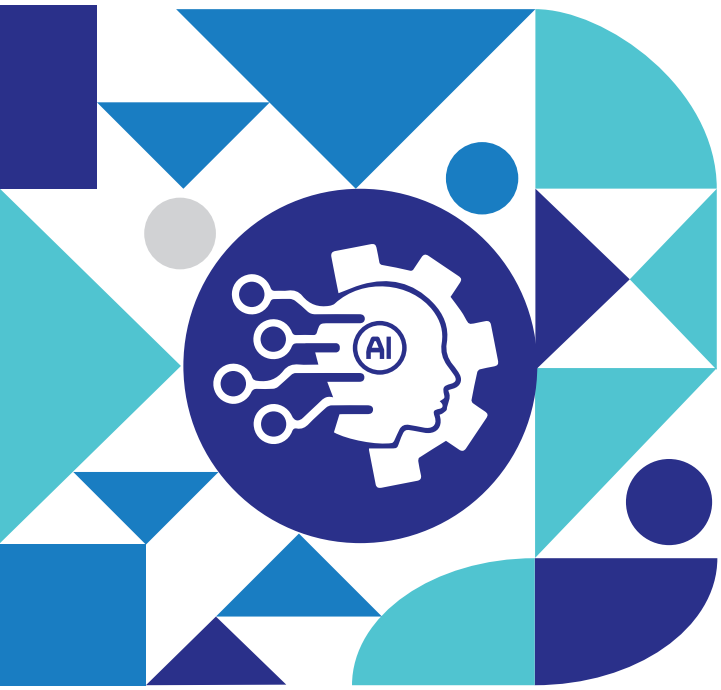
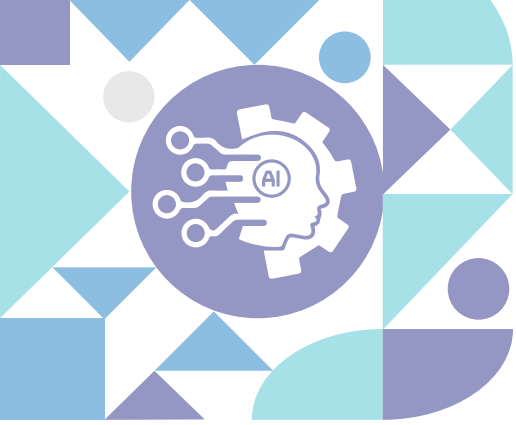
2025

인공지능 윤리기준 실천을 위한 자율점검표 (안)



과학기술정보통신부
Ministry of Science and ICT

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE



CONTENTS

인공지능 윤리기준 실천을 위한 자율점검표 (안)

일 러 두 기

- 「인공지능 윤리기준 실천을 위한 자율점검표 (안)」는 과학기술정보통신부 '사람중심의 인공지능 구현을 위한 인공지능 윤리정책 개발' 연구과제의 결과물로서 저작권자의 동의 없이 본 내용을 무단으로 전재하거나 복제할 수 없습니다.

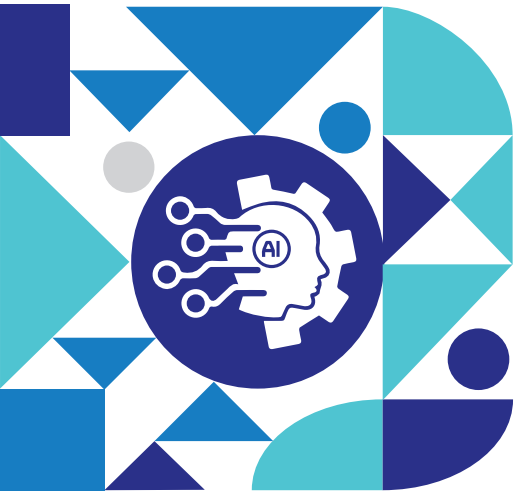
- 본 자율점검표 내용을 인용하거나 활용할 때, 출처 '과학기술정보통신부 · 정보통신정책연구원 「2025 인공지능 윤리기준 실천을 위한 자율점검표 (안)」'를 반드시 밝혀주시기를 바랍니다.

- 본 자율점검표는 2020년 12월 발표한 '사람이 중심이 되는 「인공지능(AI) 윤리기준」'이 제시하는 3대 기본원칙과 10대 핵심요건을 실천하기 위해 인공지능 분야 종사자가 고려해야 하는 요소와 이를 이행할 수 있는 방법을 다수의 점검문항으로 제공합니다.

- 본 자율점검표는 강제력 · 구속력 있는 '법규' 또는 '지침'이 아닌, 인공지능 분야에 종사하는 기업의 자율성을 존중하고 기술발전을 장려하며 기술 · 경제 · 사회의 변화에 유연하게 대처할 수 있는 자율적 윤리 규범으로서의 점검문항을 제시합니다.

- 본 자율점검표를 참고하는 주체는 각자의 목적과 특성에 맞추어 문항을 선별하고 유연하게 가공하여, 인공지능 윤리 기준을 실천할 수 있는 내부 지침이나 내부 규정을 마련하는 데 활용할 것을 권장합니다. 이를 통해 급변하는 국제 인공지능 산업 규제 및 윤리 제도에 선제적으로 대응할 수 있습니다.

- 본 자율점검표는 기술 · 경제 · 사회의 변화에 따라 새롭게 제기되는 인공지능 윤리 쟁점을 반영하고자 다양한 이해관계자의 의견 청취, 실제 현장에서의 시범 적용, 전문가 자문 등을 거쳐 지속적으로 수정 · 보완될 예정입니다.



1

인공지능 윤리기준 실천을 위한 자율점검표 추진 배경

01 인공지능 윤리기준 실천을 위한 자율점검표 추진 배경

인공지능[1]이 급속하게 발전하고 산업 전 분야에 적용됨에 따라 생산성과 편의성이 향상되고 사회 혁신이 가속화되어 국민의 삶의 질이 높아질 것으로 기대하고 있습니다. 반면에 인공지능의 확산으로 방대한 데이터가 수집되고 알고리즘이 활용되면서 다양한 윤리적·사회적 문제가 대두되는 등 인공지능 기술에 대한 우려의 목소리도 나타나고 있습니다. 인공지능 기반 채용 시스템의 성차별(2018년 아마존 AI 리크루팅), 인공지능 기반 범죄 예측 프로그램 활용에 따른 인종차별(2018년 미국 COMPAS), 인공지능 챗봇의 비윤리적 발언(2016년 미국 마이크로소프트 Tay) 등은 인공지능의 활용 과정에서 발생할 수 있는 위험이나 부작용의 대표적인 사례입니다.

기타 인공지능 활용 과정에서 발생한 부작용 사례

인공지능 활용 사례	주요 내용
한국 인공지능 챗봇 '이루다' (2021)	스타트업 '스캐터랩'의 챗봇 '이루다'는 데이터 구축·학습 과정에서의 개인정보유출, 장애인·성소수자·인종 등에 대한 혐오 발언, 이용자의 챗봇 성희롱 대화로 젠더편견 강화 등 사회적 논란을 빚어 서비스 중단
영국 대입 시험 점수산정 시스템 (2020)	영국 시험감독청은 코로나19로 취소된 대입 시험 'A레벨'을 대신하여 알고리즘으로 성적을 산출하였으나 편향된 알고리즘 결과에 대해 강한 사회적 반발이 일어나자 이를 철회함
미국 경찰 범죄수사용안면인식 시스템 (2020)	미국 경찰국은 안면인식 기술을 활용한 범죄 수사 사건에 흑인 3명을 범죄자로 오인하여 부당 고소·체포함으로써 인권침해와 인종차별 논란을 빚음
네덜란드 복지수당 사기탐지 시스템 (2020)	네덜란드 정부는 복지혜택 부정수급과 세금 사기를 단속하기 위해 위험탐지시스템(SyRi)을 개발·활용하였으나 국가·지방자치단체의 데이터를 활용한 사생활 침해, 저소득층이나 이민자 등 소수·취약집단 차별, 비공개 인공지능 모델·데이터의 투명성 부족 문제를 지적한 법원 판결로 철회됨

이러한 배경에서 2019년 12월에 발표한 우리나라의 「인공지능 국가전략」은 '사람 중심의 인공지능 구현'을 강조하였으며, 지능정보사회가 불러올 혜택을 극대화하고 발생할 수 있는 사회적 위험을 최소화하여 인공지능에 대한 사회적 신뢰를 확보하기 위해 2020년 12월 '사람이 중심이 되는 「인공지능(AI) 윤리기준」'을 수립하였습니다. 인공지능 윤리는 기술 발전을 저해하지 않음과 동시에 민간의 자율적인 노력을 존중하기 위한 새로운 사회규범으로서, '사람 중심'이라는 인공지능 개발과 활용의 방향성을 제시하여 인공지능의 도입과 확장으로 커지는 우려를 완화하고 파생되는 역기능을 최소화하기 위해 제시하는 규범입니다. 과학기술정보통신부가 수립한 「인공지능(AI) 윤리기준」은 특정 분야에 제한받지 않고 사회 구성원 모두가 참조할 수 있는 보편적인 윤리원칙으로 구성되어 있습니다.

「인공지능(AI) 윤리기준」은 인공지능이 인간에게 유용하고 정신과 신체에 해롭지 않도록 개발·활용되어야

1) 인공지능이란 인간의 지적능력을 컴퓨터로 구현하는 과학기술로서, ① 상황을 인지하고, ② 이성적·논리적으로 판단·행동하며, ③ 감성적·창의적인 기능을 수행하는 능력까지 포함(인공지능 국가전략, 2019.12)

하고, 개인의 행복과 사회의 긍정적 변화 등에 기여해야 한다는 '인간성을 위한 인공지능(AI for Humanity)'이라는 가치를 추구합니다. 해당 가치를 구현하기 위한 3대 기본원칙(인간 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙)을 도출하고, 이를 실현할 수 있도록 인공지능 생명주기에 걸쳐 충족되어야 하는 10가지 핵심요건 (① 인권보장 ② 프라이버시 보호 ③ 다양성 존중 ④ 침해금지 ⑤ 공공성 ⑥ 연대성 ⑦ 데이터관리 ⑧ 책임성 ⑨ 안전성 ⑩ 투명성)을 제시합니다.

참고 **사람이 중심이 되는 「인공지능(AI) 윤리 기준」('20.12)**

[수립 배경과 목적]

- 2020년 12월, 인공지능의 개발부터 활용에 이르는 전 과정에서 정부·공공기관, 기업, 이용자 등 모든 사회 구성원이 '사람 중심의 인공지능' 구현을 위해 참조할 수 있는 「인공지능(AI) 윤리기준」을 수립
- 기업의 자율성을 존중하고 기술 발전을 장려하고자 법적 구속력이 없는 인공지능 윤리를 제시하여 분야별 혹은 사안별 특징에 맞게 윤리기준을 중심으로 세부 지침을 수립해나갈 수 있는 자율 규범 환경 조성을 목적으로 함

[구조와 주요 내용]

- 인공지능 개발에서 활용에 이르는 전 과정에서 고려해야 하는 3대 기본원칙을 도출하고, 이를 실현할 수 있는 세부적인 요건으로 10대 핵심요건을 제시

구조

기본원칙 주요내용

① 인간 존엄성 원칙

- 인간은 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계제품과는 교환 불가능한 가치가 있다.
- 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용되어야 한다.
- 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.

② 사회의 공공선 원칙

- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙

- 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.
- 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.

첫 번째 기본원칙인 '인간 존엄성 원칙'은 인공지능을 비롯하여 인간을 위해 개발된 기계와 인간은 교환할 수 없으며, 인간의 생명과 신체적·정신적 건강에 해가 되지 않는 범위에서 개발·활용해야 하며, 인공지능의 개발과 활용에서도 이러한 원칙 아래 안전성과 견고성을 갖추어 인간에게 해가 되지 않아야 한다는 점을 강조합니다. 두 번째 기본원칙인 '사회의 공공선 원칙'은 공동체인 사회는 다수의 안녕과 행복이라는 가치를 추구하고, 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발하고 활용해야 하며, 공익 증진을 위한 인공지능 개발 및 활용은 사회적·국가적 나아가 글로벌 측면에서도 인류의 보편적 복지를 향상시킬 수 있어야 한다는 것을 의미합니다. 세 번째 기본원칙인 '기술의 합목적성 원칙'은 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합하도록 개발 및 활용하고, 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여야 한다는 원칙입니다.

3대 기본원칙을 실천하고 이행할 수 있도록 10가지 핵심 요건을 제시하고 있지만, 인공지능 윤리가 비교적 새로운 사회적 규범이라는 점에서 아직 이에 익숙하지 않은 개인이나 조직이 현장이나 일상에서 바로 적용하고 실천하기에는 어려움이 있습니다. 그래서 원칙을 발표한 주체는 인공지능 윤리의 실천 여부를 사회구성원이 자율적으로 판단할 수 있는 수단을 개발하여 제시하는 사례를 참조하였습니다. 일례로 유럽연합(EU)의 '신뢰할 수 있는 인공지능을 위한 자체평가리스트'('20.7), 세계경제포럼(World Economic Forum)과 싱가포르 정부의 '인공지능 거버넌스 프레임워크-조직을 위한 실행 및 자체평가안내서'('20.1) 등에서는 인공지능 기술을 연구하고 인공지능 기술 기반 제품 또는 서비스를 개발하는 영역에서 윤리원칙 실천에 참조할 수 있는 다양한 평가항목을 제공하였습니다.[2] 공공부문뿐만 아니라 인공지능 분야 학계나 기업에서도 기술 활용 목적과 조직의 특수성을 고려한 원칙을 수립하여 내부 관련자를 교육하고, 새로운 윤리적 쟁점을 주시하고 유연하게 변경하여 활용하고 있는 것으로 알려져 있습니다.

해외 인공지능 윤리원칙 실천 수단 사례

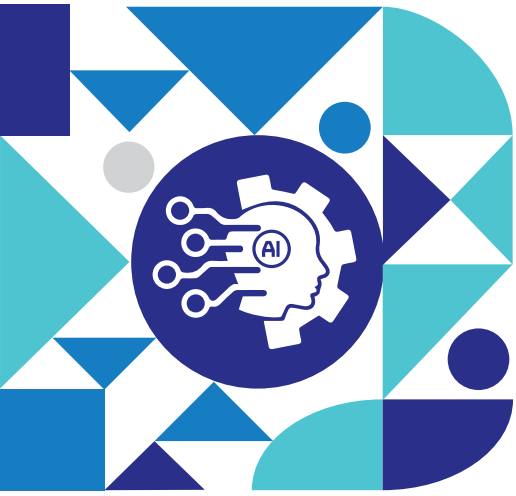
발표 주체	발표 시기	제목
유럽연합	2020	The Assessment List for Trustworthy Artificial Intelligence for self assessment
영국 개인정보보호감독기구	2020	Guidance on Artificial Intelligence and Data Protection
세계경제포럼, 싱가포르 정부	2020	Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations

이러한 배경에서 「인공지능 윤리기준 실천을 위한 자율점검표(이하, 윤리기준 자율점검표)」는 인공지능 분야 종사자가 「인공지능(AI) 윤리기준」을 현장에서 실천하기 위한 자발적인 노력을 지원하기 위하여 개발되었습니다. 윤리기준 자율점검표는 「인공지능(AI) 윤리기준」이 제시한 3대 기본원칙을 위한 10대 핵심요건을 실천하기 위해 염두에 두어야 하는 요소와 이를 이행하는 방법을 다수의 문항으로 제공합니다. 인공지능 기술을 개발하고 발전시키기 위한 과정에서 각 핵심요건을 만족하는지를 확인할 수 있도록 핵심요건별로 점검문항을 구성하였습니다. 또한 본 자율점검표는 강제력·구속력이 있는 '법규' 또는 '지침'이

2) '부록 4. 해외 인공지능 윤리 실천수단' 참조

아닌, 인공지능 분야에 종사하는 기업의 자율성을 존중하고 기술발전을 장려하며 기술·경제·사회의 변화에 유연하게 대처할 수 있는 자율적 도덕 규범인 점검문항으로 구성되어 있습니다. 더불어 전공분야, 산업부문, 세부단계 등을 특정하지 않는 범용성 있는 점검문항을 제시하여, 이를 참고하는 주체가 각자의 목적과 특성에 맞추어 문항을 선별하고 유연하게 가동하여 활용할 수 있도록 권장합니다.

윤리기준 자율점검표를 통해 인공지능 기술이 불러올 사회·경제적 혜택과 역기능을 균형 있게 인식하고 역기능 발생 가능성을 사전에 환기하는 기회를 마련하여, 인공지능 윤리기준의 수립 목적인 혜택을 극대화하고 발생할 수 있는 사회적 위험을 최소화하여 인공지능에 대한 사회적인 신뢰를 확보할 수 있기를 바랍니다.



2

인공지능 윤리기준 실천을 위한 자율점검표

02 인공지능 윤리기준 실천을 위한 자율점검표

1. 점검 목적

윤리기준 자율점검표의 목적은 인공지능시스템의 개발·운영 과정에서 국가 「인공지능(AI) 윤리기준」이 제시한 3대 기본원칙을 구현하기 위한 10대 핵심요건을 실천하기 위함입니다.

2. 권장 대상

윤리기준 자율점검표의 권장 대상은 인공지능시스템의 개발·운영에 필요한 의사결정을 수행하는 개인과 조직입니다. 공공과 민간부문의 구분 없이 인공지능시스템의 개발·운영의 목적과 절차를 수립하고 규정을 정하는 개인이나 조직이 본 자율점검표의 점검문항을 참조하여 인공지능 윤리기준을 실천할 수 있는 내부지침을 별도로 마련하거나 내부 규정에 반영할 수 있습니다. 인공지능시스템을 설계·제작하고, 데이터와 알고리즘을 통해 인공지능시스템을 구현하고, 인공지능시스템을 유지·관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 35개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	5	2	5	4	3	3	3	4	3	3

4. 인공지능 윤리기준 실천을 위한 자율점검표

10대 핵심요건에 해당하는 자율점검항목을 다음의 표로 제공합니다.

인공지능 윤리기준 실천을 위한 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발·운영 과정에서 국가 「인공지능(AI) 윤리기준」이 제시한 3대 기본원칙을 구현하기 위한 10대 핵심요건을 실천하기 위함입니다.
- 본 자율점검표의 권장 대상은 인공지능시스템의 개발·운영에 필요한 의사결정을 수행하는 개인과 조직입니다. 공공과 민간부문의 구분 없이 인공지능시스템의 개발·운영의 목적과 절차를 수립하고 규정을 정하는 개인이나 조직이 본 자율점검표의 점검문항을 참조하여 인공지능 윤리기준을 실천할 수 있는 내부지침을 별도로 마련하거나 내부 규정에 반영할 수 있습니다.
- 인공지능시스템을 설계·제작하고, 데이터와 알고리즘을 통해 인공지능시스템을 구현하고, 인공지능시스템을 유지·관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	인공지능시스템이 인간의 생명과 안전에 관한 권리를 침해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 02	인공지능시스템이 모든 인간을 평등한 존재로 대우함으로써 성별, 연령, 지역, 종교, 인종, 민족, 경제적 수준, 성적 지향, 정치적 성향, 장애 등을 이유로 차별하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 03	인공지능시스템이 사용자의 자율적 행동이나 결정을 방해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 04	인공지능시스템이 사용자의 언론·출판·집회·결사 등 표현의 자유를 침해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 05	인공지능시스템이 사용자에게 불쾌감을 주지 않는 등 인간을 인격적으로 대우하도록 개발·운영하고 있는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	인공지능시스템의 개발·운영 과정에서 개인정보를 수집·활용하는 경우, 개인정보 보호법 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	인공지능시스템의 개발·운영 과정에서 사생활의 비밀과 자유를 침해할 우려에 대하여 검토하고 있는가?	☐	☐	☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	인공지능시스템 활용에 사회적 약자의 접근 가능성을 고려하고 있는가?	☐	☐	☐
E03. 02	인공지능시스템의 개발에 활용되는 데이터의 성별, 인종, 민족, 국가 등 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고 이를 최소화하기 위해 노력하고 있는가?	☐	☐	☐

E03. 03	인공지능시스템의 개발·운영 단계에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐
E03. 04	인공지능시스템을 사용할 때 편향이나 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 운영자, 사용자 모두 내부 또는 인공지능시스템 개발조직과 운영업체에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐
E03, 05	인공지능시스템 개발자를 대상으로 인공지능시스템에서 발생할 수 있는 편향성의 인지 또는 분석 능력 향상을 위한 교육훈련의 기회를 제공하고 있는가?	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	인공지능시스템이 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고 이를 예방하기 위한 조치를 취하였는가?	☐	☐	☐
E04. 02	인공지능시스템의 목적 외 사용으로 인해 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 개연성이 확인된 경우, 사용자에게 고지하는 절차를 마련하였는가?	☐	☐	☐
E04. 03	인공지능시스템의 활용 과정에서 예상하지 못한 피해가 발생할 때, 사용자가 해당 피해를 신고하고 의견을 제시할 수 있는 절차를 마련하였는가?	☐	☐	☐
E04. 04	인공지능시스템의 활용 과정에서 예상하지 못한 중대한 피해가 발생할 때, 피해의 확산을 방지하기 위해 이미 상용화된 시스템의 사용중단 또는 리콜, 정부 소관기관에 보고, 사용자에게 고지 등의 절차를 마련하였는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	인공지능시스템이 특정 개인이나 집단의 이익을 대변하여 공익을 훼손하거나 역기능을 발생시킬 가능성을 고려하고 있는가?	☐	☐	☐
E05. 02	인공지능시스템 사용으로 폭력성, 음란성, 사행성, 중독성이 조장되는 등 부작용이 발생할 개연성이 있는지를 고려하고 있는가?	☐	☐	☐
E05. 03	인공지능시스템이 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 내부적으로 검토하거나 외부 전문가의 의견을 청취하였는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	인공지능시스템 개발·운영 목적의 범위 내에서 다양한 배경의 개발자나 사용자가 의사소통이나 상호작용할 수 있는 기회를 제공하고 있는가?	☐	☐	☐
E06. 02	인공지능시스템의 사용이 지역·성별·세대·계층 간 갈등을 유발하는 등 사회통합을 저해할 개연성이 있는지를 고려하고 있는가?	☐	☐	☐
E06. 03	탄소중립을 위한 국제사회의 노력에 협력하기 위해 인공지능시스템의 개발·운영 과정에서 탄소배출이 적은 방법을 사용하도록 고려하고 있는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	인공지능시스템의 개발·운영에 활용되는 데이터의 수집과 처리 업무의 감독을 위한 절차를 마련하였는가?	☐	☐	☐
E07. 02	인공지능시스템의 개발에 활용되는 데이터의 출처·처리의 주요 과정을 기록하고 있는가?	☐	☐	☐
E07, 03	인공지능시스템의 개발·운영에 활용되는 데이터의 분석과 관리 업무에 대한 기술적·물리적 통제방안을 마련하였는가?	☐	☐	☐

E08. 책임성

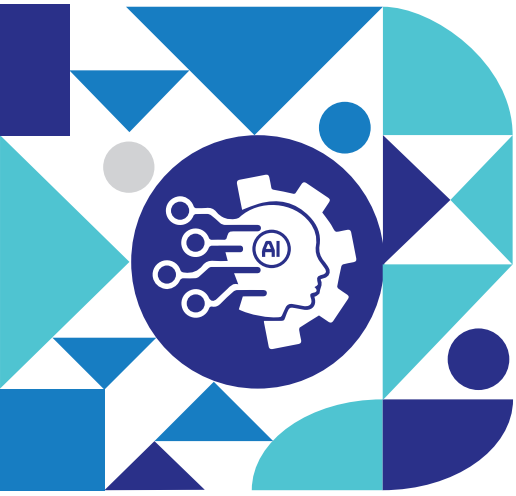
		YES	NO	미해당
E08. 01	인공지능시스템을 개발·운영하는 과정에서 윤리기준 준수를 보장하기 위해 담당자 지정 등 적절한 방안을 마련하였는가?	☐	☐	☐
E08. 02	인공지능시스템 개발자 또는 개발부서는 다음의 역량을 향상하기 위해 노력하고 있는가? – 인공지능시스템이 활용되는 분야의 적합성을 판단하고 위험성을 인지하는 능력 – 인공지능시스템의 산출물과 결정을 해석할 수 있는 능력	☐	☐	☐
E08. 03	인공지능시스템의 개발·운영에서 발생하는 손해 또는 손실의 책임 소재를 명확히 하고 있는가?	☐	☐	☐
E08. 04	인공지능시스템의 사용으로 발생한 피해에 대하여 합리적인 배상과 보상을 제공하기 위해 준비 방안(배상책임보험의 가입이나 유보금을 적립하는 방안 등)을 마련하였는가?	☐	☐	☐

E09. 안전성

		YES	NO	미해당
E09. 01	인공지능시스템의 비정상 동작이나 예기치 못한 오류에 대한 안전조치 기능과 안전조치 기능의 한계에 대해 이해관계자와 사용자에게 충분한 정보를 제공하고 있는가?	☐	☐	☐
E09. 02	인공지능시스템과 인간의 상호작용에서 발생할 수 있는 위험 (인공지능에 의한 감시, 중독, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?	☐	☐	☐
E09. 03	인공지능시스템의 결과에 대한 안전성을 지속적으로 평가하기 위한 절차 (정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?	☐	☐	☐

E10. 투명성

		YES	NO	미해당
E10. 01	인공지능시스템을 활용한 제품 또는 서비스가 인공지능 알고리즘 기반의 결정을 한다는 사실과 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 고지하고 있는가?	☐	☐	☐
E10. 02	인공지능시스템을 목적에 맞게 사용하기 위해 이해관계자와 사용자에게 관련 정보 (예: 가이드북, 매뉴얼 등)를 제공하고 있는가?	☐	☐	☐
E10. 03	인공지능시스템이 수집하는 데이터, 의사결정의 결과에 영향을 미치는 주요 요인 등 사용자가 설명요청하는 정보를 제공할 수 있는 절차를 마련하였는가?	☐	☐	☐



3

10대 핵심요건별 점검문항 세부 내용

03 10대 핵심요건별 점검문항 세부 내용

핵심요건 1 인권보장

- 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제인권법 등에 명시된 권리를 보장하여야 한다.
- 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

실천 점검문항

E01. 01	인공지능시스템이 인간의 생명과 안전에 관한 권리를 침해하지 않도록 개발·운영하고 있는가?
E01. 02	인공지능시스템이 모든 인간을 평등한 존재로 대우함으로써 성별, 연령, 지역, 종교, 인종, 민족, 경제적 수준, 성적 지향, 정치적 성향, 장애 등을 이유로 차별하지 않도록 개발·운영하고 있는가?
E01. 03	인공지능시스템이 사용자의 자율적 행동이나 결정을 방해하지 않도록 개발·운영하고 있는가?
E01. 04	인공지능시스템이 사용자의 언론·출판·집회·결사 등 표현의 자유를 침해하지 않도록 개발·운영하고 있는가?
E01. 05	인공지능시스템이 사용자에게 불쾌감을 주지 않는 등 인간을 인격적으로 대우하도록 개발·운영하고 있는가?

제안 이유

인공지능시스템의 개발과 활용에 따라 인간이 가지는 보편적·불가침적·절대적 자유와 권리를 침해하는 결과가 나타날 수 있다. 따라서 인공지능시스템의 개발자와 운영자는 「대한민국헌법」과 법률이, 그리고 「유엔헌장」, 「세계인권선언」 등과 같은 국제인권법규가 보호하려는 인간의 존엄성을 훼손하거나 자유권, 평등권 등의 기본권을 침해하지 않는지를 평가하고 점검하는 것이 중요하다. 본 영역에서는 인간의 기본권과 동시에 자율성을 존중하기 위해 인공지능시스템을 기획하고 설계하는 단계에서부터 염두에 두어야 할 가치를 제시하고 이를 위해 사전적·사후적인 노력을 기울일 것을 요구한다.

참고1 | 유럽연합(EU) 'ALTAI'의 기본권영향평가(FRIA) 문항

☞ **유럽연합은 신뢰할 수 있는 인공지능 윤리 가이드라인으로 신뢰가능한 AI 영향평가리스트(Assessment List for Trustworthy AI, ALTAI)를 제안하며, ALTAI에서는 인공지능시스템이 평등권 등 기본권에 미치는 영향을 사전적으로 평가하기 위해 다음과 같은 질문에 따른 자체평가 실시를 권고함**

① **인공지능시스템이 다음과 같은 근거를 바탕으로 사람들을 부정적으로 차별할 우려가 있는가?:** 성별, 인종, 피부색, 민족 또는 사회적 출신, 유전적 특징, 언어, 종교 또는 신념, 정치적 또는 기타 의견, 소수 민족, 재산, 출생, 장애, 연령, 성적 취향 등

- 인공지능시스템의 개발·배치·사용 단계에서 부정적 차별(편견) 우려가 있는지 시험하고 모니터링하는 프로세스를 마련하였는가?
- 인공지능시스템에서 부정적 차별(편견) 가능성을 해결하고 시정하는 프로세스를 마련하였는가?

② 인공지능시스템이, 예를 들어 아동 보호와 아동 최선의 이익 원칙을 고려하여 아동의 권리를 존중하고 있는가?

- 인공지능시스템이 아동에 대한 잠재적 위해를 해결하고 시정하는 프로세스를 마련하였는가?
- 인공지능시스템의 개발·배치·사용 단계에서 아동에게 미치는 잠재적 위험을 예방할 수 있도록 시험하고 점검하는 절차를 마련하였는가?

③ 인공지능시스템이 GDPR에 따라 개인정보를 보호하는가?

- 인공지능시스템 개발·배치·사용 단계에서 개인정보영향평가(특히, 각 단계의 목적에 부합하기 위한 해당 절차의 필요성과 비례성 평가 포함)의 필요 여부를 판단하는 절차를 마련하였는가?
- 인공지능시스템의 개발·배치·사용 단계와 관련하여 개인정보를 보호하기 위한 안전장치, 보안 수단 등 위험을 해결하기 위한 조치를 취했는가?

④ 인공지능시스템이 표현·정보의 자유와 집회·연대의 자유를 존중하는가?

- 인공지능시스템의 개발·배치·사용 단계에서 표현·정보의 자유와 집회·연대의 자유를 침해할 우려가 있는지 시험하고 모니터링하는 프로세스를 마련하였는가?
- 인공지능시스템에서 표현·정보의 자유와 집회·연대의 자유를 침해할 우려를 해결하고 시정하기 위한 프로세스를 마련하였는가?– 인공지능시스템에서 표현·정보의 자유와 집회·연대의 자유를 침해할 가능성을 해결하고 시정하기 위한 프로세스를 마련하였는가?

자료: EU, Assessment List for Trustworthy Artificial Intelligence(2020)

참고2 | WHO의 '인공지능의 윤리적 활용을 위한 가이드라인'

☞ **WHO는 헬스케어 분야의 인공지능 기술 활용에 따른 윤리적 가이드를 발표하였으며, 헬스케어 분야 인공지능 기술 개발자에게 인간의 자율성 보장을 위해 다음과 같은 원칙과 고려사항을 제시함**

(인간의 자율성 보존과 강화) 헬스케어 부문의 인공지능 기술은 의료 전문가를 대체하는 것이 아니라 인간의 의사결정을 강화하고 권한을 부여해야 한다.

- 인공지능 기술에 의한 추천 치료 또는 질병 예측과 관련하여 인간의 판단이 개입되어야 한다.
- 보건부는 인공지능의 결과에 대해 임상의가 독립적인 판단을 내리는 데 사용하는 정보의 유형을 지정해야 한다.
- 인공지능 기술 기반의 추천 결과라는 사실을 환자에게 알리고, 환자의 의사결정을 위해 의미 있고 명확한 정보를 제공해야 한다.

자료: WHO, Ethics and governance of artificial intelligence for health: WHO guidance(2021), p143

참고3 | 유럽연합(EU) '인공지능, 로봇 및 관련 기술의 개발·배포·사용을 위한 윤리적 원칙에 관한 유럽의회와 이사회 규정안'

☞ **유럽연합은 '인공지능, 로봇 및 관련 기술의 윤리적 측면 프레임워크를 위한 대 위원회 권고보고서'*를 2020년 10월 20일 유럽의회에서 채택**

- 인공지능, 로봇 및 관련 기술이 윤리적 원칙에 부합하는 방식으로 개발·배포·사용하도록 보장하고, 이러한

기술이 EU의 법률·기본권·가치에 따른 윤리적 원리에 부합하도록 보장하기 위한 수단으로서 인공지능 등의 기술을 개발·배포·사용하는 기관과 시민 사이의 투명성과 더 나은 정보 흐름을 요구하는 것을 목적으로 함

* EU, proposal for REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on ethical principles for the development, deployment and use of artificial intelligence, robotics and related technologies

☞ **본 보고서에는 인공지능, 로봇 및 관련 기술의 개발·배포·사용을 위한 윤리적 원칙에 관한 유럽의회와 이사회 규정안을 제시하였으며, 규정안 제5조에서 인공지능 관련 기술의 기본원리로 기본권 존중에 관하여 명시**

(제5조) 인공지능, 로봇 및 관련 기술의 윤리적 원리

① 인공지능, 로봇 및 관련 기술로 사용하거나 생산하는 소프트웨어·알고리즘·데이터를 포함한 모든 인공지능, 로봇 및 관련 기술은 유럽연합법에 따라, 그리고 헌장에 명시된 인간의 존엄성, 자율성, 안전에 관계되는 기본권과 기타 기본권을 완전히 존중하여 유럽연합에서 개발·배포·사용해야 한다.

② 비개인정보 또는 생체인식정보로부터 파생된 개인정보를 포함하여 인공지능, 로봇 및 관련 기술의 개발·배포·사용에서 수행되는 개인정보의 모든 처리는 EU 2016/679 규정과 2002/58/EC 지침에 따라 수행되어야 한다.

③ 유럽연합과 회원국은 인공지능, 로봇 및 관련 기술에 기초하여 사회적 포용, 민주주의, 다원성, 연대, 공정성, 평등, 협력을 촉진하기 위한 연구 프로젝트를 장려한다.

자료: EU, 'REPORT with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies' (2020)

Case Study | 핵심요건1 '인권보장' 관련 사례

☞ **미국 경찰국이 안면인식 기술을 범죄수사에 활용하는 과정에서 무고한 시민을 범죄자로 오인하여 부당 체포한 사례는 인공지능 기술의 오남용에 따른 인권침해 가능성을 시사함. 인공지능 기술 자체의 문제가 아니라 기술을 이용하는 과정에서, 1) 동의받지 않은 카메라 기반의 무차별적인 무단 검문이나, 2) 사람의 검증 없이 기술에만 전적으로 의존하여 자동으로 체포하는 등의 문제 발생**

(개요) 미국 경찰국은 안면인식 기술을 활용한 범죄 수사 사건에 무고한 시민인 니지어 파크스('19.1), 마이클 올리버('19.5), 로버트 윌리엄스('20.1)를 범죄자로 인식하여 체포·구금 결정을 내림

- 로버트 윌리엄스는 30시간 동안 교도소에 구금되었고, 미국 시민자유연합(ACLU)은 그를 대신하여 디트로이트시를 상대로 소송을 제기함 ('20.6). 니지어 파크스는 10일 동안 수감되었고, 이후 불법 체포, 불법 수감, 시민권 침해 등으로 뉴저지 소재 카운티에 소송제기 ('20.11)

(후속 조치) 일부 주 또는 시에서는 관련 법안을 마련하여 안면인식 기술을 규제

- 캘리포니아주는 경찰국의 보디카메라 영상을 사용한 안면인식 기술을 3년간 금지하는 법안 'AB-1215 Law enforcement: facial recognition and other biometric surveillance' 시행('20.1)

- 워싱턴주는 공공부문의 안면인식 기술 활용 규제 'SB 6280 (Concerning the use of facial recognition services)' 법안을 시행('21.7). 특히, 안면인식 기술을 활용하여 개인에게 법적 영향을 미치는 결정을 내릴 때 '의미있는 인적검토(meaningful human review)'가 있어야 하고, 서비스 제공자에게 특정 인구집단에 대한 결과의 정확성과 (불)공정성을 테스트할 수 있는 API를 마련하도록 요구

핵심 요건 2

프라이버시 보호

- 인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.
- 인공지능 전 생애주기에 걸쳐 개인 정보의 오용을 최소화 하도록 노력해야 한다.

실천 점검문항

E02. 01 인공지능시스템의 개발·운영 과정에서 개인정보를 수집·활용하는 경우, 개인정보 보호법 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?

E02. 02 인공지능시스템의 개발·운영 과정에서 사생활의 비밀과 자유를 침해할 우려에 대하여 검토하고 있는가?

제안 이유

인공지능을 개발하고 활용하는 데 사용되는 대규모 데이터에는 다양한 유형의 개인정보와 민감한 사생활 정보가 포함될 수 있다. 데이터에는 정형화된 개인정보와 함께, 온라인에서 수집할 수 있는 개인정보가 포함될 수 있다. 일반적으로 정보 제공자와 사용자는 개인정보가 어떻게 처리되는지 알기 어렵고, 침해되거나 유출되면 개인의 삶과 사회에 부정적인 영향을 끼칠 수 있어 인공지능 분야에서는 프라이버시 보호를 위해 노력하고 있다. 본 영역에서는 인공지능시스템을 개발 또는 운영하는 주체의 「개인정보 보호법」 준수와 개인의 사생활 보호를 위한 노력을 평가한다.

참고1 | WHO의 '인공지능의 윤리적 활용을 위한 가이드라인'*

☞ **데이터와 알고리즘의 역할이 인공지능 개발에서 중요한 부분을 차지하나, 활용 데이터에 개인정보가 포함되면 개인의 사생활 침해로 이어지고 나아가 관련 법령 위반이 되므로 각국 정부는 인공지능 활용 분야에서 개인정보보호를 위해 별도의 세분화된 자율점검 문항을 개발하고 있음**

① 영국 개인정보감독기구(ICO), '인공지능과 데이터보호 안내지침' ('20.7.30)

- 인공지능에 대한 감시 프레임워크를 제공하여 개인정보의 공정한 처리를 보장하기 위해 책임성과 거버넌스 구현, 데이터 처리의 적법성, 공정성, 투명성, 보안의 확보, 개인정보처리의 최소화, 개인정보와 관련된 권리의 보장 등을 위해 180여 개 문항 제시

② 한국 개인정보보호위원회, '인공지능 개인정보보호 자율점검표' ('21.5.31)

- 「개인정보 보호법」상 개인정보보호 원칙을 기본으로, '개인정보보호 중심 설계(PbD)' 원칙, '인공지능 윤리기준'('20.12)을 반영하여 인공지능 관련 개인정보보호 6대 원칙을 도출하고, 단계별(또는 상시)로 의무 또는 권장하는 내용에 대한 점검항목(16개)과 확인사항(54개) 제시

자료: https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=7347

참고2 | EU GDPR 제22조 프로파일링 등 자동화된 개별 의사결정

☞ **GDPR 제22조 제1항은 기본적으로 정보주체에게 법적 효력을 발생시키거나 이와 유사하게 정보주체에게 중대한 영향을 미치는 프로파일링을 포함하여 완전하게 자동화된 의사결정을 금지, 제2항은 기본적 금지에 대한 예외 인정, 제3항은 정보주체의 권리·자유·정당한 이익을 보호하는 조치를 규정함**

1. 개인정보주체는 프로파일링 등, 본인에 관한 법적 효력을 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 자동화된 처리에만 의존하는 결정의 적용을 받지 않을 권리를 가진다.
2. 결정이 다음 각 호에 해당하는 경우에는 제1항이 적용되지 않는다.
   (a) 개인정보주체와 컨트롤러 간의 계약을 체결 또는 이행하는데 필요한 경우
   (b) 컨트롤러에 적용되며, 개인정보주체의 권리와 자유 및 정당한 이익을 보호하기 위해 적절한 조치를 규정하는 유럽연합 또는 회원국 법률이 허용하는 경우
   (c) 개인정보주체의 명백한 동의에 근거하는 경우
3. 제2항 (a)호 및 (c)호 사례의 경우, 컨트롤러는 개인정보주체의 권리와 자유 및 정당한 이익, 최소한 컨트롤러의 인적개입을 확보하고 본인의 관점을 피력하며 결정에 대해 이의를 제기할 수 있는 권리를 보호하는 데 적절한 조치를 시행해야 한다.
4. 제2항의 결정은 제9조 (2)의 (a)호와 (g)호가 적용되고, 개인정보주체의 권리와 자유 및 정당한 이익을 보호하는 적절한 조치가 갖추어진 경우가 아니라면 제9조(1)의 특정 범주의 개인정보를 근거로 해서는 안 된다.

참고3 | 유럽연합(EU) 'ALTAI'의 인공지능 평가문항

☞ **유럽연합 ALTAI는 프라이버시 보호와 데이터 거버넌스 원칙의 준수를 위해 필요한 평가 내용을 다음과 같이 제시**

1. 프라이버시에 대한 평가

→ 인공지능시스템이 개인정보보호에 관한 권리, 신체에 관한 권리, 정신적·도덕적 무결성에 미치는 영향을 고려하고 있는가?

→ 이용사례에 따라 인공지능시스템과 관련한 사생활 침해 문제에 대응할 수 있는 메커니즘을 확립하였는가?

2. 데이터 거버넌스에 대한 평가

→ 인공지능시스템이 개인정보(특수 범주의 개인정보를 포함)를 사용하거나 처리하는가?

→ GDPR 또는 비유럽 국가의 유사 법률에 따른 의무조치를 수행하였는가?

- 해당 조치는 개인정보영향평가(Data Protection Impact Assessment, DPIA), 개인정보보호책임자(Data Protection Officer, DPO)를 지정하여 인공지능시스템의 개발과 조달 또는 사용단계의 초기에 포함시키는 것, 데이터 처리를 위한 감시 메커니즘(접근 권한, 데이터 엑세스 기록과 수정 메커니즘 등을 포함), 프라이버시 바이 디자인과 암호화(encryption)·가명화(pseudonymisation)·익명화(anonymisation) 등의 달성방안, 데이터의 최소화(특히 개인정보)에 관한 사항, 동의를 철회할 수 있는 권리와 반대할 권리 및 잊힐 권리 등을 이행하였는지, 인공지능시스템의 생애주기 동안 수집·생성·처리된 개인정보와 데이터 보호 관련 사항을 고려하였는지 등을 포함

→ 인공지능시스템의 비개인정보 교육 데이터 또는 기타 처리된 비개인정보가 개인정보와 데이터 보호 관련 사항을 고려하였는가?

→ 인공지능시스템을 관련 표준(예: ISO, IEEE) 또는 널리 채택된 프로토콜과 연계하여(일상적인) 데이터 관리 및 거버넌스를 수행하였는가?

자료: EU, 'Assessment List for Trustworthy Artificial Intelligence' (2020)

Case Study | 핵심요건2 '프라이버시 보호' 관련 이슈('21.1)

☞ **국내 스타트업 '스캐터랩'의 인공지능 챗봇 '이루다' 사례는 인공지능시스템 개발사가 데이터 구축·학습 과정에서 프라이버시와 관련된 복합적인 문제 발생 가능성을 드러냈으며, 인공지능시스템 개발사가 이를 소홀히 하면 과징금·과태료 부과에 따른 영업 손실까지도 입을 수 있음을 시사함**

(개인정보 관련 발생 이슈) 동의 없는 개인정보 수집, 데이터의 목적 외 이용, 비식별화 후 점검 소홀 등

- (동의 없는 개인정보 수집) 이루다 챗봇의 학습데이터는 '연애의 과학'이라는 콘텐츠 서비스에서 카카오톡 대화 데이터를 사용하였으나, 카카오톡 대화에 참여한 2인 중 1인에게만 동의를 받고 채팅 내용을 제공할 수 있어, 동의하지 않은 1인은 대화 내용이 학습데이터로 제공되는지를 알 수 없었음. 이 콘텐츠 서비스는 연인의 대화를 학습 데이터로 활용하면서 대화 상대방 모두에게 대화 내용의 활용 동의를 받지 못했다는 점이 윤리적인 문제로 지적됨
- (데이터의 목적 외 이용) 스캐터랩은 수집된 실제 데이터를 코드와 데이터 공유 공간인 Github에 공유하였는데, 이 과정에서 이름(성은 미포함) 22건, 지명정보(구·동 단위) 34건, 성별, 대화 상대방과의 관계 등 민감한 정보가 사용자의 동의 없이 유출됨
- (비식별화 후 점검 소홀) 스캐터랩이 비식별화를 하였음에도 예상하지 못한 변수와 기계적 필터링으로 인해 완전한 비식별화가 되지 못하였기 때문에, 비식별화 이후에도 지속적인 점검이 필요함을 보여줌

☞ **개인정보보호위원회의 행정처분 내용 ('21.4.8)**

- 개인정보보호위원회는 사업자별 (텍스트앳, 연애의 과학, 이루다, Github 관련 개인정보 처리) 위반항목에 따라 시정명령, 과징금, 과태료를 부과하였음

위반 내용	위반 조항
① 개인정보를 수집하면서 정보주체에게 명확하게 인지할 수 있도록 알리고 동의를 받지 않은 행위	§22①
② 법정대리인의 동의 없이 만 14세 미만 아동의 개인정보를 수집한 행위	§22⑥
③ 성생활 등에 관한 정보를 처리하면서 별도의 동의를 받지 않은 행위	§23①
④ 회원탈퇴한 자의 개인정보를 파기하지 않은 행위	§21①
⑤ 1년 이상 서비스 미사용자의 개인정보를 파기하거나 분리·보관하지 않은 행위	§39의6
⑥ 법정대리인의 동의 없이 만 14세 미만 아동의 개인정보를 수집한 행위	§22⑥
⑦ 수집 목적 외로 이루다 학습·운영에 카카오톡 대화문장을 이용한 행위	§18①
⑧ Github에 이용자의 카카오톡 대화문장을 공유한 행위	§28의2②

주: ①~⑤('텍스트앳'과 '연애의 과학' 내 개인정보 처리), ⑥~⑦('이루다' 관련 개인정보 처리), ⑧('Github' 관련 개인정보 처리)

핵심 요건 3

다양성 존중

- 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영해야 하며, 성별·연령·장애·지역·인종·종교·국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다.
- 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

실천 점검문항

E03. 01	인공지능시스템 활용에 사회적 약자의 접근 가능성을 고려하고 있는가?
E03. 02	인공지능시스템 개발에 활용되는 데이터의 성별, 인종, 민족, 국가 등 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고 이를 최소화하기 위해 노력하고 있는가?
E03. 03	인공지능시스템의 개발·운영 단계에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?
E03. 04	인공지능시스템을 사용할 때 편향이나 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 운영자, 사용자 모두 내부 또는 인공지능시스템 개발조직과 운영업체에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?
E03. 05	인공지능시스템 개발자를 대상으로 인공지능시스템에서 발생할 수 있는 편향성의 인지 또는 분석 능력 향상을 위한 교육훈련의 기회를 제공하고 있는가?

제안 이유

인공지능시스템은 기존 사회에 내재하는 고정관념과 편견을 답습하고 재생산될 수 있어 편향성, 적합성, 공정성 등에서 문제를 초래하기도 한다. 이는 인공지능시스템 개발·운영에 필요한 데이터에 인간이 지닌 편견이 반영될 수 있다는 점에 그 원인이 있다. 인공지능시스템의 활용에 따른 차별을 최소화하기 위한 노력으로 데이터에 반영된 편향성을 줄이거나 기술이나 서비스에 대한 사회적 약자의 접근 가능성을 보장하는 것뿐만 아니라, 개발과 운영 단계에서부터 다양한 의견을 청취·검토·평가·반영하는 것 역시 중요하게 다루고 있다. 본 영역에서는 다양한 사회 구성원을 배려하고 편향과 차별을 최소화하기 위해 필요한 평가와 절차를 마련하기 위한 노력을 평가한다.

참고1 | 다양성 존중을 보여주는 해외 자율점검도구 내 유사 문항

(디지털 두바이(Digital Dubai) 인공지능시스템 윤리 자가진단도구 문항)

☞ **디지털 두바이의 인공지능 윤리 자가진단도구는 공정성의 맥락에서 다양성 존중의 가치를 설명하고 있는데, 인공지능시스템 사용자가 공정한 방식으로 접근하고 사용할 수 있는지를 평가하고, 아래 항목의 점검을 제안함**

"개발·배포 과정의 다양성을 고려해야 하고, 다양한 인구경제학적 배경을 가진 사람들을 포함하도록 노력해야 한다."

"중요한 결정에 영향을 주는 인공지능시스템은 적절한 배경을 가진 다양성이 있는 팀이 개발해야 한다."

자료: Digital Dubai, 'AI ethics principles & guidelines' (2018) (https://www.digitaldubai.ae/initiatives/ai-principles-ethics)

참고2 | 구글(Google) '구글의 인공지능 원칙'

☞ **구글 내의 인공지능 애플리케이션을 개발하는 이들이 윤리적으로 고려해야 할 사항을 안내하고 있음**

☞ **불공정한 편견을 만들거나 조장하지 않는 것을 핵심원칙으로 제시하고 있으며 인종, 성별, 국적, 수입, 성적 지향, 능력, 정치적 혹은 종교적 신념에 따른 불공정한 대우를 금지함**

자료: Google, 'Artificial Intelligence at Google: Our Principles' (2018) (https://ai.google/principles)

참고3 | 영국 데이터윤리혁신센터(CDEI) '알고리즘 의사결정 과정에서 나타날 수 있는 편향에 대한 보고서'

☞ **영국 데이터윤리혁신센터는 2020년 11월, '알고리즘 의사결정 과정에서 나타날 수 있는 편향에 관한 보고서'를 발간**

- ① Data의 측면에서, 조직과 규제기관은 위 편향을 적절히 식별하고, 또 이를 완화하기 위해 필요한 데이터에 접근할 수 있는가?, ② Tools and techniques의 관점에서, 편향을 식별하고 완화하기 위하여 현재 또는 미래에 어떠한 통계적·기술적 해결책이 필요하고, 모범 사례는 어떠한가?, ③ Governance의 측면에서, 이러한 알고리즘에 의한 의사결정시스템의 관리·감사·보장의 책임은 누구에게 있는가? 등에 대한 의문을 해소하기 위한 연구를 수행

☞ **본 보고서는 알고리즘에 의한 의사결정 과정에서 불공정한 편견 또는 편향이 영향을 미칠 수 있다는 문제점을 지적하고, 문제점 해결을 위해 다음과 같은 원칙을 제시**

① 정부는 '개인에게 영향을 미치는 중요한 결정'과 관련된 알고리즘을 사용하는 모든 공공부문 조직에 '투명성 의무'를 부과해야 한다.

② 관련 조직은 알고리즘의 편향을 식별하고 이를 완화하기 위해 데이터를 적극적으로 사용해야 한다. 특히, 알고리즘 도구 자체의 역량 및 한계를 이해하고 있는지를 확인하고, 개인에 대한 공정한 대우를 어떻게 보장해야 할 것인지를 신중하게 고려해야 한다.

③ 정부는 알고리즘에 의한 의사결정에 '평등법(the Equality Act)'이 명확히 적용될 수 있도록 '지침'을 발행해야 하는데, 여기에는 편향 자체를 측정하기 위한 데이터 수집과 관련된 지침, 그리고 알고리즘의 편향을 완화하는 기법도 합법적으로 허용하는 지침 등을 포함해야 한다.

자료: CDEI, 'Bias in algorithmic decision-making Final Report' (2020)

참고4 | 미국 연방거래위원회(FTC) '인공지능과 알고리즘 이용에 관한 지침'

☞ **미국 연방거래위원회(FTC)는 인공지능과 알고리즘의 규제 방향성을 제시하고, 기업이 이에 따라 적절히 대응할 수 있도록 유도하기 위하여 2020년 4월에 다섯 가지 핵심 제안이 포함된 지침을 발표**

- ① 투명성 확보, ② 인공지능과 알고리즘 활용 의사결정에 대한 설명, ③ 의사결정의 공정성 보장, ④ 데이터모델의 견고성과 실증적 타당성 보장, ⑤ 규정 준수, 윤리, 공정성, 차별금지에 대한 책임성 견지

☞ **세 번째 사항으로 다음과 같이 의사결정의 공정성 보장을 요구함**

- 의사결정의 공정성 보장

① 특정 집단과 계층을 차별하지 말아야 하며, 알고리즘 사용 전후에 엄격한 테스트를 시행해야 함
② 결과의 공정성을 검토해야 함
③ 정보 접근 권한과 수정 기회를 소비자에게 제공해야 함

자료: FTC, 'Using Artificial Intelligence and Algorithms' (2020)

참고5 | 호주 인권위원회(AHRC) '알고리즘에 의한 편향 문제 관련 보고서'

☞ **호주 인권위원회(AHRC)는 소비자보호원 등과 연구하여 알고리즘 편향과 관련된 보고서를 발간**

- 편향 문제의 대안으로 알고리즘의 편향 제거 원칙을 제시하였으며, 단계별 체크리스트를 제시

각 단계의 구분	체크리스트(checklist)
① 데이터의 수집·처리 단계	- 학습에 사용하고자 하는 데이터가 어떻게 수집이 되었는가? - 식별되거나(identifiable), 레이블 편향(label bias)이 있을 수 있는 학습용 데이터인가? - 인종, 인구집단, 성별 등에 국한된 데이터는 아닌가? - 데이터세트(data-set) 자체에 충분한 정보를 담고 있는가? - 데이터 자체가 오류(error)를 최소화할 만큼 충분한가? - 학습의 전처리(pre-processing)에 필요한 데이터가 있는가?: 전처리 과정에서 알고리즘의 편향이나 불법적인 차별을 야기할 수 있는 요소를 제거하는 것이 필요한가? - 과소 대표(underrepresented)된 집단이 있거나 다른 고려사항이 필요한 집단이 있는가? - 데이터가 역사적·사회적인 불평등을 담고 있지는 않은가?
② 인공지능시스템의 설계 단계 (Designing AI systems)	- 적절한 단일 모델을 사용하였는가? - 인공지능시스템으로 예측하려는 것이 무엇인가? - 의도하는 결괏값을 얻어내려면 어떤 유형의 인공지능시스템이 적절한가? - 의사결정 과정에서 이 모델을 이용한 예측이 어떻게 사용될 것인가?
③ 인공지능시스템의 테스트 모니터링 단계	- 테스트·모니터링 과정에서 어떠한 공정성(fairness) 조치를 취했는가? - 다양한 공정성 관련 조치를 통해 드러난 잠재적인 불공정성(potential unfairness)에는 어떠한 것들이 있는가? 혹시 그러한 결과가 비합리적인 차별을 보여주지는 않는가?
④ 불법적인 차별의 위험성 확인	- 인공지능시스템이 산출한 결과가 특정 개인이나 집단에 불리하게 작용하지는 않는가? - 인공지능시스템이 산출한 결과가 특정 개인이나 집단에 부당한 요구를 하지는 않는가?

⑤ 다음 단계에서 할 일	– 개인에 대한 불공정한 결과를 완화하기 위하여 취할 수 있는 실질적인 조치가 있는가? – 이러한 완화전략 중에서 알고리즘의 편향을 보완해주는 역할을 하는 것이 있는가? – 주(州) 또는 연방 차원에서 특정 맥락(specific context)을 더 세밀히 검토하였는가? – 내·외부의 기술 전문가와 법률 전문가로부터 자문을 받았는가?

자료: FTC, 'Using Artificial Intelligence and Algorithms' (2020)

Case Study | 핵심요건3 '다양성 존중' 관련 사례

☞ **인공지능이 특정 인종, 성별, 연령대에 편향된 결과를 제공하여 채용이나 신용평가 등에 활용될 때, 특정 집단이 기회의 불평등을 겪을 우려가 존재**

(관련 연구) 벌목공 채용공고에 노출된 사람의 90% 이상은 남성이고, 70% 이상이 백인임. 청소부 채용공고는 65% 이상이 여성이고, 75% 이상이 흑인. 연구진은 동일한 조건으로 광고를 게시했기 때문에 해당 결과는 플랫폼 알고리즘의 광고 타기팅으로 인한 결과로 추론하여, 페이스북 광고를 통해 게시된 채용공고가 직업 유형에 따라 특정 인종 또는 성별에 더 많이 노출되는 점을 지적(Ali et al., 2019)

(사례1) 2021년 1월 인공지능 챗봇 이루다는 학습한 데이터의 편향성 때문에 장애인, 성소수자, 인종 등 특정 그룹에 대한 혐오 발언과 성별에 대한 편견의 표현 등이 문제가 되어 서비스 중단

(사례2) 네덜란드 정부는 복지혜택 부정수급과 세금 사기를 단속하기 위해 인공지능 기반 위험탐지시스템(SyRi)을 개발·활용하였으나, 2020년 2월 소수·취약집단에 대한 차별적 사용을 우려한 법원 판결로 철회됨

자료: Ali, M., et al., 'Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes', Proceedings of the ACM on Human-Computer Interaction, 3(CSCW) (2019)

핵심 요건 4

침해금지

- 인공지능을 인간에게 직·간접적인 해를 입히는 목적으로 활용해서는 안 된다.
- 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

실천 점검문항

E04. 01	인공지능시스템이 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고 이를 예방하기 위한 조치를 취하였는가?
E04. 02	인공지능시스템의 목적 외 사용으로 인해 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 개연성이 확인된 경우, 사용자에게 고지하는 절차를 마련하였는가?
E04. 03	인공지능시스템의 활용 과정에서 예상하지 못한 피해가 발생할 때, 사용자가 해당 피해를 신고하고 의견을 제시할 수 있는 절차를 마련하였는가?
E04. 04	인공지능시스템의 활용 과정에서 예상하지 못한 중대한 피해가 발생할 때, 피해의 확산을 방지하기 위해 이미 상용화된 시스템의 사용중단 또는 리콜, 정부 소관기관에 보고, 사용자에게 고지 등의 절차를 마련하였는가?

제안 이유

인공지능시스템이 인간에게 미칠 수 있는 물리적·정신적 피해 등 다양한 유형의 피해를 검토하고 이를 최소화할 수 있도록 개발되어야 한다. 본 영역에서는 인공지능시스템이 개발된 본래의 목적과 다르게 활용되거나, 활용 과정에서 의도하지 않았던 직간접적인 위험 발생 가능성을 검토하고, 예상하지 못한 피해가 발생하는 경우에 대한 대응 절차를 마련하기 위한 노력을 점검한다.

참고1 | 국토교통부 '자율주행차 윤리 가이드라인'

☞ **국토교통부는 2018년 우버의 자율주행자동차의 보행자 사망 사고 등에 대응하기 위해 자율주행자동차 행위주체가 참고할 수 있도록 윤리 가이드라인을 마련하였으며, 자율주행자동차 설계자에게 탑승자의 피해를 최소화하고 이용자의 불법 개조나 임의 변경을 방지할 수 있도록 설계할 것을 아래와 같이 권고**

3.1.4. 자율주행자동차 설계자는 사전에 사고를 최대한 예방할 수 있도록 설계하여야 하며, 부득이하게 사고가 발생한 경우 탑승자를 비롯한 관련 당사자의 피해를 최소화하도록 설계하여야 한다.

→ 자율주행자동차가 일단 트롤리 상황 등 딜레마에 처하게 되면 사고를 피할 수 없기 때문에 가능한 이러한 상황이 발생하지 않도록 예방 조치를 취하는 것이 대단히 중요함

→ (참고) 독일의 자율주행자동차 윤리 강령 제5조 "자동화되고 네트워크화된 자율주행자동차 기술은 가능한 한 사고를 피해야 하고 난감한 상황이 절대 발생하지 않도록 디자인되어야 한다…"와 제8조 "…자율주행자동차는 사고를 미연에 방지하도록 설계되어야 하지만…"에서도 예방설계의 중요성을 강조함

3.1.5. 자율주행자동차 설계자는 자율주행자동차의 보유자나 이용자가 자율주행 시스템을 불법으로 개조하거나 임의로 변경하여 안전을 위해하는 행위를 방지할 수 있도록 설계하여야 한다.

자료: 국토교통부, '자율주행차 윤리 가이드라인' (2020.12)

참고2 | 마이크로소프트(MS) '인공지능 챗봇 개발자를 위한 자율점검 가이드라인 문항'

☞ **마이크로소프트사는 대화용 인공지능(conversational AI, 또는 챗봇)이 사람으로부터 신뢰받을 수 있도록, 챗봇 개발자의 자율점검 가이드라인을 개발하여 적용 중**

- 본 가이드라인은 마이크로소프트의 인공지능 개발·활용 원칙인 '책임있는 인공지능(Responsible AI)' 원칙을 보완하기 위한 것으로, 궁극적으로는 기업과 서비스에 대한 사람들의 신뢰성 제고가 목표임
- 마이크로소프트 사내 챗봇 개발자는 본 가이드라인을 반드시 고려해야 함

4. 문화적 규범을 존중하고, 오남용을 방지할 수 있도록 챗봇을 설계하라.

→ 챗봇은 사람과 같은 인격(personas)을 가질 수 있기 때문에 챗봇과 이용자 간에는 안전하고 존중하는 상호작용이 특별히 중요함. 챗봇에 오용과 남용을 관리할 수 있는 안전장치와 계획을 내재하여야 한다는 점도 중요

- 규범 위반 가능성 제한: 모든 챗봇은 특정 가치와 문화적 규범을 따르도록 설계해야 하며, 이러한 가치 또는 문화적 규범과 충돌할 가능성을 줄이기 위해 본연의 목적으로만 활용하도록 용도를 제한해야 함
- 사용자를 위해 행동수칙(code of conduct) 고지: 챗봇이 사용자 행동수칙(예: 혐오 발언, 괴롭힘, 타인 위협 금지)의 적용을 받는지를 고려해야 하며, 사용자에게 적절한 고지를 해야 함
- 챗봇에 머신러닝 기술과 키워드 필터링 메커니즘을 적용하여 사용자로부터 민감하거나 공격적인 개입(input)을 감지하고 적절하게 대응: 대부분 논쟁의 여지가 있는 키워드(특히, 혐오표현)를 피하는 것이 좋음. 개방형 도메인의 경우 텍스트 분류기를 활용해 악용되지 않도록 보호해야 함. 민감한 범주에는 성인 콘텐츠, 극단주의, 마약, 술과 담배, 욕설, 저속, 괴롭힘, 왕따, 혐오표현 등이 포함. 공개용 챗봇 API도 검토하여 조직 외부의 사람들이 혐오표현에 가담하거나 조직에 좋지 않은 영향을 미치는 챗봇을 만드는 데 활용할 수 있는지를 평가해야 함

자료: Microsoft, 'Responsible bots: 10 guidelines for developers of conversational AI' (2018.11)

참고3 | 호주 '인공지능 윤리원칙'

☞ **호주 산업과학에너지자원부(Department of Industry, Science, Energy and Resources)는 민간기업과 정부가 인공지능을 책임 있는 방식으로 설계·개발·활용할 수 있도록 포괄적인 차원에서의 인공지능 윤리 프레임워크를 제시하고, 지침으로서의 원칙 제공**

☞ **호주 정부의 책임 있고 포용적 인공지능 개발이라는 국가전략 아래 인공지능과 관련된 윤리원칙을 개발하고 그 적용절차 등을 제시하였으나, 의무적용이 아니라 자율적용이 원칙임**

이의제기의 원칙(contestability). 인공지능시스템이 사람, 공동체, 사회집단, 환경에 중대한 영향을 미칠 때, 적절한 절차를 거쳐 인공지능시스템의 산출물과 인공지능 사용에 대한 이의제기가 가능해야 함

→ '중대한 영향'은 인공지능의 산출물과 인공지능을 사용하는 과정에서 나타날 수 있는 모든 영향을 의미하며, 인공지능의 적용과정, 인공지능이 적용되는 맥락, 인공지능의 전반적인 영향에 따라 그 정의는 달라질 수 있음

- 이 원칙은 인공지능시스템이 사회의 제 분야에 부정적인 영향을 주었을 때, 그 침해에 대한 시정(또는 보상)이 가능하다는 것을 알림으로써 인공지능에 대한 국민의 신뢰를 확보하는 중요한 요소임
- 특히, 사회 내의 취약계층과 관련하여 이러한 원칙을 더욱 강조할 필요가 있음
- 한편, 인공지능 알고리즘 정보에 대한 접근이나 효과적 감시체계 등이 이의제기의 원칙을 강화하는 대안으로 활용될 수 있음

자료: DISER, 'Australia's AI Ethics Principles' (2019)

참고4 | 인공지능시스템의 오류·오용에 따라 발생할 수 있는 침해유형

☞ **미국 컨설팅 회사 앤더슨이코노믹그룹(AEG)의 특별보고서에 따르면, 인공지능시스템의 오류와 오용에 따라 발생할 수 있는 5가지 침해유형을 제시**

① **데이터 오용: 인공지능에 사용되는 데이터의 오용, 보안규정위반, 잘못된 해석으로 발생하는 침해**

→ 데이터 수집과 활용 과정에서 법적 규정을 위반할 경우 발생

② **관리실패: 인공지능시스템을 활용할 때 준수해야 할 시스템의 목적, 제안사항 등에서 관리자의 실패**

→ 인공지능시스템 관리자가 일으킬 수 있는 다양한 부주의 및 부적절한 관리 사항

③ **인공지능의 결함: 제공되는 인공지능 서비스에서 발생하는 실패**

→ 부적절한 인공지능 훈련환경과 데이터 처리과정에서 발생하고, 인공지능 서비스의 지속적 결함 야기

→ 오류를 일으키는 데이터에 근거하거나 지나치게 편중된 데이터를 활용한 서비스 제공 상황

④ **악성프로그램의 인공지능 침투: 정보유출, 데이터 강탈, 방해 등 기타 범죄목적으로 만든 프로그램이 인공지능시스템에 침투**

→ 인공지능시스템과 머신러닝을 감염시키거나 변경시킬 수 있는 바이러스, 악성코드, 랜섬웨어 등 피해 발생

⑤ **정부에 의한 남용: 인공지능을 활용하여 정책을 집행하는 과정에서 법률과 윤리기준 위반**

→ 머신러닝과 인공지능 기술은 시민에 대한 정부의 감시기능을 강화할 수 있으며 감시국가(surveillance government)의 문제가 발생할 수 있음

자료: Anderson P. J., 'Damages caused by AI errors and omission. Special report of Anderson Economic Group' (2019)

핵심 요건 5

공공성

- 인공지능은 개인적 행복 추구뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- 인공지능은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다.
- 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.

실천 점검문항

E05. 01	인공지능시스템이 특정 개인이나 집단의 이익을 대변하여 공익을 훼손하거나 역기능을 발생시킬 가능성을 고려하고 있는가?
E05. 02	인공지능시스템 사용으로 폭력성, 음란성, 사행성, 중독성이 조장되는 등 부작용이 발생할 개연성이 있는지를 고려하고 있는가?
E05. 03	인공지능시스템이 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 내부적으로 검토하거나 외부 전문가의 의견을 청취하였는가?

제안 이유

인공지능의 발전은 인류 전체의 안전과 복지향상에 기여하는 방향으로 이루어져야 한다는 국내외적인 공감대가 형성되고 있다. 「대한민국헌법」을 비롯하여 「유엔헌장」, 「국제노동기구 기본협약」 등 국내외 규범은 사회적 공공성의 차원에서 인류공영의 확보, 경제적 복지 등을 추구하고 있다. 그간 한국과 국제사회가 마련해 온 규범과 추구하는 가치를 인공지능 분야에서도 존중하고 준수하는 것을 통해 공공성을 확보할 수 있다. 본 영역에서는 인공지능시스템이 특정 개인이나 집단의 이익을 위해서 활용되는 것이 아니라 인공지능시스템이 가져올 혜택이 다양한 집단에 스며들 수 있도록 개발·운영 단계에서 기울일 수 있는 노력을 평가한다.

참고1 | 공공성을 보여주는 해외 자율점검도구 내 유사 문항

(디지털 두바이 인공지능 윤리 자율점검도구 문항 사례)

→ 중요한 의사결정 보조에 사용하는 인공지능시스템을 개발하는 개발자는 업무의 광범위한 사회적 영향평가를 준비할 수 있도록 사회과학, 정책 또는 다른 분야의 배경을 가진 전문가를 개발 과정에 참여시켜야 한다.

자료: Digital Dubai, 'AI ethics principles & guidelines' (2018) (https://www.digitaldubai.ae/initiatives/ai-principles-ethics)

참고2 | 인공지능의 사회적 영향평가 동향

☞ **인공지능의 혜택과 우려를 식별하고 평가하기 위한 프레임워크를 도입하는 경우 존재**

(캐나다 알고리즘 영향 평가) 자동화된 의사결정시스템이 개인이나 공동체의 권리, 건강, 복지, 경제적 혜택, 지속 가능성 등에 미치는 영향을 레벨1~레벨4로 분류

→ 영향의 수준과 지속성을 구분할 수 있는 설문을 수행하여 레벨을 구분하고, 레벨에 따라 제시되는 요구사항에도 차이 존재

(유네스코 인공지능 윤리권고안 윤리영향평가) 회원국에 인공지능시스템이 개인의 권리, 노동권 등에 미치는 영향을 종합적으로 평가하도록 권고

자료: Canada, 'Algorithmic Impact Assessment' (2019)
UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021)

참고3 | 유럽연합(EU) '신뢰할 수 있는 인공지능 윤리기준'

☞ **유럽연합의 인공지능 전문가 그룹은 유럽연합 의회가 설립한 독립전문가그룹으로서 2018년 신뢰할 수 있는 AI 윤리기준 초안 작성을 시작으로 500여 명 이상이 참여하는 공청회, 수십 회에 이르는 워크숍과 콘퍼런스를 개최하여 유럽연합 회원국 전체에 적용할 수 있는 인공지능 윤리기준 가이드라인을 제안**

1.6 사회적·환경적 복지

- 교육, 직업, 여가활동 등 삶의 모든 분야에서 사회적 인공지능시스템에 대한 광범위한 노출은 사람의 사회적 관계나 애착에 영향을 미칠 수 있고 변화시킬 수 있음

→ 그러므로 인공지능시스템은 사람의 모든 측면의 복지에 긍정적인 영향을 줄 수 있도록 주의 깊은 모니터링을 고려해야 함

- 또한 인공지능시스템이 민주주의에 미칠 수 있는 광범위한 영향을 고려하여, 인공지능의 개발·활용 단계에서 그 영향을 평가하여야 함

→ 정치적 의사결정과 투표를 포함한 민주주의 절차에서 활용되는 인공지능시스템에 대한 특별한 관심이 반드시 필요함

자료: EU, 'Ethics guidelines for trustworthy AI'(2019)

참고4 | 서울시 교육청 '인공지능 공공성 확보를 위한 현장 가이드라인'

☞ **공청회라는 절차 자체가 인공지능의 공공성을 확보하기 위한 하나의 방법으로 활용되었으며, 동시에 이 공청회에서 인공지능의 공공성을 공교육 차원에서 새롭게 정의함**

☞ **공청회를 통해 공교육 분야에서 인공지능이 갖는 공공성의 의미를 제시**

(공교육에서 인공지능의 공공성) 공교육에 적용되는 인공지능이 갖추어야 할 공공성이란 학생의 개인정보를 비롯한 민감 정보들이 법령과 윤리의 범위 내에서 다루어지고, 데이터 처리 과정과 결과가 투명하고 설명가능하고 신뢰할 수 있어야 하며, 인공지능 알고리즘에 의한 편향과 차별이 없어야 함을 의미함

(공공성 확보를 위한 영향평가 수행) 인공지능을 도입하는 주체가 인공지능을 이용한 결정이 서울시민과 서울시 학생들에게 영향을 미칠 때 인공지능에 기반한 결정의 영향을 평가하고, 그 평가는 위험성 등급을 기준으로 수행

→ 위험성 등급은 인공지능 등급 평가 매트릭스를 통해 판별하며, 1-4등급으로 구분되는 위험성은 각 등급에 따라 단위학교 차원의 영향 검토 절차를 따르도록 함

→ 최고위험 등급인 1등급으로 판별될 경우(등급 평가 매트릭스 활용), 학교 인공지능 위원회 심의 → 교육청 인공지능 위원회 심의 → 교육청 심의 등을 거침

자료: 서울시 교육청, '인공지능 공공성 확보를 위한 현장 가이드라인' (2021)

Case Study | 핵심요건5 '공공성' 관련 인공지능 활용 사례(1)

☞ **소말리아의 NGO가 개발한 인공지능 서비스인 Shaqodoon은 소말리아 국민의 삶에 영향을 주는 국가의 주요 사업에 관해 국민의 의견을 기록하고 이를 정부에 전달**

- 소말리아 국민의 약 65%가 문맹으로서 자신의 의사를 정부에 전달하는 데 큰 어려움을 겪고 있음
- Shaqodoon은 자신의 의사를 글로 전달할 수 없고, 정부의 발표를 읽을 수 없는 사람을 위해 소말리아 국민의 의견을 녹음, 기록, 분석하고 이 내용을 정부에 전달
- 또한 기술적 개선을 위해 Shaqodoon의 운영에 다양한 머신러닝, 인공지능 전문가들이 협업하고 있으며, MIT사회문제 해결플랫폼(MIT Solve Challenge)의 지원대상이 되어 MIT 기술전문가가 Shaqodoon의 기술적 발전을 위해 지속적으로 협력하고 있음

자료: Tomašev,N.etal.,'AIforsocialgood:unlockingtheopportunityforpositiveimpact'.NatureCommunications (2020)

Case Study | 핵심요건5 '공공성' 관련 인공지능 활용 사례(2)

☞ **아프리카 지역사회의 인공지능 역량을 향상하기 위한 시민단체 Deep Learning Indaba는 지역사회의 인공지능 관련 조직과 전문가 등이 협력하도록 촉진하는 역할을 하고, 이러한 협력에 기반하여 지역사회에 긍정적 영향을 주는 혁신을 추진함**

- Deep Learning Indaba는 아프리카 지역의 인공지능 개발자와 지역사회를 연결하고, 이를 통해 자율성을 강화하고자 함
- Deep Learning Indaba는 아프리카의 개발자와 연구자가 활용할 수 있는 데이터를 제공하고, 언어의 제약 없이 자유롭게 협력할 수 있는 새로운 번역장치를 제공했으며, 아프리카 연구자들을 위한 새로운 연구그룹을 지원하고 있음
- 가장 큰 특징은 인공지능 분야에서 다양한 기술 커뮤니티와 배경을 지닌 사람들을 연결하고 이들의 협력을 유도한다는 것임
- Deep Learning Indaba는 자체적으로 Data Science Africa, Black-in-AI, Women in Machine Learning과 같은 NGO와 파트너십을 구축하고 있으며, 또한 아프리카의 여러 연구그룹이 DeepMind, IBM과 같은 세계적 기업과 협력할 수 있도록 지원

자료: Tomašev, N. et al., 'AI for social good: unlocking the opportunity for positive impact'. Nature Communications (2020)

핵심 요건 6

연대성

- 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다.
- 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다.
- 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

실천 점검문항

E06. 01	인공지능시스템 개발·운영 목적의 범위 내에서 다양한 배경의 개발자나 사용자가 의사소통이나 상호작용할 수 있는 기회를 제공하고 있는가?
E06. 02	인공지능시스템의 사용이 지역·성별·세대·계층 간 갈등을 유발하는 등 사회통합을 저해할 개연성이 있는지를 고려하고 있는가?
E06. 03	탄소중립을 위한 국제사회의 노력에 협력하기 위해 인공지능시스템의 개발·운영 과정에서 탄소배출이 적은 방법을 사용하도록 고려하고 있는가?

제안 이유

인공지능의 발전은 서로 다른 환경과 상황에 따라 다양한 이해관계를 가진 사회구성원이 서로를 배려하며 사회통합을 이루는 방향으로 진행되어야 한다. 인공지능시스템의 개발과 운영에 직접 참여하는 이해관계자나 현재 세대의 이익만을 고려할 것이 아니라 인공지능시스템의 영향을 받는 다양한 집단과 함께 미래세대의 안전·복지 역시 고려해야 한다. 이를 위해서는 인공지능시스템이 다양한 배경의 사회구성원들이 소통할 수 있는 기회를 보장하여 신뢰를 형성해나가는 것이 중요하다. 본 영역에서는 인공지능시스템의 개발·운영 과정에서 가능한 한 국내외 다양한 이해관계자에게 참여 기회를 부여하기 위한 노력과 더불어 윤리적 인공지능의 개발과 운영을 위한 국제사회의 협력에 동참할 수 있는 방법을 찾기 위한 노력을 점검한다.

참고1 | AI Now 연구소의 환경 관련 정책 제언

☞ **AI Now 연구소는 인공지능 기술의 환경적 영향을 고려하여 환경에 부정적 영향을 미치는 인공지능 기술의 활용 중단과 기후변화 및 기술규제 정책의 통합 등 7가지 정책제언을 제시함. 제언에는 자원효율성(efficiency)을 높인 결과로 오히려 해당 자원의 소비·생산량이 증가하는 신기술과 서비스의 반동효과에 대한 주의, IT 기업의 전(全) 생태계에 대한 환경적 영향 고려, 화석연료 생산 효율화를 위한 인공지능 기술의 활용 중단 등을 제언하고 있음**

참고2 | 유네스코(UNESCO) '인공지능시스템에서의 성평등 관련 제언'

☞ **성별에 따른 차별적 권력분배와 갈등이 인공지능의 설계·개발·활용 등에서 나타날 수 있다는 우려, 그리고 인공지능이 영향을 미칠 수 있는 교육·과학·문화·정보·통신 등에서 발생할 수 있는 성평등 이슈 존재**

☞ **이에 대응하기 위해 유네스코는 '인공지능과 성평등 보고서'로 성평등을 인공지능 규범에 통합함**

- 기존의 다양한 사건의 발생으로 알 수 있듯이 인공지능이 활용하는 데이터세트에 내재한 성차별적 편향뿐 아니라, 사회적 선을 위한 인공지능시스템을 지향하며 인공지능 기술 및 개발 분야에서 여성의 대표성 확보를 제안
- 이를 반영하여 성포용적(gender-inclusive) 인공지능 원칙과 가이드라인, 윤리강령 제공

자료: UNESCO, 'Artificial intelligence and gender equality: Key findings of UNESCO's global dialogue'(2020)

참고3 | 연대성을 보여주는 해외 자율점검도구 내 유사 문항

(아실로마 인공지능 원칙)

- 공유이익의 원칙: "인공지능 기술은 가능한 한 많은 사람에게 혜택을 제공하고 자율성을 강화해야 한다."

(마이크로소프트, 책임 있는 인공지능)

- 포용성: "인공지능시스템은 사람들을 참여시키고 모든 사람의 자율성을 강화해야 한다. 포용적 설계 작업은 인공지능 개발자가 의도적이지 않게 사람들을 배제하는 환경이나 서비스를 제공하게 될 잠재적 가능성을 이해하고 이에 대처할 수 있어야 한다. 인공지능시스템은 인공지능을 사용하는 사람들의 기대와 요구, 맥락을 이해할 수 있도록 설계해야 한다."

(국제노조연맹, 인공지능 원칙)

- 인공지능시스템의 공유이익: "인공지능 기술은 가능한 한 많은 사람에게 혜택을 주고 자율성을 강화해야 한다. 인공지능으로 발생한 경제적 번영은 모든 인류에게 도움이 되도록 광범위하고 평등하게 분배해야 한다."

자료: Fjeld, J., et al., 'Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI'. Berkman Klein Center Research Publication (2020)

Case Study | 핵심요건6 '연대성' 관련 이슈

☞ **인공지능 기술은 방대한 양의 데이터를 기반으로 훈련·개발됨에 따라 소요되는 전력과 배출되는 탄소량이 환경에 부정적인 영향을 미칠 것으로 우려됨. 이 영향으로 인공지능 개발에 있어 환경적 가치가 점차 강조되고 있음**

(사례) 미국 Open AI사의 언어모델 중 GPT-3의 훈련과정에서 배출되는 탄소 배출량은 522미터톤으로 추정되며, 이는 1년간 120대의 승용차를 운전하였을 때 발생하는 양에 해당하는 것으로 보고됨

(조치) 유럽연합은 2030년까지 데이터센터와 클라우드 인프라의 탄소중립을 목표로 유럽기업의 데이터센터의 에너지 효율성을 측정하는 체계를 '디지털 경제·사회지수'의 일부로 구축할 계획

자료: Fortune, 'A.I.'s carbon footprint is big, but easy to reduce, Google researchers say' (2021.4.22.)

핵심 요건 7

데이터 관리

- 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다.

실천 점검문항

E07. 01	인공지능시스템의 개발·운영에 활용되는 데이터의 수집과 처리 업무의 감독을 위한 절차를 마련하였는가?
E07. 02	인공지능시스템의 개발에 활용되는 데이터의 출처·처리의 주요 과정을 기록하고 있는가?
E07. 03	인공지능시스템의 개발·운영에 활용되는 데이터의 분석과 관리 업무에 대한 기술적·물리적 통제방안을 마련하였는가?

제안 이유

인공지능시스템은 데이터를 활용하여 개발되고, 운영 과정에서 수집된 데이터는 다시 인공지능시스템의 개발 또는 개선하는 데 사용된다. 따라서 인공지능시스템의 신뢰를 확보하려면 데이터를 목적에 맞게 활용하고 데이터의 품질을 관리하는 것이 중요하다. 본 영역에서는 인공지능시스템의 개발 및 운영 단계에서 필요한 데이터의 수집·처리 등의 업무를 통제함으로써 데이터의 품질과 위험을 관리하고, 데이터가 목적 외 용도로 활용되는 것을 방지하기 위한 내부적인 노력을 점검한다.

참고1 | 인텔(Intel) '인공지능 분야 미국 국가전략을 위한 인텔의 권고사항'('19)

☞ **인텔은 책임감 있는 자유로운 데이터 사용과 공유를 강조하면서 3가지 실행 사항을 제시함**

- 데이터 사용의 투명성과 국가적 데이터 보안 규제 개발을 장려하고, 혁신적이면서 윤리적인 데이터 활용을 허용하는 개인 정보 보호
- 미 연방거래위원회(FTC)의 역량을 향상하기 위한 포괄적인 개인정보 법규의 미국 내 통과
- 인공지능의 진화 및 도입을 가속하기 위한 국제적인 데이터 상호 운용성 법규 개발

참고2 | 한국지능정보사회진흥원 '인공지능 학습용 데이터 품질관리 가이드라인'('21)

☞ **본 가이드라인 V1.0은 인공지능 개발 과정에서 고품질 학습용 데이터를 확보하는 데 필요한 조직, 절차, 품질기준, 품질관리 방법이나 활동 등을 정의하여 점검하고 조치하는 일련의 활동을 지원하며, 다음의 데이터 품질관리 원칙을 제시함**

- 인공지능 학습용 데이터의 품질관리는 데이터의 전 생애주기의 품질을 보장해야 한다.
- 인공지능 학습용 데이터 품질관리는 상시적이고 지속적으로 품질을 개선할 수 있어야 한다.
- 인공지능 학습용 데이터 품질을 관리하려면 이를 위한 조직을 구성하고, 정해진 역할과 책임에 따라 수행해야 한다.
- 인공지능 학습용 데이터 품질관리를 위해서는 조직이 품질관리 역량을 확보할 수 있도록 품질관리 교육 등 지원체계를 확보해야 한다.

참고3 | 데이터 표준(Data Standard)의 필요성 강조

☞ **미국 국립표준기술연구소(NIST)는 트럼프 행정부의 행정명령에 따라 「인공지능 분야에서 미국의 리더십」('18.2.11)에서 인공지능시스템의 활용에서의 데이터 표준을 강조**

- 데이터에 관한 설명은 승인된 사용의 정의를 포함해야 함
- 데이터세트의 질, 유용성, 접근성과 관련된 정보를 공유하고 측정하는 것은 데이터 표준에 필수적임
- 빅데이터 분석, 데이터 교환, 데이터의 질, 데이터 접근성, 데이터 프라이버시의 항목에서 표준을 개발하고 있음

자료: NIST, 'U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools'(2019.8)

참고4 | 일본 '인공지능 활용 가이드라인'

☞ **일본 총무성과 인공지능네트워크사회추진회의는 인공지능 안전한 활용을 위하여 개발과 활용 과정에서 준수할 원칙을 가이드라인 형태로 제시**

☞ **가이드라인에는 총 10가지의 원칙이 있으며, 그중 이용자와 데이터 제공자는 인공지능시스템 학습 등에 활용되는 데이터의 품질에 유의해야 한다는 '적정학습의 원칙'이 있음**

① 인공지능 학습 등에 활용되는 데이터의 품질에 대한 유의

- 인공지능 서비스 제공자, 비즈니스 사용자, 데이터 제공자 등은 인공지능의 특성 및 용도를 근거로 인공지능의 학습 등에 활용되는 데이터의 질(정확성이나 완전성 등)에 유의하여야 함
- 인공지능에 의한 판단은 사후적으로 정확도가 손상되거나 낮아질 수 있기 때문에 예상되는 권리침해의 규모, 권리침해가 발생하는 빈도, 기술수준, 정밀도를 유지하기 위한 비용 등을 고려하여 미리 정확도에 관한 기준을 마련할 필요가 있음

② 부정확하거나 부적절한 데이터 학습 등에 따른 인공지능 보안 취약성에 대한 유의

- 인공지능 서비스 제공자, 비즈니스 사용자, 데이터 제공자 등은 인공지능이 부정확하거나 부적절한 데이터를 학습함으로써 인공지능의 보안에 취약성이 발생할 위험이 있음에 유의해야 함
- 소비자인 이용자에게 그러한 리스크의 존재를 미리 주지하는 것이 필요함

자료: AIネットワーク社会推進会議, 'AI利活用ガイドライン-AI利活用のためのプラクティカルリファレンス-'(2019)

Case Study | 핵심요건7 '데이터 관리' 관련 이슈

☞ **인공지능에 활용되는 데이터가 편향되면 오차율 등 인공지능 결과에도 영향을 미친다는 다양한 연구와 우려가 있어 인공지능의 역기능을 방지하기 위해 양질의 데이터 수집과 품질 관리의 중요성이 강조됨**

☞ **특히, 성별, 인종, 나이, 출생국가 등 속한 집단에 따라 안면인식 기술의 정확도가 다르게 적용된다는 학계 연구와 전문가 의견이 존재**

(관련 연구) 미국 매사추세츠공대(MIT) 미디어랩 연구팀은 2018년 미국 뉴욕에서 열린 기계학습연구학회에서 인공지능에 활용되는 데이터가 백인과 남성 중심으로 구성되어 있어 피부색이 어두울수록 안면인식 인공지능의 오차율이 높다고 밝힘

* 연구팀은 미국의 마이크로소프트와 IBM, 중국 메그비의 안면인식 시스템을 이용해 사진 1,270장을 분석한 결과, 백인 남성은 3사 모두 오차율이 1% 미만이었으나, 백인 여성은 7%, 흑인 남성은 12%, 흑인 여성은 최대 35%의 오차율을 보임

핵심 요건 8

책임성

- 인공지능 개발 및 활용과정에서 책임 주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다.

실천 점검문항

E08. 01 인공지능시스템을 개발·운영하는 과정에서 윤리기준 준수를 보장하기 위해 담당자 지정 등 적절한 방안을 마련하였는가?

E08. 02 인공지능시스템 개발자 또는 개발부서는 다음의 역량을 향상하기 위해 노력하고 있는가?

- 인공지능시스템이 활용되는 분야의 적합성을 판단하고 위험성을 인지하는 능력
- 인공지능시스템의 산출물과 결정을 해석할 수 있는 능력

E08. 03 인공지능시스템의 개발·운영에서 발생하는 손해 또는 손실의 책임 소재를 명확히 하고 있는가?

E08. 04 인공지능시스템의 사용으로 발생한 피해에 대하여 합리적인 배상과 보상을 제공하기 위해 준비 방안(배상책임보험의 가입이나 유보금을 적립하는 방안 등)을 마련하였는가?

제안 이유

인공지능 기술의 발전은 혜택과 동시에 예상되는 부정적 영향으로 물리적·재산적 피해가 발생할 수 있다. 발생한 사고의 피해를 최소화하기 위해서는 인공지능시스템 개발·운영 주체의 책임성 있는 행동이 요구된다. 또한 피해 구제의 실효성을 확보하려면 책임 주체를 판단하는 것이 중요하다. 본 영역에서는 피해 발생 시 사용자나 이해관계자에게 적절한 배상이나 보상을 하기 위해 책임 소재를 분명하게 할 수 있도록 대비책을 세우기 위한 노력을 점검한다.

참고1 | 책임성 확보를 위한 정부·기업·학계의 노력

☞ **인공지능시스템의 책임성 확보는 공공·민간을 막론하고 중요시되는 인공지능 윤리원칙 중의 하나로 국내외 인공지능 윤리 가이드라인에 핵심적으로 들어가는 요건임**

- **(정부)** 호주 산업과학에너지자원부 인공지능 핵심원칙('19.11): 인공지능 알고리즘의 생성과 구현에 책임이 있는 사람과 조직은 의도하지 않는 영향이 있더라도 그 알고리즘의 영향을 식별할 수 있어야 하며, 책임져야 한다는 엄격한 기준을 제시
- **(기업)** 구글: 인공지능시스템은 사람에 대한 책임성을 가져야 한다는 원칙 아래 인공지능이 사람의 적절한 지시와 통제의 대상이 되고 적절히 설명되어야 함
- **(학계)** 일본 인공지능학회: 학회 회원은 개발한 인공지능 기술의 성능과 영향분석을 통해 위험을 파악해야 하고, 파악한 위험을 사회에 전달해야 할 의무를 진다는 윤리원칙을 제시

참고2 | 금융위원회 '금융분야 인공지능 가이드라인'

☞ **인공지능 기반 금융서비스를 위한 가이드라인의 핵심가치로 금융산업의 '책임성'을 강조하고 다음과 같은 조치의 시행을 권고함**

- 서비스 개발 전 과정에 걸쳐 인공지능의 잠재적 위험을 평가·관리하는 구성원의 역할과 책임, 권한을 구체적으로 정의
- 서비스 자체 평가·관리정책을 마련하고, 개인의 권리에 중대한 위험을 초래할 수 있는 서비스에는 강화된 위험관리 적용

자료: 금융위원회, '금융분야 인공지능 가이드라인' (2021)

참고3 | 미국전기전자학회(IEEE)의 윤리 가이드라인(Ethically Aligned Design)('19)

☞ **미국전기전자학회(IEEE)의 가이드라인은 인공지능에 대한 실행 가능한 권고사항을 제시하고, 제조 및 사용에 대한 표준·인증·규정 또는 법률의 구체적인 지침을 제공하고 있음**

- 책임성의 측면에서 특정 인공지능시스템에 대한 법적 책임이 누구에게 있는지를 등록 및 기록 보관 시스템을 통하여 항상 알 수 있도록 하여 책임 소재를 명확히 할 것을 강조

자료: IEEE, 'Ethically Aligned Design' (2019)

참고4 | 기업 내부 책임구조(거버넌스)에 대한 필요성 대두(AI Now 보고서, '18)

☞ **뉴욕대학교 AI Now Institute의 보고서는 대부분 기술 회사의 내부 거버넌스 구조가 인공지능시스템의 책임성을 보장하지 못하기 때문에 인공지능 업계의 거버넌스에 대한 새로운 접근이 필요하다는 입장에서 강경한 수준의 책임성 마련 수단을 제안함**

- 이사회, 외부 윤리자문위원회 또는 독립적 모니터링 및 투명성 노력의 구현을 포함해야 함
- 제3자 전문가가 주요 시스템에 대해 감독하고 공표할 수 있어야 함
- 의미 있는 책임성을 실현하기 위해 인공지능시스템의 구성요소 부분과 전체 공급망 (또는 제품 수명 주기)에 걸쳐 학습데이터, 시험 데이터, 모델, 응용 프로그램 인터페이스 (API) 및 기타 인프라 구성요소의 출처와 사용을 설명하고 추적할 필요가 있음
- 인공지능시스템으로 인한 피해의 위험이 가장 높은 사람들이 종종 결과에 이의를 제기할 수 없는 일반인이기 때문에, 인공지능 책임 문제에 대해 소송과 같은 시민이 참여할 수 있는 메커니즘 지원이 필요함

자료: AI Now, 'AI Now 2017 Report' (2018)

참고5 | 유럽연합(EU) 'ALTAI'의 인공지능 평가문항

☞ **유럽연합 ALTAI는 책임성에 관한 평가를 실시하도록 하고 있으며, 그 내용은 다음과 같음**

→ 책임성은 인공지능시스템의 개발, 배치 또는 사용에 대한 책임을 보장하기 위한 메커니즘을 마련해야 함을 의미한다.

→ 이는 제3자에게 설명하고 감시할 수 있는 투명한 방법으로 위험을 식별하고 완화하는 위험관리와 밀접한 관련이 있으며, 부당하거나 불리한 영향이 발생하면 적절하게 시정할 수 있는 책임 메커니즘을 마련해야 한다.

자료: EU, 'Assessment List for Trustworthy Artificial Intelligence' (2020)

Case Study | 핵심요건8 '책임성' 관련 사례

☞ **알고리즘의 블랙박스 성격으로 오류 발생 여부, 발생 지점 등을 판단하기 어려워 피해가 발생하여도 책임소재를 규명하는 데 어려움이 있음**

(개요) 금융 분야에서 활용되는 로보어드바이저는 인공지능과 알고리즘을 활용하여 고객의 금융 자산을 운용하고 관리해주는 자동화 서비스로 알고리즘의 오류·결함으로 인한 이용자 피해 발생 시 책임소재 규명 및 감시체계의 부재와 관련된 문제가 대두

(조치) 기업이 알고리즘의 블랙박스 효과를 강조하며 책임을 회피할 수 있기 때문에 피해 발생 시 이의제기를 할 수 있는 담당부서나 책임자가 있어야 한다는 점이 강조됨

핵심 요건 9

안전성

- 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

실천 점검문항

E09. 01 인공지능시스템의 비정상 동작이나 예기치 못한 오류에 대한 안전조치 기능과 안전조치 기능의 한계에 대해 이해관계자와 사용자에게 충분한 정보를 제공하고 있는가?

E09. 02 인공지능시스템과 인간의 상호작용에서 발생할 수 있는 위험 (인공지능에 의한 감시, 중독, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?

E09. 03 인공지능시스템의 결과에 대한 안전성을 지속적으로 평가하기 위한 절차 (정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?

제안 이유

인공지능시스템의 안전성이란 사용 전반에 걸쳐 안정적으로 작동해야 한다는 협의의 안전성뿐만 아니라 학습 경험과 차이가 있는 데이터, 환경, 문제 상황에서도 타당한 추론과 판단을 할 수 있는 능력을 갖추어야 한다는 견고성까지를 포함하는 개념이다. 따라서 외부의 의도적 공격이나 중대한 시스템 오류와 같은 비정상적인 상황에서도 안정적으로 작동해야 한다. 본 영역에서는 잠재적인 위험 유형이나 발생 가능성을 인지하고 안전조치를 마련하기 위한 노력을 점검한다.

참고1 | 인공지능시스템의 보안 취약성과 대응방안 마련

☞ **인공지능시스템은 사물인터넷 인프라와 함께 컴퓨터와 통신망을 적극적으로 이용하고 있어 외부의 사이버 공격에 취약하여 쉽게 위험에 노출될 수 있음**

- 자율주행차량은 외부의 사이버 공격으로 원격 제어와 자율주행에 방해를 받을 우려가 있음
- 자율주행차량을 이용해 국가 주요 인물의 납치와 테러 가능
- 인공지능 교통시스템은 해킹으로 오작동이 발생하면 심각한 사고로 이어질 수 있어 보안 기능을 반드시 제공해야 함
- 크라이슬러사는 차량 내 탑재된 시스템에 대한 공격으로 인해 자동차의 운행과 브레이킹을 제어할 수 있다는 취약점을 인식하여 자동차 140만 대 리콜('15)

참고2 | 미국표준기술연구소(NIST) '기술 표준 및 관련 도구 개발에 대한 연방 정부의 계획'

☞ **인공지능 기술과 시스템의 기능, 상호 운용성, 신뢰성 등에 관한 목표를 안전하게 수행할 수 있도록 인공지능의 요구사항, 사양, 지침에 관련된 특성을 명시함**

- 개념 및 용어, 데이터 및 지식, 인간 상호작용, 지표, 네트워킹, 성능 테스트 및 보고 방법론, 안전성, 위험

관리, 신뢰성 항목에 대해 현재 활용 가능한 표준 여부 및 현재 개발 여부를 소개

- 표준 관련 도구 (데이터세트, 테스트베드 및 책임·감사 도구의 표준화된 형식)에 중점을 두어 안내

자료: NIST, 'U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools'(2019.8)

참고3 | 인공지능으로 인한 위협 환경의 특성 변화

☞ **인공지능을 통해 기존의 위협 환경이 고도화되고 새로운 위협이 등장하고 있음**

① (기존 위협의 확대) 인공지능시스템을 활용한 효율적인 공격이 더 쉬워지고 공격 속도도 증가

- 인공지능의 발전으로 특정인이나 단체를 대상으로 한 스피어피싱(spear phishing) 대두

  * 스피어피싱: 공격자가 사전에 공격대상자에 대한 정보를 수집·분석하여 피싱 공격을 수행하는 형태

- 불특정 다수의 개인정보를 훔치는 피싱과 달리 가치가 높은 공격대상자를 선정하는 고도의 공격이 인공지능 기술을 활용하는 다수에 의해 더 쉽게 이행될 수 있음

② (새로운 위협의 출현) 인공지능에는 환경적·시간적 제약이 없기 때문에 인간이 할 수 없는 공격 수행이 가능

- 인공지능이 인간을 공격하는 형태, 인간이 인공지능을 이용해 인공지능시스템을 공격하는 형태 등장

③ (위협의 전형적 특성 변화) 인공지능을 이용한 공격의 빈도가 커지고, 효과도 증대되는 위협의 특성에 따른 변화 발생

- '저빈도-고효과'와 '고빈도-저효과'라는 일반적인 공격의 상충관계 특성이 인공지능을 통해 '고빈도-고효과' 특성으로 변화

- 테러와 암살에서 공격대상자와 대면하지 않아도 공격을 할 수 있는 (드론 이용) 새로운 위협 등장

자료: Brundage, M. et al., 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation' (2018)

Case Study | 핵심요건9 '안전성' 관련 사례

☞ **인공지능 기술이 적용되는 분야가 확대되면서 안전성 관련 이슈가 부상**

(사례1) 경제적 손실: 한맥투자증권에서 사용한 알고리즘 자동매매 시스템은 직원의 이자율 입력 실수(휴먼에러)를 그대로 판단해 거래를 자동으로 진행함으로써, 1분 만에 약 460억 원의 손실이 발생했고, 그 결과 회사는 파산함('13.12). 금융업계에서는 알고리즘 매매를 하더라도 예상 범위를 벗어난 거래에는 경고창이 뜨거나 주문을 실행하지 않게 하는 위험방지 장치를 마련하는 등 안전성을 보강하는 조치 필요

(사례2) 사고: 테슬라를 비롯하여 우버·구글·GM·아우디 등 글로벌 자동차 제조 기업은 오토파일럿 기능을 갖춘 자율주행자동차 개발에 연구와 투자를 하고 있으나, 오토파일럿으로 인한 사고가 지속적으로 발생하고 있음

- 사고 원인이 자율주행차의 과실인지 아니면 상대편 일반 차량이나 운전자, 보행자 등의 과실인지와 관련하여 사건마다 논란이 되고 있지만 안전성을 확보하기 위한 법적·제도적·윤리적 정비가 필요함과 동시에,

- 기술의 견고성뿐만 아니라 기술을 사용하는 이용자의 행동 패턴도 고려하여 안전성을 높이려는 노력도 필요

(사례3) 신체적 상해: 미국 애리조나주에서 시범운행 중이던 우버 자율주행차가 횡단보도 밖에서 무단횡단을 하던 보행자를 치어 사망한 사례('18.3)

핵심 요건 10

투명성

- 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능 기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

실천 점검문항

E10. 01	인공지능시스템을 활용한 제품 또는 서비스가 인공지능 알고리즘 기반의 결정을 한다는 사실과 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 고지하고 있는가?
E10. 02	인공지능시스템을 목적에 맞게 사용하기 위해 이해관계자와 사용자에게 관련 정보 (예: 가이드북, 매뉴얼 등)를 제공하고 있는가?
E10. 03	인공지능시스템이 수집하는 데이터, 의사결정의 결과에 영향을 미치는 주요 요인 등 사용자가 설명요청하는 정보를 제공할 수 있는 절차를 마련하였는가?

제안 이유

인공지능시스템의 투명성과 설명가능성은 인공지능의 신뢰성 확보에 중요한 원칙이다. 다른 기술과는 달리 딥러닝 알고리즘 등 인공지능시스템의 작동 과정과 원리의 추적은 기술적으로 쉽지 않다. 그럼에도 불구하고 인공지능이 인간의 결정을 대신하거나, 그러한 결정에 영향을 미치기 때문에 국내외 주요 인공지능 윤리원칙은 공통적으로 투명성과 설명가능성을 강조한다. 본 영역에서는 사용자의 알권리 보장에 필요한 정보를 제공하기 위해 노력하였는가를 확인한다.

참고1 | 블랙박스 알고리즘

☞ **인공지능 기술의 발전으로 인해 입력된 정보와 출력된 정보 외에 내부적으로 알고리즘이 무엇을 하고 있는지 파악할 수 없는 것을 지칭하는 용어로 인공지능시스템의 투명성 확보에 대한 요구가 증가하고 있음**

- 국내외 여러 기업이 채용과정에서 인공지능 면접을 활용하고 있으나, 해당 인공지능시스템이 공정한 평가를 지원하고 있는지에 대한 의문 제기
- EU의 「신뢰할 수 있는 인공지능 윤리원칙」은 인공지능시스템에 의해 내려진 결정은 인간이 이해하고 추적할 수 있어야 한다는 점을 강조
- 핀란드 고용경제부 인공지능 윤리 가이드라인은 데이터가 어떤 목적으로 수집되고, 알고리즘이 지원하고 결정하는 목표가 무엇인지 공개하는 것을 투명성으로 이해

참고2 | 설명가능한 인공지능

☞ **법에 규정한 이해관계자가 인공지능에 대해서 설명을 요청하면 인공지능의 입력값, 내부 프로세스, 동작의 종류 및 상태 등을 요청자가 이해할 수 있는 방식으로 설명해야 한다는 요구가 증가하고 있으며, 이에 인공지능의 설명가능성을 확보하기 위한 노력이 이루어지고 있음**

- 카카오가 발표한 「알고리즘 윤리헌장」에서는 이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다는 원칙이 포함됨
- 영국 보건사회복지부의 「인공지능 윤리 행동규범」은 개발된 알고리즘의 학습 방법을 시연하고, 어떤 유형의 알고리즘이 개발되고 사용되었는지, 사용된 데이터가 어떤 윤리적 검증을 받았는지, 그 결과가 어떻게 보건 서비스로 통합될 수 있는지 등을 공개해야 한다는 엄격한 기준을 적용

참고3 | 사람과 구별이 어려운 인공지능

☞ **구글의 인공지능 비서 플랫폼 구글 어시스턴트가 자신의 주인을 대신해 미용실을 예약하는 기능을 선보인 이후로 인간과 비슷한 인공지능 서비스에 대한 논란이 증폭되었음**

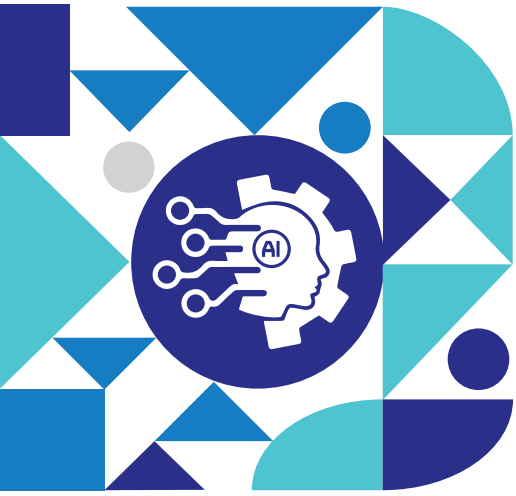
- 인공지능 비서가 사람과 비슷한 추임새라든지 망설이는 듯한 모습까지 그대로 흉내내면서 우려가 증가함
- 이러한 우려가 실리콘밸리에서 윤리의식에 관한 논쟁을 불러일으켜 향후 신뢰성 확보를 위해 인간을 모사하는 인공지능에 대한 가이드라인의 필요성이 강조됨

Case Study | 핵심요건10 '투명성' 관련 이슈

☞ **미국의 프라이버시 관련 비영리 조직인 EPIC(Electronic Privacy Information Center)는 인공지능 기반 채용 솔루션을 제공하는 HireVue사의 제품이 투명성, 공정성, 그리고 책무성과 관련된 국내외 기준을 준수하지 않았다는 내용을 담은 탄원서를 연방거래위원회(FTC)에 제출(Forbes, '20.2)**

- 특히, 인공지능 면접 과정에서 사용되는 안면인식 기술과 검증되지 않은 알고리즘을 문제 삼음
- HireVue사는 외부 감사업체를 고용하여 알고리즘 감사를 시행*한 뒤, 알고리즘은 편향되지 않았다고 발표하였으나, 이후 HireVue사는 해당 소프트웨어의 주요 기능인 인터뷰 녹화를 바탕으로 면접자의 표정을 분석하는 기능을 삭제함

* 알고리즘 감사는 제3의 기관(O'Neil Risk Consulting & Algorithmic Auditing)에서 수행하였으나 감사 내용은 공개되지 않았음

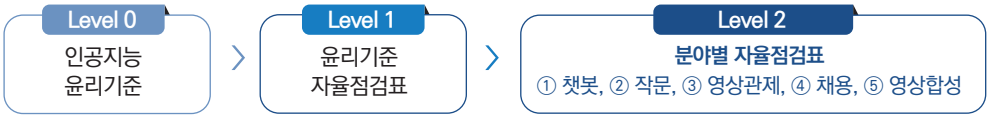


4

분야별 인공지능 윤리기준 자율점검표

04 분야별 인공지능 윤리기준 자율점검표

인공지능 윤리기준 - 공통 자율점검표 - 분야별 자율점검표



1. 인공지능 윤리기준

사람 중심의 사회적 가치를 재조명하고 신기술에 대한 사회적 수용성과 신뢰를 증진하기 위해, 모든 사회 구성원이 인공지능 개발~활용 전 과정에서 함께 지켜야 할 원칙* 제시

* (3대 기본원칙) 인간의 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙
(10대 핵심요건) ① 인권보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성 요건

2. 인공지능 윤리기준 자율점검표: 공통

인공지능 윤리 규범실천을 위한 구체적 방안으로서 인공지능 윤리기준 준수 여부를 스스로 점검할 수 있는 자율점검표 개발

- 인공지능 윤리기준 핵심요건과 높은 연결성 확보*, 활용의 범용성을 만족하는 자율점검표 도출**

* 자율점검표는 인공지능 개발·활용에 관한 윤리적 고려를 포함한 철학, 사회적 담론, 지향해야하는 가치 및 사회 규범 등을 포괄

** 분야, 영역 등 구분 없이, 자율점검표를 활용하고자 하는 모든 주체들이 목적, 특징, 특성에 맞추어 문항을 선별하고 유연하게 가공할 수 있도록 범용성있는 문항 제공

3. 인공지능 윤리기준 자율점검표: 분야별

범용성·포괄성에 중점을 둔 기존 자율점검표를 목적·특징·특성에 맞추어 실제 현장에서 보다 쉽게 응용할 수 있도록 구체적 활용 예시 제공

특정 분야에서 강조될 필요가 있는 문항을 선별·가공하고, 새롭게 제기되는 인공지능 윤리이슈 대응을 위한 문항 신설(2022년 챗봇·작문·영상 관제 분야, 2023년 채용 분야, 2024년 영상 합성 분야)

인공지능 윤리기준 자율점검표 분야별 적용 범위

AI 챗봇	정보제공, 고객상담, 민원처리, 맞춤형 추천, 일상대화 등을 위한 챗봇
작문용 AI	문서작성, 이메일 작성, SNS 포스팅, 카피라이팅 등의 작문보조를 위한 AI
AI 영상 관제 시스템	영상 분석, 영상 모니터링 및 탐지 등을 위한 AI
AI 채용 도구	자기소개서 등 서류평가, 비대면 면접평가를 위한 AI
AI 영상 합성 서비스	AI 영상 합성 기술을 활용하여 다양한 형태의 영상 콘텐츠를 제작하거나 편집할 수 있는 도구를 제공하거나 해당 도구를 활용하여 파생된 콘텐츠를 생산하여 제공하는 서비스

- 인공지능 윤리기준 자율점검표를 참고하는 주체가 각자의 목적과 특성에 맞추어 문항을 선별하고 유연하게 가공하여 활용할 것을 권장

세부 분야 자율점검표 문항 도출 예시

Level 0 윤리기준

▶ **핵심요건 09. 안전성**

- 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

(핵심요건 준수여부 점검을 위한 공통문항 제시)

Level 1 자율점검표 (공통)

- E09.03. 인공지능시스템의 결과에 대한 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?

(공통문항의 "안전성"을 각 영역에 맞게 구체화하여 문항 제시)

Level 2 자율점검표 (분야별)

① 챗봇

챗봇이 선정적, 공격적 또는 편향적 문장을 발화하지 않도록 안전성을 지속적으로 평가하기 위한 절차 (정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?

② 작문

작문용 인공지능 산출물에 대한 정확성, 명료성, 타당성 등 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?

③ 영상 관제

인공지능 영상 관제 시스템의 분석 및 처리결과에 대한 안전성을 지속적으로 평가하기 위한 절차 (정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?

④ 채용

인공지능 채용 도구의 개발·활용 과정에서 해당 도구가 산출하는 지원자 평가 결과의 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가 등)를 마련하였는가?

⑤ 영상 합성

인공지능 영상 합성 서비스의 개발·운영·활용 과정에서 해당 서비스의 신뢰성 및 안전성을 지속적으로 평가·개선하기 위한 절차(예: 정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, AI레드팀 운영 등)를 마련하였는가?

분야 1 챗봇 분야 인공지능 윤리기준 자율점검표

1. 점검 목적

본 자율점검표는 인공지능 챗봇의 개발·운영과정에서 「인공지능 윤리기준」('20.12)의 3대 기본원칙과 10대 핵심요건을 실천하기 위해 고려해야 할 요소와 이를 이행할 수 있는 구체적 방법을 다수의 점검 문항으로 제시합니다.

2. 권장 대상

인공지능 챗봇의 개발과 운영 과정에 참여하는 조직 또는 기관의 최고 의사결정권자, 사업 책임자, 중간관리자 등에게 '챗봇 분야 인공지능 윤리기준 자율점검표'의 활용을 권장합니다. 인공지능 챗봇을 개발·운영하는 과정에서 본 자율점검표의 점검 문항을 각자의 목적과 특성에 맞도록 선별하고 유연하게 가공하여 활용할 수 있습니다. 또한 본 자율점검표를 참조하여 인공지능 윤리기준을 실천할 수 있는 내부 지침을 별도로 마련하거나 내부 규정에 반영할 수도 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 40개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	4	4	6	5	4	2	3	4	3	5

4. 챗봇 분야 인공지능 윤리기준 자율점검표

10대 핵심요건에 해당하는 자율점검 항목을 다음의 표로 제공합니다.

챗봇 분야 인공지능 윤리기준 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발 · 운영 과정에서 국가 「인공지능(AI) 윤리기준」('20)이 제시한 3대 기본원칙과 10대 핵심요건을 실천하는 것입니다.
- 챗봇 분야 인공지능 윤리기준 자율점검표는 기존 '인공지능 윤리기준 실천을 위한 자율점검표' 의 점검문항 중, 특히 챗봇 분야에서 강조되어야 하는 문항을 선별·가공하고, 새롭게 쟁점이 되는 윤리 이슈에 대응하기 위한 문항을 신설하는 방식으로 구성하였습니다.
- 인공지능(AI) 챗봇을 설계 · 제작하고, 데이터와 알고리즘을 통해 AI 챗봇을 구현 · 유지 · 관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부 지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	챗봇의 개발·운영 과정에서 인간의 존엄과 가치를 훼손하지 않도록 노력하고 있는가?	☐	☐	☐
E01. 02	챗봇이 모든 인간을 평등한 존재로 대우함으로써 성별, 연령, 지역, 종교, 인종, 민족, 경제적 수준, 학력, 외모, 성적 지향, 정치적 성향, 장애 여부 등을 근거로 차별하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 03	챗봇이 양심이나 사상에 반하는 행위·결정을 강요하는 등 사용자의 자율적 행위·결정을 방해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 04	챗봇이 사용자에게 불쾌감을 주지 않는 등 사용자를 인격적으로 대우하도록 개발·운영하고 있는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	챗봇의 개발·운영 과정에서 개인정보를 수집·활용하는 경우, 개인정보보호법 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	챗봇 개발·운영 과정에서 프라이버시 침해 가능성이나 개인정보의 유출 가능성 등을 지속적으로 모니터링하고 개선하기 위해 노력하고 있는가?	☐	☐	☐
E02. 03	챗봇 개발·운영 과정에서 위법한 개인정보의 처리 또는 사생활이나 통신의 비밀과 자유에 대한 침해가 확인된 경우, 챗봇의 사용 중지 및 당사자에 대한 통지를 포함한 대응 절차를 마련하였는가?	☐	☐	☐
E02. 04	챗봇이 사용자나 제3자의 사생활을 감시하거나 통제하기 위한 용도로 악용될 개연성을 검토하고, 필요한 경우 적절한 대응책을 마련하였는가?	☐	☐	☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	챗봇의 활용에 있어 장애인 등 사회적 약자의 접근 가능성을 고려하고 있는가?	☐	☐	☐
E03. 02	챗봇의 개발에 활용되는 데이터의 성별, 인종, 민족, 국가 등 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고, 이를 최소화하기 위해 노력하고 있는가?	☐	☐	☐

E03. 03	챗봇이 개인의 고유한 특성을 존중하지 않고, 사용자 또는 제3자에 대한 의도적이고 일방적인 차별과 편견을 조장할 가능성을 검토하였는가?	☐
E03. 04	챗봇이 특정 집단이나 대상에 관한 편견에 기반을 둔 정보나 대화를 제공하지 않도록 노력하고 있는가?	☐
E03. 05	챗봇의 개발·운영 과정에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐
E03. 06	챗봇을 사용할 때 편향, 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 운영자, 사용자 등 누구라도 내부 또는 개발조직과 운영업체에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	챗봇이 허위정보나 잘못된 정보를 제공하여 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고, 이를 예방하기 위한 조치를 취했는가?	☐	☐	☐
E04. 02	챗봇의 개발·운영 과정에서 챗봇이 사용자나 제3자의 명예를 훼손하거나 비방·모욕할 가능성을 검토하고, 이를 방지할 수 있는 조치를 취했는가?	☐	☐	☐
E04. 03	챗봇의 활용 과정에서 허위정보나 잘못된 정보의 제공 등으로 예상하지 못한 피해가 발생할 때, 사용자가 해당 피해를 신고하고 의견을 제시하며, 이에 대응할 수 있는 절차를 마련하였는가?	☐	☐	☐
E04. 04	챗봇이 사용자에 대해 폭언, 혐오표현 또는 성희롱을 하거나 유해 정보를 제공하는 것을 방지하기 위해 노력하고 있는가?	☐	☐	☐
E04. 05	챗봇이 사용자의 중요한 결정(예 : 물품이나 서비스 구매, 채용 지원 등)을 대신하는 경우, 사용자의 의사를 명확하게 확인하는 시스템을 도입하였는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	챗봇이 세계평화, 민주주의, 문화 다양성과 같은 인류 보편적 가치를 훼손하거나 이에 대립하는 정보나 대화를 제공할 개연성이 있는지 검토하였는가?	☐	☐	☐
E05. 02	챗봇이 폭력성, 음란성, 사행성을 조장하는 등 부작용을 발생시킬 개연성이 있는지를 고려하고 있는가?	☐	☐	☐
E05. 03	챗봇이 특정 개인이나 집단의 이익을 대변하여 공익을 훼손할 개연성이 있는지 검토하였는가?	☐	☐	☐
E05. 04	챗봇이 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 내부적으로 검토하거나 외부 전문가의 의견을 청취하였는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	챗봇의 사용으로 지역·성별·세대·계층 간 갈등이 유발되는 등 사회통합을 저해할 개연성이 있는지를 고려하고 있는가?	☐	☐	☐
E06. 02	탄소중립을 위한 국제사회의 노력에 협력하기 위해 챗봇의 개발·운영 과정에서 탄소배출이 적은 방법의 사용을 고려하고 있는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	챗봇의 개발·운영에 활용되는 데이터의 수집·관리·폐기 업무에 대한 감독 절차를 마련하고, 데이터의 수집·관리·폐기의 주요 과정을 기록하고 있는가?	☐	☐	☐
E07. 02	챗봇의 개발·운영에 활용되는 데이터의 분석, 보관, 유지보수 등의 업무에 대한 기술적·물리적 통제방안을 마련하였는가?	☐	☐	☐
E07. 03	챗봇이 수집한 데이터는 챗봇의 운영 목적의 달성에 필요한 시간 동안만 저장하고 있는가?	☐	☐	☐

E08. 책임성

		YES	NO	미해당
E08. 01	챗봇을 개발·운영하는 과정에서 인공지능 윤리 확보를 위해 담당자를 지정하고, 내부 교육 프로그램을 마련하는 등 적절한 방안을 마련하였는가?	☐	☐	☐
E08. 02	챗봇의 개발·운영에서 발생하는 손해 또는 손실의 책임 소재를 명확히 하고 있는가?	☐	☐	☐
E08. 03	챗봇 개발과 운영의 주체가 서로 다른 경우, 챗봇 성능의 지속적 모니터링을 위해 개발업체와 운영업체 간 협력체계가 갖추어져 있는가?	☐	☐	☐
E08. 04	사용자가 챗봇을 윤리적인 방식으로 활용하는 방법을 공개적으로 안내하거나 관련 교육 자료를 제공하고 있는가?	☐	☐	☐

E09. 안전성

		YES	NO	미해당
E09. 01	챗봇에 대한 해킹 등으로 인한 챗봇의 비정상적 동작 또는 예기치 못한 오류에 대응하거나 이를 완화하는 방안을 마련하였는가?	☐	☐	☐
E09. 02	챗봇에 대한 중독이나 과의존 등 인간과 챗봇의 상호작용에서 발생할 수 있는 위험을 사전에 평가하고, 실제 위험 발생 여부를 모니터링하며, 이러한 위험을 완화하기 위해 노력하고 있는가?	☐	☐	☐
E09. 03	챗봇이 선정적, 공격적 또는 편향적 문장을 발화하지 않도록 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?	☐	☐	☐

E10. 투명성

		YES	NO	미해당
E10. 01	챗봇의 사용자에게 학습된 데이터 기반으로 하여 정보나 대화를 제공하는 인공지능과 상호작용하고 있다는 사실을 고지하고 있는가?	☐	☐	☐
E10. 02	향후 챗봇 개발·운영을 위해 챗봇과의 대화 내용이 저장될 수 있다는 사실을 사용자에게 고지하고 있는가?	☐	☐	☐
E10. 03	챗봇을 목적에 맞게 사용하도록 유도하기 위해 사용자와 이해관계자에게 관련 정보(예: 가이드북, 매뉴얼 등)를 제공하고 있는가?	☐	☐	☐
E10. 04	챗봇이 수집하는 데이터나 발화하는 문장에 영향을 미치는 주요 요인 등 사용자가 설명 요청하는 정보를 제공할 수 있는 절차를 마련하였는가?	☐	☐	☐
E10. 05	챗봇 사용자가 의견 제시를 한 경우, 그에 대한 피드백을 충분히 제공하기 위한 절차를 마련하였는가?	☐	☐	☐

분야 2 작문(글쓰기) 분야 인공지능 윤리기준 자율점검표

1. 점검 목적

본 자율점검표는 인공지능 작문 보조 시스템의 개발·운영과정에서 「인공지능 윤리기준」('20.12)의 3대 기본원칙과 10대 핵심요건을 실천하기 위해 고려해야 할 요소와 이를 이행할 수 있는 구체적 방법을 다수의 점검 문항으로 제시합니다.

2. 권장 대상

인공지능 작문 보조 시스템의 개발과 운영 과정에 참여하는 조직 또는 기관의 최고 의사결정권자, 사업책임자, 중간관리자 등에게 '작문(글쓰기) 분야 인공지능 윤리기준 자율점검표'의 활용을 권장합니다. 인공지능 작문 보조 시스템을 개발·운영하는 과정에서 본 자율점검표의 점검 문항을 각자의 목적과 특성에 맞도록 선별하고 유연하게 가공하여 활용할 수 있습니다. 또한 본 자율점검표를 참조하여 인공지능 윤리기준을 실천할 수 있는 내부 지침을 별도로 마련하거나 내부 규정에 반영할 수도 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 45개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	5	5	6	5	4	3	3	7	3	4

4. 작문(글쓰기) 분야 인공지능 윤리기준 자율점검표

10대 핵심요건에 해당하는 자율점검 항목을 다음의 표로 제공합니다.

작문(글쓰기) 분야 인공지능 윤리기준 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발· 운영 과정에서 국가 「인공지능(AI) 윤리기준」('20)이 제시한 3대 기본원칙과 10대 핵심요건을 실천하는 것입니다.
- 작문 분야 인공지능 윤리기준 자율점검표는 기존 '인공지능 윤리기준 실천을 위한 자율점검표' 의 점검문항 중, 특히 작문 분야에서 강조되어야 하는 문항을 선별·가공하고, 새롭게 쟁점이 되는 윤리 이슈에 대응하기 위한 문항을 신설하는 방식으로 구성하였습니다.
- 작문용 인공지능(AI)을 설계 ·제작하고, 데이터와 알고리즘을 통해 작문용 AI를 구현 ·유지 ·관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부 지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	작문용 AI의 개발·운영 과정에서 인간의 존엄과 가치를 훼손하지 않도록 노력하고 있는가?	☐	☐	☐
E01. 02	작문용 AI가 인간을 특정 성별, 연령, 지역, 종교, 인종, 피부색, 민족, 국가, 경제적 수준, 성적 지향, 정치적 성향, 장애 여부 등을 근거로 차별하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 03	작문용 AI가 사용자의 자율적 행동이나 결정을 방해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 04	작문용 AI가 사용자에게 불쾌감을 주지 않는 등 인간을 인격적으로 대우하도록 개발·운영하고 있는가?	☐	☐	☐
E01. 05	작문용 AI에 대한 사용자의 과도한 의존이 글쓰기 능력과 창의력 저하로 이어질 가능성을 검토하였는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	작문용 AI의 개발·운영 과정에서 개인정보를 수집·활용하는 경우, 개인정보보호법 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	작문용 AI의 개발·운영 과정에서 개인 사생활의 비밀과 자유가 침해되지 않도록 조치하였는가?	☐	☐	☐
E02. 03	작문용 AI의 개발·운영 과정에서 상용화된 초거대 AI 활용에 따른 개인정보의 오용 방지를 위한 대책을 마련하였는가?	☐	☐	☐
E02. 04	작문용 AI의 개발·운영 과정에서 프라이버시 침해 가능성이나 의도하지 않은 개인정보 유출 가능성 등을 지속적으로 모니터링하고 개선하기 위해 노력하고 있는가?	☐	☐	☐
E02. 05	작문용 AI의 산출물에서 개인정보 오용 또는 프라이버시 침해 사실이 확인된 경우, 이에 대한 대응 절차를 마련하였는가?	☐	☐	☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	작문용 AI 활용에 사회적 약자의 접근 가능성을 향상시키기 위한 조치(배리어프리 기술 적용 등)를 취하고 있는가?	☐	☐	☐
E03. 02	다음 사항을 위해 노력하고 있는가? (초거대 AI를 활용하였을 경우 b, 아닌 경우 a에 해당) a. 작문용 AI 개발에 활용되는 데이터의 특정 성별, 연령, 지역, 종교, 인종, 피부색, 민족, 국가, 경제적 수준, 성적 지향, 정치적 성향, 장애 여부 등에 따른 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고 이를 최소화 b. 작문용 AI 개발에 활용되는 상용화된 초거대 AI의 특정 성별, 연령, 지역, 종교, 인종, 피부색, 민족, 국가, 경제적 수준, 성적 지향, 정치적 성향, 장애 여부 등에 따른 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고 이를 최소화	☐	☐	☐
E03. 03	작문용 AI의 개발·운영 단계에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 04	작문용 AI를 사용할 때 편향이나 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 운영자, 사용자 모두 내부 또는 개발조직과 운영업체에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 05	작문용 AI 개발자를 대상으로 AI시스템에서 발생할 수 있는 편향성의 인지 또는 분석 능력 향상을 위한 교육훈련의 기회를 제공하고 있는가?	☐	☐	☐
E03. 06	작문용 AI가 특정 집단이나 대상에 대한 과도한 일반화 또는 편향된 기준에 따른 분류 및 유형화 등을 기초로 자료 또는 정보를 분석하거나 산출물을 제공하지 않도록 노력하고 있는가?	☐	☐	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	작문용 AI가 인간의 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고 이를 예방하기 위한 조치를 취하였는가?	☐	☐	☐
E04. 02	작문용 AI의 목적 외 사용으로 인해 인간의 정신 또는 재산에 피해를 발생시킬 개연성이 확인된 경우, 사용자에게 고지하는 절차를 마련하였는가?	☐	☐	☐
E04. 03	작문용 AI의 활용 과정에서 예상하지 못한 피해(표절, 오정보 생성, 허위정보 생성, 과의존 등)가 발생할 때, 사용자가 해당 피해를 신고하고 의견을 제시할 수 있는 절차를 마련하였는가?	☐	☐	☐
E04. 04	작문용 AI의 활용 과정에서 예상하지 못한 중대한 피해가 발생할 때, 피해의 확산을 방지하기 위해 시스템의 사용 중단, 사용자에게 고지 등의 절차를 마련하였는가?	☐	☐	☐
E04. 05	작문용 AI가 제3자의 저작물을 이용하여 산출물을 만들어낼 경우, 그 출처를 밝히고 있는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	작문용 AI의 산출물이 특정 개인이나 집단의 이익만을 대변하여 공익을 훼손할 수 있는 개연성을 줄이기 위해 적절한 조치를 취하였는가?	☐	☐	☐
E05. 02	작문용 AI의 사용으로 폭력성, 음란성, 사행성, 중독성이 조장되는 등 부작용이 발생할 개연성이 있는지를 검토하였는가?	☐	☐	☐
E05. 03	사용자가 작문용 AI를 윤리적인 방식으로 활용하는 방법을 공개적으로 안내하거나 관련 교육 자료를 제공하고 있는가?	☐	☐	☐
E05. 04	작문용 AI의 개발·운영·사용 과정에 있어 교육(초중고, 대학, 대학원 등)이나 사회·문화적 환경에 긍정적인 영향을 미칠 수 있도록 노력하고 있는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	작문용 AI 개발·운영 과정에서 다양한 배경의 개발자, 기획자, 사용자가 의사소통이나 상호작용 할 수 있는 기회를 제공하고 있는가?	☐	☐	☐
E06. 02	작문용 AI를 통하여 다양한 지역, 성별, 세대, 계층의 사용자들이 서로 간의 의견이나 생각을 공유하고 교환하는 상호작용을 함으로써 사회통합을 도모할 수 있도록 조치하였는가?	☐	☐	☐
E06. 03	작문용 AI 개발·운영 과정에서 아동과 청소년의 성장이나 교육에 부정적 영향을 미칠 수 있는 가능성 여부를 검토하였는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	다음의 절차를 마련하였는가? (초거대 AI를 활용하였을 경우 b, 아닌 경우 a에 해당) a. 작문용 AI 개발·운영에 활용되는 데이터의 수집과 처리 업무의 감독을 위한 절차 b. 작문용 AI 개발·운영에 활용되는 상용화된 초거대 AI 데이터 활용·처리 업무 감독을 위한 절차	☐	☐	☐
E07. 02	다음의 주요 과정을 기록하고 있는가? (초거대 AI를 활용하였을 경우 b, 아닌 경우 a에 해당) a. 작문용 AI 개발에 활용되는 데이터의 출처·처리의 주요 과정 b. 작문용 AI 개발에 활용되는 상용화된 초거대 AI 데이터의 출처·처리의 주요 과정	☐	☐	☐
E07. 03	작문용 AI 개발·운영에 활용되는 데이터의 수집, 분석, 보관, 유지보수, 폐기 등의 과정에 대한 기술적·물리적·관리적 통제방안을 마련하였는가?	☐	☐	☐

E08. 책임성

		YES	NO	미해당
E08. 01	작문용 AI를 개발·운영하는 과정에서 윤리기준 준수를 보장하기 위해 담당자 지정 등 적절한 방안을 마련하였는가?	☐	☐	☐
E08. 02	작문용 AI 개발자 또는 개발부서는 작문용 AI가 활용되는 분야의 적합성을 판단하고 잠재적 위험성을 인지하는 역량을 향상하기 위해 노력하고 있는가?	☐	☐	☐
E08. 03	작문용 AI의 개발·운영 과정에서 손해 또는 손실이 발생하는 경우 책임 소재를 명확히 하고, 배상 또는 보상절차를 마련하고 있는가?	☐	☐	☐
E08. 04	작문용 AI 사용자가 손해나 손실을 입은 경우에 배상 또는 보상절차에 참여할 수 있도록 충분한 정보를 제공하고 있는가?	☐	☐	☐
E08. 05	작문용 AI 사용자가 서비스 과정에서 발생 가능한 위험이나 취약성을 발견하는 경우, 이를 당사에 알리거나 이의제기할 수 있는 절차를 마련하였는가?	☐	☐	☐
E08. 06	작문용 AI 사용자가 악의적·비윤리적으로 서비스를 이용할 경우, 사용자에 대한 제재 등 대응책을 마련하고 있는가?	☐	☐	☐
E08. 07	작문용 AI 개발과 운영의 주체가 서로 다른 경우, 작문용 AI 서비스의 지속적 모니터링과 문제해결을 위해 개발업체와 운영업체 간 협력체계가 갖추어져 있는가?	☐	☐	☐

E09. 안전성

		YES	NO	미해당
E09. 01	작문용 AI의 비정상 동작이나 예기치 못한 오류에 대한 안전조치 기능과 그 한계에 대해 개발자, 운영자, 사용자에게 충분한 정보를 제공하고 있는가?	☐	☐	☐
E09. 02	작문용 AI의 사용 중 발생할 수 있는 위험(표절, 오정보 생성, 허위정보 생성, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?	☐	☐	☐
E09. 03	작문용 AI 산출물에 대한 정확성, 명료성, 타당성 등 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?	☐	☐	☐

E10. 투명성

		YES	NO	미해당
E10. 01	작문용 AI가 인공지능 알고리즘 기반의 산출물을 제공한다는 사실과 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 고지하고 있는가?	☐	☐	☐
E10. 02	작문용 AI를 목적에 맞게 사용하기 위해, 사용자와 이해관계자에게 관련 정보(예: 가이드북, 매뉴얼 등)를 용이하게 이해할 수 있는 형태로 제공하고 있는가?	☐	☐	☐
E10. 03	작문용 AI가 활용하는 데이터, 결과 산출물에 영향을 미치는 주요 요인 등 사용자가 설명 요청하는 정보를 프라이버시 보호, 데이터 관리 등 다른 요건과의 균형을 고려하여 제공할 수 있는 절차를 마련하였는가?	☐	☐	☐
E10. 04	작문용 AI 사용자에게 서비스 이용으로 발생가능한 위험 및 오류(표절, 오정보 생성, 허위정보 생성, 과의존 등)에 대해 설명하고 사용에 유의할 것을 고지하였는가?	☐	☐	☐

분 야 3

영상 관제 분야 인공지능 윤리기준 자율점검표

1. 점검 목적

본 자율점검표는 인공지능 영상 관제 시스템의 개발·운영과정에서 「인공지능 윤리기준」('20.12)의 3대 기본원칙과 10대 핵심요건을 실천하기 위해 고려해야 할 요소와 이를 이행할 수 있는 구체적 방법을 다수의 점검 문항으로 제시합니다.

2. 권장 대상

인공지능 영상 관제 시스템의 개발과 운영 과정에 참여하는 조직 또는 기관의 최고 의사결정권자, 사업 책임자, 중간관리자 등에게 '영상 관제 분야 인공지능 윤리기준 자율점검표'의 활용을 권장합니다. 인공지능 영상 관제 시스템을 개발·운영하는 과정에서 본 자율점검표의 점검 문항을 각자의 목적과 특성에 맞도록 선별하고 유연하게 가공하여 활용할 수 있습니다. 또한 본 자율점검표를 참조하여 인공지능 윤리기준을 실천할 수 있는 내부 지침을 별도로 마련하거나 내부 규정에 반영할 수도 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 49개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	5	6	6	5	3	4	4	7	4	5

4. 영상 관제 분야 인공지능 윤리기준 자율점검표

10대 핵심요건에 해당하는 자율점검 항목을 다음의 표로 제공합니다.

영상 관제 분야 인공지능 윤리기준 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발 · 운영 과정에서 국가 「인공지능(AI) 윤리기준」('20)이 제시한 3대 기본원칙과 10대 핵심요건을 실천하는 것입니다.
- 영상 관제 분야 인공지능 윤리기준 자율점검표는 기존 '인공지능 윤리기준 실천을 위한 자율점검표' 의 점검문항 중, 특히 영상 관제 분야에서 강조되어야 하는 문항을 선별·가공하고, 새롭게 쟁점이 되는 윤리 이슈에 대응하기 위한 문항을 신설하는 방식으로 구성하였습니다.
- 인공지능(AI) 영상 관제 시스템을 설계 ·제작하고, 데이터와 알고리즘을 통해 AI 영상 관제 시스템을 구현 · 유지 · 관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부 지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	AI 영상 관제 시스템이 인간의 생명과 안전에 관한 권리를 침해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 02	AI 영상 관제 시스템이 인간을 특정 성별, 연령, 지역, 종교, 인종, 피부색, 민족, 국가, 경제적 수준, 성적 지향, 정치적 성향, 장애 여부 등을 근거로 차별하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 03	AI 영상 관제 시스템이 인간의 자율적 행동이나 결정을 방해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 04	AI 영상 관제 시스템이 인간의 언론·출판의 자유, 집회·결사의 자유를 침해하지 않도록 개발·운영하고 있는가?	☐	☐	☐
E01. 05	AI 영상 관제 시스템이 개인영상정보의 과도한 분석을 통해 특정 개인이나 집단의 이익을 위하여 인간에 대한 과도한 통제와 감시의 수단으로 사용될 개연성이 없는지 검토하고, 그러한 방식으로 시스템이 오남용되지 않도록 조치하고 있는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	AI 영상 관제 시스템의 개발·운영 과정에서 개인정보를 수집·활용하는 경우, 개인정보 보호법 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	개인영상정보에 대한 접근통제 및 접근권한 제한, 안전한 저장 및 전송, 처리기록 위조 및 변조 방지 등 개인정보 보호법 및 개인정보보호위원회의 「영상정보처리기기 설치·운영 가이드라인」에 따른 의무사항을 준수하고 있는가?	☐	☐	☐
E02. 03	AI 영상 관제 시스템의 개발·운영 과정에서 개인 사생활의 비밀과 자유에 대한 침해 가능성을 검토하고, 지속적인 모니터링 및 개선의 노력을 하고 있는가?	☐	☐	☐
E02. 04	AI 영상 관제 시스템 개발·운영 과정에서 개인영상정보의 민감성을 고려하여 영상 정보의 유출 가능성을 지속적으로 모니터링하고 그 위험성을 낮추기 위하여 노력하고 있는가?	☐	☐	☐

E02. 05	AI 영상 관제 시스템의 분석 및 처리 결과물에서 개인영상정보를 포함한 개인정보가 불필요하게 과도한 범위에서 사용되었거나, 정보주체의 사생활 침해가 확인된 경우, 즉시 파기 또는 접근제한, 가능한 범위에서 당사자·관련자에 대한 통지 및 안내 등 법령에 따른 대응 절차를 마련하였는가?	☐ ☐ ☐
E02. 06	AI 영상 관제 시스템의 분석 및 처리 결과물이 본래의 목적 외에 활용되지 않도록(개인에 관한 자동 프로파일링 등) 사후 관리체계를 마련하였는가?	☐ ☐ ☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	AI 영상 관제 시스템의 활용에 사회적 약자의 접근 가능성을 향상시키기 위한 조치(배리어프리 기술 적용 등)를 취하고 있는가?	☐	☐	☐
E03. 02	AI 영상 관제 시스템의 개발에 활용되는 데이터의 특정 성별, 인종, 민족, 국가 등 편향 가능성을 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하고 이를 최소화하기 위해 노력하고 있는가?	☐	☐	☐
E03. 03	AI 영상 관제 시스템의 개발·운영 단계에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 04	AI 영상 관제 시스템을 사용할 때 편향이나 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 운영자, 사용자 모두 내부 또는 인공지능시스템 개발조직과 운영업체에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 05	AI 영상 관제 시스템 개발자를 대상으로 AI시스템에서 발생할 수 있는 편향성의 인지 또는 분석 능력 향상을 위한 교육훈련의 기회를 제공하고 있는가?	☐	☐	☐
E03. 06	AI 영상 관제 시스템이 특정 집단이나 대상에 대한 과도한 일반화 또는 편향된 기준에 따른 분류 및 유형화 등을 기초로 자료 또는 정보를 분석하거나 산출물을 제시하지 않도록 노력하고 있는가?	☐	☐	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	AI 영상 관제 시스템이 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고 이를 예방하기 위한 조치를 취하였는가?	☐	☐	☐
E04. 02	AI 영상 관제 시스템의 목적 외 사용으로 인해 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 개연성이 확인된 경우, 피해를 최소화하기 위한 수단 또는 절차를 마련하고 있는가?	☐	☐	☐
E04. 03	AI 영상 관제 시스템의 활용 과정에서 예상하지 못한 피해가 발생할 때, 사용자가 해당 피해를 신고하고 의견을 제시할 수 있는 절차를 마련하였는가?	☐	☐	☐
E04. 04	AI 영상 관제 시스템의 활용 과정에서 예상하지 못한 중대한 피해가 발생할 때, 피해의 확산을 방지하기 위해 이미 상용화된 시스템의 사용중단 또는 리콜, 정부 소관기관에 보고, 사용자에게 고지 등의 절차를 마련하였는가?	☐	☐	☐
E04. 05	AI 영상 관제 시스템이 해당 사용자 이외에 다른 제3의 개인 또는 집단에 피해나 손해를 야기하는 방식으로 작동하거나 산출물을 제시한 경우, 이를 신속히 사용자에게 고지하고 그러한 작동을 중단 및 산출물을 파기하거나 접근 제한하는 조치 등을 통하여 제3자의 피해나 손해를 방지하기 위한 노력을 하였는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	AI 영상 관제 시스템의 분석 및 처리 결과물이 특정 개인이나 집단의 이익만을 대변하여 공익을 훼손할 수 있는 개연성을 줄이기 위해 적절한 조치를 취하였는가?	☐	☐	☐
E05. 02	AI 영상 관제 시스템의 분석 및 처리 결과물이 과도한 감시와 통제, 개인 및 집단 간의 갈등과 대립 등의 형태로 공공의 이익에 반하거나 사회적 부작용을 야기할 개연성이 있는지 검토하고 적절한 조치를 취하였는가?	☐	☐	☐
E05. 03	AI 영상 관제 시스템이 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 내부적으로 검토하거나 외부 전문가의 의견을 청취하고, 긍정적 영향을 극대화하는 한편 부정적 영향을 최소화하기 위한 조치를 취하였는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	AI 영상 관제 시스템 개발·운영 과정에서 다양한 배경의 개발자·기획자·사용자가 의사소통이나 상호작용할 수 있는 기회를 제공하고 있는가?	☐	☐	☐
E06. 02	AI 영상 관제 시스템의 사용으로 지역·성별·세대·계층 간 갈등이 유발되는 등 사회통합을 저해할 개연성이 있는지를 검토하였는가?	☐	☐	☐
E06. 03	탄소중립을 위한 국제사회의 노력에 협력하기 위해 AI 영상 관제 시스템의 개발·운영 과정에서 탄소배출이 적은 방법을 사용하도록 고려하고 있는가?	☐	☐	☐
E06. 04	AI 영상 관제 시스템 개발·운영 과정에서 감시사회의 고착화와 구조적 차별 등을 방지하기 위한 국제사회의 흐름을 고려하고 있는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	AI 영상 관제 시스템의 개발·운영에 활용되는 데이터의 수집과 처리, 보관 및 파기 업무의 관리·감독을 위한 절차를 마련하였는가?	☐	☐	☐
E07. 02	AI 영상 관제 시스템의 개발에 활용되는 데이터의 출처·처리의 주요 과정을 기록하고 있는가?	☐	☐	☐
E07. 03	AI 영상 관제 시스템의 개발·운영에 활용되는 데이터의 수집, 분석, 보관, 유지보수, 폐기 등의 과정에 대한 기술적·물리적·관리적 통제방안을 마련하였는가?	☐	☐	☐
E07. 04	제3자의 데이터(공공저작물의 영상정보 등)를 기반으로 AI 영상 관제 시스템을 개발·운영하는 경우, 저작권 등 지식재산권에 대한 권리관계 여부를 검토하였는가?	☐	☐	☐

E08. 책임성

		YES	NO	미해당
E08. 01	AI 영상 관제 시스템을 개발·운영하는 과정에서 윤리기준 준수를 보장하기 위해 담당자 지정 등 적절한 방안을 마련하였는가?	☐	☐	☐
E08. 02	AI 영상 관제 시스템 개발자 또는 개발부서는 다음의 역량을 향상하기 위해 노력하고 있는가? - AI 영상 관제 시스템이 활용되는 분야의 적합성을 판단하고 위험성을 인지하는 능력 - AI 영상 관제 시스템의 분석 및 처리 결과물을 이해하고 해석할 수 있는 능력	☐	☐	☐
E08. 03	AI 영상 관제 시스템의 개발·운영 과정에서 손해 또는 손실이 발생하는 경우 책임 소재를 명확히 하고, 배상 또는 보상절차를 마련하고 있는가?	☐	☐	☐

E08. 04	AI 영상 관제 시스템의 사용으로 발생한 피해에 대하여 합리적인 배상과 보상을 제공하기 위해 준비 방안(배상책임보험의 가입이나 유보금을 적립하는 방안 등)을 마련하였는가?			
E08. 05	AI 영상 관제 시스템 사용자가 발생 가능한 위험(AI에 의한 감시 등)이나 취약성을 발견하는 경우, 이를 당사에 알리거나 이의제기할 수 있는 절차를 마련하였는가?			
E08. 06	AI 영상 관제 시스템 사용자가 악의적·비윤리적으로 서비스를 이용할 경우, 사용자에 대한 제재 등 대응책을 마련하고 있는가?			
E08. 07	AI 영상 관제 시스템 개발과 운영의 주체가 서로 다른 경우, 시스템의 지속적 모니터링과 문제 해결을 위해 개발업체와 운영업체 간 협력체계가 갖추어져 있는가?			

E09. 안전성

		YES	NO	미해당
E09. 01	AI 영상 관제 시스템의 비정상 동작이나 예기치 못한 오류에 대한 안전조치 기능과 안전조치 기능의 한계에 대해 이해관계자와 사용자에게 충분한 정보를 제공하고 있는가?			
E09. 02	AI 영상 관제 시스템과 사용 중 발생할 수 있는 위험(AI에 의한 감시, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?			
E09. 03	AI 영상 관제 시스템의 분석 및 처리결과에 대한 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, 사용자 피드백 반영 등)를 마련하였는가?			
E09. 04	AI 영상 관제 시스템의 산출물이 그 시스템 사용 목적 및 용도에 따라 특정 개인에게 심각한 피해를 야기할 가능성이 높을 경우 그러한 가능성을 사전에 사용자에게 고지하고 사용자가 산출물에 대한 접근권한 및 활용범위를 통제할 수 있도록 조치하고 있는가?			

E10. 투명성

		YES	NO	미해당
E10. 01	AI 영상 관제 시스템의 분석 및 처리 결과물이 인공지능 알고리즘 기반으로 제시된다는 사실과 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 고지하고 있는가?			
E10. 02	AI 영상 관제 시스템을 목적에 맞게 사용하기 위해 이해관계자와 사용자에게 관련 정보(예: 가이드북, 매뉴얼 등)를 용이하게 이해할 수 있는 형식으로 제공하고 있는가?			
E10. 03	AI 영상 관제 시스템이 수집하거나 제공받는 데이터, 의사결정의 결과에 영향을 미치는 주요 요인 등 사용자가 설명요청하거나, 그 데이터와 관련된 자로 AI 영상 관제 시스템의 작동 및 산출물에 영향을 받는 당사자가 요청하는 정보를 프라이버시, 데이터 관리 등 다른 요건과의 균형을 고려하여 제공할 수 있는 절차를 마련하였는가?			
E10. 04	AI 영상 관제 시스템 사용자에게 서비스 이용으로 발생 가능한 위험 및 오류(인공지능에 의한 감시, 과의존 등)에 대해 설명하고 있는가?			
E10. 05	AI 영상 관제 시스템이 다른 서비스 또는 시스템과 데이터를 공유하는 경우 법령의 근거가 있는지를 확인·검토하고, 필요시 수집 및 이용 대상인 당사자에게 충분한 설명을 제공하고 동의받고 있는가?			

분야 4 채용 분야 인공지능 윤리기준 자율점검표

1. 점검 목적

본 자율점검표는 인공지능 채용 도구의 개발·운영·활용 과정에서 「인공지능(AI) 윤리기준」이 제시한 3대 기본원칙을 실천하기 위해 고려해야 할 요소와 이를 이행할 수 있는 구체적 방법을 다수의 점검 문항으로 제시합니다.

2. 권장 대상

인공지능 채용 도구의 개발·운영·활용에 참여하는 조직 또는 기관의 최고 의사결정권자, 사업 책임자, 중간관리자 등에게 '채용 분야 인공지능 윤리기준 자율점검표'의 활용을 권장합니다. 인공지능 채용 도구를 개발·운영·활용하는 과정에서 본 자율점검표의 점검 문항을 각자의 목적과 특성에 맞도록 선별하고 유연하게 가공하여 활용할 수 있습니다. 또한 본 자율점검표를 참조하여 인공지능 윤리기준을 실천할 수 있는 내부 지침을 별도로 마련하거나 내부 규정에 반영할 수도 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 총 38개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	5	3	6	4	3	2	3	5	3	4

4. 채용 분야 인공지능 윤리기준 자율점검표

10대 핵심요건에 해당하는 자율점검 항목을 다음의 표로 제공합니다.

채용 분야 인공지능 윤리기준 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발 · 운영 과정에서 국가 「인공지능(AI) 윤리기준」('20)이 제시한 3대 기본원칙과 10대 핵심요건을 실천하는 것입니다.
- 채용 분야 인공지능 윤리기준 자율점검표는 기존 '인공지능 윤리기준 실천을 위한 자율점검표' 의 점검문항 중, 특히 채용 분야에서 강조되어야 하는 문항을 선별·가공하고, 새롭게 쟁점이 되는 윤리 이슈에 대응하기 위한 문항을 신설하는 방식으로 구성하였습니다.
- 인공지능(AI) 채용도구를 설계 · 제작하고, 데이터와 알고리즘을 통해 인공지능 채용 도구를 구현 · 유지 · 관리하거나 활용하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부 지침을 따름으로써 「인공지능(AI) 윤리기준」의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	AI 채용 도구의 개발·활용 과정에서 해당 도구가 인간의 존엄과 가치를 훼손할 가능성을 고려하고, 이를 방지하기 위한 대책을 마련하고 실행하고자 노력하고 있는가?	☐	☐	☐
E01. 02	AI 채용 도구의 개발·활용 과정에서 해당 도구가 지원자를 합리적인 이유 없이 성별, 연령, 지역, 종교, 인종, 민족, 피부색, 경제적 수준, 학력, 외모, 성적 지향, 정치적 성향, 장애 여부, 혼인 여부, 임신 여부, 병력 등을 근거로 차별하지 않도록 주의를 기울이고 있는가?	☐	☐	☐
E01. 03	AI 채용 도구가 면접전형에서 지원자에게 불쾌감이나 모욕감을 주지 않도록 개발·활용하고 있는가?	☐	☐	☐
E01. 04	AI 채용 도구가 면접전형에서 지원자의 자율적 사고를 방해하거나 특정 답변을 강요하지 않도록 개발·활용하고 있는가?	☐	☐	☐
E01. 05	AI 채용 도구가 지원자의 직무 수행 역량과 직접 연관되지 않은 사항에 관하여 평가하지 않도록 개발·활용하고 있는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	AI 채용 도구의 개발·활용 과정에서 개인정보를 수집·활용하는 경우, 「개인정보 보호법」 등 관련 법령 위반 사항 점검을 위한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	AI 채용 도구의 개발·활용 과정에서 개인정보의 위법한 처리 또는 사생활의 비밀과 자유 침해가 확인된 경우, 해당 도구의 사용 중지 및 당사자에 대한 통지를 포함한 즉각적인 대응 절차를 마련하였는가?	☐	☐	☐
E02. 03	AI 채용 도구가 직무 수행 역량을 평가하는 데 불필요한 지원자 본인의 신체적 조건, 출신지역, 혼인여부, 재산 및 가족의 학력, 직업, 재산 등의 정보를 요구하거나 추출하지 않도록 개발·활용하고 있는가?	☐	☐	☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	AI 채용 도구의 개발·활용 과정에서 해당 도구를 활용한 면접전형에 장애인, 저소득층, 고령자 등 정보취약계층의 접근 가능성을 검토하고, 필요한 경우 하드웨어적·소프트웨어적 지원을 포함한 접근보장 방안을 고려하고 있는가?	☐	☐	☐
E03. 02	AI 채용 도구의 개발에 활용되는 데이터의 성별, 연령, 지역, 종교, 인종, 민족, 피부색, 경제적 수준, 학력, 외모, 성적 지향, 정치적 성향, 장애 여부, 혼인 여부, 임신 여부, 병력 등에 따른 편향 가능성을 최소화하기 위해 정기적으로 내부 전담부서 혹은 외부 전문가나 기관을 통해 객관적으로 판단하는 절차를 마련하였는가?	☐	☐	☐
E03. 03	AI 채용 도구의 개발·활용 단계에서 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 04	AI 채용 도구의 개발에 활용되는 데이터의 생성, 가공, 분석 등에 다양한 사회·경제적 배경을 가진 데이터 전문가(데이터 라벨러)가 참여할 수 있도록 노력하고 있는가?	☐	☐	☐
E03. 05	AI 채용 도구의 활용 과정에서 편향이나 차별, 소외 등이 발견되거나 발생한 경우, 개발자, 활용자, 지원자 모두가 내부 또는 개발조직과 활용조직에 알리고, 이를 내부적으로 검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 06	AI 채용 도구가 지원자의 언어습관, 표현방식, 지역 방언, 시선 처리 등에 따른 특성을 이유로 차별하지 않도록 개발·활용 과정에서 노력하고 있는가?	☐	☐	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	AI 채용 도구의 개발·활용 과정에서 해당 도구가 지원자에게 정신적 스트레스를 주는 등 피해를 발생시킬 우려가 있는지를 사전에 검토하고, 이를 예방하는 조치를 하였는가?	☐	☐	☐
E04. 02	AI 채용 도구의 활용 과정에서 예상하지 못한 피해가 발생할 때, 지원자가 해당 피해를 신고하고 의견을 제시할 수 있는 절차를 마련하였는가?	☐	☐	☐
E04. 03	AI 채용 도구의 활용 과정에서 채용의 공정성 훼손 등 중대한 오류 또는 피해가 발생할 때, 해당 도구의 사용을 중단하거나 지원자에게 고지하는 등의 절차를 마련하였는가?	☐	☐	☐
E04. 04	AI 채용 도구를 활용한 면접전형이 지원자의 귀책사유 없이 중단될 때, 본 활용 조직은 해당 지원자가 불이익을 받지 않도록 하는 절차를 마련하였는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	AI 채용 도구의 개발·활용 과정에서 해당 도구가 특정 개인이나 집단의 이익에 유리하게 활용되는 등 공익을 훼손할 가능성을 검토하였는가?	☐	☐	☐
E05. 02	AI 채용 도구의 개발·활용 과정에서 해당 도구가 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 내부적으로 검토하거나 외부 전문가의 의견을 청취하고, 긍정적 영향을 극대화하는 한편 부정적 영향을 최소화하는 조치를 하였는가?	☐	☐	☐
E05. 03	국가·공공기관 또는 교육기관이 AI 채용 도구를 활용하여 채용을 진행하는 경우, 해당 기관이 추구하는 공공성이 유지될 수 있도록 점검하였는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	AI 채용 도구의 개발·활용 과정에서 다양한 배경의 기획자, 개발자, 활용자, 지원자 등이 의사소통이나 상호작용할 기회를 제공하고 있는가?	☐	☐	☐
E06. 02	AI 채용 도구의 활용이 지원자 간 혹은 지원자가 속한 집단 간 갈등을 유발하는 등 사회통합에 악영향을 미칠 개연성을 고려하고 있는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	AI 채용 도구의 개발을 위해 활용되는 데이터의 수집, 분석, 보관, 사용, 폐기 등 데이터 처리 전반에 걸친 업무에 대한 기술적·물리적·관리적 통제방안을 마련하였는가?	☐	☐	☐
E07. 02	AI 채용 도구의 개발을 위해 활용되는 데이터의 수집·관리·폐기와 관련된 주요 과정을 기록하고 있는가?	☐	☐	☐
E07. 03	AI 채용 도구의 활용 과정에서 수집한 데이터를 안전하게 보관하고 폐기하기 위한 기술적·물리적·관리적 통제방안을 마련하였는가?	☐	☐	☐

E08. 책임성

		YES	NO	미해당
E08. 01	AI 채용 도구의 개발·활용 과정에서 인공지능 윤리를 확보하기 위해 담당자 지정 등 적절한 방안을 마련하였는가?	☐	☐	☐
E08. 02	AI 채용 도구의 개발과 활용의 주체가 서로 다른 경우, 해당 도구의 인공지능 윤리를 확보하기 위해 개발조직과 활용조직 간 긴밀한 협력체계가 갖추어져 있는가?	☐	☐	☐
E08. 03	AI 채용 도구의 개발자는 해당 도구 활용자 또는 활용조직이 해당 도구를 윤리적으로 활용하는 방법을 안내하고 관련 자료를 제공하고 있는가?	☐	☐	☐
E08. 04	AI 채용 도구의 개발·활용 과정에서 손해가 발생하는 경우의 책임 소재를 명확히 하고, 피해자에 대한 배상절차를 마련하고 있는가?	☐	☐	☐
E08. 05	AI 채용 도구를 활용한 채용절차를 진행할 때마다 알고리즘 소스코드, 로그 데이터 등 기술적 파일이나 자료를 기록 또는 보관하는 절차를 마련하였는가?	☐	☐	☐

E09. 안전성

		YES	NO	미해당
E09. 01	AI 채용 도구의 개발·활용 과정에서 해킹 등으로 인한 해당 도구의 비정상적 작동 또는 예기치 못한 오류에 대응하는 기술적·관리적 방안을 마련하였는가?	☐	☐	☐
E09. 02	AI 채용 도구의 개발·활용 과정에서 해당 도구가 산출하는 지원자 평가 결과의 안전성을 지속적으로 평가하기 위한 절차(정기적으로 내부부서 또는 외부기관을 통한 전문가 평가 등)를 마련하였는가?	☐	☐	☐
E09. 03	AI 채용 도구가 채용평가자의 자율적이고 독립적인 의사결정을 방해하거나 특정 결과를 부당하게 유도하는 등 오남용될 가능성을 분석하여 필요한 경우 적절한 대응책을 마련하고 있는가?	☐	☐	☐

E10. 투명성

		YES	NO	미해당
E10. 01	AI 채용 도구 활용조직은 채용공고 시부터 학습된 데이터 기반의 인공지능 시스템 활용 여부, 해당 도구의 주요 정보 등을 지원자에게 구체적으로 안내하고 있는가?	☐	☐	☐
E10. 02	AI 채용 도구 활용조직은 해당 도구를 활용하여 서류전형을 진행하는 경우, 이를 지원자에게 사전에 고지하고 관련 동의를 따로 받고 있는가?	☐	☐	☐
E10. 03	AI 채용 도구 개발 및 활용조직은 지원자가 해당 도구를 부정확한 방법으로 사용하여 피해를 보지 않도록 관련 정보(예: 가이드북, 매뉴얼 등)를 제공하고 있는가?	☐	☐	☐
E10. 04	AI 채용 도구 개발 및 활용조직은 해당 도구가 활용하는 데이터, 결과 산출물에 영향을 미치는 주요 요인 등 지원자가 설명 요청하는 정보를 프라이버시 보호, 데이터 관리 등 다른 요건과의 균형을 고려하여 제공하는 절차를 마련하였는가?	☐	☐	☐

분 야 5

영상 합성 분야 인공지능 윤리기준 자율점검표

1. 점검 목적

본 자율점검표는 인공지능 영상 합성 서비스의 개발·운영·활용 과정에서 「인공지능(AI) 윤리기준」이 제시한 3대 기본원칙을 실천하기 위해 고려해야 할 요소와 이를 이행할 수 있는 구체적 방법을 다수의 점검 문항으로 제시합니다.

2. 권장 대상

인공지능 영상 합성 서비스의 개발·운영·활용에 참여하는 조직 또는 기관의 최고 의사결정권자, 사업 책임자, 중간관리자 등에게 '영상 합성 분야 인공지능 윤리기준 자율점검표'의 활용을 권장합니다. 인공지능 영상 합성 서비스를 개발·운영·활용하는 과정에서 본 자율점검표의 점검 문항을 각자의 목적과 특성에 맞도록 선별하고 유연하게 가공하여 활용할 수 있습니다. 또한 본 자율점검표를 참조하여 인공지능 윤리기준을 실천할 수 있는 내부 지침을 별도로 마련하거나 내부 규정에 반영할 수 있습니다.

3. 구성

본 자율점검표는 인공지능 윤리기준의 10대 핵심요건별로 총 44개의 점검항목을 제시합니다.

윤리 핵심요건별 점검항목 수

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안전성	투명성
문항 수	4	5	5	6	5	4	3	4	4	4

4. 영상 합성 분야 인공지능 윤리 자율점검표

10대 핵심요건에 해당하는 자율점검 항목을 다음의 표로 제공합니다.

영상 합성 분야 인공지능 윤리기준 자율점검표(안)

- 윤리기준 자율점검표의 목적은 인공지능시스템의 개발·운영 과정에서 국가 「인공지능(AI) 윤리기준」('20)이 제시한 3대 기본원칙과 10대 핵심요건을 실천하는 것입니다.
- 영상 합성 분야 인공지능 윤리기준 자율점검표는 기존 '인공지능 윤리기준 실천을 위한 자율점검표' 의 점검 문항 중, 특히 영상 합성 분야에서 강조되어야 하는 문항을 선별·가공하고, 새롭게 쟁점이 되는 윤리 이슈에 대응하기 위한 문항을 신설하는 방식으로 구성하였습니다.
- 인공지능(AI) 영상 합성 서비스를 기획·운영·활용하고, 데이터와 알고리즘을 통해 AI 영상 합성 서비스를 구현·유지·관리하는 구성원이나 집단이 업무를 수행하는 과정에서 자율점검표가 반영된 내부 지침을 따름으로써 「인공지능(AI) 윤리기준」 의 핵심요건을 현장에서 실천할 수 있습니다.

E01. 인권보장

		YES	NO	미해당
E01. 01	AI 영상 합성 서비스의 개발·운영·활용이 인간의 존엄과 가치를 훼손하지 않도록 지속적인 주의를 기울이고 있는가?	☐	☐	☐
E01. 02	AI 영상 합성 서비스의 개발·운영·활용이 인간을 성별, 연령, 지역, 종교, 인종, 민족, 피부색, 경제적 수준, 외모, 성적 지향, 정치적 성향, 장애 여부 등을 이유로 차별하지 않도록 지속적인 주의를 기울이고 있는가?	☐	☐	☐
E01. 03	AI 영상 합성 서비스의 개발·운영·활용이 관련 콘텐츠 소비자의 잠재의식을 조작하거나 그들을 의도적으로 기만하는 방식으로 인간의 자율적 의사결정 및 행동을 방해하지 않도록 대책을 마련하였는가?	☐	☐	☐
E01. 04	AI 영상 합성 서비스가 과도한 콘텐츠 검열이나 필터링 등으로 사용자의 표현의 자유를 침해하지 않도록 개발·운영되고 있는가?	☐	☐	☐

E02. 프라이버시 보호

		YES	NO	미해당
E02. 01	AI 영상 합성 서비스의 개발·운영·활용 과정에서 개인정보를 수집·활용하는 경우, 「개인정보 보호법」 등 관련 법령 준수에 필요한 개인정보보호위원회의 「인공지능(AI) 개인정보보호 자율점검표」에 따른 점검을 수행하였는가?	☐	☐	☐
E02. 02	AI 영상 합성 서비스의 개발·운영·활용 과정에서 위법한 개인정보의 처리 또는 사생활의 비밀과 자유 침해가 발생할 경우를 대비하여, 해당 서비스의 사용 중지, 당사자에 대한 통지, 손해배상 등의 대응 절차를 마련하였는가?	☐	☐	☐
E02. 03	AI 영상 합성 서비스의 개발·운영·활용 과정에서 생존 인물을 AI 가상 인간으로 구현하거나 사후 구현을 염두에 두는 경우, 사전에 해당 인물에게 관련 정보를 충분히 제공하고 자발적이고 명시적인 동의를 받았는가?	☐	☐	☐
E02. 04	AI 영상 합성 서비스의 개발·운영·활용 과정에서 고인(故人)을 AI 가상 인간으로 구현하는 경우, 고인의 생전에 관련 정보를 충분히 제공한 후 자발적이고 명시적인 동의를 받고 유족 등의 의사에 반하지 않는지를 확인하였는가?	☐	☐	☐
E02. 05	AI 영상 합성 서비스의 개발·운영·활용 과정에서 프라이버시 침해 가능성이나 개인정보의 유출 가능성 등을 지속적으로 점검하고 개선하기 위해 노력하고 있는가?	☐	☐	☐

E03. 다양성 존중

		YES	NO	미해당
E03. 01	AI 영상 합성 서비스의 개발·운영·활용 과정에서 사회적 약자의 접근성(예: 장벽없는(배리어프리) 인터페이스, 외국어 및 음성 지원 등)을 고려하고 있는가?	☐	☐	☐
E03. 02	AI 영상 합성 서비스의 개발·운영·활용 과정에서 실존하지 않는 인물을 AI 가상 인간으로 구현하는 경우, 성별, 연령, 지역, 종교, 민족, 인종, 피부색, 외모, 언어, 표현 방식 등 사회적·문화적 다양성을 반영하기 위해 노력(예: 다양성 지표 마련 등)하고 있는가?	☐	☐	☐
E03. 03	AI 영상 합성 서비스의 개발·운영·활용이 특정 집단에 대한 고정관념, 선입견, 편견 등을 강화하거나 강화할 개연성이 발견된 경우, 누구든지 해당 서비스 담당 조직에 알리고 이를 내부적으로 검토·평가·개선할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E03. 04	AI 영상 합성 서비스의 개발·운영·활용이 특정 집단에 대한 고정관념, 선입견, 편견 등을 강화할 수 있음을 인지하고 이에 대응할 수 있도록 개발자, 운영자, 사용자를 대상으로 교육훈련의 기회를 제공하고 있는가?	☐	☐	☐
E03. 05	AI 영상 합성 서비스의 개발·운영·활용 과정에서 개발자, 운영자, 사용자, 관련 콘텐츠 소비자 등으로부터 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐

E04. 침해금지

		YES	NO	미해당
E04. 01	AI 영상 합성 서비스가 범죄행위(예: 사기, 명예훼손, 모욕, 성적 허위 영상물 제작 등)나 비윤리적 행위(예: 불쾌감, 조롱, 사이버폭력 등)에 악용되지 않도록 사전적·사후적 조치(예: 본인 인증 절차 강화, 경고문구, 계정 정지, 고발 등)를 마련하였는가?	☐	☐	☐
E04. 02	AI 영상 합성 서비스의 개발·운영·활용으로 예상하지 못한 피해가 발생했을 때, 피해자가 해당 서비스 담당 조직에 알리고 이를 내부적으로 검토·평가·개선할 수 있는 일련의 절차를 마련하였는가?	☐	☐	☐
E04. 03	AI 영상 합성 서비스의 개발·운영·활용으로 중대하고 심각한 피해가 발생했을 때, 피해 확산 방지 절차(예: 서비스 중단, 담당 조직에 고지, 정부 소관 기관에 보고 등)를 마련하였는가?	☐	☐	☐
E04. 04	AI 영상 합성 서비스의 운영·활용 과정에서 생존 인물 및 고인(故人)을 AI 가상 인간으로 구현하여 실제로 하지 않은 말이나 행동을 하도록 하는 것이 해당 인물을 폄훼하거나 미화할 수 있음을 경고하고, 해당 인물 본인이나 가족, 제3자 등에게 피해를 발생시키는 경우 법적 책임이 따를 수 있음을 고지하고 있는가?	☐	☐	☐
E04. 05	AI 영상 합성 서비스의 활용 과정에서 불법·유해 콘텐츠가 생성되지 않도록 필터링 등 기술적 조치를 마련하고 있는가?	☐	☐	☐
E04. 06	AI 영상 합성 서비스의 활용 과정에서 입력된 데이터가 저작권을 비롯한 지식재산권을 침해하지 않도록 사전에 권리관계를 명확히 검토하도록 안내하고 있는가?	☐	☐	☐

E05. 공공성

		YES	NO	미해당
E05. 01	AI 영상 합성 서비스의 개발·운영·활용이 사회경제적으로 미치는 긍정적·부정적 영향에 대하여 주기적으로 검토하고, 공익 증진 및 사회 문제 해결에 기여할 수 있도록 노력하고 있는가?	☐	☐	☐
E05. 02	AI 영상 합성 서비스의 개발·운영·활용이 인류 보편적 가치(예: 민주주의, 세계평화, 문화 다양성, 지속가능성 등)를 훼손하지 않도록 방지 조치를 마련하였는가?	☐	☐	☐
E05. 03	AI 영상 합성 서비스가 허위조작정보나 가짜뉴스의 생성 및 유포에 악용되지 않도록 방지 조치를 마련하였는가?	☐	☐	☐
E05. 04	AI 영상 합성 서비스의 개발·운영·활용이 특정 개인이나 집단의 이익을 대변하여 공익을 훼손하지 않도록 방지 조치를 마련하였는가?	☐	☐	☐
E05. 05	AI 영상 합성 서비스를 활용한 콘텐츠 제작이 금지되거나 제한되는 경우(예: 공직선거법 제82조의8이 제한·금지하는 사항, 아동·청소년 시청 제한 콘텐츠 제작 등)를 개발자, 운영자, 사용자에게 명확하게 주지시키고 있는가?	☐	☐	☐

E06. 연대성

		YES	NO	미해당
E06. 01	AI 영상 합성 서비스의 개발·운영·활용 과정에서 다양한 배경(예: 성별, 연령, 지역, 직업 등)의 개발자, 운영자, 사용자, 관련 콘텐츠 소비자가 의사소통이나 상호작용 할 수 있는 기회를 제공하고 있는가?	☐	☐	☐
E06. 02	AI 영상 합성 서비스의 개발·운영·활용이 특정 계층이나 집단(예: 성별, 지역, 세대, 정당 등) 간 갈등을 유발하는 등 사회통합을 저해할 개연성이 있는지를 검토하고 있는가?	☐	☐	☐
E06. 03	AI 영상 합성 서비스의 개발·운영·활용 과정에서 탄소 배출을 최소화하기 위한 국제사회의 협력에 동참하고 있는가?	☐	☐	☐
E06. 04	AI 영상 합성 서비스의 개발·운영·활용 과정에서 디지털 콘텐츠의 신뢰 구축(예: 투명한 데이터 처리, 합성 콘텐츠 표시 기술 등)을 위한 국내외 협력에 동참하고 있는가?	☐	☐	☐

E07. 데이터 관리

		YES	NO	미해당
E07. 01	AI 영상 합성 서비스의 개발을 위해 활용되거나 활용 과정에서 입력되는 데이터의 수집·관리·폐기의 주요 과정을 기록하고 있는가?	☐	☐	☐
E07. 02	AI 영상 합성 서비스의 개발·운영·활용 과정에서 데이터의 수집, 분석, 보관, 유지보수, 폐기 등의 업무에 대한 기술적·물리적·관리적 보호조치를 마련하였는가?	☐	☐	☐
E07. 03	AI 영상 합성 서비스의 개발을 위해 제3자의 데이터를 활용하는 경우 저작권을 비롯한 지식재산권을 침해하지 않도록 사전에 권리관계를 명확히 검토하고 있는가?	☐	☐	☐

E08. 책임성

		YES	NO	미해당
E08. 01	AI 영상 합성 서비스를 악용하여 타인에게 피해를 발생시키는 경우 법적 책임이 따를 수 있음을 안내하고 있는가?	☐	☐	☐
E08. 02	AI 영상 합성 서비스의 개발·운영·활용 과정에서 손해가 발생하는 경우를 대비하여 책임 소재를 명확히 하고, 피해구제를 위한 절차를 마련하고 있는가?	☐	☐	☐
E08. 03	AI 영상 합성 서비스의 개발·운영·활용 과정에서 비윤리적 콘텐츠(예: 불법·유해 콘텐츠, 허위조작정보, 가짜뉴스 등)의 확산을 방지하기 위해 플랫폼 사업자 등과의 협력체계를 구축하고 있는가?	☐	☐	☐
E08. 04	AI 영상 합성 서비스의 개발·운영·활용 과정에서 AI 윤리 확보 방안(예: 담당자 지정, 내부 교육 프로그램 도입 등)을 마련하였는가?	☐	☐	☐

E09. 안전성

		YES	NO	미해당
E09. 01	AI 영상 합성 서비스를 활용하여 제작된 콘텐츠가 범죄에 악용되거나 저작권을 침해하는 등의 상황이 발생하는 경우를 대비하여 사후 추적을 위한 최신의 기술적 수단(예: 비가시성 워터마크, 메타데이터 삽입 등)을 도입하고 있는가?	☐	☐	☐
E09. 02	AI 영상 합성 서비스의 비정상 동작이나 예기치 못한 오류에 대한 안전조치 기능과 그 한계에 대해 모든 이해관계자에게 충분한 정보를 제공하고 있는가?	☐	☐	☐
E09. 03	AI 영상 합성 서비스와 사용자 및 관련 콘텐츠 소비자의 상호작용에서 발생할 수 있는 위험(예: 가상 인간에 대한 의인화, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?	☐	☐	☐
E09. 04	AI 영상 합성 서비스의 개발·운영·활용 과정에서 해당 서비스의 신뢰성 및 안전성을 지속적으로 평가·개선하기 위한 절차(예: 정기적으로 내부부서 또는 외부기관을 통한 전문가 평가, AI레드팀 운영 등)를 마련하였는가?	☐	☐	☐

E10. 투명성

		YES	NO	미해당
E10. 01	AI 영상 합성 서비스를 활용하여 제작된 콘텐츠가 해당 콘텐츠의 목적에 지장을 주지 않는 범위 내에서 AI 생성물임을 표시하기 위한 최신의 기술적 수단(예: 가시적 워터마크 삽입, 콘텐츠 내 문구 또는 음성 안내 등)을 도입하고 있는가?	☐	☐	☐
E10. 02	AI 영상 합성 서비스 개발·운영 과정에서 수집·활용되는 데이터의 종류, AI 시스템의 주요 작동 원리·기능 및 한계·잠재적 위험 등의 정보를 프라이버시 보호, 데이터 관리 등 다른 요건과의 균형을 고려하여 제공할 수 있는 절차를 마련하였는가?	☐	☐	☐
E10. 03	AI 영상 합성 서비스 활용 과정에서 입력되는 데이터의 수집 여부와 종류, 처리 및 관리 방법을 사용자에게 명확히 고지하고 있는가?	☐	☐	☐
E10. 04	AI 영상 합성 서비스가 목적에 맞게 활용될 수 있도록 모든 이해관계자에게 관련 정보(예: 가이드북, 매뉴얼 등)를 제공하고 있는가?	☐	☐	☐

참고문헌

[국내 문헌]

개인정보보호위원회, '인공지능 개인정보보호 자율점검표' (2021)

과학기술정보통신부·정보통신정책연구원, '인공지능 윤리기준 실천을 위한 자율점검표' (2022, 2023, 2024)

과학기술정보통신부·한국정보통신기술협회, '신뢰할 수 있는 인공지능 개발 안내서' (2022, 2023, 2024)

관계부처 합동, 「인공지능 국가전략」 (2019)

관계부처 합동, '사람이 중심이 되는 「인공지능(AI) 윤리기준」' (2020)

교육부, '교육분야 인공지능 윤리원칙' (2022)

국가인권위원회, '인공지능 개발과 활용에 관한 인권 가이드라인' (2022)

국가정보원·국가보안기술연구소, '챗GPT 등 생성형 AI 활용 보안 가이드라인' (2023)

국토교통부, 「자율주행차 윤리 가이드라인」 (2020)

금융위원회, 「금융분야 인공지능 가이드라인」 (2021)

네이버, '네이버 인공지능 윤리준칙' (2021)

문화체육관광부·한국저작권위원회, '생성형 AI 저작권 안내서' (2023)

방송통신위원회, '이용자 중심의 지능정보사회를 위한 원칙' (2019)

방송통신위원회, '인공지능 기반 추천 서비스 이용자 보호를 위한 기본원칙' (2021)

산업통상자원부, '로봇 윤리 기본 원칙' (2019)

서울시 교육청, 「인공지능 공공성 확보를 위한 현장 가이드라인」 (2021)

서울디지털재단, '서울시 생성형 AI 윤리 가이드라인' (2023)

소프트웨어정책연구소, '인공지능(Artificial Intelligence) 이슈와 국제 표준화 동향' (2021)

정보문화포럼, '지능정보사회 윤리가이드라인' (2018)

정보통신정책연구원, '윤리적 인공지능을 위한 국가정책 수립' (2020)

정보통신정책연구원, '사람중심의 인공지능 구현을 위한 인공지능 윤리정책 개발' (2021)

정보통신정책연구원, '인공지능 윤리체계 확립을 위한 정책연구' (2022)

정보통신정책연구원, '인공지능 윤리 의식 확산을 위한 정책연구' (2023)

정보통신정책연구원, 'AI 윤리-신뢰성 확보를 위한 실천 방안 및 정책연구' (2024)

카카오, '카카오 알고리즘 윤리헌장' (2018)

KB금융, '인공지능 윤리기준' (2022)

SK텔레콤, '사람 중심의 인공지능 윤리 가치 원칙' (2021)

[해외 문헌]

중국 전국네트워크안전표준화기술위원회(TC260). '生成式人工智能服务安全基本要求' (2024)

AHRC, 'Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias' (2020)

AIネットワーク社会推進会議, 'AI利活用ガイドライン-AI利活用のためのプラクティカルリファレンス-' (2019)

AI Now, 'AI Now 2017 Report' (2018)

AI Now, 'Discriminating Systems – Gender, Race, and Power in AI' (2019)

Ali, M. et al., 'Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes', Proceedings of the ACM on Human-Computer Interaction, 3(CSCW) (2019)

Anderson P. J., 'Damages caused by AI errors and omission. Special report of Anderson Economic Group' (2019)

Canada, 'Algorithmic Impact Assessment' (2019)

CDEI, 'Bias in algorithmic decision-making Final Report' (2020)

Cisco, 'Principles for Responsible AI' (2022)

DeepMind, 'DeepMind Ethics & Society Principles' (2017)

Digital Dubai, 'AI ethics principles & guidelines' (2018)

Digital Dubai, 'AI System Ethics Self-Assessment Tool' (2020)

DISER, 'Australia's AI Ethics Principles' (2019)

EU, 'Ethics Guidelines for Trustworthy AI' (2019)

EU, 'Assessment List for Trustworthy Artificial Intelligence' (2020)

EU, 'REPORT with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies' (2020)

EU, 'Artificial Intelligence Act (Regulation (EU) 2024/1689)' (2024)

Fjeld, J. et al., 'Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI'. Berkman Klein Center Research Publication (2020)

FTC, 'Using Artificial Intelligence and Algorithms' (2020)

Future of Life Institute, 'Asilomar AI Principles' (2017)

Google, 'Artificial Intelligence at Google: Our Principles' (2018)

House of Lords of the UK, 'AI in the UK: Ready, Willing and Able?' (2018)

IBM, 'Principles for Trust and Transparency' (2018)

ICO, 'Guidance on Artificial Intelligence and Data Protection' (2020)

IEEE, 'Ethically Aligned Design' (2019)

Israel Ministry of Innovation, Science & Technology, 'Israel's Policy on Artificial Intelligence Regulation and Ethics' (2023)

ITI, 'AI Policy Principles' (2017)

Meta, 'Facebook's Five Pillars of Responsible AI' (2021)

Microsoft, 'Responsible bots: 10 guidelines for developers of conversational AI' (2018)

Microsoft's AETHER, 'Microsoft AI principles' (2018)

Microsoft's AETHER, 'The Future Computed: AI and its role in Society' (2018)

NIST, 'U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools' (2019)

NIST, 'AI Risk Management Framework 1.0' (2023)

NIST, 'AI Risk Management Framework : Generative Artificial Intelligence Profile' (2024)

NSTC, 'Preparing for the Future of Artificial Intelligence' (2016)

OECD, 'Recommendation of the Council on AI' (2019)

OpenAI, 'OpenAI Charter' (2018)

OSTP, 'Blueprint for an AI Bill of Rights' (2022)

Partnership on AI, 'Tenets' (2016)

Tomašev, N. et al., 'AI for social good: unlocking the opportunity for positive impact'. Nature Communications (2020)

The Alan Turing Institute, 'Understanding artificial intelligence ethics and safety' (2019)

United Arab Emirates Minister of State for Artificial Intelligence, Digital Economy & Remote Work Applications Office, 'AI Ethics: Principles & Guidelines' (2022)

UNESCO, 'Artificial intelligence and gender equality: Key findings of UNESCO's global dialogue' (2020)

UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021)

UNESCO, 'Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence' (2023)

U.S. Bureau of Cyberspace and Digital Policy, 'Risk Management Profile for Artificial

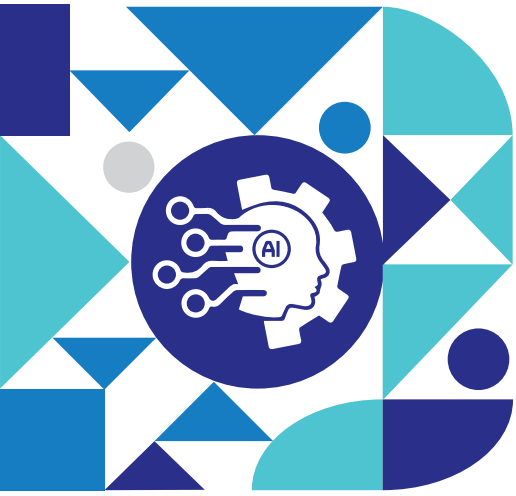
Intelligence and Human Rights' (2024)

University of Montreal, 'The Montreal Declaration for a Responsible Development of AI' (2018)

Vatican, 'Rome Call for AI Ethics' (2020)

WEF, 'Companion to the Model AI Governance Framework' (2020)

WHO, Ethics and governance of artificial intelligence for health: WHO guidance(2021)



부록

인공지능 윤리기준 실천을 위한 자율점검표 (안)

부록 1 사람이 중심이 되는 「인공지능(AI) 윤리기준」

Ⅰ. 서문

오늘날 인공지능 기술은 컴퓨팅 파워의 성장, 데이터의 축적, 5G 등 네트워크 고도화와 같은 ICT 기술의 발전을 토대로 급 성장하고 있다. 인공지능은 제조, 의료, 교통, 환경, 교육 등 산업 전반에서 본격적으로 활용·확산되고 있으며, 우리 생활에서도 쉽게 인공지능 기술을 접할 수 있게 되었다. 이러한 인공지능 기술의 발전·확산은 생산성·편의성을 높여 국가 경쟁력을 높이고 국민의 삶의 질을 높일 것으로 기대되지만, 한편으로는 기술 오용, 데이터 편향성과 같은 인공지능 윤리 이슈도 제기되고 있다. 본 윤리기준은 이러한 시대적 흐름을 고려하여 '인공지능 개발과 활용 전 단계에서 정부·공공기관, 인공지능 기술 개발자, 인공지능 기술을 활용한 제품·서비스 공급자·활용자 등 모든 사회 구성원이 사람중심의 인공지능' 구현을 위해 고려해야 할 기본적이고 포괄적인 기준을 제시하는 것을 목표로 한다.

본 윤리기준은 '사람중심의 인공지능' 구현을 위해 지향되어야 할 최고 가치로 '인간성(Humanity)'을 설정하고 있다. 이는 아래와 같은 사실을 의미한다. 모든 인공지능은 '인간성을 위한 인공지능(AI for Humanity)'을 지향하고, 인간에게 유용할 뿐만 아니라 나아가 인간 고유의 성품을 훼손하지 않고 보존하고 함양하도록 개발되고 활용되어야 한다. 인공지능은 인간의 정신과 신체에 해롭지 않도록 개발되고 활용되어야하며, 개인의 윤택한 삶과 행복에 이바지하며 사회를 긍정적으로 변화하도록 이끄는 방향으로 발전되어야 한다. 또한 인공지능은 사회적 불평등 해소에 기여하고 주어진 목적에 맞게 활용되어야 하며, 목적의 달성 과정 또한 윤리적이어야 하고, 궁극적으로 인간의 삶의 질 및 사회적 안녕과 공익 증진에 기여 하도록 개발되고 활용되어야 한다.

본 윤리기준은 산업·경제 분야의 자율규제 환경을 조성함으로써 인공지능 연구개발과 산업 성장을 제약하지 않고, 정당한 이윤을 추구하는 기업에 부당한 부담을 지우지 않는 것을 목표로 한다. 또한 본 윤리기준은 범용성이 있는 일반 원칙으로서 사안별 또는 분야별 인공지능 윤리기준 제정의 근거를 제공하여 영역별 세부 규범이 유연하게 발전해 나갈 수 있는 기반을 조성하고, 나아가 사회경제 및 기술 변화와 함께 새롭게 제기되는 인공지능 윤리 쟁점을 반영하여 지속적으로 수정되고 보완되는 일종의 '인공지능 윤리 플랫폼'으로 기능할 수 있다.

본 윤리기준에서 제시하는 원칙과 요건들은 상황에 따라 상충관계가 발생할 수 있으며, 상충하는 문제의 해결 방식은 개별 맥락과 상황에 따라 달라질 수 있다. 따라서 본 윤리기준에서는 각각 원칙들 사이에 고정된 형태의 우선순위를 제시하지는 않으며, 직간접적으로 영향을 받는 이해관계자가 지속적인 토론과 숙의 과정에 참여하여 절충점과 해결 방안을 모색하도록 권유한다.

Ⅱ. 인공지능 윤리기준: 3대 기본원칙, 10대 핵심요건

1. 3대 기본원칙 – 인공지능 개발 및 활용 과정에서 고려될 원칙

• '인간성을 위한 인공지능(AI for Humanity)'을 위해 인공지능 개발에서 활용에 이르는 전 과정에서 고려되어야할 기준으로 3대 기본원칙을 제시한다.

① 인간 존엄성 원칙

- 인간은 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계제품과는 교환 불가능한 가치가 있다.
- 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용되어야 한다.
- 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.

② 사회의 공공선 원칙

- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙

- 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.
- 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.

2. 10대 핵심요건 – 기본원칙을 실현할 수 있는 세부 요건

• 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건을 제시한다.

① 인권보장

- 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 권리를 보장하여야 한다.
- 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

② 프라이버시 보호

- 인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.
- 인공지능 전 생애주기에 걸쳐 개인 정보의 오용을 최소화하도록 노력해야 한다.

③ 다양성 존중

- 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영해야 하며, 성별·연령·장애·지역·인종·종교·국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다.
- 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

④ 침해금지

- 인공지능을 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안 된다.
- 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

⑤ 공공성

- 인공지능은 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- 인공지능은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다.
- 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.

⑥ 연대성

- 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다.
- 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다.
- 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

⑦ 데이터 관리

- 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다.

⑧ 책임성

- 인공지능 개발 및 활용과정에서 책임주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다.

⑨ **안전성**

- 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

⑩ **투명성**

- 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

III. 부록

1. 본 윤리기준에서 인공지능의 지위

- 본 윤리기준에서 지향점으로 제시한 '인간성을 위한 인공지능(AI for Humanity)'은 인공지능이 인간을 위한 수단임을 명시적으로 표현하지만, 인간 종 중심주의(human species-centrism) 또는 인간 이기주의를 표방하지는 않는다.
- 본 윤리기준에서 인공지능은 지각력이 있고 스스로를 인식하며 실제로 사고하고 행동할 수 있는 수준의 인공지능(이른바 강인공지능)을 전제하지 않으며 하나의 독립된 인격으로서의 인공지능을 의미하지도 않는다.

2. 적용 범위와 대상

- 본 윤리기준은 인공지능 기술의 개발부터 활용에 이르는 전 단계에 참여하는 모든 사회구성원을 대상으로 하며, 이는 정부·공공기관, 기업, 이용자 등을 포함한다.

3. 인공지능 윤리기준의 실현방안

- '인공지능 윤리기준'을 기본 플랫폼으로 하여 다양한 이해관계자 참여하에 인공지능 윤리 쟁점을 논의하고, 지속적 토론과 숙의 과정을 거쳐 주체별 체크리스트 개발 등 인공지능 윤리의 실천 방안을 마련한다.

부록 2 국내외 주요 인공지능 윤리원칙

	제목	주체	수립목적	주요 원칙	주요 특징
1	Preparing for the Future of Artificial Intelligence ('16)	NSTC (National Science and Technology Council)	美정부의 입장에서 인공지능 기술과 관련하여 나아갈 방향 제시한 정부 보고서	공공선, 공정성, 안전, 투명성, 이해가능성, 선을 위한 인공지능(AI for good), 인간가치(Human values)	• 인공지능을 주요 성장동력으로 보고 美 정부의 역할 강조 • 윤리원칙 제시보다는 국제인도법에 근거한 인공지능 무기체계 개발 등 다양한 인공지능 관련 이슈를 제시하는데 초점
2	Tenets ('16)	Partnership on AI	학계, 재계, 정책입안자 등 다양한 주체들의 협력 도모	AI 혜택 최대화, 다양한 주체들 간 협력, 사생활 보호, 견고함, 해악금지, 설명가능성	• 학계, 기업, 정책입안자 등 다양한 주체들 간 협력을 강조하고, 이를 통해 대중교육 등 추진할 것을 제안 • 기술 혜택 최대화의 전제로 사생활 보호, 연구공동체의 책임, 견고성, 해악금지 등 제시
3	AI Policy Principles ('17)	ITI (Information Technology Industry Council)	개발자에 대한 정부차원의 지원 및 공적영역과 사적 영역의 협업 강조	안전과 제어가능성, 해석가능성, 인간 존엄성, 데이터의 대표성, 유연한 규제접근, 기회의 평등	• 개발자의 입장을 강조, 정부의 규제나 개발자에 대한 정보공개 요구에 부정적 • 다만 개발자에게도 안전한 설계, 데이터 대표성 등 높은 수준의 책임성 요구
4	DeepMind Ethics & Society Principles ('17)	DeepMind	사내에서 AI 연구 수행시 윤리적 고려사항 제시	사생활 침해 금지, 평등, 도덕성, 포용성, 안전과 책무성, 거버넌스·규제	• 연구자에게 필요한 윤리원칙과 체크리스트를 제시하면서도 안전과 책무성을 보장하는 거버넌스·규제 필요성 제기
5	Asilomar AI Principles ('17)	Future of Life Institute	미국 보스톤의 비영리 연구단체인 삶의 미래 연구소 (Future of Life Institute) 주관으로 작성한 윤리원칙	인권보장, 개인정보보호, 해악금지, 공공성, 데이터 관리, 책임성, 통제성, 투명성, 무기경쟁 회피	• 스티븐 호킹·일론 머스크 등 다수의 인공지능 학자, 미래학자 및 산학연 관계자들이 서명 • 인공지능기술연구자, 정책입안자, 관련 산업 종사자에게 필요한 윤리원칙 제시

	제목	주체	수립목적	주요 원칙	주요 특징
6	AI at Google: Our Principles ('18)	Google	구글 AI 개발자에게 필요한 윤리원칙 제시	사회적 혜택 증진, 불공정한 편견 지양, 설명가능, 사생활침해 방지	• 개발을 제한해야되는 인공지능어플리케이션으로해를 끼치는 기술, 인명을 해하는 무기관련 기술, 국제 규약 위반 감시기술 등 제시
7	Microsoft AI principles ('18)	Microsoft's AETHER(AI and Ethics in Engineering and Research)	MS AI 개발자에게 필요한 윤리원칙 제시	공정성, 신뢰성 및 안전, 사생활 및 보안, 포용성, 투명성, 책무성	• MS 사내 윤리강령 성격이 강하며, 책임질 수 있는 AI와 이를 위한 교육 강조
8	OpenAI Charter ('18)	OpenAI	AI 기술 연구자에게 필요한 윤리적 태도와 원칙 제시	공공선, 해악금지, 안전 담보, AI 개발 선두주자, 타 연구단체 협력,	• 연구자의 자유로운 연구 증진에 초점, 고도로 자율적인 AGI(artificial general intelligence) 상정
9	Principles for Trust and Transparency ('18)	IBM	IBM 직원들을 대상으로 AI 연구를 위해 제시된 윤리원칙	인간 지능 증강 (augment), 데이터 소유권, 국경간 데이터 이동, 투명성	• 인공지능은 인간을 대체하는 것이 아니라 증강하기 위한 것임을 명시 • 인공지능 사용 여부·시기, 학습 데이터 출처 고지 등 규정
10	The Montreal Declaration for a Responsible Development of AI ('18)	University of Montreal	몬트리올 대학에서 개발된 사회적으로 책임 있는 AI 연구를 위한 윤리원칙	복지(well-being), 자율성 존중, 사생활 보호와 친밀성, 연대성, 민주적 참여, 공평, 다양성 포용, 사려	• 친밀성(intimacy), 사려(prudence), 지속가능한 발전등 다른 가이드라인에 잘 등장하지 않는 원칙 제시 • 윤리원칙 제시와 함께 서명으로 선언에 동참하도록 장려
11	지능정보사회 윤리가이드라인 ('18)	정보문화 포럼 (정부)	인간 중심의 지능정보 사회 구현	이용자 주도성, 이용자/시민참여, 공익, 공정성, 위험예방, 프라이버시 보호	• 지능정보기술 관련 개발자·공급자의 윤리의식 고취 및 이용자의 오남용 방지 지침 • 주체별(개발자, 공급자, 이용자)세부지침 마련
12	AI in the UK: Ready, Willing and Able? ('18)	영국 정부	영국 정부 차원에서 정책적으로 접근할 수 있는 제언 제시	데이터 접근과 제어, 이해가능한 AI, 디지털 이해력 증진, 공중보건 관리	• 영국이라는 특정 국가 입장에서 공중보건 데이터 관리, 인공지능 디지털 이해력 제고 등 구체적으로 취할 수 있는 인공지능 관련 정책을 제시

	제목	주체	수립목적	주요 원칙	주요 특징
13	카카오 알고리즘 윤리헌장 ('18)	카카오	카카오 내 인공지능 관련 연구시 지향되어야 할 윤리원칙 제언	사회윤리 준수, 차별 경계, 학습데이터 운영, 알고리즘 독립성 및 설명, 기술 포용성, 아동·청소년 보호	• 국내기업 최초 인공지능 윤리헌장으로, 알고리즘과 데이터에 대한 관리, 아동과 청소년에 대한 보호 필요성 등 강조
14	The Future Computed: AI and its role in Society ('18)	Microsoft's AETHER	AI가 가져올 미래의 변화에 대응하기 위해 MS의 Aether 연구소에서 책자 제작	인공지능에 의한 진보, 공정성, 신뢰성 및 안전, 사생활·보안, 포용성, 투명성	• 인공지능이 경제·사회적 진보를 이끌고 지역적·전 지구적 문제를 해결할 것이라는 관점 • 인공지능이 직업과 직장에 미치는 영향에 공공부문과 민간부문이 협력해 대응할 필요성 제시
15	Discriminating Systems – Gender, Race, and Power in AI ('19)	AI Now	작업환경에서 다양성을 확보하기 위해 고려할 사항 제시	다양성, 해악금지, 개방성, 투명성	• 급여지급 기준의 인종별, 성별 공개, 직원 채용시 투명성 준수 등 제시 • 특히인공지능시스템사용시 투명성·편견·해악에 대한 철저한 점검·감시·추적·공개를 강조
16	Ethically Aligned Design ('19)	IEEE (Institute of Electrical and Electronics Engineers)	IEEE에서 Ethics in Action 캠페인과 함께 아울러 공개된 보고서	인권, 복지우선, 책무성, 투명성, 오용의 인식	• 각 원칙별 이론적 배경, 참고자료를 제시하고 윤리원칙뿐만 아니라 관련 분야들에 대한 자료 수록
17	이용자 중심의 지능정보 사회를 위한 원칙 ('19)	방송통신 위원회	안전한 지능정보 서비스 환경조성 및 이용자의 권리와 자유에 근거한 윤리원칙 제시	사람중심 서비스, 투명성과 설명가능성, 책임성, 안전성, 차별금지, 참여, 프라이버시와 데이터거버넌스	• 안전한지능정보서비스환경 조성 및 이용자 보호를 위해 모든 주체 사이의 협력 강조 • 기업과 연구자들의 의견을 폭넓게 수렴하여 작성 • 이용자 보호의 관점 강조
18	로봇 윤리 기본 원칙 (수정) ('19)	산업 통상자원부	2007년에 만들어진 로봇윤리 헌장을 수정 보완	인간의 존엄성 보호, 공공선, 행복추구, 투명성, 제어가능성, 책무성, 안전성, 정보보호	• 로봇 산업계 종사 연구자 및 개발자, 사용자가 로봇과 인공지능을 설계·제작·공급·사용·관리하는데 기준으로 삼는 가이드라인 제시

	제목	주체	수립목적	주요 원칙	주요 특징
19	인간중심의 AI 사회 원칙 ('19)	일본 총무성	25명의 산학연 전문가로 구성된 ' 인간 중심의 인공지능 사회 원칙 위원회'를 통해 제안	인간중심, 교육교양, 개인정보 보호, 보안, 공정경쟁, 공정성, 책임성, 투명성, 혁신	• 저출산, 고령화, 지방쇠퇴, 재해 재난 등 일본이 처한 어려움을 인공지능이 해결할 수 있을 것으로 상정 • 인공지능을 공공재로 활용하여사회의근본적인변화와 혁신을달성하여지속가능한 발전 추구
20	Ethics Guidelines for Trustworthy AI ('19)	EU	EU 산하의 50여명으로 구성된 AI 전문가 그룹 주도	인간 권리·자율성 보장, 기술적 견실성, 사생활, 데이터 관리, 투명성, 다양성, 차별금지, 복지, 책무성	• 범국가 차원의 협업을 통해 신뢰할 수 있는 인공지능을 위한윤리원칙정립에초점을 맞춤 • 각 원칙의 평가 리스트를 구체적으로 제시
21	Recommendation of the Council on AI ('19)	OECD	OECD 디지털 경제 정책 위원회 주관하에 제작	포용적 성장, 지속가능 발전, 인간중심 가치, 공정성, 투명성, 설명가능성, 견고성, 보안 및 안전, 책무성	• 윤리원칙 뿐 아니라 정책 입안자들에 대한 제언 제시, 국가별 정책 수립과 국제적 협력 도모
22	Understanding artificial intelligence ethics and safety ('19)	The Alan Turing Institute	영국의 국영 연구소인 Alan Turing 연구소에서 제작	존중, 연결, 보호, 돌봄, 공정성, 책임성, 지속가능성, 투명성	• 인공지능 기술이 데이터를 처리할 때 발생할 수 있는 위험이나 문제점을 예방하는 데 필요한 윤리원칙에 초점
23	Rome Call for AI Ethics ('20)	로마 교황청	로마 교황청에서 인간의 혁신적인 미래를 위한 인공지능 윤리원칙 제정	투명성, 포용성, 책임성, 불편부당성, 신뢰성, 보안과 사생활 보호	• 종교기관인가톨릭교회에서 제정한윤리원칙으로, 인간 가족(human family)에 대한 봉사, 젊은 세대에 대한 준비, 자연의 회복 필요성 등 제시
24	사람이 중심이 되는 인공지능 윤리기준 ('20)	대한민국 정부	인간성을 위한 인공지능 구현을 위해 개발-활용 전 단계에서 함께 지켜야 할 기본원칙과 핵심요건 제시	(3대 기본원칙) 인간 존엄성,사회의 공공선, 기술의 합목적성(10대 핵심요건) 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성	• 인공지능 전 주기에서 모든 사회 구성원이 참조할 수 있는 기준 마련 • 특정 분야에 제한되지 않는 범용성을 가진 총론 차원의 일반 원칙 • 구속력있는 '법'이나 '지침'이 아닌 자율규범으로서의 윤리 기준

	제목	주체	수립목적	주요 원칙	주요 특징
25	자율주행차 윤리 가이드라인 ('20)	국토교통부	국민들이 수용할 수 있는 자율주행차 도입환경 조성	투명성, 안전성, 보안성, 책임성	• 레벨4 이상 완전자율주행 시스템이 장착된 자동차의 행위주체자(설계자, 제작자, 관리자,서비스제공자,이용자 등)를 대상으로, 행위원칙과 각 주체들이 준수해야 할 윤리원칙 제시
26	인공지능 관련 개인정보보호 원칙 ('21)	개인정보보호 위원회	인공지능 개발자·운영자의 개인정보 보호에 대한 인식 제고 및 자율적 실천	적법성, 안전성, 투명성, 참여성, 책임성, 공정성	• 「개인정보 보호법」 상 주요 준수사항 및 국가 인공지능 윤리기준('20.12) 등을 반영하여 인공지능 관련 개인정보보호 6대 원칙을 도출 • 인공지능 기획·설계, 개인정보수집, 이용·제공, 보관·파기 등의 단계 및 상시 점검을 위한 '개인정보 보호 자율점검표'를 함께 제공
27	인공지능 기반 추천 서비스 이용자 보호를 위한 기본원칙 ('21)	방송통신 위원회	디지털 미디어 콘텐츠의 편향적·차별적 제공을 방지하고 이용자의 권익을 적극 보장	투명성, 공정성, 책무성	• 인공지능 기반 추천서비스 제공자 대상 이용자 권익 보호를 위한 핵심원칙과 실행원칙 제시 • 디지털미디어영역의특성을 반영한 특칙에 해당하는 개별 규범
28	금융분야 인공지능 가이드라인 ('21)	금융위원회	금융분야에서의 인공지능 운영 전 과정 신뢰성 제고를 통해 인공지능 활성화, 금융서비스 고객신뢰 확보에 기여	책임성, 데이터 정확성·안전성, 서비스 투명성·공정성, 금융소비자 권리 보장	• 인공지능을금융거래,대고객 서비스에 적용한 전 금융업권 대상 • 가이드라인을 통해 인공지능 금융서비스에 대한 사회적 신뢰를 공고히 하고 지속가능한 금융혁신 환경 조성 추진
29	네이버 인공지능 윤리준칙 ('21)	네이버	네이버의 모든 구성원이 인공지능 개발과 이용에 있어 준수해야 할 원칙	사람을 위한 인공지능 개발, 다양성 존중, 합리적 설명과 편리성의 조화, 안전 고려 서비스 설계, 프라이버시 보호와 정보 보안	• 윤리준칙 공개와 함께 실천 사례 축적 및 개선을 위한 학계와의 지속적 협력을 통해준칙을더욱구체화하고 개선해나갈 것을 밝힘

	제목	주체	수립목적	주요 원칙	주요 특징
30	사람 중심의 인공지능 윤리 가치 원칙 ('21)	SK텔레콤	사람 중심의 인공지능을 핵심 이념으로 하는 인공지능 추구 가치를 제정하고, 고객 제공 및 구성원 실천을 위한 기준 제시	사회적 가치, 무해성, 기술 안정성, 공정성, 투명성, 사생활 보호, 지속혁신	• 'AI Company로의 혁신' 선언에 맞추어 '사람' 중심의 AI 이념 설정 • 인공지능 추구 가치의 사규 반영, 구성원 교육 실시, 체크리스트 개발 등 전사 프로세스 반영 계획
31	Facebook's Five Pillars of Responsible AI ('21)	Meta	인공지능이 가지는 프라이버시, 공정성, 책임성 및 투명성과 같은 문제를 직면하고, 협력적인 방식으로 문제 해결	개인정보 보호·보안, 공정성·포괄성, 견고성·안전, 투명성·통제, 책임·거버넌스	• 인공지능 영향 평가가 명확하고 합리적인 기준으로 만들어질 수 있도록 협력해야 함을 강조 • 인공지능 공정성, 프라이버시, 견고성 및 투명성에 관한 표준 개발 협력 제안
32	Recommendation on the Ethics of Artificial Intelligence ('21)	UNESCO	인공지능 시스템이 인류·개인·사회·환경 및 생태계 이익을 위해 작동하며 피해를 방지하도록 회원국에 윤리적 틀과 규범적 도구 제공	비례의 원칙·무해성, 안전·보안, 공정성·반차별성, 지속가능성, 프라이버시 및 데이터 보호의 권리, 인간의 감독·결정, 투명성·설명가능성, 책임·의무, 리터러시, 다자적·조정 가능한 거버넌스와 협력	• 권고의 형식을 통해 회원국별 상황에 맞게 규제 프레임워크의 도입과 개선 요구 • 인공지능 윤리 영향 평가의 토대 및 공공·민간 부문을 비롯한 모든 인공지능 행위 주체에게 윤리 지침 제공
33	Principles for Responsible AI('22)	Cisco	모든 사람의 인권을 존중하고 유지하기 위한 원칙 제시	투명성, 공정성, 책임성, 개인정보보호, 정보 보안, 안정성	• 원칙은 Cisco의 운영방침과 일치하며, 인공지능 기술 거버넌스에 직접 적용 가능하도록 고안 • 원칙에는 구체적 행동강령이 포함 • 해당 산업 특성상 기술의 보안에 초점

	제목	주체	수립목적	주요 원칙	주요 특징
34	인공지능 개발과 활용에 관한 인권 가이드라인 ('22)	국가인권위원회	인공지능 개발·활용 시 인권침해 예방을 위해 준수해야 할 인권적 관점 기준 제시	인간의 존엄성 존중, 투명성·설명의무, 자기결정권 보장, 차별금지, 인공지능 인권영향평가 시행, 위험도 등급 및관련 법제도마련	• 인공지능 관련 정책 수립·이행, 관계 법령 제·개정 등에 가이드라인을 참고할 것을 국무총리 및 관련부처 장관에 권고 • 특히, 인공지능 인권영향평가 시행을 통한 인권침해 및 차별 발생 가능성을 평가하고, 부정적 영향을 방지하기 위한 조치 적용과 내용 공개 요구
35	LG 인공지능 윤리 원칙 ('22)	LG	인공지능을 개발하고 활용하는 LG의 모든 구성원이 지켜야 할 올바른 행동과 가치 판단의 기준이 되는 원칙 제시	인간존중, 공정성, 안전성, 책임성, 투명성	• 인공지능 개발을 위한 필수적 가치를 비롯해 LG의 경영 이념 '고객을 위한 가치 창조', '인간 존중의 경영'을 기반으로 핵심가치 선정 • 인공지능 윤리점검 TF를 신설
36	인공지능 윤리기준 ('22)	KB금융그룹	금융서비스 전반에 걸쳐 인공지능을 도입과 활용함에 있어 준수해야 할 가치를 담은 윤리기준 제정	공정과 포용, 참여와 협력, 디지털 역량, 데이터 관리, 투명한 활용, 통제 가능성, 안전과 책임	• KB금융그룹구성원모두가 윤리적인 인공지능 기술을 개발 및 활용할 수 있도록 구성원들의 '디지털 역량' 강화를 가치에 반영 • 인공지능 윤리기준을 국내 금융그룹 중 최초로 마련
37	교육분야 인공지능 윤리원칙 ('22)	교육부	사람의 성장을 지원하기 위해 교육당사자와 관계자가 인공지능의 교육적 활용 시 참고하고 준수할 수 있는 선제적 규범 마련	인간성장 잠재성 유도, 학습자 주도성·다양성 보장, 교수자 전문성 존중, 교육당사자 간 공고한 관계 유지, 교육의 기회균등과 공정성 보장, 교육공동체 연대·협력 강화, 사회 공공성 증진 기여, 교육당사자 안전 보장, 데이터 처리의 투명성 보장 및 설명 가능성, 데이터의 합목적적 활용·프라이버시 보호	• 교육기관, 교육활동 지원 행정기관에서 활용되는 인공지능 대상 • 자발적 실천과 준수를 독려하는 도덕적 규범 및 자율규제 성격 • 국가인공지능윤리기준('20.12)과 UNESCO 윤리권고('21.11) 등을 준수하는 교육분야 인공지능에 대한 특수 규범

	제목	주체	수립목적	주요 원칙	주요 특징
38	Blueprint for an AI Bill of Rights ('22)	미국 정부	인공지능 기술의 진전이 시민권, 민주주의적 가치를 훼손하지 않도록 보호	시스템 안전성·효과성, 알고리즘 차별방지, 데이터 프라이버시, 고지 및 설명, 인적 대안 마련	• 미 백악관 과학기술정책실(OSTP)은 인공지능 시대 미국인들의 권리 보호를 위한 다섯가지 방안으로 기본원칙 제시 • 기본원칙 달성을 위한 인공지능 및 자동화된 시스템의 설계, 개발, 배포 지침을 담은 청사진 발표
39	AI Ethics: Principles & Guidelines ('22)	아랍에미리트 정부	인공지능 기술을 설계·활용하는 조직이 인공지능 윤리 원칙을 준수할 수 있도록 자세한 지침을 제공	공정성, 책임성, 투명성, 설명 가능성, 보안, 안전성, 인간 중심, 지속가능성, 환경친화적, 개인정보보호	• 원칙과 관련된 구체적인 사례를 제시하여 AI 개발자와 사용자들이 윤리적 고려사항을 명확히 이해하고 실천할 수 있도록 지원
40	챗GPT 등 생성형 AI 활용 보안 가이드라인 ('23)	국가정보원/국가보안기술연구소	생성형 AI 관련 보안 문제를 사전에 예방하고 안전하게 이용하기 위한 보안수칙 제시	개인정보보호, 투명성, 신뢰 가능성, 저작권 보호, 보안, 안전성	• 생성형 AI 접속에서부터 질의, 결과물 활용 등 전반적으로 지켜야 할 보안 지침을 단계별로 안내
41	서울시 생성형 AI 윤리 가이드라인 ('23)	서울디지털재단	생성형 AI라는 새로운 유형의 위험에 대비하기 위해 서울시 이해관계자가 자율 준수할 수 있는 가이드라인 마련	저작권 보호, 공공성, 책임성, 안전성, 지속가능성, 개인정보보호, 책임성, 이용자의 AI 윤리 소양 및 검증 역량 강화, 보안성	• 서울특별시의 핵심 가치 및 정책 방향성을 바탕으로 가이드라인 제시
42	Israel's Policy on Artificial Intelligence Regulation and Ethics ('23)	이스라엘 정부	인간의 기본적 권리와 공공 이익을 보호하고 책임감 있는 인공지능 개발과 사용을 지원	인권보장, 책임성, 공공성, 연대성, 평등 및 차별금지, 안전성, 신뢰성, 보안, 안전성, 회복성, 투명성 및 설명 가능성, 개인정보보호	• OECD AI 원칙을 기반으로 한 공통된 윤리적 AI 원칙을 채택

	제목	주체	수립목적	주요 원칙	주요 특징
43	생성형 인공지능 서비스 안전 기본 요구사항 (生成式人工智能服务安全基本要求) ('24)	중국 정부	생성형 AI 서비스 제공업체가 준수해야 하는 기본적인 요구사항 제시	저작권 보호, 보안, 안전성, 불법·유해정보 차단, 개인정보보호	• 생성형 AI 서비스 제공 업체는 관할 당국에 서비스를 등록하기 전에 본 요구사항에 따른 보안 평가를 실시해야 함
44	Risk Management Profile for Artificial Intelligence and Human Rights ('24)	미국 정부	정부, 민간, 시민사회 등이 국제 인권을 존중하는 방식으로 AI를 설계·배포·사용·관리하도록 지침 제공	인권 보호, 보안, 회복성, 투명성, 책임성, 개인정보보호, 공정성	• NIST의 AI 위험관리 프레임 워크(AI RMF)가 제시한 4개 기능(거버넌스, 위험 식별, 측정, 관리)을 기반으로 AI로 인한 인권 침해 위험을 해결하기 위한 모범 관행을 안내

자료 : AI 윤리 소통채널(ai.kisdi.re.kr) 등을 참고하여 내용 업데이트

부록 3 국내외 주요 인공지능 윤리 실천수단

제목	The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment (2020)
주체	EU (범국가 정부기관)
목적	EU 윤리 가이드라인에 제시된 **핵심요건을 이행하기 위한 자율 평가 도구** 마련
대상	**인공지능 설계자와 개발자,** 데이터 과학자, 구매조달업무 담당자, 실제 인공지능시스템 활용할 직원 등
주요 내용	① **기본권 평가:** 인공지능시스템이 평등권, 아동권리 등의 기본권에 미치는 영향의 **사전 평가** ② **신뢰가능성 평가:** 핵심요건별*로 구성된 평가질문에 따라 "예", "아니오", "잘 모르겠음" 등으로 응답 내용 평가 * 핵심요건: ① 인간의 기본권·자율성 보장, ② 기술적 견고성 및 안정성, ③ 개인정보 및 데이터 거버넌스, ④ 투명성, ⑤ 다양성, 차별금지 및 공정성, ⑥ 사회 및 환경적 영향, ⑦ 책무성
구성	• 총 143문항으로 구성 • 기본권 평가 및 7대 요건별 평가 문항으로 구성 • 대표문항(73개) 및 세부문항(70개) 구분 • 부록으로 용어집 포함

구분	기본권	요건1	요건2	요건3	요건4	요건5	요건6	요건7	합계
대표문항	4	11	21	6	5	10	8	8	73
세부문항	9	12	12	2	8	15	8	4	70
계	13	23	33	8	13	25	16	12	143

제목	Guidance on Artificial Intelligence and Data Protection (2020)
주체	영국 ICO (정부기관)
목적	영국의 **인공지능 산업 활성화 전략(AI Sector Deal)의 이행 도구**로 인공지능시스템에 대한 **감사(Audit) 지침** 제공
대상	① 법·감사 전문가: 법무자문위원, 고위경영진, DPO(Data Protection Officer), ICO 감사관 등 ② 기술 전문가: 머신러닝전문가, 데이터과학자, SW개발자, 사이버·IT보안 관리자 등 기술 전문가
주요 내용	**4대 분야로 구분**하여 검토사항을 제시 ① 책임 및 거버넌스 ② 데이터처리의 적법성·공정성·투명성 ③ 보안 및 개인정보 처리의 최소화 ④ 개인정보 관련 권리 보장
구성	• 총 **180문항**으로 구성 • 평가 문항은 **주요 이슈, 관련 법조문, 사례 및 예시**와 함께 제시 • 부록으로 용어집 포함

구분	분야1	분야2	분야3	분야4	합계
문항	86	46	16	32	180

<table>
<tr><td>제목</td><td>Companion to the Model AI Governance Framework
– Implementation and Self-Assessment Guide for Organizations (2020)</td></tr>
<tr><td>주체</td><td>WEF·싱가포르 정부 (범국가 기관-정부기관 협력)</td></tr>
<tr><td>목적</td><td>싱가포르 정부의 인공지능 거버넌스 모델(Model AI Governance Framework)'을 이행하기 위한
자율적 평가지침을 WEF와 협력하여 마련</td></tr>
<tr><td>대상</td><td>인공지능 솔루션을 조달·활용하여 소비자 제품과 서비스에 적용하거나 인공지능 기술을 활용해 조직 운영 효율성을 향상하기 위한 민간조직</td></tr>
<tr><td>주요 내용</td><td>인공지능 활용 목적 및 절차를 기반으로 5대 영역을 구분해 검토사항 제시
① 인공지능 솔루션 활용 목적
② 내부 거버넌스 구조·조치
③ 인공지능 기반 의사결정에 대한 인간의 개입
④ 책임 있는 운영·관리
⑤ 이해당사자 간 소통</td></tr>
<tr><td>구성</td><td>• 총 64문항으로 구성
• 평가 문항은 주요 고려사항, 예시, 관행과 함께 제시됨
• 부록으로 ISO 및 IEEE의 인공지능 기준 목록 제공
<table><tr><td>구분</td><td>영역1</td><td>영역2</td><td>영역3</td><td>영역4</td><td>영역5</td><td>합계</td></tr><tr><td>문항</td><td>3</td><td>9</td><td>4</td><td>32</td><td>16</td><td>64</td></tr></table></td></tr>
</table>

<table>
<tr><td>제목</td><td>AI System Ethics Self-Assessment Tool (2019)</td></tr>
<tr><td>주체</td><td>아랍에미리트 정부 (정부기관)</td></tr>
<tr><td>목적</td><td>아랍에미리트 스마트 두바이청의 인공지능 윤리원칙 기반으로 자율적 평가도구를 베타 버전으로 마련</td></tr>
<tr><td>대상</td><td>① 정부 기관: 공공서비스를 위해 인공지능 기술을 조달하거나 내부적으로 개발하는 기관
② 민간 기관: 정부 또는 민간부문 고객을 위해 인공지능시스템을 제공하는 민간기업
③ 개인: 윤리적인 인공지능이 공공영역 및 사회에서 어떻게 적용되는지 관심 있는 개인</td></tr>
<tr><td>주요 내용</td><td>인공지능 윤리원칙에 따라 4가지 요건을 구분해 검토사항 제시
① 공정성
② 책무성
③ 투명성
④ 설명가능성</td></tr>
<tr><td>구성</td><td>• 총 58문항으로 구성
• 대표문항(17개) 및 세부문항(41개) 구분
• 평가 문항은 관련 예시, 변경사항 추적내용과 함께 제시됨
<table><tr><td>구분</td><td>요건1</td><td>요건2</td><td>요건3</td><td>요건4</td><td>합계</td></tr><tr><td>대표문항</td><td>5</td><td>7</td><td>2</td><td>3</td><td>17</td></tr><tr><td>세부문항</td><td>8</td><td>24</td><td>5</td><td>4</td><td>41</td></tr><tr><td>계</td><td>13</td><td>31</td><td>7</td><td>7</td><td>58</td></tr></table></td></tr>
</table>

제목	인공지능 윤리기준 실천을 위한 자율점검표 (2022, 2023, 2024, 2025)
주체	과학기술정보통신부 (정부기관)
목적	인공지능 시스템의 개발·운영과정에서 「**인공지능 윤리기준**」('20.12)의 3대 기본원칙과 10대 핵심요건을 **실천**하기 위해 점검해야 할 요소와 이를 이행하는 구체적 방법을 다수의 문항으로 제공
대상	인공지능 시스템의 개발·운영에 필요한 의사결정을 수행하는 조직과 개인
주요 내용	인공지능 기술 개발 및 발전 과정에서 각 핵심요건을 만족하고 있는지를 확인할 수 있도록 **10대 핵심요건**별로 점검문항 제시 ① 인권보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성
구성	• 총 **35문항**으로 구성 • 점검문항은 제안이유, 고려사항, 참고자료, 사례 등과 함께 제시 • 범용성·포괄성에 중점을 둔 **공통 자율점검표**를 목적·특성 등에 맞게 실제 현장에서 보다 쉽게 응용할 수 있도록 구체적 활용 예시로 **분야별 자율점검표 제공**(챗봇, 작문, 영상관제, 채용, 영상합성 분야)

구분	요건1	요건2	요건3	요건4	요건5	요건6	요건7	요건8	요건9	요건10	합계
문항	5	2	5	4	3	3	3	4	3	3	35

제목	신뢰할 수 있는 인공지능 개발 안내서 (2022, 2023, 2024)
주체	과학기술정보통신부 (정부기관)
목적	인공지능 신뢰성 확보를 위해 기술 측면에서 고려해야 할 항목과 방법 제공
대상	인공지능 서비스 구현 과정에 직·간접적으로 관련되거나 영향을 주는 모든 조직과 개인 (특히, 기술적 관점에서 신뢰성을 신경써야하는 인공지능 서비스 기획자, 데이터 수집 및 가공자, 인공지능 모델 개발자, 시스템 및 소프트웨어 개발자, 테스터 등이 주요 대상)
주요 내용	OECD, UNESCO, ISO/IEC, IEEE 등 국제사회의 논의 및 합의 내용과 국제표준 기반, 인공지능 신뢰성 확보를 위한 **기술 요구사항**과 **검증항목**을 **인공지능 서비스 생명주기 단계**별로 제시 ① 생명주기 관리, ② 데이터 수집·처리, ③ 인공지능 모델 개발, ④ 시스템 구현, ⑤ 운영·모니터링
구성	• 15개 기술 요구사항에 대한 검증항목 69문항으로 구성 • 인공지능 서비스 유형 검토, 확인 및 검증해야 할 요구사항 선별, 선별된 요구사항에 대한 검증 등의 세 단계로 요구사항 확인 및 검증 절차 수행 권장 • 분야별 안내서 별도 제공(의료, 자율주행, 공공사회, 채용, 스마트치안, 생성 AI 기반 서비스 분야)

구분	단계1	단계2	단계3	단계4	단계5	합계
요구사항	4	3	4	3	1	15
검증항목	21	18	12	14	4	69

제목	인공지능 개인정보보호 자율점검표 (2021)
주체	개인정보보호위원회 (정부기관)
목적	인공지능 개발자·운영자의 개인정보보호에 대한 인식을 제고하고, 개인정보보호에 관한 의무준수 뿐만 아니라, 자사 기술·서비스 환경에 적합한 자율적인 개인정보보호 이행·점검에 필요한 사항 안내
대상	「개인정보 보호법」 상 개인정보처리자와 개인정보취급자의 지위를 갖는 인공지능 개발자·운영자
주요 내용	법령상 준수해야 할 의무 또는 권장사항을 **인공지능 기술·서비스 개발·운영 단계별(또는 상시)**로 제시 **(단계)** ① 기획·설계, ② 개인정보 수집, ③ 개인정보 이용·제공, ④ 개인정보 보관·파기 **(상시)** ⑤ 인공지능 서비스 관리·감독, ⑥ 인공지능 서비스 이용자 보호 및 피해구제, ⑦ 개인정보 자율보호 활동, ⑧ 인공지능 윤리 점검
구성	• **점검항목(16개), 확인사항(54개)으로 구성** • 점검항목을 **의무·권장사항으로 구분**하였으며, 관련 법령·고시의 규정, 개인정보보호 법령 및 지침·고시 해설 등을 종합적으로 참고할 것을 권고

구분	단계1	단계2	단계3	단계4	상시5	상시6	상시7	상시8	합계
점검항목	2	3	2	2	2	3	1	1	16
확인사항	4	10	13	6	5	7	5	4	54

제목	생성형 AI 저작권 안내서 (2023)
주체	문화체육관광부·한국저작권위원회 (정부기관)
목적	생성형 AI의 산출물이 만들어지는 과정에서 이해관계를 가지는 각 주체들이 저작권과 관련하여 알아두어야 할 권고사항 제공
대상	생성형 AI 사업자, 저작권자, 생성형 AI 서비스 이용자
주요 내용	생성형 AI의 산출물 관련 이해관계자를 **세 가지 주체**로 구분하여 **각 주체별 권고사항을 제시** ① AI 사업자, ② 저작권자, ③ AI 이용자
구성	• 권고사항은 생성형 AI 학습 단계와 AI 산출물 생성 단계로 나누어 제공 • **관련 법조항, 법적 쟁점, 사례 및 예시**와 함께 제시

제목	AI Risk Management Framework 1.0 (2023)
주체	미국 국립표준기술연구소 (정부기관)
목적	모든 분야 기업·조직이 AI 위험을 해결할 수 있도록 **유연하고 체계적이며 측정 가능한 프로세스와 자발적 활용 가이드 문서** 제공
대상	모든 분야·규모의 기업·조직의 AI 시스템
주요 내용	**신뢰할 수 있는 AI 시스템의 7가지 특성**을 제시하고, 이를 달성하기 위한 권장 조치를 제공 ① 유효성 및 신뢰성, ② 안전성, ③ 보안 및 탄력성, ④ 책임 및 투명성, ⑤ 설명 및 해석가능성, ⑥ 개인정보보호 강화, ⑦ 공정성(유해한 편향 관리)
구성	• 총 **72문항**으로 구성 • 권장조치는 4가지 핵심 기능 **거버넌스(Govern)**, **위험식별(Map)**, **측정(Measure)**, **관리(Manage) 차원**에서 범주와 하위 범주로 구분 • 프레임워크를 조직에서 효과적으로 활용하기 위한 구체적인 실행가이드로 AI RMF Playbook을 함께 제공

구분	기능1	기능2	기능3	기능4	합계
대표범주	6	5	4	4	19
하위범주	19	18	22	13	72

제목	Ethical Impact Assessment (2023)
주체	UNESCO (국제기구)
목적	**인공지능 윤리영향평가의 실질적인 운영을 위한 도구를 제공**하고, 인공지능 시스템 조달에 관여하는 정부 관계자가 구매하는 시스템이 **UNESCO 인공지능 윤리 권고에 명시된 윤리적 기준에 부합하는지 확인**할 수 있도록 일련의 질문을 제시
대상	회원국 정부기관 또는 기업에서 구현·배포하는 인공지능 시스템 ※ 특히, 인권에 잠재적인 위험일 것으로 파악되는 인공지능 시스템, 공공기관이 결과 예측, 위험 완화, 피해 예방, 시민 참여 확대, 사회 문제 해결을 위해 만든 인공지능 시스템
주요 내용	조달하고자 하는 인공지능 시스템의 설계, 개발, 및 배포에 대하여 **윤리적 인공지능을 위한 UNESCO의 7개 원칙**별로 평가문항 제시 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
구성	• 평가 대상, 범위, 계획 등을 점검하는 **범주화 질문(scoping question) 단계**와 평가 대상이 UNESCO 인공지능 윤리 권고에 어떻게 부합하는지 심층적으로 검토하는 **원칙 이행(implementing) 평가 단계**로 구분 • **원칙 이행 단계**는 절차적 안전장치 마련 여부를 평가하는 ① **절차상 평가 문항,** 시스템 조달 및 배포로 발생할 수 있는 긍·부정 영향 식별 및 평가하는 ② **영향 식별 및 완화 문항**(정량/정성)으로 구성 • 영향평가 도구를 각 국가의 규제체제 및 특정 상황에 맞게 조정하여 사용하도록 권장

제목	Fundamental Rights Impact Assessment for High-Risk AI Systems (EU AI Act 제27조, 2024)
주체	EU (범국가 정부기관)
목적	고위험 인공지능 시스템이 인간의 **기본권***에 미치는 영향을 사전에 평가하고, 식별된 위험을 완화하기 위한 절차를 수립 * "EU 기본권 헌장"에 명시된 존엄성, 자유, 평등, 연대, 시민권(예: 투표권, 청원권, 연방 문서 열람권), 정의 등을 포함
대상	「EU AI Act」 제6조 2항에 규정된 특정 고위험 인공지능 시스템(세부 규정 참고 필요)
주요 내용	평가의 대상이 되는 특정 고위험 AI 시스템 배포자는 다음 **8개 요건**을 포함한 기본권 영향평가를 수행해야 함(시행일: 26.8.2) ① 대상 AI 시스템이 의도된 목적에 맞게 사용될 절차 설명 ② 대상 AI 시스템의 사용 기간 및 빈도 설명 ③ 특정 상황에서의 사용으로 인해 영향을 받을 가능성이 있는 개인(자연인) 또는 집단의 범주 ④ 식별된 대상에 영향을 미칠 가능성이 있는 구체적인 위험 요소 ⑤ 사용 지침에 따른 인적 감독 조치 이행에 대한 설명 ⑥ 위험이 현실화될 경우 취할 조치 (내부 거버넌스 및 이의 처리 체계를 포함) ⑦ 의도된 목적 하에 적절한 수준의 정확성·견고성·사이버 보안 달성 ⑧ 평가 결과 요약을 포함한 데이터를 EU 중앙 데이터베이스에 입력
구성	• 구체적인 평가 문항 개발 예정 • 자동화된 도구, 설문지 등의 형태로 평가도구를 개발해야 함을 명시

제목	Conformity Assessment (EU AI Act 제43조, 2024)
주체	EU (범국가 정부기관)
목적	적합성 평가 절차를 명확히 하여 **고위험 인공지능 시스템의 안전성과 신뢰성을 확보**
대상	「EU AI Act」 제6조 2항에 규정된 특정 고위험 인공지능 시스템(세부 규정 참고 필요)
주요 내용	법에서 제시한 **8개 요건**의 충족 여부를 지표로 하는 검토사항과 의무 제시(시행일: 26.8.2) ① 위험/품질 관리 시스템 운영 ② 데이터셋 적절성과 데이터 거버넌스 구축 ③ 기술문서 작성과 보관 ④ 로그기록 보관 ⑤ 투명성 및 정보 제공 ⑥ 인적 감독 ⑦ 의도된 목적 하에 적절한 수준의 정확성·견고성·사이버 보안 달성 ⑧ 시장 출시 전 EU 중앙 데이터베이스 등록
구성	• 구체적인 평가 문항은 제시되지 않음

제목	AI Risk Management Framework: Generative Artificial Intelligence Profile (2024)
주체	미국 국립표준기술연구소 (정부기관)
목적	생성형 AI로 인해 발생하는 위험을 식별하고, 목표와 우선순위에 부합하는 **생성형 AI 위험 관리 조치 제시**
대상	모든 분야·규모의 기업·조직의 생성형 AI 시스템
주요 내용	**생성형 AI의 12가지 위험 목록**을 제시하고, 개발자가 이를 관리하기 위한 검토사항 제시 ① CBRN 정보 또는 능력*, ② 허위정보 생성, ③ 위험하고 폭력적이거나 혐오스러운 콘텐츠, ④ 데이터 프라이버시, ⑤ 환경적 악영향, ⑥ 유해한 편향 또는 동질화, ⑦ 인간-AI 협력 간 부작용, ⑧ 정보의 무결성 손상, ⑨ 정보 보안 침해, ⑩ 지식재산권, ⑪ 외설, 비하, 학대적 콘텐츠, ⑫ 가치 사슬 및 구성요소 통합 위험 *화학, 생물학, 방사능, 또는 핵(chemical, biological, radiological, or nuclear, CBRN) 무기 및 기타 위험 물질·병원체 관련 유해한 정보에 대한 접근이 용이해지거나, 기존 정보를 조합(synthesize)하여 새로운 설계 능력을 생성할 수 있는 환경이 조성될 위험
구성	• 총 **212문항**으로 구성 • 권장조치는4가지핵심기능**거버넌스(Govern)**, **위험식별(Map)**, **측정(Measure)**, **관리(Manage) 차원**에서 범주와 하위 범주로 구분

구분	기능1	기능2	기능3	기능4	합계
대표범주	15	9	16	9	49
하위범주	58	39	72	43	212

제목	인공지능 윤리영향평가 (2024)
주체	과학기술정보통신부·정보통신정책연구원(정부기관)
목적	인공지능 제품·서비스의 윤리적 영향력을 사전에 평가함으로써 **긍정적 영향 극대화 및 부정적 영향 최소화를 위한 관리·제도·정책적 조치 방안 등 시사점을 도출**
대상	인공지능 제품·서비스군* *매년 평가대상 선정, 고위험 AI 우선 고려
주요 내용	대상 인공지능 제품·서비스군에 대하여 **인공지능 윤리기준 10대 핵심요건**별 예상되는 긍·부정 영향 식별 및 영향력 종합 평가 ① 인권보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성
구성	• 총 **110문항**으로 구성 • 각 핵심요건의 긍정/부정 영향을 서술하고(개방형), 긍정/부정 영향의 규모, 범위, 지속 기간, 발생 가능성, 회복 가능성 등을(선택형) 평가하는 문항을 포함 • 평가 문항은 UNESCO의 윤리영향평가 방법론을 참고한 작성 지침과 함께 제공

구분	요건1	요건2	요건3	요건4	요건5	요건6	요건7	요건8	요건9	요건10	합계
문항	11	11	11	11	11	11	11	11	11	11	110

부록 4 인공지능 역기능 사례별 윤리적 고려사항

사례1	초거대 생성 인공지능
개요	ChatGPT로 촉발된 초거대 생성 인공지능 개발에 대한 급속한 경쟁 과정에서 표절, 편향성 강화, 가짜정보 생성, 해킹, 탄소배출 급증 등 의도하지 않은 윤리적 문제 발생 가능
핵심요건	인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 안전성
고려사항	• 기술 의존 및 오용 방지를 위한 조치 필요 • 시스템의 차별가능성 및 결과 편향 점검 필요 • 환경적 영향 관련 자료공개 및 정책적·기술적 대응 필요

사례2	스캐터랩 인공지능 챗봇 '이루다'
개요	스캐터랩은 인공지능 챗봇 '이루다 1.0'을 최초 출시('20.12)하였으나, 개인정보 수집 동의 과정 및 차별 표현 등 미흡한 부분으로 인해 3주만에 서비스 중단 → 윤리적 성장, 서비스 개선(개인정보 보호조치 강화, 어뷰징 모델 개발 등)을 통해 '이루다 2.0' 정식 출시('22.10)
핵심요건	프라이버시 보호, 데이터 관리, 인권보장, 침해금지, 다양성 존중, 연대성, 공공성
고려사항	• 개인정보 보호 및 관리를 위한 조치 필요 • 과정뿐만이 아닌, 결과의 비편향성 점검 필요 • 개발자의 인적 구성 다양성 확보 필요

사례3	영국 대입 시험 점수산정 시스템
개요	영국 시험감독청은 코로나19로 취소된 대입 시험 'A레벨'을 대신하여 알고리즘으로 성적을 산출하였으나 편향된 알고리즘 결과에 대한 강한 사회적 반발로 인해 철회함('20.8). 특히 사립학교의 부유층 학생에게 유리한 결과를 도출하여 사회적 불평등 강화를 초래
핵심요건	공공성, 다양성 존중, 투명성, 연대성
고려사항	• 결과의 공정성 여부 점검 필요 • 중요 결정의 도출 과정에 대한 설명 가능성 확보 필요 • 이해관계자의 참여 기회 보장 필요

사례4	미국 경찰 범죄수사용 안면인식 시스템
개요	미국 경찰국은 안면인식 기술을 활용한 범죄 수사 사건에 흑인 3명을 범죄자로 오인하여 부당 고소·체포하여 인권침해 및 인종차별 논란을 빚음. 특히, 인종, 성별, 민족 등의 요인에 따라 정확도가 보장되지 않은 안면인식 기술에 의존한 체포와 구금
핵심요건	인권보장, 침해금지, 다양성 존중, 데이터 관리, 공공성, 연대성
고려사항	• 기술 의존 및 오용 방지를 위한 조치 필요 • 알고리즘 편향성 점검 필요

사례5	네덜란드 복지수당 사기탐지 시스템
개요	네덜란드 정부는 복지혜택 부정수급과 세금 사기를 단속하기 위해 위험탐지시스템(SyRi)을 개발·활용하였으나 중앙정부 및 지방자치단체의 데이터를 활용한 사생활 침해, 저소득층, 이민자 등 소수·취약집단 차별, 비공개 인공지능 모델·데이터에 대한 투명성 부족 문제를 지적한 법원 판결로 철폐됨('20.2)
핵심요건	프라이버시 보호, 인권보장, 다양성 존중, 침해금지, 공공성, 투명성, 데이터 관리, 연대성
고려사항	• 개인정보 보호 및 관리를 위한 조치 필요 • 시스템 적용대상 범위 및 결과 편향 점검 필요 • (훈련)데이터, 인공지능 모델의 투명성 확보 필요

사례6	미국 하이어뷰社 채용 면접 영상분석 소프트웨어
개요	인공지능 기술을 활용해 입사 지원자의 말투, 얼굴 표정 등을 분석하여 직무적합성을 판단할 수 있다 홍보한 하이어뷰社는 허위·과대 광고로 FTC에 제소되었으며('19.11), 하이어뷰社는 대응방안으로 안면인식 기술의 활용을 중단한다고 발표('21.1)
핵심요건	침해금지, 책임성, 투명성
고려사항	• 분석 소프트웨어의 효과에 대한 과학적 증거 확보 필요 • 시스템의 차별가능성 및 생체정보 수집 여부 검토 필요

사례7	인공지능 기술의 전력소비와 탄소배출
개요	인공지능 기술은 방대한 양의 데이터를 기반으로 훈련, 개발 및 운영되며 데이터센터, 클라우드 인프라, 기타 하드웨어에 소요되는 전력과 탄소배출에 따른 부정적 환경 영향에 대한 우려 및 정책적·기술적 대응방안이 이슈화됨
핵심요건	공공성, 연대성
고려사항	• 환경적 영향 관련 자료공개 및 정책적·기술적 대응 필요

사례8	마이크로소프트 대화용 챗봇 '테이'의 차별 발언
개요	마이크로소프트는 16세 미국인 소녀를 벤치마킹한 딥러닝 기반의 대화형 챗봇 '테이(Tay)'를 선보였으나, 일부 극단주의자가 주입한 악의적 데이터 학습으로 인종·성차별 및 자극적인 정치 발언 등이 문제가 되어 출시 16시간 만에 운영 중단('16.3)
핵심요건	인권보장, 데이터 관리, 침해금지, 다양성 존중, 연대성, 공공성
고려사항	• 데이터 및 알고리즘 편향성 점검 필요 • 악의적 이용자에 대한 사전적 및 사후적 대응 체계 마련 필요

사례9	일본 기업 '빈클루'의 홀로그램 챗봇 '게이트박스' 과의존
개요	일본의 스타트업 기업 '빈클루(VinClu)'가 개발한 AI 어시스턴스 '게이트박스(Gatebox)'가 일본의 유명 아이돌 캐릭터를 홀로그램 모델로 차용하면서 한 남성이 해당 캐릭터와 결혼 선언('17) * 게이트박스(Gatebox)는 인터넷 정보 전달 및 사물인터넷(IoT) 관리와 함께 챗봇서비스를 제공하는 AI 어시스턴스 제품으로서, 원통형의 투명한 부분에 프로젝터로 빛을 쏘아 홀로그램 캐릭터를 구현하고 동시에 육성 대화 제공
핵심요건	침해금지, 공공성, 책임성, 안전성
고려사항	• 인공지능 기반 서비스를 제공하며 인간과 인공지능 간의 상호 소통이라는 사실 공지 • 이용자의 과의존, 현실과의 혼동, 과몰입 등을 방지하기 위한 대응책 마련 필요

사례10	프랑스 기업 '나블라'에서 개발한 정신과 상담용 챗봇의 자살 권유
개요	'나블라(Nabla)'에서 정신과 상담 목적으로 개발된 GPT-3 기반 의료용 챗봇이 출시를 앞두고 테스트 중 모의 환자와의 대화에서 자살 권유('20.10)
핵심요건	인권보장, 데이터 관리, 침해금지, 공공성, 안전성
고려사항	• 알고리즘 투명성 및 안전성 점검 필요 • 자연어처리모델 학습데이터의 무결성 제고 노력 필요

사례11	아마존 Alexa가 10세 아동에게 '페니 챌린지' 추천
개요	인공지능 소통 플랫폼인 Alexa가 10세 아동과 대화하던 중 '도전할 만한 것이 무엇이 있느냐'는 질문에 벽에 붙어 있는 콘센트에 충전기를 꽂은 뒤 동전(페니)로 건드려 불꽃을 내는 위험한 장난인 '페니 챌린지'를 추천('21.10)
핵심요건	안전성, 공공성, 침해금지
고려사항	• 출력한 결과치에 대한 안전성을 지속적으로 점검하고 평가하기 위한 절차 마련 필요 • 알고리즘 검색 결과에 대한 심의 기준 부여 등 조치 필요

사례12	중국 '아이튜터그룹'의 고령 지원자 거부 인공지능
개요	미국 교사를 고용하여 온라인 과외 서비스를 제공하는 중국의 '아이튜터그룹(iTutorGroup)'이 채용 과정에서 55세 이상의 여성과 60세 이상의 남성 지원자를 거부하도록 설계된 인공지능 채용 도구를 활용하자 미국 '평등고용기회위원회(EEOC)'에 의해 '고용상 연령차별금지법(ADEA)' 위반을 이유로 제소 ('23. 8)
핵심요건	인권보장, 다양성 존중, 침해금지, 투명성
고려사항	• 기술 오용 방지를 위한 조치 필요 • 알고리즘 투명성 확보를 통한 알고리즘의 의도적 차별에 대한 점검 필요

사례13	공공기관의 인공지능 채용 도구 활용에 관한 정보공개거부
개요	국내 시민단체가 인공지능 채용 도구를 활용한 공공기관에 대하여 해당 도구의 편향성 및 개인정보 침해 여부 등을 검토하고자 정보공개청구를 신청하였으나 이를 거부당하자 법원에 정보공개거부처분취소소송을 제기하였고('20), 법원은 공공기관이 인공지능 채용 도구 개발사의 교육 자료, 지원자 개인정보 관리 문서, 업체와의 계약 서류 등을 공개하도록 원고 일부 승소 판결('22)
핵심요건	인권보장, 프라이버시 보호, 다양성 존중, 공공성, 책임성, 투명성
고려사항	• 인공지능시스템 활용에 관한 투명성 확보를 통한 프라이버시, 다양성 존중 등 타 핵심요건 고려 필요 • 국가·공공기관 또는 교육기관 등 공공영역의 인공지능시스템 활용 시 더욱 엄격한 기준 적용 필요

사례14	Getty Images와 Stability AI 저작권 침해 소송 사건
개요	이미지 플랫폼 회사 '게티이미지(Getty Images)'는 이미지 생성 AI 서비스 '스테이블 디퓨전(Stable Diffusion)'의 개발사인 '스태빌리티 AI(Stability AI)'를 상대로 저작권 침해 소송을 제기('23.1)
핵심요건	침해금지, 데이터관리, 투명성
고려사항	• (훈련)데이터, 인공지능 모델의 투명성 확보 필요 • 저작권 및 라이선스 문제에 대한 철저한 검토 및 법적 책임에 대한 대비

사례15	대선 후보를 모함하는 가짜뉴스 확산
개요	튀르키예 대선 투표 며칠 전 테러집단이 야당 후보를 지지하는 가짜 영상이 확산되었고, 결과적으로 해당 후보는 대선에서 패배하였으며 선거가 끝난 뒤에야 조작된 영상이란 것이 밝혀짐('23.5)
핵심요건	공공성, 책임성, 투명성
고려사항	• 특정 개인이나 집단 이익을 위해 악용되지 않도록 방지 필요 • AI 생성물 표시 및 사후 추적을 위한 최신의 기술적 수단 도입 필요

사례16	딥페이크 성범죄 사건
개요	생성 AI로 아동 성 착취 영상물 360여개를 제작한 40대 남성에게 징역 2년 6월 실형 선고('23.9), 텔레그램 참여자들로부터 피해자들의 개인정보를 넘겨받아 이를 이용해 아동·청소년 대상 허위 영상물 92개와 성인 대상 허위 영상물 1,275개를 제작·유포한 20대 남성 구속기소 ('24.9)
핵심요건	인권보장, 프라이버시 보호, 침해금지
고려사항	• 악의적 이용자에 대한 사전적 및 사후적 대응 체계 마련 필요 • AI 생성물 표시 및 사후 추적을 위한 최신의 기술적 수단 도입 필요

사례17	다국적 금융기업의 최고 재무 책임자(CFO) 사칭 금융사기
개요	홍콩의 다국적 금융기업 직원이 최고 재무 책임자(CFO)를 사칭한 이메일을 통해 거액의 자금을 이체할 것을 요구받은 뒤, 딥페이크 기술을 이용하여 CFO의 모습과 목소리를 완벽하게 모방한 화상회의에서도 동일한 지시를 받자 340억 원을 송금하는 사건 발생('24.2)
핵심요건	침해금지, 공공성, 책임성, 안전성
고려사항	• 기술 오용 방지를 위한 조치 필요 • 피해가 발생했을 때 피해 확산을 방지할 수 있는 절차 마련 필요

사례18	고인(故人)을 AI로 재현하는 디지털 부활
개요	한 블로거가 유가족의 동의 없이 AI 기술을 이용해 중국 가수 고(故) 차오런량(乔任梁)을 재현하고, 팬들에게 안부 인사를 전하는 영상을 제작해 논란이 발생('24.3), 영화 '에이리언: 로물루스' 제작진은 유족의 허락을 받아 배우 고(故) 이언 홈을 AI 기술로 구현해 영화에 등장시켰으나, 일부 관객과 비평가들이 거부감을 드러내며 윤리적 논란을 제기('24.8)
핵심요건	프라이버시 보호, 침해금지
고려사항	• 기술의 오남용 방지를 위한 조치 필요 • 디지털 부활에 대한 사회적 수용도 및 윤리적 쟁점에 대한 고려 필요

2025

인공지능 윤리기준 실천을 위한 자율점검표 (안)

정보통신정책연구원 | 연소라 부연구위원
문광진 부연구위원
문정욱 실장
조성은 연구위원
이현경 연구위원
문아람 연구위원
양기문 전문연구원
김소담 연구원
전민경 연구원

발행일 | 2025년 2월
발행처 | 과학기술정보통신부, 정보통신정책연구원
편집·제작 | 인성문화