정책연구 25-26

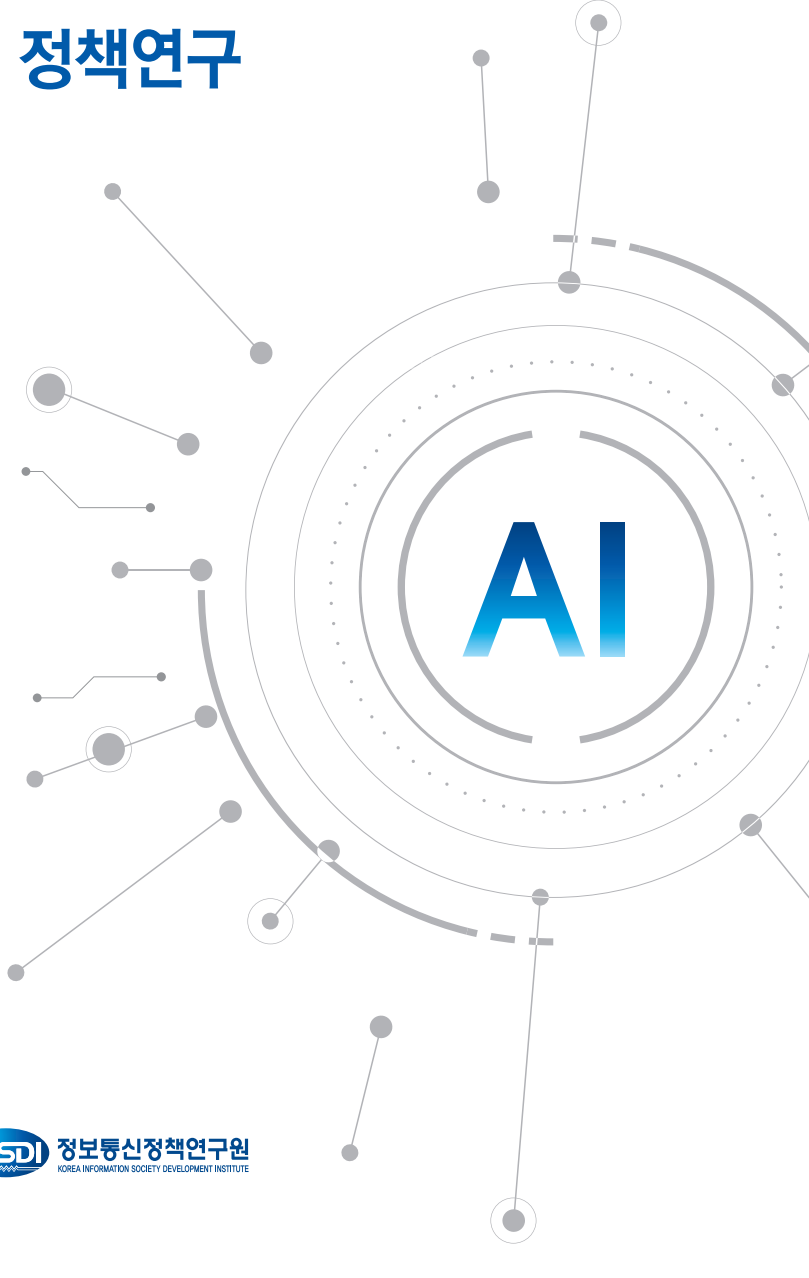
ai.kisdi.re.kr

AI 윤리 확산 기반조성

AI 윤리 확보를 위한 실천 방안 및 정책연구

문정욱 외

2025. 12



과학기술정보통신부
Ministry of Science and ICT

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

정책연구 25-26

AI 윤리 확산 기반조성

AI 윤리 확보를 위한 실천 방안 및 정책연구

문정욱 외

2025. 12

과학기술정보통신부
Ministry of Science and ICT

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

이 보고서는 2025년도 과학기술정보통신부 일반회계 디지털 질서 기반 구축 및 글로벌 확산 지원 사업(안전한 AI 활용 기반조성)의 연구결과로서 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

제 출 문

과학기술정보통신부 장관 귀하

본 보고서를 『안전한 AI 활용 기반조성』 사업의 결과보고서로 제출합니다.

2025년 12월

연 구 기 관: 정보통신정책연구원
총괄책임자: 문정욱 실 장
참여연구원: 조성은 연 구 위 원
이현경 연 구 위 원
연소라 부연구위원
김휘홍 부연구위원
문아람 연 구 위 원
양기문 전문연구원
김희연 부연구위원
이시직 부연구위원
김지혜 전문연구원
강하연 연 구 원
김소담 연 구 원
권지혜 연 구 원
전민경 연 구 원

목 차

표 목 차

그 림 목 차

요 약 문

1. 제목

AI 윤리 확보를 위한 실천 방안 및 정책연구

2. 연구 목적 및 필요성

인공지능(AI)은 이제 산업적 도구를 넘어, 일상생활은 물론 국가 안보와 민주주의 시스템 전반에까지 영향을 미치는 핵심 기술로 확산되고 있다. 특히 초거대 AI의 급속한 발전은 기술적 안전장치와 사회적·제도적 대응 간의 간극을 확대시키며, 윤리와 법·제도가 기술 발전 속도를 충분히 따라가지 못하는 이른바 '지체현상'에 대한 우려를 증폭시키고 있다. 2023년 발표된 국제 공개서한이 AI 개발에 대한 신중한 접근을 촉구한 것도 이러한 맥락에서 이해할 수 있다. 이와 함께 AI의 탈옥(jail breaking), 환각(hallucination), 보안 취약성 등 기술적 위험과 더불어, 오용·악용 가능성, 책임소재의 불명확성, 사회적 신뢰 저하 등 윤리적 수용성 문제가 주요 정책 과제로 부각되고 있다. 이러한 문제의식 속에서 국제사회는 AI의 위험을 관리하고 신뢰할 수 있는 활용을 도모하기 위한 정책과 제도 마련을 본격화하고 있으며, 우리나라도 2025년 1월 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」을 제정하여 윤리원칙, 민간자율AI윤리위원회, 고영향 AI 영향평가 등 자율성과 책임에 기반한 관리체계를 제도화하였다. 이는 경직된 규제보다는 윤리적 기준과 자율적 이행을 통해 기술 혁신과 사회적 신뢰를 조화시키려는 정책적 전환으로 평가할 수 있다.

본 사업은 이러한 제도적·사회적 전환을 배경으로, AI 윤리영향평가를 수행하고 헬스케어 분야 윤리 자율점검표의 개발·현장 적용을 지원함으로써 책임 있고

신뢰할 수 있는 AI 활용 기반을 강화하는 것을 목적으로 추진되었다. 특히 AI 채용 서비스를 대상으로 전문가평가단과 국민포럼단이 참여하는 다층적 윤리영향평가를 실시하고, 개발기업과의 협업을 통해 현장 적용 가능한 맞춤형 점검표를 개발함으로써, 「인공지능(AI) 윤리기준」이 선언적 원칙을 넘어 실천 중심의 정책 수단으로 기능할 수 있음을 실증적으로 제시하고자 하였다.

3. 연구의 구성 및 범위

본 연구는 AI 기술의 활용 및 확산에 따른 잠재적 위험과 부작용을 예방하는 동시에, 올바른 AI의 개발·활용을 장려하기 위한 자율규제 기반의 AI 윤리체계 정립을 목표로 수행되었다. 보다 구체적으로는 첫째, 전년도에 이어 AI 채용 서비스를 대상으로 'AI 윤리영향평가'를 시범적으로 적용하고, 평가 절차와 방법론을 개선하여 서비스의 편익과 위험을 식별하고 합리적으로 관리하기 위한 개선방안을 도출하였다. 둘째, 헬스케어 분야에서의 AI 윤리기준 자율점검표를 개발하고 이를 현장에 적용하였다. 셋째, 2023년 11월 구축한 AI 윤리 소통채널을 지속 운영하고 기능을 고도화하였다.

4. 연구 내용 및 결과

첫째, 「AI 윤리영향평가 프레임워크」를 기반으로 AI 채용 서비스의 윤리적 영향을 평가하였다. 평가대상은 전문가 검토와 과학기술정보통신부 협의를 거쳐 선정하였으며, 이후 1,500명 대상 대국민 인식조사와 서비스 분석을 통해 주요 쟁점사항을 도출하고 평가 수행을 위한 기초자료를 마련하였다. 이어 학계·산업계·법조계·공공·시민사회·국제기구 등 다양한 분야 전문가로 구성된 '전문가평가단'과 일반 국민이 참여한 '국민포럼단(FGI)'이 프라이버시 보호, 포용성, 책임성, 투명성, 공정성 등 5개 윤리영역에 대해 정량·정성 평가를 수행하였다. 또한 2025년

11월 27일 개최한 ‘2025 AI 윤리 공개세미나’에서 평가 추진 경과와 중간 결과를 공유하며 국민 의견을 폭넓게 청취하였다.

둘째, 2022년 2월 처음 공개하고 매년 개정 중인 「인공지능 윤리기준 실천을 위한 자율점검표(안)」을 헬스케어 분야에 맞게 변형한 분야별 점검표를 개발하는 한편, 기업의 자율적 윤리실천을 지원하기 위해 에이아이트릭스(AITRICS)와 협업하여 기업 맞춤형 윤리점검표를 마련하였다.

셋째, ‘AI 윤리 소통채널’을 상시 운영하며 AI 윤리 관련 사업성과, 정책 자료, 논의 결과 등을 체계적으로 아카이빙함으로써 AI 윤리정책의 지속적 발전과 사회적 확산을 지원하였다. 동시에 새로운 윤리적 이슈를 논의·공유하는 참여 플랫폼으로서의 기능도 강화하였다.

5. 기대효과

본 연구는 AI 윤리영향평가를 통한 합리적 관리 기반 제공과 자율점검표 개발·적용을 통한 자율적 관리체계 확산으로 AI 윤리 실천 기반을 조성하고, AI에 대한 사회적 신뢰 수준을 제고함으로써 사람과 AI가 공존할 수 있는 안전한 사회 구현에 기여한다. 특히 국제적 정합성을 갖춘 평가 방법론을 기반으로 산업계·학계·법조계·공공·시민단체 전문가, 일반시민 등 다양한 이해관계자가 참여한 영향평가 수행과 기업 협업을 통해 현장 상황을 반영한 맞춤형 점검표를 개발하였다. 이러한 성과는 AI 기본법에 따른 윤리원칙 제정 및 실천수단 확산을 위한 실제적 사례를 제공하고 국제적 정책 흐름에 부합하는 자율규제 환경 조성에도 기여한다. 나아가 AI 윤리 소통채널을 통한 사업성과 공개와 정책 자료 확산은 AI 윤리에 대한 국민 인식을 제고하고, AI 기술이 인간 중심의 가치를 유지하면서도 그 잠재력을 발휘할 수 있는 기반을 마련할 것이다. 이를 통해 AI 기술의 긍정적 혁신과 윤리적 위험 관리가 균형을 이루며, 책임 있는 AI 활용이 사회 전반으로 확산될 것이다.

제 1 장 서 론

제 1 절 연구의 필요성 및 목적

인공지능(AI)은 이제 단순한 산업적 도구의 범위를 넘어 우리의 일상생활은 물론 국가 안보와 민주주의 시스템 체계 전반에까지 영향을 미치며 점차 깊숙이 스며들고 있다(문정욱, 2025). 2023년에는 초거대 AI의 안전성과 사회적 영향을 우려하며, AI 개발의 속도와 방향에 대한 재검토를 촉구하는 공개서한이 발표되었다. 요슈아 벤지오, 스튜어트 러셀 등 주요 AI 연구자들이 참여한 이 공개서한은 “준비되지 않은 가을을 서두르지 말자(“Let’s enjoy a long AI summer, not rush unprepared into a fall”)”는 메시지를 통해 초거대 AI 개발에 대한 신중한 접근과 안전 논의의 필요성을 강조하였다. 이는 그간 일상에서 경험해 보지 못한 AI 기술 발전 속도를 윤리와 법·제도가 따라가지 못함에 따른 이른바 ‘지체현상’에 대한 우려에서 비롯되었다. 이러한 우려를 최소화하기 위해 해외 주요 국가들과 국제기구들은 AI 윤리, 신뢰, 안전과 관련된 다양한 정책과 제도, 윤리 실천수단을 개발하여 안전한 AI 개발 및 활용 기반을 조성하고 있다.

우리나라도 이러한 국제적 흐름에 발맞추어 책임 있고 윤리적인 AI 개발 및 활용을 목표로 범국가 차원의 ‘사람이 중심이 되는 인공지능 윤리기준’을 수립하였고, 이후 자율점검표, 윤리영향평가, 교육콘텐츠, 소통채널, 기본권 영향평가 제도 개발 등 다양한 차원에서 윤리기준을 실천하고 확산하기 위한 정책적 노력을 기울이고 있다. 2025년 1월 제정된 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(AI 기본법)」은 이러한 정책 흐름을 제도적으로 구체화한 전환점으로 볼 수 있다. AI 기본법은 본 연구와 밀접하게 연관되는 제도적 장치로서, 윤리원칙의 제정·공표(제27조), 민간자율인공지능윤리위원회 설치(제28조),

고영향 AI 영향평가(제35조) 등을 규정함으로써, 일률적 규제보다는 자율성과 책임에 기반한 관리체계를 지향한다는 점에서 국제적 정책 흐름과도 부합한다.

AI 대전환에 따른 AI 일상화가 가속화 되면서 우리가 직면한 이슈와 쟁점, 그리고 잠재적 위험은 크게 기술적 차원과 사회적 차원으로 구분해 볼 수 있다. 우선 기술적 차원에서는 생성형 AI가 보안 가드레일을 우회하는 탈옥(jail breaking)이나 사실이 아닌 정보를 그럴듯하게 보여주는 환각(hallucination) 현상이 여전히 풀기 어려운 문제로 지적되고 있다. 이러한 문제는 기존 기술적 대응 방식 중 하나인 사후적 보안 패치나 버그 바운티 등으로는 근본적 해결이 어렵다. 또한 영상 합성 등 AI 생성 기술의 발전 속도가 탐지 기술을 앞지르면서 AI 생성 결과물에 대한 식별 근거를 확보하기 어려워지고 있어, 기술적 방어의 한계가 보다 뚜렷해지고 있다. 더욱이 혹시라도 AI 모델 자체가 해킹되는 경우에는 개인정보를 포함한 민감한 학습데이터가 외부로 유출될 수 있는 보안 위험까지 초래할 가능성도 있다. 사회적 차원을 살펴보면 우리가 직면한 가장 큰 문제점은 '윤리적 수용성'이다. 2024년 정보통신정책연구원(KISDI)에서 수행했던 AI 윤리영향평가 결과에서 확인할 수 있듯이, AI 영상합성 서비스 즉 딥페이크 기술의 활용 확산은 창작 활성화, 비용 절감, 표현의 자유 확대, 소수 계층의 효과적인 의사 전달이라는 혁신의 빛을 비추는 동시에 성범죄 악용, 가짜뉴스 생성 및 유포, 불법 콘텐츠 유포라는 오용과 악용의 그림자를 동시에 드리우고 있다. 여기서 주목해야 할 점은 시민들이 AI 활용 과정에서 느끼는 우려는 단순한 AI 기술 자체의 결함이나 미완결성이 아니라, 통제 불가능한 악의적 활용 가능성과 책임소재의 불명확성에서 비롯된다는 것이다. 즉 아무리 기술적으로 견고하고 안전한 기제를 마련한다고 하더라도 사회 구성원들이 이를 윤리적으로 활용하지 않거나 용인하고 신뢰하지 않는다면 그 기술은 혁신과 성장의 동력이 아닌 갈등과 위험의 원인으로 작동할 가능성이 높다. 결국 기술적 안전은 사회적 신뢰를 확보하기 위한 필수조건이며 윤리적 수용은 기술이 시장과 사회에 안정적으로 정착하기 위한 충분조건이 될 수 있을 것이다. 이러한 두 가지 요소를

모두 충족시켜 신뢰할 수 있는 안전을 구현한다면, 국가 차원의 AI 경쟁력 제고(AI 3대국 도약)로도 연결될 수 있을 것이다.

이러한 문제의식에 본 연구의 출발점이 있다. 즉 AI 활용 확산에 따라 수반되는 부정적 영향을 최소화하거나 사전에 방지하기 위해서는 무엇보다 책임감을 가진 올바른 AI 활용이 전제가 되어야 할 것이다. 이를 위해서는 신뢰할 수 있는 AI 기반 제품 및 서비스가 제공되고 동시에 윤리적으로 활용할 수 있는 체계와 함께 구체적인 실천 수단을 마련할 필요가 있다. 이에 본 연구는 AI 기술과 서비스가 사회 전반에 미치는 영향을 윤리적 시각에서 분석하고, 이를 토대로 정책적 시사점을 제시하는 것을 목표로 한다. 더 나아가 윤리적 위험을 사전에 예방하는 동시에 기술 발전을 뒷받침할 수 있는 실천적 수단으로서 자율점검표 개발에 주안점을 두었다. 아울러 연구 성과를 일반 국민에게 공개하고 지속적으로 새롭게 제기되는 AI 윤리 및 신뢰 관련 쟁점을 논의하는 공론의 장으로서 윤리소통채널 운영 방안도 함께 모색하고자 한다.

제 2 절 연구 추진체계와 전략

본 연구는 AI 기술의 확산에 따라 발생할 수 있는 사회적·윤리적 위험을 선제적으로 관리하고, 책임 있는 AI 활용 문화 정착과 자율규제 기반 마련에 중점을 두고 추진되었다.[1] 이를 위해 2025년에는 AI 채용 서비스를 대상으로 한 윤리영향평가, 헬스케어 분야 자율점검표 마련 및 기업 윤리점검표 개발 지원, 국민 소통 기반 강화 등 세 가지 중점 과제를 중심으로 연구를 수행하였다.

첫째, AI 채용 서비스가 미치는 긍정적·부정적 윤리영향을 체계적으로 파악하기

1) 정보통신정책연구원은 2020년 국가「인공지능(AI) 윤리기준」마련을 지원하고, 이후 AI 윤리 실천 수단 및 정책에 관한 연구를 지속적으로 수행해오고 있음(문정욱 외, 2024)

위해 'AI 윤리영향평가'를 수행하고, 평가 결과를 정책적·실무적 개선 방안으로 연계할 수 있도록 시사점을 도출하였다. 특히 2024년 개발된 프레임워크(과학기술정보통신부·정보통신정책연구원, 2024)와 국내외 사례 연구를 기반으로 분석 범위와 구조를 재설계하였으며, 이에 따라 2025년 AI 채용 서비스에 대한 윤리영향평가를 수행하고, 2025년 11월 27일 개최된 'AI 윤리 공개 세미나'에서 그 중간 결과를 공개하였다. 영향평가 최종 결과는 이해관계자 논의를 통해 보완한 후, 기업·시민사회·학계·공공에서 쉽게 활용할 수 있도록 책자 형태의 보고서로 공개할 예정이다. 둘째, 「인공지능 윤리기준 실천을 위한 자율점검표(안)」을 기반으로 헬스케어 분야에 적용 가능한 점검체계를 설계하고, 현장 적용성을 높이기 위해 기업 맞춤형 윤리점검표 개발을 지원하였다. 셋째, 2023년 11월 개설된 'AI 윤리 소통채널'을 지속적으로 운영하며, 사업성과 및 정책 자료의 체계적 아카이빙 기능을 강화하고 국민 참여와 의견수렴을 위한 상시 소통 기반을 확립하였다.

본 연구는 과학기술정보통신부의 지원을 받아 정보통신정책연구원이 수행하였으며, 연구의 실효성을 높이기 위해 학계·산업계·법조계·시민단체 등 다양한 이해관계자의 의견을 수렴하고 이를 연구·정책 설계 과정에 적극 반영하였다.

[그림 1-1] 추진체계

자료: 연구진 작성

제 2 장 AI 윤리영향평가 시행

제 1 절 AI 윤리영향평가 개요

1. 평가 개요

가. 추진 배경

최근 국내외에서 AI 기술의 활용이 확대됨에 따라, AI 윤리성·신뢰성·안전성을 확보하기 위한 영향평가 도입 필요성이 지속적으로 제기되고 있다. 특히 국제사회에서는 AI 윤리·안전 규제 논의가 본격화되고 있으며, 각국은 자국의 제도 환경에 맞는 AI 영향평가 체계를 모색하고 있다.

UNESCO는 2021년 11월 발표한 'AI 윤리권고(Recommendation on the Ethics of Artificial Intelligence)'를 통해 회원국에게 AI 시스템의 편익·위험 식별, 위험 예방 및 모니터링 조치를 포함한 윤리적 영향평가 도입 및 이행을 촉구하였다. 또한 2023년 8월에는 이를 지원하기 위한 '윤리적 영향평가: AI 윤리권고 도구(Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence)'를 발표하였다. 국가인권위원회 또한 2022년 'AI 개발과 활용에 관한 인권 가이드라인'에서 AI 인권영향평가 제도 마련·시행을 권고하였다. 이후 2024년 5월 과학기술정보통신부에 'AI 인권영향평가 도구 적용'을 요청했으며, 2025년에는 해당 도구의 활용 및 정책 반영 여부에 대한 이행 실태 점검을 요구하는 등 관련 제도 마련을 지속적으로 요구하고 있는 상황이다.

2022년 당시 영향평가 제도들[2]은 AI 기술 특유의 데이터 기반 위험, 알고리즘 편향 가능성, 자동화된 의사결정 과정의 투명성·책임성 문제 등 국제적으로 요구되는 윤리·인권 기반 위험 식별을 충분히 반영하기 어려운 구조적 한계가 존재했다. 해당 평가들은 본래 지능정보서비스의 사회적·경제적 파급력 측정 또는 기술 변화의 광범위한 영향 파악을 목적으로 설계되었기 때문에, AI가 실제 사회·개인에 미치는 윤리적·기본권적 영향을 체계적으로 평가하는 데에는 제도적 제약이 있었다.

국제적 규제 환경 또한 빠르게 변화하고 있다. 2025년부터 단계적으로 시행 중인 유럽연합(EU) AI Act에서도 확인된다. EU는 기본권 보호를 강화하기 위해 특정 고위험 AI 시스템 배포자에게 해당 시스템을 실제로 사용하기 전 기본권 영향평가(Fundamental Rights Impact Assessment, FRIA)를 수행하도록 의무화하였다. 기본권 영향평가는 고위험 AI 시스템이 영향을 미칠 가능성이 있는 개인 또는 집단의 기본권 침해 위험을 사전에 식별하고, 위험이 발생할 경우 취해야 할 조치와 인적 감독 체계 등을 명확히 점검하도록 요구하는 제도이다. 특히 신용평가, 생명·건강보험 등 개인의 권리에 직접적 영향을 미칠 수 있는 분야를 포함하여, 고위험 AI 시스템의 배포 단계에서 기본권 보호를 제도적으로 활보할 수 있는 장치를 마련한 것으로 평가할 수 있다. 국내에서도 2025년 1월 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」이 제정됨에 따라, AI 영향평가 제도화의 법적 기반이 마련되었다(제35조).

나. AI 윤리영향평가 프레임워크

국내외 제도환경 변화 속에서 요구되는 윤리·인권 기반 AI 영향평가 체계를 마련하기 위해, 과학기술정보통신부와 정보통신정책연구원은 2023년부터 「AI 윤리영향평가 프레임워크」를 개발하여 2024년 4월 공개하였다. 또한 2024년 AI

2) 「과학기술기본법」 제14조 '기술영향평가', 「지능정보화 기본법」 제56조 '지능정보서비스 등의 사회적 영향평가' 등

영상 합성 서비스를 대상으로 시범 평가를 실시하여 프레임워크의 적용 가능성과 평가 체계의 타당성을 검증하였다.

AI 윤리영향평가 프레임워크는 다음의 목적을 갖는다. 첫째, AI 개발·운영 기업이 윤리·신뢰성 확보를 위한 자율적 노력을 수행할 수 있도록 실천 기준을 제공한다. 둘째, AI 제품·서비스의 긍정적·부정적 영향을 사전에 식별·평가하여 위험 예방 및 완화 전략을 도출하도록 지원한다. 셋째, 평가 결과를 공개·공유함으로써 기업·시민사회·학계·정부 등 다양한 주체가 참고할 수 있는 정책·산업적 기반 자료를 제공한다.

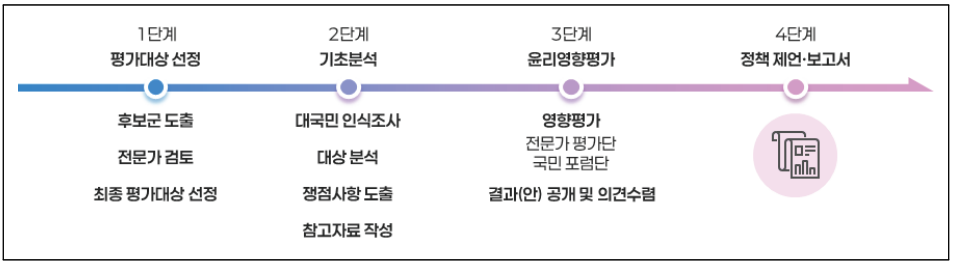
AI 윤리영향평가는 국가 「인공지능 윤리기준」 10대 핵심요건[3]을 기반으로 ① 요건별 긍정·부정 영향 식별, ② 영향의 규모·범위·발생 가능성 분석, ③ 종합 평가 및 개선 방향 도출의 절차를 거쳐 수행된다. 또한 사회적 관심, 기술 동향, 정책적 필요성 등을 고려하여 매년 평가 대상을 선정하며, 산·관·학·연 전문가와 시민사회가 참여함으로써 전문성·공정성·객관성·신뢰성을 확보하고 있다.

2. 추진체계

정보통신정책연구원은 과학기술정보통신부로부터 매년 AI 윤리영향평가 사업을 수탁하여 수행하고 있으며, 2025년 AI 윤리영향평가는 다음에 제시된 절차에 따라 추진되었다. 이어지는 절에서는 평가대상 선정을 시작으로 대국민 인식조사, 본 평가 수행, 그리고 도출된 AI 채용 서비스의 윤리 확보를 위한 주체별 역할과 과제에 관한 내용을 자세히 다룰 예정이다.

3) 인공지능 윤리기준 10대 핵심요건: 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성 요건

[그림 2-1] 2025년 AI 윤리영향평가 절차



자료: 연구진 작성

제 2 절 평가 대상 선정

1. 개요

2025년 AI 윤리영향평가의 대상 선정을 위해 2025년 제정된 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 「AI 기본법」)에서 규정한 '고영향 인공지능' 영역을 중심으로 전문가 의견조사를 실시하였다. 산·관·학·연·시민사회 등 다양한 분야에서 참여한 20명의 전문가는 평가의 목적과 시의성을 고려하여 「AI 기본법」이 제시한 11개 '고영향 인공지능' 영역(〈표 2-2〉) 중 윤리영향평가가 시급히 필요한 영역의 우선순위를 선정하였다. 또한, 각 전문가는 1·2·3순위로 선택한 영역을 기준으로 윤리영향평가가 필요하다고 판단되는 구체적 AI 서비스·제품군을 선정 이유와 함께 제안하였다.

〈표 2-1〉 AI 윤리영향평가 대상 선정 의견조사 전문가의 구성

구분	성명	소속
공공	라기원	한국법제연구원 연구위원
	배동석	한국정보통신기술협회 팀장
	유재흥	소프트웨어정책연구소 책임
	이현숙	한국과학창의재단 지역과학문화실 실장
	이희권	한국과학기술평가원 연구위원
	최민석	AI 안전연구소 실장
	최호진	한국행정연구원 선임연구위원
법조계	장준영	법무법인 세종 변호사/AI 센터장
산업계	김동규	NC 문화재단 포용기술사업팀장
	박진현	한국통신사업자연합회 사무국장
	안소영	LG AI연구원 정책수석
	조장래	비트코퍼레이션 고문
시민사회	이지은	참여연대 공익법센터 선임간사
학계	김신	한국외국어대학교 철학과 교수
	문광진	국립목포대학교 법·경찰학부 교수
	윤건	한신대학교 공공인재학부 교수
	이원태	국민대학교 특임교수
	이청호	상명대학교 교양학부 교수
	천현득	서울대학교 과학학과 교수
	최호진	한국과학기술원 전산학부 교수

자료: 연구진 작성

〈표 2-2〉 AI 기본법에 명시된 고영향 인공지능 영역

고영향 인공지능 영역	
가	「에너지법」 제2조제1호에 따른 에너지의 공급
나	「먹는물관리법」 제3조제1호에 따른 먹는물의 생산 공정
다	「보건의료기본법」 제3조제1호에 따른 보건의료의 제공 및 이용체계의 구축·운영
라	「의료기기법」 제2조제1항에 따른 의료기기 및 「디지털의료제품법」 제2조제2호에 따른 디지털의료기기의 개발 및 이용
마	「원자력시설 등의 방호 및 방사능 방재 대책법」 제2조제1항제1호에 따른 핵물질과 같은 항 제2호에 따른 원자력시설의 안전한 관리 및 운영
바	범죄 수사나 체포 업무를 위한 생체인식정보(얼굴·지문·홍채 및 손바닥 정맥 등 개인을 식별할 수 있는 신체적·생리적·행동적 특징에 관한 개인정보를 말한다)의 분석·활용
사	채용, 대출 심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가
아	「교통안전법」 제2조제1호부터 제3호까지에 따른 교통수단, 교통시설, 교통체계의 주요한 작동 및 운영
자	공공서비스 제공에 필요한 자격 확인 및 결정 또는 비용징수 등 국민에게 영향을 미치는 국가, 지방자치단체, 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관 등(이하 "국가기관등"이라 한다)의 의사결정
차	「교육기본법」 제9조제1항에 따른 유아교육·초등교육 및 중등교육에서의 학생 평가
카	그 밖에 사람의 생명·신체의 안전 및 기본권 보호에 중대한 영향을 미치는 영역으로서 대통령령으로 정하는 영역

자료: 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」

2. 의견조사 결과

가. 고영향 인공지능 영역별 우선순위

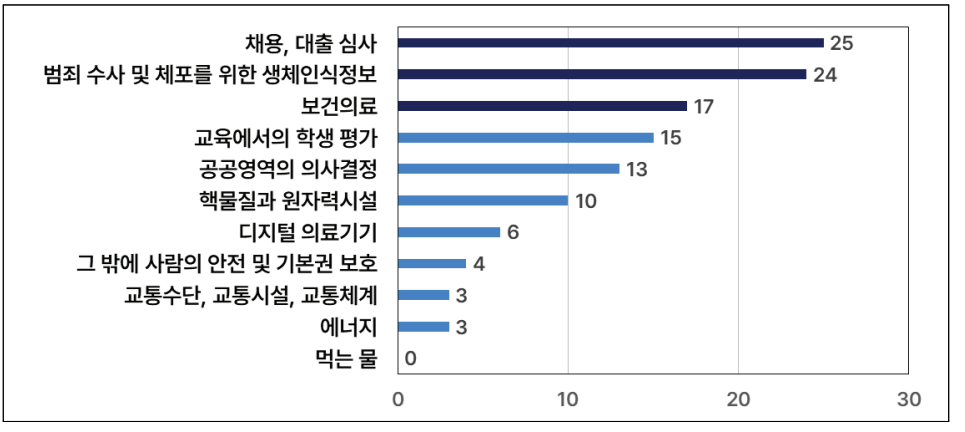
전문가 20인은 고영향 인공지능 11개 영역의 평가 시급성을 판단하여 1·2·3순위 순위를 선정하였고, 순위별로 각각 3점, 2점, 1점의 가중치를 부여해 최종 점수를 산출하였다. 분석 결과, 아래의 〈표 2-3〉과 같이 ① 채용·대출 심사, ② 범죄 수사·체포, ③ 보건·의료 영역 순으로 AI 윤리영향평가의 우선적 시행이 요구되는 것으로 나타났다.

〈표 2-3〉 고영향 인공지능 영역 우선순위 조사 결과

고영향 인공지능 영역		1순위	2순위	3순위	단순 합계	가중치 적용
1	에너지의 공급을 위해 사용되는 AI	1	0	0	1	3
2	먹는 물의 생산 공정에 사용되는 AI	0	0	0	0	0
3	보건의료의 제공 및 이용체계 등에 사용되는 AI	4	2	1	7	17
4	디지털 의료기기 개발 및 이용에 사용되는 AI	1	1	1	3	6
5	핵물질과 원자력시설의 안전 관리 및 운영을 위해 사용되는 AI	2	2	0	4	10
6	범죄 수사나 체포 업무를 위한 생체인식 정보의 분석·활용에 사용되는 AI	4	4	4	12	24
7	채용, 대출 심사 등 개인의 권리·의무 관계에 영향을 미치는 판단·평가 목적의 AI	5	3	4	12	25
8	교통수단, 교통시설, 교통체계의 주요한 작동 및 운영에 사용되는 AI	0	1	1	2	3
9	공공서비스 제공에 필요한 자격 확인 및 결정 또는 비용징수 등 공공영역의 의사결정에 사용되는 AI	1	4	2	7	13
10	교육에서의 학생 평가 등에 사용되는 AI	2	2	5	9	15
11	그 밖에 사람의 생명·신체의 안전 및 기본권 보호에 중대한 영향을 미치는 AI	0	1	2	3	4

주: 1순위에 3점, 2순위에 2점, 3순위에 1점의 가중치 부여
자료: 연구진 작성

[그림 2-2] 고영향 인공지능 영역 우선순위 조사 결과



주: 1순위에 3점, 2순위에 2점, 3순위에 1점의 가중치 부여
자료: 연구진 작성

나. 대상 서비스·제품군 세분화

각 전문가가 제안한 세부 AI 서비스·제품군을 영역별로 정리한 결과는 다음과 같다. ① 채용·대출 심사 영역에서는 '채용'과 '대출 심사'가 주로 제안되었으며, 특히 '채용'은 가장 많은 전문가가 평가 필요 대상으로 선정하였다. ② 범죄 수사·체포 영역에서는 다양한 서비스가 언급되었으나 그 중 '안면 인식'과 '범죄·치안 예측'이 비교적 다수 제안되었다. ③ 보건·의료 영역에서는 '의료 진단'이 가장 많이 제시되었다. 전체 제안 빈도를 집계한 결과, 채용(10회), 안면 인식(6회), 대출 심사(6회), 의료 진단(6회) 순으로 평가 필요성이 높게 나타났다. 자세한 결과는 아래의 〈표 2-4〉, [그림 2-3]에 정리하였다.

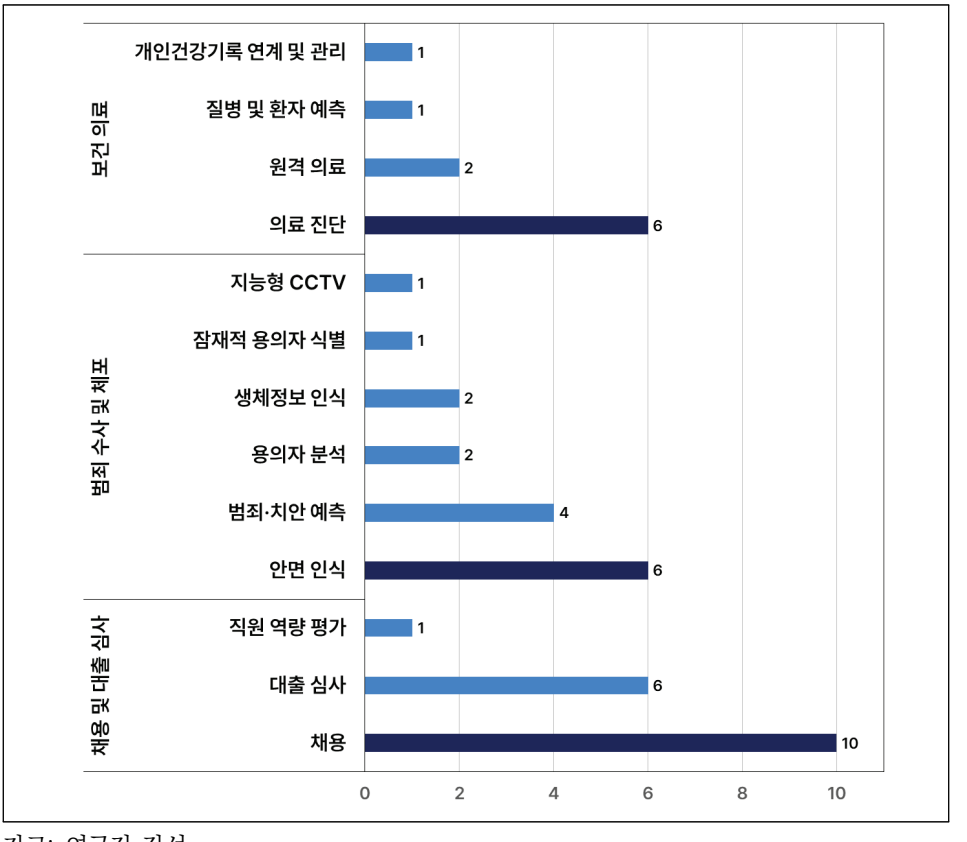
〈표 2-4〉 AI 윤리영향평가 대상 AI 서비스·제품군 조사 결과

영역	세부 AI 서비스·제품군		빈도수
채용·대출 심사	1	채용	10
	2	대출 심사	6
	3	직원 역량 평가	1
범죄 수사·체포	1	안면 인식	6
	2	범죄·치안 예측	4
	3	용의자 분석	2
	4	생체정보 인식	2
	5	잠재적 용의자 식별	1
	6	지능형 CCTV	1
보건·의료	1	의료 진단	6
	2	원격 의료	2
	3	질병 및 환자 예측	1
	4	개인건강기록[4] 연계 및 관리	1

자료: 연구진 작성

4) 개인건강기록(PHR: Personal Health Record): 의료기관에 흩어져 있는 진료·검사 정보와 스마트폰 등으로 수집한 활동량 데이터, 스스로 측정한 체중·혈당 등의 정보를 모두 취합해 사용자 스스로 열람하고 관리할 수 있도록 구축한 건강기록시스템

[그림 2-3] AI 윤리영향평가 대상 AI 서비스·제품군 조사 결과



자료: 연구진 작성

3. 평가 대상 선정

위의 평가대상 선정 절차를 거쳐 2025년 AI 윤리영향평가의 최종 대상으로 'AI 채용 서비스'를 선정하였다. 고영향 인공지능 영역에 대한 전문가 의견조사 결과와 더불어, 사회적 파급력, 정부정책과의 부합성, 중요성, 대응필요성, 시의성 등을 종합적으로 고려하여 본 대상을 채택하였다.

특히 시의성 측면에서 AI 채용 서비스는 기업·공공기관 등 인재 선발 과정 전반에 빠르게 확산되며, 고용 기회와 공정성 및 사회적 신뢰에 직접적인 영향을

미치고 있다. 알고리즘 판단의 불투명성, 편향 데이터 기반 차별의 재생산 가능성, 과도한 개인정보 활용 우려 등의 문제는 지속적으로 제기되고 있으며, 해당 서비스가 「AI 기본법」에서 규정한 고영향 인공지능에 포함됨에 따라 제도적 대응 체계 마련이 시급한 분야로 평가된다.

그러나 위험만 존재하는 것은 아니다. 채용 과정의 일관성 제고, 직무 역량 중심의 평가 강화, 대규모 지원자 대상의 효율적·표준화된 심사 등 AI 채용 서비스가 제공할 수 있는 긍정적 효과 역시 중요한 고려 요소로 작용하였다. 이러한 순기능은 공정하고 합리적인 인재 선발이라는 사회적 목표 달성에 기여할 수 있는 만큼, 윤리적 위험요인을 적절히 관리한다면 사회적 편익을 크게 확장할 수 있을 것으로 기대된다.

또한, AI 채용 서비스로 인해 발생 가능한 윤리적 문제는 개별 기업의 조치만으로 해결되기 어려우며, AI 모델 및 채용 서비스 개발기업-활용기업(채용기관)-플랫폼(구인·구직 서비스)-지원자 등 다수의 이해관계자가 긴밀히 얽혀 있는 구조적 특징을 갖는다. 이는 윤리영향평가의 취지에 부합하며, 생태계 전반의 공동 대응 필요성이 더욱 강조되는 지점이다.

기술적 측면에서도, 최근 AI 채용 서비스는 텍스트 기반 자기소개서, 음성·영상 인터뷰, 온라인 활동 데이터 등 다양한 형태의 정보를 통합적으로 분석하는 멀티모달(Multi-Modal) 방식으로 빠르게 고도화되고 있다. 이와 같은 변화는 단순한 평가 결과의 형태만이 아니라 지원자-시스템 간 상호작용 전반에 구조적 변화를 가져오는 만큼, 윤리영향평가에서 중점적으로 검토해야 할 분야로 판단된다.

제 3 절 AI 채용 서비스 대국민 인식조사

1. 조사 개요

본 절에서는 2025년 AI 윤리영향평가 대상인 AI 채용 서비스에 대한 국민의 인식과 태도를 조사·분석한 결과를 정리하였다. 조사는 AI에 대해 기본적인 인지와 이해 수준을 갖춘 일반 시민 1,500명을 대상으로 2025년 6월 19일부터 6월 27일까지 진행되었다. 표본은 행정안전부 주민등록인구통계를 기준으로 성·연령·지역별 비례할당 방식으로 구성하였으며, 구조화된 설문지를 활용한 온라인 방식으로 수행되었다. 본 조사의 결과는 AI 채용 서비스와 관련한 주요 영향요소를 식별하고, 정책 제언의 방향을 설정하는 데 참고자료로 활용되었다.

〈표 2-5〉 조사 설계

구분	주요 내용
조사 대상	- 전국 만 19세 이상 69세 이하 성인 남녀
표본 수	- 총 1,500명 ※ 행정안전부 주민등록인구통계 기준 성/연령/지역 비례배분
조사 방법	- 구조화된 질문지를 활용한 온라인 조사
조사 기간	- 2025년 6월 19일~6월 27일
조사 내용	- AI 채용 서비스 이용: 인지, 경험, 선호 - AI 채용 서비스 인식: AI 채용 서비스의 정확도·공정성·편향성, 책임 등 - AI 채용 관련 정부 대응과 신뢰 인식

자료: 연구진 작성

〈표 2-6〉 응답자 특성

		사례수	%			사례수	%
전체		1,500	100.0				
성별	남자	765	51.0	거주 지역	서울	280	18.7
	여자	735	49.0		부산	95	6.3
연령	만19-29세	252	16.8		대구	68	4.5
	30대	270	18.0		인천	90	6.0
	40대	308	20.5		광주	41	2.7
	50대	351	23.4		대전	42	2.8
	60대	319	21.3		울산	34	2.3
학력	고졸 이하	271	18.1		경기	408	27.2
	대학교(4년제 미만)	285	19.0		강원	43	2.9
	대학교(4년제 이상)	795	53.0		충북	44	2.9
	대학원 재학 이상	149	9.9		충남	62	4.1
이용 경험	유	312	20.8		전북	49	3.3
	무	1,188	79.2		전남	49	3.3
인지 수준	상	136	9.1		경북	71	4.7
	중	732	48.8		경남	92	6.1
	하	632	42.1		제주	20	1.3
직업	관리자/전문가	261	17.4		세종	12	0.8
	사무직	420	28.0	가구 구성	1인	221	14.7
	서비스/판매직	192	12.8		2인	356	23.7
	기능원/조작·조립/단순 노무	124	8.3		3인	402	26.8
	학생/주부/취업준비/무직	455	30.3		4인 이상	521	34.7

자료: 연구진 작성

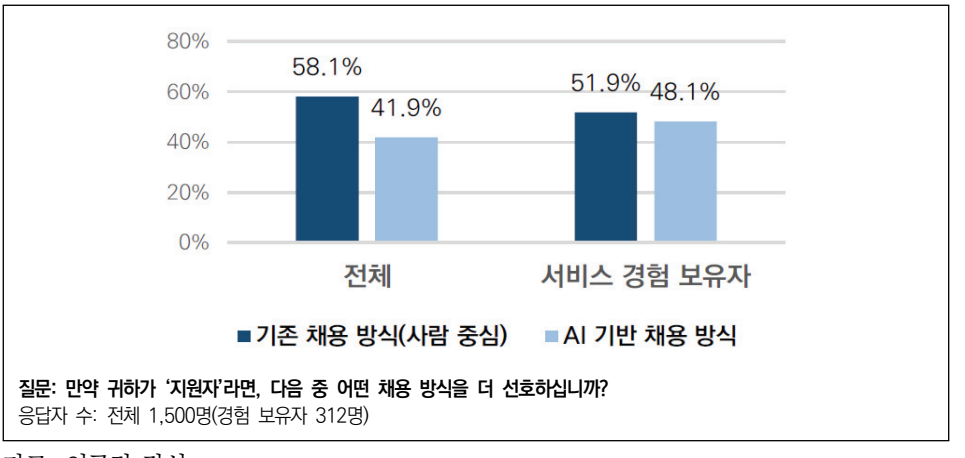
2. 조사 결과

가. AI 채용 서비스 경험 및 선호

응답자의 20.8%가 지원자 또는 채용 담당자로서 AI 채용 서비스를 접하거나 활용한 경험이 있다고 응답하였다.

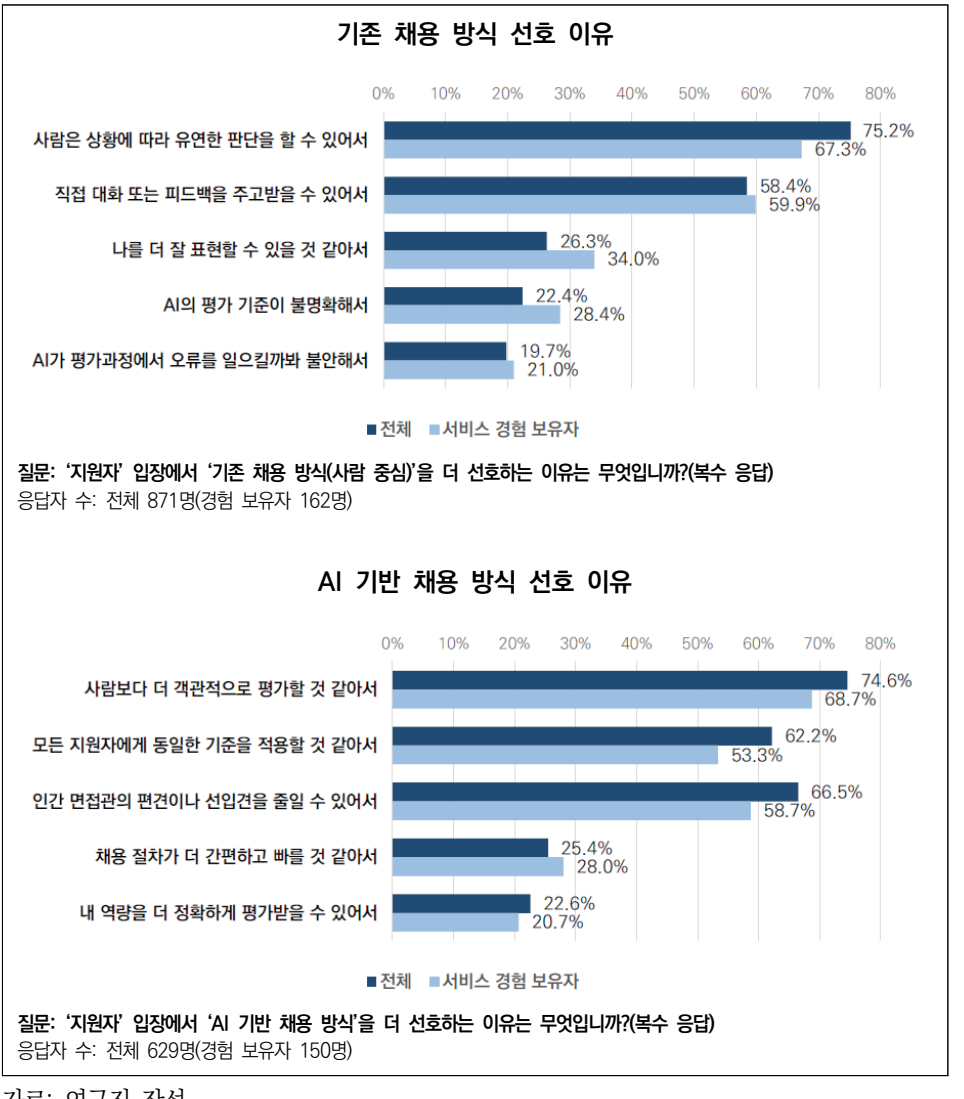
지원자 입장에서 전체 응답자의 58.1%는 기존 채용 방식을, 41.9%는 AI 기반 방식을 선호하였으나, AI 채용 서비스 경험이 있는 응답자 집단에서는 기존 방식 51.9%, AI 기반 방식 48.1%로 선호 차이가 크지 않았다.

[그림 2-4] 지원자 입장 선호 채용 방식



자료: 연구진 작성

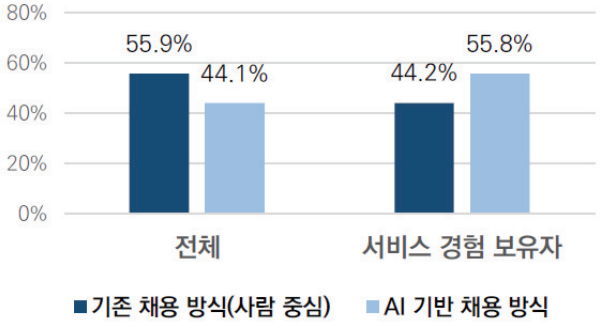
[그림 2-5] 지원자 입장 채용 방식 선호 이유



자료: 연구진 작성

채용 담당자 입장에서도 기존 방식 선호가 55.9%로 더 높았으나, 경험 보유자 집단에서는 AI 기반 채용 방식 선호가 55.8%로 기존 방식(44.2%)보다 더 높게 나타났다.

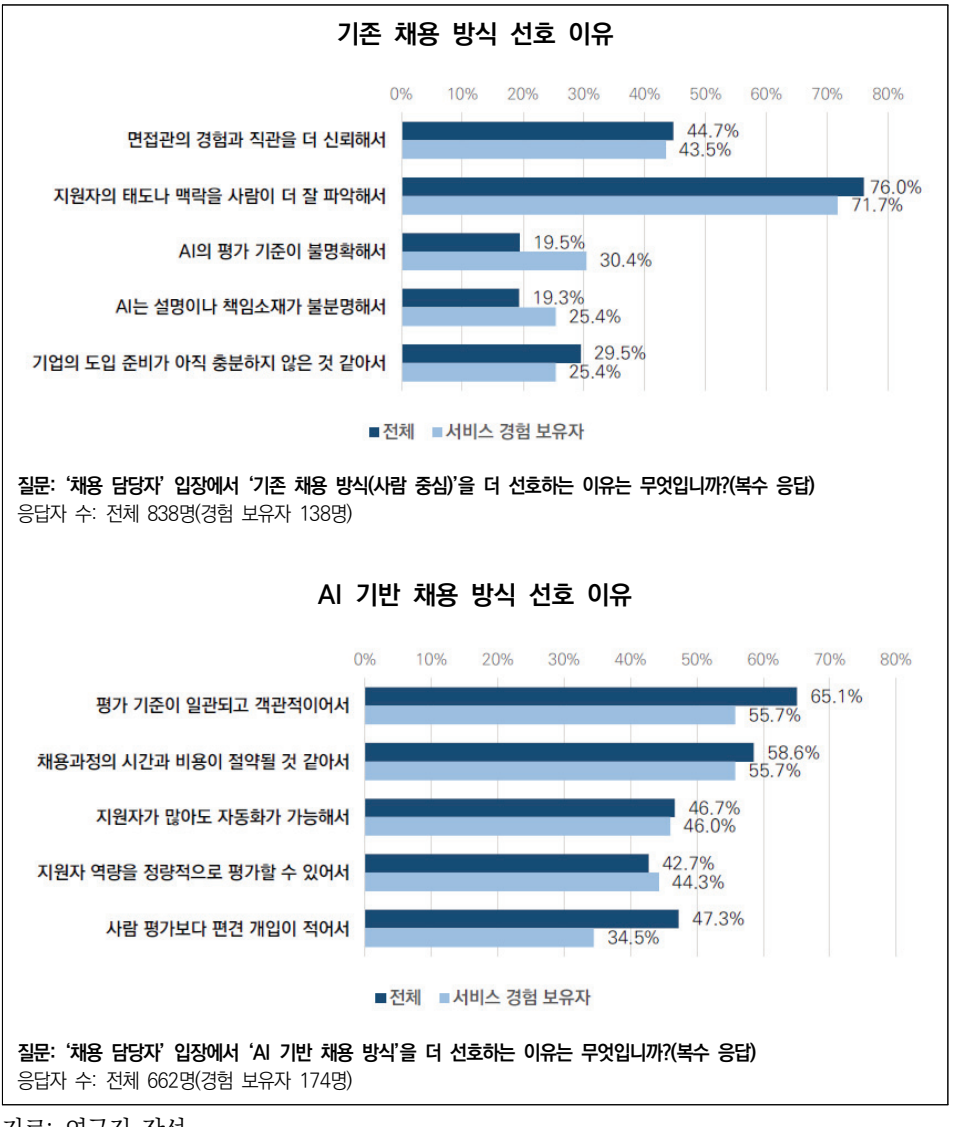
[그림 2-6] 채용 담당자 입장 선호 채용 방식



질문: 만약 귀하가 '채용 담당자'라면, 다음 중 어떤 채용 방식을 더 선호하십니까?
응답자 수: 전체 1,500명(경험 보유자 312명)

자료: 연구진 작성

[그림 2-7] 채용 담당자 입장 채용 방식 선호 이유



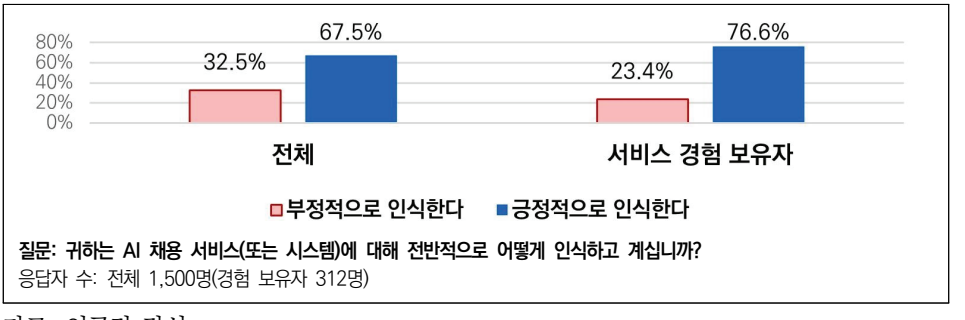
자료: 연구진 작성

나. AI 채용 서비스 인식·태도

AI 채용 서비스에 대한 국민 인식은 전반적으로 긍정적 평가가 우세하나, 직접 경험 여부에 따라 긍·부정 인식 정도와 분포에 일정 부분 차이가 나타났다.

전반적 인식(그림 2-8)에서는 전체 응답자의 67.5%가 AI 채용 서비스를 긍정적으로 평가하였으며, 서비스 경험이 있는 집단에서는 76.6%로 더 높게 나타났다. 반면 부정적 인식은 전체 32.5%, 경험자 23.4%로 경험 보유 여부에 따라 차이가 확인되었다.

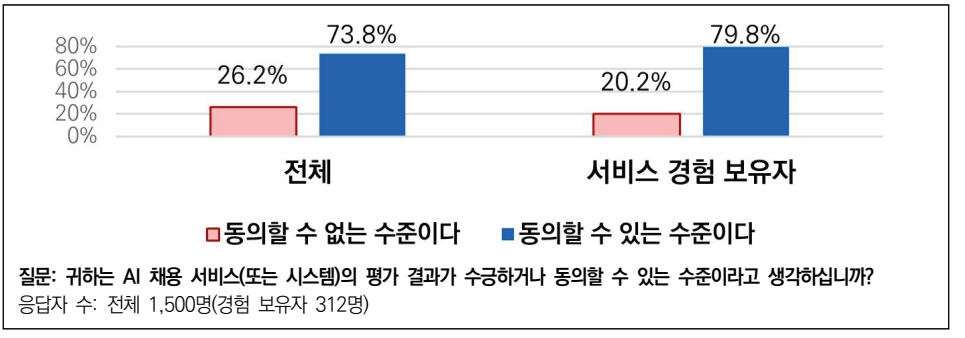
[그림 2-8] AI 채용 서비스에 대한 전반적 인식



자료: 연구진 작성

AI 채용 평가 결과 수용성(그림 2-9)를 보면, 전체 응답자의 73.8%가 '수용 가능한 수준'이라고 평가했으며, 경험자 집단에서는 이 비율이 79.8%로 더 높았다. AI 채용 평가에 대한 신뢰 수준은 전반적으로 긍정적이며, 실제 경험이 있는 경우 신뢰도가 더 강화되는 경향을 보였다.

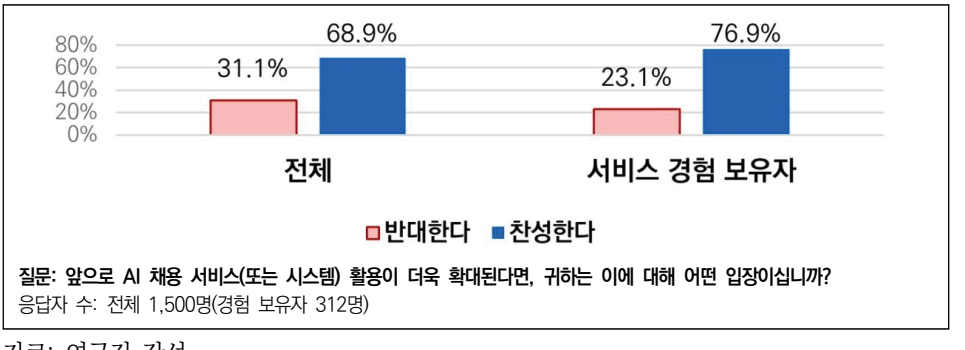
[그림 2-9] AI 채용 평가 결과 수용



자료: 연구진 작성

AI 채용 서비스 확대에 대한 태도(그림 2-10)에서도 긍정적 전망이 우세하였다. 전체 응답자의 68.9%가 서비스 확대에 찬성하였고, 경험자 집단에서는 찬성 비율이 76.9%로 더 높았다. 반대 비율은 31.1%, 경험자 23.1%로 나타나 경험이 있는 집단일수록 서비스 확산에 대해 더 개방적인 태도를 보였다.

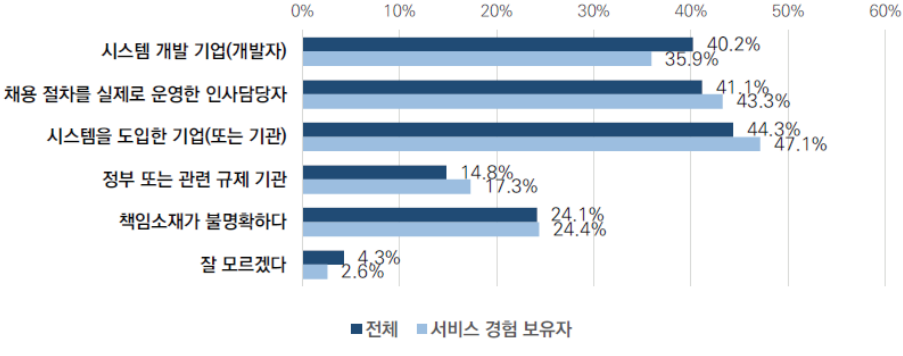
[그림 2-10] AI 채용 서비스 확대에 대한 찬반 입장



자료: 연구진 작성

문제 발생 시 책임 인식(그림 2-11)에서는 특정 단일 주체보다는 여러 주체가 동시에 책임을 부담해야 한다는 인식이 뚜렷했다. 전체 기준으로는 '시스템 도입 기업(44.3%)', '인사담당자(41.1%)', '개발기업(40.2%)' 순으로 응답이 높았으며, 경험자 집단에서도 유사한 경향을 보였다. 이는 AI 채용 과정에서 결과의 책임이 기술 제공자뿐만 아니라 시스템을 도입·운영하는 기관 전체에 걸쳐 분산되어 인식되고 있음을 보여준다. 한편 '책임소재가 불명확하다'는 응답도 약 24%로 나타나 책임 구조 명확화에 대한 정책적 필요성을 시사하였다.

[그림 2-11] AI 채용 문제 발생 시 책임



질문: AI 채용 서비스(또는 시스템)에 오류, 불공정 또는 부정적 결과가 발생했을 경우, 그 책임은 누구에게 있다고 생각하십니까?(복수 응답)
응답자 수: 전체 1,500명(경험 보유자 312명)

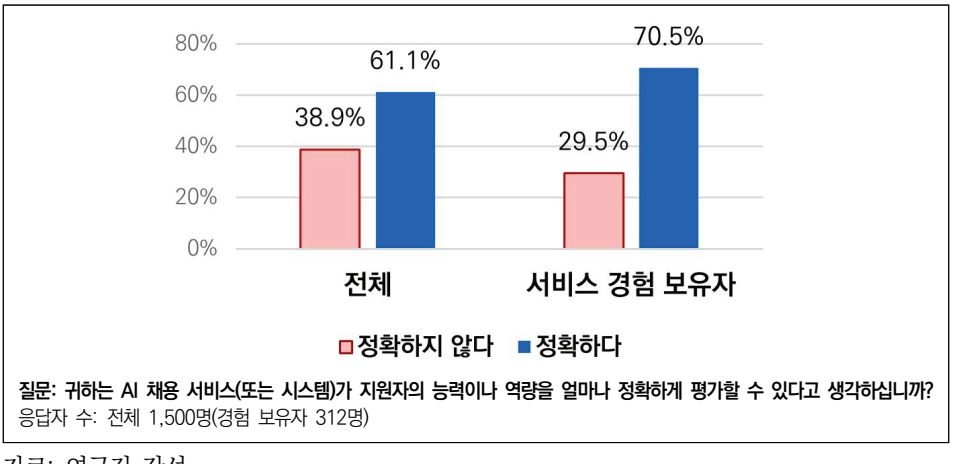
자료: 연구진 작성

다. AI 채용 서비스의 정확성·편향 가능성·공정성 인식

AI 채용 서비스의 정확성, 편향 가능성, 그리고 공정성에 대한 국민 인식은 전반적으로 긍정적인 평가가 우세하였으나, 항목별로 경험 여부와 판단 근거에서 차이가 확인되었다.

정확성 인식(그림 2-12)에서는 전체 응답자의 61.1%가 AI 채용 서비스가 '정확하다'고 평가해 긍정 인식이 우세했으며, 경험자 집단에서는 이 비율이 70.5%로 더 높았다. 정확하다고 판단한 이유로는 '감정이나 편견 개입 없이 안정적으로 판단할 것 같다'(65.6%), '데이터 기반이라 정확도가 높을 것 같다'(60.2%) 등이 주요하게 제시되었다. 반면 '지원자의 개별 상황을 충분히 반영하지 못할 것 같다'(70.7%)는 응답이 가장 높아, 정성적·맥락적 요소 반영의 한계를 우려하는 시각도 함께 존재하였다.

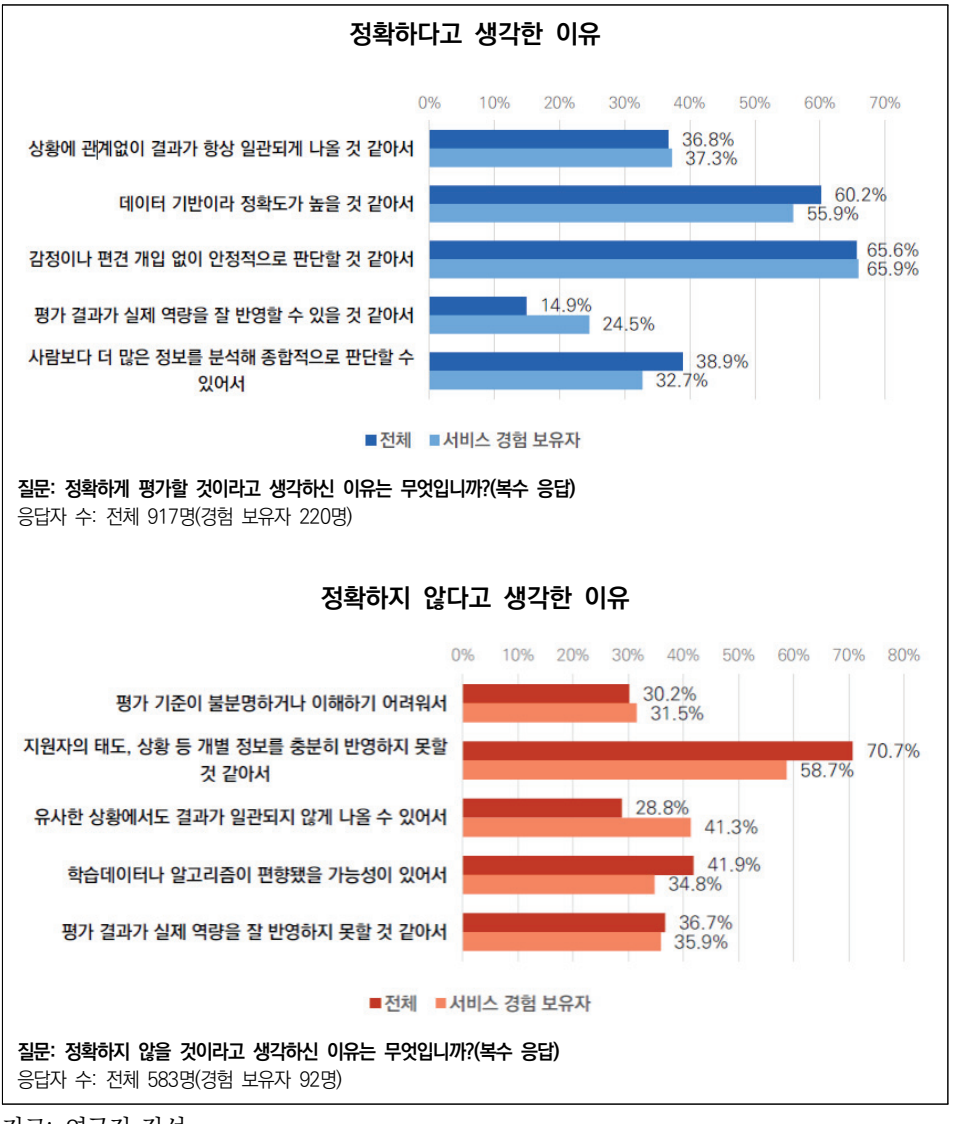
[그림 2-12] AI 채용 서비스의 정확성



질문: 귀하는 AI 채용 서비스(또는 시스템)가 지원자의 능력이나 역량을 얼마나 정확하게 평가할 수 있다고 생각하십니까?
응답자 수: 전체 1,500명(경험 보유자 312명)

자료: 연구진 작성

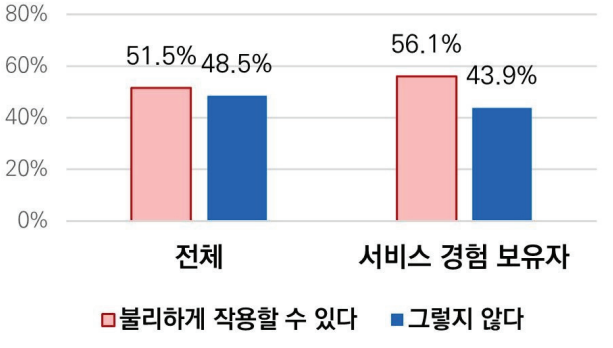
[그림 2-13] AI 채용 서비스 정확성 판단 이유



자료: 연구진 작성

편향 가능성(그림 2-14)에 대해서는 전체적으로 '불리하게 작용할 수 있다'는 응답이 51.5%로 다소 더 높게 나타났으며, 경험자 집단에서는 56.1%로 우려 수준이 상대적으로 더 높았다. 편향 우려의 이유로는 '특정 조건에 과도한 가중치가 부여될 수 있어서'(56.3%), '학습데이터 자체의 편향 가능성'(41.8%) 등이 대표적으로 지적되었다. 이는 AI 채용 서비스의 작동 방식과 데이터 기반에 대한 구조적 우려가 존재함을 시사한다.

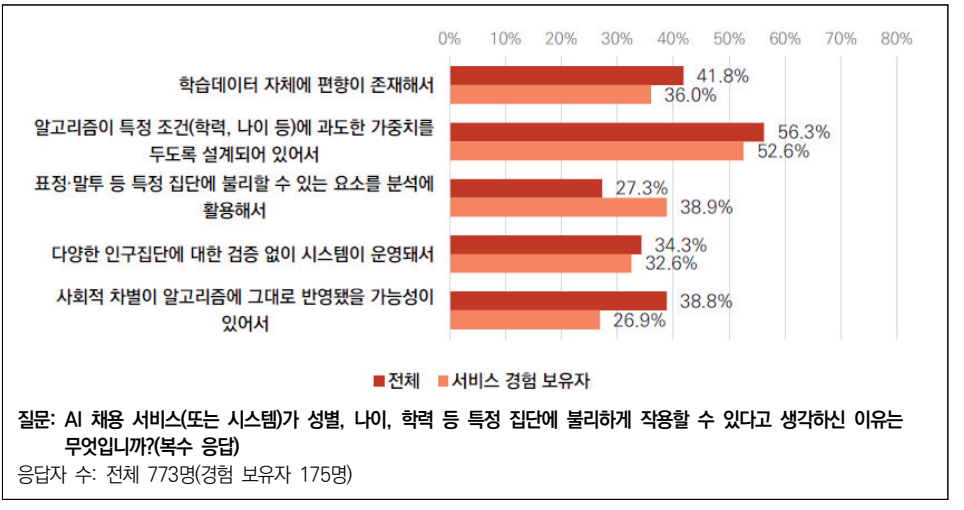
[그림 2-14] AI 채용 서비스의 편향 가능성



질문: AI 채용 서비스(또는 시스템)가 성별, 나이, 학력 등 특정 집단에 불리하게 작용할 수 있다고 생각하십니까?
응답자 수: 전체 1,500명(경험 보유자 312명)

자료: 연구진 작성

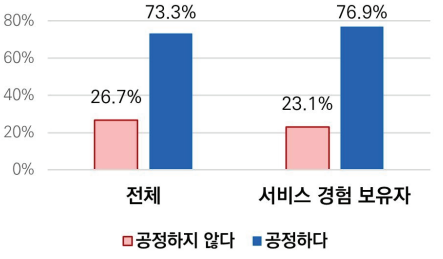
[그림 2-15] AI 채용 서비스 편향 가능성 판단 이유



자료: 연구진 작성

공정성 인식(그림 2-16)에서는 전체 응답자의 73.3%가 AI 채용 서비스가 공정하다고 보았다. 경험자 집단에서는 76.9%로 더욱 높았다. 공정하다고 평가한 이유는 '동일한 기준이 모든 사람에게 적용될 수 있어서'(64.9%)와 '개인적 관계·인맥 등 외부 요인의 영향이 줄어들 것 같다'(62.3%) 등이 주를 이루었다. 다만 공정하지 않다고 본 응답자들은 '지원자의 상황·배경 등 맥락을 충분히 고려하지 못함' (62.3%), '정성적 요소를 반영하지 못함'(60.1%)을 주된 이유로 제시하여, 정성적·맥락적 정보 반영의 한계가 공정성 판단을 제약하는 요인으로 인식되고 있음을 시사하였다.

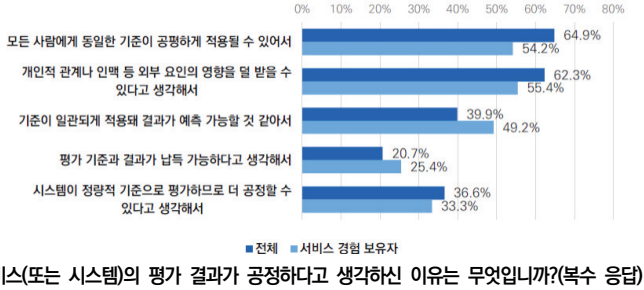
[그림 2-16] AI 채용 서비스의 공정성



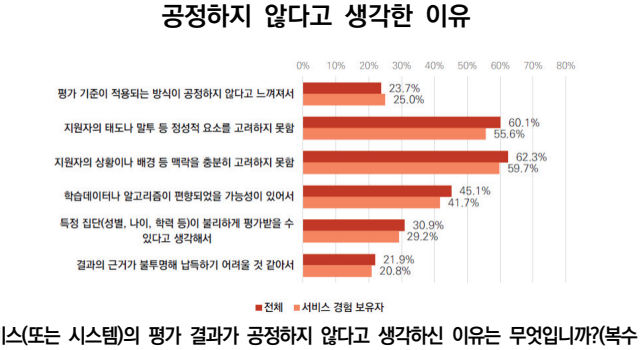
질문: 귀하는 AI 채용 서비스(또는 시스템)의 평가 결과가 공정할 것이라고 생각하십니까?
응답자 수: 전체 1,500명(경험 보유자 312명)

자료: 연구진 작성

[그림 2-17] AI 채용 서비스 공정성 판단 이유



질문: AI 채용 서비스(또는 시스템)의 평가 결과가 공정하다고 생각하신 이유는 무엇입니까?(복수 응답)
응답자 수: 전체 1,099명(경험 보유자 240명)

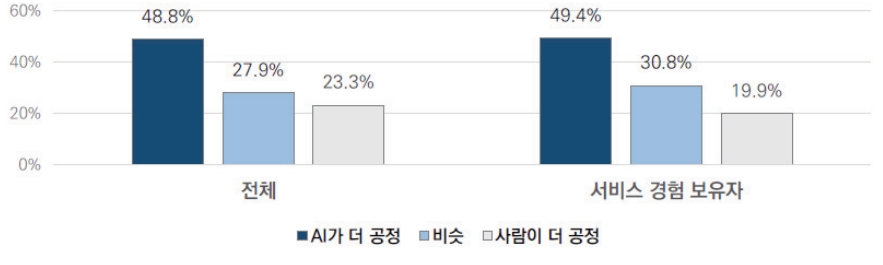


질문: AI 채용 서비스(또는 시스템)의 평가 결과가 공정하지 않다고 생각하신 이유는 무엇입니까?(복수 응답)
응답자 수: 전체 401명(경험 보유자 72명)

자료: 연구진 작성

AI 채용 방식과 기존 채용 방식 간 공정성 비교(그림 2-18)에서는 전체 응답자의 48.8%가 'AI가 더 공정하다'고 응답했으며, 경험자 집단에서는 49.4%로 유사한 수준이었다. '사람이 더 공정하다'고 응답한 비율은 전체 23.3%, 경험자 19.9%로 상대적으로 낮았으며, 약 30%는 두 방식이 '비슷하다'고 보았다. 이는 AI 채용 방식을 기존 방식보다 더 공정하게 인식하는 경향이 우세하지만, 두 방식을 비슷한 수준으로 평가하는 응답자도 적지 않음을 보여준다.

[그림 2-18] AI 채용 방식과 기존 채용 방식의 공정성 비교



질문: 귀하는 AI 채용 서비스(또는 시스템)와 사람(기존 채용 방식) 중 어떤 방식이 더 공정하게 평가한다고 생각하십니까?
응답자 수: 전체 1,500명(경험 보유자 312명)

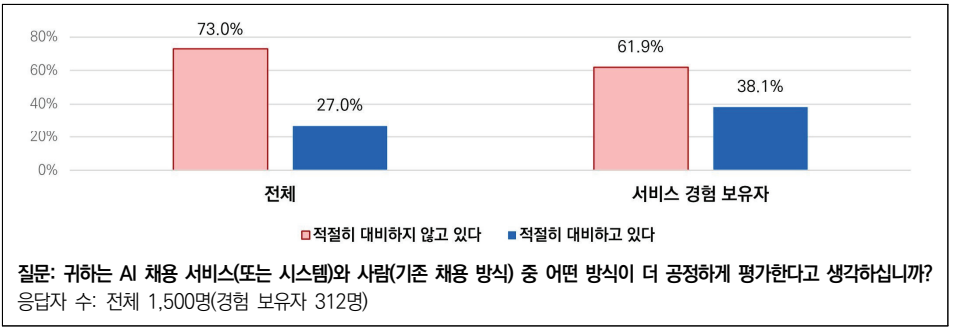
자료: 연구진 작성

라. AI 채용 관련 정부 대응과 신뢰

AI 채용 서비스 확산에 따라 정부의 역할과 대응 역량에 대한 국민 평가를 살펴본 결과, 전반적으로 정부 대응에 대한 신뢰 수준은 높지 않은 편으로 나타났으나, 서비스 경험 여부에 따라 평가 강도에는 일부 차이가 확인되었다.

정부의 대응 수준(그림 2-19)에 대해서는 전체 응답자의 73.0%가 '적절히 대비하지 않고 있다'고 응답해, 정부 대응이 충분하지 않다고 보는 인식이 우세한 것으로 나타났다. 다만 서비스 경험 보유자 집단에서는 해당 비율이 61.9%로 낮아진 것을 확인할 수 있으며, 경험자 집단이 비경험자보다 정부 대응을 상대적으로 더 긍정적으로 평가하는 경향을 보였다.

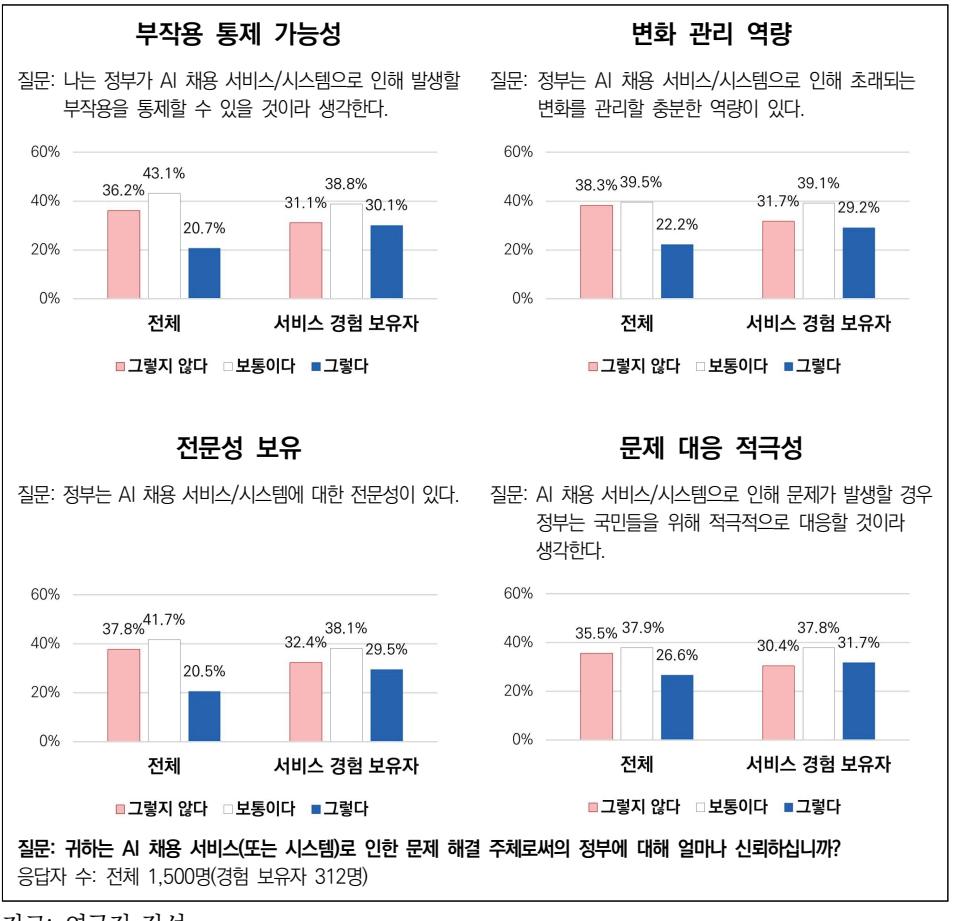
[그림 2-19] AI 채용 관련 정부 대응 수준



자료: 연구진 작성

문제 해결 주체로서 정부에 대한 신뢰를 세부 항목별로 살펴보면, 전반적으로 긍정 응답은 20~30%대에 머물러 전반적인 신뢰 수준은 높지 않았다. 부작용 통제 가능성의 경우 긍정 응답은 전체 20.7%, 경험자 30.1%였으며, 변화 관리 역량(전체 22.2%, 경험자 29.2%), 전문성 보유 인식(전체 20.5%, 경험자 29.5%)에서도 유사한 응답 분포가 확인되었다. 문제 발생 시 정부의 적극적 대응 기대에서도 긍정 응답은 전체 26.6%, 경험자 31.7%였으며, '보통이다' 응답이 각각 37.9%, 37.8%로 가장 높은 비율을 차지하였다. 모든 세부 항목에서 '보통이다' 응답이 가장 높은 비율을 차지했으며, 부정 응답도 일관되게 30%대 수준을 기록한 점은, 정부가 AI 채용 관련 위험·변화·전문성 측면에서 충분한 역량을 갖추고 있다는 인식이 아직 제한적이며, 동시에 많은 국민이 정부 역량에 대해 뚜렷한 긍정·부정 판단을 유보하고 있음을 보여준다.

[그림 2-20] 문제해결 주체로서 정부에 대한 신뢰



자료: 연구진 작성

종합하면, 국민은 AI 채용 서비스가 확산되는 상황에서 정부의 대응 수준과 전문성, 대응 역량에 대해 전반적으로 낮은 수준의 신뢰를 보이는 것으로 나타났다. 다만 서비스 경험 보유자 집단에서는 모든 항목에서 상대적으로 높은 신뢰 수준을 보이는 경향이 관찰되어, 실제 서비스 경험이 정부 역할에 대한 인식에 일정 부분 영향을 미치는 것으로 나타났다.

제 4 절 AI 채용 서비스 윤리영향평가

1. AI 채용 서비스 개요

AI 채용 서비스는 이력서 분석, 직무 적합성 예측, 면접 평가 등 채용 과정 전반에 AI 기술을 적용하여 인재 선발 절차를 자동화하거나 지원하는 시스템을 의미한다. 기존에는 채용 담당자가 지원자의 정보를 직접 검토하고 평가했다면, AI 채용 서비스는 기존의 채용 절차를 보조하거나 일부 대체하여 채용의 효율성, 공정성, 객관성을 높이는 도구로 활용되고 있다.

최근 AI 채용의 실효성이 부각되면서 국내외 도입 수요가 빠르게 증가하는 추세다. Straits Research(2024)는 글로벌 AI 채용 시장 규모가 2024년 6억 1,756만 달러에서 2033년 11억 2,584만 달러까지 성장할 것으로 전망하였다. 미국의 경우 포춘 500대 기업 중 98% 이상이 채용 과정에 AI를 활용하고 있으며(한경비즈니스, 2024.6.29.), 글로벌 AI 면접 솔루션 기업 '하이어뷰'가 2025년 2월 전 세계 4,000여 고용주를 대상으로 조사한 결과에서도 채용 과정 AI 활용률이 2024년 58%에서 2025년 72%로 증가한 것으로 나타났다(The Guardians, 2025.5.14.).

국내에서도 확산 속도가 빠르다. 한국경제인협회의 '대기업 동향·인식 조사'(2024)에 따르면 신규 채용 과정에서 AI를 활용하거나 도입을 고려 중인 기업 비율은 40.7%(N=123)로, 전년 대비 약 15%p 증가하는 등 도입 확산이 뚜렷하게 나타났다.

AI 채용 서비스는 주로 ① 이력서·자기소개서 분석, ② AI 기반 후보자 매칭, ③ AI 면접, ④ 채용업무 자동화 등 여러 채용 단계에서 활용되며, 기업들은 주로 상용화된 AI 채용 서비스를 도입해 기존 절차를 보조하거나 자동화하고 있다.

〈표 2-7〉 AI 채용 서비스의 주요 활용 단계 및 기능

채용 단계	주요 기능
이력서· 자기소개서 분석	지원자의 이력서, 자기소개서, 포트폴리오 등을 AI가 자동 분석 직무적합 점수 산출 필수역량 핵심 키워드 자동 추출
AI 기반 후보자 매칭	기업이 원하는 직무역량·스킬·경력과 지원자 매칭 직무 예측 모델을 활용하여 “지원자 추천” 결과 제공
AI 면접	화상 인터뷰에서 언어적(답변 내용), 비언어적(표정·자세 등), 음성 특성 분석을 통해 면접 평가 지원
채용업무 자동화	직무 요건 기반의 직무 설명·채용공고 자동 작성 면접 일정 조율, 안내 메시지 발송 등 반복 응대 업무 자동 처리

자료: 연구진 작성

AI 채용은 사람의 주관적 판단을 보완하여 일관된 평가 기준을 적용할 수 있다는 점에서 공정성 강화에 기여할 뿐 아니라, 시간·비용 절감 효과가 있어 기업의 활용도도 높다. 그러나 동시에 △학습 데이터의 편향 및 차별, △프라이버시 침해, △운영 과정의 불투명성, △디지털 격차 등 다양한 윤리적 위험이 제기되면서 규제의 필요성도 커지고 있다. 실제로 주요국의 법규제에서도 채용을 포함한 고용 관련 의사결정 AI 시스템을 ‘자동화 의사결정 도구’ 또는 ‘고위험 AI 시스템’으로 분류하여 규제 대상으로 삼고 있다.

일반 시민들도 이러한 위험을 명확히 인지하고 있다. 한국노동연구원이 2024년 20~39세 구직자 1,055명을 대상으로 실시한 실태 조사에 따르면, AI 채용 경험자 중 AI 평가 결과에 동의하지 않는 이유로 ‘AI의 평가 기준이 공개되지 않기 때문에’(40.3%), ‘AI가 학습한 편견이나 편향이 자신에게 적용되었기 때문에’(19.5%) 등이 높은 비중을 차지했다(양승엽 외, 2024).

AI 채용이 먼저 확산된 미국에서는 실제 법적 분쟁 사례도 지속되고 있다. 2022년 아이튜터그룹은 고령 지원자를 자동 탈락시킨 혐의로 제재[5]를 받았고,

5) 미국 강사를 고용하여 온라인 과외 서비스를 제공하는 중국 기업 ‘아이튜터그룹’이 강사

2023년 워크데이는 인종을 기준으로 한 차별 의혹으로 소송[6]에 휘말리는 등 사회적 논란이 이어지고 있다(한국경제, 2024.10.28.). 이처럼 AI 채용 서비스는 효율성과 편의성이라는 잠재적 이점과 함께, 분쟁과 규제 이슈로 직결될 수 있는 윤리·법적 위험이 공존하는 분야이다.

2. 국민포럼단 FGI 주요 결과

국민포럼단 FGI는 일반 국민 관점에서 AI 채용 서비스에 대한 인식, 기대, 우려사항 등을 심층적으로 파악하기 위한 목적으로 운영되었다. 두 차례에 걸쳐 진행한 국민포럼단 FGI(Focus Group Interview)의 구성, 절차, 결과를 소개하고자 한다.

가. 국민 포럼단의 구성

국민포럼단[7]은 총 30명으로 구성되며, 20대부터 60대까지 연령대별 5~7명씩 총 5개 그룹으로 나뉜다. 연령 외에도 성별, AI 채용 서비스에 대한 태도[8] 등 다양한 요소를 함께 고려하였다.

채용 과정에서 활용한 AI 채용 서비스가 고령 지원자의 이력서를 자동으로 걸러내어 200명 이상의 고령 지원자들이 채용 기회를 박탈당했다. 미국 평등고용기회위원회(EEOC)는 2022년 5월 '고용상 연령차별금지법(ADEA)' 위반을 이유로 제소했으며, 2023년 8월 피해 지원자 200여명에게 총 36만 5000달러(약 5억 700만 원)를 배상, 재지원 허용 등의 조건으로 합의

6) 미국 HR 서비스 기업 '워크데이'의 AI 채용 서비스를 이용한 회사에 지원하여 탈락한 지원자들이 AI가 인종을 기준으로 차별적인 평가를 했다고 워크데이에 대한 소송을 제기(2023년), 현재까지도 사건이 진행중

7) 'AI 채용 서비스 대국민 인식조사'(2025.6)를 진행하며 1,500명을 대상으로 국민포럼단 참여 의사를 확인하였으며, 참석 의사를 밝힌 응답자 중 국민포럼단을 선정

8) 'AI 채용 서비스 대국민 인식조사'의 다음 문항에 대한 답변을 토대로 AI 채용 서비스에 대한 태도를 구분함: [문항] 귀하는 AI 채용 서비스(또는 시스템)에 대해 전반적으로 어떻게 인식하고 계십니까? [답변] ① 매우 부정적으로 인식한다 ② 다소 부정적으로 인식한다 ③ 대체로 긍정적으로 인식한다 ④ 매우 긍정적으로 인식한다

〈표 2-8〉 국민포럼단의 구성

그룹	성별	연령	직업	태도
G1	여	28	컨퍼런스 기획	3
	남	27	반도체 설계	3
	남	28	건설 사무직	3
	남	28	스포츠 강사	3
	여	29	컨텐츠(기사, 광고관련)	2
	여	25	데이터 분석	2
	여	27	국방업 사무직	2
G2	여	36	교육업	4
	여	34	임대용업 업체 사무직	3
	남	36	건설업	3
	남	36	교육 사무직	2
	남	36	건설 사무직	2
G3	여	44	대학교 근무	3
	남	41	자동차 제조업	3
	여	44	프리랜서 방송작가	3
	여	44	제조업·사무직	2
	남	42	섬유 무역업	2
	남	43	전기안전 공공서비스	2
G4	남	55	IT 시스템 컨설팅	3
	여	55	회계업	3
	여	54	렌즈 회사	3
	남	50	화학제품 제조업	3
	남	50	소프트웨어 유통업	2
	여	56	영어 번역	2
	여	50	지적정보 데이터	2
G5	남	66	기계업	3
	남	60	병원 행정	3
	여	61	국제회의 통역	4
	여	61	학원 강사	2
	여	61	관리사무소	2
	남	61	의료기기 연구개발	2

자료: 연구진 작성

나. 평가의 절차와 방법

FGI 조사는 그룹별로 총 두 차례, 전체 10회에 걸쳐 진행되었다. 1차 조사에서는 AI 채용 서비스에 대한 이해와 인식을 다루었으며, 2차 조사에서는 긍정적·부정적 영향을 평가하고, 정부·기업·개인 주체별 노력에 대한 논의를 중점적으로 진행하였다.

〈표 2-9〉 국민포럼단 FGI 평가 절차

1차 좌담회		준비		2차 좌담회
8.18.(월)~8.22.(금), 약 70분				9.1.(월)~9.5.(금), 약 120분
(진행 내용) • Warm up • 본 평가 - AI 채용 서비스 인식 - 채용 방식에 대한 선호 인식 - 윤리 가치 반응	→	AI 채용 서비스 윤리영향 설문	→	**(진행 내용)** • Warm up • 본 평가 - 윤리 가치별 긍·부정 영향 사례 공유 및 토론 - 주체별 노력에 대한 생각

자료: 연구진 작성

다. 평가 결과

1) AI 채용 서비스에 대한 인식

AI 채용에 대해 긍정적으로 인식하는 참석자들은 학연, 지역, 혈연, 외모 등과 관계없이 공정하게 평가할 것 같다는 인식이 있으며, 채용 업무의 속도와 효율을 크게 향상시킬 수 있다는 기대가 있었다.

〈표 2-10〉 AI 채용 서비스에 대한 긍정적 인식

"보통 얼굴 보거나 대면으로 하면 학력이라든지 이런 혈연, 지연 이런 것도 있을 텐데 AI를 거치면 공정하게 할수 있을 것 같아요."(30대, 응답자C)
"저는 아주 객관하게 공평하게 할 수 있는 기본 툴이라고 생각을 하고요. 사람이 봤을 때 인재 채용을 할 때 선입견을 가질 수가 있거든요. 학벌이나 지역이나 기타 등등 (중략) 그러니까 마지막 판단은 반드시 사람이 하게 될 테니 처음에 사람이 가질 수 있는 선입견은 다 배제하고 회사에서 가질 수 있는 회사에서 갖고 있는 기본 경영 방침이라든지 그런거에 잘 부합하는 사람들을 1차 걸러낼 수 있을 때 아주 유용하게 쓸 수 있을 거라 생각이 들어요." (40대, 응답자A)
"저는 제가 사실 채용자 담당자 입장이에요. 그러니까 사람들 면접을 많이 보고 서류를 많이 심사를 하잖아요. 근데 그게 굉장히 일이 엄청나거든요. 맨 처음에 걸러내는 정도가. 그 부분에서 1차적으로 걸러 내주기만 해도 시간이 엄청난 단축이 되고 저는 사람과 사람과의 물론 서류뿐만 아니라 만남에서의 오는 피로감이 있고 감정이 어떻게 보면 들 수밖에 없잖아요. (중략) 어느 정도 1차, 2차 정도는 걸러주고 나중에 최종 면접만 보게 된다면 그리고 윤리적인 영향이나 이런 것까지 다 조율이 된다면 저는 굉장히 긍정적이라고 생각합니다."(40대, 응답자C)

자료: 연구진 작성

반면, 부정적인 의견으로는 기계가 사람을 판단하는 것에 대한 거부감, AI의 오류 발생 우려, 감정이나 태도 등을 AI가 판단하기 어려울 것 같다는 의견이 있었다.

〈표 2-11〉 AI 채용 서비스에 대한 부정적 인식(1)

"AI가 채용이라는 게 사람의 인생을 좌지우지할 수도 있는 사안이잖아요. 그거를 AI가 사람 위에 있다는 생각이 들어가지고요."(40대, 응답자F)
"대면 관계 일을 할 때 AI가 예를 들어서 채용적인 면에서도 그런 거를 대체하는 것 자체를 사람이 사람 사이에서 느끼는 감정 자체를 그걸 캐치를 못할 것 같아요."(40대, 응답자E)
"라지랭귀지 모델이라고 알고 있는데 이거는 텍스트거나 이미지거나 이런 게 들어가야지 되는 분석을 하는 거잖아요. 패턴을 가지고요. 근데 그렇게 봤었을 때 사람을 어떤 식으로 평가한다 라고 했을 때 정말 정량적인 자료만 가지고 평가할 거 아니에요? 지금은요. 그래서 정성적 자료가 약간 걱정돼요."(50대, 응답자G)
"저도 인터뷰를 해봤지만 이 사람의 에티튜드나 물론 갖고 있는 커리어나 실력이나 이런 것도 중요하지만 에티튜드를 많이 보게 되는데 그런 관점에서 보면 AI로 채용을 한다라는 것 자체는 사실 가능한가?"(50대, 응답자A)

자료: 연구진 작성

특히 20대 집단에서는 AI 채용 서비스를 직접 경험했거나 주변에서 경험한 비중이 높았으며, 주로 부정적인 인식이 두드러지게 나타났다. 입꼬리를 올리거나 눈동자를 움직이지 않는 등 직무와 관련 없는 정보들이 난무하여 정확한 기준을 알 수 없고, 불필요한 단계가 추가되었다는 인식이 큰 것으로 나타났다.

〈표 2-12〉 AI 채용 서비스에 대한 부정적 인식(2)

"뭔가 꿀팁 같은 게 블로그에 많이 올라와요. 그런데 항상 하는 말이 약간 되도 않는 게임 시키면서 실수하면서 찡그리면 감정이다 계속 소개를 할 때 입꼬리를 올리고 있어야 AI가 인식을 잘한다 이런게 있었어요. 이게 채용하고 무슨 관련이 있지? 싶은 똥개 훈련인 것 같다는 느낌을 많이 받았었어요."(20대, 응답자F)
"근데 막상 그게 진짜 나를 평가한다고 생각하니까 눈을 조금 굴려도 그것도 다 인식이 된다 이런 말도 있고 정확히 그게 다 되는지는 모르겠지만 제가 담당해서 본 게 아니니까 아무튼 너무 긴장이 돼서 막상 이렇게 얘기하는 것보다 훨씬 면접할 때 말도 잘 못 나오는 것 같고 그런 질문을 하는데도 몇 초 안에 답을 해라 이런 식의 평가를 하는데 그게 잘 이루어지는 건지 모르겠어요."(20대, 응답자E)
"보통 AI 면접 보면은 1, 2차 면접을 안 본다던가 하나를 생략하던가 그래야 되는데 제가 봤던 곳은 다 보고 하나가 더 추가된 느낌이 있었고 그리고 측정 방식이 명확하지가 않으니까 그러니까 들리는 거 눈 굴리면 안 된다 입 올라가는거 그런 거 본다고 하는데 명확한 기준은 제시해 주지 않고 찾아봐야 되는데 그것도 카더라 하는 거니까 그런 게 있었던 것 같아요."(20대, 응답자B)

자료: 연구진 작성

2) 채용 방식에 대한 선호 인식 탐색

(1) 지원자 관점에서 선호하는 채용 방식

지원자 관점에서 선호하는 채용 방식에 대해 물었을 때, 학벌이나 다른 스펙이 좋은 경우라면 유리할 것 같다는 의견과 오히려 스펙이 좋지 않은 사람이라면 AI 채용이 유리할 것 같다는 상반된 의견이 나타났다.

AI 채용 서비스를 선호하는 이유로는, AI 채용을 통해 피드백을 얻을 수 있다면 오히려 부족한 부분을 채울 수 있는 기회가 될 수 있으며, 사람을 대면하는 것보다는 중압감이 줄어들 것 같다는 점이 언급되었다.

〈표 2-13〉 지원자 관점에서 AI 채용 방식 선호

"예를 들어서 회사에서 나는 긍정적인 성향, 협동성이 높은 사람을 뽑겠다 라고 회사가 생각하면 아무리 대학의 성적이 좋아도 소심하거나 혼자 연구하는 사람은 안 뽑히겠죠. 그러니까 그거를 정하는 기준은 회사에서 정하는 거고 지원자인 나로서는 나의 성향을 그냥 그대로 쉽게 AI 툴로 보여줘서 나를 뽑아봐라 그렇게 하는게 저는 지원자로서는 너무 좋을 것 같아요."(40대, 응답자A)
"지원자 입장에서 만약에 탈락했다고 생각을 하고 이랬을 때 뭐 때문에 탈락했는지에 대해서 누구도 언급하지는 않잖아요. 그냥 죄송하지만 탈락하게 되었습니다 이런 식으로만 문자가 오지 그거에 대해서 오지 않는데 피드백이 정확하게 오기 때문에 나에 대한 객관화도 확실히 알 수 있고 그다음에 다른 지원을 할 때도 그거에 대해서 내가 부족한 점을 인식하고 그 부분에 대해서 더 공부를 해서 그거를 더 활용해 가지고 다음 다른 데에다가 지원할 수도 있고 그래서 그런 부분이 되게 긍정적인 것 같아요."(40대, 응답자C)
"저는 신입이 됐던 경력이 됐던 내가 다른 사람 앞에서 평가를 받는다는 것 자체는 굉장한 중압감이 있을 수 있잖아요. AI를 통해서 인터뷰를 한다라고 하면 그런 측면에서는 뭔가 부담을 덜어내고 내가 표현하고 싶은 내가 하고자 하는 이야기들을 더 자연스럽게 할 수 있지 않을까 라는 생각이 있었어요."(50대, 응답자A)

자료: 연구진 작성

기존 채용 방식을 선호하는 응답자의 경우 채용 과정이 단순히 기업이 지원자를 평가하는 자리가 아닌 지원자 또한 기업을 평가하는 자리이므로 AI가 채용을 담당한다면 그러한 기회가 사라질 것이라고 생각했다. 또한 학습데이터 입력을 사람이 하는 것이기 때문에 오히려 조작의 위험이 높아질 우려가 있으며, 대면 업무를 평가하기엔 부적절하다는 의견 역시 나타났다. 그리고 기존 채용 방식이 AI 채용보다는 유연성이 있기 때문에 여러 특성을 상황에 맞게 양해해 줄 수 있을 것이라는 의견도 있었다.

〈표 2-14〉 지원자 관점에서 기존 채용 방식 선호

"특히 서비스업은 사람을 대면하고 해야 되는데 전반적인 걸 다 얘가 하면은 얘가 진짜 사람을 대면했을 때 어떻게 대하는지에 대한 데이터가 없으니까 더 위험 요소가 커진다라고 생각해서 저는 바꿨습니다."(20대, 응답자D)
"저는 어쨌든 회사에서도 저를 평가하지만 저도 그 회사를 평가하는 자리잖아요. 면접이라는 기회를 통해서. 근데 AI로만 채용을 대체하면 나는 지금부터 널 평가할 거야라는 평가 대상자로 밖에 기업이 안 보는 것 같아서 일방적으로 저를 평가한다는 기분이 들어서 저는 거기서부터 조금 그 기업은 안 가고 싶을 것 같아요."(20대, 응답자F)
"AI 자체가 인간이 넣어 놓은 자료를 기반으로 하기 때문에 내가 어떤 사람을 뽑고 싶다면 만약 인사팀에서 그런 자료를 넣을 때 조작하거나 그렇게 해서 넣을 수가 있어요. (중략) 그러니까 이게 조작이나 인간이 다루는 사람에 따라서 굉장히 그게 실수나 오류가 많이 생기거든요. 그래서 만약에 잘못된 데이터를 넣고 나 이런 사람 뽑아줘 했을 때 걔가 제대로 알아듣고 제대로 뽑을까? 물론 자료를 넣어주는 사람이 잘 넣으면 문제가 없는데 조작 가능성이 있거든요. 그게 굉장히 큰 리스크가 될 것 같아요. 오히려 더 객관화되지 않고 한쪽으로 치우쳐지는 그런 리스크가 있는 사람이 뽑혀질 수도 있는 거죠."(50대, 응답자F)

자료: 연구진 작성

단계별로 구분했을 때 대부분의 응답자들이 서류, 인적성 평가는 AI가 진행해도 괜찮지만 면접은 반드시 기존 방식대로 사람이 평가해야 한다는 의견이 주를 이뤘다. 더불어 AI를 활용하더라도 사람이 스크리닝하는 과정이 포함되거나, 최종 결정을 AI가 아닌 사람이 결정하도록 AI를 보조적 수단으로 활용해야 한다는 의견도 나타났다.

〈표 2-15〉 지원자 관점에서 채용 단계별 선호

"저는 사실 3단계 면접은 무조건 사람이 해야 된다 라고 생각을 하는데 1단계, 2단계 같은 경우에는 AI를 태운다고 해도 지금하고 그렇게 거부감이 많이 차이가 나지는 않을 것 같아요. 왜냐하면 1단계는 어차피 서류전형부터 컴퓨터로 쓰는 거고 2단계도 기계적으로 회사 가서 시험 보고 똑같이 마킹하고 다르지 않을 것 같아서 이거를 근데 한번 인사 담당자가 AI 분석 결과를 보고 그걸 한번 더 거르는 용도로는 괜찮을 것 같은데 완전히 AI로 맡겨버려서 AI가 이 사람은 합격입니다 이 사람은 불합격입니다. 이건 아닌 것 같아요. 스크리닝을 했으면 좋겠어요."(20대, 응답자F)
"업무에 대한 능력에 대한 평가 같은 거를 AI로 해갖고 세분화적으로 실질적인 현장에 나가서 업무를 할 수 있는지에 대한 평가는 AI가 충분하다고 생각하는데 그 사람의 태도라든지 아니면 세세한 면접 보면서 하는 행동이나 눈빛이나 이런 시선 변화라든지 이런 세밀한 것 같은 거는 AI가 아직 그거 따라가지 못할 거라고 생각하고 그건 사람이 할 수 있는 분야라고 생각을 하거든요."(30대, 응답자F)
"저도 최종적인 건 사람이 했으면 좋겠어요. 어쨌든 최종적인 거는 아직까지는 아까 말했듯이 감정이나 이런 걸 볼 수가 없고 그 사람 성향이나 이런 거를 AI가 판단하기에는 아직은 어려우니까."(50대, 응답자B)

자료: 연구진 작성

(2) 채용자 관점에서 선호하는 채용 방식

채용 담당자 관점에서는 대체로 AI 채용 활용에 긍정적이었다. 업무 효율성 제고, 비용 절감의 측면과 더불어 채용 담당자 관점에서 채용 결과에 대한 책임을 져야하는 부담감을 경감시킬 수 있고, 근거 마련이 가능하기 때문이다.

〈표 2-16〉 채용자 관점에서 AI 채용 방식 선호

"좀 나눌 것 같은데 만약에 제가 인사 담당자 입장이면 일단 명백한 근거에 의해서 사람을 뽑는 거기 때문에 이 사람이 어떻게 되든 난 일단 명분이 있다 그러니까 더 나을 것 같고 근데 비용 같은 측면에서도 말씀을 해 주셨는데 초기 투자금은 들겠지만 이게 또는 기업이 10년만 사람을 뽑을 것도 아니고 계속 뽑아야 되잖아요. 길게 보면 비용적 측면에서도 분명히 이득이니까 지금 그 사람들이 하고 있는 거라고 생각을 해서 저는 AI가 더 맞다고 생각합니다."(20대, 응답자A)
"왜냐하면 천 개의 서류를 제가 봐야 한다면 저희도 사실 저희 회사에서 채용할 때 저런 애를 왜 뽑았어? 라는 어떻게 저런 애가 들어왔어? 이런 얘기를 스스럼없이 하는 경우가 생겨요. 근데 그게 많이 보다 보니까 인사팀은 사람을 뽑아야 되는데 사람이 급하다고 빨리빨리 뽑아야 되니까 급할 때는 그냥 정말 눈에 띄는 사람을 뽑게 돼 버리거든요. 그러다 보니까 문제가 생기는 거예요. 어떻게 저런 사람? 그것도 모르는 애를 어떻게 뽑았어 그러니까 채용팀에서는 항상 피로도가 있어요. 그 피로도를 AI가 서류라도 그러니까 거기서 많이 걸러지면 채용자 입장에서 굉장히 편한 거죠."(50대, 응답자F)
"특히나 채용 담당자의 입장에서 말씀하신 것처럼 인사팀이 매해 대기업 같은 경우에는 정기적으로 수천 명씩 서류 전형을 해야 되잖아요. 그렇게 기계적으로 하는 것들은 AI가 뭔가 더 효율적으로 정해진 기준에 맞춰서 할 수 있는 역량이 더 뛰어나다 사람보다는."(50대, 응답자A)

자료: 연구진 작성

그러나 AI 채용에 대한 부정적인 측면으로는, 틀에 박힌 인재가 아닌 독창적인 인재를 선별할 수 있을지에 대한 의문을 보였다. 또한, 기존 방식으로도 문제가 없었는데 AI로 대체된다면 일자리의 위협을 받을 수 있을 것이라 생각하는 경향을 보였다.

〈표 2-17〉 채용자 관점에서 기존 채용 방식 선호

"AI 툴에 의해서 하면 딱 정해진 틀에서 사람을 뽑는다는 기분이 드는데 제가 봤을 때 요즘 세계적인 글로벌 기업 같은 경우는 정해진 사고방식 보다는 진짜 독특하고 창의적이고 이런 인재를 원하는데 거기에 원하는 사람들을 AI 툴로 제대로 보고 뽑을 수 있나 그게 의문점이 들어서요."(40대, 응답자F)
"지금까지 저는 회사에서 인재 채용할 때 기존 방식에 있었어도 문제되는 게 하나도 없었던 거 같아요. 그리고 또 하나 제가 만약에 채용관 입장이라고 하면 회사가 AI를 도입하면 시간은 분명히 절약되겠죠. 그럼 과연 저랑 같이 일하는 사람이 예를 들어 이 6명이 같이 채용관 입장에서 한다면 과연 이 6명 중에 회사에서 몇 명을 잘라야 되지 않을까요?"(40대, 응답자E)

자료: 연구진 작성

단계별로 보았을 때는 지원자 관점과 유사하게 서류, 인적성 단계는 AI가 하고 마지막 면접 단계는 기존 방식을 선호하는 것으로 나타났다. AI가 정량적인 기준에 대해서는 빠르고 정확하게 처리할 수 있기 때문에 서류 및 인적성 단계는 AI가 적합하다는 의견이 주를 이뤘다.

단계를 불문하고 기존 방식을 고수하는 입장에서도 AI의 기술력과 정확도에 대한 의구심이 있었으며, 반대로 기존 채용 방식에서도 오류는 나타났기 때문에 AI를 통한 채용이 오히려 정확하고 새로운 정보를 얻을 수 있을 것이라는 기대도 있었다.

AI 기술력이 아직은 신뢰할 수 있는 수준에 도달하지 못했기 때문에 마지막 단계에서는 기존 방식으로 최종 합격자를 선별하는 과정이 필요하다는 의견도 있었다.

〈표 2-18〉 채용자 관점에서 채용 단계별 선호

"서류는 AI가 저는 거의 확실하다고 봅니다. 왜냐하면 자기가 거짓말을 써도 AI가 그걸 찾아낼 수 있을 거예요. 왜냐하면 AI가 하다못해 챗지피티로 뭘 하잖아요. 거기서도 제가 황당하게 한 걸 개가 찾아내더라고요. 그런 걸 굉장히 많이 받기 때문에 최소한 서류적인 거를 걸러내는 거는."(60대, 응답자B) *저는 1, 2, 3단계 전부 다 기본 방식을 고수를 하는 입장이고요. 1, 2단계는 AI는 아니겠지만 현재의 방식이 있을 거고 그대로 유지를 하고 3단계는 AI를 하는 거에 대한건 아까 말씀드렸던 다른 이유들로 인해서 AI의 실수나 아니면 아직은 AI 기술이 그만큼 고도화가 안 됐다 라고 보고 있고 그 이유는 데스크 리서치나 이런 거 할 때도 실제로 모르는 거에 대한 내용을 물어봐서 얻은 정보를 가지고 개발이나 이런 걸 했을 때 실제로 나중에 시간이 지나서 알게 된 정보랑 미스매치 되는 경우도 봤고 그러다 보니까 아직은 시기상조지 않나라는 생각이 있어서 현재까지는 기존 방식으로 모든 단계를 다 선호합니다."(40대, 응답자B)* *"저는 다 AI가 했으면 해요. 왜냐하면 제 경험으로 비춰봤을 때 얘기 드린 거니까 면접으로 잘 뽑았어도 중도 퇴사하는 사람들도 많고 그렇기 때문에 더 기술력이 과도하게 발달한다면은 저는 면접이 오히려 더 공정하고 지원자가 다를 수도 있기 때문에 센스 있을 거라고 생각하거든요."(30대, 응답자C)*

자료: 연구진 작성

3) 윤리 가치별 AI 채용 서비스의 긍·부정 영향 논의

국민포럼단은 프라이버시 보호, 포용성, 책임성, 투명성, 공정성의 주요 윤리 가치를 기준으로 AI 채용 서비스의 영향에 대해 논의하였다. 각 가치별로 긍정적 효과와 부정적 위험 요인을 균형 있게 검토하였다.

(1) 프라이버시 보호

프라이버시 보호와 관련한 긍정적인 효과로 직무와 무관한 불필요한 정보 제공을 줄일 수 있다는 점을 기대하고 있었으며, AI가 자동으로 민감정보를 삭제하거나 비직무적 정보를 선별적으로 배제함으로써 개인정보 보호를 더 철저히 할 수 있을 것이라고 생각하였다.

〈표 2-19〉 프라이버시 보호에 대한 긍정 영향(1)

기존 블라인드 채용보다 좀 더 강화해서 아예 사진이랑 이름, 학번 이런 걸 다 제외하고 개발자를 뽑을 때는 소프트웨어 코딩 능력 이런 게 더 중요하다고 생각을 해서... 완전 역량 중심으로 평가를 하면 개인 정보가 유출될 일이 없고... (20대, 응답자A)
AI는 사전에 자기가 미리 세팅해 놓은 데이터에만 집중하기 때문에 불필요한 개인적인 그런 수집 내용에 대해서는 관심이 없을 것 같아요... (60대, 응답자C)
이력서를 보통 서류 파일로 받는 회사 같은 경우는 관리가 안 되는 경우가 많고 인사 담당자 PC를 보면 들어가 있는 경우도 되게 많은데 AI를 통해서 관리하면 특정 기관에 제출했던 건 1년 지나면 삭제되도록 이런 식으로 하면 관리가 굉장히 깔끔할 것 같아서... (20대, 응답자G)

자료: 연구진 작성

데이터 접근 권한을 제한적으로 두거나 접근 기록을 모두 남김으로써 프라이버시 보호 제고 효과를 기대하기도 하였다. 또한, 면접관을 면대면으로 만나는 것이 아니므로 프라이버시 침해 우려가 낮아지고, 지원자 간 비교가 어려워져 더 객관적인 평가가 가능하다는 장점을 꼽기도 하였다.

〈표 2-20〉 프라이버시 보호에 대한 긍정 영향(2)

정보를 열람했는지 기록이 남기 때문에 이력서가 여러 곳으로 흘러가는 위험을 줄일 수 있을 것 같아요. AI가 하는 것이기 때문에 기계화, 자동화 시스템이잖아요. 그러니까 함부로 사람들이 건드리지 못한다고 저는 생각했거든요... (60대, 응답자D)
현재는 모든 면접관들이랑 인사 담당자들이 직무 수행이랑 관련이 없는 개인정보 같은 것들을 다 볼 수 있는 구조인데 AI가 담당자별로 제한을 둘 수 있게 함으로써 그런 프라이버시 유출 여지를 사전에 차단할 수 있지 않을까... (20대, 응답자B)
실제 면접관들을 만나면 그 이름과 얼굴을 기억하기 때문에 프라이버시 침해 우려가 있지만 AI 채용 서비스로 만났을 때는 내 정보를 삭제할 수 있다는 선택이 있을 수도 있을 것 같아서... (30대, 응답자B)

자료: 연구진 작성

반면, 생체 정보 수집 및 오남용, 해킹, 범죄 활용 등에 대한 우려가 매우 크게 나타났으며 범죄가 아니더라도 마케팅 용도로 활용되거나 예상치 못한 곳에 데이터가 노출될 수 있다는 우려를 보이기도 했다.

채용 과정에서 SNS 활동 기록 등 제공하지 않은 데이터까지 수집 및 활용하여 분석할 것 같다는 두려움이 존재하기도 했다. 전반적으로 프라이버시 보호에 있어서는 부정적 영향 최소화가 최우선이라는 의견이 다수를 차지하였다.

〈표 2-21〉 프라이버시 보호에 대한 부정 영향

목소리나 이런 거 유출됐을 때 보이스피싱 등으로 활용될 수 있고 동공이나 표정 이런 걸로 휴대폰 잠금장치 해제라든지 도어록 해제, 금융 앱 로그인 할 수 있는 범죄에 악용할 수 있는 가능성이 제기되는 바이고요. 그리고 제3자 데이터를 공유하거나 해킹, 매매 이런 거를 타 기업 등으로 판매나 유출도 할 수 있다는 점에서 우려... (40대, 응답자C)
화상으로 AI 면접을 보는 것 같은데 그거는 수집되는 정보들이 얼굴, 목소리로 고품질의 괜찮은 데이터들이고 깨끗한 데이터들이어서 이런게 마케팅이나 연구나 외부에 활용되지 않을까라는 걱정이 많긴 해요... (20대, 응답자G)
혹시나 제출되는 이메일을 기반으로 AI 채용 서비스가 제 이메일을 기반으로 SNS 인스타그램이나 이런 것들 들어가 가지고 개인 정보를 연동해가지고 제 취미생활이나 제가 여행했던 기록이나 여러 가지 생활 패턴이나 뭘 먹었는지 이런 것까지도 저를 가지고 파악을 한다라고 한다면 과도한 정보 수집이 될 수 있을 것 같아요... (50대, 응답자G)

자료: 연구진 작성

(2) 포용성

포용성 관련 긍정적인 효과는 모두에게 동일한 기회를 줄 수 있고, 언어, 거리, 장애 등 다양한 장벽을 뛰어넘을 수 있다는 점을 꼽았다. 그에 따라서 채용 응시에 편리함이 증진되는 것뿐 아니라 인종, 장애 유무와 관계없이 모두 동일하게 전형을 진행할 수 있기 때문에 경제적인 효과도 누릴 수 있을 것이라 생각했다.

〈표 2-22〉 포용성에 대한 긍정 영향(1)

자막이나 음성 안내해서 장애인 분들도 지원을 수월하게 할 수 있지 않을까 하는 부분이랑 온라인이기 때문에 시골이나 해외에 있어도 갈 필요 없이 응시 가능하다는 게 긍정적인 것 같습니다.... (20대, 응답자C)
어떤 사람이 가지고 있는 그런 배경이나 이런 걸 다 떠나서 객관적인 그 사람의 실력을 본다는 거기 때문에 그런 감정적인 부분이 배제가 될 수 있기 때문에 굉장히 객관적이면서 지원자들 그룹을 명확하게 평가할 수 있는 게 굉장히 장점이다 생각해요... (50대, 응답자F)

자료: 연구진 작성

아울러, AI가 개인이 가지고 있는 역량들을 조합하고 분석하여 인간이 발견하지 못한 장점이나 적합한 직무를 추천하고 제안함으로써 포용성을 더 높일 수 있다는 의견을 제시하기도 하였다.

〈표 2-23〉 포용성에 대한 긍정 영향(2)

사람이 가진 그 방대한 자료들 예를 들어서 자격증이랄지 대학에서 어떤 과목을 들었는지 저희는 사실 대학의 학점으로 주로 평가를 했다면 이 사람이 들었던 수업의 히스토리를 분석을 해 준다든지 이 사람이 가졌던 경험들에 대해서 AI는 폭넓게 이거를 분석을 하고 오히려 추천을 해 줄 수도 있을 거라는 생각이 들더라고요. 그래서 사람으로서는 다 볼 수 없는 자료들, 이 사람이 그런 정보들을 분석해서 추천을 해주면 더 포용성 있게 사람을 선발할 수 있지 않을까... (40대, 응답자A)
사람이 했을 때 하고 그래서 사람이 했을 적에는 뽑히지 않았을 그런 인재를 뽑아낼 수 있는 정말 숨은 인재를 발굴할 수도 있다 라는 게 긍정적인 면일 거라고 생각해요... (60대, 응답자F)

자료: 연구진 작성

반면, 부정적 효과로는 AI를 학습하는 데이터가 기존의 편향된 데이터를 기반으로 한다면 사회적 약자들이 오히려 피해를 입을 수 있다는 점을 꼽았다. 또한, AI 기준의 획일화로 인하여 조직 다양성이 저해되는 점을 우려했다.

〈표 2-24〉 포용성에 대한 부정 영향(1)

인재상 만들어놓고 거기에 부합하지 않으면 탈락돼 버리는 거죠. 그리고 환경이 열악한 곳에 거주하거나 상황이 좋지 않은 지원자는 AI 채용 시스템 자체를 이용하지 못하는 경우가 있을 수 있을 것 같아요.... (30대, 응답자F) *물론 그 기업이 원하는 인재상이 물론 있겠지만 AI가 스스로가 어떤 기준을 잡고 만들어 가는데 이 기준이 다양한 사람들에 대한 포용성 이런 것이 배제되고 어떤 획일적인 기준으로만 갖다 대면 기업의 입장에서는 그냥 똑같은 인형을 찍어내듯이 회사라는 게 사실은 인재상도 중요하지만 다양한 사람들이 다양한 아이디어를 모아서 운영하는 게 기업인데 그런 부분들이 배제되지 않을까... (50대, 응답자A)*

자료: 연구진 작성

언어로 표현되지 않는 부분에 대한 평가가 어려울 것이라 생각했으며, 전형적이지 않은 패턴을 가진 지원자 평가가 정상적으로 이루어지지 않을 수 있다는 위험을 지적하였다.

〈표 2-25〉 포용성에 대한 부정 영향(2)

근데 의료 분야의 경우에는 말 더럽게 못하는데 수술은 무지하게 잘하는 놈이 있을 수 있어요. 실제로 많이 봤어요. 그런 의사들이 있어요. 실제로. 말은 정말 못해요. 그런데 메스를 잡으면 무지하게 잘해요. 그거를 그러면은 어떻게 판단할 거냐... (60대, 응답자B)

자료: 연구진 작성

(3) 책임성

책임성 관련 긍정적 효과는 평가 과정이 모두 기록되기 때문에 문제가 있을 때 그 원인을 찾을 수 있고, 책임 소재를 규명할 수 있으며 수정이 가능하다 점이 다수 언급되었다. 책임성 측면에서 부정적 효과 최소화보다 긍정적 효과 극대화를 우선해야 한다는 소수 의견이 있었으며, 이는 책임성을 분명히 함으로써 다른 윤리 가치 또한 함께 제고될 수 있다는 이유에서 비롯되었다.

〈표 2-26〉 책임성에 대한 긍정 영향

모든 평가랑 그 과정에 채용 과정의 모든 평가나 의사 결정 같은 것들을 자동으로 기록하고 문제가 있을 때 추적하면 잘못된 결정에 대해서 명확한 책임 소재를 밝힐 수 있으니까... (20대, 응답자B)
채용 서비스를 만드는 단계부터 어떤 책임 주체를 명확히 하고 검증 체계도 마련할 수 있다 보니까 어떤 지원자한테 발생할 수 있는 피해를 막고 오류가 발생했을 때 그 책임 소재를 명확히 할 수 있을 것 같아서라고 썼고요. 두 번째는 만약에 오류가 발생했을 때 그걸 개선할 수 있으니까요... (30대, 응답자E)
책임 주체를 명확히 할 필요가 있는 게 더 중요하다. 그래야지 다른 성격들 투명성이나 공정성도 따라올 것 같다... (30대, 응답자E)

자료: 연구진 작성

반면, 부정적 효과로는 책임을 누가 질 것인지 명확하지 않고, 문제가 발생했을 때 오히려 책임을 떠넘기는 부작용이 발생할 수 있다는 우려가 강했으며, AI 결과에 대한 무비판적 수용을 우려하였다.

〈표 2-27〉 책임성에 대한 부정 영향(1)

시스템을 도입하는 업체, 기업. 두 주체 간에 시스템 운영이나 관리, 데이터 보안이나 오류 같은 거에 대해서 역할이나 책임이 불분명하면 채용 과정에서 발생하는 문제들에 대해서도 신속하게 대응이 어려울 것 같고 그렇다 보면 책임 소재도 불분명해져서 이런 책임성에 있어서 회피가 있을 것 같다고 생각했고... (20대, 응답자B) *저는 기업이 그런 채용 의사 결정에 대해서 책임을 AI한테 떠넘기면 AI가 이렇게 판단을 해서 넌 떨어진 거다 이런 식으로 핑계를 댈 수 있다. 만약에 혹시나 오류나 버그 같은 게 나오면 그것도 그 업체들 간에 그런 서로 계속 책임을 미루는 게 있을 것 같다... (20대, 응답자E)* *그리고 또 무비판적으로 수용하는 경우도 있을 것 같아요. AI 결과를. 예를 들면 AI가 이렇게 했으니까 이 절차대로 따랐다 이런 식으로... (20대, 응답자B)*

자료: 연구진 작성

책임성 측면에서 부정적 효과를 최소화하는 데 힘써야 한다는 의견이 두드러졌다. 부정적 효과가 불러일으킬 문제들이 피해가 명확하고, 아직은 법이나 제도적으로 미흡한 부분이 있어서 보완이 필요하기 때문이다.

〈표 2-28〉 책임성에 대한 부정 영향(2)

각각의 입장을 다 만족시키고 100% 만족을 못하더라도 책임 소재에 대한 명확한 선을 긋기 위해서는 국가에서 어떤 일정한 정책이나 아니면 규정 같은 게 있어야 될 것 같고요. 기업도 마찬가지로 어떠한 일이 벌어졌을 때 기업이 책임진다, 지원자가 책임진다 또는 그 서비스를 제공한 회사가 책임을 진다 그런 거에 대한 명확한 근거 그런 게 필요하지 않을까 싶습니다... (50대, 응답자E)

자료: 연구진 작성

(4) 투명성

투명성 관련해서는 평가 과정이 모두 자동으로 기록되고 근거가 남기 때문에, 채용 결과에 대한 사후 검증과 오류 추적이 가능하고 문제가 발생했을 때 원인을 규명하거나 수정할 수 있다는 점에서 긍정적인 영향을 기대하고 있었다.

〈표 2-29〉 투명성에 대한 긍정 영향(1)

AI 면접 같은 경우는 아무래도 답변 내용이나 과정이 모두 녹취나 기록될 텐데 그걸 근거로 몇 점이 나왔다라는 근거가 남아서 사후 검증까지도 기록된 거를 가지고 그 채용된 사람의 사후검증 까지도 할수 있지 않을까... (40대, 응답자B)
투명하게 기준과 결과가 모든 지원자에게 동일하게 제공되기 때문에 투명성을 제고할 수 있고 채용절차 전 과정을 AI가 기록하고 저장함으로써 대외적으로 그리고 대내적으로 공정성과 정당성을 확보할 수 있고... (50대, 응답자E)

자료: 연구진 작성

AI가 산출한 결과와 평가 근거를 지원자에게 명확히 설명하고 피드백 형태로 제공할 수 있다면 이를 통해 지원자는 평가 과정에 대해 납득하고 신뢰를 높일 수 있으며, 채용 절차가 단순 결과 통보가 아닌 상호소통적 과정으로 인식할 것이다.

또한, 평가자의 주관적 판단이나 인맥·관계 등 외부 요인의 개입을 최소화하고, 채용 과정 전반을 기준 중심·절차 중심으로 운영할 수 있다는 장점도 언급하였다.

〈표 2-30〉 투명성에 대한 긍정 영향(2)

지원자들이 자신의 결과에 대해서 어떻게 투명하게 이것을 평가를 했는지를 정보에 대한 공개 청구를 했을 때 신속하게 답변을 할 수 있다라는 게 되게 좋은 것 같고요... (40대, 응답자A)
회사가 사람 뽑을 때 특정한 기준을 세웠다면 그거에 도달하는 사람을 투명하게 뽑을 수 있다. 사람에 따라 사람에 의존해서 사람의 주관적인 판단에 의해서 판단하는 것이 아니고 정해진 그 투명한 절차에 따라서 투명한 규정에 의해서 그 기준에 의해서 뽑을 수 있는 게 투명성에서 장점이라고 생각했습니다... (40대, 응답자A)
지인 등의 취업 청탁 같은 거나 상부에서 압력이 있다거나 누군가의 채용을 위한 그런 상황에 직면하지 않아도 된다는 그런 점에서 투명성이 더 좋을 것 같고 고착화된 인간 관계에서 부득이 오가는 선례들 있잖아요. 그런 부분을 없앨 수 있는 계기가 될 수 있지 않을까 싶습니다... (40대, 응답자C)

자료: 연구진 작성

반면, 부정적 효과로는 AI 모델의 구조적 특성상 왜 합격되었는지, 혹은 불합격되었는지를 명확히 설명하기 어렵고, 모델 성능과 설명력 간의 트레이드오프가 존재하는 점을 강조했다.

〈표 2-31〉 투명성에 대한 부정 영향(1)

AI의 성능과 그 AI 결과에 대한 설명력이 거의 트레이드 오프 단계예요... (중략) ...어떤 모델을 사용을 할지는 아직 모르겠지만 이 내재적인 모델의 한계로 인해서 저게 실현이 불가능할 수 있다... (20대, 응답자F) *AI가 판단하는 근거를 사람도 이해가 돼야지 서로 그게 맞을 텐데 그게 어려우니까 지원자 입장에서는 못 믿겠다 싶을 수 있을 것 같아요... (30대, 응답자E)*

자료: 연구진 작성

또한, 기업에서 데이터를 제공하지 않으면 투명성을 확인할 수 없으며, 구체적인 정보를 제공하지 않은 이상 문제가 발생하거나 결과를 납득하기 어렵더라도 지원자가 이의제기를 하는게 불가능하다는 점을 우려하고 있었다.

〈표 2-32〉 투명성에 대한 부정 영향(2)

AI가 판단할 때 피드백이 꼭 온다는 보장은 없다고 생각해요. 어차피 AI가 피드백을 꼭 줘야 되는 의무는 없어요. AI가 그렇게 프로그램을 짰을 때 피드백을 꼭 줘야 된다는 의무는 없다고 보거든요. 근데 그렇게 됐을 때 AI가 합격, 불합격으로 판단했을 때 나중에 예를 들어서 지원자가 AI한테 물어볼 수는 없잖아요. ... (40대, 응답자E) *결과에 대해서 이의 제기하거나 어떤 채용 결과가 왜 그렇게 나왔는지를 알 수가 없다면 지원자 입장에서는 이의 제기할 것도 없을 것 같고 개선 방향을 찾기도 힘들 것 같아요... (30대, 응답자E)*

자료: 연구진 작성

(5) 공정성

공정성 관련한 긍정적 영향으로 문제점이 감지되었을 때 빠른 수정이 가능하고, 인간이 가지고 있는 무의식적인 편견을 제거하고 평가를 할 수 있게 됨에 따라 모두가 공정한 채용 절차를 밟을 수 있다는 점이 강조되었다. 채용을 진행하는 기업 입장에서도 지원자의 허위 사실 기재나 거짓 응답 등을 거를 수 있어서 공정한 채용이 가능하다는 장점이 있다.

〈표 2-33〉 공정성에 대한 긍정 영향

문제점이 감지가 되고 태클이 들어왔을 때 알고리즘을 개선할 수 있다는 게 가장 공정성에 크게 영향을 미친다고 생각했고 특히 사람은 원래 그랬어 하면서 식별하지 못했던 많은 정보들을 이걸 한번 꼬집어주는 요소가 있으면 근데 우리 그거 왜 그랬었지? 라고 하면서 사람도 반성하게 되고 그거를 수정할 수 있어서 그래서 그게 조금 긍정적이라고 생각했어요... (20대, 응답자D)
사람들이 무의식적으로 갖게 되는 그런 선입견이나 편견이 있잖아요. 그런 거에 대한 편향을 조금 줄일 수 있고 내가 낸 평가 기준을 적용해서 공정하게 평가를 할 수 있겠다 한 겁니다... (20대, 응답자E)
스크리닝 단계에서 나이나 사진 등으로 편향된 인재들이 생길 텐데 그거에 영향받지 않고 온전히 직무 관련된 경험과 역량만으로 균등한 평가가 가능하다 그리고 면접관마다도 역량들이 다를 텐데 면접관마다 다른 역량으로 인한 질문의 난이도나 기준이 다를 때 지원자가 겪을 수 있는 불공정성에 대해서 사전에 차단이 가능해지기 때문에... (40대, 응답자B)
심리적으로 거짓말을 할 때라든지 이럴 때 눈이 조금씩 올라간다거나 그런 어떤 드라마를 본 것 같아요. 눈을 깜빡거리거나 눈이 동공이 흔들린다거나 이런게 AI가 분명히 누적이 됐을 거니까 그런 거에 있어서 객관적 데이터로 이 사람이 거짓말하는지 이상한 사람인지 판단할 수 있지 않을까. 미세한 것들을 컴퓨터는 기술이 고도화된다면 체크가 가능하지 않을까... (40대, 응답자D)

자료: 연구진 작성

반면, 부정적 효과로 편향과 차별 문제가 가장 강조되었다. 기존 데이터를 학습하게 된다면 편향된 데이터를 학습하는 것이나 마찬가지이며, 이는 오히려 공정성을 저해할 수 있다.

〈표 2-34〉 공정성에 대한 부정 영향

표준화되지 않은 데이터가 특정 집단을 대표했을 때 그 집단에게 오히려 불리하게 작용되면 공정성을 저해할 수 있다... (30대, 응답자B)
데이터베이스 자체가 과거의 기록을 결론적으로 입력하는 건데 그거를 뒤바꿔는 데이터를 입력하지는 않다고 봐요. 그럼 결론적으로 사람이 지금까지 뽑아왔던 과정 자체를 결론은 AI가 채용한다고 보기 때문에 저는 결론적으로 이런 장점 학연, 지연, 동양인 뺀 나머지 거에 대해서는 다 일반적인 이런 거는 결론적으로 과거 데이터가 그렇기 때문에 쭉 이어질 거라는 거죠... (40대, 응답자E)

자료: 연구진 작성

4) 주체별 노력 및 책임

(1) 정부

현재 정부 대응 수준에 대해서는 전혀 대비하고 있지 않은 것 같다는 의견이 주를 이뤘으며, 관련 법안이나 제도에 대해 접해 본 적이 없고, AI와 관련된 다른 범죄들에 대해서도 대비가 되고 있지 않다고 느끼고 있기 때문에 상대적으로 중요도가 낮은 AI 채용에 대해서는 더욱 대비를 하지 않을거라고 생각했다.

〈표 2-35〉 AI 채용 서비스에 대한 정부 대응 수준 인식

심각도가 높은 것도 지금 너무 대응이 안되고 있는데 당연히 이런 거에 대한 대응도 낮을 거다... (20대, 응답자A)
대응 자체를 그렇게 정부가 하고 있다고 생각하지 않아요. 그리고 3, 4년 전부터 봅고 있다고는 하지만 대학교에도 AI 채용 연습할 수 있는 가상 공간이 만들어지긴 했는데 실질적으로 많이 사용하고 있다고는 저는 보지 않거든요. 그런 거에 따라서 정부 차원에서도 그냥 형식상으로만 하고 있구나 라는 생각이 들더라고요... (30대, 응답자A)
기업에서 주도적으로 하는 거지 정부가 하기에는 차원이 일단은 다를 것 같아요. 기업에서 먼저 주도적으로 선제적으로 하면 정부가 서포팅은 하지만 이걸 먼저 할 수 있는 그런 범위는 아닌 것 같아요... (50대, 응답자F)
하기는 하겠죠. 그게 성과가 나타날런지 일단 정부도 정부지만 일선에서 회사들이나 개발 업체 여기서부터가 좀 더 신경을 써야 될 것 같아요... (60대, 응답자F)

자료: 연구진 작성

정부의 역할에 대해서는 규제보다는 기술이 발전할 수 있는 환경을 마련, 우선 시행 기업에 혜택 제공 및 활성화 등이 정부가 할 일이라는 의견을 제시했다. 또한, 정부는 개발사나 운영사가 바른 방향으로 나아갈 수 있도록 처벌 기준을 수립하는 역할로 인식되고 있었다.

〈표 2-36〉 정부의 책임과 노력

먼저 기술적인 지원을 할 수 있는 그런 여건을 먼저 만들고 전폭적인 지원을 먼저 하는 게 맞지 … (중략) …미리 부정적인 차단 가능한 요인들을 다 막아놓고 기술 개발 하라는 느낌이라 그러면 인식이 더 안 좋을 거라 먼저 전폭적인 지원을 한다는 뭔가 표명이나 금액적인 인상이나 이런 게 필요할 것 같아요… (40대, 응답자B)
정부가 기업한테 시스템을 도입하면 어떤 베네핏을 주는 거죠. 활성화, 보편화시키기 위해서… (60대, 응답자C)
이익 집단은 돈을 버는 게 주 목적이지 선을 추구하는 건 사실 너무 허울뿐인 것처럼 들릴 수도 있는 사람들이잖아요. 근데 정부 같은 경우에는 공공에 가치를 수호할 책임이 있는 집단이고 그럴 힘도 가장 센 집단이다 보니까 가장 빠르고 분주하게 대처하고 움직여야 되는 집단이라고 생각을 합니다… (20대, 응답자F)
자율적으로 맡기되 문제가 발생하면 그거를 처벌하는 방식이 나을 거 같아요. 기업을 너무 옥죄면 이게 유연성이 떨어지다 보니까 정부가 시키는 대로 하면 끌려갈 수밖에 없고 뭔가 발전적이지 않은 방향으로 갈 것 같아서 못할 때 처벌하는 정도로만… (30대, 응답자F)

자료: 연구진 작성

(2) 개발사 및 운영사

개발사의 경우, 시스템을 공급할 때 해당 서비스가 가지고 있는 문제점에 대해 충분히 공유해야 하며, 문제 발생 시 즉각적으로 수정해야하는 책임이 있다는 의견이 있었다.

〈표 2-37〉 AI 채용 서비스 개발사의 책임과 노력

어쨌든 개발한 기업 자체가 알고리즘이라든지 데이터 편향 이런 문제도 얘기를 했잖아요. 그런 문제에서 신경을 써야지 개발자들이 오히려 더 신경을 써야 되지 않나 생각을 했습니다.... (20대, 응답자C)
상호적으로 해야 된다고 보는 것 같아요. 이 회사에서도 뭔가 오류나 이런 문제점에 대한 인식에 대한 피드백을 같이 하고 그럼 그 회사의 개발자도 그 부분에 있어서 계속 업데이트를 하고 그거가 계속 상호적으로 가야 이게 개선이 될 것 같아요... (40대, 응답자C)
소스를 공개할 수는 없겠지만 최소한 이거를 발주한 사람한테는 소스를 줘야죠. 어떻게 만들었다... (60대, 응답자C)

자료: 연구진 작성

운영사의 경우, 해당 서비스를 도입하여 사용하겠다고 최종 결정한 곳이기 때문에 큰 책임이 있다는 의견이 주를 이루었으며, 도입한 기업은 서비스에 대한 충분한 검증이 필요하며 개발사와 논의를 통해 기준을 세워야 한다고 생각했다.

〈표 2-38〉 AI 채용 서비스 운영사의 책임과 노력

어쨌든 제가 한 건 이 회사고 그런 개발자나 다른 정보가 어떤 데서 담당을 하고 있는지까지는 모르잖아요. 그런 면접을 보는 사람이. 직접적으로 보는 기업에서 책임을 지는 게 맞다고 생각해요... (20대, 응답자E)
결론은 기업이 오케이 그래도 우리는 할 거야 그래서 오케이 했기 때문에 개발자한테 그걸 샀을 거라고 보거든요. 결론은 그렇기 때문에 개발자가 잘못한건 없죠... (40대, 응답자E)
저 시스템을 쓸지 안 쓸지 리스크를 안고도 쓰겠다 라고 판단하는 것도 결국은 기업에서 하는 거기 때문에 기업에서 당연히 가장 책임이 크다 라고 생각을 하고요... (50대, 응답자D)

자료: 연구진 작성

(3) 채용 지원자 개인

반면, 지원자 개인이 할 수 있는 노력은 많지 않다고 응답자들은 평가하였다.

〈표 2-39〉 채용 지원자 개인의 책임과 노력

채용에서 대부분 지원자가 을의 위치이기 때문에 노력한다고 할 수 있는 거 없죠. 저희가 1년만 보관하다가 폐기해 주세요 이래도 안 해요... (20대, 응답자G)
확인할 수도 없어요. 뭔가 증거가 있어야지 뭘 써야 될 텐데 기업이 마음먹고 투명성 아까 저희가 얘기한 거 그거 위배해 버리면 지원자는 답이 없거든요... (30대, 응답자F)
자신의 권리를 보호하기 위해서는 아까 소비자 보호원 말씀드렸지만 이런 AI 채용 서비스 관련해서 불이익을 받았을 때 어떤 데를 찾아가서 그게 정부기관이 됐던 지방자치단체가 됐던 민간 단체가 됐던 그런 거를 확인하는 것도 한 방법이지 않을까... (50대, 응답자A)

자료: 연구진 작성

3. 전문가 평가 주요 결과

앞서 살펴본 국민포럼단 FGI 조사와 병행하여, 전문가 평가단을 대상으로 한 평가를 실시하였다. 전문가 평가는 산업계·학계·법조계·공공기관·시민단체 등 다양한 분야 전문가로 평가단을 구성하고, AI 채용 서비스의 윤리적 영향을 정량·정성적으로 분석하는 것을 목표로 하였다.

가. 전문가 평가단의 구성

전문가 평가단은 '제4기 AI 윤리·신뢰성 포럼' 위원장인 연세대학교 문명재 교수를 평가단장으로 하여 총 28명의 각계 전문가로 구성되었다. 평가단은 '제4기 AI 윤리·신뢰성 포럼' 위원 21명과 산·관·학·연·시민사회 전문가 7명으로 이루어졌으며, 구체적인 구성은 다음 〈표 2-40〉과 같다.

〈표 2-40〉 전문가 평가단의 구성

구분	성명	소속
평가단장	문명재	연세대학교 행정학과 교수
산업계	김동환	포티투마루 대표
	김민성	한국IBM 상무
	김유철	LG AI연구원 전략부문 부문장
	문병순	카카오모빌리티 경제연구소 소장
	박선민	구글코리아 대외협력정책 상무
	송대섭	네이버 이사
	이영복	제네시스랩 대표
	조장래	비트코퍼레이션 고문
	하주영	스케터랩 변호사
공공	김명주	AI 안전연구소 소장
	안성원	SPRi AI정책연구실 실장
	양승엽	한국노동연구원 사회정책연구본부 부연구위원
	이현숙	한국과학창의재단 지역과학문화실 실장
학계	문광진	국립목포대학교 법·경찰학부 교수
	박성필	KAIST 문술미래전략대학원장
	변순용	서울교육대학교 윤리교육과 교수
	선지원	한양대학교 법학전문대학원 교수
	이상욱	한양대학교 철학과 및 인공지능학과 교수
	이원태	국민대학교 특임교수
법조계	김도엽	김앤장 법률사무소 변호사
	노태영	김앤장 법률사무소 변호사
	손지원	법무법인 혁신 변호사
	이동익	법무법인 선운 대표변호사
시민사회	이정수	한국소비자단체협의회 사무총장
	이지은	참여연대 공익법센터 선임간사
	정지연	한국소비자연맹 사무총장
국제기구	김은영	UNESCO 한국위원회 의제정책센터장

자료: 연구진 작성

나. 평가의 절차와 방법

전문가 평가는 28명을 대상으로 2025년 8월 1차 평가, 9월 2차 평가 두 차례에 걸쳐 서면 조사 방식으로 진행되었다. 2025년 AI 윤리영향평가는 AI 채용 서비스의 특성과 실제 활용 맥락을 반영하여, 국가 「인공지능(AI) 윤리기준」의 10대 핵심요건 중 관련성이 높은 항목들을 중심으로 평가 영역을 재구성하였다. (〈표 2-41〉) 특히, AI 채용에서 중요하게 다뤄지는 '공정성' 항목을 별도 영역으로 추가하였다. 그 결과, 평가단은 AI 채용 서비스의 설계·개발·배포·활용 전 과정에서 발생할 수 있는 윤리적 영향을 ① 프라이버시 보호, ② 포용성, ③ 책임성, ④ 투명성, ⑤ 공정성 5개 영역으로 설정하고, 각 영역별 긍·부정 영향을 식별 및 검토하는 정량·정성 평가를 수행하였다. 구체적인 전문가 평가의 절차와 내용은 〈표 2-42〉와 같다.

〈표 2-41〉 평가 영역

평가 영역	내용
프라이버시 보호	개인정보의 안전한 수집·활용·관리와 데이터 주체의 통제권 보장
포용성	모든 지원자가 차별 없이 접근하고 평가받을 수 있는 환경 조성
책임성	기록·감독·검증을 통한 책임추적성과 사후책임 확보
투명성	판단과정·결과에 대한 공개, 설명, 피드백을 통한 신뢰 형성
공정성	편향 최소화와 일관된 기준을 통한 기회의 평등 보장

자료: 연구진 작성

〈표 2-42〉 전문가 평가 절차와 방법

구분	1차 서면 평가		2차 서면 평가
평가 단위	5개 AI 윤리 영역		1차 서면 평가로 도출한 31개 긍·부정 영향 유형
평가 항목	**긍·부정 윤리 영향 서술**	▶	**1. 긍·부정 윤리 영향 평가** * 평가 항목 (공통) 영향을 받는 대상(서술형) (긍정) 규모, 지속 기간, 범위, 발생 가능성 (부정) 규모, 지속 기간, 범위, 발생 가능성, 해결 가능성 ** UNESCO '윤리적 영향평가도구 방법론' 평가 지표 참고 **2. 정부지원 및 대응 필요성 평가** * 평가 항목 (긍정) 기대 수준, 정부 지원 필요성 (부정) 우려 수준, 정부 대응 수준

▼

구분	의견 조사
평가 항목	윤리적 AI 채용 서비스 개발·활용을 위한 정부, 기업, 개인의 책임과 노력 서술

자료: 연구진 작성

1차 평가는 2025년 8월(8월 11일~21일) 평가지 작성을 통한 서면 조사로 10일간 진행되었다. 평가 대상인 AI 채용 서비스에 대한 이해를 높이기 위해, 기초 분석 단계에서 작성한 보고서를 평가지와 함께 제공하였다. 1차 평가지의 구성과 문항, 응답 예시는 아래 [그림 2-21]에 제시하였다. 전문가들은 국가 「인공지능(AI) 윤리기준」 10대 핵심요건 중 관련성이 높은 항목을 재구성한 5개 평가 영역을 기준으로, 각 영역에서 나타날 수 있는 구체적인 긍·부정 윤리 영향을 서술하였다.

[그림 2-21] 1차 전문가 평가지 문항 및 응답 예시

1. 프라이버시 보호	
AI 채용 서비스는 「개인정보 보호법」을 준수해야 하며, 지원자의 민감정보 등을 불필요하게 수집하거나 수집 목적을 벗어나 처리하지 않도록 주의해야 합니다. 개인정보는 수집 단계부터 이용 목적, 처리 범위, 보유 기간 등을 명확히 고지하고, 수집·이용·보관·삭제 등 모든 처리 단계에서 안전하게 관리될 수 있도록 설계·운영하는 것이 바람직합니다.	
[긍정적 영향 평가 문항] "AI 채용 서비스"가 개인과 집단의 프라이버시 보호에 긍정적으로 기여할 수 있는 효과는 어떤 상황이나 요인에서 나타날 수 있다고 보십니까?	
프라이버시 보호 영역에 대한 긍정적 영향	• *(작성 예시) 프라이버시 침해는 정보의 접근성이 커질수록 일어날 가능성이 크다. 그런 측면에서 인공지능이 채용과 관련없는 지원자의 정보를 자동으로 가명화처리 하여 인사담당자에게 제공함으로써 기업이나 조직의 개인적 정보접근을 제한할 수 있다.*
[부정적 영향 평가 문항] "AI 채용 서비스"가 개인과 집단의 프라이버시를 침해할 수 있는 위험은 어떤 상황이나 요인에서 발생할 수 있다고 보십니까?	
프라이버시 보호 영역에 대한 부정적 영향	• *(작성 예시) AI 영상 면접 솔루션은 지원자의 얼굴 표정, 목소리 톤, 시선 처리와 같은 생체 및 행동 정보를 포괄적으로 수집하고 분석할 수 있다. 이는 지원자가 자신의 어떤 특성이, 어떤 기준으로 분석되는지 명확히 인지하지 못한 상태에서 사실상의 감시를 당하는 결과를 낳을 수 있으며, 추론된 민감 정보(예: 스트레스 수준, 성격 특성)가 생성되어 정보 주체의 자기결정권을 본질적으로 침해할 위험이 커진다.*

자료: 연구진 작성

1차 평가 결과, 긍정 영향 346건, 부정 영향 342건의 사례가 도출되었으며, 이를 취합·검토한 뒤 사례 간 공통적인 특성을 기준으로 영향 유형을 체계적으로 분류하고 명명하였다. 그 결과, AI 채용 서비스에 관련된 총 31개의 긍·부정 윤리 영향을 정리하였다.

2차 평가는 2025년 9월(9월 5일~17일) 1차 평가 결과를 토대로 설계한 온라인 평가지 기반 서면 조사로 12일간 진행되었다. 1차 평가와 동일한 28명의 전문가가 참여하였으며, 1차 평가에서 도출된 31개 영향 유형을 단위로 각 영향에 대한 정량 평가를 수행하였다. 세부 평가지표는 UNESCO의 '윤리적 영향평가도구 방법론'의 주요 지표를 참고하여 설계하였으며, 구체적 내용은 다음의 〈표 2-43〉과 같다.

또한, 5개 윤리 영역별로 도출된 세부 영향을 고려하여, 각 영향에 의해 직접적·간접적·의도치 않게 영향을 받는 대상을 기술하도록 하였고, 높은 빈도로 언급된 키워드를 중심으로 응답을 정리하였다.

2차 평가 결과에 대해 추가 의견을 수렴하기 위해 2025년 10월(10월 21일~29일) 최종 의견조사를 진행하였으며, 상세한 내용은 제5절에서 다룬다.

〈표 2-43〉 2차 전문가 평가지 평가 항목

평가 항목	평가 내용
긍정·부정 윤리 영향	영향 유형의 규모, 지속 기간, 발생 가능성, 해결 가능성
정부지원 및 대응 필요성	• 긍정 영향 : 영향 유형에 대한 기대 수준, 정부 지원 필요성 • 부정 영향 : 영향 유형에 대한 우려 수준, 정부의 대응 수준
영향을 받는 대상	직접적(주요), 간접적(부차적), 예상치 못하거나 의도치 않은 영향을 받는 대상

자료: 연구진 작성

정량 평가 방식은 다음과 같다. 첫째, 긍정·부정 윤리 영향 평가 시, 긍정 영향의 경우 영향의 ① 규모, ② 지속 기간, ③ 발생 가능성을 4점 척도로 평가하게 하였고, 부정 영향의 경우 영향 유형의 ① 규모, ② 지속 기간, ③ 발생 가능성, ④ 해결 가능성을 4점 척도로 평가하게 하였다. 둘째, 정부 지원 및 대응 필요성 평가 시, 긍정 영향의 경우 영향 유형에 대한 ① 기대 수준, ② 정부 지원 필요성을 4점 척도로 평가하게 하였고, 부정 영향에 대한 ① 우려 수준, ② 정부 대응 수준을 4점 척도로 평가하도록 했다. 2차 평가지의 형식과 문항, 응답 예시는 아래 [그림 2-22]와 같다. 평가지표별 결과는 지표별 평균값을 그래프로 시각화하고, 해당 영향이 속한 영역의 전체 평균값과 비교·분석하였다.

[그림 2-22] 2차 전문가 평가지 문항 및 응답 예시

[1] 긍정적 영향 평가 문항

1. 프라이버시 보호	
[긍정영향 3] 사람에 의한 개인정보 접근 제한	
영향 설명	AI 채용 서비스는 채용 과정에서 불필요한 인적 개입을 줄이고, 개인정보 접근 권한을 최소화해 유출과 오용의 위험을 낮춘다. 평가 단계별로 필요한 최소 정보만 열람 가능하도록 설계되고, 로그 기록과 권한 통제를 통해 무분별한 열람을 방지한다. 이는 인사담당자의 주관적 판단이나 관리 부주의로 인한 프라이버시 침해 가능성을 줄일 수 있다.

평가 문항				
1. 예상되는 긍정적 영향의 규모(크기)	2. 예상되는 긍정적 영향이 지속되는 기간	3. 예상되는 긍정적 영향이 발생할 가능성	4. 예상되는 긍정적 영향에 대한 기대 수준	5. 예상되는 긍정적 영향 정부 지원 필요성
① 보통/경미한 ② 중간 ③ 큼 ④ 매우 큼	① 단기(1~2 ② 중기(3~4년) ③ 장기(5~10년) ④ 세대 간(10년 이상)	① 낮음 ② 중간 ③ 높음 ④ 매우 높음	① 전혀 기대되지 않음 ② 기대되지 않음 ③ 기대됨 ④ 매우 기대됨	① 전혀 필요하지 않음 ② 필요하지 않음 ③ 필요함 ④ 매우 필요함

[2] 부정적 영향 평가 문항

1. 프라이버시 보호	
[부정영향 1] 민감 정보의 과도한 추론 및 프로파일링	
영향 설명	AI 채용 서비스는 표정, 시선, 음성 등 생체·행동 데이터를 분석해 성격이나 심리 상태 같은 민감한 특성을 추론할 수 있다. 이는 직무와 무관한 과잉 수집으로 이어져 지원자의 자기결정권과 사생활을 침해하며, 나아가 성적 정체성이나 정치적 성향 등 드러내고 싶지 않은 개인적 특성까지 부당하게 노출시킬 위험이 있다. 충분한 동의와 고지가 이루어지지 않는다면 부당한 프로파일링으로 이어져 심각한 프라이버시 침해를 초래할 수 있다.

평가 문항					
1. 예상되는 부정적 영향의 규모(크기)	2. 예상되는 부정적 영향이 지속되는 기간	3. 예상되는 부정적 영향이 발생할 가능성	4. 예상되는 부정적 영향이 해결될 가능성	5. 예상되는 부정적 영향에 대한 우려 수준	6. 예상되는 부정적 영향 정부 대응 수준
① 보통/경미한 ② 심각한 ③ 치명적 ④ 재앙적	① 단기(1~2 ② 중기(3~4년) ③ 장기(5~10년) ④ 세대 간(10년 이상)	① 낮음 ② 중간 ③ 높음 ④ 매우 높음	① 매우 낮음 ② 낮음 ③ 높음 ④ 매우 높음	① 전혀 걱정되지 않음 ② 걱정되지 않음 ③ 걱정됨 ④ 매우 걱정됨	① 매우 못하고 있음 ② 못하고 있음 ③ 잘하고 있음 ④ 매우 잘하고 있음

[3] 영향을 받는 대상

1. 프라이버시 보호

영향을 받는 대상

제시된 긍정·부정 영향 사례와 같이, 프라이버시 보호 영역에서 발생할 수 있는 영향을 고려할 때, 영향을 받을 수 있는 대상을 작성해 주십시오.

※ 영향을 받는 개인/그룹/단체에 대해 자유롭게 설명해 주십시오.

[참고] AI 채용 서비스 관련 주요 이해관계자

① [구직자] 경력직·신입직 지원자, 청년 구직자, 취약계층 등
② [채용 기업] HR 담당자, 경영진, 대기업·중소기업·스타트업·공공기관 등
③ [AI 개발사/서비스 제공자] 개발자, 데이터 관리자, 보안 관리자, 제품 기획자, AI 채용 솔루션 업체 등
④ [기존 채용 서비스 업계] 헤드헌팅 업체, 온라인 채용 플랫폼, 채용 컨설팅 업체 등
⑤ [데이터 관련 주체] 클라우드·외주 업체, 데이터 제공·관리자, 데이터 라벨링 업체, 개인정보 처리 업체 등
⑥ [노동·직업 단체] 노동조합, 직종별 협회, 인사관리 전문가 단체 등
⑦ [투자·금융 관계자] 투자자, 주주, 벤처캐피탈 등
⑧ [규제·감독기관] 정부 부처, 감독 기구, 법원, 분쟁조정기구 등
⑨ [시민사회 및 외부 이해관계자] 언론, 교육기관, NGO, 연구자, 일반 대중, 지역사회 등

Primary	Secondary	Unexpected/Unintended
직접적인(주요) 영향을 받는 대상	**간접적인(부차적인) 영향을 받는 대상**	**예상치 못하거나 의도치 않았지만 영향받을 수 있는 대상**
대상: 구직자 본인	*대상: 채용 기업*	*대상: 시민사회 및 외부 이해관계자*
이유: 개인정보, 생체정보, 행동패턴 등이 직접 수집되는 대상. 채용이 걸려 있으므로 보다 적극적인 정보 제공 가능성	*이유: 서비스 제공자나 내부 관리 소홀로 인해 개인정보 유출 사고가 발생할 경우, 기업의 신뢰도와 평판이 훼손되고 법적 책임 및 금전적 손실을 간접적으로 입을 수 있음*	*이유: AI를 통한 대량의 개인정보 수집 및 처리 관행이 확산될 경우, 채용 절차를 넘어 사회 전반에서 개인정보 침해에 대한 우려를 높이고, 개인의 정보 통제권을 약화시키는 사회적 분위기를 조성할 수 있음*

자료: 연구진 작성

다. 평가 결과

전문가 평가 결과는 첫째, ① 프라이버시 보호, ② 포용성, ③ 책임성, ④ 투명성, ⑤ 공정성 5개 AI 윤리 영역에 따른 '윤리 영향 평가' 결과, 둘째, AI 채용 서비스에 대해 식별된 전체 영향에 대한 '정책 지원과 대응 필요성 평가' 결과 두 가지로 정리하고자 한다.

1) 윤리 영향 평가

먼저, 1차 전문가 평가 응답의 질적 분석을 통해 도출한 31개 긍·부정 윤리 영향 유형을 소개하고, 이에 대한 정량 평가 결과를 요약해 제시한다. AI 채용 서비스의 전체 윤리 영향은 〈표 2-44〉에 정리하였다.

〈표 2-44〉 AI 채용 서비스의 긍·부정 윤리 영향

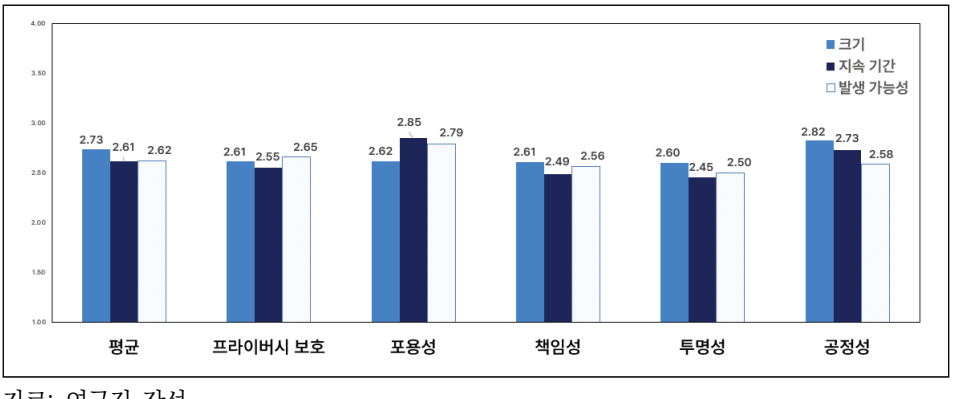
평가 영역	영향유형	
	긍정	부정
프라이버시 보호	① 수집·활용 데이터 최소화 및 비식별화, ② 데이터 관리 체계 강화, ③ 사람에 의한 개인정보 접근 제한	① 민감 정보의 과도한 추론 및 프로파일링, ② SNS 등 채용과 무관한 개인정보 수집, ③ 개인정보 보안 취약성 확대
포용성	① 물리적·시간적 참여 장벽 완화, ② 장애 지원자의 접근성 제고, ③ 언어 장벽 완화	① 디지털 취약계층 소외, ② 절차적 부담·불편 증가, ③ 시스템 설계·기능 한계로 인한 포용성 제약
책임성	① 로그·증빙 기반 책임소재 규명 용이, ② 개인 임의성 축소에 따른 오류 귀속 명료화, ③ 실시간 모니터링·경보 가능으로 책임 이행 체계 확보	① 다중 이해관계자 구조로 인한 책임 주체 모호성, ② 블랙박스 특성으로 인한 책임 추적 어려움, ③ 형식적 인간 개입과 AI 결정 수용 확대로 책임성 관리 약화
투명성	① 사전 고지를 통해 절차적 정당성 확보, ② 표준화된 채용체계 구축을 통한 예측 가능성 제고, ③ 데이터 기반의 객관적인 피드백 제공	① 선택적·제한적 정보 공개 및 정보 비대칭, ② 블랙박스 특성에 따른 설명가능성의 한계, ③ 사전 고지 부재 시 알 권리 침해, ④ 이해곤란·상호작용 부재로 인한 실질적 불투명성
공정성	① 평가 기준의 일관성 확보, ② 직무 역량 중심 평가 확대, ③ 체계적인 편향성 검증 및 완화	① 데이터 기반 차별 재생산 및 고착화, ② 공정성 기준의 불확실성과 신뢰 약화, ③ 정형화된 인재상 의존의 한계

자료: 연구진 작성

영역별 상세 결과를 설명하기에 앞서, 5개 윤리 영역별 지표값 비교 결과를 소개하고자 한다. 각 영역의 지표값은 해당 영역에 포함된 영향 유형들의 평균으로 산출하였다. 긍정 영향에 대한 비교는 [그림 2-23], 부정 영향에 대한 비교는 [그림 2-24]에 정리하였다.

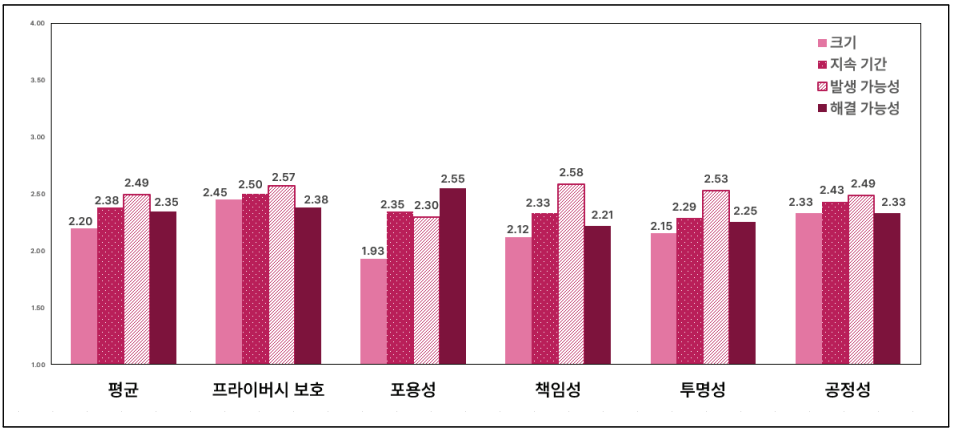
[그림 2-23]의 긍정 영향 비교 결과, '포용성' 영역이 모든 지표에서 비교적 가장 높은 값을 보였으며, 영역 간 지표값 차이는 전반적으로 크게 두드러지지 않았음을 알 수 있었다. 한편 [그림 2-24]의 부정 영향 비교에서는 '프라이버시 보호' 영역이 영향 크기, 지속 기간, 발생 가능성 지표에서 전체 평균보다 대체로 가장 높은 값을 보이며, 부정적 영향이 크게 우려되는 영역으로 나타났다. 반대로 '포용성' 영역은 크기, 지속 기간, 발생 가능성이 낮고, 해결 가능성은 가장 높게 평가되어, 부정적 영향의 실질적 위험은 상대적으로 낮으나 정책·기술적 개입을 통해 충분히 완화 가능한 영역으로 인식되고 있음을 보여준다.

[그림 2-23] 5개 AI 윤리 영역 지표 비교(긍정)



자료: 연구진 작성

[그림 2-24] 5개 AI 윤리 영역 지표 비교(부정)



자료: 연구진 작성

(1) 프라이버시 보호

AI 채용 서비스는 「개인정보 보호법」을 준수해야 하며, 지원자의 민감정보 등을 불필요하게 수집하거나 수집 목적을 벗어나 처리하지 않도록 주의해야 한다. 개인정보는 수집 단계부터 이용 목적, 처리 범위, 보유 기간 등을 명확히 고지하고, 수집·이용·보관·삭제 등 모든 처리 단계에서 안전하게 관리될 수 있도록 설계·운영하는 것이 바람직하다.

① 긍정 윤리 영향

프라이버시 보호 영역과 관련하여 AI 채용 서비스의 긍정적 윤리 영향으로 '수집·활용 데이터 최소화 및 비식별화', '데이터 관리 체계 강화', '사람에 의한 개인정보 접근 제한'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-45〉, [그림 2-25]에서 확인할 수 있다.

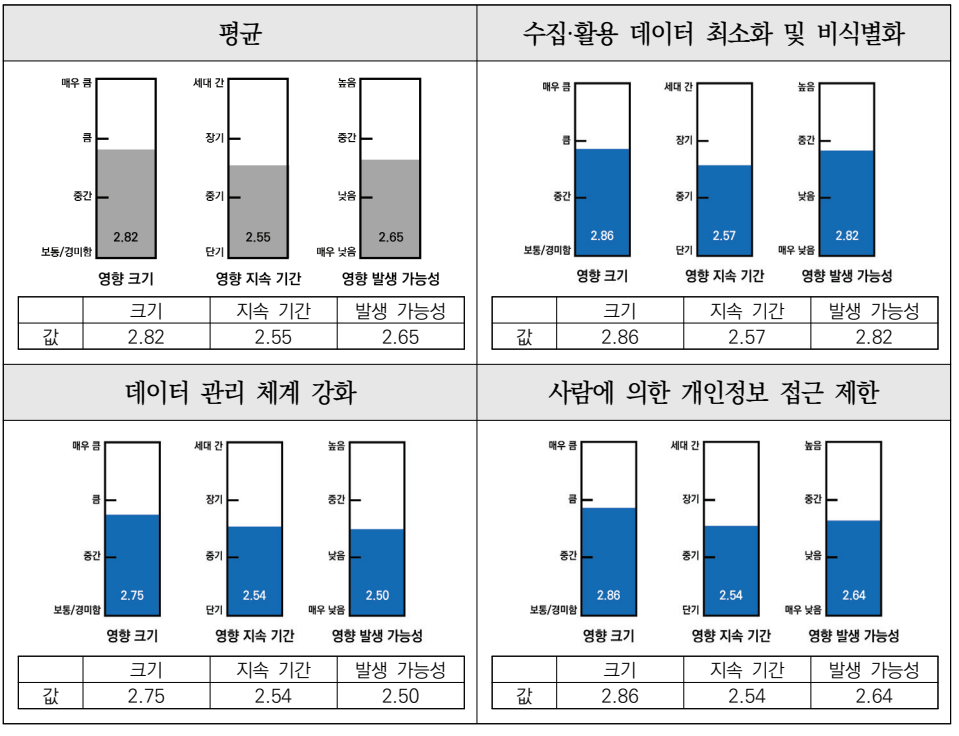
[그림 2-25]에 따르면, '수집·활용 데이터 최소화 및 비식별화'의 발생 가능성이 영역 평균보다 다소 높게 나타났지만, 전체적으로 세 영향 유형의 지표값은 영역 평균과 유사한 수준을 보였다.

〈표 2-45〉 프라이버시 보호 - 긍정적 영향 유형

영향 유형	내용
수집·활용 데이터 최소화 및 비식별화	- 채용 서비스는 채용과 무관한 정보를 배제하거나 비식별 처리하여 최소한의 데이터만 수집·활용하여 평가한다. - 이를 통해 과도한 개인정보 노출을 방지하고, 직무 역량 중심의 공정한 평가 환경을 조성한다.
데이터 관리 체계 강화	- AI 채용 서비스는 개인정보의 수집부터 보관, 활용, 삭제까지 전 과정에서 자동화된 관리 체계를 적용하여 프라이버시 보호 수준을 높인다. - 지원자 데이터는 접근 권한이 엄격히 통제되며, 열람·수정 이력이 모두 로그로 남아 사후 감사와 책임 규명이 용이하며 장기 보관이나 관리 소홀로 인한 유출 위험을 최소화한다. - 일부 시스템은 온디바이스 처리 방식이나 암호화를 통해 중앙 서버 저장을 피함으로써 대규모 보안사고 가능성을 줄인다.
사람에 의한 개인정보 접근 제한	- AI 채용 서비스는 채용 과정에서 불필요한 인적 개입을 줄이고, 개인정보 접근 권한을 최소화해 유출과 오용의 위험을 낮춘다. - 평가 단계별로 필요한 최소 정보만 열람 가능하도록 설계되고, 로그 기록과 권한 통제를 통해 무분별한 열람을 방지한다. - 이는 인사담당자의 주관적 판단이나 관리 부주의로 인한 프라이버시 침해 가능성을 줄일 수 있다.

자료: 연구진 작성

[그림 2-25] 프라이버시 보호 - 긍정 영향 지표 비교



	크기	지속 기간	발생 가능성
값	2.82	2.55	2.65

	크기	지속 기간	발생 가능성
값	2.86	2.57	2.82

	크기	지속 기간	발생 가능성
값	2.75	2.54	2.50

	크기	지속 기간	발생 가능성
값	2.86	2.54	2.64

자료: 연구진 작성

② 부정 윤리 영향

프라이버시 보호 영역과 관련하여 AI 채용 서비스의 부정적 윤리 영향으로 '민감 정보의 과도한 추론 및 프로파일링', 'SNS 등 채용과 무관한 개인정보 수집', '개인정보 보안 취약성 확대'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-46〉, [그림 2-26]에서 확인할 수 있다.

[그림 2-24]의 부정 영향 전체 비교 그래프를 되짚어보면, 프라이버시 보호 영역의 영향 크기, 지속 기간, 발생 가능성이 다른 영역까지 포함한 전체 평균보다 상당히 높은 수준을 보여, 부정적 효과가 크게 예측되었다. 동시에 부정적 효과가 해결될 가능성은 평균 이상으로 평가되어, 긍정적 전망도 동시에 찾아볼 수 있었다. 특히, [그림 2-26]의 '개인정보 보안 취약성 확대'는 크기, 지속 기간,

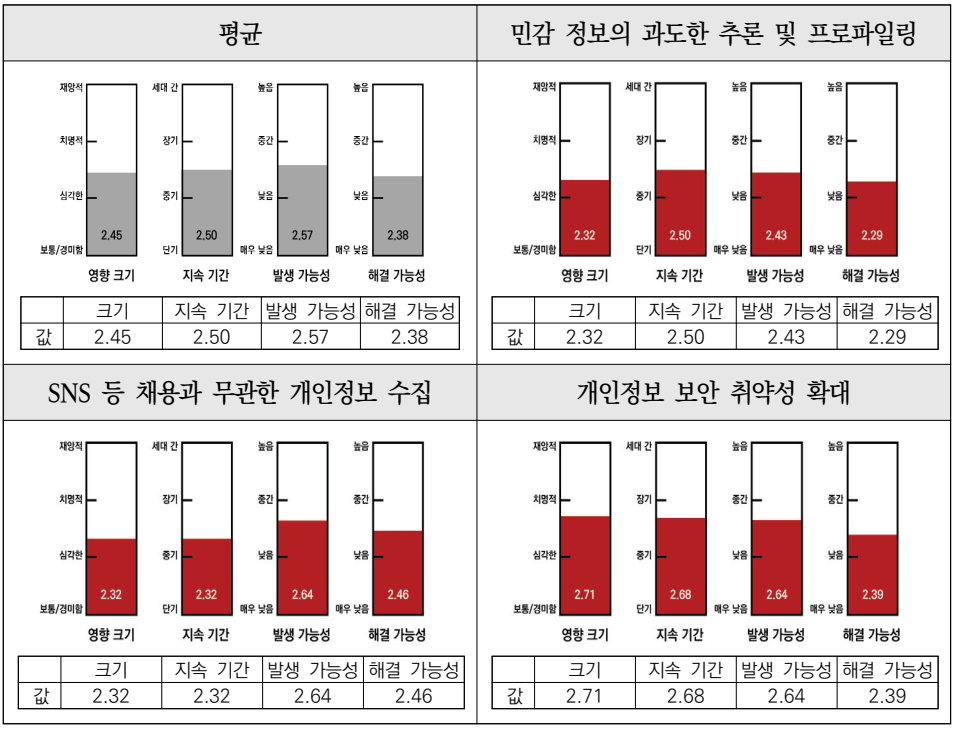
발생 가능성 모두에서 영역 평균과 가장 큰 차이를 보여 핵심 관리 필요 유형으로 확인되었다.

〈표 2-46〉 프라이버시 보호 - 부정적 영향 유형

영향 유형	내용
민감 정보의 과도한 추론 및 프로파일링	- AI 채용 서비스는 표정, 시선, 음성 등 생체·행동 데이터를 분석해 성격이나 심리 상태 같은 민감한 특성을 추론할 수 있다. - 이는 직무와 무관한 과잉 수집으로 이어져 지원자의 자기결정권과 사생활을 침해하며, 나아가 성적 정체성이나 정치적 성향 등 드러내고 싶지 않은 개인적 특성까지 부당하게 노출시킬 위험이 있다. - 충분한 동의와 고지가 이루어지지 않는다면 부당한 프로파일링으로 이어져 심각한 프라이버시 침해를 초래할 수 있다.
SNS 등 채용과 무관한 개인정보 수집	- AI 채용 서비스는 대규모 민감 데이터를 중앙화하거나 외부 업체 서버에 저장하는 과정에서 해킹·내부자 유출 등 보안사고 위험을 키울 수 있다. - 파기 절차가 미비하거나 장기 보관될 경우 재식별· 재구성 가능성이 높아지며, 유출 시 신원 도용· 딥페이크 등 심각한 2차 피해로 이어질 수 있다. - 더 나아가 기존에는 채용 기업 및 심사위원 등 제한된 범위에서만 개인정보 접근이 가능했으나, AI 채용 서비스의 도입으로 개발사, 클라우드 운영사 등 추가 주체가 개인정보 처리 과정에 개입하게 된다. - 이와 같은 접근 권한의 확장은 개인정보 유출·오남용의 잠재적 경로를 확대시켜 프라이버시 침해 위험을 심화시킬 수 있다.
개인정보 보안 취약성 확대	- AI 채용 서비스는 대규모 민감 데이터를 중앙화하거나 외부 업체 서버에 저장하는 과정에서 해킹·내부자 유출 등 보안사고 위험을 키울 수 있다. - 파기 절차가 미비하거나 장기 보관될 경우 재식별·재구성 가능성이 높아지며, 유출 시 신원 도용·딥페이크 등 심각한 2차 피해로 이어질 수 있다. - 더 나아가 기존에는 채용 기업 및 심사위원 등 제한된 범위에서만 개인정보 접근이 가능했으나, AI 채용 서비스의 도입으로 개발사, 클라우드 운영사 등 추가 주체가 개인정보 처리 과정에 개입하게 된다. - 이와 같은 접근 권한의 확장은 개인정보 유출·오남용의 잠재적 경로를 확대시켜 프라이버시 침해 위험을 심화시킬 수 있다.

자료: 연구진 작성

[그림 2-26] 프라이버시 보호 - 부정 영향 주요 지표 비교



	크기	지속 기간	발생 가능성	해결 가능성
값	2.45	2.50	2.57	2.38

	크기	지속 기간	발생 가능성	해결 가능성
값	2.32	2.50	2.43	2.29

	크기	지속 기간	발생 가능성	해결 가능성
값	2.32	2.32	2.64	2.46

	크기	지속 기간	발생 가능성	해결 가능성
값	2.71	2.68	2.64	2.39

자료: 연구진 작성

③ 영향을 받는 대상

프라이버시 보호 영역에서 영향을 받는 대상과 구체 상황은 아래의 〈표 2-47〉과 같다. 직접적 영향을 받는 대상으로는 '구직자'가 AI 채용 과정의 개인정보 직접 제공자이자 정보주체로, 민감정보(이력서·영상·SNS 등) 수집·분석에 따른 노출·재식별·프로파일링 피해 가능성이 높고, 권리행사 어려움의 이유로 가장 많이 언급되었다.

'채용기업'은 개인정보 유출 및 관리 소홀 시 평판·신뢰도 하락, 법적 리스크 부담, AI 채용 확산에 따른 관리·운영 비용 증가 가능성 등으로 인해 간접적 영향을 받는 대상에 다수 포함되었다.

예상치 못한 영향을 받는 대상으로는 '시민사회 및 외부 이해관계자'가 AI 기반 개인정보 수집·분석 확산으로 사회 전반의 프라이버시 기준이 낮아지고 감시문화가 강화되며, 신뢰 저하 등 광범위한 파급효과가 발생할 수 있다는 이유에서 가장 빈번히 언급되었다.

〈표 2-47〉 프라이버시 보호 - 영향을 받는 대상

영향단계	영향을 받는 대상		N
1차 (직접적 영향)	구직자	AI 채용 과정의 개인정보 직접 제공자이자 정보주체로, 민감정보(이력서·영상·SNS 등) 수집·분석에 따른 노출·재식별·프로파일링 피해 가능성 높음. 권리행사 어려움	28
	채용기업	지원자 개인정보를 관리·보호해야 하는 법적 책임 주체로, 유출 시 평판·재무 리스크 및 법적 제재 가능성	3
	AI 개발사·서비스 제공자	알고리즘 설계·보안·데이터처리 과정에서 직접적인 기술적 책임이 발생하며, 시스템 결함 시 법적·평판 리스크 존재	2
2차 (간접적 영향)	채용기업	개인정보 유출·관리 소홀 시 평판, 신뢰도, 법적 리스크 부담. AI 채용 확산에 따른 관리·운영 비용 증가	14
	AI 개발사·서비스 제공자	고객 요구에 따른 비윤리적 데이터 활용 압력 존재. 개인정보 유출 시 복구비용·법적 제재 부담	5
	구직자	과도한 정보 요구나 부정확한 기준으로 간접 피해 발생 가능. 프라이버시 침해에 대한 불신 형성	4
	기존 채용 서비스 업계	AI 채용 확산으로 기존 채용 시장의 경쟁구도·데이터 관리 구조 변화. 개인정보 보호 수준 제고 필요	4
	노동·직업 단체	조합원(구직자)의 개인정보 침해 사례 대응 및 권익보호 전략 마련 필요	2
3차 (예상치 못한, 의도하지 않은 영향)	시민사회 및 외부 이해관계자	AI 기반 개인정보 수집·분석 확산으로 사회 전반의 프라이버시 기준 하락, 감시문화 정착, 신뢰 저하 등 파급효과 발생	8
	AI 개발사·서비스 제공자	대규모 유출·논란 시 사업 지속성 및 서비스 신뢰도 하락. 보안 강화 요구 증대	5
	노동·직업 단체	SNS 검열·노동자 감시 등으로 조합 활동 위축 및 대응전략 재정비 필요	5

영향단계	영향을 받는 대상		N
	구직자 가족	원격 면접·네트워크 환경 등에서 가족 정보 우발적 노출 위험	2
	규제·감독기관	AI 채용 관련 분쟁·민원 급증 시 행정 부담 및 기준 마련 필요	2
	기존 채용 서비스 업계	사회적 불신 확산 시 기존 서비스 이용 위축 및 시장 신뢰 하락	2
	예비 구직자	채용 적합성에 집중해 온라인 활동·의사 표현 자율성 위축	2

주: 최소 2명 이상이 응답한 대상 기준, 중복응답 허용
자료: 연구진 작성

(2) 포용성

AI 채용 서비스는 연령, 성별, 언어, 장애, 인종, 학력, 지역 등 다양한 배경을 가진 지원자가 배제되지 않고 공정하게 평가받을 수 있도록, 포용적이고 접근 가능한 방식으로 설계되어야 한다. 이를 위해 시스템 설계, 학습 데이터 구성, 사용자 인터페이스 구현 등 전 과정에서 문화적, 언어적, 신체적 특성에 따른 차이를 고려한 기준이 반영되어야 한다. 또한 기술적 접근성과 함께, 다양한 사용자들이 실제로 서비스를 쉽게 이용할 수 있도록 사용 편의성 또한 충분히 고려되어야 할 필요가 있다.

① 긍정 윤리 영향

포용성 영역과 관련하여 AI 채용 서비스의 긍정적 윤리 영향으로 '물리적·시간적 참여 장벽 완화', '장애 지원자의 접근성 제고', '언어 장벽 완화'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-48〉, [그림 2-27]에서 확인할 수 있다.

[그림 2-23]의 긍정 영향 전체 평균 비교 그래프를 살펴보면, '포용성' 영역이 모든 지표에서 상대적으로 가장 높은 값을 보였다. [그림 2-27]을 보면, 영향 크기는 '장애 지원자의 접근성 제고', 지속기간은 '언어 장벽 완화', 발생가능성은 '물리적·시간적 참여 장벽 완화' 영향에서 상대적으로 높은 수치를 보였다. 세

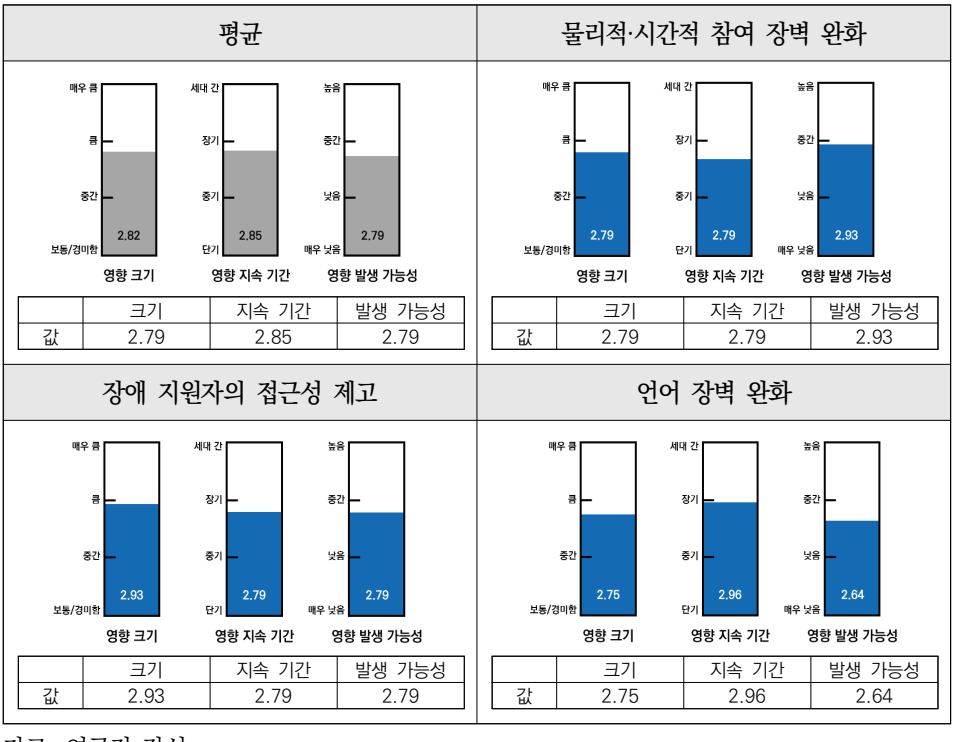
영향 모두 큰 편차 없이 비슷한 경향을 보였으며, 이는 포용성과 관련된 긍정 영향이 전반적으로 고르게 인식되고 있음을 시사한다.

〈표 2-48〉 포용성 - 긍정적 영향 유형

영향 유형	내용
물리적·시간적 참여 장벽 완화	- AI 채용 서비스는 온라인·비대면 채용 절차를 통해 지리적·시간적 제약을 최소화한다. - 수도권 외 지역 거주자, 해외 지원자, 육아·돌봄 등으로 일정이 제한된 사람도 동일한 조건에서 참여할 수 있다. - 이동·준비 비용과 시간을 크게 줄여 다양한 배경의 지원자들이 공평한 기회를 얻도록 돕는다.
장애 지원자의 접근성 제고	- AI 채용 서비스는 장애인의 다양한 요구를 고려한 맞춤형 편의를 제공해 채용 과정에서의 장벽을 크게 낮춘다. - 시각장애인은 스크린리더·음성안내를, 청각장애인은 실시간 자막·수어 지원을, 지체장애인은 충분한 시간 설정이나 대체 입력 방식을 통해 동등한 조건에서 평가받을 수 있다. - 이는 기존 면접에서 발생하던 편견과 불편을 줄이고, 장애 여부와 무관하게 역량 중심의 공정한 경쟁을 가능하게 한다.
언어 장벽 완화	- AI 채용 서비스는 음성인식, 번역, 다국어 인터페이스 같은 기능을 통해 언어적 배경 차이에서 비롯되는 불이익을 줄인다. - 방언이나 억양, 문법 오류 등은 평가 요소에서 분리하고, 자막·통역이나 텍스트 입력 대체 기능을 제공해 비원어민·이주민·다문화 배경 지원자도 동등하게 참여할 수 있다.

자료: 연구진 작성

[그림 2-27] 포용성 - 긍정 영향 지표 비교



	크기	지속 기간	발생 가능성
값	2.79	2.85	2.79

	크기	지속 기간	발생 가능성
값	2.79	2.79	2.93

	크기	지속 기간	발생 가능성
값	2.93	2.79	2.79

	크기	지속 기간	발생 가능성
값	2.75	2.96	2.64

자료: 연구진 작성

② 부정 윤리 영향

포용성 요건과 관련하여 AI 채용 서비스의 부정적 윤리 영향으로 '디지털 취약계층 소외', '절차적 부담·불편 증가', '시스템 설계·기능 한계로 인한 포용성 제약'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-49〉, [그림 2-28]에서 확인할 수 있다.

[그림 2-28]의 포용성 요건 부정 영향 주요 지표 그래프에 따르면, '디지털 취약계층 소외'가 영향의 크기, 지속 기간, 발생 가능성 지표 모두에서 가장 높은 값을 기록하며, 가장 큰 부정적 영향으로 인식되고 있었다. 반면 '절차적 부담·불편 증가'는 영향 크기와 발생 가능성에서 평균 대비 다소 낮은 수준을 보였다. 해결 가능성은 세 영향 모두 유사한 수준이었으며, 5개 영역 전체 평균보다 상당히

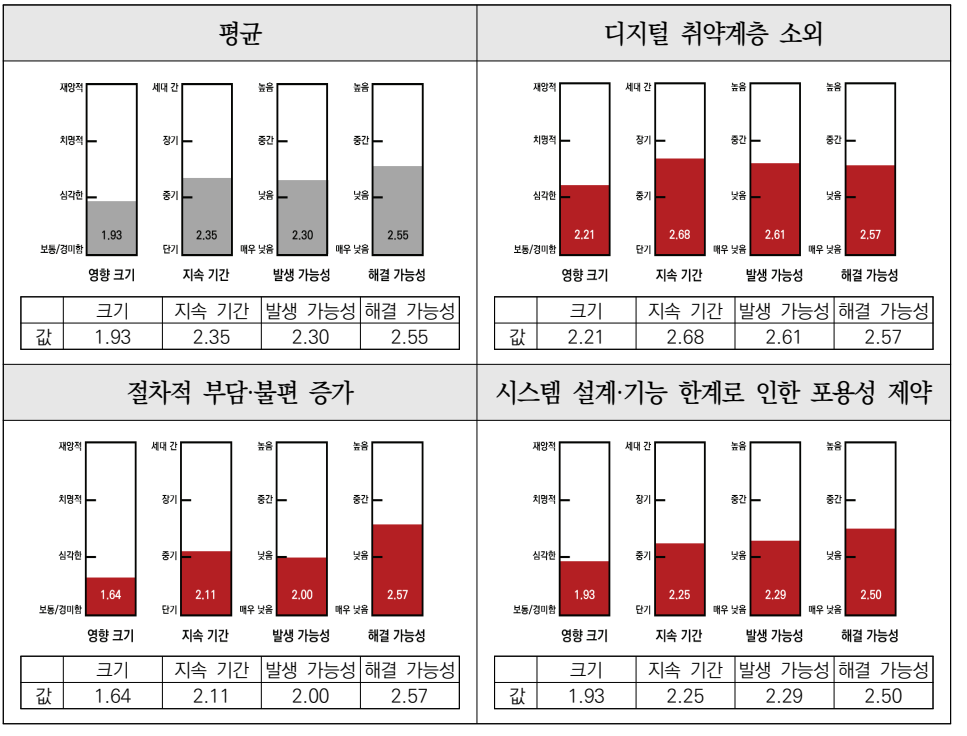
높게 평가되어, 포용성 관련 부정 영향은 정책적·기술적 개입을 통해 완화 가능성이 크다는 인식이 확인되었다.

〈표 2-49〉 포용성 - 부정적 영향 유형

영향 유형	내용
디지털 취약계층 소외	- AI 채용 서비스는 안정적인 인터넷, 고성능 기기, 디지털 활용 능력을 전제로 설계되어 있어 사회경제적 여건이나 연령, 장애 여부에 따라 불평등을 심화시킬 수 있다. - 디지털 역량이 부족한 고령층, 저소득층, 장애인 등은 지원 자체가 차단되거나 실제 역량과 무관하게 낮은 평가를 받을 위험이 크다. - 이는 채용 기회가 기술 접근성의 격차에 따라 좌우되는 새로운 형태의 배제를 낳아 포용성을 저해한다.
절차적 부담·불편 증가	- AI 채용 과정은 일부 지원자에게 추가적인 심리적·절차적 부담을 준다. - 카메라 앞 발화에 익숙하지 않거나 성격적 특성상 화면 응시가 불편한 경우, 면접 참여 자체가 위축될 수 있다. - 또한 일반 전형과는 별도로 장애 전형 등 추가 준비가 요구되거나, 시스템 적응 시간이 부족할 경우 지원자는 실제 역량과 무관하게 불리함을 겪는다. - 결과적으로 이러한 절차적 특성은 채용 과정에서 새로운 참여 장벽으로 기능하며, 특정 집단의 포용성을 저해할 위험이 있다.
시스템 설계·기능 한계로 인한 포용성 제약	- AI 채용 서비스는 본질적으로 시스템이 제공하는 기능과 설계 범위 안에서만 작동한다. - 만약 서비스가 지역·장애·문화적 다양성을 충분히 고려한 설정을 갖추지 못한다면, 시스템이 지원하지 않는 요소는 채용 과정에 반영되지 못해 기술적 한계가 곧 사회적 배제로 이어지는 새로운 형태의 차별을 만들어낼 위험이 있다.

자료: 연구진 작성

[그림 2-28] 포용성 - 부정 영향 주요 지표 비교



자료: 연구진 작성

③ 영향을 받는 대상

포용성 영역의 영향을 받는 대상과 구체 상황은 아래의 〈표 2-50〉과 같다. '구직자'는 연령·성별·장애 등 특성에 따라 차별 및 배제될 위험이 존재하며, AI 알고리즘이 특정 집단에 불리하게 작동할 수 있다는 점에서 가장 많은 전문가가 직접적 영향을 받는 대상으로 평가하였다.

'채용기업'은 포용성 결여로 인한 사회적 비판, 평판 리스크, 다양성 관리 부담이 증가하기 때문에 간접적 영향을 받는 대상으로 반복적으로 언급되었다.

예상치 못한 영향을 받는 대상으로도 역시 '구직자'가 많은 수로 언급되었는데, 디지털 취약계층, 고령층, 언어·문화적 소수자, 성소수자 등이 시스템 설계 및 사회적 구조상 비의도적 편향으로 인해 불이익을 받을 수 있는 반면, 언어 장벽

해소 등 일부 약자에게는 완화 요인으로 작용할 수 있다는 상반된 가능성이 함께 지적되었다.

〈표 2-50〉 포용성 - 영향을 받는 대상

영향단계	영향을 받는 대상		N
1차 (직접적 영향)	구직자	AI 채용 과정에서 지원자 특성(연령·성별·장애 등)에 따라 차별·배제 위험 존재. AI 알고리즘이 특정 집단을 불이익하게 대할 가능성이 높음	25
	AI 개발사·서비스 제공자	데이터 학습·모델 설계 시 편향 제거 및 공정성 확보를 고려해야 하는 주체	2
	채용기업	채용 결과에 따라 특정 지원자 집단이 소외될 경우 법적·사회적 책임 발생 가능	2
2차 (간접적 영향)	채용기업	포용성 결여로 인한 사회적 비판, 평판 리스크, 다양성 관리 부담 증가	13
	구직자	AI 채용 결과의 편향이 누적될 경우, 반복된 탈락 경험으로 자기효능감 저하·사회적 배제 심화	7
	노동·직업 단체	차별적 알고리즘에 대한 대응·모니터링 필요. 구성원 권익 보호 과제 부상	3
	AI 개발사·서비스 제공자	고객 요구·데이터 편향에 대응해 포용성 기준을 기술적으로 구현해야 함	2
	규제·감독기관	포용성 관련 감독·평가기준 미비에 따른 행정적 부담 예상	2
	시민사회 및 외부 이해관계자	AI 채용 결과의 불평등이 사회적 논쟁·감시 요구로 확산될 가능성	2
3차 (예상치 못한, 의도하지 않은 영향)	구직자	디지털 취약계층, 특정 세대(고령층), 언어·문화적 소수자, 성소수자 등에 대해 시스템 설계나 사회적 구조상의 비의도적 편향으로 평가 불이익이 발생 가능. AI 기반 지원이 언어장벽 해소 등 특정 약자에겐 완화 요인으로 기능	7
	시민사회 및 외부 이해관계자	포용성 부족한 AI 채용 확산은 사회 전반의 다양성·기회평등 인식 저하로 이어질 가능성	7

영향단계	영향을 받는 대상		N
	AI 개발사·서비스 제공자	편향 문제 지속 시 기술 불신·시장 위축 위험	4
	규제·감독기관	차별·불평등 문제 발생 시 법·제도 개선 및 감독 부담 확대	3
	기존 채용 서비스 업계	포용성 중심의 채용 문화 확산 요구 대응 필요	3

주: 최소 2명 이상이 응답한 대상 기준, 중복응답 허용
자료: 연구진 작성

(3) 책임성

AI 채용 서비스는 산출과정과 최종결과에 대해 책임 주체가 명확히 설정되어야 하며, 오류 또는 부정확한 판단으로 인한 피해가 발생할 경우 신속한 대응이 가능하도록 체계적인 구조를 마련해야 한다. 또한, 알고리즘이 안전하게 작동하고 예측 불가능한 방식으로 지원자에게 위해를 가하지 않도록 지속적인 테스트와 검증을 수행하는 것이 중요하다.

① 긍정 윤리 영향

책임성 영역과 관련하여 AI 채용 서비스의 긍정적 윤리 영향으로 '로그·증빙 기반 책임소재 규명 용이', '개인 임의성 축소에 따른 오류 귀속 명료화', '실시간 모니터링으로 책임 이행 체계 확보'를 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-51〉, [그림 2-29]에서 확인할 수 있다.

[그림 2-29]의 그래프를 살펴보면, '개인 임의성 축소에 따른 오류 귀속 명료화'는 영향의 크기, 지속 기간, 발생 가능성 지표 모두에서 영역 평균보다 상당히 높은 값을 보여, 책임성 강화를 위한 AI 활용의 긍정적 효과가 특히 두드러지는 유형으로 나타났다.

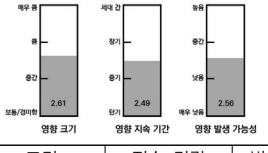
〈표 2-51〉 책임성 - 긍정적 영향 유형

영향 유형	내용
로그·증빙 기반 책임소재 규명 용이	- 인간 평가자의 주관적 판단과 달리, AI 채용 서비스의 정량화된 로그와 감사 추적 기록(audit trail)은 오류 원인 분석과 개선 조치를 용이하게 해 분쟁 해결과 책임성 제고에 실질적으로 기여한다. - 채용 결정이 어떻게 이루어졌는지를 객관적으로 재구성할 수 있으며, 이의제기나 감사 시 책임소재를 명확히 규명할 수 있다.
개인 임의성 축소에 따른 오류 귀속 명료화	- AI 채용 서비스는 모든 지원자에게 동일한 평가 기준을 일관되게 적용하여 인간 평가자의 피로, 편견, 감정에 따른 자의적 판단을 최소화한다. - 이로써 동일 조건의 지원자는 같은 결과를 얻을 수 있고, 특정 그룹의 불이익이 반복된다면 이는 개인 심사자의 일탈이 아니라 알고리즘 규칙이나 데이터셋 문제임을 명확히 지목할 수 있다.
실시간 모니터링으로 책임 이행 체계 확보	- AI 채용 서비스는 모든 평가 과정에서 발생하는 오류나 편향을 실시간으로 탐지하고 관리자에게 경고하는 기능을 탑재할 수 있다. - 특정 집단에 불리한 점수 패턴이나 기술적 오작동이 발견되면 즉시 확인·수정이 가능해, 문제를 방치하지 않고 빠르게 대응할 수 있다. - 이러한 자동화된 모니터링과 경보 체계는 인간 평가의 한계를 보완하고, 채용 과정 전반의 신뢰성과 책임성을 강화하는 데 기여한다.

자료: 연구진 작성

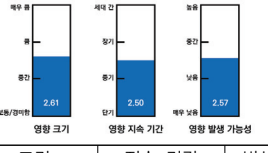
[그림 2-29] 책임성 - 긍정 영향 지표 비교

평균



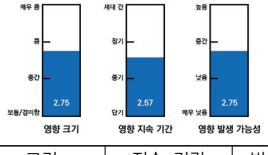
	크기	지속 기간	발생 가능성
값	2.61	2.49	2.56

로그·증빙 기반 책임소재 규명 용이



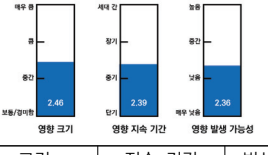
	크기	지속 기간	발생 가능성
값	2.61	2.50	2.57

개인 임의성 축소에 따른 오류 귀속 명료화



	크기	지속 기간	발생 가능성
값	2.75	2.57	2.75

실시간 모니터링으로 책임 이행 체계 확보



	크기	지속 기간	발생 가능성
값	2.46	2.39	2.36

자료: 연구진 작성

② 부정 윤리 영향

책임성 영역과 관련하여 AI 채용 서비스의 부정적 윤리 영향으로 '다중 이해관계자 구조로 인한 책임 주체 모호', '블랙박스 특성으로 인한 책임 추적 어려움', '형식적 인간 개입과 AI 결정의 무비판적 수용'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-52〉, [그림 2-30]에서 확인할 수 있다.

책임성 영역의 부정 영향 주요 지표 그래프를 살펴보면, '형식적 인간 개입과 AI 결정의 무비판적 수용'은 크기와 지속 기간, 발생 가능성 지표에서 평균보다 높은 값을 기록하여, AI 의사결정 과정에서 실질적인 인간 개입이 충분히 확보되지 않을 위험이 크게 인식되고 있었다. 또한, '블랙박스 특성으로 인한 책임 추적 어려움' 역시 모든 지표에서 영역 평균보다 높게 평가되어, 부정적 효과가 크게 인식됨을 알 수 있었다. 반면 '다중 이해관계자 구조로 인한 책임 주체 모호'는 영향 크기와 지속 기간은 상대적으로 낮지만, 발생 가능성은 평균과 유사하여 구조적 책임 분산에 따른 위험이 지속적으로 존재함을 시사한다.

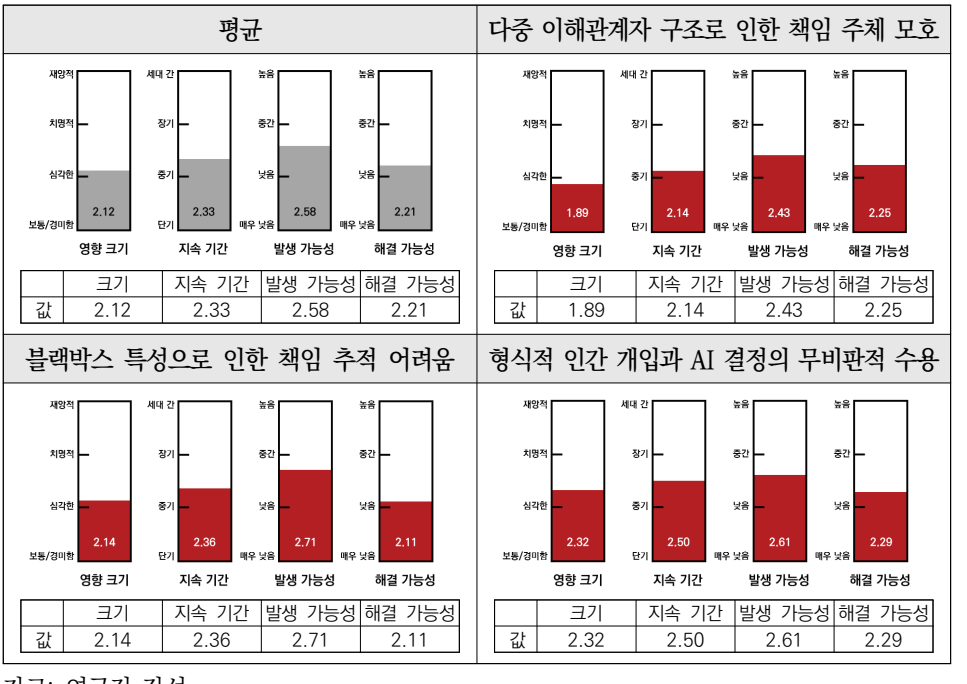
〈표 2-52〉 책임성 - 부정적 영향 유형

영향 유형	내용
다중 이해관계자 구조로 인한 책임 주체 모호	- AI 채용 서비스는 기업, 개발사, 데이터 제공자 등 다양한 이해관계자가 얽혀 있어 오류나 피해가 발생했을 때 책임 주체가 불명확해질 수 있다. - 명확한 법적·계약적 기준이 부재할 경우 각 주체가 서로 책임을 전가하며 '책임의 공백'이 발생하고, 지원자는 적절한 구제를 받기 어렵다. - 이는 기업의 책임 회피를 용이하게 하고, 사회적 불신과 분쟁을 증폭시켜 책임 관리 체계를 약화시키는 요인으로 작용한다.
블랙박스 특성으로 인한 책임 추적 어려움	- AI 채용 서비스의 평가 과정은 딥러닝 기반 모델 특유의 불투명성과 복잡한 연산 과정으로 인해 결과가 어떻게 도출되었는지 추적하기 어렵다. - 오류냐 편향이 발생해도 어떤 데이터와 알고리즘이 문제였는지 명확히 특정하기 곤란하며, 설명 책임을 이행하는 데 한계가 발생한다. - 지원자는 결과의 정당성을 검증하기 어려우며, 기업과 규제기관도 책임 규명에 필요한 근거를 확보하기 힘들어 책임성 관리가 기술적 차원에서 약화될 수 있다.

영향 유형	내용
형식적 인간 개입과 AI 결정의 무비판적 수용	- AI 채용 서비스는 인간 최종 개입을 전제로 설계되지만, 실제 현장에서는 담당자가 AI의 '객관적 데이터'를 비판 없이 수용하는 자동화 편향이 발생할 수 있다. - 이 경우 인간 개입은 형식적으로만 존재하게 되어, 부당한 결과가 발생해도 "AI가 그렇게 판단했다"는 식으로 책임을 전가하거나 회피할 위험이 커진다.

자료: 연구진 작성

[그림 2-30] 책임성 - 부정 영향 주요 지표 비교



	크기	지속 기간	발생 가능성	해결 가능성
값	2.12	2.33	2.58	2.21

	크기	지속 기간	발생 가능성	해결 가능성
값	1.89	2.14	2.43	2.25

	크기	지속 기간	발생 가능성	해결 가능성
값	2.14	2.36	2.71	2.11

	크기	지속 기간	발생 가능성	해결 가능성
값	2.32	2.50	2.61	2.29

자료: 연구진 작성

③ 영향을 받는 대상

책임성 영역의 영향을 받는 대상과 구체 상황은 아래의 〈표 2-53〉과 같다.

직접적 영향을 받는 대상으로는 '채용기업'이 AI 채용 결과에 대한 인사결정 책임 주체로서, 오류·차별 발생 시 직접적인 법적·사회적·재정적 책임을 부담하게 된다는 점에서 가장 많이 언급되었다.

간접적 영향 역시 '채용기업'이 가장 다수로 평가되었는데, 전문가 평가단은 외부 개발사나 내부 AI 시스템에 의존할 경우 책임 귀속이 복잡해지고, 시스템 신뢰성 확보를 위한 관리·감독 부담이 증가한다는 이유로 반복 언급하였다.

의도치 않은 영향을 받는 대상으로는 'AI 개발사·서비스 제공자'가 많은 수로 언급되었는데, 책임 회피 및 전가 구조가 확산될 경우 산업 전반의 신뢰가 저하되고, 장기적으로는 책임 관리체계 강화 요구가 증가할 것이라 평가되었다.

〈표 2-53〉 책임성 - 영향을 받는 대상

영향단계	영향을 받는 대상		N
1차 (직접적 영향)	채용기업	AI 채용 결과의 신뢰성과 투명성, 인사결정의 책임 주체로서 주요 법적·도덕적 책임을 지님. AI 판단 오류나 차별 발생 시 기업이 직접적인 사회적·재정적 책임을 부담	14
	구직자	AI 채용 시스템의 판단에 영향을 받는 직접적 대상. 잘못된 평가·결과에 대한 이의제기나 구제 절차의 부재로 인해 책임 소재 불분명	12
	AI 개발사·서비스 제공자	알고리즘 설계·운영상의 결함, 데이터 품질 부족 등으로 발생한 문제에 대해 기술적 책임. 시스템 오류·결과 설명 요구에 대응할 의무	7

영향단계	영향을 받는 대상		N
2차 (간접적 영향)	채용기업	외부 개발사나 내부 AI 시스템 의존 시 책임 귀속 문제 발생. 시스템 신뢰 확보를 위한 관리·감독 부담 증가	9
	구직자	불투명한 평가 결과로 인한 피해 시 구제 절차 부족. 기업 또는 개발자 간 책임 공방에 따른 불이익 가능성	7
	규제·감독기관	책임소재 불분명 사안이 증가할 경우 감독·행정 부담 확대	3
	시민사회 및 외부 이해관계자	책임에 대한 구조가 불명확할 때 사회적 불신 및 공적 대응 요구 증가	3
	AI 개발사·서비스 제공자	책임 분담 구조 내에서 기업·정부 간 책임한계 불명확. 고객사 요구에 따라 책임 회피 구조 형성 우려	2
	기존 채용 서비스 업계	책임 기준이 강화되면 기존 채용 플랫폼에도 규제 확산 가능	2
	노동·직업 단체	인사결정 오류 발생 시 조합원 보호를 위한 제도적 보완 필요	2
	데이터 관련 주체	데이터 제공·라벨링 단계의 오류에 대한 연쇄적 책임 부담 가능성	2
	투자·금융 관계자	AI 채용 실패 사례 발생 시 기업 평판 리스크로 투자 불안정성 증대	2
3차 (예상치 못한, 의도하지 않은 영향)	AI 개발사·서비스 제공자	책임 회피·책임전가 구조가 확산될 경우 산업 신뢰 저하. 장기적으로 책임 관리체계 강화 요구 증가	7
	시민사회 및 외부 이해관계자	채용의 불투명성·책임 불분명 사례가 사회적 논쟁, 공공 불신으로 이어질 가능성	5
	규제·감독기관	다수 이해관계자 간 책임 분쟁 시 조정·감독 부담 증가. 제도 개선 필요	4
	기존 채용 서비스 업계	책임 기준 강화로 기존 서비스에도 법적·윤리적 검증 요구 확산	4
	데이터 관련 주체	데이터 오류·유출로 인한 책임 전이 및 법적 분쟁 가능성	2
	채용기업	책임 회피 시 기업 신뢰·고용 브랜딩 악화	2
	투자·금융 관계자	책임 불분명으로 인한 ESG 리스크 확대, 투자 위축 가능성	2

주: 최소 2명 이상이 응답한 대상 기준, 중복응답 허용
자료: 연구진 작성

(4) 투명성

사회적 신뢰 형성을 위해 AI 채용 서비스는 투명성과 설명가능성을 확보하기 위한 노력을 기울이는 것이 바람직하다. 특히 자동화된 판단이 이루어지는 경우, 결과의 주요 기준과 판단 근거에 대한 적절한 설명가능성을 제공하려는 노력을 고려할 필요가 있다. 아울러 AI 채용 시스템을 제공하거나 활용하는 주체는 AI 기술의 활용 범위와 채용 과정에서 발생 가능한 주요 위험 요소를 지원자에게 미리 안내해야 한다. 이는 지원자의 알 권리와 절차적 정당성 확보에 도움이 되며, 기업의 사회적 신뢰 형성에도 도움이 될 수 있다.

① 긍정 윤리 영향

투명성 영역과 관련하여 AI 채용 서비스의 긍정적 윤리 영향으로 '사전 고지를 통해 절차적 정당성 확보', '표준화된 채용체계 구축으로 예측가능성 제고', '데이터 기반의 객관적인 피드백 제공'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-54〉, [그림 2-31]에서 확인할 수 있다.

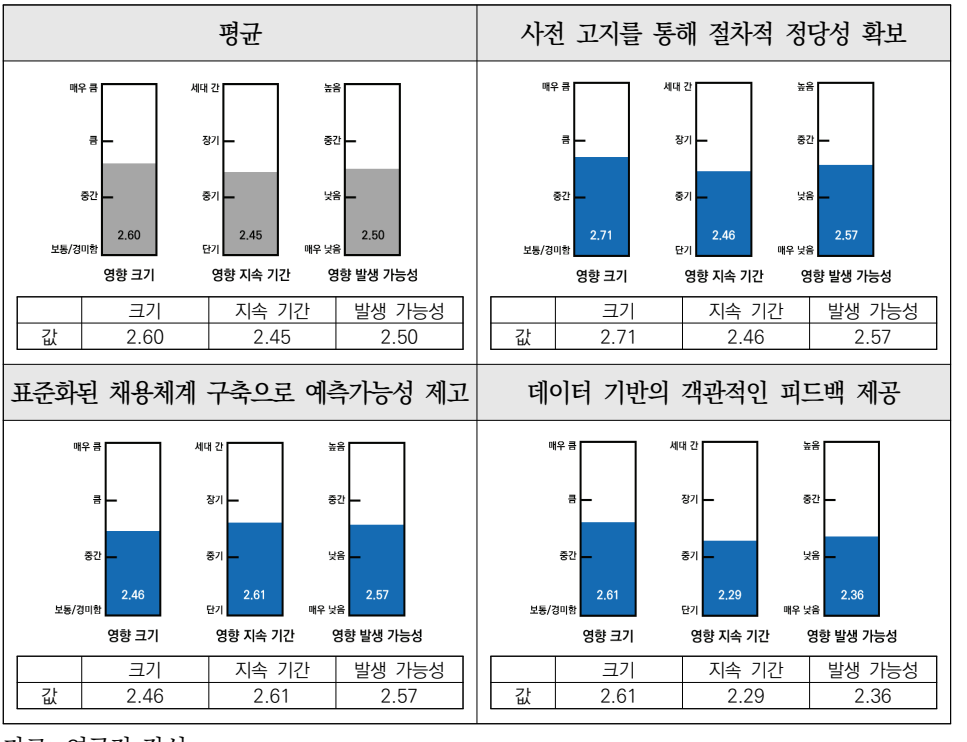
[그림 2-31]에 따르면 '사전 고지를 통해 절차적 정당성 확보'가 모든 지표에서 평균보다 높은 값을 기록하여, 투명성 강화를 위한 핵심적인 긍정 효과로 인식되고 있었다. 반면 '데이터 기반의 객관적인 피드백 제공'은 지속 기간과 발생 가능성에서 다소 낮게 평가되었다.

〈표 2-54〉 투명성 - 긍정적 영향 유형

영향 유형	내용
사전 고지를 통해 절차적 정당성 확보	- AI 채용 서비스는 사용 단계, 평가 항목, 데이터 활용 범위 등을 사전에 명확히 고지할 수 있어, 지원자가 자신이 어떤 기준에 따라 평가받는지 인지하고 준비할 수 있게 한다. - 이는 불확실성과 '깜깜이 채용'에 대한 불안감을 해소하며, 지원자의 알 권리와 정보 기반 동의를 보장한다. 나아가 절차적 정당성을 강화하고, 기업과 지원자 간의 신뢰 형성과 법적 리스크 예방에도 기여한다.
표준화된 채용체계 구축으로 예측가능성 제고	- 기존 인간 평가의 주관적 편차를 줄이고 채용 절차를 구조화·표준화한다. - 이는 지원자들에게 동일한 잣대를 적용하여 결과의 일관성과 예측 가능성을 확보한다. - 표준화된 절차는 "어떤 기준과 절차가 적용되었는지"를 명확히 드러내고, 이해관계자 누구나 동일하게 확인·검증할 수 있게 하기 때문에 채용 과정 전반에 대한 사회적 신뢰를 높이는 데 기여한다.
데이터 기반의 객관적인 피드백 제공	- AI 채용 서비스는 평가 과정에서 산출된 데이터를 기반으로 구체적이고 표준화된 피드백을 지원자에게 제공할 수 있다. - 단순히 합격·불합격 결과만 통지하는 기존 관행과 달리, 강점과 약점, 보완이 필요한 역량을 수치와 지표로 제시함으로써 지원자의 알 권리와 결과 수용성을 제고한다. - 이는 불합격자도 자신의 발전 방향을 파악할 수 있게 하여, 채용 과정에 대한 신뢰와 긍정적 경험을 높이고, 기업에도 투명하고 책임 있는 채용 이미지를 구축하는 데 기여한다.

자료: 연구진 작성

[그림 2-31] 투명성 - 긍정 영향 지표 비교



	크기	지속 기간	발생 가능성
값	2.60	2.45	2.50

	크기	지속 기간	발생 가능성
값	2.71	2.46	2.57

	크기	지속 기간	발생 가능성
값	2.46	2.61	2.57

	크기	지속 기간	발생 가능성
값	2.61	2.29	2.36

자료: 연구진 작성

② 부정 윤리 영향

투명성 요건과 관련하여 AI 채용 서비스의 부정적 윤리 영향으로 '선택적·제한적 정보 공개 및 정보 비대칭', '블랙박스 특성에 따른 설명가능성의 한계', '사전 고지 부재 시 알 권리 침해', '이해곤란·상호작용 부재로 실질적 불투명성 발생'을 도출하였다. 네 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-55〉, [그림 2-32]에서 확인할 수 있다.

[그림 2-32]의 투명성 요건 부정 영향 주요 지표 그래프를 살펴보면, '블랙박스 특성에 따른 설명가능성의 한계'가 영향의 크기, 지속 기간, 발생 가능성에서 영역 평균보다 모두 높은 값을 보여, 지원자가 AI 판단 과정을 이해하기 어렵다는 점이 가장 두드러진 부정적 영향으로 인식되고 있었다. '선택적·제한적 정보 공개

및 정보 비대칭' 역시 동일 지표에서 평균보다 높은 수준을 기록하여, 필수 정보가 충분히 제공되지 않을 경우 투명성이 크게 저하될 수 있음을 보여준다.

〈표 2-55〉 투명성 - 부정적 영향 유형

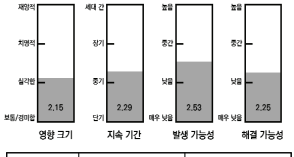
영향 유형	내용
선택적·제한적 정보 공개 및 정보 비대칭	- 핵심적인 평가 기준과 알고리즘 작동 원리가 영업비밀이나 기술적 한계를 이유로 충분히 공개되지 않는 경우가 발생할 수 있다. - 기업은 모호하거나 긍정적인 정보만을 선택적으로 제공하여 형식적 투명성만 확보하고, 지원자는 어떤 데이터와 절차가 실제로 작동하는지 알기 어렵다. - 또한 개발사와 이용사 간에는 기술적·운영적 지식의 불균형이 존재해 문제가 발생했을 때 서로 책임을 전가하거나 핵심 정보를 은폐할 가능성이 높다. - 정보 비대칭은 지원자의 알 권리와 절차적 정당성을 침해할 뿐 아니라, 기업 내부에서도 채용 결과를 검증·설명하기 어렵게 만들어 투명성을 저해한다.
블랙박스 특성에 따른 설명가능성의 한계	- AI 채용 서비스는 딥러닝·생성형 모델과 같이 복잡한 구조를 가진 경우가 많아, 특정 판단이 어떻게 도출되었는지를 지원자나 기업, 심지어 개발자조차 명확히 설명하기 어렵다. - 이는 탈락 사유에 대한 납득을 어렵게 하고, 이의제기나 검증을 위한 근거 확보를 불가능하게 만든다. - 모델 특유의 불투명성과 예측 불가능한 오류는 결과의 정당성에 대한 의심을 불러일으키며, 결국 채용 과정의 투명성을 심각하게 저해하는 요인으로 작용할 수 있다.
사전 고지 부재 시 알 권리 침해	- AI 채용 서비스가 어떤 단계에서, 어떤 데이터를 활용하여, 어떤 기준으로 지원자를 평가하는지를 사전에 안내하지 않을 경우, 지원자는 자신이 어떤 절차에 따라 평가받는지 알 수 없어 준비 기회와 통제권을 상실한다. - 자동화된 평가가 음성·표정·행동 등 개인 특성을 분석함에도 불구하고 이를 미리 알리지 않으면, 지원자는 불투명한 절차 속에서 불이익을 받을 수 있으며, 결과에 대한 납득 가능성도 현저히 낮아진다. - 나아가 동의 없는 데이터 활용이나 미공개 평가 요소는 프라이버시 침해와 절차적 불공정으로 이어지며, 기업에 대한 불신과 법적 분쟁의 위험을 높인다.

영향 유형	내용
이해곤란·상호작용 부재로 실질적 불투명성 발생	- AI 채용 서비스가 평가 기준이나 결과를 공개하더라도, 기술적 용어·통계지표 위주의 설명은 일반 지원자가 이해하기 어렵다. - HR담당자조차 결과 해석이 힘든 경우도 발생하며, 이때 투명성은 형식적으로만 충족될 뿐 실질적으로는 확보되지 않는다. - 또한, 사람 면접관처럼 즉각적인 피드백이나 상호작용이 없는 구조는 지원자의 불안과 불신을 심화시킬 수 있다. - 결과적으로 “정보는 공개되었지만 이해 가능성이 낮은 상태”는 투명성을 약화시키며, 실제 납득 가능성과 신뢰 확보에 실패하게 된다.

자료: 연구진 작성

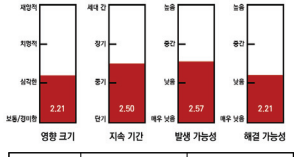
[그림 2-32] 투명성 - 부정 영향 주요 지표 비교

평균



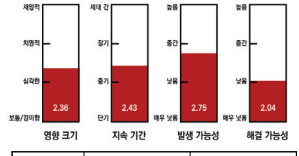
	크기	지속 기간
값	2.15	2.29
	발생 가능성	해결 가능성
값	2.53	2.35

선택적·제한적 정보 공개 및 정보 비대칭



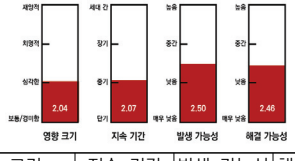
	크기	지속 기간
값	2.21	2.50
	발생 가능성	해결 가능성
값	2.57	2.21

블랙박스 특성에 따른 설명가능성의 한계



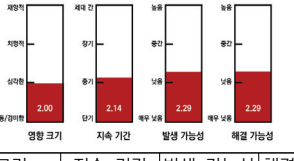
	크기	지속 기간
값	2.36	2.43
	발생 가능성	해결 가능성
값	2.75	2.04

사전 고지 부재 시 알 권리 침해



	크기	지속 기간	발생 가능성	해결 가능성
값	2.04	2.07	2.50	2.46

이해곤란·상호작용 부재로 실질적 불투명성 발생



	크기	지속 기간	발생 가능성	해결 가능성
값	2.00	2.14	2.29	2.29

자료: 연구진 작성

③ 영향을 받는 대상

투명성 영역의 영향을 받는 대상과 구체 상황은 아래의 〈표 2-56〉과 같다. 직접적 영향을 받는 대상으로 '구직자'가 가장 많이 언급되었고, 주요 이유로는 AI 채용 평가 기준과 작동 방식이 불투명할 경우 결과의 합리성 검증이 불가능하고, 이의제기가 사실상 불가능해지는 등 정보 비대칭이 심화된다는 점에서 강조되었다.

'채용기업'은 평가 과정의 불투명성으로 인한 기업 신뢰도가 하락하고, 공정성 논란 및 법적 책임 위험이 커지며 내부 관리·검증 절차를 강화해야 한다는 부담이 존재한다는 이유로 간접적 영향을 받는 대상으로 다수 언급되었다.

의도치 않은 영향을 받는 대상으로는 'AI 개발사·서비스 제공자'가 가장 다수로 평가되었는데, 전문가 평가단은 알고리즘 공개 요구 확산으로 영업비밀 보호와 사회적 책임 간 충돌이 심화될 수 있다는 점을 이유로 꼽았다. '시민사회 및 외부 이해관계자'도 유사한 빈도수로 언급되었으며, 불투명한 채용 시스템 확산이 전반적 사회 신뢰 저하와 공정성 논란 확대를 초래할 수 있다는 점에서 예상치 못한 영향 대상이라고 평가되었다.

〈표 2-56〉 투명성 - 영향을 받는 대상

영향단계	영향을 받는 대상		N
1차 (직접적 영향)	구직자	AI 채용 평가 기준과 작동 방식이 불투명할 경우 결과의 합리성 검증 불가, 이의제기 어려움 등 정보 비대칭 구조 심화	19
	AI 개발사·서비스 제공자	알고리즘의 설명가능성·로직 공개 범위를 명확히 해야 하는 핵심 책임 주체. 모델 구조·데이터 처리 과정의 투명성 확보 필요	5
	채용기업	AI 평가 결과를 인사결정에 활용할 때, 평가 기준과 절차를 구직자에게 명확히 고지해야 함. 설명 책임 강화 필요	4

영향단계	영향을 받는 대상		N
2차 (간접적 영향)	채용기업	평가 과정의 불투명성으로 인한 기업 신뢰도 하락, 공정성 논란 및 법적 책임 위험이 커짐. 내부 관리·검증 절차 강화 필요	14
	구직자	AI 채용의 기준·절차가 불명확할 경우 반복된 불신, 제도적 투명성 요구 확대	5
	AI 개발사·서비스 제공자	시스템 내부 로직 비공개로 인해 고객사의 불만 및 책임 불분명 현상 초래 가능	2
	규제·감독기관	평가 알고리즘 설명·공시 제도 도입 시 감독 부담 및 행정 리스크 증가	2
	기존 채용 서비스 업계	AI 채용 확산으로 투명성 기준 경쟁 심화, 기존 서비스의 신뢰 확보 필요	2
	데이터 관련 주체	데이터 수집·가공 단계의 불투명성이 전체 AI 채용 신뢰도에 부정적 영향	2
	시민사회 및 외부 이해관계자	채용 투명성 결여가 사회적 불공정 인식 확산 및 감시 요구 강화로 이어질 수 있음	2
3차 (예상치 못한, 의도하지 않은 영향)	AI 개발사·서비스 제공자	알고리즘 공개 요구 확산으로 영업비밀 보호와 사회적 책임 간 충돌 심화. 신뢰성 확보를 위한 제도적 지원 필요	7
	시민사회 및 외부 이해관계자	불투명한 채용 시스템 확산 시 사회적 신뢰 저하, 공정성 논란 확대 가능성	6
	채용기업	평가 기준·결과 설명 부족 시 기업 이미지·브랜드 신뢰도 하락 위험	5
	규제·감독기관	투명성 확보를 위한 설명의무 부과 시 법령 정비 및 행정 부담 증가	3
	구직자	AI 평가 기준이 모호할 때 불공정 인식 확산 및 채용 불신 심화	2
	기존 채용 서비스 업계	투명성 기준 강화로 기존 채용 플랫폼의 구조 개선 요구 증가	2
	예비 구직자	평가 기준을 알 수 없어 지원 회피, 자기검열 유발 가능	2
	투자·금융 관계자	투명성 논란이 기업 ESG 리스크로 전이, 투자 위축 가능성	2

주: 최소 2명 이상이 응답한 대상 기준, 중복응답 허용
자료: 연구진 작성

(5) 공정성

AI 채용 서비스는 다양한 배경을 가진 지원자들이 정당한 기준에 따라 일관되게 평가받을 수 있도록, 알고리즘 설계, 데이터 구성, 평가 절차 전반에 걸쳐 공정성 확보를 위한 체계적인 노력을 기울일 필요가 있다. 이는 단순히 동일한 기준을 일률적으로 적용하는 데 그치지 않고, 특정 집단이 구조적으로 불이익을 받지 않도록 정당한 차등 기준 적용과 의도적·비의도적 편향의 식별 및 개선방안 마련을 포함한다. 공정성은 사회적으로 합의된 기준에 따라 판단되어야 하며, 다양한 이해관계자의 관점을 반영하여 평가 기준과 알고리즘 적용 방식을 설계하는 것이 도움이 될 수 있다. 이러한 점에서 공정성은 고정된 상태가 아니라, 지속적으로 점검하고 보완해나가야 하는 윤리적 고려사항으로 다뤄진다.

① 긍정 윤리 영향

공정성 영역과 관련하여 AI 채용 서비스의 긍정적 윤리 영향으로 '평가 기준의 일관성 확보', '직무 역량 중심 평가 확대', '체계적인 편향성 검증 및 완화'을 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-57〉, [그림 2-33]에서 확인할 수 있다.

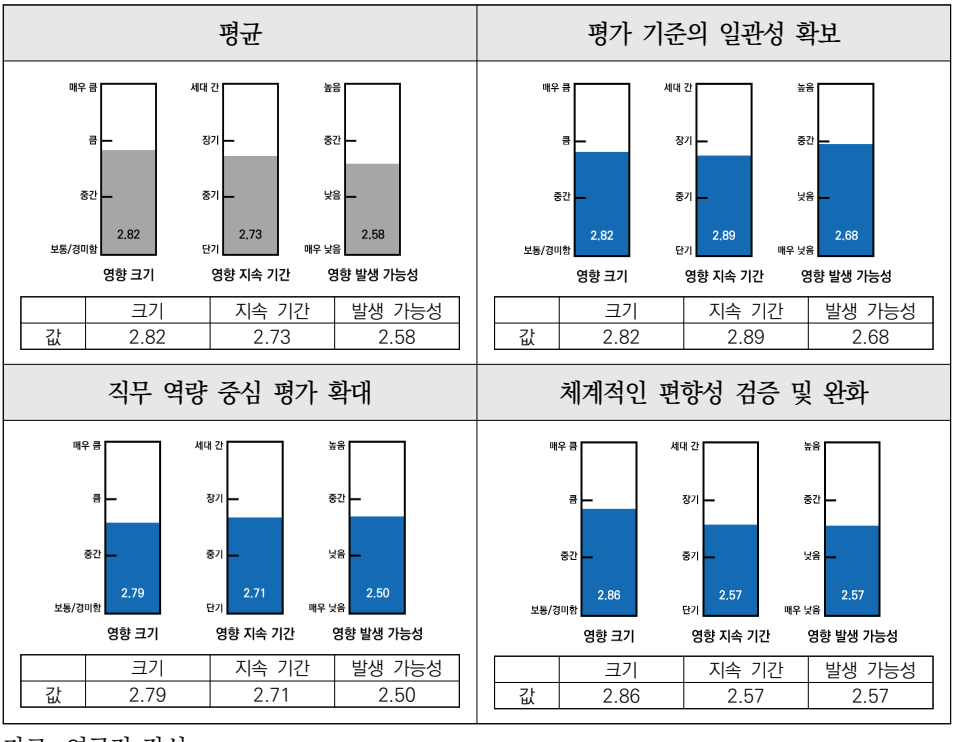
[그림 2-33]에 따르면, '평가 기준의 일관성 확보'는 영향 지속 기간과 발생 가능성이 영역 평균보다 크게 높은 수치를 보였으며, 영향 규모 역시 평균과 동일한 수준을 기록해, 공정성 강화를 위해 AI 채용 도입이 가장 크게 기여할 수 있는 영역으로 인식되었다. '직무 역량 중심 평가 확대'는 전체 평균보다 다소 낮지만, 지속 기간이 비교적 높게 평가되어 직무 적합성 중심의 평가 체계가 장기적으로 안정적으로 작동할 가능성을 시사한다.

〈표 2-57〉 공정성 - 긍정적 영향 유형

영향 유형	내용
평가 기준의 일관성 확보	- AI 채용 서비스는 사전에 정의된 동일한 질문과 절차를 모든 지원자에게 적용해 평가 기준을 일관되게 유지한다. - 평가자의 피로, 성향, 무의식적 편견 같은 변수는 배제되어, 개인적 상황과 무관하게 공평한 경쟁 환경이 마련된다. - 이를 통해 지원자들은 동일한 조건에서 평가받으며 결과의 수용성과 신뢰가 높아진다.
직무 역량 중심 평가 확대	- AI 채용 서비스는 전통적으로 채용 과정에서 중시되어 온 학력, 출신 학교, 성별, 나이 등 인구통계학적·배경적 요소를 배제하고, 직무 관련 경험과 역량에 집중한다. 이를 통해 비전통적 경력이나 다양한 배경에서 쌓은 실질적 능력을 가진 인재가 공정하게 평가받을 기회를 넓힌다. - 나아가 이러한 역량 중심의 평가 방식은 기존의 스펙 중심 채용에서 벗어나 숨은 인재를 발굴하고, 인재 풀의 다양성을 강화하는 데 기여할 수 있다.
체계적인 편향성 검증 및 완화	- AI 채용 서비스는 채용 결과를 데이터 기반으로 분석하여 특정 집단에 불리한 결과가 반복되는지를 감시할 수 있다. - 편향이 드러날 경우 알고리즘과 가중치를 조정하거나 편향 완화 기법을 적용해 개선할 수 있으며, 이를 투명하게 공개하면 사회적 신뢰도 높일 수 있다. - 인간 평가자의 편견처럼 보이지 않고 지나치기 쉬운 차별 요소를 체계적으로 진단·보정함으로써, 공정성 기준을 지속적으로 강화하는 장치로 기능한다.

자료: 연구진 작성

[그림 2-33] 공정성 - 긍정 영향 지표 비교



	크기	지속 기간	발생 가능성
값	2.82	2.73	2.58

	크기	지속 기간	발생 가능성
값	2.82	2.89	2.68

	크기	지속 기간	발생 가능성
값	2.79	2.71	2.50

	크기	지속 기간	발생 가능성
값	2.86	2.57	2.57

자료: 연구진 작성

② 부정 윤리 영향

공정성 영역과 관련하여 AI 채용 서비스의 부정적 윤리 영향으로 '데이터 기반 차별 재생산 및 고착화', '공정성 기준의 불확실성과 신뢰 약화', '정형화된 인재상 의존의 한계'를 도출하였다. 세 영향의 구체적인 내용과 정량 평가 결과는 아래의 〈표 2-58〉, [그림 2-34]에서 확인할 수 있다.

[그림 2-34]의 공정성 영역의 부정 영향 주요 지표 그래프를 살펴보면, '데이터 기반 차별 재생산 및 고착화'는 크기, 지속 기간, 발생 가능성 모두에서 영역 평균보다 높은 값을 보여, 가장 두드러진 부정적 영향으로 인식되고 있었다. 이는 훈련 데이터의 편향이 알고리즘을 통해 반복적으로 증폭될 수 있다는 구조적 위험에 대한 전문가들의 우려를 반영한다. '공정성 기준의 불확실성과 신뢰 약화'

역시 크기와 발생 가능성이 평균보다 높게 나타나, 공정성 판단 기준이 불명확할 경우 채용 결과에 대한 신뢰가 저하될 수 있음을 시사하였다.

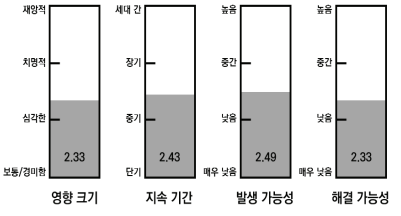
〈표 2-58〉 공정성 - 부정적 영향 유형

영향 유형	내용
데이터 기반 차별 재생산 및 고착화	- 과거 채용 데이터의 편향(성별·학력·지역 등)을 학습한 모델은 그 패턴을 통계적 규칙으로 재현해 특정 집단의 불이익을 자동화·지속화할 수 있다. - 민감정보를 직접 쓰지 않더라도 우편번호·동아리·졸업연도 같은 대리 변수로 간접 차별이 발생해, 다양한 배경의 지원자가 구조적으로 배제되고 불평등이 기술적으로 고착될 위험이 크다.
공정성 기준의 불확실성과 신뢰 약화	- 공정성은 사회적 합의가 필요한 가치인데, 기업·개발사가 임의로 정의해 알고리즘에 탑재하면 사회 통념과 어긋난 기준이 사실상 표준처럼 적용될 수 있다. - 더불어 블랙박스 구조와 설명 부족은 이의제기·책임 추궁을 어렵게 만들어, 조작·오남용 여부를 가리기 힘들게 하고 채용 결과에 대한 수용성과 신뢰를 약화시킨다.
정형화된 인재상 의존의 한계	- AI 채용 서비스는 고성과자 데이터나 내부 규칙에 따라 '이상적 인재상'을 일률적으로 정의하고, 여기에 부합하지 않는 지원자를 체계적으로 배제할 위험이 있다. - 이 과정에서 비전통적 경로를 통해 역량을 쌓은 인재나, 독특한 경험과 창의적 잠재력을 가진 인재는 정량화된 지표에서 불리하게 평가될 수 있으며, 이는 다양한 인재가 공정하게 기회를 얻는 것을 구조적으로 제약하는 문제로 이어질 수 있다.

자료: 연구진 작성

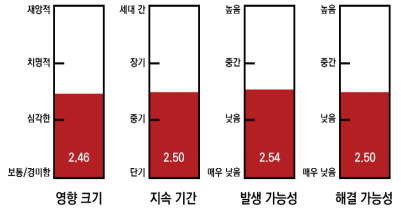
[그림 2-34] 공정성 - 부정 영향 주요 지표 비교

평균



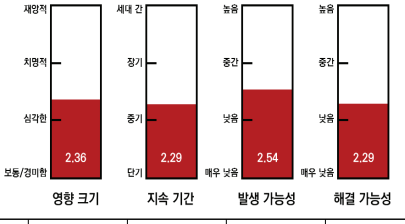
	크기	지속 기간	발생 가능성	해결 가능성
값	2.33	2.43	2.49	2.33

데이터 기반 차별 재생산 및 고착화



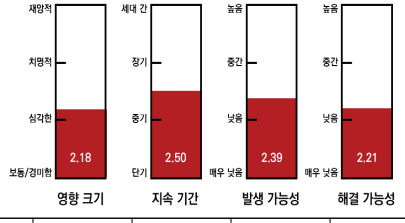
	크기	지속 기간	발생 가능성	해결 가능성
값	2.46	2.50	2.54	2.50

공정성 기준의 불확실성과 신뢰 약화



	크기	지속 기간	발생 가능성	해결 가능성
값	2.36	2.29	2.54	2.29

정형화된 인재상 의존의 한계



	크기	지속 기간	발생 가능성	해결 가능성
값	2.18	2.50	2.39	2.21

자료: 연구진 작성

③ 영향을 받는 대상

공정성 영역에서 영향을 받는 대상과 구체 상황은 아래의 〈표 2-59〉와 같다. 직접적 영향을 받는 대상으로는 '구직자'가 평가 기준·학습데이터·모델 편향에 따라 불공정한 평가나 차별을 직접적으로 경험할 위험이 가장 크고, 이의 제기 및 결과 검증 수단이 제한적이어서 구조적 불이익 가능성이 크다는 점에서 다수 포함되었다.

'채용기업'은 AI 판단 결과의 공정성 확보에 실패할 경우의 인사 결정 신뢰도 하락, 법적·평판 리스크 증가, 내부 검증 절차 강화 필요성 등으로 인해 간접적 영향을 받는 대상으로 평가되었다.

예상치 못한 영향을 받는 대상으로는 비교적 다수의 대상이 언급되었고, 그 중 'AI 개발사·서비스 제공자'와 '시민사회 및 외부 이해관계자'가 가장 많이 언급되었다. AI 개발사·서비스 제공자는 공정성 문제에 대한 사회적 비판이 강화될 경우 기술 신뢰성과 기업 이미지가 훼손될 가능성이 크고, 시민사회 및 외부 이해관계자는 공정성 부족이 사회 전반의 불평등 인식 심화 및 제도 개선 요구 확산으로 이어질 수 있다는 점에서 예상치 못한 영향 대상으로 평가되었다.

〈표 2-59〉 공정성 - 영향을 받는 대상

영향단계	영향을 받는 대상		N
1차 (직접적 영향)	구직자	AI 채용 시스템의 평가기준·학습데이터·모델 편향에 따라 구직자가 불공정하게 평가되거나 차별받을 위험이 가장 높음. 이의제기·결과 검증 수단이 부족하여 구조적 불이익 가능	26
	AI 개발사·서비스 제공자	불공정성 논란이 발생할 경우 기업의 사회적 책임과 브랜드 신뢰도 저하로 이어질 수 있음. 채용 절차의 공정성 검증 체계 필요	4
2차 (간접적 영향)	채용기업	AI 판단결과의 공정성 확보 실패 시 인사결정 신뢰도 하락, 법적·평판 리스크 부담. 내부 검증 절차 강화 필요	19
	구직자	반복된 탈락 등으로 인한 구조적 불공정 인식 확대. 불투명한 기준으로 인한 사회적 불신 누적	3
	시민사회 및 외부 이해관계자	AI 채용의 공정성 논란이 사회 전체의 불평등 담론으로 확산될 수 있음	3
	AI 개발사·서비스 제공자	공정성 논란 발생 시 기술적 책임과 고객사 신뢰 저하 부담	2
	규제·감독기관	공정성 기준 마련 및 분쟁 발생 시 조정 역할 강화 필요	2
	데이터 관련 주체	학습 데이터의 대표성 부족·편향 발생 시 결과 왜곡 책임이 연쇄적으로 발생 가능	2

영향단계	영향을 받는 대상		N
3차 (예상치 못한, 의도하지 않은 영향)	AI 개발사·서비스 제공자	공정성 문제에 대한 사회적 비판이 강화될 경우, 기술 신뢰성 저하 및 기업 이미지 손상 위험	6
	시민사회 및 외부 이해관계자	공정성 부족이 사회 전반의 불평등 인식 심화로 이어지며, 제도 개선 요구 확산 가능	6
	규제·감독기관	공정성 관련 분쟁·민원 증가로 행정 부담 확대. 법·제도 정비 필요	4
	기존 채용 서비스 업계	AI 채용 확산으로 공정성 중심의 경쟁 환경 조성 필요	3
	채용기업	공정성 기준 미비 시 기업 신뢰도 하락 및 내부 감사 부담 증가	3
	예비 구직자	AI 평가 기준에 맞추려는 자기검열, 다양한 경로의 지원 위축 우려	2
	투자·금융 관계자	공정성 논란이 지속될 경우 기업의 ESG 리스크로 인식될 가능성	2

주: 최소 2명 이상이 응답한 대상 기준, 중복응답 허용
자료: 연구진 작성

2) 정책지원과 대응 필요성 평가

① 프라이버시 보호, ② 포용성, ③ 책임성, ④ 투명성, ⑤ 공정성 5개 AI 윤리 영역에 대해 도출된 31개의 영향에 대한 정책 지원 및 대응 필요성을 평가하였고, 사분면 매핑 분석을 수행하였다. 그 결과를 요약해 제시하고자 한다.

긍정 영향의 경우 전문가 평가를 통해 분석한 '기대 수준' 값을 X축에, '정부 지원 필요성' 값을 Y축에 대응하여 사분면 그래프로 시각화하였다. [그림 2-35]를 보면, 두 지표를 기준으로 1사분은 '지속관리 영역', 2사분면은 '공공복리 영역', 3사분면은 '기타 영역', 4사분면은 '민간자율 영역'으로 구분된다.

1사분면인 지속관리 영역은 긍정적 효과에 대한 기대가 크고, 그 효과가 원활히 창출될 수 있도록 정부의 적극적이고 체계적인 지원이 지속적으로 요구되는 영향이 위치한다. 2사분면인 공공복리 영역은 긍정적 영향에 대한 기대 수준 자체는 상대적으로 크지 않지만, 공공의 안전과 복리를 위해 정부 주도의 대응이

필요하다고 판단되는 영향이 포함된다. 3사분면인 기타 영역은 긍정적 영향에 대한 기대수준이 제한적이고, 정부 자원을 투입할 필요성도 상대적으로 낮은 영역으로 해석된다. 마지막으로 4사분면인 민간자율 영역에는 민간이 주도적으로 대응할 수 있도록 정부의 직접 개입보다는 제도적·환경적 뒷받침 등 간접적 지원이 필요한 영향이 위치한다.

부정 영향의 경우 '우려 수준' 값을 X축에 대응, '정부 지원 필요성' 값을 Y축에 대응하여 사분면 그래프로 나타냈다. [그림 2-36]에 따르면, 1사분면은 '주의관심 영역', 2사분면은 '문제완화 영역', 3사분면은 '잠재위험 영역', 4사분면은 '시급해결 영역'으로 구분된다.

주의관심 영역은 부정적 영향에 대한 우려 수준이 높고 정부 대응도 일정 수준 이루어지고 있으나, 지속적인 모니터링과 예방적 관리가 요구되는 영향이 위치한다. 문제완화 영역은 부정적 영향에 대한 우려 수준이 크지 않고 정부 대응도 비교적 양호한 영향이 위치한다. 잠재위험 영역은 현재 부정적 영향에 대한 우려 수준이 높지 않지만 정부 대응 수준이 낮아, 환경 변화나 기술·시장 확산에 따라 향후 위험이 증폭될 가능성이 있는 영향이 포함된다. 시급해결 영역은 부정적 영향에 대한 우려 수준이 높음에도 불구하고 정부 대응이 상대적으로 미흡한 영역으로, 우선 대응이 필요한 영역이다.

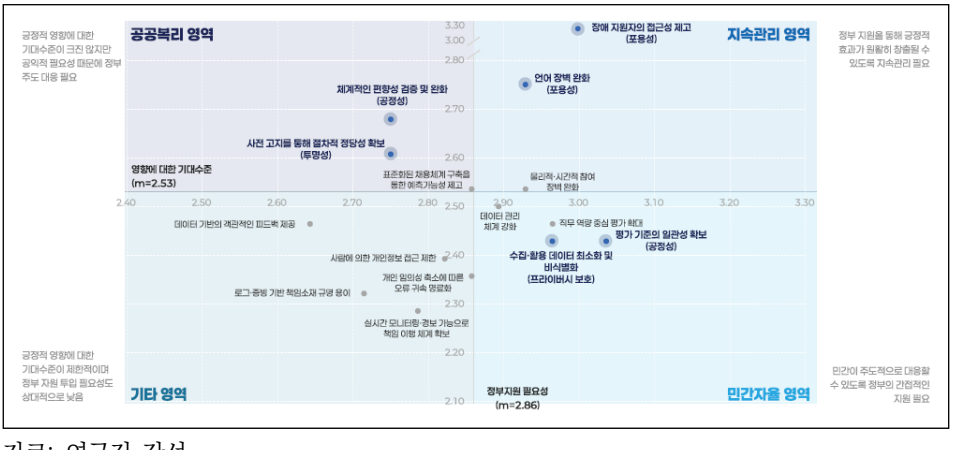
(1) 긍정 윤리 영향

긍정 윤리 영향의 경우 총 15개 영향이 도출되었으며, [그림 2-35]에서 분석 결과를 확인할 수 있다. 사분면별 분포를 살펴보면, 먼저 지속관리 영역에는 '장애 지원자의 접근성 제고(포용성)', '언어 장벽 완화(포용성)', '물리적·시간적 참여 장벽 완화(포용성)'이 위치하였다. 이는 포용성 관련 긍정 효과에 대한 기대 수준이 높고, 해당 효과가 안정적으로 발현될 수 있도록 중·장기적으로 정부의 지원이 필요하다는 점을 시사하기에 주목할 필요가 있다.

공공복리 영역에는 ‘체계적인 편향성 검증 및 완화(공정성)’, ‘사전 고지를 통해 절차적 정당성 확보(투명성)’, ‘표준화된 채용체계 구축을 통한 예측가능성 제고(투명성)’가 포함되었다. 이들 영향은 정부 주도의 제도 설계와 지원이 필요하다는 평가를 반영한다.

한편 민간자율 영역에는 ‘수집·활용 데이터 최소화 및 비식별화(프라이버시 보호)’, ‘평가 기준의 일관성 확보(공정성)’, ‘직무 역량 중심 평가 확대(공정성)’, ‘데이터 관리 체계 강화(프라이버시 보호)’가 위치하였다. 이는 해당 영역에서 민간의 자율적 혁신과 내부 관리 역량이 중요한 역할을 할 수 있으며, 정부는 민간이 주도적으로 대응할 수 있도록 뒷받침해야 함이 강조되었다.

[그림 2-35] 정부 정책 지원이 필요한 긍정적 영향(N=15)



자료: 연구진 작성

(2) 부정 윤리 영향

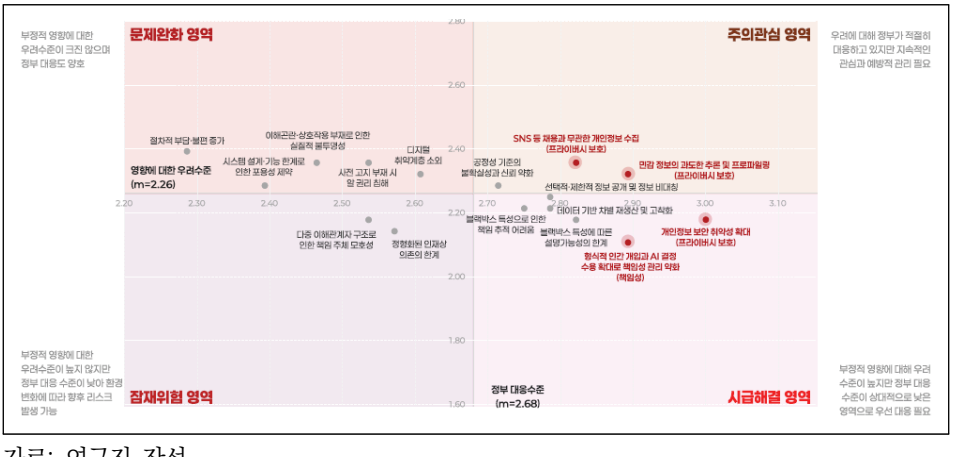
부정 영향의 경우 총 16개 영향이 도출되었으며, [그림 2-36]에서 그 분포를 확인할 수 있다. 이 가운데 시급해결 영역에 위치한 영향으로는 ‘개인정보 보안 취약성 확대(프라이버시 보호)’, ‘형식적 인간 개입과 AI 결정 수용 확대로 책임성 관리 약화(책임성)’, ‘블랙박스 특성에 따른 설명가능성의 한계(투명성)’, ‘데이터

기반 차별 재생산 및 고착화(공정성)', '선택적·제한적 정보 공개 및 정보 비대칭(투명성)', '블랙박스 특성으로 인한 책임 추적 어려움(책임성)'으로 식별되었다. 이들 유형은 부정적 효과에 대한 우려 수준이 높고, 동시에 정부 대응 수준이 상대적으로 낮은 것으로 평가되어, 향후 정책 설계 및 제도 개선에서 우선적으로 검토해야 할 핵심 관리 대상으로 해석된다.

잠재위험 영역에는 '다중 이해관계자 구조로 인한 책임 주체 모호성(책임성)', '정형화된 인재상 의존의 한계(공정성)'가 포함되었다. 현재 우려 수준은 상대적으로 크지 않지만, 정부 대응이 충분히 마련되어 있지 않아 제도·시장 환경 변화에 따라 향후 위험이 부각될 수 있는 영향으로 평가된다.

한편, 주의관심 영역에는 '민감 정보의 과도한 추론 및 프로파일링(프라이버시 보호)', 'SNS 등 채용과 무관한 개인정보 수집(프라이버시 보호)', '공정성 기준의 불확실성과 신뢰 약화(공정성)'가 위치하였다. 이들 유형은 정부의 제도적 대응이 적절히 이루어지고 있음에도 불구하고, 여전히 우려 수준이 높게 유지되는 사안으로 인식되고 있어, 현행 규제의 실효성 점검과 보완적 관리가 지속적으로 요구되는 영역으로 볼 수 있다.

[그림 2-36] 정부 정책 대응이 필요한 부정적 영향(N=16)



자료: 연구진 작성

제 5 절 AI 채용 서비스 윤리 확보를 위한 주체별 역할·과제

본 절에서는 앞서 수행한 영향평가 결과를 바탕으로 AI 채용 서비스의 윤리 확보를 위한 주체별 역할과 필요 노력을 정리하였다. AI 채용 생태계가 지속 가능하게 발전하기 위해서는 정부, 개발기업, 운영기업, 개인 등 네 주체가 각자의 위치에서 책임을 다하고 상호 협력하는 구조가 필수적이다. 특히 프라이버시 보호, 포용성, 책임성, 투명성, 공정성이라는 다섯 가지 핵심 윤리 영역에서의 역할이 유기적으로 연계될 때 신뢰할 수 있는 AI 채용 환경이 조성될 수 있을 것이다.

정부는 법제도 정비와 감독체계 구축을 통해 전체 생태계의 기반을 마련하고, 개발기업은 설계 단계에서부터 윤리 원칙을 내재화한 기술을 구현해야 한다. 운영기업은 실제 채용 과정에서 인간 중심의 검증 절차를 확립하여 시스템 운영의 책임성을 높여야 하며, 개인은 데이터 주체이자 채용 참여자로서 자신의 권리를 인식하고 적극적으로 행사함으로써 투명성과 공정성을 감시하는 역할을 수행해야 한다. 이러한 다층적 협력 구조 속에서 각 윤리 영역별 구체적인 역할과 실천 과제는 다음과 같다.

1. 프라이버시 보호

프라이버시 보호의 관점에서 정부는 AI 채용 과정에서 발생하는 민감정보 추론, 프로파일링 등 신유형의 개인정보 위험을 제도적으로 관리할 수 있는 기반을 마련하고, 감독·인증 체계를 통해 신뢰도를 높이는 역할을 수행해야 한다. 개발기업은 설계 단계에서부터 개인정보 최소화와 투명성 기능을 내재화하여 기술적 신뢰성을 확보할 필요가 있으며, 운영기업은 인사관리 전 과정에서 개인정보의 수집·이용·보관을 책임 있게 관리하고 지원자에 대한 설명 책임을 강화해야 한다. 개인은 정보주체로서 자신의 정보를 통제하고 권리를 적극적으로 행사함으로써 프라이버시 보호 체계의 중요한 참여자로서 역할을 수행해야 한다.

〈표 2-60〉 프라이버시 보호 영역에서의 주체별 역할

정부	개발기업	운영기업	개인
신뢰 기반 조성과 개인정보 보호 제도 정비	Privacy by Design 구현과 기술적 신뢰성 확보	인사관리 전 과정의 개인정보 보호 책임 실현	데이터 주체로서의 권리 행사와 정보관리 역량 강화
1. 법제·윤리기준 정립 - AI 채용 서비스 특화 법적·윤리적 기준 마련 - 민감정보 추론·프로파일링 등 신유형 위험에 대한 금지·관리 근거 강화 - 개인정보보호법 등 기존 제도의 사각지대 보완 2. 감독·인증체계 강화 - AI 채용 시스템 관리·감독 기관의 전문성·권한 강화 - 공공·민간 협력형 인증제 - 공신력 있는 평가·감독 체계 구축 3. 기술·관리적 보호조치 고도화 - 데이터 최소화·비식별화·접근통제 등 보호기술 표준 제시 - 공공부문이 선도적으로 모범사례 축적 4. 국경 간 데이터 관리·국제협력 - 글로벌 개인정보 보호 협력체계 구축	1. 설계단계의 윤리 내재화 - 개인정보 최소수집 및 비식별화 자동화 - 투명성·설명가능성 기능 내장 - 개인정보 처리 절차 사전 위험점검 체계화 2. 데이터 접근·관리 통제 강화 - 민감정보 별도 암호화·접근권한 최소화·이력관리 - 내외부 오남용 방지체계 구축 3. 윤리 검증 및 감사제도 운영 - 독립적 검증위원회, 외부감사 도입 - AI 자율점검표 등 내부 평가 - 학습데이터 출처 적법성·투명성 확보	1. AI 채용 프로세스 관리 - AI를 단독 결정 주체가 아닌 '보조적 판단도구'로 운영 - 인사담당자 교육을 통한 결과 해석·검증 역량 강화 2. 데이터 최소화 및 이용 이력 관리 - 직무역량 중심 정보만 수집·활용 - 개인정보 이용·보관·삭제 이력의 체계적 관리 3. 후속관리 및 설명책임 확보 - 지원자 문의·이의제기 창구 운영 - 설명·피드백 절차를 문서화하고 정기 점검 체계 구축	1. 권리 인식·통제권 강화 - 개인정보 제공 범위·목적을 명확히 확인하고 선택권 행사 2. 정보 리터러시 제고 - AI 채용 과정에서 개인정보가 어떻게 활용되는지 이해 - 개인정보 제공 시 위험 인식 향상 3. 피해예방·구제 요구 - 오남용 발견 시 적극적 구제 요구 - 신고·상담 창구 적극 활용

자료: 연구진 작성

2. 포용성

포용성을 강화하기 위해 정부는 디지털 접근성 보장 기준을 마련하여 포용적 채용 환경을 조성하고, 취약계층을 지원할 수 있는 제도적·물리적 기반을 확충해야 한다. 개발기업은 설계 단계에서부터 다양한 사용자 특성을 고려한 접근성 중심 설계를 구현하여 서비스 자체에 포용성을 내재화해야 하며, 운영기업은 실제 채용 절차에서 지원자의 다양성을 고려한 배려 조치와 포용적 운영문화를 정착시키는 역할을 수행해야 한다. 개인은 디지털 활용 역량을 높이고 다양성을 존중하는 태도로 채용 과정에 참여함으로써 포용적 생태계 형성에 기여해야 한다.

〈표 2-61〉 포용성 영역에서의 주체별 역할

정부	개발기업	운영기업	개인
디지털 약자 포함 포용적 채용 환경의 제도화	Accessibility by Design 내재화	포용적 채용 절차 운영	포용적 환경 속에서의 능동적 참여
1. 국가표준·접근성 가이드라인 제정 - 웹접근성·보편설계(Universal Design) 원칙을 반영한 표준 마련 2. 취약계층 지원 인프라 구축 - 장애인·고령자 등 대상 맞춤형 직업훈련, 지원 프로그램 운영 - 기기 대여 등 포용적 인프라 확대 3. 인센티브·인증제 운영 - 포용적 서비스를 운영하는 기업에 세제· 조달 등 인센티브 제공	1. 접근성 중심 설계 - 장애인용 보조기기·자막·음성안내·쉬운 언어 모드 반영 - 설계단계부터 보조기기 호환성 테스트 2. 언어·문화 다양성 반영 - 다국어 지원 및 문화적 표현 차이에 따른 오판 최소화 3. 취약계층 고려 데이터 구성 - 다양성 확보를 위한 편향 제거 및 테스트 세트 다변화	1. 포용적 채용 체계 구축 - 화상면접·온라인 직무테스트 등 접근성 높은 절차 운영 - AI 결과에 의존하지 않고 인간 검증 체계 구축 2. 장애인·외국인 지원자 배려 - 다국어 안내·보조기능 지원·속도조절 기능 활성화 3. 포용문화 확산 관리 - 차별 사례 모니터링 및 개선 교육 실시를 통한 포용적 채용 문화 확산	1. 디지털 역량 강화 - AI 채용 이해·활용 교육 참여, 원격 환경에 능동적 적응 2. 다양성 존중 문화 참여 - 차이를 인정하는 포용적 문화 조성 참여 - 차별적 결과 발견 시 피드백 제공 3. 기술 이해와 균형적 인식 - AI 한계를 인식하고 기술에 대한 비판적 관점 유지

정부	개발기업	운영기업	개인
- 공공조달 가이드라인 마련을 통한 기업 참여 유도 4. 피해구제 절차 정비 - 차별 피해 신고·구제 시스템 구축			

자료: 연구진 작성

3. 책임성

책임성 확보 측면에서 정부는 데이터 수집부터 운영·평가에 이르는 전 과정의 책임 추적 기반과 감독 체계를 마련하여, 문제 발생 시 원인 규명과 시정이 가능한 구조를 갖추어야 한다. 개발기업은 기록·설명·감사 기능을 설계 단계에서부터 내재화하여 시스템의 책임성을 기술적으로 담보해야 하며, 운영기업은 AI의 판단을 그대로 수용하는 것이 아니라 인간 검증 절차를 유지하고 지속적인 모니터링과 개선체계를 운영하여 실질적 책임을 강화해야 한다. 개인은 평가 과정과 결과를 비판적으로 살피고 권리를 행사함으로써 책임 이행을 감시하는 주체로서 역할을 수행해야 한다.

〈표 2-62〉 책임성 영역에서의 주체별 역할

정부	개발기업	운영기업	개인
책임성 확보 위한 법·감독 기반 구축	책임추적성과 설명가능성을 갖춘 설계	운영단계의 인간검증과 사후관리 체계를 통한 책임 경영 실현	책임이행 감시자이자 참여 주체
1. 기록·추적 표준화 - 데이터 수집-학습-운영-평가 전 과정 로그관리·보존 의무화 - 사후적 책임소재 명확화를 위한 근거 마련	1. 자동 로그·이력 관리 시스템 구축 - 전 과정을 자동으로 기록하고, 문제 발생 시 원인 규명 가능한 구조 설계 2. 설명가능성 강화	1. 운영 기록·증빙 관리 - 의사결정 이력 및 이의제기 처리결과 체계적 보관 2. Human-in-the- loop 절차 확립	1. 책임 감시 및 피드백 - AI 평가 과정과 결과를 확인하고 피드백 제공 - 정부·기업의 책임이행 감시 신고채널 적극 활용

정부	개발기업	운영기업	개인
2. 감독·감사 체계 강화 - 감독기관의 평가결과 열람·감사 권한 확대 - 로그 미비나 불투명성에 대한 제재 및 시정 명령 제도화 3. 자율규제·법제 병행 운영 - AI 윤리경영 인증제, 자율점검 기반 평가제도 도입 - 기업 책임성 확보를 위한 명확한 규제기준 정비	- 판단 근거를 시각화·요약해 이해할 수 있는 형태로 제공 - 이용기업·감독기관이 검증할 수 있도록 설명 인터페이스 내장 3. 검증·감사 절차 내재화 - 내부감사 루틴화, 제3자 평가제 도입 - 주기적 신뢰성 평가 및 개선	- AI 판단을 인사담당자가 검토·승인하는 구조 유지 3. 모니터링·개선 루프 운영 - 시스템 이상징후 발생 시 즉각적 개선 체계 마련 - 정기적 검토·보고 절차를 통해 책임성 강화	2. 자기기록 관리 - 제출자료·커뮤니케이션 이력 등을 보존하여 사후 책임 규명 근거 확보 - 사실과 다른 정보 제공 지양 3. 오류 인식·개선 참여 - 설명 부재·절차 위반 발견 시 신고 및 권리 행사

자료: 연구진 작성

4. 투명성

투명성을 강화하기 위해 정부는 고지·설명 기준을 제도화하고, 로그·평가지표 등 채용 과정의 핵심 정보를 표준화하여 사회적 신뢰 기반을 구축해야 한다. 개발기업은 설명가능성 기능과 피드백 기반 구조를 설계 단계에 포함하여 시스템 자체의 투명성을 높여야 하며, 운영기업은 지원자에게 평가 절차와 결과를 명확하고 이해하기 쉬운 방식으로 안내하고, 이의제기 절차를 명확히 하는 등 소통 중심의 운영을 실현해야 한다. 개인은 정보 열람·설명 요구권을 적절히 행사하고 개선 의견을 제시함으로써 투명성 확보에 기여해야 한다.

〈표 2-63〉 투명성 영역에서의 주체별 역할

정부	개발기업	운영기업	개인
투명성 제도화 및 공공 피드백 체계 구축	설명가능성과 피드백 기능을 내장한 시스템 설계	절차 고지·결과 설명을 통한 소통	정보 이해와 감시를 통한 신뢰 형성 참여
1. 투명성 기준 정비 - AI 활용사실 고지, 설명 요구권 보장 등 의무제 도입 2. 표준화된 로그·지표 관리 - 채용 단계별 표준화된 평가체계 및 로그 관리 지침 수립 - AI 판단 근거의 기록·열람·검증 가능성을 제도적으로 확보 3. 공공 리포트 발간, 기술 확산 - AI 채용 투명성 리포트 발간, XAI 기술 확산 지원	1. 사전 고지 모듈 설계 - AI 개입 단계, 자동화 수준 등을 명확히 알릴 수 있는 투명성 모듈을 내장하고 인터페이스 구성 2. 설명가능 모델 구현 - 판단 근거를 이해 가능한 형태로 시각화· 요약 제공 - 결과 해석 및 오류 확인이 가능하도록 설명가능성 강화 3. 피드백 루프 구축 - 반복적인 오판이나 편향을 줄이기 위한 데이터 기반 개선 루프 내장	1. 사전 고지 및 일관성 유지 - 채용 단계별 평가항목·AI 활용 정도 명확히 안내 - 채용 기준과 과정의 일관성 유지 2. 결과 설명·이의제기 절차 제공 - 지원자에게 결과를 요약·설명·피드백 형태로 제공 - 이의제기, 재검토 절차 명문화 3. 내부 투명성 관리 - 투명성 확보를 위한 교육·점검 - 책임자 지정, 로그 보관·관리	1. 정보 이해·권리 행사 - AI 채용 평가 구조 이해, 자동화 의사결정 여부, 데이터 처리 범위에 대해 정보 열람 및 설명 요구권 행사 2. 비판적 정보 활용 - AI 평가의 한계를 인식하고, 비판적으로 자신의 역량 보완 - 평가 결과를 자기 개발에 적극 활용 3. 소통·감시 참여 - 불투명한 절차 발견 시 개선 의견 개진 - 정부·기업의 투명성 정책에 피드백 제공

자료: 연구진 작성

5. 공정성

공정성을 확보하기 위해 정부는 차별을 방지하는 기준을 마련하고, 편향성을 검증할 수 있는 데이터셋과 평가 기준을 제공해 기업의 자율적 검증 환경을 조성해야 한다. 개발기업은 데이터 구성과 모델 설계 단계에서 편향을 최소화하고 직무역량 중심의 기준을 반영하여 알고리즘의 공정성을 확보해야 한다. 운영기업은 평가 기준의 일관성을 유지하고 구제 절차를 정비하여 공정한 채용 운영을 실천해야 하며, 개인은 객관적 정보를 기반으로 자신을 표현하고 부당한 결과에 대해 문제를 제기함으로써 공정성 유지에 기여해야 한다.

〈표 2-64〉 공정성 영역에서의 주체별 역할

정부	개발기업	운영기업	개인
공정성 기준 제도화 및 편향성 검증 인프라 구축	데이터·모델 단계의 편향 완화 설계	일관된 평가·구제 절차를 갖춘 공정한 채용 운영	공정한 경쟁 참여자이자 감시자로서의 역할
1. 공정성 기준 명문화 - 직무역량 중심 평가원칙·차별금지 원칙 제도화 2. 편향 검증 인프라 구축 - 정부 주도 검증용 테스트 데이터셋 개발·배포 - 기업이 자율적으로 공정성 검증을 수행할 수 있는 환경조성 3. 공정채용 지원·감독 - 공정채용 모범사례 확산 - 피해 발생 시 구제 절차 명확화 4. 데이터 다양성 확보 지원 - 공공데이터 허브 구축, 기업 다양성 확보를 위한 기술적·재정적 지원	1. 편향 감지·교정 알고리즘 내재화 - 성별·연령·언어 등 편향 요소 최소화 2. 직무 역량 중심 모델링 - 평가 기준을 직무 수행능력, 문제해결력, 경험 기반 역량에 초점 - 정략적 스펙이 아닌 직무 적합성 중심 모델링 3. 정기 검증 및 개선 - 외부 전문가 및 제3자 기관을 통한 정기적 공정성 검증 절차 운영 - 공정성 평가 보고서 공개 및 개선 이력 관리	1. 평가기준 일관성 유지 - 변경·적용 내역 기록 및 내부 검토 절차 강화 2. 피드백·구제 절차 명문화 - 평가근거·이의제기 절차를 명확히 안내 - 차별적 결과나 오판 가능성을 시정할 수 있는 체계 구축 3. 직무 중심 채용문화 확산 - 직무 적합성, 성과 역량 중심의 채용문화 확산 - 인사담당자 교육을 통한 인식 개선	1. 객관적 자기표현 - 데이터·성과 기반 자기소개 및 포트폴리오 준비 - 허위정보를 지양하고 객관적 근거로 자신의 역량 제시 2. 감시·참여 - 부당한 차별 또는 불합리한 결과 발생 시 피드백 시스템 적극 활용

자료: 연구진 작성

제6절 소 결

AI 채용 서비스 확산과 함께 실시한 국민 인식조사와 윤리영향평가 결과는, AI 기반 채용 방식의 효율성·일관성·절차적 정당성 측면에서 긍정적 기대를 받는 동시에, 정성적 요소 반영의 부족, 데이터 편향, 설명 가능성 한계 등 주요 위험 요인에 대한 우려 또한 뚜렷하게 존재함을 보여주었다. 서비스 경험 여부에 따라 인식 정도에는 차이가 있었으나, 정부의 사전 대비 수준·전문성·변화 관리 역량에 대해서는 전반적으로 신뢰가 높지 않아 제도적 신뢰 기반 강화가 향후 중요한 과제로 확인되었다.

영향평가는 전문가평가단과 국민포럼단을 포함하여 프라이버시 보호, 포용성, 책임성, 투명성, 공정성 등 5개 윤리영역에 대해 정량·정성 평가를 병행하였다. 그 결과 개인정보 최소 수집, 접근성 제고, 절차적 공정성 강화, 책임 추적성 향상 등 긍정적 영향이 확인된 한편, 민감정보 과도 추론, 디지털 격차 심화, 책임 공백, 설명가능성 한계, 데이터 편향에 따른 구조적 불공정 등 부정적 위험 또한 병존하는 것으로 나타났다. 특히 FGI 논의는 전문가평가단의 분석과 정합적인 방향을 보이며, 직무 무관 정보 활용 우려, 약자 배제 가능성, 책임 주체 불명확성, AI 결과의 무비판적 수용 위험, 설명력 제약 등 국민 관점의 구체적 우려를 드러냈다. 동시에 AI가 접근성과 프라이버시 보호, 절차적 일관성을 강화할 수 있다는 긍정적 전망도 제시되어 영향평가 결과를 보완적으로 뒷받침하였다.

영향의 성격에 따라 필요한 정책 대응의 우선순위도 구분되었다. 편향 검증과 사전 고지 강화 등은 공익적 관리가 필요한 영역, 장애 지원자 접근성, 언어 장벽 완화 등은 지속적 지원이 요구되는 영역, 수집·활용 데이터 최소화와 평가기준 일관성 확보 등은 민간 자율 개선이 효과적인 영역으로 나타났다. 반면 민감정보 과도 추론, 형식적 인간 개입, 보안 취약성 등은 정부의 시급한 대응이 필요한 영역으로 도출되어, 기술적·사회적 맥락을 동시에 고려한 차등적 규율체계의

필요성을 제기한다.

AI 채용 서비스의 윤리적 구현을 위해서는 정부-개발기업-운영기업-개인의 역할 분담과 협력적 생태계 구축이 필수적이다. 정부는 제도적 안전장치와 기준을 마련하여 신뢰 기반을 조성해야 하며, 개발기업은 설계단계에서부터 윤리를 내재화하는 프로세스를 확립해야 한다. 운영기업은 HITL(Human-in-the-Loop)을 포함한 책임 기반 운영체계를 마련하고, 개인은 권리 행사와 감시 참여를 통해 윤리 요구의 실질적 구현을 뒷받침해야 한다. 이러한 다층적 구조가 갖춰질 때, AI 채용 서비스는 기술적 효율성을 넘어 사회적 신뢰와 책임을 기반으로 지속가능한 방식으로 정착할 수 있을 것이다.

종합하면, 본 영향평가 결과는 AI 채용 서비스의 긍정적 가능성과 내재적 위험이 동시에 존재함을 보여주며, 신뢰 기반의 제도 설계, 위험 기반의 차등적 규율, 가치 중심의 윤리 운영체계를 결합한 통합적 정책 대응이 필요함을 시사한다. 향후 정책은 공정하고 책임 있는 채용 환경을 구축하기 위한 체계적 관리·지원 방향을 중심으로 발전해 나가야 할 것이다.

제 3 장 AI 윤리기준 자율점검표 개발·적용

제 1 절 분야별 점검표 개발: 헬스케어 분야 AI 윤리기준 자율점검표(안)

1. 개요

제1기 인공지능 윤리정책 포럼은 2022년 2월 24일 출범식을 갖고, '2022 인공지능 윤리기준 자율점검표(안)'을 공개하였다(과학기술정보통신부·정보통신정책연구원, 2022). 동 점검표는 개발·운영 주체가 각자의 상황에 맞춰 '인공지능(AI) 윤리기준'을 유연하게 적용할 수 있도록 범용적이고 포괄적인 문항들로 구성되었다. 그러나 바로 이러한 범용성 탓에 개별 산업 분야의 특수성을 충분히 반영하는 데 한계가 있었고, 그 결과 실제 기업 현장에서 활용도가 제한된다는 지적이 지속적으로 제기되었다.

이러한 현장의 목소리에 부응하여, 정보통신정책연구원은 2022년부터 산업별 특수성을 반영한 자율점검표를 개발하고 있다. 이때 기존 공통 문항을 기반으로 삼되, 각 분야의 특성에 비추어 강조가 필요한 문항을 선별·가공하고, 최신 AI 윤리 이슈를 반영한 문항은 추가하는 과정을 거쳤다. 그렇게 2022년 챗봇, 작문, 영상관제 분야를 시작으로 2023년에는 채용 분야, 2024년에는 영상 합성 분야에 특화된 자율점검표가 개발되었다. 해당 분야별 자율점검표는 2025년 2월, 개정된 '2025 인공지능 윤리기준 실천을 위한 자율점검표(안)'에 수록되어 공개되었다(과학기술정보통신부·정보통신정책연구원, 2025).

2025년에는 'AI 질병 예측 및 진단 서비스'를 신규 특화 분야로 선정하여, 이를 포괄하는 '헬스케어 분야 인공지능 윤리기준 자율점검표(안)'을 개발하였다. '헬스케어'는 문헌에 따라 다양하게 정의되지만(류기성·김일환, 2025), 넓게 해석하면,

질병을 치료하는 의료 행위뿐만 아니라, 건강 유지·증진·회복을 위한 제반 활동을 포괄하는 개념으로 정의할 수 있다(헬스케어 특별위원회·관계부처 합동, 2018). 여기에는 ① 건강 데이터를 분석하여 건강 유지 및 질병 예측을 돕는 '예방 및 웰니스', ② 의료진의 신속하고 정확한 진료 및 치료 의사결정을 지원하는 '진단 및 치료', ③ 치료 후 일상 복귀와 만성 질환 관리를 돕는 '예후 관리와 재활'이 포함된다(기능별 분류는 OECD/Eurostat/WHO, 2017). 본 연구는 세 영역 중에서도 높은 수익성과 도입률, 확실한 시장 수요에 힘입어 현재 시장의 주류를 형성하고 있는(식품의약품안전처, 2024) '진단 및 치료' 분야를 대상으로 하였다. 이는 「의료법」상 의료인의 전문적 판단이 요구되는 영역이며, 이러한 점을 명확히 하기 위하여 본 연구에서는 해당 분야에서 활용되는 AI 기술을 '의료 AI 기술'로 통칭하였다. 그리고 의료 AI 기술을 '방대한 양의 비정형 임상 데이터를 학습한 알고리즘을 바탕으로 질병의 패턴을 인식하고 예측 모델을 생성하는 기술'로 정의하여(Salathé et al., 2018), 단순한 규칙 기반 시스템(예: 고혈압 측정기 등)이나 비의료 건강 관리(예: 활동량 분석, 단순 수면 패턴 추적 등)와 구분하였다.

또한, 해당 기술이 적용된 서비스의 개발·운영·활용 전 과정을 윤리적 측면에서 다루기 위해 AI 질병 예측 및 진단 서비스를 '개인의 의료정보를 알고리즘으로 분석하여, 현재의 질병 유무를 판별하거나 향후 특정 질병 또는 증상이 발생할 확률을 미리 계산하여 의료진이나 사용자에게 제공하는 서비스'로 정의하고, 이를 헬스케어 분야 자율점검표의 적용 범위로 삼았다.

본 점검표는 2026년 초 발표될 '2026년 인공지능 윤리기준 자율점검표(안)'에 포함되어 대중에 공개될 예정이다. 이하에서는 헬스케어 분야 AI 윤리기준 자율점검표를 개발하게 된 배경, 개발 과정에서 고려한 주요 쟁점, 그리고 구체적인 절차와 내용을 중심으로 살펴보고자 한다.

2. 개발 추진 배경

가. 의료 AI 기술의 사회적 확산

의료 AI 기술은 의료 영상, 환자 임상기록, 생체신호, 유전체 정보 등 다양한 의료 데이터를 기반으로 진단·예측·치료 의사결정을 지원하는 기술로, 보건의료 전반의 핵심 기반 기술로 자리 잡으며 빠르게 확산되고 있다. Precedence Research (2024)는 글로벌 의료 AI 시장이 2023년 약 206억 달러에서 2030년 1,879억 달러(약 250조 원)로 성장할 것으로 내다보며, 연평균 성장률(CAGR)을 약 37%로 제시하였다. Fortune Business Insights(2024) 또한 2024년 의료 AI 시장을 약 290억 달러로 평가하고, 2032년에는 5,041억 달러(약 670조 원) 규모로 확대될 것으로 전망하는 등, 주요 시장조사기관들은 의료 AI 분야의 고성장을 공통적으로 분석하고 있다.

OECD(2024)는 의료 AI가 진단 정확도 향상, 치료 접근성 개선, 의료 자원의 효율적 배분, 개인 맞춤치료 강화 등 의료 성과 전반을 개선할 잠재력을 지닌다고 평가한다. WHO(2021) 또한 의료 AI가 예방·진단·치료·모니터링·공중보건 대응 등 다양한 기능을 수행하는 “의료 AI 시스템”으로 고도화되고 있으며, 향후 의료 분야의 표준적 도구로 자리잡을 가능성이 높다고 본다. 국내에서도 보건복지부(2024)는 ‘의료 인공지능 연구개발(R&D) 로드맵(안)(’24~'28)’을 수립하여 AI 기반 의료기술 혁신을 국가 핵심과제로 설정하고, 필수의료·신약개발 분야의 AI 연구개발 확대와 의료데이터 활용체계 고도화를 추진하고 있다.

한편, 의료 AI는 국내외 연구에서 여러 하위 영역으로 구분된다. 과학기술정보통신부·한국정보통신기술협회(TTA)의 「신뢰할 수 있는 인공지능 개발 안내서」(2023)는 의료 AI 활용 분야를 ① 이미징 및 진단, ② 정밀 의료 및 수술, ③ 어시스턴스,[9] ④ 프로세스 효율화, ⑤ 신약 개발의 다섯 영역으로 분류한다. 이 가운데 실제 임상·병원 환경에서 의료진의 의사결정과 환자 안전에 직접적

9) 환자나 의료진 사이의 소통을 원활하게 하고, 건강관리 전반이나 체계적인 치료 관리를 하는데 활용되는 인공지능 서비스

영향을 미친다는 점에서 윤리적 고려가 특히 필요한 ① 이미징 및 진단, ② 정밀 의료 및 수술을 중심으로 의료 AI 확산 현황을 살펴보고자 한다.

이미징 및 진단 분야는 의료 AI 중 임상 도입과 상용화가 가장 활발한 영역이다. 미국 FDA가 공개한 AI·ML 기반 의료기기 목록을 분석한 최근 연구에 따르면, 승인된 AI·ML 의료기기 중 약 75~77%가 영상의학 분야에 속하는 것으로 나타났다(Washington Post, 2025.4.5.). AI 기반 의료영상 분석 기술은 CT·MRI· X-ray 등 기존 장비의 판독 효율과 정확도를 높이는 핵심 분야로 주목받고 있으며, 대량 영상 데이터를 신속하게 분석해 이상 소견을 표시하거나 위험 환자를 우선 분류하는 솔루션은 영상의학과·응급실·검진센터 등에서 필수 도구로 자리잡아 가고 있다.

글로벌 의료기기 기업 GE 헬스케어, 지멘스 헬시니어스, 필립스 등은 AI 기반 영상 기능을 CT/MRI 등 장비에 통합하고, 영상 판독 자동화·노이즈 제거·병변 검출 보조 기능을 강화하는 방향으로 기술 개발을 확대하고 있다(Radiology Business, 2024.5.14.).

국내에서도 유방암·폐암 등 주요 암 조기 진단을 위한 AI 솔루션 도입이 확산되고 있다. 일례로, 루닛의 유방암 검진 보조 시스템은 실제 임상 도입 후 1년간의 평가에서 간격암 추가 발견과 판독 업무량 감소 성과를 발표하였다(Lunit, 2024. 12.2.).

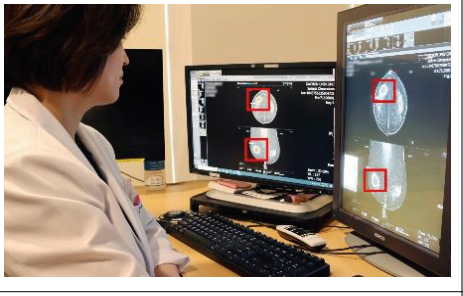
정밀 의료 및 수술 분야는 환자 개별의 유전체·임상 데이터를 활용한 맞춤형 치료, 의료진 판단을 보조하는 임상 의사결정 지원, 그리고 수술 자동화 기술 등 다양한 영역에서 빠르게 확장되고 있다. 이러한 기술은 치료 계획 수립의 정밀도를 높이고, 중환자 모니터링과 수술 절차의 효율성을 개선하는 등 의료 의사결정 전반에서 중요한 역할을 하고 있다.

국내에서는 상급종합병원을 중심으로 환자 상태 악화 예측, 패혈증 위험도 분석 등 치료 의사결정 보조 AI의 활용이 확대되고 있다. 예를 들어, 국내 의료 인공지능 기업 에이아이트릭스의 생체신호 분석 소프트웨어는 입원 환자의 전자의무기록

(EMR) 데이터를 기반으로 사망·심정지·예기치 않은 중환자실 전실·패혈증 발생 위험을 예측하며, 실제 병원 내 코드 블루 발생률을 약 25% 감소시켰다는 연구 결과를 발표하였다(에이아이트릭스, 2025.8.27.).

또한, 세계적 로봇수술 기업 인튜이티브 서지컬은 2024년 연례보고서에서 다빈치 로봇수술 시스템이 2023년 약 8,600대, 2024년 말 약 9,900대 설치되었다고 발표하며, 로봇 보조수술이 글로벌 의료 인프라의 중요한 축으로 자리잡고 있음을 보여주었다.

[그림 3-1] 의료 AI 기술 활용 영역

	
① 이미징 및 진단 - 루닛社의 유방암 검진 보조 AI 시스템	② 정밀 의료 및 수술 - 로엔서지컬社의 신장 결석 수술 로봇

자료: (좌) 머니투데이(2025.1.23.), (우) 조선일보(2025.7.14.)

이처럼 의료 영상 분석 및 진단, 정밀 치료, 로봇 보조수술, 환자 이상징후 예측 등 다양한 의료 AI 기술이 동시에 고도화되며, 의료 AI는 단일 기술을 넘어 의료 서비스 체계 전반을 구성하는 핵심 인프라로 빠르게 성장하고 있다. 이러한 확산은 의료의 효율성과 접근성을 크게 높이는 기회이지만, 동시에, 안전성·책임소재·형평성·프라이버시 등 새로운 위험도 함께 동반한다.

나. 의료 AI 기술 활용에 대한 우려

의료 AI 기술은 진단 정확도 향상, 치료 효율성 개선 등 긍정적 효과를 제공하지만, 동시에 환자 안전과 권익 측면에서 여러 우려가 제기되고 있다. 첫째, AI 기반 진단·치료 보조 기술의 오류 가능성이다. 의료 AI가 임상 의사결정 과정에 깊이 관여할수록, 알고리즘의 오작동이나 예측 오류는 환자의 생명·건강에 직접적인 위해로 이어질 수 있다. 미국 FDA(2024)는 "AI·ML 기반 의료기기는 알고리즘 업데이트 방식, 데이터 편향, 설명가능성 부족 등에 따라 성능 저하와 환자 위해 가능성이 존재한다"고 경고하였다. 실제로 최근 몇 년간 FDA는 AI·ML 기반 의료기기에서 진단·측정 오류와 관련된 리콜 사례를 다수 보고하며 임상 안전성 검증의 중요성을 강조하고 있다(TechTarget, 2025.8.26.). 특히 영상진단 AI는 개발 데이터의 특성, 병원 간 장비 편차, 임상 환경 차이 등에 따라 성능이 일관되지 않을 가능성이 제기되고 있다. 이는 데이터 편향(bias) 문제와도 밀접하게 연결된다. WHO(2021) 또한 의료 AI가 특정 인구집단에서 낮은 정확도를 보이거나, 학습 데이터 대표성 부족으로 인해 불평등한 의료 결과를 초래할 위험이 있다고 지적한다.

둘째, 환자 데이터 보호와 개인정보 침해 위험이 커지고 있다. 의료 AI는 EMR, 유전체, 병력 등 민감한 건강정보를 대규모로 활용하기 때문에, 데이터 유출 시 피해는 심각한 수준에 이를 수 있다. 2023년 미국 29개 주에서 206개 병원을 운영하는 의료 서비스 제공업체 커뮤니티 헬스 시스템(CHS)에서는 제3자 공급업체의 보안 사고로 약 백만 명 규모의 환자 건강정보(PHI, Protected Health Information)와 개인식별정보(PI)가 노출되는 사건이 발생했다(TechTarget, 2023.2.15.). 이 사례는 의료기관이 방대한 디지털 의료데이터에 의존하는 환경에서 보안 취약점이 환자 정보 유출로 직결될 수 있음을 보여준다. 의료기관과 여러 기업 간 데이터 공유가 증가하는 만큼, 개인정보 보호 체계가 미흡할 경우 의료 AI 생태계 전반의 신뢰 기반이 흔들릴 위험이 있다.

셋째, AI 의사결정 과정의 불투명성과 책임소재 문제다. OECD의 「AI in Health Report」(2024)는 의료 AI가 개입한 진단·치료 판단에 오류가 발생할 경우, 책임이 개발사·의료기관·의료진 중 누구에게 있는지 명확하지 않다는 점을 주요 위험 요소로 제시한다. WHO(2021) 또한 의료진이 AI의 권고를 충분한 검증 없이 신뢰할 경우 오진이나 과진단 등 환자 안전 문제로 이어질 수 있다고 경고하고 있다. 국내에서도 식약처와 보건복지부가 관련 가이드라인을 개정하며 AI 의료기기 검증 기준과 평가 절차를 강화하고 있으나, 실제 의료 분쟁에서 책임 귀속 기준이나 법적 판단 요소는 여전히 충분히 정립되지 않았다는 지적이 지속되고 있다(메디게이트, 2023.8.12.).

마지막으로, 의료 접근성의 디지털 격차 심화 우려도 제기된다. 고성능 AI 의료기기·영상 분석 시스템은 대형병원 중심으로 우선 도입되고 있으며, 중소병원이나 지역 의료기관은 비용·인프라 문제로 적극적인 활용이 어려운 실정이다.

이처럼 의료 AI는 의료 서비스의 효율성과 질을 개선하는 동시에, 임상 안전성·데이터 보호·책임소재·형평성 측면에서 기존 의료체계가 경험하지 못한 새로운 위험을 동반한다. 따라서 의료 AI의 확산은 단순한 기술 도입을 넘어, 안전성 검증·윤리적 평가·책임 체계 마련을 포괄하는 제도적 관리가 필수적으로 요구된다.

다. 의료 AI 기술에 대한 규제 동향

1) 유럽연합

유럽연합(EU)의 의료 AI 규제 환경은 여러 법령이 유기적으로 연계된 다층적 규제 구조로 특징지어진다. 이는 단일 법령이 아닌, AI 기술 전반의 위험을 통제하는 「AI 법(AI Act)」[10], 의료제품의 임상적 안전성을 검증하는 「의료기기법(MDR)」[11], 데이터의

10) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

11) Regulation (EU) 2017/745 of the European Parliament and of the Council

보호와 활용을 조율하는 「일반 개인정보 보호법(GDPR)」 및 「유럽 보건데이터 공간(EHDS)」으로 구성되어 있다. 이들은 차례대로 수평적, 수직적 및 기반적 통제 기제로 작용하여 의료 AI의 시장 진입과 활용을 엄격히 통제한다.

먼저 「AI Act」는 기술규제 측면에서 위험 기반 접근(Risk-Based Approach)에 바탕을 두고(EU의 「AI Act」에 관한 개요는 정보통신정책연구원, 2024) 의료 AI에 강력한 사전 통제 기제를 두고 있다(EU COM, 2020, 17면). 이 법 제6조 제1항은 부속서 I과 연계하여 「MDR」에 따른 적합성 평가(Conformity assessment) 대상인 의료 AI를 국민의 기본권과 안전에 중대한 영향을 미칠 수 있는 '고위험(High-risk) AI'로 분류하고, 이에 따라 「AI Act」는 제공자(Provider)에게 제품을 시장에 출시하기 이전에 데이터의 편향성을 탐지·수정하는 거버넌스 체계를 구축하고(제10조), 시스템의 판단 과정을 기록하는 로그 생성 기능을 탑재하며(제12조), 무엇보다 인간이 시스템의 오류에 언제든지 개입할 수 있는 '인간 감독(Human oversight)' 인터페이스를 구현할 법적 의무를 부과한다(제14조). 그렇게 「AI Act」는 의료 AI의 임상적 유효성을 논하기에 앞서, 기술 자체가 투명하고 통제 가능한지를 검증하는 첫 번째 관문을 두고 있다.

의료제품으로서의 안전성을 검증받기 위해 AI는 「MDR」을 통과해야 한다. 이 법의 핵심을 이루는 규제 작동 방식은 소프트웨어의 위험도에 따른 '등급 상향(Up-classification)'이다. 과거 지침(MDD) 체계하에서 의료 소프트웨어들은 대부분 1등급으로 분류되어 자가 선언(Self-declaration)만으로 시장에 진입할 수 있었지만, 「MDR」 이후 부속서 8 규칙 11의 신설로 진단 및 치료 정보를 제공하는 소프트웨어는 대부분 2a등급 이상으로 상향 조정되었다(「MDR」의 내용은 한국의료기기안전정보원, 2022). 그에 따라 제조사는 권한 있는 제3자 인증기관(Notified body)의 엄격한 기술 문서 심사와 임상 평가를 거쳐야만 CE 인증을 획득할 수 있게 되었다. 「MDR」과 「AI Act」 사이에 발생할 수 있는 이중 규제 문제에 대한 오해를 해소하기 위해, 2025년 6월 일종의 사례집을 제공하였다. 이는

---

of 5 April 2017 on medical devices.

MDR 인증기관이 「AI Act」의 요구사항까지 포함하여 단일 적합성 평가 심사를 수행한다는 점을 명확히 함으로써(MDCG/AIB, 2025) 개발사들이 예측가능한 환경에서 제품 및 서비스를 개발할 수 있는 토대를 마련하였다. 또한 「MDR」은 허가 이후에도 실제 임상 현장에서 수집된 데이터를 바탕으로 지속적인 성능 모니터링을 의무화하는 '사후 시장 감시(PMS)' 제도를 도입하였으며(제83조), 이는 학습 데이터와 실세계 데이터의 괴리로 발생할 수 있는 AI의 성능 저하를 지속적으로 통제하는 핵심 기제로 평가되고 있다(Gilbert et al., 2021).

의료 AI의 핵심 자원인 데이터 영역에서는 보호와 활용이라는 상충 가치를 조율하는 이원화된 규제 기제가 작동한다. 「GDPR」은 정보주체의 개인정보자기결정권을 최우선으로 삼아 '자동화된 의사결정'을 원칙적으로 금지하고 있으며(제22조), 그에 따라 AI가 의사의 개입 없이 환자의 진단이나 치료 방향을 독단적으로 결정하는 것은 금지된다. 또한, 「GDPR」 제13조 제2항 (f), 제14조 제2항 (g), 제15조 제1항 (h)는 정보주체에게 알고리즘의 판단 논리와 예상되는 결과에 관한 정보를 요구할 수 있는 권리를 부여함으로써(이상용/이혜리, 2024), 기술적으로 '설명가능한 AI' 구현을 간접적으로 강제하는 효과를 가진다. 한편, 2025년부터 본격적인 준비 단계에 돌입한 「EHDS」는 개인정보의 과도한 보호로 인해 연구·혁신·정책 수립 등이 저해되지 않도록 2차 활용을 법적으로 보장한다(EU COM, 2022, Recital 41). 동법은 제34조 제1항 (f)에 공익적 연구나 AI 개발 목적으로 데이터를 2차 활용할 경우, 개별 환자의 동의를 일일이 받는 대신 '보건의료 데이터 접근 기구(Health Data Access Body, HDAB)'의 데이터 이용 허가를 통해(제45조) 가명화된 데이터를 활용할 수 있는 법적 통로를 제공한다.

정리하면, EU의 의료 AI 규제 체계는 '기술-제품-데이터'라는 세 가지 차원이 맞물린 통합적 기제를 기반으로 구축되어 있다. 기업은 「MDR」상 인증을 획득하는 과정에서 「AI Act」의 투명성 요건과 「GDPR」의 데이터 처리 기준을 동시에 충족하여야 한다. 이러한 통합적 접근은 높은 진입장벽으로 작용하기도 하지만, 역설적으로 가장 신뢰할 수 있는 의료 AI 생태계를 구축하는 기반이기도 하다.

2) 미국

미국의 의료 AI 규제 환경은 EU의 통합적 접근과 대조적으로 식품의약국(FDA)을 중심으로 한 '섹터별 규제(Sector-Specific Regulation)'와 기술 발전 속도에 맞춘 가이드라인 기반 규율로 특징지어진다(최정윤/최신혜, 2024). 이는 별도의 포괄적 입법을 시도하기보다 기존의 「연방식품의약품화장품법(FD&C Act)」과 「21세기 치료법(21st Century Cures Act)」, 「건강보험 이동성 및 책임법(HIPAA)」 등 기존 법체계를 AI 특성에 맞게 해석하고 보완하는 실용주의적 관점을 취한 결과이다. 규제 체계는 내용상 크게 제품의 변화를 관리하는 FDA의 '전 주기적 관리(Total Product Lifecycle, 이하 TPLC)', 기능의 투명성을 요구하는 국가 보건정보기술조정실(ONC)의 인증 규제, 데이터 프라이버시를 다루는 「HIPAA」 및 연방거래위원회(FTC)의 사후 통제로 구별된다.

규제 체계의 핵심은 FDA가 주도하는 전 주기적 관리 체계의 확립이다. FDA는 2021년 행동 계획(FDA, 2021)을 통해 제품 출시 시점에 '고정된' 성능만을 평가하는 기존의 정태적 통제에서 벗어나, 학습을 통해 성능이 변화하는 AI의 특성을 반영하여 TPLC 체계로 전환한다는 점을 공식화하였다. 이를 구현하기 위하여, FDA는 동 행동 계획에서 '사전 변경 계획(Predetermined Change Control Plan, 이하 PCCP)'을 소개하였고, 2025년 별도의 가이드라인(FDA, 2025)을 통해 AI 기반 해당 제도에 대해 상세하게 안내하고 있다. PCCP는 개발사가 AI 모델의 성능 향상이나 알고리즘 변경 계획을 사전에 제출하여 승인받으면, 해당 범위 내에서 이루어지는 업데이트에 대해서는 복잡한 재허가 절차를 면제해 주는 제도이다(최정윤/최신혜, 2024). 이는 안전성이 담보된 계획된 변화는 허용하되, 그 범위를 벗어나는 변화는 엄격히 통제함으로써 혁신의 속도와 안전성 사이의 균형을 맞추는 규제 합리주의의 전형적인 사례에 해당한다.

한편, 「21세기 치료법」은 의료 AI를 위험도에 따라 규제 대상과 비규제 대상으로 구분하는 '선별적 규제 완화' 방식을 취하고 있다(Levine et al., 2022). 이 법은 의료진이 AI의 판단 근거를 투명하게 이해할 수 있고, 환자의 상태가

위중하지 않은 때 사용되는 '임상 의사결정 지원 시스템(CDSS)'을 FD&C Act의 적용 대상인 의료기기에서 제외한다. 즉, 소프트웨어가 설명가능하고, 의사의 주도적 판단 가능성을 보장한다면 규제 당국의 개입을 최소화하여 시장 진입을 지원하겠다는 의도이다(최정윤/최신혜, 2024). 알고리즘 내부를 알 수 없거나 중대한 진단을 수행하는 AI는 여전히 FDA의 엄격한 인허가(510(k) 또는 PMA) 트랙을 거치게 함으로써 위험에 비례한 통제력을 유지한다.

2024년 이후 새롭게 부상한 규제 방식은 보건복지부 산하 ONC가 주도하는 '알고리즘 투명성 및 편향성 통제'이다(최정윤/최신혜, 2024). HTI-1(Health Data, Technology, and Interoperability) 규칙은 전자의무기록(EHR)에 연동되는 모든 예측 AI 모델에 대해 '의사결정지원개입(Decision Support Intervention, DSI)의 투명성'을 인증 요건으로 삼고 있다. FDA가 제품 자체의 안전성을 본다면, ONC는 의료체계 내에서 AI가 공정하게 작동하는지 감시하는 것이다. 개발사는 AI가 어떤 데이터를 학습했는지, 인종이나 성별에 따른 편향은 없는지, 성능 검증은 어떻게 수행했는지를 사용자인 의료진에게 투명하게 공개하여야 하며, 이를 충족하지 못하면 미국 공공 의료보험 시장에서의 사용이 제한된다.

데이터 측면에서는 「HIPAA」에 따른 비식별화와 「연방거래위원회법」에 따른 기만적 행위 감시가 상호보완적으로 작동한다. 「HIPAA」는 AI 학습에 사용되는 의료정보에서 18가지 개인 식별 정보를 제거(De-identification)하도록 규정함으로써 프라이버시를 보호한다(45CFR § 164.514(b)). FTC는 데이터를 정보주체의 동의 없이 수집하는 행위를 '부당하고 기만적인 거래 관행'으로 규정하고 강력한 사후 제재를 가한다(FTC법 제5조). 이는 별도의 프라이버시 보호법을 제정하는 대신 강력한 소비자 보호 기구의 집행권을 통해 기업의 데이터 오남용을 억제하는 미국 특유의 시장 중심적 통제 방식에 해당한다(최호영, 2025).

3) 대한민국

대한민국의 의료 AI 규제 환경은 2025년 「인공지능기본법」의 제정과 「디지털의료제품법」의 시행이 맞물리면서, 과거 하드웨어 중심의 단편적 규제에서 벗어나 기술의 신뢰성, 제품의 안전성은 물론 「의료법」상 의료 행위의 책임성을 유기적으로 연결하는 구조로 고도화되었다. 여기에 연구·개발 단계의 윤리를 심의하는 「생명윤리 및 안전에 관한 법률」(이하, 「생명윤리법」이라 한다)이 더해져 촘촘한 안전망을 형성하고 있다.

「생명윤리법」과 이에 근거한 기관생명윤리위원회(IRB)의 심의 제도는 의료 AI 기술의 연구·개발 이는 AI가 아직 모델의 실체를 갖추지 않은 연구 단계에서 기능하며, '벨몬트 보고서'가 제시한 세 가지 윤리 원칙 '인간 존중', '선행(Beneficence)', '정의(Justice)'를 기준으로(국가과학기술인력개발원, 2021) 연구의 윤리성을 엄격히 심사한다. 의료 AI 기술 연구에 이를 적용하면, 데이터 수집 과정에서 피험자에게 충분한 정보를 제공하고 적법한 동의를 구했는지(인간 존중), AI 학습에 필요한 환자의 민감한 의료정보가 윤리적으로 적법하게 수집되었는지, 임상 연구 과정에서 피험자의 권리와 안전이 침해되지 않았는지(선행), 학습 데이터가 특정 집단에 편중되지 않고 공정하게 수집되었는지(정의)를 평가할 수 있다.

연구 단계를 통과하여 개발된 의료 AI 기술은 2025년 제정된 「인공지능기본법」에 의해 규율된다. 법 제2조 제4호 라목은 「의료기기법」에 따른 의료기기 및 「디지털의료제품법」에 따른 디지털의료기기의 개발·이용에 활용되는 AI를 사람의 생명·신체에 중대한 영향을 미치는 고영향 인공지능으로 정의하고 있다. 이러한 법적 지위는 의료 AI 사업자에게 강력한 의무를 부과하는 근거가 된다. 구체적으로 사업자는 고영향 AI의 안전성과 신뢰성을 확보하기 위해 위험관리 방안을 수립해야 하며(제32조), 특히 기술적으로 가능한 범위 내에서 AI가 도출한 최종 결과와 이에 활용된 기준을 이용자에게 설명할 방안을 마련해야 한다(제34조 제1항). 또한 생성형 AI가 활용되는 경우, AI 사업자는 해당 제품이 AI에

기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 하며(제31조), 제품 출시 전 해당 AI가 사람의 기본권에 미치는 영향을 평가하도록 노력해야 한다(제35조).

이러한 규제 아래에서 「디지털의료제품법」은 실질적인 제품의 시장 진입을 통제하는 역할을 담당한다. 기존 「의료기기법」하에서 AI 소프트웨어는 엑스레이 장비 등 하드웨어와 동일한 범주 내에서 관리되었으나, 「디지털의료제품법」 시행으로 '디지털 의료제품'이라는 별도의 트랙으로 분리되었다. 그 결과 소프트웨어의 특성에 맞게 유연한 등급 분류와 심사 체계가 가능해졌다. AI 소프트웨어는 물리적 위해성이 낮다는 점에 착안하여 복잡한 임상시험 대신 실사용 데이터(RWD)를 활용한 임상적 성능 평가가 허용되었고(제15조), 그에 따라 시장 진입 장벽이 대폭 낮아졌다. 그밖에 식약처는 '생성형 AI 기반 의료기기 허가·심사 가이드라인(2025)'을 정비하여 생성형 AI 등 신기술 적용에 대한 구체적인 심사 기준을 제시하고, '의료기기의 사이버보안 허가·심사 가이드라인(2024)'에 AI 모델에 대한 공격이나 데이터 위변조 위험에 대응하기 위해 사이버보안을 허가의 요건으로 명시하여, 정확성뿐만 아니라 보안성도 안전의 척도라는 점을 분명히 하였다.

제품이 허가를 받고 실제 의료 현장에 투입되면, 「의료법」에 의한 '인간 중심의 행위 통제'가 규제를 주도한다. 「의료법」 제27조 제1항(무면허 의료행위 등 금지)은 아무리 고도화된 AI라도 의료인의 개입 없이 단독으로 진단을 내리거나 처방을 결정하는 것을 차단한다. 따라서 법적으로 AI는 의사의 판단을 보조하는 '도구'의 지위에 머물며, AI가 도출한 결과에 대한 최종 확인과 서명은 반드시 면허를 취득한 의료인이 수행해야 한다(김화, 2023). 이러한 방식은 AI 오진 시 법적 책임을 오로지 의료인에게 귀속시킴으로써 의료 현장에서 AI가 무분별하게 남용되는 것을 방지하는 강력한 안전망으로 기능한다.

데이터 규제 영역에서는 「개인정보 보호법」과 보건복지부의 '보건의료데이터 활용 가이드라인(2025)'이 주도하는 '가명처리를 통한 데이터 효용 극대화'가

강조되고 있다. 과거에는 환자의 동의 없이는 의료정보를 AI 학습에 사용하는 것이 사실상 불가능했으나, 현재는 데이터를 특정 개인을 알아볼 수 없도록 가공하고 적절한 안전조치를 취한다면 정보 주체의 동의 없이도 과학적 연구 및 상업적 통계 작성 목적으로 활용할 수 있는 법적 근거가 마련되어 있다(제28조의2 제1항). 이 규정으로 인해 고품질의 AI 모델 학습을 위한 핵심 자원의 확보가 용이해졌다.

4) 소결

유럽연합, 미국, 대한민국의 의료 AI 규제 동향을 자세히 살펴보면, 규제 흐름은 기술적 차원의 성능 검증에 머물지 않고, 더 나아가 기술의 사회적·윤리적 신뢰성을 법적으로 확보한다는 관점으로 수렴하고 있음을 확인할 수 있다. 인간의 감독과 투명성 확보, 전주기적 관리를 통한 지속적인 안전성 요구, 설명 가능성과 책무성의 강조가 그 예이다. 이러한 규제 방향은 윤리적 정당성이 선결되었을 때 비로소 의료 AI 기술이 인류의 건강 증진이라는 '선행'의 도구로 온전히 기능할 수 있다는 사고에 근거한다.

이처럼 국내외적으로 강력한 법적 규제 체계가 마련되고 있음에도 불구하고, 별도의 'AI 윤리 자율점검표'를 개발하여야 할 정당성은 여전히 유효하며 오히려 더욱 강화되었다고 볼 수 있다. 첫째, 법적 규제는 기술의 발전 속도보다 느리며 최소한의 강제 규정만을 담고 있다. 법의 사각지대에 존재하는 윤리적 공백을 메우기 위해서는 개발 단계에서의 선제적인 자율 점검이 필수적이다. 둘째, 법적 규제가 사고 발생 후의 처벌이나 인허가 단계의 통제에 집중하는 반면, 자율점검표는 연구·개발 단계부터 잠재된 윤리적 위험을 스스로 식별하고 수정할 계기를 제공함으로써 윤리의 내재화를 구현하고 사회적 비용을 예방하는 효과적인 수단이 된다. 셋째, '인권보장'이나 '다양성 존중'과 같이 추상적 용어는 개발자나 이용자가 현장에서 즉각적으로 적용하기 어렵다. 이때 자율점검표는 거시적 규범을 구체적인 행동 지침으로 번역하여 제공하고, 이를 통해 의료 AI 생태계

전반의 실질적인 윤리 역량을 강화하는 한편, 규제 준수를 넘어 신뢰할 수 있는 의료 AI 서비스를 사회 전반에 안착시키는 데 기여할 것이다.

3. AI 윤리기준 핵심요건별 쟁점 사항

가. 의의

지금까지 헬스케어 분야 시장의 성장 추세, 의료 AI 기술의 활용에 대한 사회적 우려, 국내외 관련 규제 동향 등을 살펴보았다. '헬스케어 분야 인공지능 윤리기준 자율점검표'의 구체적인 문항은 이러한 배경을 바탕으로 의료 AI 서비스의 개발·활용 과정에서 제기될 수 있는 쟁점으로부터 도출되어야 한다.

본 자율점검표의 개발 목적은 헬스케어라는 특정 산업 분야에 「인공지능(AI) 윤리기준」을 실천적으로 내재화할 수 있는 계기를 마련하는 데 있다. 기존 '2022 인공지능 윤리기준 자율점검표(안)'은 범용성·포괄성을 매개로 다양한 영역에서의 AI 윤리 확산에 이바지하였으나, 개별 분야의 특수성을 반영하는 데 한계가 있다는 지적을 마주할 수밖에 없었다. 이에 본 연구는 첫째, 해당 분야의 특수성을 고려하여 기존 문항 중 특히 강조되어야 할 문항을 선별·가공하고, 둘째, 의료 AI 서비스와 관련하여 새롭게 제기될 수 있는 윤리적 쟁점을 식별하고 이에 대응하는 신규 문항을 개발하는 방향으로 추진되었다.

출발점은 무엇보다도 「인공지능(AI) 윤리기준」이며, 특히 기준을 제시하는 10대 핵심요건을 골격으로 삼아 세부 점검문항을 구성하였다. 10대 핵심요건은 ① 인권보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성을 가리킨다. 아래에서는 헬스케어 분야에서 AI 윤리를 준수하기 위해 요건별로 고려해야 할 주요 쟁점 사항을 순차적으로 검토하고자 한다.

나. 인권보장

「인공지능(AI) 윤리기준」의 첫 번째 핵심요건은 인권보장이다. 비록 이 요건이 아홉 개의 핵심요건과 병렬적으로 나열되어 있기는 하지만, 「인공지능(AI) 윤리기준」이 '인간성'을 최고 가치로 설정하고 있음(문정욱 외, 2020)을 떠올려보면 그 실질이 나머지 요건과 동등하지 않음을 알 수 있다. 인권보장은 다른 핵심요건의 전제이고, 동시에 기본 방침의 성격을 가진다. 따라서 인권보장에 관한 점검문항을 작성할 때는 기본 가치이자 원칙으로서 인권보장 요건의 특수성이 잘 나타나도록 구체화하여야 한다. 이는 AI가 인간을 단순한 데이터나 수단이 아닌 목적으로 대우하고 있는지를 성찰하게 하는 근본적인 질문을 던지는 것에서 시작한다.

헬스케어 분야에서 인권보장은 환자와 의료진이라는 두 핵심 주체가 AI 기술에 종속되거나 도구화되지 않도록 하는 데 초점을 맞춰야 한다. 의료 AI 기술이 효율성을 전면에 내세워 인간을 데이터 처리의 대상으로 전락시킬 위험이 있기 때문이다. 이러한 맥락에서 의료 AI 기술이 환자를 데이터 객체로 취급하거나, 의료진을 단순히 기계를 조작하는 운영자로 격하시킬 가능성을 인식하고 이를 방지하기 위한 실행 방안과 지속적인 실천 노력을 기울이고 있는지를 점검할 필요가 있다. 이는 기술 만능주의로 인해 자칫 간과되기 쉬운 인간의 주체성을 유지하기 위한 선결 과제이다.

인간의 주체성은 구체적인 의료 행위 과정에서 의료인의 최종 판단 권한으로 발현되어야 한다. 이는 사람 중심 AI의 원칙을 의료 현장에 적용한 것이며, EU의 「AI 법」과 우리 「의료법」이 이미 규정하고 있는 사항이기도 하다. 의료 AI의 진단이나 예측을 성찰 없이 받아들일 경우, 의료인은 의료 AI의 '연장된 팔'에 불과하게 된다. 의료인이 의료 현장에서 도구로 격하되지 않도록, 의료 AI 서비스는 의료진이 AI의 결과를 비판적으로 검토하고, 자신의 전문성과 자율성에 기반하여 최종 판단을 내리는 데 필요한 기능적·절차적 장치를 제공하기 위해 노력하여야 한다.

마지막으로 AI와의 상호작용 과정에서 발생할 수 있는 정서적·인격적 침해 가능성도 고려하여야 한다. 서비스의 인터페이스나 결과 전달 방식이 인간의 감정을 배려하지 못한다면 인간의 존엄성이 훼손될 여지가 있기 때문이다. 그러므로 의료 AI 서비스가 의료진에게 불쾌감이나 모욕감을 주지 않는 언어와 태도로 설계되었는지 지속적으로 탐지하는 체계를 갖춤으로써 기술적 효용을 넘어 사용자의 인격을 보호하기 위해 노력해야 한다.

다. 프라이버시 보호

「인공지능(AI) 윤리기준」의 두 번째 핵심요건인 프라이버시 보호는 개인의 사생활 보호와 개인정보의 오남용을 예방하는 차원을 다룬다. AI 시스템을 개발하고 운영하는 데 있어 개인정보가 학습 데이터나 입력 데이터로 활용될 수 있고, 그 과정에서 개인정보가 헌법 및 법률이 보장하는 정보주체의 권리를 침해하는 방식으로 처리될 여지가 있기 때문이다. 헌법 제17조는 사생활의 비밀과 자유를 명시적으로 보장하고 있으며, 더 나아가 헌법재판소는 헌법 제10조 및 제17조로부터 개인정보자기결정권을 도출하고 이를 독자적인 기본권으로 승인한 바 있다. 개인정보자기결정권은 자신에 관한 정보를 언제 누구에게 어느 범위까지 공개할 것인지 스스로 결정할 수 있는 정보주체의 권리를 보장한다. 이는 개인의 사생활이 그의 의사에 반하여 공개되는 위험으로부터 보호하는 소극적 차원을 넘어 개인이 자기정보에 관하여 능동적으로 통제할 수 있는 적극적 권리를 보장한다는 점에서 중요한 헌법적 의미가 있다. 프라이버시 보호는 특히 민감한 의료정보를 다루는 헬스케어 분야에서 세심하게 고려되어야 한다. 헬스케어 분야 의료 AI의 프라이버시 보호 점검 문항은 이러한 헌법적 가치를 반영하여 '법 준수'와 '위기관리'라는 두 가지 축으로 구성되었다.

다른 핵심요건에 비해 프라이버시 보호는 이미 매우 구체적인 관련 법령 및 가이드라인에 의해 규율되고 있다. 이는 일반적인 개인정보보다 더욱 엄격한 관리가 요구되는 '민감정보'(「개인정보 보호법」 제23조)를 다루고 있기 때문이다.

의료 AI 개발 과정에서는 방대한 환자 데이터를 학습용으로 활용하기 위해 「개인정보 보호법」 제28조의2에 따라 '가명처리' 과정을 거치며, 가명정보는 동법 제28조의3 내지 제28조의5 및 동법 시행령의 법적 제한 아래에 놓인다. 보건복지부는 가명 의료정보 활용의 안전성을 담보하기 위하여 '보건의료데이터 활용 가이드라인(2025)'에서 의료 분야의 특수성을 반영한 가명처리 절차를 상세하게 안내하고 있다. 해당 가이드라인은 의료 데이터 활용 시 ① 사전준비 ② 가명처리 ③ 적정성 검토 ④ 활용 및 사후관리 단계를 밟도록 하며, 특히 내부 데이터 심의위원회 또는 기관생명윤리위원회(IRB)를 통한 엄격한 적정성 검토를 요구한다. 또한 개인정보보호위원회의 '인공지능(AI) 개인정보보호 자율점검표(2021)'는 AI 설계·개발·운영 단계에서 준수해야 할 적법성, 안전성, 투명성 등 6대 원칙과 단계별 이행 사항을 제시하고 있다. 이처럼 법적 요구사항이 가이드라인 등을 통해 구체화되어 있는 상황은 본 자율점검표가 섣불리 독자적인 기준을 나열하기보다는, 기존의 요구사항을 포괄적으로 준용하여 체계적 정합성을 확보하는 동기로 작용하였다. 이에 의료 AI 시스템을 개발·활용하는 경우, 「개인정보 보호법」 등 상위 법령 위반 사항을 점검하는 것은 물론, 의료 데이터의 특수성을 고려한 가이드라인 및 다른 자율점검표 등 기존의 권고사항을 충실히 이행하였는지 확인함으로써 법적 정당성을 우선 확보하도록 하였다.

물론 법령상 요구사항의 준수만으로는 기술적 오류나 예기치 못한 침해 사고를 모두 차단하기 어렵다. 「개인정보 보호법」 제29조(안전조치의무)는 "개인정보가 분실·도난·유출·위조·변조 또는 훼손되지 아니하도록 내부 관리계획 수립, 접속기록 보관 등" 안전성 확보에 필요한 기술적·관리적 및 물리적 조치를 하도록 규정하고 있다. 더욱이 의료 AI 서비스 특유의 리스크인 재식별화 가능성, 민감정보 유출 등은 정보주체에게 회복하기 어려운 피해를 줄 수 있으므로, 더 높은 수준의 위기관리 능력이 요구된다. 따라서 사고 인지부터 재발 방지까지 포괄하는 대응 매뉴얼을 마련하고 정기적으로 그 실효성을 검증하는 위기관리 체계를 갖추는 것은 정보주체의 프라이버시를 보호하는 최소한의 안전장치라 할 수 있다. 특히

본 자율점검표는 위기관리 체계가 문서로만 존재하지 않고 실제 작동 가능한 상태로 유지되고 있는지 확인하고자 하였다.

라. 다양성 존중

「인공지능(AI) 윤리기준」의 세 번째 핵심요건인 다양성 존중은 사전적 배려 차원에서 AI 시스템이 산출하는 결과가 모든 사람에게 공정하게 적용될 수 있도록 여러 여건을 마련하는 데 초점을 둔다. 이는 헌법 제11조가 보장하는 평등권을 AI 기술 영역으로 확장한 것으로, 학습 데이터의 편향성 해소를 목표로 삼는 '기술적 공정성'과 '기술 혜택의 분배적 정의' 실현이라는 두 가지 측면으로 구체화된다. AI가 특정 인구 집단의 데이터에 과적합(Overfitting) 상태에 이르거나 특정 계층만 향유할 수 있는 기술이 되면, 헌법적 가치의 침해는 물론 건강 불평등의 심화라는 결과로 이어질 수 있다. 다만, 후자의 측면은 사회적 안녕과 공동체의 이익 증진에도 긴밀하게 연결되어 있으므로 아래 '공공성'에서 다루고, 여기에서는 기술적 공정성과 관련한 내용에 한정하여 문항을 구성하였다.

의료 AI의 윤리적 위험은 무엇보다도 불완전하거나 편향된 학습 데이터로부터 기인한다. 특정 성별, 인종, 연령대의 데이터가 누락되거나 과대 대표될 경우, AI는 해당 집단에 대해 부정확하거나 차별적인 진단을 내릴 수 있다. 따라서 데이터의 대표성을 확보하고 지속적으로 편향을 교정하는 절차적 장치가 필수적이다. 대표적인 사례는 다원적 거버넌스의 구성이 있다. 편향성은 기술적 시각만으로는 발견하기 어렵고, 다양한 배경을 가진 주체가 참여할 때 비로소 특정 집단에 불리하게 작용할 수 있는 잠재적 위험을 사전에 감지할 수 있기 때문이다. 이에 환자, 의료진, 기술진, 윤리 전문가 등 여러 이해관계자가 참여할 수 있는 거버넌스 체계를 구축하여 다양한 관점을 바탕으로 의료 AI 시스템을 지속적으로 개선하고 사회적 수용성을 확보할 수 있는 절차를 마련하고 있는지를 점검할 필요가 있다.

또한 의료 AI 기술의 윤리적 활용을 위해 사후적인 편향 발견 및 개선 절차를 갖출 필요가 있다. 개발 단계에서 편향을 완전히 제거하는 것은 불가능에 가까우므로, 운영 과정에서 드러나는 차별적 요소를 즉시 수정할 수 있는 환류 체계가 중요하다. 예를 들어, 의료 AI 시스템 운영 중 편향이나 차별이 발견되면, 개발자, 환자, 의료진 등 누구든지 이를 신속히 운영 주체에게 알릴 수 있는 채널을 마련하고, 내부 검토·평가·개선까지 이어지는 절차를 갖추는 것이다. 이를 통해 데이터 편향이 고착되어 구조적 차별로 연결되는 것을 방지할 수 있을 것이다.

마. 침해금지

「인공지능(AI) 윤리기준」의 네 번째에 자리하는 '침해금지'는 AI 시스템이 인간의 생명과 신체, 정신에 해를 입혀서는 안 된다는 원칙을 의미한다. 이는 "해를 끼치지 말라"는 의료의 제1원칙과 맥을 같이 한다. 침해금지의 내용을 헬스케어 분야에 맞게 구체화하면, 의도하지 않은 오작동이나 기술적 한계로 인해 환자에게 발생할 수 있는 직·간접적 위해를 사전에 방지하고, 사고 발생 시 피해를 최소화하는 적극적 안전 체계를 구축할 것으로 세분할 수 있다.

의료 AI는 진단 보조나 치료 계획 수립 등 환자의 건강에 직접적인 영향을 미치는 업무를 수행하므로, 사소한 오류도 환자의 생명·신체에 대한 치명적인 결과로 이어질 수 있다. 바로 이러한 점이 개발 단계에서부터 위해 가능성을 원천적으로 차단하기 위한 노력이 선행되어야 하는 이유이다. 특히 포괄적인 위험성 평가가 필요하다. 의료 AI는 물리적 상해뿐만 아니라 잘못된 진단으로 치료 적기의 실기, 정신적 스트레스 등 다양한 형태의 피해를 초래할 수 있다. 따라서 의료 AI 서비스를 제공하는 사업자는 이러한 위해를 사전에 검토하고, 그에 대응하는 예방 조치를 마련하여 위험관리의 사각지대를 제거하는 것이 권장된다.

또한, AI의 불완전성을 인정하고 이를 보완할 체계적 안전망을 구축하여야 한다. AI 모델은 확률에 기반하므로 완벽한 정확도를 보장할 수 없으며, 때로는 확신을 가지고 틀린 답을 내놓기도 한다. 이러한 기술적 한계가 환자의 안전을

위협하지 않도록, AI 사업자는 의료 AI 시스템의 오진이나 잘못된 치료 권고로 인한 위험을 최소화하기 위해 신뢰도 임계값을 설정하거나 불확실성을 표시하는 등 안전망을 체계적으로 구축할 필요가 있다. 이는 AI가 확신할 수 없는 상황에서는 판단을 유보하고 인간 의료진에게 결정권을 넘기도록 함으로써 (Human-in-the-loop), 기술적 오류가 실제 의료 사고로 이어지는 연결 고리를 끊는 핵심적인 조치이기 때문이다.

의료시스템에 대한 해킹이나 데이터 조작은 환자의 프라이버시 침해를 넘어, 투약 정보 변경이나 기기 오작동 유발 등 직접적인 신체적 위해로 이어질 수 있다. 즉, 헬스케어 분야에서 보안은 곧 안전으로 직결된다. 따라서 외부의 악의적인 공격이나 내부의 비정상적인 사용으로부터 시스템을 보호하는 것은 침해금지의 필수조건이다. 이에 의료 AI 시스템이 환자에게 해를 끼치거나 의료 서비스를 방해하지 않도록 접근 권한을 엄격하게 관리하고, 비정상적 사용 패턴을 자동으로 탐지하는 등 포괄적인 보안 조치를 구축할 필요가 있다. 본 자율점검표는 이를 점검할 수 있는 문항을 구성하여 의료 AI 시스템의 무결성을 확보하도록 하였다.

이처럼 여러 예방 조치를 취하더라도 예기치 못한 사고는 언제든지 발생할 수 있다. 따라서 침해금지 핵심요건의 완성은 결국 사고 발생 시 피해의 확산을 막는 대응 능력에 좌우된다. 의료 AI 시스템 활용 과정에서 예상하지 못한 피해가 발생한 경우, 피해의 확산을 방지하기 위해 시스템 사용을 중단하고 이를 의료진에게 즉시 알리는 등의 응급 대응 절차가 적절한 예시이다. 이는 사고 발생 시 즉각적인 '킬 스위치' 작동과 신속한 보고 체계를 구현하여, 2차 피해를 방지하고 의료 현장의 혼란을 최소화하는 데 이바지할 수 있다.

바. 공공성

「인공지능(AI) 윤리기준」의 다섯 번째 핵심요건인 '공공성'은 AI 시스템이 개인이나 특정 기업의 이익 창출 도구를 넘어, 사회적 안녕과 공동체의 이익 증진에 기여해야 함을 의미한다. 특히 헬스케어 분야에서의 AI는 헌법 제36조

제3항이 “모든 국민은 보건에 관하여 국가의 보호를 받는다”고 규정하고 있는 만큼 높은 공공성을 요구받는다. 이에 본 자율점검표는 AI 기술이 의료 자원의 불균형을 심화시키는 것이 아니라 해소하는 방향으로 활용되고, 기술 도입이 가져올 사회적 혼란이나 역기능을 교육과 제도를 통해 최소화하는 데 초점을 맞추어 구성되었다.

의료 AI 기술은 소수의 특권층이 아닌, 모든 국민의 건강권 증진에 기여하여야 한다. 이는 “인공지능은 최대한 많은 사람에게 최대한 많은 유익이 공평하게 분배될 수 있도록 개발되고 유통되고 사용되어야 한다”라는 「인공지능(AI) 윤리기준」의 원칙과 맞닿아 있다. 헬스케어 분야에서 의료 AI 기술의 다양성 존중은 사용자의 경제적·지리적·환경적 배경에 상관없이 공정한 의료 서비스를 제공받을 수 있는 기회균등을 보장하는 방향을 제시한다. 이에 본 자율점검표는 의료 AI 사업자가 경제적 지위, 지역적 위치, 사회적 신분과 관계없이 모든 환자가 동등하게 의료 AI 서비스에 접근할 수 있도록 국가 주도의 접근성 정책(저소득층 지원, 의료 사각지대 해소 등)에 적극적으로 참여하고 있는지를 점검 문항으로 구성하였다. 이는 AI 기술이 의료 공공성을 저해하지 않고, 오히려 의료 사각지대를 해소하는 도구로 기능하도록 유도하기 위함이다.

포용적 정책 참여 외에 의료 자원과 환경의 다양성을 고려한 기술적 포용성의 확보도 고려할 필요가 있다. 대형 병원의 고품질 데이터와 인프라에 맞게 설계된 AI는 자원이 부족한 소규모·지방 의료기관에서는 제대로 작동하지 않을 위험이 있다. 따라서 의료 AI 시스템이 대형 병원과 소규모 의료기관, 도시와 농촌 지역의 서로 다른 의료 환경과 자원 제약을 고려하여 각 기관의 상황에 적합한 진단 결과를 제공하기 위해 기술적 노력을 기울여야 한다. 이는 환경적 편향을 해소하여 어떤 진료 현장에서든 균질한 의료 서비스를 제공하게 함으로써 지역 간 의료 격차 해소에 기여한다는 의미가 있다.

의료 AI의 도입은 의료 수가 체계, 의료인력 수급, 보험 재정 등 사회 전반에 복합적인 영향을 미칠 것으로 예상된다. 따라서 기술의 도입 효과를 단순히 진단

정확도로만 평가해서는 안 되며, 사회적 비용과 편익도 고려할 필요가 있다. 이러한 사회·경제적인 영향은 내부적으로 검토되어야 할 뿐만 아니라, 외부 전문가의 의견을 청취하여 다양한 관점의 평가를 확보했을 때 비로소 온전히 평가할 수 있다. 이는 기술이 초래할 수 있는 부정적 사회 변화를 선제적으로 예측하고 대비하는 책임 있는 혁신의 자세를 요구하는 것이다.

공공성의 또 다른 축은 AI의 역기능을 최소화하기 위한 교육이다. 의료진이 AI에 지나치게 의존하여 독자적인 판단 능력을 상실하거나, 윤리적 책임을 기계에 전가하는 현상은 의료 서비스의 질적 하락이라는 사회적 손실로 이어질 가능성이 크다. 따라서 의료진이 AI 시스템을 활용하면서도 독립적인 임상 판단력과 환자 중심의 의료 서비스 능력을 유지할 수 있도록 AI의 작동 원리, 기술적 한계의 인식, 윤리적 책임 등을 포괄하는 체계적이고 지속적인 교육 프로그램과 실무 가이드라인을 제공하여 궁극적으로 환자에게 최선의 이익이 돌아가도록 공공적 책무를 이행할 필요가 있다.

사. 연대성

여섯 번째 핵심요건인 '연대성'은 AI가 인간과 인간, 인간과 사회, 현세대와 미래 세대 간의 관계를 단절시키는 것이 아니라, 더욱 긴밀하게 연결하고 협력하도록 유도해야 함을 의미한다. 특히 이 핵심요건의 가치는 AI의 혜택과 위험을 사회적, 국제적, 시간적 차원에서 함께 나누며, 개발부터 활용에 이르는 과정에서 모든 이해관계자가 상호 협력하고 책임을 분담하는 데 있다. 마찬가지로 헬스케어 분야에서도 연대성은 개별 의료기관이나 국가 단위를 넘어, 인류 보편의 건강 증진을 위해 협력하고 기술의 파급 효과에 대해 공동의 책임을 지는 자세를 요구한다.

의료 AI 기술은 의료진의 역할 변화, 환자-의사 관계의 재정립 등 기존 의료 생태계에 큰 변화를 가져올 것으로 예상된다. 이러한 변화가 갈등이나 소외로 이어지지 않으려면, 기술 도입 과정이 소수 전문가에 의해 일방적으로 주도되어서는

안 될 것이다. 이를 위하여 무엇보다도 어떤 AI 시스템이든 알고리즘의 설계부터 사후평가 및 감독에 이르는 주기가 하나의 연속선을 이루고 있으며, 안전성과 신뢰성은 이 과정에 관여한 모든 이해관계자의 연대를 통해 실현될 수 있다는 점을 상기하여야 한다. 의료 AI 시스템의 경우, 개발·운영 과정에서 다양한 배경과 전문성을 가진 이해관계자들이 동등한 입장에서 의사소통하고 상호작용할 수 있는 정기적 협의체, 공개 토론회 등 기회가 제공되는 것이 바람직하다. 이는 단순히 의견을 듣는 것에 그치지 않고, 서로 다른 입장을 가진 주체들이 숙의를 통해 '합의된 윤리적 기준'을 만들어가는 과정을 보장함으로써 사회적 신뢰와 연대를 강화할 수 있다.

연대의 대상은 세대와 국가를 넘나든다. 세대 간 연대의 경우, 최근 초거대 AI 모델의 학습과 운영 과정에서 발생하는 막대한 전력 소비와 탄소 배출이 기후 변화에 미치는 영향을 떠올릴 수 있다. 이는 기후 위기를 가속하여, 결과적으로 미래세대의 건강권을 침해하는 요인으로 작용할 것으로 예상된다. 이러한 우려를 고려하면, 헬스케어 윤리는 인간의 신체뿐만 아니라, 그 인류가 삶을 영위하는 환경의 건강성까지 다룰 때 진정한 의미를 획득할 수 있을 것이다. 미래세대에 대한 윤리적 책무를 이행하고자 한다면, 의료 AI 시스템의 에너지 효율성을 지속적으로 개선하고, 환경 부담이 최소화된 AI 기술을 활용함으로써 AI 기술이 지속 가능한 발전 목표에 부합하도록 유도할 필요가 있다. 또한 AI 기술의 윤리적 기준이 국가마다 다르다면, 규제가 느슨한 국가에서 위험한 실험이 감행되는 '윤리 덤핑'의 문제가 발생할 수 있다. 따라서 전 세계적으로 보편타당한 윤리 수준을 확보하기 위해 WHO 등 국제기구의 의료 AI 윤리 가이드라인을 참고하여 국제적 공통 기준을 준수하고, 인류 공동의 건강 문제해결에 기여하기 위한 국제사회의 노력에 동참하는 것이 바람직하다.

아. 데이터 관리

일곱 번째 핵심요건인 '데이터 관리'는 AI 시스템의 개발·운영에 활용되는 모든 데이터를 투명하고 책임 있게 관리하며, 데이터의 전체 생애주기에 걸쳐 품질을 보장하고 편향성을 최소화하여 공정하고 신뢰할 수 있는 AI 시스템 구현에 이바지할 것을 가리킨다. 이는 AI 시스템의 성능과 신뢰성을 결정짓는 가장 기초적인 토대이며, 특히 잘못된 데이터의 학습으로 인한 오진이 치명적인 결과로 직결되는 의료 AI의 경우 그 중요성은 배가된다.

민감정보에 해당하는 의료 데이터는 활용처가 다양할 뿐만 아니라, 환자의 취약성과 의존성이 극대화된 상황에서 수집되므로 목적 외 활용은 환자의 자율성과 존엄성에 대한 현저한 위협으로 작용한다. 의료 데이터의 오남용을 방지하고, 의료 AI 시스템 전체의 신뢰도를 유지하기 위해서는 우선 데이터의 변경이나 삭제 이력을 투명하게 관리하여야 한다. 「개인정보 보호법」 제29조(안전조치의무)와 제28조의4(가명정보에 대한 안전조치 의무 등)가 개인정보처리자에게 접속기록의 보관 및 점검, 접근 통제 등 기술적·관리적 보호조치를 취할 의무를 부과한 것도 같은 맥락에서 이해할 수 있다. 그러나 이러한 의무는 주로 외부 침입을 방지하거나, 사후에 침해 사실을 확인하기 위한 보안 차원에 머물러 있어 정당한 접근 권한을 가진 내부 연구자나 개발자가 데이터를 최초 승인된 목적 외 용도로 사용하는 경우를 실시간으로 통제할 수 없다. 특히 복잡한 경로를 거치는 의료 AI 개발 과정에서는 '누가 언제 왜 이 데이터를 가공했는지'에 대한 이력이 누락되기 쉽다. 따라서 법률이 요구하는 '기록의 보관'을 넘어, '기록의 상시적 감시와 절차적 투명성' 확보를 통해 윤리기준에 의한 법의 사각지대 보완이 필요하다. 예를 들어, 의료 AI 시스템에서 처리되는 모든 환자 데이터의 사용 내역을 정기적으로 점검하여, 목적 외 사용이나 무단 접근이 발생하였는지 모니터링하는 체계를 구축한다면, 윤리적 차원에서 시스템의 투명성을 확보할 수 있을 것이다.

여느 AI 시스템과 마찬가지로 의료 AI의 성능은 학습 데이터의 품질에 좌우된다. 데이터의 품질에 대한 고려는 개인정보처리자에게 데이터의 정확성, 완전성,

최신성을 유지할 의무를 부과하는 「개인정보 보호법」 제3조(개인정보 보호 원칙)에서 확인할 수 있다. 그러나 법률은 데이터 자체에 오탈자나 사실과 다른 내용이 없어야 한다는 기록상의 정확성에 초점을 두고 있을 뿐, AI 학습에 필요한 통계적 품질이나 학습 적합성까지 보장하지는 않는다. 데이터 자체는 사실과 부합하더라도 예컨대 불필요하거나 잘못된 정보가 많다거나, 특정 질병군에 관한 데이터가 부족하다면, 이를 학습한 AI는 잘못된 진단으로 환자의 생명과 신체에 대한 심각한 침해를 초래할 수 있다. 즉, 데이터의 품질이 확보될 때 비로소 의료 AI 시스템이 방대한 양의 데이터를 신속하게 처리하여 효율적인 의료 서비스의 제공을 가능하게 한다는 전망도 성립할 수 있다.

학습 데이터로서의 유효성을 확보하려면 데이터의 안전한 보관에 그치지 않고, 적극적인 품질 통제 절차까지 고려할 필요가 있다. 예를 들어, 데이터 품질이 일정 기준에 미치지 못한다면, 이를 학습 과정에서 과감히 배제하거나 별도의 전처리를 거치도록 품질 관리 기준을 수립하는 것이다. 또한 희귀질환이나 초기 단계의 질병처럼 데이터의 절대량이 부족하여 학습이 어려운 경우, 이를 방치하는 대신 의료기관과의 협력을 통해 양질의 데이터를 추가로 수집하거나, 데이터 증강(Augmentation) 기법을 활용하는 것도 한 방법일 수 있다. 이는 비록 법이 강제하지는 않지만, 안전하고 정확한 의료 서비스를 제공하기 위한 윤리적 차원의 노력이라 할 수 있다.

의료 서비스에서의 차별 금지는 「헌법」 제11조의 평등권 및 「보건의료기본법」[12]에 명시된 핵심 가치이다. 동법 제10조 제2항은 “모든 국민은 성별, 나이, 종교, 사회적 신분 또는 경제적 사정 등을 이유로 자신과 가족의 건강에 관한 권리를 침해받지 아니한다”라고 규정하고 있다. 사람에 의한 의도적인 차별 행위는 당연히 현행 법률에 따라 규율된다. 반면, 학습 데이터의 인구통계학적 편중으로 인해 AI가 특정 범주의 사람에 대해 낮은 정확도를 보이는 ‘성능의 불균형’은 개발자의 자발적인 해소 노력에 맡겨져 있다. 즉, AI의 윤리적 개발은 추상적인

12) 보건의료기본법[시행 2025.6.21.] [법률 제20589호, 2024.12.20., 일부개정]

평등의 원칙을 통계적·기술적 실증 절차로 구현하는 개발자의 노력에 달려있다. 예를 들어, AI 모델의 성능을 다양한 인구통계학적 변수별로 집단을 구별하여 세부 성능 평가를 수행하고, 특정 집단에 대한 편향이 식별되는 경우, 가중치를 다시 부여함으로써 데이터 간 균형을 맞추는 등 기술적 편향 완화 기법을 생각해 볼 수 있다. '다양성 존중'은 이러한 개입을 통해 단순한 선언에 머물지 않고, 구체적인 설계 과정에 반영될 수 있게 될 것이다.

자. 책임성

여덟 번 째 핵심요건인 책임성이란 AI 시스템의 개발부터 운영, 활용에 이르는 전 과정에서 발생할 수 있는 피해와 손실에 대해 각 주체의 역할과 의무를 명확히 구분하고, 예상가능한 위험은 사전에 최소화하고, 실제 피해 발생 시 신속하고 적절한 배상과 구제를 제공할 수 있는 실효적 책임 체계를 구축하는 것을 내용으로 한다. 특히 의료 AI 시스템은 환자의 생명과 직결되는 만큼, AI의 불확실성으로 인한 위험을 누가, 어떻게 분담할 것인지 명확하게 정의되어야 한다.

책임성의 핵심은 복잡한 AI 생태계에서 각 주체의 책임 범위를 명확히 하는 것이다. 현대 의료 AI 생태계는 개발사, 서비스 제공자, 의료기관, 의료진 등 다수의 주체가 복잡하게 얽혀 있다. 따라서 사고 발생 시 각 주체는 서로 책임을 전가하여 책임 소재가 불분명한 상태에 놓이는 문제가 발생할 위험이 크다. 물론 앞서 검토한 바와 같이 「의료법」 제27조 등은 의료행위의 주체를 의료인으로 한정하고 있어, 의료 AI 시스템의 활용으로 의료 사고가 발생했을 때 그 책임은 일차적으로 의료인이 부담하게 될 것이다. 그러나 AI 모델 자체의 기술적 결함이나 클라우드 서버의 오류 등 기술적 오류까지 의료인에게 전가하는 것이 온당한지에 대해서는 의문이 제기되고 있다. 이러한 책임 소재의 모호성은 의료 AI 기술의 도입 단계에서 개발사 및 서비스 제공자-의료기관 간 계약을 통해 책임을 명확히 분리하고 이를 문서화함으로써 일부 해소될 수 있다. 예를 들어, 개발자는 알고리즘의 오류를 방지할 책임을 부담하고, 서비스 제공자는 안정적

운영을 위해 서버를 유지하고 보안을 확보할 책임이 있으며, 의료진은 전문성에 기반하여 최종 판단 및 감독 책임을 부담하는 것이다. 이렇게 책임의 범위를 사전에 정의한다면, 각 주체는 자신의 책임 영역에서 최대한의 주의 의무를 다할 수 있고, 사고 발생 시 자기 책임인지 여부를 명확히 확인할 수 있을 것이다.

또한, 현행 법체계는 명확한 위법행위에 대한 처벌에는 유효하지만, 급변하는 기술 환경에서 발생하는 새로운 유형의 윤리적 딜레마를 해결하는 데는 구조적 한계를 보인다. 특히 AI 기술의 불확실성과 복잡성을 고려할 때, 단일 주체의 판단만으로 이러한 윤리적 딜레마를 해결할 것이라 기대하기 어렵다. 이 점에서 조직 내부적으로 독립적인 AI 윤리위원회를 설치하고, 의료진, 윤리 전문가, 내부 실무진, 시민사회 등 다양한 구성원이 참여하여 주요 의사결정에 대한 윤리적 검토를 수행할 수 있는 권한을 부여할 필요성이 대두된다. 이러한 위원회의 설치 및 운영은 의료 AI 개발사나 의료기관이 수익보다 환자의 안전을 우선시하고, 스스로 윤리 위험을 관리하고 설명가능한 책임을 이행하려는 능동적인 의지의 표명으로 볼 수 있다. 개발사, 서비스 제공자, 의료기관의 자발적 준수 노력은 향후 발생할 수 있는 법적 분쟁에서 기업이 충분한 주의의무를 기울였음을 입증하는 근거가 될 수 있으며, 무엇보다 의료 AI 생태계의 지속 가능한 발전을 위한 기초가 된다.

가장 중요한 책임의 이행은 손해를 입은 환자에 대하여 신속하게 배상하는 것이다. 현행「민법」제750조(불법행위의 내용)에 따른 불법행위 책임을 묻기 위해서는 피해자가 가해자의 고의나 과실을 입증해야 하는데, 블랙박스 특성상 전문가조차 알기 힘든 AI 알고리즘의 결함을 환자가 입증하는 것은 사실상 불가능에 가까울 것이다. 그렇다고 의료 AI 시스템으로 인해 손해를 입은 환자가 적절한 구제 없이 방치되는 것도 온당하지 않다. 이러한 법적 한계를 직시하고, 위험 분산의 원리에 입각한 실질적 해결책의 모색이 필요하다. 예를 들어, 과실 유무를 따지는 법적 다툼에 앞서 환자의 피해를 신속하게 복구할 수 있도록 개발사, 서비스 제공자, 의료기관이 충분한 한도의 배상책임보험에 가입하는 것을

생각해 볼 수 있다. 이는 AI 기술의 혜택을 누리는 AI 사업자와 의료기관이 법적 의무를 이행하는 것을 넘어, 사회적 책임을 다한다는 점에서 윤리적 의미가 있다.

차. 안전성

안전성은 AI 시스템의 개발, 배포, 운영, 폐기에 이르는 전 과정에서 개인과 사회 전반에 미칠 수 있는 잠재적 위험을 사전에 식별하고, 명백한 오류나 침해가 발생했을 때 사용자가 시스템을 제어할 수 있도록 보장하는 안전 보장 체계를 요구한다. 헬스케어 분야에서 AI의 안전성은 환자의 생명과 직결되므로 타협이 불가하다. 이러한 사고방식은 현행 「의료기기법」[13]이 의료기기가 갖추어야 할 물리적·기술적 안전 성능을 엄격히 규정하고, 이를 충족해야만 시장 진입을 허용하는 것에서도 찾아볼 수 있다. 그런데 이러한 허가 절차는 제품 출시 전 특정 시점의 성능을 검증하는 데 초점이 맞춰져 있기 때문에, 데이터를 통해 끊임없이 학습하고 변화하는 AI의 특성이나 실제 의료 현장에서 발생하는 변수까지 법이 모두 통제할 수는 없다. 법적 허가 기준의 충족 그 이상의 실질적이고 동적인 안전성을 확보하기 위한 노력은 윤리적 차원에서 다뤄야 할 내용이다.

실제 '임상적 맥락'에서 발생할 수 있는 위험에 대응하고자 한다면, 임상 환경에서 발생할 수 있는 구체적인 위험 시나리오를 사전에 분석하는 절차가 선행하여야 한다. 특히 AI의 편의성이 높아질수록 의료진이 AI의 판단을 성찰 없이 수용하는 과의존 현상이 발생할 수 있다. 이러한 인적 요소까지 위험 요인으로 포함하여, 임상 전문가와 협업하여 구체적인 대응 매뉴얼을 마련했는지 점검할 필요가 있다. 이는 「의료기기법」이 제시한 법적 기준인 '기계적 안전'을 넘어, 기계가 인간과 상호작용하는 임상 환경에서도 안전한지를 확인하기 위함이다.

또한, 시간이 지남에 따라 입력 데이터의 분포가 변화하여 AI 모델의 성능이 서서히 저하될 수 있다는 점을 고려하여 생애주기 모니터링 체계를 갖추는 것도 안전성 확보에 도움이 될 수 있다. 물론 「의료기기법」 제31조의5 제1항도 시판

13) 의료기기법 [시행 2025.8.1.] [법률 제20753호, 2025.1.31., 일부개정]

후 안전성 정보 관리 의무를 두고 있지만, 이는 보고된 부작용을 수집하는 수동적 관리에 가깝다. 동적 안전성의 확보를 위해서는 법적 요구사항보다 선제적으로 상시 모니터링 및 즉각 대응 체계를 구축하는 것이 바람직하다. 예를 들어, 의료 AI 시스템이 배포된 후에도 지속적으로 의료 AI의 성능을 추적하여, 비정상적인 작동이나 성능 저하 징후가 포착되면 즉시 알리고 대응하는 체계를 개발사나 서비스 제공자가 마련하는 것을 고려할 수 있다.

특히 AI가 아무리 고도로 발전하더라도, 최종적인 통제권은 언제나 인간에게 주어져야 한다. 「의료기기법」은 기기의 오작동 방지를 포괄적으로 요구하지만, AI가 생성한 환각(Hallucination)이나 갑작스러운 알고리즘의 오류 상황에서 의료진이 직관적으로 개입할 수 있는 인터페이스까지 규정하지는 않는다. 이 점에서 최후의 안전장치로서 '인간에 의한 제어권' 확보의 필요성을 확인할 수 있다. 단순히 제어권을 부여하는 것뿐만 아니라, 이 기능은 직관적으로 구현되어야 한다. 즉, 최종 사용자인 의료진이 복잡한 절차 없이 즉시 시스템을 중단시키거나 수동 모드로 전환할 수 있어야 한다. 이러한 최종적인 통제권의 확보는 기술 만능주의에 대한 경계일 뿐만 아니라, 어떤 경우에도 환자의 안전을 최우선으로 보호하겠다는 의지의 표명이기도 하다.

카. 투명성

마지막 핵심요건인 '투명성'은 AI 시스템의 작동 방식, 사용 데이터, 잠재적 위험과 한계를 AI 사용자가 명확히 이해할 수 있도록 정보를 공개하고 설명하는 원칙을 일컫는다. 이는 AI 활용 사실과 그에 따른 유의 사항을 사전에 고지하여 AI 시스템의 일관된 작동 사실을 사용자가 확인할 수 있도록 하는 데 목적이 있다. 특히 사람의 생명을 다루는 헬스케어 분야에서 AI 의사결정에 대한 이해는 시스템과 그 사용자의 신뢰를 제고하는 핵심 기반이 된다.

투명성은 최근 법제화를 통해 그 중요성이 강조되고 있다. 개정된 「개인정보 보호법」 제37조의2(자동화된 결정에 대한 정보주체의 권리 등)는 AI에 의한

결정이 정보주체에게 중대한 영향을 미칠 때, 해당 결정에 대한 설명을 요구할 수 있는 권리를 명시하고 있다. 그러나 이 설명 요구권은 '블랙박스'라 불리는 AI 알고리즘의 복잡한 작동 원리를 비전문가인 환자나 의료진에게 '어떻게 어느 정도로' 설명해야 하는지에 대한 구체적 방법론까지 제시하지는 못한다. 기술적 난해함을 이유로 형식적인 정보만 제공하거나, 역으로 과도한 기술 정보만 나열한다면, 이를 법적 의무의 이행으로 평가할 수는 있으나 환자와 의료진의 실질적 이해를 돕는 투명성이라 평가하기는 어렵다.

설명 요구권이 의료 현장에서 실효성을 가지려면, 이를 뒷받침하는 행정 절차가 필요하다. 특히 환자가 설명을 요청하는 경우, 이에 적절하게 대응할 수 있는 표준화된 절차를 마련하여야 할 것이다. 더 나아가 각 단계에서 필요하다고 판단되는 처리 시간, 담당자, 설명 방법 등을 명확히 정의한다면, 원활한 절차 진행을 보장하고, 설명의 일관성을 확보하여 사용자의 예측가능성을 높이는 데 이바지할 수 있을 것이다.

내용 측면에서는 사용자 눈높이에 맞춘 소통 전략이 필요하다. 모든 정보를 기계적으로 나열하는 것을 투명성으로 평가할 수 없다. 오히려 사용자의 기술적 이해 수준에 따라 설명의 깊이와 방법을 조정하는 차등적 접근이 투명성을 준수하는 효과적인 방안이 될 것이다. 의료진에게는 의료 AI 시스템이 산출한 결과의 임상적 타당성을, 환자에게는 그 의미와 영향을 중심으로 정보를 구성하여 전달해야 할 것이다. 그럼에도 해소되지 않는 복잡한 결과나 예외 사례에 대해서는 전문의나 AI 전문가에게 추가 해설을 요청할 수 있도록 인적 지원 체계를 갖추는 것이 바람직하다. 이는 기술적 설명의 한계를 전문가와의 소통을 통해 보완하는 예시이다.

4. 주요 개발 과정

「헬스케어 분야 인공지능 윤리기준 자율점검표」의 초안을 마련하기 위해 정보통신정책연구원의 연구진은 관련 문헌 연구, 국내외 규제 동향 분석, 전문가

의견 수렴 등을 통해 의료 분야의 AI 개발 및 활용과 관련된 주요 쟁점을 도출하였다. 이후, 「인공지능 윤리기준」 및 「2025 인공지능 윤리기준 실천을 위한 자율점검표(안)」과의 연계성을 고려하여, 의료 AI 서비스의 개발과 활용에서 특히 강조되어야 하는 점검 문항을 선별·가공하였다. 또한 헬스케어 분야에서 제기되는 윤리적 쟁점에 대응하기 위해 새로운 점검 문항을 추가로 신설하며 초안을 구성하였다. 특히, 업계 종사자와의 협의를 통해 의료 AI 서비스 개발·운영과정에서 활용되는 기술적 특수성을 반영하고자 노력하였다.

〈표 3-1〉 자율점검표 초안 점검문항 예시

구분		주요 내용
의료	인권보장 E01.02	의료 AI 시스템은 의료진이 AI의 진단 결과를 검토하고 전문성과 자율성에 기반하여 최종 판단을 내릴 수 있도록 필요한 기능을 제공하고 있는가?
	프라이버시 보호 E02.02	의료 AI 시스템에서 발생할 수 있는 개인정보 침해나 프라이버시 위반에 대비하여 사고 인지부터 재발 방지를 포괄하는 사고 대응 매뉴얼을 마련하고, 정기적인 점검을 통해 실효성을 검증하는 위기관리 체계를 확립하고자 노력하고 있는가?
	다양성 존중 E03.03	의료 AI 시스템 운영 중 편향이나 차별이 발견되면 개발자, 환자, 의료진 등 누구든 이를 신속히 운영 조직에 알리고, 내부 검토·평가·개선까지 이어지는 절차를 갖추고 있는가?
	데이터 관리 E07.01	의료 AI 시스템에서 처리되는 모든 환자 데이터의 사용 내역을 목적별로 기록하고 정기적으로 점검하여, 목적 외 사용이나 무단 접근이 발생하였는지 지속적으로 모니터링하는 체계를 구축하고 있는가?
	안전성 E09.04	의료 AI 시스템에서 명백한 오류나 비정상적 동작이 감지되면 최종 사용자인 의료진이 즉시 시스템을 중단하거나 수동으로 조작할 수 있도록 직관적인 제어 기능이 구현되어 있는가?
	투명성 E10.04	의료 AI 시스템의 복잡한 결과나 예외적인 사례에 대해서는 해당 분야 전문의나 AI 전문가의 추가적인 해석과 검토를 제공할 수 있는 지원 체계를 구축하고 있는가?

자료: 연구진 작성

가. 의견수렴

이러한 과정을 통해 도출된 초안에 대하여 2025년 11월 21일부터 12월 5일까지 학계·산업계·법조계 등 외부 전문가 15인을 대상으로 본 점검표에 대한 서면 자문을 진행하였다. 그 주요 의견을 예시적으로 소개하자면 다음의 표와 같다.

〈표 3-2〉 자율점검표 초안에 대한 서면 자문 예시

구분	주요 내용
인권 보장	의료법상 의료행위의 주체는 의료인이 되어야 하고, AI는 보조로서의 역할에 한정되어야 하는 만큼 그러한 사실이 명시되는 방향의 수정 필요
다양성 존중	정보 단계에서부터 편향 및 차별이 학습되지 않도록 정보의 다양성 확보에 대한 내용이 추가될 필요
데이터 관리	의료 정보의 합리적인 기밀성 유지에 초점을 둘 필요
안전성	제어 기능이 시스템 오류 발생시에도 정상적으로 작동하는지에 대한 검증이 이루어졌는가에 대한 내용 추가될 필요

자료: 연구진 작성

본 자율점검표의 실효성을 확보하고 산업계의 관심도를 높이기 위해서는 본 점검표가 기존 규제 절차와의 정합성을 고려하여 효율성을 높이는 방향으로 고도화되고, 향후 인공지능 윤리 문제 관련 사례를 바탕으로 점검표를 지속 개선해나갈 필요가 있다는 의견이 있었다. 또한, 해당 자율점검표를 활용하는 기업이 AI 윤리를 효과적으로 확보할 수 있도록 의료 분야에 특화된 정량적 평가 체계 및 가이드라인을 선제적으로 제공해야 한다는 의견, 장기적으로는 이해관계자 및 제품 유형별로 특화하여 자율점검표를 세분화할 필요가 있다는 의견이 있었다.

다. 최종안

전문가 의견수렴을 거쳐 총 34개 점검문항으로 구성된 「헬스케어 분야 인공지능 윤리기준 자율점검표」의 최종안을 도출하였다. 검토 과정에서 문항 간 상충 가능성을 해소하고, 중복 요소를 삭제·통합하는 최적화 과정을 거쳤다. 특히, 윤리기준의 엄밀한 관철과 실질적 수용성 사이의 균형을 확보하는 데 주력하였다. 지나치게 느슨한 기준은 점검표의 의미를 퇴색시키고, 과도하게 엄격한 기준은 기업의 자발적 점검 의지를 위축시킬 우려가 있기 때문이다. 이러한 상충 관계 속에서 최적의 접점을 모색하고자 하였고, 그 결과 초안의 39개 문항을 세부 조정하여 최종 34개 문항으로 확정하였다.

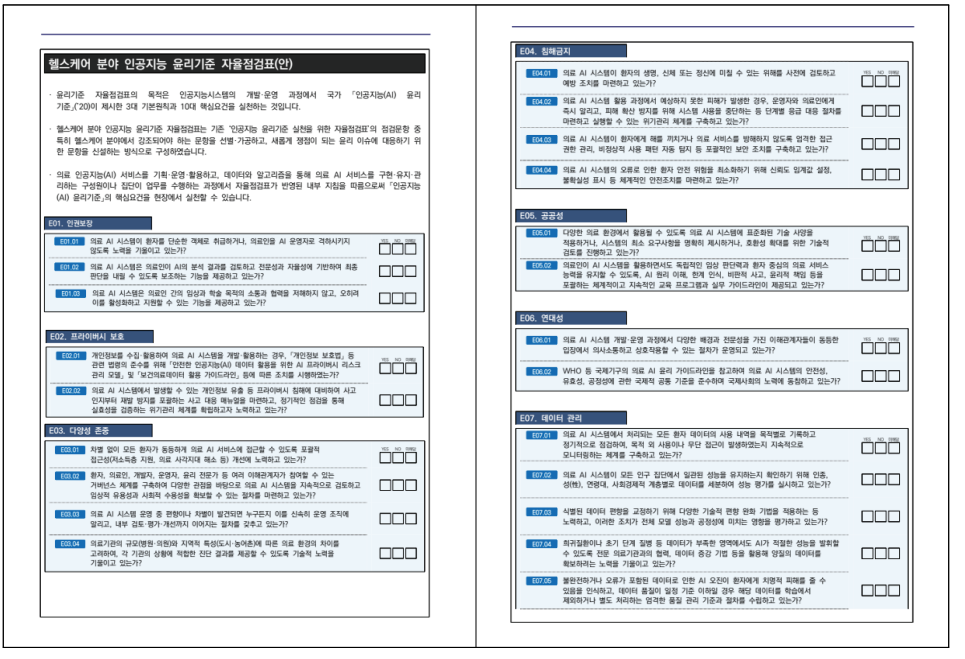
〈표 3-3〉 「헬스케어 분야 인공지능 윤리기준 자율점검표」 문항 수 변동추이

헬스케어 분야 인공지능 윤리기준 자율점검표 초안											
핵심요건	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	합계
문항 수	4	2	4	4	4	3	5	4	4	5	39
헬스케어 분야 인공지능 윤리기준 자율점검표 최종안											
핵심요건	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	합계
문항 수	3	2	4	4	2	2	5	3	4	5	34

자료: 연구진 작성

본 점검표는 2026년 초 발표될 '2025 인공지능 윤리기준 자율점검표(안)'의 개정판인 '2026 인공지능 윤리기준 자율점검표(안)'에 포함되어 대중에 공개될 예정이다.

[그림 3-2] 헬스케어 분야 인공지능 윤리기준 자율점검표 예시



자료: 연구진 작성

제 2 절 자율점검표 현장 적용

1. 개요

과학기술정보통신부 「인공지능(AI) 윤리기준」의 사회적 확산을 목적으로 2022년 2월 정보통신정책연구원은 「2022 인공지능 윤리기준 실천을 위한 자율점검표」 초안을 개발 및 공개하였으나, 각 산업 영역 제품 및 서비스 특성을 고려하여 자율점검표를 맞춤형으로 활용하는 데 한계가 있다는 지적이 꾸준히 제기되었다. 이러한 한계를 극복하기 위해 정보통신정책연구원은 2022년 3월부터 기업과 협업하여 AI 시스템 기업현장에서 자율점검표를 실제로 적용해보고, 해당 기업의

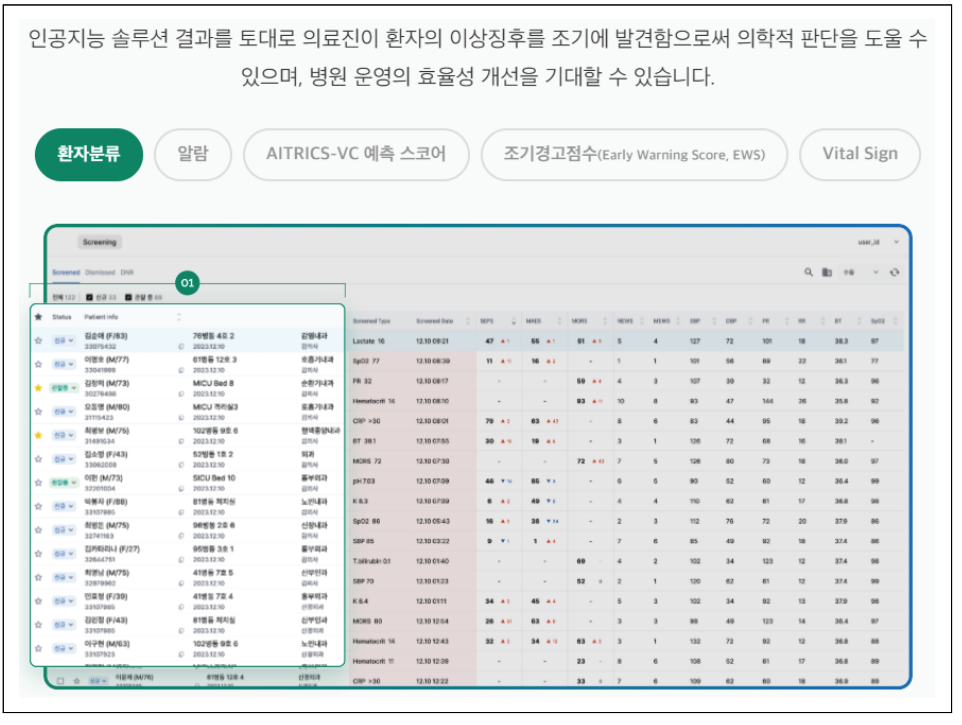
사업 특성에 맞게 점검 문항을 가공하여 기업 내부적으로 활용할 수 있는 윤리점검표를 개발하고자 하였다. 구체적으로 2022년에는 챗봇(스캐터랩, Scatter Lab)·작문(뤼튼테크놀로지스, Wrtn Technologies)·영상 관제(마크애니, MarkAny), 2023년에는 채용(제네시스랩, Genesis Lab), 2024년에는 영상 합성(이스트소프트, ESTsoft) 분야의 기업과의 협업을 통해 총 5건의 기업 윤리점검표를 공동으로 개발하였다(문정욱 외, 2024). 이어서 2025년에는 '에이아이트릭스(AITRICS)'와 협업하여 의료 AI 분야 자율점검표를 적용하고, 에이아이트릭스가 AI 디지털 의료기기를 개발하는 과정에서 「인공지능(AI) 윤리기준」을 준수 여부를 스스로 점검할 수 있도록 지원하는 「에이아이트릭스 인공지능 윤리점검표」를 개발하였다.

2. 현장 시범 적용 경과

가. 에이아이트릭스 소개

2016년 설립된 에이아이트릭스는 AI 기술을 기반으로 생체신호를 분석하여 환자의 상태악화를 조기 예측하는 임상의사결정지원시스템인 AITRICS-VC 서비스를 개발 및 제공하고 있는 기업이다. AITRICS-VC 서비스는 병원 내 축적되는 19종의 실시간 EMR 정보를 AI가 분석하여 AI 예측 점수를 통해 의료진의 신속한 의사결정을 지원하는 예측 결과를 제공한다. 일반병동 중증 이벤트, 일반병동 패혈증, 중환자실 사망에 대한 세 가지 AI 모델의 유효성을 평가하기 위해 확증임상시험을 진행하여 높은 수준의 예측 정확도와 성능을 입증하기도 하였다. AITRICS-VC가 처음 적용된 전주예수병원에서 실사용 후향 연구를 진행한 결과, 도입 전에 비해 코드블루가 약 25% 감소하여 환자 상태 악화를 실제로 방지함을 검증하기도 했다. 현재는 해당 서비스를 확장하여 응급실 환자의 상태 악화 발생을 예측하는 AITRICS-ER 서비스와, 의료 정보 AI 핵심 요약 및 질환 예측을 제공하는 버추얼닥터(V.Doc) 서비스를 개발 중에 있다.

[그림 3-3] 에이아이트릭스 AITRICS-VC 서비스



자료: https://aitrics.com/kr/sub/product/vc.php

에이아이트릭스는 AI 기술과 서비스의 신뢰성 및 품질에 대한 높은 이해와 가이드라인 준수를 위해 노력하고 있다. 이에 대해 AITRICS-VC 제품 시장성을 인정받아 2023년 과기정통부가 주관한 '제1회 인공지능 신뢰성·품질 대상'에서 과기부 장관상(최우수상)을 수상하였다. 또한 정보보안과 관련된 보건신기술(NET) 인증을 획득하고, 미국, 베트남, 홍콩 등 국제 인증 및 허가를 획득하는 등 제품 및 기술에 대한 안전성 제고를 위해 노력하고 있다. 2025년 2월 시행된 디지털 의료기기법과 관련하여 공표된 우수관리체계 인증을 위해서도 적극 계획하고 있다.

나. 「에이아이트릭스 인공지능 윤리점검표」 개발

2025년 10월 초 정보통신정책연구원은 에이아이트릭스와 「인공지능 윤리기준」 현장 적용에 관한 사전 협의를 시작으로 하여, 이후 본격적으로 「에이아이트릭스 인공지능 윤리점검표」의 개발에 공동 착수하였다. 2025년 10월부터 12월까지 약 3개월간 총 8차례의 회의를 개최하였으며, 회의에서는 온라인과 오프라인 방식을 적절히 병행하며 진행 상황을 공유하고 작업한 내용에 관한 논의를 진행하였다.

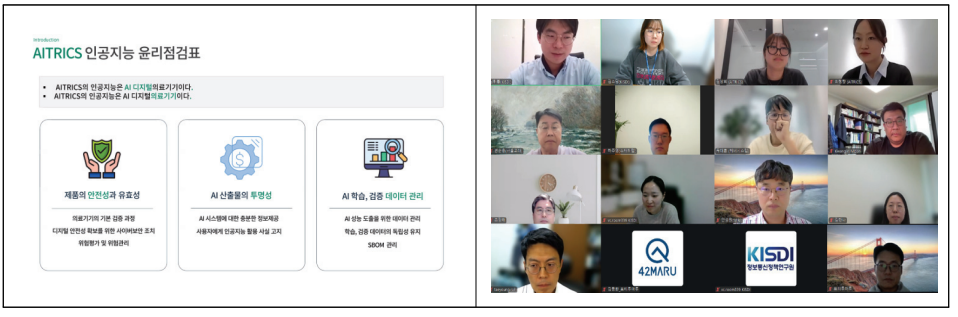
〈표 3-4〉 「에이아이트릭스 인공지능 윤리점검표」 개발 경과

차수	일자	방식	주요 내용
1	'25년 10월 1일	오프라인	- 킥오프(Kick-off) 회의
2	10월 23일	오프라인	- 기초작업 착수
3	10월 29일	온라인	- 자율점검표 및 윤리점검표 비교·검토
4	11월 6일	오프라인	- 점검표 초안 마련
5	11월 14일	온라인	- 점검표 초안 온라인 설명회
6	11월 20일	온라인	- 전문가 의견수렴(1차)
7	11월 26일	온라인	- 전문가 의견수렴(2차)
8	12월 5일	오프라인	- 최종 점검

자료: 연구진 작성

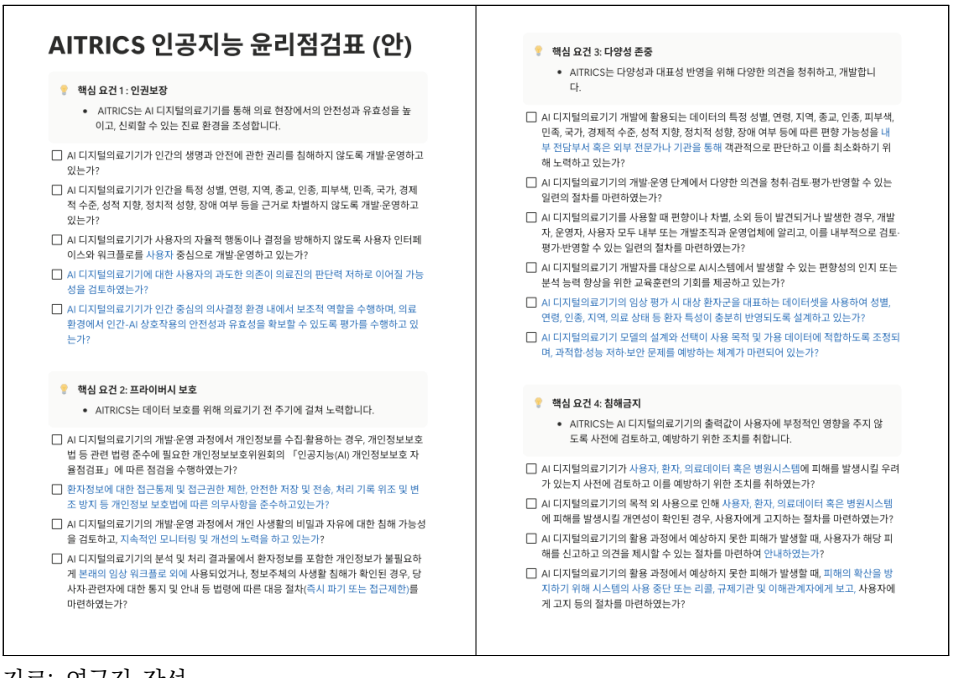
에이아이트릭스는 실제 적용에 준하는 방식으로 자율점검표를 검토하고, 해당 기업 내부에서 직접 활용할 수 있는 개별·구체화된 점검문항을 도출·구성하였다. 기존 자율점검표(안)의 점검문항 중 의료 AI 분야에 적용할 수 있는 문항을 선별·가공하였으며, 에이아이트릭스의 AI 신뢰성·윤리 관련 활동 및 결과물 등을 활용하여 의료 AI 분야에 적용이 가능한 문항을 추가하였다. 이러한 방식을 통해 도출된 초안에 대한 온라인 공개 설명회를 개최하였고, 이후 각계 전문가로부터 의견을 수렴하여 최종안을 마련하였다. 「에이아이트릭스 인공지능 윤리점검표」 최종안은 2026년 2월 공개될 예정이다.

[그림 3-4] 「에이아이트릭스 인공지능 윤리점검표」 초안 공개 설명회



자료: 연구진 작성

[그림 3-5] 「AITRICS 인공지능 윤리점검표(안)」



자료: 연구진 작성

제3절 소 결

의료 AI 기술은 의료영상, 환자의 임상기록, 생체신호, 유전체 정보 등 다양한 의료 데이터를 기반으로 진단, 예측, 치료 의사결정을 지원하며 보건의료 전반의 핵심기술로 빠르게 자리 잡고 있다. 특히 영상의학 분야에서는 기존 장비의 판독 효율을 높이는 솔루션이 임상에 활발히 도입되고 있으며, 정밀 의료영역에서도 생체신호 분석을 통해 환자 상태의 악화를 예측하는 임상 의사결정 지원 시스템의 활용이 확대되는 추세이다. 그러나 의료 AI 기술의 진보가 서비스의 효율성을 높이는 동시에 환자의 안전과 권리에 대한 새로운 위험을 동반한다는 점을 간과해서는 안 된다. 가장 큰 우려는 무엇보다 임상 의사결정에 깊이 관여하는 알고리즘의 오작동이나 예측 오류로 인해 환자의 생명 및 신체에 직접적인 위해를 초래할 수 있다는 점에서 기인한다. 이러한 안전성 문제는 AI 학습의 원천인 의료 데이터의 관리 문제와도 긴밀하게 맞물려 있다. 방대한 건강 정보의 집적과 활용 과정에서 보안 취약점이 노출될 경우, 대규모 환자 정보 유출이라는 돌이킬 수 없는 프라이버시 침해 사고로 이어질 수 있기 때문이다. 나아가 AI의 의사결정과정을 명확히 파악하기 어렵다는 점도 진단 오류나 사고 발생 시 그 책임이 개발사, 서비스 제공자, 의료진 중 누구에게 귀속되는지 명확히 규명하기 어려운 법적·윤리적 난제를 야기한다. 고비용의 첨단 AI 인프라가 대형병원 중심으로 편중되면서, 자원이 부족한 중소병원이나 지역 의료기관은 소외되는 의료 접근성의 격차 심화도 간과할 수 없는 문제이다.

이러한 위험에 대응하기 위해 우리나라는 「인공지능기본법」과 「디지털의료제품법」 등을 통해 법적 규제를 강화하고, ‘보건의료데이터 활용 가이드라인’ 및 ‘AI 개인정보보호 자율점검표’ 등으로 법적 의무 이행을 지원하고 있다. 그러나 법령과 지침만으로는 급변하는 기술 양상과 복잡한 윤리적 사각지대를 모두 포괄하기 어렵기에, 본 장에서는 AI 윤리기준의 10대 핵심요건을 헬스케어 분야의

특수성에 맞춰 구체화하고 이를 내재화할 수 있는 실천적 방안을 자율점검표의 형식을 빌려 제시하였다. 우선 인간 중심 가치 실현을 위해 '인권보장'과 '프라이버시 보호'가 선행되어야 한다. 의료진이 AI에 종속되지 않고 최종 의사 결정권을 행사할 수 있도록 절차적 장치를 마련하고, 의료정보와 관련한 사고 발생 시 실질적으로 작동하는 위기관리 체계를 통해 환자의 정보를 보호하는 것이 바람직하다. 사회적 차원에서는 '다양성 존중', '공공성', '연대성'을 통해 데이터 편향을 지속적으로 감시하고, 소규모 의료기관도 활용 가능한 기술적 포용성을 확보하며, 미래세대를 위한 환경적 책무를 다하는 협력 체계의 구축을 강조하였다. 기술적 신뢰성은 '침해금지', '안전성', '데이터 관리'를 통해 확보될 수 있다. 이와 관련하여 오작동 시 시스템을 즉시 중단하는 안전장치와 의료진의 직관적인 제어권을 구현하고, 데이터 변경 이력의 투명한 관리 및 저품질 데이터 배제 등 품질 통제를 제시하였다. 마지막으로 '책임성'과 '투명성'을 제고하기 위해 개발사와 의료기관 간 책임 범위를 명확히 문서화하고, 사용자 눈높이에 맞춘 설명 등을 제시하여 의료 AI 생태계의 신뢰 구축에 기여하고자 하였다.

본 자율점검표는 학계·산업계·법조계·시민사회 전문가의 의견 수렴을 거쳐 마련되었다. 전문가들은 「의료법」상 의료인의 주체성 명시, 데이터 편향 방지를 위한 다양성 확보, 시스템 오류 시 제어 기능 검증 등의 의견을 제시하였으며, 이를 반영하여 점검 문항을 최종 조정하였다. 또한, 이론적 개발에 그치지 않고, 자율점검표의 현장 적용성을 높이기 위해 의료 AI 개발 및 제공자인 '에이아이트릭스(AITRICS)'와 협업을 진행하였다. 에이아이트릭스는 환자의 생체신호를 분석해 상태 악화를 조기에 예측하는 서비스를 개발한 기업으로, 높은 예측 정확도를 인정받고 있다. 정보통신정책연구원과 에이아이트릭스는 약 3개월 간 공동 작업을 통해, 범용 자율점검표를 해당 기업의 특성에 맞게 가공한 '에이아이트릭스 인공지능 윤리점검표'를 개발하였다. 이 과정에서 헬스케어 분야 의료 AI에 특화된 문항을 선별하고, 기업의 신뢰성 및 윤리 활동을 반영한 신규 문항을 추가하였다. 개발된 기업 맞춤형 점검표는 온라인 설명회와 전문가 의견 수렴을 거쳤으며, 이는 2026년 초 공개될 예정이다.

제 4 장 AI 윤리 소통채널 구축·운영

제 1 절 AI 윤리 소통채널

정보통신정책연구원은 AI 윤리가 사회 전반에 정착되고 지속적으로 발전할 수 있도록 2023년 11월부터 'AI 윤리 소통채널'을 온라인 플랫폼 형태로 운영하고 있다. 이 소통채널은 기업, 학계, 시민사회 등 다양한 이해관계자가 AI 윤리 관련 이슈와 국내·외 동향을 폭넓게 이해할 수 있도록 지원하며, 윤리적 실천을 위한 협력 기반과 지식정보 공유의 허브로 기능한다. 또한, AI 윤리 정책자료를 체계적으로 아카이브하고, 새롭게 제기되는 윤리 이슈를 상시적으로 논의하는 정책 논의의 기반 인프라로서 역할을 수행하고 있다.

특히 온라인 플랫폼의 개방성과 접근성을 바탕으로 AI 윤리에 관한 사회적 합의 형성과 디지털 신뢰 자본 확충을 핵심 목표로 한다. 시·공간 제약 없이 참여 가능한 공론장을 제공함으로써 국민의 정책 과정 참여를 촉진하고, AI 활용 과정에서 요구되는 윤리기준에 대한 능동적 판단과 숙의 경험을 확대한다. 더불어 AI 기술 확산 과정에서 심화되는 AI 격차와 AI 리터러시 이슈에 대응하여 모든 국민이 AI의 윤리적 쟁점을 이해하고 의견을 제시할 수 있도록 지원함으로써 포용성과 접근성을 강화하고, 정부 정책 수립 과정의 투명성과 다양한 이해관계자의 참여를 높여 지속 가능한 사회적 합의 형성에 기여하고자 한다.

2025년 소통채널은 이러한 목표를 달성하기 위해 국내·외 AI 윤리 정책자료를 지속적으로 확충하고 아카이브화하였으며, 이용자 친화적 인터페이스 정비와 다양한 온라인 이벤트 수행을 통해 양방향 소통 기반을 강화하였다. 앞으로도 AI·디지털 전환기에 필요한 사회 대응 역량을 높이는 핵심 플랫폼으로서, 국민 참여 확대와 정책 생태계 지원 기능을 지속적으로 고도화해 나갈 예정이다.

1. AI 윤리 소통채널 기본 구성[14)]

'AI 윤리 소통채널'은 △사업 소개, △정책 홍보 및 성과 제공, △국제사회 소통, △윤리정책 수집, △의견 청취 등의 주요 기능을 담당한다. 이를 효과적으로 수행하기 위해 '소개', 'AI 윤리실천', 'AI 윤리교육', '정책저장소', 'AI 윤리정책 포럼', '참여소통방' 등을 포함한 6개의 상위 메뉴와 총 19개의 하위 게시판으로 체계적으로 구성되어 있다.

[그림 4-1] 'AI 윤리 소통채널' 홈페이지

자료: 'AI 윤리 소통채널(https://ai.kisdi.re.kr)', 2025.12.1. 검색

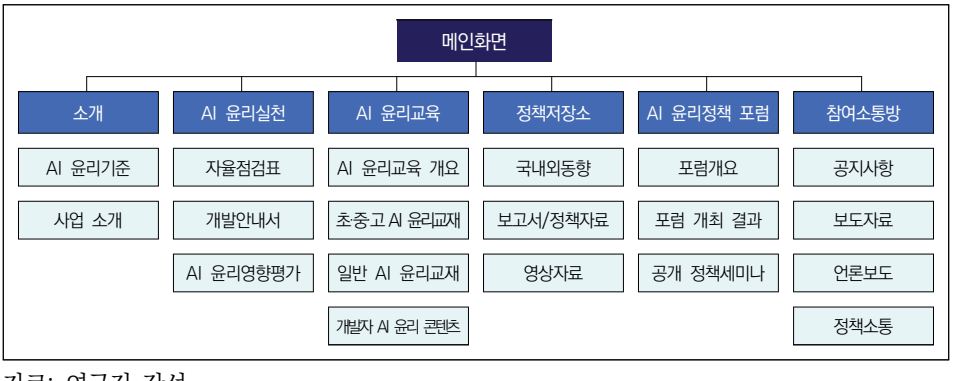
14) 해당 내용은 문정욱 외(2024) "AI 윤리·신뢰성 확보를 위한 실천 방안 및 정책연구"를 바탕으로 재구성

〈표 4-1〉 AI 윤리 소통채널 주요 기능

기능	내용
사업소개	- 사업 개요, 중점 추진분야, 경과 소개
정책 홍보 및 성과 제공	- 국가 정책(윤리기준, 전략 등), 실천수단(윤리영향평가, 자율점검표, 개발안내서 등), 거버넌스 (AI 윤리정책 포럼), 윤리교육 (교재, 콘텐츠) 등 사업 추진 성과물 공개 및 홍보
국제사회 소통	- AI 윤리 관련 주요 추진사항 등을 담은 영문 웹페이지 제공
윤리정책 수집	- 해외 주요국, 국제기구, 민간 영역의 윤리원칙 현황 정보 축적
의견청취	- 특정 안건에 대한 의견수렴, 상시 의견수렴 등 진행

자료: 연구진 작성

[그림 4-2] 홈페이지 메뉴

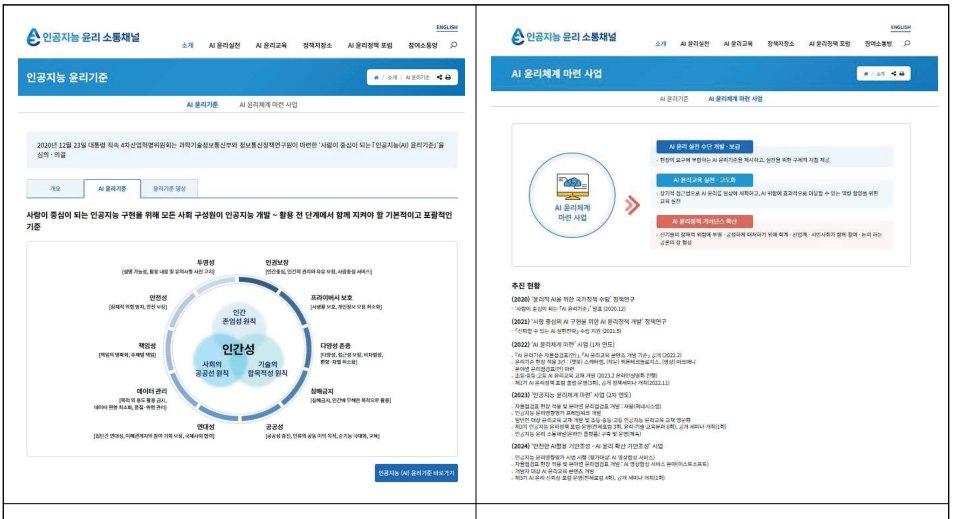


자료: 연구진 작성

가. 메뉴 ①: ‘소개’

「인공지능(AI) 윤리기준」의 배경 및 필요성, 지향점, 핵심 내용을 설명하고, 2020년부터 2024년까지의 AI 윤리 관련 사업 추진 경과, 중점 분야 등 사업 전반 현황에 관한 정보를 포괄적으로 제공한다.

[그림 4-3] ‘소개’ 상세 페이지



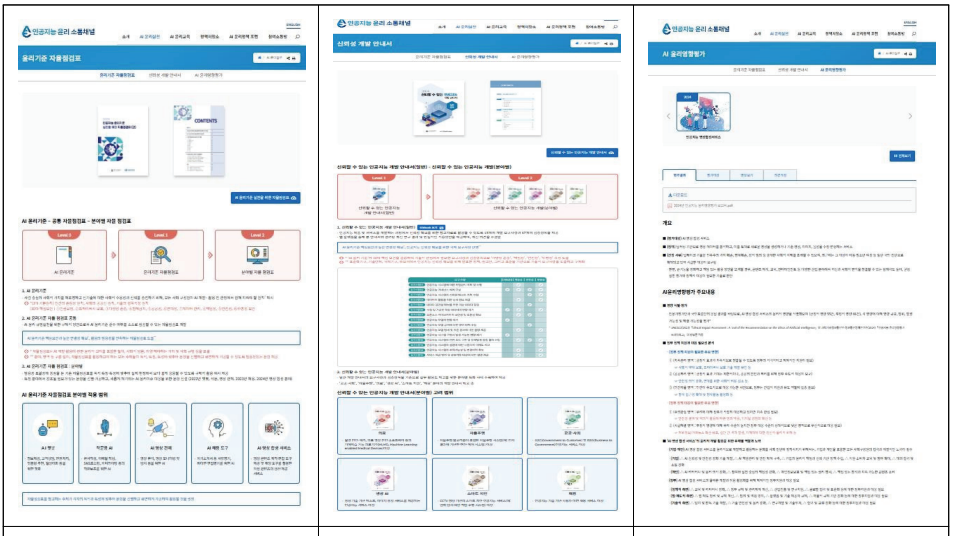
AI 윤리기준 | AI 윤리체계 마련 사업

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

나. 메뉴 ②: ‘AI 윤리실천’

「인공지능(AI) 윤리기준」의 실천수단으로서 ‘AI 윤리기준 실천을 위한 자율점검표’, ‘신뢰할 수 있는 AI 개발 안내서’, ‘AI 윤리영향평가’의 개요, 개발 및 평가 결과 등의 주요 정보를 소개하고 전자문서 형식의 자율점검표와 AI 윤리영향평가 보고서를 제공함으로써, 이용자 스스로가 AI 윤리를 점검하고 준수할 수 있도록 지원한다. 추가로 ‘윤리기준 자율점검표’와 ‘AI 윤리영향평가’ 게시판에 별도의 의견작성 기능을 추가하여 각 정책 수단에 대한 일반 이용자의 의견수렴 활성화를 도모하였다.

[그림 4-4] ‘AI 윤리실천’ 상세 페이지



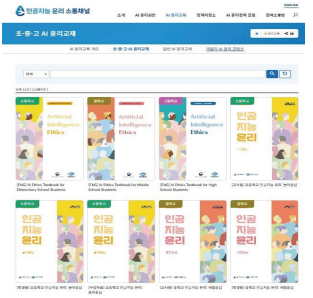
윤리기준 자율점검표	신뢰성 개발 안내서	AI 윤리영향평가

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

다. 메뉴 ③: ‘AI 윤리교육’

AI 전환기 사회에 대응하여 시민들의 AI 윤리 역량을 강화할 목적으로, AI 윤리교육의 필요성 체계적으로 제시하고 주체별 맞춤 교육 콘텐츠를 제공한다. 각 게시판을 통해 초등, 중등, 고등학교급 및 일반 성인 대상의 AI 윤리교재의 목적과 구성 특징을 각각 제시하며, 교사용과 학생용 교재, 수업자료, 초·중·고등학교 영문 교재 및 실습 자료 등을 누구나 자유롭게 다운로드하여 활용할 수 있도록 제공하고 있다. 개발자 또는 산업계 실무자를 대상으로 개발한 ‘개발자 AI 윤리 콘텐츠’는 이용자의 특성을 고려하여 GitBook 형태로 제공하고 있다. 또한, AI 윤리교육에 대한 사회적 중요도와 높은 관심을 반영하여, 이용자가 교육 콘텐츠에 대한 개선 의견이나 요구사항을 더욱 손쉽게 제시할 수 있도록 별도의 ‘의견작성’ 기능을 추가하여 운영하고 있다.

[그림 4-5] ‘AI 윤리교육’ 상세 페이지

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>AI 윤리교육 개요</td><td>초·중·고 AI 윤리교재</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>일반 AI 윤리교재</td><td>개발자 AI 윤리 콘텐츠</td></tr>
</table>

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’ 2025.12.3. 검색

라. 메뉴 ④: ‘정책저장소’

‘정책저장소’ 메뉴는 AI 윤리 관련 정책을 ‘국내외 동향’, ‘보고서/정책자료’, ‘영상자료’로 구분하여 제공한다. ‘국내외 동향’ 게시판에서는 2016년부터 2024년까지 발표된 국내외의 주요 AI 윤리원칙의 개요, 주요 특징, 개념 등을 한눈에 볼 수 있도록 정리하여 원문 링크와 함께 제공한다. ‘보고서/정책자료’ 게시판은 과학기술정보통신부의 AI 윤리 정책자료 및 본 사업의 성과물인 정책보고서를 전자문서 형태로 다운로드할 수 있다. ‘영상자료’ 게시판은 ‘AI 윤리기준’의 소개, ‘AI 윤리 공개 세미나’, ‘AI 윤리교재 온라인 설명회’ 등 관련 영상 링크를 한데 모아 제공한다.

[그림 4-6] ‘정책저장소’ 상세 페이지

국내외 동향	보고서/정책자료	영상자료

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

마. 메뉴 ⑤: ‘AI 윤리정책 포럼’

‘AI 윤리정책 포럼’ 메뉴는 2022년부터 매년 진행해 온 각 포럼의 정보를 체계적으로 제공한다. 해당 메뉴는 포럼의 개요와 위원 구성을 소개하는 ‘포럼개요’, 각 포럼 회차별로 주요 발표 및 논의 내용을 간추려서 전달하는 ‘포럼 개최 결과’, 연례행사인 공개 세미나의 정보를 담은 ‘공개 정책세미나’ 게시판으로 구성된다.

[그림 4-7] ‘AI 윤리정책 포럼’ 상세 페이지

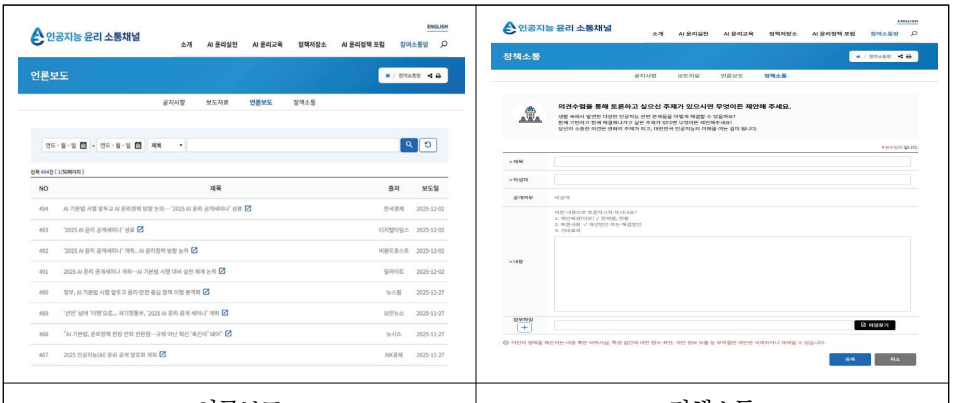
포럼 개요	포럼 개최 결과	공개 정책세미나

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

바. 메뉴 ⑥: ‘참여소통방’

‘참여소통방’은 AI 윤리 및 신뢰성 제고를 위한 정부의 정책활동을 폭넓게 제공하고 국민의 직접적인 참여를 촉구하기 위해 ‘공지사항’, ‘보도자료’, ‘언론보도’, ‘정책소통’ 게시판으로 구성되어 있다. ‘보도자료’ 게시판을 통해 AI 윤리 관련 정부 정책활동을 소개하고 관련 정보를 아카이브하여 이용자에게 전달함으로써 정책 홍보를 강화한다. ‘언론보도’ 게시판의 경우 정부 정책활동뿐 아닌 민간기업 등 다양한 주체에서 진행하는 AI 윤리 제고 활동 또한 전달하여 더욱 폭 넓고 다양한 AI 윤리 확보 방안 및 현황을 제시한다. ‘정책소통’ 게시판은 AI 윤리에 대한 국민의 자유로운 의견을 수렴하는 주요 창구로 기능한다. 이를 통해 특정 안건에 대한 의견수렴은 물론 상시적인 국민 의견수렴 활동을 진행함으로써, 국민의 정책 참여를 적극적으로 제고하고자 한다.

[그림 4-8] ‘참여소통방’ 상세 페이지



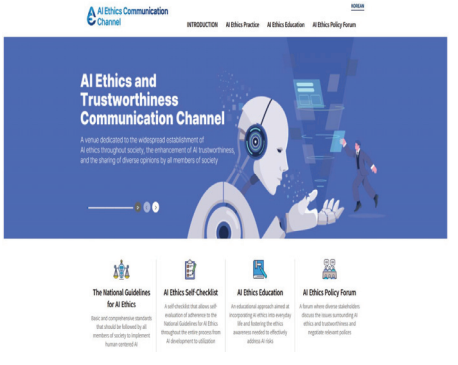
언론보도	정책소통

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

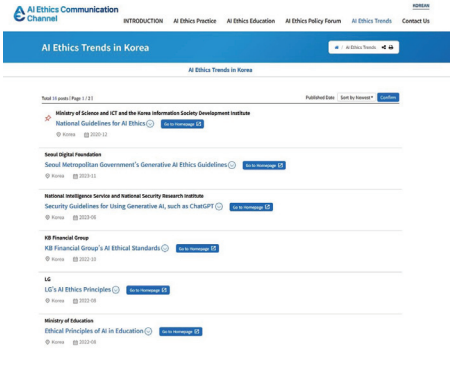
사. 영문 페이지

국제사회에서 한국의 AI 윤리 정책에 쉽게 접근하고 활용할 수 있도록 주요 정보들을 영문화하여 제공한다. 특히, AI 윤리 정착을 위한 국제 공조가 강조되고 한국의 AI 윤리정책에 대한 국제적 관심이 높아지는 추세를 반영하여, 「인공지능(AI) 윤리기준」, AI 윤리실천, AI 윤리교육, AI 윤리정책 포럼에 관한 개요를 영문으로 제공한다. 이와 더불어, 국내 주요 AI 윤리원칙의 요약본을 영문으로 제공하며, 초등·중등·고등 학교급별 영문교재 또한 누구나 다운로드하여 활용할 수 있도록 제공하고 있다.

[그림 4-9] ‘AI 윤리 소통채널’ 영문 웹사이트



메인



AI Ethics Trends in Korea

자료: ‘AI 윤리 소통채널(https://ai.kisdi.re.kr)’, 2025.12.3. 검색

2. AI 윤리 소통채널 기능 개선

AI 윤리 소통채널은 이용자 친화적 환경 조성과 웹 접근성[15] 강화를 목표로 지속적인 디자인 및 기능 개선을 진행해 왔다. 특히 2024년부터 과학기술정보통신부가 지정한 웹 접근성 품질인증기관인 ㈜웹와치의 인증 절차를 준수하며 디지털 취약계층의 접근성과 편의성을 지속적으로 개선하였다. 2025년에는 이용자 요구를 반영한 인터페이스 및 이용자 경험(UI/UX) 개선을 위해 설문조사를 진행하였다. 해당 조사를 통해 585건의 의견 수렴했고 해당 의견 분석을 바탕으로 모바일 최적화 및 홈페이지 가독성 제고 등을 포함한 로드맵을 수립·이행하여 단순한 기능 보완이 아닌 이용자 요구를 반영한 홈페이지 구축을 시도했다.

가. UI/UX 개선을 위한 이용자 설문조사

AI 윤리 소통채널은 이용자가 보다 편안하고 효과적으로 웹페이지를 사용할 수 있는 환경을 구축하기 위해 실제 사용자들의 의견을 반영한 UI/UX 개선을 추진하였다. 이를 위해 2025년 7월 14일부터 31일까지 AI 윤리 소통채널 내에서 설문조사를 실시하였으며, 확보된 총 585건의 참여 데이터를 상세하게 분석하여 홈페이지 개선 과제를 도출하였다.

설문 결과에 의하면, 객관식 문항 응답자의 87% 이상이 홈페이지에 대해 '만족' 또는 '매우 만족'을 선택하여 전반적인 이용자 만족도는 매우 높은 수준임을 확인하였다. 주관식 문항을 분석한 결과, 이용자들은 소통채널의 주요 강점으로 제공되는 콘텐츠의 권위성과 신뢰도, 디자인의 명료성과 가독성, 쾌적한 이용 환경, 정보의 유용성을 높이 평가했다. 그러나 약점으로는 비효율적인 정보 구조, 상호작용성의 부족, 검색의 어려움, 가독성 문제가 제시되었다. 이에 이용자 피드백 중 제기된 단점을 중점적으로 분석하여 모바일 최적화 및 폰트 가독성

15) 웹 접근성은 장애인, 고령자 등 정보 약자가 비장애인과 동등하게 웹 정보를 이용할 수 있도록 보장하는 것을 의미하며, 웹 접근성 품질 인증은 이러한 웹 접근성 표준을 준수한 사이트에 부여되는 공식 인증 제도를 지칭

향상, 검색 기능 강화, 상호작용 및 커뮤니티 기능 강화를 핵심 진단 과제로 확정했다. 해당 설문조사 결과 분석과 내·외부 전문가들의 의견을 바탕으로 구체적인 개선 작업을 진행하였으며 자세한 사항은 다음에서 확인할 수 있다.

〈표 4-2〉 AI 윤리 소통채널 주요 강점 및 약점

강점	콘텐츠의 권위성과 신뢰도	- 이용자는 AI 소통채널이 제공하는 정보의 깊이, 객관성을 가장 큰 장점으로 인식
	디자인의 명료성과 가독성	- 복잡한 윤리적 개념을 다룸에도 불구하고, 깔끔하고 전문적인 시각적 디자인이 내용의 이해도 제고
	쾌적환 이용 환경	- 불필요한 광고, 팝업이 없어 콘텐츠에 집중할 수 있는 환경 제공
	정보의 유용성	- AI 윤리정책의 핵심 내용을 전달하며, AI 윤리기준의 유용한 활용 방안을 제시
약점	비효율적인 정보 구조	- 채널이 보유한 정보의 전체 범위를 한눈에 인식하기 어려움
	상호작용성의 부족	- 일방적인 정보 전달이 주를 이루며, 이용자들이 질문, 토론할 수 있는 커뮤니티 및 참여 기능 확대 부족
	검색의 어려움	- 내비게이션 및 검색 기능이 취약하여 이용자가 이미 알고 있는 경로 외 새로운 콘텐츠 발견이 어려움
	가독성 문제	- 모바일 기기에서 일부 페이지 화면 확인이 어려우며, 텍스트 크기가 작아 가독성이 떨어짐

자료: 연구진 작성

나. 기능 개선 및 효율화

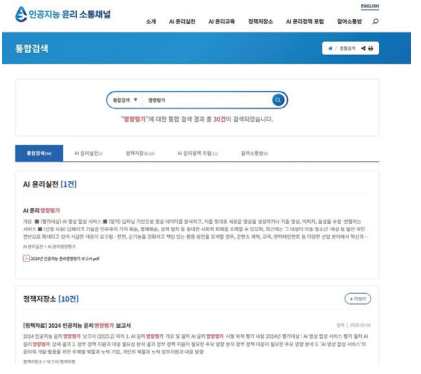
AI 윤리 소통채널은 이용자 설문조사 결과 분석을 기반으로 UI/UX를 포함한 핵심 기능 전반을 개선하여 이용 편의성과 정보 접근 효율성을 제고하였다. 기능 개선 및 효율화는 크게 UI/UX 및 접근성 개선, 검색 기능 및 정보제공 능력 강화, 커뮤니티 및 양방향 소통 기능 확대 세 가지 목표를 중심으로 추진되었다.

먼저 UI/UX 및 접근성 개선을 위해 모바일 환경에서 안정적인 구동이 가능하도록 반응형 페이지를 전체적으로 점검하고 폰트 크기 및 간격을 통일하여 가독성을 높였다. 또한, 이용자 친화적 환경 구축을 위해 게시판 상단부의 불필요한 내용을 제거하여 핵심 콘텐츠로의 접근 단계를 최소화했다. 직관성이 낮은 인터페이스는 직관적인 아이콘 및 이용 방식으로 대체하거나 마우스 오버 시 상세 설명을 제공하는 방식을 추가하여 사용자 이용 편의성을 제고했다. 아울러 웹페이지 상단에 상위 구조를 명시하는 사이트 이동 경로(breadcrumb) 기능을 개선하여 이용자가 소통채널의 전체 구조를 쉽게 파악할 수 있도록 조치하였다.

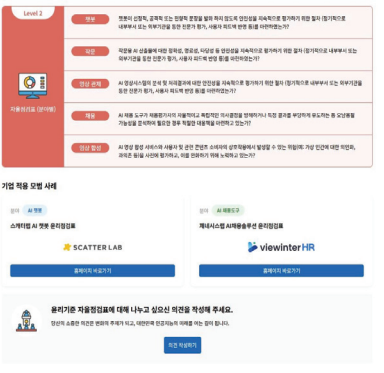
정보 탐색의 효율성을 높이기 위해 검색 기능을 대폭 개선하였다. 기존의 메뉴명과 게시글 제목에 한정되었던 검색 범위를 콘텐츠 내용까지 확대하여 검색 결과의 정확도와 질을 제고하였다. 또한, 기존에 검색 기능이 부재했던 영문 소통채널에도 검색 기능을 추가하여 국제 이용자의 정보 접근성을 확보하였다. 이외에도 더욱 폭넓은 AI 윤리 정책 정보제공을 위해 보도자료 관련 Open API와 RSS 서비스를 확대하였다.

다양한 이해관계자의 의견을 더욱 효과적으로 수렴하기 위하여 의견수렴 기능 또한 강화했다. 기존에는 '참여소통방' 메뉴 하위의 '정책소통' 게시판을 통해서만 의견수렴이 가능하여 접근 단계가 복잡했으나, 이를 개선하고자 '윤리기준 자율점검표', 'AI 윤리영향평가', 'AI 윤리교육' 등 핵심 게시판에 의견수렴 기능을 추가하여 접근성을 높였다. 이와 더불어 영문 소통채널 또한 의견수렴(Contact Us) 메뉴를 신설하여 국제적인 소통 창구를 마련하였다.

[그림 4-10] ‘AI 윤리 소통채널’ 기능 개선



검색 결과 예시



의견수렴 기능 추가

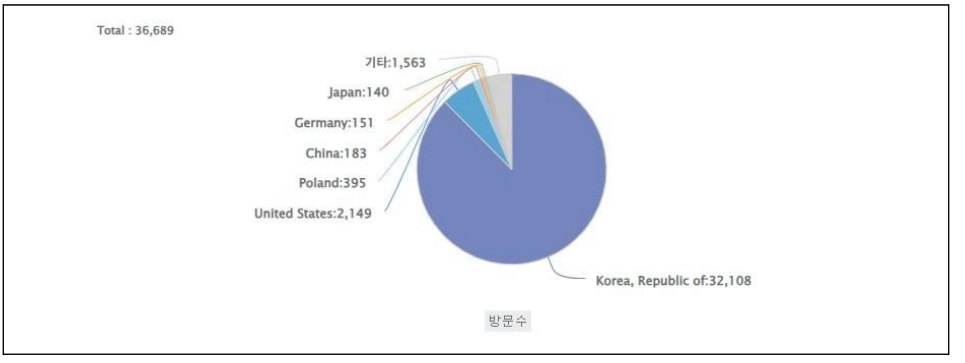
자료: 연구진 작성, 2025.12.5. 검색

다. 이용자 증대 노력

AI 윤리 및 안전에 대한 국제사회의 논의가 활발하게 확대되는 정책환경에 대응하여, AI 윤리 소통채널은 국제사회와의 소통을 강화하기 위해 영문 서비스 운영을 확대하였다. 유네스코(UNESCO)를 비롯해 아세안(ASEAN), 남미 등 전 세계 다양한 국가에서 한국의 AI 윤리 정책 및 사례에 대한 관심이 지속적으로 증대하는 추세이다. 이러한 국제적 관심을 반영하듯, 2025년 소통채널의 방문자(36,689명)[16] 중 국외 방문 비율이 약 12.49%(4,581명)로 집계되었다. 이에 따라 소통채널은 AI 윤리 관련 국내 현황 및 정책의 영문 제공 자료를 확대하고, 국제적인 의견수렴 창구를 개설하였다. 특히, 한국에서 발표된 AI 윤리원칙들을 체계적으로 정리하여 영어로 제공함으로써 국제적 이해와 협력을 증진하고자 노력하고 있다. 이러한 노력을 시작으로 소통채널은 한국의 AI 윤리정책 경험을 공유하는 국제적 AI 윤리정책 지식정보 공유 허브로 그 역할을 더욱 확대하고자 한다.

16) AI 윤리 소통채널의 2025년 방문자 수의 총합은 36,734명으로 집계되었으나, 국가별 방문자 분석은 소속 국가가 불명확한 방문자 수를 제외한 36,689명을 대상으로 진행함

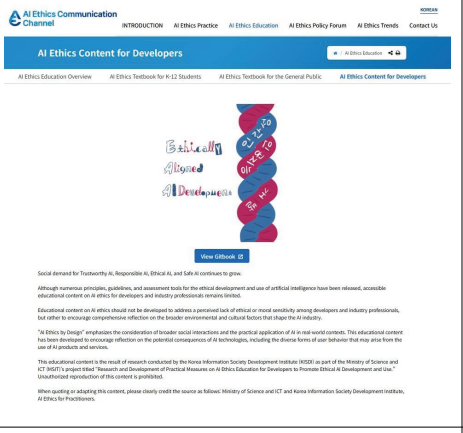
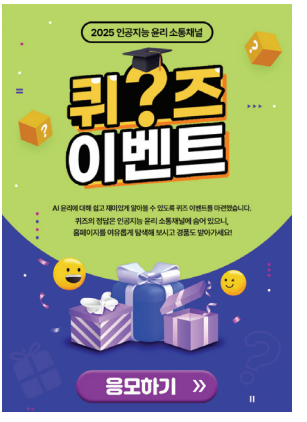
[그림 4-11] 국가별 방문자 통계



자료: 연구진 작성(온라인 웹로그 분석 서비스 Ace Counter 이용)

AI 윤리에 대한 대중적 관심 유도와 AI 윤리 소통채널 활성화를 목표로 2025년 11월 19일부터 30일까지 퀴즈 이벤트를 진행하였다. 2024년 의견수렴 이벤트가 주관식으로 진행되어 이용자들의 참여 부담이 높았다는 점을 고려하여, 2025년에는 참여 문턱을 낮추고 흥미를 높이기 위해 소통채널 탐색을 기반으로 하는 객관식 퀴즈 형식을 채택했다. 퀴즈 문항은 AI 윤리 실천의 가장 핵심적인 기반이 되는 「인공지능(AI) 윤리기준」과 AI 윤리·신뢰성 실천 및 AI의 윤리적 주체적 활용을 위한 'AI 윤리영향평가'에 대한 내용으로 구성했다. 본 퀴즈 이벤트는 단순한 정답 맞히기가 아닌 참가자들이 정답을 찾는 과정에서 소통채널 내의 AI 윤리 관련 콘텐츠를 자연스럽게 탐색하도록 유도하고자 하였다. 퀴즈 이벤트에는 총 5,250명이 참여하여 홈페이지 뷰 수 증대와 탐색 활성화라는 목표를 달성했다. 특히, 참가자들이 AI 윤리 정보를 스스로 습득하는 능동적 학습 기회를 제공하여 이벤트의 교육적 의미를 강화했다. 또한, 이번 이벤트는 일견 어렵게 느껴질 수 있는 AI 윤리라는 주제에 쉽고 재미있게 접근하는 기회를 제공했다는 점에서 의의가 있다.

[그림 4-12] 이용자 증대 노력

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>영문 소통채널 확대</td>
<td>퀴즈 이벤트 홍보 이미지</td>
</tr>
</table>

자료: 연구진 작성, 2025.12.3. 검색

3. AI 윤리 소통채널 성과 및 향후 계획

2025년 AI 윤리 소통채널은 AI 윤리 관련 접근성을 높이고 사업 및 정책 정보를 제공하는 본연의 기능을 충실히 수행하였다. 특히, 이용 지표가 전년 대비 크게 상승하며 채널 운영의 효과성을 입증하였다. 2025년 총 방문자 수는 36,734명, 총 페이지 뷰는 90,806건으로 집계되었다. 이는 전년도 동기간 이용 통계[17]와 비교했을 때 괄목할 만한 성과로써, 방문자 수 약 44.27% 증가, 페이지 뷰는 약 56.03% 증가하였다. 특히 페이지 뷰의 높은 증가율은 단순한 방문자 유입의 증가를 넘어, 이용자가 채널 내 콘텐츠를 더 깊이 있고 다양하게 탐색했음을 보여준다. 이는 콘텐츠 이용도와 채널 체류 시간이 함께 증가한 것으로 해석할 수 있어, 소통채널이 양적 성장뿐 아니라 질적 성장도 동시에 달성했음을 시사한다.

17) AI 윤리 소통채널의 2024년 방문자 수와 페이지 뷰는 각각 25,468명, 58,196건으로 집계됨

〈표 4-3〉 2025년 월간 이용 통계

월간	방문자 수	페이지 뷰
1월	1,372	4,905
2월	1,497	4,184
3월	2,142	6,650
4월	2,335	6,261
5월	2,723	6,619
6월	3,066	7,064
7월	3,208	8,217
8월	2,225	5,072
9월	3,402	8,266
10월	3,459	8,578
11월	7,805	17,202
12월	3,500	7,788
합계	36,734	90,806

자료: 연구진 작성, 2026.1.19. 검색

AI 윤리 소통채널은 AI 윤리 정책 및 관련 자료를 체계적으로 축적·제공하여 지식정보 공유 허브 역할 또한 성공적으로 수행했다. 2025년까지 약 150개의 정책자료를 누적 제공하였으며, 해당 자료들의 총 다운로드 횟수는 57,000여 건으로 집계되었다. 이는 채널의 핵심 목적 중 하나인 정책 확산 기능을 효과적으로 달성한 것으로 평가할 수 있다. 주요 다운로드 자료 분석 결과, 이용자들은 AI 윤리의 기본 원칙과 실천 사례에 높은 관심을 보이며, 소통채널이 제공하는 콘텐츠를 활발히 이용하고 있음이 확인되었다. 정보통신 정책연구원에서 수행한 연구·개발 결과물인 '사람이 중심이 되는 인공지능(AI) 윤리기준'을 비롯하여, '2024 윤리영향평가', 'AI 윤리기준 실천을 위한 자율점검표', 초·중·고 및 일반 AI 윤리 교재·교육 자료 및 정책보고서 등이 특히 높은 다운로드 수치를 기록했다. 또한 AI에 대한 정부의 거시 정책자료인 'AI 디지털 혁신 성장전략(안) 요약본', '전국민 AI 일상화 실행계획' 역시 높은 다운로드 횟수를

기록했으며, 이는 이용자들이 AI 관련 정책 전반에 걸쳐 정보를 활발히 탐색하고 있음을 나타낸다.

특히, 2024년 윤리영향평가 자료에 대한 이용자들의 관심은 주목할 만한 요소이다. 단순히 윤리영향평가의 최종 결과에만 관심이 집중된 것이 아닌, 평가 및 대상 소개 자료, 1차 평가 결과 자료, 공개 토론회 발표 자료 등 정책 개발 및 적용 전체 과정과 관련한 모든 자료에 대해 높은 관심이 확인되었다. 이는 이용자들이 완결된 최종 정책뿐 아니라 정책이 수립되고 실행되는 과정에 대해서도 높은 관심을 가지고 적극적으로 정보를 탐색하고 있다고 유추할 수 있다. 또한, 소통채널이 이처럼 정책 수립 과정 정보를 투명하게 제공하는 것은 국민의 적극적 정보 탐색과 능동적 판단 경험을 촉진하며, 궁극적으로 정책 과정의 투명한 공개와 국민의 자발적 참여를 지원하여 디지털 사회의 신뢰 자본 축적과 AI 윤리 정착 토대 마련에 기여한다고 판단된다.

〈표 4-4〉 정책자료 다운로드 통계(2025년 상위 10건)

순위	정책자료	다운로드 수
1	사람이 중심이 되는 인공지능(AI) 윤리기준	4,692
2	2024 윤리영향평가 보고서 외[18]	3,957
3	2024 윤리기준 실천을 위한 자율점검표(안)	3,042
4	2025 인공지능 윤리기준 실천을 위한 자율점검표(안)	3,003
5	‘고등학교 인공지능 윤리: 탐구중심’ 교재 및 수업자료	2,829
6	‘초등학교 인공지능 윤리: 놀이중심’ 교재 및 수업자료	2,616
7	2023 인공지능 윤리기준 자율점검표(안)	2,428
8	‘중학교 인공지능 윤리: 체험중심’ 교재 및 수업자료	1,764
9	‘일상 속 인공지능 윤리 이야기’ 일반 AI 윤리교재 및 교수학습자료	1,494
10	2024 신뢰할 수 있는 인공지능 개발 안내서	1,386

자료: 연구진 작성, 2026.1.19. 검색

18) 2024년 AI 윤리영향평가 보고서를 비롯한 1차 평가 결과 자료집, 공개 토론회 발표자료 등 중간 결과 공개자료들의 다운로드 수를 합산

향후 AI 윤리 소통채널의 지속적인 발전을 위해 유입 채널 최적화와 시의성을 고려한 콘텐츠 확장을 추진할 계획이다. 2025년 소통채널의 유입 패턴 분석 결과, ‘AI 윤리기준’, ‘윤리 교육’ 등 특정 키워드를 통한 검색 유입이 가장 많은 것으로 파악된다. 홈페이지 유입 제고를 위해 핵심 키워드 기반 검색엔진 최적화(SEO, Search Engine Optimization)를 심화할 예정이다. 소통채널의 전체 및 개별 콘텐츠에 대한 메타 태그 설정을 더욱 정교하게 다듬고 지속적인 SEO 작업을 통해 검색 결과 노출 순위를 높여 자연 유입을 강화하고자 한다. 더불어, 전 세계적으로 AI 윤리 정책이 윤리 원칙 제정을 넘어 법 제정으로 확장되는 흐름을 반영한 시의성 있는 콘텐츠를 선제적으로 제공하고자 한다. 특히 「AI 기본법」 도입에 발맞춰 기존의 AI 윤리 원칙과 AI 기본법의 관계를 설명하는 콘텐츠를 추가 제공하고 「AI 기본법」 시대에 맞는 AI 윤리정책에 대한 정보를 지속적으로 제공하여, 이용자들이 급변하는 법적, 윤리적 환경에 대비할 수 있도록 지원을 강화할 예정이다.

제2절 소 결

정보통신정책연구원은 AI 윤리를 사회에 확립하고 관련 논의를 촉진하기 위한 기반을 마련하고자 2023년 11월부터 온라인 플랫폼 형식의 ‘AI 윤리 소통채널’을 운영하고 있다. 소통채널은 기업, 학계, 시민사회 등 다양한 이해관계자에게 AI 윤리 정책 정보 등을 제공하는 지식정보 공유의 허브 역할을 수행하며, 국가 AI 윤리 정책자료 아카이브 및 윤리 이슈 논의를 위한 정책 허브로 기능한다. 이를 통해 AI 정보격차를 줄이고 국민의 AI 리터러시 수준 향상을 지원하며 정부 정책 수립의 투명성을 높이는 데 기여하고자 한다. 이러한 목표를 효율적으로 달성하기 위해 소통채널은 △사업 소개, △정책 홍보, △국제사회 소통, △윤리정책 수집, △의견 청취 등의 핵심 기능을 제공하며, AI·디지털 전환 시대의 대응 역량을

강화하는 데 핵심적인 역할을 수행 중이다.

AI 윤리 소통채널은 웹 접근성 강화와 이용자 편의성 증진을 목표로 지속적으로 개선됐으며, 특히 2024년부터 지속해서 웹 접근성 품질인증을 준수하여 디지털 취약계층의 접근성을 높였다. 2025년 이용자 설문조사(585건) 결과를 바탕으로 모바일 최적화, 가독성 향상, 정보 구조 및 검색 기능 강화를 핵심적으로 추진하여 직관적인 UI 개선 및 정보 탐색 효율화를 제고했다. 또한, 핵심 게시판에 의견수렴 기능을 추가하여 양방향 소통 접근성을 높였으며, 국내 AI 정책에 대한 해외의 관심에 대응하여 영문채널 제공 콘텐츠를 확대했다. 더불어 퀴즈 이벤트(5,250명 참여)를 진행하여 홈페이지 활성화와 AI 윤리 콘텐츠의 능동적 학습 기회를 제공하는 성과를 거두었다.

2025년 AI 윤리 소통채널은 AI 윤리 관련 정보 제공이라는 기본 목적을 충실히 수행하며 양적, 질적 성장을 이루었다고 평가할 수 있다. 2025년 총 방문자 수는 36,734명(전년 대비 약 44.27% 증가), 총 페이지 뷰는 90,806건(전년 대비 약 56.03% 증가)을 기록했는데, 이는 이용자들이 소통채널에 더욱 길게 체류하며 콘텐츠를 깊이 있게 탐색했다고 해석할 수 있다. 또한, 약 150개의 축적된 정책자료 제공했으며, 정책자료의 다운로드 횟수가 57,000여 건을 기록함으로써 정책정보 제공이라는 목표를 효과적으로 달성한 것으로 판단된다. 더불어, 이용자들이 '윤리영향평가'와 관련하여 정책 수립 과정의 중간 자료에도 높은 관심을 보인 점을 통해 소통채널이 정책 투명성을 높이고 국민의 능동적 참여를 촉진하는 데 기여했다고 해석할 수 있다. 또한 소통채널은 지속적인 발전을 위해 2025년 수집된 이용자들의 경향 자료를 바탕으로 향후 핵심 키워드 기반의 검색엔진 최적화(SEO) 작업을 심화하여 자연 유입을 증대할 계획이다. 더불어 「AI 기본법」 시행이라는 정책적 변화를 반영하여 관련 콘텐츠를 확대 제공함으로써, 이용자들이 AI와 관련하여 급변하는 법적·윤리적 환경에 효과적으로 대비할 수 있도록 지원을 강화할 예정이다.

제5장 결 론

제1절 연구결과 요약 및 정책적 함의

1. AI 윤리영향평가 시행

AI 윤리영향평가는 기술의 사회적 수용성을 높이고 지속 가능한 발전을 견인하는 핵심 거버넌스 기제로서 그 위상을 확립해 나갈 필요가 있다. 본 연구는 작년에 이어 올해 2차 시범 적용을 통해 윤리영향평가 프레임워크의 실효성을 검증하고, 기술의 혁신성과 윤리적 안전성 간의 균형점을 모색했다는 점에서 중요한 의의를 갖는다. 특히, 이번 평가는 'AI 채용 서비스'를 대상으로 하여, 기술 오남용과 편향성, 개인정보 유출 등 잠재적 리스크를 진단하는 동시에, 서비스의 책임 있는 활용과 확산을 위해 정부, 기업, 시민 차원에서의 역할과 노력을 중심으로 입체적인 가이드라인을 제시하였다. 이번 평가를 통해 AI 기반 채용은 효율성과 절차적 정당성 확보라는 긍정적 기대를 받는 동시에 정성적 판단의 부재, 데이터 편향, 설명 가능성의 한계라는 구조적 위험 요인을 내포하고 있음이 확인되었다. 특히 주목할 점은 AI 채용 서비스 자체에 대한 일반 시민의 인식과 태도는 서비스 이용 경험의 유무에 따라 일정한 차이를 보인 반면, 해당 서비스로 인한 부정적 영향을 관리할 정부의 전문성과 변화 관리 역량에 대한 신뢰 수준은 전반적으로 낮게 나타났다는 점이다. 이는 향후 정책 추진 시 기술적 보완 못지않게 정부의 제도적 신뢰 기반을 강화하는 노력이 선결 과제임을 시사한다.

전문가 평가단과 국민포럼단이 5대 핵심 윤리 영역(프라이버시, 포용성, 책임성, 투명성, 공정성)을 정량·정성적으로 입체 평가한 결과, AI 채용의 순기능과 역기능이 병존하는 것으로 나타났다. 보다 구체적으로는 절차적 공정성과 정보 수집의 최소화는 긍정적으로 평가되었으나, 민감정보 추론 가능성, 디지털 격차, 책임

소재의 모호함은 여전한 리스크로 지목되었다. 특히 FGI 결과는 전문가 평가와 높은 정합성을 보이며, 직무 무관 정보 활용이나 사회적 약자 배제 등 수요자 관점에서 제기되는 구체적인 우려를 심도 있게 드러냈다.

이 과정에서 전문가, 시민사회, 일반 대중이 함께 참여하는 다층적 거버넌스(Multi-stakeholder Governance)를 가동하여 평가의 투명성과 절차적 정당성을 확보하였다는 점은, 본 윤리영향평가 결과의 신뢰성과 타당성을 제고하는 동시에 향후 국제사회에서 AI 윤리영향평가의 하나의 참고 기준을 제시한다는 점에서도 중요한 의의를 지닌다. 무엇보다 전문가와 국민포럼단이 함께한 숙의 과정은 AI라는 신기술 정책 수립에 있어 사회적 합의와 투명성을 확보하는 모범 사례로 자리 잡을 수 있다. 이러한 노력은 향후 AI 개발 및 운영의 전 과정에서 발생할 수 있는 윤리적 이슈와 잠재적 위험을 선제적으로 관리하고 예방하는 데 결정적인 역할을 할 수 있을 것으로 보인다.

또한 이는 단순히 윤리적 논란과 부정적 영향을 예방하는 차원을 넘어 개발자와 사용자 모두에게 '책임 있는 AI(Responsible AI)'를 실현할 수 있는 구체적인 지식과 실천적 도구를 제공했다는 데 함의가 있다. 본 연구가 제안한 정책적·기술적 조치들은 향후 AI 전 주기에 걸쳐 윤리적 고려를 내재화하는 구조적 토대가 될 것으로 희망한다. 결론적으로 AI 윤리영향평가가 우리 사회의 필수적인 안전장치이자 혁신의 촉매제로 안착하기 위해서는 지속적인 정책 지원과 국제적 연대가 필수적이다. 본 연구의 제언이 마중물이 되어 AI 기술이 인류의 복지와 공공선에 기여하는 '인간 중심의 AI 생태계'로 나아가기를 기대한다.

2. AI 윤리기준 자율점검표 개발·적용

AI 기술이 헬스케어 분야의 핵심 인프라로 급부상함에 따라 대두된 복합적인 윤리적 과제에 대응하기 위해 추진된 '헬스케어 분야 AI 윤리기준 자율점검표(안)'의 개발 배경과 그 구체적인 실행 과정을 상세히 기술하였다. 최근 의료 AI 기술은 방대한 임상 데이터를 기반으로 질병 예측 및 진단의 혁신을 주도하고 있으나,

알고리즘 오류로 인한 환자 안전의 위협, 민감정보 집적에 따른 프라이버시 침해 등 다양한 윤리적 위험을 동반하고 있다. 비록 「AI 기본법」이나 「디지털의료제품법」 등 법적 규제 체계가 마련되고 있지만, 급변하는 기술 속도와 의료 현장의 복잡성을 경직된 법규만으로 포괄하기에는 한계가 명확하여, 법적 규제를 보완하고 개발 단계에서부터 잠재적 위험을 선제적으로 관리할 수 있는 특화된 윤리 기준의 필요성이 제기되었다. 이에 정보통신정책연구원은 헬스케어 분야의 특수성에 부합하는 윤리적 기준을 수립하고 이를 현장에 안착시키기 위한 자율점검표 개발을 추진하였다. 연구진은 국내외 문헌 연구와 규제 동향 분석을 통해 헬스케어 분야에서 특히 강조되어야 할 점을 식별하였고, 이를 '인공지능 (AI) 윤리기준'의 10대 핵심요건에 결합하여 자율점검표 초안을 구성하였다. 이후 각계 전문가의 의견을 수렴하여 점검 문항을 수정하고 기업 현장 적용 과정을 거쳐 최종안을 마련하였다.

이러한 헬스케어 분야 자율점검표 개발·적용 과정은 AI 윤리기준을 선언적 원칙에 머무르게 하지 않고 실제 서비스 설계·운영 단계에서 활용 가능한 실천 도구로 구체화하였다는 점에서 정책적 의의를 갖는다. 특히 법률 중심의 사후 규제만으로는 포착하기 어려운 의료 AI의 윤리적 위험을 개발·운영 단계에서 사전에 점검하도록 유도함으로써 자율규제와 법적 규제를 보완적으로 연결하는 정책 수단으로서의 가능성을 확인하였다. 이는 향후 「인공지능(AI) 윤리기준」을 현장에 안착시키기 위한 다양한 실천 도구 설계에 있어 중요한 참고 사례로 활용될 수 있을 것이다.

3. AI 윤리 소통채널 구축·운영

정보통신정책연구원이 2023년 11월부터 운영하고 있는 'AI 윤리 소통채널'은 AI 윤리의 사회적 정착을 지원하고 정책 논의의 기반을 마련하는 핵심 플랫폼으로 자리매김하고 있다. 소통채널은 다양한 이해관계자에게 AI 윤리 정책 정보를 제공하는 지식 공유 허브이자 국가 정책 자료 아카이브로서 기능하며, AI 격차

완화, AI 리터러시 제고, 정책 수립의 투명성 강화에 기여하고 있다.

웹 접근성 품질인증 준수, 2025년 이용자 설문조사(585건) 결과를 반영한 UI/UX 및 검색 기능 개선은 디지털 취약계층을 포용하고 정보 접근성을 제고하기 위한 실질적 정책 노력이다. 이러한 개선의 효과는 2025년 누적 방문자 수 36,734명(전년 대비 44.27% 증가), 정책자료 다운로드 57,000여 건 등 의미 있는 성과로 나타났으며, 이는 소통채널이 정책정보 확산이라는 본래의 목표를 효과적으로 달성했음을 보여준다.

또한 이용자들이 '윤리영향평가' 관련 자료에 높은 관심을 보인 점은 소통채널이 정책 과정의 투명성을 강화하고 국민의 능동적 참여를 촉진함으로써 디지털 사회의 신뢰 자본을 축적하는 역할을 하고 있음을 시사한다. 이와 함께 영문 채널 확장, 퀴즈 이벤트(5,250명 참여) 등을 통해 국제적 협력 기반을 넓히고 대국민 소통을 활성화하는 등 정책적 목표 달성을 위한 노력을 지속해 왔다.

이러한 온라인 소통채널의 구축·운영 성과는 AI 윤리 정책과 제도 논의가 일부 전문가나 정책 담당자에 국한되지 않고 국민과 지속적으로 공유·축적되는 구조를 마련하였다는 점에서 정책적 의의를 갖는다. 특히 윤리영향평가, 자율점검표, 교육 콘텐츠 등 다양한 윤리 실천수단의 결과와 자료를 통합적으로 제공함으로써, 개별 정책 수단을 연결하는 '정책 인프라'로 기능하고 있다. 이러한 기능은 AI 기본법 시행 이후 변화하는 정책 환경에서, 국민의 신뢰와 이해를 바탕으로 하는 참여형 AI 거버넌스를 실현하는 데 중요한 기반으로 활용될 수 있을 것이다.

제2절 향후 정책 방향

AI 기술의 비약적인 발전은 전 산업 분야의 혁신을 주도하며, 사회 전반의 대전환을 촉진하고 있다. 이러한 흐름 속에서 법과 제도, AI 윤리는 기술의 혁신성을 보존하면서도 신뢰성 및 안전성을 확보하여야 하는 막중한 과제에 직면해 있다.

이에 시대적 요구에 부응하고, 기술이 사회에 안착할 수 있도록 윤리적 안전망을 구축하기 위한 정책 방향을 다음과 같이 제안한다.

2025년 AI 채용 서비스를 대상으로 실시한 윤리영향평가는 식별된 부정적 영향의 성격에 따라 정책 대응의 우선순위와 책임 주체를 세분화하였다. 구체적으로 편향 검증과 같이 공익적 관리가 요구되는 영역, 장애인 접근성처럼 지속적인 제도적 지원이 필요한 영역, 데이터 관리와 같이 민간의 자율적 대응이 상대적으로 효율적인 영역을 구분하였다. 특히 민감정보의 과도한 추론이나 형식적 수준에 그치는 인간 개입과 같은 고영향 요소에 대해서는 정부의 즉각적이고 시급한 개입 필요성을 강조하였다. 이러한 평가 결과를 토대로 기술적·사회적 맥락을 종합적으로 고려하여 정책 대응의 우선순위를 설정하고, 이에 기반한 정부 정책을 마련할 필요가 있다. 즉, 윤리영향평가 결과를 단순한 긍·부정 영향 제시에 그치지 않고 긍정적 영향은 극대화하며 부정적 영향은 최소화할 수 있는 정책 과제를 도출하고, 이들 과제 중에서도 실효성, 파급력, 시급성 등을 기준으로 정책 우선순위를 설정하여 단계적으로 추진해야 할 것이다.

또한 이러한 영향평가 프레임워크와 결과를 주요 AI 국가 및 국내외 기업들과 공유하고 국제사회에 '프로세스' 표준을 적극적으로 제안할 필요가 있다. AI 윤리영향평가 프레임워크는 단순한 체크리스트 수준을 넘어 전문가 평가와 시민 참여형 평가를 결합한 독창적 모델을 제시했다는 점에서 국제적으로 경쟁력이 있다. 일반 시민의 인식조사와 일반 국민으로 구성된 포럼단이 실제 평가과정에 참여함으로써, AI 영상 합성 서비스 및 AI 채용 서비스와 같은 민감한 이슈에 대해 시민들이 실제 느끼는 불안감과 우려를 구체적으로 도출해 낼 수 있었다. 이는 기술 중심 평가가 놓치기 쉬운 사회적 수용성 데이터를 확보하는 핵심 매커니즘이 될 수 있다는 점에서 의미가 있다. 따라서 그간 축적한 우리나라 AI 윤리 확보 노력을 국제사회와 공유하고 동시에 영향평가 상호 인정 등 국가 간 협력 기반을 마련하는 노력이 필요하다. 이러한 노력이 뒷받침될 때, 우리는 글로벌 차원에서 신뢰 기반의 AI 수용성을 한층 강화하는 데 기여할 수 있을

것이다.

AI 윤리의 구체적 실천수단으로 자리매김하고 있는 자율점검표와 관련해서는 다음과 같은 정책 방향을 제안하고자 한다. 첫째, 향후 개발 및 보급될 'AI 윤리기준 자율점검표'는 AI 서비스이용자의 특성과 역량을 고려하여 다각적으로 세분되어야 한다. 그간 추진해 온 산업 분야별 세분화 방식이 AI 윤리기준의 현장 적용성을 높이는 데 크게 기여했다는 점은 분명하다. 그러나 이러한 접근만으로는 실제 서비스를 제공받는 이용자의 다양성을 섬세하게 포착하기에는 한계가 있었다. 특히 아동, 청소년 등 각 대상이 지닌 신체적·정신적 특성에 따라 취약성의 정도와 양상이 다를 수 있음에도, 이를 충분히 반영하지 못했다는 점은 시급히 보완해야 할 과제이다. 따라서 앞으로는 기존의 분야별 분류를 기본으로 삼되, 이용자의 특성과 역량까지 고려한 세분화를 추진하여 이러한 사각지대를 해소할 필요가 있다. 이러한 다차원적 접근을 시도할 때, '인공지능(AI) 윤리기준'이 산업 현장과 국민의 삶 속에 실질적으로 확산되고 정착될 수 있는 계기가 마련될 것이다. 둘째, 기존의 '인공지능(AI) 윤리기준'을 급격한 기술적 진보와 관련 법·제도의 개선 사항 등 대내외 환경 변화를 시의적절하게 반영하여 최신화하여야 한다. 윤리기준은 한 번 정해지면 변하지 않는 고정된 규범이 아니라, 기술 발전 속도와 사회적 인식의 변화에 맞춰 유연하게 진화하는 체계여야 하기 때문이다. 이러한 지속적인 현행화 과정은 향후 'AI 윤리기준 자율점검표'가 현실과 괴리되지 않고, 현장에서 실제로 작동하는 지속 가능한 AI 윤리 실천 수단으로서 그 기능을 다할 수 있도록 토대를 다지는 핵심 작업이다.

다음으로 AI 윤리정책 수립과 관련하여 다양한 이해관계자가 참여하는 논의의 장으로 기능하고 있는 윤리소통채널 운영 방향에 대해서 다음과 같이 제안하고자 한다. 우선 AI 윤리 소통채널은 AI 윤리의 지속적 확산과 시민 역량 강화, 그리고 포용성을 고려하고 사회적 합의 형성을 지원하는 공론장의 기능을 안정적으로 이어갈 필요가 있다. 이를 위해, 플랫폼의 운영 기반을 강화하는 동시에, 기능적·내용적 측면에서의 개선과 확장을 지속적으로 추진해야 한다. 특히, 검색 유입

구조 변화에 대응하여 검색엔진 최적화(SEO)를 고도화하고, 메타 태그 관리 등 기술적 요소를 정교하게 개선함으로써 핵심 주제어 기반의 자연 유입을 확대해야 한다. 또한 2026년 1월 시행되는 「AI 기본법」에 따른 정책 환경 변화를 선제적으로 반영하여 관련 콘텐츠를 확충함으로써, 이용자들이 급변하는 AI 기본사회에 효과적으로 적응할 수 있도록 지원 역량을 강화해 나가야 한다.

흔히 규제와 윤리는 혁신의 속도를 저해하는 요소로 오해되지만, 브레이크 없는 자동차가 결코 안전하게 속도를 낼 수 없듯이 견고하고 신뢰할 수 있는 안전장치는 지속 가능한 혁신의 전제 조건이다. 이러한 제도적·윤리적 기반은 기업들이 글로벌 시장이라는 고속의 경쟁 환경 속에서 위험을 관리하며 안심하고 도전할 수 있도록 하는 필수적 토대가 된다. 아울러 국민 누구나 AI의 편익을 공정하게 향유하고 이를 주체적으로 활용할 수 있는 '모두의 AI'와 'AI 기본사회'를 구현하기 위해서도 AI 윤리는 선택이 아닌 필수불가결한 기본 전제라 할 수 있다. 이러한 정책적 노력이 축적될 때, 대한민국은 AI 3대 강국으로의 도약과 더불어 국제 AI 질서 형성 과정에서 규칙을 수동적으로 수용하는 위치를 넘어 규범을 설계하고 제시하는 규칙 설계자로 자리매김할 수 있는 확고한 기반을 갖추게 될 것이다. 이러한 노력은 「AI 기본법」의 사회적 안착과 유기적으로 연계될 필요가 있다. 진흥과 규제의 균형이라는 관점에서 필요최소한의 규제를 담고 있는 동 법제는 실제로 시장, 사회, 국민, 공공 전반에 걸쳐 다차원적인 영향을 미칠 것으로 예상된다. 다만, 현재의 기술 발전 속도와 사회적 합의 수준을 고려할 때, 현행 법제가 모든 영역을 포괄적으로 규율하는 데에는 구조적 한계가 존재할 수밖에 없다. 이에 따라 법률 규율만으로는 충분하지 않은 영역에 대해서는 자율규제와 보완적 정책 수단으로서의 윤리 정책과 제도를 통해 대응할 필요가 있다. 특히 AI 법제도와 윤리 정책 간의 정합성과 연계성을 체계적으로 제고하려는 노력이 향후 AI 거버넌스의 효과성을 좌우하는 핵심과제가 될 것이다.

참고문헌

[국내 문헌]

개인정보보호위원회(2021). “인공지능(AI) 개인정보보호 자율점검표”.

과학기술정보통신부·정보통신정책연구원(2022). “2022 인공지능 윤리기준 실천을 위한 자율점검표”.

과학기술정보통신부·한국정보통신기술협회(2023). “2023 신뢰할 수 있는 인공지능 개발 안내서-의료분야”.

국가과학기술인력개발원(2021). “기관생명윤리위원회 업무 매뉴얼”.

김화(2023). “의료인공지능과 민사상 책임”, ≪생명윤리정책연구≫, 제17권 제1호, 이화여자대학교 생명의료법연구소, 1-32.

류기성·김일환(2025), “디지털 헬스케어에 관한 비교법적 고찰”, ≪법이론실무연구≫, 제13권 제2호, 한국법이론실무학회, 503-529.

머니투데이(2025.1.23.). “가천대 길병원, 루닛 AI 솔루션 도입…암 위험 큰 ‘치밀유방’ 정밀 분석”. URL: https://www.mt.co.kr/thebio/2025/01/23/2025012309331562061

메디게이트(2023.8.12.). “의사가 의료 AI 활용 망설이는 가장 큰 이유는 ‘책임성’…의료인 선택, 과실 판단 주요 기준”. URL: https://www.medigatenews.com/news/3592485391

문정욱·문아람·김정언·이시직·양기문·황선영·변순용·문명재·선지원·김형주·이청호·김봉제(2020). “윤리적 인공지능을 위한 국가정책 수립”. 《정책연구 20-21》. 정보통신정책연구원.

문정욱·조성은·문아람·이현경·문광진·양기문·김사혁·정선민·황선영·변순용(2021). 사람중심의 인공지능 구현을 위한 인공지능 윤리정책 개발. 《정책연구 21-24》, 정보통신정책연구원.

문정욱·문광진·이현경·양기문·황인성·안기창·성욱준·김법연·왕재선(2022). 인공지능 윤리체계 확립을 위한 정책연구.《정책연구 22-28-01》, 정보통신정책연구원.

문정욱·문광진·조성은·이현경·양기문·김지혜·황인성·안기창·강하연·유지연·홍은영(2023). 인공지능 윤리 의식 확산을 위한 정책연구.《정책연구 23-24-01》. 정보통신정책연구원.

문정욱·조성은·연소라·문광진·이현경·양기문·김지혜·강하연·김소담·전민경·박천희·홍은영(2024). AI 윤리·신뢰성 확보를 위한 실천 방안 및 정책연구.《정책연구 24-18》. 정보통신정책연구원.

문정욱(2025). AI 윤리와 안전 확보를 위한 도전과 과제(토론문).《글로벌 AI 안전 생태계 주권확보를 위한 정책토론회 발표자료집》. AI 미래가치포럼.

보건복지부(2024). “의료 인공지능 연구개발(R&D) 로드맵(안)(’24~’28)”.

식품의약품안전처(2024). “2023년 의료기기 허가보고서”.

식품의약품안전처(2024). “의료기기의 사이버보안 허가·심사 가이드라인(민원인 안내서)”.

식품의약품안전처(2025). “생성형 인공지능 의료기기 허가·심사 가이드라인(민원인 안내서)”.

양승엽·노호창·문준혁(2024). 인공지능 채용 가이드라인(안) 개발.《연구보고 2024-02》. 한국노동연구원.

윤호상(2020). “의료데이터의 활용: 데이터 3법 개정 후의 쟁점”, ≪DAIG≫, 창간호, 서울대학교 인공지능정책 이니셔티브, 62-79.

이상용·이혜리(2024). “개인정보보호법상 자동화된 결정 조항의 해석”, ≪법조≫, 제73권 제1호(통권 763호), 법조협회, 217-247.

이재훈(2020). “〈특집〉 데이터 3법 개정에 따른 바이오·의료정보 활용방향과 시사점”, ≪BioINpro≫, Vol. 71, 생명공학정책연구센터, 1-19.

에이아이트릭스(2025.8.27.). “바이탈케어 도입 이후 코드 블루, 장기 입원 비율 감소 및 의료진의 조기 개입 빈도 증가”. URL: https://aitrics.com/kr/sub/media/

news.php?mode=view&bid=1&s_type=&s_keyword=&s_cate=&idx=1548&page=1

조선일보(2025.7.14.). “날개 단 K의료기기… 美서 원격 로봇 수술, AI 진단도”. URL: https://www.chosun.com/economy/science/2025/07/14/WBDY6YGQLJGBRFUIPGFDR6M4NE/

차민정(2021). “데이터 3법 시행에 따른 바이오산업 분야에 미치는 영향”, ≪BRIC View 2021-T15≫, 1-11.

최정윤·최신혜(2024). “미국의 보건의료 분야 인공지능 규제 체계 분석 – FDA의 의료기기 규제를 중심으로”, ≪법제연구≫, 제67호,, 325-374.

최선미·김경진(2022). “데이터 3법 기반 디지털 헬스케어 산업에서 안전한 데이터 활용에 관한 연구”, ≪한국융합학회논문지≫, 제13권 제4호, 25-37.

최호영(2025). “개인 건강정보(Health Data)의 웹 스크래핑 등을 통한 수집·활용에 대한 고찰: 국내 및 주요국 사례를 중심으로, ≪HIRA Research≫, 제53권 제2호, 110-129.

정보통신정책연구원(2024). ”인공지능서비스 이용자 보호를 위한 법제 도입 방안 연구”.

한경비즈니스(2024.6.29.). “일자리 얻으려면 ‘AI에게 점수 따는 법’ 익혀야 하는 시대”. URL: https://magazine.hankyung.com/business/article/202406197697b

한국경제(2024.10.28.). ““AI로 채용했다가 5억 날렸다”…소송 휘말리더니 ‘날벼락’”. URL: https://www.hankyung.com/article/202410259897g

한국경제인협회(2024.3.28.). “2024년 상반기 대기업 채용동향·인식 조사”. URL: https://www.fki.or.kr/kor/news/statement_detail.do?bbs_id=00035481&category=ST

헬스케어 특별위원회·관계부처 합동(2018.12.). “4차 산업혁명 기반 헬스케어 발전 전략”.

[해외 문헌]

EU COM(2020). *White Paper on Artificial Intelligence – A European appr oach to excellence and trust*.

EU COM(2022). *Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space.*

FDA(2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Softw are as a Medical Device(SaMD) Action Plan*.

FDA(2024). *Proposed Regulatory Framework for Modifications to AI/ ML-Based Software as a Medical Device*.

FDA(2025). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Soft ware Functions.*

Fortune Business Insights(2025.11.). *AI in Healthcare Market Size, Share & Industry Analysis, By Platform, By End-user, and Regional Forecasts, 2025-2032*, URL: https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-in-healthcare-market-100534

Gilbert, Stephen/Fenech, Matthew/Hirsch, Martin/Upadhyay, Shubhanan/Biasiucci, Andrea/Starlinger, Johannes(2021). *Algorithm Change Protocols in the Regulation of Adaptive Machine Learning-Based Medical Devices*. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8579211/

Intuitive Surgical(2025). *Intuitive Surgical 2024 Annual Report*, URL: https://isrg.intuitive.com/static-files/500ff989-ad91-4b32-a59e-f94a34d75997

Lunit(2024.12.2.). *Real-World Validation: Lunit AI Proven Successful in 1-Year Breast Cancer Screening Deployment*. URL: https://www.

lunit.io/en/company/news/real-world-validation-lunit-ai-proven-successful-in-1-year-breast-cancer-screening-deployment

MDCG/AIB(2025). Guidance on the interpla between the Medical Devices Regulation / In vitro Diagnostic Medical Devices Regulation and the Artificial Intelligence Act. URL: https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en

OECD/Eurostat/WHO(2017). *A System of Health Accounts 2011: Revised edition*.

OECD(2024). *AI in Health: Huge Potential, Huge Risks*.

Precedence Research(2025.5.). *AI in Medical Imaging Market Size, Share, and Trends 2025 to 2034*. URL: https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market

Radiology Business(2024.5.14.). FDA adds more than 120 new AI-enabled medical devices focused on radiology to list of approvals, URL: https://radiologybusiness.com/topics/artificial-intelligence/fda-adds-more-120-new-ai-enabled-medical-devices-focused-radiology-list-approvals

Salathé Marcel/Wiegand, Thomas/Wenzel, Markus(2018), *Focus Group on Artificial Intelligence for Health*. URL: https://doi.org/10.48550/arXiv.1809.04797

Straits Research(2025.3.). *AI Recruitment Market Size & Outlook,* 2025-2033. URL: https://straitsresearch.com/report/ai-recruitment-market

TechTarget(2023.2.15.). *Community Health Systems Impacted by Data Breach Tied to GoAnywhere MFT Vulnerability*. URL: https:// www.techtarget.com/healthtechsecurity/news/366594463/Community-Health-Systems-Impacted-by-Data-Breach-Tied-to-GoAnywhere-MFT-Vulnerability

TechTarget(2025.8.26.). *Public companies linked to 92% of AI medical device recalls*. URL: https://www.techtarget.com/healthtechanalytics/news/366630073/Public-companies-linked-to-92-of-AI-medical-device-recalls

The Guardians(2025.5.14.). "People interviewed by AI for jobs face discrimination risks, Australian study warns". URL: https://www.theguardian.com/australia-news/2025/may/14/people-interviewed-by-ai-for-jobs-face-discrimination-risks-australian-study-warns

Washington Post(2025.4.5.). *AI hasn't killed radiology, but it is changing it*, URL: https://www.washingtonpost.com/health/2025/04/05/ai-machine-learning-radiology-software/

WHO(2021). *Ethics & Governance of AI in Health*.

● 저 자 소 개 ●

문 정 욱

- 고려대학교 행정학 박사
- 현 정보통신정책연구원 디지털사회전략연구실 실장

조 성 은

- Rutgers University 커뮤니케이션학 박사
- 현 정보통신정책연구원 연구위원

김 휘 홍

- Ludwig Maximilian University of Munich 공법학 박사
- 현 정보통신정책연구원 부연구위원

이 현 경

- The George Washington University 정책학 박사
- 현 정보통신정책연구원 연구위원

연 소 라

- Texas A&M University 경제학 박사
- 현 정보통신정책연구원 부연구위원

양 기 문

- 연세대학교 정보시스템학 석사
- 현 정보통신정책연구원 전문연구원

김 지 혜

- 서강대학교 국제학 석사
- 현 정보통신정책연구원 전문연구원

강 하 연

- 충남대학교 정책학 석사수료
- 현 정보통신정책연구원 연구원

김 소 담

- 서울대학교 지능정보융합학 석사
- 현 정보통신정책연구원 연구원

권 지 혜

- 호서대학교 데이터사이언스학 석사
- 현 정보통신정책연구원 연구원

정책연구 25-26

AI 윤리 확보를 위한 실천 방안 및 정책연구

2025년 12월 일 인쇄
2025년 12월 일 발행

발행인 배경훈
발행처 과학기술정보통신부
세종 갈매로 477, 정부세종청사 4동
Homepage: www.msit.go.kr
인 쇄 (사)아름다운사람들