ISSUE 1

# AI 안전의 개념과 범위

## Definition and Scope of AI Safety


- **유재흥** 소프트웨어정책연구소 AI정책연구실 책임연구원 | jayoo@spri.kr
- **노재원** 인공지능안전연구소 | jwnoh@etri.re.kr
- **조지연** 인공지능안전연구소 | jy.cho@etri.re.kr

## Executive Summary

최근 인공지능(AI)의 부상과 함께 인공지능의 기능적 한계, 잠재적 위험에 대한 우려도 고조되고 있다. 이러한 상황에서 2023년 11월에는 세계 최초로 세계 정상들이 참여한 'AI 안전성 정상회의'가 개최되었고, 2024년 5월에는 후속 회의로서 'AI 서울 정상회의'도 열렸다. 2024년 8월에는 2021년부터 발의된 EU의 「인공지능법」이 발효되었다. 앞으로 믿고 쓸 수 있는 인공지능, 인간에 무해한 인공지능의 개발과 보급을 위한 정책적 노력들이 강화될 것으로 예측된다.

하지만, 정책적 대상으로서 인공지능 안전의 개념과 범위는 모호하다. 비슷한 개념으로 인공지능 신뢰성, 인공지능 윤리, 인공지능 견고성, 책임 있는 인공지능 등이 혼재되어 사용되고 있다. 따라서, 무엇보다 '인공지능 안전성'에 대한 개념에 대한 합의가 필요한 상황이며, 그 대상이 기술적 영역에 국한되는지, 기업의 책무와 윤리적 영역까지도 포함하는지 개념의 범위에 대한 검토도 필요하다. 이에 본 연구에서는 최근 화두가 되고 있는 '인공지능 안전성'을 다른 유사 개념들과 비교하여 개념을 정의하고, 그 개념이 포함하는 범위에 대해서도 조망했다. 이를 위해 현재까지 제정된 ISO/IEC의 국제표준 정의와 EU, OECD 등 국제 협력 기구에서의 정의, 그리고 최근 설립된 미국, 영국 등 AI 안전연구소들이 준용하는 개념들을 살펴보았다. 이를 토대로 향후 AI 안전성을 확보하기 위해 정책적 시사점을 정리하였다.

---

With the recent rise of artificial intelligence(AI), concerns regarding its functional limitations and potential risks have been increasing. In response, November 2023 saw the world's first AI Safety Summit with the participation of global leaders, followed by the Seoul AI Summit in May 2024. In August 2024, the EU's Artificial Intelligence Act, initially proposed in 2021, came into effect. It is anticipated that policy efforts to develop and promote trustworthy and harmless AI will be strengthened in the future.

However, the concept and scope of AI safety as a policy target remain unclear. Similar terms such as AI trustworthiness, AI ethics, AI robustness, AI reliability, and responsible AI are often used interchangeably. Thus, it is crucial to reach a consensus on the concept of 'AI safety' and to examine whether it is limited to technical aspects or includes corporate responsibility and ethical considerations. This study compares the recently prominent concept of 'AI safety' with similar terms to define its scope and meaning. To do so, we review the definitions provided by ISO/IEC international standards, international organizations like the EU and OECD, and the concepts adopted by newly established AI safety research institutes in the US and UK. Based on this analysis, the study outlines policy implications for ensuring AI safety in the future.

# I 배경

## ● AI의 부상과 함께 AI의 안전에 대한 논의도 점차 확대

- 챗GPT(ChatGPT) 등장 이후 생성AI의 역량이 검색, 문서 생성을 비롯한 다양한 영역에 접목 중
  - 인터넷 정보 서비스, SW개발, 미디어·콘텐츠를 포함 다양한 산업에서 생산성 증대 목적으로 도입 확대
    * McKinsey는 생성AI 도입 효과를 연간 2.6~4.4조 억 달러로 보고 있으며, 생성AI 글로벌 시장은 2027년 1,511억 달러로 급성장

- 이와 함께 생성AI의 기술적 한계와 잠재적 위험 요인도 노출되면서 안전성에 대한 우려도 고조
  - 할루시네이션((Hallucination), 부정확성을 비롯한 기능적 오류, 딥페이크 음란물 생성 등 AI의 악용, AI의 자율성 확대와 통제 불가능성 등 AI의 잠재적 위험 요인이 인류 생존을 위협할 것이라는 사회적 우려 부상

- 이에 학계, 산업계, 정부, 국제 사회 등 사회 전반에서 AI의 안전성 확보를 위한 논의가 전개
  - 학계에서는 유수 AI 국제학술대회에서 AI 투명성, 공정성, 안전 등 AI 신뢰성 관련 연구가 증가*
    * AI 학회(AAAI, AIES, FAccT, ICML, ICLR, NeurIPS) AI 신뢰성 논문: 347건('19)→1,094건('23)[1]
  - 산업계*에서는 자발적인 AI 안전 확보 프레임워크 개발해 운영하고, 각국 정부**들은 AI 규제안 마련 및 AI 위험 대응을 위한 국제 협력을 강화하는 추세
    * (산업계) 오픈AI, 구글, MS, 앤트로픽, 네이버 등 주요 AI 기업들은 자체적인 AI Safety Framework를 수립해 운영 중
    ** (정부) EU는 「AI법(AI Law)」을 2021년부터 추진해 발효('24.8), 미국은 행정명령(EO 14110)으로 연방 정부 차원의 AI 안전성 확보 조치 수립, 영국은 AI 안전성 정상회의('23.11.)를 최초로 주최하면서 AI 거버넌스 리더십 확보를 위한 국제 협력 강화

## ● 하지만, AI 안전(Safety) 관련 유사 개념들이 혼재되어 사용되는 상황

- 최근, AI 안전성이 주목받고 있으나 이와 유사한 AI 신뢰성, AI 책임성, AI 윤리 등 다양한 개념들이 존재하고 있고, 사용 주체와 상황에 따라 혼용되고 있는 상황
  - AI 안전에 대한 합의된 개념은 존재하지 않으며, 최근 AI 관련 국제 논의들이 전개되면서 유사한 개념(AI 신뢰성, 책임성, 공정성, 윤리 등)과의 차이를 명확히 하고 공통적 정의를 모색하는 상황

1 장진철, 노재원, 유재흥, 조지연(2024.8.), 책임 있는 AI를 위한 기업의 노력과 시사점, <SPRi 이슈리포트>

● 정책적 대상으로 AI 안전의 개념과 범위에 대한 검토 필요

- AI 안전성 확보를 위한 국제 협력이 더욱 활성화될 것으로 예상되는바, 현시점에서 AI 안전 관련 다양한 용어들과 구분된, 국제표준 수준의 AI 안전 개념에 대한 정리가 필요

- AI 안전의 개념과 개념의 범위를 명확히 함으로써 정책적 대상을 구체화하고 기존의 관련 정책과의 중복 방지와 정책 사각지대를 발굴하는 데 활용

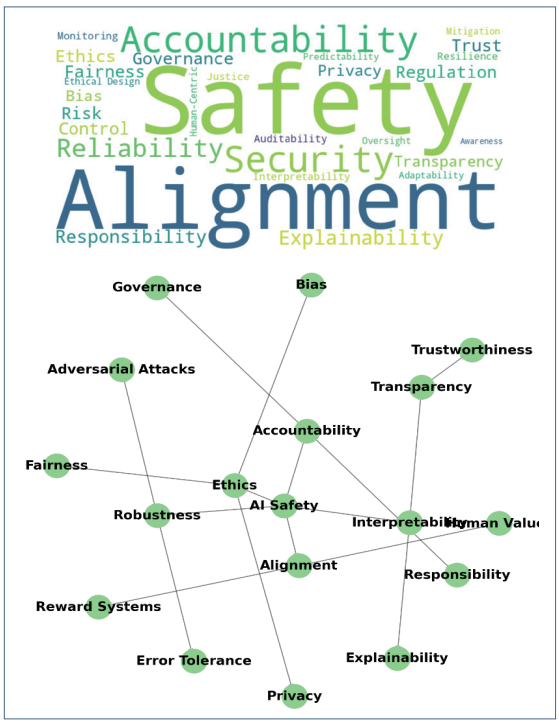
## Ⅱ AI 안전의 개념

### 1. 용어의 혼재

● AI 안전(AI Safety)은 다양한 유사 개념과 혼재되어 사용

- 최근 AI 안전에 대한 관심이 커지고 있으나 AI 안전과 관련된 유사한 개념들이 혼용되고 있어, AI 안전의 정의와 범위에 대한 정확한 인지가 어려움
  - AI 안전은 AI 신뢰성, AI 책임성, AI 윤리, AI 투명성, AI 견고성, AI 보안, AI 위험 등 다양한 유사한 개념들과 혼재되어 사용되고 있으나 각 개념 간의 차이가 존재

그림 1 - 'AI Safety' 관련 워드 클라우드 및 키워드 네트워크(예시)[2]




2 챗GPT(ChatGPT-4o)를 작성해 그림(2024.10.8.)

### ● 국제 수준의 논의를 바탕으로 AI 안전의 정의와 범위에 대한 정리 필요

- 본 보고서에서는 AI 안전 또는 AI 안전성과 유사한 다양한 용어들의 개념 차이를 구분하여, AI 안전 정책의 대상과 범위를 결정하는 데 참고할 수 있도록 하는 것을 목표로 함

- 이를 위해, 국제 표준화 단체를 통해 정의된 용어[3]와 EU, OECD, 미국의 NIST 등 인공지능 분야의 국제 논의를 주도하고 있는 국제기구 및 단체 등에서 합의된 정의를 토대로 개념을 정리[4]

## 2. AI 안전 관련 용어와 관계

### ● AI 안전과 혼용되거나 유사하게 쓰이는 개념들을 아래와 같이 도식화

- EU, OECD, 美 NIST, 및 ISO 등 국제기구 및 단체, 표준 기관에서 논의되고 정리된 정의를 중심으로 개념 간의 관계를 분석하여 아래 <그림 2> 같이 도식화하였음

- AI 안전은 AI 신뢰성, AI 윤리, AI 책무성과 주로 혼용되고 AI 위험에 대응되는 개념으로 논의되고 있음
  - AI 신뢰성(Trustworthiness)이 가장 포괄적인 개념이며, AI 신뢰성을 달성하기 위한 규범적 가이드라인으로 AI 윤리와 명문화된 제도로서 AI 법제도를 포함
  - AI 윤리(Ethic)는 AI를 개발하고 배포하는 전 과정에서 인간의 기본권, 도덕적 가치, 사회 규범, 그리고 인류 공통선에 부합하는 가치관을 가지고 만들어야 한다는 규범적 기준을 의미
  - AI 책무성(Accountability)은 특별히 AI 시스템의 개발과 배포를 수행하는 개발사 또는 사업자의 책무를 강조하는 개념으로 이러한 책무는 「AI법」을 준수하거나 AI 윤리(AI Ethics)에 바탕하고 있음
  - AI 위험(Risk)은 AI 시스템의 안전을 위협하는 기술적, 사회적 요인들을 포함하는 개념이며, 결국 AI 안전은 이러한 위험을 제거 또는 최소화하는 기술적, 제도적 수단을 확보하는 것이라 할 수 있음

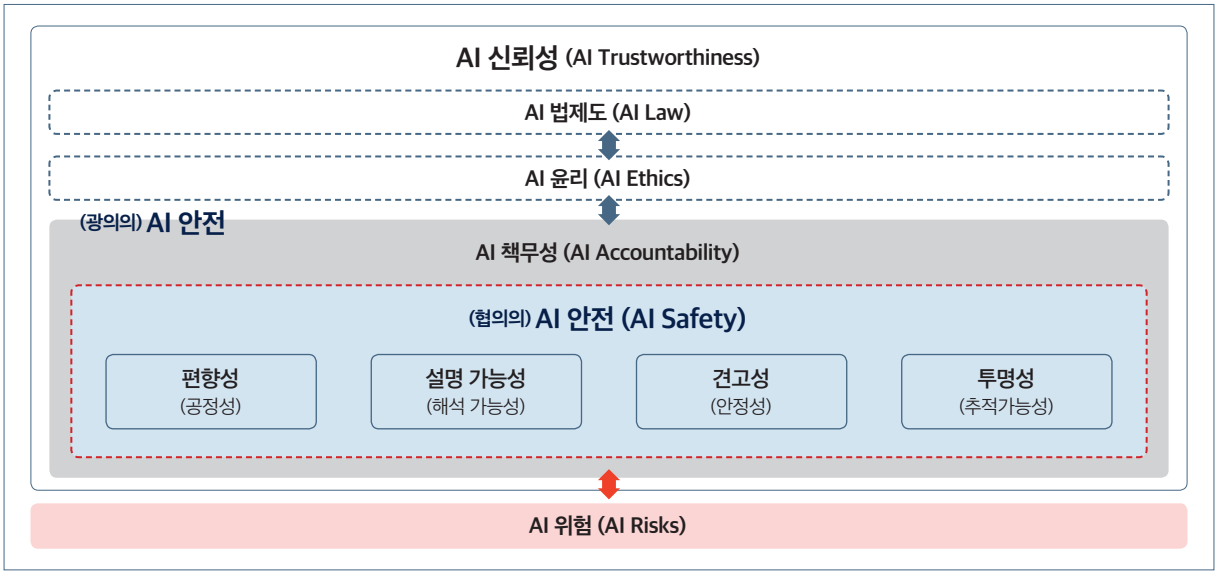
- AI 안전(Safety)은 크게 '협의의 AI 안전'과 '광의의 AI 안전'으로 구분할 수 있는 개념
  - 협의의 안전은 AI 시스템의 공정성, 설명 가능성, 견고성, 투명성을 확보하는 기술적 접근을 의미하고, 광의의 안전은 AI 시스템의 운영과 배포 과정에서 이용자 피해, 사회적 영향에 대한 사업자의 책임까지 포괄하는 개념

---

3 국제표준기구(ISO/IEC JTC1/SC42)에서 제정한 인공지능 분야 표준 문서의 용어 정의를 참고, 제정된 31건의 표준 문서 중 AI 일반 용어, AI 위험, AI 편향성 관련 문헌을 중심으로 정리('24.10. 기준)

4 주요 참고 자료로 EU AI 윤리 가이드라인(2021.4.) 및 「인공지능법('24.6.)」, OECD AI 원칙(2024), 미국 국가표준기술원(NIST)의 AI 표준 문서 관련 용어(Glossary) 데이터베이스를 참고하였음

그림 2 - AI 안전 관련 용어들의 관계




● 다음 절에서 AI 안전 관련 용어의 정의와 관련 최근의 논의에 대해 보다 구체적으로 살펴 봄

## 1) AI 신뢰성

● AI 신뢰성(AI trustworthiness)은 가장 포괄적 개념으로 준법적이고 윤리적이며 안전하고, 책임성 있는 AI를 통칭

- EU는 <신뢰할 수 있는 AI 윤리 가이드라인(2021.4.)>에서 신뢰할 수 있는 인공지능(Trustworthy AI)의 조건으로 적법성(Lawful), 윤리성(Ethics), 견고성(Robustness)을 제안하고 있음
  - AI 시스템은 구속력 있는 모든 법률을 준수하고, 구속력이 없더라도 윤리적 원칙과 가치에 부합해야 하며, 기술적·사회적 관점에서 의도치 않은 피해를 유발하지 않을 것이라는 기대에 부합해야 한다는 것이 EU의 관점(SPRi, 2021.3.[5])

- 최근 국제표준기구는 AI 신뢰성(Trustworthiness)을 '이해관계자의 기대를 검증 가능한(Verifiable) 방식으로 충족하는 능력'으로 정의(ISO/IEC TR 24028)
  - AI 신뢰성은 AI가 사용되는 상황이나 분야, 사용된 데이터와 기술, 제품 및 서비스에 따라 서로 다른 신뢰성 특성*으로 이해관계자의 기대가 충족되었음을 확인함으로써 검증 가능

5 유재흥, 추형석, 강송희(2021.3.), 유럽(EU)의 인공지능 윤리 정책 현황과 시사점: 원칙에서 실천으로, <SPRi 이슈리포트>

* AI 신뢰성 특성에는 안정성(Reliability), 가용성(Availability), 회복력(Resilience), 보안(Security), 프라이버시(Privacy), 안전성(Safety), 책임성(Accountability), 투명성(Transparency), 무결성(Integrity), 진정성(Authenticity), 품질 및 사용성 등이 포함

- 이해관계자*의 표준적 정의에 따르면 "결정이나 활동에 영향을 미치거나, 영향을 받거나, 또는 스스로 영향을 받는다고 인식할 수 있는 개인, 그룹 또는 조직"으로 정의(ISO/IEC 22989:2022, 3.5.13)

  * 세부적으로 AI 제공자(서비스 제공자, 플랫폼 제공자), AI 생산자(AI 개발자, 데이터과학자), AI 고객(서비스 사용자, 서비스 운영자), AI 파트너(데이터 공급자, 시스템 통합자), AI 영향 대상(시민단체, 데이터 영향 대상), 관계 기관(정책 입안자, 규제 기관) 등이 포함(TTA, 2023[6])

## 2) AI 윤리

**● AI 윤리(AI ethics)는 AI를 개발하고 배포하는 전 과정에서 인간의 기본권, 도덕적 가치, 사회 규범, 그리고 인류 공통선에 부합하는 가치관을 가지고 만들어야 한다는 규범적 기준**

- '윤리(Ethic)'의 사전적 정의는 '사람으로서 마땅히 행하거나 지켜야 할 도리' 또는 '도덕적 원칙들의 집합'을 의미하며, 윤리학(Ethics)은 이러한 도덕의 원리, 기원, 본질 등을 규명하는 학문을 일컬음

- AI 윤리는 "AI 기술의 개발과 사용에서 도덕적 행위를 이끌기 위해 널리 받아들여진 옳고 그름의 기준을 사용하는 가치, 원칙, 기술들의 집합"으로 정의할 수 있음(Leslie, 2019[7])
  - EU는 AI 개발과 배포 과정에서 '윤리적(Ethical)' 원칙의 준수를 권고했는데 그 원칙에는 인간의 자율성 존중, 위해(Harm) 방지, 공정성, 설명 가능성 등을 하위 요소로 포함시키고 있음(EU, 2021)

- 과학기술정보통신부는 <국가 AI 윤리기준(2020.12.)>[8]에서 AI 윤리에서 고려해야 할 3대 요소로 인간 존엄성의 원칙, 사회 공공선 원칙, 기술 합목적성의 원칙을 제시한 바 있음
  - AI 시스템의 전 생명주기에 걸쳐 △인권 보장, △프라이버시 보호, △다양성 존중, △침해 금지, △공공성, △연대성, △데이터 관리, △책임성, △안전성, △투명성 등 10대 핵심 요건을 갖춰야 할 것을 제안

- AI 윤리는 AI 시스템의 규범적 목표로서 개인, 사회, 국가의 도덕적 원칙을 의미하며, 이는 신뢰할 수 있는(Trustworthy) AI 시스템이 갖추어야 중요한 구성 요소

6 한국정보통신기술협회(TTA)(2023.12.6.), 인공지능 시스템 신뢰성 제고를 위한 요구사항, 정보통신단체표준 TTA.KO-10.1497

7 Leslie, D.(2019), Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute

8 과학기술정보통신부(2022.12.23.), 사람이 중심이 되는 인공지능 윤리 기준 마련

### 3) AI 책무성

● **AI 책무성(Accountability)은 AI 시스템 실패에 따른 영향 및 피해에 대한 개발 주체 또는 서비스 제공자의 책임을 강조하는 개념으로 광의의 AI 안전 개념의 구성 요소**

- AI 책무성(Accountability)이란 기업이 개발하거나 배포하는 AI 시스템의 위험 또는 피해에 대해 책임지도록 하는 프로세스로 정의할 수 있음(NIST[9])
  - ISO에서는 책무성을 책무가 있는 상태(State of being accountable)라고 정의하고 있으며, 여기서 책무는 어떠한 행동, 결정 및 산출물에 대해 답변할 의무가 있음을 나타냄(ISO/IEC, 2022[10])
  - 미국-EU의 무역기술위원회(TTC)의 공동 AI 용어집[11]에서도 기업의 책임을 규제나 합의에 기초하거나, 위임의 일환으로 할당된 책임(Allocated responsibility)으로 규정
  - 일반적으로 AI 윤리와 법제도를 준수하여 AI를 개발·배포하고 피해에 대한 사업자의 책임을 이행하는 것을 책임성(Responsibility), 책무성은 더 나아가 다양한 이해관계자에게 AI 시스템의 잠재적 위험과 실패, 작동 과정의 투명하게 공개까지 포함해 적극적으로 사업자의 의무를 다하는 것을 의미

- '국가 AI 윤리기준(2020)'에도 인공지능 개발 및 활용 과정에서 책임 주체를 설정함으로써 발생할 수 있는 피해를 최소화하는 노력을 핵심적 AI 윤리 요건으로 보고 있음(과학기술정보통신부, 2020)
  - AI 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 하는 것을 AI 확산의 주요 과제로 제시

### 4) AI 위험

● **AI 위험(AI Risk)은 AI의 안전이 확보되지 못했을 경우의 잠재적 위해를 의미**

- AI 위험이란 시스템이 의도된 목표를 달성하지 못했을 때의 영향을 나타내며 AI 안전의 직접적 위협 요인이 됨
  - 위험(Risk)은 일반적으로 위험 원천, 잠재적 사건, 그 결과 및 가능성 측면에서 표현(ISO/IEC, 2022[12])

9 NIST(2024.3.27.), Glossary of Terms, https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/glossary-of-terms

10 ISO/IEC 22989:2022-Information technology-Artificial intelligence-Artificial intelligence concepts and terminology, 3.5.1

11 European Commission(2023.5.), EU-U.S. Terminology and Taxonomy for Artificial Intelligence First Edition

12 ISO/IEC 24386:2022- Information technology-Artificial intelligence-Overview of ethical and societal concerns, 3.31

- 위험(Risk)은 물리적 피해, 신체 위해 또는 생명에 위협이 되는 사건·사고·결함을 의미하지만, 디지털 사회에서는 디지털 기술의 결함으로 발생하는 경제적, 정신적 피해의 범위가 매우 커 이를 포함

  ※ EU AI법에 정의된 '심각한 사고(Serious incident)'는 AI 시스템의 사고 또는 오작동으로 인해 직접적 또는 간접적으로 다음 중 하나로 이어지는 경우를 의미: ①사람의 사망 또는 심각한 건강 피해 ②중요 인프라의 관리 및 운영에 심각하고 되돌릴 수 없는 혼란 ③기본권 보호를 목적으로 하는 연합법(EU법) 하의 의무 위반 ④재산 또는 환경에 대한 심각한 피해

  - EU는 「인공지능법」은 위험을 위해(Harm)의 발생 가능성과 위해의 심각성(Severity)의 조합으로 정의
  - 오슈아 벤지오(Yoshua Bengio)의 과학보고서('24.5.)에서도 'AI 위험'을 AI 모델이나 시스템의 개발 또는 배포로 인해 발생할 수 있는 피해의 발생 확률과 해당 피해의 심각도의 조합으로 정의

- 위험(Risk)은 위해(Harm)와 위험원(Hazards)과 구분되는 개념(OECD, 2024)
  - 위해*는 위험원에 의한 사람들의 건강에 대한 손상이나 재산 또는 환경에 대한 유무형의 피해
    * 위해(Harms)는 세부적으로 물리적(Physical) 위해, 환경적(Environmental) 위해, 경제적(Economic or Financial) 위해, 명성적(Reputational) 위해, 공공이익(Public Interest) 위해, 기본권 위해, 심리적 위해로 유형화 가능
  - 위험원은 하나 이상의 AI 시스템의 개발, 사용 또는 오작동이 AI 사고*로 이어질 가능성이 있는 사건, 상황 또는 일련의 사건
    * 사고란 개인 또는 집단의 건강에 대한 상해 또는 위해, 중요 인프라의 관리 및 운영 중단, 인권 침해 또는 기본권, 노동권 및 지적재산권 보호를 위한 관련 법률 위반, 재산, 지역사회 또는 환경에 대한 피해

- AI의 위험은 크게 네 가지 측면에서 발생할 수 있으며 이에 대응 방안 마련이 필요(UK, 2024.5.)
  - 악의적 사용, 오작동, 시스템적 위험, 교차 위험으로 구분되며 요소 기술 및 프로세스의 결함 내지 오류로 발생할 수 있고, 개발자 및 서비스 제공자의 방관·무책임성에 기반하여 피해가 가중되거나 확산될 수 있음

● **다음 절에서는 AI 안전의 개념과 범위에 대해 보다 구체적으로 살펴보고자 함**

## 3. AI 안전의 개념 및 범위

● **안전(Safety)은 다양한 산업 내 존재하는 위험 특성을 반영한 개념**

- 안전의 사전적 의미는 일반적으로 '위험이 없어 피해를 입을 수 없는 상태' 또는 '마음이 편안하고 신체가 온전한 상태'를 일컫는 광의의 의미로 해석됨

- 하지만, 불완전한 기술과 도구가 다양한 영역에서 적용되면서, 특정 산업 또는 영역의 고유한 위험 요인으로부터 안전 확보를 요구하게 되는데 이는 협의의 '안전'에 해당함
  * (예) 자동차/철도/항공기 안전, 원자력 안전, 식품·의약품 안전, 화학독성물 안전, 건설 현장 안전, 농산물 안전 등

- 국제표준기구에서도 다양한 산업 내의 '안전' 확보 필요에 따라 '안전'의 개념에 대한 표준 정의를 마련
  - ISO/IEC는 일반적으로 안전을 '수용할 수 없는 위험으로부터 자유'로 정의하여 재난이나 사고의 위험으로부터 허용 가능한 수준으로 통제되어 있는 상태로 접근(ISO, 2014[13])

- AI 안전에 앞서 등장한 소프트웨어(SW)의 안전 관련 연구에서는 'SW 안전'을 협의의 개념과 광의의 개념에서 다루고 있음(SPRi, 2020)
  - 좁은 의미에서 SW 안전은 전체 시스템의 안전 보장을 위해 외부에 미치는 위험 요소를 분석하고 제거하여 SW 오류로 인한 사고를 예방하는 사전적 기술 조치에 초점을 둠
  - 반면, 광의의 SW 안전에서는 단순한 SW 오류에 의한 사고를 예방함과 동시에 사람의 신체나 생명에 위해를 미칠 수 있는 '여러 위험 요인'을 방지하고 '충분히 대비'가 되어 있는 상태로 확대하여 정의

- AI 안전 역시 내부적 위험 요인을 제거 또는 최소화에 초점을 둔 협의의 개념과, 다양한 기술적, 비기술적 위험 요인들을 고려해 그로 인해 피해에 대비한 광의의 개념으로 접근할 수 있음

● **AI 안전은 AI 시스템 자체의 오류를 막고 견고성을 보장하는 협의의 개념에서 AI 시스템의 생명주기 전 단계에서 개발자의 사회적 책임까지 포함하는 개념으로 확대**

- (협의) 일반적으로 AI 안전은 'AI 시스템이 위해가 없도록 기획, 개발, 배포될 수 있도록 위험 요소를 제거하고 안전장치를 마련함으로써 안전한 AI의 활용을 보장하는 기술적 기준'으로 정의할 수 있음

- (광의) 넓은 의미에서 AI 안전은 'AI 시스템과 성능에 영향을 미치는 다양한 위험 요소들과 AI 사용 및 배포 후 각종 부작용과 피해에 대해 기술적·제도적으로 충분히 대비된 상태'로 정의할 수 있음
  - 이는, AI 시스템의 오작동을 일으키는 기술적 요인뿐만 아니라, 사용자에 의한 악용과 AI 시스템의 활용 결과로 인한 신체적, 물리적, 경제적, 사회적 피해에 대한 책임까지 포함하는 개념임

- 최근 국제 사회에서는 AI 안전을 AI에 대한 개발자, 사업자의 책임까지를 포함하여 다양한 기술적, 비기술적 위험원에 대한 사전 예방과 대응, 사후 조치를 요구하는 광의의 개념으로 접근

13 ISO/IEC Guide 51:2014- Safety Aspects - Guidelines for their inclusion in standards, 2014

- AI 기술의 오용 및 악용과 관련된 위험 요인과 잠재적 피해를 예방하기 위해 기술적 대응을 포함한 다양한 법적·제도적 조치가 요구됨에 따라 AI 안전 개념과 논의의 범위가 확장
  * AI 안전성 정상회의와 AI 서울 정상회의를 포함해 국제 사회에서 논의되는 AI 안전은 기술적 관점의 AI 안전 확보는 물론 인류 생존을 위협할 수 있는 위험으로 떠오른 AI의 개발자 및 사업자의 책임성을 포함한 광의의 개념으로 접근

● **국제표준에서는 AI 안전을 AI 시스템이 인간의 생명, 건강, 재산 또는 환경을 위험에 빠뜨리지 않는 상태로 유지될 것이라는 기대와 관련된 기능적 개념으로 정의(ISO, 2023[14])**

- ISO의 정의에 따르면 일반적 안전의 정의에 근거해 AI 안전을 '수용할 수 없는 위험으로부터 자유'로 정의하는데, 이는 AI 위험 수준을 관리함으로써 시스템과 성능의 안전성을 확보하는 데 초점
  - 요슈아 벤지오 외(2024.5.)의 과학보고서[15]에서는 안전성(Safety)을 AI 관련 피해를 방지하거나 이로부터 보호할 때 사용할 수 있는 개념으로, AI 보안은 사이버 공격이나 AI 모델의 코드 및 가중치의 유출 등 기술적 간섭으로부터 AI 시스템을 보호하는 것으로 정의
  - 국가 AI 윤리기준(2020.12.)에도 AI 윤리의 핵심 요건 중 하나로 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있는 안전성을 제시하며, AI 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖출 것을 제언

- 협의의 AI 안전을 보장하기 위해서는 AI 개발 및 배포 전 과정에서 편향성 제거, 작동 방식 및 결과의 투명성과 설명 가능성, 시스템의 견고성 확보 등 AI 시스템의 안전한 작동 보장이 중요
  - **(편향성)** 특정 대상, 사람, 또는 그룹을 다른 것들과 비교하여 체계적으로 다르게 대우하는 것을 의미하며[16] 문화, 세대, 지역, 정치적 의견 등 맥락에 따라 달라질 수 있는 공정성(Fairness)*의 개념과 함께 언급됨
    * ISO/IEC JTC1/SC42에서는 공정성의 개념이 복잡하고 매우 맥락적이며, 논란의 여지가 있어 별도로 정의하지 않음
  - **(설명 가능성)** AI 시스템의 결과에 영향을 미치는 중요한 요소들을 사람이 이해할 수 있는 방식으로 표현하는 속성(ISO/IEC 22989[17]) 또는 시스템이 특정한 방식으로 행동한 이유와 특정한 해석을 내놓은 이유를 이해하는 기술로 정의

14 ISO/IEC 23894:2023-Information Technology-Artificial Intelligence-Guidance on risk management, 2023

15 UK Government(2024.5.), International Scientific Report on the Safety of Advanced AI: Interim Report, 요슈아 벤지오 교수가 의장으로 진행 중인 과학보고서의 중간보고서는 AI 안전성 정상회담 논의를 위한 보고서로 세계 30여 개 국가 75명의 최고 수준의 전문가가 참여

16 ISO/IEC TR 24368:2022- Information technology - Artificial intelligence - Overview of ethical and societal concerns, 2022

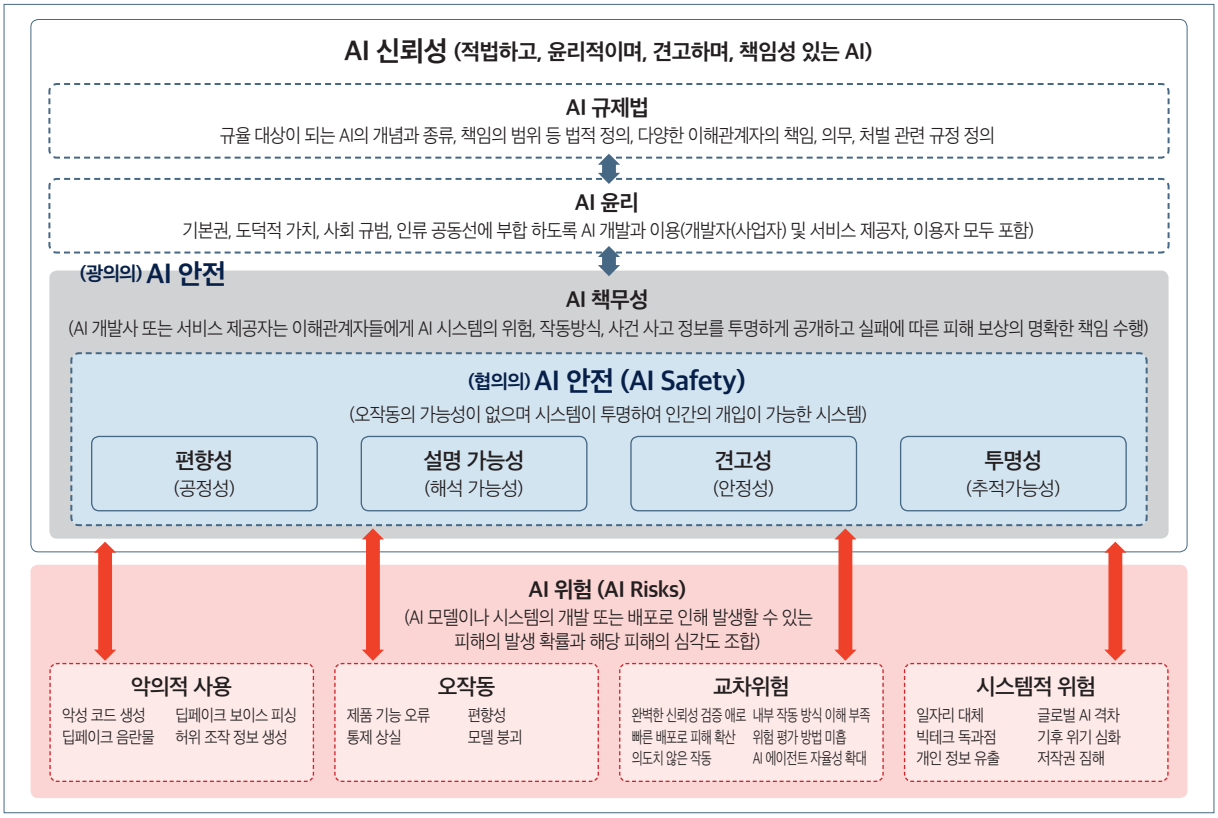
17 ISO/IEC 22989:2022-Information technology-Artificial intelligence-Artificial Intelligence concepts and Terminology, 2022

- **(투명성)** 조직적 특성 및 시스템 특성 모두를 포함하는 개념으로 조직의 활동과 결정에 관해 이해관계자에게 접근 가능하고 이해할 수 있는 방식으로 전달하고, 시스템에 대한 적절한 정보(기능, 성능, 한계, 구성 요소, 절차, 조치, 설계 목표, 가정, 데이터 소스 등)를 이해관계자에게 제공하는 능력으로 정의
  * 특히, AI 시스템의 데이터 수집·가공 및 학습 과정의 문서화를 통한 추적 가능성(Traceability)이 시스템 투명성의 중요한 요건
- **(견고성)** 어떠한 상황에서도 시스템이 성능 수준을 유지하는 능력으로 AI 시스템 또는 관련 구성 요소가 잘못된 입력이나 스트레스가 많은 환경 조건에서도 제대로 작동할 수 있는 정도와 조치 및 결과를 재현할 수 있는 능력으로 정의
  * **AI 안정성(Reliability)**은 견고성 또는 신뢰성으로 번역되기도 하나 AI 시스템에 대한 외부 공격 대응, 오류 발생 시 체계적 대처, AI 기능의 정확성과 재현성 확보, 의도치 않은 결과와 작동 오류를 최소화하는 역량으로 성능 및 서비스 제공의 안정성 의미가 강함

## 4. 소결

● 이상에서 AI 안전의 정의와 범위 관련 논의를 종합하면 아래의 <그림 3>과 같음

그림 3 - AI 안전의 개념과 범위



● **AI 안전은 신뢰할 수 있는 AI 실현의 핵심 요건으로 국제 사회의 논의는 AI 안전을 개발사, 사용자, 정부, 국제기구 등 다양한 이해관계자의 협력이 필요한 광의적 개념으로 접근**

- 이러한 의미에서 안전한 AI(Safe AI)는 '신뢰할 수 있는 AI'나 '책임 있는 AI'와는 개념상의 차이*

  * **신뢰할 수 있는 AI(Trustworthy AI)**는 가장 포괄적 개념으로 AI 윤리 및 원칙, 표준화, 국제 협력, 법규제 마련을 위해 논의할 때 주로 언급되는 최상위 수준의 개념이며, **책임 있는 AI(Responsible AI)**는 AI 윤리와 원칙을 토대로, 기술적으로 안전한 AI의 작동을 보장하고, 오류와 서비스 실패에 대한 개발자 또는 서비스 제공자의 성실한 책임성을 강조하는 개념

- AI 안전은 다양한 내외부의 AI 위험에 사전, 사후적으로 대응함으로써 안전한 AI를 실현하는 활동
  - 내부적으로 AI 오작동, 작동 방식의 투명성 제고, 통제 불능 방지, 편향성 및 기능 오류 최소화와 같은 AI 시스템의 성능에 영향을 미치는 요인들을 통제하여 AI 시스템의 견고성과 안정성을 확보하는 노력
  - 대외적으로 AI 시스템의 오용(가짜 정보 생성, 악의적 활용 등)과 예기치 못한 사회적 부작용(기후변화, 일자리 위협, 정보 유출 등)에 대해 사업자 또는 서비스 제공자의 사회적 책무를 수행

## Ⅲ 요약 및 시사

● **이 보고서에서는 AI 안전(AI Safety)과 혼재되어 사용되고 있는 여러 용어들의 개념을 비교 정리하여 정책적 대상으로 AI 안전의 개념과 범위를 보다 명확히 하고자 하였음**

- AI 안전은 '신뢰할 수 있는 AI' 실현을 위한 핵심 요건으로, 안전한 AI의 개발은 적법하고, 윤리적이며, 시스템적으로 견고하고, 개발자의 사회적 책무가 바탕이 되어야 함

● **AI 안전의 정의와 범위에 따라 정책적으로 집중할 영역이 달라질 수 있음을 시사**

- 협의의 AI 안전은 AI 시스템 자체의 오류와 잠재적 위험 요소를 제거 또는 최소화하고, 시스템의 설명 가능성, 투명성을 높이는 기술 개발에 초점을 두고 있음
  - AI 오작동, 사이버 공격으로부터 데이터 및 AI 모델 보호, AI 모델의 설명 가능성 확보 등 AI 시스템의 오류 없고 견고한 작동 방식을 보장하는 원천 기술 개발에 대한 지속적 투자 필요

- 첨단 AI 또는 프런티어 AI 모델의 잠재적 위험원, 첨단 AI에 대한 객관적 시험 및 평가 방법, AI 에이전트의 자율 증식 및 통제 가능성 등 중장기적 위험 대응 기술 개발에 선제적 대응 요구

• 반면, 광의의 AI 안전 개념에서는 다양한 AI 위험원에 대비하고, 사고의 예방과 사후 조치 관련 개발자 및 서비스 제공자의 책무성을 포함시키는 정책 방안들을 고려해야 함
- 현안으로 떠오른 생성AI가 초래하고 있는 각종 사회적 부작용(딥페이크, 허위 정보 유포, 저작권 침해 등)을 최소화하기 위한 규제 검토 및 비기술적 대응책 마련 필요
- AI가 디지털 기반 기술로 자리매김함에 따라 AI의 오류로 인한 피해가 금융, 교통, 에너지, 의료 등 사회 전반에 급속히 확산되는 이른바 시스템적 위험(Systemic risk)에 대응하기 위한 관리 체계 구축이 시급

## ● AI 안전 확보가 AI의 전면적 확산의 선결 조건으로 부상하고 있는 만큼 이해관계자들의 사회적 책무성을 바탕으로 한 AI 안전 생태계 조성 추진 필요

• AI 안전은 기술적 조치로만 완벽히 보장될 수 없으므로 제3자 검·인증 및 감리 제도 등 제도적 수단과 보험과 같은 경제적 피해 보장 체계와 연계된 생태계로 육성
- AI 생태계의 다양한 이해관계자(개발자, 사용자, 시민단체, 규제 기관, 정부 등) 간의 '허용할 수 있는 위험' 수준과 논의와 사고 시 책임 소재의 명확화를 위해 사회적 논의를 진전시키는 노력 필요
- 공공과 산업 전반의 안전한 AI 전환(AX)을 위해 체계적 계획 수립과 지원 방안 마련

• 인공지능의 안전 확보를 위한 국제 협력의 지속적 추진으로 글로벌 경쟁력 강화 필요
- AI 안전 관련 국제 협의체, 국제표준 활동에 참여함으로써 글로벌 수준에서 정립된 AI 안전 관련 개념들을 올바로 이해하고 이를 바탕으로 국제 협력, 국내 제도 마련, 정책 방안을 구체화
- 글로벌 AI 안전연구소를 중심으로 AI 안전 관련 정보 공유, 인력 교류, 인프라 공유, 및 표준화 논의 추진

## 참고 : AI 안전·신뢰성 용어 정의

### 표 - ISO/IEC AI 안전·신뢰성 관련 표준 용어

| 용어 | 정의 | 출처 |
|---|---|---|
| AI 시스템<br>AI system | AI 시스템은 명시적 또는 암시적인 목표를 위해 받은 입력으로부터 예측, 콘텐츠, 추천 또는 물리적 또는 가상 환경에 영향을 미칠 수 있는 결정을 생성하는 방법을 추론하는 기계 기반 시스템으로 다양한 AI 시스템은 배포 후 자율성 및 적응성의 수준에서 차이가 있음<br>An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment(OECD, 2024.5. 개정)[18] | OECD, 2024.5. |
| 편향<br>Bias | 특정 대상, 사람, 또는 그룹을 다른 것들과 비교하여 체계적으로 다르게 대우하는 것[19]<br>Systematic difference in treatment of certain objects, people, or groups in comparison to others | ISO/IEC 22989(3.5.4) |
| 위해<br>Harm | 사람들의 건강에 대한 손상이나 재산 또는 환경에 대한 손상<br>Injury or damage to the health of people or damage to property or the environment | ISO/IEC TR 24028(3.17) |
| 위험원<br>Hazard | 잠재적인 위해의 원천<br>Potential source of harm | ISO/IEC TR 24028(3.18) |
| 위험<br>Risk | 목표에 대한 불확실성의 영향 Effect of uncertainty on objectives<br>• 영향은 예상과의 편차일 수 있으며, 긍정적, 부정적 또는 둘 다일 수 있고, 기회와 위협을 다루거나, 만들거나, 결과로 이어질 수 있음<br>• 목표는 다양한 측면과 범주를 가질 수 있으며, 다양한 수준에서 적용될 수 있음<br>• 위험은 일반적으로 위험 원천, 잠재적 사건, 그 결과 및 가능성 측면에서 표현 | ISO/IEC 22989(3.5.11)<br>ISO/IEC TR 24028(3.31) |
| | 위해(Harm)의 발생 가능성과 위해의 심각성(Severity)의 조합으로 정의함<br>'Risk' means the combination of the probability of an occurrence of harm and the severity of that harm | EU AI Act('24.6.)<br>UK Gov't('24.5.) |
| 안전성<br>Safety | 수용할 수 없는 위험으로부터의 자유<br>Freedom from risk which is not tolerable | ISO/IEC TR 24028(3.34)<br>ISO/IEC 23894(A.10) p20 |
| 보안성<br>Security | 제품 또는 시스템이 정보와 데이터를 보호하는 정도로, 사람들이나 다른 제품 또는 시스템이 자신들의 유형과 권한 수준에 맞는 데이터 접근 정도를 갖도록 함<br>Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization | ISO/IEC TR 24028(3.35) |
| 위협<br>Threat | 시스템, 조직 또는 개인에게 손상을 초래할 수 있는 원하지 않는 사건의 잠재적 원인<br>Potential cause of an unwanted incident, which may result in harm to systems, organizations or individuals | ISO/IEC TR 24028(3.39) |
| 신뢰<br>Trust | 사용자 또는 다른 이해관계자가 제품 또는 시스템이 의도한 대로 작동할 것이라고 확신하는 정도<br>Degree to which a user or other stakeholder has confidence that a product or system will behave as intended | ISO/IEC TR 24028(3.41) |
| 신뢰성<br>Trustworthiness | 이해관계자의 기대를 검증 가능한 방식으로 충족하는 능력<br>Ability to meet stakeholders' expectations in a verifiable way | ISO/IEC TR 24028(3.42) |
| 취약성<br>Vulnerability | 하나 이상의 위협(Threats)에 의해 악용될 수 있는 자산(Assets) 또는 통제(Control)의 취약점<br>Weakness of an asset or control that can be exploited by one or more threats | ISO/IEC TR 24028(3.48) |

| 용어 | 정의 | 출처 |
|---|---|---|
| 설명 가능성<br>Explainability | AI 시스템의 결과에 영향을 미치는 중요한 요소들을 사람이 이해할 수 있는 방식으로 표현하는 속성<br>Explainability: property of an AI system to express important factors influencing the AI system results in a way that humans can understand(ISO/IEC 22989) | ISO/IEC 22989<br>ISO/IEC 23894(A.12) p20 |
| 공정성<br>Fairness | ISO/IEC에서는 공정성의 개념이 복잡하고 매우 맥락적이며, 논란의 여지가 있어 별도 정의를 내리지 않음<br>다만, 편향의 개념과는 다르며 문화, 세대, 지역, 정치적 의견 등 맥락에 따라 달라질 수 있음<br>Fairness is a concept that is distinct from, but related to bias. Fairness can be characterized by the effects of an AI system on individuals, groups of people, organizations and societies that the system influences. However, it is not possible to guarantee universal fairness. Fairness as a concept is complex, highly contextual and sometimes contested, varying across cultures, generations, geographies and political opinions. What is considered fair can be inconsistent across these contexts. This document thus does not define the term fairness because of its highly socially and ethically contextual nature.(ISO/IEC 24027) | EU 2021[20]<br>ISO/IEC 24027 p5 |
| 투명성<br>Transparency | 조직적 및 시스템 특성 모두를 포함하는 개념으로 조직의 활동과 결정과 관련해 이해관계자에게 접근 가능하고 이해할 수 있는 방식으로 전달, 시스템에 대한 적절한 정보(기능, 성능, 한계, 구성 요소, 절차, 조치, 설계 목표, 가정, 데이터 소스, 등)를 이해관계자에게 제공하는 시스템 특성<br><Organization> property of an organization that appropriate activities and decisions are communicated to relevant stakeholders. in a comprehensive, accessible and understandable manner, <System> property of a system that appropriate information about the system is made available to relevant stakeholders | ISO/IEC 22989(3.5.14)<br>ISO/IEC 23894(A.12) p20 |
| 안정성<br>Reliability | 일관된 의도된 행동과 결과의 속성<br>Property of consistent intended behaviour and results | ISO/IEC TR 24028(3.30) |
| 견고성<br>Robustness | 어떠한 상황에서도 시스템이 성능 수준을 유지하는 능력으로 AI 시스템 또는 관련 구성 요소가 잘못된 입력이나 스트레스가 많은 환경 조건에서도 올바르게 작동할 수 있는 정도와, 조치 및 결과를 재현할 수 있는 능력<br>Ability of a system to maintain its level of performance under any circumstances | ISO/IEC 22989(3.5.12)<br>ISO/IEC 23894(A.8) p19 |
| 책무성<br>Accountability | 책무성의 조직적 특성 및 시스템 특성을 모두 포함하며, 조직의 책무성은 조지의 결정과 행동에 대해 설명하고 이를 관리 기관, 법적 당국, 이해관계자에게 책임지는 것을 의미, 시스템 책무성은 엔터티(구체적이거나 추상적인 관심의 대상)의 결정과 행동을 추적할 수 있는 능력<br>Accountability refers both to a characteristic of organizations and to a system property:<br>— Organizational accountability means that an organization takes responsibility for its decisions and actions by explaining them and being answerable for them to the governing body, to legal authorities and more broadly to stakeholders.<br>— System accountability relates to being able to trace the decisions and actions of an entity to that entity(ISO/IEC 23894) | EU 2021<br>ISO/IEC 23894(A.2) p18 |

출처: EU(2021.4.), OECD(2024), ISO/IEC 표준 문서 참고하여 저자 작성

18 OECD(2024.5.), Recommendation of the Council on Artificial Intelligence

19 ISO/IEC TR 24368:2022- Information technology - Artificial intelligence - Overview of ethical and societal concerns, 2022

20 EU(2021.4.), Ethics Guidelines for Trustworthy AI

## 참고문헌

### 1. 국외문헌


- European Commission(2023.5.), EU-U.S. Terminology and Taxonomy for Artificial Intelligence First Edition
- EU(2021.4.), Ethics Guidelines for Trustworthy AI
- ISO/IEC Guide 51:2014- Safety Aspects - Guidelines for their inclusion in standards, 2014
- ISO/IEC TR 24368:2022- Information technology - Artificial intelligence - Overview of ethical and societal concerns, 2022
- ISO/IEC 23894:2023- Information Technology-Artificial Intelligence-Guidance on risk management, 2023
- ISO/IEC 22989:2022- Information technology - Artificial intelligence - Artificial Intelligence Concepts and Terminology, 2022
- ISO/IEC TR 24368:2022- Information technology - Artificial intelligence - Overview of ethical and societal concerns, 2022
- Leslie, D.(2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute
- NIST(2024.3.27.), Glossary of Terms
- OECD(2024.5.), Defining AI Incidents And Related Terms
- OECD(2024.5.), Recommendation of the Council on Artificial Intelligence
- UK Government(2024.5.), International Scientific Report on the Safety of Advanced AI: Interim Report


### 2. 국내문헌


- 권영환, 진회승, 송지환(2020.1.), 소프트웨어 안전 관리 프레임워크 연구, <SPRi 연구보고서>
- 유재흥, 추형석, 강송희(2021.3.), 유럽(EU)의 인공지능 윤리 정책 현황과 시사점: 원칙에서 실천으로, <SPRi 이슈리포트>
- 과학기술정보통신부(2022.12.23.), 사람이 중심이 되는 인공지능 윤리 기준 마련
- 장진철, 노재원, 유재흥, 조지연(2024.8.), 책임 있는 AI를 위한 기업의 노력과 시사점, <SPRi 이슈리포트>
- 한국정보통신기술협회(TTA)(2023.12.6.), 인공지능 시스템 신뢰성 제고를 위한 요구사항, 정보통신단체표준 TTA.KO-10.1497