼굴을 잘못 인식할 확률이 10배에 달했습니다.[35] 물론 사람이 내리는 결정도 차별적일 가능성이 있지만, 인공지능의 결정이 차별적인 경우, 훨씬 더 많은 사람들에게 장기간 영향을 줄 수 있습니다.[36]

2019년 10월, 미국 의료서비스에서 활용된 인공지능 알고리즘이 백인 환자들을 흑인 환자들보다 우대한다는 사실이 드러났습니다. 미국의 병원에서 2억 명 이상의 사람들에게 사용된 이 알고리즘은 어떤 환자에게 추가적인 의료 관리가 필요할지를 예측해 왔는데 인종적인 편향을 가지고 있었던 것입니다. 이 알고리즘은 인종 자체를 변수로 사용하지는 않았지만, 인종과 높은 상관관계를 가진 다른 변수를 사용했는데, 그것이 바로 '의료비 지출 내역'이었습니다. 이 변수를 사용한 이유는, 어떤 사람이 지출한 의료비는 그 사람이 가진 의료적 필요 정도를 압축적으로 보여줄 것이라는 가정 때문이었습니다. 하지만 같은 질환을 앓고 있더라도 흑인 환자들이 백인 환자들보다 평균적으로 더 낮은 의료비를 지출하는 경향이 있었습니다. 다행히 연구자들이 이 문제를 사전에 조사하여, 편향 수준을 80% 줄이는 조치를 개발하였습니다. 위와 같은 조사가 이루어지지 않았다면 인공지능 시스템의 편향성 때문에 사회 구조적으로 적은 의료비를 지출하는 흑인들이 계속 차별받는 결과를 야기하였을 것입니다.[37]

사회 불평등의 심화

미국의 법원은 피고의 재범 위험을 예측하기 위하여 '컴파스(COMPAS)'라는 알고리즘 시스템을 활용하고 있습니다. 2016년 이 알고리즘이 흑인을 차별하였다는 논란이 일었습니다.[38] 이 사례를 보도한 언론사 프로퍼블리카는 컴파스 알고리즘이 실제로는 재범을 저지르지 않은 흑인임에도 재범을 저지를 가능성이 높다고 잘못 예측하는 비율, 즉 흑인에 대한 위양성률(false positive)이 백인보다 두 배가 높았다고 비판하였습니다. 하지만 컴파스를 개발한 회사는 인

종 변수를 학습한 적이 없다고 주장하였습니다. 인구집단적으로 흑인의 재범률이 백인보다 높은 현실, 즉 진양성률(true positive)이 반영되었을 뿐이라는 것입니다. 그럼에도 컴파스 알고리즘은 피고인의 지인이 체포된 경험이 있는지, 피고인이 지난 1년 동안 여러 번 이사했는지, 피고인이 정학 또는 퇴학당한 적이 있는지 등 인종과 밀접한 상관관계가 있는 사회경제적 데이터를 사용하였다는 점에서 인종차별적 결과를 산출하였다는 지적을 받았습니다.[39]

어느 흑인 피고인은 이 예측 알고리즘이 공정하게 재판받을 권리를 침해하였다며 소송을 제기하였습니다. 법원은, 알고리즘의 점수가 판사가 고려하는 여러 요소 중 하나일 뿐이라며 기본권 침해를 인정하지 않았습니다. 그러나 개인에게 법적으로 중대한 영향을 미치는 알고리즘에 대한 구체적인 사항이 영업비밀이라는 이유로 공개되지 않은 점에 대해서는 비판이 계속되고 있습니다. 피고인이 자신의 위험 점수가 왜 그렇게 책정되었는지 알 수 없으면 자신의 혐의에 대하여 충분한 방어권을 행사할 수 없고, 판사도 정확한 논리를 이해하지 못한 채 알고리즘이 산출한 점수에 영향 받을 가능성이 있습니다. 무엇보다 이 알고리즘은 역사적으로 흑인이 과잉단속되었던 차별적 맥락을 고려하지 않고 오히려 그러한 불평등을 증폭시킨다는 비판을 받았습니다.

시간이 갈수록 상황이 악화될 수도 있습니다. 미국 경찰의 예측치안 도구는 우범지역을 예측해서 순찰 구역을 추천하는데, 우범지역은 대체로 유색인종 주민이 많은 가난한 지역일 가능성이 높습니다. 경찰이 이 지역을 집중적으로 순찰하면 가난한 유색인종 주민을 더 많이 단속하게 됩니다. 이런 패턴을 학습한 예측치안 도구는 또다시

이 지역을 우범지역으로 추천할 가능성이 높고, 이런 일이 영원히 반복될 수도 있습니다. 이런 문제를 피드백 루프(feedback loop)라고 합니다.[40]

유엔 빈곤과 인권 특별보고관은 인공지능이 개발되고 활용되는 과정 전반에 큰 인권적 문제가 있다고 분석하였습니다.[41] 첫째, 일반 인구집단의 행동에서 일반적으로 산출된 예측에 기반하여 개인의 권리에 미치는 결정을 내린다는 점에서 불공정한 측면이 있습니다. 둘째, 그런데 이 알고리즘의 기능이 대부분 불투명해서 인권침해를 발견하기가 쉽지 않습니다. 셋째, 결국 이런 인공지능의 활용이 불평등과 차별을 강화하거나 악화시킬 수 있습니다.

인공지능 편향에 대응하는 요구

유엔 사무총장은 인공지능 알고리즘이 기존 사회의 편견과 편향을 강화하여 차별과 사회적 배제를 악화시키는 경향이 있다고 지적하였습니다. 데이터 기반 도구는 종종 인간의 편견과 편향을 코드화하며, 이러한 편견과 편향의 대상인 여성과 소수자 및 취약 계층에게 불균형적인 영향을 미친다는 것입니다. 따라서 국가는 인공지능 등 신기술에 기반한 알고리즘과 자동화된 결정으로 인해 의도하지 않은 편향과 차별이 발생하지 않도록 그 원인과 영향력을 해결해야 할 시급한 필요성이 있습니다.[42]

국가인권위원회는 인공지능의 발달이 편향과 차별 등 인권 문제를 증가시키고 있다고 평가하며, 이를 방지하는 조치를 구체적으로 마련할 것을 권고해 왔습니다.[43] 인공지능의 편향성은 하나의 현상이며, 차별은 인공지능을 활용해 현실에 적용되는 결정을 내릴 때 발생

합니다. 고용, 의료, 사법 등 중요한 사회영역에서 인공지능에 기반한 결정이 편향적으로 내려지면 사회적 차별로 이어질 수 있습니다. 이는 사회 전체에 대한 개인의 신뢰를 훼손하고 사회서비스의 원활한 제공에도 지장을 초래할 수 있습니다.[44)]

국가인권위원회 〈인공지능 개발과 활용에 관한 인권 가이드라인〉

32. 인공지능을 개발하고 활용할 때는 인공지능으로 인해 영향을 받는 사람의 다양성과 대표성을 반영하기 위해 노력해야 하고, 성별, 종교, 장애, 나이, 출신 지역, 신체조건, 피부색, 성적 지향, 사회적 신분 등 개인과 집단의 특성에 따라 편향적이고 차별적인 결과가 나오지 않도록 하여야 합니다.

33. 또한 인공지능의 결정이 특정 집단이나 일부 계층에게 차별적이거나 부정적 영향을 초래하지 않기 위해 개발 단계부터 다양한 계층의 의견을 수렴하고, 차별적 결과가 발생하지 않도록 필요한 조치를 취해야 합니다.

34. 학습데이터의 수집과 선정, 알고리즘의 설계와 활용방향 설정 등 인공지능 개발 전 과정에 걸쳐 편향이나 차별의 요소가 배제될 수 있도록 점검하는 절차를 마련해야 합니다. 여기에는 학습데이터의 개별 요소를 검사하고 차별적 영향을 초래할 수 있는 데이터를 조정하는 등의 조치가 포함되어야 합니다.

35. 특히 학습용 데이터가 인공지능의 판단에 직접적인 영향을 미치는 상황을 고려할 때, 학습용 데이터의 수집 단계부터 차별적 요소를 통제하고 데이터 편향성을 최소화하여 인공지능을 통한 의사결정이 특정 집단에 부정적 영향을 미치지 않도록 해야 합니다.

36. 개발한 인공지능에 대해 주기적인 모니터링을 거쳐 데이터 품질과 위험을 관리하고, 차별적 결과나 의도치 않은 결과에 대해 개선의 조치를 주기적으로 수행해야 합니다.

37. 인공지능 기술 및 서비스에 대한 접근성과 인공지능이 주는 혜택은 사회적 약자를 포함하여 모든 사회구성원에게 평등하게 제공되어야 합니다. 또한 모든 사람의 인공지능에 대한 지식과 이해를 촉진시키고, 인공지능의 활용이 어려운 계층에 대한 교육 및 지원이 이루어져야 합니다.

가장 큰 문제는 우리 사회에 역사적이고 사회적인 불평등이 존재하는 것이 사실이고, 인공지능이 이런 편향을 학습하고 증폭시킨다는 점입니다. 따라서 궁극적인 해결책은 포괄적 차별금지법을 제정하여 우리 사회의 편견과 차별 문제를 해결하는 것입니다.

한편, 위험한 차별적 영향을 미치는 인공지능은 법률로 엄격한 의무를 부과하거나 아예 금지할 필요가 있습니다. 「유럽연합 AI법」은 장애인 등의 취약성을 악용하여 심각한 피해를 야기하는 인공지능 시스템이나 직장 및 교육기관의 감정인식 AI을 금지하였습니다. 고위험 인공지능을 배치하는 사업자는 차별적 영향에 대한 평가를 포함하여 기본권 영향 평가를 실시하여야 합니다. 미국 콜로라도주 「인공지능시스템과 상호작용을 하는 소비자 보호에 관한 법(이하 '콜로라도 인공지능법')」은 법률로 금지한 차별을 야기하는 알고리즘 시스템을 금지하고 고위험 분야에서 인공지능을 개발하고 배치할 때 차별적 영향이 나타나지 않는지 사전적인 영향 평가를 실시하고 조치하도록 하였습니다.

4. 인공지능의 사회적 영향

인권은 특정한 권리를 넘어 보편적이며, 양도 불가능하고, 불가분적이며, 상호 의존적이고, 상호 연관되어 있는 체계입니다. 앞서 인공지능의 데이터와 알고리즘이 인권에 미치는 영향을 살펴보았을 때 개인정보 자기결정권, 사생활의 권리, 차별받지 않을 권리와 같은 인권이 깊은 영향을 받는 것을 보았습니다. 그러나 인공지능이 사회적으로 인권에 미치는 영향은 더 폭이 넓고 때로는 장기적인 위험을 낳을 수 있습니다.

우선 인공지능 기술은 표현의 자유에 다양한 측면에서 영향을 미칠 수 있습니다. 개인화 알고리즘은 정보가 넘쳐나는 시대에 이용자들이 원하는 정보를 쉽게 찾을 수 있도록 할 수 있습니다. 생성형 AI 도구들은 글쓰기나 그림을 그리는 기술적 역량이 부족한 사람들도 자신이 원하는 작품을 쉽게 만들 수 있도록 지원할 수 있습니다. 반면, 플랫폼 알고리즘은 이용자가 어떠한 정보를 접할 것인지 왜곡할 수 있고 허위정보나 자극적인 콘텐츠가 더 쉽게 유통되도록 함으로써 이용자의 정보접근권을 제한하고 공론장 형성에 부정적인 영향을 미칠 수 있습니다. 또한, 콘텐츠 관리를 위한 플랫폼의 알고리즘이 불법이 아닌 표현까지 삭제하여 이용자의 표현의 자유를 침해할 수 있습니다. 생성형 AI 도구의 발전은 허위조작정보를 쉽게 생성할 수 있도록 하는데, 이는 풍자나 패러디 목적으로 활용될 수도 있지만 정치적, 경제적 목적으로 사회적인 논란을 야기할 수도 있습니다.[45] 2018년 유엔 표현의 자유 특별보고관은 인공지능의 특성 중 자동화, 데이터 분석, 적응성이 특히 표현의 자유를 비롯하여 전반적인 인권에 미치는 영향이 크다고 지적하였습니다.

한편으로는 인공지능 기술이 발전하면서 노동권에 중대한 영향을 미칠 수 있습니다. 노동시장에서 일자리가 감소하고, 노동조건이 기계적으로 결정되며, 노동환경이 전자적으로 상시 감시되는 등의 문제가 이미 발생하고 있습니다. 인공지능 기술이 인간의 노동력을 대체함으로서 인간노동의 존엄성이 훼손되고 일할 권리와 고용 안정에 위협이 발생하며, 고용·노동조건 및 평가 과정에 인공지능 기술을 활용함으로써 채용과 해고를 비롯한 노동관계 전반에서 불투명하고 편향적인 결정이 이루어질 수 있는 것입니다. 특히 국제노동기구(ILO)는 인공지능 기술의 발전으로 직장 모니터링을 위해 디지털 감시 시스템이 점점 더 많이 배치되고 노동 그 자체에 대한 일상적인 감시가 만연할 수 있다고 우려했습니다. 인공지능 감시 시스템은 노동과정 그 자체에 대한 감시는 물론, 노동자의 내밀한 생각이나 감정 등 과거보다 더 방대하고 더 사적인 정보를 수집하고, 노동자 간 친분 관계에 대한 추론으로 노동자의 단결권이나 노동조합 활동을 위축시킬 수 있습니다[46]. 인공지능 기술을 장착한 로봇 제조시스템이 확산됨에 따라 사람에게 안전한 노동환경을 위협하는 문제도 나타나고 있습니다.

환경권 또한 인공지능 기술의 발전에 깊은 영향을 받습니다. 인공지능은 기후 변화 대응, 에너지 효율 개선, 생태계 보전 등 환경 문제 해결에 활용되고 있습니다. 예를 들어, 인공지능은 위성·기상 데이터를 분석해 장기 기후 변화 시나리오를 예측하거나, 전력망을 최적화하고, 폐기물의 재질과 상태를 분석해 관리를 지원하며, 생태계 모니터링 등에 사용될 수 있습니다. 그러나 인공지능 개발과 운영에는 대규모 데이터센터와 고성능 컴퓨팅 자원이 필요하며, 이 과정에서 막대한 전력과 물과 같은 자원이 사용되는데, 이에 따라 온실가스 배

출, 전자폐기물 증가 등의 부정적 환경 영향이 발생합니다.

국가인권위원회는 표현의 자유, 정보 접근 및 의견 표명에 영향을 미치는 국가가 적절히 조치를 마련해야 한다고 지적하였습니다. 더불어 집회 및 결사의 자유, 노동권 등 인권 전반을 인공지능의 부정적 영향으로부터 보호해야 한다고 강조하였습니다.

국가인권위원회 〈인공지능 개발과 활용에 관한 인권 가이드라인〉

17. 인공지능의 개발과 활용은 개인의 행복과 사회적 공공성의 증진에 위배되어서는 안 되고, 표현의 자유, 집회 및 결사의 자유, 노동권 등 인권을 인공지능의 부정적 영향으로부터 보호해야 합니다.

50. 국가는 대량 감시와 차별로 이어지고, 집회 및 결사의 자유에 부정적 영향을 행사할 위험이 높은 얼굴인식 등 원격 생체 인식 기술의 사용을 공공장소에서 원칙적으로 금지해야 합니다. 해당 기술은 특별한 경우에만 사용하되, 인권침해나 차별의 위험성이 드러난 경우 이를 방지하거나 완화하는 조치를 취하기 전까지는 사용을 중단해야 합니다.

51. 국가는 다양한 정보가 자유롭게 유통되는 정보 환경을 조성하고, 표현의 자유, 정보 접근 및 의견 표명 등에 부정적 영향을 미치는 인공지능에 대해 적절한 조치를 마련해야 합니다.

인공지능과 인권 의무

1. 기업의 인권 의무

휴스턴 교사평가 알고리즘 사례

미국 휴스턴 지역의 학교 교사들은 2011년부터 2015년까지 '에바스(EVAAS)'라는 평가 알고리즘을 통해 직무 평가를 받았습니다. 휴스턴 교육청은 평가 점수가 나쁜 교사의 85%를 해고한다는 목표를 세웠습니다. 대상이 된 교사들은 자신의 점수가 산출된 방식과 해고되는 이유에 대하여 납득할 만한 설명을 듣고 이의를 제기하고 싶었습니다. 하지만 교육청은 교사들의 요청을 거부하였습니다. 평가 알고리즘에 대한 정보는 이 프로그램을 제공한 업체의 영업 비밀이기 때문에 공개할 수 없다는 것이었습니다. 심지어 교육청 자신도 이 민간업체의 평가 알고리즘을 이해할 수 없었습니다. 교사노동조합은 2014년 불투명한 평가 알고리즘에 대해 소송을 제기하였습니다.

2017년 법원은 교사들의 적법절차 권리가 침해되었다고 인정하였습니다. 공공기관이 알고리즘을 활용하면서 영업 비밀과 적법절차의 권리 사이에 균형을 맞추기 위해서는 중요한 공공 의사결정에 비밀 알고리즘을 활용하지 않아야 한다는 것이었습니다.[47]

이 사건은 교육청이 공공 인공지능의 평가 대상이 된 교사들에게 적법절차를 보장할 의무가 있었는데, 민간 기업의 영업 비밀을 이유로 그 의무를 위반한 사건이었습니다. 적법절차의 원칙이란, 형사절

차나 행정절차 등 국가 작용이 국민의 이해관계에 불이익한 결정을 내릴 때, 대상이 된 당사자에게 고지하고, 의견을 청취하고, 방어기회를 주는 등 절차적으로 투명하고 적정해야 한다는 헌법상 원칙입니다.[48] 적법절차의 원칙은 모든 공공기관이 자신이 운영하는 공공인공지능에도 적용해야 합니다. 결정의 이유를 투명하게 설명할 수 있어야 하고, 이의 제기 의견을 받아야 합니다.

이처럼 국가는 자신의 공권력이 그 대상이 되는 사람의 인권을 침해하지 않아야 할 소극적 의무가 있습니다. 국가는 여기서 더 나아가 적극적으로 제3자가 누군가의 인권을 침해하는 것을 방지하기 위한 조치를 취해야 합니다. 이러한 적극적 의무를 보호 의무라고 합니다. 기업 역시 그 사업 과정이나 결과 면에서 인권에 직·간접적인 영향을 미치기 때문입니다.

기업과 인권 이행지침

인권법은 본래 기업이 아닌 정부를 대상으로 발달해 왔습니다. 그러나 세월이 흐르면서 기업도 인권에 큰 영향을 미친다는 사실이 명확해지면서, 2011년 유엔은 기업의 인권 책임을 설명하는 〈기업과 인권 이행지침(UNGP)〉을 채택하였습니다. 이러한 요구는 이후 OECD 〈다국적 기업 책임 경영 지침〉에도 반영되었습니다.

이러한 국제 인권 규범에 따르면 국가에는 인권을 보호할 의무가 있고, 기업에는 국제적으로 인정되는 모든 인권을 존중해야 할 책임이 있습니다. 즉, 기업은 인권침해를 방지하고 관련된 부정적인 영향에 대처해야 할 의무가 있습니다. 이러한 인권 의무를 이행하기 위하여 기업은 인권존중을 위한 정책을 선언하고, 인권 영향 평가를 비롯

하여 인권실사 절차를 실시하며, 회사가 야기하거나 기여한 인권침해에 대하여 구제를 제공하여야 합니다.[49] 물론 국가가 통치하거나 규제하는 정부로서가 아니라 기업을 소유하거나 운영하는 행위자로서 경제 활동을 하는 경우 이 규범은 국가기관에도 적용됩니다.

국제 인권 규범에서 이야기하는 인권 의무의 핵심은 국가와 기업이 인권을 보호하는 조치를 취해야 한다는 것입니다. 따라서 국가와 기업은 자신이 개발하고 활용하는 인공지능이 누군가의 인권에 미치는 부정적인 영향에 대해서도 이를 예방하거나, 완화하거나, 구제할 수 있는 조치를 취해야 합니다. 유엔 인권최고대표는 인공지능 제품과 서비스도 〈기업과 인권 이행지침〉을 지켜야 한다고 여러 차례 말했습니다. 유엔 사무총장은 인공지능 신기술을 개발하고 활용하는 국가기관과 기업들이 〈기업과 인권 이행지침〉을 비롯한 국제 인권법을 준수할 의무가 있다고 강조하였습니다.[50]

유엔 사무총장 (2020)

기술 발전의 혜택을 온전히 누리고 피해 가능성을 최소화하기 위해서는, 신기술의 개발과 배치가 견고한 인권 기반에 뿌리를 두어야 합니다. 국제 인권법은 국가 간에 합의된 내용이고 국가적, 지역적, 국제적 메커니즘이 모니터링하고 있습니다. 따라서 끊임없이 변화하는 기술 환경 문제에 대응할 때 핵심적인 지침이 될 수 있습니다. 인권법은 실체적 권리와 절차적 권리를 담고 있으며, 이 권리 침해 피해를 예방하거나, 완화하거나, 구제할 것을 요구합니다. 또한, 인권법은 인권을 존중하고, 증진하며, 보호해야 할 국가의 의무를 부과하고, 기업 또한 이와 같은 책임을 이행할 수 있는 프레임워크를 제시합니다.

인공지능 인권 영향 평가

기업과 인권에 대한 국제 규범에서 국가와 기업이 인권 의무를 이행하기 위해서는 '인권실사(due diligence)'라고 불리는 주의 의무를 사전에 다하는 것이 매우 중요합니다. 기업이 인권 영향 평가를 비롯한 인권실사를 수행하면서 자신이 부정적인 인권 영향을 야기하거나 기여한 사실을 확인하면 그 영향을 방지하거나 완화해야 하고, 확인된 피해에 대해서는 구제를 제공하거나 이에 협력해야 한다는 것입니다. 이때 인권 영향 평가는 자신의 사업이 실제적 또는 잠재적으로 인권에 부정적인 영향을 미치는지 평가하고, 이러한 부정적인 영향을 중단, 예방 또는 완화하기 위한 조치를 취하는 과정입니다. 국제 인권 규범은 잠재적으로 영향을 받는 권리주체들에게 인권 영향 평가 과정과 결과를 공개하고 참여를 보장할 것을 요구해 왔습니다.

인공지능 머신러닝은 매우 불투명합니다. 유엔 인권최고대표(2021)는 "많은 인공지능 시스템의 의사결정 과정은 불투명하다. 인공지능 시스템의 개발 및 운영을 뒷받침하는 정보 환경, 알고리즘, 모델의 복잡성은 물론 정부와 민간 행위자들의 의도적인 비밀주의는, 인공지능 시스템이 인권과 사회에 미치는 영향을 일반 대중이 이해할 수 있는 뜻 깊은 여정을 방해하는 요인이다."고 지적하였습니다.[51]

휴스턴 사례에서처럼 인공지능 시스템을 개발한 기업이 영업비밀을 고수하면, 인공지능 시스템의 대상이 된 사람은 물론, 이를 조달받아서 업무에 활용하는 공공기관 또는 다른 기업들조차 그 작동원리를 이해하기가 어렵습니다. 또한 머신러닝 시스템은 사람이 설명

하기 어렵거나 불가능한 방식으로 패턴을 인식하고 결과물을 산출할 수 있습니다. 이를 흔히 '블랙박스' 문제라고 합니다. 이런 불투명성은 인권 의무를 이행하기 어려운 상황을 초래할 수 있습니다. 인공지능으로 인한 인권침해가 발생하더라도 공공기관이 이를 조사하거나 해당 인권침해에 관련된 기업에게 책임을 묻기가 어려울 수 있기 때문입니다.

> 유엔 인권최고대표 (2021)
>
> 머신러닝 시스템은 불투명성의 핵심적인 요인입니다. 이 시스템은 설명하기 어렵거나 불가능한 방식으로 패턴을 식별하고, 설명하기 어렵거나 불가능한 처방을 내릴 수 있습니다. 이를 흔히 "블랙박스" 문제라고 합니다. 불투명성으로 인해 인공지능 시스템을 유의미하게 조사하는 것이 어려워지고, 인공지능 시스템이 위해를 야기하는 경우 불투명성이 효과적인 책무성 확보에 장벽이 될 수 있습니다.

이와 같은 인공지능의 불투명성을 고려해 보면, 인공지능으로 인한 인권침해가 일어나기 전에 이를 예방하는 일이 특히 중요할 것입니다. 인권 영향 평가는 사전적인 조치를 취하는 데 핵심적인 과정입니다.

국가인권위원회에 따르면 "인권 영향 평가는 국가와 기업 등 공적·사적 주체가 시행하거나 추진하는 정책이나 사업과정 등이 인권에 미치는 부정적 영향을 사전에 평가하여 이를 식별하고 방지, 완화하는 조치를 취하는 과정"입니다.[52] 인공지능에 대한 인권 의무가 요

구되면서 유럽연합 AI법을 비롯한 국제 규범에서 인공지능이 인권에 미치는 영향을 평가하는 제도들이 도입되기 시작했습니다.[53]

국가인권위원회는 2024년 7월 8일에 인공지능 인권 영향 평가 도구를 공개하였습니다.[54] 국가인권위원회 인공지능 인권 영향 평가 도구는 4단계별에 걸쳐 72개 문항으로 구성되어 있습니다.

첫 번째 단계는 인권 영향 평가를 계획하고 준비하는 단계입니다. 영향 평가의 수행은 관련된 사업부서와 독립된 조직 또는 전문성을 갖춘 외부기관이 수행하는 것이 바람직합니다. 두 번째 단계는 인권 영향을 분석하고 평가하는 단계입니다. 데이터, 알고리즘, 심각도별로 인권영향의 정도를 평가하기 때문에 인권 영향 평가에서 핵심적인 단계입니다.

세 번째 단계는 인권영향을 개선하거나 구제하는 단계입니다. 인권 영향 평가는 인권에 미치는 부정적 영향에 대하여 방지, 완화 및 구제 조치를 이행하는 것을 중요하게 여깁니다. 네 번째 단계는 인권 영향 평가의 결과를 공개하고 점검하는 단계입니다. 인권 영향 평가를 마친 이후에도 해당 인공지능 시스템의 활용 과정에서 문제가 발생하지 않는지 모니터링하며, 문제가 발생할 경우 이에 대응할 수 있도록 지속적으로 점검한다는 점에서 반복적인 과정이라고 볼 수 있습니다.

인권 영향 평가에서는 인권에 부정적인 영향을 받을 수 있는 개인이나 단체를 비롯한 이해관계자를 참여시키는 것이 필수적입니다. 이러한 참여는 특정 단계가 아니라 모든 단계에서 고려되고 이행해야 할 사항입니다.

2. 영향 받는 사람

영향 받는 사람의 개념

국제 인권 규범은 국가와 기업이 인공지능의 인권 위험으로부터 사람을 보호해야 할 의무가 있으며, 특히 부정적인 인권 영향을 받게 될 사람을 보호해야 한다고 강조합니다. 인공지능으로부터 이러한 '영향을 받는 사람'은 인공지능 환경의 권리주체입니다.

예를 들어 어떤 공공기관이 활용하는 인공지능 채용도구가 사투리를 잘 인식하지 못하여 사투리를 사용하는 사람에 대하여 부당한 불합격 결정을 내렸을 경우, 특정 지역 출신인 그 사람 뿐 아니라 유사한 사투리를 사용하는 지역의 사람들 모두가 이 인공지능이 내리는 결정에 의해 실제적 또는 잠재적으로도 부정적인 영향을 받는 사람들이 됩니다. 환자의 질병을 진단하고 수술 필요 여부를 제안하는 병원 인공지능의 경우에도, 여성 환자에 대해서 학습이 잘 되어 있지 않아 오진을 내렸을 경우 이 병원 인공지능으로 인해 부정적인 영향을 받는 것은 여성들입니다.

국가인권위원회는 영향 받는 사람을 주목하면서 그 권리를 보장할 것을 요구했습니다. 인공지능의 위험으로 인하여 인권에 부정적인 영향 받는 사람은 특히 인공지능이 자신의 생명, 안전, 기본권에 중대한 영향을 미치는 결정을 내릴 때 그에 대한 법적인 보호를 받을 수 있어야 합니다.

국가인권위원회 〈인공지능 개발과 활용에 관한 인권 가이드라인〉

4. 인공지능으로 영향을 받는 당사자들은 인공지능의 도입, 운영, 결정에 대하여 참여의 기회를 보장받고 있지 못하며, 인공지능으로 인한 인권침해가 발생한 경우에도 적절하고 효과적인 권리구제를 받을 수 있는 절차와 방법이 미흡한 상황입니다.

13. '영향을 받는 당사자(affected individuals)'는 국가나 기업의 규정 또는 행위 등으로 인하여 인공지능의 적용대상이 되고, 직·간접적으로 인권에 영향을 받는 개인 또는 집단을 의미합니다.

19. 이러한 알 권리의 보장과 인공지능이 미치는 영향력과 중요성을 감안할 때, 인공지능의 판단과정과 그 결과에 대한 적절하고 합리적인 설명이 보장되어야 합니다. 학습 및 추론, 판단의 과정과 결과에 이른 이유를 설명하기 어려운 인공지능은 이에 대한 대응의 불확실성과 영향을 받는 당사자의 불안감을 유발하고, 인권 및 안전에 관한 법령과 정책의 집행 효과를 불분명하게 할 수 있습니다.

23. 또한, 인공지능에 의한 자동화된 의사결정이 예정되어 있는 경우, 영향을 받는 당사자들은 사전에 그 사실을 알아야 합니다. 자동화된 의사결정에 의하여 영향을 받는 당사자는 그 결정의 이유에 대하여 설명을 듣고, 당사자 진술을 할 수 있으며, 이의를 제기할 수 있어야 합니다.

32. 인공지능을 개발하고 활용할 때는 인공지능으로 인해 영향을 받는 사람의 다양성과 대표성을 반영하기 위해 노력해야 하고, 성별, 종교, 장애, 나이, 출신 지역, 신체조건, 피부색, 성적 지향, 사회적 신분 등 개인과 집단의 특성에 따라 편향적이고 차별적인 결과가 나오지 않도록 하여야 합니다.

인공지능의 위험으로부터 영향 받는 사람을 보호하는 제도는 이제 도입 단계이지만 우리나라 뿐 아니라 국제적 규범으로 계속 발달하고 있습니다. 자동화된 결정과 같이, 누군가의 인권에 영향을 미치는 인공지능이 개발되고 활용될 때에는, 그 영향을 받는 사람을 보호해야 하는 의무를 다해야 합니다. 사람에게 중대한 영향을 미치는 인공지능을 개발하고 활용하는 사업자는 당사자에게 그 과정과 결과를 일정 수준 이상으로 설명할 수 있도록 조치해야 합니다. 설명할 수 없는 인공지능은 그 대상이 된 사람에게도, 사회 전체적으로도 부정적인 영향을 미치고 인공지능에 대한 신뢰를 잃게 만들 수 있습니다.

유엔 사무총장은 인공지능을 개발하고 활용하는 일을 결정하는 과정에도 영향 받는 사람이 참여하는 것이 중요하다고 말합니다. 국가적인 결정 뿐 아니라 직장 등 인공지능이 활용되는 현장에서도 영향을 받는 당사자가 참여하여 인공지능의 활용에 대한 결정이 이루어지는 것이 중요합니다.

> 유엔 사무총장 (2020)
>
> 국제 의무를 준수하는 신기술의 개발, 확산 및 채택은 권리주체의 효과적이고 의미 있는 참여를 통해 향상될 수 있습니다. 이를 위해 국가는 권리주체, 특히 가장 큰 영향을 받거나 부정적인 결과를 겪을 가능성이 높은 권리주체가 개발 과정에 효과적으로 참여하고 기여할 수 있는 기회를 창출하고, 특정한 신기술의 채택을 촉진해야 합니다. 국가는 참여 보장과 포용적 의견수렴을 통해서, 경제적 효율성, 환경적 지속가능성, 포용성 및 형평성을 갖춘 균형적이고 통합적인 지속가능개발목표에 있어 어떤 기술이 가장 적절하고 효과적인지 결정할 수 있습니다.

유럽연합 AI법은 인공지능 사업자가 영향 받는 사람들에게 인공지능 활용 사실을 공개하고 협의하는 절차를 두고 있습니다. 특히 직장에서 고위험 인공지능을 도입하는 회사는 사전에 노동자 대표자와 대상이 되는 노동자에게 그 사실을 알리도록 하였습니다.

영향 받는 사람의 구제

2025년 유엔 인권최고대표사무소는, 인공지능 기업에 〈기업과 인권 이행지침〉을 적용하는 문제에 대해 보고서를 발표하였습니다.[55] 인공지능을 개발하고 활용하는 공공기관이나 민간기업은 인권침해가 발생하였을 때 적절하고 효과적인 구제를 보장해야 합니다. 하지만 인공지능 환경에서 구제가 쉽지 않은 이유는 인공지능 기술과 환경이 매우 복잡하고 불투명하기 때문입니다.

유엔 인권최고대표사무소 (2025)

인공지능 기술이 초래하는 인권 위험은 편향성, 차별, 사생활 침해에 그치지 않고 건강 위험, 복지 문제 및 표현의 자유와 정보접근권 등 여타 인권 문제까지 포괄합니다. 모든 위험을 인공지능 배치 전에 완전히 예측할 수 있는 것이 아니며, 의도하지 않았거나 예견할 수 없는 위험도 존재합니다. 이를 해결하기 위해서는 포괄적이고 다각적인 접근이 필요합니다. 인공지능 관련 피해자들이 구제수단에 접근할 때는 복잡성과 불투명성 등 인공지능 시스템의 기술적 특성으로 인하여 추가적인 어려움을 겪을 수 있습니다. 이는 의사결정 과정과 다양한 이해관계자의 관여방식을 이해하기 어렵게 만들어 책임 소재에 대한 판단을 극도로 복잡하게 만듭니다.

OECD도 유사한 문제를 지적합니다.[56] 결과물이 어떻게 생성되었는지 설명하기 어려운 인공지능은, 해로운 편향을 감지하거나 완화하기 어렵고, 문제가 발생했을 때 책임 소재를 판단하는 것도 어렵게 만든다는 것입니다. 이런 깜깜이 인공지능이 많은 사회 분야에서 사용되면, 개인은 물론 사회적으로도 인공지능의 인권 위험이 악화될 수 있습니다. 개인과 사회가 겉으로는 효율적이지만 잠재적으로 편향이나 결함이 있는 인공지능에 과도하게 의존할 수 있고, 겉에서 보기에는 이런 결함들을 발견할 수 없기 때문에 인공지능의 위험과 편향이 계속될 수 있는 것입니다.

인권침해에 대한 책임 문제를 판단하려면 복잡한 인공지능 환경 속에서 여러 기술적 요소와 여러 행위자의 역할에 대해 파악할 수 있어야 합니다. 예를 들어 경찰이 활용하는 얼굴인식 인공지능이 잘못된 체포를 수행한 경우,[57] 경찰에게 우선적인 책임이 있겠지만, 인공지능이 문제의 원인을 제공했을 수도 있습니다. 센서가 잘못 인식했을 수도 있고, 센서가 수집한 데이터를 분석해야 하는 알고리즘이 잘못된 추론을 하였을 수 있고, 알고리즘이 기반하고 있는 모델이 이미 편향적이었을 수 있고, 알고리즘이 추가적으로 학습한 데이터에 인종 편향이 있었을 수도 있습니다.

게다가 인공지능 시스템이 개발하는 주체로부터 활용하는 주체로 이전되고 또 시간이 지나는 동안 언제 어떤 데이터를 학습하였고 언제 어떻게 알고리즘을 정렬하였는지, 기록이 남아있지 않거나 작동원리가 불투명한 경우도 있습니다. 이런 상황에서는 인공지능으로 인한 피해를 찾아내고 입증하는 일이 쉽지 않습니다.

그래서 인공지능의 책임 문제를 해결하기 위하여 세계 여러 나라에서 인공지능 사업자에게 여러 주의 의무를 부과하는 인공지능법들이 마련되기 시작했습니다. 국제 인권 규범은 인공지능 관련 입법에서 영향 받는 사람을 보호하기 위하여 설명, 인간의 관리·감독, 문서의 작성 및 보관 등 주의 의무를 다하고 피해 구제를 위한 조치를 의무적으로 취할 의무가 있다는 점을 분명히 밝히고 있습니다.

유엔 인권최고대표사무소는 국가와 기업에 피해구제를 위한 구체적인 조치를 취할 것을 요구합니다. 우선 국가는 인공지능으로 그 권리가 침해된 사람에게 효과적인 구제수단과 완전한 배상을 받을 수 있도록 보장하는 제도를 마련해야 한다는 것입니다. 특히 사람에 대해 결정을 내리는 고위험 인공지능에 대해서는 투명성을 보장하고 인간의 감독을 보장하는 조치를 취하도록 요구해야 합니다. 당사자 영향 받는 사람들이 자신의 권리와 구제수단에 대한 이해가 가능하도록 국가가 리터러시나 대중적 교육을 지원하면 더욱 좋을 것입니다. 장애인이나 노인 등 취약한 사람들에 대한 특별한 지원 또한 필요합니다. 이와 더불어 인공지능을 개발하거나 활용하는 기업도 구제에 필요한 체계를 갖추거나 협력하고 영향 받는 사람이 접근할 수 있도록 보장해야 합니다.

> 유엔 인권최고대표사무소 (2025)
>
> 55. 국가는 국제 인권법 및 〈기업과 인권 이행지침〉을 포함한 국제 기준에 따라 다음을 이행해야 합니다.
>
> (b) 인공지능 제품과 서비스로 인해 개인의 권리가 침해될 경우,

효과적인 구제수단과 완전한 배상을 받을 수 있도록 보장해야 하며, 여기에는 인공지능이 지원한 의사결정 과정에 대해 투명성을 요구하고 해당 결정에 대해 의미 있는 인간 감독을 요구하는 등의 조치가 포함됩니다.

(f) 영향 받는 이해관계자의 디지털 리터러시를 지원하여, 이들이 이용 가능한 구제수단에 대한 정보에 포용적이고 명확한 언어와 형식으로 접근할 수 있도록 보장해야 합니다.

(g) 저소득층 및 소외 계층이 구제 제도에 접근하는 데 불균형적인 영향을 미치는 비용적, 절차적 장벽을 제거해야 합니다. 인공지능과 관련된 잠재적 피해 및 이용 가능한 구제수단에 대하여 대중적 인식 제고 및 홍보 전략에 투자하고, 가장 큰 영향을 받는 지역사회와 이를 공동으로 개발하여야 합니다.

57. 인공지능을 개발하고 배치하는 기업은 〈기업과 인권 이행지침〉을 포함하여 해당되는 국제 기준에 따라 다음을 이행해야 합니다.

(c) 인공지능 제품과 서비스에 대한 효과적인 고충 처리 메커니즘을 운영 수준에서 수립하거나 이에 참여하고, 국가적 사법 제도와의 협력을 포함하여 영향 받는 개인 및 지역사회에 효과적인 구제수단을 제공해야 합니다.

국가인권위원회 또한 인공지능 기업들이 문서화 등 절차적 조치를 취해야 하고, 인권침해 피해 구제를 위하여 이러한 자료에 감독 기관과 피해자가 접근할 수 있어야 한다고 설명하였습니다.

국가인권위원회 〈인공지능 개발과 활용에 관한 인권 가이드라인〉

47. 감독 기관은 공공기관과 민간의 위법한 인공지능 개발과 활용 여부를 조사하고 피해 구제 및 조치를 취하기 위하여 상세 정보에 접근할 수 있어야 합니다. 이를 위하여 공공기관 인공지능 및 민간 고위험 인공지능 개발자 및 운영자는 사용된 데이터와 알고리즘의 주요 요소 등을 기록하고 문서화하여 일정 기간 보관하여야 합니다.

49. 국가는 인공지능으로 인하여 인권을 침해당하거나 차별을 받은 사람이 진정을 제기하여 권리를 구제받을 수 있는 기회를 보장하는 등 국가 기관의 구제수단에 대한 접근을 제공해야 합니다. 인공지능을 개발하고 활용하는 공공기관과 민간은 언제든지 구제가 가능하도록 그 책임자에 대한 정보는 물론, 이의를 제기할 수 있는 기관과 방법에 대한 정보를 일반에 공개하여야 합니다.

유럽평의회는 2024년 9월 4일, 인공지능에 대하여 구속력 있는 국제협약으로 「인공지능과 인권·민주주의·법치주의에 관한 기본협약」을 세계 최초로 공개하였습니다. 이 AI 국제협약은 협약 비준국들에게 영향 받는 사람에 대한 구제 제도를 갖출 것을 요구하였습니다.[58] 첫째, 인권에 중대한 영향을 미칠 가능성이 있는 인공지능 관련 정보를 관할 당국에 제공해야 하고, 해당하는 경우 영향 받는 사람에게도 제공해야 합니다. 둘째, 이 정보는 영향 받는 사람이 이의를 제기하거나 이해하기에 충분해야 합니다. 셋째, 관련된 사람은 국가에 진정을 제기할 수 있어야 합니다.

인공지능과 인권·민주주의·법치에 관한 기본 협약

제14조 - 구제 수단

1. 각 당사국은 국제적 의무에 따라 구제 수단이 요구되고 국내 법률 체계에 부합하는 범위 내에서, 인공지능 시스템의 수명 주기 내 작동으로 인한 인권침해에 대하여 접근 가능하고 효과적인 구제 수단을 이용할 수 있도록 보장하는 조치를 채택하거나 유지해야 합니다.

2. 위 1항을 지원하기 위해 각 당사국은 다음과 같은 조치를 채택하거나 유지해야 합니다.

a. 인권에 중대한 영향을 미칠 가능성이 있는 인공지능 시스템과 그와 연관된 사용에 관한 관련 정보를 문서화하여, 해당 정보에 접근할 권한이 있는 기관에 제공하고, 적절하고 해당하는 경우 영향을 받는 사람에게 제공하거나 전달할 수 있는 조치

b. 전 a호에 언급된 정보가, 영향을 받은 사람이 시스템 사용에 의해 내려진 결정에 이의를 제기하거나 시스템 사용에 대하여 실질적인 정보를 제공하기에 충분하고, 관련 있고 적절한 경우 시스템 자체에 대한 실질적인 정보를 제공하기에 충분하도록 보장하는 조치

c. 관련된 사람이 관할 당국에 진정을 제기할 수 있는 효과적인 방법

인공지능법과 인권 과제

1. 유럽연합

유럽연합 AI법의 주요 내용

유럽연합 AI법[59]은 세계 최초로 인공지능을 포괄적으로 규제한 법입니다. 유럽연합 AI법은 위험 기반 접근(risk-based approach)에 기반하여 인공지능 시스템을 4가지 위험등급으로 나누고 차등적으로 규제합니다.

우선 '허용할 수 없는 위험'을 가진 인공지능 시스템, 즉 국민의 안전, 생계, 그리고 권리에 명백한 위협으로 간주되는 인공지능 시스템의 사용은 금지됩니다. 여기에는 △ 잠재의식 조작으로 심각한 피해를 초래하는 인공지능 시스템, △장애·연령·사회경제적 취약성을 악용하여 심각한 피해를 야기하는 인공지능 시스템, △ 인종·정치적 의견·노동조합 가입 여부·종교적 신념·성생활 또는 성적 지향을 유추하는 생체 인식분류 시스템, △ 사회신용시스템, △ 법집행기관이 공공장소에서 실시간 원격으로 얼굴이나 동작 등 생체 인식으로 감시하는 인공지능 시스템, △ 예측 치안, △ 무작위 수집을 통한 얼굴인식 데이터베이스 생성, △ 직장·교육 기관에서 감정을 추론하는 인공지능 시스템의 제공과 배치가 포함됩니다.

건강, 안전, 기본권 등에 심각한 위험을 초래하는 시스템은 '고위험' 인공지능 시스템으로 분류됩니다. 고위험 인공지능 시스템은 크

게 두 가지 종류로 구분되는데, 하나는 유럽연합 제품 안전법 적용 제품에 사용 되는 인공지능 시스템이고 다른 하나는 인권과 안전에 고위험을 미치는 인공지능 시스템입니다. 안전에 고위험을 미치는 인공지능 시스템은 자율주행차나 의료기기 등에 사용되는 인공지능 시스템입니다. 권리에 고위험을 미치는 인공지능 시스템은 △ 원격 생체인식 식별, 민감 속성 생체인식 분류 또는 감정 인식, △ 도로, 수도, 가스 등 중요 인프라, △ 교육 분야에서 입학 및 학습 성과 평가, △ 채용 및 노동자 관리, △ 의료, 금융, 보험 등 필수 서비스의 적격성 평가, △ 법 집행, △ 이주, 망명 및 국경 통제 관리, △ 사법 행정 및 선거 등 민주적 절차 등에서 사용되는 인공지능 시스템이 포함됩니다.

고위험 인공지능 시스템은 시장에 출시하기 전에 엄격한 요구조건을 충족해야 합니다. △ 적절한 위험 평가 및 완화 시스템, △차별적인 출력 위험을 최소화하기 위한 양질의 훈련 데이터, △결과의 추적 가능성을 보장하기 위한 작동 로그, △법 준수 여부를 평가하는데 필요한 상세한 문서화, △고위험 인공지능 시스템의 배치자(deployer)에게 명확하고 적절한 정보 제공, △인간의 감독 조치, △견고성, 사이버보안, 정확성의 준수 등입니다.

또한 고위험 인공지능 시스템의 제공자, 수입 및 유통업자, 배치자 등은 각자의 역할에 따른 다양한 의무를 부여받습니다. 고위험 인공지능 시스템의 제공자는 사전 적합성 평가를 통과해야 유럽연합 시장에 출시할 수 있습니다.

그 외에는 '제한된 위험', '최소 위험'이 있는데, 제한된 위험을 가진 인공지능 시스템을 비롯하여 특정 인공지능 시스템은 투명성 의

무를 준수해야 합니다. △ 인공지능 시스템의 제공자는 사람이 자신과 상호작용하는 대상이 AI라는 것을 알 수 있도록 설계해야 하며, △ 생성형 인공지능 시스템의 제공자는 결과물이 AI 생성물임을 기계판독 가능한 방식으로 인식될 수 있도록 해야 합니다. 또한, △ 감정인식이나 생체인식 인공지능 시스템의 배치자는 자연인에게 그 사실을 알려야 하며, △딥페이크를 생성하는 인공지능 시스템의 배치자는 결과물이 AI에 의해서 생성되었다는 사실을 공개해야 합니다. 예술적 작품에 해당할 경우 해당 작품의 감상을 저해하지 않는 방식으로 알릴 수 있습니다.

한편, 유럽연합 AI법은 범용 AI(General Purpose AI)에 대해 특별히 규정하고 있습니다. 이 규정은 초안에는 없었지만 2022년 말 챗GPT가 등장하는 등 범용 AI, 또는 생성형 AI, 최첨단 AI라고 불리는 AI 모델 및 시스템의 급속한 발전이 이루어지면서 이에 대한 규율이 새롭게 추가되었습니다. 범용 AI 모델 제공자는 기술 문서를 작성해야 하고, 유럽연합 인공지능 사무국과 국가 당국의 요청시 이를 제공해야 합니다. 또한 저작권 보호와 관련한 유럽연합법을 준수하고 AI 모델 훈련에 사용된 데이터의 충분히 자세한 요약을 공개해야 합니다. 특히, 시스템적 위험을 가진 범용 AI 모델 제공자는 적대적 테스트 등 시스템적 위험의 식별 및 완화, 심각한 사건에 대해 인공지능 사무국과 국가 당국에 보고, 적절한 사이버 보안 조치 등을 수행해야 합니다.

유럽연합 AI법의 위험 기반 접근법은 제품 안전을 위한 여러 법률들의 규제 방식을 반영한 것입니다. 하지만 입법 과정에서 유럽의회와 시민사회는 AI법에 대하여 인권 기반 접근법을 수용할 것을 강하

게 요구하였습니다. 그 결과 최종적으로 입법된 AI법에서는 △ 금지된 인공지능 규정, △ 인권에 고위험을 미치는 인공지능 시스템을 배치하는 공공기관·금융기관에 대한 기본권영향평가, △ 침해에 대한 국가 진정과 구제 절차, △ 고위험 인공지능의 의사결정의 대상이 된 사람의 설명요구권에 대한 규정 등을 보완하였습니다. 이러한 기본권 보호 조항들은 시민사회나 유럽의회가 제안했던 내용의 일부만 수용하였지만, 그럼에도 인공지능 기술로부터 인권과 신뢰를 보호하기 위한 진전을 이루었다고 평가할 수 있습니다.

유럽연합 인공지능 규제의 후퇴

유럽연합 AI법은 2024년 8월 1일 발표되었지만 실제 시행은 단계적으로 이루어집니다. 2025년 2월 2일, 금지된 인공지능 시스템 및 AI 리터러시 관련 조항의 시행이 먼저 시작되었습니다. 2025년 8월 2일에는 범용 AI 모델과 거버넌스에 관련된 조항이 시행되기 시작했습니다. 대부분의 다른 조항들은 발효 후 2년의 유예기간을 거쳐 2026년 8월 2일부터 시행될 예정입니다.

그런데 범용 AI를 둘러싼 전 세계적인 시장 경쟁이 격화되고 미국의 트럼프 2기 정부가 유럽연합의 디지털 규제 법률들에 대한 규제 완화를 요구하면서, 유럽연합은 규제 완화를 요구하는 미국과 산업계 요구를 따르기 시작하였습니다. 유럽 집행위원회는 2025년 11월 19일, 유럽연합 디지털 간소화 규칙을 발표하였습니다. 이 규칙에는 데이터(개인정보), 사이버보안, 인공지능에 대한 규제를 완화하는 디지털 옴니버스(Digital Omnibus) 법안이 포함되어 있습니다.[60] 특히 디지털 옴니버스 법안은 인공지능 학습 데이터에 쓰일 수 있는 개인

정보에 대한 규제를 완화하고, AI법에서 예정되어 있었던 고위험 인공지능 규제의 시행 시점을 최대 16개월 유예하였습니다.

이러한 규제 완화 흐름에 대해 유럽의 정보인권 단체들은 강하게 저항하고 있습니다. 이들은 디지털 옴니버스 법안이 "유럽연합의 핵심 디지털 보호체계를 대대적으로 다시 개정·재협상하려는 시도이며, 이는 유럽연합의 인권 및 디지털 정책의 기초를 무너뜨릴 위험이 있다"고 비판하고 있습니다. 그러면서 유럽연합 지도부에 미국 트럼프 대통령과 빅테크의 압박에 맞서 유럽연합의 디지털 규칙을 수호할 것을 촉구하였습니다.

2. 미국

미국의 인공지능 법제

미국은 아직 연방 차원에서 인공지능을 포괄적으로 규율하는 법률을 가지고 있지 않습니다. 다만 기존의 분야별 법률과 규제기관의 권한이 인공지능에도 적용됩니다. 2023년 4월, 연방거래위원회(FTC), 법무부, 소비자금융보호국(CFPB), 동등교육기회위원회(EEOC) 등 4개 기관은 공동성명을 발표하여, 인공지능 시스템을 포함한 자동화 시스템이 시민권, 공정 경쟁, 소비자 보호, 기회 균등에 영향을 미치고 있으며, 현행법에는 인공지능에 대한 예외 조항이 없기 때문에 시민들을 보호하기 위해 기존 법을 적극적으로 집행할 것이라고 강조하였습니다.[61]

미국은 자율 규제에 대한 규범과 원칙을 발전시켜 왔습니다. 2023년 1월 국립표준기술연구소(NIST)에서 발표한 AI 위험관리 프레임워크 1.0(AI RMF 1.0)의 경우, 인공지능 사업자가 AI 제품, 서비스 및 시스템의 설계, 개발, 사용 및 평가에서 신뢰성을 제고할 수 있도록 만들어진 자율규범입니다.[62]

2023년 10월 30일, 바이든 정부는 〈안전하고 보안이 되며 믿을 수 있는 인공지능의 개발 및 이용에 대한 행정 명령〉을 발표하였습니다.[63] 이 AI 행정명령은 사기, 차별, 허위정보, 국가안보 위험 등 인공지능의 위험을 인정하고 이를 책임감 있게 사용하기 위하여 연방정부 차원의 대책을 마련하는 데 초점을 맞추었습니다. 연방정부의 대책 마련을 촉구하는 8대 지침 영역은 △ 안전과 보안, △ 혁신과 경쟁 촉진, △ 근로자 지원, △ 평등과 인권 보호, △ 소비자 보호, △ 개인정보보호, △ 연방정부의 AI 활용 증진, △ 해외에서 미국의 리더십 강화 분야이었습니다.

트럼프 정부의 인공지능 규제 완화

바이드 정부의 AI 행정명령은 트럼프 2기 정부가 들어서자마자 철회되었습니다. 2025년 1월 23일, 트럼프 대통령은 〈인공지능에서 미국 리더십을 위한 장벽 제거〉 행정명령에 서명함으로써, AI 행정명령에 따라 취해진 모든 정책, 지침, 규정, 명령에 대해 중단, 취소, 수정 검토 및 조치할 것을 지시했습니다. 새로운 행정명령은 시장중심, 규제완화, '미국 우선주의' 기반의 AI 리더십 강화 및 국가 안보를 강조하였습니다. 또한 "이념적 편향이나 조작된 사회적 의제로부터 자유로운 인공지능 시스템을 개발"하고 "미국이 인공지능 분야에서 세

계적 리더십을 유지하기 위해" 단호하게 행동해야 한다고 밝혔습니다.[64]

한편, 트럼프 대통령의 지시로 수립된 'AI 행동계획'은 미국의 경제적 번영, 국가 안보, 그리고 인류 발전을 보장하기 위한 로드맵을 강조하였습니다. 그 가운데 'AI 혁신 가속화 전략'은 규제 장벽 제거 등의 내용을 담고 있으며, '미국 AI 인프라 구축 전략'은 데이터센터·반도체·전력망 건설 가속화를 위한 인허가 간소화 등을 포함하고 있습니다. '국제 AI 외교 및 안보 주도 전략'에서는 중국 리스크 견제와 글로벌 AI 기준 선점 등의 내용을 다룹니다.

특히 연방 조달지침을 개편하여 "이데올로기 편향 없는" AI를 개발한 기업과만 계약하도록 하였습니다. 또한, 규제 완화를 따르지 않고 AI 개발 및 도입에 장벽이 되는 주 정부 차원의 규제는 연방 조달 자금 지원에서 제외하도록 하였습니다.

콜로라도주의 인공지능법

미국은 연방 차원의 포괄적인 인공지능법이 없지만, 주 정부 차원에서 인공지능을 규율하기 위한 다양한 입법 시도가 이루어지고 있습니다. 그 중 2026년 6월 30일 시행이 예정되어 있는 콜로라도 인공지능법(SB205)은 미국에서 최초로 마련된 포괄적 인공지능 규제 법률로 평가됩니다.

이 법은 콜로라도주 내 모든 '고위험' 인공지능 시스템의 개발자와 배포자를 적용 대상으로 합니다. 법률은 특히 자동화된 의사결정 시스템을 규율 대상으로 삼아, 시스템이 소비자에게 '중대한 결정'을

내리거나 그 결정에 실질적으로 기여하는 경우를 고위험으로 정의하였습니다. 여기서 중대한 결정이란 교육, 고용, 필수 정부 서비스, 의료, 주택, 보험, 법률 서비스 등에서 소비자에게 제공되는 서비스를 제공 또는 거부하거나, 비용 또는 조건을 설정함에 있어 소비자에게 실질적으로 법적 또는 중대한 영향을 미치는 행위를 의미합니다.

콜로라도 인공지능법은 개발자에게 위험 정보 제공의 의무를, 배포자에게는 소비자에 대한 고지 및 정보 제공 의무를 부여합니다. 배포자는 고위험 인공지능 시스템에 대한 영향평가를 수행해야 하고, 개발자는 이에 필요한 정보를 제공해야 합니다.

콜로라도 인공지능법은 현재 미국 주 단위 인공지능 규제의 핵심 모델로 자리매김하고 있으며, 코네티컷, 매사추세츠, 뉴멕시코, 뉴욕, 버지니아 등 여러 주 의회가 이를 참고해 고위험 인공지능 시스템의 편향성과 차별 문제를 통제하는 법안을 검토하고 있습니다.[65] 다만 이 법에 대한 연방정부와 사업자의 압력이 계속되고 있기 때문에 2026년 6월 30일에 예정대로 발효될 수 있을지 지켜볼 필요가 있습니다.

캘리포니아주의 인공지능 규제

캘리포니아주는 미국 첨단 기술 산업의 중심지이면서 소비자 개인정보 보호법(CPRA) 등 소비자 보호를 위한 법제를 미국 내에서 선도해 왔습니다. 아직 인공지능을 규율하기 위한 포괄적인 법률은 없지만, 분야별로 인공지능의 위험으로부터 소비자를 보호하기 위한 다양한 법률을 제정해오고 있습니다.[66]

우선 AI 학습데이터 투명성 강화법(AB 2013)은 2026년 1월 1일부터 캘리포니아 주민에게 제공되는 모든 생성형 인공지능 시스템 또는 서비스가 개발 과정에서 사용한 학습데이터의 출처, 데이터의 성격, 개인정보 포함 여부 등을 투명하게 공개하도록 의무화하였습니다. 이는 생성형 AI의 데이터 기반 위험에 대한 권리주체의 통제권을 확대하려는 조치입니다. 또한, 식별 가능한 개인의 성적 딥페이크 생성 자체를 불법화하고(SB 926), 소셜미디어 플랫폼에 피해자 신고 채널 개설 및 신속한 삭제 조치를 의무화(SB 981)함으로써, AI를 악용한 성적 이미지 조작에 대응하는 보호기준을 제시하였습니다.

AI 투명성법(SB 942)은 월간 이용자 또는 방문자가 100만 명 이상인 대규모 공개 인공지능 시스템 제공자를 대상으로, 인공지능 시스템이 콘텐츠를 생성하거나 수정한 경우 그 사실을 이용자에게 명확히 알리도록 의무를 부과하였습니다. 단순한 문구 표기뿐 아니라 워터마크, 메타데이터, 시각·청각적 신호 등 다양한 방식이 허용되며, 위반 시 하루 최대 5,000달러의 과태료가 부과됩니다. 이는 AI 생성 콘텐츠의 출처 인식 가능성을 높여 이용자의 혼란을 방지하고, 허위정보·조작 콘텐츠 확산을 억제하려는 시도로 볼 수 있습니다.

또한 AI 기술을 이용한 개인의 외모·목소리 등에 대한 디지털 복제물 문제에 대응하기 위해, 배우 등 개인의 디지털 복제물을 무단 사용하지 못하도록 규정하고(AB 2602) 계약 과정의 투명성을 강화하며, 그 사용권리를 사후 70년까지 상속인에게 부여하도록 법적으로 보호하였습니다(AB 1836).[67] 이와 함께 선거 관련 AI 딥페이크 표시·삭제 의무(AB 2655), 소비자 프라이버시법 개정(AB 1008), 의료·교육 분야의 AI 규제 강화 등도 병행하면서 소비자 보호 측면에서

AI가 오용되는 것을 방지하는 체계가 확장되었습니다.

다만 캘리포니아 주의회가 유럽연합의 '범용 AI 모델'에 해당하는 '프론티어 AI 모델'을 규제하는 프론티어 AI 안전법(SB 1047)을 통과시켰으나 주지사가 거부권을 행사하여 시행되지 못했습니다. SB 1047은 일정 규모 이상의 프론티어 AI 모델에 사전 위험평가, '킬 스위치', 연 1회 외부감사, 위험 사전신고 등의 안전조치를 요구하고 있었습니다.[68] 이후 캘리포니아주는 첨단 AI 투명성법(SB 53)을 새롭게 통과시켜, 2026년 1월 1일부터 대규모 생성형 AI 개발 기업에 대해 제품 안전성과 관련한 정기적 영향 평가 보고를 의무화했습니다.[69]

더불어, 캘리포니아 시민권위원회는 2025년 10월 1일부터 시행되는 고용 분야 AI 사용 규정을 확정하여, 고용주가 채용·평가 등에서 인공지능 시스템을 사용할 경우 투명성, 차별 방지, 설명 가능성 등을 갖추도록 요구하였습니다. 아울러 캘리포니아는 미국 최초로 AI 챗봇 규제법을 제정해 미성년자 보호와 인공지능 제공자의 안전 의무를 대폭 강화하였습니다.

3. 한국

2024년 12월 26일 한국의 인공지능기본법이 국회를 통과하였고 2026년 1월 22일 시행됩니다. 이 법은 인공지능과 관련한 사항을 포괄적으로 규정한 기본법으로서, 기존에 인공지능 진흥을 소관했던 지능정보화 기본법보다 인공지능에 더 구체화된 진흥과 규율에 관한 사항을 규정하였습니다. 이 법을 소관하는 과학기술정보통신부는

2025년 11월 12일 현재 시행령안을 입법예고하였고 하위법령안의 의견을 수렴하고 있습니다.[70]

한국의 인공지능기본법은 국내·외적으로 위험 기반 접근을 따르는 것으로 알려졌습니다. 이에 시민사회는 인공지능기본법이 인공지능의 위험을 실효적으로 규제해야 한다고 요구하였습니다. 하지만 인공지능기본법은 입법 과정에서 첨단기술 산업진흥을 소관하는 부처인 과학기술정보통신부의 요구와 인공지능 산업 진흥을 위하여 규제를 최소화해야 한다는 산업계의 요구를 주되게 수용하였습니다.

금지하는 인공지능이 없음

인공지능기본법의 가장 큰 문제는 금지 인공지능에 대해서 아무것도 규정하고 있지 않다는 점입니다. 또한 국방과 국가안보 목적으로만 개발·이용되는 인공지능에 대해서는 법 전체의 적용을 배제하였습니다.

따라서 유럽연합 AI법과 달리 장애·연령·사회경제적 취약성을 악용하는 인공지능, 인종·정치적 의견·노동조합 가입 여부·종교적 신념·성생활 또는 성적 지향을 유추하는 생체 인식분류, 범죄수사와 무관한 경찰의 실시간 공공장소 생체 인식, 예측 치안, 직장·교육 기관에서 감정을 추론하는 인공지능 시스템을 개발하고 사용하는 것을 전혀 금지하고 있지 않습니다. 이와 같은 인공지능들은 인권에 허용할 수 없는 위험을 야기하지만 인공지능기본법은 이를 고영향 인공지능으로 지정하여 규제하고 있지도 않습니다.

또한 국방 또는 국가안보 목적의 인공지능에 대한 적용 배제는 이

중용도(dual use)로 사용될 수 있는 인공지능의 특성상 광범위한 의무 면제로 이어질 수 있습니다. 적용을 배제하는 인공지능에 대해서는 국방부장관, 국가정보원장, 경찰청장이 자체적으로 판단할 것이 아니라 최소한 국가인공지능위원회에서 공개적으로 심의하여야 합니다.

고영향 인공지능의 대상과 책무가 부족함

인공지능기본법은 유럽연합 AI법이나 미국 콜로라도 인공지능법의 '고위험'과 유사한 '고영향' 인공지능을 정의하고 인공지능을 개발하거나 이용하는 사업자에게 몇 가지 책무를 부과하였습니다.

인공지능기본법은 고영향 인공지능에 대하여 "사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능 시스템"이라고 정의했습니다. 고영향 인공지능 시스템의 구체적인 영역으로 열거된 분야로는 ① 에너지의 공급, ② 먹는 물의 생산 공정, ③ 보건의료의 제공 및 이용체계의 구축·운영, ④ 의료기기 및 디지털의료기기의 개발 및 이용, ⑤ 핵물질과 원자력시설의 안전한 관리 및 운영, ⑥ 범죄 수사나 체포 업무를 위한 생체 인식 정보(얼굴·지문·홍채 및 손바닥 정맥 등 개인을 식별할 수 있는 신체적·생리적·행동적 특징에 관한 개인정보를 말한다)의 분석·활용, ⑦ 채용, 대출 심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가, ⑧ 교통수단, 교통시설, 교통체계의 주요한 작동 및 운영, ⑨ 공공서비스 제공에 필요한 자격 확인 및 결정 또는 비용징수 등 국민에게 영향을 미치는 국가기관 등의 의사결정, ⑩ 유아교육·초등교육 및 중등교육에서의 학생 평가입니다. 그 밖의 고영향 인

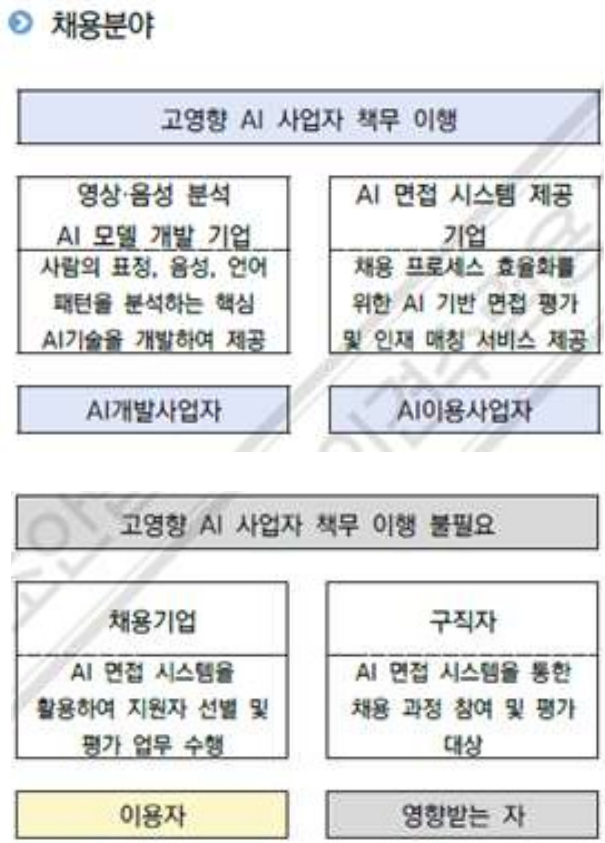
공지능은 사람의 생명·신체의 안전 및 기본권 보호에 중대한 영향을 미치는 영역으로서 대통령령으로 정하도록 하였습니다.

하지만 “개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가”가 무엇인지 법률에서는 더 자세하게 규정하지 않았습니다. 과학기술정보통신부가 입법예고한 시행령안도 추가적인 고영향 인공지능 영역을 아무 것도 규정하지 않았습니다. 이를테면 유럽연합 AI 법에서는 고위험에 해당하는 생체 인식, 거짓말탐지기를 포함한 감정인식, 학교와 직장의 감시 시스템, 출입국 감시 시스템, 재판이나 선거 절차에 쓰이는 인공지능, 인공지능 프로파일링이 한국 인공지능기본법에서 고영향에 해당하는지 예측할 수 없는 상태입니다.

인공지능기본법은 고영향 인공지능을 시장에 제공하거나 이용하는 사업자에 대한 몇 가지 책무를 규정하였습니다. 고영향 인공지능 사업자는 ① 위험관리방안의 수립·운영, ② 기술적으로 가능한 범위에서의 인공지능이 도출한 최종결과, 인공지능의 최종결과 도출에 활용된 주요 기준, 인공지능의 개발·활용에 사용된 학습용데이터의 개요 등에 대한 설명 방안의 수립·시행, ③ 이용자 보호 방안의 수립·운영, ④ 고영향 인공지능에 대한 사람의 관리·감독, ⑤ 안전성·신뢰성 확보를 위한 조치의 내용을 확인할 수 있는 문서의 작성과 보관, ⑥ 그 밖에 고영향 인공지능의 안전성·신뢰성 확보를 위하여 국가인공지능위원회에서 심의·의결된 사항 등을 조치할 책무가 있습니다.

하지만 업무에 고영향 인공지능을 배치하는 사업자라 하더라도 최종 ‘이용자’로서 인공지능 제품과 서비스를 단순 이용하는 경우에는 고영향 인공지능 사업자의 책무를 적용받지 않습니다. 이로 인하여 인공지능도구를 단순 이용하는 병원, 금융기관, 채용기업의 경우 고

영향 인공지능사업자에 해당하지 않는다는 것이 과학기술정보통신부의 해석으로 하위법령집에 명시되었습니다.



* 출처: 과학기술정보통신부, 한국지능정보사회진흥원. (2025. 9. 17).

이 점은 유럽연합이나 미국 콜로라도주 등 인공지능법을 제정한 지역들에서 고위험 인공지능을 업무에 배치하여 사용하는 모든 사업자를 '배치자(deployer)'로 정의하고 설명, 인간의 관리·감독, 문서화, 영향 평가 등 일정한 책무를 부과하고 있는 것에 비하여 매우 규제를 완화한 것입니다.

영향 받는 사람의 보호와 구제를 강화하여야 함

인공지능기본법에서 주목할 부분은 '영향받는 자'에 대한 정의를 두고 있다는 점입니다. 즉, 인공지능에 의하여 "생명, 신체의 안전 및 기본권에 중대한 영향을 받는 자"는 "인공지능의 최종결과 도출에 활용된 주요 기준 및 원리 등에 대하여 기술적·합리적으로 가능한 범위에서 명확하고 의미 있는 설명을 제공받을 수 있"는 권리가 있다고 명시하였습니다. 따라서 사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치는 인공지능, 즉 고영향 인공지능을 개발하거나 이용하는 사업자는 원칙적으로 설명 제공 등 영향 받는 사람의 권리를 보장하는 체계를 갖추어야 할 것입니다.

> 인공지능기본법
>
> 제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.
>
> 9. "영향 받는 자"란 인공지능제품 또는 인공지능서비스에 의하여 자신의 생명, 신체의 안전 및 기본권에 중대한 영향을 받는 자를 말한다.

> 제3조(기본원칙 및 국가 등의 책무)
>
> ② 영향 받는 자는 인공지능의 최종결과 도출에 활용된 주요 기준 및 원리 등에 대하여 기술적·합리적으로 가능한 범위에서 명확하고 의미 있는 설명을 제공받을 수 있어야 한다.

하지만 이 법률은 영향 받는 사람이 어떻게 이 설명요구권을 행사할 수 있을지 더 이상 아무런 언급을 하지 않고 모호한 상태로 남겨두었고 시행령안에서도 이에 대해 아무런 규정을 하지 않았습니다. 법률이 어떤 권리를 원칙적으로 선언하였지만 이를 보호할 수 있는 요건, 내용, 절차에 대한 실체적 규정을 가지고 있지 않다면, 권리주체로서는 이 권리를 실질적으로 행사하기가 매우 어려울 것입니다. 게다가 인공지능기본법은 고영향 인공지능을 이용하는 공공기관이나 민간기업이라 하더라도 최종 '이용자'로서 인공지능 제품과 서비스를 단순 이용하는 경우에는 고영향 인공지능 사업자의 책무를 적용받지 않는다고 보고 있습니다. 이러한 해석에 따르면 인공지능기본법에 기반해서 영향 받는 사람들이 권리를 행사하는 것이 사실상 불가능에 가깝습니다.

장애인이 어느 민간기업의 채용 인공지능으로부터 부당한 영향을 받았을 때 설명을 요구할 수 있는 방법을 찾을 수 있을까요? 사회복지 인공지능으로부터 부당한 영향을 받은 수급자가 지방자치단체에 설명을 요구하였을 때 충분한 설명을 들을 수 있을까요? 병원 진단 인공지능으로부터 부당한 영향을 받은 여성이 인간이 충분히 재검토

해줄 것을 요구할 수 있을까요? 대출 심사 인공지능으로부터 부당한 영향을 받은 이주민이 금융기관을 상대로 이의를 제기할 수 있을까요? 한국의 인공지능기본법 하에서는 이들 사업자에 대하여 설명을 요구하거나, 인간의 관리·감독을 요구하거나, 이의를 제기하거나, 문서를 확보하는 일이 쉽지 않을 것으로 보인다는 점에서 매우 우려가 됩니다.

이러한 문제는 영향 받는 당사자 개인 차원의 문제로 그치는 것이 아닙니다. 장애인, 여성, 외국인 노동자 등에 대한 차별을 금지하는 법률들을 집행하는 기관들이나 국가인권위원회가 차별 시정 또는 구제 업무를 집행할 때 불투명한 인공지능 시스템에 대한 자료를 확보하는 일도 매우 어려워질 것입니다. 단순 이용자로 분류되는 병원, 금융기관, 채용기업 등의 사업자에는 아무런 책무가 적용되지 않기 때문입니다.

한편, 인공지능기본법은 고영향 인공지능에 대한 검·인증과 영향평가 제도도 규정하고 있습니다. 다만 고영향 인공지능 사업자의 의무는 사전에 검·인증을 받도록 '노력'하여야 하고, 영향평가의 경우 사전에 사람의 기본권에 미치는 영향을 평가하기 위하여 '노력'하여야 하는 데 그칩니다. 국가기관등의 경우에만 고영향 인공지능을 이용할 때 검·인증등을 받은 인공지능에 기반한 제품 또는 서비스를 우선적으로 고려하여야 하고, 영향평가를 실시한 제품 또는 서비스를 우선적으로 고려하여야 합니다.

하지만 영향평가의 경우 인공지능법이 입법된 다른 지역에서는 고위험 인공지능을 배포하는 사업자의 필수적인 의무사항이고 유럽연합은 고위험 인공지능을 시장에 제공하는 모든 사업자에게 검·인증

도 의무화하고 있습니다. 비록 우리나라 인공지능기본법에서 국가기관등이 고영향 인공지능에 대한 검·인증과 영향평가를 의무적으로 고려하도록 했지만, 고영향 제품과 서비스를 제공하는 일반 기업이 검·인증과 영향평가를 받지 않아도 된다는 사실은 해당 제품과 서비스의 영향을 받는 일반 시민들에게 큰 위험을 낳을 수 있습니다.

한편, 법률상으로 고영향의 정의나 영향평가의 대상에 '기본권'에 대한 영향을 고려대상으로 포함하였다는 점에서 고영향 인공지능을 제공하거나 이용하는 공공기관이나 민간기업은 '기본권'에 대하여 이해할 필요가 있습니다.

그러나 법률이나 시행령안에서 구체적인 고영향 목록에 영향 받는 사람의 기본권에 높은 위험을 미치는 영역을 충분히 포함하고 있지 않고, 기본권에 미치는 영향에 대한 평가조차 인권기구와의 협업 없이 과학기술정보통신부가 일방적으로 소관하고 있습니다. 인권기구나 영향 받는 사람이나 관련 단체의 참여를 배제하는 제도 하에서는 인공지능이 기본권에 미치는 위험이 소홀히 다루어질 것이 우려됩니다.

인공지능기본법은 미약하나마 피해자의 구제를 위해 접근할 수 있는 절차를 규정하고 있습니다. 인공지능의 영향을 받은 사람은 이 법을 위반한 사항에 대하여 소관부처인 과학기술정보통신부장관에게 신고 및 민원을 제기할 수 있고, 부처에는 사실을 조사하고 중지 또는 시정을 명할 수 있는 권한이 있습니다.

그런데 이 법의 신고 대상에는 영향 받는 자에 대한 설명 의무를 이행하지 않은 사업자에 대한 신고는 포함되어 있지 않습니다. 또한 사실조사와 시정명령은 부처가 할 수도 있고 안 할 수도 있는 재량사

항으로 규정되어 있고, 사업자가 부처의 시정 명령을 이행하지 않을 때에만 3천만원 이하의 행정적인 과태료를 부과할 뿐입니다. 따라서 이런 제재 조항들이 사업자가 이 법을 준수할 수 있도록 충분한 압력을 발휘할 수 있을지 의문입니다. 게다가 2025년 9월 과학기술정보통신부는 '계도기간'이라는 명목으로 과태료 부과를 상당기간 유예할 계획을 밝혔습니다. 이러한 상황에서는 고영향 인공지능 사업자라 하더라도 노력과 비용을 들여 책무를 준수하기 위한 준비를 갖추도록 유인할 수 있을 것인지 의문입니다.

현재로서는 인공지능의 불투명성과 복잡성을 극복하기 위한 기술적이고 제도적인 과제가 많이 남아 있는 것이 사실입니다. 국가인권위원회가 지적하였듯이, 인공지능으로 영향을 받는 당사자들이 인공지능의 도입, 운영, 결정에 대하여 참여의 기회를 보장받고 있거나 인권침해에 대하여 효과적인 권리 구제를 보장받고 있다고 보기 어려운 것이 우리 현실입니다.[71] 인공지능에 대한 인권 기반 접근으로 이런 문제들을 계속 해결해 가야 할 것입니다.

나오며

「유럽연합 AI법」은 위험 기반 접근을 취한 것으로 널리 알려져 있지만, 다른 한편으로 기업의 인권 책무를 강조하는 인권 기반 접근에 대한 요구사항 역시 입법 과정에서 반영하였습니다. 인권에 미치는 고위험을 금지하거나 강력한 주의 의무를 부과하고 있고, 기본권 영향평가와 구제에 대한 내용을 포함한 것입니다. 그러나 최근 국제적인 산업 경쟁이 치열해 지면서 인권에 대한 관심과 의무사항이 축소되고 있는 듯합니다.

한국의 인공지능기본법에 대한 입법 논의도 처음에는 채용 AI의 불투명성과 책무성 부족 문제를 고민하고 일자리를 비롯하여 사람과 사회에 미치는 영향을 살펴가며 시작되었습니다. 그러나 아쉽게도 법률과 시행령의 시행을 앞둔 지금 시점에는 'AI 3대 강국'을 목표로 한다는 정부와 산업계의 목소리만 크게 들리고 있습니다.

그럼에도 인공지능의 인권적 책임을 확보하고 인공지능의 위험으로부터 영향을 받는 사람을 보호하기 위한 노력을 멈출 수는 없습니다.

인권에 허용될 수 없는 위험을 미치는 인공지능을 금지하고, 일부 고위험 영역의 인공지능을 제공하거나 이용하는 사업자에게는 설명, 인간의 관리·감독, 문서화 등 주의 의무를 다하도록 요구해야 합니다.

특히 중요한 사회 영역에서 불투명한 인공지능이 부당하거나 편향된 데이터와 알고리즘에 기반하여 인권을 침해하거나 차별적인 결과를 낳는 것을 방지해야 합니다. 이를 위해서는 인공지능 인권 영향평가가 사전에 의미 있게 수행되어야 하며, 인권 침해나 차별이 일어났을 때에는 그 피해자를 구제할 수 있는 체계를 갖추어야 합니다. 또한 이러한 과정에는 영향을 받는 당사자들이 참여하여 의견을 제시할 수 있어야 합니다.

우리가 인공지능이 사람에게 가져다 줄 혜택에 대해 기대한다면, 인공지능이 사람에게 미치는 영향에 더욱 초점을 맞추어 살펴야 할 것입니다. 이것이야말로 인공지능에 대한 인권 기반 접근의 핵심적인 목표입니다. 인권 기반 접근범에 기반한 인공지능 시대가 되어야 평범한 사람들의 삶과 노동이 첨단 기술과 조화를 이루며 민주주의를 향해 갈 수 있을 것입니다.

주 석

1) The Guardian. (2017. 10. 24). Facebook translates 'good morning' into 'attack them', leading to arrest. 〈https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest〉.
2) 동아일보. (2020. 12. 30). "닮은 건 수염뿐인데.." 안면인식 AI 오류로 무고한 흑인男 체포.
3) ENNHRI. Key human rights challenges of AI. 〈https://ennhri.org/AI-resource/key-human-rights-challenges/〉.
4) FRA. (2019). Facial recognition technology: fundamental rights considerations in the context of law enforcement.
5) 중앙일보. (2018. 8. 12). 중국서 가장 무서운 말 신용불량자 … 자녀 대학 합격도 취소.
6) Unni Koroth. (2025. 6. 2). The Target Pregnancy Prediction: Analytics Power and Ethics Collide. Medium Blog. 〈https://blog.othor.AI/the-target-pregnancy-prediction-analytics-power-and-ethics-collide-3177cc7955f7〉.
7) 유엔문서 A/73/348. (2018. 8. 29). Promotion and protection of the right to freedom of opinion and expression..
8) 유엔문서 A/HRC/43/29. (2020. 3. 4). Report of the Secretary-General: Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights.
9) 유엔문서 A/HRC/59/32. (2025. 6. 16). Practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, including activities relating to artificial intelligence: Report of the Office of the United Nations High Commissioner for Human Rights.
10) 유엔문서 A/HRC/48/31. (2021. 9. 15). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights.
11) "'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate cert AIn personal aspects relating to a natural person, in particular to analyse or predict aspects

concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;". GDPR Art. 4(4).

12) Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01), pp.6-8.

13) 개인정보 보호위원회. (2021. 4. 29). 개인정보위, '이루다' 개발사 (주)스캐터랩에 과징금·과태료 등 제재 처분; 개인정보 보호위원회. (2021. 4. 28). 제2021-007-072호 심의·의결서.

14) 경향신문. (2021. 1. 11). 이루다 말 속에 실존 이름·주소 버젓이…이용자들 "법적 대응"; 디지털투데이. (2024. 1. 30). 챗GPT서 개인정보 유출했다?…피해자 발생.

15) Yoshua Bengio, et al. (2025). International AI Safety Report: The International Scientific Report on the Safety of Advanced AI Scientific Report. pp.139-143.

16) 중앙선데이. (2025. 8. 30). 가족애까지 해킹… AI 보이스피싱이 '신뢰의 위기' 불렀다.

17) 한겨레. (2024. 8. 27). "혹시 내 사진도?"…학교 덮친 딥페이크 범죄 공포; BBC. (2024. 9. 3). 한국 학교를 집어삼킨 '딥페이크 음란물' 사태를 들여다보다.

18) OECD. (2024b). Assessing potential future artificial intelligence risks, benefits and policy imperatives. OECD Artificial Intelligence Papers, No.27. p.24.

19) ILO. (2025). Navigating workers' data rights in the digital age: A historical, current, and future perspective on workers' data protection. ILO Working Paper 149.

20) 유엔문서 A/HRC/48/31.

21) Yoshua Bengio, et al. (2025). pp.139-143.

22) 동아일보. (2020. 12. 31). 흑인 얼굴 구별못한 AI… 안면인식 인종차별 논란.

23) H. Moraes, M. Reis. (2024). Privacy attacks on AI systems: A current concern for organizations. 〈https://iapp.org/news/a/privacy-attacks-on-AI-systems-a-current-concern-for-organizations〉.

24) R. Staab, M. Vero, M. Balunovic, M. Vechev. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models. 〈https://openreview.net/forum?id=kmn0BhQk7p〉.

25) OECD. (2024a). AI, Data Governance and Privacy: Synergies and Areas of International Co-operation. OECD Artificial Intelligence Papers, No.22. p.21.

26) 국가인권위원회 상임위원회. (2020. 4. 2). 「인공지능산업 진흥에 관한 법률안」에

대한 의견표명 결정.

27) 조선일보. (2018. 10. 11). "이력서에 '여성' 들어가면 감점"…아마존 AI 채용, 도입 취소.

28) Yoshua Bengio, et al. (2025). pp.92-99.

29) Norwegian Consumer Council. (2023). Ghost in the Machine: Addressing the consumer harms of generative AI. 〈https://www.forbrukerradet.no/side/new-report-generative-AI-threatens-consumer-rights/〉. pp.29-30.

30) 뉴시스. (2024. 8. 29). AI가 보는 CEO는 '백인남성', 사회복지사는 '여성'…"젠더편향 심각".

31) Luke Haliburton, Jan Leusmann, Robin Welsch, Sinksar Ghebremedhin, Petros Isaakidis, Albrecht Schmidt, Sven Mayer. (2025). Uncovering labeler bias in machine learning annotation tasks. AI and Ethics. 5:2515-2528. 〈https://link.springer.com/article/10.1007/s43681-024-00572-w?utm_source=chatgpt.com〉.

32) Declan Humphreys. (2025). AI's Epistemic Harm: Reinforcement Learning, Collective Bias, and the New AI Culture War. Philosophy & Technology. 38:102. 〈https://link.springer.com/article/10.1007/s13347-025-00928-y?utm_source=chatgpt.com〉.

33) Denis Newman-Griffis, Jessica Sage Rauchberg, Rahaf Alharbi, Louise Hickman, Harry Hochheiser. (2022). Definition drives design: Disability models and mechanisms of bias in AI technologies. arXiv:2206.08287. 〈https://doi.org/10.48550/arXiv.2206.08287〉.

34) TUC Cymru. (2025. 5. 22). AI Inequalities: Disabilities. 〈https://www.tuc.org.uk/blogs/wales/AI-inequalities-disabilities?utm_source=chatgpt.com〉.

35) 동아일보. (2020. 12. 30).

36) European Commission (2020). White Paper On Artificial Intelligence - A European approach to excellence and trust. p.11.

37) Scientific American. (2019. 10. 24). Racial Bias Found in a Major Health Care Risk Algorithm. 〈https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/〉.

38) ProPublica. (2016. 5. 23). Machine Bias. 〈https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing〉; 오요한, 홍성욱. (2018). 인공지능 알고리즘은 사람을 차별하는가?.

과학기술학연구 제18권 제3호. pp.153-215; 이슬아. (2023). 인공지능 판사 앞의 7가지 숙제 -재범위험성 예측 알고리즘을 둘러싼 과학기술적·법적 논의 분석. 사법, 1(64), pp.665-714.

39) Medium. (2023. 11. 7). How to avoid the COMPAS problem in healthcare.

40) 캐시 오닐. (2017). 대량살상 수학무기. 흐름출판. 제5장.

41) 유엔문서 A/74/493. (2019. 10. 11). Report of the Special Rapporteur on extreme poverty and human rights.

42) 유엔문서 A/HRC/43/29.

43) 국가인권위원회 상임위원회. (2020. 4. 2); 국가인권위원회 전원위원회. (2022. 6. 13). 국가별 인권상황 정기검토(UPR) 관련 인권위 독립보고서(안)의 건 결정; 국가인권위원회 상임위원회. (2023. 7. 13). 「인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률안」에 대한 의견표명.

44) 보안뉴스. (2023. 8. 8). 인공지능의 편향성 문제, 얼마나 심각하고 어떻게 해결하나?

45) 유엔문서 A/73/348.

46) ILO. (2025).

47) Courthouse News Service. (2017. 5. 8). Houston Schools Must Face Teacher Evaluation Lawsuit. <https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/>.

48) 헌법재판소 1992. 12. 24. 선고 92헌가8 결정.

49) BSR. (2025). "Fundamentals of a Human Rights-Based Approach to Generative AI". Guide 1 of the Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI. p.11. <https://www.bsr.org/en/reports/human-rights-across-the-generative-AI-value-ch AIn>.

50) 유엔문서 A/HRC/43/29.

51) 유엔문서 A/HRC/48/31.

52) 국가인권위원회. (2022. 10. 21). <인공지능 개발과 활용에 관한 인권 가이드라인> 권고, 국무총리 및 관련 부처 장관·기관장 수용.

53) BSR. (2025).

54) 국가인권위원회. (2024. 7. 8). 인공지능 개발과 활용에 있어서 인권 보호를 위해 인권 영향 평가 실시 필요.

55) 유엔문서 A/HRC/59/32.

56) OECD. (2024b).

57) 동아일보. (2020. 12. 30).

58) Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

59) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

60) European Commission. (2025.11.19). [Press Release] Simpler 유럽연합 digital rules and new digital wallets to save billions for businesses and boost innovation,

61) FTC. (2023. 4. 25). FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI.

62) NIST. AI Risk Management Framework. 〈https://www.nist.gov/itl/ai-risk-management-framework〉.

63) Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

64) REMOVING BARRIERS TO AMERICAN LEADERSHIP IN ARTIFICIAL INTELLIGENCE,

65) 한국지능정보사회진흥원. (2024. 8. 19). [법제Brief(24-12)] 美 콜로라도 주, 인공지능법의 주요 내용 및 시사점.

66) 세계법제정보센터 법제동향. (2024. 10. 16). 미국 캘리포니아주, 인공지능 관련 18개 법률 제정.

67) The Verge. (2024. 9. 18). California governor signs rules limiting AI actor clones.

68) ZDNet Korea. (2025. 6. 18). "무산된 SB 1047, 부활하나"…캘리포니아, AI 외부 규제안 다시 꺼냈다.

69) Hunton (2025. 10. 9). California Governor Newsom Signs Groundbreaking AI Legislation into Law. 〈https://www.hunton.com/privacy-and-information-security-law/california-governor-newsom-signs-groundbreaking-ai-legislation-into-law〉.

70) 과학기술정보통신부, 한국지능정보사회진흥원. (2025. 9. 17). 인공지능기본법 하위법령 및 가이드라인(안) 대국민 의견수렴(080-137-1300). 〈https://nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99835&bcIdx=28600&parentSeq=28600〉.

71) “인공지능으로 영향을 받는 당사자들은 인공지능의 도입, 운영, 결정에 대하여 참여의 기회를 보장받고 있지 못하며, 인공지능으로 인한 인권침해가 발생한 경우에도 적절하고 효과적인 권리구제를 받을 수 있는 절차와 방법이 미흡한 상황입니다.” 국가인권위원회. 2022. 4. 11. 결정. 4문.

2025년 11월
정보인권연구소 이슈보고서

인공지능에 대한 인권기반접근

정보인권연구소
Institute for Digital Rights

HEINRICH BÖLL STIFTUNG
서울
동아시아 | 국제 대화