인공지능 윤리(AI Ethics) 기술 및 표준화 동향 연구

김혜정, *박용주
한국전자기술연구원
스마트네트워크 연구센터
e-mail : *h426jung@keti.re.kr, suede8247@keti.re.kr*

A Study on the Trend of AI Ethics Standardization

Hyejung Kim, *Yongju Park
Smart Network Research Center
KETI(Korea Electronics Technology Institute)

Abstract

In this paper, we examine the advancement of artificial intelligence technology, which is highly likely to develop into a superintelligent body having an artificial ego, autonomy, and self-learning ability within 100 years amid the rapid evolution of artificial intelligence. In addition, as artificial intelligence technology develops, it is intended to find out an institutional device that can prevent various risks in advance and ensure the safety of humanity, which may occur due to the emergence of autonomous super-intelligences that are out of human control. In order to secure such safety devices, we will study the current state of artificial intelligence ethics standardization in order to standardize artificial intelligence ethics.

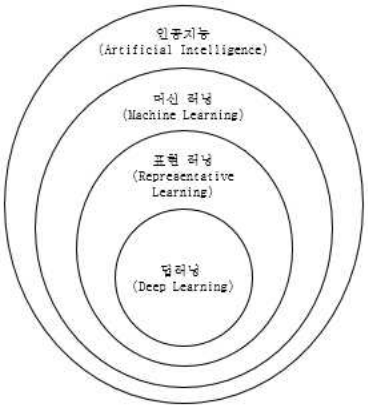
I. 서론

인공지능(AI)에 대한 사전적 정의는 '인간의 지능으로 할 수 있는 사고, 학습, 자기 개발 등을 컴퓨터가 할 수 있도록 하는 방법을 연구하는 컴퓨터 공학 및 정보 기술의 한 분야로서, 컴퓨터가 인간의 지능적인 행동을 모방할 수 있도록 하는 것'으로 '인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램으로 실현한 기술'로 요약된다. 이러한 인공지능(AI)에 대한 정의와는 별도로 인공지능(AI) 부문 연구자들과 엔지니어, 업계 관계자들은 인공지능(AI) 기술의 적용 수준에 따라 인공지능(AI) 개념을 <표 1>과 같이 나누며, 이러한 기술적 단계 또는 범주를 통한 인공지능(AI) 개념을 설명한 구성도는 아래 <그림 1>과 같다. [1]

<표 1> 인공지능(AI) 기술 개발 단계별 정의

구분	정의
인공지능 (AI)	인간의 인지능력(언어, 음성, 시각, 감성 등), 학습능력, 추론능력 등 인간지능을 구현하는 모든 기술
머신러닝 (Machine Learning)	데이터를 자체적으로 학습하여, 판단이나 예측 결과를 제공하는 알고리즘
표현러닝 (Representative Learning)	형상 검출이나 분류를 위해 필요한 데이터를 스스로 결정하고 최적화된 표현, 산출을 위한 방식 선택 등의 모든 작업을 진행한는 기술
딥러닝 (Deep Learning)	여러 층으로 쌓인 인공 뉴런의 조합을 통해 학습을 진행하는 모든 기계학습 기술

현재 인공지능의 빠른 발달은 경제적, 사회적 효과에 대한 기대뿐 아니라 자동화로 인한 일자리 대체, 통제 불능 문제 등 부정적 영향에 대한 우려감도 높이고 있는 상황이다.



<그림 1> 기술적 범주에 따른 인공지능 개념도

본 논문에서는 인공지능의 빠른 진화 속에서 100년 이내에 인공자아와 자율성, 그리고 자기학습 능력까지 갖는 초지능체(Superintelligence)로 발전할 가능성이 높아짐에 따라 인간의 통제를 벗어난 자율적 초인공지능체 등장으로 인한 여러 위험을 사전에 예방하고 인류의 안전성을 보장할 수 있는 제도적 장치가 인공지능 윤리 표준화 작업을 위해서 현재 인공지능 윤리 표준화 현황에 대해서 연구해 보고자 한다.

II. 본론

2.1 인공지능(AI) 윤리

인공지능의 진화방향은 약한 인공지능에서 스스로 사고 판단 예측, 스스로 학습 진화, 두뇌를 모사하는 인지컴퓨팅 등 강인공지능 기술으로 진화할 것으로 전망되고 있다. 전략 게임, 번역, 자율주행, 영상 인식 분야에 적용되는 약인공지능은 그동안 놀라운 발전을 거두었으며, 약인공지능은 여행 계획, 소비자 추천 시스템, 광고 타겟팅과 같은 많은 상업 서비스들을 가능하게 하였다. 현재 연구는 강한 인공지능을 이용하여 의학 진단, 교육, 과학 연구 분야에서 더욱 중요한 응용 기회를 모색하면서 다양한 변화들을 촉발시키고 있다.

현재, 전반적인 인공지능 기술 진화 과정에서 부각되고 있는 이슈들은 크게 인류와의 관계성 이슈로 부상하고 있는 상황이다. 이러한 기술의 진화와 인류와의 관계를 조화롭게 만들면서 인간에게 보다 편리하고 친근하며 안전한 인공지능 기술로의 개발을 보장하고 담보하기 위한 보편적, 제도적 장치가 바로 인공지능 윤리라고 할 수 있다. [2]

2.2 인공지능(AI) 윤리 이슈

자율주행자동차, 의료 분야 등에서 이미 인공지능, 로봇 기술이 상용화 단계에 진입하고 있으며, 머지않은 미래에 인공지능 기술은 단지 인간의 생활을 조력하는 단계를 넘어 인간의 삶의 일부로 편입되는 이른바 포스트휴먼(post-human)의 시대가 도래할 것으로 전망되고 있다. 인공지능 기술은 단순히 집적된 데이터의 활용을 넘어서서, 스스로 데이터를 분석 및 학습하고 이에 기반해 특정 목표를 실행에 옮길 것으로 예견되고 있다. 이러한 인공지능의 진화와 확산은 규범 환경 또는 구조적 측면에서 <표 2>과 같이 과거와는 다른 새로운 차원의 문제를 제기하면서 새로운 규범체계 정립의 필요성을 높아지고 있는 상황이다. [3]

<표 2> 인공지능의 주요 윤리적 이슈

구분	내용
인공지능 윤리적 안정성 확보	인공지능 관련 소프트웨어, 프로그램의 윤리적 오류 또는 오판을 기술적으로 최소화하고 이를 제어할 수 있는 제도적 장치 구비
사생활 침해 및 프라이버시 보호	인공지능 기반 데이터 연결, 확장, 공유가 보편화됨에 따른 프라이버시 침해 가능성을 최소화하고 피해발생 시 이에 대한 보상 제도 구축
인공지능 윤리적 책임성 강화	인공지능 적용 기술 및 제품의 설계, 생산 과정에서 알고리즘 자체의 윤리적 책임성을 부여 또는 부과하는 원칙과 기준 마련
인공지능 관련 윤리적 갈등 조정	인공지능 윤리 관련 이해관계자 간의 이익갈등 및 권리상충 문제에 대한 조정, 해결 방안 마련

2.3 인공지능(AI) 윤리 개발 동향

인공지능 기술을 이용한 이미지 혹은 동영상 합성 기술을 의미하는 딥페이크(Deep Fake)에 의해 만들어진 이러한 가짜 영상은 미국, 인도, 멕시코 등에서 정치적으로 악용된 사례가 있어 규제 필요성이 크게 대두되고 있다. 미국에서는 딥페이크에 대해 기술적 해결책을 우선하여 입법 움직임이 없는 가짜뉴스와 달리 적극적 입법 시도가 진행되고 있으며, EU는 허위정보에 대해 적극적 입법으로 대응하고 있으므로, 딥페이크는 허위정보 관련법에 의해 규제되고 있는 상황이다.

또한, 인공지능 기술의 판단 결과에 따른 차별성, 편향성이 일반인들의 대출, 채용, 형사 사법 및 광고를

포함한 다양한 응용 분야에서 조정없이 사용된다면, 몇몇의 소수 집단은 인공지능에 의한 명시적인 불이익을 받게 될 것으로 예상되고 있는 상황이다. 이를 위해서 데이터 차별성, 편향성에 기반을 둔 차별성과 편향성을 어떻게 보정할 것인지에 대한 연구와 함께 기계학습 모델의 동작 방식이나 결과에 대한 설명을 제공하는 설명 가능한 인공지능에 대한 연구 역시 윤리적 측면에서 지속적으로 수행 중에 있다.

2.4 인공지능 윤리 표준화 현황

인공지능(AI) 표준화 작업은 현재 초기 단계로 바람직한 인공지능 기술 개발 및 제품 적용을 위한 표준화 작업이 활발히 논의 중에 있다. 인공지능(AI) 표준화 대표 글로벌 기구로는 공적 표준화 기구인 ISO/IEC 기술위원회와 사실상 표준화 기구인 IEEE 및 Khronos Group으로 구분되며 각 기구가 추구하는 인공지능(AI) 개발 방향에 따라 표준화 활동이 진행 중인 상황이다.

2.4.1 공적 표준화 기구(ISO/IEC) 동향

인공지능(AI)에 대한 공적 표준화 기구는 해당 분야의 기술이 사회에서 공통적, 통상적으로 활용되고 적용되며 통용될 수 있는 보편적인 참조 표준을 개발하는 것을 목적으로 하고 있다. 인공지능 기술과 관련한 기본적인 개념과 용어에 대한 정의, 참조 구조, 적용가능 분야 별 원칙과 기준 등에 관한 표준화 개발을 주요 작업으로 진행되고 있다. ISO/IEC JTC 1에서는 인공지능(AI) 분야에 있어 <표 3> 같은 개념 및 용어에 대한 글로벌 기준 수립 작업을 공동으로 진행을 완료하였다.

<표 3> ISO/IEC JTC 1의 인공지능 개념 및 용어 글로벌 표준화 작업 결과

구분	내용
ISO/IEC 2382-28:1995	Information technology - Vocabulary - Part 28: Artificial intelligence - Basic concepts and expert systems
ISO/IEC 2382-29:1999	Information technology - Vocabulary - Part 29: Artificial intelligence - Speech recognition and synthesis
ISO/IEC 2382-31:1997	Information technology - Vocabulary - Part 31: Artificial intelligence - Machine learning
ISO/IEC 2382-34:1999	Information technology - Vocabulary - Part 34: Artificial intelligence - Neural networks

ISO/IEC에서는 2017년 11월에 인공지능(AI) 분야 기술 표준의 중요성을 인지하고 표준화 작업을 전문적으로 진행하기 위해 ISO/IEC JTC 1 내 신규 분과위원회 SC42를 신설하여 관련 표준화를 진행하고 있다. ISO/IEC JTC 1/SC 42가 개발하고 있는 표준 문서 <표 4>에 제시하였으며, 2020년 기준 총 10건으로 점차 표준개발 방향이 신뢰성과 사회적 관심 등 윤리적 이슈로 향하고 있음이 확인되고 있다. [4]

<표 4> ISO/IEC JTC 1 SC 42 표준 개발 현황

표준 번호	표준명
ISO/IEC WD 22989	Artificial intelligence - Concepts and terminology
ISO/IEC WD 23053	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC NP 23894	Information Technology - Artificial Intelligence - Risk Management
ISO/IEC NP TR 24027	Information technology - Artificial Intelligence (AI) - Bias in AI systems and AI aided decision making
ISO/IEC NP TR 24028	Information technology - Artificial Intelligence (AI - Overview of trustworthiness in Artificial Intelligence
ISO/IEC NP TR 24029-1	Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview
ISO/IEC NP TR 24368	Overview of ethical and societal concerns
ISO/IEC NP TR 24030	Information technology - Artificial Intelligence (AI) - Use cases
ISO/IEC NP TR 24372	Overview of computational approaches and AI systems
ISO/IEC NP 38507	Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations

2.4.2 사실상 표준화 기구 동향

사실상 표준화 기구의 경우에는 시장 매커니즘 하에서 해당 기술의 필요성이 합리적으로 인정될 경우, 이에 필요한 표준을 개발하는 것을 목적으로 하고 있다. 인공지능(AI)에 대한 사실상 표준화 기구는 인공지능 데이터 교환 및 공유 포맷, 시스템 설계 규정 등 실제 기술의 개발 및 적용 방향과 밀접한 표준 개발을 중점 진행 중이며 인공지능에 관한 글로벌 대표 사실상 표준화 기구는 IEEE와 Khronos Group이다.

먼저, IEEE는 인공지능 기술의 설계 단계에서 필요한 윤리 척도를 위한 프로젝트 그룹인 P7006, P7010 등을 운영하여 구체적인 표준안 발표하고 기계학습 프로토콜, 의료 서비스에서의 인공지능을 위한 신규 프로젝트 그룹들을 신설하여 시장활용 가능성이 높은 기술 분야에서 표준 개발 진행 중이다. IEEE의 윤리적 설계

문서의 구성 및 내용은 <표 5>에 제시하였다.

Khronos Group에서는 인공지능 기술의 구현 작업에 필요한 데이터 및 신경망 포맷 표준 개발 진행 중이다. 또한, 문서 개발뿐만 아니라 실제 적용 가능한 코드를 Github를 활용하여 배포하는 기술 개발 중에 있음과 더불어 NNEF v1.0를 배포하여 다양한 머신러닝 프레임워크에서 상호호환 가능한 신경망 모델 포맷을 개발하여 출시하였다.

<표 5> IEEE 윤리적 설계 문서 구성 및 내용

구분	내용
1. 일반원칙	모든 유형의 인공지능 및 자율시스템에 적용되는 윤리적 관심사 구분 및 목표 구성
2. 자율지능시스템을 통한 가치창출	자율지능시스템에 가치를 구현하는 목표 달성 위한 접근방법 제시
3. 윤리적 연구 및 설계 가이드라인	윤리적으로 건전한 인공지능 기술 개발 및 적용을 통한 경제적 영향력과 사회적 영향력 간의 균형을 유지 위한 쟁점 권고사항 제안
4. 범용인공지능(AGI)과 초인공지능(AS) 안전 및 혜택	인공지능 시스템은 산업의 변혁을 초래할 것으로 예상됨에 따라, 인공지능으로 인한 안전과 혜택 관련 다수의 쟁점과 권장사항 제시
5. 개인 데이터 및 개인접근 통제	데이터 비대칭성으로 인한 윤리적 딜레마 해결을 위한 개인 데이터 정의, 접근, 관리를 위한 쟁점 및 권고사항 제시
6. 자율무기 시스템 재구축	자율시스템 기반 공격무기는 전통적인 공격무기보다 많은 윤리적 문제를 내포하고 있음에 따라 관련 작업의 권고사항 제시
7. 경제적 및 인도주의적 쟁점	인간과 기술 간 글로벌 생태계를 형성하는 주요 동력을 확인하여, 경제적이고 인도주의적인 파급효과 연구 및 분석함
8. 법률	인공지능 및 자율시스템의 개발은 복잡한 윤리적 문제를 내포하고 있기 때문에 해당 문제에 대한 규제, 거버넌스, 국내법 및 국제법 제정에 대한 의견과 입장 제시

Ⅲ. 결론 및 향후 연구 방향

인공지능 기술의 급속한 발전 속도에 대응하기 위한 공적 표준화 및 사실상 표준화 제정 움직임이 국제적으로 진행되고 있는 현황을 감안한 국내 대응 체계의 구축이 절실한 상황이다. 이에 따라 인공지능 윤리 관련 선진국 및 표준화 동향에 대한 선제적 파악을 통해 국내 전기, 전자 산업계 중심의 협의체 창설 및 인공지능 윤리 표준화에 대한 입장 정리 및 통합이 필요할 것으로 판단된다. 그리고 나아가 국내 인공지능 기술 발전에 친화적이고, 기여가능한 표준이 글로벌 표준으로 채택될 수 있도록 하여 국내 인공지능 기술 육성을 지원하여, 향후 인공지능 기술 및 산업 진흥을 통하여 국가 발전에 기여하는 인공지능 생태계를 조성해야 할 것이다.

Acknowledgement

본 연구는 2021산업통상자원부의 재원으로 한국산업기술평가 관리원(KEIT) 지원을 받아 수행된 연구임 (No. 20012594, AI 윤리적 영향력 표준화를 위한 포럼 협의체 구성 및 가이드라인 마련 등 표준화 기반조성)

참고문헌

[1] 정보통신정책연구원, 4차 산업혁명시대 산업별 인공지능 윤리의 이슈 분석 및 정책적 대응방안 연구, 2018.

[2] 한국정보통신기술협회, 인공지능-표준화 동향 분석 (기술보고서), 2019.

[3] 서울대학교인공지능정책이니셔티브, 윤리적 인공지능의 실현과 과제, 2019.

[4] 산업 인공지능 윤리 표준화 포럼, 인공지능 윤리 유스케이스 표준 양식, 2021.