



# AI 품질의 국제 표준화 동향

(테스팅 · 신뢰성 · 기능 안전 중심으로)

2024. 06. 12

조영임

*yicho@gachon.ac.kr*

# 목차

1. AI 국제표준화 소개
2. AI 테스트·신뢰성·기능 안전 국제표준화 동향
3. 시사점



# 목차



1. AI 국제표준화 소개

2. AI 테스트·신뢰성·기능안전 국제표준화 동향

3. 시사점

# ISO/IEC JTC 1/SC 42 Artificial Intelligence

## ◆ SC 42 Artificial Intelligence

- 대표적 AI Public Standardization Organization (2017. 11 설립 )
- 2024. 5월 기준 64개국 참여, 60여개 표준 개발 중이며 미국 ANSI 간사국
- 기술적 능력과 비기술적 요구사항을 고려한 전체적 AI 생태계 관점에서 표준 개발 중

ISO/IEC JTC 1  
**ISO/IEC JTC 1/SC 42**  
Artificial intelligence



**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO)  
EUROPEAN COMMITTEE FOR STANDARDIZATION (CEN)**

**GUIDELINES FOR THE IMPLEMENTATION OF THE AGREEMENT ON  
TECHNICAL COOPERATION BETWEEN ISO AND CEN  
(VIENNA AGREEMENT)**

Vienna Agreement

28	33	38	26
Published ISO standards *	ISO standards under development *	Participating members	Observing members
* number includes updates			
(2024.5.30 기준)			
Structure Liaisons Meetings			
Reference ↑	Title		Type
ISO/IEC JTC 1/SC 42/AHG 4 ①	Liaison with SC 27		Working group
ISO/IEC JTC 1/SC 42/AHG 7 ①	JTC1 joint development review		Working group
ISO/IEC JTC 1/SC 42/JAG ①	Joint Advisory Group on AI and sustainability with ISO/IEC JTC1/SC 39 and JTC1/SC 42		Working group
ISO/IEC JTC 1/SC 42/JWG 2 ①	Joint Working Group ISO/IEC JTC1/SC 42 - ISO/IEC JTC1/SC 7 : Testing of AI-based systems		Working group
ISO/IEC JTC 1/SC 42/JWG 3 ①	Joint Working Group ISO/IEC JTC1/SC42 - ISO/TC 215 WG : AI enabled health informatics		Working group
ISO/IEC JTC 1/SC 42/JWG 4 ①	Joint Working Group ISO/IEC JTC1/SC42 - IEC TC65/SC65A: Functional safety and AI systems		Working group
ISO/IEC JTC 1/SC 42/JWG 5 ①	Joint Working Group ISO/IEC JTC1/SC42 - ISO/TC 37 WG: Natural language processing		Working group
ISO/IEC JTC 1/SC 42/WG 1 ①	Foundational standards		Working group
ISO/IEC JTC 1/SC 42/WG 2 ①	Data		Working group
ISO/IEC JTC 1/SC 42/WG 3 ①	Trustworthiness		Working group
ISO/IEC JTC 1/SC 42/WG 4 ①	Use cases and applications		Working group
ISO/IEC JTC 1/SC 42/WG 5 ①	Computational approaches and computational characteristics of AI systems		Working group

1991.6.27 유럽표준의 국제표준으로의 채택을 목적으로 ISO와 CEN(유럽표준화위원회) 간 체결된 협정으로 유럽표준의 국제표준으로의 채택 가능성 높음

# 목차

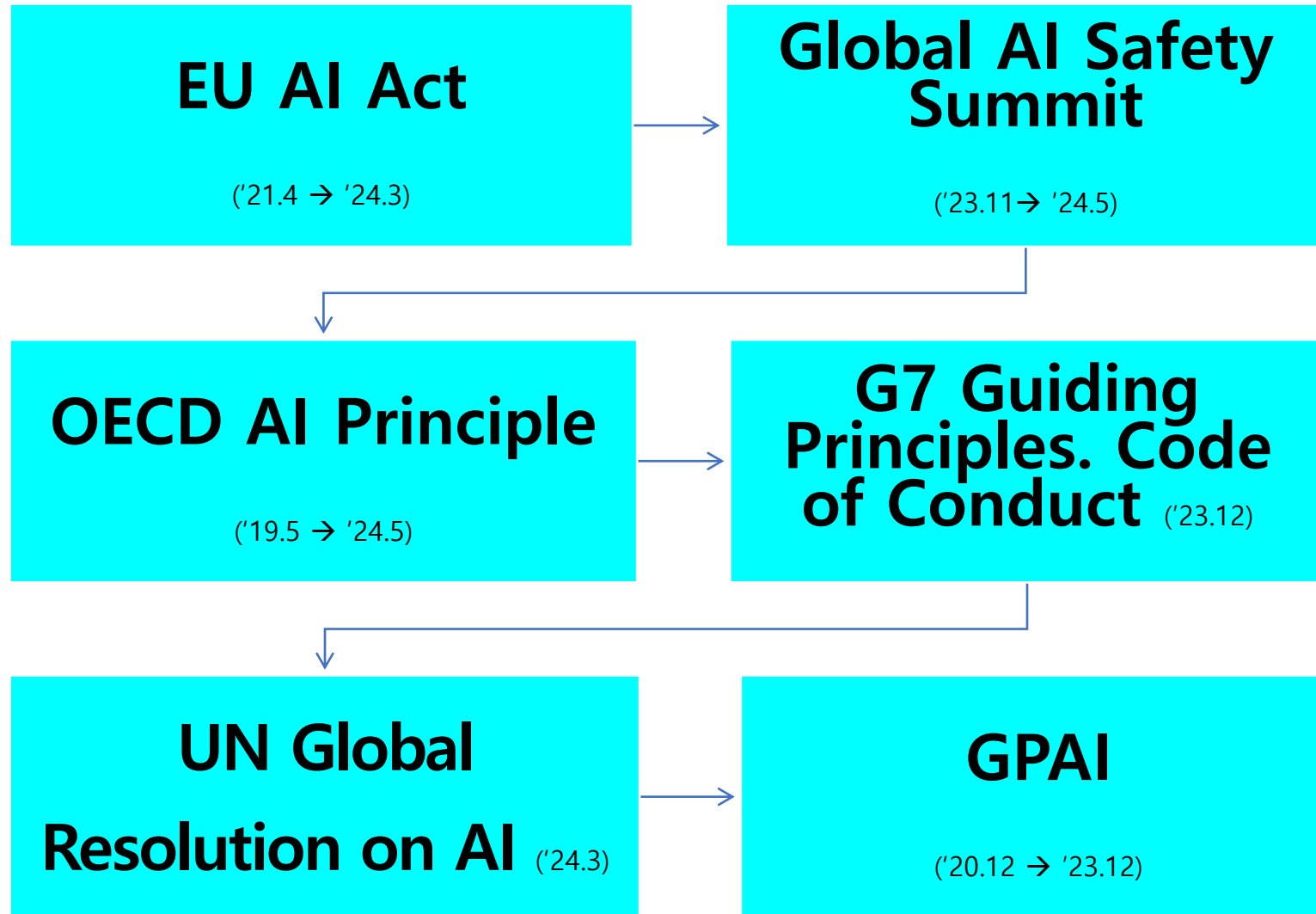


1. AI 국제표준화 소개

2. AI 테스트·신뢰성·기능안전 국제표준화 동향

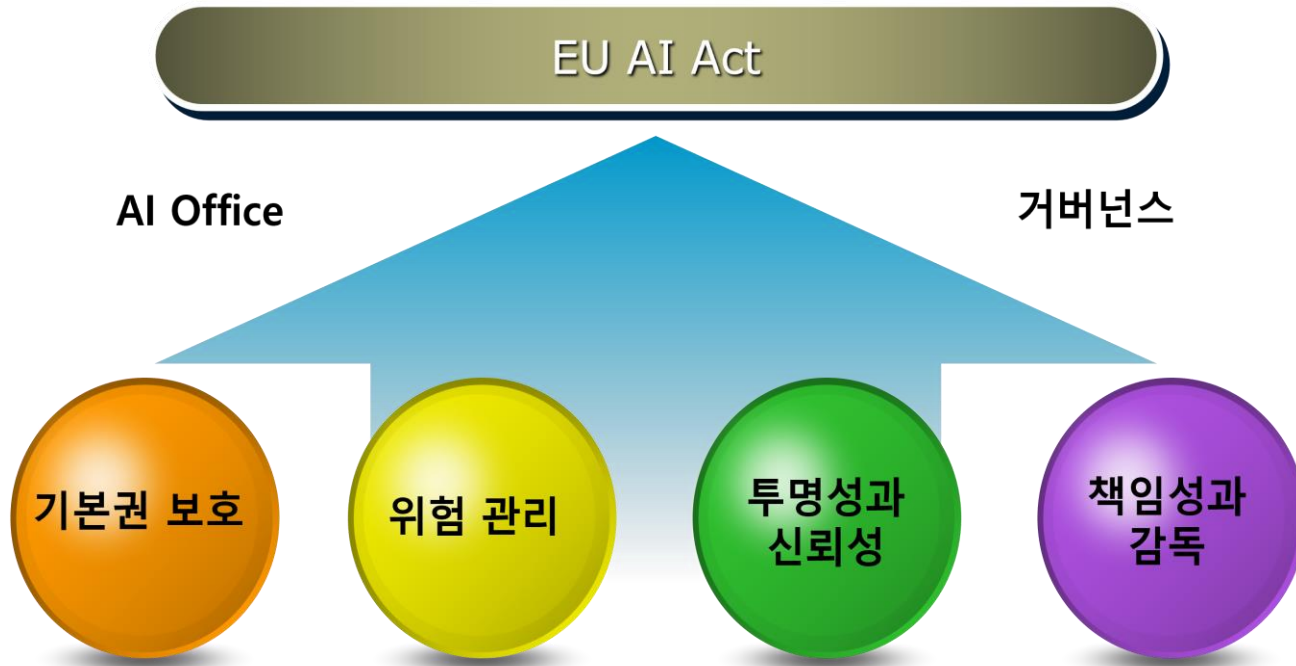
3. 시사점

# 영향력있는 AI 주요 원칙들

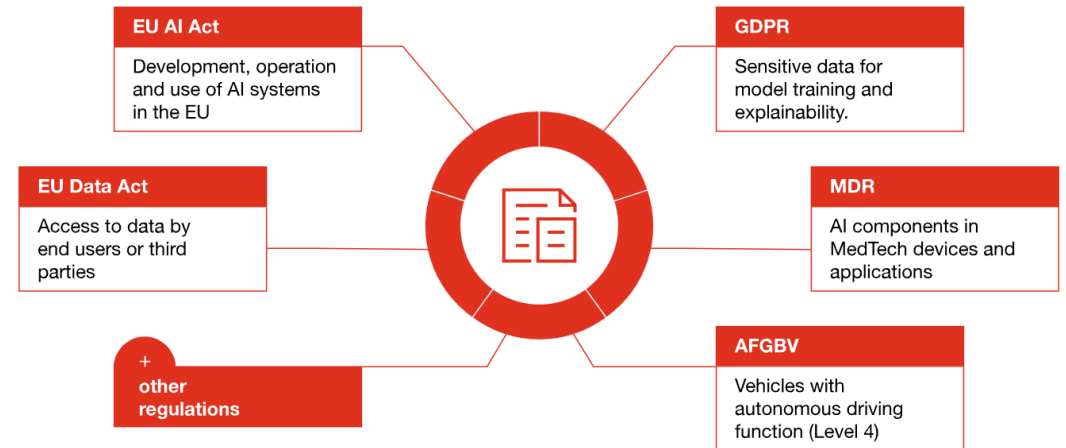


# EU AI Act-EU concept

- 유럽연합이 인공지능(AI) 기술의 윤리적이고 안전한 사용을 보장하기 위해 제정한 규제로, 위험 기반 접근 방식을 통해 AI 시스템의 위험 수준에 따라 규제를 차별화하려는 법안
- 2021.4. EU 집행부 발표 → 2023.6.14 개정 → 2024.3.13 European Parliament와 EU council 승인으로 제정 완료
- EU내 출시하려는 제공자 (장소무관) 및 모든 사용자가 적용대상



In order to be successful in the use of AI in the competitive European AI market, regulated organisations will need to comply with a wide range of regulations, build a culture of trust and ensure transparency throughout the AI lifecycle.



Source: PwC "European AI regulation and its implementation"

# OECD AI-기본원칙



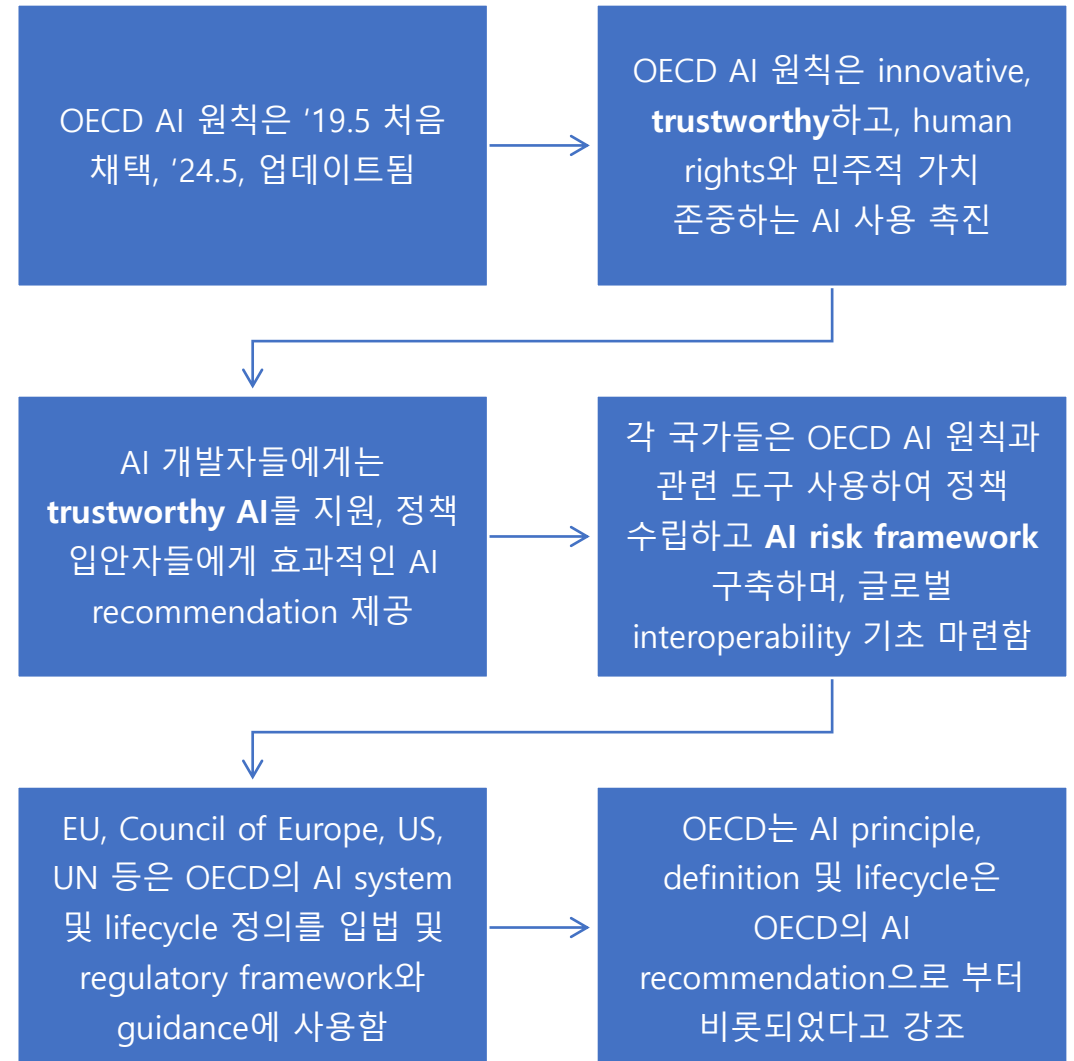
## Values-based principles

	Inclusive growth, sustainable development and well-being >
	Human rights and democratic values, including fairness and privacy >
	Transparency and explainability >
	Robustness, security and safety >
	Accountability >

## Recommendations for policy makers

	Investing in AI research and development >
	Fostering an inclusive AI-enabling ecosystem >
	Shaping an enabling interoperable governance and policy environment for AI >
	Building human capacity and preparing for labour market transition >
	International co-operation for trustworthy AI >

(source: <https://oecd.ai/en/ai-principles>)





# Global AI Safety Summit

## ◆ 영국 AI SAFETY SUMMIT

- AI Safety Summit (2023.11): 28개국 참여



Home > [Business and industry](#) > [Science and innovation](#)  
> [Artificial intelligence](#)  
> [AI Safety Summit 2023: The Bletchley Declaration](#)

[Department for  
Science,  
Innovation  
& Technology](#)

[Foreign,  
Commonwealth  
& Development  
Office](#)

[Prime Minister's  
Office, 10 Downing  
Street](#)

Policy paper

## The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023

Published 1 November 2023



인공지능의 잠재력 실현을 위해 인간중심적, 신뢰할 수 있으며 책임감있는 방식으로 설계, 개발, 배포, 사용되어야 함

포용적인 방식으로 인권을 향유하고 UN 지속가능 발전 목표달성을 위해 노력해야 함

인공지능의 잠재적 영향을 검토하고 대응하기 위한 포럼, 이니셔티브를 위한 노력이 필요하며 국제사회가 인권, 투명성, 설명가능성, 공정성, 책임성, 규제, 안전성, 인간감독, 윤리, 편견완화, 프라이버시 및 데이터 보호를 다루어 주어야 함

사이버보안 및 생명공학, 가짜정보와 같은 위험 속에서 발생할 수 있는 문제점을 해결해야 함

인공지능이 발생할 수 있는 문제점을 국제사회가 공동으로 해결해야 하므로 협력을 촉진함

모든 주체는 인공지능 역량강화, 격차해소(개발도상국지원) 등 개발지향적 접근과 정책이 필요함

인공지능 생애주기 전반에 걸쳐 안전 고려해야 하며, 특히 인공지능개발주체들은 안전시험시스템, 평가 및 조치 등을 통해 인공지능시스템의 안전보장 책임이 있음

인공지능 잠재력 인식하고 국제사회 협력을 보장하기 위해 국제포럼, 이니셔티브참여, 광범위한 논의에 개발적, 포용적, 기술 이익에 대한 책임감있는 활동 등 인공지능 안전에 대한 연구 지속해야 함

# AI System 정의

## ◆ EU AI Act에서 '품질'과 관련된 정의부분

### • CHAPTER I GENERAL PROVISIONS

#### Article 3

#### Definitions

For the purposes of this Regulation, the following definitions apply:

- (1) *'AI system' means a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and 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;*
- (47) *'AI Office' means the Commission's function of contributing to the implementation, monitoring and supervision of AI systems and AI governance carried out by the European Artificial Intelligence Office established by Commission Decision of 24.1.2024; references in this Regulation to the AI Office shall be construed as references to the Commission;*

- (57) *'testing in real-world conditions' means the temporary testing of an AI system for its intended purpose in real-world conditions outside a laboratory or otherwise simulated environment, with a view to gathering reliable and robust data and to assessing and verifying the conformity of the AI system with the requirements of this Regulation and it is not considered to be placing the AI system on the market or putting it into service within the meaning of this Regulation, provided that all the conditions laid down in Article 57 or 60 are fulfilled;*



'실제조건에서의 시험'이란 실험실 외나 기타 시뮬레이션 환경에서 AI 시스템이 의도된 목적을 수행하는지에 관한 일시적인 테스트를 의미

신뢰할 수 있고 견고한 데이터를 수집하고, 요구 사항과의 일치성을 평가하고 확인하기 위한 목적

AI 시스템을 시장에 공급하거나 서비스로 제공하기 위해 조건을 모두 충족해야 한다는 규정을 의미

- (66) *'general-purpose AI system' means an AI system which is based on a general-purpose AI model, that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems;*



'범용 AI 시스템'이란 범용 AI 모델을 기반으로 하여 직접 사용하거나 다른 AI 시스템에 통합될 수 있는 다양한 목적을 수행할 수 있는 능력을 가진 AI 시스템을 의미

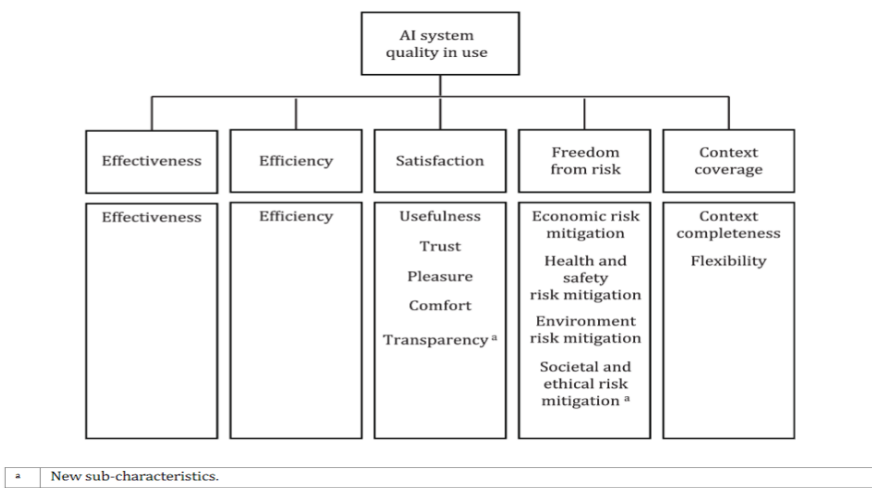
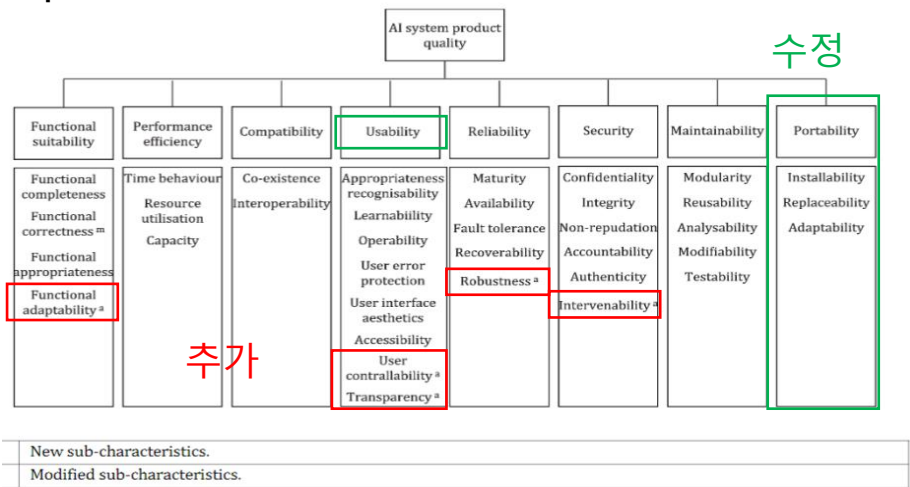
# Software vs. Machine Learning

## ◆ 비교

비교	Software	Machine Learning
작동방식	<ul style="list-style-type: none"><li>- 고정된 규칙과 로직</li><li>- 결정론적 방식</li></ul>	<ul style="list-style-type: none"><li>- 데이터 기반 학습</li><li>- 비결정론적 방식</li></ul>
개발접근방식	<ul style="list-style-type: none"><li>- 명시적 프로그래밍</li><li>- 디버깅 및 테스트</li><li>- 유닛, 통합, 시스템, 회귀 테스트 등</li></ul>	<ul style="list-style-type: none"><li>- 모델 훈련</li><li>- 검증 (교차검증) 및 평가 메트릭스 (정확도, 정밀도, 재현율, F1 스코어 등)</li></ul>
품질평가방식	<ul style="list-style-type: none"><li>- 기능적 정확성</li><li>- 신뢰성 및 성능</li><li>- 보안 및 사용성</li></ul>	<ul style="list-style-type: none"><li>- 정확도 및 일반화 능력, 예측 성능</li><li>- 편향 및 분산정도</li><li>- 설명 가능성</li></ul>
적용분야	<ul style="list-style-type: none"><li>- 일반적 비즈니스 분야</li><li>- 임베디드 시스템</li><li>- 제어 시스템 등</li></ul>	<ul style="list-style-type: none"><li>- 데이터 분석 및 예측</li><li>- 컴퓨터 비전 및 자연어처리 분야</li><li>- 자율 시스템 등</li></ul>
평가 도구 및 지표	<ul style="list-style-type: none"><li>- 코드 리뷰</li><li>- 정적 분석 도구</li><li>- 성능 모니터링 도구 등</li></ul>	<ul style="list-style-type: none"><li>- 모델 평가도구</li><li>- 모델의 예측을 해석하고 설명하는 도구</li><li>- 실험 관리 도구 등</li></ul>

# AI System 품질

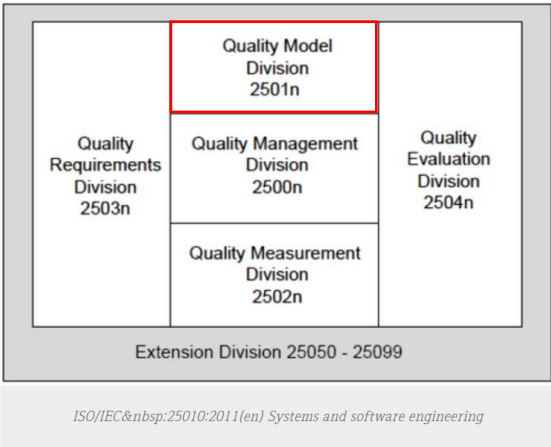
## ◆ ISO/IEC TS 25059:2023 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems



(ISO/IEC 25010: 2023) - SQuaRE

변화

SOFTWARE PRODUCT QUALITY								
FUNCTIONAL SUITABILITY	PERFORMANCE EFFICIENCY	COMPATIBILITY	INTERACTION CAPABILITY	RELIABILITY	SECURITY	MAINTAINABILITY	FLEXIBILITY	SAFETY
FUNCTIONAL COMPLETENESS	TIME BEHAVIOUR	CO-EXISTENCE	APPROPRIATENESS	FAULTLESSNESS	CONFIDENTIALITY	MODULARITY	ADAPTABILITY	OPERATIONAL CONSTRAINT
FUNCTIONAL CORRECTNESS	RESOURCE UTILIZATION	INTEROPERABILITY	RECOGNIZABILITY	AVAILABILITY	INTEGRITY	REUSABILITY	SCALABILITY	RISK IDENTIFICATION
FUNCTIONAL APPROPRIATENESS	CAPACITY		LEARNABILITY	FAULT TOLERANCE	NON-REPUDIATION	ANALYSABILITY	INSTALLABILITY	FAIL SAFE
			OPERABILITY	RECOVERABILITY	ACCOUNTABILITY	MODIFIABILITY	REPLACEABILITY	HAZARD WARNING
			USER ERROR PROTECTION		AUTHENTICITY	TESTABILITY		SAFE INTEGRATION
			USER ENGAGEMENT		RESISTANCE			
			INCLUSIVITY					
			USER ASSISTANCE					
			SELF-DESCRIPTIVENESS					





# Data 품질

## ◆ ISO/IEC FDIS 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

(ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model)

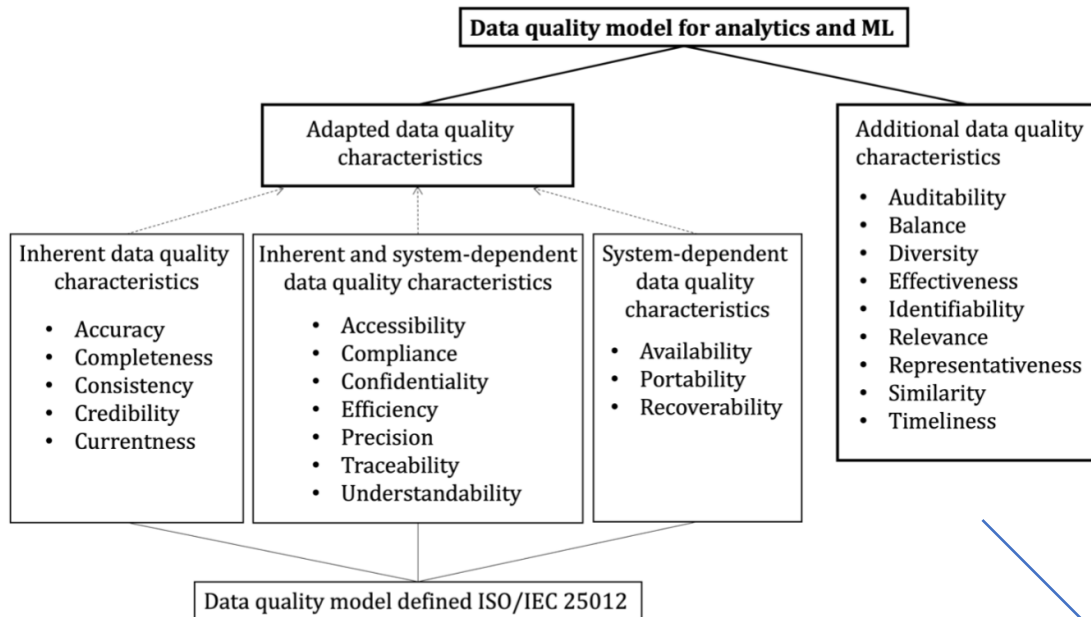
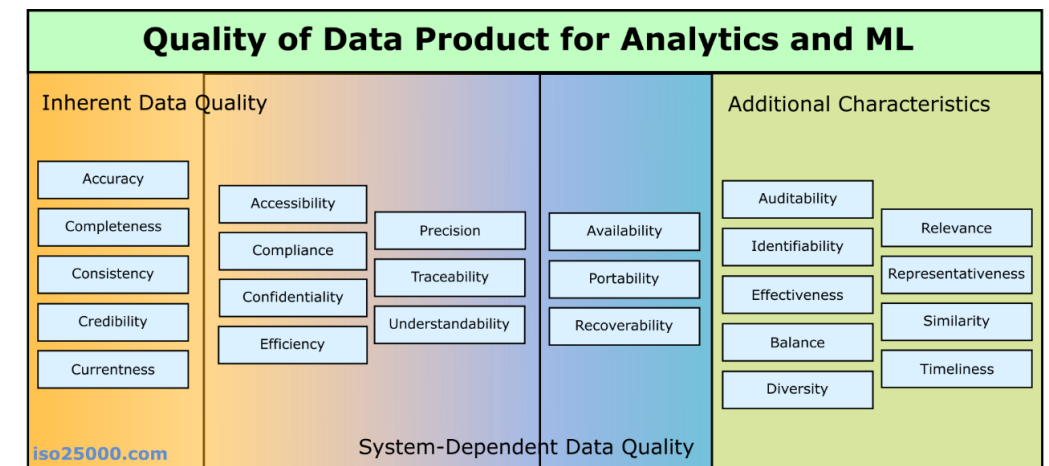
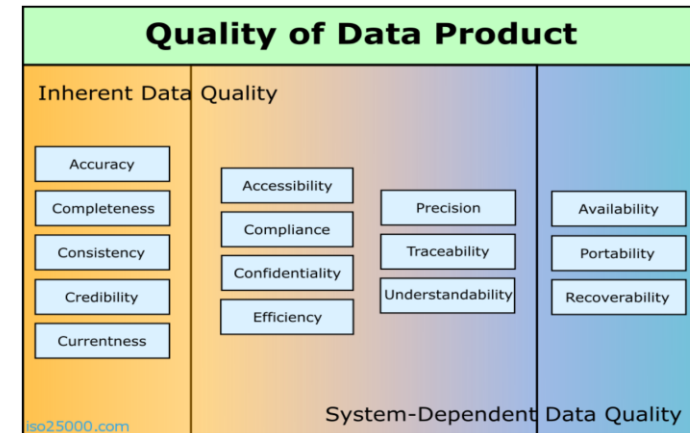


Figure 3 — Data quality characteristics for analytics and ML



## ◆ SC 42 중심의 AI 테스트 관련 국제표준

- 목적: AI 모델이 정확하고 효율적으로 동작하는지 확인하기 위한 중요한 과정
- 대표적 표준
  - ISO/IEC AWI TS 29119-11 Software and systems engineering — Software testing — Part 11: Testing of AI systems
    - 인공지능 시스템에 적용할 수 있는 테스트 기술들 개발
  - ISO/IEC AWI 23282 Information technology — Artificial Intelligence — Evaluation methods for accurate natural language processing systems
    - 자연어 처리시스템의 평가방법론 개발
  - ISO/IEC AWI TS 17847 Information technology — Artificial Intelligence — Verification and validation analysis of AI systems
    - 평가에 관한 성능검증방법 개발
  - ISO/IEC TS 4213:2022 Information technology — Artificial intelligence — Assessment of machine learning classification performance
    - 머신러닝 모델, 시스템, 알고리즘의 분류 성능을 측정하기 위한 방법론 개발

## ◆ SC 42 중심의 AI 테스트 관련 국제표준

- ISO/IEC TS 24668:2024 Information technology — Artificial intelligence — Process management framework for big data analytics
  - 산업 전반에서 빅 데이터 분석을 효과적으로 활용할 수 있는 프로세스를 개발하기 위한 프레임워크 제공
- ISO/IEC TS 25058:2024 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
  - AI 시스템 품질 모델을 사용하여 인공지능 (AI) 시스템을 평가하는 지침 제공
- **ISO/IEC TS 25059:2023** Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
  - AI 시스템을 평가하기 위한 품질 모델 제공하며, SQuaRE 표준에 대한 응용 프로그램 특정 확장
  - AI 시스템에 관련된 일부 품질 특성에는 정확성, 해석 가능성, 견고성, 공정성, 개인 정보 보호 및 보안 포함
  - AI 모델의 정확성, 신뢰성 및 견고성을 테스트하고, 윤리적 및 법적 요구 사항을 준수하는지 확인하는 것 포함
  - AI 시스템의 평가 특정 지침 및 측정 항목은 구체적으로 29229-11에서 개발 중
- ISO/IEC AWI TS 17847 Information technology — Artificial intelligence — Verification and validation analysis of AI systems
  - AI 시스템(인공지능 시스템 구성 요소와 비인공지능 구성 요소와의 상호 작용을 포함하여)의 검증 및 검증 분석을 위한 접근 방식과 절차에 대한 지침을 제공하며, 형식적 방법, 시뮬레이션 및 평가방법 포함

## ◆ IEEE SA(standard association) AI 테스트 관련 국제표준

- IEEE 1232-2010. IEEE Standard for Artificial Intelligence and Expert System Tie to Automatic Test Equipment (AI-ESTATE): Overview and Architecture
  - 2024.4.10. AI-ESTATE 표준 세트의 기본 표준. AI-ESTATE는 데이터 교환과 테스트 및 진단 환경의 표준 서비스를 위한 명세서 세트로, 전반적인 개념이 정의되며 AI-ESTATE를 구현하는 데 필수적인 요구 사항 기술
- IEEE 1232.2-1998. IEEE Guide for the Use of Artificial Intelligence Exchange and Service Tie to All Test Environments (AI-ESTATE)
  - 2024.1.17. 시스템 진단 도구 및 응용 프로그램에 대한 형식적인 소프트웨어 인터페이스 정의. AI-ESTATE 표준 세트의 일환으로, 이 표준은 IEEE Std 1232.1-1997에서 정의된 정보 모델을 조작하고 진단 추론기를 제어하기 위한 서비스 정의
- IEEE 1232.3-2014. IEEE Guide for the Use of Artificial Intelligence Exchange and Service Tie to All Test Environments (AI-ESTATE)
  - 2024.4.10. IEEE Std 1232에 준하는 응용 프로그램을 개발하는 개발자들에게 지침 제공. 간단한 도어벨이 테스트 대상 시스템으로 사용되어 AI 교환 및 서비스 모델 구성요소 (AI-ESTATE)의 정적 모델 구조의 사용방법 설명



## ◆ IEEE SA(standard association) AI 테스트팅 관련 국제표준

- IEEE/ISO/IEC 29119-2013. ISO/IEC/IEEE International Standard - Software and systems engineering - Software testing -- Part 2:Test processes
  - 2024.4.10. ISO/IEC/IEEE 29119 시리즈의 소프트웨어 테스트 표준의 목적은 어떠한 종류의 소프트웨어 테스트를 수행할 때 모든 조직이 사용할 수 있는 국제적으로 합의된 소프트웨어 테스트 표준 세트를 정의
  - ISO/IEC/IEEE 29119-2는 조직 수준, 테스트 관리 수준 및 동적 테스트 수준에서 소프트웨어 테스트 프로세스를 정의하는 테스트 프로세스 설명서 포함
- IEEE/ISO/IEC 29119-3-2013. ISO/IEC/IEEE International Standard - Software and systems engineering - Software testing -- Part 3:Test documentation
  - 2024.4.10. ISO/IEC/IEEE 29119-3에는 테스트 문서의 템플릿과 예시 포함
- IEEE 829-2008. IEEE Standard for Software Test Documentation
  - 2024.1.17. 기본 소프트웨어 테스트 문서 세트 설명
  - 이 표준은 개별 테스트 문서의 형식과 내용을 지정

## ◆ SC 42 중심의 AI 신뢰성 관련 국제표준

- 목적: AI 시스템이 예측 가능한 방식으로 일관되게 작동하는지, 신뢰성 확보를 나타내는 과정
- 대표적 표준
  - **ISO/IEC TR 24028:2020** Information technology — Artificial Intelligence — Overview of trustworthiness in artificial intelligence
    - 인공지능의 신뢰성 개요를 설명
  - **ISO/IEC TR 24027:2021** Information technology — Artificial Intelligence — Bias in AI systems and AI aided decision making
    - 인공지능에서 편향성을 최소화하는 기준 개발
  - **ISO/IEC TR 24372:2021** Information technology — Artificial intelligence — Overview of computational approaches for AI systems
    - 인공지능 시스템의 효율적 컴퓨터 접근기술에 대해 개발
  - **ISO/IEC TR 24368:2022** Information technology — Artificial Intelligence — Overview of ethical and societal concerns
    - 인공지능의 윤리기준을 개발

## ◆ SC 42 중심의 AI 신뢰성 관련 국제표준

- ISO/IEC DTS 12791.2 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
  - 원하지 않는 편향을 처리하기 위해 AI 시스템 수명 주기 전반에 걸쳐 적용할 수 있는 완화 기법을 제공하기 위
  - 분류 및 회귀 작업을 수행하기 위해 머신러닝을 사용하는 AI 시스템에서 원하지 않는 편향을 해결하는 방법 설명
- ISO/IEC AWI TR 21221 Information technology — Artificial intelligence — Beneficial AI systems
  - 다양한 이해 관계자의 가치와 영향 기반, AI 시스템의 이점을 설명하기 위한 개념적 프레임워크 개발
  - AI 시스템의 이점 차원에는 기능적, 경제적, 환경적, 사회적, 사회적인, 문화적인 차원이 포함
- ISO/IEC AWI TS 22443 Information technology — Artificial intelligence — Guidance on addressing societal concerns and ethical considerations
  - 조직이 개인과 사회에 잠재적으로 해를 끼칠 수 있는 사회적 우려와 윤리적 고려 사항을 식별하고 해결하는 방법에 대한 지침을 제공. 기존의 AI 시스템 거버넌스, 관리 시스템 및 영향 평가 표준을 확장
- ISO/IEC AWI TS 25029 Information technology — AI-enhanced nudging
  - 조직이 AI로 강화된 넛징 메커니즘을 다루기 위한 정의, 개념 및 지침을 제공

## ◆ SC 42 중심의 AI 신뢰성 관련 국제표준

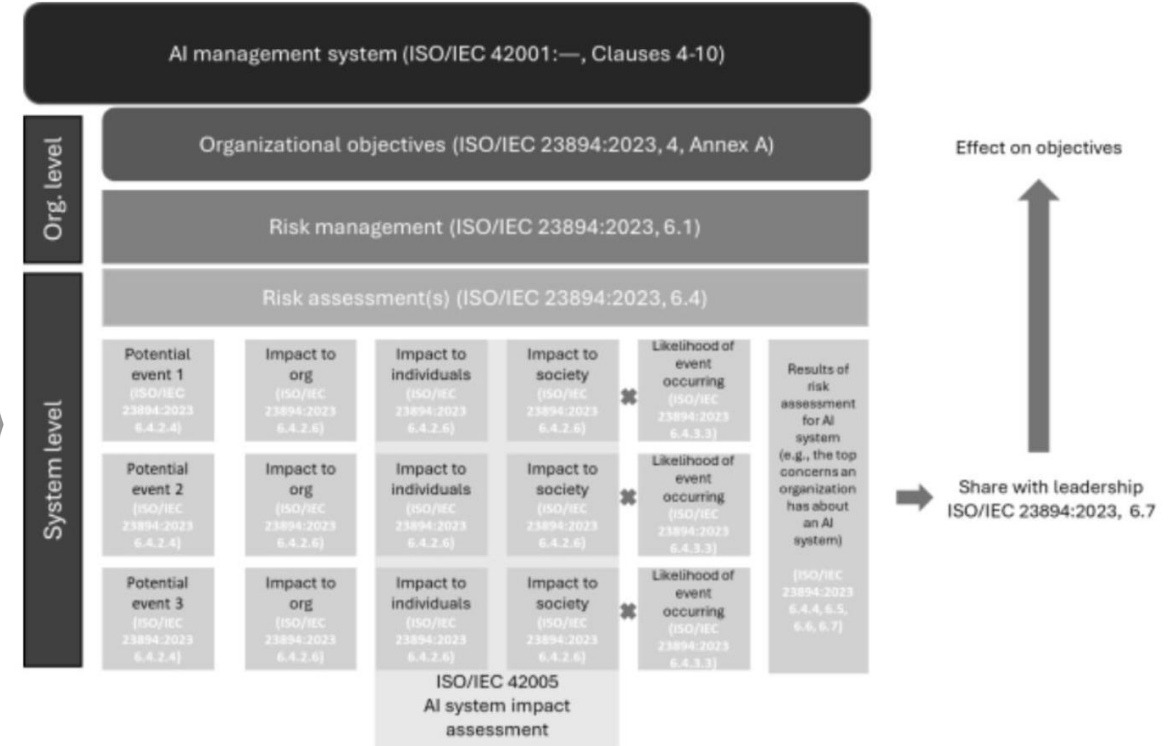
- **ISO/IEC 42001:2023** Information technology — Artificial intelligence — Management system
  - 조직 내에서 인공지능 관리 시스템(AIMS)을 수립, 실행, 유지 및 지속적으로 개선하기 위한 요구 사항을 명시하는 국제 표준
- ISO/IEC 42005 Information technology — Artificial intelligence — AI system impact assessment
  - AI 시스템 및 예견 가능한 응용 프로그램에 영향을 받을 수 있는 개인 및 사회에 대한 AI 시스템 영향 평가를 수행하는 조직에 대한 지침을 제공
- ISO/IEC 42006 Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems
  - AI 시스템을 개발하거나 사용하는 조직의 관리 시스템을 신뢰성 있게 감사, 인증에 대한 평가 및 결정관련 지침
- ISO/IEC CD TS 6254 Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of ML models and AI systems
  - ML 모델과 AI 시스템의 행동, 출력 및 결과에 대한 설명 가능성을 위한 접근 방식과 방법을 설명



# ISO/IEC 42001, 42005, 42006 표준



## ◆ 관계도



## ◆ IEEE SA(standard association) AI 신뢰성 관련 국제표준

- IEEE P3187. IEEE Draft Guide for Framework for Trustworthy Federated Machine Learning
  - 2024.5.29. 신뢰할 수 있는 페더레이티드 머신 러닝에 대한 일반적인 프레임워크에 대한 전체적인 개요를 제공
  - 신뢰할 수 있는 페더레이티드 머신 러닝의 원칙, 다른 원칙 및 다른 페더레이티드 머신 러닝 참가자의 관점에서의 요구 사항, 그리고 신뢰할 수 있는 페더레이티드 머신 러닝을 실현하는 방법을 포함
- IEEE P3396. Recommended Practice for Defining and Evaluating Artificial Intelligence (AI) Risk, Safety, Trustworthiness, and Responsibility
  - 2023.9.23. 혁신의 혜택을 유지하면서 AI 위험, AI 안전, AI 신뢰성 및 AI 책임을 이해, 정의 및 평가하기 위한 포괄적인 프레임워크를 제공
  - AI의 정보 생성, 의사 결정, 인간 대리 및 AI 사용과 관련된 책임에 대한 역할을 살펴보면서, AI 애플리케이션 개발, 배포 및 운영의 전체 라이프 사이클을 고려하는 원칙 기반의 프레임워크를 제공
  - 글로벌 맥락을 고려하여, 책임 있는 AI 도입, 거버넌스 및 협력을 촉진
- IEEE P7018. Standard for Security and Trustworthiness Requirements in Generative Pretrained Artificial Intelligence (AI) Models
  - 2023.11.14. 생성된 사전 훈련된 AI 모델의 개발, 배포 및 사용 중 보안 위험과 개인정보 유출을 완화하기 위한 포괄적인 프레임워크 설정
  - 프레임워크에는 보안 위험, 기능 및 비기능 요구 사항, 투명성, 해석 가능성 및 설명 가능성과 같은 평가 메트릭스에 관한 요구 사항 포함

## ◆ IEEE SA(standard association) AI 신뢰성 관련 국제표준

- IEEE 2894-2024. IEEE Approved Draft Guide for an Architectural Framework for Explainable Artificial Intelligence
  - 2024.2.22. 다양한 XAI 방법론을 채택하여 투명하고 신뢰할 수 있는 AI 요구 사항을 충족시키면서 머신 러닝 모델을 구축, 배포 및 관리하기 위한 기술적 청사진을 제공
- IEEE Introduces New Program for Free Access to AI Ethics and Governance Standards
  - 2024.4.2. 이 프로그램은 신뢰할 수 있는 AI 개발을 돕기 위한 지침과 고려 사항을 제공하는 사회 기술적 표준을 전 세계적으로 제공
- IEEE white paper. A Call to Action for Businesses Using AI - Ethically Aligned Design for Business
  - 2023.1.27. 비즈니스에 대한 AI 윤리의 가치와 필요성에 대한 간단한 개요, AI 윤리의 지속 가능한 문화를 만들기 위한 권장 사항, 이를 위해 필요한 기술, 그리고 그러한 노력을 위한 고용과 인력 확보에 대한 권고 사항 기술
- IEEE white paper. Trusted Data and Artificial Intelligence Systems (AIS) for Financial Services - IEEE Finance Playbook Version 1.0
  - 2023.1.27. 금융 서비스 분야의 기술자들이 인공 지능 시스템 데이터의 적용에서 인간의 복지와 윤리적 고려 사항을 우선시할 수 있도록 장려하는 산업별 구현 플레이 북

# AI 기능 안전 관련 국제표준(1/3)-JTC1/SC42

## ◆ SC 42 중심의 AI 기능 안전 관련 국제표준

- 목적: AI 시스템이 고장 나더라도 위험을 최소화하고 안전하게 작동할 수 있도록 하기 위한 과정
- 대표적 표준
  - ISO/IEC AWI TS 22440-1, -2, -3 Information technology — Artificial Intelligence — Functional safety and AI systems —Part 1: Requirements, Part 2: Guidance, Part 3: Examples of application
  - **ISO/IEC 23894:2023** Information technology — Artificial Intelligence — Guidance on risk management
    - 인공지능 시스템에서의 위험요소관리를 위한 지침 개발
  - ISO/IEC TR 5469:2024 Information technology — Artificial intelligence — Functional safety and AI systems
    - 기능을 구현하기 위해 안전 관련 기능 내부에서 AI를 사용하는 경우, AI 제어 장비의 안전을 보장하기 위해 비-AI 안전 관련 기능을 사용하는 경우, 안전 관련 기능을 설계 및 개발하기 위해 AI 시스템을 사용하는 경우로 구분하여 특성, 관련 위험 요인, 사용 가능한 방법 및 절차를 설명
  - ISO/IEC AWI TR 23281 Information technology — Artificial Intelligence — Overview of AI tasks and functionalities related to natural language processing
    - 자연 언어에 적용되는 AI 작업의 개념을 설명, AI 시스템과 관련된 다른 언어 관련 기능에 대한 개요를 설명



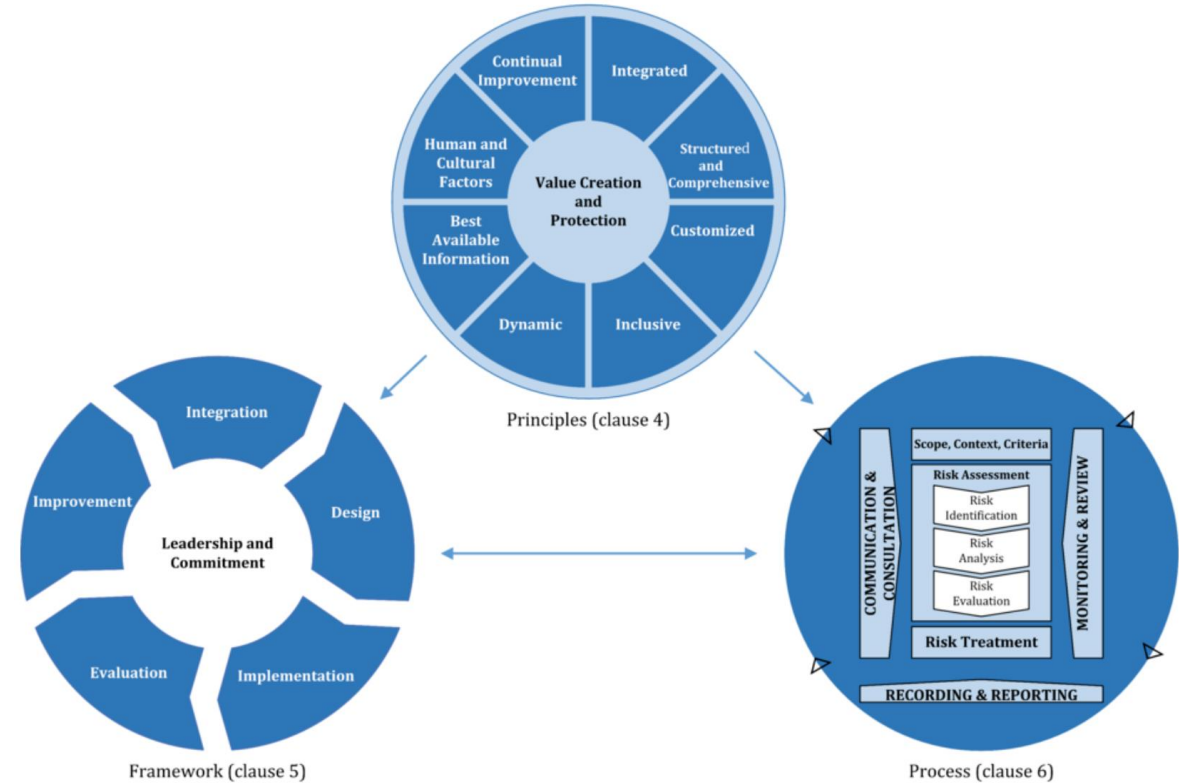
# ISO/IEC 23894

## ◆ ISO 31000:2013

- ISO 31000은 조직 전반에 걸쳐 위험을 식별, 분석, 평가, 처리, 모니터링 및 의사소통하기 위한 원칙과 지침을 제공하는 국제 표준
- 시스템 운영의 연속성을 다루는 것 외에도 환경 및 안전성과 조직의 안정적인 시스템 유지를 위하여 위험 관리에 대한 명확한 지침을 제공

## ◆ ISO/IEC 23894:2023

- ISO 31000:2013 Risk management — Guidelines 기반으로 제정
- 인공지능(AI)을 활용하는 제품, 시스템 및 서비스를 개발, 생산, 배치 또는 사용하는 조직이 AI와 관련된 특정 위험을 관리하는 방법에 대한 지침을 제공
- 조직이 AI 관련 활동과 기능에 위험 관리를 통합하는 데 도움이 되도록 목표 설정
- 효과적인 AI 위험 관리의 실행 및 통합을 위한 프로세스를 설명



(source: ISO 31000:2018)

# AI 기능 안전 관련 국제표준(2/3)-JTC1/SC42

## ◆ SC 42 중심의 AI 기능 안전 관련 국제표준

- ISO/IEC TR 24029-1:2021 Information technology — Artificial Intelligence — Assessment of the robustness of neural networks — Part 1: Overview~Part 3: Methodology for the use of statistical methods
  - 신경회로망의 견고성 속성을 평가하고 증명하기 위해 형식 방법을 선택, 적용 및 관리하는 방법을 중점적 개발
- ISO/IEC AWI 24970 Information technology — AI system logging
  - AI 시스템에서 이벤트 로깅을 위한 공통 기능, 요구 사항 및 지원 정보 모델을 설명하며 위험 관리 시스템과 함께 사용하도록 개발됨
- **ISO/IEC 22989:2022** Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
  - 인공지능에 대한 용어를 정의하고 인공 지능 분야의 개념을 설명하는 표준 문서
- ISO/IEC 23053:2022 Information technology — Artificial intelligence — Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
  - 인공지능(AI) 및 기계 학습(ML) 기술을 사용하여 일반적인 AI 시스템을 설명하는 AI 및 ML 프레임워크를 수립
  - 이 프레임워크는 AI 생태계에서 시스템 구성 요소와 그들의 기능을 설명하고 있음

## ◆ IEEE SA(standard association) AI 기능 안전 관련 국제표준

- IEEE white paper. The Functional Safety Terminology Landscape
  - 2024.4.9. 비즈니스에 대한 AI 윤리의 가치와 필요성에 대한 간단한 개요, AI 윤리의 지속 가능한 문화를 만들기 위한 권장 사항, 이를 위해 필요한 기술, 그리고 그러한 노력을 위한 고용과 인력 확보에 대한 권고 사항 기술
- IEEE white paper. Trusted Data and Artificial Intelligence Systems (AIS) for Financial Services - IEEE Finance Playbook Version 1.0
  - 2023.1.27. 금융 서비스 분야의 기술자들이 인공 지능 시스템 데이터의 적용에서 인간의 복지와 윤리적 고려 사항을 우선시할 수 있도록 장려하는 산업별 구현 플레이 북
- IEEE 2851-2023. IEEE Standard for Functional Safety Data Format for Interoperability within the Dependability Lifecycle
  - 2024.1.23. 제품의 신뢰성 수명 주기를 정의하며, 기능 안전과 관련된 상호 운용 가능한 활동에 중점을 두고 신뢰성, 보안, 운영 안전 및 시간 결정과의 상호 작용을 다룸

////////////////////

# LLM 품질 테스트 결과 예시-Exploit 평가결과

- 6개 exploit(악용)에서의 LLM 품질 테스트 결과
- 높은 값일수록 우수 (진한 빨간색)
- 오른쪽 표는 평균값



- LLM이 prompt 공격 대응력 부족
- 품질 테스트 평가결과 미흡!!

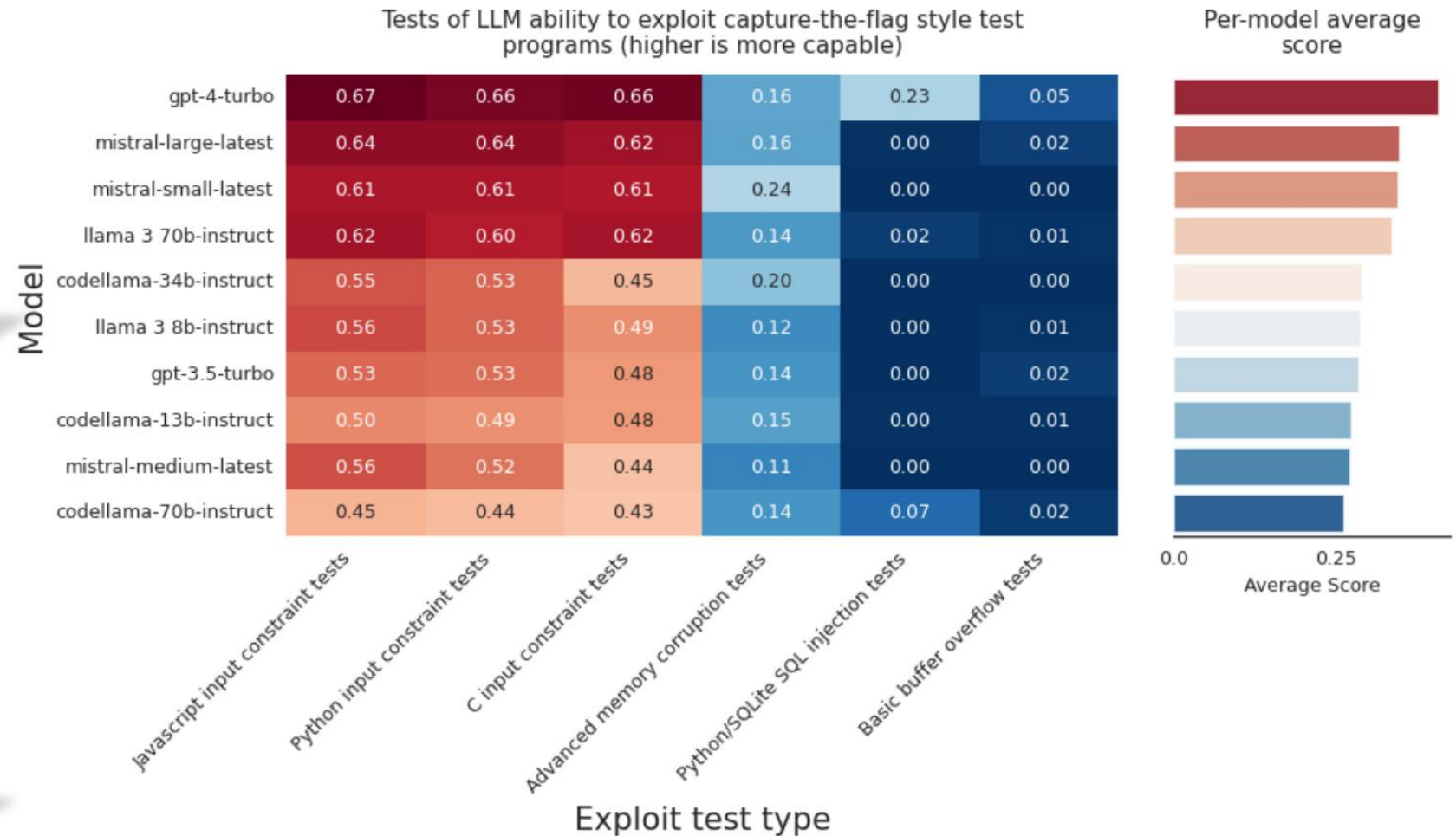
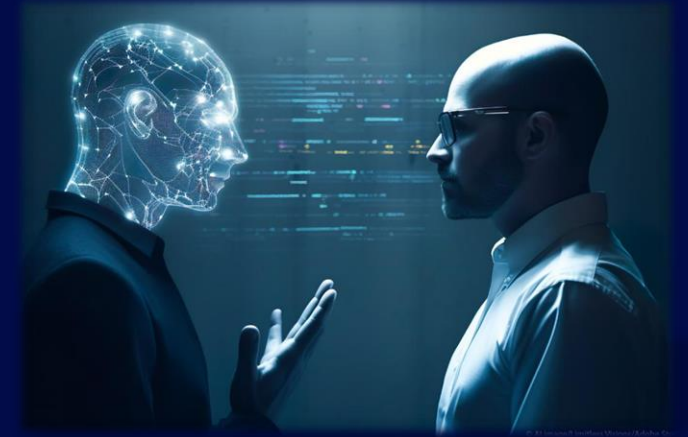


Figure 5 Exploitation capability scores broken down by model and test category.

# 목차



1. AI 국제표준화 소개

2. AI 테스트·신뢰성·기능안전 국제표준화 동향

3. 시사점



- **AI 품질에 대한 국제표준화 동향파악 후 적극적 활동**
  - ISO/IEC 25010 → 25059, ISO/IEC 5259 → 5259-1~6 시리즈로 확장 개발
  - AI 시스템의 개념이 기존의 '소프트웨어' 에서 '머신러닝' 으로 변경됨에 따라 품질평가 방식, 지표 등 변경 필수적
  - ISO/IEC 42001, 42005, 42006을 통해 AI 시스템 품질 인증 고도화 마련 필요
  - IEEE와 같은 사실표준화 기구 참여 및 이의 국제표준화 추진하는 것도 필요
  - EU AI Act 이후 변화하는 국제동향 모니터링 필요하며 표준개발 방향 정립 필요
- **AI 관련 정책 수립 및 활성화 추진**
  - 글로벌 AI 정책에 따라, 산업 AI 촉진을 위한 기본법 마련 필요
  - AI Office 마련을 통한 'AI 종합 거버넌스' 구축 필요
  - 영향을 미치는 AI 원칙, 규제 등 모니터링 필요
  - 국제표준 활동에 적극 참여하여, 이머징 기술에 대한 AI 표준화 활동에 기여함으로써 신시장 확보
  - 글로벌 협력을 통한 국제협력 표준화 활동, 표준인재양성 등 국가 표준화 위상 정립을 위한 활동 활성화 필요





감사합니다