인공지능 프레임워크 신뢰성 표준 현황에 관한 연구

김성한, 최영환
한국전자통신연구원

sh-kim@etri.re.kr, yhc@etri.re.kr

A Study on the Artificial Intelligence Trust Standardization

Kim Sung Han, Choi Young Hwan
ETRI

요 약

본 논문은 국내외 인공지능 프레임워크 신뢰성 관련 표준 현황을 분석하고, 특히 공적 표준화 기구인 ITU-T 및 JTC 1/SC 42 에서의 표준 개발 주요 사항에 대해 기술하고 있다. 이를 통해 향후 인공지능 신뢰성 관련 국제표준화 활동을 위한 참고로 활용 가능하다.

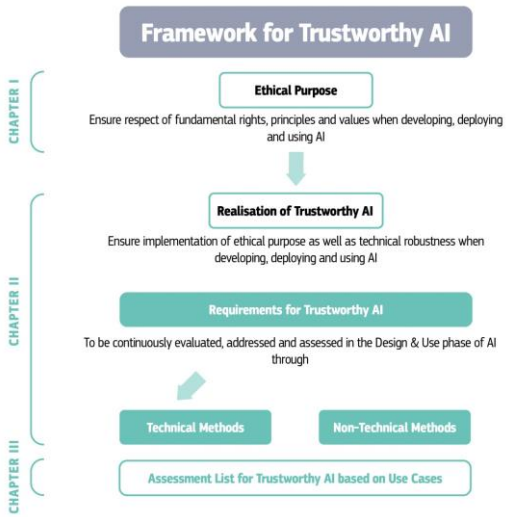
Ⅰ. 서 론

일반적인 트러스트(trust)는 사전적 의미로 타인의 미래 행동이 자신에게 호의적이거나 또는 최소한 악의적이지는 않을 가능성에 대한 기대와 믿음을 말한다. 즉, 사람을 믿고 그 사람의 미래 행동이 좋은 결과로 이어질 것이라는 믿음에 따른 행동에 대한 약속이다. 기술적인 측면에서 기존에 보안 및 프라이버시와 함께 사용자와 시스템 간의 자신감이나 확신, 신뢰하거나 의지, 믿음, 강도, 무결성, 의존성, 기대, 보증 등을 판단하고 이를 바탕으로 의사결정이 이루어질 수 있도록 제공한다.

신뢰할 수 있는 AI 를 위한 프레임워크에 대한 연구는 국제기구 및 단체에서 진행되고 있다. 이중에서 유럽에서 만든 AI 윤리 가이드 라인은 세 개의 주제로 구성되며, 각각은 추가 추상화 수준에 대한 지침을 제공하며 신뢰할 수 있는 AI 를 달성하기위한 프레임워크를 구성한다.

- 윤리적 목적: 본 주제는 AI를 다루는 모든 사람들이 준수해야하는 핵심 가치와 원칙에 중점을 둔다. 이는 EU 수준에서 EU 조약과 유럽 연합 기본권 헌장에 규정된 가치와 권리에 명시 되어있는 국제 인권법을 기반으로 한다.
- 신뢰할 수 있는 AI의 실현: 좋은 의도만으로는 충분하지 않다. AI 개발자, 배포자 및 사용자도 이러한 원칙과 가치를 기술 및 사용에 실제로 구현하기 위해 조치와 책임을 취하는 것이 중요하다. 또한 기술적 인 관점에서 시스템이 가능한 한 견고하다는 예방 조치를 취하여 윤리적 목적이 존중되더라도 AI가 의도하지 않은 해를 입히지 않도록 해야 한다. 따라서 신뢰할 수 있는 AI에 대한 요구 사항을 식별하고 이를 실현하는 데 사용할 수 있는 잠재적인 방법 (기술적 및 비 기술적)에 대한 지침을 제공한다.
- 평가 목록 및 사용 사례: 앞서 제시된 윤리적 목적과 구현 방법을 기반으로 AI 개발자, 배포자 및 사용자가 신뢰할 수 있는 AI를 운영 할 수 있는 예비 및 비포괄적 평가 목록을 설정한다.

이 지침의 구조는 아래 (그림 1) 신뢰 AI 프레임워크 가이드라인과 같다[1].



(그림 1) 신뢰 AI 프레임워크 가이드라인

Ⅱ. 표준화 현황

본 절에서는 ITU-T 및 JTC 1/SC 42 표준화 기구에서 AI 트러스트 관련 표준화 현황에 대해 기술한다.

ㅇ ITU-T Q16/13

본 Question은 네트워크의 기능을 고도화하기 위하여 기존의 상황인지 기술 등을 확장하여 지식기반으로 동적 네트워크 제어 및 관리가 가능토록 하는 표준개발을 추진하고, 미래 ICT 인프라에서 신뢰성 있는 통신

네트워크 및 서비스 제공을 위한 중점 기술의 표준화를 담당하고 있다.

ㅇ 미래 트러스트 ICT 인프라 표준화를 위한 CG

ICT융합, 사물인터넷 및 Connect 2020 결의가 채택되었으며, "미래 트러스트 및 지식 인프라" 표준의 필요성을 논의하기 위한 워크숍 및 트러스트 위한 CG-Trust를 만들었고 트러스트 정의, 유즈 케이스, 주요 기술 항목 발굴 및 향후 표준화 전략 등을 담은 기술 보고서 개발 작업 중이다.

ITU-T CG-Trust는 정량적 및 정성적 지표를 바탕으로 트러스트 관리가 이루어질 수 있도록 트러스트 레벨, 트러스트 품질 및 트러스트 인덱스 등과 같은 개념을 정립하고, 주요 기술 이슈를 도출하고, 핵심 유스 케이스 등에 대한 분석을 진행하는 등 향후 트러스트 표준화에 대한 큰 방향을 제시하고 있다[2].

번호	작업 아이템	권고 초안 명	추진일정
1	Y.3051 (ex Y.trusted-env)	The basic principles of trusted environment in ICT infrastructure	2017-02
2	Y.3052 (ex Y.trust-provision)	Overview of trust provisioning for ICT infrastructures and services	2017-02
3	Y.3053 (ex Y.trustnet-fw)	Framework of trustworthy networking with trust-centric network domains	2018-01
4	Y.3054 (ex Y.trustworthy-media)	Framework for trust-based media services	2018-05
5	Y.3053.Amendment	Trustworthy networking deployment architecture and procedures	2018-12
6	Y.trust-index	Trust index for ICT infrastructures and services	2021-07
7	Y.trust-arch	Functiaonl architecture for trust enabled service provisioning	2020-07
8	Y.SNS-trust	Framework for evaluation of trust and Quality of Media in Social Networking Services	2021-07
9	Y.trust-pdm	Framework for trust-based personal data management platform	2020-07
10	Y.PII-Did	Prioritization based de-identification methods for personally identifiable information	2020.07
11	Y.OBF_trust	Open bootstrap framework enabling trustworthy networking and services for distributed diverse ecosystem	2020.12

JTC 1/SC 42 WG3는 신뢰성 관련 표준 문서를 다수 개발하고 있으며 현재 진행중인 문서는 아래와 같다.

번호	작업 아이템	권고 초안 명
1	ISO/IEC CD 23894	Artificial Intelligence - Risk Management
2	ISO/IEC AWI TR 24027	Artificial Intelligence (AI) - Bias in AI systems and AI aided decision making
3	ISO/IEC TR 24028:2020	Artificial Intelligence (AI - Overview of trustworthiness in Artificial Intelligence
4	ISO/IEC DTR 24029-1	AI-Assessment of the robustness of neural networks - Part 1: Overview
5	ISO/IEC AWI 24029-2	AI-Assessment of the robustness of neural networks - Part 2: Formal methods methodology
6	ISO/IEC AWI TR 24368	Artificial intelligence (AI) – Overview of ethical and societal concerns
7	ISO/IEC WD 5059	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality Model for AI-based systems
8	ISO/IEC AWI TR 5469	Artificial intelligence – Functional safety and AI systems

여 지식기반으로 동적 네트

Ⅲ. 결론

본 논문에서는 인공지능 신뢰성 표준 이슈에 대해 공적 표준 기구인 ITU-T 와 JTC 1/SC 42 에서 진행중인 표준 현황에 대해 기본적인 소개를 하였다. 인공지능 신뢰성 이슈는 사회적, 제도적인 측면부터 기술적인 측면까지 다양한 현안 사항이 있지만 본문에서는 표준화 측면에서 진행되는 부분에 대해 일부 언급하였다. 본 논문은 향후 인공지능 신뢰성 이슈에 대해 구체적이고 체계적인 접근을 위한 가이드로 활용 가능리라 사료된다.

.

참 고 문 헌

[1] "A Layered Model for AI Governance," Urs Gasser and Virgilio A.F. Almeida, Harvard University, IEEE Internet Computing, 2017.

[2] ITU-T Y.3052, "Overview of trust provisioning in information and communication technology infrastructures and services", 2017.