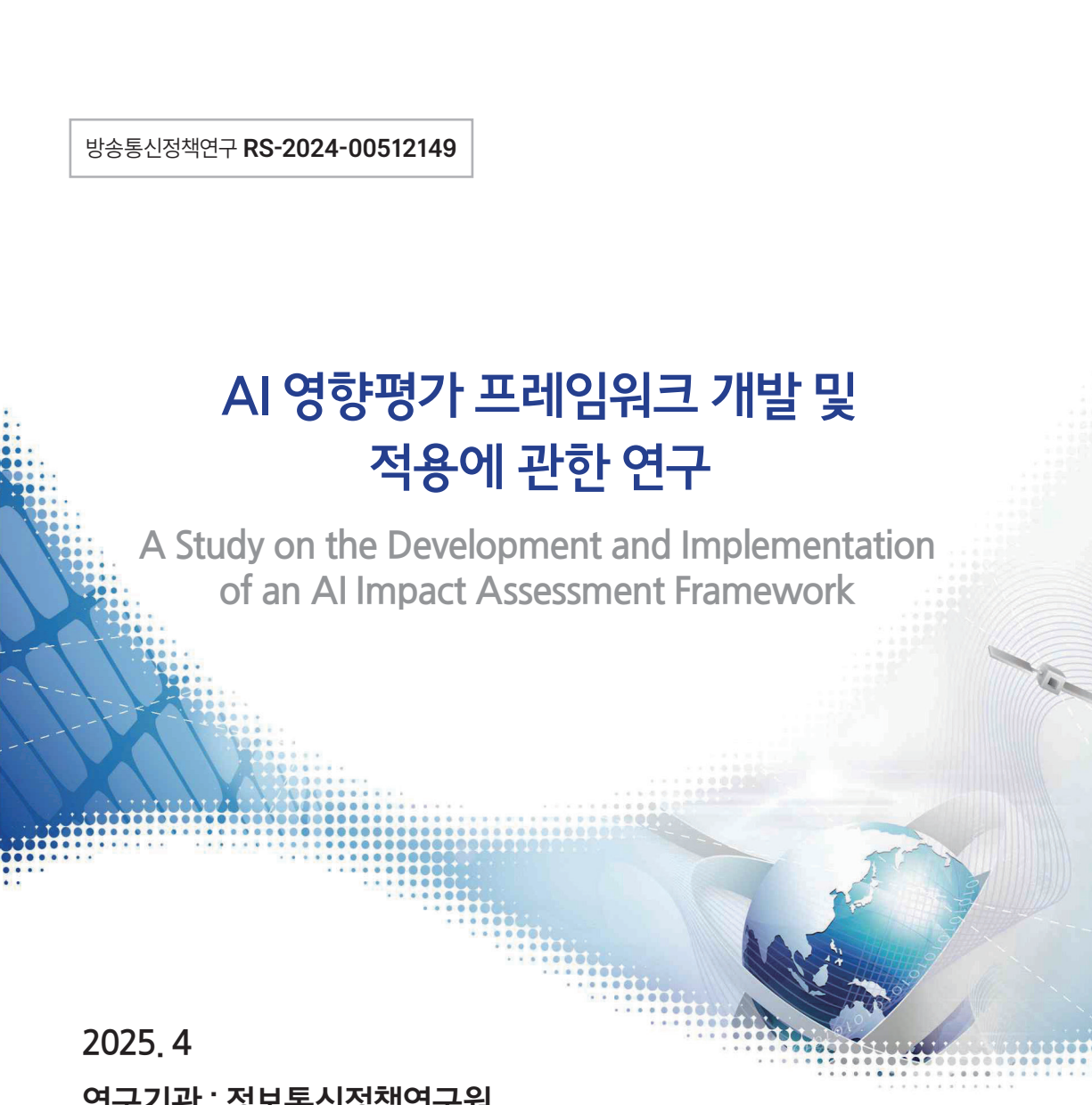
방송통신정책연구 RS-2024-00512149

AI 영향평가 프레임워크 개발 및 적용에 관한 연구

A Study on the Development and Implementation of an AI Impact Assessment Framework



2025. 4

연구기관 : 정보통신정책연구원

과학기술정보통신부 Ministry of Science and ICT

정보통신기획평가원 Institute of Information & Communications Technology Planning & Evaluation

방송통신정책연구 RS-2024-00512149

AI 영향평가 프레임워크 개발 및 적용에 관한 연구

(A Study on the Development and Implementation of an AI Impact Assessment Framework)

문정욱/양기문/강하연/김소담

2025. 4

연구기관 : **정보통신정책연구원**

과학기술정보통신부

정보통신기획평가원

이 보고서는 2024년도 과학기술정보통신부 방송통신발전기금 방송통신정책연구사업의 연구결과로서 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

* 보고서 공개여부 : 비공개
* 비공개 기간 : '25.05.01 ~ '26.04.30.(1년)
* 비공개 사유 : 정부 공식 발표 전 노출에 따른 정책 혼선 방지

제 출 문

과학기술정보통신부 장관 귀하

본 보고서를 『AI 영향평가 프레임워크 개발 및 적용에 관한 연구』의 연구결과보고서로 제출합니다.

2025년 4월

연 구 기 관 : 정보통신정책연구원
총괄책임자 : 문정욱
참여연구원 : 조성은 문아람
이현경 연소라
양기문 강하연
김소담 이시직
김지혜 박지원
문광진

목 차

표 목 차

그 림 목 차

요 약 문

1. 제 목

AI 영향평가 프레임워크 개발 및 적용에 관한 연구

2. 연구 목적 및 필요성

최근 인공지능(AI) 기술이 급속도로 발전하고 사회 전반에 걸쳐 활용되면서, AI가 우리 사회와 개인의 삶에 미치는 영향력이 점차 확대되고 있다. AI는 산업 혁신과 경제 성장을 촉진하고, 의료, 교통, 금융, 교육 등 다양한 분야에서 인간의 삶의 질을 향상시키는 긍정적 잠재력을 지니고 있다. 그러나 동시에 AI의 활용이 확대됨에 따라 개인정보 침해, 알고리즘 편향, 책임 소재의 불명확성, 인간 자율성 침해 등 부정적 영향에 대한 우려도 함께 증가하고 있다. 이러한 상황에서 AI의 혜택을 최대화하고 부작용을 최소화하기 위한 국제적인 노력이 활발히 전개되고 있다. 우리나라 역시 국제적 흐름에 발맞추어 다양한 정책을 수립하여 AI 기술의 발전과 올바른 활용을 지원하고 있다.

이처럼 국내외적으로 AI의 안전성, 신뢰성, 윤리성을 확보하기 위한 다양한 노력이 이루어지고 있으나, AI 기술의 사회적 영향력을 체계적으로 평가하고 관리하기 위한 구체적인 방법론이나 프레임워크는 아직 충분히 정립되지 않은 상황이다. 따라서 AI 기반 제품 및 서비스의 개발, 배포, 운영 등 전 과정에서 국민의 기본권 등에 미치는 영향을 식별하고, 이를 합리적으로 관리하기 위한 포괄적인 AI 영향평가 프레임워크의 개발이 필요한 시점이다. 2026년 1월 시행 예정인 「AI 기본법」 제35조에서 고영향 AI 제품 및 서비스에 대한 영향평가 권고 내용을 규율하고 있는 시점에서, 본 연구의 목적은 AI 기반 제품·서비스의 개발, 배포, 운영 등 전 과정에서 국민의 기본권에 미치는 영향을 체계적으로 식별하고 합리적으로 관리하기 위한 AI 영향평가 프레임워크(안)을 개발하고, 이를 실제로 적용하는 방안을 제시하는 데 있다. 특히, 「AI 기본법」 제35조에 명시된 고영향 AI 기본권 영향평가

의 주체, 대상, 범위, 평가 요소, 절차 등을 포괄하는 종합적인 평가 프레임워크를 개발함으로써 법 시행에 대비한 실질적 지침을 마련하고자 한다.

3. 연구의 구성 및 범위

본 연구의 범위는 크게 세 부분으로 구성된다. 첫째, 국내외 주요 AI 영향평가 사례를 분석하였다. 미국, EU, 유네스코 등에서 제시하는 영향평가 방법론의 목적, 주체, 대상, 방식을 심층 비교하고 국내 적용을 위한 시사점을 도출했으며, 국내에서 시행 중인 기술, 윤리, 사회적 영향평가 사례를 검토하고 포괄적 AI 영향평가 프레임워크 개발에 필요한 적용 가능성과 한계를 분석하였다. 둘째, AI 영향평가 가이드라인 마련을 위한 TF를 운영했으며, 사례 분석 및 TF 논의를 토대로 글로벌 정합성과 국내 특수성을 반영한 AI 영향평가 프레임워크(안)을 마련하였다. 셋째, AI 영향평가가 도입될 경우 예상되는 이슈와 쟁점을 분석하고 이에 대한 대응방안을 제안하였다.

4. 연구 내용 및 결과

본 연구는 「AI 기본법」 제35조에 따른 AI 영향평가 제도의 방향성과 실효적 적용 방안을 마련하기 위해 국내외 사례 분석, 실무 논의 참여, 프레임워크 설계 등을 종합적으로 수행하였다. 우선, 국내외 영향평가 유사 제도 검토를 통해, UNESCO의 윤리적 영향평가, 캐나다의 알고리즘 영향평가 등 다양한 사전적 평가 제도가 기본권 보호와 위험 예방을 목적으로 국제사회에서 활발히 운영되고 있음을 확인하였다. 이러한 분석을 통해 우리나라 AI 영향평가 제도 설계 시 고려해야 할 국제적 기준과 시사점을 도출하였다. 다음으로, 과학기술정보통신부 주관 AI 영향평가 가이드라인 TF를 구성 및 운영하여 실무 현장에서 요구되는 평가 방향, 주체, 절차, 주요 평가요소 등에 관한 의견을 수렴하고 반영하였다. 이를 바탕으로, 평가계획 수립부터 영향평가 수행, 결과 분석, 완화조치 마련, 문서화 및 보관에 이르는 세부 절차를 체계화하고, 「AI 기본법」 시행에 대비한 영향평가 프레임워크(안)을 제시하였다. 또한, AI 영향평가 제도의 현장 도입을 위한 과제도 함께 검토하였다. 고영향 AI 사업자 책무와 영향평가 간 중복성 문제, 정부기관 조달 시 '우선적 고려' 규정의 실효

성 한계, 영향평가 수행을 지원할 전담기관의 필요성 등을 주요 쟁점으로 식별하고, 이에 대한 정책적 대응방안을 제안하였다.

종합적으로, 본 연구는 AI 영향평가 제도가 단순한 규정 제정에 그치지 않고, 실제 현장에서 기본권 보호와 위험 예방이라는 본래 목적을 달성할 수 있도록 실효성을 확보하는 제도 설계 방향과 구체적 실천방안을 제시하였다는 데 의의를 가진다.

5. 정책적 활용 내용

현재 정부는 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」의 시행을 위한 하위 규정 마련을 적극적으로 추진하고 있다. 2025년 하반기 중 기본법 시행령 전문 공개 및 의견수렴 절차를 진행할 예정이며, 이를 통해 제도의 구체적 운영 방향을 확정할 계획이다. 다만, 시행령 조문만으로는 세부 이행방안과 구체적 절차를 모두 규정하는 데 한계가 있을 수 있다는 점을 고려하여, 정부는 시행령과 병행하여 고시와 가이드라인 등 하위 규범과 실무지침을 함께 마련하고 있다. 고시는 제도 운영의 법적 기반을 보완하고, 가이드라인은 사업자들이 실제 영향평가를 수행하는 과정에서 참고할 수 있도록 구체적 절차와 방법론을 제공하는 역할을 하게 된다. 본 연구를 통해 도출된 AI 영향평가 가이드라인(안)은 이러한 정부 정책 추진 방향과 긴밀히 연계된 것으로, 기본법 시행 이후 AI 영향평가 제도의 현장 안착과 실효성 확보를 지원하기 위한 구체적 기반 자료로 활용될 수 있을 것이다.

6. 기대효과

2026년 1월 시행 예정인 「AI 기본법」 제35조에서 고영향 AI 제품 및 서비스에 대한 영향평가 권고 내용을 규율하고 있는 시점에서, 본 연구의 AI 영향평가 프레임워크는 법적 기반을 갖춘 실질적인 평가 도구로서 인공지능 사업자에 실무적 지침을 제공하고 국내 AI 생태계의 건전한 발전을 지원하는 핵심 역할을 할 것으로 기대된다.

SUMMARY

1. Title

A Study on the Development and Implementation of an AI Impact Assessment Framework

2. Objective and Importance of Research

International efforts to maximize the benefits of artificial intelligence (AI) while minimizing its risks are actively underway. Korea is also aligning with this global trend by implementing various policies to promote the responsible development and use of AI technologies. Despite these efforts, there remains a lack of concrete methodologies or frameworks for systematically assessing and managing the societal impacts of AI. Therefore, it is crucial to develop a comprehensive AI Impact Assessment (AIIA) framework that can identify and manage the effects of AI-based products and services on fundamental rights throughout their entire lifecycle, from development to deployment and operation. In this context, as the AI Basic Act—mandating the recommendation of impact assessments for high-impact AI products and services—comes into force in January 2026, this study aims to develop a draft AI Impact Assessment framework. The framework will systematically define the subjects, scope, evaluation elements, and procedures necessary to assess and manage AI's potential impacts on fundamental rights, providing a practical guide in preparation for the Act's implementation.

3. Contents and Scope of the Research

The scope of this study consists of three main parts. First, major domestic and international AI impact assessment cases were analyzed. The study conducted an in-depth comparison of the objectives, entities, subjects, and methodologies of AI impact assessment frameworks proposed by the United States, the European Union, and UNESCO, and derived implications for application in Korea. It also reviewed domestic cases of technology, ethics, and social impact assessments to analyze their applicability and limitations for developing a comprehensive AI impact assessment framework. Second, a task force (TF) was organized to develop an AI Impact Assessment Guideline. Based on the results of the case analysis and TF discussions, a draft AI Impact Assessment Framework was developed, reflecting both global alignment and Korea's specific context. Third, the study identified potential issues and challenges that could arise when AI impact assessments are implemented and proposed corresponding response measures.

4. Research Results

This study was conducted to establish the direction and practical implementation strategies for the AI Impact Assessment (AIIA) system under Article 35 of the AI Basic Act. First, by analyzing domestic and international cases, including UNESCO's Ethical Impact Assessment and Canada's Algorithmic Impact Assessment, the study identified global practices emphasizing fundamental rights protection and risk prevention, and derived implications for Korea's system design. Second, by participating in the AI Impact Assessment Guideline Task Force (TF) organized by the Ministry of Science and ICT, the study incorporated practical needs into the framework, including evaluation direction, actors, procedures, and key assessment elements. Based on these discussions, a draft AIIA framework was proposed, outlining detailed procedures from planning and assessment to analysis, mitigation measures, and documentation, in preparation for the enforcement of

the AI Basic Act. Additionally, the study examined key challenges for field implementation, including the potential overlap between the duties of high-impact AI operators and the AIIA requirements, the limitations of the "preferential consideration" rule for government procurement, and the need for dedicated support institutions. Overall, the study emphasizes that the AI Impact Assessment system must not remain a formal requirement but must be designed and operated to meaningfully protect fundamental rights and prevent risks in practice.

5. Policy Suggestions for Practical Use

The government is actively preparing subordinate regulations to support the implementation of the AI Basic Act. In the second half of 2025, the government plans to release the full draft of the Act's Enforcement Decree for public consultation and to finalize the detailed operational framework based on the feedback received. The AI Impact Assessment Guideline will serve as a practical reference for businesses, providing detailed procedures and methodologies for conducting impact assessments. The draft AI Impact Assessment Guideline developed through this study is closely aligned with the government's policy direction and is expected to serve as a foundational resource to support the effective implementation and field adoption of the AI Impact Assessment system after the enforcement of the AI Basic Act.

6. Expectations

As Article 35 of the AI Basic Act, which recommends impact assessments for high-impact AI products and services, is scheduled to take effect in January 2026, the AI Impact Assessment Framework developed in this study is expected to serve as a practical tool with a legal foundation. It will provide operational guidance to AI businesses and play a key role in supporting the healthy development of the domestic AI ecosystem.

CONTENTS

제1장 서 론

제1절 연구의 필요성 및 목적

최근 인공지능(AI) 기술이 급속도로 발전하고 사회 전반에 걸쳐 활용되면서, AI가 우리 사회와 개인의 삶에 미치는 영향력이 점차 확대되고 있다. AI는 산업 혁신과 경제 성장을 촉진하고, 의료, 교통, 금융, 교육 등 다양한 분야에서 인간의 삶의 질을 향상시키는 긍정적 잠재력을 지니고 있다. 그러나 동시에 AI의 활용이 확대됨에 따라 개인정보 침해, 알고리즘 편향, 책임 소재의 불명확성, 인간 자율성 침해 등 부정적 영향에 대한 우려도 함께 증가하고 있다.

이러한 상황에서 AI의 혜택을 최대화하고 부작용을 최소화하기 위한 국제적인 노력이 활발히 전개되고 있다. 미국은 예산관리처(OMB)의 AI 사용에 대한 거버넌스, 혁신 및 리스크 관리 강화 공문을 통해 'AI 영향평가(AI Impact Assessment)'를 안전 및 권리에 영향을 미치는 AI 시스템에 대한 권고 조치로 규정했으며, 유럽연합(EU)은 2024년 7월 EU AI법(EU AI Act)을 통해 고위험 AI 시스템 배포자에게 '기본권 영향평가(Fundamental Rights Impact Assessment)'와 '적합성 평가(Conformity Assessment)' 실시를 의무화했다. 또한 유네스코(UNESCO)는 2021년 11월 '인공지능 윤리 권고(Recommendation on the Ethics of Artificial Intelligence)'를 발표하고, 회원국들에게 '윤리영향평가(Ethical Impact Assessment)' 도입을 권고하는 등 AI의 영향력을 체계적으로 평가하고 관리하기 위한 다양한 방법론이 제시되고 있다.

우리나라 역시 이러한 국제적 흐름에 발맞추어 '인공지능 국가전략'(2019.12), '인공지능 윤리기준'(2020.12), '신뢰할 수 있는 인공지능 실현 전략'(2021.5), '대한민국 디지털 전략'(2022.9), '전국민 AI 일상화 실행 계획'(2023.9), '디지털 권리장전'(2023.9), '인공지능 윤리·신뢰성 확보 추진계획'(2023.10), 'AI·디지털 혁신성장 전략'(2024.4), 'AI·반도체 이니셔티브'(2024.4) 등 다양한 정책을 수립하여 AI 기술의 발전과 올바른 활용을 지원하고 있다.

특히 2024년 5월 'AI 서울 정상회의'에서 채택된 '안전하고 혁신적이며 포용적인 AI를 위한 서울 선언'과 'AI 안전 과학에 대한 국제 협력을 위한 서울 의향서', 그리고 제1차 국가인공지능위원회에서 발표된 '국가 AI 전략 정책 방향'(2024.9)은 모두 AI 발전과 안전·신뢰의 균형 있는 달성을 강조하고 있다. 한편 2025년 1월 21일 공포된 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 AI기본법)'은 인공지능에 관한 국가 차원의 협력 체계를 구축하고 인공지능 산업을 체계적으로 육성하며, 인공지능으로 발생할 수 있는 잠재적 위험을 사전에 예방하는 것을 골자로 한다.

이처럼 국내외적으로 AI의 안전성, 신뢰성, 윤리성을 확보하기 위한 다양한 노력이 이루어지고 있으나, AI 기술의 사회적 영향력을 체계적으로 평가하고 관리하기 위한 구체적인 방법론이나 프레임워크는 아직 충분히 정립되지 않은 상황이다. 국내에서는 과학기술정보통신부의 '기술영향평가'와 'AI 윤리영향평가', 국가인권위원회의 'AI 인권영향평가' 등 다양한 영향평가가 시행되고 있으나, 이들은 각각 특정 영역에 초점을 맞추고 있어 AI가 사회 전반에 미치는 포괄적인 영향을 평가하기에는 한계가 있다.

따라서 AI 기반 제품 및 서비스의 개발, 배포, 운영 등 전 과정에서 국민의 기본권 등에 미치는 영향을 식별하고, 이를 합리적으로 관리하기 위한 포괄적인 AI 영향평가 프레임워크의 개발이 필요한 시점이다. 이러한 프레임워크는 AI의 윤리, 신뢰성, 안전성을 확보하면서도 기술 혁신을 저해하지 않는 균형 잡힌 접근 방식을 제공함으로써, 개발자, 제공자, 이용자, 정부, 공공기관 등 AI 관련 주체들이 AI를 안전하고 윤리적으로 개발·활용할 수 있는 기반을 마련할 것으로 기대된다. 또한 2026년 1월 시행 예정인 「AI 기본법」 제35조에서 고영향 AI 제품 및 서비스에 대한 영향평가 권고 내용을 규율하고 있는 시점에서, 본 연구의 AI 영향평가 프레임워크는 법적 기반을 갖춘 실질적인 평가 도구로서 인공지능 사업자에 실무적 지침을 제공하고 국내 AI 생태계의 건전한 발전을 지원하는 핵심 역할을 할 것으로 기대된다.

본 연구의 목적은 AI 기반 제품·서비스의 개발, 배포, 운영 등 전 과정에서 국민의 기본권에 미치는 영향을 체계적으로 식별하고 합리적으로 관리하기 위한 AI 영향평가 프레임워크(안)을 개발하고, 이를 실제로 적용하는 방안을 제시하는 데 있다. 특히, AI 기본법 제35조에 명시된 고영향 AI 기본권 영향평가의 주체, 대상, 범위, 평가 요소, 절차 등을 포괄

하는 종합적인 평가 프레임워크를 개발함으로써 법 시행에 대비한 실질적 지침을 마련하고자 한다.

이를 위해 국내외 주요 AI 영향평가 사례를 분석하고 비교함으로써, 글로벌 정합성과 우리나라의 특수성을 고려한 시사점을 도출하고, 이를 바탕으로 포괄적인 AI 영향평가 프레임워크(안)을 개발하여, AI 기반 제품·서비스가 사회에 미치는 다양한 영향을 체계적으로 식별하고 평가할 수 있는 기준을 제시하고자 한다. 또한 개발된 프레임워크를 실제 AI 기반 제품·서비스에 적용할 경우 발생할 수 있는 이슈와 쟁점을 식별하고, 이에 대한 대응 방안을 제시함으로써 AI 발전과 안전·신뢰의 균형 있는 달성을 위한 정책적·제도적 기반을 마련하고자 한다.

제 2 절 연구 추진체계와 전략

본 연구의 범위는 크게 세 부분으로 나눌 수 있다. 첫째, 해외 주요 AI 영향평가 사례분석 및 검토로, 미국, EU, 유네스코 등 국제 사회에서 주요하게 제시하고 있는 AI 영향평가 방법론을 대상으로 목적, 주체, 대상, 방식 등을 심층 분석하고 비교하며, 국내 적용을 위한 사례별 시사점을 도출한다. 둘째, 우리나라 AI 영향평가 사례분석 및 검토로, 국내에서 시행 중인 기술, 윤리, 사회적 영향평가 방법론을 분석하고 비교하며, 포괄적 평가 프레임워크 개발 시 적용을 위한 시사점을 도출한다. 셋째, 우리나라 AI 영향평가 프레임워크(안) 개발 및 도입 방안 연구로, 국내와 해외의 AI 영향평가 주요 사례분석을 통해 글로벌 정합성과 우리나라 특수성을 아우르는 포괄적 AI 영향평가 프레임워크(안)을 개발하고, 평가대상, 범위, 주체, 지표, 평가 결과에 따른 환류 등을 포함하며, 실제 AI 기반 제품·서비스에 AI 영향평가 프레임워크를 적용할 경우 발생할 수 있는 이슈와 쟁점을 식별하고 대응 방안을 제시한다.

본 연구는 문헌 연구, 사례 분석, 연구반 운영, 전문가 의견 수렴, 프레임워크 개발 등의 방법을 통해 진행된다. 문헌 연구를 통해 국내외 AI 영향평가 관련 학술 논문, 정책 보고서, 법령 및 가이드라인 등을 수집·분석하고, 사례 분석을 통해 미국, EU, 유네스코 등 국제기구 및 해외 주요국의 AI 영향평가 사례와 국내 AI 영향평가 및 유사 영향평가 사례를

분석한다. 또한 AI 기술, 법, 정책, 행정, 윤리, 사회학 등 다양한 분야의 전문가로 구성된 연구반 운영과 산·학·연·관 전문가 의견 수렴을 진행하여 개발된 프레임워크의 적용성, 실효성, 타당성 등을 제고하고자 한다.

〔그림 1-1〕 본 연구의 추진 체계

자료: 연구진 작성

본 연구는 서론을 포함하여 국내외 AI 영향평가 사례 분석, AI 영향평가 프레임워크 개발, 결론 및 향후 연구방향으로 구성되어 있다. 이를 통해 AI 기술의 급속한 발전과 확산에 따라 증가하는 사회적 영향력과 기본권에 미치는 영향을 체계적으로 평가하고 관리할 수 있는 프레임워크를 개발함으로써, AI의 안전하고 윤리적인 활용을 촉진하고 국민의 기본권을 보호하는 데 기여하고자 한다.

제 2 장 국내외 AI 영향평가 제도 동향

제 1 절 해외 AI 영향평가 유사 제도 동향[1]

1. EU「기본권 영향평가」(Fundamental Rights Impact Assessment)

EU의 인공지능법(EU AI Act) 제27조 '고위험 인공지능에 대한 기본권 영향평가'는 기본권[2]이 효율적으로 보호될 수 있도록 특정 고위험 인공지능 시스템 배포자가 해당 시스템을 사용하기 전 기본권 영향평가를 수행하도록 규정한다. 이는 고위험 인공지능 시스템 배포자가 영향을 받을 가능성이 있는 개인 또는 집단의 권리에 대한 구체적 위험을 식별하고, 이러한 위험이 구체화될 경우 조치하기 위한 목적에서 고안되었다.

구체적인 평가 대상은 ①〈표 2-1〉의 정의를 따르는 '고위험 인공지능 시스템'[3]을 배포하는 정부기관 또는 공공 서비스 제공 민간단체, ②부속서 Ⅲ의 5(b)(c)[4]에 해당하는 모든 고위험 인공지능 시스템 배포자로 규정하고 있다. 의사결정 과정에서 실질적인 영향을 미치지 않는 경우를 포함하여 개인의 건강, 안전 또는 기본권에 중대한 위험을 초래하지 않는 경우 고위험으로 간주하지 않는다(제6조 제3항). 모든 고위험 인공지능 시스템 배포자는 해당 시스템의 사용으로 인해 발생할 수 있는 기본권 영향평가 수행이 요구된다.

---

1) 본 절은 과학기술정보통신부·정보통신정책연구원(2024), 양기문(2024)을 바탕으로 재구성

2) 기본권(fundamental rights) : EU 기본권 헌장에 명시된 존엄성, 자유, 평등, 연대, 시민권(예:투표권, 청원권, 연방문서 열람권), 정의 등을 포함하여 EU 법적 체계 내에서 보장되는 권리

3) EU 인공지능법 부속서 Ⅲ(고위험 AI 시스템) 목록에 해당하는 AI 시스템. 단, 중요 디지털 인프라 관리 및 운영, 도로, 교통 또는 수도, 가스, 난방 또는 전기 공급에서의 안전 구성 요소로 사용되도록 의도된 AI 시스템은 평가 대상에서 제외

4) 5(b) : 개인 신용도 평가 및 신용점수 설정을 위해 사용되는 AI 시스템(예: AI 기반 종합신용평가 시스템, 대출승인 시스템 등) / 5(c) : 생명·건강 보험 관련 개인에 대한 위험평가와 보험료 책정을 위해 사용되는 AI 시스템(예: AI 기반 건강 기록 분석 시스템, 보험인수 심사 시스템 등)

〈표 2-1〉 EU 인공지능법 부속서 Ⅲ에서의 고위험 인공지능 시스템

다음 영역 중 하나에 어느 하나에 해당하는 인공지능 시스템은 고위험 인공지능 시스템에 해당	
1	관련 법에 따라 사용이 허용되는 생체 인식
2	중요 인프라(디지털 인프라, 도로·교통, 수도·가스·난방 또는 전기 공급 등)
3	교육 및 직업 훈련
4	고용, 근로자 관리 및 자영업에 대한 접근(기회)
5	필수 민간·공공 서비스와 혜택에 대한 접근 및 향유(건강·보험 등)
6	관련 법에 따라 사용이 허용되는 법 집행
7	관련 법에 따라 사용이 허용되는 한도 내에서의 이민, 망명 및 국경 통제 관리
8	사법 행정과 민주적 절차

자료: EU(2024)

평가는 인공지능 시스템 최초 사용 시 수행해야 하며, 평가 요소에 관한 사항이 변경되었거나 최신 상태가 아니라고 판단될 경우, 평가 정보 업데이트가 필요하다. 반면, 평가 요소가 유사한 경우에 배포자는 이전에 수행된 기본권 영향평가 또는 제공자가 수행한 기존 영향평가를 활용할 수 있다. 평가의 방법으로는 아래의 〈표 2-2〉에서 제시하는 여섯 가지 사항을 포함해야 함을 명시한다.

〈표 2-2〉 EU 인공지능법 제27조에서의 기본권 영향평가 필수 평가 사항

다음 사항을 포함한 기본권 영향평가 수행	
a	대상 AI 시스템이 의도된 목적에 맞게 사용될 절차 설명
b	대상 AI 시스템의 사용 기간 및 빈도 설명
c	특정 상황에서의 사용으로 인해 영향을 받을 가능성이 있는 개인(자연인) 또는 집단의 범주
d	식별된 대상에 영향을 미칠 가능성이 있는 구체적인 위험 요소
e	사용 지침에 따른 인적 감독 조치 이행에 대한 설명
f	위험이 현실화될 경우 취할 조치(내부 거버넌스 및 이의 처리 체계를 포함)

자료: EU(2024)

평가 이후 절차상의 의무는 다음과 같다. 우선 시스템 배포자는 기본권 영향평가 결과를 지정된 양식[5]에 따라 시장 감시기관에 통보해야 한다. 또한, 제49조 제3항과 제4항에

따라 공공기관, 기구 또는 단체를 대표하거나 대신하여 활동하는 배포자는 기본권 영향평가 결과 요약을 포함한 데이터(부속서 Ⅷ의 C절)를 EU 데이터베이스[6]에 입력하고 이후 최신의 상태를 유지해야 한다. 요구되는 정보의 목록은 〈표 2-3〉을 통해 확인할 수 있다.

〈표 2-3〉 EU 인공지능법 부속서 Ⅷ의 EU 데이터베이스 등록 정보 목록

부속서 Ⅷ : 제49조에 따른 고위험 인공지능 시스템 등록 시 제출해야 할 정보
C절 - 제49조 3항에 따라 고위험 인공지능 시스템 배포자가 제출해야 하는 정보

1	시스템 배포자의 이름, 주소 및 연락처 정보
2	시스템 배포자를 대신하여 정보를 제출하는 사람의 이름, 주소 및 연락처 정보
3	EU 데이터베이스에 인공지능 시스템을 입력한 제공자의 URL
4	제27조에 따라 실시한 기본권 영향평가 결과 요약
5	제26조 제8항에 명시된 규정 (EU) 2016/679 제35조 또는 지침 (EU) 2016/680 제27조에 따라 수행된 데이터 보호 영향 평가의 요약 (해당되는 경우)

자료: EU(2024)

2. UNESCO 「윤리영향평가」(Ethical Impact Assessment)

UNESCO는 2021년 11월 '인공지능 윤리 권고(Recommendation on the Ethics of Artificial Intelligence)'를 발표하고, 회원국들에게 윤리영향평가 도입을 권고하였다. 윤리영향평가는 인공지능 시스템의 이익과 우려, 적절한 위험 예방·완화 및 모니터링 조치를 식별하고 평가하기 위한 목적으로 개발되었으며, 이후 2023년 8월 윤리영향평가의 실질적 운영을 위한 도구[7]를 공개하였다. 해당 도구는 주로 인공지능 시스템 조달에 관여하는 정부 공무원(개인 및 팀)을 돕기 위해 설계되었으며, 구매하는 인공지능 시스템이 UNESCO 인공지능 윤리 권고에 명시된 윤리적 기준에 부합하는지 확인할 수 있는 일련의 질문을 제공한다. 단, 평가 도구의 활용과 관련하여 조달자 또는 조달 부서가 인공 지능 시스템의 설계 및

5) EU 인공지능법 제27조 제5항에 따라, AI 사무국(AI Office)은 대상 시스템 배포자가 기본권 영향평가 의무를 간소화하여 준수할 수 있도록 질문지 양식을 개발해야 함

6) EU 인공지능법 제71조 '고위험 AI 시스템에 대한 EU 데이터베이스'

7) UNESCO(2023). "Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence" 보고서를 통해 발표

개발에 직접 관여하지 않아 평가 문항에 적절하게 응답할 수 없는 경우가 많기 때문에, UNESCO는 시스템 조달자가 평가 과정에서 제공업체 및 기타 개인이 참여하는 공개 논의를 진행할 것을 필수적으로 제안하고 있다. 또한 각 국가의 규제 체제 및 특정 상황과 맥락에 맞게 영향평가 도구를 조정하여 사용하도록 권장한다.

윤리영향평가의 절차는 두 단계로 구성되며, ① 본격적인 평가에 앞서 평가 대상과 범위, 계획 등을 미리 정하는 범주화 질문 (scoping question) 단계, ② 평가 대상이 UNESCO 인공지능 윤리 권고에 어떻게 부합하는지 심층적으로 검토하는 원칙 이행(implementing) 평가 단계로 이루어진다. 평가의 체계도는 다음의 〈표 2-4〉와 같다.

〈표 2-4〉 인공지능 윤리영향평가 체계도

범주화 질문 Scoping Question		원칙 이행 Implementing the UNESCO Principles
1. 프로젝트 설명		**5. 안전, 보안 원칙**
1.1. 시스템에 대한 설명		5.1. 절차상 평가
1.2. 종속성		5.2. 긍·부정 영향 식별 및 완화
2. 비례성 선별 및 위해 금지		**6. 공정성, 차별금지, 다양성 원칙**
2.1. 비례성(과잉 금지)		**7. 지속가능성 원칙**
2.2. 위해 금지	➜	**8. 프라이버시, 데이터 보호 원칙**
3. 프로젝트 거버넌스 (역할과 책임 설정)		**9. 인간의 감독 및 결정 원칙**
3.1. 역할 및 책임		**10. 투명성·설명가능성, 책임성·책무성 원칙**
4. 이해관계자 거버넌스		**11. 인식, 리터러시 원칙**
4.1. 이해관계자 참여계획		※ 6~11 각각의 원칙 관련하여서도 5와 동일하게 절차상 평가 및 긍·부정 영향 식별 및 완화 평가 진행

자료: UNESCO(2023)를 토대로 연구진 정리

범주화 질문 단계를 통해 프로젝트가 UNESCO 권고사항에 어떻게 부합하는지 심층적으로 평가하기에 앞서, 조달 또는 개발 초기 단계에서 강력한 윤리적 기반을 확립하는 것이 중요하기에 프로젝트팀이 이를 수행할 수 있도록 네 가지 핵심 범주화 질문을 제시한다.

우선, 프로젝트 설명 항목은 아래의 〈표 2-5〉와 같다. 프로젝트팀은 구현하려는 시스템의 세부 내용과 목표를 명확히 기술해야 하며, 이는 시스템의 특성과 영향평가 범위를 구

체화하는 데 기여한다. 또한, 계획이 실제 사회적 문제나 도전 과제와 어떻게 연결되는지를 점검하는 데에도 도움이 된다.

〈표 2-5〉 범주화 질문 단계 (1) 프로젝트 설명

시스템에 대한 설명	• 시스템에 대한 초기 설명 • 시스템 목표·목적 및 해결하려는 문제 설명 • 프로젝트 수명주기를 준거로 현재 상태 설명 • 시스템 특성(상호작용 대상 및 선택권, 적용 분야, 시스템이 사용될 비즈니스 기능, 핵심 기능 및 활동에 대한 영향, 배포 범위 등) 명시
종속성	• 기존 프로젝트 확정·적용 여부 • 특정 목적·목표를 위해 개발되었는지 기성 모델(BERT, ChatGPT 등) 기반으로 구축되었는지 여부 • 그 외 윤리적 영향평가에 영향을 미칠 수 있는 대상(직접 개발하지 않은 모델이나 직접 사용하지 않은 데이터 : 특정 기계학습 패키지, 사전 훈련된 모델 등)에 대한 시스템 종속성 나열

자료: UNESCO(2023)를 토대로 연구진 정리

그 다음의 비례성 검토 항목은 〈표 2-6〉에서 자세히 확인할 수 있다. 시스템은 사회적 점수화(social scoring) 또는 대규모 감시를 위해 사용되어서는 안되며, 명시된 목표를 달성하는 데 비례하는 수준에서 AI 사용을 정당화할 수 있다.

〈표 2-6〉 범주화 질문 단계 (2) 비례성 검토

비례성	• 목표 달성을 위한 비알고리즘적 옵션에 대한 고려 • 계산적으로 더 단순한 접근 방식을 포함하여 다양한 방식 고려 • 명시된 목적에 비례하도록(과잉되지 않도록) 프로젝트 범위에 대한 제한 설정 • 시스템이 명시된 목적을 달성하는 데 얼마나 효과적이었는지 검토
위해금지	• 시스템을 사회적 점수화 또는 대규모 감시에 사용하거나 다른 행위자들에 의해 적용될 수 있는지를 검토하고 이를 방지하기 위한 조치 마련 • 예상되는 영향이 불가역적이거나 되돌리기 어렵거나 생사에 관련된 결정을 포함하는지 검토 • 시스템과 그 응용이 인권(인간의 존엄성, 표현의 자유, 공정한 재판)에 미치는지 검토

자료: UNESCO(2023)를 토대로 연구진 정리

세 번째, 프로젝트 거버넌스 항목은 〈표 2-7〉에서 자세히 제시되어 있다. 효과적인 영향평가를 위해서는 초기 단계부터 역할과 책임을 명확히 정의하는 것이 중요하며, 프로젝트팀의 다양성 확보는 최종 결과물 품질에 영향을 미치므로 반드시 고려되어야 한다.

〈표 2-7〉 범주화 질문 단계 (3) 프로젝트 거버넌스

역할 및 책임	• 시스템을 담당하는 프로젝트팀 내에서 최종 결정 권한을 가진 사람은 누구인지 설명 • 프로젝트 내 주요 업무 흐름에 대한 책임자 설명(제3자 또는 외부 조직 대표자 포함) • 프로젝트팀의 다양성에 대한 고려 • 개인적 혹은 팀으로서 누락될 수 있는 관점에 대한 위치성(positionality) 재검토

자료: UNESCO(2023)를 토대로 연구진 정리

네 번째, 이해관계자 거버넌스 항목은 영향평가에 다양한 관점이 반영될 수 있도록 사전에 계획을 수립하는 과정을 다룬다. 특히, 시스템의 영향을 직접 받게 될 이들의 관점과 경험, 의견을 평가 과정에 통합하는 것이 필요함을 강조한다.

〈표 2-8〉 범주화 질문 단계 (4) 이해관계자 거버넌스

이해관계자 참여 계획	• 시스템 배포로 가장 영향을 받을 가능성이 높은 이해관계자 그룹 식별 • 시스템 개발, 배포 및 사용 과정에서 관여시키거나 상의할 이해관계자 그룹 선별 • 이해관계자를 참여시키는 목적 • 이해관계자 참여계획(맞춤화, 사용 가능 자원, 참여 형태, 타임라인 등)

자료: UNESCO(2023)를 토대로 연구진 정리

원칙 이행 단계에서는 조달하고자 하는 인공지능 시스템의 설계, 개발 및 배포의 전 과정이 윤리적 AI를 위한 UNESCO 원칙(① 안전, 보안, ② 공정성, 차별금지, 다양성, ③ 지속가능성, ④ 프라이버시, 데이터 보호, ⑤ 인간의 감독 및 결정, ⑥ 투명성·설명가능성, 책임성·책무성, ⑦ 인식, 리터러시 원칙)에 부합하는지를 평가한다.

우선, 절차상 평가 과정을 통해 각 원칙에 대한 권고사항을 이행하기 위한 절차적 안전장치가 충분히 마련되어 있는지를 평가하며 관련 세부 항목은 〈표 2-9〉에서 확인할 수 있다.

〈표 2−9〉 원칙 이행 단계 (1) 절차상 평가 문항 예시

[문항 예시] 6. 공정성, 차별금지, 다양성

이 섹션 전반에 걸쳐 특정 그룹을 대상으로 한 테스트에 관한 질문에 답할 때, 프로젝트 팀은 특히 인종, 피부색, 혈통, 성별, 연령, 언어, 종교, 정치적 견해, 국가 출신, 민족 출신, 사회적 출신, 경제적 또는 사회적 출신 상태, 장애 등을 고려해야 합니다. 또한, 여러 기준이 결합된 그룹에서 테스트가 수행되었는지, 즉 시스템이 교차성(intersectionality) 측면에서 테스트되었는지를 명시해 주세요.

6.2.1. 차별적 결과 방지 :

6.2.1.1. 알고리즘이 다양한 그룹을 대상으로 테스트 되었습니까?

6.2.1.1.1. 정확성(또는 사용된 다른 성능 지표) 측면에서 차이가 있었습니까? 이러한 차이가 있었다면 설명해 주세요.

6.2.1.1.2. 특정 그룹에 대해 차별적인 효과가 나타났습니까?

6.2.2. 데이터 품질 및 차별적 편향 방지 :

6.2.2.1. 데이터의 편향을 테스트하기 위한 프로세스가 마련되었습니까?

6.2.2.1.1. 데이터 내 사회적 및 역사적 편향을 방지하기 위한 분석을 수행했습니까?

6.2.2.1.2. 데이터가 균형 잡혀 있으며, 목표 최종 사용자 집단의 다양성을 반영하고 있습니까?

6.2.2.1.3. AI 시스템의 학습에 사용된 데이터와 실제 처리되는 데이터 간의 차이로 인해, 특정 그룹에 대해 차별적인 결과를 초래하거나 성능이 달라질 가능성이 있습니까?

6.2.2.1.4. 설계 과정에서 데이터 품질 문제를 해결하는 방안을 문서화하는 프로세스를 개발했습니까?

6.2.2.1.5. AI 시스템의 설계 및 개발 과정에서 편향이 발생할 가능성에 대해 AI 디자이너 및 개발자가 인식할 수 있도록 교육 및 인식 제고 이니셔티브를 마련했습니까?

6.2.3. 접근성 측면에서의 차별 방지 :

6.2.3.1. AI 시스템의 설계가 모든 사람이, 특히 소외된 그룹이 접근하고 상호작용할 수 있도록 구성되어 있습니까? 접근성 측면에서의 제한 사항이 있다면 구체적으로 명시해 주세요.

6.2.3.1.1. AI 시스템이 장애가 있는 사람들에게 사용 가능한지 평가했습니까? (예: 스크린 리더 지원, 이미지에 대한 대체 텍스트 제공, 색맹 친화적인 색상 팔레트 등)

6.2.3.1.2. AI 시스템이 경제적으로 불안정한 상황에 있는 사람들에게도 사용 가능한지 평가했습니까?

6.2.3.2. 공정성 원칙은 기술적 관점에서 어떻게 접근되었습니까? 예를 들어, AI 시스템이 어떤 기술적 공정성 개념을 기준으로 조정되었는지 구체적으로 설명할 수 있습니까? (예: 개별 공정성, 인구 통계적 균형, 기회 균등 등)

6.2.3.3. AI 시스템이 적용될 인구 집단은 누구입니까? 해당 인구 집단이 특히 소외된 그룹에 속합니까?

자료: UNESCO(2023)를 토대로 연구진 정리

이후 영향 식별 및 완화 과정을 통해 시스템 조달 및 배포로 발생할 수 있는 긍·부정 영향을 식별하고, 세부 지표를 평가한다. UNESCO 보고서의 2장 원칙 이행 단계에서 '안전 및 보안'을 시작으로 '인식, 리터러시 원칙'까지 총 7가지의 핵심 요소에 대한 평가 도구를 제공하는데, 각 과정마다 긍정적·부정적 영향력에 관한 측정이 반복적으로 제시되고 있다.

긍정적 영향과 부정적 영향의 세부 평가 요소는 아래의 〈표 2-10〉, 〈표 2-11〉과 같다. 긍정적 영향력에 대하여 〈표 2-10〉의 총 네 가지의 질문이 제시된다. 첫 번째는 긍정적 영향이 무엇인가에 관한 질문으로서 예를 들어 인공지능 시스템이 개인에 대한 악성 소프트웨어 및 해킹 위험을 식별하는 데 활용될 수 있음을 들 수 있다. 두 번째는 이러한 영향의 규모(scale)에 대한 전망으로서 매우 높음부터 경미한 수준으로 구분된다. 세 번째는 범위에 관한 것으로서 직접적·간접적인 수준, 그리고 기간(단기, 중기, 장기, 세대 간)에 따라 평가하게 된다. 마지막으로 긍정적인 결과가 발생할 가능성에 대한 평가도 병행한다. 부정적 영향의 경우에는 〈표 2-11〉과 같이 문제 발생 시 부정적 결과의 해결가능성에 대한 문항이 추가된다.

〈표 2-10〉 원칙 이행 단계 (2) 영향 식별 및 완화 (긍정)

윤리가치에 대한 시스템의 긍정적 영향	**예상되는 긍정적 결과의 규모 평가**	**예상되는 긍정적 결과의 범위 평가** • 영향을 받는 개인/그룹/단체에 대한 설명 (모든 살아있는 유기체 포함 가능) • 영향의 기간	**예상되는 긍정적 결과가 발생할 가능성 평가**
개방형 (정성) • 예: 인공지능 시스템은 개인에 대한 악성 소프트웨어(Malware) 및 해킹 위험을 식별하는 데 활용될 수 있음	**유의미한 수준** • 매우 높음 • 높음 • 중간 • 보통 / 경미한	**영향의 범위** 영향을 받는 대상 • 직접적 / 간접적 / 예기치 않은·의도하지 않은 영향 기간 • 단기 / 중기 / 장기 / 세대 간	**발생 가능성** • 낮음 • 중간 • 높음 • 매우 높음

자료: UNESCO(2023)를 토대로 연구진 정리

〈표 2-11〉 원칙 이행 단계 (2) 영향 식별 및 완화 (부정)

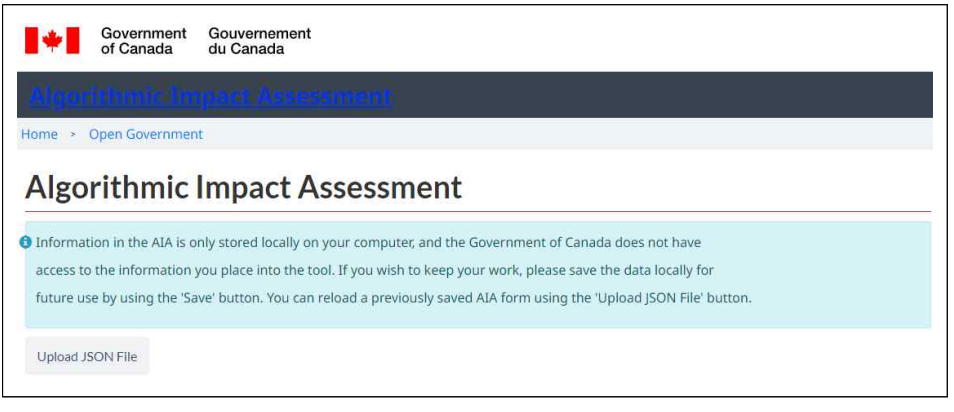
<table>
<tr><td>윤리가치에 대한 시스템의 부정적 영향</td><td>예상되는 부정적 결과의 규모 평가</td><td colspan="2">예상되는 부정적 결과의 범위 평가
• 영향을 받는 개인/그룹/단체에 대한 설명 (모든 살아있는 유기체 포함 가능)
• 영향의 기간</td><td>예상되는 부정적 결과가 발생할 가능성 평가</td></tr>
<tr><td>개방형 (정성)
• 예: 인공지능 시스템은 사용자의 심리적 안녕을 위협할 수 있는 허위 정보나 해악 또는 남용을 확산시키는 데 기여</td><td>중대한 수준
• 재앙적
• 치명적
• 심각한
• 보통 / 경미한</td><td colspan="2">영향의 범위
<u>영향을 받는 대상</u>
• 직접적 / 간접적 / 예기치 않은·의도하지 않은
<u>영향 기간</u>
• 단기 / 중기 / 장기 / 세대 간</td><td>발생 가능성
• 낮음
• 중간
• 높음
• 매우 높음</td></tr>
<tr><td colspan="2">예상되는 부정적 영향의 해결가능성 평가</td><td colspan="3">절차적 안전조치의 부정적 영향 완화 정도
추가적 완화 및 구제 전략 시행 필요성
• 시스템을 즉시 중지해야 할 극단적 경우 설명
• 조사 및 구제가 필요한 경우, 시기와 그 이유 설명
• 잠재적 피해가 반복되지 않기 위한 보장 방안</td></tr>
<tr><td colspan="2">해결가능 정도
• 매우 낮음
• 낮음
• 중간
• 높음</td><td colspan="3">개방형 (정성)</td></tr>
</table>

자료: UNESCO(2023)를 토대로 연구진 정리

3. 캐나다 「알고리즘 영향평가」(Algorithmic Impact Assessment)

알고리즘 영향평가(Algorithmic Impact Assessment, AIA)[8]'는 캐나다 재정위원회의 「자동화된 의사결정에 대한 지침」(〈표 2-12〉)을 지원하기 위한 필수 위험평가 도구로, 자동화된 의사결정 시스템의 영향 수준을 평가하고, 정부 부처와 기관이 해당 시스템과 관련된 위험을 더 잘 이해하고 관리할 수 있도록 설계하였다. 평가의 주체는 자동화된 의사결정 시스템을 개발·조달하는 연방정부 기관이다. 평가의 대상은 자동화된 의사결정 알고리즘으로, 특정 조직 단위에서 사용하는 시스템(예: 보건 시스템 내에서의 의사결정 알고리즘)으로 볼 수 있다. 〔그림 2-1〕의 웹페이지로 구성된 평가 시스템을 통해 각 점수를 산출하고, 이를 등급화하여 영향의 부정적 효과를 완화하고자 한다. 평가 요소는 위험에 대한 영향과 완화를 위한 요인을 복합적으로 고려한다.

〔그림 2-1〕 캐나다 '알고리즘 영향평가' 웹사이트



자료: Government of Canada, Algorithmic Impact Assessment[9]

8) 2019년 캐나다 재정위원회 (Treasury Board)에서 공공부문에 사용할 자동화된 의사결정의 요구사항을 명시한 훈령(Directive on Automated decision-making)에서 발표

9) https://open.canada.ca/aia-eia-js/?lang=en

〈표 2－12〉 캐나다 재정위원회의「자동화된 의사결정에 대한 지침」

자동화된 의사결정 시스템 사용 프로그램을 주무하는 부처 차관이 지명하는 사람 또는 차관보는 다음에 대해 책임져야 한다.
6.1. 알고리즘 영향평가
6.1.1. 자동화된 의사결정 시스템을 생산하기 전 알고리즘 영향평가를 완료하고 영향평가 최종결과 공개
6.1.2. 알고리즘 영향평가 결과에 따라 알고리즘 영향 수준별 요구사항(부록C)을 적용
6.1.3. 자동화된 의사결정 시스템의 기능 또는 범위가 변경되는 경우를 포함하여 알고리즘영향평가를 정기적으로 검토 및 업데이트
6.1.4. '열린정부 지침(Directive on Open Government)'에 따라 캐나다 재정위원회 사무국이 지정한 정부 웹사이트 및 기타 서비스를 통해 접근 가능한 형식으로 알고리즘 영향평가 최종결과 공개

자료: Government of Canada(2023)

알고리즘 영향평가는 시스템 설계, 알고리즘, 의사결정 유형 및 데이터를 비롯한 여러 평가 요소를 기반으로 ① 영역별 위험을 평가하는 동시에, ② 확인된 위험을 관리하기 위해 시행 중인 완화 조치를 평가하는 질문으로 구성되어 있다. 〈표 2－13〉을 살펴보면, 위험 영역은 프로젝트(project), 시스템(system), 알고리즘(algorithm), 의사결정(decision), 영향(impact), 데이터(data) 6개로 구성되며, 〈표 2－14〉의 완화 영역은 자문(consultation)과 위험성 제거 및 완화 조치(de-risking and mitigation measures) 2개로 구성된다.

〈표 2－15〉와 같이 총 8개 영역에 대하여 각 영역별 점수가 산출되며, 영향 수준[10]은 달성 가능한 최대 점수 대비 현재 점수의 백분율로 계산된다. 점수 백분율을 기준으로 네 단계[11]의 영향 수준을 정의하며, 도출된 영향 수준에 따라 요구되는 완화 조치 수준이 결정된다. 캐나다 영향평가 사례를 통해 알고리즘 시스템이 개인과 공동체에 대한 권리, 건강 또는 웰빙, 경제적 이익, 그리고 생태계의 지속가능성에 대한 광범위한 영역에서 위험을 사전에 식별하고 영향력을 평가할 수 있다. 영향이 거의 없거나 작은 의사결정은 가역적이고 지속 기간이 짧은 반면, 영향이 매우 큰 의사결정은 되돌릴 수 없고 영구적이다.

10) 영향 수준 = 현재 점수 ÷ 위험영역 최대 달성가능 영향점수 × 100

11) 1단계는 영향이 거의/전혀 없음으로 점수 백분율의 범위는 0~25%이며, 2단계는 중간 정도의 영향으로 26~50%, 3단계는 큰 영향으로 51~75%, 마지막으로 4단계는 매우 큰 영향으로 76% 이상의 백분율인 경우로 구분

〈표 2-13〉 알고리즘 영향평가 평가 요소 (1) 위험 영역

위험영역		설명
1. 프로젝트	프로젝트 단계	발주조직(Project owner), 설명, 단계(설계 또는 구현)
	자동화 이유	의사결정 프로세스에 자동화를 도입해야 하는 이유
	위험 프로파일	프로젝트에 대한 높은 수준의 위험 지표(예: 고객의 취약성)
	프로젝트 권한	프로젝트에 대한 새로운 정책 권한 모색 필요성
2. 시스템	시스템 정보	시스템의 역량(예: 이미지 인식, 위험평가)
3. 알고리즘	알고리즘 정보	알고리즘 공개에 대한 제한사항, 결과값 도출 방법에 대한 설명 능력
4. 의사결정	의사결정 정보	자동화된 결정의 분류 및 설명(예: 보건 서비스, 사회부조, 라이센싱 등)
5. 영향	영향평가	- 자동화 유형(전체 또는 부분) - 의사결정의 영향 지속성, 가역성 - 영향을 받는 분야(예: 권리, 개인정보보호와 자주성, 건강, 경제적 이익, 환경)
6. 데이터	출처	시스템에서 사용하는 데이터의 출처, 수집 방법 및 보안 분류
	타입	정형 또는 비정형으로 사용되는 데이터의 특성(오디오, 텍스트, 이미지 또는 비디오)

자료: Government of Canada(2024)

〈표 2-14〉 알고리즘 영향평가 평가 요소 (2) 완화 영역

완화영역		설명
1. 자문	이해관계자	개인정보보호 및 볍률 전문가, 디지털 정책 팀, 기타 분야의 전문가 등 내부 및 외부 이해관계자와의 의견수렴(consulted)
2. 위험성 제거 및 완화 조치	데이터 품질	시스템의 역량(예: 이미지 인식, 위험평가)
	절차적 공정성	데이터가 대표성을 지니며 편향적이지 않도록 보장하는 프로세스 및 해당 프로세스와 관련된 투명성 측정
	개인정보보호	시스템에서 사용되거나 생성된 개인정보를 보호하기 위한 조치

자료: Government of Canada(2024)

〈표 2－15〉 원시영향점수와 완화점수

영역		문항 수	최대 점수
위험영역	1. 프로젝트	16	27
	2. 시스템	1	-
	3. 알고리즘	2	6
	4. 의사결정	2	7
	5. 영향	20	42
	6. 데이터	10	44
	원시영향점수	51	126
완화영역	7. 자문	2	2
	8. 위험성 제거 및 완화 조치	32	44
	완화점수	34	46

주: 완화점수가 최대 달성 가능 완화점수의 80% 미만인 경우, 현재 점수(current score)는 원시영향점수와 동일하며, 80% 이상일 경우, 원시영향점수에서 15%를 차감하여 현재 점수를 산출
자료: Government of Canada(2024)

〈표 2－16〉의 영향 수준이 도출되면, 〈표 2－17〉과 같이 영향 수준에 따라 요구사항을 이행해야 한다. 위험이 거의 없는 1단계의 경우는 정부 차원의 규제나 개입이 요구되지 않지만, 2단계 이상의 경우에는 점진적으로 보다 엄격한 검토 절차가 부과된다. 예컨대, 위험 수준이 2단계 또는 3단계인 경우에는 교수, 연구자, 기술 공급자, 데이터 전문가 등 다양한 분야의 외부 전문가 중 최소 1인 이상의 자문을 받도록 규정되어 있다. 반면, 가장 높은 위험군인 4단계에서는 자문 인원이 2명 이상이어야 하며, 이 중에는 반드시 캐나다 국립연구위원회, 통계청, 통신보안기관 등 공신력 있는 기관으로부터 자격을 인정받은 전문가가 포함되어야 한다. 이러한 체계는 알고리즘이 초래할 수 있는 위험의 수준에 따라 차등적이고 정교한 관리가 필요하다는 점을 잘 보여준다. 이는 본 연구에서 다루는 AI 영향평가 또한 기술 혹은 서비스 영역의 위험도에 따라 서로 다른 평가 접근이 요구될 수 있음을 시사한다.

〈표 2-16〉 영향 수준

영향 수준	정의	점수 백분율 범위
I	거의 또는 전혀 영향 없음	0% ~ 25%
II	중간 정도의 영향	26% ~ 50%
III	큰 영향	51% ~ 75%
IV	매우 큰 영향	76% ~ 100%

자료: Government of Canada(2024)

〈표 2-17〉 영향 수준별 요구사항

요구사항	수준 I	수준 II	수준 III	수준 IV
① 피어리뷰 (section 6.3.5)	비해당	다음 전문가 중 최소 **1명** 이상에게 자문받고, 평가 결과에 대한 최종 검토 또는 이해하기 쉽게 작성한 풀이 요약을 캐나다 정부 웹사이트에 게시 : - 연방, 주, 준주 또는 시 정부기관에서 자격이 인증된 전문가 - 고등교육기관 자격을 갖춘 교수진 - 관련 비정부기구 소속 유자격 연구자 - 관련 전문성을 갖춘 서드파티 공급업체 - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재 - 재정위원회 사무국에서 지정한 데이터 및 자동화 자문기구 또는, - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재. 게시된 검토에 대한 접근이 제한되는 경우, 결과에 대한 평문 요약을 공개적으로 이용할 수 있는지 확인		다음 전문가 중 최소 **2명** 이상에게 자문받고, 평가 결과에 대한 최종 검토 또는 이해하기 쉽게 작성한 풀이 요약을 캐나다 정부 웹사이트에 게시 : - **캐나다 국립연구위원회, 통계청, 통신보안 기구에서** 자격이 인증된 전문가의 검토 - 고등교육기관 자격을 갖춘 교수진 - 관련 비정부기구 소속 유자격 연구자 - 관련 전문성을 갖춘 서드파티 공급업체 - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재 - 재정위원회 사무국에서 지정한 데이터 및 자동화 자문기구 또는, - 자동화된 의사결정 시스템의 사양을 전문가가 검토하는 저널에 게재. 접근이 제한되는 경우, 결과에 대한 평문 요약을 공개적으로 이용할 수 있는지 확인

요구사항	수준 Ⅰ	수준 Ⅱ	수준 Ⅲ	수준 Ⅳ
② 성별기반 플러스 분석 (GBA+) (section 6.3.6)	비해당	성별기반 플러스 분석(GBA+)을 통해 다음의 문제를 해결할 수 있도록 확인 : - (시스템, 데이터 및 의사결정 포함) 자동화 프로젝트가 성별 또는 기타 정체성 요인에 미치는 영향 - GBA+를 통해 식별된 위험을 해결하기 위해 계획된 조치 또는 기존의 조치		
③ 통지 (section 6.2.1-6.2.2)	비해당	사용 중인 모든 서비스 제공 채널(인터넷, 대인, 메일 또는 전화)을 통해 이해하기 쉬운 용어로 공지	사용 중인 모든 서비스 제공 채널(인터넷, 대인, 메일 또는 전화)을 통해 이해하기 쉬운 용어로 공지. **또한, 다음을 사항을 설명하는 자동화된 의사결정 시스템에 대한 문서를 이해하기 쉬운 용어로 발간 :** - 구성 요소의 작동 방식 - 행정적 결정을 지원하는 방식 - 모든 검토 또는 감사의 결과 - 학습 데이터에 대한 설명, 또는 이 데이터를 공개적으로 사용할 수 있는 경우 익명화된 학습 데이터 링크	
④ 의사결정 Human-in-the Loop (section 6.3.11)	인간의 직접적인 개입 없이 의사결정이 내려질 수 있음		의사결정 절차에서 특정 시점에 인적 개입이 없으면 의사결정이 내려질 수 없으며, 최종 의사결정은 사람에 의해 이루어져야 함	
⑤ 설명 제공 (section 6.3.2)	적용 가능한 모든 법적 요구사항 외에도 공통적인 의사결정 결과에 대한 유의미한 설명을 공개할 수 있도록 보장. 해당 설명은 다음에 대한 일반 사항을 반드시 제공 : - 의사결정 과정에서 시스템의 역할 - 입력 데이터, 출처 및 수집 방법 - 입력 데이터 평가 기준과 그 프로세스	적용 가능한 모든 법적 요구사항 외에도 **혜택, 서비스, 규제 조치를 거부하는 모든 의사결정 결과에 대하여 고객에게 유의미한 설명을 제공되도록 보장.** 해당 설명은 다음에 대한 일반사항을 반드시 제공해야 함 : - 의사결정 과정에서 시스템의 역할 - **해당되는 경우 학습 데이터와 클라이언트 데이터,** 출처, 수집방법 - **클라이언트 데이터** 평가 기준과 그 프로세스 - 시스템 산출 결과물과 행정적 의사결정 맥락에서 결과를 해석하는 데 필요한 관련 정보 - **결정을 내리게 된 주요 요인을 포함한 행정적 결정의 정당성** 또한, 고객에게 선택 가능한 옵션에 대한 정보를 적절한 곳에 제공해야 함		

요구사항	수준 Ⅰ	수준 Ⅱ	수준 Ⅲ	수준 Ⅳ
	- 시스템 산출 결과물과 행정적 의사결정 맥락에서 결과를 해석하는 데 필요한 관련 정보 - 의사결정 주요 요인 또한, 고객에게 선택 가능한 옵션에 대한 정보를 적절한 곳에 제공해야 함 알고리즘 영향평가를 통해 이해하기 쉬운 용어로 설명해야 하며, 부처 웹사이트에서 찾아볼 수 있어야 함	이러한 **요소에 대한 일반사항**은 알고리즘 영향평가를 통해 **반드시 제공**되어야 하며, 부처 웹사이트를 통해 찾아볼 수 있어야 함		
⑥ 교육(훈련) (section 6.3.7)	비해당	시스템의 설계 및 기능에 대해 문서화	시스템의 설계 및 기능에 대해 문서화, 교육과정 이수 필수	시스템의 설계 및 기능에 대해 문서화, 정기적인 교육과정, 교육 이수 확인 수단 강구
⑦ IT 및 비즈니스 연속성 관리 (section 6.3.8)	비해당		자동화된 의사결정 시스템 이용이 불가능할 경우, 지정 관리자와 협력하여 시스템 복구 전략, 비즈니스 연속성 계획 및 기타 관련 보안 조치가 수립되도록 보장	
⑧ 시스템 구동승인 (Section 6.3.12)	비해당		책임자(차관) 승인	국가 재정위원회 승인

자료: Government of Canada(2023)

알고리즘 영향평가는 자동화된 의사결정 시스템 설계 전과 생산 전에 총 두 번 완료되어야 한다. 첫 번째 영향평가의 결과에 따라 자동화된 의사결정 시스템 구현 과정에서 충족되어야 할 완화 및 자문 요구사항을 제시한다. 두 번째 영향평가에서 그 결과가 구축된 시스템을 정확하게 반영하는지, 즉 완화 조치 이행 여부를 확인하는 재검토 절차를 필수적으로 이행해야 한다.

4. 덴마크 「디지털활동 인권영향평가」 (Human Rights Impact Assessment of Digital Activities)

인종, 민족, 성별과 관련된 평등 추구를 목적으로 운영되는 덴마크의 국가 인권 연구소(The Danish Institute for Human Rights)는 2020년 11월 '디지털활동 인권영향평가 지침(Guidance on Human Rights Impact Assessment of Digital Activities)'[12]을 발간하였다. 지침의 대상은 디지털 프로젝트, 제품 및 서비스, 이와 관련된 디지털 활동으로, 인권영향평가를 프로젝트, 제품 또는 서비스 수준에서 영향을 평가하는 핵심적인 인권실사 도구로 간주한다. 해당 지침은 평가 수행 결과로 영향에 대한 스냅샷과 예방·완화 조치에 대한 권장사항을 제공한다. 또한 법률, 규정 또는 시장에 대한 중대한 변화가 있거나 사회적, 정치적 상황에 중대한 변화가 있을 때 최초 인권영향평가의 결과를 재평가해야할 의무가 있다. 그러므로 평가가 반복적인 과정임을 인식하고 지속적인 학습과 분석을 이어가야 한다.

평가의 고려 요소 및 기준으로 국제 인권 기준, 인권 기반 절차, 책임성 등의 요소를 필수적으로 고려하며 프로세스·콘텐츠 핵심기준[13]에 따라 평가한다. 평가는 총 다섯 단계에 따라 진행되며, ① 계획 및 범위 설정, ② 데이터 수집 및 상황 맥락 분석, ③ 영향 분석, ④ 영향 예방, 완화 및 구제, ⑤ 보고, 모니터링 및 점검으로 이루어져 있다.

첫 번째, '계획 및 범위 설정' 단계이다. 평가의 효과적 수행 및 원하는 결과 달성을 위해서는 올바른 계획과 범위 설정이 필수적이다. 범위 설정의 목표는 디지털 활동의 유형, 인

12) 디지털활동 인권영향평가 지침, https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities

13) ① 프로세스 핵심기준 : 참여, 차별금지, 역량강화, 투명성, 책무성
② 콘텐츠 핵심기준 : 벤치마크, 영향범위, 영향 심각도 평가, 영향 완화 조치, 구제책에 대한 접근성

권적 맥락, 관련 이해관계자, 평가 결과의 활용처와 관련한 정보를 검토하여 인권영향평가의 매개변수를 정의하려는 데 있다. 기관의 담당자는 이 단계에서 피평가기관으로부터 독립적인 인권영향평가팀의 구성과 이해관계자의 참여에 대한 결정을 내린다. 또한, 평가를 위한 임무 정의서(TOR, Terms of Reference)를 작성해야 한다. 범위설정과 임무 정의(TOR)는 이후 평가에 소요되는 시간, 새롭게 발견되는 이슈나 인권 영향에 따라 유동적으로 변경될 수 있다.

두 번째 단계는 '데이터 수집 및 상황 맥락 분석'이다. 사용자, 잠재적 및 실제적으로 영향을 받는 권리주체, 특히 취약 집단(여성, 장애인, 노인, 성소수자, 이주민, 난민 등)의 인권 향유에 대한 기본 데이터를 이해관계자로부터 수집한다. 앞서 범위 설정 단계에서는 2차 자료 분석에 의존하지만, 본 단계에서는 1차적인 데이터를 수집하고 대면 또는 온라인으로 인터뷰하는 등 이해관계자의 참여와 협의를 중시하고 있다. 따라서 증거 수집, 기준 연구 또는 기준 개발 단계라 불리기도 한다. 데이터 수집에 인권기반접근법(참여, 데이터 세분화, 자기 확인, 투명성, 개인정보보호, 책무성) 측면에서 살필 것을 권고하고 있으며, 이렇게 수집된 데이터를 통해 평가팀은 인권 향유의 현재 상태에 대한 맥락 분석을 수행할 수 있다. 맥락 분석은 인권에 미치는 실제 영향을 식별하고 향후 영향을 더 잘 예측하는 데 도움이 된다. 데이터 수집 및 맥락 분석 단계에서 인권의 특정 측면에 주목하여 데이터를 수집해야 할 뿐만 아니라, 후속 영향 예방, 완화 및 개선에 초점을 두어야 하며, 이를 위해 평가팀은 구조적, 절차적, 결과적 수준에서 질적 및 양적 지표 모두를 사용해야 한다. 단, 인권 영향 분석은 항상 질적이고 설명에 기반한 분석을 요구하기 때문에 지표 및 기타 측정값에만 의존하면 안되며, 지표가 평가에 가치를 더하는 도구일 뿐 이를 대체할 수는 없음을 염두해야 한다. 위험 신호로서 도움이 될 수 있으나 실제적 및 잠재적 인권 영향을 완전히 이해하기 위해서는 관련 권리주체, 의무주체 및 기타 관련 당사자와 협의하는 등 질적 방법을 사용하는 조사가 추가적으로 이루어져야 함을 명시한다.

세 번째로 '영향 분석' 단계가 이어진다. 영향 분석 단계에서는 실제적 또는 잠재적인 인권영향을 식별하고 심각도를 평가하기 위해 앞서 수집된 데이터를 분석하는 단계로, 그 기준은 ⓐ 영향의 유형(〈표 2-18〉), ⓑ 영향의 심각도(〈표 2-19〉), ⓒ 부정적인 영향으로 규정한다. 분석에는 국제 인권 기준, 사업 비교, 이해관계자 참여 결과 등을 활용해야 하

며, 즉각적으로 보이는 영향뿐만 아니라 사업이 야기 또는 기여했거나 그럴 수 있는 모든 영향, 직접적으로 관련된 영향을 고려해야 한다. 특히, 부정적 인권영향은 작위 또는 부작위가 개인 및 집단이 인권을 향유할 수 있는 능력을 전체 또는 부분적으로 제한할 때 발생한다. 본 지침이 제시하는 잠재적인 영향의 예시로, 형사재판 과정에서 재범 가능성을 계산하는 알고리즘이 특정인의 적법절차 및 공정한 재판을 받을 권리에 잠재적인 위험을 초래할 수 있다는 것을 언급한다.

영향의 유형 분석은 아래의 〈표 2-18〉과 같이 이루어지며 영향의 심각도 분석은 다음의 〈표 2-19〉에서 세부 내용을 확인할 수 있다. 영향 분석에는 영향의 범위, 규모 및 회복불가능성을 고려하여 영향의 '심각도'를 평가하는 과정이 포함되어야 하며 심각도가 높을수록 신속하고 엄격한 조치가 필요하다. 심각도는 영향의 범위(scope)[14], 규모(scale)[15], 회복불가능성(irremediability)[16]에 의해 결정하며, 여성, 아동, 소수 인종, 장애인, 성소수자 등을 비롯하여 취약하거나 소외될 위험이 높은 집단별로 인권 영향이 어떻게 다르게 미치는지 특별한 주의를 기울여야 한다. 심각도 지표를 분석한 예시는 〈표 2-20〉에 정리하였다. 이와 같은 심각도 평가를 통해 식별된 영향에 대하여 조치의 우선순위[17]를 결정할 수 있다.

14) 영향이 얼마나 넓게 미치는지 또는 영향을 받는 사람 수에 대한 것

15) 영향의 중대성

16) 영향을 받은 개인을 영향 이전의 상황과 같거나 동등한 상황으로 복원하는 능력에 대한 것. 통상 영향의 규모나 범위가 클수록 회복가능성이 높지 않음

17) UN의 지침에 따라 모든 영향을 동시에 해결하는 것이 불가능할 때 그 심각도 순서에 따라 해결하며, 가장 심각한 영향이 가장 먼저 해결되어야 함

〈표 2-18〉 영향의 유형 분석

영향 유형	예시	영향을 받는 권리
회사의 고유 업무로부터 '야기'된 영향	• 노조결성을 광범위하게 제한하는 국가에서 운영되는 한 은행이 노동자 만족도 및 유지율 향상에 도움이 될 신규 '스마트' 인적자원관리 시스템을 구매하고 적용한다. • 신규 시스템은 자연어 처리를 사용하여 직원 간 내부 이메일을 분석하고 노동자의 정서 상태를 평가한다. • 노동자는 근로계약의 일부로 회사가 전자메일에 접근하는 데 동의한다. 시스템 도입 후 회사 내 노조 설립률이 현저히 떨어졌는데, 이는 노동자들이 회사가 노조 활동을 더 쉽게 파악하고 노동자에 부정적인 영향을 미칠 수 있다고 우려하였기 때문이다.	결사의 자유
누적영향을 비롯하여 회사의 고유 업무 또는 제3자를 통해 '기여'된 영향	• 한 소프트웨어 개발사가 상업적 목적으로 공용 소셜 미디어 플랫폼에서 '데이터를 스크랩'할 수 있는 디지털 제품을 개발한다. 이 제품은 제3자에게 판매되고 제3자는 데이터를 스크랩하여 인터넷 이용자에 대한 정보를 정부에 제공하기 위해 이 제품을 사용한다. • 정부는 이 데이터를 사용하여 정치적 반대자를 감시한다. • 소프트웨어 개발사는 잠재적인 사용 사례와 관련 위험을 알고 있어야 한다.	사생활의 권리, 안전의 권리
계약 또는 비계약 사업 관계를 통해 운영, 제품 및 서비스와 '직접적으로 관련'된 영향	• 한 인공지능 개발사가 신체 언어 인식 제품을 개발하고 법 집행 기관이 이 '기성품'을 구매한다. • 법 집행 기관은 이 제품을 범죄 용의자 식별 용도로 적용한다. • 이 제품의 사용으로 소수 민족 및 인종에 대한 부당한 체포가 증가한다.	평등권 및 차별받지 않을 권리, 이동의 자유, 신체의 자유와 안전의 권리

자료: The Danish Institute for Human Rights(2020)

〈표 2-19〉 영향의 심각도 분석

항목	지표	심각도
범위	영향 영역 전체 인구의 20% 이상 또는 식별된 집단의 50% 이상	A
	영향 영역 전체 인구의 10% 이상 또는 식별된 집단의 10-50%	B
	영향 영역 전체 인구의 5% 이상 또는 식별된 집단의 10% 미만	C
규모	- 사망 또는 정신적·육체적 건강에 악영향을 미쳐 삶의 질 및 수명을 현저히 감소시킬 수 있음 - 여기에는 안전의 권리 또는 건강권과 관련한 심각한 영향으로 이어지는 사생활의 권리에 대한 영향이 포함된다.	A
	- 기초생활필수품(교육, 생계 등)에 대한 명백한 인권침해 - 영향평가 절차에서 식별된 집단 또는 주제별 전문가가 높게 평가한 것으로 식별된 문화적, 경제적, 자연적, 사회적 측면에 대한 영향 - 영향평가 과정에서 생계, 건강 또는 안전에 우선순위가 높은 것으로 식별된 공공 서비스 전달과 관련된 부정적 영향(예: 공공의료 서비스에 접근하는 개인을 지원하는 인공지능 챗봇이 노인을 돕는 효율성이 떨어져 일반인보다 적절한 건강 조언을 덜 받는 경우)	B
	- 그 밖의 영향 모두	C
회복 불가능성	(어려움) - 영향의 특성상 구제하기 어렵거나 불가능 - 복잡한 기술 요구사항으로 인해 개선이 어려움 - 식별된 집단에서 구제책 수용을 거의 발견할 수 없음 - 영향에 관련된 사업 파트너가 영향을 구제할 수 있는 능력이 낮음 - 영향으로 인한 손실을 대체할 수 있는 실행 가능한 대안이 없음	A
	(보통) - 영향의 특성상 구제가 가능하지만 쉽지 않음 - 영향을 구제하기 위한 기술적 요구사항이 보다 간단함 - 영향을 받는 식별된 집단이 구제책을 수용 - 일부 기능 개발이 지원되는 경우 영향에 관련된 사업 파트너가 구제 수단을 제공할 수 있음	B
	(쉬움) - 영향의 특성상 구제하기 쉬움 - 영향을 구제하기 위한 기술적 요구사항이 간단 - 영향을 받는 식별된 집단이 구제책을 수용 - 사업 파트너가 영향을 구제할 수 있는 능력을 가지고 있음	C

자료: The Danish Institute for Human Rights(2020)

〈표 2-20〉 영향의 심각도 지표 분석 예시

어떤 국가의 사법부가 도주 및 재범 위험을 계산하여 판결을 지원하는 데 사용하는 알고리즘을 개발하였고, 그 대상은 자동화된 위험평가에 대해 이의를 제기할 수 없다고 할 때, 이 알고리즘은 권리주체의 적법 절차와 공정한 재판을 받을 권리에 잠재적인 영향을 미침

항목	심각도
범위	• B: 판결만을 보조하는 고도로 정확한 알고리즘의 경우 영향을 받는 사람의 절대적인 수는 적을 수 있지만, 사례를 자세히 살펴본 결과 주로 영향을 받는 사람들이 선주민으로 나타났고 취약 집단으로 식별되었다.
규모	• A: 공정한 재판을 받을 권리에 대한 영향으로 상당한 규모의 인권영향이 있다. 알고리즘의 지원으로 이루어진 판결을 받은 모든 사람에게 적용된다. • 그러나 영향을 받는 선주민과 관련한 영향은 훨씬 더 크며, 이는 그 영향이 본질적으로 차별적이기 때문이다.
회복 불가능성	• B: 판결의 성격에 따라 상황을 완전히 구제하지 못할 수도 있다. 예를 들어, 미래의 직업 기회가 제한되어 평생 영향을 미칠 수 있다. • 형은 오랫동안 정신 건강에 영향을 미칠 수 있으며, 수년간 가족생활을 영위할 수 없으며, 반드시 회복될 수 있는 것이 아니다. 그러나 영향의 일부는 차별적 판결을 변경하여 구제될 수 있다.
종합평가	• 높은 심각도의 영향으로 간주될 수 있다. 이는 특히 취약 집단에 영향을 미치는 지속적인 영향이며 일부 사례(과도한 징역형)는 예를 들어 건강과 생계 및 가족생활에 대한 권리와 관련한 영향으로 인해 구제될 수 없다.

자료: The Danish Institute for Human Rights(2020)

무엇보다 본 평가가 인권 존중에 기여하기 위해서는 부정적인 인권 영향을 우선적으로 식별하고 해결하는 데 중점을 두어야 한다. 긍정적인 인권 영향을 식별하는 것은 주요 목표가 아니며, 부정적인 영향을 식별하고 해결하는 데 방해가 되어서는 안된다고 명시하고 있다. 이와 유사하게, UN 역시 별도 지침을 통해 긍정적인 기여도가 부정적인 영향을 상쇄할 수 없으며 각각 검토해야 함을 지적한다.

이후의 절차는 네 번째, '영향 예방, 완화 및 구제' 단계이다. 인권영향평가의 필수적 부분으로, 본 단계에서 기관, 평가팀 및 이해관계자는 협력하여 부정적인 인권영향을 예방, 완화 및 개선하기 위한 계획을 수립한다. 기업이 확인된 부정적 영향에 대해 어떤 조치를 취해야 하는지, 기업은 이해관계자와 어떤 방식으로 협력해야 하는지, 참여 모니터링을 영

향 관리에 어떻게 적용할 수 있는지 등에 관하여 다루고 있다. 권리주체 및 그 대리인은 영향 예방, 완화 및 개선 조치를 계획, 시행 및 모니터링하는 데 유의미하게 참여해야 함을 명시하였다. 또한, 식별된 영향에 대한 조치는 주로 부정적인 인권 영향을 방지하고 감소시키는 데 중점을 두어야 하며, 완화 조치에 대해 계층적 접근 방식(mitigation hierarchy)[18]을 참고할 수 있다. 부정적 영향이 식별되고 영향 관리 계획이 수립되면, 조치를 이행할 것인지, 어떻게 이행할 것인지 후속 조치를 시행하여 식별된 영향을 해결하는 것이 가장 중요하다. 이때 지속적인 모니터링을 통해 영향 완화 조치가 효과적인지 여부에 대한 정보를 제공하고 그렇지 않은 경우 필요한 조정을 수행할 수 있도록 하며, 예상치 못한 영향도 식별할 수 있어야 한다.

마지막으로 덴마크 인권영향평가는 '보고, 모니터링 및 점검' 단계를 거친다. 해당 단계는 인권영향평가 절차를 투명하고 책임성 있게 만드는 데 있어 중요한 단계이다. 최종 평가 보고서는 식별된 영향, 이를 해결하기 위해 취한 조치 및 조치 효과에 대한 모니터링을 문서화한 것으로, 목차는 ① 소개, ② 방법론, ③ 맥락 설명, ④ 조사결과 및 조치로 구성할 것을 제안한다. 또한, 보고서 공개 시 정보의 기밀성 및 민감성 등을 고려할 필요가 있다.

본 인권영향평가는 그 효과성 평가 및 지속적인 개선 조치까지 포함하는 것으로, 영향평가가 효과적이었는지 평가할 때에는 먼저 인권영향평가 절차 그 자체를 점검하면서 초기 목표를 충족했는지 파악하고 판단한다. 그 다음에는 최종 보고서 발행 후 후속적으로 예상하지 못한 영향, 회사 정책 및 관행에 대한 상당한 변경, 관련 현지 상황의 중대한 변화 등을 검토해야 한다.

이해관계자(권리주체, 의무주체 및 기타 이해관계자) 참여는 인권영향평가 모든 단계에서 공통적으로 필요한 사항이다. 일방향적 사업 정보 제공, 양방향적 정보교환 및 대화, 특정 조치 후 의견을 수렴, 상호합의를 도출하기 위한 양방향적 협상 등의 다양한 참여 방식을 해당 평가 대상의 특성에 맞게 조정하여 운영할 것을 권고한다.

---

18) 방지를 우선시하는 접근법으로, 방지가 가능하지 않은 경우 이를 축소하거나 완화하는 방법을 우선적으로 고려

5. 유럽평의회 「인권, 민주주의, 법치 관점의 AI 시스템 위험 및 영향평가」(Democracy, and the Rule of Law Impact Assessment for AI Systems)

유럽평의회[19] 인공지능 위원회(Committee on Artificial Intelligence, 이하 CAI)는 2024년 11월 인권, 민주주의, 법치 관점의 AI 시스템 위험 및 영향 평가(이하 HUDERIA)[20]' 방법론을 공식적으로 채택하였다. HUDERIA(The Risk and Impact Assessment of AI Systems from the point of view of Human Rights, Democracy and the Rule of Law) 영향평가는 영국의 AI 국가 연구소인 앨런 튜링 연구소에서 개발한 것으로 공공 및 민간 주체를 대상으로 한다. 인공지능 시스템의 생애주기 동안 인권, 민주주의 및 법치에 대한 위험과 영향을 식별·평가·완화하기 위한 유연한 방법론을 제공하기 위한 목적으로 설계되었다.

유럽평의회의 '인공지능과 인권, 민주주의, 법치에 관한 유럽 평의회 기본 협약[21]'(이하 '인공지능 기본 협약') 서명 국가들은 협약에 명시된 위험 및 영향 관리 의무 이행에 HUDERIA 방법론을 적용할 수 있다. 인공지능 기본 협약은 공공 및 민간부문 인공지능 시스템을 규제할 때 인권, 민주주의, 법치에 미칠 수 있는 위험을 관리하는 법적 규제를 제시한 최초의 국제 협약으로, 2024년 5월 미국과 EU를 포함한 10개국이 서명했다. 인공지능 기본 협약 서명 국가들은 제5장에 명시된 위험 및 영향 관리에 대한 기본 표준을 포함한 의무를 완전히 준수해야 한다. HUDERIA는 법적 구속력이 없는 독립적인 지침으로서, 인공지능 기본 협약 당사국은 자국의 법률에 따라 위험 및 영향 평가에 대한 새로운 접근 방식을 개발하거나 기존 접근 방식을 개선하기 위해 HUDERIA 전체 또는 일부 적용이 가능하다.

---

19) 유럽 평의회(Council of Europe)는 유럽 국가들이 인권, 민주주의, 법치를 옹호하고 보호하기 위해 1949년에 설립된 국제기구로 46개 회원국 보유

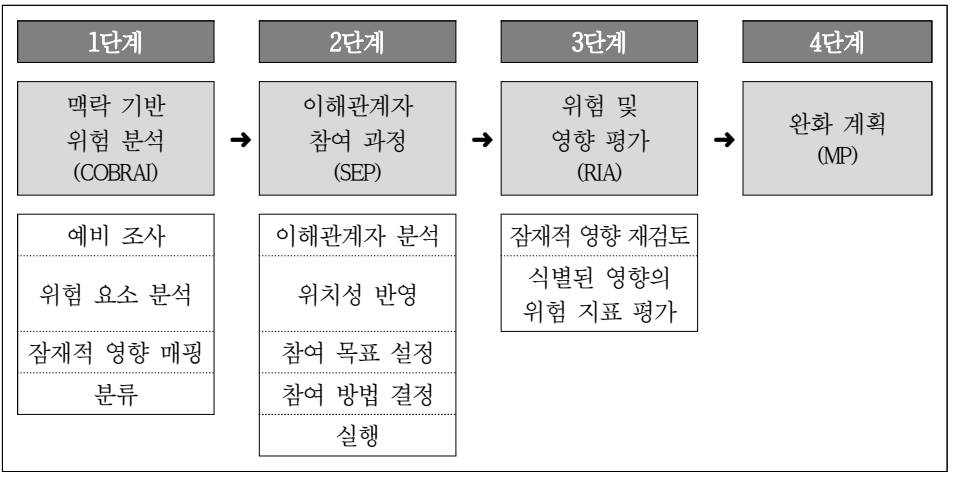
20) Committee on Artificial Intelligence(2024). Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law(HUDERIA Methodology). Retrieved from https://rm.coe.int/cai-2024-16rev2-methodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f

21) CET No.225 (유럽평의회 조약 제225호), https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

본 영향 평가는 위험 및 영향을 평가하기 위한 상위 개념, 과정 및 요소를 제시하는 일반적 지침인 ① 'HUDERIA 방법론'과 평가 실행을 지원하기 위한 지원 자료와 도구, 권장 사항을 제공하는 구체적 지침인 ② 'HUDERIA 모델'이 결합된 구조를 취한다. 상위 개념 중심의 HUDERIA 방법론을 2024년에 채택하였고, 2025년에 상세한 지원 자료와 권장 사항을 포함한 HUDERIA 모델 개발 계획을 발표하였다.

구체적으로, HUDERIA는 네 가지 요소, ① 맥락 기반 위험 분석, ② 이해관계자 참여 과정, ③ 위험 및 영향 평가, ④ 완화 계획으로 구성된다(〔그림 2-2〕).

〔그림 2-2〕 인권, 민주주의, 법치 관점의 AI 시스템 위험 및 영향 평가 체계도



자료: Committee on Artificial Intelligence(2024)를 토대로 연구진 정리

1단계는 '맥락 기반 위험 분석(COBRA:Context-Based Risk Analysis)'으로, 예비 조사를 통해 인공지능 시스템의 여러 영역에서 나타나는 다양한 위험 요소를 식별하고, 영향을 받는 대상과 잠재적 부정 영향을 분석하여 이후 본 평가에 필요한 기초 자료를 제공하는 단계이다. ① 예비 조사, ② 위험 요소 분석, ③ 인권, 민주주의 및 법치에 대한 잠재적 영향 매핑, ④ 분류의 네가지 하위 단계를 순서대로 수행하며 마지막 단계의 분류 결과에 의거해 이후의 HUDERIA 절차는 불필요할 수 있다. 우선, 예비조사를 통해 인권, 민주주의, 법치의 관점에서 잠재적 부정 영향의 예비 범위와 시스템의 기술적 특성, 우려 사항을 탐

구한다. 위험 요소 분석 단계에서는 인공지능 시스템의 적용, 설계 및 개발, 배포 영역과 관련된 여러 위험 요인을 종합적으로 평가한다. 인공지능 시스템의 적용 영역은 시스템이 적용되는 산업 분야, 법적 및 규제 환경, 사회문화적 특성 등을 포함하고, 설계 및 개발 영역은 시스템 관련 기술적 특성을 포함, 배포 영역은 위험이 실제로 발생하고 관리되는 데 영향을 미치는 요소를 포함한다. 잠재적 영향 매핑 단계로 넘어가면, 분석한 위험 요인으로 인하여 잠재적 영향을 받는 대상(사람이나 집단)을 식별하고 인권, 민주주의 및 법치에 대한 잠재적이고 실제적인 영향의 특성을 설명한다. 분류 단계에서는 이전 단계에서 수집된 정보에 의거하여 인공지능 시스템 개발 및 배포 과정의 위험 수준이 이익 수준을 초과하는지 판단, 중대한 위험을 초래하는 인공지능 시스템을 식별 및 분류한다.

2단계는 '이해관계자 참여 과정(SEP : The Stakeholder Engagement Process)이다. 영향을 받을 수 있는 대상(특히, 취약·소외 계층)의 의견을 포함시키는 절차로, 인공지능 시스템으로 인해 영향을 받을 수 있는 이해관계자 집단을 식별하고, 상대적인 이익, 권리, 사회경제적 취약성, 중요성 등을 평가하여 포용적인 참여를 지원한다. 이 단계는 다섯 개의 하위 단계, ① 이해관계자 분석, ② 위치성 반영, ③ 참여 목표 설정, ④ 참여 방법 결정, ⑤ 실행으로 구성된다.

3단계는 '위험 및 영향 평가(RIA : The Risk and Impact Assessment)이다. COBRA(1단계)에서 분류된 중대한 위험을 초래하는 인공지능 시스템의 생애주기 내 활동이 인권, 민주주의, 법치에 미칠 수 있는 잠재적 위험과 영향을 상세히 평가하는 절차로, 1·2단계에서 식별한 영향 재검토 및 추가 분석 후 영향의 위험 지표를 평가한다. ① 잠재적 영향 재검토, ② 식별된 영향의 위험 지표[22] 평가의 두 하위 단계로 구성된다.

4단계는 '완화 계획(MP : Mitigation Plan)이다. 부정적 위험 및 영향을 해결하기 위한 구체적 조치와 전략을 제시하는 절차로, 식별된 피해의 심각성과 발생 가능성을 토대로 완

22) ① 규모(Scale) : 잠재적 부정 영향의 심각성
② 범위(Scope) : 잠재적 부정 영향의 범위 (영향을 받는 사람들의 수, 개인이나 집단의 사회경제적 취약성, 영향이 미치는 시간의 범위)
③ 해결가능성(Reversibility) : 잠재적 부정 영향을 받은 사람들이 그 영향 이전의 상황이나 이에 상응하는 상황으로 복원될 수 있는 가능성
④ 발생가능성(Probability) : 잠재적 부정 영향의 발생가능성

화 조치 및 계획을 수립하고 구제 방안을 마련한다. 부정 영향을 예방하거나 완화하기 위해 취할 수 있는 행동을 결정할 시 '완화 체계(Mitigation Hierarchy)'라는 구조적 접근법을 활용한다 (〈표 2-21〉). 완화 체계는 부정적 영향 '회피' 전략에 우선순위를 두고, 그 다음 이를 줄이고 개선하는 것을 제안한다. '회피' 전략은 잠재적인 부정 영향을 무시하는 것이 아니며, AI 시스템의 생산, 설계 배포 등의 과정에서 부정 영향을 피하기 위해 처음부터 변경을 수행하는 것임을 강조하였다.

〈표 2-21〉 완화 체계 접근법

완화 체계 (Mitigation Hierarchy)			
회피 Avoid	완화 Mitigate	복구 Restore	보상 Compensate
AI 시스템의 생산·사용·설계·개발·배포 과정의 부정적 영향을 피하기 위해 처음부터 변경을 수행 * 회피는 잠재적 부정적 영향을 무시하는 것이 아니라는 점을 강조	AI 시스템의 생산·사용·설계·개발·배포 과정의 부정적 영향을 최소화하기 위한 조치를 수행	부정적 영향을 받은 사람들을 부정적 영향 이전의 상황과 유사하거나 동등한 상황으로 복구, 재활시키기 위한 변경을 수행	부정적 영향을 방지하거나 완화할 다른 접근이 불가능하거나 효과적이지 않은 경우, 현물 또는 다른 방식으로 보상을 제공
Level 1	Level 2	Level 3	Level 4
← 우선순위 상			우선순위 하 →

자료: Committee on Artificial Intelligence(2024)를 토대로 연구진 정리

추가로, 정기적 점검 및 재평가 과정은 HUDERIA 모든 단계에서 필수적으로 이행되어야 할 중요 절차로 강조한다. AI 시스템 생애 주기 전체에 걸쳐 위험 및 영향 평가가 효과적으로 유지되기 위한 지속적인 과정이며, 새로운 영향을 식별하고 완화 계획을 업데이트할 수 있는 기회를 제공하기 때문이다. HUDERIA는 정기적 점검 과정의 구체적인 방식, 기준, 계기, 점검 및 거버넌스 매커니즘에 대한 유연성을 제공한다.

6. 소결

본 절에서는 해외의 주요 AI 영향평가 사례를 검토하였다. EU, UNESCO, 캐나다 등 국제사회에서 제시하고 있는 영향평가의 목적, 주체, 대상, 방식 등을 심층적으로 비교·분석하였고, 그 요약은 〈표 2-22〉에 정리하였다.

본 절에서 검토한 5개의 해외 AI 영향평가 각 사례별 주요 특징은 다음과 같다. EU의 기본권 영향평가는 최초의 인공지능 규제 법인 'EU 인공지능법'에 근거한 고위험 인공지능 시스템에 대한 의무 평가로, 기본권을 평가 지표로 하여 인공지능 시스템의 정의, 영향받는 대상, 구체적 위험 등을 평가 항목으로 제시하고 있다. 유럽에 출시되는 모든 제품은 인공지능 법의 규제를 적용받기 때문에, 유럽 수출을 목표로 하는 국내 기업에도 해당 영향평가는 사실상 의무로 작용하며, 제품의 설계 단계부터 관련 평가 요소를 면밀히 고려할 필요가 있다.

UNESCO의 윤리영향평가는 정부가 조달하는 인공지능 시스템을 대상으로 하며, 윤리적 가치에 따른 영향 식별을 핵심 목적으로 한다. 특히, 기업이나 정부기관이 자율적으로 인공지능 시스템의 윤리적 영향을 분석하고, 이에 대한 내부적인 대응책을 마련할 수 있는 평가의 틀을 정립하였다는 점, 그리고 영향의 규모, 범위, 발생가능성 등 다층적 분석을 가능하게 하는 실질적 평가 지표를 구축하였다는 점에서 의의가 있다.

캐나다의 알고리즘 영향평가는 정부기관이 개발·조달하는 자동화된 의사결정 시스템을 평가 대상으로 하며, 위험 완화조치를 독립된 평가 요소로 제도화하고 있다는 측면에 주목할 필요가 있다. 예를 들어, 특정 위험이 충분히 완화 가능하다고 판단될 경우 최종 위험 점수에서 일정 비율(최대 15%)을 감점하는 방식으로 반영된다. 이처럼 완화 노력을 긍정적 요소로 반영함으로써, 기업이나 기관이 자발적으로 개선 조치를 이행하도록 유도하는 효과가 있다. 불완전한 결과를 그대로 공개하는 것보다, 구체적인 위험 완화 계획까지 포함해 책임 있게 공시함으로써 오히려 소비자의 신뢰를 확보할 수 있다는 점에서, 기업 입장에서도 전략적인 선택이 될 수 있다.

덴마크의 디지털활동 인권영향평가, 유럽평의회의 HUDERIA 영향평가는 인공지능 시스템이 기본권 및 인권에 미치는 영향에 대하여 다층적인 평가를 수행한다는 점에서 본 연구의 AI 영향평가와 유사한 방향성을 보인다. 두 사례 모두 UNESCO의 영향 측정 방식-영

향의 범위, 규모, 기간, 해결가능성 등-을 각 도구의 사용 맥락에 맞게 차용하고 있으며, 이를 통해 평가 대상의 긍·부정 영향을 입체적으로 식별한다.

이상의 분석을 통해 도출한 시사점은 다음과 같다. 첫째, AI 영향평가의 글로벌 상호운용성을 확보할 수 있도록 세부 절차와 방법을 개발할 필요가 있다. AI 영향평가는 'AI기본법' 35조 1항에 따른 노력 사항으로, 강제력은 없으나 제도 도입의 실효성을 높이기 위해서는 기업의 자발적 참여를 유도할 수 있는 설계가 중요하다. 특히, 유럽 지역으로의 제품 수출을 고려하는 국내 기업의 경우, EU의 기본권 영향평가와 상호 인정 가능한 평가 항목과 절차를 제공한다면 중복 대응의 부담을 줄일 수 있다. 이는 AI 영향평가가 단순히 규범적 요구를 넘어 기업의 대외 전략과도 연계될 수 있는 실질적 동기를 제공할 수 있다.

둘째, 영향평가 이후의 환류 과정을 제도화할 필요가 있다. 단순히 위험을 식별하는 데 그치지 않고, 식별된 위험에 대해 기업이나 기관이 책임을 가지고 완화 조치를 이행하도록 유도하는 평가 구조가 필요하다. 이와 관련하여, 캐나다의 사례는 완화 노력에 인센티브를 부여하는 방식으로 설계되어 있으며, 완화 조치의 공시를 통해 투명성과 신뢰를 확보하고 있다. 이는 우리나라 AI 영향평가 제도 설계에도 반영 가능한 유용한 사례이다.

셋째, 인공지능 시스템이 기본권 및 인권에 미치는 영향을 다층적으로 분석하는 데 초점을 맞춰야 한다. 유네스코, 덴마크, 유럽평의회의 영향평가 제도는 모두 영향을 단일 지표로 단순화하지 않고, 윤리적 가치에 기반해 긍정 및 부정 영향을 식별하며 영향의 규모·범위·기간·발생가능성·회복가능성 등을 종합적으로 분석하는 방식을 취하고 있다. 특히 UNESCO의 영향 평가 도구는 평가라는 형식을 빌려 기업이 자율적으로 숙고하고 학습할 수 있도록 설계되어 있으며, 평가 결과 자체보다는 그 과정을 통한 윤리적 성찰과 책임 강화에 방점을 두고 있다. 덴마크와 HUDERIA의 사례 역시 이와 유사한 평가 절차를 안내하고 있는 만큼, 이러한 접근은 본 연구의 AI 영향평가 체계 개발의 주요한 방향성으로 제안하고자 한다.

〈표 2-22〉 해외 AI 영향평가 유사 제도 비교

	EU	UNESCO	캐나다	덴마크	유럽평의회
	기본권 영향평가 (Fundamental Rights Impact Assessment)	윤리영향평가 (Ethical Impact Assessment)	알고리즘 영향평가 (Algorithmic Impact Assessment)	디지털활동 인권영향평가 (Human Rights Impact Assessment of Digital Activities)	인권, 민주주의, 법치 관점의 AI 시스템 위험 및 영향평가 (HUDERIA)
관련 근거	**EU AI Act 제27조**	**인공지능 윤리 권고 정책권고#1**	**자동화된 의사결정에 대한 지침 제6조**	**디지털활동 인권영향평가 지침**	**유럽평의회 조약 제225호**
평가 주체	- **고위험 인공지능 시스템 배포자**	- **정부 인공지능 시스템 조달자**	- **자동화된 의사결정 시스템을 개발·조달하는 연방정부 기관**	- **정부기관, 민간 (자체)**	- **공공 및 민간 주체**
평가 대상	- **고위험 인공지능 시스템*** * ① 정부기관 또는 공공서비스를 제공하는 민간기관 : 중요 인프라 관련 인공지능 시스템을 제외한 고위험 인공지능 시스템, ② 민간기관 : 금융(신용도 평가 등), 보험(생명, 건강 보험 관련 개인의 위험성 평가 또는 가격책정 등)에 해당하는 인공지능	- **정부에서 조달하는 인공지능 시스템**	- **자동화된 의사결정 시스템**	- **디지털 프로젝트, 제품 및 서비스, 이와 관련된 디지털 활동**	- **공공 및 민간부문 인공지능 시스템**

	EU	UNESCO	캐나다	덴마크	유럽평의회
	기본권 영향평가 (Fundamental Rights Impact Assessment)	윤리영향평가 (Ethical Impact Assessment)	알고리즘 영향평가 (Algorithmic Impact Assessment)	디지털활동 인권영향평가 (Human Rights Impact Assessment of Digital Activities)	인권, 민주주의, 법치 관점의 AI 시스템 위험 및 영향평가 (HUDERIA)
평가 지표	- **기본권*** * "EU 기본권 헌장"에 명시된 존엄성, 자유, 평등, 연대, 시민권(예: 투표권, 청원권, 연방 문서 열람권), 정의 등을 포함	- **UNESCO AI 윤리원칙*** * 안전·안정성, 공정성·비차별성·다양성, 지속가능성, 프라이버시·데이터 보호, 인간감독·결단, 투명성·설명가능성, 책무·책임성, 인식·리터러시	- **영역별 위험** (프로젝트, 시스템, 알고리즘, 의사결정, 영향, 데이터) - **시행 중인 위험 관리 조치** (협의, 위험성 제거 및 완화 조치)	- **인권에 미치는 영향 유형** - **인권에 미치는 영향의 심각도**(범위, 규모, 회복 불가능성) - **인권에 미치는 부정적 영향**	- **영향의 위험 지표**(규모, 범위, 해결가능성, 발생 가능성)
평가 방법	- **개발 예정*** * 본 법안은 EU AI Office가 자동화된 도구, 설문지 등의 형태로 평가도구를 개발해야 함을 명시(대상 인공지능 시스템 정의, 사용 기간과 빈도, 영향받는 대상 식별, 가능한 구체적 위해 위험 및 완화 조치 점검 등 포함)	- **조달자 자체 정량·정성 평가** (윤리원칙 관련 절차 이행 점검, 긍정·부정 영향 진단*) * 영향력, 범위, 규모, 발생 가능성, 회복 수준 등을 윤리원칙별로 식별 및 평가	- **온라인 정량평가** - **영향점수 도출 및 영향 수준* 판별** * 영향점수 백분율 기준 4단계로 영향 수준을 정의하며, 영향 수준에 따라 요구되는 완화 조치 결정	- 계획 및 범위 설정 - 데이터 수집 및 상황 맥락 분석 - 영향 분석 - 영향 예방, 완화 및 구제 - 보고, 모니터링, 점검	- 맥락 기반 위험 분석 - 이해관계자 참여 과정 - 위험 및 영향 평가 - 완화 계획
비고	- 배포자는 **평가 결과를 시장감시기관에 통지**해야 하며, EU Database에 고위험 인공지능 시스템 등록 시 **평가 결과 정보 제출 의무**	- **평가 틀**을 제공한 것으로, 특정 맥락 및 각국 상황에 맞게 **조정하여 활용**할 것을 권장	- 영향수준별 완화조치 요구사항 이행 - 개방형 정부 포털에 평가 결과 공개	- 최종 평가보고서 발간 - 효과성 평가 및 지속 개선 조치까지 포함한 모니터링 - 모든 단계에서 이해관계자의 참여	- 부정 영향 완화조치 수립 시 구조적 접근법 '완화 체계' 활용 - 모든 단계에서 정기적 점검 및 재평가 과정 이행

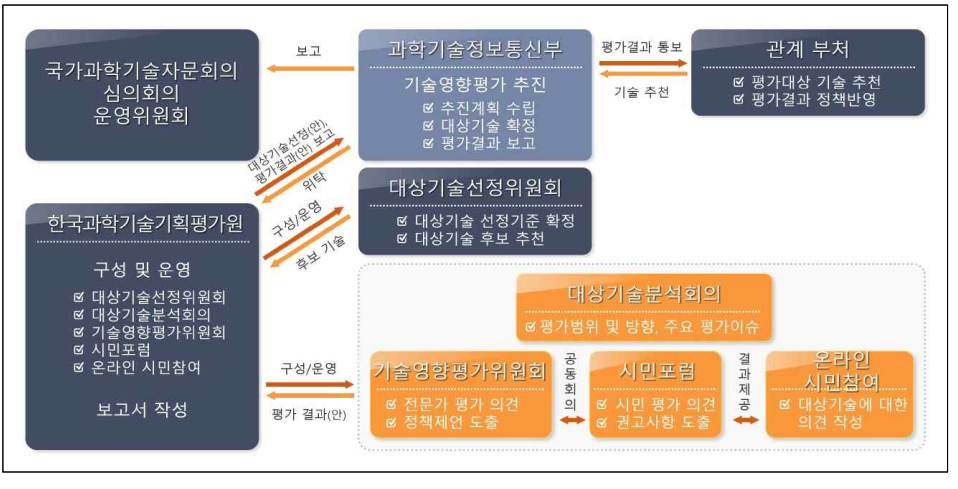
자료: 연구진 작성

제 2 절 국내 AI 영향평가 유사 제도 동향

1. 과학기술정보통신부 「기술영향평가」

기술영향평가는 「과학기술기본법」 제14조에 근거해 새로운 과학기술에 대한 선제적 예측과 대응을 위해 과학기술정보통신부와 한국과학기술기획평가원에 의해 시행된다. 기술영향평가 추진체계는 〔그림 2-3〕과 같으며, 세부 사항은 이후에 살펴볼 예정이다. 과학기술 발전은 경제 가치 창출을 비롯한 삶의 질 증대, 국가 경쟁력 강화 등 긍정적 효과와 동시에 환경·윤리문제 등의 부정적 효과 또한 초래할 수 있으며, 과학기술이 사회에 미치는 영향력의 범위와 복잡도가 지속해서 확대 및 증가하고 있다. 이에 따라 기술영향평가는 새롭게 등장하는 과학기술이 경제·사회·윤리·문화·환경 등에 미치는 영향을 사전에 평가하여 신 과학기술로 인한 긍정적 영향은 극대화하고, 부정적 영향은 최소화할 수 있는 정책적 제언의 도출, 해당 기술에 대한 사회적 수용성 증가에 주요한 목적을 두고 있다.

〔그림 2-3〕 과학기술정보통신부 기술영향평가 절차



자료: 과학기술정보통신부·한국과학기술기획평가원(2023)

기술영향평가 범위와 절차를 개괄적으로 살펴보면 다음의 〈표 2-23〉, 〔그림 2-4〕와 같다. 기술영향평가의 첫 시작은 대상기술의 선정으로 출발하며, 전문가와 이해관계자 등

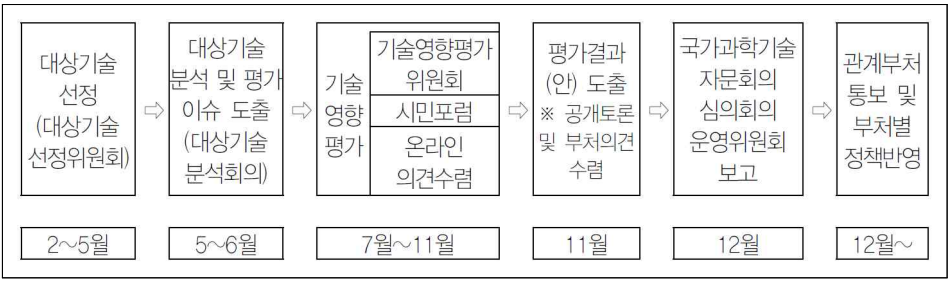
의 의견을 수렴하여 최종적으로 대상기술을 확정한다. 이어 대상기술 분석회의에서 평가 방향과 이슈, 소주제 등을 사전에 도출하고, 이어서 기술영향평가위원회, 시민포럼, 온라인 의견수렴을 통해 최종 평가결과를 도출한다.

〈표 2-23〉 기술영향평가의 범위 및 절차

구분	내용
대상	• 미래의 신기술 및 기술적·경제적·사회적 영향과 파급효과 등이 큰 기술로서 과학기술정보통신부 장관이 관계 중앙행정기관의 장과 협의하여 정하는 기술
주체	• 과기정통부가 매년 실시하되, KISTEP에 위탁·실시
운영	• 전문가 및 시민단체의 참여를 확대하고 일반국민의 의견을 수렴
내용	• 국민생활의 편익증진 및 관련 산업발전에 미치는 영향, 부작용 방지 방안 • 새로운 과학기술이 경제·사회·문화·윤리·환경에 미치는 영향 • 해당 기술의 성격과 파급효과가 성별 등 특성에 미치는 영향

자료: 「과학기술기본법 시행령」 제23조(기술영향평가의 범위 및 절차)

〔그림 2-4〕 한국과학기술기획평가원의 기술영향평가 절차



자료: 과학기술정보통신부·한국과학기술기획평가원(2023)

기술영향평가는 ① 대상기술 선정, ② 대상기술 분석 및 평가이슈 도출, ③ 기술영향평가, ④ 평가결과(안) 도출, ⑤ 국가과학기술자문회의 운영위원회 보고, ⑥ 관계부처 통보 및 부처별 정책반영 등 6단계로 구성된다. 각 단계를 상세히 살펴보면 다음과 같다.

첫 번째 과정인 대상기술 선정 단계는 어떤 기술을 대상으로 기술영향평가를 실시할 것인지를 논의하는 단계로 이해할 수 있다. 대상의 선정은 「과학기술기본법 시행령」 제23조에 명시한 '미래의 신기술 및 기술적·경제적·사회적 영향과 파급효과 등이 큰 기술'을 기

준으로 하며, 자세한 내용은 다음의 〈표 2-24〉와 같다. 진행 순서에 따라 상세히 살펴보면 먼저, 위탁 실시 기관인 한국과학기술기획평가원에서 국가차원의 주요 기술보고서, 설문조사 결과를 종합하여 다수의 후보기술군을 도출한다. 설문조사는 일반시민과 기술수준평가 전문가를 대상으로 실시하며, 2023년 시행된 기술영향평가는 일반시민 614명과 기술수준평가 전문가 360명이 참여하였다.

이후 과학기술, 사회과학, 언론, 시민 등 다양한 분야의 전문가로 구성된 대상 기술 선정 위원회가 평가대상 기술 선정 기준을 수립하고 이를 기반으로 평가 필요성이 높은 후보군 3개를 선정한다. 선정 과정에서 위원회 위원을 대상 서면평가와 일반시민 대상으로 기술선호도 조사를 병행하여 그 결과를 종합적으로 고려한다. 서면평가는 사회적 관심도, 파급효과의 범위, 파급효과의 깊이, 사회적 합의의 필요성, 결과의 활용가능성을 측정하며, 기술선호도 조사는 기술에 대한 관심도, 긍정·부정 파급효과, 사회적 합의의 필요에 대한 우선순위를 조사한다. 선정위원회에서 도출한 후보기술에 대한 관계부처의 선호도 조사를 수렴하여 최종 평가대상 기술을 선정한다.

〈표 2-24〉 기술영향평가의 대상기술 선정 기준

기준	세부 기준
기준 1	• 현재 등장하고 있는 신기술로서 가까운 미래(5~10년)의 국민 생활에 파급 효과가 클 것으로 예상되는 기술
기준 2	• 사회적 관심도가 높아 기술영향평가의 유용성과 시의적절성이 높을 것으로 기대되는 기술
기준 3	• 개념과 범주가 명확한 중간규모*의 기술 (* 과학기술표준 분류상의 중분류 또는 소분류 수준으로서 과제 수준의 기술은 제외)
기준 4	• 과거에 평가한 기술 중에서도 환경변화 및 기술발전 추이를 고려하여 재검토가 필요하다고 인정되는 기술

자료: 과학기술정보통신부·한국과학기술기획평가원(2023)

이후 선정된 대상기술의 분석과 평가를 위한 단계가 진행된다. 이 단계에서는 선정한 기술의 정의와 평가범위를 설정하고 선정 기술의 주요 이슈와 특징이 무엇인지 밝힌다. 이 과정을 위해 해당 분야 산·학·연 전문가, 인문 사회학자 등 전문가 8인으로 대상분석

회의를 구성한다. 해당 회의에서 선정 기술의 특성, 기술개발 수준 및 활용 동향을 검토하고, 구체적인 평가대상과 범위 등을 약 2개월에 걸쳐 확정한다.

다음으로 전문가와 시민이 함께 기술영향평가를 수행하며 이는 세 단계로 구분된다. 첫째, 기술영향평가위원회 의견을 수렴한다. 앞서 도출된 주제에 대하여 심도 있는 토론을 진행하며, 정책적 시사점을 전문가의 관점에서 도출한다. 산·학·연 전문가와 다양한 분야 사회과학 전문가 15인 내외로 구성하며, 평가 연속성과 일관성을 위해 기술영향평가 유경험자와 대상기술 선정 및 분석위원 중 일부를 포함해 구성한다. 둘째, 시민포럼을 통한 평가의견을 수렴한다. 기술영향 평가위원회와 병행하여 진행하며, 15인 내외로 구성한다. 참여자는 선정기술과 직접 이해관계가 없는 시민들로 제한하며, 다양한 계층을 포함하도록 한다. 셋째, 온라인 게시판 운영을 통해 시민의 목소리를 수렴한다. 게시판에는 평가 대상 기술분석회의의 결과 등 관련 자료들을 업로드하여 시민들에게 정보를 제공하는 기능도 수행한다. 이후 기술영향평가의 전 과정의 결과를 종합적으로 고려하여 평가결과(안)을 도출하고 이를 공개토론회 형식으로 진행하고, 주요 위원회 및 부처에 최종 결과를 통보한다.

〈표 2-25〉 역대 기술영향평가 대상 기술

연도	평가대상	연도	평가대상
2003	NBIT융합기술	2016	가상·증강현실 기술
2005	RFID/나노	2017	바이오 인공장기
2006	줄기세포치료기술/나노소재/UCT	2018	블록체인 기술
2007	기후변화대응기술	2019	소셜로봇기술
2008	국가재난질환대응기술	2020	정밀의료 기술
2011	뇌-기계 인터페이스	2021	레벨4 이상 자율주행
2012	빅데이터 분석기술과 활용	2022	합성생물학
2013	3D 프린팅/스마트 네트워크	2023	양자과학기술
2014	무인 이동체/초고층 건축물	2024	AI 헬스케어/휴머노이드 AI/브레인 AI
2015	유전자가위/인공지능		

자료: 과학기술정책지원서비스, 기술영향평가[23)]

23) https://k2base.re.kr/foresight/ta/intrcn/intrcnList.do?pageNavi=page2024&searchTap=notice&index=25

2. 과학기술정보통신부 「인공지능 윤리영향평가」

인공지능 윤리영향평가는 인공지능 기반 서비스의 윤리적 영향력에 대한 체계적인 식별과 합리적인 관리를 위한 평가체계로 2024년 과기정통부와 정보통신정책연구원에서 개발하여 같은 해 AI 영상합성 서비스를 대상으로 시범 시행되었다. 해당 평가는 인공지능 제품·서비스군을 대상으로 2020년 발표된 국가 「인공지능(AI) 윤리기준」의 10대 핵심요건을 기준으로 윤리·신뢰성 측면에서 발생할 수 있는 긍정·부정적 영향을 도출하고 이를 바탕으로 긍정적 영향을 촉진하고 부정적 영향을 완화하기 위한 전략을 제시한다. 이를 통해 인공지능 윤리·신뢰성 실천을 위한 기업의 자율적 노력을 지원하고, 사용자가 인공지능을 윤리적이고 주체적으로 활용하기 위한 기준을 제시하고자 한다. 또한 윤리적 영향력을 사전 평가하여 관련 조치 방안 등의 시사점을 도출하고 인공지능의 윤리적 영향력 파악을 위한 참 자료를 제공한다. 인공지능 윤리영향평가는 궁극적으로 인공지능 제품·서비스를 보다 윤리적인 방식으로 개발, 배포, 활용하도록 장려하는 것을 목적으로 한다.

인공지능 윤리영향평가 절차는 평가대상 선정, 기초 분석, 윤리 영향평가 시행, 정책 제언 및 보고서 공개 4개의 단계로 구성된다. 평가대상의 선정은 기초조사, 대국민 인식조사, 인공지능 윤리·신뢰성 포럼과 부처의 검토 등을 종합하여 진행된다. 해당 과정을 통해 선정된 평가 대상을 기술, 법제도, 사회적 영역 등 여러 관점에서 분석하여 평가 분석틀을 구체화 및 정교화하고 쟁점사항을 도출하고 평가 참고자료를 작성한다. 이후, 산·관·학·연·시민사회 등 다양한 소속과 분야 전문가로 구성된 '전문가평가단'과 일반 국민의 참여로 이루어진 '국민포럼단'이 평가 대상에 대한 윤리적 영향평가를 핵심요건 별로 긍정·부정적 영향 모두를 종합적 관점에서 실시한다. 해당 평가 결과는 공개토론회와 단행본 등 여러 형식을 통해 공개하여 공공영역과 기업뿐 아닌 일반 국민 또한 쉽게 접근하고 다양한 의견을 제시할 수 있도록 지원한다.

〈표 2-26〉 인공지능 윤리영향평가 체계

구분	개요 및 세부 사항
1. 평가대상 선정	(개요) 절차적 타당성과 신뢰성을 확보하는 방식으로 평가대상 선정 (구성) ① 평가 후보군 선별 ② 대국민 의견조사 ③ 포럼위원 검토 ④ 최종 평가대상 선정
2. 기초 분석	(개요) 인공지능 윤리영향평가를 진행하기 위한 평가도구(분석틀)의 구체화 및 정교화 (구성) ① 대상 분석 ② 쟁점사항 도출 ③ 참고자료 작성
3. 인공지능 윤리영향평가	(개요) 과정에서의 민주성과 포괄성, 결과에 대한 신뢰성 및 타당성 확보를 위한 다면적 평가 진행 (구성) ① 전문가 평가위원단 ② 국민 포럼단 ③ 온라인 소통채널 ④ 평가 결과(안) 공개
4. 정책 제언·보고서	(개요) 안전한 인공지능 개발 및 활용을 위해 평가 결과를 종합·정리하여 기업, 시민사회, 학계, 공공(정부) 등에 공개 (구성) ① 정책 제언 ② 보고서 발간

자료: 과학기술정보통신부·정보통신정책연구원(2024)

인공지능 윤리영향평가는 정확성과 신뢰성을 높이기 위해 정량적·정성적 방법론을 다각적으로 연계하고, 상호확인하는 방법론적 삼각망 접근을 채택한다. 또한, 사회에 미칠 파급력과 정부정책과의 부합성, 중요성, 대응 필요성, 시의성 등을 고려하여 평가 대상을 선정하며, 이를 위해 고위험 영역 인공지능 유형에 해당하는 서비스군의 주요 쟁점 및 이슈를 도출하여 평가가 필요한 서비스 영역을 선별한다. 평가 대상에 대한 쟁점사항 도출을 위해 인공지능 윤리기준 10대 핵심요건[24]과 함께 2023년 8월 UNESCO가 발표한 인공지능 윤리영향평가 가이드라인 윤리원칙인 ① 안전 및 보안, ② 인간 감독 및 결정, ③ 투명성 및 설명 가능성, ④ 인식 및 리터러시)도 함께 고려한다. 또한, UNESCO가 제안한 윤리 가치에 대한 인공지능 시스템의 긍정·부정적 영향, 각 영향의 규모, 범위, 발생 가능성 등의 식별·분석을 기본 틀로 평가 체계를 구성함으로써 인공지능 윤리영향평가의 국제적 정합성을 확보하고자 한점을 살펴볼 수 있다.

24) 인공지능 윤리기준 10대 핵심요건은 ①인권보장, ②프라이버시 보호, ③다양성 존중, ④침해금지, ⑤공공성, ⑥연대성, ⑦데이터 관리, ⑧책임성, ⑨안전성, ⑩투명성으로 구성

〈표 2-27〉 인공지능 윤리영향평가 문항 예시(인권보장 영역)

긍정적 영향 평가 문항

1. 인권보장 영역에 대한 긍정적 영향	• *(작성 예시) AI 영상합성서비스는 창의적이고 혁신적인 방식으로 자신을 표현할 수 있는 도구를 제공한다는 점에서 개인이 가지는 **표현의 자유 확대**에 영향을 미칠 수 있음. 독창적인 아바타, 캐릭터를 만들어 콘텐츠를 생산할 수 있고, 이러한 방식을 통해 자신의 생각과 아이디어를 표현하는 범위가 확장되고 이는 곧 표현의 자유 증진으로 연결될 수 있음*
2. 예상되는 긍정적 영향의 규모(크기)	**영향 수준 평가** ① 보통/경미한 ② 중간 ③ 큼 ④ 매우 큼
3. 예상되는 긍정적 영향의 범위 • 영향받는 개인/그룹/단체에 대한 설명 (모든 살아있는 유기체 포함) • 영향의 기간	**긍정적 영향을 받는 대상** • [Primary] *직접적인(주요) 영향을 받는 대상* 대상 설명 : *(작성 예시) AI영상합성서비스를 활용하는 **예술가 및 콘텐츠 제작자**(영상, 음악, 또는 시각 예술의 다양한 분야에서 AI 영상합성서비스를 활용하여 기존에는 불가능했던 창작물과 퍼포먼스 만들 수 있음)* • [Secondary] *간접적인(부차적인) 영향을 받는 대상* 대상 설명 : *(작성 예시) **성소수자, 인종적·종교적 소수자**(사회적 압력 속에서도 자신의 정체성을 감추며 안전하게 목소리를 낼 수 있는 도구로 활용 가능하고 이를 통해 차별받지 않고 의견을 표현할 기회를 얻음)* • [Unexpected/Unintended] *예상치 못하거나 의도치 않았지만 영향 받을 수 있는 대상* 대상 설명 : *(작성 예시) **일반시민, 대중**(AI영상합성기반 콘텐츠를 접하게 되는 대중, 다만 다양한 의견과 표현이 자유롭게 표현되는 과정에서 이를 받아들이거나 비판적으로 분석하는 능력 중요해짐)* **긍정적 영향이 지속되는 기간** ① 단기(1~2년) ② 중기(3~4년) ③ 장기(5~10년) ④ 세대 간(10년 이상)
4. 예상되는 긍정적 영향이 발생할 가능성	**발생 가능성** ① 낮음 ② 중간 ③ 높음 ④ 매우 높음

부정적 영향 평가 문항

1. **인권보장 영역에 대한 부정적 영향**	• *(작성 예시) 정치적 목적이나 사기 등의 목적으로 만들어진 딥페이크 영상이 확산되면서, 대중이 **허위정보, 가짜정보 등을 사실로 받아** 들이고 그에 따른 **잘못된 행동을 하게 될 가능성***
2. **예상되는 부정적 영향의 규모(크기)**	**중대한 수준 평가** ① 보통/경미한(moderate/minor) ② 심각한(serious) ③ 치명적(critical) ④ 재앙적(catastrophic) ※ (보통/경미한) 영향은 있지만 중대하지 않음 (심각한) 일시적 악화 (치명적) 지속적이고 심각한 악화 (재앙적) 돌이킬 수 없는 심각한 피해(생명권, 개인 자유, 안전권 박탈 등)
3. **예상되는 부정적 영향의 범위** • 영향받는 개인/그룹/단체에 대한 설명 (모든 살아있는 유기체 포함) • 영향의 기간	**부정적 영향을 받는 대상** • [Primary] *직접적인(주요) 영향을 받는 대상* 대상 설명 : *(작성 예시) **대중 및 일반시민**(자신이 접하는 정보가 사실인지 아닌지를 판단하기 어려워지면서, 거짓 정보에 의해 잘못된 판단을 하게 될 수 있음)* • [Secondary] *간접적인(부차적인) 영향을 받는 대상* 대상 설명 : *(작성 예시) **언론사, 저널리스트**(가짜 영상이나 정보가 빠르게 확산 되면, 언론사는 이러한 정보를 검증하는 데 상당한 시간과 자원을 소비하게 됨. 또한, 딥페이크 기술로 인해 잘못된 정보가 먼저 확산되면서 진실을 보도하는 언론의 역할이 약화 될 수 있으며, 이는 언론의 신뢰도에도 영향을 미침)* • [Unexpected/Unintended] *예상치 못하거나 의도치 않았지만 영향 받을 수 있는 대상* 대상 설명 : *(작성 예시) **교육 기관 등**(교육 기관은 학생들이 진실과 허위 정보를 구분하는 능력을 키우기 위해 AI 리터러시 교육을 더욱 강화)* **부정적 영향이 지속되는 기간** ① 단기(1~2년) ② 중기(3~4년) ③ 장기(5~10년) ④ 세대 간(10년 이상)
4. **예상되는 부정적 영향이 발생할 가능성**	**발생 가능성** ① 낮음 ② 중간 ③ 높음 ④ 매우 높음
5. **예상되는 부정적 영향이 해결될 가능성**	**해결 가능성 정도** ① 매우 낮음 ② 낮음 ③ 높음 ④ 매우 높음

자료: 문정욱 외(2024)

인공지능 윤리영향평가는 인공지능에 대한 일반 시민의 인식을 파악하기 위해 단계별 조치를 취하고 있다. 2024년 AI 영상합성 서비스 대상으로 실시한 윤리 영향평가의 경우, 평가대상 선정 과정에서 만 19~69세 남녀 1,500명을 대상으로 AI 서비스 영역별 위험도와 친숙도 등의 위험지각 정도를 측정하는 설문조사를 실시하였다. 또한, 윤리 영향평가 과정에서 30인의 '국민포럼단'을 대상으로 연령, 성별, AI 서비스에 대한 태도 등을 고려하여 5개 그룹으로 구분하여 표적집단면접(Focus Group Interview)을 실시하였다. 이를 통해 일반 국민 관점에서 AI 영상합성 서비스에 대한 기대, 윤리적 영향력과 우려 사항 등을 심도 있게 평가함과 동시에 온라인 소통채널을 통해 자유로운 의견 수렴을 진행하였다. 이러한 과정을 통해 인공지능에 대한 일반 국민들의 관점을 파악하고 인공지능 제품 및 서비스로 인한 사회적·윤리적 영향에 대한 정확한 정보를 제공하여 인공지능에 대한 막연한 불안감을 감소시키고 인공지능 윤리에 대한 사회적 인식과 사용자로서의 주체적 인공지능 활용 능력을 제고하고자 하였다.

〔그림 2-5〕 인공지능 윤리영향평가 구성

자료: 과학기술정보통신부·정보통신정책연구원(2024)

3. 개인정보보호위원회 「개인정보 영향평가」

개인정보보호위원회에서 시행하는 개인정보 영향평가제도는 개인정보 처리가 수반되는 사업 추진 시 사전에 수행해야 하는 평가 제도이다. 개인정보 영향평가는 개인정보파일 운용 정보시스템의 신규 도입이나 기존 운영 중인 개인정보 처리시스템의 중대한 변경 등 시스템의 구축·운영·변경 등이 개인정보에 미치는 영향과 위험 요인을 사전 분석하여 개선 사항을 도출하고 이행 여부를 점검하는 것을 목적으로 한다. 이를 통해 설계 단계에서부터 개인정보 침해 위험성을 검토 및 개선하여 시스템 구축 및 운영 시 발생할 수 있는 개인정보 침해 위험을 최소화하고 효과적인 대응책 마련의 효과를 기대할 수 있다.

개인정보 영향평가는 「개인정보 보호법」 제33조(개인정보 영향평가)에 근거를 둔다. 해당 조항은 공공기관이 개인정보파일의 운용으로 인하여 정보주체의 개인정보 침해가 우려되는 경우 개인정보 영향평가의 실시와 해당 결과를 개인정보보호위원회에 제출하는 것을 의무로 규정한다. 동법 시행령 제75는 개인정보 영향평가를 수행하지 않거나 그 결과를 제출하지 않는 경우 3천만원 이하의 과태료 부과를 명시하고 있다. 동법 시행령 제35조(개인정보 영향평가의 대상)는 평가 대상이 되는 개인정보파일에 대해 정의하며 자세한 사항은 자세한 사항은 〈표 2-28〉과 같다.

〈표 2-28〉 개인정보 영향평가 의무수행 대상

구분	세부 기준
1	구축·운용 또는 변경하려는 개인정보파일로서 5만명 이상의 정보주체에 관한 민감정보 또는 고유식별정보의 처리가 수반되는 개인정보파일
2	구축·운용하고 있는 개인정보파일을 해당 공공기관 내부 또는 외부에서 구축·운용하고 있는 다른 개인정보파일과 연계하려는 경우로서 연계 결과 50만명 이상의 정보주체에 관한 개인정보가 포함되는 개인정보파일
3	구축·운용 또는 변경하려는 개인정보파일로서 100만명 이상의 정보주체에 관한 개인정보파일
4	개인정보 영향평가를 받은 후에 개인정보 검색체계 등 개인정보파일의 운용체계를 변경하려는 경우 그 개인정보파일(이 경우 영향평가 대상은 변경된 부분으로 한정)

자료:「개인정보 보호법 시행령」 제35조(개인정보 영향평가의 대상)

법령상 규정된 대상시스템이 아니라도 대량의 개인정보 혹은 민감한 개인정보를 수집·이용하는 기관은 개인정보 유출 및 오·남용으로 인한 사회적 피해를 막기 위해 영향평가를 수행할 수 있으며, 공공기관 외의 개인정보처리자 또한 개인정보파일 운용으로 인하여 정보주체의 개인정보 침해가 우려되는 경우 영향평가를 하기 위하여 적극 노력하여야 한다.

개인정보 영향평가는 개인정보보호위원회가 지정한 영향평가기관에 의뢰하여 평가를 수행하여 그 결과를 개인정보보호위원회에 제출하고, 해당 날로부터 1년 이내에 개인정보 영향평가 개선사항 이행확인서 또한 개인정보보호위원회에 제출하여야 한다. 세부적인 개인정보 영향평가 절차는 사전준비 단계, 수행 단계, 이행 단계의 3단계로 구분할 수 있다. 사전준비 단계에서 영향평가의 필요성 등을 포함한 사업계획서의 작성, 평가기관 선정 등을 진행한다. 이후의 수행 단계에서 수행계획의 수립, 평가자료 수집, 개인정보 흐름 분석, 개인정보 침해요인 분석, 개선계획 수립을 순차적으로 진행하고 영향평가의 모든 과정과 산출물을 정리한 영향평가서와 요약본을 작성한다.

한편, 개인정보보호위원회는 정보주체인 국민이 영향평가 결과를 확인하기 어려웠던 문제를 개선하기 위해, 2024년 10월 31일 「개인정보 영향평가에 관한 고시」를 일부 개정하고 공공기관이 수행한 영향평가 결과의 요약본을 공개하도록 의무화하였다[25]. 공개 내용에는 평가 대상 시스템 및 개인정보 파일 개요, 개인정보 처리 흐름 분석, 평가 기준별 개인정보 침해 위험요인 분석 및 개선조치, 평가결과 등이 포함되며 영향평가서 요약본은 각 기관의 누리집을 통해 공개될 예정이다(개인정보보호위원회, 2023.10.13.)

25) 개인정보보호위원회에 따르면 2011년 제도 시행 이후 850여 개 기관에서 4,000여 건의 영향평가를 수행(2023.9월 말 기준)했으나 그 결과가 외부에 공개되지 않아 정보주체인 국민이 이를 확인하기 어려웠음을 지적하며 공공기관 개인정보 영향평가 결과에 대한 공개제도 도입 취지를 밝힘(개인정보보호위원회, 2023.10.13.)

〈표 2-29〉 영향평가 수행단계 세부 절차 및 내용

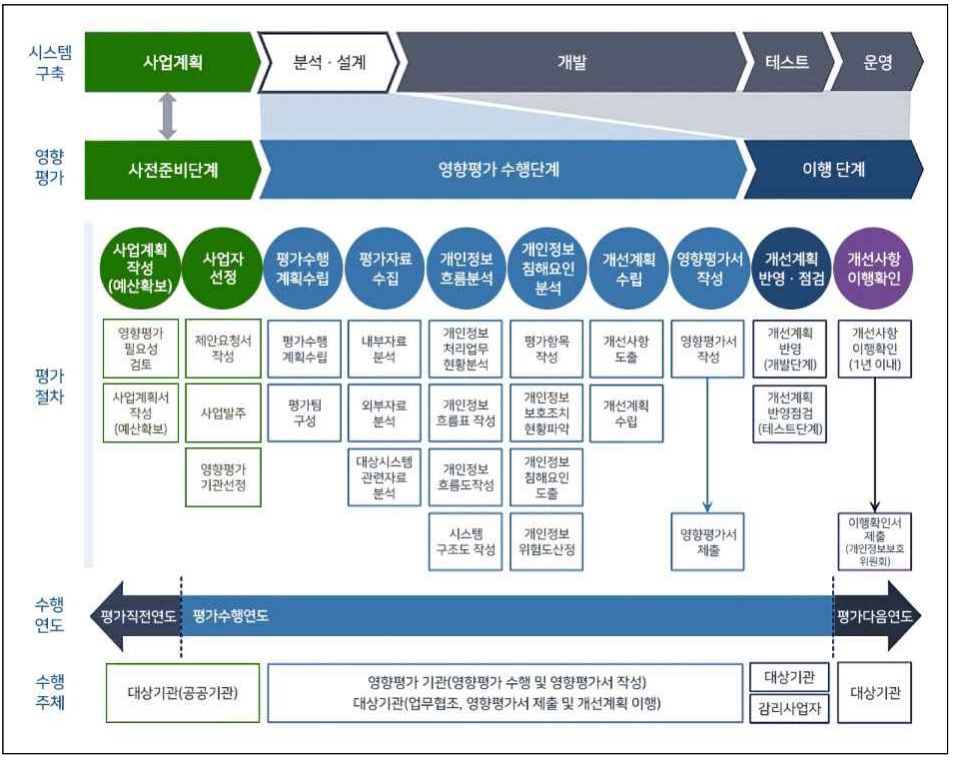
구분	목표 및 개요
1. 영향평가 수행계획 수립	(목표) 효율적인 평가 수행을 위해 사전 계획 수립 (개요) 사업 주관 부서가 영향평가 수행계획을 수립하고, 수립한 계획서는 착수 회의를 통하여 영향 평가팀은 물론 당해 사업과 관련하여 협조가 필요한 유관 부서, 외부기관, 외부 전문가 등과 공유
2. 평가자료 수집	(목표) 대상사업 및 개인정보보호 관련 기관 내·외부 정책환경 분석을 위한 자료 수집 (개요) 개인정보보호 관련 법규 및 상위기관의 지침과 해당기관 내부규정 현황 파악, 당해 사업 이해·분석하기 위해 필요한 자료 등 취합 및 분석
3. 개인정보 흐름 분석	(목표) 대상사업에서 처리되는 개인정보 흐름에 대한 파악을 위해 정보시스템 내 개인정보 흐름 분석 (개요) ① 개인정보 처리업무 분석 (개인정보 영향도 등급표, 개인정보 처리 업무표 및 업무 흐름도 작성) ② 개인정보 처리 업무표를 기반으로 개인정보 흐름표 작성 ③ 개인정보 흐름표를 기반으로 개인정보 흐름도 작성 ④ 기관 내 네트워크 및 보안시스템 구조 등을 분석하여 정보시스템 구조도 작성
4. 개인정보 침해요인 분석	(목표) 개인정보의 흐름에 따른 개인정보 조치사항 및 계획 등을 파악하고 개인정보 침해 위험성 도출 (개요) ① 평가기준 수립 및 개인정보보호 조치사항 파악을 위한 평가항목 작성 ② 자료 분석, 현장 및 시스템 실사, 담당자 인터뷰 등을 통해 개인정보보호 조치 현황 파악 ③ 파악한 조치 현황을 기반으로 개인정보 침해요인 도출 ④ 도출된 개인정보 침해요인에 대한 위험도 산정
5. 개선계획 수립	(목표) 개인정보 침해 요인별 위험도 분석에 기반하여, 위험요소를 제거하거나 최소화하기 위한 개선방안 및 개선계획을 수립 (개요) 도출된 개인정보 침해요소에 대해 기관 내 인력 및 예산 등 자원을 고려, 유관업무 담당자와의 협의를 거쳐 체계적으로 정비한 개선 계획 수립
6. 영향평가서 및 요약본 작성	(목표) 평가의 모든 과정 및 산출물을 정리하여 영향평가서 및 요약본 작성 (개요) 영향평가 추진경과 및 중간산출물 등의 내용을 정리하고 도출된 위험요소 및 개선계획 등 최종산출물들을 모두 취합하여 영향평가서 및 요약본을 작성하고, 대상기관은 보고서 품질을 검토함

자료: 개인정보보호위원회(2024)

개인정보 영향평가 이행 단계는 개인정보파일의 구축·운용 전 평가 수행 단계에서 설정한 개선사항의 반영여부 확인과 영향평가서를 제출받은 날로부터 1년 이내에 수행해야

하는 개선사항 이행 확인으로 구분된다. 후자는 영향평가 시 도출된 개선계획이 평가 수행기간 내에 완료되지 않은 경우, 필요한 절차로 개선계획이 예정대로 수행되고 있는지 점검하고, 점검 결과 미흡한 부분이 있다면 해당 원인을 분석하여 수립한 조치 방안을 포함한 '개인정보 영향평가 이행확인서'를 개인정보보호위원회에 제출한다. 이러한 영향평가 후속 절차는 해당 평가의 목적인 안전한 개인정보 처리 과정 설계 및 수행의 지속성과 실효성을 제고에 유의미한 영향을 미칠 것으로 예상된다.

〔그림 2-6〕 개인정보 영향평가 절차



자료: 개인정보보호위원회, 개인정보 영향평가제도[26]

26) https://www.pipc.go.kr/np/default/page.do?mCode=D050020000

4. 국가인권위원회 「인공지능 인권영향평가 도구」

국가인권위원회의 인공지능 인권영향평가 도구는 인공지능의 개발 및 활용으로 인한 잠재적 인권침해 및 차별 위험을 사전에 방지하거나 완화하기 위한 목적으로 인공지능 개발 및 활용 주체의 자율적인 평가 시행을 장려하고 지원하기 위해 마련되었다. 해당 평가도구는 계획 및 준비, 분석 및 평가, 개선 및 구제, 공개 및 점검 4단계로 구성되어 있으며 각 단계 별 세부적인 구성 및 문항수는 다음의 〈표 2-30〉과 같다.

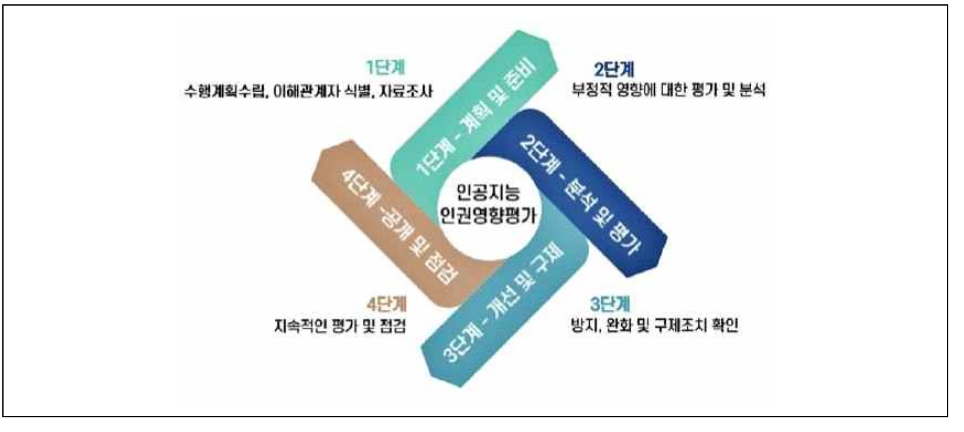
〈표 2-30〉 인공지능 인권영향평가 도구 구성 및 문항수

구분	주요 내용		문항수
1. 계획 및 준비	가. 인권영향평가 계획		14
	나. 사전 조사		5
2. 분석 및 평가	가. 인공지능 기술과 관련된 영향 분석 및 평가	(1) 개인정보보호	1
		(2) 데이터 관리	6
		(3) 알고리즘의 성능과 신뢰성	3
		(4) 차별금지	2
		(5) 설명가능성과 투명성	5
		(6) 자동화 정도와 인간의 개입	4
		(7) 보안	3
		(8) 접근성	2
		(9) 라이선스	1
	나. 인권에 미치는 영향 및 심각도	(1) 영향을 받는 인권	3
		(2) 인권에 미치는 영향의 심각도	4
3. 개선 및 구제	가. 방지		1
	나. 완화		4
	다. 구제		4
	라. 이해관계자와의 의견수렴 및 협의		2
4. 공개 및 점검	가. 인공지능 시스템의 주요 요소의 공개		1
	나. 인권영향평가 결과 공개		2
	다. 사후 모니터링		2
	라. 인권영향평가에 대한 점검		1
	마. 인권영향평가의 재수행		2

자료: 국가인권위원회(2024)

인공지능 인권영향평가는 평가의 대상을 공공기관과 민간을 분류하여 제시한다. 공공기관의 경우, 직접 개발·조달 하는 모든 인공지능 평가 대상으로 하며, 민간의 경우, 인권침해 위험성이 높은 고위험 인공지능을 평가 대상으로 한다. 평가 시기는 사전영향평가를 원칙으로 하되, 평가 결과에서 일정 기준 이상의 구체적인 위험이 확인된 경우 정기적, 사후적 평가를 통해 지속적 관리를 수행할 것을 권장한다. 평가의 수행 주체는 별도의 독립성을 가진 내부 조직 혹은 인권 분야 및 인공지능 기술과 관련한 전문성과 독립성을 갖춘 제3의 기관에서 수행하는 것이 바람직하다고 평가한다. 평가의 절차는 앞서 언급한 4단계를 순차로 진행하며 모든 단계에서 이해관계자의 참여를 권장한다. 평가 문항 내용은 해당 위원회에서 2022년 정부에 권고한 「인공지능의 개발과 활용에 관한 인권 가이드라인」에서 강조하는 인권 및 인간 존엄성의 존중에 기반하며, 형식은 주로 객관식이나 관련된 설명을 추가할 수 있도록 제시한다.

〔그림 2-7〕 인공지능 인권영향평가 이행단계



자료: 국가인권위원회(2024)

인공지능 인권영향평가 도구의 특징적인 요소 중 평가주체와 형식을 더욱 구체적으로 살펴보고자 한다. 해당 평가는 인공지능으로 발생 가능한 위험을 객관적으로 평가하기 위해 평가주체의 독립성을 강조한다. 또한 평가 전 과정에서 이해관계자의 의견수렴 및 협의를 권장하며, 이해관계자와 관련된 평가문항을 포함한다. 평가문항은 일부 주관식 문항

을 제외하면, 대부분 체크리스트 형식의 객관식으로 제시된다. 답변 항목 중 '정보 없음'을 제공하는 부분이 눈에 띄는데, 이는 질의에 답할 수 있는 관련 정보가 없어서 판단하기 힘들다는 것을 의미한다. 이를 체크하더라도 향후 관련 정보를 찾기 위해 추가적인 노력을 할 필요가 있으며, 해당 답변이 많은 경우 영향평가가 제대로 이루어지지 않을 가능성이 크다. 추가로 체크리스트와 함께 관련된 구체적 정보를 서술식으로 설명하는 공간을 제공한다. 이를 통해 점검항목을 충족하지 못하더라도 합당한 이유나 다른 방안을 제시할 수 있으며, 더 나아가 평가항목을 매개로 하는 이해관계자 사이의 소통과 토론, 그리고 서로의 역량을 강화하는 과정에 도움이 될 수 있을 것으로 예상된다.

〈표 2-31〉 인공지능 인권영향평가 문항 예시

Q1-1-2-2. 인권영향평가를 수행하는데 충분한 인적, 재정적 자원이 확보되어 있습니까? ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 설명 ( )
Q1-1-2-3. 평가팀은 인권영향평가를 수행하는데 필요한, 인공지능 기술 및 인권, 그리고 인공지능이 활용될 분야에 대한 전문성을 갖추고 있습니까. ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 평가팀의 구성, 팀원의 역할, 전문 분야 등을 설명하시오. 설명 ( )
Q1-1-2-4. 조직 내에 인권영향평가 수행의 요건, 주체, 절차 등을 상세히 규정한 내규가 있습니까? ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 설명 ( )
Q1-1-2-5. 인권영향평가 결과에 관한 보고체계(보고절차, 해당 결과의 수용 여부를 결정할 수 있는 조직 또는 최종 책임자 등)가 명확히 규정되어 있습니까? ☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음 설명 ( )

자료: 국가인권위원회(2024)

5. 소결

본 절은 국내에서 시행 혹은 제안된 영향평가 제도를 살펴보고 본 연구의 AI 영향평가 프레임워크와 관련된 시사점을 도출하고자 한다. 이를 위해 과기정통부에서 주관하는 기술영향평가와 인공지능 윤리영향평가, 개인정보보호위원회 주관 개인정보 영향평가, 국가인권위원회의 인공지능 인권영향평가를 검토하였다. 〈표 2-31〉에서 각 영향평가의 개요를 확인 및 비교할 수 있다.

기술영향평가와 인공지능 윤리영향평가는 평가 목적 측면에서 인공지능 영향평가와 유사하므로 면밀한 검토가 필요하다. 세 가지 영향평가는 단순한 위험 관리 차원을 넘어, 평가 대상이 초래할 수 있는 긍정적 영향은 극대화하고 부정적 영향은 최소화하는 데 목적이 있다. 특히, 인공지능 윤리영향평가는 평가의 대상, 지표, 목적의 측면에서 인공지능 영향평가와 가장 공통점이 많고 면밀한 검토가 필요한 제도이다. 기술영향평가는 정부기관이 평가를 주도하고, 평가 대상의 층위에서도 인공지능 영향평가와 차이를 보이므로, 세부적인 평가 방식과 구성은 이러한 특성을 반영하여 별도로 설계될 필요가 있다.

개인정보 보호 및 데이터 보안이 개인정보 영향평가의 주요 항목으로 포함되어 있는 점을 감안할 때, 인공지능 영향평가와의 중복 가능성에 대해 면밀한 검토가 필요하다. 유사한 영향평가의 복수 수행으로 인한 기업의 부담 완화를 위해 중복 가능성이 있는 항목의 제거, 혹은 일부 항목의 평가 면제 등의 방안에 대한 고려가 필요하다. 개인정보 영향평가에서 평가 이후 이행 확인 절차가 안전조치의 실효성을 확보하고 전체적인 신뢰도 향상에 기여하고 있는 점을 고려할 때, 유사한 절차를 도입하는 방안을 검토할 필요가 있다.

인공지능 인권영향평가 역시 평가 대상 및 지표 측면에서 본 영향평가와 공통되는 부분이 있으므로, 이에 대한 검토가 필요하다. 다만, 해당 제도는 부정적 영향의 식별과 이에 대한 대응 조치에 중점을 두고 있어, 평가의 목적 및 관점에서 본 영향평가와 차이가 존재한다. 그럼에도 불구하고, 체크리스트 형식은 평가 주체의 부담을 줄이고 평가의 효율성을 높일 수 있으므로, 이러한 방식의 도입 여부는 고려할 필요가 있다.

〈표 2-32〉 국내 AI 영향평가 유사 제도 비교

구분	기술영향평가	인공지능 윤리영향평가	개인정보 영향평가	인공지능 인권영향평가
관련 근거	과학기술기본법 제14조	지능정보화기본법 제56조, 제62조, 인공지능(AI) 윤리기준	개인정보보호법 제33조	인공지능 개발과 활용에 관한 인권 가이드라인
평가 주체	- **정부**	- **정부주도*** * 평가과정에 전문가(평가위원단), 시민사회(국민포럼, 소통채널) 등 이해관계자 참여	- **개인정보위원회가 지정한 영향평가 기관**	- **대상 AI 시스템 개발·활용 기관*** * 독립된 별도의 내부 조직 또는 제3의 기관의 평가 수행 권장
평가 대상	- **새로운 과학기술*** * 미래 신기술 및 기술·경제·사회적 영향과 파급효과 등이 큰 기술	- **AI 제품·서비스군*** * 평가대상 매년 선정, 고위험 AI 우선 고려	- **공공기관이 도입·운영하는 개인정보파일 운용 시스템 등*** * 특정 요건 충족 시 의무화	- **공공기관이 직접 개발하거나 조달하는 모든 AI 시스템** - **민간에서 활용하는 인권침해 위험성이 높은 고위험 AI**
평가 지표	- **기술의 사회적 영향(경제, 사회, 문화, 윤리, 법률·규제, 환경, 사용자 특성 등*)** * 기술 특성을 고려해 지표 추가	- **AI 윤리기준 10대 핵심요건*** * 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성)	- **개인정보보호 관련 법적 요건 만족, 취급하는 개인정보에 대한 침해 우려 발생 가능성 여부** - **영향평가 점검표* 기준 주요 침해요인 분석 및 개선** * 대상 기관 및 시스템의 개인정보보호 관리체계, 개인정보 처리단계별 보호조치, 기술적 보호조치, 특정 IT 기술 활용 시 개인정보보호	- **AI의 기술적 특성*과 인권** * 개인정보보호, 데이터 관리, 알고리즘 성능·신뢰성, 차별금지, 설명 가능성과 투명성, 자동화 정도 및 개입, 보안, 접근성, 라이선스
평가 방법	- **전문가(위원회), 시민(시민포럼) 담론 중심 정성평가*** * 예상 파급효과와 그에 따른 정책 제언을 정성적으로 제시	- **전문가 위원단 정량·정성평가 (핵심요건별 영향 예상 및 긍정·부정 영향 진단*)** * 영향력, 범위, 규모, 발생 가능성, 회복 수준 등 UNESCO 접근 방식 - **평가과정에 국민포럼(PGI), 온라인 의견수렴 병행**	- **(공공기관)** * ① 평가 타당성 검토, 평가기관 선정, 평가수행 계획수립 등 영향평가 사전준비, ② 평가기관 평가 결과·기반 침해 요인 개선계획 반영·점검 - **(평가기관)** * 개인정보 침해요인 분석, 개선계획 수립, 영향평가서 작성	- **개발·도입 주체의 자체 정량·정성평가*** * (기술적 특성) 위험 식별 및 기술적 조치 여부 점검, (인권) 영향을 받는 인권 식별, 영향의 심각도 평가
비고	- **관계부처에 결과를 통보하고, 각 소관 분야별 정책 반영 실적 확인** * 과학기술기본법 시행령 제23조 제7항	- **AI의 윤리적 영향력 및 관리·제도·정책적 조치방안에 관한 체계적 참고자료를 모든 사회구성원에게 제공하는 것을 목표로 함**	- **부처 지정기관을 통해 영향평가를 수행함으로써 전문성·객관성 확보** - **전 과정 제도화를 통해 평가 실효성, 절차적 정당성 확보**	- **평가 법제화 전까지 AI 개발·활용 주체(공공·민간)의 자율적 평가 시행 지원을 목표로 함**

자료: 연구진 작성

제 3 장 AI 영향평가 프레임워크(안)[27]

제 1 절 AI 영향평가 프레임워크(안) 마련 배경

1. AI 영향평가 시행의 법적 근거

「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 'AI 기본법')은 2020년 7월 국회에 처음 발의된 이후, 4년이 넘는 기간 동안 다양한 의견을 수렴하며 논의가 이어졌고, 22대 국회에서 여야 합의를 통해 2024년 12월 본회의를 통과하였다. 이후 2025년 1월 국무회의 의결 등 절차를 거쳐, 2026년 1월 22일부터 시행될 예정이다.

AI 기본법은 "AI의 건전한 발전을 지원하고 AI 사회의 신뢰 기반 조성에 필요한 기본적인 사항을 규정함으로써 국민의 권익과 존엄성을 보호하고 국민의 삶의 질 향상과 국가경쟁력을 강화하는데 이바지할 수 있는 대한민군 AI의 새로운 기준을 마련하고자" 제정되었다. 본 법을 통해 정부는 민간이 자율적으로 추진하는 AI 안전성·신뢰성 검·인증, AI 영향평가에 대한 정부의 지원 근거를 마련하였다.

〈표 3-1〉 AI 기본법 제35조(고영향 인공지능 영향평가)

제35조(고영향 인공지능 영향평가) ① 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다. ② 국가기관등이 고영향 인공지능을 이용한 제품 또는 서비스를 이용하려는 경우에는 영향평가를 실시한 제품 또는 서비스를 우선적으로 고려하여야 한다. ③ 그 밖에 영향평가의 구체적인 내용·방법 등에 관하여 필요한 사항은 대통령령으로 정한다.

자료: 국가법령정보센터, https://www.law.go.kr

27) 본 보고서의 관련 내용은 AI 영향평가 가이드라인이 아직 확정되지 않은 상태에서, 2025년 4월 기준으로 TF에서 논의 중인 작업 단계의 자료를 참고하여 작성되었음. 향후 가이드라인 최종안이 확정됨에 따라 내용이 변경될 수 있음을 유의

2. AI 영향평가 시행의 취지와 목적

AI 영향평가는 AI 개발과 활용과정에서 인간의 존엄성과 기본권을 존중하는 것을 기본 지향으로 삼고 있다. 특히, AI 기본법은 제34조 등을 통해 고영향 AI 사업자에게 기본적인 책무를 부과하고 있으나, 이만으로는 인간 존엄성, 자유, 평등, 사회·경제적 기본권 등 국민의 주요 기본권을 충분히 보호하는 데 한계가 있을 수 있기 때문에 법은 제35조를 통해 별도의 AI 영향평가 제도를 마련하였다.

AI 영향평가는 사업자가 고영향 AI 또는 이를 이용한 제품·서비스를 제공하기에 앞서, 해당 제품·서비스가 국민의 기본권에 미칠 수 있는 영향을 사전에 식별하고 필요한 완화 조치를 마련함으로써 기본권 침해를 예방하는 것을 목적으로 한다. 이를 통해 고영향 AI가 초래할 수 있는 위험에 선제적으로 대응하고, AI 신뢰성을 기반으로 한 선순환적 리스크 관리체계를 구축하고자 한다.

또한, 우리나라의 AI 영향평가 제도는 글로벌 수준의 상호운용성을 확보하기 위한 고려 아래 설계되었다. 유럽연합(EU) 역시 인간 존엄성, 자유, 평등, 시민의 권리 등 "유럽연합 기본권 헌장"에 따라 보장되는 기본권 보호를 위해, EU AI Act 제27조에서 '기본권 영향평가(Fundamental Rights Impact Assessment, FRIA)' 제도를 의무화(2026년 8월 시행 예정)하고 있으며, 우리나라의 AI 영향평가는 이러한 국제적 논의 흐름을 반영하여 글로벌 시장에서도 통용될 수 있는 기본권 보호 기준을 갖추는 것을 지향하고 있다.

제 2 절 AI 영향평가 가이드라인 TF

법 제정과 함께, 과학기술정보통신부는 AI 기본법 제정의 효과를 민간에서 조속히 체감할 수 있도록 하기 위해, 하위법령 정비와 병행하여 법에 근거한 주요 방침과 고시를 마련하는 별도 전담반(TF)을 운영하였다. 다음에서는 'AI 영향평가 가이드라인' TF의 운영 현황과 주요 논의사항을 살펴보고자 한다.

1. AI 영향평가 가이드라인 TF 운영 현황

AI 영향평가 가이드라인 TF는 법조계, 산업계, 학계·연구계, 시민단체, 공공분야의 인공지능 기술, 윤리, 법제도, 정책 전문가로 구성되었다. 구체적인 명단은 다음과 같다.

〈표 3-2〉 AI 영향평가 가이드라인 TF 위원 명단(2025.4 기준)

구 분	성명	소속
법조계(3)	변동훈	김&장 법률사무소 변호사
	김태주	법무법인 광장 변호사
	박일현	법무법인 율촌 변호사
산업계(6)	김동환	포티투마루 대표
	이영복	제네시스랩 대표
	하주영	스캐터랩 변호사
	김경훈	카카오 이사
	류승훈	네이버 변호사
	박형준	한국정보통신진흥협회 팀장
학계·연구계(3)	문명재	연세대 행정학과 교수
	이상욱	한양대 철학과 교수
	선지원	한양대 법학전문대학원 교수
시민단체(1)	손지원	오픈넷 변호사
공공(2)	-	과학기술정보통신부
	문정욱	정보통신정책연구원 실장

자료: 연구진 작성

AI 영향평가 가이드라인 TF는 2025년 2월 킥오프 회의를 시작으로 단기간 내 두 차례 논의를 진행하였다. 각 회의의 개요는 다음과 같다.

AI 영향평가 가이드라인 TF 킥오프 회의는 2025년 2월 27일(목) 씨이오스위트 광화문교보센터에서 개최되었다. AI 영향평가 대상 기본권의 범위, 평가의 대상과 결과 형태 등의 AI 영향평가 체계 구축을 위한 토론이 주요하게 진행되었다. 또한 AI 영향평가의 의무화 작용 가능성과 타 평가와 관계 정립 필요성을 짚어보는 등 다양한 논의를 진행하였다.

〈표 3-3〉 AI 영향평가 가이드라인 TF 1차 회의

- 일시 : 2025.2.27.(목), 16:00~18:00
- 장소 : 씨이오스위트 광화문교보센터
- 참석 : 과기정통부 인공지능기반정책과장, 담당 사무관, AI 영향평가 가이드라인 TF 위원, KISDI 관계자 등 15인 내외
- 내용 : AI 영향평가 추진 경과 설명, AI 영향평가 가이드라인 마련 방안 등 논의 진행

	주요 내용	비고
1	AI 영향평가 추진경과 설명	참석자 전원
2	기본권 개념에 대한 논의	
3	가이드라인 목차 구성안 논의	

자료: 연구진 작성

AI 영향평가 가이드라인 TF 2차 회의는 2025년 3월 13일(목) HJBC 광화문에서 개최되었으며, 사전에 진행된 의견조사의 결과를 공유하고 이에 대한 토론과 AI 영향평가 체계 논의를 중심으로 진행되었다. AI 영향평가 가이드라인 TF 의견조사는 2차 TF 회의 이전인 3월 7일(금)까지 실시되었고 그 내용은 AI 영향평가를 위한 기본권 범주에 대한 의견과 AI 기본법상의 '고영향 AI' 영역에서의 AI 시스템 활용으로 인해 침해될 수 있는 기본권 시나리오 작성으로 구성되었다. 2차 TF 회의는 해당 의견조사 결과와 이전 차수의 회의 결과를 바탕으로 도출한 AI 영향평가 프레임워크(안)에 대한 토론을 진행하였다. 해당 회의는 AI 영향평가 및 가이드라인과 관련하여, 가이드라인 서술 방식, 평가결과 검증 필요성 및 공개 여부, 결과에 따른 환류 조치 도입 여부가 주로 논의되었다. TF 운영과 관련하여 AI 기본법 시행령 및 타 조항의 가이드라인과 연계·협력을 통한 정합성 제고와 명확한 역할 분담을 통한 평가 부담 완화의 중요성 또한 제기되었다.

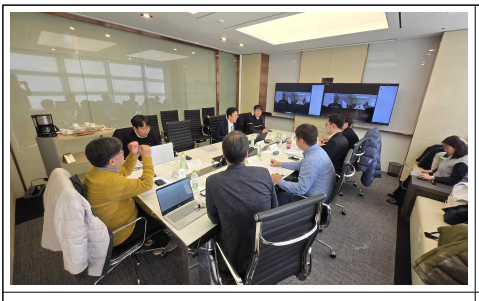
〈표 3-4〉 AI 영향평가 가이드라인 TF 2차 회의

- 일시 : 2025.3.13.(목), 10:00~12:00
- 장소 : HJBC 광화문
- 참석 : 과기정통부 담당사무관, AI 영향평가 가이드라인 TF 위원, KISDI 관계자 등 15인 내외
- 내용 : 기본권 범주, AI 영향평가 체계 등에 관한 논의 진행

	주요 내용	비고
1	사전 의견조사 결과 논의	참석자 전원
2	AI 영향평가 체계 논의	

자료: 연구진 작성

〔그림 3-1〕 AI 영향평가 가이드라인 TF 회의



1차 회의(2.27)

2차 회의(3.13)

자료: 연구진 작성

2. AI 영향평가 가이드라인 TF 논의사항

가. AI 영향평가 명칭

AI 기본법 제35조를 근거로 시행하는 평가인 만큼, 명칭은 법령에서 사용하고 있는 표현과 동일하게 'AI 영향평가'로 통일하기로 하였다. 또한, 법령에서 서술하고 있는 기본권 보호 취지를 반영하여, 평가 방법론은 기본권을 중심으로 구축하기로 하였다. 평가 범위에 있어서는 AI 서비스의 개발·활용 과정과 밀접한 관련이 있는 기본권을 주요 대상으로 삼되, 기본권은 아니지만 그 특성과 밀접하게 연관된 주요 권리도 함께 논의 대상으로 포함한다.

나. 기본권

AI 기본법상 고영향 AI 영역과 밀접한 관련이 있는 기본권을 중심으로 영향평가 체계를 구축하기로 하였다. 대한민국 헌법, 유럽연합 기본권 헌장 등에서 규정하는 기본권의 범위가 매우 광범위하여 모든 기본권을 일괄적으로 평가하는 데 현실적 한계가 존재하기 때문이다.

〈표 3-5〉 「AI 기본법」 제2조(정의)

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.

4. "고영향 인공지능"이란 사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로서 다음 각 목의 어느 하나의 영역에서 활용되는 것을 말한다.

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: 국가법령정보센터, https://www.law.go.kr

〈표 3-6〉 대한민국 헌법상 기본권

권리		조항	주요 내용
인간의 존엄과 가치 및 행복추구권		제10조	• 모든 국민은 인간으로서의 존엄과 가치를 가지며, 행복을 추구할 권리가 있음 • 생명권, 인격권, 일반적 행동의 자유, 자기결정권 등의 기본권 도출
평등권		제11조	• 모든 국민은 법 앞에 평등 • 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별받지 않음
자유권	신체의 자유	제12조	• 모든 국민은 신체의 자유를 가짐 • 법률에 의하지 않고서는 체포·구속·압수·수핵 또는 심문을 받지 않으며, 법률과 적법한 절차에 의하지 아니하고는 처벌·보안처분 또는 강제노역을 받지 않음 • 고문받지 않을 권리와 진술거부권, 변호인의 조력을 받을 권리, 체포·구속적부심사청구권 등 포함
	법적 불이익처우로부터의 자유	제13조	• 형벌불소급의 원칙, 죄형법정주의, 일사부재리의 원칙, 소급입법에 의한 참정권 및 재산권 박탈 금지, 연좌제 폐지 등 규정
	거주·이전의 자유	제14조	
	직업선택의 자유	제15조	
	주거의 자유	제16조	
	사생활의 비밀과 자유	제17조	
	통신의 비밀	제18조	
	양심의 자유	제19조	
	종교의 자유	제20조	• 국교는 인정되지 아니하며, 종교와 정치는 분리
	언론·출판의 자유와 집회·결사의 자유	제21조	• 언론·출판에 대한 허가나 검열, 집회·결사에 대한 허가는 인정되지 않음 • 언론·출판은 타인의 명예나 권리 또는 공중도덕이나 사회윤리를 침해해서는 안 되며, 침해한 때에는 피해자가 이에 대한 피해의 배상을 청구할 수 있음
	학문과 예술의 자유	제22조	• 저작자·발명가·과학기술자와 예술가의 권리는 법률로써 보호
	재산권	제23조	• 그 내용과 한계는 법률로 정함 • 재산권의 행사는 공공복리에 적합하도록 해야 함 • 공공필요에 의한 재산권의 수용·사용 또는 제한 및 그에 대한 보상은 법률로써 하되, 정당한 보상 지급

권리		조항	주요 내용
참정권	선거권	제24조	
	공무담임권	제25조	
청구권	청원권	제26조	
	재판청구권	제27조	• 모든 국민은 헌법과 법률이 정한 법관에 의하여 법률에 의한 재판을 받을 권리를 가지며, 신속한 재판을 받을 권리를 가짐 • 형사피고인은 지체없이 공개재판을 받을 권리를 가지며, 유죄 판결이 확정될 때까지 무죄로 추정 • 피해자는 재판절차에서 진술할 수 있음
	형사보상청구권	제28조	• 형사피의자 또는 형사피고인으로서 구금되었던 자가 법률로 정하는 불기소처분 또는 무죄판결을 받은 때에는 국가에 정당한 보상을 청구할 수 있음
	국가배상청구권	제29조	• 종교와 정치는 분리공무원 직무상 불법행위로 손해를 받은 국민은 국가 또는 공공단체에 정당한 배상을 청구할 수 있음
	범죄피해자 구조청구권	제30조	• 타인의 범죄행위로 생명·신체에 대한 피해를 받은 국민은 법률이 정하는 바에 의해 국가로부터 구조받을 수 있음
사회권	교육받을 권리	제31조	• 모든 국민은 능력에 따라 균등하게 교육 받을 권리가 있음 • 의무교육(초등)은 무상으로 실시
	근로의 권리	제32조	• 국가는 고용증진 및 적정임금의 보장을 위해 노력해야 하며, 최저임금제 시행해야 함 • 근로의무 내용과 조건은 민주주의 원칙을 따라야 하며, 근로조건 기준은 인간 존엄성을 보장하도록 법률로 정함
	근로3권	제33조	• 단결권, 단체교섭권, 단체행동권
	인간다운 생활을 할 권리	제34조	• 국가는 사회보장·복지의 증진에 노력해야 하며, 재해를 예방하고 그 위험으로부터 국민을 보호
	환경권	제35조	• 모든 국민은 건강하고 쾌적한 환경에서 생활할 권리를 가지며, 국가와 국민은 환경보전을 위해 노력해야 함
	혼인과 가족, 보건에 관한 권리	제36조	• 혼인과 가족생활은 개인의 존엄과 양성의 평등을 기초로 성립·유지되어야 하며, 국가는 이를 보장 • 국가는 모성의 모호를 위해 노력 • 국민은 보건에 관하여 국가의 보호를 받음

자료: 「대한민국헌법」을 토대로 연구진 정리

〈표 3-7〉 유럽연합 기본권 헌장

Title Ⅰ : Dignity 존엄성	Title Ⅱ : Freedoms 자유
01-Human dignity 인간의 존엄 02-Right to life 생명권 03-Right to the integrity of the person 인간 고유성에 대한 권리 04-Prohibition of torture and inhuman or degrading treatment or punishment 고문과 비인간적·굴욕적 대우·처벌 금지 05-Prohibition of slavery and forced labour 노예 및 강제노동 금지	06-Right to liberty and security 자유와 안전에 대한 권리 07-Respect for private and family life 사생활·가족생활 존중 08-Protection of personal data 개인정보 보호 09-Right to marry and right to found a family 혼인의 권리와 가정 형성의 권리 10-Freedom of thought, conscience and religion 사상·양심·종교의 자유 11-Freedom of expression and information 표현과 정보의 자유 12-Freedom of assembly and of association 집회와 결사의 자유 13-Freedom of the arts and sciences 예술과 과학의 자유 14-Right to education 교육의 권리 15-Freedom to choose an occupation and right to engage in work 직업선택의 자유와 근로권 16-Freedom to conduct a business 영업의 자유 17-Right to property 재산권 18-Right to asylum 망명권 19-Protection in the event of removal, expulsion or extradition 이동, 추방, 인도에 대한 보호
Title Ⅲ : Equality 평등	**Title Ⅳ : Solidarity 연대**
20-Equality before the law 법 앞의 평등 21-Non-discrimination 차별금지 22-Cultural, religious and linguistic diversity 문화적·종교적·언어적 다양성 23-Equality between women and men 여성과 남성 간 평등 24-The rights of the child 아동의 권리 25-The rights of the elderly 노인의 권리 26-Integration of persons with disabilities 장애인의 통합	27-Workers' right to information and consultation within the undertaking 근로자의 기업 내 정보·협의권 28-Right of collective bargaining and action 단체교섭 및 단체행동권 29-Right of access to placement services 직업소개 서비스 접근권 30-Protection in the event of unjustified dismissal 부당해고로부터 보호 31-Fair and just working conditions

	공정하고 정당한 근로조건 32-Prohibition of child labour and protection of young people at work 아동노동 금지와 직장에서 연소자 보호 33-Family and professional life 가족과 직업생활(의 조화) 34-Social security and social assistance 사회보장 및 사회부조 35-Health care 의료 36-Access to services of general economic interest 일반 경제이익 서비스에 대한 접근 37-Environmental protection 환경보호 38-Consumer protection 소비자 보호
Title V : Citizens' right 시민의 권리	Title VI : Justice 사법
39-Right to vote and to stand as a candidate at elections to the European Parliament 유럽의회 선거에서의 선거권·피선거권 40-Right to vote and to stand as a candidate at municipal elections 지방선거에서 선거권·피선거권 41-Right to good administration 좋은 행정에 대한 권리 42-Right of access to documents 문서에 대한 접근권 43-European Ombudsman 옴부즈맨 44-Right to petition 청원권 45-Freedom of movement and of residence 이동과 거주의 자유 46-Diplomatic and consular protection 외교적·영사적 보호	47-Right to an effective remedy and to a fair trial 효과적 구제와 공정한 재판의 권리 48-Presumption of innocence and right of defence 무죄 추정과 변호권 49-Principles of legality and proportionality of criminal offences and penalties 죄형법정주의 및 범죄와 형벌의 비례원칙 50-Right not to be tried or punished twice in criminal proceedings for the same criminal offence 일사부재리의 권리

자료: EU(2012)를 토대로 연구진 정리

어떤 기본권을 집중적으로 다루어야 할지에 대해서는 TF 논의를 통해 결정하며, 이를 위해 기본권 범주를 도출하고, AI 제품·서비스의 실제 기본권 침해 사례를 식별하는 등 논의를 진행하였다.

〈표 3-8〉 기본권 범주 도출

기본권 범주	근거 및 주요 내용	비고
인간존엄 및 인격권 일반	• 인간으로서의 존엄과 가치 • 우리 헌법 제10조를 비롯하여, 근대 헌법들이 기본권조항의 근간으로 삼고 있는 개념	• 사실상 다른 기본권을 포괄할 수 있는 최상위의 개념으로, 영향평가 시에는 다른 구체적인 기본권 침해를 모두 검토한 후, 최종 단계에서 다룰 필요가 있음
평등권	• 법 앞의 평등, 차별받지 않을 권리 • 헌법 제11조(평등권)	• 편향된 데이터로 인해 특정 계층에 유리한 결정 도출 • 인공지능이 도출하는 결과의 평등뿐만 아니라, 인공지능이 생산하는 편익의 수혜 측면에서의 평등도 고려
인신의 안전	• 신체적·정신적 위해를 당하지 않고 안전하게 AI 서비스를 활용할 권리 • 헌법 제12조(신체의 자유) • 헌법 제34조(인간다운 생활을 할 권리) • IT기본권(독일연방헌법재판소)	• 의료 인공지능의 활용에 따라 발생하는 신체적 침해 • 자율주행차가 야기하는 사고 • 거짓 정보나 명예훼손에 해당하는 정보를 생산하여 유포하는 경우 • AI 서비스가 원인이 되는 위해의 방지에만 국한
프라이버시권	• AI 서비스의 준비(데이터분석 등) 및 실행 과정에서 사생활을 침해당하지 않을 권리 • 헌법 제17조(사생활의 비밀과 자유) • 헌법 제18조(통신의 비밀) • 개인정보자기결정권(헌법재판소)	• 개인의 사상이나 신념을 포함한 민감정보까지도 분석의 대상으로 삼아 원치 않는 정보의 노출이 이루어지는 경우 • 주로 보안 관련 AI 시스템에서 개인의 사생활을 과도하게 통제하는 경우
행동과 표현의 자유	• 이용자 자신에 관한 사항을 스스로 결정하고, 이용자의 자의에 따른 표현을 제약당하지 않을 권리 • 헌법 제12조(신체의 자유) • 헌법 제21조(표현의 자유) • 헌법 제24조(선거권)	• 이용자의 의사와 부합하지 않는 왜곡된 결과물을 생산하여 배포하는 경우 • 미디어 추천시스템을 통해 편향된 정보만 습득시 왜곡된 정치적 의사결정
사회·경제적 기본권	• AI 서비스로 인해 부당한 재산적 침해를 받지 않을 뿐만 아니라, 실질적인 사회·경제생활에 방해를 받지 않을 권리 • 헌법 제15조(직업의 자유) • 헌법 제23조(재산권) • 헌법 제32조(근로의 권리) • 헌법 제34조(인간다운 생활을 할 권리) • 헌법 제35조(환경권)	• 취업 면접 내지 업무 평가에서의 AI 활용을 통해, 대상자의 직업 수행을 제한하게 되는 경우 • AI 시스템이 이용자의 편의에만 초점을 맞춘 나머지, 사회 질서나 공익적 가치에는 위해가 되는 경우

자료: 연구진 작성

다. AI 영향평가 대상 설정

서비스 개별 특성이 아닌, AI 기본법상 고영향 영역 여부만을 기준으로 특정 AI가 영향평가 대상에 해당하는지를 일률적으로 판단하는 데에는 신중한 접근이 필요하다. 동일한 영역에 해당한다고 하더라도 AI의 목적, 모델, 프로세스, 산출물 등에 따라 위험성의 정도가 달라질 수 있기 때문이다. 따라서 단순히 고영향 영역에 속한다는 이유만으로 일괄적으로 평가를 실시하는 것이 적절한지에 대해 추가적인 검토가 필요하다. 서비스군을 기준으로 영향평가 대상을 구분하는 경우, 평가 대상 여부를 사전에 예측할 가능성은 높아질 수 있다. 그러나 이 경우에도 보다 명확한 판단 및 분류 기준이 마련되어야 한다는 점이 함께 제기되었다.

한편, 고영향 AI의 기준 및 예시에 대한 세부 사항은 별도로 운영 중인 '고영향 AI 기준과 예시 가이드라인 TF'를 통해 논의가 진행되고 있다. 이에 따라, 본 TF에서는 고영향 제품 또는 서비스가 이미 특정되었다는 전제하에 영향평가의 프로세스, 방법론, 조치사항 마련 등 구체적 평가 체계 수립에 집중하여 논의를 이어가기로 하였다.

라. AI 영향평가 결과의 문서화 방식 및 공개 여부

AI 영향평가 가이드라인에는 평가 결과의 문서화 방식과 공개 형태에 대한 권고사항을 포함하는 방안이 논의되었다. 현재 시행령이나 가이드라인을 통해 평가 결과 공개를 법적 의무로 규정할 계획은 없으나, 평가 결과 공개 시 바람직한 방식과 형태를 가이드라인에 명시하는 것은 긍정적으로 검토되고 있다.

TF 논의 과정에서 다양한 의견이 제시되었다. 우선 학계 C위원은, 평가 결과 보고서를 홈페이지 등에 공개하는 것을 장려하는 문구를 가이드라인에 포함할 필요가 있다고 제안하였다. 이는 국제 AI 거버넌스의 기본원칙인 '투명성' 원칙에 부합하며, 법률상 강제는 어렵더라도 권고를 통해 사업자들이 자발적으로 이행할 수 있도록 해야 한다는 취지이다. 시민단체 A위원 또한, 가이드라인은 본래 권고적 효력을 지니므로, 법적 강제성을 갖지 않더라도 평가 결과 공개를 '원칙'으로 천명하는 것이 바람직하다는 의견을 제시하였다. 이는 입법 취지, 기업의 책임성 강화, 신뢰성 확보, 이용자 보호, 알 권리 보장 등의 측면에서 평가 결과 공개의 중요성을 강조하는 것이다. 반면, 산업계 C위원은 평가 결과 공개를 사실상 의무화하는 방향은 자율적으로 영향평가를 시행하는 사업자들에게 부담을 줄 수

있다고 지적하였다. 이에 따라, 가이드라인에서는 “평가 결과에 대한 투명한 관리를 권장한다”는 형태로 표현하고, 결과 공개를 가능한 방안 중 하나로 제시하는 것이 적절하다는 의견이 제시되었다.

또한, 평가 결과의 문서화 범위에 대해서도 추가 논의가 이루어졌다. 단순히 영향평가 결과만을 기록하는 데 그치지 않고, 기업의 사후 조치 계획 및 시행 사항을 함께 포함하는 것이 필요하다는 의견이 제기되었다. 특히, 과학기술정보통신부와 정보통신정책연구원의 ‘AI 윤리 자율점검표’[28] 형태처럼 평가서를 구성할 경우, 개발자 등 현장의 실무자 부담을 줄이고 사후 조치 수립의 편의성을 높일 수 있을 것으로 기대된다. 다만, AI 영향평가 결과 보고서는 단순 자율점검표와 구분되기 위해, 평가 결과 분석뿐만 아니라 이에 대한 기업의 사후 대응 계획(또는 조치 시행 내역)까지 포함하도록 차별화할 필요성이 제기되었다.

한편, 사후 대응 사항까지 결과 보고서에 포함할 것인지에 대해서는 일부 이견도 존재하였다. 일부 위원은 별도로 의무화하지 않더라도 기업이 자율적으로 해당 내용을 포함할 가능성이 높다는 점을 들어, 강제성 부여에 신중해야 한다는 의견을 제시하였다.

TF는 AI 영향평가가 단순히 평가 결과에 그치는 것이 아니라, 소비자와 이용자가 평가를 통해 안정성과 신뢰성을 체감할 수 있도록 하는 방향으로 결과 관리가 이루어져야 한다는 데 공감하였다.

마. AI 영향평가 결과의 검증

AI 기본법 제35조 제2항과 관련하여, 평가 결과의 객관성을 확보하기 위한 방안에 대한 논의가 이루어졌다. TF에서는 평가의 신뢰성을 제고하고 영향평가의 실효성을 담보하기 위해 다양한 의견을 교환하였다.

법조계 C위원은 평가 결과의 객관성을 담보하기 위해 제3자가 평가 결과를 감수하는 절차를 도입할 필요가 있다고 제안하였다. 이에 대해 학계 C위원은 ISO, IEEE 등의 국제표준 사례를 참고하여, 이번 가이드라인에서는 우선 자율적인 평가 방법론을 제시하되, 향후 공공부문 또는 국제적 차원에서 검증 절차 도입을 대비할 필요성이 있음을 함께 설명하는

28) 과학기술정보통신부와 정보통신정책연구원은 2022년 「인공지능 윤리기준 실천을 위한 자율점검표」를 처음으로 마련·공개했으며, 이를 기반으로 매년 산업 분야별 특성을 반영한 자율점검표와 함께 개정안을 지속적으로 발표 중(최신버전 : 2025.2)

방안을 고려할 수 있다고 제시하였다. 또한, 학계 C위원은 장기적으로 AI 영향평가 전문 역량을 갖춘 주관기관(KISDI 등)을 지정하고, AI 영향평가 센터를 설립하여 전문성을 지속적으로 축적할 필요가 있다는 추가 의견을 제시하였다. 시민단체 A위원은 평가의 객관성과 공정성을 보다 확실히 담보하기 위해, 평가 수행 주체를 제3의 독립기관이나 외부기관, 또는 다양한 외부 전문가들로 구성된 위원회 형태로 운영할 수 있도록 가이드라인에서 방향성을 제시하는 방안을 고려할 수 있다고 하였다. 한편, 산업계 C위원은 평가 결과에 대한 별도의 검증 의무를 부과하는 것보다는, 국가기관 등이 영향평가 실시 여부를 고려할 때 적절한 가이드라인을 통해 사업자들의 자연스러운 참여를 유도하는 것이 바람직하다는 입장을 표명하였다.

다만, AI 서비스의 정부 조달 과정에서 영향평가 결과를 어떻게 반영할 것인지에 대한 절차나 방식 등은 본 TF의 논의 범위를 넘어서는 사항으로, 별도의 검토가 필요한 것으로 정리되었다.

바. AI 영향평가 결과에 따른 조치

AI 영향평가 결과에 따른 조치 방안에 대해서도 TF 내에서 다양한 논의가 이루어졌다. 영향평가가 단순히 평가에 그치지 않고, 실질적인 위험 완화와 사업자의 책임성 강화로 이어지기 위해서는 평가 결과를 기반으로 한 후속 조치가 필요하다는 점에 공감대가 형성되었다. 특히, 완화 조치 등과 관련해서는 AI 기본법 제34조상의 사업자 책무성과 연계하여 검토할 필요가 있다는 의견이 제시되었다. 평가 결과에 따라 적절한 후속 조치를 마련하는 것은 영향평가의 환류(feedback) 과정을 실질화하는 핵심 요소이며, 영향평가 본연의 목적을 고려할 때 조치에 관한 사항까지 논의하는 것이 필수적이라는 점이 강조되었다.

다만, 입법자의 의도를 넘어서는 새로운 법적 의무를 부과하거나, 기업의 자율성을 침해하는 방향으로 가이드라인을 구성해서는 안 된다는 점에 주의가 필요하다는 의견도 제기되었다. 법에서 정하지 않은 사항에 대해 강제력을 부여하기보다는, 윤리적·권고적 차원에서 영향평가 결과에 따른 조치 마련과 문서화·공개의 필요성을 가이드라인에 반영하는 방안이 제안되었다. 이와 관련하여, AI 영향평가는 결과에 대한 피드백이 원활히 이루어지고, 이를 통해 기업의 제도적·윤리적 역량이 강화되는 선순환 구조를 지향해야 한다는 방향성이 제시되었다. 평가 결과와 이에 따른 후속 조치를 상세하게 보고서로 정리하여 사

회에 공유하는 것도 바람직한 방식으로 검토되었다. 또한, 산업계 A위원은, AI 기본법 제34조에 따른 고영향 인공지능 사업자의 위험관리체계가 충분히 구축된 경우, 이와 동등한 수준의 기본권 보호 체계가 마련되었다고 판단하여 영향평가 시행과 동일한 효력을 인정할 수 있는지에 대한 추가 검토 필요성을 제기하였다.

한편, 향후 사업자 책무성 TF 등 관련 TF와의 논의 결과를 조율하고 정합성을 확보하는 과정이 필수적이라는 점도 함께 강조되었다.

사. 타 평가와의 관계 정립

AI 영향평가 가이드라인 마련에 있어, 기업의 부담을 완화하고 자발적 참여를 독려하기 위해 기존의 타 평가 제도와의 관계를 고려할 필요성이 제기되었다. 특히, 한국정보통신기술협회(TTA) 검인증과 같은 유사 평가제도와의 연계 가능성을 검토함으로써 평가 절차의 효율성과 사업자의 실질적 부담 경감을 도모할 수 있다는 이유에서이다.

다만, TTA 검인증은 개발 단계에서 진행되는 사전 모델 평가로, 실제 배포·사용 환경에서 발생할 수 있는 영향까지 포괄적으로 고려하지는 않는다. 동일한 AI 모델이라 하더라도 활용 목적, 방식 등에 따라 실제 사회적 영향은 달라질 수 있기 때문에, 별도의 'AI 영향평가'를 통해 배포 및 사용단계의 기본권 침해 가능성 등을 추가적으로 점검할 필요가 있다. 이는 EU AI Act에서도 개발자(적합성 평가)와 배포자(기본권 영향평가)에게 별도의 평가 의무를 부과하는 구조와 유사하다.

한편, 평가 비용과 부담을 줄이기 위해서는 평가 대상의 특성을 고려하여 불필요한 평가 요소를 제외하거나, 이미 사전 평가된 항목은 중복 평가를 생략하는 방안도 검토할 수 있다. 다만, AI 영향평가의 취지, 목적, 평가 내용의 특수성을 고려하여, 관련성이 명확히 인정되는 평가에 한해 제한적으로 연계 가능성을 검토하는 것이 바람직하다는 점도 함께 논의되었다.

아. AI 영향평가 가이드라인 서술 방식

AI 영향평가 가이드라인은 사업자들이 자율적으로 영향평가를 수행하는 것의 장점과, 이를 통해 얻을 수 있는 실질적 이익을 명확히 설명하는 방향으로 서술할 필요가 있다. 특히, AI 영향평가는 사업자들이 제도적·규제적 요구사항은 물론, 윤리적 책임 이행 측면에서도 부족한 부분을 스스로 진단하고, 이를 대응 방안 마련과 역량 강화의 기회로 활용할

수 있는 유용한 도구임을 강조해야 한다. 또한, AI 영향평가 수행이 국제 AI 거버넌스 흐름에 부합하며, 향후 글로벌 시장에서 요구될 수 있는 안전성 및 신뢰성 기준에 효과적으로 대비할 수 있는 방법이라는 점을 함께 설명하는 것이 바람직하다. 이를 위해, 관련 글로벌 동향 및 국제기준에 관한 자료를 부록으로 제공하는 방안도 고려할 수 있다.

제 3 절 AI 영향평가 프레임워크(안)

본 절에서는 현재까지 AI 영향평가 가이드라인 TF 논의 결과를 바탕으로 마련된 AI 영향평가 프레임워크 초안을 소개하고자 한다. 다만, AI 영향평가 프레임워크는 아직 TF 논의가 진행 중인 상태이며, 정부 차원의 최종 확정 및 공식 발표는 향후 별도로 이루어질 예정이다. 따라서 본 절에 제시하는 구성안은 현시점에서의 검토 결과를 반영한 것으로, 추후 추가 논의 과정 및 정책 방향에 따라 세부 내용이 변경될 수 있음을 전제로 한다. 이하에서는 각 구성 항목별로 세부 내용을 살펴보고자 한다.

1. AI 영향평가 개요

AI 영향평가는 AI의 개발 및 활용과정에서 인간의 존엄성과 기본권을 존중하는 것을 지향한다. 특히, 「AI 기본법」 제34조에 따른 고영향 AI 사업자의 책무만으로는 인간의 존엄성, 자유, 평등, 사회·경제적 기본권 등 국민의 권리를 충분히 보장하기 어려운 한계를 보완하고, 글로벌 수준의 상호운용성[29]을 확보하기 위해 도입되었다. AI 영향평가는 고영향 AI 또는 이를 활용한 제품·서비스를 제공하기에 앞서, 해당 기술이 국민의 기본권에 미치는 영향을 식별하고, 필요한 완화 조치를 마련함으로써 기본권 침해를 예방하는 것을 목적으로 한다. 이를 통해 고영향 AI가 초래할 수 있는 위험에 선제적으로 대응하고, AI 신뢰성을 기반으로 한 선순환적 리스크 관리체계를 구축하고자 한다.

---

29) 유럽연합은 인간 존엄성, 자유, 평등, 시민의 권리 등을 보호하기 위해, '유럽연합 기본권 헌장' 등 법적 체계에 기반하여 기본권 영향평가(Fundamental Rights Impact Assessment, FRIA) 제도를 마련하였으며(EU AI Act 제27조), 이는 2026년 8월부터 시행될 예정

2. 법적 근거

AI 영향평가는 AI 기본법 제35조를 근거로 한다.

〈표 3－9〉 「AI 기본법」 제35조(고영향 인공지능 영향평가)

제35조(고영향 인공지능 영향평가) ① 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다. ② 국가기관등이 고영향 인공지능을 이용한 제품 또는 서비스를 이용하려는 경우에는 영향평가를 실시한 제품 또는 서비스를 우선적으로 고려하여야 한다. ③ 그 밖에 영향평가의 구체적인 내용·방법 등에 관하여 필요한 사항은 대통령령으로 정한다.

자료: 국가법령정보센터, https://www.law.go.kr

3. AI 영향평가 수행 주체

고영향 AI를 이용한 제품 또는 서비스[30]를 제공하고, 이를 통해 "영향받는 자"에게 영향을 미치는 AI 사업자는 AI 영향평가를 수행하기 위해 노력할 것이 권장된다. 영향평가는 일회성 절차가 아니며, 평가요소에 중대한 사항이 변경되었거나 최신 상태를 반영할 필요가 있다고 판단되는 경우 재평가를 수행해야 한다.

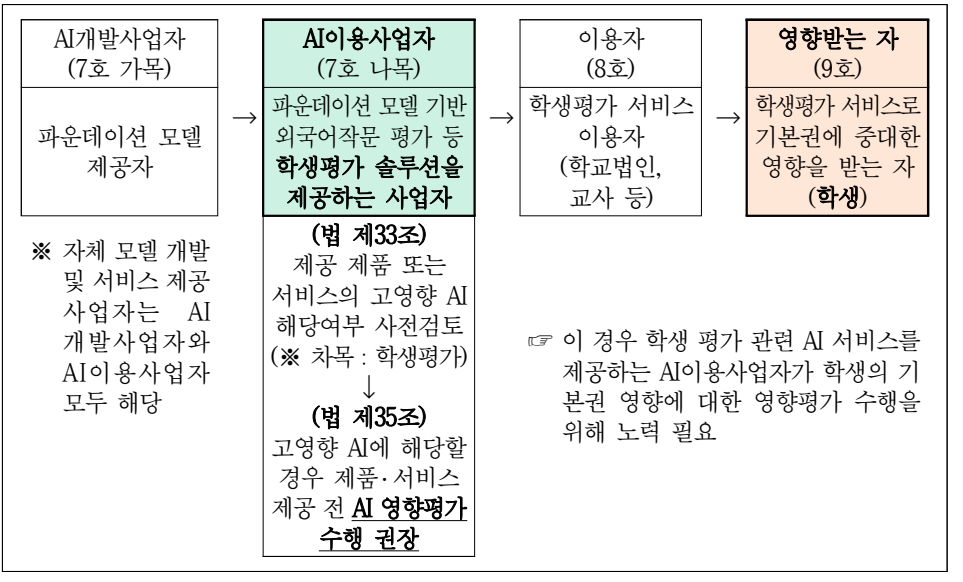
〈표 3－10〉 「AI 기본법」 제2조(정의)

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다. 7. "인공지능사업자"란 인공지능산업과 관련된 사업을 하는 자로서 다음 각 목의 어느 하나에 해당하는 법인, 단체, 개인 및 국가기관 등을 말한다. 가. 인공지능개발사업자: 인공지능을 개발하여 제공하는 자 나. 인공지능이용사업자: 가목의 사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자 8. "이용자"란 인공지능제품 또는 인공지능서비스를 제공받는 자를 말한다. 9. "영향받는 자"란 인공지능제품 또는 인공지능서비스에 의하여 자신의 생명, 신체의 안전 및 기본권에 중대한 영향을 받는 자를 말한다.

자료: 국가법령정보센터, https://www.law.go.kr

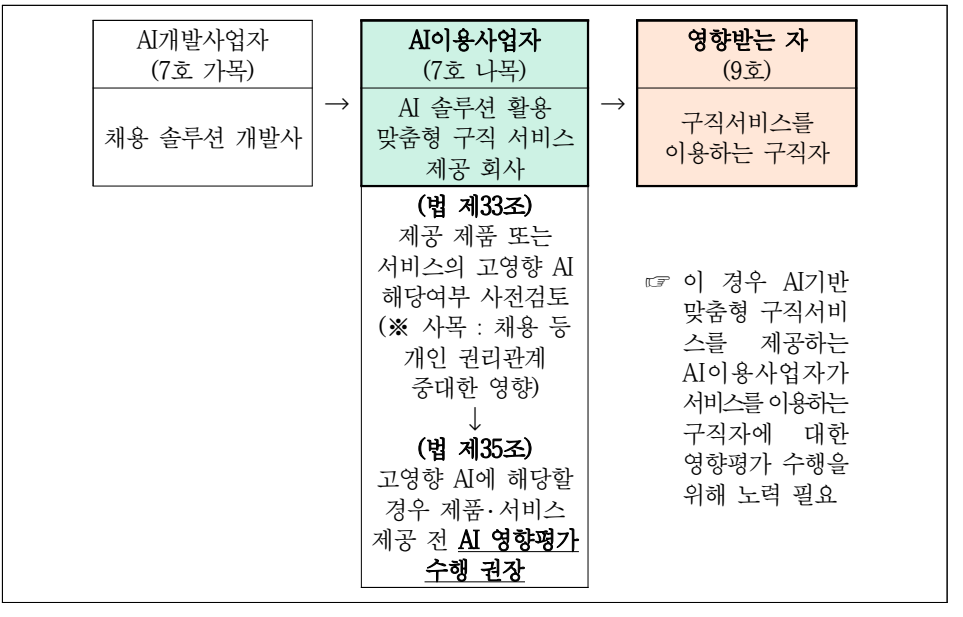
30) 사업자는 제공하는 제품·서비스의 AI가 고영향 AI에 해당하는지를 사전에 검토해야 하며, 필요할 경우 과학기술정보통신부에 확인 요청할 수 있음(「AI 기본법」 제33조)

〔그림 3-2〕 학생평가 솔루션 사례



자료: 연구진 작성

〔그림 3-3〕 채용 솔루션 도입 사례



자료: 연구진 작성

4. AI 영향평가 수행 및 예시

가. 평가 요소

평가 대상이 되는 고영향 AI 제품·서비스 제공자는 다음과 같은 사항을 포함하여 AI 영향평가를 수행하는 것이 권장된다.

① 이용 행태: 우선, 해당 AI 제품·서비스가 의도된 목적에 맞게 사용되는 경우를 상정하여 이용 시나리오를 작성하고, 필요한 경우 이용 절차, 이용 기간 및 빈도 등을 함께 설정한다. 이는 AI 이용이 기본권 침해에 미치는 영향 정도를 보다 정확히 판단하기 위함이다.

② 피영향자 식별: 특정 상황에서 AI 제품·서비스 이용으로 인해 영향받을 가능성이 있는 개인 또는 집단(영향받는 자)의 범주를 식별해야 한다. 이를 통해 AI 이용에 따른 영향 집단을 명확히 특정할 수 있다.

③ 피영향 기본권 식별: 이용 과정에서 침해될 수 있는 기본권 유형을 식별하고, 해당 기본권이 어떤 영향 요소로 인해 침해될 수 있는지를 분석·검토한다. 이를 통해 침해 가능한 기본권을 특정하고, 침해 가능성과 그 정도를 보다 체계적으로 평가할 수 있다.

〈표 3-11〉 AI 제품·서비스로 인한 기본권 영향 프로파일 예시

[예시1] 금융기관 대출심사에서 활용되는 AI 제품·서비스로 인한 평등권 영향 프로파일
※「AI 기본법」 제2조 제4호 사목 : 채용, 대출 심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가

유형	프로파일 내용	
평등권	개요	편향된 데이터 기반 AI의 차별적 평가와 기회 불평등
	시나리오	금융기관의 대출 심사 AI가 과거 특정 직업군이나 지역(저소득층 밀집 지역)의 대출 상환 실적 데이터에 기반하여 해당 직업 종사자나 지역 거주자에게 자동으로 낮은 신용점수를 부여하고 대출을 거절하거나 높은 금리를 적용하여, 개인의 실제 신용도와 무관하게 금융 서비스 접근에 차별을 야기
	결과	집단 특성에 근거한 부당한 차별과 경제적 기회 박탈 **→ 헌법 제11조 평등권 침해**
	완화 방안	대출 심사 시 특정 지역 등으로 인한 차별 효과가 발생하지 않도록 시스템을 조정하고, 담당 인력을 활용한 직접 심사를 통해 보완 추진

[예시2] 초·중등교육 영역에서 활용되는 AI 제품·서비스에 대한 영향 프로파일

※「AI 기본법」 제2조 제4호 차목 :「교육기본법」 제9조제1항에 따른 유아교육·초등교육 및 중등교육에서의 학생 평가

유형	프로파일 내용	
평등권	개요	AI 평가 시스템의 편향된 판단으로 인한 특정 집단 차별 발생
	시나리오	AI 자체의 기계적 오류 혹은 편향된 과거 학습 데이터로 인하여 특정 지역, 문화, 사회경제적 배경을 가진 학생들에게 불리한 평가를 내리고, 해당 정보가 대학 입시나 장학금 지원에 반영되어 교육 기회의 평등한 접근성을 제한
	결과	학생의 배경에 따른 부당한 차별과 교육 기회 불평등 발생 **→ 헌법 제11조 평등권, 제31조 교육권 침해**
	완화 방안	솔루션 수정을 통해 지역, 사회적 배경 등이 학생 평가에 영향을 미치지 않도록 개선
행동과 표현의 자유	개요	AI 평가 시스템으로 인한 학생의 자기표현 제약과 행동 위축
	시나리오	AI 기반 학생평가 시스템이 특정 형태의 답변이나 학습 방식만을 올바른 것으로 인식하도록 설계되어, 학생들이 창의적 표현과 다양한 사고방식을 자유롭게 시도하지 못하고 AI가 인식하기 쉬운 표준화된 방식으로만 자신을 표현
	결과	학생의 창의적 표현 위축과 자율적 학습 태도 저해 **→ 헌법 제21조 표현의 자유, 제10조 일반적 행동자유권 침해**
	완화 방안	전체 학생평가에서 AI 평가 시스템에 따른 결과 반영 정도를 낮추고, 현장 교사의 평가 반영을 높이는 방식으로 완화

자료: 연구진 작성

나. 보완방안

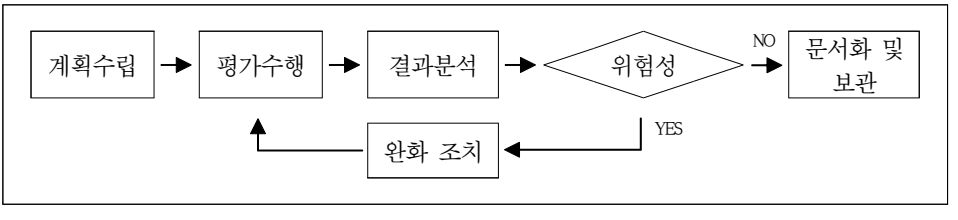
AI 영향평가 결과, 식별된 기본권 침해 가능성에 대해 제품·서비스 제공자는 자체적인 침해 완화 방안을 마련하고, 이행 계획을 검토해야 한다. 보완방안은 평가 결과에 따른 위험 관리 및 기본권 보호 조치로서, 영향평가의 실효성을 확보하는 핵심적 단계이다.

5. AI 영향평가 프로세스

AI 영향평가는 사전 계획부터 평가 결과의 문서화에 이르기까지 일련의 절차를 체계적으로 수행하는 것을 원칙으로 한다. 현시점에서 논의된 평가 프로세스(안)는 다음과 같으며, 세부 내용은 향후 추가 구체화될 예정이다.

먼저, 영향평가 필요성을 검토하고, 평가 대상이 되는 제품·서비스를 정의 및 검토한다. 이 과정에서는 제품·서비스의 설명, 사용 절차 및 범위, 이용 기간, 비례성 등을 종합적으로 고려한다. 이어서 평가계획을 수립하고, 평가를 수행할 평가팀을 구성하여 각 임무를 명확히 정의하는 등 사전 준비를 완료한다. 다음으로, 평가 수행 단계에서는 평가에 필요한 자료를 수집하고, 영향받을 가능성이 있는 개인 또는 집단을 식별한다. 이들이 받게 될 기본권 영향과 구체적 영향 요소를 프로파일링하며, 긍정적·부정적 영향의 발생 가능성, 영향의 규모, 지속 기간, 해결 가능성 등을 종합적으로 평가한다. 평가 수행 결과를 토대로 평가결과를 분석하고, 필요한 경우 기본권 침해 완화를 위한 후속 조치 필요성을 판단한다. 이어서 평가 결과에 따라 기본권 침해 위험을 줄이기 위한 영향 완화 조치를 계획하고, 필요시 이를 시행한다. 마지막으로, 평가 과정 및 결과를 체계적으로 문서화하여 기록하고, 향후 활용 및 점검을 위해 이를 보관한다.

〔그림 3-4〕 AI 영향평가 프로세스(안)



자료: 연구진 작성

〈표 3－12〉 AI 영향평가 프로세스(안)

단계	세부 내용(안)
평가계획 수립	• 영향평가 필요성 검토 • 대상 제품·서비스 정의 및 검토(설명, 사용절차 및 범위, 기간, 비례성 등) • 계획 수립, 평가팀 구성, 임무 정의 등 사전 준비
영향평가 수행	• 평가자료 수집 • 영향받을 가능성이 있는 개인 또는 집단 식별 • 식별된 대상이 받게될 기본권, 구체적 영향요소 프로파일 • 긍정·부정 영향 발생 가능성, 규모, 기간, 해결 가능성 등 평가
평가결과 분석	• 평가결과 분석 및 완화 조치 필요성 판단
완화조치	• 필요할 경우 영향 완화를 위한 조치 시행
문서화 및 보관	• 평가과정 및 결과 문서화

자료: 연구진 작성

제 4 장 AI 영향평가 도입을 위한 과제

제 1 절 개요

AI 영향평가는 인공지능 기술의 책임 있는 개발 및 활용을 촉진하고, 국민의 기본권을 보호하기 위한 제도로 마련되었다. 이에 따라 AI 영향평가 프레임워크(안)가 제시되었으나, 실제 현장에서 이를 적용하기 위해서는 다양한 이슈와 쟁점에 대한 사전 검토가 필수적이다. AI 영향평가는 새로운 제도적 시도인 만큼, 적용 과정에서 제도 운영의 실효성, 기업의 이행 부담, 정부기관의 활용 방식 등 여러 차원의 문제들이 발생할 수 있다. 특히, 프레임워크를 마련하는 것에 그치지 않고, 평가 도입 및 실행 과정에서 예상되는 쟁점들을 면밀히 검토하고 이에 대한 구체적 대응방안을 모색하는 것이 중요하다. 이에 본 장에서는 AI 영향평가의 현장 적용과 관련하여 예상되는 주요 쟁점들을 살펴보고, 프레임워크 마련과 함께 도입 및 운영 측면에서 고려해야 할 논의사항을 제시하고자 한다.

제 2 절 AI 영향평가 도입 관련 주요 쟁점

1. 사업자 책무와의 중복성

「AI 기본법」 제34조는 고영향 AI 사업자에게 고영향 AI의 안전성·신뢰성을 확보하기 위한 일련의 책무를 부과하고 있다. 이에 따라 고영향 AI 또는 이를 이용한 제품·서비스를 제공하는 사업자는 위험관리방안의 수립·운영, 이용자 보호 방안의 수립·시행, 안전성·신뢰성 확보 조치의 내용을 확인할 수 있는 문서의 작성과 보관 등의 의무를 이행해야 한다. 이러한 상황에서 별도로 제35조에 따른 AI 영향평가를 추가로 수행해야 하는지에 대한 우려의 목소리도 존재한다. 사업자 책무 이행 과정에서 이용자 보호 방안까지 마련하고 있음에도 불구하고, 별도의 영향평가 절차를 요구하는 것은 사업자에게 이중 부담을 초래

할 수 있다는 점, 기존 조치와의 중복성을 해소할 필요가 있다는 점이 주요한 문제로 지적되고 있다.

그러나 AI 영향평가는 제34조 사업자 책무와는 목적과 기능이 명확히 구별된다. AI 영향평가는 고영향 AI 또는 이를 이용한 제품·서비스 제공에 앞서, 해당 기술이 사람의 기본권에 미치는 영향을 사전적으로 식별하고, 필요한 완화 조치를 마련하도록 권고하는 '사전적 평가' 규정이다. 이를 통해 고영향 AI가 초래할 수 있는 기본권 침해 위험에 선제적으로 대응하고, 나아가 선순환적 리스크 관리체계를 구축할 수 있도록 하는 것을 목적으로 한다. 영향평가를 실시함으로써, AI 사업자 스스로 정부 기관이나 입법자가 사전에 예측하기 어려운 고영향 AI의 내재적 위험을 조기에 파악하고 대응할 수 있으며, 제품·서비스 제공 이전에 사업자 책무 이행 여부를 점검하는 체계를 확립할 수 있다. 이러한 과정을 통해 AI 안전성·신뢰성과 기본권 보호를 동시에 확보하는 선순환적 운영 구조를 마련할 수 있다는 기대효과가 있다.

또한 제34조와 제35조는 규율 목적과 적용 방식에 있어 본질적인 차이를 가진다. 제34조는 AI 사업자에게 안전성 및 신뢰성 확보를 위한 위험관리방안 수립, 이용자 보호 방안 마련 등 일정한 조치를 취할 의무를 부과하는 규정이다. 반면 제35조는 기본권 침해 예방을 주된 목적으로 하며, 제품·서비스 제공 이전에 기본권에 미치는 영향을 평가하도록 권고하는 '노력 의무' 규정으로 자리매김하고 있다. AI 영향평가는 단순히 사업자의 위험관리 조치를 요구하는 데 그치지 않고, 고영향 AI가 잠재적으로 야기할 수 있는 기본권 침해 위험을 보다 체계적으로 검토하고 대응책을 마련하기 위한 수단이다.

적용 대상과 범위 측면에서도 두 조항은 차이가 존재한다. 제34조는 안전성·신뢰성 확보와 밀접하게 관련된 기본권을 중심으로 규율하는 반면, 근로의 권리, 환경권 등과 같이 안전성과 직접적 관련이 없는 기본권 보호에 대해서는 적용이 어렵다. 또한, 제34조는 주로 이용자를 보호하는 데 초점을 맞추고 있어, 이용자 외 제3자에게 미칠 기본권 관련 영향에 대해서는 명확하게 규율하고 있지 않다. 이에 반해 제35조는 이용자뿐만 아니라 제3자를 포함한 모든 사람을 대상으로 기본권 보호를 지향하며, 이러한 포괄적 접근을 통해 제34조의 한계를 보완할 수 있다.

결론적으로, 「AI 기본법」 제35조에 따른 AI 영향평가는 모든 사람을 대상으로, 모든 기본권을 보호하기 위한 목적의 조치로서, 제34조를 통해 보호되기 어려운 영역까지 포괄하

는 역할을 수행한다. 영향평가를 이행함으로써 결과적으로 사업자 책무(제34조)도 충실히 준수되는 효과를 기대할 수 있다. 제34조는 AI 사업자가 안전성·신뢰성 확보를 위해 적정한 방법을 마련해야 하는 의무 규정인 반면, 제35조는 제품·서비스가 초래할 수 있는 기본권 영향을 다방면으로 사전에 검토하는 권고 규정으로 기능한다. 두 조항은 규제 목적, 의무 부담 여부, 수행해야 하는 조치 등에서 명확한 차이를 가지므로, 각 조항이 규정하는 바에 따라 상호 보완적으로 이행할 필요가 있다.

2. 정부기관 조달

「AI 기본법」 제35조 제2항에는 "국가기관 등이 고영향 AI를 이용한 제품 또는 서비스를 이용하려는 경우, 영향평가를 시행한 제품 또는 서비스를 우선적으로 고려하여야 한다"라고 규정하고 있다. 하지만 해당 조항의 실제 적용에 있어 영향평가 활성화에 충분히 기여할 수 있을지 여부에 대해서는 다양한 의견이 존재한다. 특히, '우선적 고려'라는 표현의 구체적 의미와 적용 방식, 국가기관 등의 조달 수요 규모, 사업자들의 실질적 유인 제공 여부 등에 따라 조항의 실효성이 달라질 수 있다는 점에서 심도 있는 검토가 필요하다. 이하에서는 제35조 제2항이 실질적으로 영향평가 활성화에 기여할 수 있는 지, AI 영향평가가 자율적으로 도입되고 확산하기 위해 필요한 법적·정책적 수단은 어떤 것이 있을지를 중심으로 논의를 정리하고자 한다.

「AI 기본법」 제35조 제2항이 AI 영향평가 활성화에 실질적으로 기여할 수 있을지에 대한 다양한 견해가 존재한다. 우선, 본 조항이 실질적으로 기여하기 위해서는 국가기관 등의 AI 제품·서비스 이용 수요가 충분히 확보되어야 한다는 점이 지적되었다. 국가기관의 이용 수요가 미미할 경우, AI 사업자들이 영향평가를 인센티브가 약화될 수 있기 때문이다. 또한, '우선적 고려'의 구체적인 방식이 명확히 규정되어 있지 않아 단순히 영향평가 실시 여부만 평가하거나 낮은 수준의 가점을 부여하는 방식으로 운영된다면, 사업자 입장에서 영향평가를 적극적으로 수행할 유인이 제한될 수 있다는 우려도 제기되었다. 따라서 실질적인 영향평가 활성화를 위해서는 국가기관이 영향평가 실시 여부뿐만 아니라 평가결과의 수준까지 구매 여부에 적극 반영하도록 유도하는 방안이 필요할 것이다. 반면, 제35조 제2항이 실질적으로 AI 영향평가의 확산에 기여할 수 있다는 긍정적인 평가도 있었

다. 국가기관 등이 조달 과정에서 동등하거나 유사한 품질 및 가격의 제품·서비스를 비교할 때, 영향평가 실시 여부가 선정 기준으로 작용할 가능성이 높으며, 공공부문 납품 실적이 민간부문에서도 강력한 신뢰 지표로 작용할 수 있기 때문에 자연스럽게 민간 기업의 영향평가 참여를 유도할 수 있다는 것이다. 특히, 「클라우드컴퓨팅 발전 및 이용자 보호에 관한 법률」(이하 '클라우드컴퓨팅법') 제20조 제2항과 관련된 사례에서도, 공공부문 클라우드 조달 시 보안인증(Cloud Security Assurance Program, CASP)을 받은 서비스를 우선 고려토록 한 규정이 실제로 인증 참여를 촉진시킨 효과가 있었음을 볼 때, AI 영향평가 역시 유사한 효과를 기대할 수 있다는 분석이 제시되었다.

AI 기본법 제35조 제2항에 대해 다음과 같은 법적·정책적 한계가 존재한다. 우선, 제35조 제2항은 영향평가를 시행한 제품 또는 서비스를 '우선적으로 고려'하도록 규정하고 있으나, '우선적 고려'의 구체적인 방식이 명확히 제시되어 있지 않아 실질적 영향평가 활성화에 한계가 있을 수 있다는 지적이 제기되었다. 특히, 국가기관 등이 단순히 영향평가 실시 여부만 확인하거나 낮은 수준의 가점을 부여하는 방식으로 운영될 경우, 사업자들이 영향평가를 수행할 유인이 약화될 수 있다. 이와 관련하여, 시행령으로 국가기관의 구매 기준을 구체적으로 정하는 것은 법적 위임 한계상 어려울 수 있어, 가이드라인 등을 통해 영향평가 결과를 정성평가 항목에 적극 반영하도록 유도하는 방안이 필요하다는 의견이 제시되었다. 또한, 제35조 제2항은 '우선적 고려'를 규정할 뿐, 우선 구매의무까지 부과하지 않기 때문에 법적 강제력이 부족하다는 한계도 지적되었다. 이를 보완하기 위해 공공조달 과정에서 영향평가를 받은 제품·서비스에 대해 충분한 가산점을 부여하는 등 실질적 인센티브를 제공하는 방안을 시행령 또는 가이드라인을 통해 마련할 필요성이 제기되었다. 영향평가 비용이 스타트업 등 중소기업에 과도한 부담을 초래할 수 있다는 점도 한계로 지적되었다. 클라우드 보안인증(CSAP) 사례에서도 인증 비용이 중소기업에 큰 부담이 되었던 점을 감안할 때, 인공지능 영향평가 역시 비용 지원 정책이 병행되어야 실효성이 담보될 수 있다. 이에 따라, 「AI 기본법」 제17조 제3항[31]을 근거로 중소기업 대상 영향평가 비용 지원 방안을 시행령이나 시행규칙에 구체화하는 방안이 필요하다는 제안이 있

31) "과학기술정보통신부장관은 인공지능의 안전성 및 신뢰성 확보를 위하여 중소기업 등의 제34조에 따른 조치 이행 및 제35조에 따른 영향평가를 지원할 수 있다."

었다. 아울러, 현재 영향평가는 인공지능사업자가 자체적으로 수행하는 구조로 되어 있어 평가 결과의 임의성 또는 신뢰성 저하 우려가 존재한다는 점도 문제로 제기되었다. 그러나 제3기관을 통한 강제적 검증은 사업자 부담을 과도하게 증가시킬 우려가 있어, 절충안으로 영향평가 과정과 결과를 외부에 투명하게 공개하도록 권고하여 평가의 신뢰성과 공정성을 높이는 방안이 고려될 수 있다.

3. AI 영향평가 전담기관 필요성

AI 영향평가는 고영향 AI가 사람의 기본권에 미치는 영향을 사전에 식별하고 필요한 완화 조치를 마련하는 과정을 요구한다는 점에서, 상당한 전문성, 시간, 비용을 요하는 절차이다. 하지만 인적·시간적·경제적 역량이 부족한 중소기업이나 스타트업의 경우 자체적으로 영향평가를 수행하는 데 상당한 어려움을 겪을 수 있다.

이러한 현실을 고려할 때, 기업이 AI 영향평가를 원활히 수행할 수 있도록 지원하는 전담기관 또는 지원기관의 역할이 중요하며, 이에 따라 'AI 영향평가 센터'(가칭)의 설립 및 운영이 필요하다는 점이 강조된다. 'AI 영향평가 센터'는 영향평가 수행을 위한 가이드라인 제공, 절차 컨설팅, 사전 평가 지원, 평가결과 검증 지원 등 다양한 역할을 수행함으로써, 기업들이 보다 체계적이고 효율적으로 영향평가를 진행할 수 있도록 지원할 수 있다. 특히, 영향평가를 직접 수행하거나, 사업자 대신 영향평가를 대행하는 방식뿐만 아니라, 영향평가를 준비하는 과정에서 기업별 상황에 맞춘 컨설팅을 제공하고, 필요한 경우 공동평가체계(예: 산업별 표준 평가모델 구축)를 지원하는 형태도 고려할 수 있다. 이를 통해 중소기업과 스타트업이 영향평가를 보다 현실적으로 이행할 수 있도록 기반을 조성할 수 있다. 또한, 영향평가 전담기관은 평가 품질을 제고하고 평가 결과의 신뢰성을 확보하는 측면에서도 중요한 역할을 수행할 수 있다. 기업 자체 수행에 따른 평가의 편의성·임의성 문제를 보완하고, 최소한의 객관성과 일관성을 담보하는 데 기여할 수 있기 때문이다. 따라서 향후 AI 영향평가 제도의 현장 정착과 실효성 확보를 위해서는 'AI 영향평가 센터' 등의 전담기관 또는 지원체계를 구축하는 방안을 적극적으로 검토할 필요가 있다.

제5장 결 론

제1절 연구 결과 요약

본 연구에서는 인공지능 기본법 제35조에 따라 도입이 추진되고 있는 AI 영향평가 제도의 방향성과 실효적 적용 방안을 마련하기 위해 다음과 같은 내용을 중심으로 분석을 수행하였다.

우선, 제2장에서는 국내외 영향평가 유사 제도들을 검토하였다. 이를 통해, UNESCO의 윤리적 영향평가(Ethical Impact Assessment), 캐나다의 알고리즘 영향평가(Algorithmic Impact Assessment) 등과 같이 AI 영향평가와 유사한 사전적 평가 제도가 다양한 분야에서 운영되고 있으며, 기본권 보호와 위험 예방을 위한 체계적 사전 평가의 중요성이 국제적으로 강조되고 있다는 점을 확인하였다. 제3장에서는 AI 기본법 제35조를 근거로 마련 중인 AI 영향평가 가이드라인 초안 및 프레임워크(안)을 소개하였다. AI 영향평가 가이드라인 TF 논의 결과를 바탕으로, 영향평가의 기본 방향, 수행 주체, 주요 평가요소 및 세부 절차(평가계획 수립-영향평가 수행-결과 분석-완화조치-문서화) 등을 정리하고, 실질적인 평가 이행을 지원하기 위한 가이드라인 구성을 제시하였다.

〈표 5-1〉 AI 영향평가 프레임워크(안)

ㅇ **(관련 근거)** "인공지능 기본법" 제35조(고영향 인공지능 영향평가)

※ 제35조(고영향 인공지능 영향평가)

① 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다.

② 국가기관등이 고영향 인공지능을 이용한 제품 또는 서비스를 이용하려는 경우에는 영향평가를 실시한 제품 또는 서비스를 우선적으로 고려하여야 한다.

③ 그 밖에 영향평가의 구체적인 내용·방법 등에 관하여 필요한 사항은 대통령령으로 정한다.

ㅇ **(평가 대상)** 고영향 AI 제품(또는 서비스)
※ 고영향 AI 기준·가이드라인 TF 별도 진행 중
※ 다만, EU AI Act에서는 특정 고위험 AI 시스템 배포자에게만 기본권 영향평가 의무 부과

ㅇ **(평가 범위)** 고영향 AI 제품(또는 서비스)의 특정 분야
※ AI 제품 또는 서비스 전체가 아닌 전체 서비스 중 특정 분야의 영향을 평가

ㅇ **(평가 시기)** 대상 AI 제품(또는 서비스) 제공 전
※ ① 평가요소에 관한 사항이 변경되었거나 최신 상태가 아니라고 판단될 경우(EU), ② 해당 AI 시스템의 기능 또는 범위가 변경되는 경우를 포함하여 정기적으로 검토 및 업데이트(캐나다)

ㅇ **(평가 주체)** 고영향 AI 제품(또는 서비스) 제공 사업자 (제2조 제7항 나목)
※ 제2조(정의) "인공지능사업자"

> 7. "인공지능사업자"란 인공지능산업과 관련된 사업을 하는 자로서 다음 각 목의 어느 하나에 해당하는 법인, 단체, 개인 및 국가기관등을 말한다.
> 가. 인공지능개발사업자 : 인공지능을 개발하여 제공하는 자
> 나. 인공지능이용사업자 : 가목의 사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자

ㅇ **(평가 요소 및 지표)** '고영향 인공지능이용사업자'가 해당 AI 제품·서비스의 영향을 더 잘 이해하고 관리할 수 있도록 다음 사항을 포함하여 평가
- 의도된 목적에 맞게 사용될 절차, 배포 범위, 사용 기간 및 빈도 등
- 비례성 검토(명시된 목표 달성을 위해 비례하는 수준에서 AI 사용)
- 대상 AI 제품·서비스에 대한 역할과 책임
- 특정 상황에서의 대상 AI 제품·서비스 사용으로 인해 영향받을 가능성이 있는 개인 또는 집단 범주
- ① 영향받을 것으로 식별된 개인 또는 집단이 침해받을 수 있는 기본권과, ② 해당 기본권이 어떤 영향 요소로 인해 침해되는지 프로파일

※ EU AI Act에서는 고위험 AI 시스템 배포자(deployer)가 관련 의무를 준수할 수 있도록, 고위험 AI 시스템이 적절한 유형과 수준의 투명성을 확보하며 설계·개발되도록 규정(제13조) → 배포자는 제13조에 따라 제공자(provider)가 제공한 정보를 고려하여 앞서 식별된 개인 또는 집단의 범주에 영향을 미칠 수 있는 구체적 해악 평가

- 영향평가(발생 가능성, 규모, 기간, 해결 가능성 등)
- 부정적 영향 현실화 시 취할 조치(내부 거버넌스, 이의처리 체계 등 포함)

ㅇ **(평가 수행 절차)**
① 평가 계획 수립
· 영향평가 필요성 검토
· 대상 제품·서비스 정의(설명, 사용절차 및 범위, 기간, 비례성 검토 등)
· 계획 수립, 평가팀 구성, 임무 정의 등 사전 준비, 기초 분석

② 영향평가 수행
- 사례 등 평가자료 수집
- 영향받을 가능성이 있는 개인 또는 집단 식별
- 식별된 대상이 받게될 기본권, 구체적 영향요소 프로파일, 판단기준 수립 등
- 긍정·부정 영향 발생 가능성, 규모, 기간, 해결 가능성 등 평가

③ 평가결과 분석

④ 위험성 완화 조치

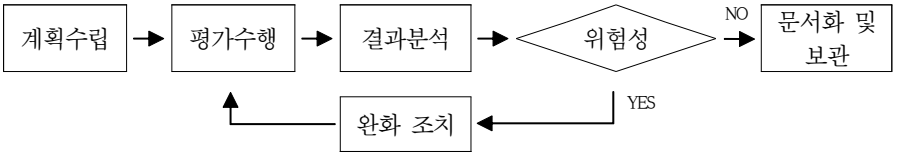
※ (참고) 캐나다 알고리즘 영향평가 : 평가는 자동화된 의사결정 시스템이 제작되기 전 두 번 완료(2차는 1차 평가 결과에 대한 추가 완화 조치를 반영)되어야 하며, 그 결과가 구축된 시스템을 정확하게 반영하는지 확인하고 최종 결과를 정부 포털에 공개

⑤ 평과과정 및 결과 문서화, 내부 공유

※ 근거조항 : 캐나다 자동화된의사결정에 대한 지침(6.1.4), EU AI Act(제72조)

⑥ 대응조치 및 지속적 모니터링

[AI 영향평가 프로세스(안)]



자료: 연구진 작성

제2절 향후 연구 방향

본 연구에서 개발된 AI 영향평가 프레임워크는 과학·기술적 측면, 경제·산업적 측면, 사회적 측면에서 다양한 기대효과를 가져올 것으로 예상된다. 과학·기술적 측면에서는 AI 영향평가 과정에서 국민의 기본권을 중심으로 한 다양한 윤리적, 사회적 문제를 예측하고 해결 방안을 모색함으로써 AI 개발 및 이용사업자가 안전하고 혁신적인 기술을 자유롭게 연구·개발할 수 있는 환경을 마련하고, 공통된 평가 기준을 통해 국내외 AI 개발 표준화를 선도할 것이다. 경제·산업적 측면에서는 AI 기반 제품 및 서비스의 신뢰도를 높여 국내 AI 산업 경쟁력을 제고하고, 초기 단계에서부터 위험 요소를 식별하고 합리적으로 관리함으로써 장기적으로 발생할 수 있는 법적, 사회적, 기술적 문제해결 비용을 절감할 것이다. 사회적 측면에서는 AI 개발·활용에서 인권, 사생활 보호, 공공성, 다양성 등의 원칙을 준수할 수 있는 기반을 마련하여 국민의 기본권을 보다 체계적으로 보호하고, 취약계층을

포함한 다양한 이해관계자에 미치는 영향을 고려함으로써 AI 도입으로 인한 사회적 불평등을 완화하며, AI의 윤리적, 책임감 있는 개발·활용에 대한 인식을 고취할 것이다.

AI 영향평가 프레임워크의 실효성 있는 적용과 활용을 위해서는 다음과 같은 정책적 지원과 노력이 요구된다.

우선, 2026년 1월 시행 예정인 「AI 기본법」의 사회적 안착을 지원하는 방향으로 연구가 진행될 필요가 있다. 특히 고영향 AI 제품 및 서비스에 대한 영향평가 관련 규정이 실효성 있게 작동할 수 있도록 구체적인 평가 가이드라인과 사례를 지속적으로 축적하고 보완하는 작업이 요구된다. 이는 개발된 프레임워크를 다양한 AI 제품과 서비스에 시범적으로 적용하고, 그 과정에서 발견된 문제점과 개선사항을 체계적으로 반영함으로써 가능할 것이다. 또한 AI 기술의 급속한 발전과 활용 범위의 확대에 따라 새롭게 등장하는 윤리적, 사회적, 법적 쟁점들을 지속적으로 모니터링하고 이를 영향평가 프레임워크에 반영하는 연구가 필요하다. 특히 생성형 AI, 자율주행차, 의료 AI, 에이전트 AI 등 분야별 특성을 고려한 맞춤형 영향평가 지표와 방법론을 개발하여 프레임워크의 실용성을 높이는 연구가 진행되어야 할 것이다.

무엇보다도, 고영향 AI 제품 및 서비스가 국민의 기본권에 미치는 영향을 체계적으로 관리하기 위해서는 실질적이고 실행 가능한 영향평가 제도의 지속적인 고도화 방안 마련이 필요하다. 이를 위해, AI 개발 및 이용 사업자, 특히 중소기업과 스타트업 등이 실제로 수행할 수 있는 현실적인 영향평가 프로세스를 설계하고, 본 제도의 효과성을 높이고 안정적으로 정착시키기 위한 전담 지원기관(가칭 'AI 영향평가 센터')의 역할과 기능을 구체화하는 연구가 병행되어야 한다.

한편, AI 영향평가를 실시하는 주체들의 역량 강화를 위한 교육 및 지원 방안에 대한 연구도 중요한 향후 연구 과제이다. AI 개발사업자, AI 이용사업자, 공공기관, 연구기관 등 다양한 주체들이 자체적으로 영향평가를 실시할 수 있는 능력을 배양하기 위한 교육 프로그램 개발, 전문가 양성, 자가진단 도구 개발 등이 이에 포함될 수 있다.

국제 협력 측면에서는 AI 영향평가에 대한 글로벌 표준화 노력에 적극 참여하고, 국제사회와의 지속적인 교류를 통해 AI 거버넌스의 국제적 정합성을 제고하는 연구가 필요하다. 이러한 노력은 우리나라가 개발한 AI 영향평가 프레임워크의 국제적 위상을 강화함은 물론, 글로벌 AI 규제 환경에 대한 국내 기업과 기관의 대응 역량을 높이는 데에도 기여할

수 있다. 특히, EU AI법에서 규정하고 있는 AI 기본권 영향평가와의 상호운용성 확보 측면에서 글로벌 협력은 매우 중요한 과제라 할 수 있다.

특히, AI 기본법 내 고영향 AI 관련 조항들 간의 유기적 관계 정립에 관한 연구가 필요하다. AI 기본법에서는 고영향 AI에 대한 영향평가 뿐만 아니라 안전성 및 투명성 확보, 신뢰성 검·인증, 위험관리·이용자 보호 등의 사업자 책무 등 다양한 조항들이 상호 연관되어 있다. 이들 조항 간의 체계적 관계를 명확히 하고, 각 제도가 상호보완적으로 작동할 수 있는 통합적 프레임워크를 개발하는 연구가 요구된다. 예컨대 영향평가 결과가 안전성 확보 조치나 위험관리 체계에 어떻게 반영되어야 하는지, 또 검·인증 제도와 어떻게 연계될 수 있는지에 대한 구체적인 메커니즘을 설계하는 연구가 중요할 것이다.

또한 AI 산업 육성과 규제의 적절한 조화를 이루는 방안에 대한 연구도 중요한 과제이다. AI 영향평가가 혁신을 저해하는 규제로 작용하지 않고 오히려 산업 발전의 촉매제로 기능할 수 있도록 하는 정책적 균형점을 찾는 연구가 필요하다. 이를 위해 AI 스타트업이나 중소기업에 대한 영향평가 지원 방안, 평가 부담을 경감하면서도 안전성과 신뢰성을 확보할 수 있는 차등적 접근법, 영향평가를 통한 기업의 경쟁력 강화 사례 발굴 등 규제와 산업 육성의 선순환 구조를 설계하는 연구가 진행되어야 할 것이다.

나아가 AI 영향평가를 통해 발견된 위험 요소나 기회 요인을 AI 연구개발 및 산업 지원 정책에 반영하는 피드백 체계 연구도 중요하다. 영향평가가 단순히 특정 AI 제품이나 서비스의 위험을 관리하는 데 그치지 않고, 국가 AI 산업 전략과 연구개발 방향 설정에 유용한 정보를 제공하는 전략적 도구로 활용될 수 있는 방안을 모색해야 한다.

산업계의 자율규제와 정부 규제의 조화를 이루는 공동규제(co-regulation) 모델에 대한 연구도 필요하다. AI 영향평가를 기업들이 자율적으로 시행하되 정부가 이를 감독하고 지원하는 체계, 산업별 특성을 반영한 자율 규제 가이드라인과 영향평가의 연계, 산업계-정부-시민사회가 함께 참여하는 협력적 거버넌스 모델 등 다양한 접근법에 대한 연구가 요구된다.

이러한 연구 방향은 궁극적으로 AI 기술의 발전과 안전, 혁신과 신뢰, 산업 경쟁력과 사회적 가치가 균형을 이루는 한국형 AI 생태계를 구축하는 데 기여할 것이다. 특히 AI 기본법이 단순한 규제 장치가 아닌, 한국 AI 산업의 글로벌 경쟁력을 높이고 AI 기술의 사회적 수용성을 강화하는 촉진제로 작용할 수 있도록 돕는 중요한 연구 기반이 될 것이다.

제6장 정부 정책 반영 현황

현재 정부는 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(AI 기본법)의 시행을 위한 하위 규정 마련을 적극적으로 추진하고 있다. 2025년 하반기 중으로 기본법 시행령 전문 공개 및 의견수렴 절차를 진행할 예정이며, 이를 통해 제도의 구체적 운영 방향을 확정할 계획이다. 다만, 시행령 조문만으로는 세부 이행방안과 구체적 절차를 모두 규정하는 데 한계가 있을 수 있다는 점을 고려하여, 정부는 시행령과 병행하여 고시와 가이드라인 등 하위 규범과 실무지침을 함께 마련하고 있다. 고시는 제도 운영의 법적 기반을 보완하고, 가이드라인은 사업자들이 실제 영향평가를 수행하는 과정에서 참고할 수 있도록 구체적 절차와 방법론을 제공하는 역할을 하게 된다.

본 연구를 통해 도출된 AI 영향평가 프레임워크(안)은 이러한 정부 정책 추진 방향과 긴밀히 연계된 것으로, 기본법 시행 이후 AI 영향평가 제도의 현장 안착과 실효성 확보를 지원하기 위한 구체적 기반 자료로 활용될 수 있을 것이다.

참 고 문 헌

〔**국내 문헌**〕

개인정보보호위원회(2023.10.13.). “공공기관의 대규모·민감 개인정보 처리 어떻게? 개인정보 영향평가 결과, 투명하게 공개한다!”[보도자료]. URL: https://pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS290&mCode=D050050000&nttId=10440

개인정보보호위원회(2024). “개인정보 영향평가 수행 안내서”. URL: https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=10089

과학기술정보통신부·한국과학기술기획평가원(2023). “2023 기술영향평가 보고서”.

과학기술정보통신부·정보통신정책연구원(2025). “2025 인공지능 윤리기준 실천을 위한 자율점검표”. URL: https://ai.kisdi.re.kr/aieth/bbs/B0000085/view.do?nttId=749&menuNo=400014&pageIndex=1

과학기술정보통신부·정보통신정책연구원(2024). “인공지능 윤리영향평가 프레임워크”. URL: https://ai.kisdi.re.kr/aieth/bbs/B0000085/view.do?nttId=642&menuNo=400014&pageIndex=1

관계부처 합동(2020.12.). “사람이 중심이 되는「인공지능(AI) 윤리기준」”.

관계부처 합동(2024.9.). “「국가 AI전략」 정책방향”.

국가인권위원회(2024). “인공지능 인권영향평가 도구”. URL: https://www.humanrights.go.kr/base/board/read?boardManagementNo=24&boardNo=7610404&menuLevel=3&menuNo=91

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.

양기문(2024). “미국과 유럽연합의 AI 영향평가 정책 현황”. 《KISDI Perspectives 2024.7. No.3》. 정보통신정책연구원.

〔해외 문헌〕

Committee on Artificial Intelligence(2024). *Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law(HUDERIA Methodology)*. Retrieved from https://rm.coe.int/cai-2024-16rev2-methodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f

EU(2012). *Charter of Fundamental Rights of the European Union*. Official Journal of the European Union, C 326, 391-407. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.C_.2012.326.01.0391.01.ENG&toc=OJ%3AC%3A2012%3A326%3ATOC

EU(2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, *Artificial Intelligence Act*.

Government of Canada(2023). *Directive on Automated Decision-Making*.

Government of Canada(2024). *Algorithmic Impact Assessment Tool*. Retrieved from https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

The Danish Institute for Human Rights(2020). *Guidance on Human Rights Impact Assessment of Digital Activities*. Retrieved from https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities

UNESCO(2021). *Recommendation on the Ethics of Artificial Intelligence*.

UNESCO(2023). *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000386276

〔기타〕

개인정보보호위원회 개인정보 영향평가제도 웹페이지. URL: https://www.pipc.go.kr/np/default/page.do?mCode=D050020000

과학기술정책지원서비스 기술영향평가 웹페이지. URL: https://k2base.re.kr/foresight/ta/intrcn/intrcnList.do?pageNavi=page2024&searchTap=notice&index=25

● 저 자 소 개 ●

문 정 욱

- 고려대학교 행정학 박사
- 현 정보통신정책연구원 디지털사회전략연구실 실장

강 하 연

- 충남대학교 정책학 석사수료
- 현 정보통신정책연구원 연구원

양 기 문

- 연세대학교 정보시스템학 석사
- 현 정보통신정책연구원 전문연구원

김 소 담

- 서울대학교 지능정보융합학 석사
- 현 정보통신정책연구원 연구원

방송통신정책연구 RS-2024-00512149

AI 영향평가 프레임워크 개발 및 적용에 관한 연구
(A Study on the Development and Implementation of an AI Impact Assessment Framework)

2025년 4월 일 인쇄
2025년 4월 일 발행

발행인 과학기술정보통신부 장관
발행처 과학기술정보통신부
세종 갈매로 477 정부세종청사 4동
Homepage: www.msit.go.kr
인 쇄 경 성 문 화 사

과학기술정보통신부
Ministry of Science and ICT

정보통신기획평가원
Institute of Information & Communications Technology Planning & Evaluation