# 인공지능 안전에 대한 미국과 유럽 동향

## Trends of AI Safety Policy in US and Europe

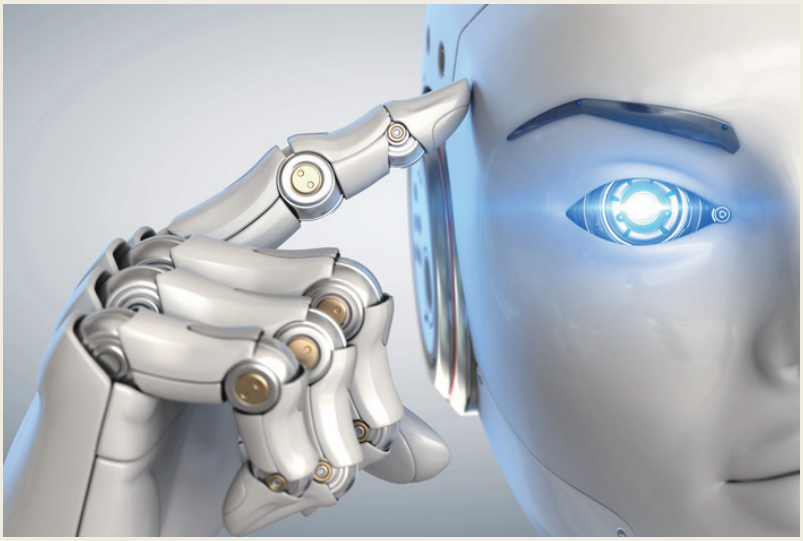
**진회승** CHIN, Hoeseung • **책임연구원** Principal Researcher, SPRi • hschin@spri.kr

미국 정부와 유럽연합에서는 트럼프 대통령의 행정명령, EC의 인공지능 윤리지침을 통해 인공지능 안전 연구에 투자하고 있다. 해외 대학과 연구소에서도 인공지능 안전 전략과 도구를 개발하고, 관련 연구를 추진하고 있다. 인공지능이 현실 세계에서 구현되기 위해서는 기술 발전과 병행하여 안전이 필수적이며, 국내에서도 인공지능 안전에 대한 준비가 필요하다.

The US government and the European Union are investing in AI safety research through the Trump's executive order and EC's AI ethics guidelines. Leading overseas universities and research institutes are also developing AI safety strategies and tools and promoting related research. In order for the true implementation of AI in the real world, safety considerations are essential in line with technological development, and preparation for AI safety is a must in Korea.

## 인공지능 시대의 도래와 안전 문제

그림 1 인공지능의 의미



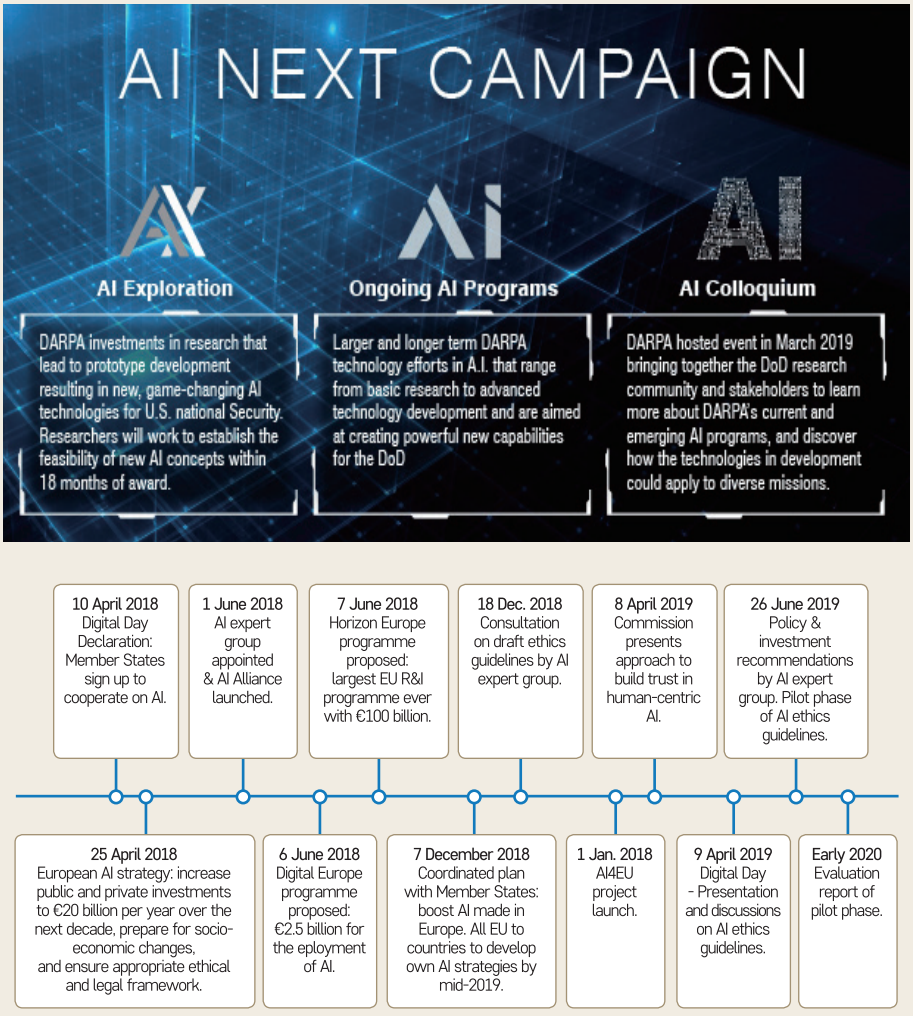
※ 자료 : The BCI 그룹(2018), Artificial Intelligence- A risk or a revolution?

인공지능 이론이 개발된 것은 컴퓨터 개발시기와 비슷하다고 해도 과언이 아니다. 그러나 지난 50여 년 동안 인공지능은 견고하지 못한 알고리즘, 컴퓨터 계산능력의 한계, 데이터 부족 등으로 실용화에 이르지 못했다. 현재는 이러한 문제점이 하나씩 해결되면서, 인공지능 구현 가능성을 긍정적으로 보고 있다. IDC에 따르면 세계 인공지능 소프트웨어 시장 규모는 연평균 35%의 성장을 통해, 2022년에는 335억 달러(약 39조)가 될 것으로 추정하고 있다.[1]

2018년 미국 방위고등연구계획국(DARPA)은 "AI Next Campaign"에 3년간 20억 달러(약 2.4조 원) 이상의 투자 계획을 발표하였다. DARPA는 20여 개의 연구 프로젝트를 운영하며, 데이터에 종속적인 머신러닝의 한계를 극복할 수 있는 기초연구와 언어 번역을 군사, 의료 등에 사용하기 위한 응용 연구들을 수행하고 있다. 같은 해 유럽연합집행위원회(EC)는 공공과 민간에서 2020년까지 인공지능에 200억 유로(약 23조 원)를 투자하겠다고 발표했다.[2] 이처럼 세계 주요국들은 미래사회를 더욱 발전시키고, 삶을 편리하게 만들 수 있는 인공지능 개발에 많은 연구비를 투자하며 박차를 가하고 있다.

1 IDC(2018.2.), SPRi(2019.4.), SW산업 주요통계
2 EU EC(2019.7.), Digital Single Market Policy Artificial Intelligence

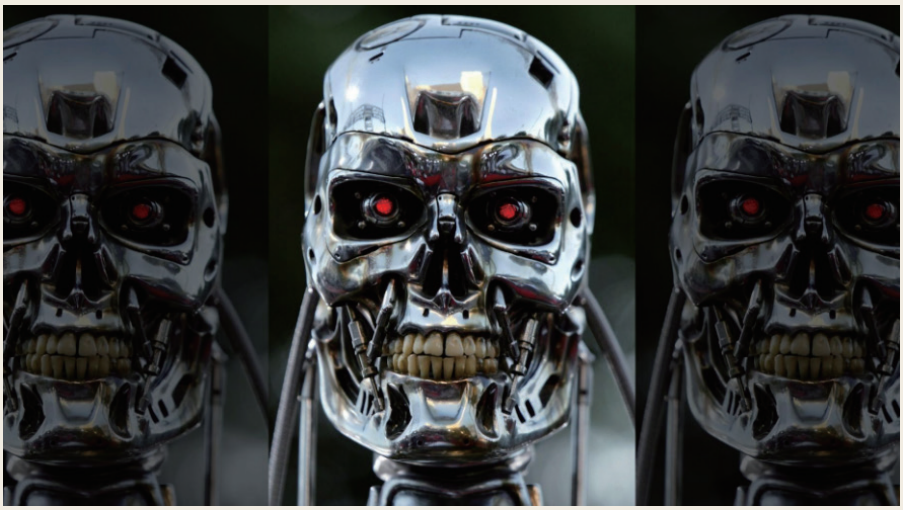
그림 2 미국 DARPA AI Next 캠페인




※ 자료 : DARPA(2018), AI Next 캠페인
※ 주 : AI 도전적인 과제 탐색, AI 연구 확대, AI 이해관계자 모임을 통한 기술 역량 제고

한편, 인공지능의 위험에 대한 우려가 높아지고 있는 것도 사실이다. 스티븐 호킹이나 유발 하라리 등이 인공지능의 위험성을 경고하는 대표적인 인물이다. 따라서 이에 대한 인식이 높아지며 인공지능 위험에 대비하기 위한 각국의 움직임이 빨라지고 있다. 트럼프 미국 대통령의 인공지능 진흥을 위한 행정명령에는 인공지능 안전표준 개발이 포함되어 있고, 유럽 EC가 발표한 인공지능 윤리지침에는 안전에 대한 검토를 포함하고 있다.

그림 3 인공지능 위험 경고



※ 자료 : fox news(2019.2.) / Mike Blake(Reuters), Is Skynet a reality? As Trump signs executive order on artificial intelligence, tech giants warn of danger

## 인공지능 안전에 대한 미국 정부와 EU의 논의 동향

트럼프 대통령은 2019년 2월 '인공지능 분야에서 미국의 리더십 유지'라는 행정명령을 발표하고 5대 원칙을 제시했다.[3] 트럼프는 5대 원칙 중 하나로 인공지능 시스템에 적합한 '기술 및 안전 표준'을 개발하도록 미국표준기술연구소(NIST)에 요청하였고 그에 따라 NIST는 '기술 표준 및 관련 도구 개발에 대한 연방 정부의 계획'초안[4]을 발표하였다. NIST 보고서에서 인공지능 안전을 위한 각 기관의 노력으로써 미국 교통국(DOT)의 '미래 교통의 준비 : 자율주행차 3.0(Preparing for the Future of Transportation : Automated Vehicles 3.0)'과 식품의약국(FDA)의 '인공지능/기계 학습(AI/ML)기반 의료기기 소프트웨어(SaMD) 수정을 위한 규제 프레임워크[5]'를 소개하고 있다.

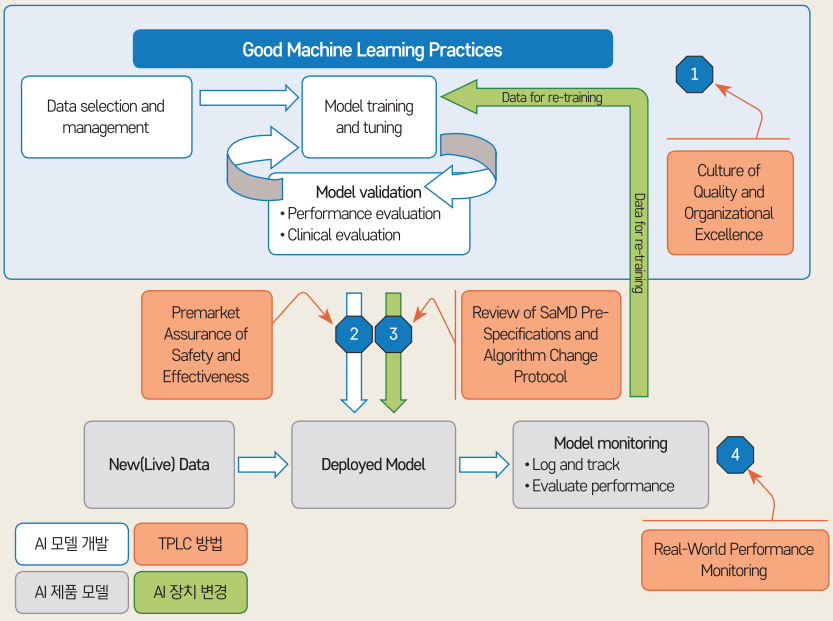
FDA 보고서를 살펴보면 질병의 치료, 진단, 완화 또는 예방하기 위한 목적의 AI/ML 기반 소프트웨어를 의료기기로 정의하고, 이를 'AI/ML기반 SaMD(이하 SaMD)'로 명명했다. 적절한

3 The White House Fact Sheets(2019.2.11.), President Donald J. Trump Is Accelerating America's Leadership in Artificial Intelligence
4 NIST(2019.7.2.), U.S. LEADERSHIP IN AI : A PLAN FOR FEDERAL ENGAGEMENT IN DEVELOPING TECHNICAL STANDARDS AND RELATED TOOLS DRAFT FOR PUBLIC COMMENT
5 FDA(2019.4.2.), Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device(SaMD)

규제 감독을 통해 안전하고 효과적인 소프트웨어 기능을 제공할 때 SaMD가 의료기기로서 역할이 가능하다고 한다. 안전한 SaMD 구현을 위해서 소프트웨어의 위험 수준을 정하고 위험이 일정 수준 이상인 소프트웨어의 사전검증이 시행된다. 사전검증은 소프트웨어 개발 전체 주기(TPLC, Total Product Life Cycle)에서 시행된다. TPLC 접근법은 시판 전 개발에서 시판 후 성과에 이르기까지 소프트웨어 제품의 평가 및 모니터링을 가능하게하고 해당 SaMD제품사의 탁월성에 대한 지속적인 검증을 가능하게 한다.

다음 그림은 안전한 의료 소프트웨어 제품을 개발하는 방법을 도식화한 것이다. 제조회사의 품질 및 안전 목표에 따라 인공지능 모델을 개발한다. 이러한 모델은 안전과 유효성에 대한 시판 전 검증과 제품 변경 시 재검증을 통하여 안전을 확보한 제품으로 개발된다. 개발된 제품은 시판 후에도 검증을 통하여 안전과 제품 성능을 확보한다.

그림 4 인공지능/머신러닝 개발에 관한 FDA의 전체 생명주기 방법 적용




※ 자료 : FDA(2019), Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML) based Software as a Medical Device(SaMD)
※ 주 : 1. 제품 품질에 대한 목표 설정, 2. 안전과 유효성에 대한 시판 전 검증, 3. SaMD와 알고리즘 변경에 대한 검증, 4. 현실 세계에서 검증

한편, 유럽연합집행위원회(EC)에서는 2019년 4월 '신뢰할 수 있는(Trustworthy)[6] 인공지능을 위한 윤리지침'을 발표했다.[7] 신뢰할 수 있는 인공지능을 개발하고 사용하기 위한 7가지 요구사항 중 하나가 기술적 '안전성'이다. 신뢰할 수 있는 인공 지능을 달성하는 데 중요한 구성 요소는 위해(Hazard) 예방을 구현하는 기술적인 견고성이다. 의도하지 않은 위험을 최소화하고 허용 가능한 수준으로 위험을 방지함으로써 인공지능 시스템이 안정적으로 작동되도록 개발되는 것이 기술적 견고성이다. 이를 통해 사람의 신체적, 정신적 안전을 보장한다.

EC 보고서는 인공지능 안전은 대체계획, 기술적 안전성·정확성·신뢰성(Reliability)·재현성을 구현함으로써 확보할 수 있다고 했다. 신뢰성은 정의된 입력과 상황 범위에서 올바르게 작동하는 것을 의미한다. 재현성은 인공지능 실험이 동일한 조건하에서 반복될 때 동일한 행동을 보이는지 여부를 나타낸다. 대체계획이란 인공 지능 시스템에는 문제가 생길 경우 대체가 가능한 안전장치를 추가하는 것이다. 이것은 인공지능 시스템이 통계에서 규칙 기반 절차로 전환하거나 행동을 계속하기 전에 인간 운영자의 개입을 요구한다는 것을 의미할 수 있다. 인공지능 안전 확보를 위해서는 종합적으로 시스템이 모든 생명체와 환경을 해치지 않고 작업을 수행하는지 검토되어야 하며 의도하지 않은 결과와 오류의 최소화를 구현해야 한다. 또한 다양한 응용 분야에 걸쳐 인공지능 시스템의 잠재적인 위험을 규정하고 평가하는 프로세스가 수립되어야 한다. 이러한 대체계획, 기술적 안전성, 정확성, 신뢰성, 재현성은 인공지능 시스템의 위험이 높을수록 완성도가 높아야 한다.

## 인공지능 안전에 대한 대학과 연구기관의 논의 동향

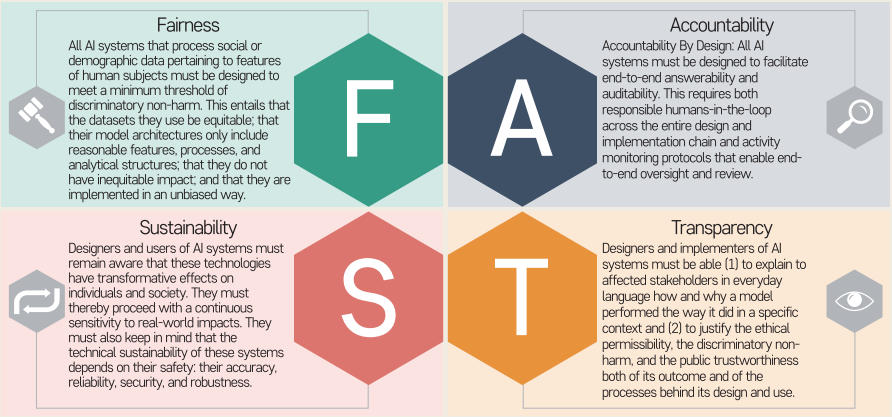
미국 스탠포드 대학은 인공지능안전센터(Center for AI Safety)를 운영하고 있다. 이 센터는 견고성 검증, 안전 중요 자동 시스템에 대한 검증 등의 연구를 시행하고, 'DNN 분석과 검증을 위한 Marabou 틀(2019)', '인간과 로봇의 상호 영향(2019)', '신경만 기반 항공기 충돌회피시스템의 안전 보장(2019)' 등의 인공지능을 적용한 시스템의 안전 확보에 대한 연구보고서를 발간하고 있다.[8] 또한 정형 검증, 안전한 로봇, 인공지능 안전 세미나 등 인공지능안전 관련 과목들을 컴퓨터과학과·항공우주학과 학부생, 컴퓨터과학과 대학원생 대상으로 운영하고 있다.

6 법률 및 규정을 준수하는 합법성, 윤리적 원칙과 가치를 준수하는 윤리성, 기술적·사회적 관점에서 신뢰성을 모두 포함
7 EC(2019.4.8.), Ethics Guidelines for Trustworthy AI
8 Guy Katz et al.(2019.7.), The Marabou Framework for Verification and Analysis of Deep Neural Networks; Kwon, M. et al.(2019.6.), Influencing Leading and Following in Human-Robot Team; Julian, K. D et al.(2019), Deep neural network compression for aircraft collision avoidance systems

영국 옥스퍼드 대학과 FHI(Future of Humanity Institute)에서 공동으로 운영하는 전략 인공지능 연구소(Strategic AI Research Center)는 인공지능이 안전하고 유익한지 확인하기 위한 전략과 도구를 개발하고 있으며, 일반 인공지능(AGI, Artificial General Intelligence)이 자체 보상 기능을 수행할 위험과 기능의 부작용으로 인한 위험을 연구한 '실수 없는 시도 : 인간 개입을 통한 안전한 강화학습 구현(2017), 존재하는 위험과 희망의 정의(2015)' 등의 안전 관련 보고서를 발간했다.[9]

영국의 앨런튜링연구소(The Alan Turing Institute)는 2015년에 케임브리지, 에든버러, 옥스퍼드 등 5개 대학과 영국의 공학과 물리과학 연구위원회가 설립하였다. 이 연구소는 2018년 8개 대학이 새로 참여하여 데이터 과학과 인공지능에 대해 연구하고 있다. 앨런튜링연구소에서 발간한 '인공지능 윤리와 안전의 이해(2019)'[10] 지침서에는 인공지능 시스템으로 인해 발생할 수 있는 잠재적인 위험을 식별하고 방지하기 위한 구체적이고 운영 가능한 조치를 제안하고 있다. 이 지침서는 공공부문 조직이 혁신 문화를 이해하고 윤리적이고 공정하고 안전한 인공지능 시스템의 설계 및 구현을 지원하는 거버넌스 프로세스를 수립함으로써 이러한 잠재적 위험을 예견하고 예방하려는 의도로 제작된 것이다.

그림 5 FAST 원칙




※ 자료 : 앨런튜링연구소(2019), 인공지능 윤리와 안전의 이해

9 William Saunders et al.(2017), Trial without Error: Towards Safe RL with Human Intervention, Owen Cotton-Barratt & Toby Ord(2015), Existential Risk and Existential Hope: Definitions

10 Leslie,D.(2019), Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute

지침서에서 제안하고 있는 FAST(Fairness, Accountability, Sustainability, and Transparency) 원칙은 인공지능 기술의 견고한 설계와 사용을 위한 실용적인 원칙이다. FAST 중 “S”인 기술적 지속가능성(Sustainability)의 확보를 위해서는 안전성, 정확성, 신뢰성, 보안, 견고성이 보장되어야 한다는 주장은 EU 보고서와 같은 맥락으로 파악된다.

인공지능 시스템은 불확실성과 변동성이 심한 현실 세계에서 작동하기 때문에 안전한 인공지능을 구축하는 것은 어렵다. 그러나 앨런튜링연구소는 프로젝트에서 “안전하고 신뢰할 수 있는 인공지능을 제작하는 목표”를 세우는 것만으로도 실제 세계의 알려지지 않았고 예상 못했던 사건에 직면했을 때 나타날 수 있는 시스템의 위험을 감소시킬 수 있다고 한다. 특히 안전에 대한 위험은 강화학습에서 나타나는데, 강화학습이 충분한 제어가 없이 이루어지기 때문이다. 지침서에서는 강화학습의 위험을 제거하기 위한 전략으로 다음을 제안하고 있다. 첫째 테스트 단계에서 광범위한 시뮬레이션 실행을 통해 도출된 제약 조건을 시스템에 프로그래밍 한다. 둘째 시스템을 해석하기 쉽게 설계하여, 시스템 안전을 평가하기 용이하게 한다. 셋째 인위적으로 시스템을 종료하는 메커니즘을 포함한다. 인공지능 프로젝트에서 나타나는 위험 정도는 알고리즘 및 기계학습 기술의 종류, 응용 프로그램의 유형, 데이터, 관련 산업에 따라 다르다. 그러나 기술 및 환경의 다양성에 관계없이 인공지능 시스템 안전성 평가는 프로젝트팀의 설계 및 구현 방법이 정확성, 신뢰성, 보안 및 견고성과 관련된 안전 목표와 일치하는지를 확인하는 작업이다.

## 시사점

인공지능이 적용된 제품을 사용하는 것은 앞으로 아주 자연스러운 일이 될 것이다. 인공지능 스피커를 이용하여 텔레비전을 켜고, 제품을 주문한다. 무인 지하철을 타고 출근하며, 자율주행 기능이 탑재된 자동차가 거리를 누빌 것이다. 그러나 안전이 기본적으로 보장되어야 함은 주지의 사실이고, 각국은 이를 위해 연구비를 투자하고 관련법을 정비하고 있다.

우리나라도 인공지능 선두그룹에 포함되기 위해서는 인공지능 기술뿐만 아니라, 인공지능 기술을 적용한 산업과 제품의 안전에 대한 검토가 필수적이다. 안전 선진국에서는 항공, 자동차, 철도 등 기존 안전 중요 시스템의 안전을 위한 표준을 만들고, 인증 기업들이 제품의 안전성을 검증해주고 있다. 인공지능이 적용된 제품에서도 유사한 움직임이 일어나고 있다. 인공지능의 성공여부가 인공지능 기술뿐만 아니라 안전과 윤리적 문제에도 좌우될 수 있다는 점을 인식하고 균형이 있게 접근할 필요가 있다. 인공지능이 가지는 혁명적 파급력을 고려할 때 인공지능의 안전성이 인류의 안전과 직결될 수 있기 때문이다.