ETRI AI 실행전략 7: AI로 인한 기술·사회적 역기능 방지

ETRI AI Strategy #7: Preventing Technological and Social Dysfunction Caused by AI

김태완 (T.W. Kim, twkim@etri.re.kr) 지능화정책연구실 책임연구원
최새솔 (S.S. Choi, saesol.choi@etri.re.kr) 지능화정책연구실 선임연구원
연승준 (S.J. Yeon, sjyeon@etri.re.kr) 지능화정책연구실 책임연구원/실장

ABSTRACT

Because of the development and spread of artificial intelligence (AI) technology, new security threats and adverse AI functions have emerged as a real problem in the process of diversifying areas of use and introducing AI-based products and services to users. In response, it is necessary to develop new AI-based technologies in the field of information protection and security. This paper reviews topics such as domestic and international trends on false information detection technology, cyber security technology, and trust distribution platform technology, and it establishes the direction of the promotion of technology development. In addition, the development of international trends in ethical AI guidelines to ensure the human-centered ethical validity of AI development processes and final systems in parallel with technology development are analyzed and discussed. ETRI has developed AI policing technology, information protection, and security technologies as well as derived tasks and implementation strategies to prepare ethical AI development guidelines to ensure the reliability of AI based on its capabilities.

KEYWORDS DeepFake, Infordemic, Cyber Attack, AI Guardian, Realtime detection, AI Ethics

Ⅰ. 서론

1. 배경 및 필요성

인공지능 기술의 급속한 발전과 이를 이용한 지능화 시스템의 확산에 힘입어 인공지능의 활용 분야가 다양화되고 인공지능 기능을 중심으로 한 제품 및 서비스가 조직 및 개인에 속속 도입되는 과정에서 새로운 보안 위협 및 인공지능의 역기능이

* DOI: https://doi.org/10.22648/ETRI.2020.J.350708
* 이 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[20ZR1400, 국가지능화 기술정책 및 표준화 연구].
* 이 논문은 ETRI 기술정책연구본부 주관으로 담당 부서와의 워크숍 및 전문가 심층회의 등을 통해 수립된 'ETRI AI 실행전략'의 동향분석을 중심으로 작성되었다. 이 논문을 쓸 수 있도록 도움을 주신 ETRI 지능정보연구본부, 정보보호연구본부, 미디어연구본부, 표준연구본부 담당자분들께 감사드린다.

다음과 같이 현실 문제로 대두되고 있어 이에 대한 대응이 필요하다.

첫째, 인공지능 기술을 이용하여 정교한 합성 사진이나 영상을 만들어 내는 딥페이크(Deepfake), 전염병처럼 허위 정보가 미디어나 인터넷 등을 통해 급속하게 퍼져나가는 인포데믹(Infodemic) 등과 같이 허위사실임을 알면서 정치적·경제적 이익 등을 얻을 목적으로 정보 이용자들이 사실로 오인하도록 하는 허위조작정보의 생산과 유포가 새로운 사회문제로 대두되고 있다[1].

Deeptrace에 따르면, 2019년 딥페이크 영상증가율은 전년대비 약 100% 수준이며 영상의 96%가 성인물이 차지[2]하여 합성 성인물 피해자의 인권 침해를 야기할 뿐 아니라, 가짜 정치 뉴스로 민주주의를 위협하거나 새로운 지능범죄 출현 등 다양한 부작용을 초래하고 있다. 또한 가짜뉴스로 일컫는 허위조작정보는 뉴미디어의 발전과 더불어 전방위로 빠르게 확산하고 파급력이 매우 크다는 특징을 보이는데, 코로나-19로 인해 인류는 유례가 없는 심각한 허위 정보가 양산되는 인포데믹 현상을 경험 중이다.

이같이 인공지능(이하, AI)이 타인의 인권을 침해하거나 사실 정보를 왜곡하고 조작하는 선동 도구로 악용되면서 사회적 불안감 및 사회적 비용이 증가하는 반면, 현재 딥페이크나 허위조작정보를 자동으로 판단할 수 있는 기술은 부재하고 수작업으로 판별하는 상황이어서, AI를 활용한 허위조작정보의 진화속도 및 파급력을 고려할 때 진위 판별과 신뢰 기반의 정보유통 인프라 구축을 위한 기술개발이 시급하다고 하겠다.

둘째, 최근 AI 기술의 발전에 힘입어 새로운 형태의 지능화된 알고리즘을 활용한 신·변종 사이버 공격의 위협이 증가하고 있다[3].

초연결 사회로 진입함에 따라 다양한 기기와 네트워크의 상호 연결, 트래픽 폭증으로 사이버 공간은 확대되고 복잡도가 급격히 증가하여 안전한 관리를 점차 어렵게 만들고 있다. 또한, 전방위적인 대규모 AI 기반의 기계 해커 공격의 위험이 증가하고 피해 범위와 심각성이 커지는 양상이다. 즉, 이동통신 기술 및 IoT(Internet of Things)의 확산으로 사회 인프라, 생산 및 생활 시설이 사이버 공간으로 연결되어 사이버 공간의 취약성이 물리적 위협으로 전이되는 양상이 심화되고, 기밀 절취·금전 취득에서 정치적 목적의 사회 혼란 야기, 기반시설을 마비·파괴와 같은 사이버 테러로 사이버 공격의 성격과 영향력이 거대화하고 있다[4].

고도화되고 다양해지는 사이버 공격에 효율적으로 대응하기 위하여 AI 기술을 적용하여 방대한 양의 사이버 위협 정보를 빠르게 수집·분석하고 사이버 공격의 실시간 예측, 탐지 및 선제적 방어가 필요하다. 정보보호 및 사이버보안은 국민 생활과 직결되며, 안전한 AI 이용환경 조성은 AI의 전면화에 필수적인 전제조건인 만큼 AI에 따른 새로운 보안 위협에 대응하기 위한 기술적 제도적 방안 마련이 시급하다.

셋째, AI의 고도화 및 활용 확산에 따른 악용 및 사이버 공격의 위협에 더하여 AI가 인간관계 및 인간성 존중에 있어서 위험을 촉발할 가능성이 제기되고 있다.

AI는 우리 삶의 편리성을 향상할 수 있는 기술이지만 편향성, 불투명성, 책임소재 불분명으로 인해 편익 못지않게 위험도 초래할 것으로 전망된다. 세계경제포럼(WEF: World Economic Forum)에 따르면, 인지된 편익이 가장 큰 4개 미래유망기술 중에서 인공지능 및 로봇공학이 부정적 결과가 가장 클 것으로 조사되었다[5]. 이러한 AI가 유발하는 위험 가능성에는 편견과 차별, 개인의 자율성·의지·권리 보장에 대한 부정, 불투명하거나 불명

확하거나 정당하지 않은 결과 초래, 사생활 침해, 사회적 관계의 고립과 붕괴, 신뢰할 수 없거나 위험하거나 혹은 질적으로 낮은 수준의 결과 초래 등이 있다. 실례로 아마존의 AI 기반 채용시스템에서 여성이 차별된 경우, 미국의 재범예측시스템에서 흑인이 백인보다 재범예측률에서 불리한 결과를 초래한 경우 등이 있다[6].

이러한 AI의 한계로 인한 문제점들은 정책적, 문화적 접근도 병행되어야 할 필요가 있다. 즉, AI의 안전한 사용을 위한 사용 원칙과 AI 설계과정과 최초 개발목적대로 사용하고 있는지를 점검할 수 있는 기준 등의 마련이 요구된다.

2. 그간의 ETRI AI 연구 및 추진 방향

그간 ETRI는 AI의 보안 위협을 방지하고 설계 및 개발 과정에서의 AI의 신뢰성을 향상시키기 위한 기술 개발을 통해 역기능 방지 역량을 확보하였다.

가짜뉴스 식별 및 추론 기술은 언어, 음성, 영상처리 기술이 기반 기술이 되는데, ETRI는 국내 최고의 언어, 음성, 영상처리 기술을 보유하고 있다. 장학퀴즈에서 우승한 바 있는 엑소브레인 언어이해 기술, 이미지넷 챌린지에서 2위를 수상한 딥뷰 영상인식 기술, 세계최고 수준의 종단형 음성인식 기술 등 단일지능 기반 기술을 이미 보유하고 있으며, 특히 엑소브레인 과제에서 허위 정보 탐지의 원천기술인 문장 간 의미 비교 기술을 개발 중이며, 동의 문장 판별 정확도 85.03%로 한국어 최고 성능의 기술을 확보하였다.

미디어 신뢰성 판별 기술 분야에서는 ICT 인프라상에서 제공되는 서비스 및 미디어에 대한 신뢰성 판별을 위한 트러스트 정보 검증 체계에 대한 국제표준(2건)을 개발하였고, 미디어의 시맨틱 분석을 통해 뉴스 토픽을 추출하고 토픽 간 관계 및 연결성을 분석하는 기술을 개발하였다.

AI 기반 모의해킹 · 방어 시뮬레이션 프레임워크 분야에서는 네트워크 구성과 호스트 정보를 분석하여 자동화된 공격경로 예측 및 보안성 평가 기술을 개발하였다.

대규모 IoT 변종 악성코드 실시간 탐지 및 행위 기반 클라우드 백신 분야에서는 기존 상용 백신의 문제점 극복을 위한 클라우드 기반의 다중 AI 백신 기술을 개발하여 세계적 수준 악성코드 탐지율(신종 악성코드 탐지율: 98.6%, 오탐율: 1.27%)의 딥러닝 기반 바이러스 백신 엔진 기술을 확보하였다.

데이터 암호화 기술 분야에서는 단일 암호문에 다수의 평문 정보를 압축 암호화하여 저장 효율성을 극대화한 동형암호 기술 및 암호화된 데이터에 대한 자유로운 활용이 가능한 CipherData 기반 기술을 확보하였다.

지능정보 기술 확산을 위한 AI 데이터 표준 분야에서는 JTC 1/SC 42 데이터의 생애주기에 따른 품질기준/개요 표준개발에 착수하였고, ITU-T SG13 데이터의 전처리 및 레이블링 등 데이터 이력 정보 관리 기능/시스템 아키텍처를 통한 데이터 투명성(Transparency) 표준을 개발 중이다.

3. ETRI AI 연구 추진 방향

ETRI는 'AI R&D 전략'과 '2035 기술로드맵 계획' 등 AI 연구전략을 수립하고 ETRI의 추진과제를 도출하였다. ETRI의 주요 역할 및 기술력을 고려하여 전문가 회의를 통해 도출된 ETRI의 AI 역기능 방지 추진과제는 다음과 같다.

1) AI 보안관 기술

2) 정보보호 및 보안 기술

3) 윤리적 AI 개발 가이드라인

Ⅱ장에서는 이 3가지 추진과제 분야에 대한 기술 개념 및 국내·외 동향을 살펴보고, Ⅲ장에서는 ETRI 추진과제별 세부내용을 소개한다. Ⅳ장에서는 추진전략의 방향을 간략히 소개한다.

Ⅱ. 기술 개념 및 동향

1. AI 보안관 기술

가. 개념 및 필요성

AI 보안관 기술은 사람 또는 AI에 의해 악의적으로 생성되는 가짜뉴스, 보이스피싱, 딥페이크 등 허위조작정보를 추론 및 탐지하고 신뢰 기반의 정보유통 플랫폼을 통해 전파하는 기술이다. 문장·문서 분석기반의 사실·의미·의도 파악을 통해 사실관계를 추론하여 허위조작정보 근거를 제시하고 강화·전이학습과 다중 인식 기술 기반으로 보이스피싱 및 딥페이크 영상을 판단하며 블록체인의 분산합의 알고리즘에 기반하여 유통정보의 위·변조 및 오남용을 방지한다.

현재 딥페이크나 허위조작정보를 자동으로 판단할 수 있는 기술은 부재한 상황이며, 사회적 파급력을 고려할 때 공공성을 강조한 기술 개발이 필요하다. 또한, 사람이 하기 어려운 허위조작정보의 판별과 온라인상에 유통되는 기사의 신뢰성 평가에 대한 AI 기반 근거 추론 기술 개발은 대규모 허위조작정보의 대응에 필요하다.

나. R&D 방향

허위조작정보의 식별은 온라인상에 유통되는 기사의 신뢰성 평가를 위하여 기사 내용과 동일한 정보 및 반대되는 정보에 대한 근거 추론이 필요하지만, 이는 현재 AI 기술 수준으로는 해결이 어려운 상황이다.

표 1 허위조작정보 식별 기술발전 방향

구분	AS-IS	To-Be
허위조작 정보 수준	검색 기반 수작업 식별	AI 기반 자동 감지
추론 기술	사람의 추론에 의존	AI 판단 근거 추론 기술
판단 근거 제시	개별적 문서 단위 근거 제시	통합적 판단 근거 자동 요약 제시

출처 ETRI 인공지능연구소, "실행전략 워크숍," 2020. 2.

1) 해외

가짜뉴스 탐지는 전문가 및 집단지성 수작업으로 이루어지고 있지만, 최근 Google, 아마존, 알리바바를 중심으로 이상 확산 패턴 감지, 언어처리 및 시맨틱 기반 등 AI 기술을 적용한 가짜뉴스 탐지 기술 개발이 추진 중이다.

마이크로소프트, 페이스북, 아마존 등 빅테크 기업을 중심으로 2016년 결성된 단체인 'AI 파트너십(Partnership on AI)'은 허위조작된 미디어 식별하고 얼굴 또는 음성이 조작된 비디오를 판별하는 혁신적 AI 기술 개발을 촉진하기 위해 딥페이크 탐지대회(Deepfake Detection Challenge)를 개최하여 딥페이크 탐지에 대한 기술적 접근을 시도하였다[7]. 또한, 미국의 DARPA는 이미지나 동영상이 AI를 이용해 수정되거나 변형되었는지를 판단할 수 있는 기술인 '미디어 포렌식(MediFor: Media Forensics)' 프로그램을 진행 중이다. 이는 이미지 수집 및 분석 등 전통적인 디지털 포렌식 기법을 자동화하기 위해 만들어졌으나 2018년 하반기에 AI가 작성한 위조 이미지 및 영상 감별로 초점을 전환하고 눈 깜빡임, 머리 움직임, 눈 색깔 등 기존에 알려진 것뿐만 아니라 AI를 활용해 인간이 알 수 없는 변화 등을 복합적으로 판단해 딥페이크 기술에 의한 조작 여부를 판단하고 있다[8].

2) 국내

국내에서는 AI 기술을 활용하여 산발적으로 가짜뉴스, 허위음성 및 허위영상 식별 기술에 관한 연구 개발을 시작하는 초기 단계이다.

과기부에서 2017년 가짜 뉴스 탐색을 위한 "인공지능 R&D 챌린지"를 개최하여 가짜뉴스를 가려낼 AI 원천기술 개발의 중요성을 강조하였지만, 단어 수준의 딥러닝 기반의 어휘 통계 분석에 그치는 한계를 보였다. SK텔레콤, KT, LG유플러스 등 통신사별로 보이스피싱 및 스팸을 검출하는 기술 및 서비스가 운용되고 있으나 이를 회피하는 범죄 유형도 고도로 지능화되고 있어서 기업별 산발적인 연구가 진행 중이다. KISA는 금융감독원 및 금융기관에 피싱사이트 탐지 노하우를 공유하는 노력을 진행 중이다. 또한, 국내 일부 기업을 중심으로 학습 데이터 구축 및 음란물 유해 사이트에 대한 대처를 위한 기술개발하고 있고, 학계에서는 KAIST, 고려대, 성균관대 등에서 영상 위주의 학습데이터 구축 및 기초적인 탐지 알고리즘 연구가 진행되는 초기 진입 단계이다.

KAIST는 Twitter와 인터넷상에서 루머가 전파되는 양상을 관찰하고 특성을 파악하여 가짜뉴스 판별하는 연구를 진행하고 있으며, 서울대학교는 네이버와 함께 뉴스 소비자의 참여를 통해 뉴스의 거짓 정도를 파악하는 플랫폼 기술을 개발하고 있다.

2. 정보보호 및 보안 기술

가. 개념 및 필요성

정보보호 및 보안 기술은 AI 기술의 대중화 · 보편화에 따라 개인정보 유출과 프라이버시 침해, AI 기반 사이버 공격 등으로 악용되는 치명적 역기능을 방지하기 위한 사이버보안에 AI 기술을 활용하는 정보보호 지능화 기술이다.

AI 기술발전은 고도로 지능화된 사이버 공격을 양산하며 AI 기반 해커로 인한 기존 보안 시스템의 무력화 등의 국가와 사회의 보안체계를 위협할 가능성이 있어 이에 대비할 필요가 있다. 미래위협을 선제적으로 예방하고 AI 해커 등 지능형 위협에 실시간 대응 및 방어를 통해 국가 인프라를 보호하고 사회안전을 확보하여야 한다. 또한, AI 기반 정보보호 및 보안은 고도로 발달하는 기술 간의 창과 방패의 싸움으로, 지속적인 연구를 통해서만 선제적 예방과 적시적 대응력 확보가 가능하다. AI가 지금껏 우리가 경험하거나 상상해 보지 않았던 새로운 공격 패턴을 만들어 내는 등 AI의 발전이 보안 시스템의 강화뿐 아니라 공격 시스템의 진화도 촉발시킬 수 있음을 염두에 두어야 한다. 아울러, AI 기반 데이터 분석 및 활용이 크게 증가하여 발생 가능한 개인 민감정보 침해를 방지하면서 AI 학습이 가능한 활용성 보장 데이터 암호화 기술에 대한 요구도 점증되고 있다.

나. R&D 방향

딥러닝 등의 기법을 활용하여 공격자의 진화에 예방적으로 대응하는 지능형 방어, 정상행위의 학습을 통해 비정상 행위를 탐지하는 행태분석탐지 등의 기술이 발전[9]하여 발생하는 각종 디지털 정보에 대해 빅데이터 및 머신러닝 기반 이벤트 분석을 통하여 실시간으로 알려지지 않은 위협 탐지, 패킷 수준의 네트워크 기반 침입 탐지 및 예방, 지능화된 신종 악성코드의 실시간 분석 등에 적용될 것이다[5]. 또한, 대량의 빅데이터에 대한 AI 학습 과정에서 발생 가능한 개인 민감정보 유출 등의 사회적 역기능 해소를 위해 개인의 프라이버시를 보장하면서도 정확한 AI 학습이 가능하도록 비식별화 기술, 암호화 기반기술, 동형암호 기술에 대한 연구가 지속될 전망이다.

1) 해외

공격예측 기술 분야에서는 공격자가 타겟 네트워크에 침입하는 데 사용할 수 있는 경로를 모델링하기 위해 공격 그래프 기술이 2000년대 초부터 지속적으로 연구되고 있다. 공격 시뮬레이션 기술은 연구 초기 단계로 미국, 이스라엘을 중심으로 AttackIQ, SafeBreach, Cymulate, Treatcare, Verodin, Picus 등 10여 개의 벤더들이 솔루션을 제공하고 있다.

암호 기술 분야에서는 미국의 DARPA가 2020년 완전동형암호 구현 고속화를 위한 SoC레벨 HW가속기 개발을 위해 3천3백만 달러 규모의 예산을 투입하는 DPRIVE 사업에 착수하였다.

악성코드 탐지 기술로는 미국 미시간대학에서 개발한 클라우드 환경에서 여러 백신 엔진을 사용한 악성코드 탐지 기능을 제공하는 CloudAV 등이 있다.

2) 국내

공격예측 기술은 ETRI, 고려대 등에서 Attack Graph 기반의 사이버 공격 경로 예측 기술이 연구되고 있으나, 상용제품은 보고된 바 없다.

암호 기술 분야에서는 서울대학교는 근사연산이 가능한 동형암호 라이브러리 HEAAN을 개발하여 국제 표준화 및 실증 연구를 진행 중이다.

악성코드 탐지 기술 분야에서는 안랩이 클라우드 기반의 악성코드 위협 분석 및 대응 기술로 ASD(AhnLab Smart Defense)를 개발하여 다수의 PC에서 수집된 수십억 개 파일의 신·변종 악성코드 탐지를 구현하였다. 이스트시큐리티는 딥러닝 기반의 악성코드 위협 대응 솔루션인 Threat Inside를 통해서 엔드포인트 보안 및 악성코드 분석 서비스를 제공하고 있다.

3. 윤리적 AI 개발 가이드라인

가. 개념 및 필요성

윤리적 AI 개발 가이드라인은 AI 시스템의 개발 과정 및 최종 시스템의 윤리적 타당성이 보장되도록 인간 중심가치, 공정성, 투명성 및 설명 가능성 등의 적합 설계 및 구현을 위한 지침을 일컫는다.

AI는 우리 삶의 편리성을 향상할 수 있는 기술이지만 편향성, 불투명성, 책임소재 불분명으로 인해 위험도 증대하고 AI 기술의 적용 범위와 영향력을 고려할 때, 위험관리는 기술적 접근뿐만 아니라 기존 사회적 관념과 인식의 전환이 수반되어야 하는 만큼 이해관계 등에 대한 합의 과정이 뒷받침되어야 하며 윤리 헌장을 통해 과학기술의 책임성을 부여할 필요가 있다.

최근 지능정보기술이 상용화되기 시작하면서 국내외에서 각종 윤리 헌장, 가이드라인이 발표되고 있으나, 필수 공통적인 사항에 대한 합의 및 확산이 이루어지지 못하고 있으며 각 헌장의 체계성, 포괄성, 대표성 등이 명확하지 않으며, AI 기술 개발의 전주기를 담은 표준화된 개발사례 및 가이드라인도 부재한 상황으로 윤리적 AI 개발 체계 정립 및 지침 마련이 시급한 상황이다[10].

나. R&D 방향

1) 해외

최근 주요국과 국제기구는 AI가 갖는 편향성 및 불공정 문제 및 악의적 오용 가능성을 차단하기 위한 기술 개발 및 활용의 윤리원칙 또는 가이드라인을 발표하였는데 세부 항목은 표 2와 같다.

2) 국내

국내에서는 정부와 공공기관이 제정한 윤리 헌

표 2 국외 AI 윤리 헌장/가이드라인

국가/기구	제목	내용
EU [11]	신뢰할 수 있는 인공지능을 위한 윤리 가이드라인 (Ethics Guidelines for Trustworthy AI) (2019. 4.)	① 인간의 선택의지와 감독, ② 인공지능 기술의 견고성과 안전성, ③ 개인정보 및 데이터 거버넌스, ④ 투명성, ⑤ 다양성, 비차별, 공정, ⑥ 사회 및 환경 복지, ⑦ 책무성
일본 (총무성)	AI 연구개발 가이드라인 (2017. 10.)	① 협력성, ② 투명성, ③ 통제가능성, ④ 안정성, ⑤ 보안성, ⑥ 프라이버시, ⑦ 윤리성, ⑧ 사용자 지원성, ⑨ 책임성
OECD [12]	Principles for responsible stewardship of Trustyworthy AI (2019. 6.)	① 포용성·지속가능성, ② 인간중심가치·공정성, ③ 투명성·설명가능성, ④ 견고성·안전성, ⑤ 책무성
IEEE [13]	Ethically Aligned Design edition 1 (2019)	① 인간권리 보호, ② 복지, ③ 데이터 대행, ④ 유효성, ⑤ 투명성, ⑥ 설명가능성, ⑦ 오용의 자각, ⑧ 적합성

표 3 국내 AI 윤리 헌장

발표내용	일자	발표 주체
로봇윤리 헌장 초안	2007. 3.	산업자원부 (현 산업통상자원부)
카카오 알고리즘 윤리헌장	2018. 1.	카카오
지능정보사회 헌장	2018. 6.	과학기술정보통신부/ 한국정보화진흥원(NIA)
지능형 정부 인공지능 활용 윤리 가이드라인(안)	2018. 12.	한국정보화진흥원(NIA)
인공지능 윤리헌장	2019. 10.	한국인공지능윤리협회 (KAIEA)
이용자 지능정보사회 원칙	2019. 11.	방송통신위원회/정보통신 정책연구원(KISDI)
자율주행차 가이드라인	2019. 12.	국토교통부 첨단자동차기술과

출처 오윤경 외, "혁신과 위험관리: 사람중심 기술혁신을 위한 추진과제," 2020; 이순기, "인공지능의 윤리적 사용을 위한 개선과제," 2020. 9. 재인용

장 5개와 비영리기관과 카카오가 마련한 윤리 헌장 2개 등 총 7건의 AI 관련 윤리 헌장 및 가이드라인이 발표되었다(표 3 참조)[10].

3) 표준화

ISO/IEC는 응용 및 산업에 적용될 AI 표준 개발을 위해 2017년 10월 JTC1/SC_42가 설립되어 AI 기반표준, 빅데이터, AI 신뢰성, AI 사례 및 응용, AI 시스템 등의 표준화를 추진 중이다. 2020년 신뢰성/윤리 분야에서 AI 신뢰성 개요(ISO/IEC 24028:2020) 표준을 제정하고 후속 표준으로 AI 시스템 및 AI 기반 의사결정 지원에서의 편향(ISO/IEC WD TR 24027), 뉴럴 네트워크의 견고성 평가(ISO/IEC 24029 시리즈), AI 윤리 및 사회적 관심사(ISO/IEC WD TR 24368), 기능 안전 및 AI 시스템(ISO/IEC AWI TR 5469), 데이터 품질, 요구사항, 가이드 (ISO/IEC WD 5259 시리즈), AI 시스템 라이프사이클(ISO/IEC WD 5059), AI 애플리케이션 개발 가이드라인(ISO/IEC WD 5339) 등의 표준을 개발 중이다. 최근 10월에는 설명가능한 인공지능(XAI)의 표준화의 일환으로 머신러닝 모델 및 AI 시스템의 설명가능성(Explainability)의 표준화 논의 중에 있다.

IEEE에서는 2016년부터 IEEE 윤리적 설계 원칙을 구체화하는 표준 P7000 시리즈(P7001~P7014)를 개발 중이다.

ETRI는 감염병 탐지, 예방, 대응, 회복 등 방역 단계별 AI 활용/응용사례를 통해 표준화 체계를 만들어 감염병 대응을 위한 의료 AI 표준 정립의 가이드라인을 제시하였다[14].

III. ETRI 추진 과제

1. AI 보안관 기술 개발

ETRI에서 추진하는 'AI 보안관' 기술 개발은 세부 핵심 기술로 (1) 가짜뉴스 식별을 위한 판단 근거 추론 기술, (2) 자가성장이 가능한 DeepFake 대응 기술, (3) 강화·전이학습을 통한 위조음성 식별 기술, (4) 멀티모달 미디어 신뢰성 판별 기술, (5) 블록체인 활용 신뢰 기반의 정보유통 플랫폼 기술

을 포함한다.

(1) 가짜뉴스 식별을 위한 판단 근거 추론 기술에서는 문장/문서 분석기반의 사실/의미/의도 파악 기술, 신뢰도 측정 기반 허위 정보 판단 기술, 사실관계 추론에 기반한 근거 제시 기술, 가짜뉴스 배포 및 확산 양상, 유포자 신뢰성 검증 등을 위한 정보확산 모델링 기술, 머신러닝을 위한 허위 정보 학습 데이터 구축 및 공유 기술 등을 개발한다.

(2) 자가성장이 가능한 허위영상(DeepFake) 대응 기술은 영상을 파편화하고, 각 파편에 대한 개별 인식 기술을 적용하여 다층/다형상의 인지 기술이 결합된 딥페이크 판별 기술, 빛의 움직임, 사물의 움직임, 조각의 결합에 따른 부분의 균질성 등의 다양한 파편화 인식 기술, 사람 대비 20% 이상의 딥페이크 검출성능을 확보 가능한 판별 네트워크 기술, 파편화되어 있는 판별 기술을 하나의 네트워크 앙상블로 결합하는 시스템 기술, 소량의 레이블된 데이터로부터 대량의 영상 데이터를 자가 증식하는 미디어 데이터 자가 증식 기술을 포함한다.

(3) 강화·전이학습 기반 위조음성 식별 기술 개발은 세부기술로 강화·전이학습 등을 통해 위조 음성 및 음향에 대응하는 학습 기술, 발성 스타일 위조, 음성합성 기반 위조에 대응하는 고성능 탐지 기술, 신호처리 및 딥러닝 하이브리드 알고리즘 기반의 고정도 성문 및 화자 검증 기술, 원시 음성의 복사를 방지하는 오디오 워터마킹 기술, 고도로 암호화된 개인별 음원(보이스 스킨) 기술 등을 포함하여 원천기술 확보를 목표로 추진한다.

(4) 멀티모달 미디어 신뢰성 판별 기술 개발은 미디어 신뢰도 측정을 위한 지능형 미디어 지식 베이스 구축 기술, 미디어의 멀티모달리티 분석을 통한 스토리 이해 기술, 미디어 고의적 편집 및 악의적 정보 판별 기술, 미디어 신뢰도 실시간 측정 플랫폼 기술 개발을 목표로 한다.

(5) 블록체인 활용 신뢰 기반의 정보유통 플랫폼 기술은 무결점의 원본/진본 정보서비스를 신속하고 정확하게 제공하는 블록체인 기반의 플랫폼 기술 등을 포함한다.

2. 정보보호 및 보안 기술 개발

ETRI에서 추진하는 '정보보호 및 보안' 기술은 개인정보 유출과 프라이버시 침해, AI 기술의 사이버 공격으로 악용되는 치명적 역기능을 방지하기 위한 정보보호 지능화 기술로 (1) AI 기반 모의 해킹·방어 시뮬레이션 프레임워크, (2) 대규모 IoT 변종 악성코드의 AI 기반 실시간 탐지 기술, (3) 암호화된 상태에서 원데이터의 처리·분석이 가능한 데이터 암호화(Cipher DB) 기술 개발을 추진한다.

(1) AI 기반 모의 해킹·방어 시뮬레이션 프레임워크 기술 개발에는 실제 환경에서 수집된 이상 행위 빅데이터 확보를 통한 AI 기반 사이버 보안 프레임워크, 기반시설(스마트팩토리, 스마트에너지, 스마트헬스, 국방망 등)의 정상/비정상(모의 공격) 데이터 수집 기술, 주요 기반시설 ICT 인프라 가상 공간 복제 기술 및 AI 기반 사이버 공격 예측 기술 개발 및 보안성 강화정책 수립 등이 포함된다.

(2) 대규모 IoT 변종 악성코드의 AI 기반 실시간 탐지 기술 개발에서는 변종/신종 악성코드 탐지 및 대응을 위한 AI 악성코드 분석 및 다중 AI 백신 기술, 5G 환경에서 정상행위 분석기반 이상 행위 탐지를 위한 비지도학습 모델 및 탐지 신뢰도 극대화를 위한 다중 머신러닝 통합형 앙상블 모델 기술, 군집 IoT 상황의 보안 위협전파 패턴 AI 분석기반 신뢰 연결관리/자율제어 기술, 국방망 등 공공서비스 적용을 위한 클라우드 기반 경량 에이전트 및 다중 AI 백신 및 지원하는 클라우드 플랫폼 응용 기술 개발 등을 포함한다.

(3) 암호화된 상태에서 원데이터의 처리 · 분석이 가능한 데이터 암호화(CipherData) 기술의 세부기술로 암호화된 상태를 유지하면서 데이터의 처리 및 연산이 가능한 검색 기술, 중복처리 기술, 소유권 증명 기술, 동형암호 기술 등의 원천기술 확보를 목표로 한다.

3. 윤리적 AI 개발 가이드라인

ETRI는 AI의 인간중심 가치, 공정성, 투명성, 설명가능성 등의 확보를 보장할 수 있도록 글로벌 규범에 부합하는 윤리적 AI 개발 헌장(기준)을 마련하고 오픈 플랫폼 등을 통한 국내외 개발자 생태계 배포 등을 추진한다. 또한 고신뢰 AI 시스템을 위해 요구사항 표준, 신뢰성 모델링 분석, 측정, 검증 방식, AI 시스템 신뢰성 제공 기능 및 구조 표준과 빅데이터/AI의 활용 목적에 따른 데이터 용도 및 생애주기별 품질 측정 표준 개발을 추진한다.

Ⅳ. 결론

AI으로 인한 기술적, 사회적 역기능을 방지하기 위하여 누구나 믿고 사용할 수 있는 AI 기술과 서비스 기반을 마련하여야 한다. 이를 위하여 ETRI는 AI 시스템의 개발 전 주기에 걸친 맞춤형 전략으로 신뢰 가능한 인간존중의 개념을 적용할 수 있도록 구현함으로써 최종 AI 시스템이 윤리적으로 적합성을 보장받을 수 있게 추진한다.

첫째, AI 시스템 기획단계에서부터 공정하고 투명한 AI 개발을 위한 표준(ISO/IEC JTC 1/SC 42 등) 및 윤리적 AI 개발 가이드라인을 준수한다.

둘째, AI 설계 및 개발 단계에서는 기술 및 서비스 개발의 원천인 데이터 진위 여부에 대하여 자가지도 및 스스로 성장하는 신뢰가능 AI(Trustworthy AI) 기술 기반의 허위 정보 대응을 위한 AI 보안관 기술을 적용하여 허위 데이터를 판별하고 제거할 수 있도록 한다. 또한 가짜뉴스 및 딥페이크를 방지함으로써 신뢰사회 기반을 마련하기 위해 원본에 대한 가짜를 판별하고 초신뢰 기반 거래를 매개할 수 있는 블록체인 기반 암호화 증명 기술을 활용할 계획이다. 프라이버시 침해 방지를 위한 민감정보 등 개인정보의 보호 측면에서는 데이터 암호화 기술을 고도화하여 암호화된 데이터를 학습 가능하게 함으로써 데이터 활용성을 확대하고 AI 기술의 활용에 따른 역기능 방지기술을 선도한다.

셋째, AI 기술의 보급 및 확산 단계에서는 IoT에서 발생되는 실시간 데이터의 처리하고 클라우드 기반의 서비스 구현을 위하여 악의적 공격 · 위협에 대응하기 위한 AI 화이트 해커 기술과 AI 클라우드 기반 백신 기술을 개발하고 사이버 위협 공격 방어를 위한 실증 프레임워크를 구축하여 새로운 보안 위협 증대에 대응 및 안전한 AI 이용환경 구축을 통해 AI 기술 활용을 촉진한다.

용어해설

딥페이크(DeepFake) AI 기술을 이용해 유명인 등 특정인의 얼굴 등을 합성하여 만든 가짜 사진이나 영상 합성물('Deep learning'과 'Fake'의 합성어)

인포데믹(Infodemic) 잘못된 정보가 전염병처럼 급속히 확산되어 바로잡기 어려운 혼란을 야기하는 현상('Information'과 'epidemic'의 합성어)

약어 정리

ASD	AhnLab Smart Defense
DB	Database
IoT	Internet of ings
MediFor	Media Forensics
WEF	World Economic Forum

참고문헌

[1] 김정현, "코로나 사태로 온국민이 가짜뉴스 문제점 실감…전문가 해법은?" 뉴스1, 2020. 3. 12.

[2] H. Ajder et al., "The State of Deepfakes: Landscape, threats, and impact," Deeptrace, Sept. 2019.

[3] 국경완, 공병철, "인공지능을 활용한 보안 기술 개발 동향," 주간기술동향, 1913호, 2019. 9. 11.

[4] 청와대, "국가 사이버 안보 전략," 2019. 4., pp. 6-7.

[5] WEF, "Global Risks Report 2017," 2017, p. 44.

[6] 이호연, "[AI의 역습②] 해외서는? '검색 조작' 벌금 받은 구글, 윤리 실현 앞장," 데일리안, 2020. 10. 15.

[7] 최창현, "페이스북 AI, 딥페이크 꼼짝마라!…100,000개 딥페이크 식별 '데이터 세트' 공개한다," 인공지능신문, 2020. 6. 14.

[8] 정은령, 고예나, "인터넷 신뢰도 기반 조성을 위한 정책방안 연구," 서울대학교 산학협력단, 방송통신위원회 방통융합정책연구 KCC-2018-26, 2018. 12., p. 111.

[9] 이대성, "차세대 사이버 보안 기술 동향," 주간기술동향, 1916호, 2019. 10. 2., p. 9.

[10] 오윤경 외 3인, "혁신과 위험관리: 사람중심 기술혁신을 위한 추진과제," 한국행정연구원, 2020. 6., p. 17; 이순기(국회입법조사처), "인공지능의 윤리적 사용을 위한 개선과제," 이슈와 논점 제1759호, 2020. 9. 25. 재인용

[11] 김정민, "인공지능 윤리 이슈와 교육과정 동향," SW 월간 중심사회 2019년 7월호, 소프트웨어정책연구소.

[12] OECD, "G20 Ministerial Statement on Trade and Digital Economy," 2019. 6., p. 11.

[13] IEEE, "Ethically Aligned Design, First Edition," 2019.

[14] 전종홍(ETRI) 외, "감염병 재난에 대응하기 위한 의료인공지능의 기술표준동향," ETRI Insight 표준화동향 2020-01, 2020.