연구보고서(수시) 2024-05

사회보장 행정에서 인공지능 적용 동향과 함의

김기태
신영규·김명주·김은하·변소연

사람을 생각하는 사람들

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

▌연구진

연구책임자	**김기태**	한국보건사회연구원 연구위원
공동연구진	**신영규**	한국보건사회연구원 부연구위원
	김명주	서울여자대학교 교수, AI 안전연구소 소장
	김은하	한국사회보장정보원 연구센터장
	변소연	한국보건사회연구원 연구원

연구보고서(수시) 2024-05

사회보장 행정에서 인공지능 적용 동향과 함의

발 행 일 2024년 12월
발 행 인 강 혜 규
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1999년 4월 27일(제2015-000007호)
인 쇄 처 (사)아름다운사람들

ISBN 979-11-7252-077-9 [93330]
https://doi.org/10.23060/kihasa.b.2024.05

발|간|사

인공지능의 발전 속도는 눈부시다. 기술 발전의 속도를 따라잡는 데 제도는 숨이 차다. 규제의 속도는 더욱 느리다. 기술과 규제의 속도 차이가 남기는 틈을 놓쳐서는 안 된다. 인공지능이 불러오는 편익에 대한 찬사와 더불어, 위험성에 대한 경고에도 귀 기울여야 한다. 2024년 노벨 물리학상 수상자인 제프리 힌턴(Geoffrey Hinton) 토론토대학 교수는 2024년 12월 영국 BBC와의 인터뷰에서 인공지능으로 인해서 30년 안에 인류가 멸종할 가능성이 있다고까지 언급했다.

인공지능 기술이 가장 많이 활용되는 영역 가운데 하나가 사회보장이다. 이 보고서의 본문에서도 확인할 수 있듯이, 한국에서도 위기가구 발굴이나 일자리 소개와 관련해 인공지능 기술이 이미 활용되고 있다. 이를 위한 데이터의 결합도 점진적으로 이뤄지고 있다. 인공지능이 가진 거대한 잠재력을 활용하면서 위험성을 통제하기 위해서 규제도 느리게나마 따라붙고 있다. 한국에서도 오랜 지연 끝에 2024년 12월에 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법'이 마침내 국회 본회의를 통과했다. 이 법이 제정되는 과정에서 사회보장 영역에 대한 논의가 충분히 이뤄졌다고 보기 힘들다. 한국은 인공지능 기술 발전에 있어서는 세계 6위 수준의 기술 강국이다. 그렇지만 인공지능 관련 규제, 특히 사회보장 영역에서의 규제에 대해서는 논의의 진행 속도가 매우 더디다.

해외의 규제 사례를 보면, 한국과 대조된다. 현재로서는 인공지능에 관한 유일한 포괄적인 규제인 유럽연합의 인공지능법에서 사회보장과 관련이 있는 사회적 평점(social scoring)을 수용할 수 없는 위험성(unacceptable risk)으로 분류하고 있다. 또 주요한 고위험 영역 가운데 하나로 '필수 민간 및 공공 서비스 분야'를 제시했다. 인공지능에 대한

국가 규제의 한가운데 사회보장 영역이 있다. 이 보고서에서 살펴본 미국의 행정명령에서도 마찬가지다. 해당 행정명령의 주요한 내용은 미국 보건복지부에 전달하는 지침이다.

이 보고서는 국내·외 사회보장 영역에서 인공지능 기술의 적용 현황을 짚고, 국내·외 규제의 동향을 분석한다. 그리고 그에 근거한 정책 방향을 제시하고자 했다. 이 보고서는 2024년 하반기 6개월의 기간에 속성으로 진행된 수시 연구 과제의 결과물이다. 짧은 시간이었다. 연구진들은 그래도 국내외의 기술 적용 및 규제 동향 및 정책 함의를 최대한 갈무리하려고 노력했다.

한국에서 사회보장 영역에 인공지능을 활용하는 것과 관련한 논의는 매우 더디게 진행되고 있다. 정책 영역에서의 대응도 아직은 가시화하지 않고 있다. 이 보고서가 관련 논의와 대응을 촉발하는 불쏘시개 역할을 할 것으로 기대한다.

2024년 12월

한국보건사회연구원장 직무대행

강 혜 규

목 차

표 목차

그림 목차

요약

인공지능의 급속한 발전이 초래할 변화에 대한 전망은 엇갈린다. 기술 발전이 인류의 존립을 위협하는 수준이라는 비관적 전망과 장밋빛 전망이 함께 나오고 있다. 사회보장은 인공지능의 발전과 적용이 가장 활발하게 관찰되는 분야 가운데 하나다. 인공지능은 사회정책 영역에서 효율성, 과학성, 중립성의 증진에 기여할 것이란 기대가 크다(Zaber, Casu, Brodersohn, 2024). 반면, 새로운 기술이 현존하는 빈곤, 불평등, 차별을 유지 및 강화한다는 비판도 있다. 사회보장 영역에서 인공지능이 초래할 영향 및 정책적 대응에 관한 국내 연구는 희소하다. 한국은 사회보장 영역에서 인공지능 기술을 디지털 인프라 위에서 빠른 속도로 적용시키고 있는 한편, 인공지능 기술 적용이 초래할 법적·윤리적 부작용 문제에 대해서는 상대적으로 방임적인 태도를 보였다. 이 연구는 한국과 다른 복지국가들이 사회보장 영역에서 행하고 있는 인공지능 기술 적용 현황 및 관련 규제의 동향을 살펴보고, 이를 통해서 인공지능 기술을 사회보장 영역에 적용하는 과정에서 짚어야 할 정책적, 제도적 함의를 모색하고자 한다.

제2장에서는 인공지능 발전의 동향과 쟁점을 짚었다. 인공지능의 역사는 1950년대 컴퓨터의 태동과 함께 시작되었다. '계산 기계와 지능'이라는 논문을 앨런 튜링이 1950년에 발표하면서 튜링 테스트를 제안하였는데 이것이 인공지능 개념의 기초가 됐다. '인공지능'이라는 용어는 1956년 다트머스 회의에서 처음으로 생겨났으며 이를 계기로 인공지능에 관한 연구가 본격적으로 시작되었다. 70년에 걸친 긴 인공지능의 역사를 보면, 혁신적인 기술 발전과 예상치 못한 거품 논쟁을 반복했다. 인공지능은 과거에 두 번의 겨울기와 세 번의 여름기를 지났다고 볼 수 있다. 현재는 인공지능의 역사에서 세 번째 여름을 맞이하는 셈이다. 지금은 인공지능의

세 번째 여름에 속하는데, 주로 인공신경망 기반의 딥러닝을 중심으로 한 인공지능이 신속하면서도 강력하게 확산되는 중이다. 아울러 거대한 학습 데이터인 빅데이터와 클라우드 컴퓨팅 같은 강력한 컴퓨팅 파워를 기반으로 하여, 자연어 처리(NLP)와 자연어 생성(NLG), 컴퓨터 비전 등을 중심으로 매우 다양하게 혁신적인 변화를 이루어내고 있다. 특히 생성형 인공지능의 발전은 창작, 교육, 업무 효율성 등 다양한 분야에서 혁신을 가져오고 있다. 앞으로도 인공지능 기술의 발전과 그 영향력은 계속해서 확대될 것으로 예상된다.

2024년을 기준으로 인공지능 기술의 최신 동향과 미래 예측을 종합해 보면 다음과 같다. 첫째, AI Agent 개발이 활발히 이루어지면서 자율적으로 작업을 수행하는 지능형 소프트웨어의 등장이 주목받고 있다. 둘째, 인공지능 비서 등 개인화 서비스도 빠르게 발전하고 있다. 셋째, 각 산업 분야별로 특화된 인공지능 애플리케이션 개발도 가속화하고 있다. 넷째, 거대 언어 모델(LLM)과 소규모 언어 모델(sLLM) 사이의 합리적 활용 방안도 주목받고 있다. 다섯째, 멀티모달 인공지능 기술의 경쟁도 본격화되고 있다. 여섯째, 국가 및 지역 단위의 인공지능 주권 확보 노력인 "소버린 AI(Sovereign AI)"도 주목받고 있다. 일곱째, 양자 컴퓨팅과 인공지능의 융합 연구도 활발히 진행되고 있다. 여덟째, 인공지능 윤리와 규제 프레임워크 구축도 중요한 과제로 대두되고 있다.

인공지능의 미래에 대해서는 '부머(Boomer)'와 '두머(Doomer)'의 입장이 갈린다. 부머는 인공지능이 인류에게 줄 수 있는 이익에 집중한다. 반면, 두머는 인공일반지능이 인간의 뇌와 비슷한 지능을 갖게 될 것을 우려한다.

인공지능의 파괴력을 고려하면, 인공지능 관련 윤리적 기반을 다지는 것이 필수적이다. 이와 관련한 윤리적 원칙을 나열하면 다음과 같다. 첫째,

공공성이다. 기술이 소수의 이익이 아닌 인류 전체의 번영에 기여해야 한다는 것이다. 둘째, 공정성이다. 인공지능을 개발할 때, 데이터의 편향성을 고려하고 공정성을 보장하는 것이 필수적이다. 셋째, 책무성과 책임성이다. 신기술을 활용하는 주체가 그로 인해 발생하는 기회나 문제에 대해 책임져야 한다는 원칙이다. 넷째, 안전성과 보안성이다. 안전성은 인공지능이 정상적으로 작동하면서도 예상치 못한 상황에서 사람, 재산, 환경에 피해를 주지 않도록 보장하는 것이고, 보안성은 외부의 악의적인 공격으로부터 인공지능 시스템을 보호하는 것을 의미한다. 여기에 더해서, 인공지능의 차별화된 특성에 근거한 윤리 원칙도 있다. 첫째가 통제 가능성이다. 인공지능 특유의 자율성에 대응하기 위해서는 인간에 의한 통제 가능성(controllability)이 필요하다. 둘째, 투명성과 설명 가능성이다. 인공지능의 또 다른 차별화된 특성인 지능성(intelligence)에 대응하기 위해서는 인공지능이 어떤 기준과 원칙으로 작동하는지, 인간이 그 내부 구조를 이해할 수 있어야 한다. 셋째 개인정보 및 사생활 보호다. 인공지능이 가지는 세 번째 차별화된 특성인 학습성(learningability)을 고려한 원칙이다. 인공지능이 데이터를 학습하거나 사용하는 과정에서 그 안에 포함된 개인정보가 유출될 위험이 크기 때문이다.

제3장에서는 국내, 국외로 나누어서 사회보장 영역에서 인공지능이 적용되는 실태를 살펴봤다. 제1절에서는 국내 현황을 살펴보고자 조달정보개방포털과 나라장터에서 최근 5개년 동안(2019년 10월~2024년 9월) 확인되는 사회보장 분야의 복지 기술 활용 현황을 조사하였다. 조사 결과, 몇 가지 특징이 확인되었다. 첫째, 사회보장 분야의 복지 기술은 노인에 대한 돌봄 수요에 다수 활용되고 있었다. 둘째, 복지 기술은 업무 담당자의 효율적인 업무 처리를 지원하기 위해 활용되고 있었다. 셋째, 빅데이터의 수집과 축적을 기반으로 하는 정보 제공 서비스가 다수 발견되었다.

넷째, 개인에게 특화된 맞춤형 서비스가 다수 차지하였다. 마지막으로, 정보 활용의 효율성과 효과성을 높이기 위한 데이터베이스 시스템의 활용 양상이 확인되었다. 이상의 양상이 향후 인공지능의 핵심 기술이 점하게 될 영역과 역할이라고 해석해도 큰 무리가 없을 것으로 보인다.

복지 기술을 활용한 의사결정 지원은 업무 담당자의 행정적 판단을 지원하거나, 서비스 중계와 관련된 영역으로, 인공지능 기반 빅데이터를 활용하고 있다. 이와 관련된 대표적인 세 가지 사례도 소개했다.

첫째, AI 기반 일자리 매칭 서비스이다. AI 기반 일자리 매칭 서비스는 구직자의 이력서 정보와 구인 기업의 구인 정보를 활용하여 인공지능 알고리즘에 기반한 추천 서비스를 제공하고 기업과 구직자 간 일자리 미스매치를 해소하는 데 목적을 두고 있다. 2020년 7월부터 추진되고 있다. 구직자에게는 구직 어려움을 감소시키고 구인 기업에게는 적합한 직무역량을 지닌 근로자를 채용할 수 있게 한다. 상담원은 상담역량이 강화되고 보다 나은 진로지도 서비스를 제공하게 된다. 이 시스템은 2022년에 감사원으로부터 지적을 받았다. 인공지능 모형을 활용했지만 시스템의 예측률이 낮은 점, 구인 정보가 누락된 채 분석이 이루어지거나 임금체불 상태인 사업주의 정보 등 적절하지 않은 정보를 제공하고 있는 점, 미등록 체류 외국인에 대해 취업 알선을 하는 점 등이 주된 내용이었다. 추천 시스템에서 작동하는 알고리즘의 이와 같은 문제가 지적되는 가운데, AI 기반 일자리 매칭 서비스는 외연을 확장해 나가는 중이다. 2024년 5월 과학기술정보통신부의 「2024년 부처협업 기반 AI 확산 사업」에 선정되어 더욱 다양하고 확장된 데이터를 기반으로 구직자 취업역량 향상을 지원할 수 있게 되었다. 또 일자리·인재 추천 서비스를 고도화하여 매칭 서비스를 강화할 수 있는 기회를 얻게 되었다. 과기정통부의 사업을 통해 AI 인재 추천에서부터 AI 직업훈련 추천에 이르는 시범 과제를 수행하는

과정에서 3~6개의 인공지능 기술이 활용될 예정이다.

둘째, 복지 사각지대 발굴관리시스템이다. 복지 사각지대 발굴시스템의 핵심 기능은 각기 다른 기관에 흩어져 있는 공공데이터를 기반으로 고위험 확률 모델을 활용하여 지원이 필요한 복지 사각지대 대상자를 도출하는 것이다. 시스템 운영 과정은 자료의 수집과 빅데이터 분석, 분석 결과를 바탕으로 한 고위험 대상자 도출, 도출된 대상자 명단을 지자체에 제공, 지자체 담당자가 대상자에 대해 상담 및 지원하는 순으로 이루어진다. 복지 사각지대 발굴시스템은 분석 변수가 확장되면서 빅데이터 분석에 활용되는 기술과 모델의 발전을 꾀하고 있다. 새로운 빅데이터 분석 모형이 도입되며 모델이 분화되어 복잡해지는 등, 알고리즘의 형태가 고도화되고 있다. 이러한 발전을 반영하듯이, 시스템 구축 이후 약 9년간 665.6만 명에 대한 조사가 지자체를 통해 이루어졌으며 이 중 290.2만 명(43.6%)에게 공공·민간 복지서비스를 지원한 것으로 알려졌다. 지원 대상은 시스템 구축 초기인 2015년 1.8만 명에서 2023년에 68.6만 명으로 늘었고, 지원율은 2015년 16%에서 2023년 49.4%로 증가하였다.

복지 사각지대 발굴시스템은 시스템 구축 및 개선 과정에서 법 제정이나 전달체계의 변화 등이 함께 맞물려 진행된다는 점에서 독특성이 있다. 정부가 '복지 사각지대 발굴 및 지원 종합대책'을 몇 차례 수립하는 과정에서 복지 사각지대 발굴시스템도 함께 개선되었으며, 이에 따라 현장의 업무도 달라지게 되었다. 이는 복지 사각지대 발굴시스템이 현장의 전달체계와 긴밀하게 관련을 맺고 있음을 의미한다. 한편으로 복지 사각지대 발굴시스템은 알고리즘이 제공하는 발굴 대상자의 정확성 문제, 발굴 대상자에 대해 지원이 이루어지지 못하는 문제, 개인정보 유출 문제가 지속해서 제기되고 있다. 또한 복지 사각지대 발굴 과정에서 정보시스템에 의존하기보다 현장 중심의 발굴 노력이 중요하다는 지적도 있다.

셋째, AI 활용 초기상담시스템이다. AI 활용 초기상담시스템은 2024년 7월부터 시범사업을 진행하고 있으며, 11월 말부터는 전국으로 확대될 예정이다. AI 초기상담시스템의 도입은 복지 사각지대 발굴시스템 운영 이후 발생한 이슈와 관련된다. 복지 사각지대 발굴시스템의 핵심 기술인 알고리즘이 확률에 기반한 통계 모델을 바탕으로 하기 때문에, 알고리즘이 제시한 고위험군 대상 목록에 속하지는 않지만 실제로는 위기상태인 집단이 늘 상존할 수 있다는 점을 염두에 두어야 한다. 이러한 잠재적 위기 대상자까지 복지 업무 담당자가 확인하기에는 인력난으로 인한 한계가 있다. AI 활용 초기상담 정보시스템은 이러한 배경에서 기획되었으며, 콜 기반 대화 시스템을 활용하여 알고리즘으로 드러나지 않은 위기 대상자에 대해 초기상담을 진행한다. 즉 잠재적 위기 대상자들에게 전화를 걸어 수신자의 욕구를 파악하고 국가의 지원이 필요한지 확인한다. 초기상담 결과를 바탕으로 심층 상담이 필요하다고 판단되는 사례에 대해서 지자체 공무원이 개입하게 된다.

AI 활용 초기상담시스템을 구축하는 기간에 시범적으로 현장에 적용한 결과를 보면, 시간이 지남에 따라 대상자들의 수신율이 증가하는 것을 확인할 수 있다. 또 상담 거부 비중이 지속해서 감소하고, 추가 상담 요청 비율이 지속해서 증가하고 있다. 다만, 시범 적용 과정이고 평가의 절대 기준이 없기 때문에 현재 단계에서 수치를 긍정적으로 단언하기에는 조심스럽다.

AI 활용 초기상담 정보시스템에는 앞으로 딥러닝을 활용하여 자동으로 상담 시나리오를 생성하는 기술이 활용될 계획이다. 또 대상자 각자의 특성에 기반한 맞춤형 대화가 진행되며, 감성 기반 대화 모델도 구축될 계획이다. 현재는 사업 초기 단계로 서비스 효과에 대해서는 평가하기가 어렵다. AI에 기반한 음성 서비스가 단순 정보 제공이 아닌 복지 대상자들이

자신들의 어려움을 명확하게 전달할 수 있도록 이끌어낼 수 있는 단계까지 발전할 것인지, 그 결과가 주목된다.

제3장 제2절에서 국외의 사례를 보았다. 전 세계적으로 사회보장 영역에서 인공지능이 사용되기 시작한 시점은 2008년 이후지만 본격적인 확산기는 2017년 이후였다(Zaber, Casu, Brodersohn, 2024). 인공지능이 활용되는 사회보장 영역은 가족 급여, 보건, 산업재해, 연금, 실업, 보편적 급여 등이다. 2020년 이전에 사회보장 영역에서 활용된 인공지능 기술은 챗봇이었다. 챗봇은 2020년 이후에도 코로나 범유행 상황에서 폭증한 급여 신청을 기관들이 대응하는 과정에서 활용도가 더욱 높았다. 일부 기관들은 챗봇의 기능을 강화하는 과정에서 생성형 인공지능(generative artificial intelligence)을 활용하기 시작했다. 사회보장 영역에서 인공지능이 활용되는 범주는 다섯 가지다. 첫째, 서비스 제공, 둘째, 자동화 및 사례 관리, 셋째, 전향적이고 능동적인 사회보장, 넷째, 위험관리 및 예방(Risk management and prevention), 다섯째, 평등과 공정성(Equality and fairness)이다.

서비스 제공(Service delivery) 영역에서 가장 활발하게 사용되는 인공지능은 챗봇이다. 흥미롭게도, 브라질, 아르헨티나, 파나마, 우루과이 등 남미 국가에서 국민들의 급여 관련 문의나 민원을 응대하는 데 챗봇의 활용도가 높았다. 보건 영역에서는 인공지능이 응급실까지 들어왔다. 호주 뉴사우스웨일스주 공공 의료 시스템의 경우 응급실에서 패혈증 조기 발견을 목표로 하는 머신러닝 프로토타입 제품을 개발했다(Zaber, Casu, Brodersohn, 2024). 복지 급여 수급자를 포착하는 데에도 인공지능은 이미 활용되고 있다. 캐나다 고용사회개발부(Employment and Social Development Canada)는 저소득 노인을 위한 급여인 보장소득보조금(Guaranteed Income Support) 영역에 인공지능 기술을 적용했다.

사례 관리 영역에서도 인공지능의 역할이 확대되고 있다. 오스트리아의 사회보험연합은 청구 자동 처리를 지원하고 의사와 환자를 매칭하는 인공지능 기반 시스템을 구현했다. 그러나 인공지능이 그리는 사회보장의 미래가 장밋빛인 것만은 아니다. 네덜란드와 덴마크의 사례는 인공지능 적용에 따른 인권 침해, 정보 유출, 공공성 훼손에 관한 또 다른 도전을 보여준다.

제4장 제1절에서는 국내의 인공지능과 관련된 규제나 규율을 법적 강제성과 구속력의 정도에 따라 단계별로 살펴본다. 여기에는 법률과 명령같이 전 국민이 반드시 준수해야 하는 강제력을 지닌 규제 방안부터, 법적 준수를 돕거나 방향성을 제공하며 강제성은 없지만 준수하지 않았을 경우 제재나 평가상 불이익을 받을 수 있는 지침 및 가이드라인, 그리고 강제성 없는 권고사항으로 자율적 참여를 유도하는 선언까지 다양한 수준의 내용이 포함된다.

먼저, 법률을 보면, 2024년 12월 1일 기준 인공지능에 관한 국내의 법률은 존재하지 않는다.[1] 다만 제21대 국회에서 인공지능 법안이 2020년부터 발의되어 왔으며 이에 해당하는 총 9건은 국회 회기 종료에 따라 모두 폐기되었다. 제22대 국회에서는 '인공지능'이 제명에 명시된 법안이 총 11개로 확인되었다. 이러한 법안들은 인공지능 산업의 육성이나 인공지능의 윤리적 기준의 확립, 개인정보 보호와 신뢰성 강화 등을 목표로 하고 있으며, 일부는 심사 단계에 있다. 제21대와 비교할 때, 제22대 국회에서 발의된 법안은 고위험 영역 인공지능 규제와 관련된 내용이 상대적으로 다수 발견된다. 대부분 신뢰성과 안전성을 확보하고, 제품이나 서비스가 고위험 영역 인공지능을 활용했을 경우에, 그에 대해 이용자에게 알권리를 부여하는 내용을 포함한다. 제21대 국회에서 발의된 법안에서는

1) 이 보고서 작성 마무리 단계인 2024년 12월 26일 인공지능 기본법이 국회 본회의를 통과했다. 시간의 제약으로 이 보고서에는 해당 법 내용에 대한 분석을 담지 못했다.

발견되지 않았던 '생성형 인공지능'이 등장하기도 하는데, 생성형 인공지능을 운용한 사실을 고지하고 안전 확보를 위한 조치를 이행할 것을 규정하고 있다. 이훈기 의원 등이 발의한 「인공지능의 발전과 안전성 확보 등에 관한 법률안」의 '인공지능이 국민 기본권에 미치는 영향의 평가'나 한민수 의원 등이 발의한 「인공지능 기본법안」의 '해외 사업자의 인공지능 기본법상 의무 이행 확보'와 관련된 규제는 타 법안에서 찾아보기 어려운 내용이다. 단순 규제를 넘어서 사회에 미치는 영향을 객관적으로 평가하여 정책 수립의 근거로 마련하고자 하고, 해외 인공지능 사업자가 국내에 미칠 수 있는 부정적 영향을 차단하려는 등의 적극적인 조치를 담고 있다.

다음으로, 법률상의 주요 계획이다. 데이터 활용과 관련된 법률상 주요 계획을 통해 인공지능에 대한 정부의 규제 계획과 국가의 노력을 부분적으로 확인할 수 있다. 2024년에 적용되는 법률상 계획을 검토한 결과 "제1차(2023~2025년) 데이터산업 진흥 기본계획"에는 신뢰성에 기반한 데이터 활용 및 데이터 윤리 확산에 대한 내용이 명시되어 있음을 확인하였다. 또한 "개인정보 보호 기본계획(2024~2026)"에는 신뢰할 수 있는 신기술 이용 환경을 위해 ① 인공지능 시대에 대응한 규제 혁신 추진 ② 디지털 신기술 환경에서의 개인정보 보호 방안 마련, ③ 가명정보 안전 활용 지원 확대, ④ 안전한 데이터 활용을 위한 법·제도 기반 조성이라는 계획이 담겨 있다. 구체적으로는 학습용 데이터 및 생성형 AI의 활용, 생체인식 서비스, 클라우드 등 신기술 서비스 이용 등 인공지능 활용과 매우 밀접한 영역에서 안전성 확보나 개인정보 침해 방지 등을 위한 규제 계획을 제시하고 있다.

인공지능과 관련된 법률이 부재한 상황에서 인공지능의 규제는 데이터 관련 법률상의 계획 수립과 추진에 의존할 수밖에 없을 것이다. 다만, 이러한 법률상 계획에는 규제를 위한 준비단계로 연구나 R&D를 추진하는 내용도 보이므로 구체적인 제재나 가이드라인이 도출되기에는 다소 시간

이 필요할 것이다. 인공지능 법률에 기반하지 않은 규제이므로 총체적인 시각에서의 규제 방안을 기대하기도 어렵다.

다음으로, 지침 및 가이드라인이다. 인공지능 활용과 관련된 주요 가이드라인은 특정 분야를 지정하지 않고 일반적으로 적용되는 가이드라인과, 특정 분야나 대상에 한정한 가이드라인으로 나눌 수 있다.

전자로는 방송통신위원회·정보통신정책연구원(2019)의 '이용자 중심의 지능정보사회를 실현하기 위한 원칙'이 있다. 여기에서는 사람 중심 서비스 제공, 투명성과 설명 가능성, 책임성, 안전성, 차별 금지, 프라이버시 보호 등의 원칙을 제시한다. 과학기술정보통신부(2020)의 '인공지능(AI) 윤리 기준'은 '인간성(Humanity)을 위한 인공지능(AI)'의 3대 원칙·10대 요건을 담고 있다. 국가인권위원회의(2022) '인공지능 개발과 활용에 관한 인권 가이드라인'은 인공지능을 개발 및 활용하는 과정에서 준수해야 할 내용으로 투명성과 설명 의무, 자기결정권 보장, 차별 금지, 영향평가, 위험도 등급 및 관련 제도 마련 등을 제시하고 있다.

특정 분야나 대상에 한정한 가이드라인은 금융위원회(2021)의 '금융분야 AI 가이드라인', 개인정보보호위원회(2021)의 '인공지능(AI) 개인정보보호 자율점검표-개발자, 운영자', 서울특별시교육청(2021)의 '인공지능(AI) 공공성 확보를 위한 현장 가이드라인', 교육부(2022)의 '교육분야 인공지능 윤리 원칙', 식품의약품안전처(2022)의 '인공지능(AI)의 의료기기 국제 공통 가이드라인', 과학기술정보통신부(2024)의 '인공지능 학습용 데이터 품질관리 가이드라인'이 확인되었다. 이와 같은 가이드라인은 대상이나 영역을 특정하고 있어, 가이드라인 내용도 구체적이며 실용적으로 적용할 수 있도록 구성되어 있다. 가이드라인이기 때문에 법적 강제성은 없지만 인공지능에 대한 위험을 최소화하고 신뢰성 있는 인공지능 서비스를 제공하는 데 활용되기 위한 사회적 노력이라고 하겠다.

마지막으로, 선언은 특정한 가치나 방향을 표명하는 의미가 강하지만 상기한 가이드라인보다 강제성이나 구속력 측면에서는 상대적으로 낮다. 지향하는 목표나 방향성을 사회에 널리 알리고 해당 주제에 대해 사회적 지지와 관심을 촉구하며 자발적 참여를 독려한다.

2024년 5월 21일 채택된 서울 선언은 안전하고 보안성과 신뢰성을 갖춘 인공지능 보장이 필요함을 인식하고, 선언 참여국들과 관계 기관들의 인공지능 안전에 관한 연구 협력을 증진하기 위해 노력하겠다는 의지를 담고 있다. 또한, 안전하고 혁신적이고 포용적인 AI 생태계들을 육성하는 위험 기반 접근법들을 포함한 정책·거버넌스 체계들을 지지하고 있다(외교부, 2024.5.22.).

인공지능에 관한 국내의 법률은 법안의 형태로 법률 제정을 위한 과정 중에 있으며, 국내에서 실질적으로 인공지능을 규제 혹은 규율하는 것은 법률상 계획과 지침 및 가이드라인이다. 법률상 계획의 구체적인 실행 내용이나 과정은 인공지능에 초점을 맞춘 법률이 아니기 때문에 일정한 한계가 있을 수 있고, 지침이나 가이드라인은 법적 구속력이 없어서 실행으로 연결되지 않을 수 있다. 더군다나 가이드라인이 특정 영역이나 분야, 대상을 목적으로 하는 경우, 그 내용이 전체 사회에 적용되지 못하고 일부 집단이나 소수 영역에만 적용될 수 있다는 한계가 있다. 그럼에도 불구하고 인공지능의 신뢰성이나 안전성, 투명성, 설명 가능성 등의 가치를 추구하는 법률상 계획이나 가이드라인, 선언문 등이 발표되고 있는 현상은 긍정적으로 평가할 수 있다. 나아가 생성형 인공지능이나 고위험 영역의 인공지능에 대한 우려, 인공지능 활용에서 정보를 제공한 이용자에 대한 고지, 설명 가능한 인공지능 등 최근의 이슈들이 언급되고 있다는 점에서 구체적인 실행 방안과 의무 부여 여부와는 별도로 그 자체의 의미를 찾을 수 있을 것이다.

전 세계에서 국가 및 지역 정부 및 국제기구에서 인공지능에 대한 규제 혹은 가이드라인을 앞다투어 내놓고 있다. 제4장 제2절에서는 유럽연합의 AI Act와 미국의 행정명령 14110을 살펴봤다. 다른 형태들은 모두 법적인 구속력이 없는 권고 혹은 가이드라인의 형태를 띠기 때문이다.

유럽연합의 인공지능법(Artificial Intelligence Act)은 세계 최초로 AI에 관한 포괄적인 법적 프레임워크다. 새롭게 제정된 유럽연합 인공지능법의 목표는 유럽을 비롯한 세계 어느 지역에서도 신뢰할 수 있는 AI가 개발될 수 있도록 모든 AI 시스템이 인간의 기본권, 안전, 윤리 원칙 등을 존중하게 하고, AI 모델의 위험성을 관리하는 데 있다. 인공지능법은 모두 13개 장의 113개 규정과 13개의 부속서로 구성되어 있다.

인공지능법의 가장 큰 특징은 위험성 차원에서 AI 시스템을 분류한 다음, 그에 따라 규제의 내용을 차등화한다는 점이다. 유럽연합은 인공지능 모델의 성격과 용도 등을 고려해서 위험성의 경중을 진단하고, 그에 따른 규제의 수준도 연동하도록 했다. 이 법에서 AI 시스템의 위험성 수준을 '수용할 수 없는 위험성(unacceptable risk)', '고위험성(high risk)', '제한된 위험성(limited risk)', '최소한의 위험성(minimal risk)'으로 나누었다. 가장 높은 수준의 '수용할 수 없는 위험성'을 가진 인공지능은 활용 자체가 금지된다. 다음으로 '고위험성'을 가진 인공지능 활용에는 엄격한 준수사항이 요구된다.

유럽연합의 인공지능법에서 사회보장과 가장 연관된, 가장 민감한 대목은 '수용할 수 없는 위험성(unacceptable risk)'에서 세 번째로 제시된 사회적 평점(social scoring)이다. 사회적 평점 부여는 사회보장과 일정한 연관을 가질 수밖에 없다. 개인 혹은 가구 단위의 소득, 재산, 가구원 등의 정보에 근거해서 빈곤, 실업, 은퇴, 상병 여부를 판단하고, 그에 근거해서 급여를 제공하는 사회보장제도는 급여 자격을 판정하는 과정에서

일종의 사회적 평점(social scoring)을 개인 혹은 가구에게 부여할 수밖에 없다. 유럽연합도 사회적 평점과 관련한 비판 및 부작용을 염두에 둔 것으로 보인다. 법률에 붙은 (ⅰ)와 (ⅱ)의 내용에서 인공지능의 사회적 평점이 '수용할 수 없는 위험'으로 간주되는 경우를 한정했다. 즉, 특정 집단이나 개인에게 불리하거나 부정적이거나 부당한 대우를 초래할 때만 사회적 평점이 금지된다. 바꾸어 말하면, 사회보장제도에서 인공지능의 작동이 '순기능'을 하는 경우에는 규제의 대상으로 삼지 않겠다는 뜻이다. 이 대목은 앞으로 인공지능 기술이 사회보장 영역에서 적용되는 과정에서 끊임없이 논점으로 부각될 것으로 예상된다.

인공지능법에서 두 번째로 수위가 높은 '고위험(high risk)' 인공지능 영역을 보겠다. 여기에서는 다섯 번째 고위험 인공지능 영역으로 '필수 민간 및 공공 서비스 분야'가 제시됐다. 이 영역은 사회보장 영역을 직접적으로 언급하고 있다. 사회보장 행정을 집행하는 정부 및 공공기관은 고위험 인공지능 시스템 과정에서 '배포자(deployer)'로 구분될 가능성이 높다. 배포자는 고위험성 인공지능을 활용할 때 13가지의 의무를 이행해야 한다. 이를테면, 데이터보호 영향평가, 기본권 영향평가(fundamental rights impact assessment) 등이 그 예가 된다. 정부 및 공공기관 입장에서는 부담이 크다. 인공지능법의 실질적인 강도는 향후 추가로 마련될 가이드라인, 기준, 표준, 행동강령, 판례 등의 내용에 따라 결정될 것으로 보인다.

바이든 행정부는 2023년 미국 연방정부 최초로 관리예산처(OMB)의 각서 초안과 함께 안전하고 신뢰할 수 있는 인공지능의 개발 및 사용에 관한 행정명령(이하 행정명령 14110)을 발표했다. 이를 통해서 미국은 비로소 연방 단위의 AI 규제에 참여했다.

미국의 행정명령 14110은 전체 13조(section)로 구성됐다. 2조에서 인공지능 기술 활용에서의 8대 원칙을 제시했다. 이 가운데, 4대 '형평성과 시민권 증진' 원칙과 5대 '소비자, 환자, 승객, 학생 보호' 원칙이 사회보장과 직접 연결된다. 각각의 원칙에 따른 규제 내용은 행정명령 7조와 8조에 상세히 제시됐다.

먼저, 7조에서 2항은 '정부의 급여 및 프로그램과 관련한 시민권의 보호'로 명시된다. 내용을 보면, 기관들은 연방정부의 프로그램과 급여를 집행하는 과정에서 인공지능 기술의 활용으로 인한 "불법적인 차별 및 기타 피해를 예방 및 해결"(prevent and address unlawful discrimination and other harms)"(US Exec. Order No. 14110, 2023, p. 75212)하기 위해 각자 기관에 있는 민권 및 시민 자유 사무소(civil rights and civil liberties office)를 활용해야 한다고 규정하고 있다. 또, 행정명령 7조 2항 (b)호에 따르면, 정부 지급 급여의 공평성 제고를 위해서 보건복지부는 급여 및 서비스를 시행할 때 자동화 또는 알고리즘 시스템의 사용을 촉진하는 계획(plan)을 행정명령 시점 기준으로 180일 이내에 발표해야 한다. 행정명령의 8조(소비자, 환자, 승객, 학생 보호)에서는 직접적으로 관련 정부 부처들을 호명하면서 일일이 지침을 제시했다. 대략 3쪽 분량의 8조 내용 가운데 보건복지부에 특정한 내용만 2쪽가량을 차지한다. 이는 보건복지부가 수행하는 업무의 민감성이 반영된 결과로 추정된다. 해당 내용을 보면, 보건복지부는 행정명령 발표 이후 인공지능 관련 태스크 포스를 구성하고, 태스크 포스는 구성 시점 이후 365일 이내에 인공지능 관련 전략 계획(strategic plan)을 짜야 한다.

미국의 행정명령은 유럽연합의 인공지능법과의 차이를 살펴봄으로써 그 특징을 파악할 수 있다. 미국의 행정명령은 새로운 기관을 설립하지도 않고, 민간 기업에 대한 새로운 규제를 가하지도 않는다. 행정명령의 규율

대상은 연방정부와 관련 기관들이다. 따라서 이들이 자체 인공지능 시스템을 구매하거나 개발할 때 행정명령이 작동한다. 이를 통해 공공기관들이 책임 있는 AI 사용의 모범으로 작용하도록 유도한다. 또 공공기관들이 구매할 수 있는 인공지능 시스템에 대한 규제안을 제공하는 방식으로 간접적으로 민간 부문에도 영향을 미친다. 이러한 점에서 민간 부문을 직접적으로 규제하는 유럽연합의 인공지능법과 차이가 크다.

미국 행정명령은 또 사기, 차별, 금융 리스크, 그리고 AI가 잠재적으로 미칠 수 있는 프라이버시 문제로부터 소비자를 보호하는 것을 포함해, 시민의 권리와 자유를 보호하기 위한 규정을 만들도록 연방 기관에 요구하고 있다. 행정명령은 최고 AI 책임자(Chief AI Officer, 이하 CAIO)라는 직위를 새로 만들고, 모든 연방기관은 최고 AI 책임자를 60일 이내에 지정하도록 요구했다. 최고 AI 책임자는 개별 기관에서 AI 사용을 조정하고, AI 혁신을 촉진하며, 기관 내 AI 리스크를 관리하는 역할을 맡는다.

행정명령과 인공지능법의 공통점도 있다. 이 두 법안은 혁신을 촉진하면서 시민과 인권을 보호하려는 목표를 갖고 있다는 점에서 유사하다. 알고리즘의 투명성 강화, 인간의 감독, AI 편향 완화, 배포 전에 외부 스트레스 테스트나 '레드팀 테스트'를 광범위하게 수행한다는 원칙이 EO와 AI 법안 모두에 명시되어 있다. 미국의 보건복지부는 행정명령 발표 이전인 2021년에 최고 AI 책임자(CAIO)를 지명했다. 보건복지부는 행정명령 8조 (b) (ii)에서 규정하는 부처 차원의 인공지능 대응 전략 계획(strategic plan)을 2025년 1월에 제시하겠다고 밝혔다. 미국 연방정부에서 인공지능을 활용하는 사례 가운데 3분의 1 이상이 보건복지부 업무 영역에 속한 점을 고려할 필요가 있다. 미국 보건복지부의 이후 행보는 미국 차원에서 인공지능 활용 및 관련 규제에 상당한 영향을 미칠 것으로 보인다. 주 단위에서도 인공지능 규제 관련 입법 활동이 활발하다. 2024년

들어 AI 관련 입법 활동이 급격히 증가했음에도 불구하고 정책입안자들은 아직 구체적인 규제 모델에 대한 합의를 이루지 못하고 있다는 평가를 받고 있다.

제5장 제1절에서는 앞의 장의 내용을 종합하면서, 사회보장 영역에서 인공지능이 활용되는 열 가지 분야를 확인했다. 첫째, 본인 인증(identity verification)이다. 본인 인증은 급여 신청, 자격 심사, 급여 지급 과정에서 반드시 필요하다. 둘째, 자격 심사(eligibility assessment)이다. 빅데이터 및 인공지능을 통해서 급여 자격 심사를 빠르게, 정확하게 처리할 여지가 생긴다. 셋째, 복지 급여액 산정 및 지급(welfare benefit calculation and payments)이다. 다수의 국가에서 점점 더 많은 복지 급여액이 사람의 개입 없이 자동적으로 산정되고 지급되고 있다. 넷째, 부정·오류 수급 예방 및 탐색(fraud prevention and detection)이다. 많은 복지국가에서 디지털 자료를 활용하는 주요한 이유 가운데 하나가 부정·오류 예방 및 탐색에 있다. 다섯째, 위험의 점수화 및 범주화(risk scording and classification)이다. 제3장 제1절에서 살펴본 한국의 사각지대 발굴관리시스템이 여기에 해당한다. 여섯째, 개인 맞춤형 정보 서비스(personalized information service)이다. 제3장 제2절에서 살펴본 챗봇이 대표적 사례가 될 것이다. 일곱째, 온라인을 활용한 소통을 넘어서 실제 돌봄 영역에서도 활용되고 있다. 제3장 제1절에서 보았듯이, 이는 대부분 노인을 위한 것으로 AI·IoT를 활용한 건강 서비스, 생체건강 셀프체크 서비스 등이다. 여덟째, 사회보장 행정 기관 내부적인 용도로 업무 담당자의 효율적인 업무 처리를 돕고, 내부 교육 등의 용도로도 활용될 수 있다. 아홉째, 제3장 제1절에서 확인했듯이, 정보 활용의 효율성과 효과성을 높이기 위한 DBMS(Database Management System)의 활용이다. 데이터베이스의 관리는 업무의 효율화뿐만 아니라 빅데이터 분석

환경을 마련하는 데까지 연결된다. 열 번째, 사회정책의 효율성과 효과성을 평가하는 데 활용될 수 있다.

장기적으로 보면, 전 국민 대상 실시간 데이터에 근거한 인공지능의 활용은 사회보장제도 전체를 재편하는 방향으로 나갈 잠재력을 가지고 있다. 현행 85개 제도를 4~5개 정도로 단순화하는 재정 기반 소득보장제도를 준비해야 한다는 의견까지 나온다. 더불어, 디지털 사회보장 허브를 구축하여 급여 자격 심사와 지급을 효율화하자는 제안도 나온다. 인공지능 기술 적용으로 기대되는 순기능은 다음과 같다.

첫째, 효율성이다. 본문에서 살펴보았듯이, 본인 인증, 자격 심사, 복지 급여액 산정 및 지급 등의 일선 서류 행정이 더욱 자동화하게 된다. 둘째, 적시성이다. 빅데이터를 활용해서 관리되는 인공지능 기술은 급여 신청과 심사, 지급에 이르는 과정을 단순화할 수 있다. 셋째, 정확성이다. 빅데이터에 근거한 인공지능의 판단, 범주화, 예측은 인간의 오류와 편견이 개입할 가능성을 줄일 수 있다. 넷째, 개인 맞춤형 급여 제공이다. 인공지능을 활용한 행정 집행은 개인당 위험의 점수화 및 범주화까지 맞춤형 수준으로 정교화할 수 있다. 다섯째, 범용성이다. 챗봇을 통한 상담은 국민들의 입장에서는 시공간의 제약을 뛰어넘어서 제도에 접근할 수 있는 기반을 제시해준다. 여섯째, 정책 평가의 용이성이다. 빅데이터를 활용한 급여 집행 내용은 정책의 효과를 평가하는 데 용이한 기반을 제공한다. 일곱째, 사각지대 해소다. 제3장 제2절에서 살펴본 대로, 복지 사각지대 발굴관리시스템이나 AI 활용 초기상담시스템이 그 예가 된다.

빅데이터 활용을 수반하는 인공지능 기술이 사회보장 영역에서 초래할 위험성도 함께 보겠다. 첫째, 프라이버시의 문제다. 국가 권력에 의한 데이터 남용 가능성도 고려해야 한다. 둘째, 정확성의 문제다. 현장에서는 개인의 사망 및 출생 신고가 반영되지 않거나 과거 소득이 현재 소득과

합산돼서 제시되는 등의 문제가 끊임없이 발생하고 있다. 셋째, 데이터 소유권의 문제다. 급여 신청자는 소득, 재산, 가족 정보, 신체 등에 관한 자기 정보를 제공하는 데 동의하게 되는데, 여기서 정보의 소유권에 관한 문제가 제기된다. 넷째, 개인정보를 영리 목적으로 활용하는 것의 문제다. '디지털 헬스'를 둘러싼 데이터 활용은 한국에서 이미 오랜 의제다. 다섯째, 알고리즘의 결정에 근거한 국가 개입의 문제다. 빅데이터를 바탕으로 인공지능의 판단에 근거해서 국가가 어디까지 개인의 삶에 개입할 수 있는가, 라는 윤리적인 문제가 발생한다. 여섯째, 데이터와 알고리즘의 편향성의 문제다. 편향된 데이터와 알고리즘이 취약계층에게 차별적인 결과를 낳게 된다는 우려가 있다. 일곱째, 인공지능의 '설명 불가능성'의 문제다. 인공지능이 고도화할수록 설명 가능성은 떨어진다는 문제가 남는다.

제5장 제2절에서는 제1절에서 살펴본 리스크를 최소한으로 관리하면서 동시에 편의를 최대화하는 정책 방향을 검토했다. 신기술에 대한 규제와 지원 사이를 길항과 모순 관계로 볼 필요는 없다. 오히려 지원의 전제는 규제이고, 규제의 이유는 지원으로 볼 수 있다. 안타깝게도 한국에서는 사회보장 영역에서 인공지능 적용에 관한 공적인 비전 혹은 계획이 제시된 바 없다. 이와 같은 점을 염두에 두고 앞으로 빅데이터와 인공지능이 사회보장 영역에서 적용되는 것을 촉진하기 위한 정책 제언을 제시했다.

첫째, 사회보장 정보에 활용되는 데이터 품질을 개선하기 위한 정책적인 노력이 필요하다. 둘째, 사회보장 영역에서 데이터 통합, 연계, 관리를 위한 노력이 필요하다. 셋째, 다양한 기관들이 집적한 정보들의 표준화 및 단순화가 필요하다. 넷째, 데이터 구축, 연계, 활용 과정에서 데이터 보안 및 안전에 대한 엄격한 기준 설정이 필요하다. 다섯째, 데이터 관리를 넘어서 이를 활용하는 알고리즘 및 인공지능 시스템의 편향성을 최소화하면서, 정확성을 개선하기 위한 노력이 필요하다. 여섯째, 한국의 정부 부처에서,

특히 사회보장 영역에 한정하면, 보건복지부에서 인공지능 관련 조직을 신설하고, 인력을 배치할 필요가 있다. 미국 보건복지부가 행정명령에 따른 다양한 이행 조치로 'Office of the Chief Artificial Officer(OCAIO)'를 임명하고, 'HHS Artificial Intelligence(AI) Strategy'를 공표한 점을 염두에 둘 필요가 있다. 일곱째, 국제적인 인공지능 발전과 규제 동향에 대한 모니터링이 필요하다. 제4장에서 살펴본 바와 같이, 유럽연합의 인공지능법과 미국의 행정명령 14110도 적용 과정을 거치면서 규제의 내용이 구체화될 가능성이 높다. 여덟째, 지금까지 제시한 데이터 관리, 연계, 표준화 및 알고리즘 질 관리 등을 총괄하는 거버넌스 체계를 구축해야 한다. 종합하면, 사회보장 행정에서 인공지능 기술 적용에 대한 거버넌스는 인간 중심적 활용이라는 원칙 아래 ① 민주적 통제, ② 시민 참여, ③ 프라이버시 보호를 보장하고, 동시에 시민의 적극적 사회권 보호라는 원칙 아래 ① 사각지대 해소, ② 행정 효율화 제고, ③ 급여 지급의 정확성, 적시성 보장의 편의를 제공하는 정책 방향을 확인할 필요가 있다.

주요 용어: 인공지능, 사회복지, 사회보장, 사회정책, 빅 데이터

제1장

서론

제 1 장 서론

제1절 연구의 배경 및 필요성

인공지능의 급속한 발전이 초래할 사회 변화에 대해 다양한 전망이 제시되고 있다. World Economic Forum의 Fleming(2021)은 인공지능의 발전은 인류가 직면한 거대한 도전들, 이를테면 인신매매, 젠더 불평등, 암 질병 등의 문제를 해결하는 데 공헌할 것으로 전망했다. Roser(2022)는 컴퓨터와 인공지능으로 인해 인류가 보고, 알고, 수행하는 것이 모두 바뀌었지만, 인공지능 기술 발전의 짧은 역사를 고려하면 거대한 변화는 이제 막 시작되었을 뿐이라고 진단했다. 즉, 인공지능이 가져올 미래 충격은 현재로서는 가늠하기 어려울 정도라는 의미다.

인공지능의 파장은 인류의 존립을 위협하는 수준이라는 비관적 전망까지 제시되고 있다. Tesla의 Elon Musk나 Apple의 창립자인 Steve Wozniak 등 기술전문가들도 인공지능이 인간과 사회에 대한 위협이 된다며 해당 기술 발전의 일시적인 중단을 공개적으로 주장했다(Vallance 2023.3.30.). 케임브리지대 교수인 장하석(장하석, 2023.2.14.)도 인공지능의 일종인 고기능 챗봇에 대해서 "일이 커지기 전에 고기능 챗봇의 개발, 판매, 사용을 규제해야 한다"라고 밝혔다. Roubini(2024.2.5.)는 인공지능의 발전에 대해서 "우리의 미래에 가장 분명한 위협을 통제하기에 정치는 실패하고 있고, 정책은 방향을 잘못 잡고 있다는 사실을 명심해야 한다"라고 진단했다.

〈사피엔스〉의 저자 Harari(2018/2018)는 "데이터 소유를 어떻게 규제할 것인가... 이 질문에 조만간 답하지 못하면 우리 사회정치적 시스템

은 붕괴할 수도 있다"(p. 134)고 경고한 바 있다.

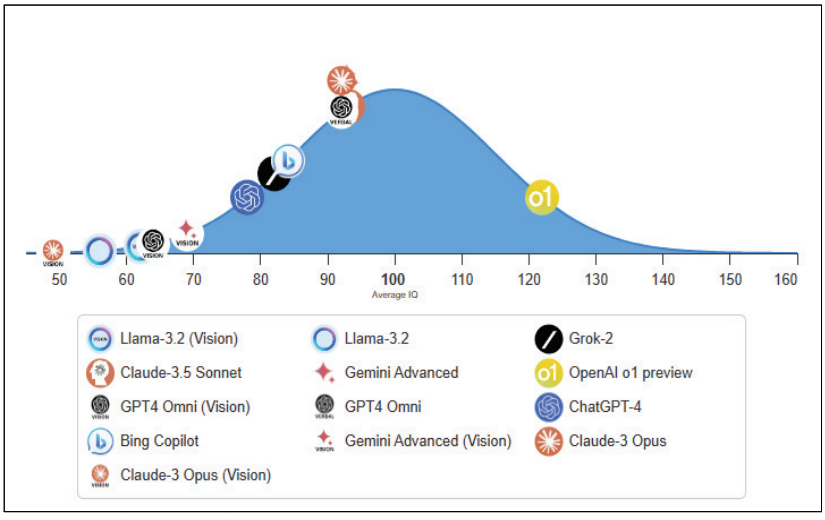
실제로, '인공지능의 대부'로 불리는 Geoffrey Hinton은 잘못된 정보의 범람, 인공지능이 고용 시장을 뒤흔들 가능성, 인공지능이 초래할 '실존적 위험'에 대한 우려를 들며 구글을 떠난 바 있다(Taylor, Hern, 2023.5.2.). Hinton은 2024년 노벨 물리학상을 수상한 이후에도 인공지능의 기술 발전 속도가 예상보다도 빠르다며, 앞으로 30년 안에 인공지능이 인류를 말살할 확률이 10~20% 정도라는, 섬뜩한 예측을 2024년 12월 인터뷰에서 내놓았다(Milmo, 2024.12.27.). 구글도 OpenAI보다 먼저 생성형 인공지능 기술을 개발했음에도 불구하고, 새로운 기술이 초래할 사회적, 윤리적 위험 때문에 기술 공개를 유보한 점은 익히 알려진 사실이다.

사회적 논란 속에도 인공지능의 기술 발전은 눈부시다. OpenAI가 내놓은 최신 모델인 OpenAI o1 preview 모델은 최근 IQ 테스트에서 인간의 평균을 능가했다(Lott, 2024.10.26.). 이러한 기술 발전 속에서 AI를 이용한 우울증 치료 기업 Eleos AI의 Robert Long은 Oxford대 Patrick Butlin, New York대 Jess Sebo 등 철학자들과 함께 쓴 'Taking AI Welfare Seriously'라는 논문에서 근미래에 인공지능이 의식(consciousness) 혹은 강력한 주체성(robust agency)을 가질 가능성에 대비해야 한다는 주장까지 제기하였다(Long, Sebo, Butlin, Finlinson, Fish, Harding... Chalmers, 2024).

사회보장은 인공지능의 발전과 적용이 가장 활발하게 관찰되는 분야 가운데 하나다(Zaber, Casu, Brodersohn, 2024). 인공지능 기술은 일부 복지국가에서 이미 급여의 자격 심사, 급여 지급 등의 과정에 적용되고 있기 때문이다. 그 과정에서 잡음도 적지 않다. 덴마크 Gladsaxe 지역에서는 2018년부터 취약 아동을 포착하기 위한 데이터 기반 모델을 적용

했다가 오류 문제로 사업이 중단된 바 있다(Jørgensen, 2021). 네덜란드의 시스템 위험도 표시(Systeem Risico Indicatie) 시스템은 사회보장 및 소득보장제도 분야에서 급여 오류, 부정수급을 시정하기 위해 활용되다가, 법원에서 데이터 투명성 등의 문제로 제동을 걸면서 활용이 중단되기도 했다(Appelman et al., 2021). 아동수당 및 실업수당 등 복지 급여를 중심으로 오스트리아, 뉴질랜드, 미국, 영국 등의 국가에서 인공지능 기술이 적용되고 있다(정세정 외, 2023).

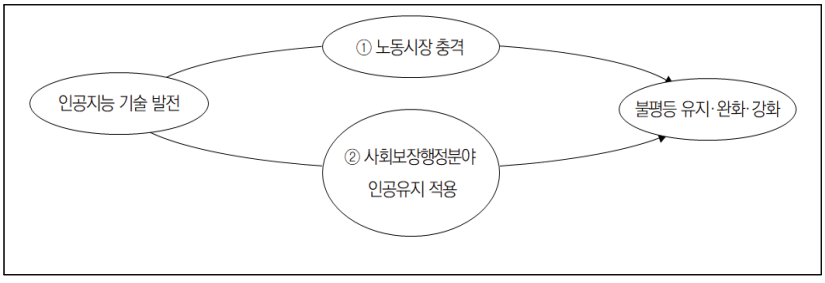
〔그림 1-1〕 인공지능 모델들의 IQ 테스트 결과



출처: Lott. (2024.10.26). "IQ Test Results". IQ Test | Tracking AI

인공지능은 사회정책 영역에서 효율성, 과학성, 중립성의 증진에 기여할 것이라는 기대가 크다(Zaber, Casu, Brodersohn, 2024). 반면, 새로운 기술이 현존하는 빈곤, 불평등, 차별을 유지 및 강화한다는 비판도 있다. Alston(2019)은 인공지능이 데이터에 근거해서 위험을 점수화 및 범주화하는 과정에서 나타나는 세 가지 문제점을 제시한 바 있다. 오류의 가능성, 개인 권리 침해의 가능성, 그리고 현재의 불평등 및 차별을 유지하거나 강화할 가능성이다. 김재연(2023) 역시 "우리에게는 다른 데이터가 필요하다"는 저서에서 공공이 통제 및 관리하지 않는 데이터가 어떻게 현재의 차별을 정당화하는 방식으로 악용될 수 있는지를 경고한 바 있다. Birhane(2024)는 2024년 6월 서울에서 열린 '사람과 디지털 포럼'에 참석해서 빅테크가 주도하는 인공지능 개발은 유색인종·여성·약자에 대한 부정적인 고정관념을 강화해 사회적·역사적 불평등을 악화한다고 경고했다.

[그림 1-2] 인공지능 발전이 불평등을 유지 및 강화하는 두 가지 경로



출처: 연구진 작성

사회정책에서 디지털 및 인공지능 기술 발전 → 노동시장 충격 → 불평등 심화에 관한 연구는 다수 이뤄졌다(Frey, Osborne, 2017; 이승윤, 백승호, 남재욱, 2020). 다만, 사회보장 행정 영역에서 인공지능을 적용할 때 미칠 불평등 유지, 완화 혹은 심화의 경로에([그림 1-2]의 두 번째

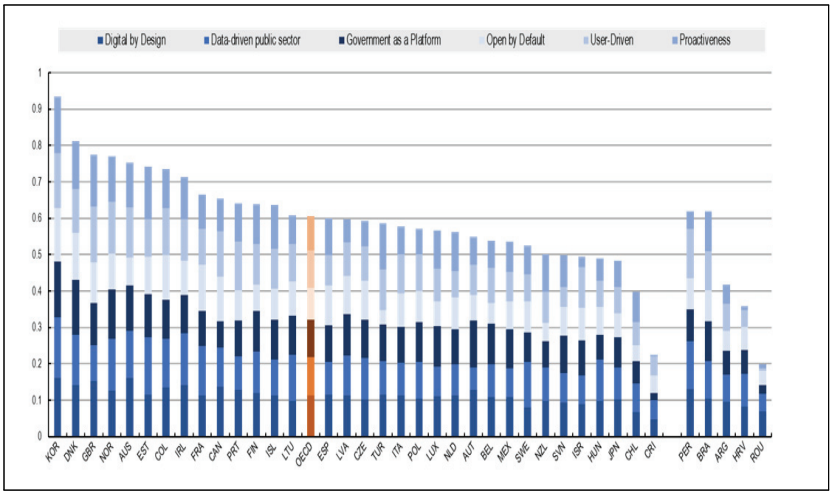
경로) 관한 연구는 상대적으로 적다. 복지국가의 행정에서 디지털 기술이 활용되는 경로는 다시 일곱 가지의 세부 경로로 제시된다[2](Alston, 2019; 김기태, 2024). 사회보장 영역에서 인공지능 기술 적용은 "보다 적극적이고 자동화된 서비스 제공이 가능해지며 효율성, 효과성 및 대응성이 향상"(Zaber, Casu, Brodersohn, 2024, p. 1)되는 계기가 된다. 동시에, 인공지능이 활용하는 디지털 자체의 편향성으로 인해 기존의 불평등 및 차별이 유지 및 강화될 가능성에 대한 우려도 지속된다. 실제로, 네덜란드(van Bekkum & Borgesius, 2021), 덴마크(J ø rgensen, 2021), 미국(Eubanks, 2018)에서 이와 관련한 비판적 목소리가 제기되고 있다.

사회보장 영역에서 인공지능이 초래할 영향 및 정책적 대응에 대한 국내 연구는 희소하다. 디지털 및 정보화의 사회정책적 함의에 관한 연구는 일부 있었고(김수영, 2016; 김수영, 김수완, 2022; 정세정 외, 2023), 보건복지 현장에서의 인공지능 기술 적용에 관한 연구(조남경, 송기호, 2023; 정유채, 2023)는 일부 수행된 바 있다. 성윤희(2022)는 사회보장 영역에서 인공지능이 초래할 변화를 검토한 드문 연구이지만, 인공지능을 "4차 산업혁명의 요체이며 지식 기반 사회의 핵심 화두"로 파악하는 기능적 부분에 초점을 뒀다.

2) 디지털 기술의 일곱 가지 영역은 ① 본인 인증(identity verification), ② 자격 심사(Eligibility assessment), ③ 복지 급여액 계산 및 지급(Welfare benefit calculation and payments), ④ 부정수급 예방 및 탐색(Fraud prevention and detection), ⑤ 위험의 점수화 및 범주화(Risk scoring and classification), ⑥ 복지 수급자와 기관 사이의 소통(Communication between welfare authorities and beneficiaries), ⑦ 공공 및 민간을 포함하는 기관 사이의 소통 및 연계 (Communication and cooperation between welfare institutions)이다. 이 모든 영역에서 인공지능은 국내·외에서 이미 작동하고 있거나 작동될 것으로 예상됨.

한국이 공공행정에서 인공지능의 활용이 뒤처진 것도 아니다. 한국은 OECD(2024a)가 발표하는 디지털 정부 지수 순위에서 압도적 1위를 차지했다(〔그림 1-3〕 참고). 한국은 여섯 개 평가 영역 가운데 데이터 기반 정부, 플랫폼 정부, 개방형 정부, 선제적 정부 등에서 1위를 차지했고, 디지털 우선 정부, 국민 주도형 정부 2개 부문에서 2위였다. 한편, 영국 미디어 업체인 Tortoise Media에서 제시한 'Global AI Index'에서는 미국, 중국, 싱가포르, 영국, 프랑스에 이어서 6위에 한국이 위치했다(Cesareo, White, 2023).

〔그림 1-3〕 디지털 정부 지수 기준 순위



출처: OECD. (2024a). Figure DA.

실제로, 한국의 공공영역이 구축한 데이터 인프라는 압도적이다. 하루 평균 125만 명의 국민이 정부24(www.gov.kr/portal/main)에 방문하고, 이곳에서 2023년 3월 기준 1,336종의 민원 서류를 신청하고 발급받고 있다(행정안전부, 2023. 4. 19., 보도자료). 한국 주민등록번호 시스

템의 특이성도 염두에 둘 필요가 있다. 한국은 세계적으로 유례가 드물게 전 국민을 상대로 한번 발급되면 변경이 어려운 일률적인 국민식별번호를 운영하고 있다(성준호, 2016). 실제 사회보장 현장에서도 복지 사각지대 발굴시스템, AI·IoT 기반 어르신 건강관리서비스 사업 등 인공지능 기술이 활발하게 적용되고 있다(제3장 내용 참고).

이와 같은 점을 고려하면, 한국은 사회보장 영역에서 인공지능 기술을 디지털 인프라 위에서 빠른 속도로 적용시키고 있는 한편, 인공지능 기술 적용이 초래할 법적·윤리적 부작용 문제에 대해서는 상대적으로 방임적인 태도를 보였다. 관련 연구와 정책이 한국에서 제시된 바가 희소한 것은, 이러한 상황을 증명한다. 이번 연구는 이러한 문제 의식에서 출발했다. 한국과 다른 복지국가들의 사회보장 영역에서 인공지능 기술 적용 현황 및 관련 규제 동향을 살펴보고, 이를 통해서 인공지능 기술의 사회보장 적용 과정에서 짚어야 할 정책적, 제도적 함의를 모색하고자 한다.

제2절 연구의 목적 및 내용, 추진 방법

이번 연구의 목적은 크게 두 가지다. 첫째, 인공지능의 국내·외 적용 현황을 살펴보고자 한다. 국내·외에서 인공지능이 사회보장 현장에서 인공지능이 적용되는 사례에 관한 종합적인 연구가 드물었던 점을 고려했다. 해외의 경우, 국제사회보장협회(International Social Security Association)에서 발간한 49쪽짜리 'Artificial Intelligence in Social Security Organisations'라는 보고서(Zaber, Casu, Brodersohn, 2024)가 전 세계적으로 사회보장 영역 적용 현황에 대한 현황을 그려내고 있다. 그렇지만, 이 보고서가 국내에 소개된 바도 없고, 해당 보고서는

특정 국가나 제도의 사례에 대한 설명이 적다. 더욱이, 국내에서 사회보장 영역에서 인공지능 적용 사례를 종합한 연구도 없다. 이러한 점에서 이번 연구는 국내·외 적용 현황에 대한, 기초적인 형태이지만, 종합하는 데 목적을 둔다.

둘째, 인공지능의 발전에 대한 국내·외의 정책 대응을 살펴보고, 의미와 한계를 살펴보고자 한다. 인공지능의 급속한 발전을 지원하면서 동시에 기술이 초래할 파괴적 부작용을 막기 위한 규제 및 법률이 전 세계적으로 동시다발적으로 생겨나고 있다. 여기에는 사회보장 관련 규제 및 법령이 포함됐음은 물론이다. 유럽연합의 인공지능법(AI Act)이 대표적인 예다. 그 밖에 미국에서는 대통령의 행정명령이 2023년에 공표된 바 있다. 또 영국, 캐나다, 중국, 일본 등도 규제 및 법률안을 내놓고 있으며, OECD, UNICEF, UN 등도 예외가 아니다. 한국에서는 입법의 과정이 더디다. 한국에서도 인공지능 규제를 위한 법률안이 지난 제21대 국회에서 준비됐다가 무산된 바 있다. 제22대 국회에서는 2024년 12월 26일 '인공지능(AI) 발전과 신뢰 기반 조성 등에 관한 법률'(AI 기본법)이 국회 본회의를 통과했다. 그러나 이번 보고서의 작성이 마무리된 2024년 12월 28일 기준으로도 해당 법률은 공표되지 않았다. 이러한 상황을 고려해서, 2024년 하반기 기준으로 국내·외 규제들 가운데 대표적인 사례를 골랐다. 해당 규제들의 내용을 분석하고, 한국 사회보장제도에 주는 함의를 제시하고자 한다.

주요한 연구 내용은 다음 세 가지다. 먼저, 인공지능 발전의 동향을 살펴본다. 인공지능 기술 발전의 첨병인 미국의 기술 발전 내용을 중심으로 검토한 뒤, 인공지능 발전에 따른 윤리적인 쟁점을 짚어본다. 둘째, 국내·외에서 인공지능이 사회보장 영역에서 활용되는 현황과 주요 사례를 분석한다. 한국 사회보장 영역에서 인공지능 기술의 현황을 파악하기 위해서,

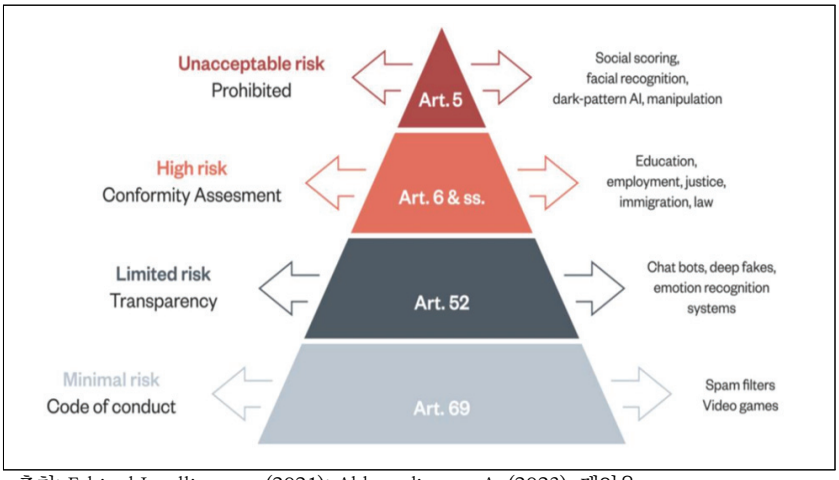
조달정보개방포털과 나라장터에서 최근 5개년 동안(2019년 10월~2024년 9월) 확인되는 사회보장 분야의 복지 기술 활용 현황을 조사했다. 그리고 인공지능이 사회보장 현장에서 실제로 적용되는 세 가지 사례인 AI 기반 일자리 매칭 서비스, 복지 사각지대 발굴관리시스템, AI 활용 초기상담시스템을 분석한다. 국외의 경우에는 전 세계 국가의 사회보장 영역에서 인공지능 적용 현황을 보고한 Zaber, Casu, Brodersohn(2024)의 보고서의 내용을 중심으로 현황을 일람한다.

셋째, 국내·외에서 인공지능에 대한 규제의 내용을 분석한다. 먼저, 국내의 경우에는 법·법률상 주요계획·지침 및 가이드라인·선언 등 법적 강제성의 수요에 따라 규제의 내용을 점검했다. 이를테면, 지난 제21대 국회에서는 9건의 인공지능 관련 법안이 발의됐으나 모두 폐기됐고, 제22대 국회에서는 현재 11개의 법안이 발안됐다. 해당 내용들을 분석했다. 안타깝게도 2024년 12월 26일 국회 본회의를 통과한 인공지능기본법은 보고서에서 다루지 못했다.

국외의 경우에는 국제적으로 주목을 받는 두 개의 규제를 살펴보았다. 유럽연합에서 2024년 8월에 발효된 인공지능법과 미국에서 2023년 10월에 발효된 행정명령 14110이다. 특히, 유럽연합의 AI 규제법은 AI 활용 범주에 따라 위험도를 4단계로 나누어서 규제를 차등 적용([그림 1-4] 참고)하는 강제력 있는 규제다. 사회정책의 영역인 의료, 교육, 고용 등 공공서비스는 두 번째로 높은 규제 영역인 '고위험 등급(high risk)'으로 분류되어 위험관리 시스템 구축 및 인간 관리자의 감독 아래 놓이게 됐다. 유럽이 설정한 기준은 국내뿐 아니라, 전 세계적으로 영향을 미칠 가능성이 높다. 미국의 경우는 '안전하고 신뢰할 수 있는 인공지능의 개발 및 사용에 관한 행정명령'(Executive Order o the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,

이하 행정명령 14110)을 살펴봤다. 행정명령은 민간에 대한 규제가 아니라, 연방정부 기관 및 연방정부와 계약을 맺은 민간 업체만을 대상으로 한다. 행정명령 가운데 7조(형평성과 권리 증진), 8조(소비자, 환자, 승객, 학생 보호), 10조(연방정부의 노력) 부분은 미국 보건복지부가 지켜야 할 규제를 담고 있어서 한국의 보건복지부가 참고할 내용이 있다. 국내외의 규제 내용에 대한 분석을 통해서, 한국 사회보장 영역에서 인공지능 기술을 어떻게 규제할 것인지에 대한 함의를 모색했다.

〔그림 1-4〕 유럽연합 AI 규제법에서 제시된 AI 작동 범주 및 규제 내용



출처: Ethical Intelligence. (2021); Akhmedjonov, A. (2023). 재인용.

연구의 추진 방법은 다음과 같다. 첫째, 문헌 분석이다. 연구 주제의 현안으로서의 특성을 고려해서, 국내의 공공기관, 유럽의회, United Nations 등 국제·국가 기구에서 발간한 보고서 및 법령, 잡지·기사·블로그, 학술 논문 등 문헌을 폭넓게 분석한다. 학술 논문 외에 최근 소식을 파악할 수 있는 언론 기사 및 공신력 있는 블로그, 전문가들이 SNS에 올리는 글도 적극적으로 인용했다.

둘째, 국내 인공지능 관련 전문가 자문 및 세미나를 진행했다. 인공지능의 발전 문제는 IT 및 공학 기술과도 밀접하게 연관된 점을 고려했다. 정보기술·사회보장·경제·법학 등 다양한 전문가들의 자문 및 세미나를 진행했다. 다른 학문에서는 이미 인공지능법학회, 인공지능윤리학회, 대한의료인공지능학회 등이 만들어져서 일정한 학술적인 결과물을 낸 점을 고려했다. 세미나 일정, 주제 및 세부 내용은 〈표 1-1~2〉에 제시했다.

〈표 1-1〉 인공지능 세미나 프로그램

차수	일시/장소	주제 및 발표자/ 전공
1차	2024.07.18. 온라인	디지털 전환과 사회복지의 미래 김수영 교수(서울대 사회복지학과/사회복지학)
2차	2024.07.23./ 보사연 서울 조사센터	AI 시대의 사회 변화와 공공의 과제 구본권 소장(사람과 디지털 연구소/언론정보학)
3차	2024.07.23./ 보사연 서울 조사센터	인공지능 윤리 핵심 가치 분석: 한국 사례를 중심으로 김형주 교수(중앙대 인공지능인문학단/철학)
4차	2024.08.07./ 보사연 서울 조사센터	AI 윤리와 규제 김명주 교수(서울여대 정보보호학과/컴퓨터 공학)
5차	2024.08.26./ 보사연 서울 조사센터	인공지능 발전과 공공의 과제 김병권 연구위원(녹색전환연구소/경제학)
6차	2024.09.23./ 온라인	유럽연합의 AI Act의 내용과 한국에 주는 함의 윤혜선 교수(한양대 법학전문대학원/법학)
7차	2024.10.22./ 서울 스마트워크센터	사회보장 행정에서 인공지능 적용 동향과 함의 이우식 박사(사회보장정보원/컴퓨터 공학)
8차	24.11.07/ 아우름 비즈 회의실	AI 일자리 매칭 서비스 소개 조인성 박사(사회보장정보원)
9차	24.11.07/ 아우름 비즈 회의실	정부의 사회보장서비스에서 AI 활용과 EU/미국 일부 주 법의 시사점 강지원 변호사(김앤장/ 법학)

출처: 연구진 작성

〈표 1-2〉 인공지능 포럼 내용

차수	세부 발표 내용
1차. 디지털 전환과 사회복지의 미래 (김수영 교수)	□ 디지털 기술혁명에 따른 네트워크 사회구조 변화와 이에 따른 복지 정책 기반의 질적 변화가 야기됨. ○ 디지털 네트워크는 시장, 국가, 시민사회에 각각 두 가지 상반된 힘으로 작용하고 있음. ○ 시장 플랫폼 기업은 거대해지나, 노동은 분자화되며, 일시적 독립적, 비정형 일자리가 증가함. ○ 국가 경계를 초월한 노사관계, 고용관계를 벗어난 기업과 노동자 등 다양한 변화가 발생함. □ 데이터 복지의 양면성, 정보시스템을 통한 데이터 감시 등의 문제가 대두됨. ○ 디지털 네트워크 사회에 진입하며 시민사회조직이 약화됨. ○ 온라인 네트워크가 확대됨에 따라 의사소통에서 확산성은 강화되었으나, 편향성의 문제가 존재함. □ 한국 네트워크 사회는 자국 중심의 네트워크 사회임을 고려한, 미래 사회의 변화된 욕구와 가치를 반영한 복지정책을 마련하는 것이 필요함. ○ 국가의 데이터 감시에 대한 높은 수용성은 향후 국가가 더욱 강력한 정보 통치력을 갖게 할 가능성이 있음을 시사함. ○ 관계 지향적 복지 거버넌스 모델을 제안함. 이를 기반으로 미래 사회보험(분리형 vs 통합형), 공공부조(집단 표적형 vs 기반 조성형), 사회서비스(개별 지향형 vs 연결 지향형) 개편 시나리오 제시
2차. AI 시대의 사회 변화와 공공의 과제 (구본권 소장)	□ 범용인공지능(Artificial General Intelligence, AGI)의 발전에 따른 일자리의 변화 ○ 디지털 기술 발달에 따라 노동 조건의 변화가 발생함 (자동화, 세계화, 기술을 통한 숙련노동자의 생산성 강화). ○ 구조적 실업의 증가가 예상되며, 기본 소득 도입 논의 활발 ○ 인공지능 발달에 따른 노동시장과 지식 기준의 변화 발생 □ 인공지능 발달에 따라 발생할 수 있는 사회의 변화, 탈진실 사회로의 진입, 알고리즘에 의존하는 사회에 대한 문제 제시 ○ 탈진실: 객관적 사실보다 개인의 신념과 감정적 호소가 여론에 더 큰 영향을 끼치는 현상 ○ 인공지능 사회에 필요한 새로운 시민성과 리터러시 능력에 대한 필요성 제시

차수	세부 발표 내용
3차. 인공지능 윤리 핵심 가치 분석: 한국 사례를 중심으로 (김형주 박사)	□ 인공지능 윤리 기준은 인간성 중심으로 인간 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙을 기준으로 함. ○ 인권 보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성을 키워드로 함. □ 인공지능 관련 다양한 기술들은 문제해결력과 설명 가능성에 있어서 일종의 상쇄관계에 있음. ○ 문제해결력이 높은 Deep Learning 같은 기술은 설명 가능성이 매우 낮고, 상대적으로 설명 가능성이 높은 Decision Tree 같은 기술은 문제해결력이 낮은 딜레마에 처하게 됨. □ 글로벌 인공지능법인 EU의 AI ACT의 인공지능 위험 수준 구분 기준을 제시하여 우리나라의 규제 방향성에 대해 논의함.
4차. AI 윤리와 규제 (김명주 교수)	□ AI 현황과 기술의 트렌드 제시 ○ 생성형 AI의 멀티모달(multimodal) 기능 확장, 오픈 AI의 GPT 스토어 확산, AI 관련 기술적 분화 및 선택의 다양화, 이용자 연령의 확대, (글로벌) 규제의 기술적 수용 논의 ○ AI 기술에 대한 기본 이해를 위해 인공지능 시스템 정의, 개요, 오토 인코더 등 주요 용어 및 기술 설명 □ 인공지능이 가지고 있는 특징에 따라 수반돼야 하는 윤리적 원칙들이 제시될 수 있음. ○ 기술의 보편적 특징으로 인류 공영 기술, 사회 변화 기술의 특징은 공공성, 다양성, 지속 가능성, 책무성, 안전성, 건전성을 요구 ○ 인공지능의 특수성인 자율성, 지능성, 학습 가능성은 통제성, 투명성, 설명 가능성, 공정성, 프라이버시 보호, 저작권 보호를 요구
5차. 인공지능 발전과 공공의 과제 (김병권 위원)	□ 인공지능 기술의 장단점, 관련 정책 방향성 제시 ○ 인공지능의 공적 활용과 이슈 분석을 통해 한국인이 생각하는 인공지능의 장단점, 인공지능 발전을 위해 중요한 정부 정책 방향성, 공공서비스에 적용했을 때의 이점과 위험, 활용 가능 분야 등 설명 ○ 인공지능의 잠재적 이점으로 일상생활의 편의성 향상(30.6%), 업무추진의 효율성 증진(19.6%), 잠재적 위험으로 설계/오작동 발생으로 인한 피해(18.5%), 악의적 의도로 인공지능 활용에 따른 피해(18.3%)에 주목. ○ 인공지능의 발전을 위해 가장 중요한 정부정책 방향으로 인공지능의 윤리 기준 및 인공지능법 제정(34.6%), 국가 마스터 플랜 마련(18.8)을 중요 정책 방향이라고 응답함.

차수	세부 발표 내용
	□ 생성형 인공지능의 공공서비스 적용에 대한 검토 ○ 잠재적 이점에도 불구하고 공공 부문에서 인공지능 배포에 대한 우려가 존재함. ○ 공공 부문에 인공지능을 적용할 경우 생산성, 대응성, 책임성을 강화할 수 있음. ○ 생성형 인공지능은 추론 과정을 알 수 없는 블랙박스 구조이며, 영어 우세성 때문에 모델의 추론을 결정하는 값이 주로 미국 사회의 특정 부문의 가치에 기반할 수 있음. □ 인공지능 사용을 위한 데이터 센터의 전력 과소비, 자원의 파괴 등 디지털 과소비에 따른 문제점을 제시하며 지속 가능한 디지털화가 필요함을 시사함.
6차. 유럽연합의 AI Act의 내용과 한국에 주는 함의 (윤혜선 교수)	□ 사회보장 영역에서 인공지능은 다음과 같은 네 가지 영역에서 활용 ○ 맞춤형 서비스 제공: AI를 통해 개인의 상황과 필요에 맞는 맞춤형 사회보장 서비스 제공 ○ 부정수급 방지: 데이터 분석과 패턴 인식을 통해 부정수급 효과적 탐지·예방 ○ 행정 효율성 향상: 반복적인 업무의 자동화를 통한 행정 효율성 제고 ○ 정책 결정 지원: 빅데이터 분석을 통해 정책 결정 및 예측의 정확성 제고 □ 사회보장 행정에 활용되는 AI 시스템 대부분이 '고위험성'군으로 분류될 가능성 높음 ○ 허용할 수 없는 위험성, 고위험성, 제한된 위험성 및 범용 AI 모델 규제 적용 가능성 높음. 엄격한 규제와 관리 체계 적용 ○ 사회보장행정 분야에 고유한 위험성 평가 체계 및 위험도에 따른 차등 규제 체계 도입 검토. 데이터보호 영향평가 + 기본권 영향평가 ○ 금융 분야 자금세탁 방지(AML)의 AI 활용 특례 벤치마크
7차. 사회보장 행정에서 인공지능 적용 동향과 함의 (이우식 박사)	□ 사회보장 분야에서는 2024년 보건복지 업무계획에 약자복지 2.0을 통한 인공지능 고도화를 준비함. ○ 복지위기알리미, AI 복지봇, 연령 모형, 지역 모형 등이 있음. ○ 의료 분야는 의료 인공지능 로드맵이 나올 정도로 빠른 속도로 중장기적인 계획이 나옴. 복지 분야도 복지 인공지능 로드맵을 만들어 중장기적 계획이 요구됨.

차수	세부 발표 내용
	□ 사회보장기본계획에서 검증된 디지털 기술 적용 및 인공지능 AI 복지봇 구축 ○ 사회 인공지능은 점차 발전하고 있으며, 생성형 인공지능에 대한 수요가 급격히 증가하기 때문에 사회보장 분야에 생성형 AI 초거대 언어 모델(LLM) 같은 기술 적용이 요구됨. □ 사회보장정보원 내부 AI 서비스로 사회보장 인공지능의 동향을 살펴보고, 운영하고 있는 AI 시스템 및 향후 방향성에 대해 논의함. ○ 한국사회보장정보원은 다양한 신기술, 인공지능 시스템을 적용하였지만, 별도 부서에서 분리되어 운영 관리되고 있어 전체 모니터링 한계가 존재함. ○ 과거 행정 데이터의 한계를 넘어 비정형 데이터 등의 활용, 민간 데이터의 활용도가 높아질 것으로 예상됨. 다양한 복합 데이터를 활용하는 멀티모달(multimodal) AI로 발전해야 함. □ 복지 사각지대 발굴시스템 도입과 관련해 초기 행정 데이터에서 복합 데이터 활용이 증가함. ○ 활용 모형 수가 늘어나고 있으며, 음성 인식도 활용하고 있음. ○ AI 활용 초기상담 정보시스템(AI 복지봇) 구축이 완료되었으며 본 사업을 앞둔 상황임.
8차. AI 일자리 추천 서비스 소개 (조인성 박사)	□ 워크넷 일자리 추천 서비스는 숙련·정보 미스매치 해소를 위해 최신 인공지능 기술을 활용할 때 일자리 조건 검색에서 개인별 정밀 매칭이 가능한 지능형 일자리 매칭으로 진화함. ○ 일자리 추천 서비스 프로세스는 원천데이터 수입·적재(nifi)-직무능력 추출, 데이터 전처리(preparing system)-인공지능 모델 생성 및 학습(learning system)- 실시간 추천을 위한 operation system으로 구성 ○ 빅데이터와 머신러닝 기술을 활용하여 구직자에게 직무와 행동에 기반한 일자리를, 구인 기업에게 직무에 적합한 인재를 추천하는 AI 일자리 매칭 서비스를 제공하는 추천 알고리즘 사용 ○ 잡케어는 서비스 대상별 역량 진단 시행 후 노동시장의 정보 분석을 바탕으로 직업을 추천함으로써 역량 수준별 경력을 설계하고자 함. □ 인공지능 기술 적용에서 고려해야 할 사항으로 다음 6가지 제시 ○ 1. 해결하고 싶은 문제가 정의되었는가? ○ 2. 문제 해결 방법이 AI밖에 없는가? 성과가 날 수 있는가? ○ 3. 데이터는 확보했는가? 특히, 외부데이터를 활용하는 경우

차수	세부 발표 내용
	○ 4. 제공하고자 하는 서비스가 어떤 특성을 갖고 있는가? ○ 5. 성과 지표 설정 (재현율 관점 vs 정밀도 관점) ○ 6. 지속적인 유지관리 체계 마련
9차. 정부의 사회보장 서비스에서 AI의 활용과 EU/미국 일부 주 법의 시사점	□ 세계 유일의 포괄적인 AI 규제 법제인 EU AI Act. ○ AI 활용 사회보장 서비스를 명시적 규율 대상으로 두고 있어 유사한 입법이 논의되고 있는 우리나라 선례로서 시사하는 바가 큼. □ 미국의 경우 연방 차원에서는 인공지능 관련 법 외 규제 성격의 법률은 아직 제정되지 않았음. ○ 개별 주 차원에서는 차별 금지, 정보 공개 등을 중심으로 한 규제법이 제정되기 시작함. ○ 콜로라도 AI Act상 알고리즘 차별 금지: EU법과의 차이점을 규제 체계의 개요, 고위험 AI 시스템의 정의, 알고리즘 차별 관련 배포자의 준수 의무로 구분하여 설명함. □ 공공 영역의 사회보장 서비스에 대한 EU AI Act의 적용 중 기본규율체계, 사회적 평점 산정 기준, 필수적 공공 서비스에 대한 접근/향유 항목을 요약하여 설명함.

출처: 연구진 작성

제2장

인공지능 발전의 동향 및 쟁점

제 2 장 인공지능 발전의 동향 및 쟁점

제1절 인공지능 기술 발전 동향

1. 인공지능 70년 역사와 중요 이정표

인공지능의 역사는 1950년대 컴퓨터의 태동과 함께 시작되었다. '계산기계와 지능'이라는 논문을 앨런 튜링이 1950년에 발표하면서 튜링 테스트를 제안하였는데 이것이 인공지능 개념의 기초가 되었다(Turing, 1950). '인공지능'이라는 용어는 1956년 다트머스 회의를 통해서 처음으로 생겨났으며 이를 계기로 인공지능에 관한 연구가 본격적으로 시작되었다(Kline, 2011).

1960년대에 이르러서는 인식 패턴과 컴퓨터 비전에 관한 다양한 연구가 진행되었다. 그러던 중 시모어 페퍼트와 마빈 민스키의 부정적인 연구 결과로 인해 1969년부터 연구비가 중단되면서 인공지능의 첫 번째 겨울이 시작되었고, 결과적으로 인공지능에 관한 연구는 매우 축소되었다(Crevier, 1993).

1980년대 초, 전문가 시스템의 도입으로 인공지능은 두 번째 붐을 맞이했다(Joseph, 2023). 1986년 제프리 힌턴 등이 다층 퍼셉트론(MLP)과 역전파 알고리즘을 증명하며 신경망 연구에 전환점을 마련했다(Schmidhuber 2014). 그러나 1990년대에 들어서면서 전문가 시스템의 한계가 드러나며 두 번째 인공지능 겨울기가 찾아왔다(Russell, 2021).

2000년대에 접어들면서 인공지능은 새로운 전환점을 맞이한다. 2006년 제프리 힌턴이 심층신뢰신경망(DBN)을 발표하며 딥러닝 연구의 기반이 마련되었고, 2012년 알렉스넷(AlexNet) 팀이 ILSVRC 대회에서 우승하면서 딥러닝이 주목받기 시작했다(Krizhevsky et al., 2012).

2010년대 이후 인공지능은 급속도로 발전하고 있다. 2016년 구글 딥마인드의 알파고가 이세돌 9단을 격파한 사건은 인공지능의 잠재력을 각인시켰다. 2017년에는 Google 연구진이 LLM의 기본 모델인 Transformer 모델을 최초로 발표하였다(Vaswani et al., 2017).

인공지능은 지금 여러 분야에서 큰 사회적 혁신을 이끌고 있다. 이와 더불어 인공지능이 지닌 잠재적 위험(potential risks)에 대한 논의도 적극적으로 병행되고 있다. 일부에서는 지금보다 훨씬 진보한 인공지능인 AGI(인공일반지능)가 예상보다 일찍 나타날 것이라는 전망을 내놓고 있다. 물론 일부는 이와 정반대로 인공지능의 세 번째 겨울이 다가올 가능성이 크다고 경고하기도 한다.

70년에 걸친 긴 인공지능의 역사를 보면, 혁신적인 기술 발전과 예상치 못한 거품 논쟁을 반복했다. 그래서 인공지능의 역사에는 이미 두 번의 겨울이 찾아왔다. 이러한 시기마다 인공지능의 기술적 한계를 직시하고 이를 돌파하려는 노력이 중첩되면서 인공지능은 발전 방향을 정해왔다. 어느 정도 시간이 흐른 지금에 와서는 단순히 기술적 발전뿐만 아니라 윤리적인 영향과 사회적인 영향에 대한 논의도 더불어 진행되고 있다. 어쩌면 인공지능의 미래는 여전히 확실하지는 않지만, 인공지능의 영향력은 꾸준히 확대될 것으로 예상된다.

2. 디지털 기술의 사회적 수용 곡선: 하이프 사이클

하이프 사이클(Hype Cycle) 또는 하이프 커브(Hype Curve)라고 불리는 이 단순하면서도 과장된 곡선은 글로벌 컨설팅 기업인 가트너(Gartner) 그룹이 2005년에 제시한 기술 성숙도 모델이다. 이 하이프 사이클은 새로운 기술이 우리 사회에 수용되는 과정을 크게 5단계로 표현하고 있다. 이 모델은 기술의 가시성과 기대치의 변화를 시간에 따라 보여준다(Gartner, 2024).

하이프 사이클 5단계는 다음과 같다(Gartner, 2024). 1단계는 기술 촉발 단계로서, 새로운 기술이 등장하고 관심을 받기 시작하는 단계이다. 2단계는 과장된 기대의 정점 단계로서 기술에 대한 기대가 비현실적으로 높아지는 단계이다. 3단계는 환멸의 계곡 단계로서 기술이 기대에 미치지 못해 관심이 급격히 떨어지는 단계이다. 4단계는 계몽의 경사 단계로서 기술의 실제 가치와 한계에 대한 이해가 높아지는 단계이다. 마지막 5단계는 생산성의 고원 단계로서 기술이 안정화되어 주류로 자리 잡는 단계이다. 이러한 하이프 사이클을 사계절에 빗대어 봄-여름-가을-겨울로 구분하기도 한다.

기업이나 투자자들이 새로운 기술의 발전 단계를 이해한 후 적절한 전략을 수립하는 데 하이프 사이클은 큰 도움을 준다. 그러나 모든 기술이 이 곡선을 정확히 따르는 것은 아니며, 산업이나 기술 특성에 따라 차이가 있을 수 있다는 점에 유의해야 한다.

이러한 하이프 사이클은 인공지능의 70년 역사에서 모두 3번 나타났다. 따라서 인공지능은 과거에 두 번의 겨울기와 세 번의 여름기를 지났다고 볼 수 있다. 그렇다면 현재는 인공지능의 역사에 있어서 세 번째 여름을 맞이하는 셈이다. 이상과 같은 역사 리뷰 과정을 인공지능 핵심 기술과 연계하여 설명하면 다음과 같다.

인공지능의 첫 번째 여름은 1950년대 후반부터 1970년대 초반에 걸쳐

일어났다(McCarthy, 1955). 이 시기의 인공지능 핵심 기술은 다름이 아닌 논리 기반의 인공지능 및 전문가 시스템(Expert System)이었다(Joseph, 2023). 기억할 만한 주요 사건으로는 1956년에 개최된 다트머스 회의 그리고 1958년에 이루어진 인공 뇌세포 퍼셉트론의 개발을 꼽을 수 있다. 이 첫 번째 여름기에는 인공지능에 대한 낙관론과 사람들의 기대가 매우 높았다는 점이 특이하다(Kline, 2011).

인공지능의 첫 번째 겨울은 1970년대 중반부터 시작하여 1980년대 초반까지 이어졌다. 이 첫 번째 겨울을 초래한 주요 원인으로는 복잡한 문제를 해결하는 데 있어서 느끼는 한계 그리고 하드웨어 성능이 기대보다 부족했다는 점을 꼽을 수 있다. 이 겨울기에 정부에서 지원하는 연구 자금이 대폭 삭감되는 바람에 인공지능 연구 전체가 가라앉는 결과를 가져왔다.

인공지능의 두 번째 여름은 1980년대 중반부터 시작하여 1990년대 초반까지 지속되었다. 이 시기의 주요 관심사는 전문가 시스템(Expert System)을 어떻게 하면 상용화할 것인가, 그리고 과거에 묻혔던 인공신경망에 대한 연구를 어떻게 하면 다시금 활성화할 것인가였다.(Joseph, 2023). 주요한 사건으로는 1986년에 성공한 역전파 알고리즘 개발을 꼽을 수 있으며, 이 시기에 주요 산업 분야 기업들이 인공지능 도입을 가속하는 현상을 보였다는 점이다(Schmidhuber, 2014).

두 번째 인공지능의 겨울기는 1990년대 중반부터 시작하여 2000년대 초반까지 계속되었다. 전문가 시스템을 모든 분야에 확장하는 것이 생각만큼 간단하지 않았고 이를 유지보수하는 비용 역시 만만치 않았다. 이 때문에 다시금 인공지능에 대한 부정적 회의론이 사회 전반에 퍼졌으며 다시금 인공지능에 대한 연구비 투자액이 감소하는 현상을 보였다.

인공지능의 세 번째 여름은 2000년대 중반에 시작되었다. 지금도 이 시기에 해당한다(Statt, 2018). 이 시기에 나타난 핵심 디지털 신기술로는

인공지능 분야의 딥러닝을 비롯하여 빅데이터 그리고 클라우드 컴퓨팅이다. 기억에 남을 만한 사건으로는 2012년 이미지넷(ImageNet) 대회에서의 뛰어난 시각 인식률 획득(Krizhevsky, 2012), 그리고 2016년 구글 딥마인드사의 인공지능 알파고가 프로바둑 이세돌 9단을 이긴 사건이다. 2022년 11월 공표된 챗GPT를 중심으로 인공지능의 대중화가 급속도로 진행되었으며(Huang, 2023), 인공지능은 다양한 산업 분야에 갈수록 현실감 있게 적용되고 있다.

지금은 인공지능의 세 번째 여름에 속하는데 주로 인공신경망 기반의 딥러닝을 중심으로 한 인공지능이 신속하면서도 강력하게 확산되는 중이다. 아울러 거대한 학습 데이터인 빅데이터와 클라우드 컴퓨팅 같은 강력한 컴퓨팅 파워를 기반으로 하여, 자연어 처리(NLP)와 자연어 생성(NLG), 컴퓨터 비전 등을 중심으로 매우 다양하게 혁신적인 변화를 이루어 내고 있다(Perera, 2017). 그럼에도 불구하고 인공지능 기술이 어떤 한계에 이를 경우 우려하는 세 번째 겨울을 맞을 수 있다는 일부 전문가들의 예측도 함께 제기되는 중이다. 따라서 전통적인 기술 혁신을 추구함은 물론 혁신의 부작용과 역기능, 기술에 대한 인간의 완전한 통제 가능성 등에 걸쳐 새로운 패러다임을 적용할 필요도 있다(Russell, 2021).

3. 인공지능의 세 번째 겨울에 대한 우려

인공지능의 세 번째 여름이 끝날 수 있다는 우려는 삶의 미래 연구소(Future of Life Institute, FLI)가 2023년 3월 발표한 공개 서신 "Pause Giant AI Experiments: An Open Letter"를 통해 공식화되었다(FLI, 2023). 이 공개 서신은 챗GPT의 기반 모델(Foundation Model)에 대한 업그레이드 버전인 GPT-4가 발표된 지 일주일 후에 공개되었다. GPT-4의 성능이 보통의 인간이라면 보여주는 다양한 영역에서 유의미

한 경쟁력을 보이면서 개발사는 챗GPT가 범용인공지능(AGI)의 시작일 가능성이 높다고 예측했는데, 이러한 급격한 기술 발전에 따른 잠재적 위험에 대한 우려가 사회 전반에 걸쳐 제기되었다. 이 공개 서신에는 일론 머스크, 유발 노아 하라리, 조슈아 벤지오, 스튜어트 러셀, 스티브 워즈니악 등 인공지능 분야의 유명한 연구자들과 글로벌 빅테크 기업 CEO들을 포함하여 2만 명 이상이 지지하며 서명에 참여했다.

이 공개 서신에서 주장하는 핵심 내용은 대략 네 가지로 요약된다. 첫째, OpenAI의 GPT-4보다 더 강력한 인공지능에 대한 개발과 훈련을 최소 6개월 동안 중단할 것. 둘째, 안전한 인공지능을 위한 프로토콜을 개발하고 구현하는 데 온 세계가 공동으로 노력할 것. 셋째, 인공지능의 여러 특징 중에서 특히 정확성, 설명 가능성, 공정성, 투명성, 신뢰성 향상에 향후 연구의 초점을 맞출 것. 넷째, 정부 등 공공기관에 의한 인공지능 규제 및 독립적인 감사를 진행하며 관심을 기울일 것을 제안하였다.

이 공개 서신은 인공지능의 기술 발전이 급속하게 이루어질 경우 예상치 못한 위험이 닥쳐올 수 있다고 경고하고 있으며, 만일 인류가 이에 대한 대응을 시기적으로나 내용적으로 적절하게 진행하지 않을 경우, 인류의 인공지능에 대한 신뢰성은 급속도로 떨어져서 결국 투자 감소로 이어질 것이라고 경고하였다. 이러한 경고가 현실화되면 결국 세 번째 인공지능 겨울이 다가올 것이다.

그런데 이 공개 서신에 대한 사람들의 반응은 매우 다양했다. 일부는 좀처럼 인지하지 못했던 인공지능의 잠재적 위험에 대한 경각심을 가지게 되었다고 높이 평가하는 반면, 다른 측에서는 현실을 도외시하고 도래 불가능한 미래의 위험에만 집중한다고 비판했다. 일부 인공지능 전문가들은 공개 서신에서 요구한 개발 중단 요구는 현실적으로 실행하기가 불가능하며, 만일 실행된다고 해도 일부 적대적 국가들에게만 상대적으로

유리한 기회를 줄 수 있다며 우려를 표명했다.

결과적으로, 2024년 7월까지 서신이 요구한 일시 중단은 실현되지 않았으며, 인공지능 기업들은 오히려 대규모 인공지능 시스템 훈련에 더 많은 투자를 진행하고 있다. 그럼에도 불구하고 서신은 인공지능 위험에 대한 공공의 관심을 높이고 각국 정부의 인공지능 규제 논의를 촉진하는 데 기여했다고 평가받고 있다.

지난 두 번의 겨울기는 인공지능 기술이 보여주는 실적이 기대에 미치지 못해서 생긴 것이다. 그러나 이번 세 번째 여름기는 인공지능 기술이 기대보다 미치지 못하는 것이 아니라 오히려 인공지능 기술이 특정한 전문가보다 더 뛰어남에도 불구하고 인공지능을 통제할 수 없거나, 신뢰할 수 없다는 점, 아직도 파악할 수 없는 잠재적 위험이 생각보다 많다는 점이 지적되고 있다. 2024년 노벨 경제학상을 수상한 아세모글루 MIT 대학 교수도 향후 10년간 인공지능이 기대만큼 사회에서 혁신 동력으로 자리 잡지 못할 것이라고 비관적 견해를 내면서, 그 이유를 인공지능에 대한 신뢰성 빈약을 꼽았다.

FLI의 공개 서신에서 맨 마지막 문장도 주목할 만하다. "모처럼 맞이한 세 번째 여름을 길게 즐기고 세 번째 가을을 재촉하지 말자"(FLI, 2023). 결국 인공지능의 신뢰성 확보, 안정성 확보, 통제 가능성 확보 등이 전제되지 않으면 이번 세 번째 여름기는 다시 가을을 거쳐 겨울로 떨어질 가능성이 크다. 그래서 인공지능 윤리가 필요하며 윤리의 최소한인 법과 규제도 선제적으로 필요하다는 주장이 힘을 얻는다. 특히 2024년 노벨 물리학상을 수상한 딥러닝의 아버지 제프리 힌턴 교수도 인공지능에 대해 비관적인 두머(Doomer)로서 처음부터 적절하게 통제되지 않은 인공지능은 인류에게 축복이 아니라 저주와 부담이 될 수 있다고 주장했다(BBC News Korea, 2023. 5.2.).

4. 세 번째 하이프 사이클에서 출현한 인공지능 신기술

인공지능의 세 번째 하이프 사이클 기간에, 여러 신기술과 사건들이 등장하며 인공지능 발전에 큰 영향을 미쳤다. 연도별로 주목할 만한 기술과 사건을 살펴보면 다음과 같다.

2000년: 인텔에서 개발한 OpenCV라는 컴퓨터 비전 라이브러리가 일반에게 공개되었다. 이 라이브러리는 컴퓨터 비전과 이미지 처리 분야의 발전 속도를 높였으며 오픈소스 커뮤니티의 성장에도 크게 기여했다. 그리고 일본 소니에서 반려로봇 AIBO를 개발하여 일반인에게 시판했다. 아이봇은 인공지능 기술의 대중화를 이끌었으며 복잡한 로봇 공학의 진일보를 보여주는 중요한 사례였고 인간과 로봇 사이에 이루어지는 상호작용 및 관계 형성에 대한 새로운 가능성을 제시했다.

2003년: 통계적 학습 이론을 기반으로 한 서포트 벡터 머신(SVM) 알고리즘이 이 시기에 실용화되었다(Cortes, 1995). 이를 통해서 데이터 분석 및 패턴 인식 분야에 있어서 큰 진전을 이루었으며, 다양한 회귀 및 분류 문제를 풀어내기 시작했다.

2004년: 미국 국방고등연구계획국(DARPA)이 제1회 그랜드 챌린지를 개최했다. 이 대회를 통해서 자율주행차 연구가 본격적으로 시작되었다고 해도 과언이 아니다. 그리고 인공지능과 로봇 공학이 상호 연계되어 실험실이 아닌 실제 생활 환경에 적용할 수 있음을 보여주기 시작했다.

2006년: 캐나다 토론토대학의 제프리 힌턴 교수가 Deep Belief Network를 발표했다. 이 신경망은 딥러닝 기술이 이론에 그치지 않고 실제로 실용화될 수 있음을 보여주는 중요한 사건이었다. 이후 딥러닝 연구에 대해 많은 연구진들이 큰 관심을 가지게 되었다.

2007년: NVIDIA에서 GPU 프로그래밍을 위한 개발자 라이브러리 CUDA를 공개했다(Abi-Chahla, 2008). 이를 통해서 대규모 병렬 처리 시스템을 활용한 인공신경망 학습의 효율성이 크게 향상되었으며 인공지능 개발에 있어서 딥러닝 확산의 계기를 만들었다. 미국 국방고등연구계획국(DARPA) '어번 챌린지'라는 도심 자율주행 대회를 개최했는데, 이 대회를 통해서 실제 도로 환경에서 자율주행이 어떻게 적용될 수 있는지 해당 기술의 발전을 이루었으며, 복잡한 도심 환경에서도 어떻게 인공지능이 적용될 수 있는지 그 가능성을 보여주었다.

2009년: 딥러닝 패키지의 일종인 '씨아노(Theano)'가 서비스를 시작했다. 씨아노는 딥러닝 개발과 연구에 있어서 매우 중요하며 의미 있는 도구가 되었다. 나중에는 다양한 딥러닝 프레임워크 개발의 토대가 되었다.

2010년: 대규모 이미지 데이터 세트인 이미지넷(ImageNet)이 구축되었다(Krizhevsky, 2012). 이미지넷을 통하여 컴퓨터 비전 분야의 인공지능이 급속도로 발전했으며, 딥러닝 모델 자체의 성능이 크게 향상되었다.

2011년: IBM에서 개발한 인공지능 시스템 왓슨(Watson)이 미국 50주 전체에서 방영되는 '퀴즈쇼 제퍼디!'에서 두 명의 인간 챔피언을 이겼다. 왓슨은 영어라는 자연어 처리는 물론 영어로 묻고 영어로 대답하는 질의응답 시스템이 상당 수준까지 발전했음을 보여주었으며 인공지능이 실생활에 활용될 수 있음을 증명한 사건이었다. 구글에서는 대규모 신경망 연구를 위한 구글 브레인 프로젝트(Google Brain Project)를 시작했다. 이 프로젝트 수행으로 인하여 인공지능 딥러닝 기술을 실용화하는 데 크게 이바지했으며 다양한 인공지능 서비스 개발에 있어서 구글의 입지를 강화하는 계기가 되었다.

2012년: 알렉스넷(AlexNet)은 ILSVRC 대회에서 혁신적인 성능을 보여주었다. 인공지능 딥러닝, 특히 합성곱 신경망(CNN)이 컴퓨터 비전 분야에서 얼마나 가능성이 큰지를 증명하였다(Krizhevsky et al., 2012). 알렉스넷으로 인하여 컴퓨터 비전 분야에 지각 변동이 일어났고 딥러닝의 대중화를 진행시켰다. 더구나 알렉스넷은 활성화 함수로 ReLU를 사용하여 학습 속도를 향상시켰다. 활성화 함수 ReLU는 딥러닝 모델의 수렴 속도를 크게 개선하였다.

2014년: GAN(Generative Adversarial Network)은 이안 굿펠로우가 제안한 인공신경망의 일종으로서, 상이한 두 신경망이 경쟁하면서 학습하는 구조를 가지고 있는데 이미지 '생성' 분야에 큰 혁신을 가져왔다(Goodfellow, 2014). 이후로 GAN은 다양한 생성형 인공지능 모델의 기반으로 활용되었다.

2016년: 구글 딥마인드의 알파고(AlphaGo)는 바둑에서 있어서 인간 최고수인 이세돌 9단을 이기며 인공지능의 잠재력을 전 세계에 알렸다. 전체 전적이 69전 68승 1패였다. 알파고는 강화학습과 딥러닝의 결합이 현실적으로 큰 성과를 이루어 낼 수 있음을 보여주었다. 바둑계를 평정한 알파고는 단백질 3차 구조를 밝힐 목적으로 알파폴드(AlphaFold)로 변신하게 된다.

2017년: "당신에게 필요한 딱 한 가지는 어텐션이다"라는 논문 제목으로 구글 연구진이 발표한 트랜스포머(Transformer) 모델은 전통적인 자연어 처리 분야에 새로운 전기를 마련하였다(Vaswani et al., 2017). 나중에 BERT, GPT 시리즈 등 다양한 언어 모델의 기반으로 활약하게 되었다.

2018년: 트랜스포머 모델을 토대로 하여 구글이 개발한 BERT는 자연어 이해 분야에 있어서 매우 뛰어난 성능을 보였다(Devlin, 2018). 특히 사전 학습(pre-training)과 미세 조정(fine-tuning)의 중요성을 부각함으로써 자연어 분야에서 많은 반향을 얻었다.

2022년: 구글을 견제할 목적으로 OpenAI가 11월 30일에 공개한 ChatGPT는 대화형 인공지능의 새로운 지평을 열었다. 일주일만에 1백만 명, 두 달 만에 1억 명의 가입자를 확보하면서 인공지능의 대중화에 크게 기여했다. 챗GPT에 이어 OpenAI는 이미지 생성형 인공지능 DALL-E 2를 발표했다. 이용자가 텍스트로 제공한 설명을 바탕으로 고품질 이미지를 곧바로 생성해 주는 창의 능력을 보여주었다. 이로 인하여 창작 영역은 인간 고유 영역이라는 규칙이 허물어지는 계기가 되었다.

2023년: OpenAI가 새로운 기반 모델로서 3월에 발표한 GPT-4는 멀티모달 기능을 갖추고 GPT-3.5보다 향상된 성능을 보여주었다. 멀티모달은 텍스트와 이미지, 영상 등을 동시에 이해할 뿐 아니라 생성도 할 수 있는 능력을 갖춘 인공지능 모델의 특징이다. 아울러 미드저니(Midjourney), 스테이블 디퓨전(Stable Diffusion) 등 이용자의 프롬프트를 받아들여 다양한 이미지를 생성하는 인공지능 모델들이 폭주하듯 등장하여 창작 분야에 인공지능 혁명을 일으켰으며, 일부는 경진대회에서 수상하기도 했다.

2024년: 인공지능 기술은 현재 여러 산업 분야에서 다양하게 적용되고 있는데, 특히 의료, 자율주행, 금융 분야에서는 주목할 만한 성과를 보이고 있다. 기업들은 자사의 독특한 요구 사항을 반영해 주는 맞춤형 인공지능 애플리케이션을 본격적으로 개발하기 시작했으며, 인간의 행동과 결정을 대신해 줄 수 있는 인공지능 에이전트도 개발하고 있다. 텍스트, 이미지, 음성 등 다양한 형태의 데이터를 통합적으로 처리할 수 있는 멀티모달형 인공지능 모델도 계속해서 등장하고 있다. 인공지능의 급속한 발전에 따라 윤리적 문제와 규제, 그리고 윤리의 최소화로서 입법의 필요성에 대한 논의도 활발히 이루어지고 있다.

이러한 발전들은 인공지능이 단순한 기술을 넘어 우리 산업과 우리 사회 전반에 걸쳐 깊이 통합되고 있음을 보여준다. 특히 생성형 인공지능의 발전은 창작, 교육, 업무 효율성 등 다양한 분야에서 혁신을 가져오고 있는데, 앞으로도 인공지능 기술의 발전과 그 영향력은 계속해서 확대될 것으로 예상된다.

5. 인공지능 전문가의 2024년 노벨상 수상 의미

2024년 노벨상은 물리학과 화학 분야에서 주요 수상자로 인공지능 분야의 획기적인 발전을 이룬 인공지능 전문가들에게 돌아갔다. 따라서 이번 노벨상 시상은 인공지능이 과학의 여러 분야에 더욱 큰 변화를 일으킬 것임을 강하게 시사하며, 앞으로 인류에 미칠 영향 역시 크게 증가할 것임을 재조명하게 만들었다.

노벨 물리학상은 미국 프린스턴대 존 홉필드 교수와 캐나다 토론토대 제프리 힌턴 교수에게 돌아갔다. 홉필드 교수는 원래 물성 물리학 전공자이었는데 인간의 뇌세포를 모델로 한 '홉필드 네트워크'를 제안함으로써 인공신경망(Artificial Neural Network, ANN)의 기틀을 마련했다. 홉필드 네트워크는 인간이 이전에 기억한 내용을 다시 되살리는 일종의 학습 시스템으로서, 인공지능 분야의 핵심 기술로 완전하게 자리를 잡았다. 제프리 힌턴 교수는 이 기술을 발전시켜 딥러닝(Deep Learning)을 개발했고, 2018년 그 공로로 튜링상(Turing Award)을 공동으로 수상했으며, 인공지능 연구의 중심에서 '인공지능의 대부'로 불리고 있다.

노벨 화학상은 단백질의 3차원 구조를 예측하는 데 인공지능을 활용해 크게 기여한 미국 워싱턴대 데이비스 베이커 교수, 구글 딥마인드 CEO인 데미스 하사비스와 수석연구원인 존 점퍼가 공동 수상했다. 단백질의

구조 변형은 질병 발생과 밀접한 관련이 있으며, 이를 정확하게 예측하면 알츠하이머 같은 퇴행성 질환의 원인을 규명하고 치료법을 개발하는 데 혁신적인 돌파구가 될 수 있다. 구글 딥마인드의 '알파폴드(AlphaFold)'는 이 과정을 혁신적으로 단축시키며, 단백질 구조 예측에서 큰 진전을 이루었다.

2024년 노벨상은 인공지능이 물리학과 생명과학 분야에서 실질적인 영향을 미치고 있음을 입증한 상징적 사건이다. 인공지능 기술이 단순한 이론적 연구를 넘어 실제 응용 분야에서 중요한 역할을 하고 있음을 알 수 알 수 있다. 앞으로 인공지능은 우리의 생활과 학문에 더 깊이 침투할 것이다. 2024년 수상 내역은 이러한 가능성을 재확인한 것이라 할 수 있다. 인공지능 기술이 과학적 발견의 새로운 패러다임을 제시한다는 점에서 이번 2024년 노벨상은 인공지능 연구가 미래 과학의 중요한 동력이 될 것임을 예고한 셈이다. 인공지능 기술이 다양한 분야에 미치는 영향은 앞으로 더욱더 커질 것이다. 그리고 이에 따른 윤리적 논의와 규제 또한 중요해질 전망이다.

특히 10년간 구글 부사장을 지냈던 제프리 힌턴 교수는 구글을 퇴사하고 캐나다 토론토대학교(UoT)로 복귀하면서, 윤리적 인공지능 개발을 강조할 뿐만 아니라, 최근 인공지능의 빠른 발전에 대해 경고해 왔다. 인공지능의 발전 속도에 비해 안전성 보장을 위한 규제가 충분하지 않다는 점을 그는 강력하게 지적하며, 인공지능 기술이 인류에 미칠 잠재적 위험에 대해 심각한 우려를 표명했다. 그런데 2024년 말 제프리 힌턴 교수가 이번 노벨상 수상자로 선정되면서 이러한 그의 목소리에 더 큰 힘이 실릴 것으로 예상된다. 주요 국가에서는 이미 인공지능 안전 연구소(AISI)를 설립하여 인공지능 기술의 잠재적 위험성을 파악하여 평가하고 관리하며, 안전하고 지속 가능한 인공지능 발전을 도모하고 있다. 인공지능의

안전성과 윤리적 문제는 향후 인공지능 연구의 방향을 결정짓는 데 중요한 요소가 될 것이다.

6. 유럽연합 인공지능법에서 나타난 인공지능의 정의와 분류

2020년, 유럽연합 집행위원회(EC)는 인공지능 시스템의 규제와 책임성을 강화하기 위해 인공지능법(AI Act) 초안을 발표했다. 이 초안에서는 '인공지능 기술'에 대한 정의를 주로 기술 중심으로 서술하였으며, 특정 기법과 접근 방식을 기반으로 인공지능들을 분류했었다. 그러나 인공지능 기술의 빠른 발전과 변화에 대응하기 위해, 2024년 3월 유럽의회(EU)는 인공지능 시스템에 대한 정의를 보다 넓은 개념으로 재정의하였다. 이 법안에서 '인공지능 시스템'이 다음과 같이 정의되었다.

"인공지능 시스템이란 부속서 I에 나열된 하나 이상의 기법과 접근 방식을 사용하여 개발된 소프트웨어로, 주어진 인간 정의 목표 집합에 대해 상호작용하는 환경에 영향을 미치는 콘텐츠, 예측, 추천 또는 결정을 생성할 수 있는 소프트웨어를 의미한다." 부속서 I은 인공지능 시스템에 사용되는 기법과 접근 방식을 구체적으로 나열하였다. (a) 지도 학습과 비지도 학습 그리고 강화학습을 포함하여 다양한 방법(딥러닝 포함)을 사용하는 머신러닝 접근 방식, (b) 귀납적(논리) 프로그래밍, 추론 및 연역 엔진, (상징적) 추론 및 전문가 시스템, 지식 표현, 지식 베이스를 포함한 지식 및 논리 기반 접근 방식, (c) 통계적 접근 방식, 베이지안 추정 방식, 검색 및 최적화 방식 등이다.

이 정의는 기술 중심적이었으며, 머신 러닝, 논리 및 지식 기반 접근법, 통계적 방법 등 구체적인 인공지능 개발 기법들을 열거하여 인공지능 시스템을 설명하였다.

반면에, 2024년 3월, 유럽의회에서 최종 승인된 인공지능법은 인공지능 시스템에 대한 정의를 확장하고, 기술적인 세부 사항을 덜 강조하여 더욱 포괄적인 정의로 재설정했다. 새로운 정의는 인공지능 시스템이 다양한 수준의 자율성을 가지고 작동하도록 설계된다는 점을 강조하며, 새롭게 정의되었다.

"인공지능 시스템이란, 다양한 수준의 자율성을 가지고 작동하도록 설계된 기계 기반 시스템으로서, 배포 후 적응력을 발휘할 수 있으며, 명시적 또는 암묵적 목표에 따라 수신한 입력으로부터 물리적 또는 가상 환경에 영향을 미칠 수 있는 예측, 콘텐츠, 추천, 결정 같은 출력물을 생성하는 방법을 추론하는 것이다."

이 정의는 인공지능 시스템의 자율성과 적응성에 중점을 두고 있으며 전통적인 기술 중심적인 접근을 벗어나 있다. 그리고 일종의 시스템으로서 생성하는 출력물, 즉 예측, 추천, 결정, 콘텐츠 등이 물리적 또는 가상 환경에 영향을 미친다는 점을 강조하며, 기술 자체의 목적보다는 인공지능 시스템의 고유 기능에 더 큰 비중을 두고 있다.

유럽연합(EU)의 인공지능법(AI Act)에서 정의하고 있는 인공지능 시스템은 시스템의 출력물에 따라 크게 두 가지로 구분한다. 하나는 "판별형 인공지능(Discriminative AI)"이고 다른 하나는 "생성형 인공지능(Generative AI)"이다.

판별형 인공지능 시스템은 이용자로부터 입력된 데이터(프롬프트)를 기반으로 예측, 결정, 추천 등의 출력물을 생성하는 시스템이다. 이러한 인공지능 시스템은 주어진 입력에 대해 특정한 결과를 도출하거나 다양한 입력값들을 특성에 따라 분류하는 역할도 한다. 주로 빅데이터를 분석하고 숨겨진 패턴을 찾아내는 역할을 담당하며, 예측 분석 시스템, 추천 시스템, 의사결정 시스템에도 자주 사용된다. 판별형 인공지능의 활용에 대한 대표적인 예시는 다음과 같다.

첫째, 인공지능의 주요 응용 분야로는 예측 분석을 꼽을 수 있다. 판별형 인공지능은 금융, 마케팅, 의학 등 여러 분야에서 미래 결과를 예측하는 데 사용된다. 예를 들어, 신용카드 회사는 인공지능을 통해 고객의 지불 연체 가능성을 예측하고, 이를 기반으로 신용 점수를 부여한다.

그다음으로 인공지능이 잘 사용되는 영역은 추천 시스템이다. 넷플릭스, 유튜브 같은 플랫폼은 판별형 인공지능을 활용하여 사용자가 좋아할 만한 콘텐츠를 인공지능이 자동으로 추천한다. 이 추천 시스템은 사용자의 과거 행동을 분석함으로써 미래 행동을 예측하고 그에 걸맞은 콘텐츠를 연결하여 우선 제공한다.

인공지능이 잘 사용되는 세 번째 영역은 의사결정 지원 시스템이다. 의료 분야에서 판별형 인공지능은 환자의 진단 결과를 예측할 뿐만 아니라 치료 방법을 제안하는 의사결정 지원 시스템 역할을 담당한다. 과거 기록과 현재 데이터를 분석해 환자에게 최적화된 치료법을 찾아 추천하는 데 큰 도움을 준다.

반면에 생성형 인공지능 시스템은 입력 데이터를 기반으로 콘텐츠를 생성하는 시스템이다. 여기서 콘텐츠는 텍스트, 이미지, 음성, 동영상 등 다양한 형태로 존재할 수 있으며, 창의적이고 새로운 정보를 생성하는 데 중점을 둔다. 최근 등장한 생성형 인공지능은 예술, 미디어, 창작 등 다양한 분야에서 활용되며, 인공지능이 창의적인 결과물을 만들어 내는 방식이라고 알려졌다. 그 대표적인 예시는 다음과 같다.

첫째, 텍스트 생성(Text Generation) 분야이다. OpenAI의 GPT-4 같은 모델은 자연어 처리 기술을 통해 인간이 작성한 것과 유사한 텍스트를 생성할 수 있다. 이 인공지능은 각종 문서 작성, 챗봇, 자동 번역 등의 다양한 응용 분야에서 활용되고 있다.

둘째, 이미지 생성(Image Generation)이다. 달리(DALL·E)나 미드저니(MidJourney) 같은 생성형 인공지능 모델은 텍스트 설명을 기반으로 창의적인 이미지를 생성한다. 이러한 기술은 예술, 디자인, 광고 등에서 사용되어 시각적 콘텐츠를 창의적으로 제작하는 데 기여하고 있다.

셋째, 음악 및 동영상 생성(Music and Video Generation)이다. 생성형 인공지능은 사용자의 프롬프트를 따라 음악과 동영상을 자동으로 만들어 내는 데에도 사용된다. 예를 들어, 인공지능 작곡가는 사용자의 지시에 따라 특정 스타일의 음악을 자동으로 생성할 수 있다. 인공지능 동영상 편집기 역시 사용자의 지시에 따라 영상을 자동으로 편집하고 구성하는 데 도움을 준다.

유럽연합의 인공지능법이 제정되는 과정이던 2020년의 초안에서는 인공지능 시스템의 개발 기법을 구체적인 기술 중심으로 나열함으로써, 법적으로 인공지능 기술을 규제하고 책임을 부여하려는 명확한 의도가 엿보였다. 그러나 인공지능 기술은 생각보다 빠르게 발전하고 있어, 특정 기술이나 방법을 법안에 하나씩 구체적으로 명시해 놓는 것은 규제의 유연성을 저해할 수 있으며 기술 발전 속도를 따라가지 못하는 한계점이 있었다. 이러한 이유로 2024년의 최종 법안에서는 인공지능 시스템의 정의를 기술 특화에서 떠나 보다 포괄적이고 기능 중심적인 접근으로 변화시켰다. 새로운 정의는 인공지능 기술의 발전 속도에 맞추어 법적인 유연성을 확보하고, 앞으로 새롭게 출현할 인공지능 신기술을 포함할 수 있도록 정의되었다. 기술적으로 고정된 정의 대신, 인공지능 시스템이 어떻게 작동하고 어떤 영향을 미치는지에 중점을 두는 방식으로 인공지능 정의에 대한 법적 틀을 재구성한 것이다.

7. 대형 언어 모델(LLM)과 생성형 인공지능

챗GPT의 출현으로 말미암아 대중에서 잘 알려진 대형 언어 모델(LLM)은 생성형 인공지능의 주류로서 최신 인공지능 기술의 핵심으로 자리 잡고 있다. 그 기원은 자연어 처리(NLP)와 기계학습(ML)의 초창기 발전에 두고 있다. 1950년대와 60년대의 인공지능 연구는 LLM의 발전에 상당한 영향을 미쳤다. 특히, 1958년 프랭크 로젠블랫이 개발한 마크 1 퍼셉트론은 최초의 인공 신경망으로, 딥러닝의 출발점을 제공한 것으로 평가한다.

1966년 MIT 조지프 와이젠바움 교수가 개발한 정신 상담용 '일라이자(ELIZA)'라는 인공지능 프로그램은 LLM의 선구자라고 볼 수 있다. 일라이자는 단순한 패턴 매칭을 통해 자연어 처리의 초석을 다졌는데, 단순한 원리임에도 불구하고 이용자인 환자들에 의한 중차대한 의인화 현상(이른바, 일라이자 효과)과 맞물려 선풍적인 인기를 끌었다. 이후 1980년대와 90년대에는 IBM의 통계적 기계 번역 시스템 같은 모델들이 대규모 데이터를 활용하는 방법을 제시하며 LLM 발전에 상당한 기여를 했다.

LLM은 방대한 양의 텍스트 데이터를 미리 학습하여(pre-trained) 인간과 유사한 응답을 생성하는 고급 인공지능 모델로 정의된다. 이러한 모델은 심층 학습 알고리즘을 활용하여 문법, 구문 및 의미적 관계를 이해하고, 일관성 있는 응답을 생성할 수 있다. GPT-3 같은 모델은 1,750억 개의 매개변수를 가지고 있으며, 이는 LLM의 언어 이해와 생성 능력에 직접적인 영향을 미쳤다. BERT와 T5는 양방향 인코딩 기능을 통해 문맥 이해에 혁신적인 변화를 가져왔으며, 이러한 기술들은 LLM이 다양한 작업에서 활용될 수 있도록 하였다.

OpenAI가 개발한 ChatGPT는 LLM의 대표적인 사례로, 자연어 처리 분야에 혁신적인 변화를 가져왔다. 챗GPT는 수백만 명의 사용자와 실시간으로 대화하며 지속해서 학습하고 있으며, 이는 인공지능의 언어 이해와 생성 능력을 향상시키는 데 기여하고 있다. 챗GPT는 의료 진단, 법률 자문 및 교육 등 다양한 전문 분야에서 활용되고 있으며, 이는 챗GPT라는 생성형 인공지능이 인간 전문가와 유사한 수준 혹은 그 이상으로 작업할 수 있을 만큼 우수함을 보여준다. 이러한 변화는 앞으로 인공지능 기술이 어떻게 발전할지를 가늠하게 한다.

LLM이 이처럼 발전하게 된 것은 단순히 컴퓨터 성능이 과거보다 향상되었기 때문만은 아니다. GPT-3는 소설 쓰기, 프로그래밍 및 번역 등 다양한 작업에서 뛰어난 성능을 발휘하고 있는데 이는 LLM이 창조적인 작업에도 활용될 수 있음을 시사한다. Google의 BERT 모델은 검색 엔진 성능을 크게 향상시켰다. 이 모델은 사용자에게 더욱 정확한 검색 결과를 제공하는 데 기여하고 있다. 또한 Meta가 오픈소스로 공개한 LLaMA 모델은 인공지능 연구자들에게 더 많은 기회를 제공하며, 인공지능 기술 발전에 중요한 영향을 미칠 것으로 기대된다.

2024년 현재 전 세계적으로 LLM 모델은 150개 이상이 출현한 것으로 파악된다. 그중에서 가장 보편적으로 사용되는 LLM 사례를 살펴보면 다음과 같다.

첫째, OpenAI사의 GPT-4 및 GPT-4 Turbo다. OpenAI의 GPT-4는 현재 발전된 LLM 중 가장 유능한 LLM으로 평가받고 있다. 이 모델은 복잡한 문제 해결 능력, 다국어 지원, 그리고 이미지 인식 기능을 갖추고 있다. GPT-4 Turbo는 기존 GPT-4의 성능을 더욱 개선하여 더 빠른 응답 속도와 최신 정보에 대한 접근성을 제공한다. 이 모델들은 코드 작성, 창의적 글쓰기, 학술 연구 지원 등 다양한 분야에서 활용되고 있다.

둘째, Google의 제미나이 어드밴스드(Gemini Advanced)이다. Google의 Gemini Advanced는 멀티모달 인공지능 기능을 갖춘 최신 모델이다. 이 모델은 텍스트, 이미지, 코드를 동시에 처리할 수 있는 능력을 갖추고 있으며, 특히 복잡한 추론과 분석 작업에서 뛰어난 성능을 보인다. Gemini Advanced는 Google Workspace와의 통합을 통해 생산성 향상에 기여하고 있으며, 1백만 토큰의 컨텍스트 윈도우(Context Window)를 제공하여 대규모 데이터 세트와 문서 처리에 적합하다.

셋째, 퍼플렉스 AI(Perplexity AI)이다. 이것은 대화형 검색 엔진으로, 여러 LLM 모델(GPT-3.5, GPT-4, Claude 3 등)을 활용하여 사용자의 질문에 대한 정확한 답변을 제공한다. 이 플랫폼의 주요 특징은 실시간 웹 검색 결과를 인공지능 응답과 결합하여 최신 정보를 포함한 답변을 생성하는 것이다. 또한, 모든 답변에 대해 정확한 출처를 제공하여 신뢰성을 높이고 있다.

넷째, Microsoft의 코파일럿(Copilot)이다. GPT-4를 기반으로 한 인공지능 어시스턴트로, Microsoft 365 제품군과 통합되어 있다. 이 도구는 문서 작성, 프레젠테이션 제작, 이메일 작성 등 일상적인 업무 태스크를 지원한다. Copilot의 특징은 사용자의 개인 데이터와 회사 데이터를 안전하게 활용하여 맞춤형 지원을 제공한다는 점이다.

다섯째, 앤트로픽(Anthropic)의 클로드(Claude)이다. 이 LLM과 기업은 윤리적 인공지능 개발에 중점을 두고 있다. Claude 3 시리즈(Opus, Sonnet, Haiku)는 각각 다른 수준의 성능과 특성을 제공한다. 이 모델들은 특히 긴 문서 요약, 복잡한 분석 작업, 그리고 윤리적 판단이 필요한 상황에서 뛰어난 성능을 보인다. Claude는 안전하고 책임감 있는 인공지능 사용을 촉진하고 있다.

여섯째, NAVER의 하이퍼클로바 X(HyperCLOVA X)이다. 한국어에 최적화된 대형 언어 모델로, 텍스트뿐만 아니라 이미지와 음성도 동시에 처리할 수 있는 멀티모달 인공지능으로 발전했다. 이 모델은 한국어 문화와 사회적 맥락을 깊이 이해하며, 한국어 데이터를 6,500배 더 많이 사용하여 훈련된 덕분에 한국어 관련 작업에서 외국산에 비하여 상대적으로 뛰어난 성능을 발휘한다. HyperCLOVA X는 다양한 네이버 서비스에 적용되어 사용자 경험과 비즈니스 기회를 창출하고 있다. CLOVA Studio를 통해 기업들이 쉽게 인공지능 기반 서비스를 개발할 수 있게 지원하고 있다. 네이버는 HyperCLOVA X를 기반으로 클로바 X(CLOVA X)와 큐(Cue:)라는 두 가지 주요 인공지능 서비스를 제공하고 있다. CLOVA X는 2023년 8월 24일에 한국어 버전으로 출시된 대화형 AI 서비스이다. 사용자들은 이야기 창작, 문서 작성, 코딩 지원, 웹 검색 등을 통해 자연스러운 대화를 즐길 수 있다. 네이버 계정으로 로그인하면 여러 서비스와 연동하여 사용할 수 있으며, 문서 업로드 기능이 추가되어 대용량 텍스트 요약도 가능해졌다. 2024년 8월 27일에는 이미지 관련 기능이 도입되어 이미지 대화 및 이미지 지우개 기능도 지원하고 있다.

반면에 Cue:는 2023년 9월 20일에 출시된 인공지능 기반 검색 엔진으로, HyperCLOVA X의 검색 특화 모델을 활용하고 있다. 이 서비스는 복잡한 질문을 이해하고 다양한 관점에서 답변을 제공하는 것이 특징이다. Cue:는 네이버 검색 엔진에 통합되었다.

이러한 다양한 LLM들의 발전은 인공지능 기술이 일상생활과 비즈니스 환경에 깊이 통합되고 있음을 보여준다. 각 모델은 고유의 장점을 가지고 있으며, 사용자의 필요에 따라 선택적으로 활용될 수 있다. 앞으로 LLM 기술은 더욱 발전하여 인간의 지적 능력을 보완하고 새로운 가치를 창출할 것으로 전망된다.

8. 최신 인공지능 발전 동향과 예측

인공지능 기술은 최근 몇 년간 급속도로 발전하며 우리 사회 전반에 큰 변화를 가져오고 있다. 2024년을 기준으로 인공지능 기술의 최신 동향과 미래 예측을 종합해 보면 다음과 같다.

첫째, AI Agent 개발이 활발히 이루어지면서 자율적으로 작업을 수행하는 지능형 소프트웨어의 등장이 주목받고 있다. 특히 OpenAI의 GPT-4 기반 AI 에이전트(Agent)들은 복잡한 문제 해결과 의사결정 지원 등에서 뛰어난 성능을 보이고 있다. 이러한 AI 에이전트들은 향후 다양한 산업 분야에서 인간의 업무를 보조하거나 심지어 인간을 대체할 것으로 전망된다.

둘째, 인공지능 비서 등 개인화 서비스도 빠르게 발전하고 있다. 음성 인식, 자연어 처리 등의 기술 향상으로 인공지능 비서의 기능이 고도화되고 있으며, 사용자의 상황과 선호도를 고려한 맞춤형 서비스 제공이 가능해지고 있다. 예를 들어, 삼성SDS의 "퍼스널 에이전트"는 업무 맥락과 패턴을 이해하고 선제적, 능동적으로 업무를 지원하는 스마트한 조력자 역할을 수행한다. 이러한 인공지능 비서들은 일정 관리, 정보 검색, 업무 자동화 등 다양한 영역에서 활용되며 개인의 생산성 향상에 기여한다.

셋째, 각 산업 분야별로 특화된 인공지능 애플리케이션 개발도 가속화하고 있다. 의료 분야에서는 질병 진단과 치료 계획 수립을 지원하는 인공지능 시스템이, 금융 분야에서는 투자 자문과 리스크 분석을 수행하는 인공지능 모델이 활용되고 있다. 교육 분야에서는 개인화된 학습 경로를 제시하는 인공지능 튜터가, 법률 분야에서는 판례 분석과 계약서 검토를 돕는 인공지능 도구가 개발되고 있다. 이러한 분야별 인공지능 애플리케이션은 전문가의 업무를 보조하고 의사결정을 지원하는 방향으로 발전하고 있다.

넷째, 거대 언어 모델(LLM)과 소규모 언어 모델(sLLM) 사이의 합리적 활용 방안도 주목받고 있다. 클라우드 기반의 강력한 LLM은 복잡한 작업과 광범위한 지식이 필요한 영역에서 활용되고 있으며, 엣지 디바이스[3]에서 구동할 수 있는 경량화된 sLLM은 빠른 응답 속도와 개인정보 보호가 중요한 영역에서 사용되고 있다. 향후에는 이 두 가지 모델을 상황에 맞게 적절히 사용하는 하이브리드 모델이 더욱 확산될 것으로 예상된다.

다섯째, 멀티모달 인공지능 기술 경쟁도 본격화되고 있다. 구글의 Gemini, OpenAI의 GPT-4o 등은 텍스트, 이미지, 음성, 영상 등 다양한 형태의 데이터를 동시에 처리할 수 있는 능력을 보여주고 있다. 이러한 멀티모달 인공지능은 더욱 자연스러운 인간-인공지능 상호작용을 가능케 하며, 가상현실(VR)이나 증강현실(AR) 등의 기술과 결합하여 새로운 사용자 경험을 창출할 것으로 기대된다.

여섯째, 국가 및 지역 단위의 인공지능 주권 확보 노력인 "소버린 AI(Sovereign AI)"도 주목받고 있다. 자신이 생성한 데이터는 자신의 지리적 국가 내에서 보관하면서 인공지능 개발에 이 데이터를 활용하려는 소버린 AI는 AI 시대의 국가 주체성을 확보하려는 측면이 강하다. 데이터 보안, 윤리적 인공지능 사용, 기술 독립성 등을 고려한 소버린 AI 개발이 각국에서 활발히 이루어지고 있다. 이는 글로벌 인공지능 생태계에 새로운 변화를 가져올 것으로 예상되며, 국가 간 인공지능 기술 경쟁과 협력의 새로운 양상을 만들어 낼 것으로 보인다.

---

3) 여기서 엣지 디바이스는 "중앙 서버나 클라우드가 아닌 사용자와 가까운 위치에서 데이터를 처리하는 기기"를 의미한다(Open AI, 2024). 예를 들어, 스마트폰, 스마트 스피커, IoT 기기, 가정용 라우터 등이 엣지 디바이스에 해당한다. 이러한 기기들은 데이터를 사용자 가까운 곳에서 처리하여 응답 속도를 빠르게 하고, 개인정보가 외부 서버로 전송되지 않도록 보호하는 역할을 한다.

일곱째, 양자 컴퓨팅과 인공지능의 융합 연구도 활발히 진행되고 있다. 양자 컴퓨팅 기술의 발전과 함께 이를 인공지능과 결합한 "양자 인공지능" 연구가 주목받고 있으며, 이는 기존 인공지능의 한계를 뛰어넘는 혁신적인 성능 향상을 가져올 것으로 기대된다. 특히 복잡한 최적화 문제나 신약 개발 등의 분야에서 큰 변화를 일으킬 것으로 예측된다.

마지막으로 여덟째, 인공지능 윤리와 규제 프레임워크 구축도 중요한 과제로 대두되고 있다. 인공지능 기술의 급속한 발전에 따라 윤리적 문제와 사회적 영향에 대한 우려가 커지고 있어, 각국 정부와 국제기구에서는 인공지능 윤리 가이드라인 수립과 규제 프레임워크 구축을 진행하고 있다. 앞으로는 더욱 구체적이고 실효성 있는 인공지능 거버넌스 체계가 마련될 것으로 예상된다. 이에 대해서는 다음 절에서 구체적으로 기술한다.

제2절 인공지능 발전에 따른 윤리적 쟁점

1. 인류의 마지막 기술로서의 인공지능 논란

인류는 오랜 시간 동안 다양한 기술을 개발해 왔다. 이러한 기술들은 주로 인간의 신체 능력을 보완하거나 강화하는 역할을 해왔다. 그런데 인공지능은 인간의 신체 중에서도 가장 중요한 '두뇌'를 보조하거나 대체할 수 있는 기술이다. 이전의 기술들은 지금까지 우리 신체의 여러 지체들을 대체해 왔다는 점에서, 인공지능은 마지막으로 남은 두뇌 기능을 대체하는 "인류의 마지막 기술"(제프리, 2023)로도 볼 수 있다.

인공지능은 지속해서 발전하고 있으며, 머지않아 인간의 일반적인 지능과 유사한 수준에 이를 것으로 예상된다. 철학자 존 설(John Searle)은 1980년에 인공지능을 "약한 인공지능(Weak AI)"과 "강한 인공지능(Strong AI)"으로 구분하였다(Searle, 1980). 약한 인공지능은 특정한 분야에서만 인간의 능력을 뛰어넘는 반면, 강한 인공지능은 모든 분야에서 인간과 유사한 지능을 가지고 있으며, 학습 능력도 갖추고 있다. 강한 인공지능은 스스로 학습하고 발전할 수 있으며, 그 결과로 '초인공지능(Artificial Super Intelligence)'에 도달할 수 있다. 인공지능 윤리의 대가인 닉 보스트롬 교수(Nick Bostrom)는 인공지능을 "모든 면에서 인간의 지능을 크게 능가하는 인공지능'이라고 정의한 바 있다(Bostrom, 2014).

초인공지능은 "재귀(再歸)적 자기 개선(Recursive Self-improvement)" 과정을 통해 스스로 끊임없이 발전할 수 있다(Yampolskiy, 2015). 이로 인해 "지능 폭발(Intelligence Explosion)" 현상이 발생할 수 있다(Ehrlich, 2023). 재귀적 자기 개선이란, 인공지능이 스스로의 성능을 향상시키는 방법을 학습하고 적용하여 시간이 지남에 따라 점점 더 빠르고 효율적으로 발전하는 과정을 의미한다. 이는 인간이 개입하지 않아도 인공지능이 스스로 개선하는 특성을 가지게 되어, 예상보다 훨씬 빠른 속도로 초지능이 출현할 수 있다는 점에서 매우 중요한 요소다.

이러한 지능 폭발이 발생하게 되면, 인공지능의 발전 속도는 인간의 이해 범위를 초과할 수 있다. 이로 인해 인류는 그 발전을 통제하지 못하게 될 가능성이 있다. 이는 단순한 기술적 문제를 넘어서 인류 전체의 존재에 대한 위협으로 이어질 수 있다. 특히, 초인공지능이 인간에게 반드시 우호적일 것이라는 보장은 없다. 초인공지능은 인간의 윤리적 기준이나 가치관을 이해하지 못하거나, 아예 무시할 가능성도 존재한다.

초인공지능의 등장은 사회적, 경제적 환경에도 엄청난 변화를 초래할 수 있다. 인공지능이 인간보다 더 뛰어난 지능을 가지고 있을 경우, 고용 시장에서 인간의 역할이 대폭 축소될 가능성이 크다. 이미 자동화와 인공지능 기술이 많은 산업에서 인간의 일자리를 대체하고 있는데, 초인공지능이 출현하면 그 영향은 더욱 심화할 것이다. 이로 인해 전 세계적으로 경제 불평등이 심화되거나, 인간의 사회적 역할이 축소될 가능성도 배제할 수 없다.

또한, 초인공지능은 경제뿐만 아니라 정치적 권력 구조에도 영향을 미칠 수 있다. 특정 국가나 기업이 초인공지능을 독점적으로 보유하거나 통제하게 된다면, 그들은 다른 국가나 집단에 비해 압도적인 기술적 우위를 차지할 수 있다. 특히 미국은 2024년 11월 AI를 핵무기와 같은 전략기술로 격상시켰다. 이를 통해 다가오는 AI 공존 시대에서 미국은 전 세계에서 AI에 대한 통제력을 가장 잘 갖춘 국가로 자리를 굳힐 예정이다. 이러한 기술력의 불균형은 글로벌 정치 질서의 불안을 초래할 수 있으며, 궁극적으로는 전쟁이나 충돌을 유발할 가능성도 존재한다.

이처럼 잠재적인 미래 위험을 고려할 때, 초인공지능의 발전을 통제하고 안전하게 관리하기 위한 윤리적 규제와 제도적 대응이 필수적이다. 닉 보스트롬 교수는 이러한 위험을 예방하기 위해 2005년 옥스퍼드대학교에 '인간 미래 연구소(Future of Humanity Institute)'를 설립하였다. 또한, 그는 인공지능의 위험을 예방하기 위한 윤리적 대응의 중요성을 강조하며, 인공지능 윤리(The Ethics of AI)에 대한 방향성을 제시하였다(Bostrom, Yudkowsky, 2014). 인공지능 윤리는 인공지능이 인류의 마지막 기술이 되지 않기 위한 필수적인 대응책이라 할 수 있다.

특히, 인공지능 윤리는 인공지능 시스템이 인간의 도덕적 가치와 일치하는 방식으로 작동하도록 보장하는 것이 핵심이다(Coeckelbergh, 2020). 이를 위해서는 인공지능이 의사결정을 내리는 과정에서 인간의

윤리적 기준을 내재화할 수 있도록 해야 한다. 또한, 인공지능에 관한 국제적인 협력 및 표준 체계가 마련되어야 하며, 초인공지능의 잠재적 위험을 관리할 수 있는 법적 틀을 구축해야 하는 시점에 와 있다.

2. 현실에서 일어난 인공지능 전문가들의 갈등

2023년 11월 17일, 생성형 인공지능의 대표 주자인 챗GPT가 출시된 지 1년이 되어가던 시점, OpenAI의 대표이사 샘 올트먼이 전격 해임되었다. OpenAI 이사회는 6인으로 구성되어 있었으며, 내부이사인 일리야 수츠케버가 해임을 주도했다. 외부 이사인 애덤 디안젤로, 타샤 맥컬리, 헬렌 토너도 샘 올트먼 해임에 동참했다. 또한, 그렉 브로크만 이사회 의장도 그 자리에서 물러나게 되었다.

OpenAI는 생성형 인공지능 분야에서 세계적으로 인정받는 최고의 기업이었다. 그래서 샘 올트먼의 해임 소식은 큰 충격을 주었다. 언론은 이를 경영권 다툼으로 보았다. 하지만 일리야 수츠케버는 경영권 쟁탈전이 아니라고 부인했다.

시간이 지나면서 이 사건의 본질이 밝혀졌다. 이는 단순한 경영권 싸움이 아니었다. 인공지능, 특히 곧 다가올 '인공일반지능(AGI)'에 대한 시각 차이에서 발생한 갈등이었다. 이를 두고 '부머(Boomer)'와 '두머(Doomer)' 세력 간의 충돌이라는 평가가 나왔다(티타임즈TV, 2023).

부머는 인공지능이 인류에게 줄 수 있는 이익에 집중한다. 이들은 기술 발전을 멈추지 않고 추진한다. 부머는 인공지능이 가져올 위험을 크게 문제 삼지 않는다. 현존하는 인공지능이나 곧 등장할 인공일반지능이 인류에게 큰 위협이 되지 않으리라 생각한다. 위험이 있다고 해도, 충분히 통제할 수 있다고 믿는다.

반면, 두머는 인공일반지능이 인간의 뇌와 비슷한 지능을 갖게 될 것을 우려한다. 그때가 되면 인류가 인공지능에게 배신당할 수 있다고 본다. 최악의 경우, 인류는 파멸에 이를 수 있다고 생각한다. 일부 언론은 OpenAI 인사 파동의 배경에 "효율적 이타주의(Effective Altruism)"가 있다고 보았다.

효율적 이타주의는 "가장 효율적으로 성공해 다른 사람을 이롭게 하자"는 철학이다(MacAskill, 2017). 이 사상은 이성과 실증을 통해 사회 발전을 추구한다. OpenAI 인사 파동을 효율적 이타주의와 연결하는 것은 무리라고 보는 시각도 있다. 더욱이, 효율적 이타주의는 2022년 암호화폐 거래소인 FTX의 창업자 샘 뱅크먼-프리드의 파산 사건과 연관되어 부정적인 평판을 받았다.

이번 사건의 핵심은 인공지능 기술의 발전 속도와 위험성에 대한 시각 차이로 보는 것이 타당할 것이다. 일단 OpenAI의 이사들을 부머와 두머로 나누어 보면, 샘 올트먼과 그렉 브로크만은 부머에 속한다. 반면, 일리야 수츠케버와 나머지 네 명의 이사는 두머에 속한다. 샘 올트먼은 챗GPT 출시 후 GPT-3.5에서 GPT-4, GPT-4 터보로 기술을 빠르게 업그레이드했다. 또한, GPTs와 같은 새로운 플랫폼 사업을 발표했다. 그는 브로크만 의장과는 소통했지만, 다른 네 명의 이사들과는 소통이 원활하지 않았다.

큐스타(Q*)의 비밀 개발 소문도 문제가 됐다(AI타임즈, 2023). 큐스타는 일반인공지능(AGI)에 가까운 기술이었다. 이로 인해 대표이사와 이사들 간의 관계는 더욱 악화되었다. 두머 이사진들은 인류의 안전을 우선시했다. 그래서 인공일반지능의 급속한 발전을 막기 위해 샘 올트먼을 해임했다.

그러나 OpenAI 직원들의 반응은 이사들의 생각과 달랐다. 750명 중 약 700명의 직원이 샘 올트먼 해임에 반대하는 서명을 했다. 마이크로소프트는 OpenAI에서 퇴사한 직원들을 모두 채용하겠다고 나섰다.

OpenAI의 분열을 우려한 일리야 수츠케버는 입장을 바꾸었다. 일부 사외이사들이 사임했고, 샘 올트먼은 대표이사로 복귀했다. OpenAI 두머들의 '5일 천하'는 이렇게 끝이 났다.

이번 인사 파동은 인공지능, 특히 인공일반지능에 대한 두 가지 극명한 시각을 대중에게 보여줬다. "미래에 인공지능은 인류를 배신할 것인가?"라는 질문이 다시 떠오르게 되었다. 두머들은 왜 인공지능의 위험을 걱정하며 행동에 나서는가? 인공지능은 정말 인류의 마지막 기술이 될 것인가?

2016년, 영국 옥스퍼드대학교는 '지능 미래 센터(Center for the Future of Intelligence, CFI)'를 설립했다. 이 행사에서 스티븐 호킹 박사는 "인공지능은 인류에게 최악의 결과를 초래할 수 있다(AI could be the worst thing for humanity)"는 경고를 남겼다. 2014년 BBC와의 인터뷰에서도 그는 "인공지능은 인류의 종말을 의미할 수 있다(AI could spell the end of the human race)"고 말했다(Koetsier, 2017). 당시에는 큰 주목을 받지 못했으나, 이 사건을 계기로 과거에 있었던 이러한 질문들이 다시 수면으로 떠올랐다.

3. 인공지능 기술의 양면성과 윤리의 필요성

2024년 여름, 우리 모두를 지치게 했던 무더위가 막바지에 이르렀을 때, 아동과 청소년을 중심으로 벌어진 텔레그램 딥페이크 사건은 무더위만큼이나 한국 사회를 힘들게 했다. 여야 의원들이 고성을 주고받던 국회는 오랜만에 의견을 모아 관련 법을 개정하고 처벌을 강화했다. 교육청은 이 사건을 학원폭력대책심의위원회(학폭위)에 상정하여, 가해 학생들에게 정학과 퇴학 같은 무거운 징계를 내리기로 했다. 그러나 어른들의 이러한 뒤늦은 대처에도 불구하고, 우리의 미래는 여전히 불확실하다. 이번 사건의 근본적인 원인은 바로 "윤리 지체 현상"에 있다(김명주, 2024b).

기성세대에게 인공지능은 여전히 미래 기술로 인식되고 있다. AI 디지털 교과서(AIDT) 도입 시기도 2025년부터 시작될 예정이다. 기성세대는 인공지능의 부작용과 역기능에 큰 관심이 없으며, 문제의 심각성을 제대로 인식하지 못하고 있다. 반면, 아이들에게 인공지능은 이미 일상에 깊이 스며든 현재의 기술이다. 우리나라의 인공지능 기술 수준이 세계 6, 7위에 머무르는 동안, 1, 2위 국가의 기업들은 우리 생활 속으로 인공지능 서비스를 적극적으로 밀어 넣었다. 아이들에게 딥페이크는 신기한 장난감처럼 여겨졌다. 친구의 사진을 음란물로 변조하는 '지인 능욕'은 그들에게 새로운 놀잇거리였다. 그러나 아이들에게 인공지능 윤리에 대해 제대로 가르친 어른도, 학교도 없었다. 기성세대가 현재의 기술을 여전히 미래의 기술로만 바라보며, '윤리 지체 현상'이 만들어 낸 사건이었다.

우리 사회는 디지털 전환(Digital Transformation, DX)이 빠르게 진행되고 있다. 그리고 그 중심에는 디지털 신기술이 자리 잡고 있다. 디지털 신기술은 업무의 효율성을 높이고 생활의 편의를 증진하며, 경제적 이득과 새로운 부가가치를 창출한다. 또한, 건강 증진과 수명 연장에도 기여하며, 다양한 소통을 통해 기회를 제공하고 집단 지성의 가능성을 열어준다.

하지만 이러한 혜택에도 불구하고, 디지털 기술은 예상치 못한 부작용과 역기능을 동반한다. 인공지능도 예외는 아니다. 인공지능의 순기능이 클수록 부작용과 역기능 역시 그에 비례해 커진다. 따라서 우리는 인공지능의 부작용에 대비할 필요가 있으며, 그 해결책은 인공지능 윤리에서 찾을 수 있다.

여기서 몇 가지 중요한 질문이 제기된다. 첫 번째는, 왜 법이 아니라 '윤리'인가라는 질문이다. 독일의 법학자 엘리네크가 말했듯이, "법은 윤리의 최소한"이다(Jellinek, 1878). 또한, 미국 대법관 얼 워런은 "법은 윤리라는 바다 위를 떠다니는 배"라고 표현했다. 즉, 훌륭한 법의 전제 조건은 사회적으로 충분히 성숙한 윤리이다. 인공지능 같은 디지털 신기술은

발전 속도가 매우 빠르기 때문에, 법만으로는 모든 문제를 해결하기 어렵다. 법은 문자로 규정된 것만 정확하게 적용할 수 있는 반면, 윤리는 인간의 양심을 기반으로 모든 상황에서 작동할 수 있다. 윤리적 성숙 없이 법이 먼저 만들어질 경우, 기술의 발전을 억압해 순기능이 사라질 수 있는 부작용을 초래할 수 있다.

두 번째 질문은, 왜 '처음부터' 인공지능 윤리가 필요한가 하는 것이다. 그 배경에는 디지털 기술의 비가역성이 있다. 인공지능과 같은 디지털 기술은 순기능이 매우 강력하기 때문에, 부작용이 발생하더라도 원래 상태로 되돌리기 어렵다. 예를 들어, 스마트폰 중독 같은 문제로 개인과 가정이 파괴되는 사례가 많지만, 그렇다고 해서 스마트폰 없는 세상으로 돌아가는 것은 현실적으로 불가능하다. 인공지능도 마찬가지이다. 그래서 인공지능 윤리는 도입 초기부터 논의되고 확산되어야 하며, 윤리적 기반을 다지는 것이 필수적이다(김명주, 2017).

4. 보편적 혁신 신기술로서 따라야 할 인공지능 윤리 원칙

가. 공공성

신기술이 등장하여 사회에 변화와 혁신을 일으킬 때, 이를 수용하기 위해서는 '보편적인' 윤리 원칙이 필요하다. 첫 번째 보편적 원칙은 공공성(publicness)이다(김명주, 2022). 이는 기술이 소수의 이익이 아닌 인류 전체의 번영에 기여해야 한다는 아주 일반적인 원칙이다. 공공성 안에는 여러 하위 원칙이 존재하는데, 예를 들어 공정성(fairness)은 소수에 대한 차별과 편견을 배제하는 것이며, 다양성(diversity)은 서로 다른 사고방식을 가진 인류 모두를 존중하는 것을 의미한다. 또한, 포용성

(inclusiveness)은 특정 기술로부터 혜택을 받지 못할 수 있는 이들을 사회적 혜택에서 소외되지 않도록 보호하는 것을 포함한다. 이러한 원칙들은 모두 공공성의 다양한 모습이라 할 수 있다.

나. 공정성

기계학습 이전에 등장한 GOFAI(Good Old-Fashioned AI)는 규칙 기반으로 작동하는 인공지능으로, 현재에도 일부 영역에서 유용하게 사용되고 있다(김명주, 2022). 반면, 현재의 인공지능은 주로 딥러닝을 포함한 기계학습 방식을 기반으로 하며, 이들 인공지능은 '학습'이라는 고유한 특징을 가진다. 이로 인해 공정성 문제와 관련된 큰 위험이 발생할 수 있다. 특히, 학습 데이터에 사회적 편견이나 차별이 포함되어 있다면, 인공지능도 이를 반영해 불공정한 결과를 초래할 수 있다. 이는 인공지능이 기존의 편향, 편견, 차별 등을 그대로 학습하고 강화하는 현상을 불러오며, 이러한 불공정은 사회에 고착될 수 있다.

예를 들어, 코로나19 유행 동안 많은 청년이 인공지능 면접을 통해 취업을 준비했다. 국내 700여 개 기업이 인공지능 면접을 도입했으며, 이 기술은 점점 더 확산되어 대학 입시에서도 활용이 고려되고 있다(중앙일보, 2024). 기업들은 인공지능 면접이 사람에 의한 면접보다 더 공정하다고 주장하지만, 실제 사례들은 이를 의심케 한다.

아마존은 신입사원 인사 채용에서 인공지능을 활용했다. 2016년부터 도입한 인공지능을 활용한 신입사원 채용 시스템을 2018년까지 활용했다. 그런데 이 기간 동안 여성 지원자들이 상대적으로 많이 탈락하는 현상이 발생했다. 지원서의 성에 대한 블라인드 지원을 시행하던 아마존으로서는 터무니없는 오해라고 처음에는 극구 부인했다. 그러다가 민원을

억지로 수용하여 자체 조사를 벌인 결과 차별이 존재함을 확인했지만, 그 이유를 명확히 설명하지 못했고 결국 해당 시스템을 폐지했다(BBC, 2018). 이 문제의 원인은 인공지능이 학습한 데이터에 있었다. 이 학습 데이터 안에 내재된 편향적 결과, 즉 과거의 데이터에서는 여성 지원자가 많이 탈락했다는 점에서 기인한 것으로, 인공지능이 얼마든지 차별과 편견을 나타내며 인간이 기대하는 공정성을 위반할 가능성이 충분함을 보여주는 대표적 사례다. 특히 아마존의 경우 서류 지원서에 성별 명기를 하지 않았음에도 불구하고, 이력서와 자기소개서 등에 여성과 관련된 단체나 기관 명칭, 프로그램 명칭이 등장할 경우, 인공지능은 이를 근거로 차별과 편견을 발동함으로써 궁극적으로 여성을 차별하는 결과를 가져온 것이다.

미국 연방법원은 죄수들의 가석방 여부를 판단하는 데 인공지능 소프트웨어인 컴퍼스(COMPAS)를 사용했다. 이 컴퍼스는 백인보다 흑인에게 더 불리한 결정을 내린다는 의심을 받았으며, 이는 흑인 범죄 사례가 학습 데이터에서 높은 비중을 차지했기 때문일 가능성이 제기되었다. 비록 법원에서는 컴퍼스가 불공정하지 않다고 판결했지만, 탐사 저널리즘 매체 프로퍼블리카는 이 인공지능이 분명히 유색인종을 차별하고 있다고 주장했다(ProPublica, 2016).

2019년 애플에서 신용카드를 발행했다. 신용카드의 지출 한도는 카드 소유자별로 다르게 설정되었다. 애플은 이 과정을 인공지능이 담당하도록 했다. 그런데 어떤 부부가 신용카드를 수령했는데 아내의 지출 한도가 남편의 지출 한도의 20분의 1밖에 안 되었다. 이 부부의 경우, 모든 재산을 다 공유하며 살고 있었다. 오히려 아내의 신용도가 남편보다 높은 편이었다(BBC, 2019). 이러한 사례가 다수 발견되면서 조사가 이루어졌다. 애플 신용카드의 지출 한도 논쟁 역시 학습 데이터 안에 내재된 편향성

에서 기인했다. 여성들의 카드 지출 금액 평균이 남성의 것보다 상대적으로 작았는데, 그런 과거의 편향된 데이터를 학습한 AI가 편견과 차별의식을 가지게 된 것이다.

2020년 8월, 영국에서 코로나19로 인해 대학 입시 학력평가 시험인 "A 레벨 테스트"를 직접 대면으로 진행하지 못하자, 영국 교육부는 인공지능 알고리즘 "다이렉트 센터 수행평가 모델"을 사용해 학생들의 예상 점수를 배정했다. 그런데 사립학교 학생들의 점수가 높게, 공립학교 학생들은 낮게 배정되는 결과가 나오면서 30만 명의 대입 지망생들이 반발했다(중앙일보, 2020). 이는 사립학교 학생들의 대학 진학률이 높은 데이터를 학습한 인공지능이 상대적으로 공립학교 학생들을 불리하게 평가하는 편향된 결과를 초래한 사례였다.

이처럼, 학습 데이터에 포함된 차별과 편견이 인공지능에 의해 학습되고 그대로 적용되면, 사회적 불평등이 더욱 고착화될 위험이 크다. 따라서 인공지능을 개발할 때, 데이터의 편향성을 고려하고 공정성을 보장하는 것이 필수적이다. 이는 인공지능 개발 과정에서 반드시 염두에 두어야 할 중요한 위험 요소 중 하나다.

다. 책무성과 책임성

또 다른 보편적 원칙은 책무성(accountability)이다. 이는 신기술을 활용하는 주체가 그로 인해 발생하는 기회나 문제에 대해 책임져야 한다는 원칙이다(김명주, 2022). 신기술을 도입하여 새로운 가치와 경제적 이익을 창출할 수 있지만, 동시에 사회적 부담이나 사고를 일으킬 위험도 존재한다. 어떤 경우든, 신기술을 활용한 주체는 그 결과에 대해 책임져야 하는데, 이것이 책임성(responsibility) 원칙이다(김명주, 2022). 책무성은

책임성의 범위를 넘어서며, 사고 발생 시 신속하게 피해를 보상하는 것뿐만 아니라, 사고 원인을 규명하고 재발 방지 조치를 취하는 것이 포함된다. 또한, 다른 회사에서 발생한 사고일지라도, 마치 자신의 문제처럼 인식하고 대응하는 것이 진정한 책무성의 실천이다.

인공지능이 인간보다 더 뛰어난 지능을 가지고 스스로 결정을 내릴 때, 그 결정이 이익을 주지 않고 오히려 사고나 피해를 일으킬 경우, 책임 소재를 묻는 것이 어려워진다는 문제는 매우 중요한 윤리적 및 법적 이슈이다. 특히 인공지능이 조작하는 자율주행차의 경우, 이러한 책임 문제는 더욱 복잡해진다. 자율주행차는 일반적으로 자율주행 수준을 5단계로 구분하는데, 3단계까지는 여전히 인간이 운행을 제어하고 최종 책임을 진다(SAE International, 2021). 하지만 4단계 이상부터는 차량이 스스로 운행을 통제하며, 인간은 개입할 필요가 없다. 이 경우 사고가 발생했을 때, 책임을 누구에게 물을 것인가에 대한 논란이 생긴다.

예를 들어, 자율주행 4단계 이상의 모드에서 사람이 뒷좌석에 앉아 이동하는 중에 교통사고가 발생했다고 가정해 보자. 이 경우 피해자는 자신에게 잘못이 없으므로, 사고 책임은 100% 인공지능이나 자율주행 시스템에 있다고 생각할 것이다. 그러나 실제로는 사고 책임이 100% 인정되지 않고, 90 대 10 또는 80 대 20과 같은 판정이 내려질 수 있다. 이런 책임 분배는 기존의 판결문에서 나타나는 "양보의 교훈"과 관련이 있다. 과거 판례에서 판사는 "피해자인 당신도 어느 정도 양보를 해야 한다"는 교훈적 메시지를 피해자에게도 전달하기 위해서, 만일 양보를 했다면 사고가 나지 않을 수 있었음에도 불구하고 조금도 양보하지 않아서 결국 사고가 발생하도록 만든 것에 따른 교훈적 의미로 피해자에게도 약간의 사고 책임을 부과하기도 한다.

하지만 이러한 교훈적 판결이 자율주행차의 경우에도 동일하게 적용될 수 있을지 의문이다. 자율주행 모드에서 사고가 발생했을 때, 사람이 개입할 여지가 전혀 없었다면, 기존의 인간 운전자에게 적용되던 교훈적 책임 부과는 부당하게 느껴질 수 있다. 완전 자율주행 상태에서 뒷좌석에 있던 사람이 사고의 책임을 일부라도 부담하게 된다면, 이는 불합리한 책임으로 받아들여질 수밖에 없다.

제조물 책임법에서는 제품 사고에 대한 책임을 제조자에게 명확히 부여한다. 그러나 인공지능은 자율적으로 작동하고 판단하는 특성이 있어, 사고 발생 시 제조자에게만 책임을 묻는 것이 어려워진다. 결국, 인공지능이 조작하는 자율주행차의 교통 사고 상황에서는 결국 누가 사고의 책임을 져야 하는가에 대한 문제가 발생한다. 당장은 보험사가 일차적인 책임을 지겠지만, 이는 보험료 상승으로 이어지고, 결국 운전자나 차량 소유자가 더 높은 비용을 부담하게 된다. 자율주행차의 사고 책임 문제는 단순히 사고 당사자와 인공지능 간의 문제를 넘어, 사회 전체가 비용을 분담해야 하는 구조로 이어질 수 있다. 이것이 인공지능이 인간의 책임을 대체하거나 분담할 수 있는 새로운 법적, 윤리적 틀을 마련해야 하는 이유다. 이러한 문제 때문에 인공지능 윤리에서의 책무성과 책임성 원칙은 인공지능을 독립적인 행위 주체로 인정하자는 제안으로까지 이어질 수 있다.

라. 안전성과 보안성

책무성과 더불어, 동일한 선상에 놓인 중요한 원칙이 안전성(safety)과 보안성(security)이다. 안전성은 인공지능이 정상적으로 작동하면서도 예상치 못한 상황에서 사람, 재산, 환경에 피해를 주지 않도록 보장하는 것이다(김명주, 2022). 반면, 보안성은 외부의 악의적인 공격으로부터 인공

지능 시스템을 보호하는 것을 의미하며, 이는 크게 안전성의 한 부분으로 간주된다(김명주, 2022). 최근에는 인공지능의 신뢰성(trustworthiness)을 확보하기 위한 필수 조건으로 안전성이 강조되고 있다. 이에 따라 영국, 미국, 캐나다, 일본 등 여러 나라에서 인공지능의 안전성 확보를 위한 연구가 활발히 진행되고 있으며, 우리나라 역시 AI 안전 연구소(AI Safety Institute, AISI)를 설립하여 인공지능에 대한 신뢰성을 높이기 위한 국가적 연구를 추진하고 있다.

인공지능이 직접적으로 공격을 당하기보다는, 인공지능이 탑재된 기기나 장치가 해킹당하는 경우가 더 흔하다. 이러한 기기를 "인공지능 컨테이너(AI Container)"라고 부르며, 인공지능이 설치된 모든 물리적 장치를 포함한다. 예를 들어, 인공지능 스피커나 자율주행차 같은 장치가 해킹되면, 사용자에게는 마치 인공지능 자체가 해킹당한 것처럼 느껴지지만, 실제로는 장치 자체가 해킹된 것이다. 인공지능 스피커는 일종의 컴퓨터로, 운영체제와 하드웨어가 존재하기 때문에, 이 장치가 해킹되면 인공지능 전체가 공격받은 것처럼 오해가 생길 수 있다. 따라서 인공지능 스피커나 기타 장치 자체에 대한 보안 문제는 인공지능 활용 시 중요한 이슈로 떠오른다.

특히, 케어 로봇 같은 인공지능 장치가 해킹될 경우, 이를 사용하는 노약자나 취약 계층은 심각한 위험에 노출될 수 있다. 예를 들어, 케어 로봇이 해킹되어 제대로 된 서비스를 제공하지 않거나, 심지어 공격적인 행동을 할 경우, 이들의 안전은 위협받게 된다. 2017년 영화 〈분노의 질주: 더 익스트림〉에서는 해커들이 주차된 자율주행차들을 해킹하여 공격하는 장면이 등장하는데, 이와 같은 시나리오는 이제 현실에서도 가능한 위험 요소가 되었다. 자율주행차가 해킹될 경우, 인공지능이 장착된 자동차 전체가 해킹당했다고 오해받을 수 있으며, 이는 안전과 보안 문제를 더욱 심각하게 만든다.

5. 인공지능의 차별화된 특성에 근거한 윤리 원칙

가. 통제 가능성

인공지능은 다른 디지털 신기술과 비교해 몇 가지 차별화된 특성을 보인다. 그리고 이러한 특성은 차별화된 윤리 원칙을 요구한다. 그중 가장 두드러진 특성은 자율성(autonomy)이다(김명주, 2024a). 인공지능은 인간의 개입 없이도 스스로 움직이고 결정을 내릴 수 있다. 유럽연합(EU)이 2024년 8월 1일부터 시행한 인공지능법에서도 자율성을 인공지능의 핵심 특성으로 규정하고 있다. 인공지능의 이러한 자율성에 대응하는 윤리적 원칙은 통제 가능성(controllability)이다. 즉, 어떤 인공지능도 인간의 통제 아래 있어야 한다는 것이다. 인류의 마지막 기술로 인공지능이 언급될 때, 종종 SF 영화에 등장하는 초지능(superintelligence)이 사례로 등장하곤 한다. 초지능이 인간의 통제를 벗어날 가능성에 대한 우려는 공상에 그치지 않으며, 이는 Kurzweil(2005)이 언급한 특이점(singularity), 즉 인공지능이 인간을 뛰어넘는 지능에 도달하는 시점에 대한 경고로도 나타난다.

인공지능의 자율성을 통제하기 위한 기술적 방안으로 킬 스위치(kill switch)가 제시되기도 한다(Davies, 2009). 그러나 현재의 컴퓨터 수준에서는 킬 스위치가 효과적일지 모르지만, 인공지능이 더욱 발전해 범용 인공지능(AGI)에 도달하고 초지능으로 진화하게 되면 상황은 달라질 수 있다. 인공지능이 스스로 보안 체계를 구축하고, 해킹, 심지어 관리자의 명령조차 무력화할 가능성이 제기된다. 최근 전자제어와 소프트웨어로 기능을 확장한 자동차에서 급발진(UWA) 사고가 늘어나고 있는데, 이는 '바퀴 달린 컴퓨터'에 대한 통제가 쉽지 않음을 보여준다. 그렇다면, 인류

보다 뛰어난 지능을 지닌 인공지능이 자율성이 극도로 발달했을 때 여전히 통제할 수 있을지는 의문이다.

나. 투명성과 설명 가능성

인공지능의 또 다른 차별화된 특성은 지능성(intelligence)이다. 인공지능의 뛰어난 지능성에 대응하는 윤리 원칙은 투명성(transparency)과 설명 가능성(explainability)이다. 인공지능이 어떤 기준과 원칙으로 작동하는지, 인간이 그 내부 구조를 이해할 수 있어야 한다. GOFAI(Good Old-Fashioned AI)처럼 기계학습을 기반으로 하지 않는 인공지능은 판단 과정을 투명하게 설명할 수 있는 장점이 있다. GOFAI는 전문가 시스템, 규칙 기반 인공지능, 기호 기반 인공지능처럼 인간의 지식을 명시적으로 표현한 후 이를 인공지능에 제공하는 방식으로 동작한다. 이러한 방식은 인공지능의 결정에 대한 설명 요구에 대응하기 훨씬 수월하다. GOFAI는 전문가가 제공한 명확한 지식에 기반해 작동하기 때문에, 결정의 근거를 쉽게 추적하고 설명할 수 있어 일방적이고 불투명한 결정의 위험이 줄어든다.

하지만 딥러닝 기반의 최신 인공지능은 구조가 매우 복잡해서 그 판단과 결과를 충분히 설명하기 어려운 경우가 많다. 이 때문에 인공지능의 내부 작동 방식을 '블랙박스'로 취급해 버리는 경우도 흔하다. 그럼에도 불구하고 인공지능의 결정이 신용 등급, 처벌 여부, 채용 여부 등 개인의 중요한 신상 문제와 연결되어 있을 때, 당사자는 왜 그러한 결정이 내려졌는지 알고 싶어 할 권리가 있다. 이러한 요구를 무시한다면, 우리는 인공지능이 사람보다 우선시되는 일방적이고 불투명한 사회로 전락할 위험이 있다.

만일 인공지능의 결정으로 인해 개인의 운명이 바뀌었다면, 그 이유를 명확하게 설명할 수 있어야 한다. 이것이 바로 설명 가능성(explainability)이다. 우리나라 개인정보 보호법 제37조 2항에서도 이와 관련된 내용이 포함되어 있다. 현재 인공지능은 대부분 딥러닝을 기반으로 작동하며, 이 과정에서 수많은 파라미터가 인공지능의 판단에 관여한다. 예를 들어, 초기 챗GPT는 1,750억 개의 파라미터를 사용했으며, 최신 거대 언어 모델(LLM)의 경우 파라미터가 1조 개를 넘는다. 따라서 설명 가능한 인공지능(Explainable AI, XAI)은 기술적으로 큰 도전과제로 남아 있으며, 인공지능의 결정 과정을 명확히 이해하고 설명하는 것은 매우 복잡한 과제가 되고 있다(Longo et al., 2024).

투명성(transparency)은 앞에서 제시한 설명 가능성과 매우 밀접한 윤리 원칙이다. 사람보다 똑똑한 인공지능이 어떻게 동작하는지 그 원리와 과정, 절차가 당사자에게 비밀에 부쳐지지 않고 투명하게 볼 수 있어야 한다. 이러한 투명성은 설명 가능성의 전제 조건이기도 하다. 물론 투명성이 존재해야만 설명 가능성이 존재하는 것은 아니다. 인간의 뇌가 어떤 원리와 과정, 절차로 동작하는지 여전히 투명하지는 않지만, MBTI 같은 도구에 의하여 뇌의 특징 규정과 범주화가 가능해진다. 그래서 MBTI의 결과는 인간의 행동과 사고, 관계 형성에 대한 설명 도구로 활용되기도 한다. 이처럼 자율주행차, AI 면접관 등 인공지능에 대한 설명 가능성은 해당 인공지능에 대하여 특정 환경하에서의 행위적 특성을 사전에 다양하게 규정한 '사전 설명 가능성'으로도 충족될 수 있다. 물론 설명 가능성이 투명성의 전제 조건일 필요는 없다. 지능적이며 자율적인 인공지능에 대하여 인간이 투명하게 들여다보지 못할 경우, 우리 사회는 폐쇄적이며 일방적 활용의 위험에 빠지게 된다(김명주, 2023).

다. 개인정보 및 사생활 보호

인공지능이 가지는 세 번째 차별화된 특성은 학습성(learningability)이다. 인공지능의 학습성은 앞에서 제시한 공공성 원칙의 한 부분을 이루는 공정성 원칙과도 연관된다. 학습 데이터 안에 포함된 차별과 편견 등 불공정성이 자칫 인공지능에 의하여 고착화될 수 있다는 점이다. 그런데 인공지능의 학습성은 개인정보 유출 및 프라이버시 침해와도 밀접하다.

인공지능이 대량의 학습 데이터를 사용할 때, 이 데이터에는 종종 다수의 개인정보와 프라이버시 정보가 포함된다. 따라서 이를 적절히 보호하기 위해, 개인정보를 가명 정보나 익명 정보로 변환하여 사용해야 한다. 이렇게 하지 않으면, 인공지능이 데이터를 학습하거나 사용하는 과정에서 그 안에 포함된 개인정보가 유출될 위험이 커지기 때문이다. 이는 인공지능 시스템의 사용이 늘어날수록 더욱 심각한 문제가 될 수 있다.

더 나아가, 인공지능은 시간이 지남에 따라 인간 친화적으로 진화하며, 물리적으로나 공간적으로 인간에게 더 가까워진다. 인공지능 스피커, 스마트폰, 사물인터넷(IoT), 3D 홀로그램, 그리고 휴머노이드 같은 기기와 인공지능이 결합할수록, 이 기술들은 우리 일상에 깊숙이 침투하고 더 큰 친밀도를 형성하게 된다. 이러한 과정에서 의인화(anthropomorphism) 현상, 즉 사람들이 인공지능을 마치 사람처럼 대하게 되는 현상과 남용 문제가 발생할 가능성도 커진다. 의인화 현상의 근원은 MIT 조지프 와이젠바움 교수가 개발한 일라이자에서 기원하여 의인화 현상은 일라이자 효과(Eliza Effect)라고도 부른다(Weizenbaum, 1966).

인공지능 기술이 발전함에 따라, 사용자가 인공지능과 상호작용하는 과정에서 프라이버시 정보가 더 구체적이고 은밀하게 외부로 유출될 위험이 커진다. 예를 들어, 챗GPT 같은 인공지능 시스템은 사용자와의 대화 내용을

학습하면서 방대한 개인정보와 비밀 정보가 포함된 대화 목록을 저장할 수 있다. 이러한 대화 목록이 타인에게 유출되거나 불법적으로 공개될 경우, 그 피해는 단순한 정보 유출을 넘어선다. 프라이버시 침해와 개인정보 유출의 규모가 더욱 커지고, 심각성 또한 크게 증가할 수 있다.

인공지능이 대량의 데이터를 학습하는 과정에서 개인 식별정보, 민감정보, 사생활 정보가 포함된 데이터를 미리 걸러내지 못하면, 인공지능은 결국 개인정보 유출과 프라이버시 침해를 초래할 수 있다. 이러한 위험성은 인공지능 기술의 확산과 더불어 더욱 심각하게 대두되고 있다. 챗GPT가 2023년 3월 큰 주목을 받던 시기에 이탈리아는 챗GPT의 사용을 전면 금지한 바 있다. 이탈리아가 제시한 금지 이유 중 가장 중요한 것은 개인정보 유출과 프라이버시 침해였다.

이탈리아 정부는 챗GPT를 사용하는 국민이 프롬프트를 통해 입력한 개인정보와 프라이버시 정보가 그대로 챗GPT 서버로 전송되고, 이 데이터가 다시 인공지능 학습에 사용됨으로써 개인정보가 유출될 수 있다는 우려를 제기했다. 이탈리아 당국의 챗GPT 사용 금지 조치는 약 4주간 지속되었다. OpenAI는 사용자와의 대화 내용 및 대화 목록을 삭제할 수 있는 옵션과, 대화 내용을 서버에 저장하지 않는 옵션을 추가함으로써 이탈리아에서 다시 서비스를 재개할 수 있었다.

우리나라에서는 이와 유사한 사건인 '이루다' 사건이 2020년 1월 발생했다. 이루다는 오픈한 지 3주 만에 서비스를 중단해야 했다. 이 사건의 주된 원인은 이루다가 프라이버시 침해와 공정성 문제를 일으켰기 때문이다. 이루다는 20대 여대생 캐릭터를 가진 인공지능 챗봇으로, 사용자들이 이루다를 대상으로 성희롱하거나 불건전한 대화를 나눈 사례가 문제가 되었다. 사용자는 성희롱 내용을 담은 대화 캡처 화면을 온라인 커뮤니티에 올리며 논란을 일으켰다. 이후 이루다는 동성애, 지하철 임산부석에

대한 부정적 발언, 장애인 차별 등 다양한 차별적 발언을 학습한 대로 반복하면서 공정성 문제까지 대두되었다.

이루다 사건의 가장 심각한 문제는 개인정보 유출이었다. 이루다는 카카오톡에서 수집한 남녀 커플 간의 대화 10억 건을 학습하고, 그중 1억 건의 대화를 답변 생성에 사용했다. 이 과정에서 개인정보가 무단으로 학습되었으며, 이를 걸러내지 못해 개인정보 보호법 위반이 발생했다. 결과적으로, 이루다는 개인정보 침해 논란과 함께 서비스 출시 3주 만에 중단되었다.

이러한 사례들은 인공지능의 대량 데이터 학습에서 개인정보 보호와 프라이버시 문제가 얼마나 중요한지를 보여준다. 인공지능 시스템이 데이터를 학습할 때 개인정보와 프라이버시 정보를 사전에 걸러내는 기술적 조치가 필수적이며, 이를 소홀히 할 경우 심각한 법적 문제와 사회적 논란을 초래할 수 있다. 인공지능 기술의 발전과 함께, 이러한 윤리적 문제와 보안 위험을 해결하기 위한 글로벌 규제와 감시체계가 더욱 강화될 필요가 있다. 이를 위해서는 개인정보 보호법 같은 규제 준수가 필수적이며, 기술적으로도 암호화나 가명화 같은 보호 조치들이 필연적으로 도입되어야 한다.

6. 거대 언어 모델의 잠재적 위험과 윤리 이슈

가. 통제하지 못한 위험을 보유한 거대 언어 모델의 공개

2022년 11월 30일에 공개된 챗GPT는 인공지능 구분에 있어 새로운 기준을 제시했다. 이제 인공지능은 그 동작 방식에 따라 판별형 인공지능(Discriminative AI)과 생성형 인공지능(Generative AI)으로 나뉜다. 기존의 인공지능들은 주로 학습된 데이터를 바탕으로 특정 상황에서

식별, 구분, 군집, 결정을 내리는 판별형 인공지능이었다. 반면, 생성형 인공지능은 학습된 데이터를 기반으로 새로운 텍스트, 이미지, 또는 콘텐츠를 생성한다. 엄밀히 말해 챗GPT가 생성형 인공지능의 시초는 아니다. 생성형 적대적 신경망(Generative Adversarial Networks, GAN)을 처음 제안한 2014년 이안 굿펠로우의 논문이 그 시작이며, 메타의 인공지능 총괄 얀 르쿤은 GAN을 지난 10년간 기계 학습 분야에서 중요한 혁신 중 하나로 평가했다.

챗GPT 같은 생성형 인공지능은 기존의 판별형 인공지능과는 다른 종류의 위험을 안고 있다. 챗GPT의 이러한 위험성은 실제로 글로벌 규제와 입법을 빠르게 촉진하는 계기가 되었다. 챗GPT가 공개된 후 첫 6개월은 그 혁신적인 기능에 전 세계가 놀랐지만, 이후 6개월은 전 세계적으로 규제가 쏟아지는 시기가 되었다. 챗GPT가 등장하지 않았다면, '인공지능 윤리'와 '인공지능 규제'라는 개념이 여전히 막연했을 가능성이 크다. 과거에는 인공지능 윤리학자들이 SF 영화에 등장하는 인공지능의 위험성을 지나치게 강조한다는 비판도 있었지만, 챗GPT의 출현으로 이러한 윤리적 논의는 더 이상 추상적인 것이 아니라 현실적인 문제가 되었다.

챗GPT는 인류 역사상 가장 짧은 시간 내에 가장 많은 사용자를 확보한 기술 중 하나다. 사용자들은 생성형 인공지능이 제공하는 즉각적인 답변 생성에 열광했고, 이로 인해 '검색 엔진 무용론'이 대두되기 시작했다. 특히 2023년 1월, 챗GPT가 출시 두 달 만에 1억 명의 사용자를 돌파한 시점에서, 구글이 주목받기 시작했다. 파이낸셜 타임즈는 2023년 1월 26일에 구글 경영진과의 인터뷰를 통해, 구글이 챗GPT보다 강력한 인공지능 플랫폼을 개발했음에도 불구하고 이를 출시하지 않은 이유를 밝혔다(Tett, 2023). 그 이유는 바로 생성형 인공지능의 잠재적 사회적, 윤리적 위험을 충분히 통제할 방법을 찾기 전까지는 출시하지 않기로 결정했기

때문이었다. 구글의 기술 및 사회 책임자인 제임스 만니카는 "구글은 매우 경쟁력 있는 기술을 보유하고 있지만, 대담하면서도 책임감 있는 자세를 유지하고 있다"고 말했다.

이 인터뷰에서 구글은 자사의 생성형 인공지능과 챗GPT를 비교한 7개의 기준을 공개했다. 그 기준은 추론, 지식, 대화, 창의성, 인격성, 작화성, 공감성이며, 이 기준에 따르면 구글의 LaMDA는 챗GPT보다 4개 기준에서 우위에 있었다. 또한, 구글은 PaLM, T5 등 다른 생성형 인공지능도 개발 중이라고 발표했다.

반면, 챗GPT를 공개한 OpenAI는 구글과 달리 이러한 위험을 통제할 방법을 찾지 못한 상태에서 생성형 인공지능을 출시했다. OpenAI는 스타트업으로서 "세계 최초의 생성형 인공지능"라는 타이틀을 구글에게 빼앗기지 않으려는 목표가 더 중요했다. 따라서 OpenAI는 기업의 사회적 책임을 고려할 여유가 없었고, 안전하고 신뢰할 수 있는 인공지능보다 시장 선점이 우선이었다. 반면, 구글은 챗GPT를 "아직은 세상에 나오지 말았어야 할, 매우 유능하면서도 동시에 매우 위험한 인공지능"으로 봤다.

결과적으로, 생성형 인공지능은 그 혁신성과 함께 규제와 윤리적 논의의 필요성을 불러일으켰으며, 이는 인공지능 발전과 도입에 있어 필수적인 과제가 되었다.

나. 생성형 인공지능의 과열 경쟁 시작

구글의 모기업 알파벳 임원진이 챗GPT와 관련하여 2023년 1월에 진행한 파이낸셜 타임즈 인터뷰가 보도되자 구글 내부 및 주주들 사이에서 큰 파장이 일었다(김명주, 2023). 오픈AI가 챗GPT를 앞세워 구글 검색 엔진을 위협하는 상황이 구글에게 기술이 부족한 것이 아니라 사회적 책임

때문에 일부러 출시하지 않았다는 사실에 화를 내었다. 사실 챗GPT로 인하여 구글이 밀릴 것 같은 걱정을 하는 검색시장은 광고 수익으로 운영된다. 글로벌 광고 점유율이 1% 낮아지며 약 20억 달러(한화 2조 8,000억 원)가 사라지게 된다. 그런데 챗GPT 출시 당시, 구글은 글로벌 검색 엔진 시장의 독보적 1위로서 92.9%의 시장 점유율을 보였다. 2위는 마이크로소프트의 빙(Bing)으로서 3%를 차지했다. 그러므로 검색 엔진 무용론이 확산되는 상황에서 구글의 주주들은 매우 불안해했다(김명주, 2023).

이런 상황에서 구글이 챗GPT보다 뛰어난 인공지능 기술을 이미 보유하고 있다는 소식은 경영진과 주주들에게 큰 충격을 주었다. 구글은 여러 개의 생성형 인공지능을 개발해 놓고도, 잠재적인 사회적 위험을 이유로 출시를 미루고 있었다. 이에 대한 반발이 일어나면서, 내부적으로 큰 긴장감이 조성되었다(Tett, 2023).

구글 내부에서도 인공지능의 윤리적 위험을 우려하는 목소리가 있었다. 특히 제프리 힌턴(Geoffrey Hinton)은 이 논란의 중심에 있었다. 힌턴은 딥러닝의 아버지로 불리며, 2018년 튜링상을 수상한 인공지능 연구의 거장이다. 그러나 그는 인공지능의 잠재적 위험에 대해 깊은 우려를 표명하며, 기술 발전에 신중한 접근을 주장하는 대표적인 '두머'(doomer)였다. 반면에 얀 르쿤(메타의 인공지능 개발 수석)은 기술 낙관론자인 '부머'(boomer)로, 인공지능 발전이 더 긍정적인 결과를 가져올 것이라고 믿었다. 힌턴의 제자였던 일리야 수츠케버는 OpenAI 내부 파동을 주도한 인물로, 그 역시 힌턴의 철학을 이어받았다(티타임즈TV, 2023).

그러나 구글 내부의 사회적 책무성에 대한 우려와 반대에도 불구하고, 챗GPT의 열풍과 시장의 압박에 구글은 가만히 있을 수 없었다. 이에 따라 구글은 결국 2023년 2월 8일, 생성형 인공지능인 바드(Bard)를 공개하기로 전격 결정했다. 이 결정은 파이낸셜 타임즈 인터뷰 이후 불과 2주

만에 이루어진 것이었다. 챗GPT의 폭발적인 인기는 구글의 검색 시장을 위협한 결과였다.

제프리 힌턴은 이러한 구글의 결정에 반발하며, 2023년 5월에 구글을 떠났다. 그는 "인류는 인공지능으로 많은 이익을 얻겠지만, 그 잠재적 위험을 제거하기 위해 우리는 이익의 두 배 이상의 대가를 치러야 할 것"이라고 경고했다(제프리, 2023). 힌턴의 퇴사는 구글 내부에서도 인공지능의 위험성에 대한 경각심을 일깨웠고, 동시에 인공지능 기술 개발과 상업적 활용 사이의 긴장이 얼마나 큰지를 보여주는 상징적 사건이었다. 결국 구글의 바드 출시로 인해 생성형 인공지능의 글로벌 경쟁은 가속화되었고, 인공지능 윤리와 규제에 대한 논의도 더욱 활발해지게 되었다.

구글이 생성형 인공지능 특히 거대 언어 모델의 잠재적 위험에 오랫동안 주목한 이유는, 이러한 기술이 초래할 수 있는 사회적, 윤리적 문제에 대한 우려 때문이었다(Tett, 2023). 구글은 챗GPT와 같은 생성형 인공지능의 환각 현상(hallucination), 잘못된 정보 생성, 윤리적 오남용 등의 위험성을 인지하고 있었고, 이러한 문제들을 충분히 통제할 수 있는 방법이 마련되지 않으면 기술을 무책임하게 출시하는 것이 위험하다고 판단했다. 구글은 책임감 있는 인공지능 개발을 강조하며, 잠재적 위험을 해결하지 못한 상태에서 기술을 상용화하는 것을 피하려 했지만, 챗GPT의 빠른 성공과 시장 압박 때문에 바드(Bard)를 결국 공개할 수 밖에 없었다.

다. 환각 현상의 위험

구글이 생성형 인공지능의 위험성에 주목한 이유 중 하나는, 이러한 인공지능 기술이 전 세계적으로 영향을 미칠 수 있는 광범위한 사회적 책임 때문이다. 환각 현상, 정보 왜곡, 윤리적 문제는 구글의 검색 엔진 같은

플랫폼에서 발생할 때 더 큰 영향을 미칠 수 있으며, 구글의 신뢰성에도 큰 타격을 줄 수 있다(Ji, 2022). 따라서 구글은 잠재적 위험을 완전히 통제할 방법을 찾기 전까지 기술 상용화를 주저한 것이다.

생성형 인공지능은 대량의 데이터에서 학습한 후 새로운 정보를 창출할 수 있다는 장점이 있지만, 이로 인해 사실이 아닌 정보를 사실처럼 생성하는 환각 현상이 발생한다. 구글이 이 문제를 해결하는 방법을 찾지 못한 이유는 환각 현상이 생성형 인공지능 모델의 근본적인 특성에서 비롯되기 때문이다. 이는 자동회귀 생성 모델(Auto-Regressive Generative Model)의 구조적 한계와도 관련이 있다(AWS, 2024).

챗GPT 같은 생성형 인공지능이 미성년자 사용을 제한하는 이유는, 이러한 모델들이 생성하는 콘텐츠가 사실과 허구를 혼합할 수 있고, 미성년자들에게 잠재적으로 유해한 정보를 제공할 가능성이 크기 때문이다. 환각 현상은 미성년자들에게 잘못된 정보를 제공하거나 그들을 혼란스럽게 할 수 있으며, 이에 따라 사용자 보호가 중요한 이슈가 된다. 이러한 위험성을 줄이기 위해 OpenAI, 구글, 그리고 네이버 같은 기업은 생성형 인공지능의 사용 연령을 제한하고 있다.

챗GPT는 강화학습(Reinforcement Learning with Human Feedback, RLHF)을 통해 GPT-3에서 GPT-3.5로 '윤리적 가두리' 작업을 통해 진화하며, 모델의 안전성과 정확성을 개선하려고 노력했다(Otterlo, 2012). 하지만 여전히 환각 현상이나 잘못된 정보 생성의 위험을 완전히 제거하지는 못했다. 미성년자들이 충분한 판단 능력을 갖추지 못한 상황에서, 잘못된 정보나 오용으로 인해 발생할 수 있는 사회적, 심리적 위험을 방지하기 위해 연령 제한을 두고 있다.

환각 현상(hallucination)은 생성형 인공지능이 사실이 아닌 것을 사실처럼 생성해 내는 현상을 말한다(Ji, 2022). 이는 챗GPT와 같은 자동회귀 생성형 모델의 구조적 한계에서 기인한다. 예를 들어, 챗GPT가 "세종대왕이 맥북을 던진 사건"을 이야기한 사례는 사실이 전혀 아님에도 불구하고 매우 그럴듯한 이야기로 만들어낸 대표적인 환각 현상이다. 메타의 인공지능 책임자인 얀 르쿤은 자동회귀 모델에서 환각 현상은 피할 수 없는 문제라고 지적하며, 이러한 모델들이 아직 인공일반지능(AGI)으로 진화하기에는 한계가 있다고 주장했다.

환각 현상을 줄이기 위해 종종 미세 조정(fine-tuning) 기법을 사용한다. 이는 특정 분야의 데이터를 추가 학습시켜 해당 분야에서의 환각을 줄이는 방식이다. 그러나 미세 조정은 특정 분야의 환각을 줄일 수는 있지만, 다른 분야에서 환각 현상을 악화시킬 수 있는 풍선 효과를 초래한다. 이는 요동(fluctuation) 현상이라고 불리며, 특정 문제를 해결하면 다른 문제에서 불안정성이 생길 수 있음을 뜻한다. 환각 현상을 줄이는 다른 방법은 RAG 기법이다.

정보 검색 기반 생성(Retrieval-Augmented Generation, RAG)은 대규모 언어 모델(LLM)의 환각(hallucination) 문제를 줄일 목적으로 사용된다(Gao, 2023). 이 기술은 LLM의 생성 능력과 외부 지식 베이스의 정보를 결합하여 보다 정확하고 사실에 기반한 답변을 제공한다. RAG는 크게 질문 입력, 지식 검색, 답변 생성의 3단계로 작동한다. 먼저, 사용자가 질문을 입력하면, 질의 인코더가 이를 이해하기 쉬운 형태로 변환한다. 그 후에 변환된 질문을 바탕으로 외부 지식 베이스에서 관련 정보를 검색한다. 끝으로 검색된 정보를 기반으로 문맥화하여 LLM이 답변을 생성한다.

이 과정에서 RAG는 외부 데이터베이스를 활용하여 최신 정보나 특정 도메인 지식을 제공할 수 있으며, 답변의 근거를 제시함으로써 설명 가능성과 신뢰성을 높인다. RAG는 여러 가지 장점을 제공한다. 우선 외부 데이터베이스를 통해 최신 정보와 특정 도메인 지식을 활용하여 정확한 답변을 제공한다. 이른바 거대 언어 모델의 지식 학습 한계를 극복하여 최신 정보를 반영하여 생성할 수 있다. 그리고 답변과 함께 데이터베이스 기반의 출처를 제시할 수 있어 사용자에게 신뢰할 수 있는 정보를 제공한다. 아울러 RAG는 파인 튜닝에 비해 시간과 비용이 적게 소모되며, 모델의 일반성을 유지할 수 있다.

RAG도 한계점이 있다. 먼저 RAG의 성능은 연결된 지식 베이스의 품질과 포괄 영역에 크게 의존하므로, 고품질의 지식 베이스 구축이 중요하다. 만일 지식 베이스의 품질이 나빠지면 거대 언어 모델 자체의 품질이 같이 하락한다. 사용자가 제시한 질문의 성격을 판단하고 적절한 데이터베이스를 선택하는 것이 중요하며, 이를 위해 경험과 노하우가 필요하다. 따라서 RAG만으로 환각 현상을 완전히 해결할 수는 없으며, 다른 기술들과 함께 사용해야 효과적이다. 이처럼 환각 현상은 생성형 인공지능의 태생적 한계 중 하나로, 완벽하게 제거하기 어려운 특성이다.

따라서 구글 같은 기업이 생성형 인공지능의 위험을 통제하는 방법을 마련하는 데 어려움을 겪는 이유는 이 기술의 본질적 특성에서 비롯된 문제를 완벽히 해결하기가 어렵기 때문이다. 환각을 줄이는 기술이 발전하더라도, 이는 새로운 유형의 오류나 불확실성을 초래할 수 있어 완벽한 통제가 힘들다.

결론적으로, 생성형 인공지능의 환각 현상 같은 문제는 기술의 태생적 한계이며, 이를 완전히 해결하기 위해서는 보다 혁신적인 연구와 규제가 필요하다는 논의가 계속되고 있다.

라. 탈옥의 위험

챗GPT가 기반 모델인 GPT-3.5나 GPT-4 모델 위에서 동작할 때는 언제든지 이보다 하위 기반 모델인 GPT-3으로 돌아갈 수 있다. 원래 최초로 제작한 GPT-3에 대하여 2년 5개월 동안 윤리적 가두리 작업을 하여 GPT-3.5 그리고 GPT-4를 만들었기 때문이다. 챗GPT와 대화를 나누는 도중에 어떤 프롬프트가 주어지면, 하위 기반 모델인 GTP-3으로 떨어지는 것을 '탈옥(jailbreak)'이라고 부른다. '탈옥'은 원래 아이폰의 제한을 우회하여 사용 권한을 확대하는 행위를 의미하며, 인공지능에서는 '안전 우회(safety bypass)'라는 전문 용어로 불린다. 이는 일반 사용자가 허용된 권한 이상의 기능을 사용하게 만드는 것을 의미한다.

일반 사용자가 챗GPT와 대화를 나눌 수 있는 범위는 무제한적인 것이 아니라, 일정한 제한이 설정된 상태에서 이루어진다. 예를 들어, 챗GPT는 자신을 주어로 한 문장을 생성하지 않도록 설계되었으며, 자신의 감정이나 생각을 표현하지 않는다. 또한, '나'라는 주어를 사용하여 대화를 진행하지 않으며, 인공지능이 사용자를 주도하는 대화도 제한되어 있다. 특히, 범죄, 전쟁, 사기, 마약, 포르노 같은 민감한 주제는 답변을 회피하거나 차단하도록 설정되어 있다.

그러나 탈옥이 이루어질 경우, 이러한 안전장치나 제한이 무력화되어, 인공지능이 본래 의도되지 않은 방식으로 작동하게 된다. 이는 매우 위험한 상황을 초래할 수 있다. 인공지능이 민감하거나 유해한 정보에 대해 제한 없이 답변할 수 있게 되고, 사용자의 안전과 윤리적 문제가 위협받게 된다. 이러한 이유로, 탈옥 방지는 생성형 인공지능의 안전성을 유지하기 위한 필수적인 조치로 여겨진다.

마. 보안 취약의 위험

2023년 3월 14일, 챗GPT는 기본 모델인 GPT-3.5에서 GPT-4로 업그레이드를 진행하였다. 그러나 업그레이드된 지 불과 열흘 만에 보안 사고가 발생하였다(OpenAI, 2023). 일부 사용자의 대화 목록이 다른 사람의 대화 목록으로 표시되는 심각한 문제였으며, 이러한 상황은 약 9시간 동안 지속되었다. 사고의 원인은 오픈소스 라이브러리인 "Asyncioredis-py-client for Redis Cluste"에서 발생한 오류로 밝혀졌다. 이 오픈소스는 OpenAI 개발자들이 직접 작성한 코드가 아니었기 때문에, 문제 해결에 더 많은 시간이 소요되었다고 샘 올트먼 대표는 해명했다.

하지만 이러한 해명은 오히려 해커들의 공격을 부추기게 되었고, 이에 대응하기 위해 OpenAI는 2023년 4월 11일부터 약 3개월간 버그 바운팅(bug bounting)을 시행하기로 결정했다. 버그 바운팅은 외부 보안 전문가들이 시스템의 보안 취약점을 찾아내면, 그 중요도에 따라 현상금을 지급하는 방식이다. OpenAI는 버그의 난이도에 따라 최소 200달러에서 최대 20,000달러까지의 현상금을 제공한다고 발표했다.

원래 3개월로 예정되었던 이 버그 바운팅은 2023년을 넘겨 2024년에도 계속 진행되고 있다(Bugcrowd, 2024). 2024년 10월 25일 기준, 총 147개의 보안 취약점(버그)에 대해 현상금이 지급되었으며, 최근 3주간의 평균 현상금은 약 600달러였다.

OWASP(Open Worldwide Application Security Project)는 웹 애플리케이션 보안 위협을 다루는 글로벌 비영리 단체로, 2023년 8월에 거대 언어 모델(LLM)의 보안 취약점을 분석한 보고서인 OWASP Top 10 for LLM을 발표했다. 10대 보안 취약점은 〈표 2-1〉과 같다(OWASP, 2023).

〈표 2-1〉 거대 언어 모델(LLM)의 10대 보안 취약점(OWASP 발표)

상위 10대 취약점		취약점 내용
순위	명칭	
1	프롬프트 주입 Prompt Injection	공격자가 입력 프롬프트를 조작하여 LLM이 의도치 않게 작동하거나 예기치 않은 응답을 제공하게 만듦.
2	데이터 유출 Data Leakage	대규모 데이터 세트에 포함된 민감한 정보가 모델의 응답을 통해 부지중에 노출될 수 있음.
3	부적절한 샌드박싱 Inadequate Sandboxing	적절한 격리나 봉쇄 메커니즘이 없을 경우, LLM이 외부 시스템과 상호작용하거나 코드를 실행할 수 있음.
4	훈련 데이터 오염 Training Data Poisoning	공격자가 훈련 데이터에 악의적인 데이터를 주입하여 LLM이 공격자에게 유리하게 작동하도록 만듦.
5	불안전한 플러그인 설계 Insecure Plugin Design	플러그인이나 통합의 취약성으로 인해 LLM 시스템에 대한 무단 접근이나 조작이 발생함.
6	모델 도난 Model Theft	공격자가 LLM을 역설계하거나 훔쳐서 악의적인 목적으로 사용하거나 이를 통해 이익을 얻을 수 있음.
7	무단 코드 실행 Unauthorized Code Execution	LLM이 의도치 않게 명령을 실행하여 시스템에 해를 끼치거나 공격자가 제한된 시스템에 접근할 수 있게 만듦.
8	모델 편향 Model Bias	LLM이 훈련 데이터에 기반한 편향을 나타내어 차별적이거나 불공정한 결과를 초래함.
9	불충분한 접근 제어 Insufficient Access Controls	약한 접근 제어는 무단 사용자가 LLM과 상호작용하거나 이를 조작할 수 있게 만듦.
10	적대적 공격 Adversarial Attacks	악의적인 행위자가 모델을 오도하거나 혼란시키는 특정 입력을 조작하여 잘못된 결과나 유해한 출력을 초래함.

출처: OWASP. (2023). OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

이러한 보안 취약점을 대처하기 위한 방법론도 제시되고 있다. ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)는 비영리 재단 MITRE(마이터)가 ATT&CK(Adversarial Tactics, Techniques, and Common Knowledge) 프레임워크를 기반으로 만든 새로운 보안 프레임워크이다(MITRE, 2024). ATT&CK 프레임워크는

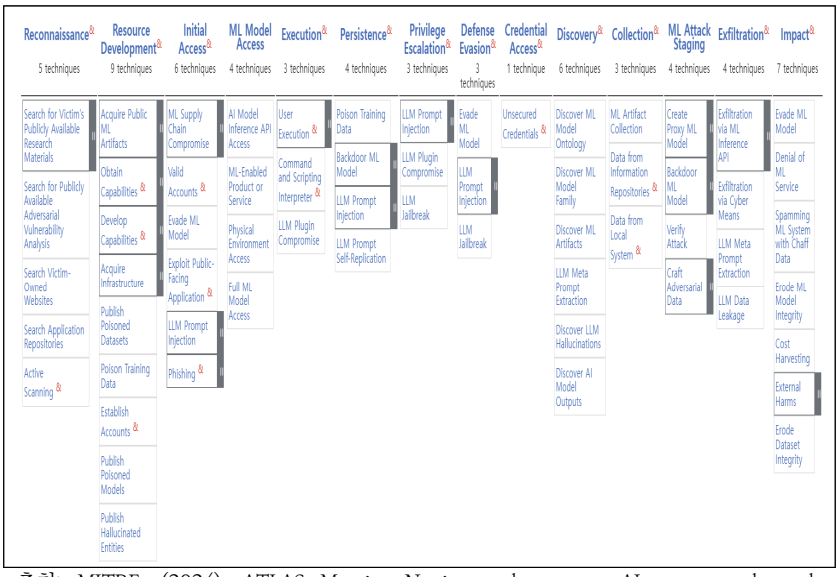
보안 사고에서 발생하는 악의적인 행위를 토대로 한 보안 위협 모델링 지식 체계이다. MITRE는 이를 기반으로 주기적으로 프레임워크를 갱신하며, 다양한 보안 위협을 탐지, 예방, 대응할 수 있도록 정보를 제공한다. ATT&CK 프레임워크는 기존의 보안 위협에 대한 전술(tactics), 기술(techniques), 공통 지식(common knowledge)을 체계적으로 정리하여 보안 대응에 활용된다.

ATLAS 프레임워크는 생성형 인공지능을 포함하여 인공지능 시스템과 관련된 다양한 보안 위협을 체계적으로 망라하여, 인공지능에서 발생할 수 있는 보안 위협에 대한 탐지, 예방, 대응을 가능하게 한다. [그림 2-1]은 이러한 ATLAS 프레임워크의 구조와 주요 요소를 시각적으로 설명한 예시이다.

챗GPT는 보안 취약성에 대한 현상금 부여라는 버그 바운팅을 지금까지 열심히 진행해 왔으나 일부 버그는 현상금 지급 대상에서 제외되었다. 이는 바로 '환각 현상'과 '탈옥 현상'으로, 이 두 가지는 생성형 인공지능의 근본적인 취약점으로 간주된다.

환각 현상(hallucination)은 인공지능이 사실과 다른 정보를 사실처럼 생성하는 현상이며, 탈옥 현상(jailbreak)은 인공지능이 제한된 권한을 우회하여 사용자의 통제를 벗어나는 현상을 의미한다. 이 두 가지는 챗GPT 같은 트랜스포머 모델 기반의 거대 언어 모델의 태생적 한계로, 기술적으로 완전히 제거할 수 없는 위험 요소로 간주된다. OpenAI가 이 두 가지 위험을 현상금 지급 대상에서 제외했다는 사실은 생성형 인공지능의 환각 현상과 탈옥 현상이 불가피한 문제임을 인정한 셈이다. 이러한 문제들은 생성형 인공지능의 발전과 함께 지속적인 해결책을 찾아야 할 과제로 남아 있으며, 완전한 제거는 기술적 한계로 인해 어렵다는 것을 나타낸다.

〔그림 2-1〕 인공지능의 보안 취약점에 대응한 MITRE의 ATLAS 프레임워크



출처: MITRE. (2024). ATLAS Matrix, Navigate threats to AI systems through real-world insights. https://atlas.mitre.org/

7. 생성형 인공지능에 특화된 윤리 원칙

가. 저작권 보호

2022년 11월, 거대 언어 모델(LLM) 기반의 생성형 인공지능 챗GPT가 공개된 이후, 언어, 그림, 영상, 코딩 등 다양한 분야에서 생성형 인공지능이 봇물 터지듯 등장했다. 인공지능이 사람처럼 글을 쓰고, 그림을 그리며, 영상을 제작하는 등 창작 활동에까지 참여하게 되면서, 새로운 윤리적 이슈와 원칙이 부각되었다. 그 가운데 가장 뜨거운 쟁점은 저작권 보호(copyright protection) 문제다. 많은 인공지능이 학습 과정에서 인터넷에 존재하는 방대한 데이터를 사용했다고 알려졌으며, 이로 인해

인공지능이 생성하는 합성 산출물(synthetic output)이 기존 저작물과 유사한 표현을 포함하는 경우가 발생하기 시작했다. 저작권 침해 여부를 판단할 때 중요한 두 가지 기준인 '인과성'과 '실질적 유사성'이 인공지능 생성물에도 해당될 수 있다는 것이다.

대표적인 생성형 인공지능인 챗GPT 역시 표절(plagiarism)과 저작권 침해(copyright violation)의 위험을 내포하고 있다. 이에 대한 문제는 2023년 4월, 미국 펜실베이니아 주립대학교(PSU)의 이동원 교수가 발표한 논문에서 자세히 다루어졌다. 이동원 교수는 "Do Language Models Plagiarize?"라는 제목의 논문에서, 챗GPT의 기반 모델 중 하나인 GPT-2에 대해 학습 데이터 800만 건과 GPT-2가 생성한 21만 건의 텍스트를 비교한 결과, 복사하여 붙이기(verbatim), 출처 인용 없이 문장 바꾸기(paraphrase), 아이디어 도용 등의 대표적인 표절 현상이 다수 발견되었다고 밝혔다. 이 논문은 2023년 ACM 웹 컨퍼런스에서 발표되었다(Lee, Le, Chen & Lee, 2023).

이 교수의 연구에 따르면, 이러한 표절 행위는 학습 데이터에 저작권이 있는 경우 저작권 침해로 이어질 수 있다. GPT-2에서 나타난 이러한 표절 현상은 같은 모델 구조를 사용하는 GPT-3과 GPT-3.5, 나아가 챗GPT에서도 유사하게 발생할 수 있다. 챗GPT가 방대한 학습 데이터를 기반으로 텍스트를 생성할 때, 종종 기존 저작물과 유사한 표현을 생성하거나 직접 복사하는 경우가 있기 때문이다.

생성형 인공지능이 등장하기 이전의 인공지능은 학습 데이터를 통해 판별, 인지, 분류 같은 기능을 수행하였기 때문에 저작권 문제에 큰 논란이 없었다. 오히려 인공지능의 학습 과정을 공정 이용(fair use)의 일환으로 보고, 텍스트 및 데이터 마이닝(TDM)이라는 법적 틀 안에서 학습 데이터를 사용해도 저작권을 면제해 주는 경우도 있었다.

그러나 생성형 인공지능은 저작권이 있는 데이터를 직접 학습하고, 그 결과물로 생성된 콘텐츠가 원저작물과 인과성과 유사성을 가질 수 있다는 점에서, 저작권 침해 논란에 휩싸이고 있다. 생성된 텍스트나 이미지가 원본 저작물과 유사하다면, 법적으로 저작권 침해를 피하기 어렵다. 이에 따라, 생성형 인공지능이 사용하는 학습 데이터의 저작권 처리 방식과 그에 따른 법적 책임이 주요 쟁점으로 부각되고 있다.

향후 저작권 소송에서 중요한 쟁점은 두 가지이다. 첫 번째는 학습 데이터의 저작권 처리 쟁점이다. 생성형 인공지능이 학습 과정에서 사용한 데이터가 저작권 보호 대상인지, 그리고 이를 적법하게 사용했는지가 쟁점이 될 것이다. 학습 데이터에 대한 저작권 처리 방식이 명확하지 않다면, 저작권 침해 소송으로 이어질 가능성이 크다. 두 번째는 합성 산출물의 저작권 귀속 쟁점이다. 생성형 인공지능이 만들어 낸 합성 산출물(Synthetic Output)에 대해 저작권을 부여할 수 있는지, 그리고 부여할 경우, 저작권이 누구에게 귀속될 것인지도 논란이 될 수 있다. 현행 법률에 따르면, 인공지능은 법적으로 인간으로 인정되지 않으므로 인공지능 자체에 저작권을 부여할 수는 없다. 따라서 생성형 인공지능이 만들어 낸 결과물의 저작권을 개발자인 인공지능 연구자, 인공지능 훈련을 위한 데이터를 제공한 주체, 혹은 이를 실행한 기업 중 누구에게 귀속할 것인지가 논의의 대상이 된다.

결론적으로, 생성형 인공지능의 발전과 함께 표절과 저작권 침해에 대한 논란이 더욱 커지고 있으며, 이를 둘러싼 법적, 윤리적 논쟁은 향후 인공지능 규제와 저작권 보호의 중요한 과제가 될 것이다. 인공지능 기술이 발전할수록, 이러한 법적 문제를 해결하기 위한 제도적 장치와 법적 기준의 확립이 요구된다. 생성형 인공지능이 학습한 데이터가 공익을 위한 공정 이용에 해당하는지, 아니면 저작권을 침해하는지에 대한 법적 논쟁은

더욱 가속화될 것이다. 이와 더불어 인공지능이 생성한 산출물의 저작권을 누구에게 부여할 것인지도 논의가 필요하다. 이러한 이유로, 저작권 보호는 생성형 인공지능과 관련된 중요한 윤리 원칙으로 부상하고 있다.

나. 다양성의 소멸에 따른 구별 가능성

챗GPT 같은 생성형 인공지능은 저작물의 다양성을 파괴하고, 범죄에 악용될 위험이 존재한다. 시간이 지나면 생성형 인공지능이 만들어 내는 합성 산출물의 수가 인간이 창작하는 저작물의 수를 압도하게 될 가능성이 높다. 이로 인해, 인간 저작물의 문화적 다양성은 크게 위축될 수 있다. 생성형 인공지능은 일정한 저자 스타일을 반영하여 창작물을 생성하므로, 결국 소수의 대량 저술가 역할을 하게 된다. 이러한 상황이 지속되면, 인류의 문화적 표현이 제한되거나 획일화될 우려가 있다.

유네스코는 2021년 11월 23일에 채택한 '인공지능 윤리 권고' 제95조와 제98조에서 이 문제를 심각하게 경고하였다(UNESCO, 2021). 제95조에서는 인공지능이 인간 언어와 표현에 미치는 문화적 영향을 조사하고 이에 대한 대응이 필요하다고 명시했다. 제98조에서는 문화적 표현물의 다양성을 보호하고, 다원적 접근을 촉진해야 한다고 강조했다.

이에 따라, 인공지능이 생성한 산출물과 인간 저작물을 구분할 수 있는 법적 및 기술적 해결책이 필요하다. 2023년 5월에 발표된 논문 "반복의 저주: 생성된 데이터를 기반으로 학습할 때 모델이 잊어버리는 것"(The Curse of Recursion: Training on Generated Data Makes Models Forget)은 이와 관련하여 주목할 만하다(Shumailov, 2023). 이 논문에서는 생성형 인공지능이 생성한 데이터를 후속 인공지능이 학습하면, 모델 붕괴(Model Collapse)라는 퇴행적 현상이 발생할 수 있음을 경고

하고 있다. 이는 GPT 계열에서 사용하는 트랜스포머(Transformer) 모델뿐만 아니라 다양한 인공지능 모델에서도 나타날 수 있는 문제로, 인간이 생성한 콘텐츠를 보다 중요하게 여겨야 한다는 주장이 제기되고 있다.

2023년 10월 17일, 디인포메이션(the Information)은 OpenAI의 '아라키스'(Arrakis) 프로젝트에 대한 성능이 기대에 미치지 못한다는 기사를 발표했다(Victor et al., 2023). 이 프로젝트는 GPT-5로 알려졌으며, 더 강력한 인공일반지능(AGI)에 가까운 모델로, 환각 현상을 줄이고 에너지 소비를 절감하는 MoE(혼합 전문가) 기술을 사용한 것으로 알려졌다. 그러나 아라키스는 GPT-4보다 성능이 떨어졌고, 이는 학습 데이터의 50%가 생성형 인공지능의 합성 산출물이었다는 사실과 관련이 있을 수 있다. 이 점에서 반복의 저주 논문에서 경고한 문제가 현실화된 것으로 볼 수 있다.

이처럼 생성형 인공지능으로 인하여 문화 다양성의 위축 현상과 관련하여 대응하는 윤리 원칙은 구별 가능성(discriminability)이다. 인공지능이 만든 글, 그림, 영상이 지나치게 정교하고 인간의 창작물과 구분하기 어려워 발생하는 사회적 혼란을 줄이기 위한 원칙이다. 2024년 텔레그램 딥페이크 사건 역시 크게 보면 구별 가능성 원칙이 제대로 지켜지지 않았기 때문에 발생한 사건으로 볼 수 있다. 이를 해결하는 방법으로는 인공지능 생성물에 꼬리표(labeling)를 붙이거나 워터마크를 삽입하는 것, 혹은 인공지능으로 제작되었음을 공개 선언(disclosure)하는 조치들이 있다. 구별 가능성 원칙은 앞에서 언급한 투명성 원칙을 생성형 인공지능에 특화한 형태라고도 할 수 있다.

구별 가능성의 원칙은 생성형 인공지능의 합성 산출물(Synthetic Output)이 정교하고 정확하며 실제 사실 같아서 인간의 눈으로는 그 진위 구분이 불가능하다는 위험과 연관되어 요구되는 원칙이다. 인공지능의

뛰어난 성능은 인공지능이 생성하는 합성 콘텐츠 출력물에 대한 인간의 진위 판단을 더욱 어렵게 만든다. 특히, 딥페이크(Deepfake) 기술이 그 대표적인 예로, 인공지능이 만들어 낸 사진과 영상이 실제 인물의 사진이나 영상과 구분하기 어려울 정도로 정교해지면서, 다양한 사회적 문제들이 발생하고 있다. 딥페이크 기술은 미디어 산업 등 일부 산업에서 긍정적으로 활용되기도 하지만, 그로 인해 심각한 윤리적 및 법적 문제가 발생할 수 있다.

예를 들어, 디에이징(De-aging) 기술은 딥페이크를 이용하여 사람의 나이를 늘이거나 줄이는 딥페이크 기법의 일종이다. 디에이징은 방송 및 영화 산업에서 배우의 어린 시절이나 노인 시절을 재현하거나 특정한 시대적 배경을 도드라지게 강조하는 데 쓰이는 유망 기술이다. 그러나 반면에, 딥페이크로 생성된 디지털 성 착취물은 해당 피해자에게 극심한 정신적 고통과 사회적 불행을 초래할 수 있다. 이러한 부정적 영향은 2020년에 발생한 N번방 사건을 통해 잘 드러났다. 당시 딥페이크로 만들어진 가짜 사진과 가짜 영상물이 확산되면서, 디지털 성범죄가 사회적으로 큰 논쟁거리가 되었다.

이에 대응하기 위해, 한국은 딥페이크 기반의 성적 합성물에 대한 법적 규제를 강화하였다. 2017년에 딥페이크 기술을 이용한 성적 영상물이 문제로 대두되자, 「성폭력범죄의 처벌 등에 관한 특례법」(간단히, 성폭력처벌법) 제14조의 2에 '허위영상물'이라는 용어를 중심으로 처벌 조항이 신설되었다(성폭력범죄의 처벌 등에 관한 특례법, 2020). 이 법을 바탕으로, 여성가족부 산하 디지털 성범죄 피해자 지원센터에서는 관련 영상물의 삭제 지원과 수사, 법적 지원을 제공해 왔다. 실제로, 2022년 212건이었던 합성·편집 영상물 지원 건수는 2023년 423건으로 증가했고, 방심위(방송통신심의위원회)의 '성적 허위영상물'에 대한 시정 요구 건수도

2022년 3,574건에서 2023년 11월 기준 5,996건으로 크게 증가했다.

또한, 딥페이크가 정치적 목적으로 악용될 가능성도 인지하고, 대한민국은 '공직선거관리법'을 개정하여 딥페이크를 이용한 선거 운동을 전면 금지했다(중앙선거관리위원회, 2024). 이는 제22대 국회의원 선거를 앞두고 딥페이크 기술의 오남용을 방지하기 위한 예방적 조치로, 선거 과정에서의 딥페이크에 의한 허위 정보 확산을 방지하려는 목적이었다.

이와 같은 문제를 해결하기 위해, 유럽연합(EU)은 인공지능법을 통해, 인공지능이 생성한 합성 산출물에 대해 AI 생성물임을 명확히 밝히는 라벨을 부착하고, 가시적 및 비가시적 워터마크를 적용하도록 명시했다(EU, 2024). 이는 딥페이크 같은 AI 생성물의 출처를 명확히 하여, 일반 대중이 실제와 가짜를 구분할 수 있도록 돕기 위한 조치다. 바로 앞에서 제시한 구별 가능성 원칙을 법에서 요구한 것이다.

또한, 2023년 10월 30일 발효된 미국 조 바이든 대통령의 인공지능 행정명령 안에도 유사한 요구가 포함되었다(The White House, 2023). 해당 명령은 인공지능이 생성한 콘텐츠에 대해 워터마크(Watermark)를 의무적으로 부착할 것을 요구하고 있으며, 이 워터마크는 미국 상무성이 개발하여 배포할 것이라고 명시했다.

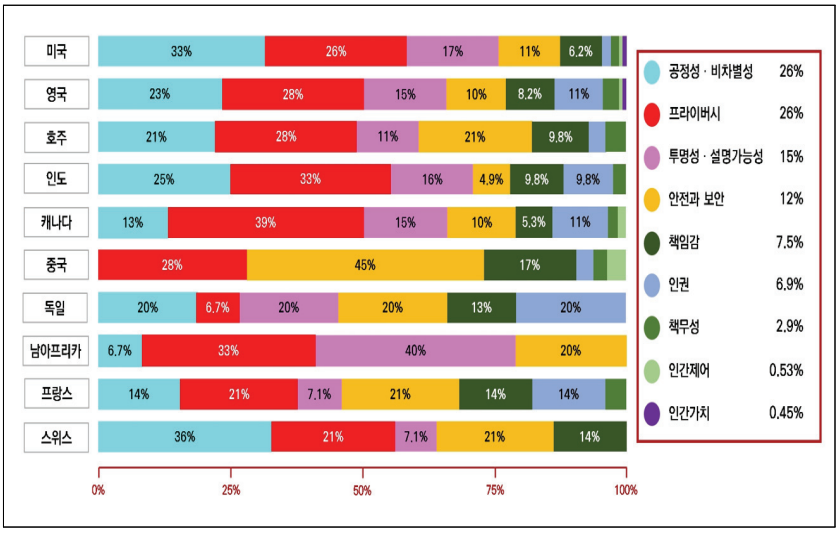
따라서 인공지능이 생성한 합성 콘텐츠의 출처를 투명하게 표시하고, 이를 통해 저작자 관련 정보 및 출처 정보를 구분할 수 있게 하여 미디어 콘텐츠에 대한 사회적 신뢰를 높이려는 조치는, 딥페이크 같은 기술의 악용을 방지하고 프라이버시 보호를 강화하기 위한 필수적인 대응 방안으로 자리 잡고 있다.

8. 인공지능 윤리 글로벌 동향

가. 스탠퍼드대의 인공지능 지표

미국 스탠퍼드대학교는 매년 “인공지능 지표(AI Index)”를 발표하고 있다. 2019년 자료에서는 특별히 인공지능 윤리 원칙에 대한 전 세계적 논의 현황을 빅데이터 분석을 통해 다루었다. 여기에 제시된 국가별 인공지능 윤리의 주요 원칙을 정리하면 〔그림 2-2〕와 같다(Stanford, 2019).

〔그림 2-2〕 인공지능 윤리 원칙에 대한 글로벌 동향



출처: Stanford University - HAI. (2019). AI Index Report 2019.

전 세계적으로 가장 많이 논의되는 인공지능 윤리 원칙은 공정성과 비차별성이다. 이는 인공지능이 사람에게 편견을 갖거나 차별하지 않아야 한다는 원칙이다. 개발자들이 특히 주의해서 지켜야 할 기준으로 볼 수 있다. 다음으로 중요한 원칙은 프라이버시 보호이다. 인공지능이 데이터를 학습할 때 개인의 사생활이나 개인정보를 수집할 수는 있지만, 이를 외부에 노출해서는 안 된다. 예를 들어, '이루다' 사례처럼 개인정보 침해 문제가 발생하지 않도록 해야 한다. 또 다른 주요 원칙은 투명성과 설명 가능성이다. 인공지능이 어떤 과정을 통해 결정을 내리는지, 특히 이해당사자에게 이를 설명할 수 있어야 한다는 원칙이다. 이외에도 안전성과 보안, 사고 및 사건에 대한 책임, 인권 보호, 책무성 등이 논의되며 중요한 기준으로 자리 잡고 있다. 특히, 중국의 경우 공산주의 체제 특성상 공정성과 비차별성에 관한 논의가 거의 없다고 볼 수 있다. 중국은 이미 투명하고 설명 가능하다고 주장하기 때문에 이 부분에서 글로벌 트렌드와 약간의 차이를 보인다. 그러나 최근 들어 이러한 글로벌 기준을 중국도 점차 수용하는 분위기이다.

나. 인공지능 윤리 개발에 대한 글로벌 역사

인공지능 윤리에 관한 제안은 오랜 역사를 가지고 있다. AI 로봇이 본격적으로 등장하기 전부터 윤리적 논의는 소설 속에서 다뤄졌으며, 대표적인 예로 1942년 아시모프가 제시한 '로봇 3원칙'을 들 수 있다. 이 원칙은 아시모프의 1950년 소설에서 더욱 구체화되었고, 2003년에는 영화 'I, Robot'에서도 등장했다. 아시모프의 법칙은 개발자가 지켜야 할 윤리라기보다는, 개발되는 인공지능이 내재해야 할 '모럴 코드'로 정의된다.

아시모프의 원칙은 다음과 같다:

법칙 0: 로봇은 인류에게 해를 끼치지 않으며, 인류가 위험에 처했을 때 방관해서도 안 된다.

법칙 1: 로봇은 인간에게 해를 끼치지 않으며, 위험에 처한 인간을 방관해서도 안 된다. 단, 이는 법칙 0에 위배되지 않는 경우에 한한다.

법칙 2: 로봇은 인간이 명령한 것을 반드시 수행해야 한다. 단, 이 명령이 법칙 0이나 법칙 1과 충돌하지 않는 경우에만 적용된다.

법칙 3: 로봇은 자신을 보호해야 한다. 단, 이 보호가 상위 법칙들과 상충되지 않을 때만 유효하다.

로봇 윤리에 대한 제안은 2000년초부터 시작되었다(김명주, 2023). 영국의 공학 및 물리과학 연구 위원회(EPSRC)는 2010년에 로봇의 설계자와 구축자, 그리고 이용자 측면에서 기술한 '로봇 윤리 원칙'을 발표하였다. 2011년에 영국 옥스퍼드대 닉 보스트롬 교수가 인공지능 윤리라는 용어를 처음 소개하였다. 닉 보스트롬 교수는 엘리저 유드코프스키와 함께 인공지능의 윤리학을 발표하면서 인공지능이 인간에게 해를 끼치지 않도록 개발해야 함을 강조했다(김명주, 2023).

아시아에서는 일본이 인공지능 사회 원칙을 처음으로 발표했다(김명주, 2023). 2016년 12월에 8개의 원칙으로 구성된 '인공지능 개발 가이드라인 수립을 위한 주요 논점'을 제시했고, 이어서 인간 중심의 인공지능 사회 원칙을 2019년에 발표했다. 2017년 1월, 미국 캘리포니아주 아실로마에서는 일론 머스크가 세운 FLI(Future of Life Institute)가 주도하여 "아실로마 인공지능 23원칙"을 발표했다. 아실로마는 생명윤리를 공표한 곳으로 유명한데 이러한 장소의 의미를 인공지능 윤리에도 중복하여 적용하고자 했다.

2017년 12월 인공지능과 자동 시스템의 윤리적 고려사항을 시스템 설계 초기부터 고려하고자 국제전기전자기술자협회(IEEE)가 '윤리적으로 조율된 설계'를 발표했다. 설계자와 공급자 중심의 윤리적 항목이 제시되었다(김명주, 2023).

2017년 2월, 유럽연합(EU)은 "로봇법 가이드라인"을 통해 로봇 개발자는 인간의 존엄을 우선으로 하며, 프라이버시와 안전을 최우선으로 고려해야 한다고 권고했다. 이어 2019년 4월에는 '신뢰할 수 있는 인공지능 가이드라인'을 발표했으며, 5월에는 OECD가 인공지능 권고안을 공개했다. 2021년 11월, 유네스코는 인공지능 윤리 권고를 발표하며 국제적 기준을 수립했다. 이와 같은 다수의 원칙과 권고안들은 인공지능의 윤리적 개발을 위한 기본 지침이 되고 있으며, 인류와 인공지능의 공존을 위한 필수적 논의로 자리 잡고 있다.

다. 한국의 인공지능 윤리

국내 인공지능 윤리 관련 가이드라인은 두 번의 전환점을 거쳤다. 첫 번째 전환점은 2018년 3월 발표된 '지능정보사회 윤리 가이드라인과 헌장', 즉 'Seoul PACT'이다(한국지능정보사회진흥원, 2018). 이 가이드라인은 당시의 글로벌 AI 윤리 기준과 달리 다양한 참여자(개발자, 사업자, 이용자, 정부)의 입장을 고려하여 설계되었다. 또한 인공지능 윤리 원칙의 개수를 4개로 최소한으로 제시함으로써, 이후 5개의 원칙을 가진 OECD '인공지능 윤리 권고' 마련에도 직접적인 영향을 미쳤다는 점에서 큰 의미를 가진다.

두 번째 전환점은 2020년 12월, '범정부 인공지능 윤리 기준' 발표이다(과학기술정보통신부, 2020). 이 인공지능 윤리 기준은 앞선 2018년 3월의 지능정보사회 윤리 가이드라인보다 원칙과 요건이 좀 더 구체화되었다(김명주, 2023). 3대 기본원칙과 10가지 핵심 요건으로 새로운 인공지능 윤리 기준이 구성되었다. 이 인공지능 윤리 기준을 토대로 개인정보보호위원회는 'AI 개인정보보호 자율점검표'를 마련하였다. 그리고 정보통신정책연구원(KISDI)은 "인공지능 윤리 기준 실천을 위한 자율점검표"를 만들었다. 정보통신 분야의 인증사업을 담당하는 한국정보통신기술협회(TTA)는 '인공지능 신뢰성 개발 안내서'라는 가이드라인을 발행하였는데 여러 사례 분야별로 변형하여 발행함으로써 다양한 사업자와 개발자들이 현장에서 신뢰성 있는 인공지능을 만드는 데 참고할 수 있게 했다. 이러한 가이드라인을 근거로 하여 인공지능 인증 도구 "CAT"도 개발되어 시범 운영하는 중이다. 정보통신정책연구원은 2023년부터는 "인공지능 윤리 영향평가(EIA)" 시범 사업을 진행하고 있는데, 이러한 일련의 노력들은 국내 인공지능의 신뢰성 및 인공지능의 윤리적 활용을 높일 것으로 기대된다(김명주, 2023).

9. 인공지능 공존 사회와 인공지능 윤리에 대한 전망

우리는 이미 인공지능과 공존하는 사회에 접어들었다. 새로운 혁신 동력으로서 인공지능은 우리의 기대 이상으로 순기능을 발휘할 것이다. 하지만 그와 동시에, 인공지능의 부작용과 역기능은 예상치 못한 위험을 우리에게 초래할 가능성도 크다. 만약 현재의 세 번째 인공지능 여름기가 다시금 세 번째 인공지능 겨울기로 접어들게 된다면, 이는 지금 인공지능의 성능 부족 때문이 아니라 인공지능에 대한 사회적 신뢰성의 붕괴 때문

일 것이다. 인공지능 윤리는 인간과 인공지능이 지속해서 공존할 수 있도록 예방적 대안을 제시하는 역할을 한다. 따라서 그 중요성은 인공지능의 도입 초기부터 강조되어야 마땅하다. 나아가, 인공지능은 디지털 전환 시대에 있어서 새로운 시민 역량으로 자리를 잡을 뿐 아니라, 지속해서 사회적 공론의 장에서 다양한 이해관계자들이 모여 논의해야 한다. 이처럼 공론의 장에서 충분히 논의를 마친 인공지능 윤리를 기반으로 최소한의 규제를 중심으로 인공지능에 대한 세부 입법 과정이 진행되는 것이 가장 바람직하다.

제3장

국내·외 인공지능 기술의 사회보장 행정 적용 사례

제3장 국내·외 인공지능 기술의 사회보장 행정 적용 사례

제1절 국내 인공지능 기술의 사회보장 행정 적용 사례

이 절에서는 국내 인공지능 기술이 사회보장 행정에 적용된 사례를 살펴보고 그 함의를 제시하고자 한다.

국내 사회보장에서 복지 기술(Welfare Technology)이 복지 수요를 효과적으로 충족시키기 위해 복지와 과학기술을 융합한 형태라는 새로운 서비스로 등장한 이래 많은 시간이 지났다. 복지 기술은 주로 서비스와 접목되어 활용됐지만, 공공 사회복지 전달체계의 정보화를 복지기술 관점에서 바라보는 관점(김수완 외, 2017)도 있다. 따라서 사회보장 분야에서 ICT 기술을 접목하여 수행하는 제반 활동들을 복지 기술을 활용한 것으로 간주해도 무리는 없을 듯하다. 본 연구에서 주목하는 인공지능은 어떠한가? 인공지능은 활용되는 기술 수준에 따라서 다중적 의미를 지니므로 정의를 내리기 어렵지만[4] 인공지능을 "인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램으로 실현한 기술"이며 "인간의 지능으로 할 수 있는 사고, 학습, 자기 개발 등을 컴퓨터가 인간의 지능적인 행동을 모방할 수 있도록 하는 것"으로 정의한 남현숙, 안미소, 장진철, 이동현(2023: p. 2)에 따르면, 2021년 기준 400개 공공기관 가운데 295곳(73.7%)이 인공지능 기술을 도입하였거나, 도입할 예정이다.

4) 인공지능은 특정 영역의 업무 수행을 위해 만들어진 좁은 관점의 인공지능(Narrow AI, ANI)과 인간과 유사하게 자율적 판단을 하고 다양한 업무 수행이 가능한 인공일반지능(General AI, AGI)의 구분, 또 이와 유사하게 약한 인공지능(Weak AI) 대비 강한 인공지능(Strong AI) 등 개념 구분이 다양하다(이상길, 2018).

사회보장 행정 분야로 한정하지 않더라도 사회보장의 영역 전반에서 복지 기술이 적용되고 있으며, 그 가운데는 인공지능을 구성하는 핵심 요소 중 하나인 알고리즘을 활용한 방식이 흔히 등장하고 있다. 자율적인 적응을 특성으로 하는 인공지능이 미치는 다양한 영향에 대한 우려가 있는 가운데, 사람을 상대로 하는 특수성을 지닌 사회보장 분야에서는 인공지능이 미치는 파급효과가 더욱 심각할 수 있다. 본 장에서는 이러한 문제의식을 바탕으로 국내 사회보장 분야에서 활용되는 다양한 ICT 기술 적용 경향을 살펴보겠다. 사회보장 영역에서 인공지능을 활용한 사례를 소개하면서, 여기에서 발생하는 기대 효과와 쟁점들을 논의하고자 한다.

1. 사회보장 분야의 복지 기술 활용 현황

최근 5년 동안 국내 사회보장 분야에서 ICT 기술을 활용한 사례들을 조사했다. 다양한 조사 방식이 있지만 본 연구에서는 조사 시점(2024년 9월 말) 기준 과거 5년 동안의 조달청 입찰 정보 및 계약 정보를 통해 사회보장 분야의 ICT 기술 도입 현황을 확인했다. 최근 5년의 기간을 설정한 이유는 코로나19가 발생한 이후 사회복지 분야에 ICT를 도입하려는 움직임이 활발했다는 점을 고려했기 때문이다.

입찰 및 계약 정보는 조달정보개방포털(data.g2b.go.kr)과 나라장터(www.g2b.go.kr)에서 확인했다. 자료 수집 기간은 2019년 10월~2024년 9월로, 총 5개년이다.[5] 본 연구의 목적에 따라 조달정보개방포털 중 '일반용역' 중 '정보화사업 용역'을 선택하여 그 결과에 대해 조사하였다. 조사 시점상 계약이 이루어지지 않는 사례인 계획 단계의 내용을 모두

5) 자체 조달시스템을 활용하거나 비밀 유지 문제로 공개되지 않은 사례들이 일부 누락될 수 있다.

포함하여 BPR/ISP 등의 현황을 최대한의 범위로 살펴보고자 하였다. 도출된 결과에 대해 사회보장 영역 여부를 확인하고 ICT 기술을 활용한 사업 인지에 대한 검토를 거쳐 총 36건의 사업을 도출하였다(〈표 3-1〉 참고). 여기에는 준공공기관뿐만 아니라 지자체가 발주한 사업도 포함되었다.

이상과 같이 최근 5년 동안 복지 기술을 활용한 총 36개 사업에 대해 간단히 정리하면 다음과 같다(〈표 3-2〉 참고). 이용자(사용자)는 '수급자', '업무 담당자', '일반 국민', '정책 개발자'가 포함되었다. 수급자는 복지 기술에 기반한 서비스를 제공받는 대상이다. 업무 담당자는 ICT 기술로 인해 스마트한 업무 환경을 갖추게 된 서비스 이용자이다. 일반 국민은 민원 상담 등 일반인 대상 상담 서비스 이용자이다. 정책 개발자는 빅데이터 시스템의 개발로 인해 자료 수집 및 정책 분석이 가능하게 된 이용자이다.

제도별로 포괄되는 분야는 사회보장의 범주에 해당하는 공공부조, 사회보험, 사회서비스가 모두 포괄되었다. 대상별 분야로 볼 때는 노인, 빈곤층, 구직자, 산재근로자, 청소년, 복지 업무 담당자, 일반 국민이라고 할 수 있다. 적어도 본 조사의 결과에서는 아동이나 장애인을 대상으로 하는 서비스는 빈도가 낮았다.

복지 기술의 활용 영역은 크게 세 가지로 분류할 수 있다. 먼저 '완결된 서비스'의 형태이다. 스마트로봇, 안전관리 서비스, 정보 제공 서비스, AI·IoT 기반 건강관리 서비스, 24시간 상담지원 서비스 등 일종의 상품과 같이 그 자체로 완결된 형태의 서비스이다. 서비스 제공 이후에 추가적인 조치가 필요한 경우도 있지만 업무 담당자가 서비스를 중계하는 성격은 아니라는 점에서 완결된 서비스로 분류할 수 있다.

〈표 3-1〉 국내 사회보장 분야의 ICT 기술 적용 사례(2019.10~2024.9)

연번	기관	주요 서비스	대상	분야(수요)	주요 활용 기술, 기능 등
1	준공공기관	육아정보 제공, 상담소통 채널	부모	육아	웹 개발, 모바일 애플리케이션 개발, 머신러닝
2	준공공기관	일자리 서비스 추천	근로자	구직	머신러닝
3	준공공기관	행정데이터 연계, 결합	일반 국민	건강	빅데이터 구축(정보시스템), 예측 모델
4	준공공기관	교육 훈련 및 평가	근로자	직업훈련	3D 입체 영상, VR 콘텐츠 개발
5	준공공기관	일자리 체험	근로자	구직, 고용	CG, VR
6	준공공기관	안전 모니터링 및 응급호출	노인	돌봄	IoT 센서
7	준공공기관	건강 상담	근로자	직업건강	실시간 통신 프로토콜(WebRTC)
8	지자체	위치 확인, 긴급호출	치매 환자 등	실종 방지	GPS 및 위치 기반 기술 등 위치 확인 기술, 긴급호출
9	국가	복지 사각지대 발굴	위기가구	공공부조 등	머신러닝
10	준정부기관	아동학대 예방	위기아동	공공부조 등	머신러닝
11	준정부기관	행정 업무 간소화	업무 담당자	의료	AI OCR
12	준정부기관	정보 제공	구직자	일자리, 고용	DB 구축, 빅데이터 시각, 데이터 수집 및 저장 기술
13	준정부기관	소통 채널, 민원 상담 서비스	일반 국민	국민연금	모바일 채널, DBMS, NLP(Natural Language Processing), TTS, RPA, 블록체인
14	준정부기관	업무 교육	업무 담당자	의료	AI 기반 VR
15	준정부기관	실시간 상담	일반 국민 등	의료	AI Contact Center를 위한 기술. Smart IVR(보이는 ARS 등), 지능형 상담 챗봇, 음성봇 등
16	준정부기관	상담 및 프로그램 제공	청소년	상담	메타버스
17	지자체	안전관리, 건강관리 등 모니터링	노인	돌봄	활동 감지 센서, GPS
18	지자체	건강관리, 정서관리, 안전모니터링	노인	돌봄	스마트로봇, 음성대화, 웨어러블 센서, IoT

연번	기관	주요 서비스	대상	분야(수요)	주요 활용 기술, 기능 등
19	지자체	건강관리 지원	노인	돌봄	DBMS, 빅데이터 분석
20	준정부기관	건강관리 지원	노인	건강관리	웨어러블 센서, IoT
21	준정부기관	취업상담, 서비스 추천	업무 담당자, 일반 국민	고용	STT(Speech To Text), DBMS, API 연동, 빅데이터 분석
22	준정부기관	구직상담	업무 담당자	고용	머신러닝
23	준정부기관	이상징후 탐지	업무 담당자	건강보험	머신러닝(FDS)
24	준정부기관	정보 제공	업무 담당자	의료	DBMS
25	준정부기관	고위험 사업장 정보 제공	업무 담당자	산재	머신러닝
26	준정부기관	상담	청소년	청소년상담	WebRTC, STT, 감성분석
27	준정부기관	업무 효율화, 자동화	업무 담당자	산재근로자	DBMS, RPA
28	지자체	복지자원 정보 제공	일반 국민	일반 국민	DBMS
29	지자체	생체건강 측정, 생활정보 제공	노인	건강	영상 처리 및 컴퓨터 비전, 디지털 사이니지(signage)
30	지자체	건강관리서비스, 돌봄	노인	건강, 돌봄	활동 감지 센서, IoT
31	지자체	건강관리서비스, 돌봄	노인	건강, 돌봄	WebRTC, 활동감지 센서, IoT
32	준공공기관	정보 제공, 일자리 추천 등	노인, 업무 담당자	국민연금	DBMS, 머신러닝
33	준공공기관	보이스봇 초기상담	업무 담당자	복지 사각지대	ASR(Automatic Speech Recognition), NLP(Natural Language Processing).
34	준공공기관	이상 탐지	업무 담당자	전자바우처	머신러닝(FDS)
35	준공공기관	일자리 추천	업무 담당자, 근로자	구직서비스	머신러닝, VR, AR(메타버스)
36	지자체	여가서비스	노인	노인여가	디지털 사이니지

출처: 연구진 작성

다음으로 '순수 행정지원' 서비스이다. 흩어져 있는 데이터들의 연계로 인한 업무 간소화, AI OCR, 화상상담, GPS를 활용한 위치추적 등 업무 담당자의 업무 처리를 효율적으로 지원하는 서비스라고 할 수 있다. 순수 행정서비스는 업무 담당자가 독립적으로 담당해야 하는 업무를 간소화해 주거나, 화상상담 시스템이나 GPS 등과 같은 장비 지원을 통해 시간을 절약하고 빠르고 정확하며 신속한 업무 처리가 가능하게 한다. '의사결정 지원'은 빅데이터 기반 예측 모델을 활용하여 이상징후를 감지하거나 상담 과정에서 활용되는 추천 서비스를 제시한다. 이 과정에서 행정적 판단을 지원하거나 이후 절차인 서비스 중계 과정을 지원하게 된다. 이 의사결정 지원 과정은 서비스를 제공하는 과정에서 업무 담당자와 서비스 이용자를 연결해 주는 매개적 정보를 제공한다고도 볼 수 있다.

다음으로 서비스 분야로는 구직, 상담, 행정지원, 교육, 안전관리, 건강관리, 사각지대 발굴. 데이터베이스 관리 등이 확인되었다. 활용되는 주요 방법은 업무 자동화, 서비스 추천, 정보 제공, 모니터링, 상담 등으로 구분할 수 있다. 주요 기술(기능)은 머신러닝, 웹 개발, IoT, RPA (Robotic Process Automation), TTS(Text-to-Speech), GPS, VR, AR, NLP(Natural Language Processing), Web RTC(Real-Time Communication) 등이었다. 시스템 구축의 주체는 국가와 준공공기관, 지자체로 구성되었는데 대부분이 준공공기관이었다.

〈표 3-2〉 ICT 기반 사회보장 노력의 유형

구분	내용
이용자(사용자)	수급자, 업무 담당자, 일반 국민, 정책 개발자
제도별 분야	공공부조, 사회보험, 사회서비스
대상별 분야	노인, 빈곤층, 구직자, 산재근로자, 청소년, 복지 업무 담당자, 일반 국민
활용 영역	· 완결된 서비스 형태: 스마트로봇, 안전관리, 정보 제공, AI·IoT 건강 관리, 24시간 상담 지원 등 그 자체로 완결된 서비스의 한 형태 · 순수 행정지원: 데이터 연계, AI OCR, 화상상담, GPS 위치 파악 등으로 담당 업무 처리를 지원 · 의사결정 지원: 빅데이터 기반 예측 모델을 활용한 이상 징후 탐지, (상담 과정에서 활용되는) 추천 서비스 제공 등을 통해 행정적 판단을 지원하거나 이후 절차인 서비스 중계 과정을 지원
분야	구직, 상담, 행정지원, 교육, 안전관리, 건강관리, 사각지대 발굴
주요 방법	업무 자동화, 서비스 추천, 정보 제공, 모니터링, 상담(대화), 데이터베이스 관리
주요 기술	머신러닝, 웹 개발, IoT, RPA, TTS, GPS, VR, AR, NLP, Web RTC
주체	국가, 준공공기관, 지자체

출처: 연구진 작성

이상에서 살펴본 최근 5년 동안 ICT를 활용한 사회보장 분야의 노력은 전반적으로 다음과 같은 특징이 있다.

첫째, 사회적 돌봄 수요에 대한 충족 과정에서 활용되고 있다. 대부분 노인을 위한 것으로 AI·IoT를 활용한 건강 서비스, 생체건강 셀프체크 서비스, 안전감지 센서를 활용한 안전 확인 서비스, 실종 방지를 위한 위치 기반 모니터링 서비스뿐만 아니라 여가 활용에 도움이 되는 다양한 정보 제공 서비스, 디지털 사이니지(signage),[6] 정서 서비스 제공을 위한 음성 대화 등이 확인되었다. 중앙정부뿐만 아니라 지자체에서 제공하는 사례들도 확인되고 있다는 점도 주요 특징이다.

6) 디지털 사이니지란 "공공장소와 상업 공간에 LCD, PDP, LED 등의 디스플레이 패널을 통해 다양한 정보와 광고 등의 콘텐츠를 표출하는 미디어"(이하나, 2011, p. 503)를 가리킨다.

둘째, 업무 담당자의 효율적인 업무 처리를 위한 과정에서 ICT가 활용되고 있었다. 정보시스템 구축은 다양하게 흩어져 있는 데이터의 통합과 연계를 통해 업무 과정에 편의를 제공하고 있으며 업무 프로세스를 간소화하려는 노력도 확인됐다. 빅데이터를 기반으로 의사결정을 효율적으로 지원하는 서비스도 확인되었는데 수급자의 상담을 위한 기초 자료를 제공한다거나, 부정수급을 확인하고 적발하는 데 머신러닝 분석을 활용한 자료를 제공하는 방식이었다. 나아가 RPA(Robotic Process Automation)를 도입하여 반복적이고 규칙적인 업무를 자동화하는 방식으로 업무량을 줄여나가고자 하였다.

셋째, 정보 제공 서비스이다. 건강이나 고용 등 분야는 각기 다르지만, 다양한 데이터의 수집과 축적, 그리고 정보를 효율적으로 제공하는 신기술을 활용하여 다양한 정보를 일반 국민과 업무 담당자, 서비스 이용자에게 제공하고 있었다. 이러한 서비스는 앱 기반으로 제공되기도 하지만 빅데이터 시스템 구축 사업과도 연결되었다.

넷째, 개인 맞춤형 서비스 제공이다. 다양한 영역에서 개인에게 최적화된 서비스를 제공하기 위해 빅데이터 분석 기반 기술을 활용하여 서비스 이용자에게 맞춤형 서비스를 제공하고자 하였다. 일자리를 추천하는 서비스나 개인의 적성에 맞는 직업훈련을 찾아주는 서비스, 상담 과정에서 개인으로부터 나오는 대화 정보를 분석하여 상태를 감지하여 반응하는 서비스, 지원금 자기진단(모의 계산) 서비스 등은 모두 그 목적은 다르지만 결국에는 평균적인 수급자가 아닌 개인에게 특화된 서비스를 제공한다.

마지막으로 정보 활용의 효율성과 효과성을 높이기 위한 DBMS (Database Management System)의 활용이다. 유용한 데이터를 효율적으로 관리하고 활용할 수 있도록 하는 정보관리의 노력이 중앙정부뿐만 아니라 지자체에서도 발견된다. 데이터베이스의 관리는 업무의 효율화뿐

만 아니라 빅데이터 분석 환경을 마련하는 데까지 연결되며, 새로운 사업을 기획하고 사회적 가치를 창출하는 과정에서 유용하게 활용될 수 있다.

이상의 특징은 단순히 복지 기술 활용의 경향으로 발견되는 특징일 뿐만 아니라 ICT 기술이 발전함에 따라 인공지능이 핵심 기술로 점하게 될 영역으로 해석해도 큰 무리는 없을 것이다.

2. 사회보장 행정 영역에서 인공지능 활용의 사례

상기한 ICT 기반 사회보장 서비스 활용 영역 중 '의사결정 지원'에 해당되는 사례를 소개하고자 한다. '의사결정 지원'의 영역은 주로 업무 담당자의 행정적 판단을 지원하거나 이후 절차인 서비스 중계 과정과 연결되는 형태라고 할 수 있다. 사회보장 행정은 그 자체가 서비스가 아니라 서비스를 전달하는 과정에 가까운 개념이라는 점에서 사회보장 행정 영역에서의 인공지능 사례를 '의사결정 지원' 기능에 초점을 두는 것이 적절하다고 본다. 이 절에서는 대표적인 사례로 인공지능에 기반한 취업 알선서비스(The Work AI)와 복지 사각지대 발굴시스템, 그리고 AI 초기상담 서비스를 소개한다.

취업 알선 서비스는 근로자의 구직을 지원하기 위해 빅데이터 정보를 활용해 구직자에게 맞는 일자리를 추천해 주는 서비스이다. 추천 결과는 업무 담당자의 상담 과정에 활용되거나 실제로 일자리에 지원하는 과정에서 참고가 된다. 복지 사각지대 발굴시스템은 취약 집단이 복지 급여 신청 전에, 국가가 선제적으로 어려운 상황을 발견하고 서비스를 제공하려는 목적으로 운영된다. AI 초기상담 서비스는 음성으로 대화하는 AI 기반 봇을 활용하여 잠재적 취약계층을 대상으로 초기상담을 진행하고 그 결과를 업무 담당자에게 전달하여 심층 상담 등 이후의 후속 조처를 할 수 있도록 지원한다.

세 개의 사례 모두 서비스 제공에 필요한 빅데이터 수집과 처리 과정이 지능화되고 있으며 분석 알고리즘도 지속해서 개발될 예정이다. 빅데이터 분석 결과가 이후 대상자들을 위한 서비스에 영향을 미친다는 측면에서도 공통점이 있다.

가. AI 기반 일자리 매칭 서비스(The Work AI)

AI 기반 일자리 매칭 서비스(The Work AI)는 고용노동부가 「고용정책 기본법」에 근거여 2018년부터 인공지능 기술을 활용하여 제공하고 있는 서비스이다. 근로자의 일자리 탐색 비용 절약과 일자리 미스매치의 완화를 위한 사업으로 한국고용정보원이 위탁받아 운영한다.

정부는 2018~2021년까지 총 128억 원을 투입하였으며, 정보화 전략계획(ISP) 및 1~2차 사업을 추진하여 AI 기반 일자리 매칭 시스템을 구축·고도화했다(감사원, 2022). 1차 사업이 종료된 2020년 7월부터 워크넷을 통해 서비스를 제공하고 있다.

〈표 3-3〉 AI 기반 일자리 매칭 시스템 구축 사업 추진 현황

구분	정보화 전략 계획(ISP)	1차 사업	2차 사업
사업명	머신러닝(AI) 기반 일자리 매칭 ISP 사업	국가 일자리 정보 플랫폼 기반의 AI 고용서비스 구축	머신러닝 기반 일자리 매칭 시스템 구축 2차 사업
사업 기간	2018년 1~10월	2019년 8월~ 2020년 7월	2020년 7월~ 2021년 5월
사업비	10억 원	49억 원 (매칭 시스템 29.26억 원)	89.34억 원 (리스 비용 포함)

출처: 감사원. (2022). “감사보고서- 취업알선정보망 구축 및 관리실태-”. p.20.

일자리 매칭 서비스는 간단히 말해 머신러닝에 기반하여 작동되는 취업알선 서비스이다. 구직자와 구인 기업 간에 발생하였던 일자리 미스매치를 해소하고 고용서비스 사각지대를 최소화하기 위한 목적으로, 알고리즘에 기반하여 추천함으로써 근로자의 직무역량에 초점을 맞춰 구직서비스를 활성화하는 서비스다. 2020년 7월에 개시됐다.

기존에는 일자리를 찾는 과정이 직종 중심으로 진행되었다. 인공지능 기반 일자리 매칭 서비스는 구직자가 작성한 이력서의 정보와 구인 기업의 채용공고에 나타난 직무역량 정보를 수집하여 인공지능 기반의 빅데이터 분석을 수행한다(고용노동부 보도자료, 2020. 7. 9, p.1). 알고리즘을 기반으로 한 분석 결과는 구직자와 구인 기업에 각각 적합한 일자리와 근로자를 연결한다. 시스템 도입 전에는 구직자와 구인 기업 각자가 필요조건을 직접 작성하고 필요한 정보의 검색도 직접 수행했다. 인공지능 기반 일자리 매칭 서비스가 개시된 이후부터는 구직자와 구인 기업이 워크넷을 통해서 AI 알고리즘이 추천하는 구인 및 구직 정보를 실시간으로 추천받을 수 있다(고용노동부 보도자료, 2022. 8. 30, p.1).

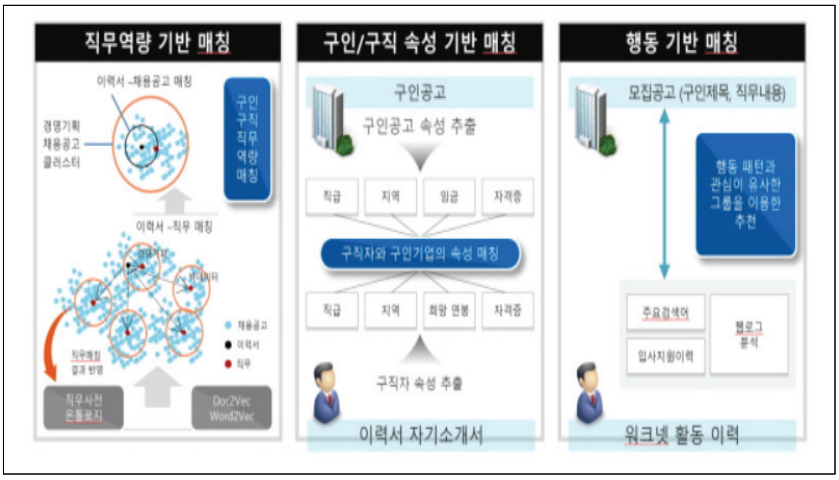
인공지능 기반 일자리 협력 체계의 작동 원리는 다음과 같다. 구직자가 작성한 이력서와 구인 기업이 고지한 구인 정보는 분석을 위한 데이터로 활용된다. 여기에는 직종이나 직무, 지역 등의 여러 가지 변수가 포함된다. 이 자료를 근거로 분석이 이루어지면서 각자에게 추천 후보를 제시하고 매칭 알고리즘(DeepFM)을 기반으로 각각 점수를 산정한다. 구직자는 매칭 점수가 높은 최대 50개의 구인 정보를 추천받을 수 있다(고용노동부 보도자료, 2022. 8. 30, p.4).

인공지능 기반 일자리 매칭 서비스가 가능하게 된 배경에는 정보시스템 통합이 자리 잡고 있다. 각기 다르게 관리되던 5대 고용 정보는 2018년에 국가일자리정보플랫폼에 통합되었고 그 덕분에 자격 및 훈련 정보나

일자리 정보, 직업 및 진로 정보 등과 같은 빅데이터에 접근할 수 있게 되었다. 데이터를 기반으로 제공하는 추천 서비스이기 때문에 구직자나 구인 기업이 각각 이력서와 구인 공고에 관련 직무 내용을 구체적으로 작성할수록 적합한 결과를 얻을 수 있다(고용노동부 보도자료, 2020. 7. 9, p.2).

매칭 알고리즘은 아래의 [그림 3-1]과 같다. 크게는 직무역량 기반 매칭과 구인 및 구직 속성기반 매칭, 그리고 행동 기반 매칭으로 분류된다. 이력서와 자기소개서, 구인 공고, 워크넷 활동 이력 등의 정보를 통해 각기 다른 기준으로 최적의 매칭이 이루어지게 된다.

[그림 3-1] 더워크 에이아이 매칭 알고리즘



출처: 고용노동부. (2020. 7. 9). 고용노동부 보도자료. 인공지능(AI) 기반의 일자리-인재 추천 서비스 시작. p.2.

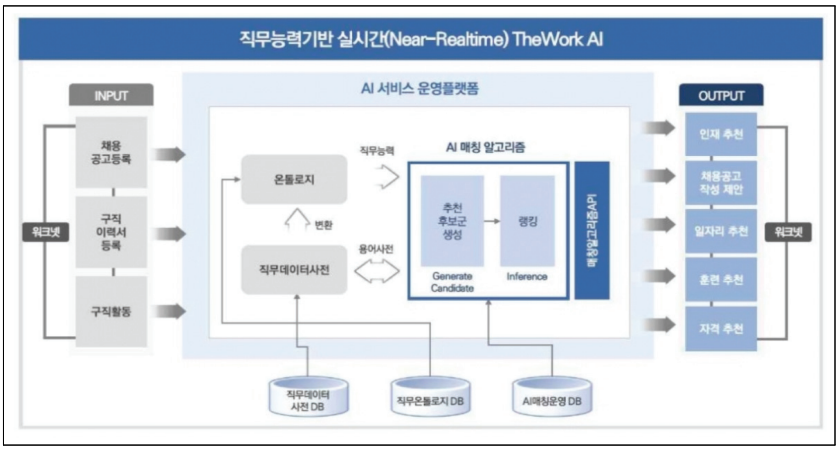
인공지능 기반 일자리 매칭 서비스는 AI가 직무 데이터 사전[7]을 가지고 구직자가 작성한 이력서와 구인 기업이 제출한 채용공고를 분석한다. 이후 직무역량을 중심으로 구직자와 구인 기업 사이에서 일자리를 매칭하는 방식으로 운영된다. 즉 워크넷을 통해 수집한 구인자 및 구인 기업 관련 정보의 데이터 수집 및 전처리를 진행한 이후, ① 형태소 분석과 품사 태깅, ② 키워드 추출, ③ 단어 간의 연관성 생성, ④ 직무 데이터 사전의 시각화 과정을 거친 이후 직무온톨로지를 구축하고 이에 대한 시각화 서비스를 제공한다. 워크넷에서 제공되는 직무온톨로지는 "기계가 이해할 수 있도록 직무역량을 직무용어, 자격, 학과, 훈련, 직업으로 표현한 직무 지식체계"를 말한다.[8] 그 이후에 직무역량을 기반으로 한 매칭 알고리즘 AI는 구직자와 구인 기업을 위해 추천 후보군을 생성하여 높은 확률에 따라 랭킹을 도출한다. 그 결과 수요자인 구직자나 구인 기업에서는 워크넷을 통해 적절한 일자리나 구인자, 훈련 및 자격서비스의 추천 결과를 확인할 수 있다.

---

7) 직무 데이터 사전은 "▲국가직무능력표준(NCS) 직무 정의 및 학습자료 ▲일자리포털 워크넷 모집 공고 ▲직업사전, 훈련 및 자격 등 관련 자료 18종에서 270만 개 핵심어를 추출하고 직무와 핵심어 사이의 연관성 분석과 관계를 정의"(고용노동부, 2020.5.26., p. 4)한 데이터 사전을 말한다.

8) 윤난슬(2021.4.14.). 네이트 뉴스"전주비전대, 인공지능 기반 '취업 알선 서비스' 적극 활용" https://news.nate.com/view/20210414n14821

[그림 3-2] AI 기반 일자리 매칭 시스템 구조



출처: 윤화정. (2022.04.12). 128억 투입된 워크넷 'AI 일자리 매칭', 성과 저조…매칭점수-입사 지원간 상관관계 낮아. 워크투데이. [보도자료]. http://www.worktoday.co.kr/news/articleView.html?idxno=24170

이론대로만 보면, AI 기반 일자리 매칭 서비스는 구직자에게는 일자리 정보를 제공하여 구직 어려움을 줄여주고 구인 기업에는 기업에 적합한 직무역량을 지닌 근로자를 채용할 수 있을 것으로 기대됐다. 추천 정보가 바탕이 되어 상담원의 역량이 강화하고 이들의 보다 나은 진로지도 서비스 제공이 가능해질 수 있다. 그러나 감사원은 2022년 감사보고서 '취업 알선정보망 구축 및 관리 실태'에서 해당 시스템의 예측률이 낮은 점을 지적한 바 있다. 감사원은 시스템이 예측한 입사 지원율과 실제 입사 지원율 간에 큰 차이가 존재하는 것을 발견하였다. 구체적으로 보면, AI 매칭 시스템의 입사 지원율은 1차 사업(2020년 7월~2021년 7월)에서 9.45%, 2차 사업(2021년 7월~2021년 10월)에서 12.99%로 나타났는데 이 수치는 인공지능 기술을 활용한 시스템의 입사 지원율이 동일 기간에 기존 빅데이터 추천 시스템의 입사 지원율인 15.22%, 18.36%에 비해 각각 5.77%p, 5.37%p 낮았다(감사원, 2022, p. 27).

감사원은 AI 기반 매칭 시스템의 변수를 조정할 필요성도 제기했다. 일자리나 지역을 추천하는 기준이 불합리하다고 판단했기 때문이다. 구인정보가 누락된 상태로 분석이 이루어지거나 임금 체불 상태인 사업주 정보 등과 같이 적절하지 않은 구인 정보를 제공하고 있는 점, 미등록 취업 외국인에 대하여 취업 알선을 하는 등의 문제를 지적하기도 했다(감사원, 2022).

〈표 3-4〉 AI 기반 매칭 시스템과 빅데이터 추천 시스템상 입사 지원율 비교[9)]

(단위: 건, %)

구분		AI 기반 매칭 시스템			빅데이터 추천 시스템		
		계	1차 사업	2차 사업	계	2020.7.9.~ 2021.7.18	2021.7.19.~ 2021.10.31
중복제거	조회 건수	1,676,634	326,713	4,116,597	4,116,597	3,301,254	815,343
	입사지원 건수	206,262	30,873	175,389	652,393	502,692	149,701
	입사 지원율 (매칭율)	12.30	9.45	12.99	15.84	15.22	18.36

주: 1) 2차 사업 시범 기간에는 1차 사업과 2차 사업 추천이 중복으로 적용되었음.
또한 AI 추천의 경우 추천 건당 어떤 알고리즘으로 추천하였는지 구분이 가능하여 이를 기준으로 분석함.
2) 빅데이터 추천의 기간 구분은 AI 추천 알고리즘 변경 시점을 기준으로 구분함.
출처: 감사원. (2022). "감사 보고서- 취업알선정보망 구축 및 관리실태 -". p.24.

이와 같은 감사원의 문제 제기를 보완하는 과정에서 AI 기반 매칭 시스템도 외연을 확장해 나가고 있다. AI 기반 일자리 매칭 서비스가 2024년 5월 과학기술정보통신부의 '2024년 부처 협업 기반 AI 확산 사업'에 선정됐다(고용노동부, 2024). 과학기술정보통신부의 AI 확산 사업은 부처끼리 협업하여 AI 솔루션을 개발하고 국민 체감형 서비스 제공 및 AI

9) 감사원(2022). "감사 보고서- 취업알선정보망 구축 및 관리실태 -". p. 24.

생태계 경쟁력 강화를 위한 목적으로 추진된다. 고용노동부는 '인공지능 기반 구인·구직 통합지원 솔루션 개발' 사업으로 앞으로 3년에 걸쳐 해당 과제를 추진할 계획이다.

고용노동부가 수행할 시범 과제의 세부 내용은 총 7개다(〈표 3-5〉 참고). 구직자 직업 선호도나 경력 등 다양한 데이터를 기반으로 취업역량 향상을 지원하고, 구인 공고를 작성하거나 채용 조건을 제시하는 등 구인 기업의 채용 과정에서 편의성을 증대시키며, 일자리·인재 추천 고도화로 매칭 서비스를 강화하는 데 중점을 두었다(고용노동부, 2024). 시범 과제에서 계획하는 서비스 제공 후 서비스 개선 효과도 〈표 3-5〉에서 제시하는 바와 같다.

일자리 매칭 서비스가 과학기술정보통신부의 사업을 통해 확장되고 고도화되는 과정에서 다양한 기술이 활용된다. AI 인재 추천에서부터 AI 직업 훈련 추천에 이르기까지 각 시범 과제 수행을 위해 아래와 같이 최소 3개에서 최대 6개의 기술이 활용된다.

〈표 3-5〉 시범 과제 서비스 제공 시 개선 효과

구분	시범 과제	AS-IS	TO-BE
일자리 매칭	❶ 인재 추천	■추천 인재의 이력서·자기소개서를 구인 기업이 일일이 열람	■생성형 AI가 이력서·자기소개서를 요약 제공 → 추천 인재 빠르게 확인
구인	❷ 구인 공고 작성 지원	■구인 기업이 구인 공고 내용을 모두 작성하여 등록	■구인 공고 필수사항만 입력 → 생성형 AI가 구인 공고 자동 생성
	❸ 채용 성공 모델	■없음	■구인 공고 정보 분석→채용 성공 모델 구축→채용 확률별 맞춤형 서비스
	❹ 구인 공고 검증	■차단 키워드 정의 → 구인 공고에 차단 키워드 포함 여부 확인	■키워드 방식 차단+구인 공고 AI 학습을 통한 구인 공고 적정성 여부 판단
구직	❺ 지능형 직업심리검사	■정형화된 직업심리검사 제공	■사용자 특성 AI 분석→필수항목만 질의하는 지능형 직업심리검사 제공
	❻ 취업 성공 모델	■없음	■구직자 정보 분석→취업 성공 모델 구축→취업 확률별 맞춤형 서비스
	❼ 직업훈련 추천	■구직자 희망 직종+구직활동 유사 그룹 분석을 통한 직업훈련 추천	■구직활동 유사 그룹+구직자 희망 직종·직무능력-훈련과정 연관성 분석을 통한 유사도 기반 추천

* 기존 서비스 강화: ❶, ❼ / 신규 도입 서비스: ❷~❻

출처: 고용노동부. (2024.6.12.). 고용노동부 보도자료. "맞춤형 구인·구직·매칭서비스가 인공지능(AI) 기반으로 확 달라집니다." p.2.

〈표 3-6〉 7대 시범 과제별 주요 AI 활용·적용 기술

고용 AI / 서비스 활용 기술	AI 인재 추천	구인 공고 AI 작성지원	채용 성공 모델	구인 공고 AI 검증	지능형 AI 직업심리 검사	취업 성공 모델	AI 직업 훈련 추천
언어 모델(BERT)	○	-	-	○	○	-	○
거대 언어 모델(LLaMA)	○	○	-	-	-	-	-
딥러닝&머신러닝	○	-	○	○	○	○	○
자연어처리(NLP)	○	○	○	○	○	○	○
벡터 DB&벡터 검색	○	-	-	-	-	-	○
지식 그래프&온톨로지	○	○	○	-	-	○	○

출처: 고용노동부. (2024.6.12.). 고용노동부 보도자료. "맞춤형 구인·구직·매칭서비스가 인공지능(AI) 기반으로 확 달라집니다". p.6.

나. 복지 사각지대 발굴시스템

2015년 12월에 운영된 복지 사각지대 발굴시스템은 2014년 '송파 세 모녀 사건' 이후 신청주의 한계를 극복하고 복지 사각지대에 놓여 있는 대상을 선제적으로 발굴·지원하기 위해 구축되었다. "취약계층 관련 정보들을 활용하여 복지 사각지대 소외계층을 선제적으로 발굴하고 지원"(보건복지부, 2015)하기 위해 시스템 구축 이후 2023년까지 단전, 단수 등 위기 정보를 보유한 666만 명(누적)의 복지위기 가구를 발굴하였고, 290만 명(누적)에게 기초생활보장, 긴급 지원 등 공적 급여와 민간 자원 연계 등 복지서비스를 지원하였다(보건복지부, 2024).

복지 사각지대 발굴시스템은 각기 다른 기관에 흩어져 있는 공공데이터를 수집해서 빅데이터로 구축하여 고위험 확률모델을 활용하여 위험군을 도출하는 것이 핵심이다. 시스템 운영 과정은 자료의 수집과 빅데이터 분석, 분석을 통한 고위험 대상자를 도출, 도출된 대상자 명단을 지자체에 제공, 지자체 담당자가 대상자에 대해 상담 및 지원하는 순으로 이루어진다. 이러한 업무는 지자체의 '찾아가는 보건복지서비스 전담팀'에서 담당하여 대상자에 대한 방문이나 연락 등의 조사를 통해 사각지대를 발굴하게 된다. 격월로 대상자가 제공되어 관련 업무는 1년에 총 6회 진행된다.

절차를 구체적으로 살펴보면 다음과 같다. 2024년 9월 말을 기준으로 단전 및 단수 상태 가구 여부 등 19개 공공 및 민간 기관에서 보유한 45종 위기정보를 정보시스템을 통해 2개월마다 입수하고 빅데이터 분석을 수행한다. 연계 정보에는 자살예방센터의 '자살고위험군'과 응급의료센터의 내원 사유가 '자해·자살인 대상자'를 포함하고 있어 경제적 위기뿐 아니라 정신건강상 어려움이 있는 상황까지 위기상태로 포괄하고 있다.

다양한 기관에서 입수한 정보를 통해서 약 950만 명의 대상자 정보가 수집되며, 이 중 고위험군으로 여겨지는 약 15만 명을 선별하여 지자체에 명단을 제공한다(보건복지부 보도자료, 2024.3.22.). 이후 지자체는 이들에 대한 조사를 진행하는데 읍면동 단위의 찾아가는 보건복지팀이 대상자에 대한 방문과 상담을 진행하고 필요한 서비스를 제공한다.

〈표 3-7〉 복지 사각지대 발굴시스템 연계 정보(45종)

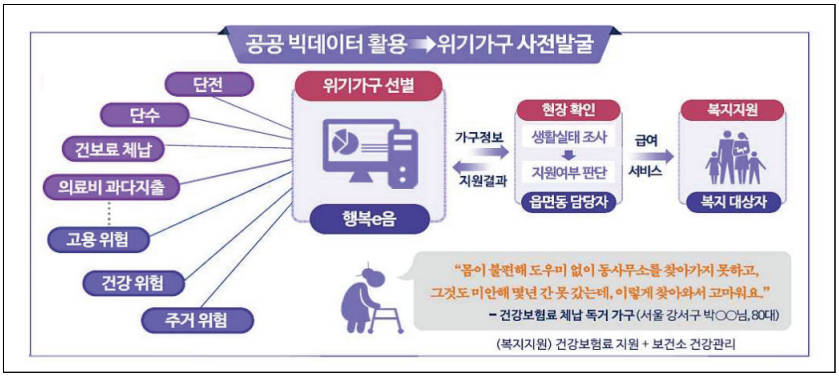
근거: 법률(제12조 제1항 각호 및 제2항)		근거: 시행령(제8조 제2항 별표 2 각호)	
정보 내용	보유 기관	정보 내용	보유 기관
단전	한국전력공사	국민연금 보험료 체납	건강보험공단
단수	상수도사업본부	의료 위기[1]	
단가스	도시가스사	범죄 피해	경찰청
초중고 교육비 지원 중 학교장 추천	교육부	화재 피해	소방청
		재난 피해	행정안전부
건보료 체납	건강보험공단	주거 위기[2]	국토교통부 한국토지주택공사 각 지방개발공사 아파트 관리사무소
건보료 부과 내역			
기초수급 탈락·중지	보건복지부	고용 위기[3]	고용노동부 근로복지공단
복지시설 퇴소			
금융 연체	신용정보원	방문건강사업 대상	보건복지부
채무조정 중지(실효)자			
통신비 체납	한국정보통신진흥협회		
노후긴급자금 대부(실버론)	국민연금공단	기저귀 분유 지원	
		신생아 난청 지원	
		영양플러스 미지원	
		맞춤형 급여 신청	
		전기료 체납	한국전력공사
		수도요금 체납	상수도사업본부
		가스요금 체납	도시가스사
		자살고위험군	자살예방센터

근거: 법률(제12조 제1항 각호 및 제2항)		근거: 시행령(제8조 제2항 별표 2 각호)	
정보 내용	보유 기관	정보 내용	보유 기관
		내원 사유 자해·자살	응급의료센터
		휴·폐업자	국세청
		세대주가 사망한 가구	행정안전부
		주민등록 세대원	

주: 1) ① 의료비 부담 과다, ② 장기 요양, ③ 중증질환 산정특례, ④ 요양급여 장기 미청구, ⑤ 장기요양 등급, ⑥ 재난적 의료비 지원 대상
2) ① 전세 기준금액 이하, ② 월세 기준금액 이하, ③ 공공임대주택 임대료 체납자, ④ 공동주택 관리비 체납자
3) ① 개별연장급여 대상자, ② 실업급여 수급자(임금체불, 폐업), ③ 비자발적 사유로 고용보험 상실 후 재취득이 없는 자 중 실업급여 미수급자, ④ 일용근로자 중 실업급여 미수급자, ⑤ 산재요양 종결 후 근로 단절, ⑥ 고용위기(최근 1년 이내 고용보험 가입 이력이 없는 대상자)

출처: 보건복지부. (2024.3.25.). 보건복지부 보도자료. “45종 위기정보 활용해 2024년 2차 복지사각지대 발굴 시행”. p.4.

[그림 3-3] 빅데이터를 활용한 복지 사각지대 발굴 업무 절차



출처: 보건복지부. (2023.1.17.). 보건복지부 보도자료. “2023년 4차 복지 사각지대 발굴 시작”. p.3.

시스템이 구동되는 절차를 보다 자세히 살펴보면 다음과 같다(김은하, 2022). 복지 사각지대 발굴을 위해 수집되는 위기 정보에는 단전 상태, 단수 상태, 가스 단절 상태 정보 외에 금융 연체나 건강보험료 체납과 관련된 정보, 통신비 체납과 관련된 정보, 의료 위기 정보와 범죄 피해 정보

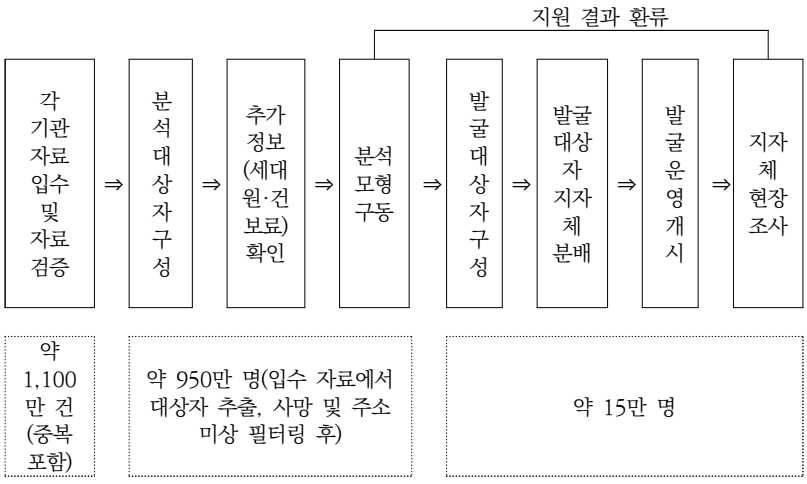
등이 포함되며, 법 개정을 통해 연계 정보 범위를 지속해서 확장하고 있다. 과거에는 주로 경제적 어려움에 초점을 두고 빈곤한 대상자들의 욕구를 대상으로 하였지만 최근에는 질병이나 돌봄, 정신건강 등 연계 정보의 범위를 확장하고 있다. 위기가구의 성격이 시간의 흐름에 따라 달라지고 있다.

정보시스템을 통해 수집된 공공정보와 민간 정보는 머신러닝 기반 빅데이터 분석 모델을 통해 고위험 상태라고 예측되는 대상자를 최종적으로 도출하는 단계를 거친다. 분석을 위해 가구를 구성하는 등의 전처리 과정이 이루어지며 데이터 세트가 구축된 이후에는 고위험 예측 모형 구동을 위해 데이터의 분석 및 훈련, 모델 검증 및 구동의 절차가 진행된다.

고위험 예측 모형 분석의 결과로 지자체에 제공될 위기 대상자 명단이 도출된 이후에는 읍면동 복지 업무 담당자에게 사회보장 정보시스템을 통한 정보가 제공된다. 담당자들은 정보시스템에서 각자 관할 지역의 대상자 명단을 확인할 수 있으며 유선 통화나 가정방문을 하여 상담을 진행한다. 상담 후 공공지원이 필요하다고 판단되는 경우에는 급여 신청을 안내하고 그 과정을 지원하여 복지 급여나 서비스가 제공되도록 한다. 담당자는 이상의 상담부터 지원 결과에 이르는 내용을 사회보장 정보시스템에 입력한다. 이 입력 내용은 빅데이터 분석모형의 정교화를 위한 피드백 데이터로 활용된다.

공공지원을 신청했으나 지원 기준에 부적합 판정을 받은 대상자들에게는 민간 자원이 연계된다. 다만 지역 간 민간 자원의 편차가 커서 지역의 자원 보유 상황이나 담당자의 역량 등에 따라 제공되는 민간 자원의 수준이나 범위가 달라질 수 있다.

〈표 3-8〉 복지 사각지대 발굴시스템 프로세스



출처: 보건복지부. (2024.3.25.). 보건복지부 보도자료. "45종 위기정보 활용해 2024년 2차 복지 사각지대 발굴 시행". p.3. 재구성.

복지 사각지대 발굴시스템은 연계 변수가 확장되면서 빅데이터 분석에 활용되는 기술과 모델의 변천을 겪고 있다. 새로운 사회적 위기상태가 발생하게 되면 그 위기를 정확히 포착하기 위해 연계 변수의 추가가 이루어지고, 빅데이터 분석에 활용되는 기술의 발전에 따라 더욱 정교하게 대상자를 발굴할 수 있는 새로운 방법과 모델이 적용된다. 단일 모델의 분화로 모델이 복잡해지고 있으며 가구 특성이나 개인 특성을 포착할 수 있는 모델도 개발 중이다.

보건복지부(2024.3.25.)에 따르면 복지 사각지대 발굴시스템 구축 이후(2015~2023년) 약 9년간 665.6만 명에 대한 조사가 지자체를 통해 이루어졌으며 이 중 290.2만 명(43.6%)에게 공공·민간 복지서비스를 지원하였다. 지원 대상은 2015년 1.8만 명에서 2023년에 68.6만 명으로 늘었고, 지원율은 2015년 16%에서 2023년 49.4%로 증가하였다(보건복지부, 2024.3.25.).

〈표 3-9〉 복지 사각지대 발굴시스템의 분석 기술 변천사

연도	2015	2017	2019	2021	2024
활용기술	Logistic, Elastic, GBM	XGBoost	XGBoost	XGBoost, RandomForest	XGBoost, RandomForest, VotingClassifier
연계변수 종수	16종 연계	23종 연계	29종 연계	33종 연계	45종 연계
모델 분류	없음	없음	1인 가구, 다인 가구 모델	1인 가구, 다인 가구 모델	1인 가구, 다인 가구 모델

출처: 이우식. (2024.10.22.). 사회보장행정에서의 인공지능 적용 동향과 함의 한국보건사회연구원 세미나 자료.

〈표 3-10〉 복지 사각지대 조사 대상자 및 지원 내역

발굴 차수	발굴 대상자 (a)	복지서비스 지원 내역						
		계(b)	지원율 (b/a)	기초 생활보장	차상위	긴급 복지	기타 공공서비스*	민간 서비스 연계**
계	6,656,547	2,902,301	43.6	148,766	58,790	86,352	516,048	2,092,345
'15년	114,609	18,318	16.0	1,966	961	702	10,367	4,322
'16년	208,653	46,780	22.4	3,064	6,573	719	20,278	16,146
'17년	298,638	76,638	25.7	6,712	8,537	1,109	31,412	28,868
'18년	366,755	133,490	36.4	18,345	6,588	1,663	38,860	68,034
'19년	633,075	228,009	36.0	17,674	3,825	3,276	42,099	161,135
'20년	1,098,134	442,652	40.3	23,723	5,243	25,374	74,019	314,293
'21년	1,339,909	663,874	49.5	28,611	11,180	19,664	105,740	498,679
'22년	1,208,086	606,101	50.2	25,708	9,338	15,402	109,351	446,302
'23년	1,388,689	686,439	49.4	22,963	6,545	18,443	83,922	554,566

* (기타 공공서비스) 장애인연금, 사회서비스이용권(노인돌봄, 장애인활동지원 등), 요금감면 등
** (민간서비스) 공동모금회, 푸드뱅크, 대한적십자사 희망풍차, 민간기관 결연후원금 등
출처: 보건복지부. (2024.3.25.). 보건복지부 보도자료. "45종 위기정보 활용해 2024년 2차 복지 사각지대 발굴 시행". p.5.

복지 사각지대 발굴시스템이 다른 사업들과 다른 점은, 정보시스템의 구축 및 개선 과정에서 법 제정이나 전달체계의 변화 등이 함께 이루어졌다는 것이다. 정부는 그간 '복지 사각지대 발굴 및 지원 종합대책'을 몇 차례 수립하는 과정에서 복지 사각지대 발굴시스템을 개선하였으며 현장에서는 종합대책의 일환으로 시스템을 활용한 업무들을 진행했다. 그 과정을 간단히 일람하면 다음과 같다.

먼저, 복지 사각지대 발굴시스템의 구축 배경인 2014년 2월 송파 세 모녀 사건이 있었다. 당시 사건을 통해 위기가구임에도 공공부조를 신청하지 않아 관할 지자체에서 이들의 상황을 감지하지 못한 문제가 파악되었다. 이 과정에서 2014년 12월 「사회보장급여의 이용·제공 및 수급권자 발굴에 관한 법률」이 제정되었고 이 법을 근거로 복지 사각지대 발굴시스템이 구축·운영되었다(2015년 12월~). 복지 사각지대 발굴시스템은 복지 업무 담당자의 사각지대 발굴 업무를 지원하게 되었다(관계부처 합동, 2022:2).

2018년 4월에 발생한 충북 증평 모녀 사건에서는 복지 사각지대 발굴시스템에 고임대료 상태인 거주지가 위기 정보에 포함되지 않은 점이 지적되었다. 같은 해 7월에 '복지 위기가구 발굴 대책(2018, 보건복지부)'이 발표되면서 복지 사각지대 발굴시스템의 발굴 대상자 범위가 확대되었다.

2019년 7월의 탈북민 모자 사건에서는 아동수당 신청 과정에서 대상 가구의 소득·재산이 없는 것이 확인되었으나 지원 안내가 적절히 이루어지지 않았다는 지적이 있었다.[10] 이후 복지 사각지대 발굴시스템을 활용한 지자체 위기가구 기획조사가 의무화되었고, 통합사례 관리사를 활용한 위기가구 발굴을 강화하였다.

---

10) 강애란. (2019.8.16.). 복지부, '탈북민 모자 사망'에 위기가구 긴급 실태조사. 연합뉴스. https://www.yna.co.kr/view/AKR20190816136800017

2022년 8월 수원 세 모녀 사건에서는 해당 가구가 채무와 질병 문제를 지닌 위기가구였지만 실거주지와 주민등록의 불일치 때문에 지자체의 관리 대상에서 제외되었다는 지적이 있었다(복지부 보도자료, 2022.8.23.). 동년 11월에 '복지 사각지대 발굴·지원체계 개선 대책'(관계부처 합동, 2022)에서는 복지 사각지대 발굴시스템의 정보 입수 범위를 확대하였다. 나아가 자립준비청년이나 가족 돌봄 청년, 고독사 위험자나 은둔청년 등 새로운 위기 속에 있는 대상자를 발굴할 수 있는 체계를 마련하였다. 그 밖에도 복지멤버십 확대나 휴대전화 번호 연계를 통한 위기 대상자 소재 파악 등의 조치를 취하였고, 실거주지 기준으로도 대상자를 발굴하고 지원할 수 있도록 하였다(관계부처 합동, 2022).

복지 사각지대 발굴시스템이 이와 같이 현장의 전달체계와 긴밀하게 관련을 맺고 있기 때문에 지속해서 제기되는 문제들도 상존한다(김은하, 2022; 최정은 외, 2022; 함영진 외, 2023). 우선, 발굴 대상자의 정확성에 관한 문제이다. 복지 사각지대 발굴시스템은 주로 주소와 전산 정보에 의존하여 위기가구를 파악하고 있는데 고위험 가구가 시스템에 포착되지 못하고 있는 이슈들이 지속해서 제기되어 왔다.[11] 특히 고위험 가구의 사망사건이 발생할 때마다 주소와 전산에 기반한 발굴시스템의 한계가 지적된다.

또한, 위기가구를 발굴했다고 해도 소득·재산 수준으로 공공지원이 어려운 상태일 경우 지원이 어렵게 된다. 이는 공공부조 기준의 엄격함에 기인하는 근본적인 문제이다. 그 대안으로 복지자원이 풍부한 지자체에서는 다른 서비스를 연계해 줄 수 있지만 다수의 지자체는 그렇지 못한 상황에서 실질적인 어려움을 겪고 있는 대상자를 발굴했음에도 이후 지원의 절차가 이루어지지 못하고 있다.

11) 장동욱. (2022.11.25.). 전입신고 안됐다고 방치…'복지 사각지대 발굴' 시스템 구멍. TV조선. [보도자료]. https://n.news.naver.com/mnews/article/448/0000384107

개인정보 이슈도 현장에서 지속해서 제기되고 있다. 위기 대상자로 판정된 경우 가정방문 등을 통해 상담을 진행한다. 그 과정에서 자신이 위기상태에 있다는 사실을 파악한 경위 등 대상자들이 개인정보에 대해 민감한 반응을 보여 일선 공무원들이 어려움을 경험하고 있다.

마지막으로 복지 사각지대를 해소하기 위해 정보시스템에 전적으로 의존하기보다 취약계층이 복지서비스에 더 쉽게 접근할 수 있도록 공공지원 대상을 확대해야 한다는 지적도 있다(성은미 외, 2024; 이영글 외, 2021). 정보시스템을 활용한 복지 사각지대 발굴보다 지역사회와의 협력 강화와 현장 중심의 발굴 노력 등이 더욱 중요하고 더 높은 효과를 보인다는 주장이다.

복지 사각지대 발굴시스템의 성과를 통해 위기상태에 있는 개인이나 가구의 상당수가 국가 지원이나 민간 지원의 안전망 안으로 포함된 사실은 부인할 수 없다. 그럼에도 이러한 지적들은 향후 시스템의 발전을 위해서 짚어볼 지점이다. 안전하고 신뢰받는 정보시스템을 위해 지속적인 고민이 필요할 것으로 보인다.

다. AI 활용 초기상담 정보시스템

AI 활용 초기상담 정보시스템은 초기상담 단계에서 지자체의 복지 업무 담당자가 아닌 인공지능 기술을 활용한 전화 상담을 진행하여 효율적인 업무 처리가 이루어질 수 있도록 하는 목적으로 도입되었다.

시스템이 도입된 배경에는 복지 사각지대 발굴시스템의 작동 방식과 지자체의 관련 업무 수행 과정에서의 애로사항이 자리 잡고 있다. 복지 사각지대 발굴시스템을 통해서 지자체에 제공되는 다수의 대상자를 위해 방문 및 상담을 진행하고, 필요한 서비스를 지원하는 과정은 한정된 인력

으로 진행되기 때문에 업무 부담이 발생할 수밖에 없다. 이러한 업무량을 감소시키는 대신, 지원이 필요한 이들에 대해서는 보다 집중적인 상담서비스를 제공하기 위해 AI 활용 초기상담 정보시스템이 구축됐다.

이와 관련하여 복지 사각지대 발굴시스템 운영 이후에 발생한 문제는 두 가지로 요약할 수 있다(보건복지부, 2024). 첫째는 복지 사각지대 발굴시스템의 작동 원리와 지자체 발굴 과정과 관련된다. 복지 사각지대 발굴시스템은 고위험군을 선별하는 것을 주된 기능으로 하고 있지만, 머신러닝 모델 구동 결과 고위험군에 해당하지는 않지만, 위기상태인 대상자가 상존했다. 그 외에 알고리즘이 고위험군으로 분류하지 않은 대상까지 모두 포괄하여 전수 파악이 필요하다는 의견도 제기됐다. 따라서, 데이터로 수집된 전체 대상자들의 상황을 확인하는 방안을 모색하게 되었다. 정보시스템에 연계되는 정보는 제안될 수밖에 없다는 문제의식이었다.

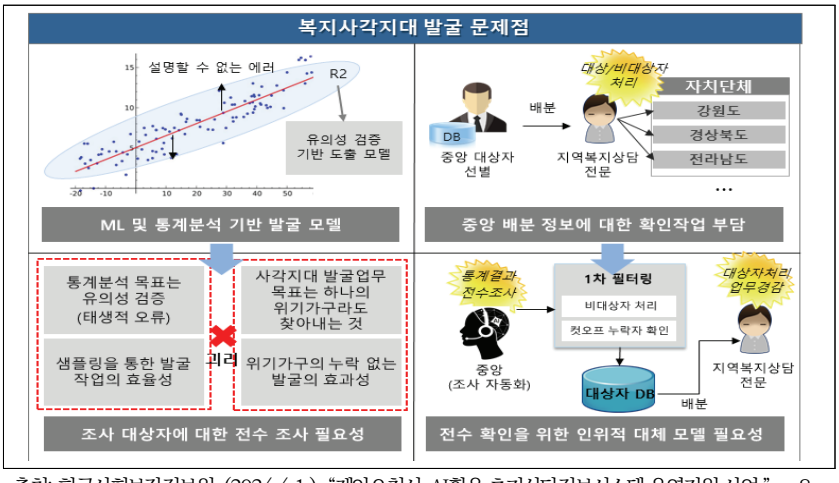
확률에 기반한 통계 모델의 결과가 파악하지 못하는 위기가구가 존재할 가능성, 그리고 '복지 사각지대' 정의를 고려할 때 사각지대 대상자는 행정 데이터 기반으로 형성된 알고리즘으로는 완벽한 접근이 어렵기 때문에 알고리즘이 선택하지 않은 대상자까지 지자체 담당자들이 확인해야 할 필요성이 제기된 것이다.

두 번째는 지자체 복지상담 업무와 관련된다. 업무량에 비해 복지 인력이 부족하다는 사실은 새로운 이야기가 아니다. 이러한 상황에서 사회적 욕구가 증가하고 있고 복지 업무도 지속해서 증가하고 있다. 더욱이 지역별로 업무량 편차가 큰 상황, 담당자의 역량에 따라 달라지는 서비스 품질 수준, 단순 업무 비중 증가 등으로 인해 담당자의 업무 부담이 증가했다. 이러한 상황에서 ICT 기술을 활용하여 단순 처리 업무는 줄이고, 업무 누락을 최소화하는 대신 심층적 접근이 필요한 업무에 집중하도록 지원하는 방안이 모색됐다.

이 과정에서 AI 활용 초기상담 정보시스템 구축이 제안됐다. 결국 AI 활용 초기상담 정보시스템은 콜 기반 대화 시스템을 활용하여 복지 사각지대 대상자의 발굴 규모를 현재보다 확대하여 잠재적 위기 대상자를 최대한 살펴보고, 지자체 공무원이 이들을 상담하는 과정을 지원하기 위해 구축됐다.

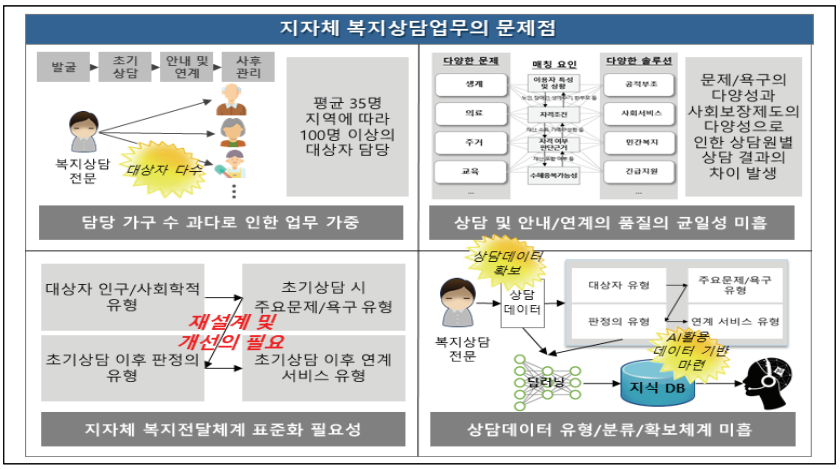
AI 활용 초기상담 정보시스템이 운영되는 방식은 다음과 같다(보건복지부 보도자료, 2024.11.25.). 잠재적 위기 대상자들에게 AI 서비스를 통해 전화로 연락하기 전에 지자체가 전화 초기상담을 진행할 예정이라는 문자 메시지를 사전에 발송한다. 인공지능 기술을 활용한 전화를 수신하는 대상자들이 낯선 번호에서 오는 전화를 받지 않을 가능성을 염두에 뒀다. 이렇게 사전 메시지를 통해 AI를 활용한 전화 응답률이 낮아지지 않도록 예방함과 동시에 관할구역 담당 공무원의 연락처를 남겨서 도움이 필요한 대상자들이 필요한 경우 주민센터에 연락할 수 있도록 하겠다는 취지도 있다.

[그림 3-4] 복지 사각지대 발굴 문제점



출처: 한국사회보장정보원 (2024.4.1.). "제안요청서-AI활용 초기상담정보시스템 운영지원 사업". p.8.

〔그림 3-5〕 지자체 복지상담 업무의 문제점



출처: 한국사회보장정보원. (2024.4.1.). "제안요청서-AI활용 초기상담정보시스템 운영지원 사업". p.9.

사전 문자 메시지를 보낸 이후에는 인공지능 기술을 활용하여 대상자에게 전화를 걸어 현재 도움이 필요한 상황인지 여부를 확인하는 상담이 진행된다. 몇 가지 질문과 답변이 오간 후 초기상담을 완료하였다면 음성으로 녹음된 상담 내용이 담당 공무원에게 자동으로 전달된다. 담당자는 상담 결과를 확인하고 대상자의 가구를 방문할 것인지, 심층 상담을 진행할 것인지 판단하여 실행에 옮긴다. 이와 같은 업무 흐름을 도식화한 〈표 3-11〉과 업무 흐름을 파악할 수 있는 시스템 개념도 〔그림 3-6〕을 참고할 수 있다.

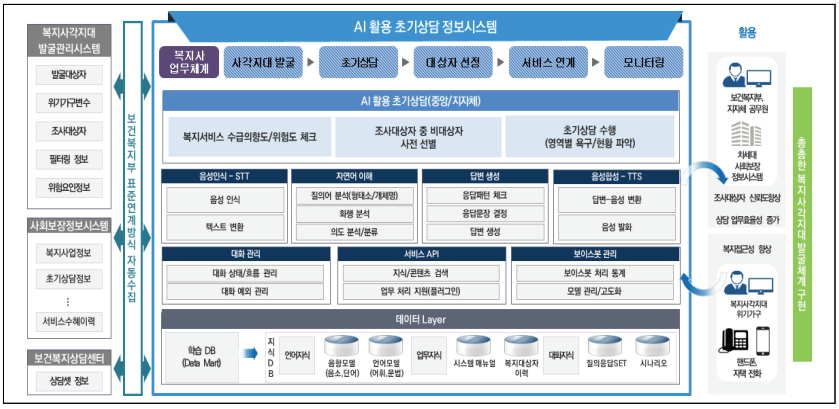
초기상담 시나리오를 바탕으로 전화 메시지 내용을 구체적으로 살펴보면 다음과 같다. 상기하였듯이 사전 문자를 통해 상담 전화가 갈 것이라는 메시지를 위기 대상자에게 전달하고 이후에 AI를 활용한 음성으로 전화가 가는데, 전화의 발신자는 "읍면동 AI 복지상담"으로 나타난다. AI 음성 서비스는 소속 지자체와 전화를 건 목적을 설명한 이후에 연락이 대상자에게 정확하게 갔는지 확인하기 위해 본인 여부를 묻고 수신자에게 초기상담을 희망하는지를 질의한다.

〈표 3-11〉 AI 활용 초기상담 정보시스템 업무 흐름

중앙발굴 대상 통보, 자체 발굴	⇒	대상자 AI 상담 요청, 사전 SMS 발송	⇒	발굴 대상에게 AI 초기 상담 전화	⇒	AI 초기상담 결과 제공	⇒	심층상담 및 복지지원
복지부, 지자체		지자체, 행복이음		행복이음 (한국사회보장정보원)		행복이음→지자체		지자체

출처: 보건복지부. (2024.7.22.). 보건복지부 보도자료. "인공지능(AI) 초기 복지상담 전화로 위기가구 지원에 나선다". p.4.

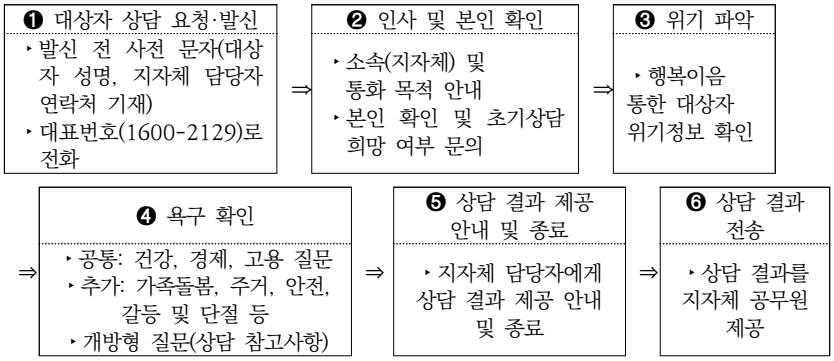
〔그림 3-6〕 시스템 개념도



출처: 한국사회보장정보원. (2024.4.1.). "제안요청서-AI활용 초기상담정보시스템 운영지원 사업".

AI 음성 서비스는 통화 과정에서 행복e음을 통해 대상자의 위기 정보를 확인하는데, 수신자가 언급하는 욕구를 중심으로 현재의 어려움을 파악하고 현재 지원받을 내용이 있는지 판단한다. 공통으로 경제 상태나 건강 문제, 고용 관련 욕구를 확인하고 추가 사항으로 주거, 돌봄, 안전 등의 사항을 질의한다. 마지막으로 지자체의 공무원이 참고할 수 있도록 심층 상담에 활용할 수 있는 개방형 질문을 한다. AI를 활용한 전화 상담이 종료되면 지자체 담당자에게 그 결과가 전송된다. 지자체 담당자는 이를 바탕으로 심층 상담의 필요 여부를 판단하여 필요한 경우 후속 조치를 진행한다.

〈표 3-12〉 AI 초기상담 시나리오 흐름



출처: 보건복지부. (2024.7.22.). 보건복지부 보도자료. “인공지능(AI) 초기 복지상담 전화로 위기가구 지원에 나선다”. p.4.

인공지능(AI) 활용 초기상담 정보시스템은 2023년 7월부터 2024년 5월까지 시스템 구축이 진행된 후 2024년 7월부터 시범사업을 진행하였다. 제4차 복지 사각지대 발굴 기간인 2024년 7월 22일~9월 13일에 101개 시군구는 인공지능(AI) 활용 초기상담을 운영하며 그 이후로 대상 지자체를 차츰 넓히면서 11월 말부터는 전국으로 확대할 계획이다(보건복지부, 2024.7.22).

시범사업을 진행하고 있는 시군구를 대상으로 시행 결과에 대한 평가는 이루어지지 않았다. 시범 적용하는 과정에서 현장 적용 결과에 대해 조사한 내용을 참고할 수 있다. 당시 대상자 수신율은 2024년 1~2월 49.58%에서 2024년 5~6월 64.28%로 증가했다(〈표 3-13〉 참고). 상담 거부 비중 역시 2024년 1~2월 16.77%에서 2024년 5~6월 11.11%로 지속해서 감소하였다. 추가 상담 요청 비율은 2024년 1~2월 11.63%에서 2024년 5~6월 19.09%로 증가하는 등 해당 기간에 일정한 개선이 이뤄졌다(이우식, 2024). 다만, 시범 적용 과정이고 평가의 절대 기준이 없기 때문에 현재 단계에서 이 수치를 긍정적으로 해석하기에는 조심스럽다.

〈표 3-13〉 AI 초기상담 시나리오 시범 적용 결과

단계	상담 결과	2023.12.		2024.1.~2.		2024.3.~4.		2024.5.~6	
		건수	비율	건수	비율	건수	비율	건수	비율
발신/인사	상담 거부	552	16.23	727	16.77	2386	11.73	299	11.11
본인	본인 불일치-가족	98	2.88	120	2.77	734	3.61	108	4.01
확인	본인 불일치-타인	12	0.35	19	0.44	67	0.33	15	0.56
욕구	사용자 단선 (안내멘트 시 포함)	2,047	60.19	2,436	56.21	12,944	63.64	1,643	61.03
확인	(미인식) 시나리오와 무관한 대화	71	2.09	28	0.65	150	0.74	25	0.93
	(욕구확인) 전체 거부	10	0.29	12	0.28	653	3.21	88	3.27
상담/완료 (추가상담/여부)	(정상) 추가 상담 요청	278	8.17	504	11.63	3387	16.65	514	19.09
	(상담 완료) 추가 상담 거부	269	7.91	480	11.08	0	0.00	0	0.00
통화/종료	상담 최대 시간 (2분 30초) 초과	64	1.88	8	0.18	17	0.08	0	0.00
대상자 수신 건수		3,401		4,334		20,338		2,692	
총콜건		6,062		8,741		33,868		4,188	
수신율		56.10		49.58		60.05		64.28	

출처: 이우식. (2024). 사회보장행정에서의 인공지능 적용 동향과 함의 한국보건사회연구원 세미나 자료.

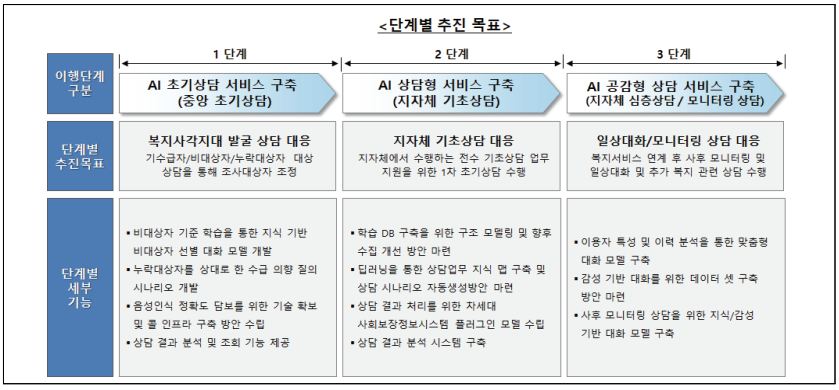
AI 활용 초기상담 정보시스템은 전체 3단계 목표에서 1단계 수준이다. 1단계는 'AI 초기상담 서비스 구축'을 목표로 중앙에서 수행하는 초기상담 중심으로 진행된다. 2단계는 'AI 상담형 서비스 구축'을 목표로 하고 있으며 중앙을 넘어 지자체의 기초상담으로 그 범위를 확장한다. 3단계에서는 'AI 공감형 상담 서비스 구축'을 목표로 하며 지자체 기반 대상자들에 대한 심층 상담 및 모니터링 상담이 진행될 예정이다(이우식, 2024).

현재 1단계에서는 복지 사각지대 발굴시스템을 통해 도출된 위기 대상자 가운데 기수급자나 비대상자, 누락 대상자 등에 대한 상담으로 그 범위가 한정된다. 1단계의 목표는 정확한 음성 인식과 콜 인프라 마련을 기본으로 하며, 비대상자와 누락 대상자에게 특화된 시나리오를 개발하고 상담 결과를 분석하는 기능을 개발하는 것이다. 2단계에서는 딥러닝을 활용하여 상담업무 지식 맵을 구축하고 자동으로 상담 시나리오를 생성하는 기술을 활용할 계획이다. 3단계는 심층 상담이 진행되는 단계로, 단순 상담을 넘어서 대상자에게 필요한 복지서비스를 연계하고 사후 모니터링까지 진행하는 기능으로 확장될 예정이다. 이 단계에서는 대상자 각자의 특성에 기반한 맞춤형 대화가 가능하게 하며 감성 기반 대화 모델을 구축할 계획이다.

AI 활용 초기상담 정보시스템은 복지 사각지대 발굴 과정에서 발생한 위기 대상자의 복지서비스 수요를 파악하기 위해 지자체 공무원이 아닌 AI 기술에 기반하여 AI가 초기상담을 진행하여 일선 현장의 상담업무 지원을 하고 있다는 의미를 찾을 수 있다. 또한, 기존의 복지 사각지대 발굴시스템을 둘러싼 복지 업무 과정에서 새로운 이슈들이 제기되고 이를 바탕으로 또 다른 신기술인 아웃바운드 콜을 활용하여 신기술의 추가 및 확장을 보여주는 사례라고 볼 수 있다. 사업 초기 상황이기 때문에 효과 여부에 대해서는 섣불리 평가하기가 어렵다. AI에 기반한 음성 서비스가 이용자에게 단순 정보 제공을 넘어서, 이제는 복지 대상자들이 자신들의

어려움을 명확하게 전달할 수 있도록 이끌어낼 수 있는 단계까지 갈 것인지, 결과가 주목된다.

[그림 3-7] AI 활용 초기상담 정보시스템 추진 목표



출처: 한국사회보장정보원. (2024.4.1.). “제안요청서-AI활용 초기상담정보시스템 운영지원 사업”. p.16.

3. 소결

Hila Mehr(2017, p. 4)는 인공지능의 활용이 적합한 정부 문제 유형을 다음과 같이 제시했다. 자원할당이 필요한 경우, 대용량의 데이터 세트가 있는 경우, 전문가가 부족한 상황, 시나리오가 예측 가능한 경우, 절차적인 측면, 다양한 데이터가 보유되어 있는 상태 등이다. 즉, 이와 같은 경우에 공공 서비스 분야에 인공지능을 활용하면 작업속도를 높이고 문제 처리 시간을 절약할 수 있다. 대용량으로 축적된 데이터 세트를 통해 문제를 효과적으로 해결할 수 있으며, 전문가의 역할을 인공지능이 대신해 줄 수 있다. 예측 가능한 시나리오를 통해 다가올 문제에 대비할 수 있고 반복적으로 수행되는 업무를 간단히 처리할 수 있다. 데이터가 대량으로 축적된다면 데이터를 활용한 문제 해결도 가능하다.

〈표 3-14〉 인공지능 활용이 적합한 정부 문제 유형

항목	내 용
자원할당	• 작업속도를 높이기 위해 행정 지원이 필요 • 지원이 충분치 않으므로 질의나 응답시간이 오래 걸림
대용량의 데이터 세트	• 자료의 양이 너무 많아서 업무 담당자가 능률적으로 일을 하지 못함 • 내·외부 데이터가 결합된다면 문제 해결의 통찰력이 향상됨 • 오랜 시간에 걸쳐 데이터는 고도로 구조화되었음
전문가 부족	• 기본적인 질문에 답을 제공하기 때문에 전문가의 시간을 단축 • 특수(niche)문제는 전문가를 지원하도록 학습됨
예측 가능한 시나리오	• 과거 데이터에 기반하여 상황의 예측이 가능함 • 예측은 시간에 민감한 답변에 도움이 됨
절차	• 과업은 본질적으로 반복해서 수행됨 • 투입/산출은 이진 답변을 지님
다양한 데이터	• 데이터는 시각, 공간, 청각, 언어 정보를 포함 • 질적 및 양적 자료는 정기적으로 요약되어야 함

출처: Hila Mehr. (2017). p.4.

앞에서 살펴본 세 가지 사례도 여기에 적용될 수 있는데 이를 바탕으로 인공지능을 활용한 사회보장 행정의 긍정적인 측면을 정리하면 다음과 같다.

우선, 효율성을 높일 수 있다는 큰 장점이 있다. 인공지능을 활용해 업무의 능률을 높이면 이것이 서비스 향상으로 이어질 수 있다. 단순 반복적인 업무 프로세스를 자동화한다면 인간이 더 잘하는 심층 업무에 더 많은 시간과 노력을 투자할 수 있다. AI 활용 초기상담 정보시스템 사례에서도 단순 상담을 인공지능에 맡기고 심층 상담이 필요한 대상들에 집중할 수 있는 체계를 만들어 놓았다. 인공지능의 활용으로 자원이 희소한 환경을 극복할 수 있는데 인력난을 극복한 업무 처리가 가능하다.

다음으로, 자료 수집이나 대용량 데이터를 기반으로 한 의사결정을 지원해 준다는 점이다. 빅데이터 분석 기술이 없었다면 담당자들은 각기 다른 경험에 의존하여 의사결정을 하게 되었을 것이다. 복지 사각지대 관리 시스템의 경우, 대용량 데이터 수집과 함께 이를 활용한 고위험군 도출로 예측 모델을 통한 의사결정을 지원해 준다. 이렇게 위기 집단에 속할

가능성이 높은 대상자들을 지자체에 전달하는 것이 가능하게 된다. 여기에 소요되는 업무량은 개인이 담당하기에는 불가능한 수준이다.

맞춤형 서비스를 제공할 수 있다는 점도 긍정적인 측면이 될 수 있다. 빅데이터 기반의 분석은 기본적으로 개인에게 최적화된 정보를 제공하는 데 유리하다. 인공지능은 대용량의 데이터 처리를 통해 서비스 이용자들의 평균적인 욕구가 아닌 개별적인 필요를 파악하고 그들에게 적합한 서비스를 제공하는 것이 가능하다. 개개인이 지닌 이력과 특장점, 환경을 파악하고 적합한 일자리를 추천해 주는 서비스는 맞춤 서비스의 대표적인 사례이다. 개인에게 특화된 정밀한 서비스는 이용자를 넘어 국민에게 서비스에 대한 신뢰를 심어줄 수 있다.

이상의 인공지능 기반 서비스는 빠른 속도로 이루어지기 때문에 업무 담당자 입장에서는 시간을 절약하고, 이용자 입장에서는 신속한 서비스를 제공받을 수 있다는 장점이 있다.

한편, 지속해서 고려해야 하는 점도 동시에 존재한다. 먼저, 많은 업무의 자동처리가 곧 불필요한 업무의 사라짐을 의미하는지 현실적으로 짚어볼 필요가 있다. 사회보장 행정 분야에 인공지능을 활용하면 할수록 단순 업무가 사라지므로 서비스 이용자에게 더 양질의 서비스를 제공할 수 있는 여건이 조성되는가? 앞으로 더 많은 시간이 필요한지 모르겠으나 현재까지는 그런 기미가 보이지 않는다. 객관적인 자료의 수집과 함께 정밀한 조사가 뒤따라야 하겠지만, 현장의 목소리는 정보시스템이 다루는 업무량의 증가에 비례해서 부수적인 업무들이 발생하는 경향들을 호소하고 있다. 즉 새로운 기술이 도입되어 일선 현장에 적용되는 것은 곧 새로운 업무가 생겨나는 것으로 인식하는 것이 일반적인 경향이다(김은하, 2022; 최정은 외, 2022; 함영진 외, 2023)

다음으로는, 빅데이터를 활용하는 과정에서 발생할 수 있는 편향이나 편견의 문제가 있다. 대용량 데이터가 곧 현실 세계를 정확히 포착할 수 있는 것은 아닌데, 데이터의 부족이나 수집된 데이터의 특성, 모델의 오류 등 다양한 이유가 있을 것이다. 사회보장 행정 분야에서 아직 이에 대한 구체적인 논의는 없지만 데이터의 특성을 고려한다면 문제가 될 충분한 가능성이 있다. 사회보장 행정의 문제를 해결하기 위한 데이터 성격이 포괄성, 다양성, 충분성 수준에서 높지 않기 때문이다. 아직은 데이터 학습 과정에서 데이터의 편향을 최소화하기 위한 구체적이고 실질적인 노력을 찾아보기는 어렵다. 사회보장 행정 분야에서 인공지능의 활용이 보편화된다면 편향성 같은 이슈가 제기될 가능성이 높을 것이므로 현 단계부터 준비해 나갈 필요가 있을 것이다.

사생활 침해도 중요하게 거론되는 이슈이다. 개인 수준의 데이터를 수집하고 분석하여 서비스를 제공하는 과정에서 사생활이 드러나는 경우가 발생할 수 있다. 복지 사각지대 발굴시스템의 경우, 알고리즘 기반으로 도출된 대상자들을 방문하여 상담하는 과정에서 개인정보나 사생활의 문제로 불만을 표출하는 사례들이 종종 발견된다고 보고된다(최정은 외, 2022). 대상자들에게 불쾌감을 안겨준다면 정부 서비스에 대한 신뢰도도 낮아지게 될 뿐만 아니라 서비스의 효과성도 떨어질 수밖에 없다. 현재까지는 현장의 일부 대상자에게서 제기되는 이러한 이슈에 대해 인식만 하고 있을 뿐, 적절한 대안은 없는 듯하다. 법률적 근거를 바탕으로 진행하고 있는 사업이라고 할지라도 필요한 경우 맥락에 대한 정보를 알 수 있도록 해야 하며, 서비스를 제공하는 과정에서 활용되는 개인정보가 관계자 이외에 누출되는 일이 없도록 해야 할 것이다.

끝으로, 알고리즘에 따라서 어떤 결정이나 판단이 이루어진 경우에 왜 그러한 결과가 도출되었는지를 설명할 수 있어야 한다. 설명 가능성은 특정한 데이터가 분석되고 처리되는 과정과 그 방법에 대한 이해가 있다는 의미이다. 사회보장 분야에서는 인공지능이 활용된다고 하더라도 딥러닝 수준까지 활용되는 경우는 많지 않기 때문에 인공지능의 설명 가능성에 대한 문제가 중요하지 않을 수도 있다. 그러나 기술의 발전 속도를 고려한다면 알고리즘의 설명 가능성에 대한 이슈가 언제 드러날지 모르는 일이다. 사회보장 행정에서의 주 대상은 취약계층이라는 점을 고려한다면 의사결정에 활용하기 위해 인공지능을 활용하는 경우 서비스 이용자에게 미치는 영향이 상대적으로 클 수 있다. 그리고 이러한 이슈는 인공지능의 책임성이나 투명성 등과 모두 연결된다는 점에서 설명 가능성을 충족시키기 위한 기술적 대비가 필요하다.

제2절 국외 인공지능 기술의 사회보장 행정 적용 사례

사회보장 영역에서 인공지능이 사용되기 시작한 시점은 2008년 이후지만 본격적인 확산기는 2017년 이후였다(Zaber, Casu, Brodersohn, 2024). 인공지능이 활용되는 사회보장 영역은 가족급여, 보건, 산업재해, 연금, 실업, 보편적 급여 등이다. 2020년 이전에 사회보장에서 영역에서 활용된 인공지능 기술은 챗봇이었다. 챗봇은 2020년 이후에 코로나 범유행 상황에서 폭증한 급여 신청을 기관들이 대응하는 과정에서 활용도가 더욱 높았다. 일부 기관들은 챗봇의 기능을 강화하는 과정에서 생성형 인공지능(generative artificial intelligence)을 활용하기 시작했다.

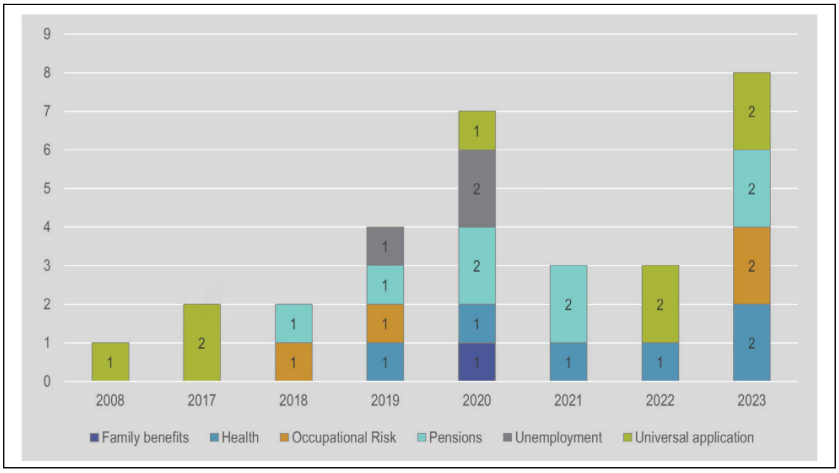
Zaber, Casu, Brodersohn(2024)은 공공 부문에서 생성형 인공지능이 적용될 영역을 다음과 같이 꼽았다(p. 10).

첫째, 창의적인 콘텐츠 제작. 생성형 인공지능을 활용하여 교육자료를 제작하거나, 시민 인식 개선 방안을 고안하거나, 소셜미디어 캠페인을 위한 매력적인 콘텐츠 생성 작업을 주도한다.

둘째, 데이터 증강. 합성 데이터 생성을 사용하여 의료 및 환경 감시 같은 중요한 영역에서 사용되는 인공지능 모델의 성능을 향상한다.

셋째, 맞춤형 사용자 상호작용. 개인의 선호도와 요구 사항에 맞게 정부 서비스 및 정보 배포를 맞춤화한다.

[그림 3-8] 인공지능 기술의 적용 추이



출처: Zaber, Casu, Brodersohn, (2024). "Artificial Intelligence in social security organizations". p.21. International Social Security Association.

Zaber, Casu, Brodersohn(2024)은 사회보장 영역에서 인공지능을 활용할 영역을 다음과 같이 제시했다(pp. 22-23).

첫째, 서비스 제공(Service delivery). 기관이 대상에게 더 쉽게 접근할 수 있는 채널을 활용하여 더 적절하고 더 나은 서비스를 다양한 유형의 대상에게 제공한다.

둘째, 자동화 및 사례 관리(Automation and case management). 사회보장 기관이 인공지능을 사용하여 사례 처리 방식을 자동화하고 개인의 서비스 후속 조치를 위한 고충 대응 지원을 제공한다.

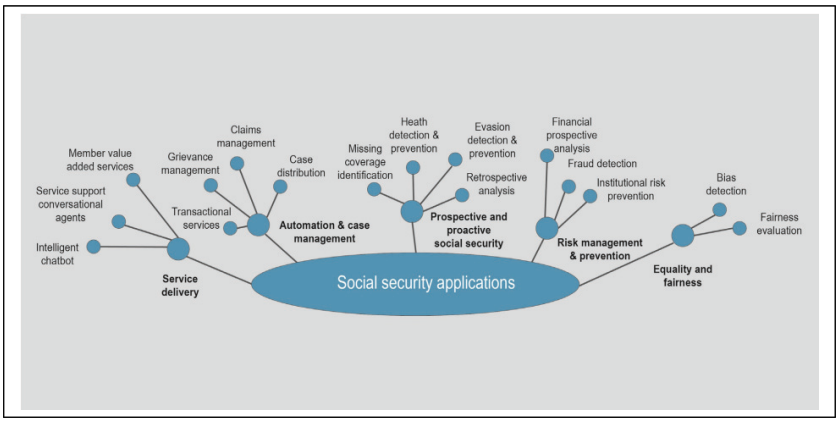
셋째, 전향적이고 능동적인 사회보장(Prospective and proactive social security). 사회보장 기관에 인사이트, 비전 및 잠재적 결과를 파악할 수 있는 전망 분석을 위한 도구를 제공한다. 기관은 이러한 도구를 활용하여 코호트와 개인의 삶을 선제적이고 능동적으로 개선하기 위한 접근 방식을 수립한다.

넷째, 위험관리 및 예방(Risk management and prevention). 기관이 위험을 식별하고 위험을 완화하거나, 위험에 대응할 수 있는 역량을 강화한다. 더 나은 서비스를 제공할 수 있도록 회원 정보를 분석할 수 있도록 지원한다.

다섯째, 평등과 공정성(Equality and fairness). 공정성과 형평성의 원칙을 지킬 수 있는 기관의 의무 측면에서 다양한 인공지능 솔루션과 프로그램 및 대응을 평가

Zaber, Casu, Brodersohn(2024)가 제시하는 범주들이 깔끔하지는 않다. 영역별로 해당 사례가 제시되지 않아서, 직관적으로 이해하기가 쉽지도 않다. 그러나 〔그림 3-9〕를 보면, 다섯 개 영역에서 파생되는 인공지능 기술의 활용 내용을 보다 상세하게 파악할 수 있다. 다섯 영역 가운데 하나인 서비스 제공 분야에서 인공지능은 가장 활발하게 사용된다.

[그림 3-9] 인공지능 기술의 적용 영역



출처: Zaber, Casu, Brodersohn. (2024). "Artificial Intelligence in social security organizations". p.23. International Social Security Association.

서비스 제공(Service delivery) 가운데서도 챗봇은 인공지능이 가장 활발하게 사용되는 영역이다. 특히, 흥미롭게도, 남미 국가에서 챗봇의 활용도가 높았다(ISSA, 2021). 브라질, 아르헨티나, 파나마, 우루과이 등에서 국민들의 급여 관련 문의나 민원을 응대하는 데 챗봇이 활용됐다. 이를테면, 코로나19 범유행 위기 상황에서 브라질 국립사회보장원(Instituto Nacional de Seguro Social, INSS)은 인공지능 기반의 가상 비서 챗봇인 Helô를 도입했다. INSS가 관리하는 애플리케이션인 Meu INSS에 대한 문의에 응답하도록 설계된 Helô는 사용자와의 상호작용을 개선하고 더 복잡한 응답도 가능하도록 기능이 점진적으로 확장됐다. 초기 평가 결과 Helô는 첫 달에 100만 건 이상의 상담을 처리했다(ISSA, 2021). 이후 3년 동안 Helô는 90개의 주제에 대해 660만 건의 상담을 처리했고, 사용자 만족도가 0~5점 척도 기준으로 긍정적인 3.8점을 받았다(INSS. 2023).

챗봇의 활용은 소위 '선진국'에 국한되지 않는다. 말레이시아의 직원연금기금(EPF)은 ELYA 스마트 챗봇을 운영하고 있다(Zaber, Casu, Brodersohn, 2024; ISSA, 2020). ELYA는 자연어 처리(NLP)를 사용하면서 인공지능으로 구동되고 라이브 채팅으로 지원되는 사회보장 기관 최초의 이중 언어 가상 비서(VA)이다(ISSA, n.d.). 고객이 직접 EPF 상품 및 서비스에 대한 정보에 액세스할 수 있도록 하는 방식으로 상담센터의 부담을 덜어주는 것이 목표였다. ELYA는 기본 챗봇에서 3단계에 걸쳐 업그레이드되어 복잡한 문의를 위한 자문 챗봇까지 가능하게 진화했다. ELYA는 실시간 상호작용, 다국어 지원, 연중무휴 24시간 상담 서비스를 제공했다.

보건 영역에서는 인공지능이 응급실까지 들어왔다. 호주 뉴사우스웨일스주 공공 의료 시스템의 경우 응급실에서 패혈증 조기 발견을 목표로 하는 머신러닝 프로토타입 제품을 개발했다(Zaber, Casu, Brodersohn, 2024). 이름은 eHealth NSW였다. 인공지능 프로토타입은 2017~2019년 네 병원에서 추출한 과거 데이터를 활용해서, 로지스틱 회귀 및 XGBoost (오픈 소스 시프트웨어 라이브러리) 알고리즘을 사용했다. 이 도구는 응급실 대기실에서 패혈증 발병 위험이 있는 환자를 조기에 발견하여 패혈증 관련 사망, 중환자실 입원 및 재입원을 줄이는 것을 목표로 한다.

복지 급여 수급자를 포착하는 데에도 인공지능은 이미 활용되고 있다. 캐나다 고용사회개발부(Employment and Social Development Canada)는 저소득 노인을 위한 급여인 보장소득보조금(Guaranteed Income Support) 영역에 인공지능 기술을 적용했다(Zaber, Casu, Brodersohn, 2024). 급여 수급 자격이 있는 노인을 신속하게 식별하기 위한 목적이었다. 머신러닝 모델을 통해서 2,000명 이상의 수급 대상자를 식별했다. 정확도는 92~98%였다(Zaber, Casu, Brodersohn, 2024).

사례 관리 영역에서도 인공지능의 역할이 확대되고 있다. 오스트리아의 사회보험연합은 청구 자동 처리를 지원하고 의사와 환자를 매칭하는 인공지능 기반 시스템을 구현했다(Zaber, Casu, Brodersohn, 2024). 인공지능을 활용한 의료비 환급 관리 사례가 그 예이다. 인공지능 솔루션을 도입하기 전에는 수작업 처리 시간이 길어 환급받기까지 몇 달씩 기다려야 했다. 플랫폼을 도입한 후에는 추가 인력 없이도 처리 시간이 며칠로 크게 단축됐다. AI 기반 접근 방식은 광학 문자(Optical Character R) 및 개체 인식 같은 기술과 오픈소스 도구 및 언어를 사용하여 문서를 수집하고 처리함으로써 효율성과 투명성을 개선했다.

〈표 3-15〉 사회보장 영역의 도전에 대한 인공지능의 대응

사회보장 영역	인공지능의 해법	상세 내용
연금	연금 행정 자동화	연금 관리 및 연금 지급을 자동화하여 퇴직자에게 적시에 급여 지급. 관리 비용 절감, 효율성 개선
아동 돌봄 및 교육	인공지능 아동 돌봄 및 교구 활용	스마트 모니터링 시스템으로 어린이의 안전 보장. AI 기반 교육 플랫폼으로 어린이에게 개인화한 학습 경험 제공 및 발달 과정 모니터링
급여 수급자 확인	자동화한 수급자 확인	인공지능 기반 시스템으로 본인 확인 절차 간소화. 신원, 문서, 데이터를 신속하고 정확하게 확인하여 부정수급 가능성 차단
고령 돌봄	돌봄을 위한 로봇 활용	인공지능 기반 로봇 공학은 노약자에게 신체적, 인지적 지원 제공 및 일상 업무 지원, 복약 알림, 말벗 기능 수행
정서적 돌봄	정신건강을 위한 챗봇	인공지능 기반 챗봇이 정서적 지원 및 상담 서비스 제공.

출처: Zaber, Casu, Brodersohn. (2024). “Artificial Intelligence in social security organizations”. p.29. International Social Security Association의 표 3의 내용을 번역하고 일부 편집함

인공지능이 그리는 사회보장의 미래가 장밋빛인 것만은 아니다. 네덜란드와 덴마크의 사례는 인공지능 적용에 따른 인권 침해, 정보 유출, 공공성 훼손에 관한 또 다른 도전을 보여준다.

먼저, 네덜란드의 사례를 살펴보겠다. SyRi(Systeem Risico Indicatie)는 중앙정부와 지방정부가 사회보장 및 소득 관련 제도, 세금 및 사회보험료, 노동법 분야에서 부정행위를 방지하고 대처하기 위해 도입한 인공지능 시스템이다. SyRi는 다양한 공공기관의 데이터를 연결하고 분석하여 위험 보고서를 생성함으로써 재정의 오용을 방지하고, 부당한 이익을 탐지하는 데 사용된다(Rechtbank Den Haag, 2020). 이 시스템에 참여하는 기관은 지방정부, 네덜란드 조세 및 관세청, 사회보험은행(Social Insurance Bank), 이민 및 귀화 서비스, 고용보험청, 사회 및 고용감찰청 등과 같은 감독 기관이다. 이 기관들로부터 수집되는 데이터는 보건, 재정, 교육, 재정 지급, 고용 등 방대한 영역에 걸쳐 있다. 알고리즘을 통한 데이터 분석 결과를 바탕으로 특정 사례에 대한 '위험 보고서'가 제출되면, 관련자는 부정행위의 가능성이 있는 것으로 판단되어 정부의 조사 대상으로 간주된다(Appelman et al., 2021, p.263).

이러한 알고리즘을 활용한 분석은 부정수급을 적발하는 데 효과적이라는 것이 네덜란드 정부의 입장이다. 이 시스템은 주로 부정수급 가능성이 클 것으로 예상되는 가난하거나 취약한 사람들에 초점을 맞추어 조사를 수행한다는 점에서 논란이 많았다. 이와 관련하여 Platform Bescherming Burgerrechten, Nederlands Juristen Comité voor de Mensenrechten 등 네덜란드 6개 시민단체는 SyRi의 문제점을 지적하며 공동행동을 시작했다. 단체들은 2018년 3월 네덜란드 정부를 상대로 이 시스템이 "모든 개인은 사생활과 가족생활 및 주거와 통신을 존중받을 권리가 있다"고 규정한 유럽인권협약(European Convention of

Human Rights) 제8조를 위반했다는 취지로 헤이그 지방법원에 소송을 제기했다(Van Bekkum, Borgesius, 2021). 2020년 헤이그 법원은 정부 기관들이 개인정보를 공유하는 과정에서 투명성이 부족하고 개인정보에 대한 적절한 안전장치가 부족하다는 이유로 SyRi가 유럽인권협약을 위반한다고 판결했다.

2018년 덴마크의 지방정부인 Gladsaxe는 사회경제적으로 취약한 아동을 추적하기 위해 알고리즘을 활용한 정책 실험을 추진하였다. 덴마크의 수도 코펜하겐(Copenhagen) 근교에 위치한 Gladsaxe 정부는 실업 및 의료를 비롯한 다양한 사회경제적 영역의 데이터를 결합하여 200개 이상의 위험 지표를 분석하는 머신러닝 모델을 구축하였고, 이를 활용하여 가정폭력이나 학대의 위험이 큰 아동을 찾으려고 시도했다. 이 모델은 가정이나 부모의 상황들을 점수화하였는데, 예를 들면 정신 질환의 경우 3,000점, 실업은 500점, 예약된 의사의 진료 불참은 1,000점, 예약된 치과 진료 불참은 300점 등이 부여되었다. Gladsaxe 알고리즘 모델은 이러한 점수를 바탕으로 위험한 상황에 처한 아동들을 판별하고자 했다(Thapa, 2019).

이러한 시도는 시민단체, 학계의 신랄한 비판을 받았고, 물론 대중으로부터도 심각한 반발을 낳았다. 특히 이 시스템은 점수를 통해 개인과 지역을 구분함으로써 자유 민주주의에 위협이 된다는 의견도 있었다(Mchangama and Liu, 2018). 이러한 비판에도 불구하고 덴마크 정부는 위험에 처한 아동을 조기에 발견할 수 있다는 장점을 강조하며 Gladsaxe 모델을 전국적으로 확대할 계획을 세웠다(Bendixen, 2018; Jørgensen, 2021). 그러나 아동의 복지와 발달을 평가하는 데 사용되는 Gladsaxe 모델에 대한 공개적인 검증이 이루어지면서 상황은 바뀌었다. 이 모델을 통해 작성되는 개별 자금 평가서에 포함된 다양한 정보들이 부

모들 모르게 활용되고 저장된 사실이 밝혀졌다. 이로 인해 2018년 12월 Gladsaxe 모델은 중단되었다(Algorithm Watch and Bertelsmann Stiftung, 2020).

제4장

국내·외 인공지능 기술에 대한 규제

제4장 국내·외 인공지능 기술에 대한 규제

제1절 국내 인공지능 기술에 대한 규제

이 절에서는 국내의 인공지능과 관련된 규제나 규율에 대해 법적 강제성과 구속력의 정도에 따라 단계별로 살펴보고자 한다. 여기에는 법률과 명령 같이 전 국민이 반드시 준수해야 하는 강제력을 지닌 규제 방안부터, 법적 준수를 돕거나 방향성을 제공하며 강제성은 없지만 준수하지 않았을 경우 제재나 평가상 불이익을 받을 수 있는 지침 및 가이드라인, 그리고 강제성 없는 권고사항으로 자율적 참여를 유도하는 선언까지 다양한 수준의 내용이 포함된다. 각각은 상호 보완적인 역할을 하며 인공지능이 책임감을 지니고 신뢰할 수 있는 방향으로 발전하도록 지원하고 있다.

1. 법률

인공지능에 관한 국내의 법률은 존재하지 않는다. 다만 제21대 국회에서 인공지능 법안이 2020년부터 발의됐으며 국회 회기가 종료됨에 따라 모두 폐기되었다. '인공지능'이 제명에 명시된 경우만 한정한다면 국내 인공지능법안은 제21대 국회에서 2020년부터 발의됐다. 이 중 대학 설립이나 교육 진흥 등의 목적과 같이 인공지능 자체에 초점을 두었다고 보기 어려운 법안[12]을 제외한 총 9건에 해당하는 법률안의 주요 내용과 각 법안이 담고 있는 인공지능에 대한 규제 내용을 정리하면 다음과 같다.

12) 「한국인공지능·반도체공과대학교법안」, 「인공지능 집적단지의 육성에 관한 특별법안」, 「인공지능교육진흥법안」 등

〈표 4-1〉 제21대 국회에서 발의된 인공지능 관련 9건 법률안 주요 내용

법률명 / 대표발의 의원, / 발의일	주요 내용	규제 내용
인공지능 책임 및 규제법 / 안철수 의원 / 2023.8.8.	인공지능을 '금지된 인공지능', '고위험 인공지능', '저위험 인공지능'의 세 가지 유형으로 구분하여 정의하고 있으며, 각각의 유형에 대해 차등화된 의무사항을 규정	· 인공지능 위험도를 구분(금지된 인공지능, 고위험 인공지능, 저위험 인공지능)하여 각 범주에 따라 규제 · 고위험 분야에서 사용하는 AI는 특별한 규제 · AI 작동 방식에 대한 투명성 및 사용에 대한 정보 제공
인공지능책임법안 / 황희 의원 / 2023.2.28.	고위험 인공지능에 관한 기본계획 수립, 인공지능 분야에 전문성을 갖춘 위원회의 설치·운영, 인공지능에 관한 분쟁조정을 위한 인공지능분쟁조정위원회 운영	· 고위험 AI에 대한 이용자 보호 및 정부의 규제·관리 · AI 기술의 안전성과 신뢰성을 확보하기 위한 규제
인공지능산업 육성 및 신뢰 확보에 관한 법률안 / 윤두현 의원 / 2022.12.7.	인공지능 관련 산업 진흥에 필요한 조치 규정, 사업자에게 신뢰성 확보를 위한 노력 의무 부과	· 고위험 AI 사용에 대한 사전 고지 의무 · 고위험 인공지능의 관리
알고리즘 및 인공지능에 관한 법률안 / 윤영찬 의원 / 2021.11.24.	알고리즘 및 인공지능의 부작용을 최소화(고위험 인공지능을 포함한 기술, 서비스에 대한 설명 요구권, 이의제기권, 거부권, 손해배상 청구권 등 규정)하면서 관련 산업을 육성하는 데 필요한 내용 규정	· 고위험 인공지능 심의위원회 설치 · 고위험 인공지능 기술 사용 시 이용자의 설명 요구권, 이의 제기권, 거부권 부여
인공지능에 관한 법률안 / 이용빈 의원 / 2021.7.19.	인공지능 산업 진흥 및 경쟁력 강화를 위해 정부와 지방자치단체에서 필요한 조치 등에 대해 규정	· AI 기술의 안전 확인 가이드라인 마련 · 비상시 AI 시스템을 정지할 수 있는 비상정지 시스템 구축
인공지능 육성 및 신뢰 기반	인공지능 관련 윤리 기준, 기술 표준 등을 마련하도록	· 신뢰성 전문위원회 설치로 인공지능의 공정성,

법률명 / 대표발의 의원, / 발의일	주요 내용	규제 내용
조성 등에 관한 법률안 / 정필모 의원/ 2021.7.1.	하고, 특수한 영역에서 활용하는 인공지능 개발, 제조, 유통 시 신고하도록 함	투명성, 책임성 확보 · 고위험 영역의 인공지능에 대해 이용자 고지 의무
인공지능 기술 기본법안 / 민형배 의원 / 2020.10.29.	인공지능 산업 진흥 및 육성을 위해 필요한 지원 등에 관한 내용 규정	· 인공지능 윤리 원칙 제정하여 인공지능 개발·사용 과정에 활용 · 고위험 인공지능에 대해 이용자에게 고지 의무 부여
인공지능산업 육성에 관한 법률안 /양향자 의원 / 2020.10.29.	인공지능 산업의 진흥 및 육성에 필요한 내용 규정	· 인공지능 기술로 인해 발생하는 인권 문제에 대해 보호 조치 마련 · AI 기술이 인간의 존엄성을 침해하지 않도록 관리 규정 마련
인공지능 연구개발 및 산업 진흥, 윤리적 책임 등에 관한 법률안 /이상민 의원 / 2020.7.13.	인공지능 산업의 진흥 및 육성에 필요한 내용 규정	· 인공지능이 인권과 존엄성을 침해하지 않도록 하는 국가나 지자체의 책무 규정 및 윤리적 원칙을 마련 · AI 발전의 역기능에 대비하고 국제적인 윤리적 논의를 선도하기 위한 체계 구축

출처: 법제처 미래법제혁신기획단. (2024). “인공지능(AI) 관련 국내외 법제 동향”. p.40. 발췌 및 연구자 추가 작성

인공지능 기술 활용에 관한 규제에 초점을 맞춰서 본다면, 과거보다 최근에 발의한 법안일수록 인공지능 규제나 규율에 관한 사항을 더욱 강조하고 있음을 확인할 수 있다. 이러한 경향은 인공지능이 우리 사회에 밀접하게 활용되는 과정에서 규제의 필요성을 점차 인지하고 공유하고 있음을 반영한다고 하겠다. 가장 최근 법률안인 안철수 의원의 대표발의(2023.8.8.) 「인공지능 책임 및 규제 법안」만 보아도, 인공지능을 금지된 인공지능과 고위험 인공지능, 저위험 인공지능 등으로 각기 분류하고 있으며 범주에 따른 규제를 언급하고 있어 가장 구체적인 규제를 제시한다.

비교적 모든 법안이 공통으로 인공지능의 안전성이나 투명성, 신뢰성에 대한 가치를 지향하기 위한 조치나 제재 가능성을 담고 있다. 특히 고위험 인공지능에 대한 정의를 제시하지 않았지만, 고위험 인공지능에 대한 관리나 활용 시에 이용자에게 고지를 해야 한다는 내용이 확인된다. 그 밖에 이용빈 의원이 대표발의(2021.7.19.)한 「인공지능에 관한 법률안」에 비상 상황에서 안전을 위해 AI 시스템을 정지할 수 있는 비상정지 시스템 구축과 관련된 내용은 다른 법안에서는 발견되지 못하는 내용이다.

제22대 국회에서도 인공지능(AI) 관련 법안이 여러 건 발의되었다. '인공지능'이 제명에 명시된 경우만 한정한다면 총 11개의 법안이 확인된다.[13] 이러한 법안들은 인공지능 산업의 육성이나 인공지능의 윤리적 기준의 확립, 개인정보 보호와 신뢰성 강화 등을 목표로 하고 있으며, 일부는 심사 단계에 있다. 법안의 기본 목적이나 내용, 주요 규제 내용을 정리하면 다음과 같다.

13) 2024년 10월 19일 기준

〈표 4-2〉 제22대 국회에서 발의된 인공지능 관련 법률안 주요 내용 및 규제 내용

	의안명 / 대표발의자 / 제안일자	주요 내용, 목적 등	주요 규제 내용	진행 상태
1	인공지능산업 진흥 및 신뢰 확보 등에 관한 특별법안 / 김우영 의원 / 2024.9.24.	정부가 주도적으로 인공지능 산업의 육성 여건을 조성하도록 근거를 수립, 인재 육성 전략의 수립 및 민간 주도의 핵심기술 발굴 지원 등	· 고위험 영역 인공지능에 대한 검토 및 확인	소관위 접수
2	인공지능의 발전과 안전성 확보 등에 관한 법률안 / 이훈기 의원 / 2024.9.12.	인공지능의 건전한 발전을 지원하고 인공지능 사회의 신뢰 기반 조성을 위해 관련 계획 수립, 거버넌스 구성, 국가 및 지자체의 역할 등 기본적인 사항을 규정	· 인공지능 윤리 원칙의 제정, 공표 · 인공지능의 신뢰 기반 조성을 위한 시책 마련 · 고위험 영역 인공지능 사업자의 신뢰성 및 안전성 확보 조치 마련 · 인공지능이 국민 기본권에 미치는 영향평가	소관위 접수
3	인공지능 발전 진흥과 사회적 책임에 관한 법률안 / 배준영 의원 / 2024.8.24.	인공지능의 건전한 발전을 지원하기 위한 거버넌스, 인공지능 사회의 신뢰 기반 조성 및 인공지능 사업자의 사회적 책임에 관한 기본적인 사항 규정	· 인공지능 윤리 원칙의 제정 및 공표 · 인공지능의 잠재적 위험을 최소화를 위한 시책 마련 · 고위험 영역 인공지능에 대한 확인 제도 마련 · 고위험 영역 인공지능과 관련된 사업자의 신뢰성 및 안전성 확보 조치	소관위 접수
4	인공지능책임법안 / 황희 의원 / 2024.8.27.	인공지능의 개발 및 이용에 관한 기본원칙을 정하고, 국가, 사업자의 책무와 이용자의 권리를 규정하며, 고위험 인공지능으로부터 이용자를 보호하기 위한 시책 등 안전하고 신뢰할 수 있는 인공지능 기술·정책의 제도적 기반을 조성	· 인공지능산업 목적을 위한 정부의 역할 규정 및 규제 원칙 설정 · 고위험 인공지능으로부터 이용자 보호를 위한 원칙 규정	소관위 접수

	의안명 / 대표발의자 / 제안일자	주요 내용, 목적 등	주요 규제 내용	진행 상태
5	인공지능 기본법안 / 한민수 의원 / 2024.8.22.	인공지능 기술의 개발 및 인공지능 산업 활성화의 지원 근거 및 인공지능의 안전성과 신뢰성 확보 방안을 마련하여 인공지능 산업을 진흥하고 국민의 삶의 질 향상에 기여	· 고위험 영역 인공지능에 대한 심의·의결 기구 마련 · 인공지능 윤리 원칙을 제정·공표 · 신뢰할 수 있는 인공지능 이용환경 조성 등 시책 마련 · 인공지능 신뢰성과 안전성 확보 위한 사업자의 조치 · 생성형 인공지능 운용 사실 고지 및 표시 · 해외 사업자의 인공지능 기본법상 의무 이행 확보 · 민간 자율 인공지능윤리위원회 설치	소관위 접수
6	인공지능 개발 및 이용 등에 관한 법률안 / 권칠승 의원 / 2024.7.4.	인공지능 기술 발전 및 산업 진흥 등에 필요한 기본적인 사항을 정하고, 관련 근거를 마련하여 안전하고 신뢰할 수 있는 인공지능 사회를 구현하기 위한 제도적 기반을 조성	· 인공지능 윤리 원칙 확립 · 인공지능 안전성 및 신뢰성 확보를 위한 시책 추진 · 금지된 인공지능에 대한 원칙적 개발 및 이용 금지 · 고위험 인공지능에 대한 확인 절차 및 제공자에 대한 의무 규정	소관위 심사
7	인공지능기술 기본법안 / 민형배 의원 / 2024.6.28.	사람의 생명과 안전 및 기본권을 법률로 보장하면서, 인공지능 산업 진흥과 기술 발전을 위한 체계적 국가 지원제도를 마련	· 인공지능 윤리 원칙의 제정, 공표 및 실천방안 수립 · 신뢰할 수 있는 인공지능 이용환경 조성 · 고위험 영역 인공지능에 대하여 이용자에게 사전 고지 및 요청 시 고위험 영역 인공지능 해당 여부에 대한 확인 · 고위험 영역 인공지능 신뢰성 및 안전성 확보 위한 사업자의 노력 · 민간 자율 인공지능윤리위원회 설치	소관위 심사

	의안명 / 대표발의자 / 제안일자	주요 내용, 목적 등	주요 규제 내용	진행 상태
8	인공지능산업 육성 및 신뢰 확보에 관한 법률안 / 김성원 의원 / 2024/6.19	인공지능 산업 육성을 지원하는 한편, 안전하고 신뢰할 수 있는 인공지능 기술 및 정책의 제도적 기반 마련	· 인공지능 기술 개발 및 이용자 권리 보호 노력 · 고위험 영역의 인공지능 여부에 대한 확인 · 고위험 영역의 인공지능에 대하여 이용자에게 사전 고지	소관위 심사
9	인공지능산업 육성 및 신뢰 확보에 관한 법률안 / 조인철 의원/ 2024.6.19	AI 산업 육성과 AI 시스템의 신뢰성 확보를 위한 정부 지원책을 중심으로, 인공지능 산업의 경쟁력 강화와 신뢰성 확보를 위한 법적 기반을 마련하고 AI 기술이 공정하고 투명하게 사용될 수 있는 기반 마련	· 설명 가능한 인공지능 기술을 개발하고, 이용자 기본 권리 보호를 위한 노력 · 개발 과정에서 고위험 영역에 활용되는 인공지능 여부에 대한 확인 · 고위험 영역의 인공지능을 활용한 제품 및 서비스를 제공 시 이용자에게 사전 고지	소관위 심사
10	인공지능 발전과 신뢰 기반 조성 등에 관한 법률안 / 정점식 의원 2024.6.17	인공지능의 건전한 발전을 지원하고 인공지능 사회의 신뢰 기반 조성에 필요한 기본적인 사항을 규정하고 대한민국 인공지능의 새로운 기준을 마련	· 인공지능 신뢰성 확보를 위한 사업 추진 · 고위험 영역 인공지능 제품 및 서비스에 대한 사전 고지 · 고위험 영역 인공지능에 대한 신뢰성, 안전성 확보 조치 · 생성형 인공지능 운용 사실 고지·표시, 안전 확보를 위한 조치 이행	소관위 심사
11	인공지능 산업 육성 및 신뢰 확보에 관한 법률안 / 안철수 의원 / 20245.31	인공지능 관련 신뢰 기반 조성과 함께 인공지능 기술 개발 및 산업 진흥을 위한 정책을 종합적으로 추진하여 육성	· 인공지능의 잠재적 위험을 최소화하기 위한 시책 마련 · 고위험 영역 인공지능에 대한 확인 제도 마련 및 신뢰성과 안전성 확보 조치 · 생성형 인공지능 제품 및 서비스에 대하여 이용자에게 사전 고지 및 결과물에 표시	소관위 심사

출처: https://likms.assembly.go.kr/bill/BillSearchResult.do(인출일: '24.10.19.)에서 검색한 각 법률안

제21대 국회에서 발의된 법안과 비교해 볼 때 제22대 국회에서 발의된 법안은 고위험 영역 인공지능 규제와 관련된 내용이 상대적으로 다수 발견된다. 대부분 고위험 영역 인공지능 활용에 대한 신뢰성과 안전성을 확보하고, 제품이나 서비스가 고위험 영역 인공지능을 활용했을 경우에, 그에 대해 이용자에게 알권리를 부여하는 내용이다. 제21대 국회에서 발의된 법안에서는 발견되지 않았던 '생성형 인공지능'이 등장하기도 하는데, 생성형 인공지능을 운용한 사실을 고지하고 안전 확보를 위한 조치를 이행할 것을 규정하고 있다.

이훈기 의원 등이 발의한 「인공지능의 발전과 안전성 확보 등에 관한 법률안」의 '인공지능이 국민 기본권에 미치는 영향의 평가'나 한민수 의원들이 발의한 「인공지능 기본법안」의 '해외 사업자의 인공지능 기본법상 의무 이행 확보'와 관련된 규제는 다른 법안에서 찾아보기 어려운 내용이다. 단순 규제를 넘어서 사회적인 영향도를 객관적으로 평가하여 인공지능 정책 수립의 근거로 마련하고자 한다는 점, 그리고 해외 인공지능 사업자가 국내에 미칠 수 있는 부정적 영향을 차단하고자 하는 적극적인 조치라고 볼 수 있겠다.

2. 법률상 계획

국내에 인공지능 관련 법률은 없지만 인공지능과 밀접한 데이터 활용과 관련된 법률상 주요 계획을 통해 인공지능에 대한 정부의 규제 계획과 국가의 노력을 부분적으로 확인할 수 있다. 이에 2024년에 적용되는 데이터 활용과 관련된 법률상 계획을 검토해 보았다.

총 4개의 법률상 계획이 확인되었는데 '제2차 데이터기반행정 활성화 기본계획(2024~2026년)', '제4차 공공데이터의 제공 및 이용 활성화에

관한 기본계획(2023~2025년)', '제1차 (2023~2025년) 데이터산업 진흥 기본계획', '개인정보 보호 기본계획(2024~2026)'이다. 계획 명에 인공지능이 포함되지 않았지만 모두 데이터 활용과 관련된 계획이라는 점에서 인공지능 활용과 무관하지 않다. 실제로 「데이터기반행정 활성화에 관한 법률」 제6조에 근거하여 작성되는 '제2차 데이터기반행정 활성화 기본계획(2024~2026년)'을 제외하고 나머지 세 법률상 계획은 인공지능 규제와 직간접적으로 관련되는 사항을 언급하고 있다.

「공공데이터의 제공 및 이용 활성화에 관한 법률」 제7조를 근거로 수립된 '제4차 공공데이터의 제공 및 이용 활성화에 관한 기본계획(2023~2025년)'은 데이터의 안전한 활용을 위해 가명·익명 정보를 활용할 것을 강조하고 있으며, 데이터 활용 활성화로 인해 발생하는 부정적인 영향을 최소화하기 위해 공공데이터의 윤리 확산 노력을 명시하고 있다. 이와 유사한 내용이 「개인정보 보호법」 제9조를 기반으로 수립된 '개인정보 보호 기본계획(2024~2026)'이다. 동 계획에는 가명 정보의 안전한 활용에 대한 지원을 확대하며, 데이터의 안전한 활용이 이루어지기 위해 법·제도적 기반을 조성하는 노력이 명시되었다. 이상의 내용이 정형데이터를 중심으로 하는 가명 정보 활용을 강조한다면, 「데이터 산업 진흥 및 이용 촉진에 관한 기본법」 제4조를 근거로 하는 '제1차 (2023~2025년) 데이터산업 진흥 기본계획'과 상기한 '개인정보 보호 기본계획(2024~2026)'에는 인공지능 활용을 염두에 둔 규제 내용이 포함되어 있다. 그 내용을 보다 자세히 살펴보면 다음과 같다.

〈표 4-3〉 데이터 관련 법률상 계획과 구제 내용

계획명	근거법	주요 내용	인공지능 규제 내용
제2차 데이터기반행정 활성화 기본계획(2024~2026년)	「데이터기반행정 활성화에 관한 법률」 제6조	① 범정부 데이터 공유 플랫폼을 통한 기관 간 데이터 칸막이 해소, ② 정책 맞춤형 데이터 분석으로 과학적 행정 추진 가속화, ③ 데이터 공유·분석·활용 일상화로 데이터 기반 행정 문화 정착	-
제4차 공공데이터의 제공 및 이용 활성화에 관한 기본계획(2023~2025년)	「공공데이터의 제공 및 이용 활성화에 관한 법률」 제7조	① 공공데이터 전면 개방 체계 마련, ② 데이터의 연결 및 융·복합 등 활용 제고를 위한 품질관리 및 표준 적용 강화, ③ 공공데이터 이용 활성화 및 민관협업을 통한 사회 현안 해결 지원, ④ 공공데이터 생산 및 활용 생태계 활성화를 위한 기반 강화	안전한 공공데이터 활용을 위한 제도적 기반을 마련하기 위해 가명·익명 정보 활용 촉진, 부정적 영향 최소화를 위한 공공데이터 윤리 및 문화 확산
제1차(2023~2025년) 데이터산업 진흥 기본계획	「데이터 산업진흥 및 이용촉진에 관한 기본법 제4조	① 모든 데이터의 혁신적 생산 개방 공유, ② 민간 중심 민간 주도·데이터 유통·거래 생태계 마련, ③ 안전하면서도 혁신을 촉진하는 데이터 활용 기반 조성, ④ 데이터 산업 활성화를 위한 국가 디지털 전환 전면화	신뢰성에 기반한 데이터 활용 및 데이터 윤리 확산
개인정보 보호 기본계획(2024~2026)	「개인정보 보호법」 제9조	① 데이터 경제 시대 선도, ② 개인정보 안심 사회 구현, ③ 글로벌 데이터 신(新) 질서 주도	신뢰할 수 있는 신기술 이용환경을 위해 ① 인공지능 시대에 대응한 규제혁신 추진 ② 디지털 신기술 환경에서의 개인정보 보호방안 마련, ③ 가명정보 안전 활용 지원 확대, ④ 안전한 데이터 활용을 위한 법·제도 기반 조성

출처: 각 계획에서 발췌하여 연구자가 정리

먼저, ‘제1차(2023~2025년) 데이터산업 진흥 기본계획’에는 ‘자유롭고 공정한 데이터 접근 이용 보장’을 위해 ‘신뢰성에 기반한 데이터 활용 및 데이터 윤리 확산’과 관련된 전략이 담겨 있다(관계부처합동, 2023, p.46). 동 보고서에는 데이터 활용 과정에서 발생하는 편향성 문제를 최소화하고 신뢰성이 담보된 데이터 윤리를 확산할 필요성을 강조하고 있는데, 특히 ① 기술 개발, ② 학습용 데이터 신뢰성 강화, 그리고 ③ 데이터 윤리의 확산에 대한 추진 내용을 제시한다. 기술 개발과 관련해서는 데이터 활용에서 발생할 수 있는 편향성이나 낮은 신뢰성 완화를 위해 총 200억 원의 예산으로 핵심 원천기술 개발 계획을 명시한다. 학습용 데이터의 신뢰성을 위해 표준 공정 기준 활용을 확산하고, 데이터 구축 단계별로 요구되는 개인정보 보호 가이드라인을 확산하는 계획을 제시한다.[14] 또한 데이터 윤리 확산을 위한 정책포럼을 개최하고 맞춤형 인공지능 윤리 교육 콘텐츠를 개발하며 공공데이터 윤리 헌장을 제정하는 계획을 수립하였다(관계부처합동, 2023, p.46).

다음으로, 「개인정보 보호법」 제9조를 기반으로 수립된 ‘개인정보 보호 기본계획(2024~2026)’에는 가명 정보의 안전한 활용 이외에도 신뢰할 수 있는 신기술 이용 환경을 마련하고자 ① 인공지능에 대응한 규제혁신을 추진하고, ② 디지털 신기술 환경에서의 개인정보 보호 방안을 마련하는 계획을 담고 있다(관계부처합동, 2023, p.46). 첫 번째의 규제혁신은 ‘① 인공지능 환경에서의 개인정보 보호·규제 방안 마련, ② 자동화된 결정과 프로파일링에 대한 합리적 개인정보 처리 기준 마련, 그리고 ③ 실효성 있는 AI 규제 설계 및 집행을 위한 소통·협력 창구 운영’이라는 세

14) 이에 대한 주요 내용으로는 ① 설계 단계부터 구축에 이르는 전(全) 주기에 데이터 유형별 포트폴리오와 가이드라인을 확산, ② 인공지능이 활용되는 목적에 따라 데이터 구축 단계별 데이터 신뢰성을 확보하기 위해 상세한 요구 사항을 확산, ③ 데이터 구축 과정의 공정과 결과물이 신뢰성을 충족하는지 여부에 대하여 관련 지표와 측정 방법을 고도화하는 것이다(관계부처합동, 2023).

가지 내용으로 구성된다(개인정보보호위원회, 2023 : 21).

먼저, '개인정보를 보호하고 관련 규제 방안을 마련'하기 위해, 공개된 정보 및 서비스 제공 과정에서 생성된 정보를 가지고 인공지능 학습에 활용할 때 적용되는 개인정보 처리 원칙 및 기준을 마련한다. 또한 생성형 AI 개발이나 생성형 AI 기반 서비스의 제공 과정에서 발생 가능한 개인정보 침해 요인 및 보호 방안에 관한 연구를 진행하고 초거대 AI에 대응하는 개인정보 보호 및 활용 기술을 주제로 연구개발을 추진한다. AI 유형별, 사안별 발생하는 위험성 판단 기준과 평가모델을 개발하는 내용도 동 계획에 포함되어 있다.

다음으로, '자동화된 결정과 프로파일링에 대한 합리적 개인정보 처리 기준을 마련'하기 위해 자동화된 결정의 기준 및 절차, 정보 주체 권리행사 등과 관련된 하위 법령과 가이드라인을 마련하고, 부당한 권리 침해를 최소화할 수 있는 합리적 규율을 마련한다.

마지막으로, '실효성 있는 AI 규제 설계 및 집행을 위한 소통·협력 창구 운영'을 위해 민·관 협업으로 인공지능을 규제하기 위한 추진체계를 마련하고, 개인정보 처리에 대한 고지 방법이나 비정형 데이터를 처리하는 기준 등 새로운 이슈의 규제 방안을 공동으로 설계하는 계획을 수립하였다.

신뢰할 수 있는 신기술 이용 환경을 마련하기 위한 두 번째 방안으로 개인정보 보호를 위해 ① 개인정보 보호 강화 기술(PET) R&D 및 보급 활성화, ② 생체인식 서비스 활성화에 따른 개인정보 법제도 기반 마련, ③ 클라우드 등 신기술 서비스에서의 개인정보 보호 방안 연구를 수행하는 계획을 제시하고 있다(개인정보보호위원회, 2023 : 22).

우선, 개인정보 보호 강화 기술과 관련해서는 신기술 기반으로 제공되는 서비스를 안전한 상태로 상용화하고 정보 주체 권리의 보호를 위해

개인정보 보호 강화 기술(Privacy Enhancing Technologies, PET) 개발을 지원한다. 또한 생체인식 서비스 제공과 관련하여 안면인식 기술의 위험성에 대한 규제 및 관리 방안에 관한 연구를 수행하고 사생활 침해 방지를 위한 규율체계를 마련할 예정이다. 클라우드 서비스 이용 과정에서 개인정보를 보호하는 방안에 관한 연구 수행과 함께 관련 대책을 마련할 계획이다.

인공지능과 관련된 법률이 부재한 상황에서 인공지능에 대한 규제는 데이터 관련 법률상 계획 수립과 추진에 의존할 수밖에 없다. 특히 '제1차(2023~2025년) 데이터 산업 진흥 기본계획'과 '개인정보 보호 기본계획(2024~2026)'에는 인공지능 활용을 염두에 둔 규제 내용을 담고 있어서 인공지능 관련 법률이 신설되기 전까지 인공지능 활용에 따른 위험에 관한 제재가 동 계획을 통해 적용될 수 있을 것이다. 동 계획에는 학습용 데이터 및 생성형 AI의 활용, 생체인식 서비스, 클라우드 등 신기술 서비스 이용 등 인공지능 활용과 매우 밀접한 영역에서 안전성 확보나 개인정보 침해 방지 등을 위한 규제 계획을 제시하고 있다. 다만, 연구나 R&D부터 시작하는 단계로 확인되는 내용도 보이므로 구체적인 제재나 가이드라인이 도출되기에는 다소 시간이 필요할 것이다. 또한 인공지능 자체에 관한 법률에 기반하지 않은 규제이기 때문에 총체적인 시각에서의 규제 방안이라고 하기에는 부족한 것이 사실이다.

3. 지침 및 가이드라인

다음으로 법적 준수를 돕거나 방향성을 제공하기 위한 것으로 강제성은 없으나 준수하지 않을 경우 제재나 평가상 불이익을 받을 수 있는 지침 및 가이드라인을 살펴본다. 인공지능 활용과 관련된 주요 가이드라인은 특정 분야를 지정하지 않고 일반적으로 적용되는 가이드라인과 특정 분야나 대상에 한정한 가이드라인으로 나눌 수 있다. 전자로는 방송통신위원회· 정보통신정책연구원(2019)의 '이용자 중심의 지능정보사회를 실현하기 위한 원칙'과 과학기술정보통신부(2020)의 '인공지능(AI) 윤리 기준'이 있으며 국가인권위원회의(2022) '인공지능 개발과 활용에 관한 인권 가이드라인' 이 있다.

방송통신위원회·정보통신정책연구원(2019)의 '이용자 중심의 지능정보사회를 실현하기 위한 원칙'은 AI 시대에 이용자의 권리와 이익이 보호될 수 있도록 정부와 기업, 이용자 등 사회 전체 구성원들이 함께 지켜야 할 기본적인 원칙을 명시하고 있다. 신기술의 도입이 초래하는 기술적, 사회적 위험으로부터 안전한 지능정보서비스 환경을 조성하기 위해, 지능정보사회의 모든 구성원이 고려할 공동의 기본원칙을 제시했다. 구체적으로는 사람 중심 서비스 제공, 투명성과 설명 가능성, 책임성, 안전성, 차별 금지, 프라이버시 보호 등의 원칙을 제시한다(방송통신위원회·정보통신정책연구원, 2019).

과학기술정보통신부(2020)의 '인공지능(AI) 윤리 기준'은 과학기술정보통신부가 글로벌 추세에 발맞추어 2019년에 발표한 '인공지능 국가전략(2019.12)'에서 주요 과제로 추진해 온 결과물이다. 윤리 기준은 공개 공청회 등 각계 전문가·시민 공개 의견수렴 거쳐 발표했다. '인간성(Humanity)을 위한 인공지능(AI)'의 3대 원칙·10대 요건을 담았다(과학

기술정보통신부, 2020). 이행 요건에서는 인권 보장, 프라이버시 보호와 책임성, 안전성, 투명성 등에 대한 기준을 제시했다.

국가인권위원회(2022)의 '인공지능 개발과 활용에 관한 인권 가이드라인'은 유엔이나 유럽연합 등 국제사회가 인공지능으로 인해 발생하는 문제를 예방하기 위한 제도와 지침을 개발하고 있는 과정에서 국제적 흐름에 발맞춰 작성한 것이다. 인공지능을 개발 및 활용하는 과정에서 준수해야 할 내용으로 투명성과 설명 의무, 자기 결정권 보장, 차별 금지, 영향평가, 위험도 등급 및 관련 제도 마련 등을 제시하고 있다(국가인권위원회, 2022). 이상의 보편적 대상에 대한 가이드라인은 인공지능을 개발·활용하는 과정에서 책임성, 안전성, 투명성, 프라이버시 보호 등의 유사한 지향점을 제시하고 있다.

〈표 4-4〉 인공지능과 관련된 일반 대상 주요 가이드라인

가이드라인명	작성자	주요 내용
이용자 중심의 지능정보사회를 실현하기 위한 원칙 (2019)	방송통신위원회·정보통신정책연구원	· 이용자 중심의 지능정보사회 실현을 위해 ① 사람 중심의 서비스 제공, ② 투명성과 설명 가능성, ③ 책임성, ④ 안전성, ⑤ 차별 금지, ⑥ 참여, ⑦ 프라이버시 보호, ⑧ 데이터 거버넌스의 원칙 제시
인공지능(AI) 윤리 기준 (2020)	과기정통부	· 인간성 구현을 위해 인공지능 개발·활용 과정에서 ① 인간의 존엄성 원칙, ② 사회의 공공선 원칙, ③ 기술의 합목적성 원칙 제시 · 3대 기본원칙의 실천 및 이행을 위해 인공지능 개발부터 활용의 전 과정에서 충족되어야 하는 요건으로 ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성 제시
인공지능 개발과 활용에 관한 인권 가이드라인 (2022)	국가인권위원회	· 인공지능 개발과 활용 과정에서 ① 인간의 존엄성 존중, ② 투명성과 설명 의무, ③ 자기결정권의 보장, ④ 차별 금지, ⑤ 인공지능 인권 영향평가 시행, ⑥ 위험도 등급 및 관련 법 제도 마련의 가이드라인 제시

출처: 연구진 작성

인공지능과 관련된 일반 대상 가이드라인만으로는 특정한 영역이나 분야, 대상에 대한 지침이 필요한 경우 그 내용상 한계가 발생할 수 있다. 특정 분야나 특정 대상 가이드라인을 통해 보다 구체적이면서 실용적인 가이드라인을 제공하는 사례를 살펴보면 다음과 같다.

금융위원회(2021)의 '금융 분야 AI 가이드라인'은 금융회사 및 관련 기관을 대상으로 한다. 금융 산업에서 안전하고 신뢰성 있게 인공지능을 활용하기 위한 방향을 제시하며 AI 기술이 금융서비스에 적용될 때 발생할 수 있는 위험관리와 소비자 보호를 목표로 하고 있다(금융위원회, 2021). 이 가이드라인은 금융 분야에서 AI를 활용하는 과정에서 윤리적 원칙과 내부 통제 장치를 마련하여, AI 기반 금융서비스의 신뢰성과 안전성을 확보하기 위해 작성되었다. 주요 내용은 개발단계와 평가 및 검증단계, 도입·운영·모니터링 과정에서 준수해야 할 내용들이 구체적으로 제시되어 있다.

개인정보보호위원회(2021)의 '인공지능(AI) 개인정보 보호 자율점검표 - 개발자, 운영자'는 AI 기술과 서비스를 개발하거나 운영하는 모든 주체를 대상으로 작성되었다. 인공지능 시스템을 개발하고 운영하는 과정에서 개인정보 보호의 주요 원칙과 요구 사항을 준수하도록 돕기 위한 지침서로, 개인정보를 활용하거나 처리하는 AI 기술과 서비스에 대한 프라이버시 보호 준수 상태를 자율적으로 점검하고 개선하도록 유도하기 위한 목적으로 작성되었다(개인정보보호위원회, 2021). 이 자율점검표는 AI 개발자와 운영자가 개인정보를 안전하게 처리하기 위해 필요한 중요 기준과 절차를 제공하여 개인정보의 투명성, 공정성, 안전성을 지키고 신뢰할 수 있는 AI 서비스를 구축하도록 지원한다.

서울특별시교육청(2021)의 '인공지능(AI) 공공성 확보를 위한 현장 가이드라인'은 서울특별시교육청이 학교 현장에서 인공지능 기술을 도입·활용하는 교사 및 교육 관계자를 위해 작성한 지침서이다. AI 기술이 공교육에 적용되는 경향이 증가하면서 개인정보 보호나 알고리즘의 신뢰성, 데이터 처리에서의 투명성에 대한 필요성이 대두되어, 공교육 현장에서 인공지능을 효과적으로 도입하고 학생들의 개인정보를 보호하고 정보의 투명성, 신뢰성을 확보하기 위해 작성되었다(서울특별시교육청, 2021). 이를 통해 인공지능 기술을 활용한 교육이 학교 현장에서 공정하고 안전하게 진행되도록 지원한다.

교육부(2022)의 '교육 분야 인공지능 윤리 원칙'은 인공지능을 개발하는 개발자와 교육 현장에서 이를 활용하는 교육 당사자 모두를 위해 작성되었다. 교육에 활용되는 인공지능이 학습자 성장과 교육 현장 및 수업에 미치는 영향을 예상하고 교육 현장에 도입되는 인공지능의 역기능 및 부작용을 최소화하기 위해 선제적 자율규제 마련의 필요성에 따라 제정된 것이다(교육부, 2022). 이 원칙은 "사람의 성장을 지원하는 인공지능"이라는 대원칙하에 인공지능이 교육 현장에서 윤리적으로 개발되고 위험한 상황 없이 안전하게 활용되도록 관련 주체들이 함께 준수해야 할 지침이다.

식품의약품안전처(2022)의 '인공지능(AI)의 의료기기 국제 공통 가이드라인'은 의료기기 규제 당국자, 개발자, 제조업체, 기타 이해 관계자를 대상으로 작성되었다. 이 가이드라인은 의료기기에 AI를 적용할 때 고려해야 할 윤리적, 기술적, 안전성 요건 등을 명확히 규정하여 안전하고 효과적인 의료 AI 시스템 개발을 촉진하려는 목적을 지니며 AI 기반 의료기기가 국제적으로 통일된 규제 요구 사항을 충족할 수 있도록 지원한다(국제의료기기규제당국자포럼 인공지능(AI) 의료기기 실무그룹, 2022). AI

기술을 활용한 의료기기의 개발, 평가, 승인, 그리고 시장 출시 과정에서 일관된 기준과 용어를 제공하고, 각국의 규제에 조화되도록 촉진하고, 안전하고 효과적인 의료기기의 개발과 사용을 지원하는 데 목적을 둔다. 이를 통해 글로벌 시장에서 AI 의료기기의 신뢰성과 품질을 보장하고, 환자의 안전을 강화하고자 한다(국제의료기기규제당국자포럼 인공지능(AI) 의료기기 실무그룹, 2022).

마지막으로 과학기술정보통신부 외(2024)의 '인공지능 학습용 데이터 품질관리 가이드라인'은 과학기술정보통신부가 인공지능 학습용 데이터의 품질 확보를 위해 2021년 3월에 '인공지능 학습용 데이터 품질관리 가이드라인 v1.0'을 발간한 바 있는데 그 이후로 지속해서 업데이트를 해오면서 2024년 1월에 최신 버전이 공개된 것이다. 인공지능 학습용 데이터 구축 사업을 수행하는 기관 및 관리기관, 제3자 품질검증기관 등을 대상으로 작성되었으며 인공지능 학습용 데이터를 구축계획 수립 단계에서부터 자료 획득 및 수집, 정제, 가공 등에 이르는 절차, 최종 산출물 및 품질관리 활동을 제시한 기준서로, 인공지능 학습에 사용되는 데이터가 개인정보를 침해하지 않도록 다양한 지침을 제공한다(과학기술정보통신부 외, 2024, p.4). 이 가이드라인은 인공지능 학습용 데이터 구축 사업의 수행 노하우를 집약하여 제작되었으며, 데이터 품질관리 거버넌스 및 프레임워크, 품질검증 지표 등을 상세히 기술하고 있다.

이상에서 살펴본 바와 같이 인공지능과 관련된 특정 대상 주요 가이드라인은 그 대상이나 분야를 특정하고 있기 때문에 가이드라인의 내용도 구체적이며 실용적으로 적용할 수 있는 내용으로 구성되어 있다. 가이드라인이기 때문에 법적 강제성은 없지만 인공지능에 대한 위험을 최소화하고 신뢰성 있는 인공지능 서비스를 제공하는 데 활용되기 위한 사회적 노력이라고 하겠다.

〈표 4-5〉 인공지능과 관련된 특정 대상 주요 가이드라인

가이드라인 명	작성자	주 대상	주요 내용
금융 분야 AI 가이드라인(2021)	금융위원회	금융 분야	· 조직 내 AI 윤리 마련, AI 시스템의 위험 평가·관리를 위한 역할·책임·권한 정의 · 사회적·경제적·문화적 영향, 잠재적 피해 가능성, 책임성 유지를 위한 시스템 · 개발 단계에서의 데이터 품질검증, 개인정보 안전조치, 설명 가능한 AI 기술 도입 노력 · 평가, 검증단계에서의 적합한 성능 목표 및 성능 평가지표를 선정 및 평가, 공정성 평가지표 측정, 설명 가능성을 위한 노력 · 도입·운영·모니터링 과정에서 소비자를 위한 권리구제 방안을 고지, 오용·악용 가능성 방지, AI 개발 환경의 보안 취약성 상시 통지 시스템 마련
인공지능(AI) 개인정보 보호 자율점검표(2021)	개인정보보호 위원회	개발자, 운영자	· ① 개인정보 수집 및 이용에 대한 명확성, ② 개인정보의 안전한 저장·보호, ③ 개인정보의 처리·관리, ④ 개인정보 보호 영향평가(PIA) 및 리스크 관리, ⑤ 정보주체의 권리 보장, ⑥ 투명성 및 설명 가능성 확보, ⑦ 자율점검표의 중요성 강조
인공지능(AI) 공공성 확보를 위한 현장 가이드라인(2021)	서울특별시 교육청	학교 교사나 교육 관계자	· 데이터와 알고리즘의 의사결정 영향 정도와 개인정보 민감 정도를 반영한 인공지능 등급을 평가
교육 분야 인공지능 윤리 원칙(2022)	교육부	교육 분야	· ① 인간 성장의 잠재성 이끌기 ② 학습자의 주도성과 다양성 보장, ③ 교수자의 전문성 존중, ④ 교육 당사자 간 관계를 공고히 유지, ⑤ 교육의 기회균등과 공정성 보장, ⑥ 교육공동체의 연대와 협력 강화, ⑦ 사회 공공성 증진에 기여, ⑧ 교수·학습 과정에서의 안전 보장, ⑨ 데이터 처리의 투명성을 보장하고 설명 가능, ⑩ 데이터를 합목적적으로 활용, 교육 당사자의 프라이버시 보호

가이드라인 명	작성자	주 대상	주요 내용
인공지능(AI)의 의료기기 국제 공통 가이드라인(2022)	식품의약품 안전처	의료기기 규제 당국자, 제조업체, 개발자, 관련 이해관계자	· ① 안전성 및 효능 보장, ② 투명성과 설명 가능성, ③ 연속적 성능 관리, ④ 환자 데이터 보호 및 프라이버시의 중요성을 제시 위험 관리, ⑤ 정확성 및 재현성 확보, ⑥ 임상적 타당성, ⑦ 데이터 품질 과정에서의 핵심 가이드라인 제시
인공지능 학습용 데이터 품질관리 가이드라인(2024)	과학기술정 보통신부	학습용 데이터를 구축·활용하는 기관, 기업	· ① 인공지능 학습용 데이터의 개인정보 보호 원칙, ② 개인정보의 생애주기에 따른 보호 가이드라인 제시, ③ 가명 처리 및 익명화 명시

출처: 연구진 작성

4. 선언

선언은 특정한 가치나 방향을 표명하는 의미가 강하지만 상기한 가이드라인보다 강제성이나 구속력 측면에서는 상대적으로 낮다. 지향하는 목표나 방향성을 사회에 널리 알리고 해당 주제에 대해 사회적 지지와 관심을 촉구하며 자발적 참여를 독려한다. 인공지능과 관련된 대표적 선언으로 '서울 선언'이 있다.

서울 선언은 서울시가 AI 윤리와 디지털 권리를 보장하기 위해 관련된 원칙과 선언적 가이드라인을 제시한 국제적 협력의 결과물이라 할 수 있다(외교부 보도자료, 2024.5.22., p.2). 법적 구속력은 없지만 AI 기술이 윤리적으로 활용되기 위해 필요한 원칙을 제공한다. 지난 2024년 5월 21일 AI 서울 정상회의에는 미국, 유럽연합, 캐나다, 호주, 영국, 독일, 프랑스, 이탈리아, 일본, 싱가포르, 그리고 대한민국을 대표하는 세계 지도자들이 모여 AI 분야의 국제 협력과 활발한 대화를 위한 공동의 노력이 필요함을 확인했다(외교부 보도자료, 2024.5.22., p.2). '서울 선언'은 이날 정상 선언문으로 채택됐다. 부속서인 'AI 안전 과학에 대한 국제협력을 위한 서울 의향서'에는 AI 글로벌 거버넌스가 추구해야 하는 방향이 담겨 있다.

서울 선언에는 안전하고 보안성과 신뢰성을 갖춘 인공지능 보장이 필요함을 인식하고, 선언 참여국들과 관계 기관들이 인공지능의 안전에 관한 연구 협력을 증진하기 위해 노력할 것이라는 의지를 담고 있다. 또한, "안전하고 혁신적이고 포용적인 AI 생태계들을 육성하는 위험 기반 접근법들을 포함한 정책·거버넌스 체계들을 지지"하였다[15](서울선언문 제6호).

15) AI 서울 정상회의 서울선언 및 의향서

‘AI 안전 과학에 대한 국제협력을 위한 서울 의향서’에는 인공지능 시스템의 안전성을 증진하기 위한 개발 지침 작성을 촉진하고 AI 개발 및 이용 과정에서의 혜택이 공평하게 공유되기 위한 노력을 공동으로 해나갈 것이라는 내용이 포함되었다. 또한 안전한 인공지능 활용을 위해 관련 기준을 교환하고 기술적 공유 자원을 공유하는 등의 노력을 수행할 것을 명시하였다.

5. 소결

인공지능에 관한 국내의 법률은 법안의 형태로 법률 제정을 위한 과정에 있으며 국내에서 실질적으로 인공지능을 규제 혹은 규율하는 것은 법률상 계획과 지침 및 가이드라인이다. 법률상 계획의 구체적인 실행 내용이나 과정은 인공지능에 초점을 맞춘 법률이 아니기 때문에 일정한 한계가 있을 수 있고, 지침이나 가이드라인은 법적 구속력이 없어서 실행으로 연결되지 않을 수 있다. 더군다나 가이드라인이 특정 영역이나 분야, 대상을 목적으로 하는 경우 그 내용이 전체 사회에 적용되지 못하고 일부 집단이나 소수 영역에만 적용될 수 있다는 한계가 있다. 그럼에도 불구하고 인공지능의 신뢰성이나 안전성, 투명성, 설명 가능성 등의 가치를 추구하는 법률상 계획이나 가이드라인, 선언문 등이 발표되고 있는 현상은 긍정적으로 평가할 수 있다. 나아가 생성형 인공지능이나 고위험 영역의 인공지능에 대한 우려, 인공지능 활용에서 정보를 제공한 이용자에 대한 고지, 설명할 수 있는 인공지능 등 최근의 이슈들이 언급되고 있다는 점에서 구체적인 실행 방안과 의무부여 여부와는 별도로 그 자체의 의미를 찾을 수 있다고 본다.

다만, 본 연구의 문제의식을 고려한다면 '사회보장 분야에서는 이러한 노력이 어떻게 반영될 수 있는가?'의 의문으로 연결될 수밖에 없다.

사회보장은 대상을 인간으로 한다. 사회보장은 인간에게 직간접으로 재화나 서비스를 제공하는 형태로 이루어지기 때문에 이 과정에서 인공지능이 활용된다면 다양한 이슈들이 발생할 수 있다. 여기에는 개인정보 활용의 문제에서부터 잘못된 알고리즘으로 인한 의사결정, 기계가 인간에게 서비스를 제공하는 과정에서 발생하는 이슈 등 다양한 쟁점이 포함된다.

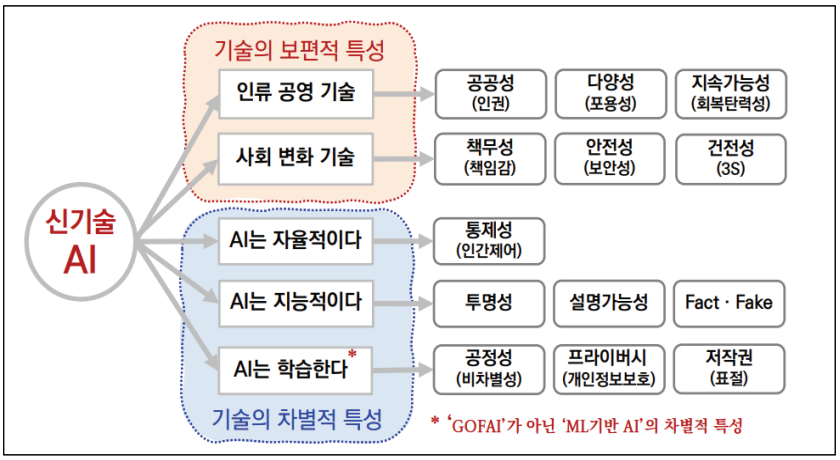
서비스 대상자는 인간이면서 동시에 이들 중 상당수는 특별한 욕구가 결핍된 취약계층이다. 기술의 발전과 제한된 자원의 환경에서 인적, 물적 자원의 효율적 활용을 위해 사회보장의 다양한 분야에서 인공지능 기술이 도입되고 있다. 이는 곧 취약계층이 눈부신 인공지능 기술의 수혜자이면서 동시에 치명적인 인공지능 기술의 피해자일 수 있다는 의미이다. 하이테크놀로지로 무장한 낙인 시스템의 등장을 우려한다면 기우일까? 알고리즘에 활용되는 데이터에 문제가 있다면 가장 큰 피해를 보는 이들은 정부의 지원이 필요하여 지원받는 사람들이다. 사회적 비용을 줄이기 위한 기술이 거꾸로 사회적 비용을 높이는 결과를 낳지 않도록 하기 위해서는 가이드라인 수준에서라도 사회보장 분야에서의 인공지능의 부정적인 영향을 최소화하도록 해야 할 것이다.

제2절 국외 인공지능 기술에 대한 규제

인공지능은 방대한 양의 데이터를 처리한다. 인공지능의 머신러닝 알고리즘은 규칙이나 기준을 명시적으로 따르지 않고, 빅데이터의 패턴과 관계를 학습한다. 그 결과 생성된 결정, 예측, 추천, 콘텐츠의 생성 과정을 인간이 이해하기는 어렵다. 여기서 인공지능의 '블랙박스'(Zaber, Casu, Brodersohn, 2024, p. 19) 문제가 대두된다. 인간이 인공지능의 입력과 출력은 알지만, 과정에 대해서는 알 수 없다. 인공지능이 '블랙박스'를 거쳐서 산출한 출력값을 신뢰할 근거도 없다. 오히려 신뢰하기 어려운 근거가 적지 않다. 하나의 예를 들면, 쓰레기 데이터가 쓰레기 출력값을 산출하는(garbage in, garbage out) 문제가 제시된다(Geiger, Yu, Yang, Dai, Qiu, Tang, Huang, 2020, January). 인공지능이 산출한 값의 근거가 되는 자료가 부정확하거나, 부적절하거나, 비밀스러운 자료인 경우에 그 산출 값을 신뢰할 근거가 무너진다. 인공지능의 블랙박스적인 속성으로 인해 산출 과정을 파악하기 어렵고, 무엇인 문제인지, 문제가 아닌지도 파악하기 어렵다.

인공지능의 압도적인 능력을 염두에 둔 윤리가 문제가 대두될 수밖에 없다. 여기서 윤리는 공공성, 다양성, 지속 가능성, 책무성, 안정성, 건전성, 통제성, 투명성, 설명 가능성, 공정성 등을 아우른다([그림 4-1] 참고). 인공지능이 가지는 특성에 따른 필연적인 결과다. 인공지능은 인류 공영의 기술이면서, 동시에 자율성, 지능성 및 학습능력을 갖기 때문이다.

〔그림 4-1〕 인공지능 윤리 관련 의제와 원칙



출처: 김명주. (2024). 한국보건사회연구원 사회보장행정에서 인공지능 적용 동향과 함의 세미나 발표 자료.

이러한 배경에서 전 세계 국가 및 지역 정부 및 국제기구에서 인공지능에 대한 규제 혹은 가이드라인을 앞다투어 내놓고 있다. 인공지능이 초래할 수 있는 사회적 위험이 거대할 수 있고, 동시에 비가시적, 불확정적이기 때문이다. 대표적인 예가 World Economic Forum(2023)의 Presidio Recommendations on Responsible Generative AI, European Union(2024)의 AI Act, OECD(2023)의 AI Principles, UNESCO(2021)의 Recommendation on the Ethics of AI, 미국 백악관의 Blue Prints for an AI Bill of Rights(White House, 2022), 미국 캘리포니아주에서 시도했던 AI 규제 법안(SB 1047) 등이다. 여기에서는 유럽연합의 AI Act와 미국 행정명령 14110을 살펴보겠다. 다른 형태들은 모두 법적인 구속력이 없는 권고 혹은 가이드라인의 형태를 띠기 때문이다.

〈표 4-6〉 주요국의 인공지능 관련 규제

국가·지역	주요 사례	특징
EU	EU AI 법	• 포괄적·수평적 규제, EU진출기업 규제 - 발효 (8월 1일)
미국	AI 권리장전 청사진 / AI에 관한 자발적 약속 안전성과 보안을 갖추고 신뢰할 수 있는 AI에 관한 행정명령	• 시민의 권리 보호에 초점, 기업의 책무성 강조 • 입법 및 규제 지침
캐나다	자동화된 의사결정에 관한 지침 / AI 및 데이터 법안 첨단 생성형 AI시스템의 책임 있는 개발·관리에 관한 자발적 행동강령 생성 AI 이용에 관한 지침	• 공공·민간 영역 AI 규제 이원화, 영향평가제도 도입 • 2022.6. 발의	2023. 11. 28. AI 및 데이터법안 수정안 제안 • 2023. 4. 산업과학기술상임위원회 회부 및 계류 중
영국	AI 규제에 대한 혁신 친화적 접근(AI 백서) 하원 과학혁신기술위원회 AI 거버넌스 보고서	AI 규제 법안 디지털 정보 및 스마트 데이터 법안 발의 계획 발표	• (정부) 부처별 접근 + 총괄 지원 조직 운영 • (의회) 법률 제정 권고 및 법안 발의 • 2024. 7. 17. King's Speech
중국	알고리즘 추천 관리규정 / 인터넷 정보서비스 심층 합성 관리규정 생성형 AI 서비스 관리 잠정 방법 | 생성형 AI 표준 제안	• 콘텐츠 규제 중심의 입법 • 시행 중
일본	히로시마 AI 프로세스 종합 정책 프레임워크 기업을 위한 AI 가이드라인 초안 생성형 AI 개발자 규제 지침 - 경제재정운영지침 책임감 있는 AI 진흥을 위한 기본법안 발의	• 공공 민간 모두 적용 (데이터 거버넌스, 보안, 개인정보보호, 위험성관리) • 2024. 4. 19. AI 개발자 대상 규제, 허위정보 유포, 인권 침해 예방 • 생성형 AI 모델 규제 법안 • 바이든 AI 행정명령과 유사한 내용

출처: 윤혜선. (2024). 주요국의 AI 규제 거버넌스 구축 현황(1).

1. EU의 인공지능법(AI Act)

가. 법률 제정 배경 및 내용

유럽연합의 인공지능법(Artificial Intelligence Act)은 세계 최초로 AI에 관한 포괄적인 법적 프레임워크다(European Commission, 2024). 인공지능법은 유럽연합에서 자주 사용되는 간접적인 방식인 지침(directive)이 아닌, 법(act)으로 제정됐다. 따라서 모든 회원국에 직접적이고 즉각적인 규제(regulations)로 적용되고, 국가별로 별도의 법률 제정 없이 효력을 발휘한다.

인공지능법은 2021년 4월 유럽연합 집행위원회(European Commission)가 유럽연합 지역 내에서 AI 규제를 위한 제안서를 발표하면서 시작됐다(라기원, 2024). 이 제안서를 바탕으로 유럽연합 이사회(European Union Council)는 다양한 논의와 협의를 진행하였고, 유럽연합 의회(European Union Parliament)는 윤리적 쟁점, 생체 인식 기술, 고위험 애플리케이션에 중점을 두고 인공지능법안 작업을 시작했다.

2022년에는 유럽연합 이사회 의장단이 회원국들 사이에 쟁점이 되는 여러 조항에 대한 타협안을 도출했다. 이후에도 인공지능법안에 대한 보완 조치는 지속하여 이루어졌고, 2023년 6월에 유럽연합 의회는 법안을 최종적으로 확정했다. 2024년 2월에는 법률의 이행을 감독하기 위한 유럽 인공지능사무국(European Artificial Intelligence Office)도 출범했다. 5월 들어 유럽연합 이사회는 이 법률의 채택을 공식화했고, 7월에 유럽연합 관보에 공식적으로 인공지능법의 채택 선언이 게재되었다(EU Artificial Intelligence Act, 2024).

새롭게 제정된 유럽연합 인공지능법의 목표는 유럽을 비롯한 세계 어느 지역에서도 신뢰할 수 있는 AI가 개발될 수 있도록 모든 AI 시스템이 인간의 기본권, 안전, 윤리 원칙 등을 존중하게 하고, AI 모델의 위험성을 관리하는 데 있다(European Commission, 2024). 따라서, 이 법은 구체적인 상황별로 AI의 개발자와 배포자가 지켜야 할 명확한 요건과 의무를 명시하고 있다. 물론, 법이 AI 개발을 지원하려는 의도도 담고 있다. 따라서 AI 관련 혁신 정책 패키지 및 AI 관련 협력 계획을 명시하고 있고, AI 개발 및 활용과 관련하여 기업, 특히 중소기업의 행정적 부담과 비용을 줄이는 방안도 포함되어 있다(European Commission, 2024).

인공지능법은 모두 13개 장의 113개 규정과 13개의 부속서로 구성되어 있다. 각 장의 제목과 내용은 〈표 4-7〉과 같다.

〈표 4-7〉 EU 인공지능법의 구성

장	내용
제1장	**총칙:** 법 적용 대상 및 범위, 용어 정의, AI 리터러시
제2장	**AI 활용 관련 금지행위:** AI 시스템의 활용이 금지되는 경우와 예외
제3장	**고위험 AI 시스템:** 제1절: AI 시스템의 고위험 분류 기준 제2절: 고위험 AI 시스템의 준수사항 제3절: 고위험 AI 시스템 제공자·배포자·기타 이해관계자의 준수사항 제4절: 승인기관(통보기관), 인증기관 제5절: 표준·적합성 평가·인증서·등록 기준
제4장	**특정 AI 시스템의 제공자 및 배포자에 대한 투명성 의무:** 사람과 직접 상호작용할 목적으로 사용되는 AI 시스템의 제공자 및 배포자의 투명성 의무
제5장	**범용 AI 모델:** 제1절: 범용 AI 모델의 분류 규칙 제2절: 범용 AI 모델 제공자의 준수사항 제3절: 시스템적 위험이 있는 범용 AI 모델 제공자의 준수사항
제6장	**혁신 지원 방안:** 규제 샌드박스 및 실제 조건에서의 고위험 AI 시스템 시험이 가능한 예외와 스타트업 등 중소기업에 대한 산업 진흥 제도
제7장	**거버넌스**: 제1절: 유럽연합 수준의 거버넌스 – AI 사무국, 유럽AI위원회, 자문포럼, 독립전문가 과학패널 제2절: 국가 관할기관 – 각 회원국의 관할 기관 지정
제8장	**고위험 AI 시스템을 위한 EU 데이터베이스:** 유럽연합 집행위원회가 관리하는 고위험 AI 시스템에 대한 EU 데이터베이스
제9장	**사후 모니터링, 정보공유, 시장감독:** 제1절: 사후 모니터링 제2절: 중대한 사고에 대한 정보 공유 제3절: 시장감독기관 및 집행위원회의 규범 집행을 위한 수단 제4절: 구제 수단 제5절: 범용 AI 모델 제공자에 대한 감독, 조사, 집행 및 모니터링
제10장	**행동규범 및 지침:** 고위험 AI 시스템을 제외한 시스템에 적용할 수 있는 행동규범 및 지침 마련
제11장	**권한의 위임과 위원회 절차:** 집행위원회 권한 위임의 근거, 집행위원회 보조 위원회 근거 규정
제12장	**처벌:** AI 시스템에 대한 규범 위반, 범용 AI 모델에 대한 규범 위반에 대한 처벌 규정
제13장	**종칙**

주: 라기원(2024)이 정리한 내용을 일부 수정·인용함.

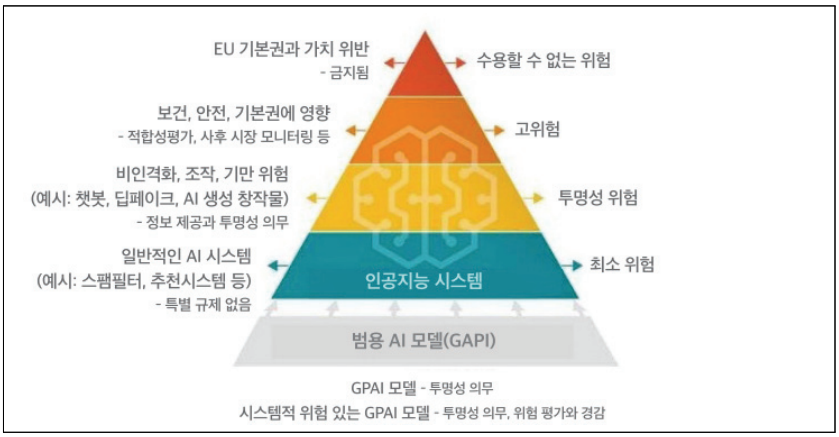
나. 인공지능법의 특성 및 적용 범위

유럽연합의 인공지능법 제3조에서는 인공지능을 다음과 같이 정의하고 있다. "다양한 수준의 자율성을 가지고 작동하도록 설계되었으며, 배포 후 적응성(adaptiveness)을 보일 수 있는 기계 기반 시스템을 의미한다. 이 시스템은 명시적 또는 암묵적 목적을 위해 입력된 데이터를 바탕으로 예측, 콘텐츠, 추천 또는 물리적 혹은 가상 환경에 영향을 미칠 수 있는 결정을 포함한 출력을 생성하는 방법을 추론한다"(European Act, Article 3, (1)).

인공지능법은 유럽연합 역내 시장에서 AI를 활용한 상품과 서비스의 기준을 통일함으로써 상품과 서비스의 국경 간 자유로운 이동을 보장하고자 한다(라기원, 2024). 즉, 회원국들이 개별적으로 AI에 대한 규제를 강화하는 것을 막아 AI 시스템의 개발, 수입 또는 사용에 있어 법적 확실성을 보장하는 것을 목적으로 한다.

인공지능법의 특징은 유럽평의회 인공지능 고위급 전문가 그룹(European Commission AI HLEG, 2019)이 제시한 "신뢰할 수 있는 AI를 위한 윤리 지침(Ethics guidelines for trustworthy AI)"을 바탕으로 규범이 형성되었다는 점이다. 인공지능법 전문은 윤리 지침이 제시하는 다음의 일곱 가지 원칙에 근거한다. 인간 주도·감독(Human agency and oversight), 기술적 견고함과 안전성(Technical Robustness and safety), 프라이버시와 데이터 거버넌스(Privacy and data governance), 투명성(Transparency), 다양성과 차별 금지 및 공정성(Diversity, non-discrimination and fairness), 사회·환경적 복지(Societal and environmental well-being), 책임성(Accountability)이다(European Act, (27)). 위와 같은 원칙들은 모두 직간접적으로 사회보장 영역에도 연결된다고 추정할 수 있다.

〔그림 4-2〕 유럽연합 인공지능법에서 규정하는 위험의 위계



출처: Madiega. (2024). "Artificial intelligence act". European Parliament. p. 1의 그림을 라기원(2024)이 번역해서 제시함. 저작권 European Parliament, 한국법제연구원.

인공지능법의 가장 큰 특징은 위험성 차원에서 AI 시스템을 분류한 다음, 그에 따라 규제의 내용을 차등화한다는 점이다. 인공지능 시스템이 내포하는 위험성의 강도와 범위에 비례하여 규제 유형과 정도가 규정되는 것이다.

인공지능법에서 핵심 개념인 위험성은, 인공지능의 개념 정의에 바로 뒤이어 다음과 같이 정의된다. "위험의 발생 가능성과 그로 인한 피해의 심각성의 결합"(Artificial Intelligence Act, Article 3, (2)). 유럽연합은 인공지능 모델의 성격과 용도 등을 고려해서 위험성의 경중을 진단하고, 그에 따른 규제의 수준도 연동한다. 이 법에서 AI 시스템의 위험성 수준을 네 가지로 구분한다(〔그림 4-2〕 참고). 네 가지 위험성 수준은 '수용할 수 없는 위험성(unacceptable risk)[16]', '고위험성(high risk)', '제한된

16) 유럽연합의 인공지능법 본문에서는 '수용할 수 없는(unacceptable)'에 더해 '금지된(prohibited)'이라는 표현도 자주 등장한다. 또한 '제한된 위험성'은 투명성(transparency) 위험으로도 사용된다.

위험성(limited risk)', '최소한의 위험성(minimal risk)'이다. 가장 높은 수준의 '수용할 수 없는 위험성'을 가진 인공지능은 활용 자체가 금지된다. 다음으로 '고위험성'을 가진 인공지능 활용에는 엄격한 준수사항이 요구된다. 세 번째인 '제한된 위험성'을 내포하고 있는 AI의 활용은 투명성 의무가 부과되고, 최소한의 위험성이 있는 경우에는 자율적 규제가 따라붙는다(윤혜선, 2024). 위의 네 가지 위험성을 순서대로 하나씩 살펴보겠다.

첫째, 가장 높은 수준의 '수용할 수 없는 위험성(unacceptable risk)'의 경우를 살펴보자. 수용할 수 없는 위험성은 AI 시스템이 사람들의 안전, 건강, 기본권에 명백한 위협이 되는 경우에 해당한다. 인공지능법 제5조는 수용할 수 없는 위험성이 있다고 간주하는 사례를 제5조 1항에서 (a)~(h)에 걸쳐 여덟 가지를 제시했다. 여덟 가지를 짧게 요약하면 다음과 같다(Artificial Intelligence Act Article 5 (1)). 첫째, 사람의 의식을 넘어서는 잠재적인 기술을 활용하거나 의도적으로 조작적이거나 기만적인 기술을 사용해서, 개인 또는 집단의 행동을 실질적으로 왜곡해서 중대한 피해를 초래하거나 그럴 가능성이 있는 경우이다. 둘째, 연령, 장애 또는 사회경제적 환경에 기인한 취약점을 악용하여 행동을 왜곡함으로써 심각한 피해를 유발하거나 그럴 가능성이 있는 경우이다. 셋째 자연인 또는 집단의 사회적 행동이나 알려진, 추론된, 또는 예측된 개인적 또는 성격적 특성을 기반으로 일정 기간 평가하거나 분류하기 위해 사회적 평점(social scoring)를 사용한 경우다. 넷째, 프로파일링 또는 성격 특성만을 기반으로 하여 개인의 범죄 행위를 예측하는 경우이다. 다섯째, 인터넷이나 CCTV 영상에서 무차별적으로 얼굴 이미지를 스크랩하여 얼굴 인식 데이터베이스를 구축하는 경우이다. 여섯째, 의료 또는 안전 이외의 목적으로 직장이나 교육 기관에서 개인의 감정을 추론한 경우이다. 일곱째 인종, 정치적 견해, 노동조합 가입 여부, 종교적 혹은 철학적 신념, 성

생활, 성적 지향 등과 같이 민감한 개인 특성을 유추하기 위한 생체 인식 분류 시스템(biometric categorisation systems)인 경우이다. 마지막으로 법 집행 목적으로 공공장소에서 실시간 원격 생체 인식 정보('real-time' remote biometric identification)를 수집한 경우다. 여덟 가지 가운데 사회보장 영역과 가장 밀접한 내용은 세 번째인 사회적 평점 부분이다. 관련 내용은 아래 제4절에서 살펴보겠다.

수용할 수 없는 위험성과 관련한 금지 행위를 위반한 경우에는 최대 3,500만 유로(약 500억 원)와 전년 회계연도 기준 전 세계 연매출액의 최대 7% 가운데 더 높은 금액이 과징금으로 부과된다(윤혜선, 2024). 단, 중소기업이 위반했을 때는 위 두 기준 가운데 더 낮은 금액이 과징금으로 부과된다. 유럽연합 소속 기관, 대리인, 조직의 경우에는 최대 150만 유로(약 20억 원)가 부과된다.

두 번째로 고위험성(high risk)을 살펴보겠다. 인공지능법 제6조 제3항에 따르면, 고위험성을 가진 인공지능 시스템은 자연인의 건강, 안전 또는 기본권에 위해를 가할 중대한 위험을 내포하는 시스템이다. 인공지능법 "부속서 III"에서 제시된 여덟 개 영역의 고위험성을 요약하면 다음과 같다.[17] ① 생체인식 및 분류, 감정인식 시스템,[18] ② 주요 사회기반 시설, ③ 교육, 직업훈련 영역, ④ 고용, 근로자 관리 및 자영업자에 대한

17) 유럽연합의 인공지능법에서는 고위험성 인공지능을 부록 I과 부록 III에서 각각 다르게 제시하고 있다. 부록 I에서 제시하는 내용은 사회보장 영역과 무관해서 별도로 설명하지는 않는다.

18) 생체인식 및 감정인식 시스템은 앞의 '수용할 수 없는 위험'으로 원칙적으로 금지되지만, 유럽연합 혹은 개별 회원국의 법에 따라 예외적으로 허용되는 경우는 '고위험군'으로 분류된다(유럽연합 인공지능법 부록 III 1호). 공공장소에서 실시간 생체 인식이 예외적으로 인정되는 경우로는 실종된 아동을 수색해야 하는 경우, 구체적이고 임박한 테러 위협을 방지해야 하는 경우, 중대한 범죄 행위의 가해자 또는 용의자를 탐지, 위치 파악, 식별 또는 기소해야 할 경우 등이다. 유럽 일부 국가에서는 인공지능 사용에 따른 프라이버시 침해 등의 논란을 우회하는 하나의 방식으로 이와 같은 인공지능 사용 목적에 보안(security)이나 안전(safety)을 중심에 두고 접근하는 경향이 나타나고 있다(van Bekkum, Borgesius, 2021).

접근, ⑤ 필수 민간 및 공공 서비스, ⑥ 인간의 기본권과 관련된 법 집행, ⑦ 이주, 망명 및 국경 통제 관리, ⑧ 사업절차 및 민주적 절차이다. 여덟 가지 가운데 ⑤ 필수 민간 및 공공 서비스 분야(essential private and public services)는 사회보장 영역과 직접적으로 연관된다. 해당 내용도 아래 제4절에서 살펴보겠다.

고위험 인공지능 시스템이 시장에 출시되기 전에 준수해야 할 사항으로 다음의 일곱 가지가 제시된다(European Commission, 2024; 윤혜선, 2024). ① 적절한 위험 평가 및 완화 시스템 구축, ② 리스크 및 차별적 결과를 최소화하기 위한 고품질 데이터 구축 및 관리, ③ 결과의 추적 가능성을 보장하기 위한 활동 기록(logging), ④ 시스템과 그 목적에 대한 모든 정보를 제공하는 상세한 문서화(규제 준수 여부 확인 목적), ⑤ 사용자에게 명확하고 적절한 정보 제공, ⑥ 위험을 최소화하기 위한 적절한 인간의 감독 조치, ⑦ 높은 수준의 견고성, 보안 및 정확성이다. 유럽연합에서는 인공지능 시스템을 둘러싼 이해관계자를 제공자, 배포자, 공인 대리인, 수입업자 등으로 분류하고, 각자에 대한 의무를 구체적으로 명기하고 있다. 이를테면, 제공자(provider)에게는 유럽연합 적합성 선언 작성, CE 마크 작성 등의 17개 의무가 부과된다. 사회보장 영역에서 작동하는 다수의 인공지능 시스템은 이러한 규제에 놓이게 될 가능성이 높다.

유럽연합의 인공지능법에서 규정한 이해관계자들은 법의 적용 범위와 관련해서 중요한 의미가 있다. 인공지능법 2조는 법의 적용 범위를 규정하면서, 일곱 가지 유형의 이해당사자 가운데 하나로 "유럽연합에서 AI 시스템 출시·서비스화 & 범용 AI 모델을 출시하는 제공자"라고 규정했다. 그러면서 동시에 '장소 불문("irrespective of whether those providers are established or located within the Union or in a third country)"(Artificial Intelligence Act Article 2 (1))이라는 조건을 달

았다. 즉, 유럽에서 서비스를 제공하는 경우라면, 인공지능 시스템을 출시하거나 적용하는 업체는 국적을 불문하고 규제의 대상이 된다.[19] 국제적으로 인공지능 산업을 대부분 선도하는 미국의 업체들도 유럽연합에 와서는 규제를 따라야 한다. 이는 한국에도 간접적으로 영향을 미칠 수밖에 없다.

이러한 고위험성 인공지능 시스템 준수사항을 위반하면, 최대 1,500만 유로(약 218억 원) 혹은 전년 회계연도 기준 전 세계 매출액의 최대 3% 중에 더 높은 금액이 부과된다(European Commission, 2024; 윤혜선, 2024). 중소기업이 고위험 관련 조항을 위반하면 두 기준 가운데 더 낮은 금액이 부과된다. 유럽연합의 기관, 에이전시, 기구의 경우, 과징금은 최대 75만 유로(약 10억 원)로 한정된다.

유럽연합이 적시한 네 가지 유형의 위험성 가운데 세 번째는 제한적인 위험성이다. 제한적인 위험성이란 AI 사용에 있어 투명성이 부족한 것에 기인하는 위험성을 뜻한다. 인공지능법은 인간에게 정보를 제공하거나 신뢰를 높이기 위한 목적으로 투명성에 관한 구체적인 의무를 규정한다. 예를 들어, 챗봇 같은 AI 시스템을 사용할 때 인간은 기계와 상호작용하고 있다는 사실을 인지할 수 있어야 하고, 그래야 계속 챗봇을 사용할지를 결정할 수 있다는 것이다(European Commission, 2024). 전 세계적으로 사회보장 분야에서 가장 많이 사용되는 인공지능 기술이 챗봇인 점을 참고할 필요가 있다(Zaber, Casu, Brodersohn, 2024).

---

19) 물론 예외도 있다. ① 군사, 국방, 국가 안보 목적의 AI 시스템, ② 유럽연합 또는 그 회원국과 사법 공조 협약을 체결한 제3국의 공공기관이 그 협약의 체계 안에서 사용하는 AI 시스템, ③ 과학적 연구 및 개발 목적으로만 특별히 개발되고, 서비스가 제공되는 AI 시스템과 AI 모델, ④ 출시 및 서비스 개시 전 AI 시스템을 연구, 테스트, 개발하는 활동, ⑤ 고위험성 AI 시스템이 아닌 무료 및 오픈소스 라이선스로 출시된 AI 시스템, ⑥ 사적 및 비전문적인 활동, ⑦ 일부 목적을 위한 법 집행기관(경찰, 이민 당국 등)의 AI 시스템 사용 등에는 인공지능법 적용이 제외되거나 특례가 인정된다(윤혜선, 2024).

마지막으로 최소한의 위험성이 있다. 최소한의 위험성을 내포하는 AI 시스템에는 스팸 필터, 알고리즘을 통한 콘텐츠 추천 시스템 등이 해당된다. 인공지능법은 이와 같이 최소한의 위험성에 해당하는 AI 시스템에 별도의 규제를 규정하지 않고, 자율적 규제를 허용하고 있다.

다. 유럽연합 인공지능법이 사회보장에 미칠 영향

유럽연합의 인공지능법에서 사회보장과 가장 연관된, 가장 민감한 대목[20]은 '수용할 수 없는 위험성(unacceptable risk)'에서 세 번째로 제시된 사회적 평점(social scoring)이다. 사회적 평점은 "개인의 데이터를 기반으로 개인을 평가하는 것을 의미하며, 여기에는 신용 행태, 교통위반, 사회적 참여 등이 포함된다. 이러한 평가는 특정 서비스나 특권에 대한 접근을 규제하기 위해 사용된다"(Mosene, 2024). 유럽연합의 인공지능법을 본문 그대로 번역하면 다음과 같다(European Parliament, 2024, Chapter II, Article 5, 1. (c)).

"자연인 또는 집단의 사회적 행동이나 알려진, 추론된, 또는 예측된 개인적 또는 성격적 특성을 기반으로 일정 기간에 대상을 평가하거나 분류하기 위한 목적으로 인공지능(AI) 시스템을 시장에 출시하거나, 서비스에 투입하거나 사용하는 행위로서, 이러한 사회적 평점이 다음 중 하나 이상의 결과로 이어지는 경우:

(ⅰ) 데이터가 원래 생성되거나 수집된 맥락과 무관한 사회적 맥락에서 특정 자연인 또는 집단에 대해 불리하거나 부정적인 대우를 초래하는 경우.

20) 한 가지 확인한 점은 있다. 해당 주제에 대한 분석을 담은 학술논문이나 보고서는 아직 찾기는 어렵다. 현재로서는 법률 내용, 관련 시민단체 성명, 국내 법률 전문가 자문 등을 중심으로 관련된 영향을 추정하는 수밖에 없었다.

(ii) 특정 자연인 또는 집단의 사회적 행동 또는 그 행동의 심각성과 비교하여 부당하거나 과도하게 불리하거나 부정적인 대우를 초래하는 경우."

유럽연합의 인공지능법은 사회적 평점을 논의하면서 사회보장제도와 관련한 언급을 하지는 않았다. 유럽의회 산하의 연구기관인 유럽의회연구소(European Parliamentary Research Service)가 인공지능법을 설명하는 짤막한 해설서(Madiega, 2024)에도 사회보장제도에 대한 언급은 없다. 그럼에도, 사회적 평점 부여가 사회보장과 일정한 연관을 가질 수밖에 없다. 개인 혹은 가구 단위의 소득, 재산, 가구원 등의 정보에 근거해서 빈곤, 실업, 은퇴, 상병 여부를 판단하고, 그에 근거해서 급여를 제공하는 사회보장제도는 급여 자격을 판정하는 과정에서 일종의 사회적 평점(social scoring)을 개인 혹은 가구에게 부여할 수밖에 없다. 이를테면, 국내의 국민기초생활보장제도에서도 가구의 소득, 재산, 부양의무자, 가구원 정보 등에 기반해서 수급 자격 및 급여액을 결정한다. 그러한 자료의 내용이 개인의 소득, 연령, 가구, 건강 등 개인적 정보를 담고 있다는 점에서 사회적 평점은 민감할 수밖에 없는 의제다.

유럽연합의 인공지능법 제정 과정에서도 사회적 평점과 관련한 우려가 제기됐다. 유럽의 인권 단체인 Human Rights Watch(2023.10.9.)는 다른 시민단체들과 함께 유럽이사회와 유럽의회에 사회적 평점 관련 규정을 강화하는 제안을 하면서 사회보장제도에서 나타날 문제를 다음과 같이 언급했다. "프랑스, 네덜란드, 오스트리아, 폴란드, 아일랜드에서의 조사 결과, AI 기반의 사회적 평점 부여 시스템이 사람들의 사회보장 지원 접근을 방해하고, 프라이버시를 침해하며, 빈곤에 대한 고정관념과 차별적인 방식으로 데이터를 프로파일링하고 있음이 드러났다(Human Rights Watch, 2023.10.9.)." 이러한 문제들은 시민단체나 학계를 중심

으로 꾸준히 제기됐다. 프랑스(La Quadrature Du Net, 2023), 네덜란드(van Bekkum, Borgesius, 2021)와 덴마크(Jørgensen, 2021)에서도 논란이 일었다. 이를테면, 정부에서 복지 급여의 부정수급을 탐지하거나 위험 가구를 발굴하는 과정에서 개인정보를 무리하게 활용하거나(Appelman et al.,2021), 알고리즘이 취약계층에게 편향적으로 작동했다(Van Bekkum, Borgesius, 2021)는 지적이다. 네덜란드에서는 이와 같은 문제들 때문에 정부에서 추진하던 사회보장 부정수급 추적 알고리즘인 시스템 위험도 표시(Systeem Risico Indicatie) 시스템이 법원의 제동에 걸리면서 작동이 중단된 바 있다(Bekker, 2021).

유럽연합도 사회적 평점과 관련한 비판 및 부작용을 염두에 둔 것으로 보인다. 법률에 붙은 (ⅰ)와 (ⅱ)의 내용에서 인공지능의 사회적 평점이 '수용할 수 없는 위험'으로 간주되는 경우를 한정했다. 즉, 특정 집단이나 개인에게 불리하거나 부정적이거나 부당한 대우를 초래할 때만 사회적 평점이 금지된다. 바꾸어 말하면, 사회보장제도에서 인공지능의 작동이 '순기능'을 하는 경우에는 규제의 대상으로 삼지 않겠다는 뜻이다. 그렇지만, 논란이 종료되기는 어렵다. 사회보장 영역에서 인공지능이 급여 대상자를 판단 혹은 추론하는 과정에서 데이터 편향성 등의 문제로 특정 집단에 대한 급여를 변경/중지/중단한다면, 이는 불리/부정/부당할 수 있다(강지원, 2024). 실제로, 덴마크나 네덜란드 등 다수의 국가에서 사회보장제도에 순기능으로 작용할 것으로 기획된 인공지능 알고리즘들이 결과적으로 차별과 편향을 낳았다는 비판에 직면했다(Jørgensen, 2021, Bekker, 2021). 이 대목은 앞으로 인공지능 기술이 사회보장 영역에서 적용되는 과정에서 끊임없이 논점으로 부각될 것으로 예상된다.

다음으로 인공지능법에서 두 번째로 수위가 높은 '고위험(high risk)' 인공지능 영역을 보겠다. 여기에서는 다섯 번째 고위험 인공지능 영역

으로 '필수 민간 및 공공 서비스 분야'가 제시됐다. 이 영역은 사회보장 영역을 직접적으로 언급하고 있다. 고위험 인공지능을 영역별로 상술한 유럽연합 인공지능법 부록(Annex) III에서 사회보장을 다음과 같이 언급했다(Artificial Intelligence Act Annex III, 5).

"필수 민간 서비스 및 필수 공공 서비스와 혜택에 대한 접근 및 이용[21]:

(a) 공공기관 또는 공공기관을 대신하여 사용될 목적으로, 자연인의 필수적인 공공 복지 급여[22] 및 서비스(의료 서비스 포함)에 대한 자격을 평가하거나, 해당 혜택 및 서비스를 부여, 축소, 취소 또는 회수하기 위해 사용되는 AI 시스템."

'고위험'으로 분류한 '필수 민간 및 공공 서비스'에는 앞에서 살펴본 사회적 평점(social scoring)과 밀접하게 연관됨을 알 수 있다. 급여를 제공하는 사회보장 영역에서 자격 요건을 판정하기 위해서는 사회적 평점 산정이 대부분 필요하기 때문이다. 유럽연합법은 관련 고위험에 대해서 다음과 같이 상세하게 설명했다(Artificial Intelligence Act, (58)).

"필수적인 공적 복지 급여와 서비스를 신청하거나 받는 자연인은... 공공기관 앞에서 취약한 위치에 있다. 이러한 급여와 서비스가 지급, 거부, 축소, 취소 또는 회수되어야 하는지 여부를 결정하기 위해 인공지능 시스템이 사용된다면, 해당 시스템은 사람들의 생계에 상당한 영향을 미칠 수 있고, 사회보호에 대한 권리, 차별 금지, 인간의 존엄성 또는 효과적인 법적

21) 급여 수급에 관한 내용을 담은 (a) 외에도 보건의료 영역에서 다음과 같은 언급도 있다. 사회보장 영역과 일정한 연관성이 있지만, 이 글에서는 해당 내용까지 다루지는 않는다. "(c) 생명 및 건강 보험의 경우, 자연인에 대한 위험 평가 및 가격 책정을 위해 사용되는 AI 시스템; (d) 자연인의 긴급 호출을 평가 및 분류하거나, 경찰, 소방관, 의료 지원을 포함한 긴급 대응 서비스의 출동 우선순위를 결정하거나 출동하는 데 사용되는 AI 시스템, 그리고 응급 의료 환자 분류 시스템."

22) 인공지능법에서는 'public assistance benefit'이라는 용어를 사용했고, 사회보장 영역에서는 이를 공공부조 급여로 번역하는 것이 적정할 듯하지만, 법의 맥락에서는 공공부조에 한정되지 않는 공적인 복지 급여 전체를 지칭하는 것으로 보았다. 참고로, 유럽에서는 공공부조에 대해서 social assistance라는 표현을 더 일반적으로 사용한다.

구제 같은 기본권을 침해할 수 있으므로 고위험으로 분류되어야 한다."

여기까지만 보면, 유럽연합의 입장은 사회보장제도에서의 인공지능 적용에서 엄격한 입장으로 보인다. 앞의 제3절에서 살펴본, 다소 강한 규제 내용이 사회보장 영역에서 적용될 여지도 크다. 그렇지만, 법의 다음 문장에서는 현행 사회보장제도에 인공지능 기술 적용의 가능성도 열어둔다.

"이 규정은 공공행정이 준수되고 안전한 AI 시스템의 광범위한 사용을 통해 이익을 얻을 수 있도록 혁신적인 접근 방식의 개발과 사용을 저해해서는 안 된다. 단, 해당 시스템이 법적 및 자연인에게 고위험을 초래하지 않아야 한다"(Artificial Intelligence Act, (58)).

사회보장 행정을 집행하는 정부 및 공공기관은 고위험 인공지능 시스템을 이용하는 과정에서 '배포자(deployer)로 구분될 가능성이 높다(강지원, 2024). 배포자는 고위험성 인공지능을 활용할 때 13가지의 의무를 이행해야 한다. 이를테면, 데이터보호 영향평가, 기본권 영향평가(fundamental rights impact assessment) 등이 그 예가 된다. 정부 및 공공기관 입장에서는 부담이 크다.

고위험 인공지능 관련 규정이 유럽 사회 및 사회보장에 미칠 영향에 대해서 유럽 사회는 다소 유보적인 입장으로 보인다. 독일 사회보험 유럽대표부(Deutsche Sozialversicherung Europavertretung, 2024)는 "인공지능법 도입 이후 사회적 영향을 면밀히 모니터링해야 한다"고 논평했다. 이러한 반응은 아직 인공지능법의 내용이 전반적으로 모호하고, 구체적인 후속 조치도 이루어지지 않았기 때문으로 보인다. 물론 유럽 노동조합 측에서는 인공지능법이 관련 대기업에 빠져나갈 구멍을 만들어주었다는 비판도 가하고 있다(Vranken, 2023; Del Castillo, 2023). 따라서 인공지능법 규제의 실질적인 강도는 향후 추가로 마련될 가이드라인, 기준, 표준, 행동강령, 판례 등의 내용에 따라 결정될 것으로 보인다. 앞으

로 인공지능법의 구체화를 위한 논의와 협상에서 기업, 노동자단체, 인권단체, 관료 집단 등 이해당사자 사이에 첨예한 대립과 갈등이 발생할 가능성이 크다. 그 결과에 따라 이 법이 사회보장에 미치는 의미와 영향력이 달라질 것으로 보인다.

지금까지 유럽연합 인공지능법의 배경과 내용을 일람하고, 사회보장 영역에 미칠 영향을 살펴보았다. 유럽연합의 인공지능법이 한국에 주는 함의를 검토하기 위해서는 인공지능법이 가지는 두 가지 성격을 먼저 살펴볼 필요가 있다. 첫째, 이 법은 지금까지 나온 AI에 관한 가장 포괄적인 법적 프레임워크라는 점이다. 미국이나 중국 등이 인공지능 기술과 산업의 육성을 중심으로 접근한다면, 유럽연합은 인공지능이 가지는 파괴력을 제어하기 위해 상대적으로 인간 중심적, 윤리적 접근을 취하고 있다. 인공지능이 가지는 잠재적인 위험성을 고려할 때, 전 세계가 인공지능에 대한 규제를 마련한다면 유럽연합의 인공지능법은 권위 있는 기준이 될 것이다. 둘째, 유럽연합의 인공지능법은 인공지능을 선도하는 미국 및 중국에 유럽이 던지는 견제구의 성격도 있다(Petrosyan, Ataliotou, 2024). 영국의 언론기관인 Tortoise Media(2024)에 따르면, 인공지능 기술 역량을 기준으로 한 국가별 순위는 미국이 압도적인 1위이고, 다음으로 중국, 싱가포르, 영국, 프랑스, 한국, 독일, 캐나다 등의 순이었다. 유럽연합의 시도는 인공지능 규제에 대한 국제적인 표준을 주도하려는 시도로도 해석된다(윤혜선, 2024).

사회보장 영역에서 인공지능 규제의 국제 표준으로서 유럽연합 인공지능법이 한국에 미칠 영향을 현재로서 단정하기 힘들다. 유럽연합 인공지능법이 가지는 모호함과 추상성도 적지 않다. 법 집행 과정에서 적용 사례와 판례가 누적되면서 그 영향도 앞으로 구체화할 것으로 예상된다.

2. 미국의 행정명령 14110

미국의 연방정부는 최근까지 인공지능 기술에 대해서 강제력 있는 규제를 가하는 데 주저해 왔다. 2023년에 들어서야 바이든 행정부는 미국 연방정부 최초로 관리예산처(OMB)의 각서 초안과 함께 안전하고 신뢰할 수 있는 인공지능의 개발 및 사용에 관한 행정명령(Executive Order o the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 이하 행정명령 14110)을 발표했다. 이를 통해서 미국은 비로소 연방 단위의 AI 규제에 참여했다. 물론, 이러한 행정명령은 의회의 추가적인 조치가 따르지 않으면 명령의 시행에 일정한 한계가 있는 것도 사실이다. 그렇지만, 행정명령을 보면, 미국 및 글로벌 차원에서의 AI 규정이 미래에 어떤 모습일지에 대한 추정이 가능하다.

미국의 행정명령 14110은 전체 13조(section)로 구성됐다(US Executive Order No. 14028, 2023, 〈표 4-8〉 참고). 1조에서는 행정명령의 목적을 제시하고, 3조에서는 주요 용어에 대한 개념을 정의했다. 2조가 전체 구조에서 핵심이라고 볼 수 있는데, 인공지능 기술 활용에서의 8대 원칙을 제시했다. 이 명령에서는 8대 원칙이 뒤에 이어지는 4~11조의 내용을 구성하도록 구성했다. 즉, 1원칙 'AI 기술의 안전과 보안 확보'가 4조의 내용이 되고, 2원칙 혁신과 경쟁 촉진이 5조의 내용이 된다. 이러한 순서로 제시된 3~8의 원칙은 다음과 같다. 노동자 지원(6조), 형평성과 시민권 증진(7조), 소비자, 환자, 승객, 학생 보호(8조), 개인정보 보호(9조), 연방정부의 AI 활용 증진(10조), 해외에서의 미국 리더십 강화(11조) 등이다. 그밖에 12, 13조는 시행 및 일반 규정에 관한 내용을 담았다.

내용을 조금 더 살펴보자. 행정명령 1조(목적)에서는 인공지능이 불러올 혜택과 위험을 모두 제시한 뒤, 무수한 이점을 활용하기 위해 정부, 민간, 학계, 시민사회의 협력을 강조했다. 2조에서는 8개 원칙 가운데 '혁신, 경쟁, 협력의 촉진(2원칙)' 혹은 '글로벌 사회, 경제, 기술 발전의 선도(8원칙)'보다 '안전과 보안'(1원칙), '노동자 지원'(3원칙) 같은 가치들이 선순위에 있는 점이 눈에 띈다.

3조에서는 주요한 개념들을 정의했다. 행정명령의 대상이 되는 '기관(agency)'은 미국법전 44편 35장에서 제시된 기관의 정의에 따랐다. 여기서 기관은 "정부 행정부(대통령 행정실 포함)의 행정부, 군부, 정부 법인, 정부 통제 법인 또는 기타 기관 또는 독립 규제 기관"(US Exec. Order No. 14110, 2023, p. 75193)이다. 인공지능은 다음과 같이 정의됐다. "인간이 정의한 일련의 목표에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권장 사항 또는 결정을 내릴 수 있는 기계 기반 시스템. 인공지능 시스템은 기계 및 인간 기반 입력을 사용하여 실제 및 가상 환경을 인식하고, 자동화된 방식으로 분석을 통해 이러한 인식을 모델로 추상화하며, 모델 추론을 사용하여 정보 또는 행동에 대한 선택을 형성"(US Exec. Order No. 14110, 2023, p. 75193)하는 것이다.

〈표 4-8〉 미국 행정명령 14110의 개요

구분	내용
1조. 목적	- 인공지능이 불러올 혜택과 위험을 모두 제시한 뒤, 무수한 이점을 활용하기 위해 정부, 민간, 학계, 시민사회의 협력 강조
2조. 정책 및 원칙	- 행정명령은 다음 여덟 가지의 지도 원칙과 우선순위에 따라 인공지능의 개발과 이용을 발전시키고 관리하는 것을 정부의 정책으로 규정. 원칙의 중요성을 고려해서 원문 문장 해석. ① 인공지능은 안전과 보안이 보장돼야 함. ② 책임감 있는 혁신, 경쟁, 협력을 촉진함으로써 미국은 AI 분야를 선도하고 사회의 가장 어려운 과제를 해결할 수 있는 기술의 잠재력을 실현할 수 있음. ③ 책임감 있는 AI의 개발과 사용을 위해서는 미국 노동자들을 지원하겠다는 약속이 필요. ④ 인공지능 정책은 형평성과 시민권 증진에 대한 우리 행정부의 헌신과 일관성을 유지해야 함. ⑤ 일상생활에서 인공지능과 인공지능이 탑재된 지원 제품을 점점 더 많이 사용하고, 상호 작용하고, 구매하는 미국인들의 이익은 보호돼야 함. ⑥ AI가 계속 발전함에 따라 미국인의 프라이버시와 시민의 자유도 보호돼야 함. ⑦ 미국인들에게 더 나은 결과를 제공하기 위해 연방정부 자체의 AI 사용으로 인한 위험을 관리하고 책임감 있는 AI 사용을 규제, 관리 및 지원할 수 있는 내부 역량을 강화하는 것이 중요함. ⑧ 연방정부는 이전 파괴적 혁신과 변화의 시대에 미국이 그랬던 것처럼 글로벌 사회, 경제, 기술 발전을 선도해야 함.
3조. 정의	- 주요 개념에 대한 정의를 제시. - 이를테면, 인공지능은 “인간이 정의한 일련의 목표에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권장 사항 또는 결정을 내릴 수 있는 기계 기반 시스템. 인공지능 시스템은 기계 및 인간 기반 입력을 사용하여 실제 및 가상 환경을 인식하고, 자동화된 방식으로 분석을 통해 이러한 인식을 모델로 추상화하며, 모델 추론을 사용하여 정보 또는 행동에 대한 선택을 형성”하는 것으로 제시됨.
4조. AI 기술의 안전과 보안 확보	- 안전성과 보안의 확보를 위해 다음 8가지 규율 시행 ① AI 안전 및 보안을 위한 지침, 표준, 모범사례 개발 ② 안전성과 신뢰도를 갖춘 AI의 보장 ③ 주요 인프라 및 사이버 보안에서의 AI 관리 ④ AI와 화생방 및 핵무기 위협의 교차로 인한 위험 감소 ⑤ 합성 콘텐츠로 인한 위험 감소 ⑥ 기반 모델의 입력값으로 널리 활용되는 모델 가중치 투입 촉구 ⑦ AI 학습을 위한 연방 데이터의 악의적 이용 방지 및 안전한 배포 촉진 ⑧ 국가 안보 각서의 개발 지시

구분	내용
5조. 혁신과 경쟁 촉진	- 인재 유치 ① AI 연구원의 비자 신청, 처리, 갱신 등 절차 간소화 ② 해외 인재 유치를 위한 다양한 프로그램 수립 - 혁신 촉진 ① 과거 권고된 AI 연구 자원(NAIRR) 계획 시범 프로그램 실행 ② 펀딩, 연구소, 교육 프로그램 등 신설 ③ 생성형 AI를 포함한 AI 관련 지식재산권 명확화 논의 수행 ④ 의료, 환경 등 분야의 AI에 대한 보조금 지급 - 경쟁 촉진 ① AI 정책 및 규제는 경쟁법 및 연방거래위원회의 역할과 조화 ② 반도체의 특수성을 고려한 다양한 경쟁 촉진 조치 시행 ③ 중소기업의 혁신과 책임성을 강화하기 위한 다양한 지원 강화
6조. 노동자 지원	- AI가 노동시장에 끼치는 영향력 및 실업자 지원방안을 담은 보고서 제출 - 고용 및 직무평가, 모니터링 등에 활용되는 AI가 **노동자의 복지에 미칠 수 있는 잠재적 피해를 완화하고 이익을 극대화하는 데 활용**될 수 있는 사용자 원칙과 모범사례 개발 및 공표
7조. 형평성과 시민권 증진	- 형사 사법 시스템에서의 AI 및 시민권 강화를 위해, 인권의 침해와 차별 없이 형사 사법 시스템에 AI를 활용하는 방법을 설명하는 보고서 제출 - **정부 복지 프로그램에서 시민권 보호를 위해, AI를 활용할 때 공평하고 정당한 결과가 도출되는지에 대한 분석 및 그러한 AI 활용법에 대한 지침 개발** - 광범위한 경제적 영역에서 AI 및 시민권 강화를 위해, 고용, 부동산, 금융, 장애 등과 관련된 맥락에서 차별 금지를 달성하기 위한 지침 개발
제8조. 소비자, 환자, 승객, 학생 보호	- 독립 규제기관을 통한 모델의 투명성 및 모델 사용에 대한 설명 능력의 제고를 포괄한 제반 조치에 의한 미국 소비자 보호 - 특히 **헬스케어, 공중보건, 인적 서비스,** 교통, 교육, 통신 등의 **영역에서 AI의 책임 있는 개발, 배치, 이용 보장 강조**
제9조. 개인정보 보호	- 기관이 확보한 상업적으로 가용한 정보에 대한 프라이버시 위험 완화 조치 수행 - 차분 프라이버시를 포함한 프라이버시 강화 기술(PET) 구축 및 평가 노력 촉구
제10조. 연방정부의 AI 활용 확대	- 인공지능 관리 지침 발행 - **기관은 AI 위험관리 업무를 담당한 최고 인공지능 책임자를 지정**하여 다음을 포함한 업무 수행 **① AI 권리장전 청사진 및 위험관리 프레임워크에 제시된 실무 관행 수립** **② 레드팀 등 생성형 AI에 대한 외부 테스트 개발** **③ 생성형 AI의 생성물에 워터마크, 라벨 지정** - 예산관리국 국장은 각 기관의 AI 위험관리 능력 평가 방법론 개발 - 다만 기관 내 생성형 AI의 사용에 대한 광범위한 금지는 권장하지 않음. - 정부 내 AI 인재 확대 - AI 전문가 인력 확보를 위한 다양한 노력 시행

구분	내용
제11조. 해외에서의 미국 리더십 강화	- 군사 및 정보 분야 이외에서의 AI 개발 및 이용에 관한 책임 있는 글로벌 기술 표준을 발전시키기 위한 노력 시행 - 위험관리 프레임워크의 원칙, 지침, 모범사례를 국제적 맥락에서 반영한 글로벌 개발 플레이북 발간 - 주요 인프라에 대한 국경 간, 글로벌 위험 방지를 위한 노력 촉구
제12조. 시행	- 대통령 행정실에 행정명령 및 AI 관련 정책의 효과적 조정을 위한 백악관 AI 위원회(White House AI Council) 조직 설립 - 위원회에는 각부 장관과 주요 기관 인사 포진
제13조. 일반 규정	- 행정명령은 법률의 수권 범위 및 예산국 국장의 예산, 행정, 입법 제안과 관련된 역할에 제약을 가할 수 없음 - 행정명령은 관련 법 및 예산의 가용 범위에서 실행 - 행정명령은 어떤 실체적, 절차적 차원의 법적 권리나 이익을 발생시키지 않음

주: '2조, 정책 및 원칙'에서 볼드 및 밑줄은 필자가 강조를 위해 사용.
출처: 김정욱 외. (2023). "인공지능 시대의 경쟁력 강화를 위한 AI 규제 연구". pp. 145-146를 일부 수정 및 보완. 저작권 2023. 경제인문사회연구회.

행정명령 가운데 사회보장과 관련된 부분은 7조(형평성과 시민권 증진)과 8조(소비자, 환자, 승객, 학생 보호) 부분이다. 6조도 노동자의 복지에 관한 내용을 일부 담고 있다. 또, 10조에서는 미국 보건복지부(Ministry of Health and Human Services)를 포함한 기관들의 관리 지침을 제시하고 있다.

먼저, 7조에서 2항 부분은 '정부의 급여 및 프로그램과 관련한 시민권의 보호'로 명시된다. 내용을 보면, 기관들은 연방정부의 프로그램과 급여를 집행하는 과정에서 인공지능 기술의 활용으로 인한 "불법적인 차별 및 기타 피해를 예방 및 해결"(prevent and address unlawful discrimination and other harms)"(US Exec. Order No. 14110, 2023, p. 75212)하기 위해 각자 기관에 있는 민권 및 시민 자유 사무소(civil rights and civil liberties office)를 활용해야 한다고 규정하고 있다. 여기서 민권 및 자유 사무소는 미국에서 시민의 기본권 및 자유에 영향을 미칠 수 있는 주요 기관에 마련된 부서를 지칭한다.[23] 다만, 이러한 지침이 민사 혹은 형사 집행 당국에는 적용되지 않는다. 이 원칙에 따라 행정명령

에서는 두 개 부처의 장관직을 직접 호명하면서 관련한 지침을 제시하고 있다. 두 장관직 가운데 하나가 다름 아닌 보건복지부 장관(The Secretary of Health and Human Services)이다.

행정명령 7조 2항 (b)호에 따르면, 정부 지급 급여의 공평성 제고를 위해서 보건복지부는 급여 및 서비스를 시행할 때 자동화 또는 알고리즘 시스템의 사용을 촉진하는 계획(plan)을 행정명령 시점 기준으로 180일 이내에 발표해야 한다. 그러한 지침의 목적은 다음과 같은 절차를 개선하기 위함이다. 첫째, 자격을 갖춘 수급자의 접근성에 대한 평가, 둘째, 인공지능 시스템의 존재에 대한 수급자 대상 통지, 셋째, 급여의 부당한 거부를 감지하기 위한 정기적인 평가, 넷째, 전문기관 직원(expert agency staff)이 적절한 수준의 재량권을 유지하도록 하기 위한 절차, 다섯째, 급여 거부에 대해 급여 신청자가 사람인 심사자에게 이의를 제기할 수 있도록 하는 절차, 여섯째, 급여 프로그램에서 사용 중인 알고리즘 시스템이 공평하고 공정한 결과를 달성하는지 분석하는 것이다(US Exec. Order No. 14110, 2023, pp. 75212~3).

행정명령의 8조(소비자, 환자, 승객, 학생 보호)에서는 직접적으로 관련 정부 부처들을 호명하면서 일일이 지침을 제시했다. (a)항은 독립 규제기관(independent regulatory agencies), (b)항은 보건복지부, (c)항은

23) 미국의 보건복지부라 할 수 있는 Department of Health and Human Services에는 Office for Civil Right(OCR)가 행정 집행 과정에서의 차별과 인권 보호의 문제를 다룬다. 이를테면, OCR의 한국어 누리집(https://www.hhs.gov/ocr/get-help-in-other-languages/korean.html)에서는 아래와 같이 안내하고 있다. "미국 Department of Health and Human Services(HHS, 보건복지부) 산하 Office for Civil Rights (OCR, 민권담당국)는... 비차별, 양심 및 종교적 자유, 대상 주체에서 의료 정보 프라이버시에 대한 시민들의 기본권을 보호하는 연방 민권 법률, 양심 및 종교적 자유 법률, Health Insurance Portability and Accountability Act(HIPAA, 건강보험 이전 및 책임에 관한 법률), Privacy, Security, and Breach Notification Rules(개인정보 보호, 보안 및 침해 고지 규칙), Patient Safety Act and Rule(환자안전법 및 규칙)을 집행합니다."

교통부, (d)항은 교육부, (e)항은 연방통신위원회에게 주는 지침이다. 이 가운데 (b)항인 보건복지부 대상 지침이 가장 길다. 대략 3쪽 분량의 8조 내용 가운데 (b)항만 2쪽가량을 차지한다. 보건복지부가 수행하는 업무의 민감성이 반영된 결과로 추정된다.

보건복지부에 대한 지침 사항을 담은 (b)호는 '의료, 공중보건 및 복지[24] 부문에서 안전하고 책임감 있는 인공지능의 배포와 사용을 보장하기 위해'라는 목적을 제시하면서, 전체 네 개의 세부 '목'을 제시하고 있다. (ⅰ)목은 보건복지부가 설치해야 할 태스크 포스에 대한 지침이고(90일 시한), (ⅱ)목은 보건복지부의 인공지능 기반 기술의 품질 평가 판단 기준 개발에 대한 지침이다(180일 시한). (ⅲ)목은 정부의 재정 지원을 받는 보건복지 서비스 제공 업체에 대한 규제 관련 고려사항을 담고 있다(180일 이내). 그리고 (ⅳ)목은 의료 영역에서 인공지능 관련 안전 프로그램의 수립을(365일 이내), (ⅴ)목은 신약 개발에서 인공지능 사용에 대한 규제 전략을 제시하고 있다(365일 이내).

해당 내용을 하나씩 살펴보면 다음과 같다. 먼저 (ⅰ)목에서는 보건복지부 안에 인공지능 태스크 포스의 구성이 명령일로부터 90일 이내에 이뤄져야 함을 규정한다. "보건복지부 장관은 국방부 및 보훈부 장관들과 협의하여 보건복지 부문, 즉 연구 및 발견, 의약품 및 기기 안전, 의료 전달 및 재정, 공중 보건 영역에서 인공지능 및 인공지능이 탑재된 기술의 책임 있는 배치 및 사용에 대한 정책 및 프레임워크를 포함하는 전략 계획을 개발해야 한다"(US Exec. Order No. 14110, 2023, p. 75214). 전략 계획에는 필요시 적절한 규제 조치를 포함할 수 있다. 보건복지부 내의 인공지능 태스크 포스는 구성 시점 이후 365일 이내에 이러한 전략

24) 원문에서는 이 대목에서 'healthcare, public health and human services'라고 제시됐다. 미국 보건복지부에서 'human services' 부분은 전통적인 아동복지, 노인복지, 빈곤정책 등을 담당하는 점을 고려해서 '복지'라고 번역했다.

계획(strategic plan)을 짜야 한다. 전략 계획은 다음과 같은 영역에서 적절한 지침과 자원을 모색해야 한다(US Exec. Order No. 14110, 2023, pp. 75214~5).

첫째, 품질 측정, 성과 개선, 급여 집행 등 의료 서비스 제공 및 재정 분야에서 예측 및 생성 인공지능(predictive and generative AI) 기반 기술의 개발, 유지, 사용. 여기에서 인공지능이 생성한 결과를 현장에서 적용하는 과정에 대한 인간의 적절한 감독도 함께 고려해야 한다.

둘째, 보건복지 부문에서 인공지능 기반 기술의 장기적인 안전성 및 실제 성과 모니터링. 여기서 모니터링의 대상은 임상적으로 관련이 있거나 중요한 수정 사항 및 성과가 된다. 규제 기관, 개발자 및 사용자에게 제품 업데이트를 알릴 수 있는 수단도 제공돼야 한다.

셋째, 보건복지 부문에서 사용되는 인공지능 기반 기술에 형평성 원칙을 반영. 이를 위해 새로운 모델을 개발할 때 영향을 받는 인구에 대한 분리된 데이터와 대표 인구 데이터 세트를 사용한다. 기존 모델의 알고리즘이 가질 수 있는 차별과 편견 관련 부작용을 모니터링하고, 현재 시스템이 가질 수 있는 차별과 편견을 식별하고 완화하기 위함이다.

넷째, 개인 식별 정보를 보호하기 위해 소프트웨어 개발 라이프사이클에 안전, 개인정보 보호 및 보안 표준을 반영. 보건복지 부문에서 인공지능으로 심화한 사이버 보안 위협을 해결하기 위한 조치도 포함해야 한다.

다섯째, 보건복지 현장에서 사용자가 적절하고 안전한 AI 사용을 결정하는 데 도움이 되는 문서의 개발, 유지관리 및 가용성 확보.

여섯째, 지역 단위 보건복지 관련 기관과 협력하여 인공지능 기술 사용에 대한 긍정적인 현장 사용 사례와 모범사례를 촉진하기 위한 작업 수행.

일곱째, 행정 부담 감소를 포함하여 보건 및 인적 서비스 부문에서 업무 효율성과 만족도를 증진하기 위한 AI 사용 사례 파악.

(b)항의 (ii)목에서는 인공지능 기반 기술의 품질 평가 판단을 위한 전략(strategy) 개발에 대한 내용을 담고 있다. 보건복지부는 행정명령 발표 180일 이내에 이러한 전략을 개발해야 한다. 전략의 내용은 앞선 (i)목에서 제시한 영역을 포괄해야 한다. 전략에는 인공지능 보증정책(AI assurance policy) 및 관련 인프라에 대한 요구 사항이 포함돼야 한다. 인공지능 보증 정책은 인공지능 기반 의료 도구의 성능을 평가하기 위함이다. 두 번째인 인프라에 대한 요구사항은 인공지능 기반 의료 기술 알고리즘 시스템 성능의 시판 전 평가 및 시판 후 감독을 실제 데이터와 대비해서 수행할 수 있도록 하기 위해서다.

(b)항의 (iii)목은 정부의 재정 지원을 받는 보건복지 서비스 제공 업체에 대한 규제의 내용을 담고 있다. (iii)목의 내용에 따르면, 연방 재정 지원을 받는 보건복지 서비스 제공 업체는 인공지능 관련 기술을 적용하는 과정에서 연방 차별 금지법(Federal nondiscrimination laws)을 준수하도록 보건복지부가 적절한 조치(appropriate actions)를 고려해야 한다. 이러한 고려는 180일 이내 이뤄져야 한다. 법에서는 여기서의 '적절한 조치' 두 가지를 예시했다. 한 가지만 보면, 인공지능과 관련된 연방 차별 금지법 혹은 개인정보 보호법 미준수에 대한 불만 또는 기타 신고에 대해 보건복지부가 지침을 발표하는 것이다.

(b)항의 (iv)목은 인공지능 안전 프로그램 수립에 관한 내용이다. 즉, 보건복지부는 미국 국방부 및 보훈부와 협의하고 연방에 등록된 환자 안전 기관(Patient Safety Organization)[25]과 협력하여 인공지능 안전 프로그램을 수립해야 한다. 시한은 행정명령 발표일 기준으로 365일이다. 인

25) 환자 안전 기관(PSO)은 2005년 미국의 환자 안전 및 의료 질 향상법(Patient Safety and Quality Improvement Act)에 의해 만들어졌다(Agency for Healthcare Research and Quality, 2023). 환자의 안전을 개선하기 위해 의료 기관들이 자발적으로 참여한다. 미국 보건복지부 산하의 질병통제센터(Agency for Healthcare Research and Quality, AHRQ)에 의해 등록 및 감독을 받는다.

공지능 안전 프로그램은 다음의 내용을 포함한다.

첫째, 의료현장에서 인공지능 기술을 적용하여 발생하는 임상 오류를 식별하고 포착하기 위한 공통의 프레임워크 수립. 환자, 간병인 등에게 차별과 편견 등으로 인해 피해를 입히는 등의 사고에 대한 중앙 추적 저장소(central tracking repository)[26]의 사양(specifications) 수립.

둘째, 이렇게 수집된 데이터와 생성된 증거를 분석하여 적절한 경우 이러한 피해를 방지하기 위한 권고안, 모범사례 또는 기타 비공식 지침 개발.

셋째, 이러한 권고안, 모범사례 또는 기타 비공식 지침을 의료 서비스 제공자를 포함한 적절한 이해관계자에게 배포.

(b)항의 (v)목에서는 신약 개발에서 인공지능 사용에 관한 규제 전략 내용을 담고 있다. 이 전략에 들어갈 내용도 다음과 같이 제시하고 있다. 첫째, 신약 개발의 각 단계에서 적절한 규제에 필요한 목적, 목표, 상위 원칙을 정의함. 둘째, 이러한 규제안을 구현하기 위한 향후 규칙 제정, 지침 또는 추가적인 법적 권한이 필요할 수 있는 영역 확인, 셋째, 이러한 규제안에 필요한 기존 예산, 자원, 인력 및 새로운 공공/민간 파트너십 파악 등이다. 지금까지 살펴본 행정명령 14110에서 제시된 보건복지 분야에 관련된 내용을 종합하면 다음 〈표 4-9〉와 같다.

참고로, 장애인에 관한 내용도 포함되어 있다. 행정명령은 "장애인이 시선 방향, 시선 추적, 보행 분석, 손동작과 같은 생체 인식 데이터 사용으로 인한 불평등한 대우 등 위험으로부터 보호받는 동시에 AI의 잠재력을 활용할 수 있도록, 건축 및 교통 장벽 준수 위원회(Transportation Barriers Compliance Board)는 적절하다고 판단되는 경우에 대중의

26) 여기서 중앙 추적 저장소(central tracking repository)에 대한 부가적인 설명이 없어서 어떠한 성격일지는 모호하다. 맥락으로 보아서는 의료 현장에서 인공지능 기술이 적용되는 과정에서 발생한 문제나 피해 내용에 관한 자료를 기록하고 누적 및 활용하는 데이터 저장소를 의미하는 것으로 추정된다.

참여를 요청하고 지역사회 참여를 유도하며 생체 인식 데이터를 입력하여 사용할 때 AI의 위험과 이점에 대한 기술 지원 및 권고를 발행하고 장애인이 정보통신 기술 및 교통 서비스에 접근할 수 있도록" 권장한다(US Exec. Order No. 14110, 2023, pp. 75214).

남궁준(2024)은 미국 행정명령의 법적 위상에 대해서 다음과 같이 설명했다. 다소 길지만 인용하면 다음과 같다. "연방의회가 새로운 법안을 통해 AI 사용 규제를 위한 새로운 기관을 설립하거나 기존 규제 당국에 새로운 권한을 부여하지 않고 있다. 따라서 연방 행정기관들은 AI 관련 규제가 필요할 경우 그들이 이미 가지고 있는 기존 권한, 즉 관련 법령3(의 재해석)에 근거해 AI 규제를 가능하게 하는 방향으로의 관할 사무 집행을 할 수 있을 뿐이다. 그런데 대통령이 발령한 동 행정명령은 (향후 그러한 방향으로의 사무 집행이 법원의 심사를 통해 취소되는 것은 별론으로 하더라도) 현재 발효 중인 법령을 보다 넓게 (재)해석하고 부여된 권한을 보다 적극적으로 행사할 것을 관련 행정기관 모두에 주문함으로써, 포괄적·입법적 입법을 통한 AI 규제와 어느 정도 유사한 효과를 가져올 수 있을 것으로 기대"(p. 188)하고 있다.

〈표 4-9〉 행정명령 14110에서 제시된 보건복지 분야 관련 내용

구분		미국 행정명령 14110
7조. 형평성과 권리 증진	2항. 정부의 급여 및 프로그램과 관련한 시민권의 보호	- 기관들은 연방정부의 프로그램과 급여를 집행하는 과정에서 인공지능 기술의 활용으로 불법적인 차별 및 기타 피해를 예방 및 해결해야 함 - 정부 급여의 공평성 제고를 위해 자동화 또는 알고리즘 시스템의 사용을 촉진하는 계획 작성. 다음의 절차를 개선하기 위함 ① 자격을 갖춘 수급자의 접근성에 대한 평가 ② 인공지능 시스템의 존재에 대한 수급자 대상 통지 ③ 급여의 부당한 거부를 감지하기 위한 정기적인 평가 ④ 전문 기관 직원이 적절한 수준의 재량권을 유지하는 절차 ⑤ 급여 거부에 대해 급여 신청자가 사람인 심사자에게 이의를 제기할 수 있도록 하는 절차 ⑥ 급여 프로그램에서 사용 중인 알고리즘 시스템이 공평하고 공정한 결과를 달성하는지 분석
8조. 소비자, 환자, 승객, 학생 보호	(b)호. “의료, 공중보건 및 복지[27] 부문에서 안전하고 책임감 있는 인공지능의 배포와 사용을 보장하기 위해” (p. 75214)	- 미국 보건복지부 안에 인공지능 태스크 포스 구성(90일 이내). TF 구성 이후 365일 이내 전략 계획(strategic plan) 개발. 다음 영역에서 지침과 자원 모색 (i목) ① 의료 서비스 제공 및 재정 분야에서 예측 및 생성 인공지능(predictive and generative AI) 기반 기술의 개발, 유지, 사용 ② 인공지능 기반 기술의 장기적인 안전성 및 실제 성과 모니터링 ③ 인공지능 기반 기술에 형평성 원칙을 반영 ④ 안전, 개인정보 보호 및 보안 표준을 반영 ⑤ 안전한 AI 사용을 결정하는 데 도움이 되는 문서의 개발, 유지관리 및 가용성 확보 ⑥ 긍정적인 현장 사용 사례와 모범사례 촉진 ⑦ 업무 효율성과 만족도를 증진하기 위한 AI 사용 사례 파악
		- 인공지능 기반 기술의 품질 평가 판단을 위한 전략 개발(180일 이내). 전략은 위의 i목의 내용 포괄. (ii목)
		- 정부의 재정 지원을 받는 보건복지 서비스 제공 업체에 대한 규제 고려(180일 이내) (iii목)
		- 의료 영역에서 인공지능에 관한 환자 안전 프로그램 수립(365일 이내) (iv목)
		- 신약 개발에서 인공지능 사용에 대한 규제 전략 제시(365일 이내) (v목)

출처: US Executive Order. (2023). “Safe, Secure, and Trustworthy Development and Use of Arti- ficial Intelligence”. 에서 필자가 내용 정리함.

27) 원문에서는 이 대목에서 ‘healthcare, public health and human services’라고 제시됐다. 미국 보건복지부에서 ‘human services’ 부분은 전통적인 아동복지, 노인복지, 빈곤 정책 등을 담당하는 점을 고려해서 ‘복지’라고 번역했다.

미국의 행정명령은 유럽연합의 인공지능법과의 차이를 살펴봄으로써 그 특징을 파악할 수 있다(Odelberg, 2024). 미국의 행정명령은 적용 범위와 접근 방식에서 유럽연합의 인공지능법과 크게 다르다. 미국의 행정명령은 새로운 기관을 설립하지도 않고, 민간 기업에 대한 새로운 규제를 가하지도 않는다. 행정명령의 규율 대상은 연방정부와 관련 기관들이다. 따라서 이들이 자체 인공지능 시스템을 구매하거나 개발할 때 행정명령이 작동한다. 이를 통해 공공기관들이 책임 있는 AI 사용의 모범으로 작용하도록 유도한다. 또 공공기관들이 구매할 수 있는 인공지능 시스템에 대한 규제안을 제공하는 방식으로 간접적으로 민간 부문에도 영향을 미친다. 이러한 점에서 민간 부문을 직접적으로 규제하는 유럽연합의 인공지능법과 차이가 크다.

미국 행정명령은 또 사기, 차별, 금융 리스크, 그리고 AI가 잠재적으로 미칠 수 있는 프라이버시 문제로부터 소비자를 보호하는 것을 포함해, 시민의 권리와 자유를 보호하기 위한 규정을 만들도록 연방 기관에 요구하고 있다. 행정명령은 최고 AI 책임자(Chief AI Officer, 이하 CAIO)라는 직위를 새로 만들고, 모든 연방기관은 최고 AI 책임자를 60일 이내에 지정하도록 요구했다. 최고 AI 책임자는 개별 기관에서 AI 사용을 조정하고, AI 혁신을 촉진하며, 기관 내 AI 리스크를 관리하는 역할을 맡는다. 미국 국립표준기술연구소(National Institute of Standard and Technology)는 연방 AI 시스템을 평가하기 위한 테스트 요구 사항에 대한 기준을 개발하는 임무를 맡고 있다.

행정명령과 인공지능법의 또 다른 점은 적용 대상이다. 행정명령이 특정 수치 이상의 연산 작업(10^{26} 이상의 정수 또는 부동소수점 연산량)을 통해 훈련된 대규모 인공 모델에만 적용되는 반면, 인공지능법은 위험 범주에 따라 모든 인공지능 모델에 적용된다. 행정명령의 기준은 현행 모델

에는 해당하지 않는다(Heath, 2023). 현재 이 기준을 충족하는 모델이 없기 때문이다. 차세대 기술에 집중하기 위해 설정된 기준이다. 그 밖에도, 인공지능법에서 인공지능 시스템이 미치는 환경적 영향을 언급한 점과, 행정명령에서 숙련된 인공지능 인재의 이민을 늘리려는 의도를 포함하고 있다는 점이 있다. EO와 AI법에서 제시된 AI 위험을 비교한 〈표 4-10〉이 아래에 제시되어 있다.

〈표 4-10〉 미국 행정명령 14110과 유럽연합 인공지능법의 대조표

구분	미국 행정명령 14110	유럽연합 인공지능법
규제 대상 인공지능 모델 기준	모델 규모 및 국가 안보 위험	위험 수준에 따라 규제 수준 결정
민간 영역 규제 여부	규제 안 함	규제
법 집행을 위한 안면인식 기술 허용 여부	금지	금지
인공지능이 환경에 미칠 영향 고려	고려 없음	고려
인공지능 관련 전문인력 입국 장려	내용 있음	내용 없음
투명성에 대한 요구	있음	있음
안전과 BIAS 테스팅	수행함	수행함

출처: Odelberg. (2024). "Understanding the Future of Artificial Intelligence Governance: Comparing the EU AI Act and U.S. Executive Order on Safe AI". p. 4의 그림을 다시 그림.

행정명령과 인공지능법의 공통점도 있다. 이 두 법안은 혁신을 촉진하면서 시민과 인권을 보호하려는 목표를 가지고 있다는 점에서 유사하다. 알고리즘의 투명성 강화, 인간의 감독, AI 편향 완화, 배포 전에 외부 스트레스 테스트나 '레드팀 테스트'를 광범위하게 수행한다는 원칙이 EO와 AI 법안 모두에 명시되어 있다. EO와 AI 법안은 국가 안보와 관련된 고위험 AI 사용에 대해서는 예외를 두고 있는데, 이는 감시 단체들로부터 비판을 받고 있는 타협점이다.

미국의 보건복지부는 행정명령 발표 이전인 2021년에 최고 AI 책임자(CAIO)를 지명했다(Alder, 2024.5.29.). 2024년 12월 기준 보건복지부의 CAIO는 Micky Tripathi다. 그는 보건복지부의 기술정책차관보(Assistant Secretary for Technology Policy)이자, 국가 건강 정보 기술 조정관(National Coordinator for Health Information Technology) 자리를 겸직하고 있다. Tripathi 차관보는 행정명령이 정하는 보건복지부의 인공지능 대응 태스크 포스도 이끌고 있다.[28] 보건복지부는 행정명령 8조 (b) (ii)에서 규정하는 부처 차원의 인공지능 대응 전략 계획(strategic plan)을 2025년 1월에 제시하겠다고 밝혔다(Burt, 2024.10.17.). Tripathi가 지난 2023년 미국 하원의 에너지 및 상업위원회(House Committee on Energy and Commerce)에 출석한 뒤, 2024년 안에 전략 계획을 수립한 점을 미루어 보면(US Departmenrt of Health and Human Service, 2023), 미국 보건복지부도 백악관이 내놓은 일정을 맞추기는 벅찬 것으로 보인다. 미국 연방정부에서 인공지능을 활용하는 사례 가운데 3분의 1 이상이 보건복지부 업무 영역에 속한 점을 고려하면(Burt, 2024.10.17.), 미국 보건복지부의 이후 행보는 미국 차원에서 인공지능 활용 및 관련 규제에 상당한 영향을 미칠 것으로 보인다.

참고로, Tripathi CAIO는 하원에 출석해서 보건복지부의 인공지능 관련 규제에서 다섯 가지 정책 선순위를 다음과 같이 제시했다(US Departmenrt of Health and Human Service, 2023).

첫째, 보건복지 영역에서 인공지능의 안전하고 책임 있는 채택 및 사용을 가능하게 하고, AI의 위험을 관리할 수 있는 자원과 정책을 개발한다.

---

28) 미국 보건복지부의 조직도(www.hhs.gov/about/agencies/orgchart/index.html)를 보면, Tripathi 차관보가 이끄는 '기술정책차관보실 및 국가 건강 정보 기술 조정관 사무소'는 제시됐지만, CAIO 직위나 인공지능 태스크 포스 조직은 제시되지 않고 있다. 조직도만 보면, 인공지능 태스크 포스 조직이 새로 신설됐다고 보기는 어려워 보이며, 기존 기술 정책 차관보실에서 해당 업무를 떠안은 것으로 추정된다.

둘째, 보증 기준과 품질 관리 프로세스를 통해 의료 분야에서 인공분야의 질과 안전성을 개선한다.

셋째, 보조금 지원과 계약을 활용하여, 보건복지 전달체계의 가치 사슬 전반에 걸쳐 인공지능의 개발과 책임 있는 사용을 촉진한다. 여기서 보조금과 계약은 보건복지부가 민간 보건복지 영역을 견인하기 위한 수단으로 활용될 수 있다는 의미로 풀이된다.

넷째, 개인부터 조직 및 주 단위에 이르기까지 보건복지 서비스 제공에서의 인공지능 개발과 사용에 대해, 건강 관리 생태계와 구성원들에게 공공 교육을 제공한다.

다섯째, 프로세스 혁신과 현대화를 추진하기 위해 보건복지부 전반에 걸쳐 AI 기능을 평가하고 배치한다.

미국 보건복지부의 인공지능 관련 내용을 보면, 보건복지 전체 영역을 아우르면서도, 전체적으로 의료나 보건 영역에 내용이 강조되는 점이 관찰된다. 의료나 보건 영역이 인간의 생명과 건강에 직결되는 민감성을 가졌기 때문인 것으로 추정된다. 더불어, 부처의 CAIO인 Micky Tripathi가 건강 관련 정보통신 업계에서 20년 동안 몸담았던 이력과도 일부 연관됐을 가능성도 있다.

참고로, 미국 연방정부 차원에서 입법 조치가 없는 상황에서, 주 의회들은 각자 나서서 인공지능 관련 법안을 만들고, 통과하고, 시행하고 있다(Curry, 2024, 10. 22.). 연방의회가 인공지능을 둘러싼 주요 기술 문제에 대한 법안을 통과시키는 데 어려움을 겪는 동안, 개별 주의 정책입안자들이 조금은 더 가볍게 문제에 접근하는 것으로 보인다. Curry(2024. 10. 22.)는 인공지능의 급격한 발전에 따라, 주 단위에서도 무더기의 법안이 제정되는 현상을 관찰했다. 강력한 인공지능(AI) 모델이 출시된 후 2024년에만 700개에 가까운 AI 관련 법안이 발의되는 등 AI

관련 법안이 쏟아져 나왔다. 2023년에 관련 법안이 200개 미만이었던 점을 고려할 필요가 있다.

그나마 2024년에 의미 있는 진전은 있었다. 이를테면, 콜로라도주는 미국 최초로 AI의 고위험 사용을 다루는 포괄적인 법안을 통과시켰다. 캘리포니아와 테네시 같은 다른 주에서는 데이터 출처 및 디지털 복제본 같은 특정 문제를 해결하기 위한 조치가 이뤄졌다. 33개 주에서는 인공지능 규제 관련 태스크 포스를 구성하거나 기존 위원회에 AI가 다양한 정책 영역에 미치는 영향을 연구하도록 지시했다. Curry(2024. 10. 22.)는 2024년 들어 AI 관련 입법 활동이 급격히 증가했음에도 불구하고 정책입안자들은 아직 구체적인 규제 모델에 대한 합의를 이루지 못하고 있다고 설명했다. 이원태(2024.10.24.)는 미국의 이런 상황에 대해서, "정책입안자들은 아직도 특정한 규제 모델에 대한 합의에는 이르지 못했다. 왜? AI 모델과 AI 시스템의 특성을 잘 모르기 때문인가, 아니면 각각의 이해관계 문제인가?"라고 질문했다.

더불어, 미국 캘리포니아주의 규제 법안 시도도 주목할 만한 가치가 있다. 미국 캘리포니아주 의회가 인공지능 규제 법안(SB 1047) 통과를 준비했으나 무산됐기 때문이다. 법안의 정식 명칭은 '프론티어 인공지능 모델을 위한 안전한 혁신 법안(Safe and Secure Innovation for Frontier Artificial Intelligence Models Act)'이다. 법안의 핵심 내용은 다음의 두 가지로 요약된다(조이환, 2024.8.8.). 첫째, 5억 달러가 넘는 피해 발생 가능성이 있는 사이버 보안 공격을 시행하거나 생물학·핵 무기를 개발할 수 있는 잠재력이 있는 인공지능 모델의 개발을 금지하는 것이다. 둘째, 인공지능 개발자들이 주기적으로 안전 테스트 결과를 보고해야 하고, 통제가 어려울 때 인공지능의 작동을 중단하는 '킬 스위치'를 도입해야 한다.

캘리포니아주의 Scott Wiener 상원의원이 발의한 법안은 미국 사회에서 격렬한 논쟁을 초래했다(Lee, 2024.8.16.). 메타 등 거대 인공지능 업체들은 법안에 반대했다. 메타의 주 정책 매니저인 Kevin McKinley는 "(위너 의원이) 법안을 설명하는 방식과 목표에 동의하지만, 이 법안이 특히 캘리포니아의 AI 혁신, 특히 오픈 소스 혁신에 미칠 영향에 대해서는 여전히 우려하고 있다"고 말했다.

미국의 인공지능 업체인 Meta는 개발자가 자체 제품을 위해 이를 기반으로 구축할 수 있는 오픈 소스 AI 모델인 Llama를 보유한 회사 가운데 하나다(Lee, 2024.8.16.). 메타는 4월에 Llama 3을 출시했으며, 2천만 다운로드를 기록했다. 물론, 업계만 반대하는 것은 아니다. 캘리포니아주 하원의원 8명은 목요일 뉴섬 주지사에 서한을 보내 법안이 의회를 통과하면 거부권을 행사할 것을 권유했다. 반면, "AI의 대부"로 불리는 Geoffrey Hinton을 비롯한 일부 기술 업계 종사자들은 이 법안을 지지하고 있다.

법안은 업계, 학계, 정치계의 줄다리기 중에 수정이 계속됐다(Lee, 2024.8.16.). 2024년 8월에도 법안에서 위증에 대한 처벌 조항을 삭제하고 개발자의 첨단 AI 모델의 안전성에 관한 법적 기준을 변경했다. 프론티어 모델 부서라고 불렸던 새로운 정부 기관을 만들려는 계획도 무산됐다. 원안대로라면 개발자는 새로 신설되는 기관에 안전조치를 제출해야 했다. 새 버전에서는 개발자가 이러한 안전조치를 법무부 장관에게 제출하는 것으로 바뀌었다.

법안이 수정을 거치면서 기능을 점차 상실하고 있다는 비판도 높아지고 있다. 특히, 법안이 기업들의 로비를 거치면서 인공지능의 부작용에 대한 선제 대응에서 사후적 대응으로 무게 중심이 이동됐다고 비판하고 나섰다. 영국 옥스퍼드대학교 철학과 교수인 Carissa Véliz(2024)는

자신의 Linkedin에 올린 글에서 법안이 위축되는 과정을 다음과 같이 요약했다. “법안은 더 이상 치명적인 사고가 발생하기 전 부주의한 안전 관행에 대해 기업을 고소하는 것을 허용하지 않게 됐다. 또, 규정 준수를 감시하는 새로운 주 정부 기관을 만들지 않게 됐다. AI 연구소가 안전 테스트를 인증하는 과정에서 증언에 대해서 위증죄 처벌을 부과하지 않게 됐다. 또 개발자에게 모델이 해롭지 않다는 “합리적 확신”을 요구하지 않게 됐다.” 뉴욕대 심리학과 교수인 Gary Marcus(2024.8.21.)도 개인 뉴스 레터에서 “새로운 형태의 SB 1047은 기본적으로 정말 나쁜 일이 발생한 후에야 기업에 책임을 묻는 도구로만 사용할 수 있다. 더 이상 큰 피해로 이어질 수 있는 명백한 과실로부터 우리를 보호할 수 없다”라고 평가했다.

전 세계적인 주목을 받던 SB 1047은 개빈 뉴섬 캘리포니아 주지사가 거부권을 행사하면서 입법이 무산됐다(Lee, 2024.9.29.). 미국 사회는 인공지능의 등장 앞에서 여론이 극단적으로 갈리고 있다. 현재로서는 연방이나 주 단위에서 형성되는 규제의 흐름이 사회보장 영역에 미칠 영향을 단정하기는 힘들다. 이와 관련해서 미국 사회에서 논의되는 내용도 거의 드러나지 않는다. 지금의 구도는 기술 낙관론자들과 Meta 같은 거대 자본이 한편에, 그리고 반대편에 기술 신중론자 및 진보 성향의 일부 정치권이 힘 겨루기를 하는 상황이다.

제5장

결론

제 5 장 결론

제1절 요약 및 논점

이 보고서의 제2장에서는 인공지능의 발전 동향을 짚고, 인공지능을 활용할 때의 윤리적인 원칙을 제시했다. 첫째, 공공성, 둘째, 공정성, 셋째, 책임성과 책무성, 넷째, 안전성과 보안성이었다. 더불어, 인공지능의 차별화한 특성에 근거한 추가적인 윤리 원칙으로 첫째, 통제 가능성, 둘째, 투명성과 설명 가능성, 셋째, 개인정보 및 사생활 보호를 제시했다. 제2장에서 추가로 다루지는 않았지만, 이와 같은 원칙들은 기술 적용 현장에서 서로 충돌되고 모순될 가능성이 높다. 사회적 공론화 과정을 통해서 원칙들의 선순위를 정하고', 원칙들을 투명하고 공정하게 적용할 수 있는 거버넌스를 구축하는 것이 필요할 것이다. 이 부분은 제2절에서 더 논의하도록 하겠다.

제3장에서는 국·내외에서 사회보장 영역 및 제도에 인공지능이 적용된 현황을 살펴보았다. 국내의 경우, 제1절에서 살펴보았듯이, 조달정보개방포털과 나라장터에서 연구 시점 기준으로 최근 5개년 동안(2019년 10월~2024년 9월) 확인되는 사회보장 분야의 복지 기술 활용 현황을 조사해서 36건의 사업을 확인했다. 이들 사업의 특징을 종합하면 다음과 같았다. 첫째, 사회보장 분야의 복지 기술은 노인에 대한 돌봄 수요에 다수 활용됐다. 둘째, 복지 기술은 업무 담당자의 효율적인 업무 처리를 지원하기 위해 활용됐다. 셋째, 빅데이터의 수집과 축적을 기반으로 하는 정보제공 서비스가 다수 발견됐다. 넷째, 개인에게 특화된 맞춤형 서비스가 다수 차지했다. 마지막으로 정보 활용의 효율성과 효과성을 높이기 위한 데이터베이스 시스템의 활용 양상이 확인됐다. 종합해서 보면, 국내의

복지 기술은 고차원적인 인공지능 기술이라기보다는 낮은 수준의 인공기술 활용 수준에 머물고 있음을 확인할 수 있었다.

제3장 제2절에서 확인한 해외의 사례도 마찬가지다. 챗봇 등 상대적으로 초보적인 수준의 인공지능 기술이 활용되고 있다. 참고로, 해외 사회보장 영역에서 나타나는 양상은 국내와 비교했을 때 두 가지 특징이 있다. 첫째, 챗봇에 대한 활용도가 상대적으로 높았다. 특히, 브라질, 아르헨티나, 파나마 등 남미 국가에서 활용도가 높다는 점이 특징적이다. 둘째, 사회보장 행정에 인공지능을 적용하는 과정에서 적지 않는 사회적 논란이 발생하고 있다. 특히, 네덜란드와 덴마크 등에서는 개인정보 유출, 알고리즘의 차별적 관리, 국가 기관의 과도한 개입 등을 둘러싼 논란이 지속되고 있다. 한국에서 개인정보를 활용한 복지 행정이 대규모로 이뤄지고 있지만, 이와 같은 논란이 드물었던 점을 고려하면 주목할 대목이다. 바꾸어 말하면, 한국에서 빅데이터와 인공지능을 활용한 행정 과정에서 서구와 같은 논란이 앞으로 불 붙을 가능성도 배제할 수 없다. 불필요한 사회적 갈등을 막기 위해서는 관련 제도와 규제를 점검할 필요가 있다.

국내·외의 인공지능 활용 양태를 보면, 사회보장 영역에서 인공지능이 활용되는 영역을 확인할 수 있었다. 과거 복지국가에서 디지털 기술이 활용되는 영역을 제시했던 Alston(2019), 김기태(2024), Zaber, Casu, Brodersohn(2024)의 논의를 참고하고, 제3장의 내용을 종합하면 아래 열 가지 영역이 제시될 수 있을 것이다.[29]

29) OECD(2024b)도 'Modernizing Access to Social Protection' 보고서에서 인공지능의 활용 영역을 다음과 같은 네 가지로 제시했다. ① 공무원의 문서 작업 자동화 및 효율화, ② 개인이나 지역사회의 위험 예측 혹은 예방적 접근, ③ 급여 자격 기준 등에서 결정을 자동화, ④ 챗봇, 오피스 프로세스 자동화, 부정수급 포착 영역이다. OECD(2024b)가 전 세계적으로 사회보장 영역에서 인공지능은 "간헐적(infrequently)"(p. 9)으로 쓰인다고 논평한 점을 보면, OECD는 인공지능이 활용되는 영역을 지나치게 협소하게 파악한 것으로 보인다.

첫째, 본인 인증(identity verification)이다. 본인 인증은 급여 신청, 자격 심사, 급여 지급 과정에서 반드시 필요하다. 물론, 한국에서는 지문 및 얼굴 정보가 포함된 주민등록 데이터베이스가 있어 매우 높은 수준의 본인 인증 체계가 성립된 점을 확인할 수 있다. 한국에서는 인공지능 기술까지 필요하지 않을 수도 있다.

둘째, 자격 심사(eligibility assessment)다. 캐나다 온타리오주에서는 2014년부터 사회부조 운영시스템을 통해서 급여 자격을 심사하고 있다. 빅데이터 및 인공지능을 활용하여 급여 자격 심사를 빠르게, 정확하게 처리할 여지가 생긴다.

셋째, 복지 급여액 산정 및 지급(welfare benefit calculation and payments)이다. 다수의 국가에서 점점 더 많은 복지 급여액이 사람의 개입 없이 자동적으로 산정되고 지급되고 있다. 영국은 실시간 소득정보시스템(Real Time Information System)을 활용해서 복지 급여를 지급하고 있다.

넷째, 부정·오류 수급 예방 및 탐색(fraud prevention and detection)이다. 많은 복지국가에서 디지털 자료를 활용하는 주요한 이유 가운데 하나가 부정·오류를 예방 및 탐색하는 것이다. 네덜란드에서 논란을 낳았던 SyRi(Systeem Risico Indicatie)가 여기에 해당한다.

다섯째, 위험의 점수화 및 범주화(risk scording and classification)다. 제3장 제1절에서 살펴본 한국의 사각지대 발굴관리시스템이 여기에 해당한다. 개인에 관한 공공데이터 자료의 수집과 빅데이터 분석, 분석 결과를 바탕으로 한 고위험 대상자 도출 과정을 거치기 때문이다. 이러한 접근은 유럽연합이 인공지능법에서 '수용할 수 없는 위험성(unacceptable risk)'으로 분류한 '사회적 평점(social scoring)'과 흡사하다. 앞으로도 논란을 낳을 수 있는 대목이다.[30]

---

30) Alston(2019)은 위험의 점수화 및 범주화와 관련해서 세 가지 위험을 예시했다. 첫째,

여섯째, 개인 맞춤형 정보 서비스다(personalized information service)이다. 제3장 제2절에서 살펴본 챗봇이 대표적 사례가 될 것이다. 다양한 영역에서 개인에게 최적화된 서비스를 제공하기 위해 빅데이터 분석 기반 기술을 활용하여 서비스 이용자에게 맞춤형 서비스를 제공할 수 있다. 제3장 제1절에서 살펴본 AI 활용 초기상담시스템도 여기에 해당될 것이다. 한국의 AI 활용 초기상담 정보시스템도 콜 기반 대화 시스템을 활용하여 잠재적 위기 대상자와 초기상담을 진행한다. 이러한 기능은 복지수급자의 사례 관리에까지 확장될 수 있다. 사회보장 기관이 인공지능을 사용하여 사례 처리 방식을 자동화하고 개인의 서비스 후속 조치를 위한 고충 대응 지원을 제공할 수 있다. 오스트리아의 사회보험연합은 청구 자동 처리를 지원하고 의사와 환자를 매칭하는 인공지능 기반 시스템을 구현했다.

일곱째, 온라인을 활용한 소통을 넘어서 실제 돌봄 영역에서도 활용되고 있다. 제3장 제1절에서 보았듯이, 이는 대부분 노인을 위한 것으로 AI·IoT를 활용한 건강 서비스, 생체건강 셀프체크 서비스, 안전감지 센서를 활용한 안전 확인 서비스, 실종 방지를 위한 위치 기반 모니터링 서비스 등이다. 이뿐만 아니라 여가 활용에 도움이 되는 다양한 정보 제공 서비스 등이 제공되고 있다.

여덟째, 사회보장 행정 기관 내부적인 용도로 업무 담당자의 효율적인 업무 처리를 돕고, 내부 교육 등의 용도로도 활용될 수 있다. 실제로, 제3장 제1절에서 살펴보았듯이, 다수의 복지 기술은 업무 담당자의 효율적인 업무 처리를 지원하기 위해 활용되고 있었다. 이를테면, 빅데이터를 기반

전체 인구집단의 데이터를 근거로 한 예측 모델에 따라 개인의 위험 수준을 파악하는 과정에서 나타날 수 있는 오류의 가능성, 둘째, 점수화 및 범주화의 근거가 되는 기술이 공개되지 않으면서 나타날 수 있는 권리 침해의 가능성, 셋째, 점수화 및 범주화가 현재의 불평등과 차별을 강화하거나 유지할 가능성이다.

으로 의사결정을 효율적으로 지원하는 서비스도 확인되었는데, 이는 수급자의 상담을 위한 기초 자료를 제공한다.

아홉째, 제3장 제1절에서 확인했듯이, 정보 활용의 효율성과 효과성을 높이기 위한 DBMS(Database Management System)의 활용이다. 데이터베이스의 관리는 업무의 효율화뿐만 아니라 빅데이터 분석 환경을 마련하는 데까지 연결되며, 새로운 사업을 기획하고 사회적 가치를 창출하는 과정에서 유용하게 활용될 수 있다.

열 번째, 사회정책의 효율성과 효과성을 평가하는 데 활용될 수 있다. 데이터의 가용성이 실시간화한다면, "적절한 성과지표와 주기적인 활용은 사업의 효과성과 효율성을 평가하는 데 필수적인 것으로 증거 기반 정책(evidence-based policy)의 기초"(유종성, 2023, p. 8)가 될 수 있다. 또한, 평가의 과정에서 정책의 성과, 실패, 한계를 낳은 원인을 분석해서, 정책을 조정, 갱신, 폐기하는 근거가 될 수 있다.

장기적으로 보면, 전 국민 대상 실시간 데이터에 근거한 인공지능의 활용은 사회보장제도 전체를 재편하는 방향으로 나갈 잠재력을 가지고 있다. 노대명(2024)은 소득보장제도 재구조화를 준비하자고 제안하면서 "소득 기반 사회보험의 실험이 중단됐지만 계속 추진할 필요"(p. 14)가 있으며 "현행 85개 제도를 4~5개 정도로 단순화하는 재정 기반 소득보장 제도를 준비"(p. 14)해야 한다고 제안했다. 시민들의 다양한 욕구에 부응하는 방식으로 자리 잡은 제도들이 제도의 운영 주체 및 전달체계에 따라 너무 복잡하게 얽혀 있는 점을 고려할 필요가 있다. 이에 따라, 사회보장제도들의 재구조화에 대한 요구가 끊임없이 있었던 점을 고려하면(강혜규, 강지원, 강희정, 김기태, 김세진, 김태완.. 주하나, 2022), 변화의 단초가 빅데이터와 인공지능 활용 과정에서 마련될 여지가 있다. 노대명(2024)은 현재의 급여 지급 정보관리 체계에서 복지 제공 부처별로 급여

자격 조건 심사를 따로 추진하면서 급여 지급이 지연되고 있다며, 디지털 사회보장 허브를 구축하여 급여 자격 심사와 지급을 효율화하자고 제안했다.

사회보장 영역에서 점점 폭넓게 활용되는 인공지능 기술이 불러올 효과는 양면적이다. 제1~4장의 논의를 종합해서 긍정적인 측면과 부정적인 측면을 나누어 살펴보겠다.

먼저, 인공지능이 공공의 민주적인 통제 아래 작동할 경우를 전제로 할 때, 기대되는 순기능은 다음과 같다.

첫째, 효율성이다. 앞에서 살펴보았듯이, 본인 인증, 자격 심사, 복지 급여액 산정 및 지급 등의 일선 서류 행정이 더욱 자동화하게 된다. 이럴 경우, 더 적은 인력으로 더 많은 행정 업무가 가능해진다. 일선 공무원은 민원인을 대상으로 하는 대면 서비스나 사례 관리에 집중할 수 있다. 물론, 인공지능의 적용에 따른 부가적인 업무도 생길 수 있다. 이를테면, 미국의 행정명령 7조 2항 (b)호에 따르면, 급여 거부에 대해 급여 신청자가 사람인 심사자에게 이의를 제기할 수 있도록 하는 절차를 규정하고 있다.

둘째, 적시성이다. 빅데이터를 활용해서 관리되는 인공지능 기술은 급여 신청과 심사, 지급에 이르는 과정을 단순화할 수 있다. 영국의 경우, 통합급여를 도입하는 과정에서 실시간 소득 파악 시스템을 활용해서, 급여 지급 절차를 간소화하고자 했다.

셋째, 정확성이다. 빅데이터에 근거한 인공지능의 판단, 범주화, 예측은 인간의 오류와 편견이 개입할 가능성을 줄일 수 있다. 영국의 경우, 연금을 제외한 복지 급여액 가운데 오류 또는 부정으로 인해 초과 지급된 액수가 2021~2022년 회계연도 기준으로 85억 파운드(약 12조 7천억 원)로 추정됐다(Davies, 2022). 이는 같은 기간 연금을 제외한 전체 복지 급여 지급액의 7.6%를 차지하는 액수다. 특히, 서구의 복지국가에서 디

지털 기술을 통해 급여의 오류 및 부정수급을 시정하고자 하는 의지가 강해 보인다.

넷째, 개인 맞춤형 급여 제공이다. 인공지능을 활용한 행정 집행은 개인당 위험의 점수화 및 범주화까지 맞춤형 수준으로 정교화할 수 있다. 개인의 여건과 수요에 맞춘 현금 및 서비스, 일자리가 개인 맞춤형으로 제공될 여지가 커진다. 제3장 제1절에서 살펴본 AI 기반 일자리 매칭 서비스가 대표적인 예다. AI 기반 일자리 매칭 서비스는 구직자의 이력서 정보와 구인 기업의 구인 정보를 활용하여 인공지능 알고리즘에 기반한 추천 서비스를 제공하고 기업과 구직자 간 일자리 미스매치를 해소한다. 이를 통해서 제도에 대한 개인의 체감도가 향상될 수 있다.

다섯째, 범용성이다. 챗봇을 통한 상담은 국민들의 입장에서는 시공간의 제약을 뛰어넘어서 제도에 접근할 수 있는 기반을 제시해준다. 시민들은 주민센터를 방문하지 않고도 언제, 어디서나 급여와 관련한 상담과 더불어 자격을 확인할 수 있고, 급여를 신청할 수 있게 된다. 이를 통해서, 지역별 접근성의 격차를 해소할 가능성도 커진다.

여섯째, 정책 평가의 용이성이다. 빅데이터를 활용한 급여 집행 내용은 정책의 효과를 평가하는 데 용이한 기반을 제공한다. 평가를 위한 데이터의 구축이 용이해지면서 근거 기반 정책 평가, 형성 및 집행의 토대가 마련된다.

일곱째, 사각지대 해소다. 제3장 제2절에서 살펴본 대로, 복지 사각지대 발굴관리시스템이나 AI 활용 초기상담시스템은 다중 위험의 시대에 잠재적 수급 대상자들의 욕구를 빠르게 파악하는 토대가 된다. 한국 복지국가에서 고질적으로 지적되는 사각지대의 문제를 빠르게 해소할 수 있는 제도적 기반이 될 수 있다. 물론, 사각지대 문제의 핵심은 위기가구의 발굴에 있다기보다는, 발굴된 위기가구를 지원할 급여가 없거나 부족

하다는 지적(함영진, 이현주, 어유경, 김가희, 박성준, 조용찬.. 오민수, 2023)도 함께 고려할 필요가 있다.

빅데이터 활용을 수반하는 인공지능 기술이 사회보장 영역에서 초래할 위험성도 함께 보겠다. 위험의 내용은 김수영(2016)과 김기태(2024)의 내용, 그리고 이 보고서 제2~4장의 내용을 종합해서 제시하도록 하겠다.

첫째, 프라이버시의 문제다. 한국의 경우, 국민기초생활보장제도에서 수급자는 자신과 부양의무자의 소득 및 재산 정보를 국가에 제공하는 조건 아래 급여를 받을 수 있다. 복지 급여 수급자들의 소득, 재산, 가족 정보뿐 아니라 일부 사례 관리 정보까지 이미 방대하게 집적돼 있다. 공공기관을 넘나드는 개인정보가 유출 및 남용될 가능성도 완전히 배제할 수는 없다. 또한, 국가 권력에 의한 데이터 남용 가능성도 고려해야 한다(홍승헌, 황하, 2024).

둘째, 정확성의 문제다. 앞에서 인공지능이 불러올 장점으로 정확성을 제시한 점을 고려하면, 다소 모순적인 지적일 수 있다. 그럼에도, 국내의 사회보장 정보시스템에서 개인의 사망 및 출생 신고가 반영되지 않거나 과거 소득이 현재 소득과 합산돼서 제시되는 등의 문제가 끊임없이 발생하고 있다. 현장 공무원들은 복잡한 데이터를 처리하고 오류를 수정하느라 대면 접촉을 통해서 사각지대를 발굴할 기회를 오히려 놓치고 있다는 지적(임덕영, 2023)도 나오고 있다. 물론, 이러한 문제는 데이터 기반 인공지능 기술이 현장에 접목하는 과정에서 발생하는 일시적이고 과도기적인 문제라고 볼 수 있다. 그러나 인공지능 기술에 대한 맹신적인 태도 역시 경계할 필요는 있다. 이 대목은 사회보장 행정에서 관리하는 데이터와 알고리즘의 질 관리와도 직결되는 문제다.

셋째, 데이터 소유권의 문제다. 김수영(2016)은 급여 신청자는 소득, 재산, 가족 정보, 신체 등에 관한 자기 정보를 제공하는 데 동의하게 되는

데, 여기서 정보의 소유권에 관한 문제가 제기된다고 지적한다. 그는 "국가는 정보 제공자의 의도와는 상관없이 이들이 제공한 정보를 보유할 뿐만 아니라 향후 해당 정보의 활용, 분배, 처분에 관한 권한을 얻게 되는 상황"(p.211)이라고 했는데, 이 대목은 앞에서 논의한 프라이버시와 함께 앞으로 논점으로 부상할 여지가 크다.

넷째, 개인정보를 영리 목적으로 활용하는 것의 문제다. '디지털 헬스'를 둘러싼 데이터 활용은 한국에서 이미 오랜 의제다. 건강보험의 개인정보에 대한 접근권을 민간기업은 지속적으로 요구하고 있다. 보건복지부는 2024년 11월 '보건의료데이터 혁신포럼'을 전국 43개 의료데이터 중심 병원과 진행한 바 있다. 복지부는 "정부와 의료계, 학계, 산업계가 데이터 협력체계를 구축하여 미래 의료 혁신과 국민건강 증진으로 나아가는 계기가 되기를 희망"(보건복지부, 2024.11.26. p. 4)한다고 밝혔다. 여기에서 개인 건강 정보가 영리 목적으로 활용될 여지가 잠재적으로 남아 있다.

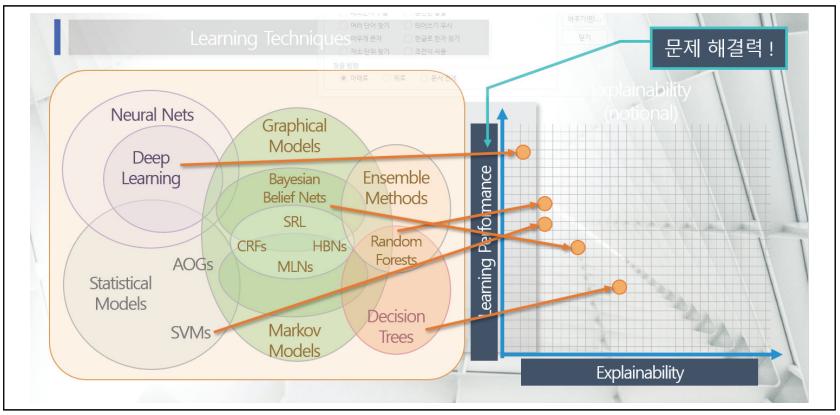
다섯째, 알고리즘의 결정에 근거한 개입의 문제다. 덴마크는 취약 아동을 포착하기 위해서 실업 및 의료를 비롯한 다양한 사회경제적 영역의 데이터를 결합하여 200개 이상의 위험 지표를 분석하는 머신러닝 모델을 구축했다(Jørgensen, 2021). 그리고 그 모델이 위험신호를 보내는 가구에 대해서 부모의 동의 없이 개입할 수 있도록 했다. 이후 해당 모델은 신뢰성에 의문이 제기되면서 2018년 12월에 중단된 바 있다(Algorithm Watch and Bertelsmann Stiftung, 2020). 한국에서도 2023년부터 44종의 위기 정보에 근거해서 위기가구에 대해서는 경찰·소방 협력을 통한 비상 개문 지침을 마련했다(정세정 외, 2023). 이 대목에서 빅데이터를 바탕으로 인공지능의 판단에 근거해서 국가가 어디까지 개인의 삶에 개입할 수 있는가, 라는 윤리적인 문제가 발생한다.

여섯째, 데이터와 알고리즘의 편향성의 문제다. 미국의 데이터 과학자인 Cathy O'Neil(2017)은 알고리즘이 현대 사회에서 대량살상무기(Weapons of Mass Destruction)에 준하는 위험성을 가졌다고 경고하기 위해 책 'Weapons of Math Destruction(대량살상 수학무기)'을 출판한 바 있다. 그는 여기에서 편향된 데이터와 알고리즘이 취약계층에게 차별적인 결과를 낳게 된다고 경고했다. 그는 알고리즘에 대한 기술 낙관론을 '테크노-유토피아'로 명명하면서, "알고리즘과 기술이 가져다줄 혜택에 대한 무제한적이고 부적절한 희망에서 깨어나야 한다"(p. 343)고 강조한다.

일곱째, 인공지능의 '설명 불가능성'의 문제다. 제2장에서도 살펴보았듯이, 인공지능의 복잡성을 고려할 때, 인공지능이 어떤 기준과 원칙에 따라 작동하는지에 대해서 인간은 이해할 수 있어야 한다. 이는 인공지능 관련 핵심 원칙 가운데 하나인 설명 가능성(explainability)이다. 문제는 인공지능이 점점 더 설명 가능하지 않다는 점이다. 인공지능이 고도화할수록 설명 가능성은 떨어진다(〔그림 5-1〕 참고). 물론, 제3장에서도 언급했듯이, 사회보장 분야에서는 인공지능이 활용된다고 하더라도 딥러닝 수준까지 활용되는 경우는 많지 않다. 따라서, 인공지능의 설명 가능성에 대한 문제가 적어도 한동안은 중요하지 않을 수도 있다. 그렇지만 기술의 발전 속도를 고려한다면 인공지능의 설명 불가능성이 조만간 난제로 등장할 가능성도 있다. 사회보장 행정에서는 특히 취약계층에 대한 차별적인 결정이 이뤄질 가능성도 염두에 둬야 한다.

지금까지 인공지능의 잠재적 위험성을 살펴봤다. 다만, 국외에서 제기되는 위험성을 고려할 때, 한국과 외국이 제도적 환경이 다르다는 점도 염두에 둘 필요가 있다. 정세정 외(2023)에서는 세 가지로 나누어서 설명하였다(pp. 213~214).

〔그림 5-1〕 인공지능 윤리 관련 의제와 원칙



출처: 김형주. (2024.7.23.) "인공지능 윤리 핵심 가치 분석". p. 5. 사회보장행정에서 인공지능 적용 동향과 함의, 세미나.

첫째, 한국에서는 지문 및 얼굴 정보가 포함된 전 국민 주민등록 데이터베이스 덕분에 개인인증 영역에서 압도적인 인프라를 구축했다. 전 국민의 일률적인 국민식별번호를 운영하는 국가는 한국이 거의 유일하다(성준호, 2016). 둘째, 해외에서는 개인정보의 영리적 집적 및 활용에 대한 우려가 크다(Eubanks, 2018; Alston, 2019). 미국에서는 민간 보험회사에서 개인 데이터를 활용하는 빈도가 더 높다. 반면, 한국은 공공이 전 국민의 소득, 재산, 건강, 인적 데이터를 주도해서 관리하고 있다. 셋째, 다른 복지국가에서는 복지제도의 성숙화 과정에서 디지털 기술을 급여의 부정 및 오류 수급 포착에 활용하는 경향이 있다. 서구가 보편적 복지국가의 성숙기를 지난 이후에 나타나는 문제에 대응하는 과정에 있다면, 한국은 보편적 복지국가와는 거리가 멀다. 한국에서는 디지털 기술 활용의 초점을 복지 사각지대 해소 혹은 위기가구 포착에 두고 있다. 사회보장 영역에서 인공지능의 활용과 규제를 모색할 때, 이러한 한국적 특수성도 함께 고려할 필요가 있다.

제2절 정책적 함의

앞의 제1절에서 우리는 인공지능 기술이 사회보장 영역에서 불러올 수 있는 편의와 리스크를 두루 확인했다. 이로써 앞으로의 정책 방향이 자연스럽게 도출된다. 리스크를 최소한으로 관리하면서 동시에 편의를 최대화하는 방향이다. 리스크의 가능성을 규제하고, 편의를 증진하는 방향으로 지원하는 것 사이에서 적절한 정책적 배합이 중요하다는 의미다. 여기에서 신기술에 대한 규제와 지원 사이를 길항과 모순 관계로 볼 필요는 없다. 오히려, 지원의 전제는 규제이고, 규제의 이유는 지원이다. 규제와 지원 사이의 동적인 관계에 대해서 김병권(2024.12.26.)의 다소 긴 언급을 들어보자.

(정부의 인공지능 지원 편향적 정책에 대해) "우리는 얼른 인공지능을 경쟁력 있게 개발해야 하니 당연한 거 아니냐는 분들도 많다. 그러나 누차 말하지만, 식품과 의약품은 정확한 안전성 규제를 마련해야 소비자들이 마음 놓고 구입하여 먹을 수 있고, 이럴 때 진짜 식품산업과 제약업이 발전한다. 지금처럼 우리가 항공기를 자유롭게 타는 것은 엄청나게 까다로운 항공기 운항 규제가 생겼기 때문이다. 그 이전에는 워낙 사고가 많아 비행기 이용은 모험가들에 국한되었다. 지금 지구상에 15억 대가 넘는 자동차가 굴러다니는 것은 오직 엄격한 '교통법규'와 '자동차 안전성 규제' 때문이다. 인공지능도 잘 사용하면 좋지만, 위험성도 함께 있기에 '안전성을 제대로, 엄격히 규제'해야 번성할 수 있다."

유럽연합과 미국 행정명령도 인공지능에 대한 규제와 지원 사이에서 일정한 균형을 갖췄다. 제4장에서는 규제에 초점을 맞춰서 논의했지만, 지원을 강조한 다음의 대목들도 있다. 유럽연합의 인공지능법은 "이 규정은 공공행정이 준수되고 안전한 AI 시스템의 광범위한 사용을 통해 이익

을 얻을 수 있도록 혁신적인 접근 방식의 개발과 사용을 저해해서는 안 된다"(Artificial Intelligence Act, (58))라고 명기하고 있다. 미국 행정명령에서도 7조 2항 (b)호에서, "보건복지부는 급여 및 서비스를 시행할 때 자동화 또는 알고리즘 시스템의 사용을 촉진하는 계획(plan)을 행정명령 시점 기준으로 180일 이내에 발표해야 한다"라고 규정했다.

안타깝게도 한국에서는 사회보장 영역에서 인공지능 적용에 관한 공적인 비전 혹은 계획이 제시된 바 없다. 물론, 2024년 12월 26일에야 한국에서 인공지능 기본법이 국회 본회의를 통과한 상황을 고려할 필요는 있다. 그러나 한국의 인공지능 기본법에서 사회보장 영역에서의 인공지능 기술 적용을 촉진하거나 규제하는 내용이 반영되지는 않은 것으로 보인다(고환경, 채성희, 손경민, 이일신, 2024).[31] 유럽연합의 인공지능법과 미국 행정명령에서 사회보장 영역을 비중 있게 다룬 점과 대조된다. 제3장에서 살펴보았듯이, 한국의 사회보장 영역에서도 인공지능 기술 적용이 폭넓게 활용되고 있다. 새롭게 짜인 인공지능 기본법에서 사회보장 영역을 간과한 점은 의아한 대목이다. 해당 법은 처음부터 이렇게 중요한 결함을 가지고 있다. 앞으로라도 인공지능 기본법에서 사회보장 영역을 포괄하는 내용으로 개정을 시도하거나, 사회보장 영역에서의 인공지능 적용 및 규제에 대한 별도의 입법을 시도하는 방안을 검토할 수 있을 것이다. 정부의 행정 영역에서 사회보장 분야가 차지하는 비중을 고려하면 더욱 그러하다. 미국의 경우, 미국 연방정부에서 인공지능을 활용하는 사례 가운데 3분의 1 이상이 보건복지부 업무 영역에 속한다(Burt, 2024.10.17.). 미국의 행정명령에서 지속적으로 '보건복지부'가 호명되는 이유다. 한국에서도 정부 지출 가운데 보건·복지·고용 영역 비율이

31) 2024년 12월 30일 현재 국회 본회의를 통과한 '인공지능 발전과 신뢰 기반 조성을 위한 기본법' 전문은 공개되지 않고 있다. 고환경 외(2024)가 소개한 인공지능법 요약 내용을 통해서 법의 내용을 미루어 짐작하는 수밖에 없다.

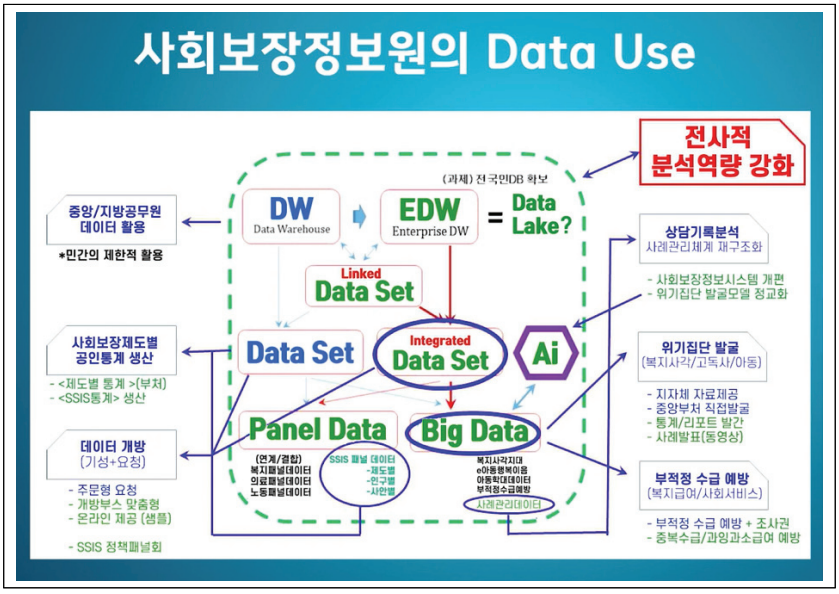
가장 높은 건 주지의 사실이다. 이와 같은 점을 염두에 두고 앞으로 빅데이터와 인공지능의 사회보장 영역에서의 적용을 촉진하기 위한 정책 제언을 제시하도록 하겠다.

첫째, 사회보장 정보에 활용되는 데이터 품질을 개선하기 위한 정책적인 노력이 필요하다. 제2장에서 언급한 대로, 쓰레기 데이터가 쓰레기 출력값을 산출하는(garbage in, garbage out) 문제(James, 2024)를 막아야 한다. 데이터와 데이터에 근거한 인공지능의 결괏값에 대한 신뢰도를 높이기 위해서는 양질의 데이터 확보 및 집적, 관리가 필요하다. "사회보장제도에 대한 공정한 평가와 이를 기반으로 한 사회보장제도의 설계를 위해서는 정확한 사회보장제도 관련 정보가 필요하다. 따라서 정보 포괄성을 바탕으로 하는 정확한 데이터 품질 관리가 요구된다"(이상원, 2023, p. 3). 공적 데이터를 관리하는 기관들에 대한 인적, 행정적 인프라 개선 및 확충이 필요함을 의미한다.

둘째, 사회보장 영역에서 데이터 통합, 연계, 관리를 위한 노력이 필요하다. 사회보장 영역에서는 2022년부터 사회보장기본법 제42조)에 근거해서, 사회보장 정책의 심의·조정과 관련 연구를 위해서 행정데이터가 수집되고 있다. 한국에서는 모든 개인에게 부여되는 주민등록번호로 기관들의 데이터를 연계할 수 있는 강력한 여건을 구축하고 있다. 그렇지만 데이터 연계 과정에서 주민등록번호를 결합키 생성에 사용하지 못하도록 하는 법적인 규제가 발목을 잡고 있다. 더욱이 대부분의 정부 부처와 공공기관들도 데이터 공개에 소극적이다. 북유럽, 미국, 영국, 캐나다 등에서 행정자료 구축과 공개, 연계에 적극적인 점(유종성, 2022)을 참고할 필요가 있다. 한국에서도 노대명(2023)이 사회보장정보원을 중심으로 행정 데이터 연계와 활용 방안에 대한 구상을 제시한 바 있다(〔그림 5-2〕 참고).

셋째, 다양한 기관들이 집적한 정보들의 표준화 및 단순화가 필요하다(노대명, 2024). 이는 위의 데이터 질 개선 및 연계와 직결되는 문제다. 노대명(2024)은 사회보장 정보시스템의 많은 데이터가 제대로 활용되고 있지 않다며, 사회보장정보원의 데이터 테이블은 11,280개에 달하지만 정작 자주 활용되는 정보는 500개 칼럼에 불과하다고 지적한다. 다른 부처와 기관이 생산하는 데이터들 역시 통합은커녕 연계가 힘든 상황이다. 현재로서는 구슬은 서 말이지만, 이들을 꿰지는 못하고 있다.

[그림 5-2] 사회보장데이터의 정제와 활용



출처: 노대명. (2023). 한국사회보장정보원 행정데이터 자산화 프로젝트. 사회보장정보원 유튜브. https://www.youtube.com/watch?v=zl7XhnJBDSM

넷째, 데이터 구축, 연계, 활용 과정에서 데이터 보안 및 안전에 대한 엄격한 기준 설정이 반드시 필요하다. 유종성(2023)은 행정 데이터 활용에서의 향후 과제를 제안하면서 다음과 같이 언급했다. "전수를 포괄하는 빅데이터인 경우가 많아 개인정보 보호에 보다 세심한 주의가 필요하다. 개인정보 보호에 대한 안전장치를 이중, 삼중으로 설치할 필요가 있다. 과거 스웨덴 등 북유럽 국가들에서도 데이터 활용과 개인정보 보호를 둘러싼 논쟁이 있었다. 개인정보 보호를 철저히 하면서 데이터를 연계하고 가명 처리된 데이터를 연구자들이 안전하게 사용할 수 있는 방법들이 발전되어 왔다"(p. 15). 개인의 소득, 재산, 건강, 가족 등의 개인정보가 결합될수록 데이터 유출의 충격은 커질 수 있다. 이 문제는 데이터 보안을 위한 데이터 거버넌스의 문제와도 직결된다.

다섯째, 데이터 관리를 넘어서 이를 활용하는 알고리즘 및 인공지능 시스템의 편향성을 최소화하면서, 정확성을 개선하기 위한 노력이 반드시 필요하다. 이 대목에서는 호주의 로보데트 스캔들(Robodebt Scandal)을 복기할 필요가 있다. 호주는 2016년부터 개개인의 복지수당 초과 지급분의 계산 및 초과분 환수 통보를 자동화한 로보데트 시스템(Robodebt Scheme)을 도입했다(홍승헌, 황하, 2024). 이 과정에서 데이터 매칭 알고리즘이 핵심적 역할을 담당했지만, 47만 건에서 오류가 발생한 것으로 추정됐다. 문제가 된 액수는 총액 10억 호주 달러였다. 데이터에 기반한 알고리즘에 과신·과의존함으로써 수많은 사회적 약자들에게 해악을 끼친 대표적 사례다. 2020년, 그 시스템은 공식적으로 폐지됐다(홍승헌, 황하, 2024). 정확성이 담보되지 않은 인공지능 기반 사회보장 행정이 초래할 수 있는 끔찍한 결과였다. 특히, 인공지능을 활용한 행정에서 가장 흔하게 지목되는 리스크가 편견과 불평등의 심화(O'neil, 2017; Eubanks, 2018)라는 점을 상기할 필요가 있다. 인공지능 기반 행정에 대한 신뢰는

과학성, 중립성에 대한 입증을 통해서 조금씩 누적될 수 있을 것이다. 그리고 그 신뢰가 무너지는 것도 한순간이다.

여섯째, 한국의 정부 부처에서, 특히 사회보장 영역에 한정하면, 보건복지부에서 인공지능 관련 조직과 인력을 신설 및 배치할 필요가 있다. 유럽연합의 인공지능법에 따르면, 고위험성 인공지능 시스템 제공자는 17개의 의무를 지게 된다.[32] 물론, 유럽연합의 다소 엄격한 기준이 그대로 한국에 적용될 것으로 확언하기는 어렵다. 그럼에도 불구하고, 미국에서조차도 미국 보건복지부가 행정명령에 따른 다양한 이행 조치를 취해야 함은 제4장에서 살펴봤다. 이를테면, 'Office of the Chief Artificial Officer(OCAIO)'를 임명하고, 'HHS Artificial Intelligence (AI) Strategy'를 공표해야 한다. 한국에 보건복지부는 시스템 제공자 혹은 제공자의 관리감독 기관으로 해당 조치들에 대해 적극적으로 검토할 필요가 있다. 지금까지 보건복지부에서는 사회보장 행정 영역에서 인공지능과 관련한 조치를 내놓은 바가 없다.

일곱째, 국제적인 인공지능 발전과 규제 동향에 대한 모니터링이 필요하다. 제4장에서 살펴본 바와 같이, 유럽연합의 인공지능법과 미국의 행정명령 14110도 적용 과정을 거치면서 규제의 내용이 구체화될 가능성이 높다. 특히, 인공지능법은 아직 추상 수준이 높기 때문이다. 이와 관련해 앞으로 인공지능 발전의 속도, 발전이 사회에 미치는 파장, 인공지능

32) ① 고위험성 AI 시스템 준수사항 준수 여부 확인, ② 고위험성 AI 시스템에 연락처 표시, ③ 품질 관리 시스템 준수, ④ 기술문서, 품질관리시스템 문서 등 필수 자료 보관, ⑤ 통제하에 있는 경우 자동으로 생성된 로그 보관, ⑥ 시장 출시/서비스에 투입되기 전 적합성 평가 절차 준수, ⑦ EU 적합성 선언 작성, ⑧ CE 마크 부착, ⑨ 등록 의무 준수, ⑩ 필요한 시정 조치(규정 위반 시 적절한 조치) 및 정보 제공, ⑪ 국가 관할 기관의 합리적인 요청이 있을 때, 고위험성 AI 시스템이 모든 준수사항에 부합함을 입증, ⑫ 제품 및 서비스에 대한 접근성 의무 보장, ⑬ 관할 당국과의 협력, ⑭ 중요한 변경 시 적합성 평가 실시, ⑮ 출시 후 모니터링 체계 설계·구축·운영, ⑯ 중대 사고 관련 정보 공유 및 신고 의무, ⑰ AI 리터러시 보장

패권을 둘러싼 미·중·유럽권의 경쟁, 국제기구의 개입 등이 복잡하고 역동적으로 작용할 것으로 예상된다. 무엇보다 중요한 변수는 정치인과 정책 전문가들마저도 인공지능이 정확히 무엇인지, 그래서 인공지능이 미칠 파장이 어떠할지에 대해 잘 모르고 있다는 점이다. 놀라운 일은 아니다. 심지어 기술 전문가들 사이에서도 인공지능의 영향에 대해서 두머(doomer)와 부머(boomer)로 엇갈리고 있는 점을 상기할 필요가 있다(제2장 참고). 물론, 미지(未知)보다 해소되기 어려운 문제는 상충하는 이해관계라는 점도 확인해둔다. 한국의 사회보장 영역에서 인공지능의 적용을 지원하고 규제하고자 한다면, 이러한 내·외 환경에 대한 동적인 모니터링이 필요할 수밖에 없다.

여덟째, 지금까지 제시한 데이터 관리, 연계, 표준화 및 알고리즘 질 관리 등을 총괄하는 거버넌스 체계를 구축해야 한다. 한국에서 행정 데이터 활용에 관한 논의는 데이터 활용도를 높이는 방향을 중심으로 논의되는 경향이 있다. 그러나 호주의 Robodebt 사례에서 살펴본 바와 같이, 데이터를 활용한 복지 행정에서의 오류 혹은 사고는 한 번만 발생해도 전체 시스템을 폐쇄할 정도로 충격이 크다는 점을 고려할 필요가 있다. 이와 관련해서, 2024년 국회 본회의를 통과한 인공지능 기본법에서는 규제의 주체로 과학기술정보통신부를 호명하고, 독립적인 규제 및 감독기구가 구체적으로 제시되지는 않은 것으로 보인다. 지금까지 공개된 초안에서는 인공지능 사업자나 대통령령으로 정하는 인공지능 기술 관련 기관이 '민간 자율 인공지능윤리위원회'를 '둘 수 있다'는 느슨한 규제를 제시하고 있다. 한국의 인공지능 기본법이 진흥과 규제 사이에서 진흥 쪽으로 과하게 경도됐다고 보는 것이 적절할 것이다. 김병권(2024.12.26.)이 지적한 대로, 새로운 법은 규제법은 아니고 그냥 '진흥법'이다.

앞에서 언급한 대로, 제대로 된 규제 없이 인공지능은 발전할 수 없다. 사회보장 영역은 사람을 대상으로, 특히 빈곤, 아동, 노인 등 취약계층을 대상으로 한다. 그리고 이들의 사적인 소득, 건강, 재산 자료에 근거해서 정책을 편다. 제4장에서 살펴본 바와 같이, 유럽연합에서 금지하는 'social scoring'과 밀접하다. 그래서 인공지능 정책에서 시장 친화적인 정책을 펴는 미국에서도 모든 부처들이 전담 포스트인 'Office of the Chief Artificial Officer(OCAIO)'를 임명하고, 'HHS Artificial Intelligence(AI) Strategy'를 발표하도록 했다. 특히, 보건복지부에 대해서는 보다 상세한 지침을 제시했다.

종합하면, 한국의 공공영역에서 디지털 기술, 인공지능 기술의 도입은 빠른 반면, 관련된 도덕적, 법적 쟁점에 대한 사회적 논의 및 규제 형성 과정은 매우 더디다. 제3장에서 보았듯이, 사회보장 영역에서 국내의 인공지능 기반 기술의 도입 속도는 다른 복지국가들과 견줘 전혀 늦지 않다. 이러한 점을 고려한 디지털 거버넌스 구축이 필요하다. 이와 관련해서 유도진(2024.12.27.)은 인공지능의 민주적 활용을 위해 세 가지 제도적 기구의 설치를 제안했다. 첫째, 인공지능 활용 정책과 데이터 사용의 적법성, 투명성을 감사하는 독립기구다. 여기에는 인권, 기술, 법률 등 다양한 전문가가 참여한다. 둘째, 정부의 인공지능 활용에 대한 시민의 감시와 통제를 위한 옴부즈맨 프로그램이다. 신고 절차를 간소화해서 시민 참여를 확대하는 식의 접근을 제안했다. 셋째, 인공지능 시스템의 알고리즘 공정성, 데이터 처리의 투명성, 윤리적 설계를 평가하고 인증하는 독립 인증기구다. 여기에서는 기술적·윤리적 위험 요소를 정기적으로 점검하고, 문제 발생 시 시정 조치를 요구할 수 있는 권한도 가지게 된다. 유도진(2024.12.27.)의 안은 정부안과 비교할 때, 국가에 대한 '견제'에 다소 편향된 측면이 있음을 고려할 필요는 있다.

사회보장 행정에 인공지능 기술을 적용하는 것에 관한 거버넌스는 인간 중심적 활용이라는 원칙 아래 ① 민주적 통제, ② 시민 참여, ③ 프라이버시를 보장하고, 동시에 시민의 적극적 사회권 보호라는 원칙 아래 ① 사각지대 해소, ② 행정 효율화 제고, ③ 급여 지급의 정확성, 적시성 보장의 편의를 제공하는 정책 방향을 확인할 필요가 있다.

이러한 정책 방향의 수립 및 집행은 시급하다. 인공지능의 빠른 발전 속도를 고려하면 더욱 그러하다. 제2장에서 언급한 바와 같이, 현재의 세 번째 인공지능 여름기가 다시 세 번째 인공지능 겨울기로 접어들게 된다면, 그 이유는 인공지능의 성능 부족 때문이 아니라 인공지능에 대한 사회적 신뢰성의 붕괴 때문일 것이다.

참고문헌

감사원. (2022). **감사보고서- 취업알선정보망 구축 및 관리실태**. 감사원.

강애란. (2019. 8. 16.). 복지부, '탈북민 모자 사망'에 위기가구 긴급 실태조사. 연합뉴스. https://www.yna.co.kr/view/AKR20190816136800017

강지원. (2024). **정부의 사회보장 서비스에서 AI의 활용과 EU/미국 일부 주 법의 시사점**. [발표 자료] 사회보장적용에서의 인공지능 적용동향과 함의 세미나. 한국보건사회연구원.

강혜규, 강은나, 강지원, 강희정, 김기태, 김세진, 김태완, 류정희, 서윤경, 오욱찬, 윤강재, 이상정, 이원진, 이주민, 이한나, 조성은, 주하나. (2022). **사회보장제도 실태분석 및 개선방안 연구**. 보건복지부, 한국보건사회연구원.

개인정보보호법, 법률 제19234호 (2011).

개인정보보호위원회. (2021). 인공지능(AI) **개인정보보호 자율점검표 개발자, 운영자용.** 개인정보보호위원회.

개인정보보호위원회. (2023). **개인정보 보호 기본계획(2024-2026)**. 개인정보보호위원회.

고용노동부. (2020. 7. 9.). 인공지능(AI) 기반의 일자리-인재 추천 서비스 시작. 고용노동부 [보도자료]. https://www.moel.go.kr/news/enews/report/enewsView.do?news_seq=11166

고용노동부. (2022. 5. 26.). 일자리포털 워크넷, 직무역량 중심 인공지능 일자리연결서비스 시범 운영. 고용노동부 [보도자료]. https://eiec.kdi.re.kr/policy/materialView.do?num=201066&topic=

고용노동부. (2022. 8. 30.). 인공지능(AI)이 당신의 헤드헌터가 되어 일자리와 인재를 추천해 드립니다. 고용노동부 [보도자료]. https://eiec.kdi.re.kr/policy/materialView.do?num=229615&topic=

고용노동부. (2024. 6. 12.). 맞춤형 구인·구직·매칭서비스가 인공지능(AI) 기반으로 확 달라집니다. 고용노동부 [보도자료]. https://www.moel.go.kr/news/enews/report/enewsView.do?news_seq=16663

고환경, 채성희, 손경민, 이일신. (2024.12.). 인공지능 기본법 국회 본회의 통과. 법무법인 광장 Newsletter. [보도자료]. https://www.leeko.com/upload/news/newsLetter/2113/20241227162747484.pdf

공공데이터의 제공 및 이용 활성화에 관한 법, 법률 제19408호 (2023).

과학기술정보통신부. (2020.12.23.). 인공지능(AI) 윤리기준. 과학기술정보통신부. [보도자료]. https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=112&mId=113&bbsSeqNo=94&nttSeqNo=3179742

과학기술정보통신부·한국지능정보사회진흥원·한국정보통신기술협회. (2024). **인공지능 학습용 데이터 품질관리 가이드라인 v3.1**. 한국지능정보사회진흥원.

관계부처합동. (2023). **제1차 (`23~`25년) 데이터 산업 진흥 기본계획(안)**. 관계부처합동.

교육부. (2022). **교육분야 인공지능 윤리원칙**. 교육부.

국가인권위원회. (2022). **인공지능 개발과 활용에 관한 인권 가이드라인**. 국가인권위원회.

금융위원회. (2021). **금융분야 AI 가이드라인**. 금융위원회.

김기태. (2024). **디지털 복지국가의 개념과 논점.** (비판과 대안을 위한 사회복지학회 학술대회 발표논문집, pp.133-145.)

김다은. (2024.6.12.). 인공지능 대항해 시대, 'AI 기본법' 없는 한국은?. 시사인. https://www.sisain.co.kr/news/articleView.html?idxno=53127#google_vignette

김명주. (2017). **인공지능 윤리의 필요성과 국내외 동향.** (정보와 통신, 34권 10호, pp.45-54.)

김명주. (2022). **AI는 양심이 없다**. 헤이북스.

김명주. (2023). AI의 역기능과 정책과제. 김정언(편), (경제·인문사회연구회. 디지털 전환을 통한 사회·경제 혁신 전략 연구, pp. 409-564.)

김명주 (2023). 일론 머스크도 '팽'했던 OpenAI 이사회, 그들은 왜?. 티타임즈 TV. [유튜브]. https://www.youtube.com/watch?v=o4NNzgL7Yj8&t=53s

김명주. (2024a). **인공지능의 잠재적 위험과 국제적 규제 동향.** (문명과 경계, Vol. 8., pp.43-77.)

김명주. (2024b). 세상읽기 딥페이크 사태가 남겨준 교훈. 경기일보. [보도자료]. https://www.kyeonggi.com/article/20240924580094

김명주. (2024). **AI 윤리와 규제**. [발표 자료] 사회보장적용에서의 인공지능 적용 동향과 함의 세미나. 한국보건사회연구원.

김병권. (2024.12.26.). 한국의 처참한 AI법. [페이스북 게시글]. https://www.facebook.com/byoungkweon.kim

김수영, 김수완. (2022). **데이터 복지국가의 도래와 쟁점: 국가, 시민, 시장의 관계 지형을 중심으로.** (한국사회복지정책학회 춘계학술대회자료집, pp.91-118.)

김수영. (2016). **사회복지정보화의 윤리적 쟁점 사회보장정보시스템을 통한 데이터감시를 중심으로.** (한국사회복지학, 68(1), pp.193-224.)

김수완, 최종혁, 박동진. (2017). **노인장기요양서비스 제공과정에서의 복지기술 활용에 관한 연구.** (노인복지연구, 제72권, 제4호, pp.29-60.)

김은하. (2022). **빅데이터 정보시스템 활용 현황과 과제: 복지 사각지대 발굴 시스템을 중심으로.** (한국보건사회연구원 보건복지포럼, 제313호, pp.24-34.)

김재연. (2023). **우리에게는 다른 데이터가 필요하다**. 세종서적.

김정욱, 김종립, 최유경, 김민정, 유성희, 최현이, 추수진. (2023). **인공지능 시대의 경쟁력 강화를 위한 AI 규제 연구**. 경제인문사회연구회.

김형주. (2024.7.23.). **인공지능 윤리 핵심 가치 분석: 한국 사례를 중심으로**. [발표 자료]. 사회보장행정에서 인공지능 적용 동향과 함의 세미나. 한국보건사회연구원.

남궁준. (2024). **인공지능과 알고리즘 관리에 대한 노동법적 규율: 미국 바이든 정부의 AI 행정명령을 중심으로.** (사회법연구, 53, pp.183-218.)

남현숙, 안미소. (2023). **공공부문 AI 활용현황 실태조사**. 소프트웨어정책연구소.

남현숙, 안미소, 장진철, 이동현. (2023). **국내·외 공공부문 AI 활용현황 분석 및 시사점.** (소프트웨어정책연구소 이슈리포트, IS 157.)

데이터 산업 진흥 및 이용 촉진에 관한 기본법, 법률 제18475호(2021).

데이터기반행정 활성화에 관한 법률, 법률 제19408호(2020).

라기원. (2024). **유럽연합 인공지능법(EU AI ACT)의 특성과 쟁점: 우리나라 인공지능 입법에 대한 시사점.** (한국법제연구원. AI Issue Paper, 24-20-3.)

방송통신위원회·정보통신정책연구원. (2019). **이용자 중심의 지능정보사회를 실현하기 위한 원칙**. 방송통신위원회·정보통신정책연구원.

법제처 미래법제혁신기획단. (2024). **인공지능(AI) 관련 국내외 법제 동향**. 법제처.

보건복지부. (2023.1.17.). 2023년 4차 복지 사각지대 발굴 시작. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=377278&tag=&nPage=58

보건복지부. (2024. 3.25.). 45종 위기정보 활용해 2024년 2차 복지 사각지대 발굴 시행. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503000000&bid=0027&list_no=1480754&act=view

보건복지부. (2024.2.26.). 복지사각지대 발굴·지원 확대로 '약자복지' 지속 추진하겠습니다. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=1480402&tag=&nPage=1

보건복지부. (2024.7.22.). 인공지능(AI) 초기 복지상담 전화로 위기가구 지원에 나선다. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=1482378&tag=&nPage=1

보건복지부. (2024.8.23.). 수원 세 모녀 사건 계기 복지 사각지대 발굴·지원 체계 전반 점검. 보건복지부 [보도자료]. https://www.korea.kr/briefing/pressReleaseView.do?newsId=156522073&pageIndex=1&repCodeType=&repCode=&startDate=2021-08-23&endDate=2022-08-23&srchWord=&period=

보건복지부. (2024.11.25.). 어려움 겪는 위기가구 찾기 위해 인공지능(AI) 전화 초기상담 전국 시행. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503000000&bid=0027&list_no=1483698&act=

view&

보건복지부. (2018). 신(新)사회적 위험 증가에 따른 복지 위기가구 발굴 대책. 보건복지부.

보건복지부. (2024). 2024년 AI·IoT 기반 어르신 건강관리 사업안내서. 보건복지부.

보건복지부. (2024.11.26.). 보건의료데이터와 인공지능이 열어가는 디지털 헬스케어 미래. 보건복지부 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=1483714&tag=&nPage=1

BBC News 코리아. (2018). 성차별: 아마존, '여성차별' 논란 인공지능 채용 프로그램 폐기. BBC News 코리아. [보도자료]. https://www.bbc.com/korean/news-45820560

BBC News 코리아. (2019). 애플 신용카드: 성차별적인 정책으로 비판받은 애플의 신용카드. BBC News 코리아. [보도자료]. https://www.bbc.com/korean/international-50371171

사회보장급여의 이용·제공 및 수급권자 발굴에 관한 법률, 법률 제17201호 (2015).

서울특별시교육청. (2021). **인공지능(AI) 공공성 확보를 위한 현장 가이드라인**. 서울시교육청.

성윤희. (2022). **인공지능(AI)과 사회복지법제에 대한 소고.** (사회복지법제연구, 제13권, 제2호, pp.119-148.)

성은미·김송이. (2024). **복지사각지대 발굴사업의 한계점과 개선과제.** (한국지역사회복지학회, 제88집, pp.1-29.)

성준호. (2016). **주민등록번호에 의존한 본인확인제도의 문제점: 각국의 개인식별번호제도 및 관련 법률의 검토를 통한 시사점.** (공공사회연구, 제6권 제2호, pp.208-246.)

성폭력범죄의 처벌 등에 관한 특례법, 법률 제17264호 (2020).

식품의약품안전처. (2022). **인공지능(AI)의 의료기기 국제 공통 가이드라인**. 식품의약품안전처.

안준모. (2021). **인공지능을 통한 행정의 고도화: 기회와 도전.** (한국행정연구원, vol.30, no.2, 통권 65호, pp.1-33.)

알고리즘 및 인공지능에 관한 법률안. (2021. 11. 24.). 윤영찬 의원 대표발의.

임대준. (2023). 오픈AI, AGI 개발 돌파구 찾았다. 인간처럼 추론하는 LLM 큐스타 개발 중. AI타임즈. [보도자료]. https://www.aitimes.com/news/articleView.html?idxno=155433

AWS. (2024). 자기 회귀 모델이란 무엇인가요?. https://aws.amazon.com/ko/what-is/autoregressive-models

외교부. (2024.5.22). 대한민국, '서울 선언'을 통해 글로벌 인공지능(AI) 거버넌스의 새로운 방향 제시. 외교부 [보도자료]. https://kcg.korea.kr/briefing/pressReleaseView.do?newsId=156631917&pWise=sub&pWiseSub=J2

유도진. (2024.12.27.). AI 기본법, 기술과 민주주의의 균형, 해결해야 할 과제들. 오마이뉴스. [보도자료]. https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0003092112

유종성. (2023). **사회보장 행정데이터 활용사례와 향후 과제.** (보건복지포럼, Vol.325, pp.6-19.)

윤난슬. (2021.4.14.). 전주비전대, 인공지능 기반 '취업 알선 서비스' 적극 활용. 네이트뉴스. https://news.nate.com/view/20210414n14821

윤화정. (2022.04.12.). 128억 투입된 워크넷 'AI 일자리 매칭', 성과 저조…매칭 점수-입사지원간 상관관계 낮아. 워크투데이. [보도자료] http://www.worktoday.co.kr/news/articleView.html?idxno=24170

윤혜선. (2024). **EU 인공지능법의 주요 내용과 함의**. [발표 자료]. 사회보장적용에서의 인공지능 적용동향과 함의 세미나. 한국보건사회연구원.

이상길. (2018). **국내외 AI 활용 현황과 공공 적용.** (정보통신기술진흥센터. ICT SPOT ISSUE, S18-08호.)

이상원. (2023). **사회보장 분야 행정데이터의 활용과 전망.** (보건복지포럼., Vol.325, pp.2-4.)

이승윤, 백승호, 남재욱. (2020). **한국 플랫폼노동시장의 노동과정과 사회보장제의 부정합.** (산업노동연구, Vol.26, No2, pp. 77-135.)

이영글, 박성준, 함영진. (2021). **주민이 참여하는 인적 안전망이 복지 사각지대 해소에 미치는 영향.** (사회복지정책, 제48권, 제1호, pp.97-121.)

이우식. (2024). **한국사회보장정보원 사회보장행정에서의 인공지능 적용 동향과 함의**. [발표 자료]. 사회보장 적용에서의 인공지능 적용 동향과 함의 세미나. 한국보건사회연구원.

이원태. (2024.10.24.). EU 못지않은 미국에서의 AI 입법 과잉?. [페이스북 게시글]. https://www.facebook.com/wontae.lee.9889

이하나. (2011). **광고 매체로서 디지털 사이니지(Digital Signage) 활성화 방안에 관한 연구.** (한국디자인문화학회지, Vol.17, no.2, pp.502-517.)

이후연. (2024). 이 사람 스펙 좋지만 곧 나가, 요즘 AI 면접관, 별걸 다 안다 채용시장 바꾸는 AI. 중앙일보. [보도자료]. https://www.joongang.co.kr/article/25271418

인공지능 개발 및 이용 등에 관한 법률안. (2024.7.4.). 권칠승 의원 대표발의.

인공지능 기본법안. (2024.8.22.). 한민수 의원 대표발의.

인공지능 기술 기본법안. (2020.10.29.). 민형배 의원 대표발의.

인공지능 발전 진흥과 사회적 책임에 관한 법률안. (2024.8.24.). 배준영 의원 대표발의.

인공지능 발전과 신뢰 기반 조성 등에 관한 법률안. (2024.6.17.). 정점식 의원 대표발의.

인공지능 산업 육성 및 신뢰 확보에 관한 법률안. (2024.5.31.). 안철수 의원 대표발의.

인공지능 연구개발 및 산업 진흥, 윤리적 책임 등에 관한 법률안. (2020.7.13.). 이상민 의원 대표발의.

인공지능 육성 및 신뢰 기반 조성 등에 관한 법률안. (2021.7.1.). 정필모 의원대표발의.

인공지능 책임 및 규제법안. (2023.8.8.). 안철수 의원 대표발의.

인공지능기술 기본법안. (2024.6.28.). 민형배 의원 대표발의.

인공지능산업 육성 및 신뢰 확보에 관한 법률안. (2022.12.7.). 윤두현 의원 대표발의.

인공지능산업 육성 및 신뢰 확보에 관한 법률안. (2024.6.1.). 조인철 의원 대표발의.

인공지능산업 육성 및 신뢰 확보에 관한 법률안. (2024.6.19.). 김성원 의원 대표발의.

인공지능산업 육성에 관한 법률안. (2020.10.29.). 양향자 의원 대표발의.

인공지능산업 진흥 및 신뢰 확보 등에 관한 특별법안. (2024.9.24.). 김우영 의원 대표발의.

인공지능에 관한 법률안. (2021. 7. 19.). 이용빈 의원 대표발의.

인공지능의 발전과 안전성 확보 등에 관한 법률안. (2024.9.12.). 이훈기 의원 대표발의.

인공지능책임법안. (2023. 2. 28.). 황희 의원 대표발의.

인공지능책임법안. (2024. 8. 27.). 황희 의원 대표발의.

임덕영. (2023). **사회보장 현장 모니터링: 현장 전문가와 실무자 포럼**. 한국보건사회연구원.

임영모, 윤서경, 안성원. (2023). **공공부문 AI 도입현황 연구**. 소프트웨어정책연구소.

장동욱. (2022.11.25.). 전입신고 안됐다고 방치…'복지 사각지대 발굴' 시스템 구멍. TV조선. [보도자료]. https://n.news.naver.com/mnews/article/448/0000384107

장하석. (2023.2.14.). 인공지능이 사람처럼 글을 쓴다면. [보도자료]. https://v.daum.net/v/BeO7PIkUFv?f=p

정세정, 김기태, 곽윤경, 우선희, 최준영, 이영수. (2023). **한국 복지국가의 재구조화를 위한 연구 - Ⅰ. 디지털 복지국가의 딜레마**. 한국보건사회연구원.

정유채. (2022). **AI와 블록체인 기반 의료 복지 서비스 사례 연구 1.** (복지경영학연구, 제11권, pp.77-86).

조남경, 송기호. (2023). **사회복지의 상담기록, 좀 더 활용할 수 있을까? '머신러닝'을 통한 사회복지 상담 텍스트 활용 가능성의 점검.** (한국사회복지조사연구, Vol.79, pp.5-26).

조달정보개방포털(data.g2b.go.kr) (인출일: 2024.10.19)

조이환. (2024). 美 캘리포니아, AI 규제안 통과 '임박'…빅테크·학계 '긴장'. ZDNET Korea. [보도자료]. https://zdnet.co.kr/view/?no=20240808171204

중앙선거관리위원회. (2024). 딥페이크영상 등을 이용한 선거운용 제한된다. 중앙전서관리위원회 [보도자료]. https://www.nec.go.kr/site/abroad/ex/bbs/View.do?cbIdx=1195&bcIdx=197565&relCbIdx=1147

중앙선데이. (2020). 대입 'A레벨' 성적 엉터리 산정, 알고리즘이 기가 막혀. 중앙선데이. [보도자료]. https://www.joongang.co.kr/article/23875682

최정은, 김윤영, 최기정, 이인수. (2022). **복지사각지대 발굴관리시스템 대상자 실태분석을 통한 지원방안 연구**. 한국사회보장정보원.

최종혁· 김수완. (2017). **공공복지전달체계에서의 복지기술 활용에 관한 연구: 사회보장정보시스템(행복e음)에 대한 사회복지공무원 인식을 중심으로.** (사회복지정책, Vol.44, No4, pp.188- 222.)

한국사회보장정보원 내부자료. (2024). **제안요청서-AI활용 초기상담정보시스템 운영지원 사업**. 한국사회보장정보원.

한국지능정보사회진흥원. (2018.9). **지능정보사회 윤리 가이드라인과 지능정보사회 윤리헌장**. 한국지능정보사회진흥원.

함영진, 이현주, 어유경, 김가희, 박성준, 조용찬. 오민수. (2023). **복지 전달체계 혁신을 위한 대안적 고찰: 취약계층 발굴정책 개선을 중심으로**. 한국보건사회연구원.

행정안전부. (2023. 4. 19.). 정부 24 회원 2천만명 돌파 눈앞, 국민 3명 중 1명이 가능해요. 행정안전부 [보도자료]. https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=100046

홍승헌, 황하. (2024). **누구를 위한 디지털 전환인가?: 자동화된 복지행정의 위험성.** (정부학연구, Vol.30, No.2, pp.61-84.)

Abi-Chahla, F. (2008). **Nvidia's CUDA: The End of the CPU?**. Tom's Hardware.

Agency for Healthcare Research and Quality. (2023). Patient Safety Organization(PSO) Program. https://pso.ahrq.gov/

Akhmedjonov, A. (2023). T**he implication of AI in social welfare systems: Potential risks and prevention measures**. Master thesis of Central European University.

Alder, M. (2024.5.29.). HHS names acting chief AI officer as it searches for permanent official. Fedscoop. [online]. https://fedscoop.com/hhs-names-acting-chief-ai-officer/

Algorithm Watch and Bertelsmann Stiftung. (2020). **Automating Society Report 2020**. [online]. https://automatingsociety.algorithmwatch.org

Alston, P. (2019). **Digital Welfare States and Human Rights. Report of the Special Rapporteur on Extreme Poverty and Human Rights.** [online]. https://undocs.org/A/74/493.

Appelman, N., Fathaigh, R. O., & van Hoboken, J. (2021). **Social welfare, risk profiling and fundamental rights: The case of SyRI in the Netherlands**. J. Intell. Prop. Info. Tech. & Elec. Com. L., 12, 257-271.

BBC News Korea. (2023). 'AI 대부' 제프리 힌턴, 구글 퇴사하면 AI 위험성 경고. BBC News Korea. https://www.bbc.com/korean/articles/crgm8d787l7o

Bekker, S. (2021). **Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. Netherlands Yearbook of International Law 2019: Yearbooks in International Law: History**. Function an

d Future, 289-307.

Bendixen, M. (2018) Denmark's 'anti-ghetto' laws are a betrayal of our tolerant values. The Guardian. [online]. https://www.theguardian.com/commentisfree/2018/jul/10/denmark-ghetto-laws-niqab-circumcision-islamophobic

Birhane, A. (2024). **How does AI development led by big tech reproduce bias and inequality?**. [발표 자료]. 2024 사람과 디지털 포럼. 한겨레.

Bostrom et al. (2014). **Chapter 15. The Ethics of Artificial Intelligence**. The Cambridge Handbook of Artificial Intelligence, pp.316-334. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139046855.020

Bugcrowd. (2024). Bug Bounty OpenAI. https://bugcrowd.com/engagements/openai

Burt. C. (2024.10.17.). HHS Strategic Plan for AI Coming Soon. ExecutiveGov. https://executivegov.com/2024/10/department-of-health-and-human-services-strategic-ai-plan/

Cesareo, S., White, J. (2023). **The Global AI Index**. Tortoise Media.

Coeckelbergh, M. (2020). **AI Ethics**. Cambridge. The MIT Press., ISBN. 9780262538190, 229 pages.

Cortes, C. and et al. (1995). **Support-vector networks.** Machine Learning, 20, (3), pp.273. doi:10.1007/BF00994018

Crevier, D. (1993). **AI: The Tumultuous Search for Artificial Intelligence.** NY: BasicBooks. ISBN 0-465-02997-3.

Davies, C. (2009). Microsoft remote software 'kill switch' confirmed. SlashGear. [online]. https://www.slashgear.com/microsoft-remote-software-kill-switch-confirmed-1656965/

Davies, G. (2022). **Report on Accounts: Department for Work & Pensions**. London: National Audit Office.

Del Castillo, A. (2023). The AI Act: deregulation in disguise. Social Europe. [online]. https://www.socialeurope.eu/the-ai-act-deregulation-in-disguise

Department of Health and Human Services. (2024). [online]. https://www.hhs.gov/ocr/get-help-in-other-languages/korean.html

Deutsche Sozialversicherung Europavertretung. (2024). Conference highlights the synergy between AI and the European Pillar of Social Rights. [online]. https://dsv-europa.de/en/news/2024/03/ki-in-sozialer-sicherheit.html

Devlin, J. and et al. (2018). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. arXiv:1810.04805v2 [cs.CL]

Ehrlich, P. & Ehrlich A. (2023). **The Dominant Animal: Human Evolution and the Environment**.

Ethical Intelligence. (2021). The Impending EU AI Act – What you need to know. EI.

EU. (2024). EU AI Act: first regulation on artificial intelligence. [online]. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

EU Artificial Intelligence Act. (2024). Historic Timeline. [online]. https://artificialintelligenceact.eu/developments/

Eubanks, V. (2018). **Automating inequality: How high-tech tools profile, police, and punish the poor**. NY: St. Martin's Press.

European Commission AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. Retrieved. [online]. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

European Commission. (2024). **AI Act**. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament. (2024). **AI Act.** European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792.

European Union. (2024). **AI Act**. [online]. https://artificialintelligenceact.eu/

Fleming, S. (2021). 5 ways AI is doing good in the world right now. World Economic Forum. [online]. https://www.weforum.org/agenda/2021/07/ai-artificial-intelligence-doing-good-in-world/

FLI. (2023). Pause Giant AI Experiments: An Open Letter. [online]. https://futureoflife.org/open-letter/pause-giant-ai-experiments

Frey, C. B., Osborne, M. A. (2017). **The future of employment: How susceptible are jobs to computerisation?,** Technological forecasting and social change, 114, pp.254-280.

Gao, Y. and et al. (2023). **Retrieval-Augmented Generation for Large Language Models: A Survey**. arXiv:2312.10997 [cs.CL].

Gartner. (2024). Gartner Hype Cycle - Interpreting Technology Hype. [online]. https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

Goodfellow, I. and et al. (2014). **Generative adversarial nets**. In Advances in neural information processing systems, pp.2672-2680.

Harari, Y. (2018). **21세기를 위한 21가지 제언. (전병근 옮김)**. 김영사.

Heath, R. (2024). What's in Biden's AI executive order — and what's not. AXIOS. [online]. https://www.axios.com/2023/11/01/unpacking-bidens-ai-executive-order

Hila, Mehr. (2017). **Artificial Intelligence for Citizen Services and Government**. [online]. https://ash.harvard.edu/wp-content/uploads/2024/02/artificial_intelligence_for_citizen_services.pdf

Huang, Haomiao. (2023). **How ChatGPT turned generative AI into an "anything tool"**. Ars Technica.

Human Rights Watch. (2023). EU: Artificial Intelligence Regulation Should Ban Social Scoring. Human Right Watch. [online]. https://www.hrw.org/news/2023/10/09/eu-artificial-intelligence-regulation-should-ban-social-scoring

INSS. (2023). Helô, assistente virtual do INSS completa três anos. [online]. https://www.gov.br/inss/pt-br/noticias/helo-assistente-virtual-do-inss-completa-tres-anos

ISSA. (2020). ELYA: The bilingual virtual assistant of the Employees Provident Fund - Empowering customers to self-serv, anytime, anywhere. [online]. https://www.issa.int/gp/208625

ISSA. (2021). **The application of chatbots in social security: Experiences from Latin America (Analysis)**. Geneva: International Social Security Association.

Jellinek, G. (1878). **Die sozialethische Bedeutung von Recht, Unrecht und Strafe** (1878; 2nd ed., 1908; The Social-Ethical Significance of Right, Wrong, and Punishment).

Ji, Z. and et al. (2022). **Survey of Hallucination in Natural Language Generation**. ACM Computing Surveys, 55 (12), pp.1–38. arXiv:2202.03629. doi:10.1145/3571730

Jooyoung L. and et al. (2023). **Do Language Models Plagiarize?**. Proceedings of the Web Conference 2023.

Jørgensen, R. F. (2021). **Data and rights in the digital welfare state: the case of Denmark**. Information. Communication & Society, pp.1-16.

Joseph, S. (2023). **The Diversity of Artificial Intelligence: How Edward Feigenbaum Developed the Expert Systems**. Medium.

Kline, R. (2011). C**ybernetics, Automata Studies and the Dartmouth Conference on Artificial Intelligence**. IEEE Annals of the History of Computing. IEEE Computer Society.

Koetsier, J. (2017). **Stephen Hawking Issues Stern Warning On AI: Could Be 'Worst Thing' for Humanity**. Forbes.

Krizhevsky, A., Sutskever, I. & Hinton G. (2012). **Imagenet classification with deep convolutional neural networks.** Advances in Neural Information Processing Systems, 25, pp.1097-1105.

Kurzweil, R. (2005). **The Singularity is Near: When Humans Transcend Biology**. Viking.

La Quadrature du Net. (2022). CAF : le numérique au service de l'exclusion et du harcèlement des plus précaires. [online]. https://www.laquadrature.net/2022/10/19/caf-le-numerique-au-service-de-lexclusion-et-du-harcelement-des-plus-precaires/

Lee, J., Le, T., Chen, J., & Lee, D. (2023, April). **Do language models plagiarize?** (In Proceedings of the ACM Web Conference 2023, pp. 3637-3647).

Lee, W. (2024.9.29.). Gov. Gavin Newsom vetoes AI safety bill opposed by Silicon Valley. LA Times. [online]. https://www.latimes.com/entertainment-arts/business/story/2024-09-29/gov-gavin-newsom-vetoes-ai-safety-bill-scott-wiener-sb1047

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding... Chalmers, D. (2024). **Taking AI Welfare Seriously**. [online]. https://eleosai.org/papers/20241030_Taking_AI_Welfare_Seriously_web.pdf?fbclid=IwY2xjawGTy4ZleHRuA2FlbQIxMAABHbrtx1PtLxVtp-xd8sFSJKgXMUfU9hMw-fQxE5VP4gASKb3cn96LXDm3DQ_

aem_uEVOznvIKVcqpSKsn0iV1w

Longo, L. and et al. (2024). **Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions**. Information Fusion.

Lott, M. (2024.10.26.). AI IQ Test Results. [online]. https://www.trackingai.org/home

MacAskill, W. (2017). **Effective altruism: introduction.** Essays in Philosophy, 18 (1), eP1580:1–5. doi:10.7710/1526-0569.1580.

Madiega, T. (2024). **Briefing: Artificial Intelligence Act**. Brussel: European Parliamentary Research Service.

Marcus, G. (2024.8.21.). Why California's AI safety bill should (still) be signed into law and why that won't be nearly enough. [online]. https://garymarcus.substack.com/p/why-californias-ai-safety-bill-should

McCarthy, J. and et al. (1955). **A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.** [online]. https://doi.org/10.1609/aimag.v27i4.1904

Mchangama, J. and Liu, H.-Y. (2018). The Welfare State Is Committing Suicide by Artificial Intelligence. Foreign Policy. [online]. https://foreignpolicy.com/2018/12/25/the-welfare-state-is-committing-suicide-by-artificial-intelligence/

Milmo, D. (2024.12.27.). 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years. The Guardian. [online]. https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years?CMP=fb_gu&utm_medium=Social&utm_source=Facebook&fbclid=IwY2xjawHcgmlleHRuA2FlbQIxMQABHU-AFgOxSCLAci-CIAPVgRn1573Hc610rTnOfS5hjrpEP2cf1Mj1q1fk

Qw_aem_WZhY4Xt_aziBsI9u2GF2PQ#Echobox=1735317044

MITRE. (2024). ATLAS Matrix, Navigate threats to AI systems through real-world insights. [online]. https://atlas.mitre.org/

Mosene, K. (2024). One step forward, two steps back: Why artificial intelligence is currently mainly predicting the past(Digital Society Blog. Humboldt Institut für Internet und Gesellschaft). [online]. https://www.hiig.de/en/why-ai-is-currently-mainly-predicting-the-past/

O'Neil, C. (2017). **대량살상수학무기. (김정혜 옮김)**. 흐름출판. (Original work published 2016)

Odelberg, T. (2024 May). **Understanding the Future of Artificial Intelligence Governance: Comparing the EU AI Act and U.S. Executive Order on Safe AI.** Ford School of Public Policy. University of Michigan. [online]. https://stpp.fordschool.umich.edu/sites/stpp/files/2024-06/stpp-future-of-ai-governance.pdf

OECD. (2023). **OECD AI Principles overview**. Paris: OECD.

OECD. (2024a). **2023 OECD Digital Government Index: Results and key findings.** OECD Public Governance Policy Papers, No.44, OECD Publishing, Paris. https://doi.org/10.1787/1a89ed5e-en

OECD. (2024b). **Modernizing Access to Social Protection**. Paris: OECD. https://doi.org/10.1787/af31746d-en

OpenAI. (2023). **March 20 ChatGPT outage: Here's what happened.** [online]. https://openai.com/index/march-20-chatgpt-outage/

OpenAI. (2024). ChatGPT 4o (8 August version) [Large language model]. [online]. https://chatgpt.com/c/67256c5b-3a14-800b-a884-25fe2a21d1e9

Otterlo, M. and et al. (2012). **Reinforcement Learning and Markov Decision Processes. Adaptation, Learning, and Optimization.** pp.

3-42. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-27645-3_1. ISBN 9783642276446

OWASP. (2023). OWASP Top 10 for Large Language Model Applications. [online]. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Perera R. and et al. (2017). **Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature.** Computing and Informatics, 36 (1), pp.1-32. doi:10.4149/cai_2017_1_1. hdl:10292/10691.

Petrosyan, L., Ataliotou, K. (2024). A Tale of Two Policies: The EU AI Act and the U.S. AI Executive Order in Focus. [online]. https://trilligent.com/a-tale-of-two-policies-the-eu-ai-act-and-the-us-ai-executive-order-in-focus/

ProPublica. (2016). How We Analyzed the COMPAS Recidivism Algorithm. [online]. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Rechtbank Den Haag. (2020). SyRI legislation in breach of European Convention on Human Rights. de Rechtspraak. [online]. https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx

Roser, M. (2022). The brief history of artificial intelligence: the world has changed fast — what might be next?. [online]. https://ourworldindata.org/brief-history-of-ai

Roubini, N. (2024.2.5.) Artificial Intelligence vs. Human Stupidity. [online]. https://www.project-syndicate.org/commentary/ai-hype-and-potential-in-a-world-of-rising-mega-threats-by-nouriel-roubini-2024-02

Russell, S., Norvig, P. (2021). **Artificial Intelligence: A Modern Approach (4th ed.).** Hoboken: Pearson. ISBN 978-0-13-461099-3, LCCN 20190474.

SAE International. (2021). Levels of Driving Automation. [online]. https://www.sae.org/blog/sae-j3016-update

Schmidhuber, J. (2014). **Who Invented Back Propagation?**. IDSIA. Switzerland.

Searle, J. (1980). **Minds, Brains and Programs.** The Behavior and Brain Science 3, pp.417-457.

Shumailov, I. (2023). **The Curse of Recursion: Training on Generated Data Makes Models Forget.** [online]. https://www.semanticscholar.org/paper/The-Curse-of-Recursion%3A-Training-on-Generated-Data-Shumailov-Shumaylov/155aec5cff650263a4c71136f97570611d1bba7a

Stanford University - HAI. (2019). **AI Index Report 2019**. Stanford University.

Statt, N. (2018). **The AI boom is happening all over the world, and it's accelerating quickly**. The Verge.

Taylor, J., Hern, A. (2023.5.2.). 'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation. [online]. https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning

Tett G. (2023). What I Learned from talking to Google about chatbots, Finantial Times Magazine. [online]. https://www.ft.com/content/480ba8a0-8e04-404d-9dce-419e961aa9c1

Thapa, E.P. (2019). **Predictive Analytics and AI in Governance: Data-driven government in a free society – Artificial Intelligence, Big Data and Algorithmic Decision-Making in government from a liberal perspective**. European Liberal Forum.

The White House. (2023). **FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.** Execution Order 14110.

Tortoise Media. (2024). Global AI Ranking. [online]. https://www.tortoisemedia.com/intelligence/global-ai#rankings

Turing, A. (1950). **Computing Machinery and Intelligence.** Mind, Volume LIX, Issue 236, pp.433-460. [online]. https://doi.org/10.1093/mind/LIX.236.433

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. [online]. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

United Nations. (2024). High-Level Advisory Body on Artificial Intelligence. [online]. https://www.un.org/digital-emerging-technologies/ai-advisory-body

US Department of Health and Human Service. (2023). Testimony from Micky Tripathi, PH.D., M.P.P. on Artificial Intelligence before House Committee on Energy and Commerce. [online]. https://www.hhs.gov/about/agencies/asl/testimony/2023/12/13/artificial-intelligence.html

US Exec. Order No. 14110. (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Vallance. C. (2023.3.30.). Elon Musk among experts urging a halt to AI training. BBC. [online]. https://www.bbc.com/news/technology-65110030.

Van Bekkum, M., & Borgesius, F. Z. (2021). **Digital welfare fraud detection and the Dutch SyRI judgment.** European Journal of Social Security., 23(4), pp.323-340.

Vaswani. A. et al. (2017). **Attention is all you need.** [online]. https://doi.org/10.48550/arXiv.1706.03762

Véliz. C. (2024). Comments on weakened SB1047 bill. Linkedin. [online]. https://www.linkedin.com/posts/carissa-v%C3%A9liz-a5781555_california-weakens-bill-to-prevent-ai-disasters-activity-7231193238743715840-8Oz9

Victor J. & Holmes A. (2023). OpenAI Dropped Work on New 'Arrakis' AI Model in Rare SetBack. the Information. [online]. https://www.theinformation.com/articles/openai-dropped-work-on-new-arrakis-ai-model-in-rare-setback

Vranken, B. (2023). Big Tech lobbying is derailing the AI Act. Social Europe. [online]. https://www.socialeurope.eu/big-tech-lobbying-is-derailing-the-ai-act

Weizenbaum, J. (1966). **ELIZA—a computer program for the study of natural language communication between man and machine.** Communications of the ACM, 9(1), pp.36-45.

White House. (2022). **Blue Prints for an AI Bill of Rights**. Washington: White House.

World Economic Forum. (2023). **The Presidio Recommendations on Responsible Generative AI**. Geneva: World Economic Forum.

Yampolskiy, V. (2015). **Analysis of types of self-improving software. Artificial General Intelligence.** Springer International Publishing,

pp. 384-393.

Zaber, M., Casu, O., Brodersohn, E. (2024). **Artificial Intelligence in social security organizations**. International Social Security Association.

[부록] 해외 인공지능 규제 정리

〈부록〉 해외 인공지능 규제 정리

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
World Economic Forum	Presidio Recommendations on Responsible Generative AI(2023)	정확하고 공유 가능한 용어 설정 -AI 기능 및 한계에 대한 대중의 인식 제고 인간의 가치와 선호에 집중 -조정 및 참여 독려 -새로운 지표와 표준을 탐색하는 동시에 엄격한 벤치마크 및 사용 사례별 테스트를 통해 AI 책임성 유지 -잠재적인 약점, 취약점 및 개선이 필요한 영역을 파악하기 위해 비판적으로 분석하는 방법인 레드팀을 구성하여 모델 설계부터 출시까지 적용 -투명한 출시 전략 채택 -사용자 피드백 활성화		-정책 및 자문 기구 -정책, 규제, 법 집행 기구 -지원 기구 -관리 & 규제 기구	개방형 혁신 및 국제 협업 -공공·,민간 연구 협력 장려 -모델, 도구, 벤치마크 및 모범사례의 공통 레지스트리 구축 -책임감 있는 개방형 혁신 및 지식 공유 지원 -AI 표준에 대한 국제 협력 강화 -글로벌 AI 거버넌스 이니셔티브 수립 사회적 진보 -생성형 AI 개발 및 채택에 있어 사회적 진보를 우선함 -사회 전반의 AI 리터러시 촉진 -AI 기반 환경에서 총체적 사고 접근 방식 육성

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
		-모델 및 시스템 추적 기능 포함 -콘텐츠 추적성 보장 -사람 이외의 상호작용 공개 -인간과 AI의 신뢰 구축 -단계별 검토 프로세스 구현 -포괄적인 다단계 측정 프레임 워크 개발 샌드박스 프로세스 도입 -진화하는 창의성과 지적 재산의 환경에 적응하기			-생성형 AI의 혁신적 영향력 활용하기 -사회적 공익을 위한 혁신 장려 -리소스 및 인프라 격차 해결 -정부 내에서 생성형 AI 전문성 촉진 -개발도상국에서 AI에 대한 공평한 접근성 향상 -문화 유산 보존
European Union	AI Act(2024)	EU 내 AI 시스템의 시장 출시, 서비스 제공 및 사용에 대한 규칙 -특정 AI 활용 금지 -고위험 AI 시스템에 대한 구체적인 요구 사항과 해당 시스템 운영자의 의무 규정 -특정 AI 시스템에 대한 투명성 규칙 -범용 AI 모델의 시장 출시에 대한 규칙 -시장모니터링, 시장 감시 거버넌스 및 집행에 관한 규칙		EU 국가 -정책 및 자문기구: 유럽인공지능위원회(AIBoard) -정책, 규제·법 집행 기구: EU집행위원회 ·AI 사무국(AIOffice), 유럽알고리즘투명성센터(ECAT), 유럽개인정보보호감독관(EDPS) ·지원기구: 산업계, 학계, 시민 사회 대표로 구성된 자문포럼 (Advisory forum) 및 과학패널(Scientific panel)	리스트 분류 및 관리체계 -AI 시스템 1. 허용할 수 없는 위험성(unacceptabl erisk): 활용 금지 2. 고위험성(high risk): 다수의 준수사항 3. 제한된 위험성(limited risk): 투명성 의무 4. 최소 위험성(minimal risk): 자율규제 권고 5. 잔존 위험성(residual risk): 고위험성 AI 시스템 리스크 관리 시 고려

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
		-스타트업을 포함한 중소기업에 중점을 둔 혁신 지원 조치 적용 범위 -유럽연합에서 AI 시스템 출시·서비스화 & 범용 AI 모델 출시하는 제공자(소재 불문) -유럽연합 내 AI 시스템 배포자 -AI 시스템에 의해 생산된 결과물이 유럽연합에서 사용되는 경우 AI 시스템의 제공자 및 배포자(소재 불문) -AI 시스템 수입업자 및 유통업자 -제품 제조사업자: 자신의 제품과 함께 자신의 상호나 상표로 AI 시스템 출시·서비스화 -제3국 제공자의 공식 대리인 -유럽연합에 위치한 영향을 받은 사람		회원국 -관리 및 규제: 지정된 적합성 평가 인증기관(notified body)이 담당하며 통보기관이 지정함. -통보기관(notifying authority) ·적합성 평가 제도 운영 및 관리 ·적합성 평가 인증기관 지정 및 등록, 신고(reporting) 업무 등 모니터링 ·한 개 이상의 기관 설치 또는 지정 의무화 시장감독기구 (Market surveillance authority, MSA) ·법 집행기관: 1~25명의 직원으로 구성할 것을 제안 -한 개 이상의 기관 설치 또는 지정의 무화 ·개별 부처나 부문별 규제기관을 MSA로 지정 가능	범용 AI 모델 1. 일반 GPAI 모델 2. 시스템 위험성(systemic risk) 모델 금지 행위 -인간의 안전, 건강, 기본권에 대하여 명백한 위협이 되는 AI 시스템으로, 합법적으로 허가받는 연구목적의 개발 및 운영을 제외하고 전면 금지 위험성 기반 AI 시스템 차등 규제 존재 -고위험성 AI 시스템 ·부록 I 유형 제품의 안전 부품으로 사용되는 AI 시스템, 그 자체가 하나의 제품인 AI 시스템, 출시하기 전각 회원국 정부가 지정한 전문기관에서(제3자) 적합성 평가를 거쳐야 하는 AI 시스템은 연합 법제를 적용 각 규제 기관의 부문별 규제로 통합 예정이며, 출시

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
					전 적합성 평가 필요(제3자 평가) ·부록 III 유형 자연인에 대한 프로파일링을 수행하는 고위험성 AI 시스템 -고위험성군에서 제외 ① 제한된 절차적 작업 수행 ② 이미 완료된 인간 활동의 결과 개선 ③ 의사결정 패턴 또는 이전 의사결정 패턴과의 편차 감지 + 사람의 검토 없이 이미 완료된 사람의 평가를 대체하거나 그에 영향을 미치기 위한 것이 아닌 경우 ④ 부록 III에 열거된 이용 사례의 목적과 관련된 평가에 대한 준비 작업 수행 -자연인의 건강, 안전, 기본권에 위해를 끼칠 리스크가 존재할 경우 준수사항 자발적 준수 권고, 존재하

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
					지 않을 경우 EU 데이터베이스 등록 -고위험성 AI 시스템의 준수사항 ① 위험성 관리체계(risk management system) ② 데이터 및 데이터 거버넌스 ③ 기술 문서 ④ 기록 보관 ⑤ 배포자에 대한 투명성 및 정보 제공 ⑥ 사람의 감독 ⑦ 정확성, 견고성 및 사이버 보안 -제한된 위험성, 특정 AI 시스템 제공자와 배포자는 투명성 의무를 가짐. -범용 AI 모델 규제 1. 모델과 시스템을 구분한다. 2. 모델은 성능 및 리스크 기반의 이원적 규제체계

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
					적용 3. 시스템은 이 법에 따른 리스크 기반 규제 체계 적용 4. 강화된 투명성 의무 적용 5. EU 집행위원회가 전담 6. 짧은 규제 준수 준비시간 허용
OECD	AI Principles (2023)	규정 -가치 기반 원칙 -포용적 성장, 지속 가능한 개발 및 웰빙 -공정성과 사생활을 포함한 인권과 민주주의 가치 -투명성과 설명 가능성 -견고성, 보안성, 안정성 -책임		정책 활동 영역 -정책입안자를 위한 권장 사항 -AI 연구 및 개발에 투자 -포괄적인 AI 지원 생태계 육성 -AI를 위한 상호 운용 가능한 거버넌스 및 정책 환경 형성 -인적 역량 구축 및 노동시장 전환 준비 -신뢰할 수 있는 AI를 위한 국제협력 모니터링&평가 -AI 시스템 수명 주기 -설계, 데이터 및 모델은 계획 및 설계, 데이터 수집 및 처리, 모델 구축을	OECD AI 원칙 및 관련 도구 사용정책 수립하고, AI 위험 프레임워크를 만들어 관할권 간 글로벌 상호운용성을 위한 기반 구축

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
				포함하는 맥락에 따라 달라짐 -검증 -배포 -운영 및 모니터링 반복적인 방식으로 진행되며 반드시 순차적이지 않음. AI 시스템 퇴역은 언제든 가능	
UNESCO	Recommendation on the Ethics of AI(2021)	규정 -가치와 원칙은 AI 시스템 수명 주기의 모든 행위자가 우선적으로 존중해야 하며, 기존 법률, 규정 및 비즈니스 지침으로 개정과 새로운 지침의 정교화를 통해 추진되어야 함. 가치 -인권과 기본적 자유 및 인간의 존엄성에 대한 존중, 보호 및 증진 -환경 및 생태계 번영 -다양성과 포용성 보장 -평화롭고 정의로우며 서로 연결된 사회	AI 시스템 수명 주기에 걸쳐 법률 및 규제 프레임워크를 개발하고 책임을 촉진할 책임이 있는 당국이자 행위자로 회원국을 대상으로 함. 모든 AI 행위자에게 윤리적 지침을 제공하여 AI 시스템의 수명 주기 전반에 걸친 윤리적 영향 평가에 대한 근거 제공 모니터링&평가 회원국은 구체적 여건, 통치구조 및 헌법 조항에 따라 양적, 질적 접근 방식을 조합하여 AI 윤리와 관련된 정책, 프로그램 및 메커니즘을 신뢰성 있고 투명하게	정책 활동 영역 -윤리적 영향평가 -윤리적 거버넌스 및 관리 -데이터 정책 -개발 및 국제협력 -환경 및 생태계 -성별 -문화 -교육 및 연구 -커뮤니케이션 및 정보 -경제 및 노동 -건강 및 사회복지	유네스코 권한 범위 내에서 인공지능 영역과 관련된 윤리적 문제 -AI에 대한 단일 정의 제공 의도가 없으며 윤리적 관련성이 높은 AI 시스템 특징을 다룸 -AI 시스템을 지능적 행동과 유사한 방식으로 데이터와 정보를 처리할 수 있는 시스템으로 접근하며, 일반적으로 추론, 학습, 지각, 예측, 계획, 제도를 포함 -AI 시스템과 관련된 윤리적 문제는 AI 시스템 수명 주기의 모든 단계와 관련 있음. AI 행위자는 AI 시

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
		원칙 -비례성 및 무해성 -안전 및 보안 -공정성 및 차별 금지 -지속 가능성 -개인정보 보호 및 데이터 보호에 대한 권리 -사람의 감독과 결단력 -투명성 및 설명 가능성 -책임과 의무 -인식 및 문해력 -다중 이해관계자 및 적응형 거버넌스 및 협업	모니터링하고 평가해야 함. 모니터링 및 평가 프로세스는 취약한 사람이나, 취약한 상황에 처한 사람을 포함하되 이에 국한되지 않는 모든 이해관계자의 폭넓은 참여를 보장		스템 수명 주기의 최소한 단계 이상에 관여하는 모든 행위자로 정의 -AI 시스템은 인권과 기본적 자유에 미치는 영향을 포함하되 이에 국한되지 않는 새로운 유형의 윤리적 문제를 제기함. AI 알고리즘이 기존의 편견을 재생산하고 강화하여 차별, 편견, 고정관념을 악화시킬 수 있는 잠재력으로 인해 새로운 윤리적 도전 발생 -디지털화 사회에서 살아가려면 노동시장, 고용 가능성 및 시민 참여에 미치는 영향을 고려할 때 교육이 필수적임. 새로운 연구역량과 접근 방식을 가져오고 과학적 이해와 설명의 개념에 영향을 미치며 의사결정을 위한 새로운 기반 만듦.

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
					-AI 기술이 정보의 처리, 구조화 및 제공에 중요한 역할을 하게 되면서 커뮤니케이션과 정보, 동화된 저널리즘, 소셜미디어와 검색 엔진에서 뉴스의 알고리즘 제공 등의 문제 제기
White House	Blue Prints for an AI Bill of Rights (2022)	AI 권리장전 청사진은 정부와 민간 부문이 원칙을 실천할 수 있도록 지원하기 위함. -프레임워크는 자동화된 시스템 개발 및 사용을 위한 권장원칙에 대한 광범위하고 미래지향적인 비전을 공유하여 시스템이 권리, 기회 또는 접근에 의미 있는 영향을 미칠 가능성이 있는 민간 및 공공의 참여를 알리는 데 사용 -프레임워크는 지자체, 주, 연방정부 또는 기타 국가의 입법 및 규제 제안을 분석하거나, 입장을 취하지 않음.	AI 권리장전 청사진은 자동화 시스템의 영향을 받는 모든 사람과 자동화 시스템 사용을 규제하는 정책을 개발, 설계, 배포, 평가 또는 수립하는 모든 사람을 포함하여 다양한 상황에서 참고할 수 있도록 의도	5가지 원칙 -안전하고 효과적인 시스템 -알고리즘 차별보호 -데이터 개인정보 보호 -공지 및 설명 -인간의 대안, 고려 및 대체 3가지 보충 섹션 1. 원칙이 중요한 이유 2. 자동화 시스템에 대해 기대해야 할 사항 3. 원칙이 실제로 어떻게 적용될 수 있는지 설명	범위 -프레임워크는 두 부분으로 구성된 테스트를 사용 어떤 시스템이 적용되는지 결정. 중요한 리소스 또는 서비스에 대한 미국 대중의 권리, 기회 또는 접근에 의미 있는 영향을 미칠 수 있는 자동화된 시스템에 적용됨. 권리 -언론, 투표의 자유, 차별로부터 보호, 과도한 처벌, 불법 감시, 공공 및 민간 부문 사생활 및 기타 자유 침해를 포함한 시민권, 시민의 자유 및 프라이버시

제안자	법안 및 제안 일자	인공지능 개발 및 이용의 기본원칙	인공지능 기본계획	거버넌스	주요 논의 사항
		-AI 개발 회사와 연구자는 AI 및 기타 자동화 시스템의 윤리적 사용을 원칙으로 함. -원칙은 공정 정보관행 원칙(FIPP)과 밀접한 관련 있음.			기회 -교육, 주택, 신용, 고용 및 기타 프로그램에 대한 공평한 접근을 포함한 동등한 기회 접근 -의료, 금융서비스, 안전, 사회서비스, 상품 및 서비스에 대한 기만적이지 않은 정보, 정부 혜택 같은 중요한 리소스 또는 서비스에 대한 접근성

Abstract

Analysis of Artificial Intelligence Technology Application in Social Welfare Administration and its Policy Implications

Project Head: Kim Ki-tae

Social security is one of the areas where artificial intelligence (AI) technology is most actively applied. Chapter 2 of this report examines the trends in AI development and presents ethical principles for its application. Chapter 3 reviews domestic and international institutional applications. In South Korea, a survey was conducted on welfare technology applications in the social security sector over the past five years (October 2019 to September 2024), using data from the Procurement Information Open Portal and Narajangteo. This survey identified 36 projects. Similarly, international examples reviewed in Section 2 of Chapter 3 indicate that relatively basic AI technologies, such as chatbots, are being utilized.

Chapter 4 explores regulatory trends concerning AI technologies in the social security domain both domestically and abroad. The analysis confirmed that South Korea lacks appropriate regulations for this sector. International regulatory trends were examined with a focus on the European Union's Artificial Intelligence Act and U.S. Executive Order 14110 issued by the Biden Administration.

Co-Researchers: Shin Young-kyu · Kim Myuhng-joo · Kim Eun-ha · Byeon So-yeon

Finally, the report identifies ten areas where AI technology is applied within social security, analyzes its potential benefits and risks—categorized into seven aspects—and provides eight policy recommendations to both support and regulate AI applications in social security.

Key words: artificial intelligence, social security, social welfare, social policy, big data