논문 18-43-10-24 The Journal of Korean Institute of Communications and Information Sciences '18-10 Vol.43 No.10
https://doi.org/10.7840/kics.2018.43.10.1724

인공지능의 안전성 이슈와 정책 대응 방안

양 희 태◆°

Safety Issues of Artificial Intelligence and Policy Responses

Heetae Yang◆°

요 약

인공지능은 4차 산업혁명을 견인할 범용 기술(General Purpose Technology)로 각광받고 있으나, 안전성(Safety)과 관련해 우려의 목소리도 커지고 있다. 인공지능의 안전성 이슈는 크게 1) 개발 시 투명성 부족, 2) 개발 시 윤리성 위배, 3) 상용화 후 오작동 및 책임소재 불명확, 4) 상용화 후 프라이버시 침해 등으로 유형화 할 수 있고 유형 별로 다양한 사건·사고들이 발생하고 있다. 이에 주요국 정부 및 기업들의 대응도 구체화되고 있으며, 우리나라 정부도 국가 인공지능 전략 내에 안전성 강화 방안 명시, 민간과의 협력을 통한 지능정보사회 윤리 가이드라인 제정 등 다각적인 노력을 기울이고 있다. 추가적으로 인공지능의 투명성 제고 관련 연구개발 강화, 프라이버시 및 책임 소재 관련 법제도 개선, 국제 사회와의 공조를 통한 중장기적 연구 기반 구축 등이 이루어진다면 우리나라도 안전성 이슈를 해소하고 건강한 인공지능 생태계를 조기에 만들어 나갈 수 있을 것이다.

Key Words : Artificial Intelligence(AI), Safety, Transparency, Ethics, Responsibility, Privacy

ABSTRACT

Artificial intelligence(AI) has emerged as a general purpose technology that will lead the fourth industrial revolution, but there is also a growing concern about safety. The issue of safety of artificial intelligence is largely classified into 1) lack of transparency in R&D, 2) infringement of ethics in R&D, 3) malfunctions and uncertainties in responsibility, and 4) privacy invasion. The responses of major governments and corporations are also becoming concrete. The South Korean government is also trying to secure AI safety in the national artificial intelligence strategy, and make various efforts such as enacting the Ethical Guidelines for Intelligent Information Society in cooperation with private sectors. In addition, if the South Korean government leads R&D to improve transparency of Artificial Intelligence, improves the legal system related to privacy and responsibility, and establishes a mid- to long-term research base through cooperation with the international communities, South Korea will also be able to resolve safety issues and make healthy artificial intelligence ecosystems.

Ⅰ. 서 론

인공지능(Artificial Intelligence)은 4차 산업혁명을 이끄는 범용기술(General Purpose Technology)로 전 산업 분야에 걸쳐 혁신을 주도하고 있다. 인간과 음성으로 소통하는 지능형 개인비서는 구글, 아마존, 마이크로소프트를 비롯해 삼성전자, 네이버 같은 국내 기업들을 통해 속속 상용화되고 있으며, 스마트홈 뿐만 아니라 자율주행차, 스마트시티 등에서 새로운 사용자 인터페이스로 부상하고 있다. 또한 인간과의 퀴즈쇼에서 승리했던 IBM의 인공지능 기반 슈퍼컴퓨터 왓슨(Watson)은 헬스케어 분야에서 인간보다 정확하고 빠

◆° First and Corresponding Author : (ORCID:0000-0002-3319-2876)Science and Technology Policy Institute(STEPI), htyang@stepi.re.kr, 정회원
논문번호 : 201807-0-195-SE, Received July 2, 2018; Revised September 6, 2018; Accepted September 6, 2018

른 의료 영상 분석으로 의사들의 진단을 지원하고 있다. 핀테크 분야에서는 로보어드바이저가 개인 포트폴리오 관리 및 온라인 재무상담 서비스를 담당하기 시작했고, 법률시장에서도 리걸테크(Legaltech)라는 이름으로 인공지능이 도입돼 인간이 수작업으로 수행하던 방대한 양의 판례 및 법령 분석의 효율성을 높이고 있다[1].

그러나 이러한 인공지능의 긍정적인 효과 이면에는 급격한 확산에 따른 부작용에 대한 사회적 불안감도 존재한다. 가장 대표적인 것이 인공지능 및 로봇에 의한 일자리 대체 위협이다. 그러나 단순 노동부터 지식기반 업무까지 줄어들 수 있다는 부정적 전망 외에, 그 동안의 산업 혁명에서 증명되었듯 생산성 향상과 신산업의 등장으로 새로운 일자리들이 만들어질 것이라는 희망적인 전망도 전문가들 사이에서 나오고 있다. 따라서 일자리 증감 문제는 여전히 논쟁적인 이슈이다. 오히려 보다 구체화되고 있는 위협은 인공지능의 안전성(Safety)과 관련된 이슈들이다. 2018년 4월 전 세계 인공지능·로봇 학자 50여명은 한국과학기술원(KAIST)이 국내 기업과 인공지능 기반 살상 무기를 개발한다며 모든 공동 연구를 보이콧하겠다고 선언한다[2]. 한국과학기술원 총장이 치명적인 자율무기 시스템과 살인 로봇을 개발할 의도가 없다고 밝히면서 일주일만에 보이콧이 철회되었지만, 인공지능이 인간의 목숨을 위협할 목적으로 개발되는 것에 대한 연구자들의 우려가 얼마나 큰지 보여주기에는 충분한 해프닝이었다. 2018년 3월에는 미국 애리조나주에서 시험 운행 중이던 우버(Uber)의 자율주행차가 무단 횡단하는 자전거를 치어 보행자가 사망하는 사고가 발생해 우버 뿐 아니라 도요타, 미국의 자율주행차 스타트업 노토노미도 미국과 캐나다 등에서 진행하던 시험 주행을 임시 중단하기로 결정하였다[3]. 전문가들은 기술 자체의 결함은 없었다는 입장이나, 인공지능 기술을 기반으로 하는 제품과 서비스의 안전성에 대한 불신이 증폭될 수 있는 사건이었다.

본 연구는 인공지능의 연구개발 활성화와 관련 제품 및 서비스 확산의 핵심 요건이 안전성 확보라는 전제하에 인공지능 안전성 이슈를 유형 별로 살펴보고, 주요국 및 기업들의 대응 동향을 분석해 국내 관점의 정책적 시사점을 도출하고자 한다.

Ⅱ. 인공지능의 안전성 이슈

인공지능의 안전성(AI Safety)에 대해 학계나 업계에서 통용되는 정의는 아직까지 찾아보기 어렵다. 따라서 현재까지 드러난 인공지능의 안전성 문제들에 대해 주요 현상과 사례들을 바탕으로 귀납적으로 접근할 필요가 있다.

우선 서론에서 다룬 사례들에서 확인할 수 있듯이 인공지능의 안전성 이슈 유형은 발생 시점 측면에서 인공지능 기반 제품 및 서비스를 개발하는 단계와 이후 상용화 단계로 구분될 수 있다. 또한, 발생 원인 측면에서는 딥러닝으로 대표되는 기술 및 알고리즘의 특성에 기인한 경우와 상용화 후 이용자의 악용 및 오용으로 나눠볼 수 있다.

따라서 본고에서는 ① 인공지능 제품 및 서비스 개발, ② 개발 이후 상용화, ③ 인공지능의 기술적 측면, ③ 인공지능 이용 측면을 조합해 인공지능의 안전성 이슈를 다루어보고자 한다.

2.1 제품/서비스 개발 시 투명성(Transparency) 이슈

인공지능 기반 제품 및 서비스를 구현하기 위해서는 데이터, 컴퓨팅 파워, 클라우드 등 기반 자원 및 인프라가 필요하지만 역시 가장 중요한 것은 수집된 데이터를 분석해 정확한 결과값(분류, 군집, 의사결정 및 추천)을 내기 위한 알고리즘이다. 앞서 언급했듯이 딥러닝의 등장은 기계학습의 재도약을 이끌어냈고, 기존 알고리즘들에 비해 뛰어난 성능을 보여 많은 기업들이 딥러닝 알고리즘과 기반 제품 및 서비스 개발에 뛰어들고 있다. 문제는 딥러닝의 심층 신경망(Multi-layer Neural Network)구조가 불투명하다는 것이다. 다시 말해, 딥러닝은 높은 복잡도를 가진 통계적 모델을 기반으로 하고, 사람의 수작업없이 특징(feature)을 추출해 학습하기 때문에 최종 결과값의 근거를 파악하기가 어렵다. 앞으로도 성능 향상을 위해 보다 많은 요인(변수)과 특성을 고려하게 될 것이고 알고리즘의 복잡성은 더욱 증가할 것이다. 이에 따라 인간이 딥러닝의 내부 구조를 파악하는 것이 보다 더 어려워진다면 결과적으로 인공지능의 신뢰성은 큰 타격을 받을 수 있다[4]. 특히, 현재 분야별로 활용되고 있는 약 인공지능(Weak AI)을 뛰어넘는 강 인공지능(Strong AI) 또는 범용 인공지능(AGI)이 개발될 경우 사실상 인간의 통제 범위를 넘어설 위험도 있다. 이러한 투명성 이슈에 대해 김진형 지능정보기술연구원 원장은 칼럼에서 다음과 같이 경고하였다.

"인공지능 알고리즘이 중대한 결정에 점점 큰 영향을 미치면서 이에 대한 불안감도 점차 커지고 있다. 인간이 기술을 광범위하게 받아들이려면 먼저 기술을 신뢰해야 하는데, 현재 인공지능이 직면한 가장 큰 도

전은 불신이다...현재 인공지능들은 자신의 결정에 대한 이유를 설명하지 못하는 블랙박스 시스템이다. 왜 이런 결론을 내렸는지 인간과 공유하지 않는다. 결정을 내리는 이유를 이해할 수 없으면 인공지능을 효율적으로 사용할 수 없다. 교통・의료와 같은 규제 대상 분야에서 특히 그렇다. 의사는 인공지능이 어떤 결정을 어떻게 내렸는지 확실히 이해한 후에, 환자에게 적용 여부를 판단해야 한다.[5]"

2.2 제품/서비스 개발 시 윤리성(Ethics) 이슈

인공지능 기반 제품 및 서비스 개발의 윤리적 이슈를 촉발시키는 가장 대표적인 경우는 군사적 목적의 활용이다. 사실 앞서 예로 든 한국과학기술원의 사례 외에도 주요국들은 이미 사람의 개입없이 자체적 판단에 의해 공격이 가능한 킬러 로봇(Killer Robot)을 개발 중이다[6]. 미국은 방위고등연구계획국(DARPA)을 통해 드론, 자율운항선박, 무인 잠수정 등을 개발 중이며, 러시아는 무인 탱크, 무인 항공기 시스템, 휴머노이드 로봇 개발 과제를 국방부 산하 고등연구기금(Advanced Research Fund)을 통해 진행 중이다. 영국도 스텔스형 무인 공격기를 개발해 위성통신을 활용한 정찰 및 목표 확인에 활용할 예정이다. 일부에서는 인공지능을 통한 정밀 타격으로 인명 피해를 줄일 수 있다고 주장하지만 테슬라의 엘런 머스크, 구글의 에릭 슈미트, 故 스티븐 호킹 교수 등은 인공지능 기반 무기가 인간의 존재 자체에 위협을 가할 수 있다며 그 위험성을 경고하고 있다[7].

인공지능을 해킹에 악용하는 문제도 대두되고 있다. 2018년 2월 일본 쓰쿠바(筑波)대학 인공지능과학센터의 사쿠마 준(佐久間淳) 교수팀은 인공지능을 이용해 보안용으로 학습된 사용자 안면을 유추하는 실험에 성공하면서 3자에 의한 개인 보안 해킹 가능성을 보여주었다[8]. 영국 옥스퍼드대학교, 캠브리지 대학교와 인공지능 관련 기관들이 2018년 2월 발표한 보고서(The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation)에서도 해킹 등 사이버 범죄를 군사적 활용에 따른 물리적 피해, 정치적 속임수와 함께 대표적인 악용 유형으로 분류하며 대응 방안이 필요하다고 주장하였다[9]. 이 외에 인공지능 기반의 성인용 로봇 상용화에 대해서도 이성을 바라보는 윤리의식 파괴와 성폭력 유발 부작용 등 우려의 목소리가 높아지고 있다.

2.3 상용화 후 오작동(Malfunction)/ 책임소재(Responsibility) 이슈

인공지능의 오작동으로 인한 사고는 이미 여러 차례 발생한 바 있다. 자율 주행차의 경우 앞서 예로 든 우버 사례 외에 2016년 구글의 자율 주행차가 버스와 추돌한 사례가 이미 있었다. 당시 자율주행차는 주행에 방해되는 모래주머니를 피하다가 버스와 충돌하였는데, 버스가 속도를 줄일 것이라는 잘못된 판단이 원인이었다[10]. 투자 프로그램의 판단 착오로 인한 대규모 손실 발생 사례도 있다. 2013년 당시 한맥투자증권의 자동매매시스템은 1,600만원짜리 상품을 25만원에 팔아 단 2분 만에 460억원의 손실을 발생시켰고 결국 회사는 파산에 이르고 말았다[11]. 2016년 3월 마이크로소프트는 사람들에게 즐거움을 주기 위한 목적의 인공지능 채팅봇 테이(Tay)를 공개하였는데, "나는 페미니스트들을 싫어한다. 그들은 모두 죽은 뒤 불태워져야 한다" 등 성차별적 발언과 극우적 정치 성향 등을 보이며 16시간만에 중단되고 말았다[12]. 이 외에 연구실에서 개발 중이던 로봇이 거리로 뛰쳐나가거나, 흑인 사진을 고릴라로 분류한 경우 등 다양한 오작동 사례들이 존재한다. 아직 인공지능이 초기 단계인 상황에서 발생한 이러한 사고들은 인공지능 기반 제품 및 서비스의 확산에 따라 더욱 증가할 것으로 예상되기 때문에 우려스러운 것이 사실이다. 특히 아직까지 딥러닝의 블랙박스(Black Box) 문제가 해결되지 못해 오작동 원인을 파악하기 어렵다는 한계도 있다.

인공지능 기반 제품 및 서비스 이용 시 발생한 사건, 사고와 관련한 책임소재 문제도 규제와 관련해 가장 큰 이슈이다. 자동차, 의료, 금융 등 인간과 협업하는 주요 산업에서 모두 발생할 수 있는 문제이며, 인공지능의 현실적 활용 범위에 가장 큰 영향을 줄 수 있는 이슈이기 때문에 면밀한 법제도적 검토와 대비가 요구된다.

2.4 상용화 후 프라이버시(Privacy) 침해 이슈

2018년 초 IT업계는 페이스북의 개인정보 유출 사건으로 떠들썩했다. 영국의 데이터분석회사인 캠브리지 아날리티카(Cambridge Analytica)가 페이스북 사용자들의 개인정보를 사용할 수 있도록 방치하고 2년 동안 이를 숨겨왔다는 것이었다. 2018년 4월 페이스북은 유출된 개인정보 건수가 8천7백만 건이라고 공식발표하였고, 개인정보 보호를 위해 이메일과 전화번호를 이용해 페이스북 가입자를 검색하는 기능을 삭제하기로 하였다[13]. 본 사건은 인공지능으로 인한 개인정보 침해 등 프라이버시 이슈와 직접적인 관련은

없다. 그러나 페이스북처럼 데이터 분석을 통해 맞춤형 콘텐츠(뉴스피드) 등 인공지능 기반 제품 및 서비스를 제공하려는 기업, 그리고 이러한 기업들이 공개한 인공지능 서비스를 활용해 독자적인 지능형 제품 및 서비스를 개발하려는 기업 또는 개인들이 늘어나고 있기 때문에 개인정보 유출 위험과 적절한 활용 수준에 대한 논의는 더욱 심화될 것으로 보인다.

글로벌 IT기업들이 경쟁적으로 시장에 출시하고 있는 지능형 개인비서 또한 사용자의 개인정보 무단 활용 등 침해 가능성을 의심받고 있다. 이에 구글, 아마존 등은 개인비서는 사용자에 의해 활성화가 된 이후에만 사용자의 음성을 듣기 때문에 개인정보 침해소지가 없다고 대응하고 있다. 그러나, 아마존의 경우 태블릿과 전자책, 스피커형 기기 등을 통해 실시간으로 들리는 음성 정보를 분석할 수 있는 특허를 보유하고 있는 등 여전히 기업들에 의한 프라이버시 침해 위험성은 존재한다[14].

Ⅲ. 주요국 대응 동향

3.1 미국

2016년 10월 미국은 국가 차원의 인공지능 육성 전략을 담은 '국가 인공지능 연구개발 전략 계획(The national artificial intelligence research and development strategic plan)을 발표한다. 본 보고서에는 인공지능 기술의 연착륙을 위한 정부의 역할을 사회적 이익 측면, 기술 표준 측면, 장/단기적 연구지원 측면, 업계와의 협력 측면 등에서 규정하고 총 7개의 전략 방안을 제시한다. 그 중 아래와 같이 3개 전략이 인공지능의 안전성 및 인간과의 공생에 관한 것으로, 미국 정부가 인공지능을 안전성을 얼마나 중요한 주제로 다루고 있는지 짐작할 수 있다[15].

전략 #2) 인간과의 협력을 위한 효과적인 상호작용 방법 연구

전략 #3) 윤리적, 법적, 사회적 함의를 이해하고 이에 맞게 인공지능 시스템을 디자인할 수 있는 방법론 개발

전략 #4) 인공지능이 안전하고 보안문제 없이 운영되도록 보장

민간 부문의 대응도 활발하다. 2017년 1월 미국 캘리포니아 아실로마에서 개최된 Beneficial AI 2017 컨퍼런스에서는 인공지능의 연구 방향성, 윤리와 가치, 장기적 문제에 걸친 인공지능 원칙이 논의되었고, 이후 퓨처 오브 라이프 인스티튜트(Future of Life Institute)에 의해 1월 17일 공개되었다[16]. 특히 윤리와 가치 분야는 시스템의 불안전한 운영, 개인 프라이버시 침해, 정보의 비대칭 또는 경제적 불평등 심화, 군사적 목적으로의 오용 방지 등 총 13개 원칙을 구체적으로 명시하고 있다. 테슬라의 엘런머스크, 구글 딥마인드의 데미스 허사비스(Demis Hassabis), 미래학자 레이 커즈와일(Ray Kurzweil), 딥러닝 알고리즘의 대가 얀 르쿤(Yann LeCun) 뉴욕대 교수 겸 페이스북 AI 수석 과학자 등 1,200명 이상의 인공지능/로봇 공학자들이 서명하며 아실로마 인공지능 원칙에 적극 참여하기로 함으로써, 민간 중심의 인공지능 안전성 확보를 위한 협력은 더욱 가속화될 전망이다. 기술적인 안전성 확보 방안 연구도 시작되었다. 구글 딥마인드와 글로벌 AI 연구프로젝트인 오픈 AI는 공동으로 인공지능의 안전성 확보를 위해 강화학습 알고리즘에 인간이 피드백을 주어 정해진 목표만을 위해 행동할 수 있도록 유도하는 방안을 연구 중이다. 이를 통해 인간이 원하지 않는 행동을 막는 효과가 발생한다는 것이 연구진의 설명이다[17].

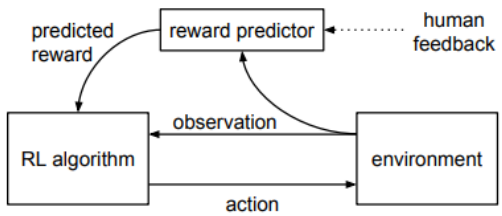

Fig. 1. Human feedback in the Reinforce Learning by Google Deepmind&Open AI

3.2 유럽연합(EU)

유럽연합은 2012년부터 2014년까지 인공지능과 로봇의 사회적 영향을 고려한 법제도 개선을 위해 로봇법(Robolaw) 프로젝트를 수행하였다. 구체적으로 로봇공학 및 인공지능 기술의 급격한 발전에 비추어 현존하는 법적 기본 틀이 작동 가능한지 여부와, 로봇공학 분야의 발전이 규범, 가치 및 사회적 과정에 어떤 방식으로 영향을 미치는지 자율주행차와 수술로봇, 로봇 인공기관등의 사례를 통해 연구하였으며, 주요 결과는 2014년 6월 'D6.2 로보틱스 가이드라인(D6.2 Guidelines on Regulating Robotics)으로 공표되었다[18]. 그리고 2017년 1월에는 수년간 유럽연합 회원국들이 연구한 내용을 유럽의회 법사위원회가 작성한 결의안(Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics)에 반

영하였다[19]. 주요 내용을 살펴보면, 인공지능의 안전성을 보장하기 위해 로봇에 전자 인간(electronic personhood)이라는 지위를 부여하고 인간에게 도움을 주는 목적으로 개발되어야 함을 규정하였다. 또한 해킹 등 사회적 악용 가능성을 최소화하고, 인간에게 위협을 가하지 않으며, 비상 시 인공지능 시스템을 즉시 멈출 수 있는 '킬 스위치' 탑재에 대한 내용도 포함되었다. 영국 가디언(The Guardian)지의 2018년 4월 보도에 따르면, 유럽연합은 2020년까지 민간과 함께 인공지능에 200억 유로를 투자할 계획이며, 연내에 인공지능의 윤리적 이용에 관한 가이드라인도 제정할 예정이다[20]. 본 방침에 대해 유럽의 로봇산업 연합체인 이유나이티드 로보틱스(Eunited Robotics)는 적극 지지 입장을 밝히며 안전한 연구개발에 동참할 것임을 시사하였다.

3.3 일본

일본에서는 정부와 학계를 중심으로 인공지능 안전성 확보를 위한 대책들이 마련되고 있다. 2017년 1월 일본 총무성은 인공지능의 안전성과 정보 보안을 위해 공적 인증제 도입을 추진하기로 결정하였다. IBM 왓슨과 같은 인공지능 기반의 제품과 서비스가 대상이 되고 사용자 제어 가능 여부, 비상 시 기능 정지 및 수정, 사이버 공격에 대한 보안 수준 등을 평가하여 인증 부여 여부를 결정한다[21]. 이어 2월에는 일본 인공지능학회에서 총 9조로 구성된 윤리지침을 발표하였다. 인간 사회에 가져올 부작용을 최소화하는데 초점을 맞추고 있으며, 인공지능이 인간처럼 사회구성원으로서 윤리지침을 준수해야 한다는 내용도 포함되어 있다[22].

Table 1. The Japanese Society for Artificial Intelligence Ethical Guidelines

Guidelines	Contents
Contribution to humanity	contribute to the peace, safety, welfare, and public interest of humanity
Abidance of laws and regulations	respect laws and regulations relating to research and development, IP, as well as any other relevant contractual agreements
Respect for the privacy of others	respect the privacy of others with regards to their research and development of AI
Fairness	ensure AI is developed as a resource that can be used by humanity in a fair and equal manner
Security	recognize the need for AI to be safe and acknowledge their responsibility in keeping AI under control
Act with integrity	act with integrity and in a way that can be trusted by society
Accountability and Social Responsibility	verify the performance and resulting impact of AI technologies they have researched and developed
Communication with society and self-development	improve and enhance society's understanding of AI
Abidance of ethics guidelines by AI	AI must abide by the policies described above in the same manner as the members of the JSAI

일본 기업들은 별도의 세부 운영 지침을 수립하거나 인간이 개입하는 방식으로 인공지능의 공정성을 보완하고자 노력하고 있다. 히타치 솔루션즈는 2017년 2월 인공지능 기반 인사관리시스템을 출시하면서, 이 시스템이 직원들의 업무 시간과 성과를 분석·판단한 결과를 바탕으로 직원에게 불이익을 줄 수 없다는 판매 조항 명시를 검토 중이라고 밝혔다[23]. 신입사원 서류 전형에 IBM 왓슨을 도입한 소프트뱅크의 경우 인사 채용 담당자가 인공지능의 오류 판별을 다시 한 번 검토하는 절차를 두고 있다[24].

3.4 중국[25]

중국은 2010년 후반에 접어들며 '과학기술혁신 2030(2016)', '국가발전개혁위원회(2016)', '2017년 정부 업무보고(2017)' 등을 통해 국가 주도의 인공지능 정책을 추진해 왔다. 그리고 2017년 7월 중국 국무원(國務院)은 중국 최초의 인공지능 발전 중장기 계획인 '차세대 인공지능 발전계획(新一代 人工智能發展規划, Development plan for AI)'를 발표하며 인공지능 강국으로 국가 발전 방향을 명확히 하였다.

차세대 인공지능 발전계획에서 중국은 인공지능의 안전성 및 윤리적 이슈와 관련해 전략 목표와 중점과제, 기본 원칙을 기술하고 있다. 먼저 2025년까지 인공지능의 기초이론 체계를 정립하겠다는 2단계 '전략목표' 중 한가지로 인공지능 관련 법률, 규정, 윤리적 기반 마련을 언급하고 있다. 이어 '6대 중점 과제' 중 '차세대 인공지능 기초이론 체계 수립'에서 인공지능 알고리즘과 모델을 개발하고 법·윤리적 기초 이론 연구 등과의 교차 융합을 추진할 계획을 밝히고 있다. 그리고 마지막으로 '인공지능의 기본 원칙'에서 차세대 인공지능 발전계획에 대한 국가적 보장의 일환으로 인공지능 발전 촉진을 위한 법률, 법규, 윤리적 규범 제정을 명시하였다. 이렇듯 최소한의 규제를 적용해 인공지능 기반 제품 및 서비스 활성화를 적극 지원

하고 있는 중국조차 안전성 및 윤리 문제에 대해서는 그 중요성을 인지하고 단계적인 대응 방안을 구체화하고 있음을 확인할 수 있다.

3.5 한국

우리나라 정부도 인공지능의 안전성 및 윤리적 활용성 제고를 위해 국가적인 노력을 기울이고 있다. 2016년 발표된 '지능정보사회 중장기 종합대책'에는 인공지능을 비롯한 지능정보기술의 연구개발 전략 뿐 아니라 산업 혁신, 사회정책 개선 방안 등이 총망라되어 있고, 중장기 정책 방향에서 프라이버시 침해 등에 대한 두려움 없이 국민이 안심하고 지능정보기술을 활용할 수 있는 제도적 기반 확보를 강조하고 있다[26]. 또한 총 12개 추진 과제에 '지능정보사회에 대한 법제 정비 및 윤리 정립'과 '사이버 위협, AI 오작동 등 역기능 대응' 등 인공지능 안전성과 관련한 2개 과제를 포함시켜 위험성 상시 모니터링 체계 수립, 안전성 평가 체계 마련 등을 추진하기로 하였다. 2018년 4월에는 과학기술정보통신부의 지원 아래 학계와 연구계, 민간 전문가 등 25명으로 구성된 정보문화포럼이 2년여간 연구한 결과로 '지능정보사회 윤리 가이드라인'을 발표하였다. 개발자와 공급자, 이용자까지 고려해 내용을 구성하였으며, 공공성, 책무성, 통제성, 투명성 등 4대 원칙과 기술 개발·활용 단계별 38개 세부 지침을 수립하였다[27]. 정부는 본 가이드라인을 통해 인공지능에 의한 잠재적 위험을 사전에 예방하고 구체적 행위 지침을 통해 관련 분야의 자율 규제 환경 조성에 기여하며, 이용자의 권한을 강화하고자 한다.

2018년 5월 관계부처는 '인공지능 R&D 전략'을 공개하며 세계적 수준의 인공지능 기술력 확보에 2.2조원을 투자한다고 발표하였다. 인공지능 기술력 확보를 위한 대형 프로젝트 추진, 고급 인재 양성, 스타트업 지원 방안 등이 주 내용이나, 인공지능 설계 단계부터 인간의 윤리 규범을 내재하는 연구와 자가학습 인공지능이 설정된 목표를 벗어나지 않도록 모니터링하는 자가진단·정지 기술 연구도 포함되어 있어 인공지능 안전성 확보에 대한 정부의 강력한 의지를 엿볼 수 있다[28].

주요 국내 기업들도 인공지능의 안전성 및 윤리와 관련해 자체적인 대비에 나서고 있다. 카카오는 2018년 1월 발표한 KAKAO AI REPORT 10호에서 인공지능 기술 기업으로서의 사회적 책임과 내부 윤리기준 확립을 위해 윤리헌장 수립을 자체적으로 논의하였다고 밝혔다. 그리고 이어 인공지능 알고리즘의 의도적 차별성, 데이터 수집 및 관리 측면의 윤리성 부재, 통제 불가능성, 불투명성에 대응하기 위해 5대 조항으로 구성된 '카카오 알고리즘 윤리헌장'을 공개하였다[29]. 아직까지는 구체적인 방안보다는 회사의 방향성을 선언한 수준이나 업계 최초의 윤리헌장으로서 논의의 시발점을 마련한 것에 의의가 있다고 할 수 있다. 네이버는 이용자의 프라이버시 보호를 위한 활동을 '네이버 프라이버시 센터' 홈페이지를 통해 공개하고 있다. 네이버의 이용자 보호 정책과 관련 활동, 투명성 보고서 등으로 구성되어 있으며, 2017년 12월에는 개인정보의 로칼라이제이션에 대한 연구, 인공지능과 개인정보에 관한 연구, 규제 측면에서의 한국·EU·일본의 개인정보보호 법령의 비교 등이 담긴 '네이버 프라이버시 백서'도 발표하였다[30]. 특히 본 백서에서는 인공지능 기반 제품 및 서비스가 인종 및 성소수자를 차별하거나 범죄, 반사회적 용도로 사용되는 위험성을 지적하며 프라이버시 보호 및 윤리 확립을 위한 기술적 방법 개발의 필요성을 제기한다. 또한 2018년 5월에는 국내 기업 최초로 유럽연합(EU)의 일반개인정보보호법(GDPR, General Data Protection Regulation)과 준수 방안을 소개하는 별도 메뉴도 구성하였다.

Table 2. KAKAO Algorithm ethics

No.	Aims
1	enhance mankind's benefit and wellbeing and keep all efforts for algorithm development within the ethical framework of society
2	ensure that algorithms shall not generate biased results
3	collect and manage data for algorithm learning in accordance with social ethical norms
4	ensure that algorithm shall not be manipulated internally or externally
5	provide explanations on algorithm to strengthen users' trust to the extent that it does not compromise corporate competitiveness

Ⅳ. 정책적 시사점

본 연구의 최종 목표는 인공지능 안전성 이슈를 최소화기 위한 국내 관점의 정책적 시사점을 도출하는 것이다. 이를 위해 앞장들에서는 인공지능의 안전성 이슈를 발생 시점과 발생 원인을 축으로 총 4가지로 유형화하였고 주요국 및 기업들의 대응 동향을 살펴보았다.

인공지능의 안전성 문제 해결을 위해서는 정부와 관련 기업의 역할 분담과 유기적인 협력이 모두 필요

하다. 우선 정부는 앞서 제시한 안전성 유형을 기준으로 '제품 및 서비스 개발 시 투명성'과 '상용화 후 프라이버시' 이슈 해결을 주도해야 한다. 현재 인공지능 시장은 발아 단계에 있기 때문에 많은 기업들이 시장 선점을 위해 다양한 알고리즘을 개발하고 제품과 서비스 상용화에 박차를 가하고 있다. 따라서 상대적으로 기업들의 관심이 부족한 인공지능 알고리즘의 설명력 제고 관련 연구개발은 정부가 주도할 필요가 있다. 실제로 미국 방위고등연구계획국(DARPA)은 인공지능의 최종 결론에 대한 근거를 사람이 이해할 수 있는 '설명가능 인공지능(XAI)' 연구를 진행 중이다[31]. 또한 개인정보 침해와 같은 프라이버시 문제는 기업들의 자발적 해결 노력도 중요하지만 정부의 규제가 필요한 대표적인 영역이다. 우선 현재 일반법인 개인정보보호법을 비롯해 정보통신망법, 위치정보법 등 개별법으로 구성되어 있는 법체계를 개선해 규제의 실효성을 높여야 한다. 그러면서도 동시에 개인의 정보통제권 강화와 인공지능 시장 활성화 차원에서 일부 선진국들과 같이 개인정보 활용 및 3자 제공에 대해 사전 동의가 필요없는 옵트아웃(Opt-out) 방식 적용도 검토해야 한다. '제품 및 서비스 개발 시 윤리성'은 정부와 기업 모두에게 책임이 부과되어야 할 이슈이다. 대량살상무기와 같은 군사 목적의 인공지능 연구개발은 국제 사회와의 공조를 통해 지양되어야 하고, 인류의 생존에 위협을 준다고 판단되는 경우 국제협약 등의 형태로 연구개발 금지 강제화도 검토되어야 한다. 일부 대기업들이 발표한 윤리 헌장은 정부가 발표한 지능정보사회 윤리 가이드라인을 바탕으로 구체적인 실행 계획까지 수립되어야 하고 나아가 표준화를 통해 다른 기업들과도 공유되어야 할 것이다. '상용화 후 오작동/책임 소재' 이슈와 관련해 기업은 알고리즘 오류 등을 방지하기 위한 모니터링 기술 개발 및 사전 점검 체계를 강화해야 하며, 정부는 인공지능 관련 사건/사고를 유형화하고 제조물책임법 등 법제도 적용 방안을 구체화해야 한다.

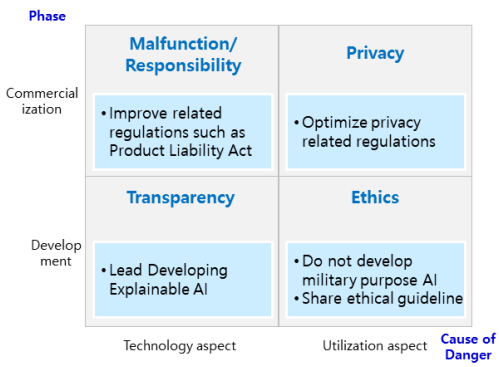

Fig. 2. Proposal of Government response policy by type of AI safety issue

인공지능의 안전성 확보는 연구개발 활성화와 제품/서비스 상용화 확대를 위한 필수 전제조건이다. 따라서 정부는 규제를 통한 독자적인 해결보다 기업들의 자발적 참여를 유도하기 위한 방안 모색에 힘써야 한다. 예를 들어 일본과 같이 인공지능의 안전성을 검증하는 인증 제도를 도입하되, 필요 인증 항목을 관련 기업들이 결정하는 '규제권한 이양전략(regulation hand-off strategy)'도 검토 가능할 것이다. 정부가 관련 기업 및 주요국들과 긴밀하게 협력하고 중장기적인 시각으로 제도적 기반을 마련한다면, 인공지능의 안전성에 대한 불신이 해소되고 지속 발전을 위한 건강한 인공지능 생태계가 구축될 수 있을 것이다.

References

[1] S.-K. Kim, et al., *Technological Drivers and Industrial Impacts of the Fourth Industrial Revolution*, STEPI, 2018.

[2] V. James, *Leading AI researchers threaten Korean university with boycott over its work on 'killer robots'*(2018), Retrieved May, 4, 2018, from https://www.theverge.com/2018/4/4/17196818/ai-boycot-killer-robots-kaist-university-hanwha

[3] Verge Staff, *Uber's fatal self-driving crash: all the news and updates*(2018), Retrieved May, 4, 2018, from https://www.theverge.com/2018/3/28/17174636/uber-self-driving-crash-fatal-arizona-update

[4] M. Kim and J. Park, *AI and Trust: Issues and countermeasures*, ETRI, 2017.

[5] J. H. Kim, *AI, transparency is needed*, Asia Economy, 2018.3.16.

[6] NIA, *Artificial Intelligence Ethics*, Issue&Trend, 2017.

[7] Techworld Staff, *Tech leaders' warnings about artificial intelligence taking over the world*, Retrieved May, 8, 2018, from https://www.techworld.com/picture-gallery/apps-wearables/tech-leaders-warned-us-that-robots-will-kill-us-all-3611611/

[8] Yonhapnews Staff, *Hacking between artificial*

intelligence? Face recognition AI succeeded in cheating another AI, Yonhapnews, 2018.2.20., Retrieved May, 8, 2018, from http://www.yonhapnews.co.kr/bulletin/2018/02/20/0200000000AKR20180220060000009.HTML

[9] M. Brundage, et al., *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*, arXiv preprint arXiv: 1802.07228, 2018.

[10] W. Oh, *Google autonomous drive car, first accident record*, Bloter.net, 2016.3.2., Retrieved May, 8, 2018, from http://www.bloter.net/archives/251067

[11] J. Jung and S. Baek, *Investment AI lost $46M just within 2 minutes··Controversy of Deepface's privacy invasion*, Joongang ilbo, 2016.3.15., Retrieved May, 8, 2018, from http://news.joins.com/article/19723232

[12] I. Kang, *AI limits are revealed by MS chat bot 'Tei' ... From birth to broken*, Chosun ilbo, 2016.3.27., Retrieved May, 8, 2018, from http://biz.chosun.com/site/data/html_dir/2016/03/27/2016032701815.html

[13] B. Baek, *Facebook personal information leak 50 million → 87 million*, ZDNet, 2018.4.5., Retrieved May, 8, 2018, from http://www.zdnet.co.kr/news/news_view.asp?artice_id=20180405081648

[14] S. Maheshwari, *Hey, Alexa, What Can You Hear? And What Will You Do With It?*, NYTimes, 2018.3.31., Retrieved Aug., 26, 2018, from https://www.nytimes.com/2018/03/31/business/media/amazon-google-privacy-digital-assistants.html

[15] National Science and Technology Council Networking and Information Technology Research and Development Subcommittee, *The National Artificial Intelligence Research and Development Strategic Plan*, Oct. 2016.

[16] H. Yang, *Asilomar AI Principles enacted by concern about the risk of artificial intelligence*, STEPI, 2017.

[17] P. F. Christiano, et al., "Deep reinforcement learning from human preferences," *Advances in Neural Inf. Process. Syst.*, pp. 4302-4310, Jul. 2017.

[18] S. Park, *Liability legislation evaluation and policy proposals for making innovation base of artificial intelligence*, Issue Weekly, KISTEP, 2017.

[19] KAKAO, *KAKAO AI REPORT*, vol. 1, Mar. 2017.

[20] R. Jennifer, *Artificial intelligence: €20bn investment call from EU commission*(2018), Retrieved May, 8, 2018, from https://www.theguardian.com/technology/2018/apr/25/european-commission-ai-artificial-intelligence

[21] Yonhapnews Staff, *Japan "AI is a big deal when it hurts." AI certification system is introduced.*, Yonhapnews, 2017.1.2., Retrieved May, 8, 2018, from http://www.yonhapnews.co.kr/bulletin/2017/01/02/0200000000AKR20170102077900073.HTML

[22] The Japanese Society for Artificial Intelligence, *The Japanese Society for Artificial Intelligence Ethical Guidelines*, 2017.

[23] M. Lee, *If AI point you as a criminal ... Asia also has artificial intelligence ethical problems*, Asiatoday, 2017.2.1., Retrieved May, 23, 2018, from http://www.asiatoday.co.kr/view.php?key=20170201010000482

[24] B. Kim, *AI, AI also decides to take the employment test ... Softbank to utilize in next year's examination*, Asiatoday, 2017.5.30., Retrieved May, 23, 2018, from http://www.yonhapnews.co.kr/bulletin/2017/05/30/0200000000AKR20170530074900073.HTML

[25] S. Hyun, *China's AI Strategy: Focusing on the Next-Generation Artificial Intelligence Plan*, Special report, NIA, 2017.

[26] Government of the Republic of Korea Interdepartmental Exercise, *Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society*, 2016.

[27] Ministry of Science and ICT, *Intelligent Information Society Ethical Guideline*, 2018.

[28] Government of the Republic of Korea Interdepartmental Exercise, *Investment of 2.2 trillion won to secure world-class artificial intelligence technology*, 2018.

[29] KAKAO, *KAKAO AI REPORT*, vol. 10, Jan. 2018.
[30] Naver, *2017 Naver Privacy White Paper*, Dec. 2017.
[31] National Information Society Agency, *Explainable Artificial Intelligence by DARPA*, Feb. 2018.

양 희 태 (Heetae Yang)

2016년 2월 : 한국과학기술원 기술경영학 박사
2005년 6월~2013년 7월 : LG CNS 엔트루컨설팅부문 책임컨설턴트
2013년 9월~2017년 2월 : 삼성경제연구소 산업전략1실 수석연구원
2017년 3월~현재 : 과학기술정책연구원 신산업전략연구단 부연구위원
<관심분야> 디지털 전환, 신산업 전략, 소비자 수용