ISSUE REPORT | 2025.1.20. IS-196

미국의 AI안전·신뢰성 정책 추진 현황과 시사점

Current Status and Implications of U.S. Policy for Implementing Trustworthiness AI

이해수, 유재홍

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

이 보고서는 「과학기술정보통신부 정보통신진흥기금」에서 지원받아 제작한 것으로
과학기술정보통신부의 공식 의견과 다를 수 있습니다.
이 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문 사항 또는 수정·보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.

소프트웨어정책연구소 AI정책연구실
이해수 선임연구원 (hs.lee@spri.kr)

CONTENT

요 약 문

본 연구에서는 그간의 미국의 인공지능(AI) 안전·신뢰성 정책 흐름을 살펴보고, 트럼프 2.0 시대에 AI 기술의 안전·신뢰성 확보하기 위한 정책 흐름이 어떻게 이어질지 예측하고자 하였다. 분석 결과, 미국은 오바마 행정부부터 트럼프 1기 행정부, 바이든 행정부에 걸쳐 AI 기술을 국가 안보와 직결된 전략 기술로서 인식하고 관련 정책을 수립·추진해왔다. 오바마 행정부에서는 AI가 안보와 글로벌 리더십 확보에 중요한 전략 과제임을 인식하고 안전한 AI 시스템을 보장하기 위한 연구 개발 계획 및 지침을 수립하였고, 트럼프 1기 행정부는 자국의 AI 리더십 확보를 위한 연구 강화와 혁신 투자, 그리고 신뢰할 수 있는 AI 개발 및 보급을 지시하였다. 바이든 행정부에서는 연방정부 차원에서 안전하고 신뢰할 수 있는 AI 혁신에 관한 행정명령을 발동하였다. 공통적으로 AI 개발 및 보급에서 세계를 선도함과 동시에 안전성과 신뢰성을 강화할 수 있는 정책을 마련하는 데 초점을 맞추고 있다. AI의 안전·신뢰성 확보 정책은 국가적으로 일관된 정책 흐름이며, 오히려 국가 안보와 미국 우선주의를 강조하는 트럼프2.0 시대에서는 더욱 강화될 수 있을 것으로 전망된다.

Executive Summary

This study examines the trajectory of U.S. Policy for Implementing Trustworthiness artificial intelligence (AI) safety and reliability policies and seeks to predict how these policies may evolve Trump 2.0 era. The analysis reveals that since the Obama administration through the first Trump administration, the United States recognized AI as a strategic technology closely tied to national security, developing and implementing policies to support its advancement. The Obama administration acknowledged AI as a critical strategic task for national security and global leadership, establishing research and development plans and guidelines to ensure safe AI systems. The first Trump administration emphasized strengthening U.S. AI leadership by enhancing research, investing in innovation, and directing the development and deployment of trustworthy AI. The Biden administration issued executive orders at the federal level to advance trustworthy AI innovation. All administrations have consistently focused on leading global AI development and deployment while simultaneously reinforcing trustworthiness. Policies to ensure AI trustworthiness thus reflect a unified national strategy. Furthermore, with the Trump 2.0 era's emphasis on national security and an "America First" agenda, these policies are anticipated to be further strengthened.

Ⅰ. 미국의 AI 안전·신뢰성 정책 흐름

■ 오바마 행정부에서 국가 AI R&D 전략 계획을 최초로 수립

● 오바마 행정부는 2016년 10월 '인공지능의 미래에 대한 행정부의 보고서'를 발간*하며 인공지능의 기회, 고려사항과 향후 과제를 전반적으로 검토

* '미래 인공지능의 준비(NSTC[1], 2016.10)'와 '국가 AI 연구개발 전략'이라는 두 개의 보고서 발간

- '미래 인공지능의 준비' 보고서에서는 AI의 현재 상태, 기존 및 잠재적 응용 분야, AI의 발전이 사회와 공공 정책에 제기하는 질문을 조사하고 구체적인 추가 조치에 대한 권고 사항도 제시
- '국가 AI 연구개발 전략' 보고서는 인공지능 기술을 활용해 사회적 공익을 증진하고 정부 운영을 개선하는 방법, 혁신을 장려하는 동시에 대중을 보호하는 여러 정책 권고사항*을 제시

* 2016년 백악관 과학기술정책실이 주도한 전국 공동 워크샵 5회, 161건의 정보요청(RFI) 내용 포함, 23개의 정부 정책 권고안 제시

● 특히, 국가 AI 연구개발전략에서 제시한 7대 전략 중 절반 정도가 AI의 윤리적, 법적, 사회적 영향을 고려하여 안전한 AI 시스템을 보장하기 위한 연구 개발 전략과 관련*

- (전략 3) 설명가능성, 투명성, 신뢰성 제고, 검증 및 타당성(V&V) 확보, 자율 성장 역량을 갖출 것으로 예상됨에 따라 장기적 AI 안전 확보, (전략 6) AI 기술 측정 및 평가를 위한 표준(시스템의 위험 관리 및 위해 요소 분석, 인간-컴퓨터 상호작용, 제어 시스템, 그리고 규제 준수 평가) 및 벤치마크 개발

■ 트럼프 1기 행정부의 행정 명령과 American AI 이니셔티브 발족해 국가전략기술로서 관리

● 트럼프 대통령은 2019년(American AI Initiative), 2020년(American AI Initiative) 두 건의 행정명령을 통해 인공지능에 대한 연구 강화와 혁신 투자, 그리고 신뢰할 수 있는 인공지능 개발 및 보급을 지시

- 2019년 2월 행정명령(EO 13859, 인공지능 분야에서 미국의 리더십 유지)에서는 AI 연구 투자 확대(비국방 AI R&D 2배 확대 포함), 연방 AI 컴퓨팅 및 데이터 자원 활용, AI 혁신 장벽 제거 (AI 기술 표준 수립, AI 안전한 개발, 검증, 배포 기술 마련), AI 인력 양성(STEM[2] 교육, 근로자 AI 역량 강화, AI전문연구인력 육성 등), 신뢰할 수 있는 AI 관리를 위한 국제 협력 노력을 담고 있음(OSTP[3], 2020.2.)
- 2020년 12월 행정명령(EO 13960, 연방정부에서 신뢰할 수 있는 인공지능의 사용 촉진)에서는 행정명령 13859에 따라 국가 안보 및 국방 이외의 목적으로 연방 정부에서 AI를 사용하는 경우에도 AI의 개발 및 사용이 미국의 가치에 부합하고 대중에게 이익이 되도록 보장하기 위해 추가 원칙 및 기관 전반에 걸친 공통 정책 지침을 통해 이러한 원칙을 이행하기 위한 프로세스 수립을 명령

1 NSTC, National Strategy for Advanced Manufacturing
2 STEM, Science, technology, engineering, and mathematics
3 OSTP, The White House Office of Science and Technology Policy

● 국가 인공지능 이니셔티브실(NAIIO[4]) 출범(The Whitehouse, 2021.1.)

- 행정명령(EO 13859)에 근거해 지속적인 AI 투자와 혁신 촉진, AI 연구를 위한 자원 액세스 강화, 차세대 AI 연구 인력 육성을 통한 미국의 인공지능 리더십 유지를 위한 미국 AI 이니셔티브를 발족
- 이를 위한 거버넌스로 백악관 과학기술정책실(OSTP) 산하에 국가인공지능 이니셔티브실(NAIIO)*을 설치

* NAIIO는 국가 인공지능 이니셔티브법(National AI Initiative Act of 2020)에 따라 설립되었으며, 국가 AI전략을 감독, 구현하며, 민간과 학계를 포함한 이해관계자들과 함께 AI 정책 입안에서 조정과 협력의 구심점 역할 수행

■ 바이든 행정부에서는 행정명령과 AI안전연구소 설립으로 AI안전 확보 실행력 강화

● 바이든 행정부에서는 2023년 1월 트럼프 정부에서부터 상무부 산하 국가표준기술연구소(NIST)가 진행한 인공지능 위험관리 프레임워크(AI RMF[5] 1.0)를 발표

- AI RMF는 연방정부의 AI개발 및 활용을 위한 표준 가이드라인으로서 역할

● 한편, AI안전과 보안 강화, 국제협력을 강조한 국가 AI R&D 전략 계획 업데이트(NSTC, 2023.5.)

- 2016년 오바마정부에서 최초로 수립된 국가 AI R&D 전략 계획은 7개 전략과제에서 2019년 트럼프 정부에서는 '민관 AI R&D 파트너십을 확대하는 8개 전략 과제로 확장
- 2023년 바이든 정부에서는 기존 8대 전략을 유지하고, AI R&D 국제 협력 전략*을 새롭게 추가하였으며 AI의 안전성, 신뢰성, 공익에 기여하는 책임성 측면의 전략**을 강조

* [전략9] 새롭게 추가된 AI R&D 국제협력 추진 전략의 세부내용으로는 △신뢰할 수 있는 AI개발과 사용을 위한 글로벌 문화 조성, △글로벌 AI 시스템, 표준 및 프레임워크 개발 지원, △ AI 전문 인력 및 정보 교류 지원, △글로벌 이익을 위한 AI 개발 촉진

** [전략4] AI 시스템의 안전 및 보안 보장 측면에서 △안전한 AI 구축, △AI 보안 확보를 새로운 세부 전략으로 추가(2023)[6]

● 신뢰할 수 있는 AI 행정명령(EO 14110) 발표와 함께 AI안전연구소를 설립하고 국제협력 강화

- 연방정부 차원에서 AI의 안전한 개발과 보급에 초점을 둔 행정명령을 발동하여 단순히 AI윤리 원칙 수립을 넘어 연방 기관의 실질적 조치들을 지시하고 이행사항을 관리
- 또한 영국에서 개최한 AI안전성정상회의('23.11)을 통해 AI안전 이슈를 글로벌 논의주제로 부상시키고, 선제적으로 AI안전연구소 설립('23.11)하고 국제 AI안전연구소 네트워크를 출범('24.11)시키는 등 관련 국제규범 리더십 확보의 거점으로 활용

⇨ 다음 장에서는 바이든 행정부에서 진행된 안전하고 신뢰할 수 있는 AI 행정명령(EO 14110)의 주요 이행사항을 국가기술표준연구소(NIST[7])와 주요 연방 기관들을 중심으로 분석함

4 NAIIO, National Artificial Intelligence Initiative Office

5 AI RMF, AI Risk Management Framework

6 '안전'은 새로운 피해를 생성하는 시스템에 대한 완화, '보안'은 시스템의 무결성에 대한 모니터링으로 정의

7 NIST, National Institute of Standards and Technology

Ⅱ. 바이든 행정부의 행정명령과 이행 현황

2.1 바이든 정부의 안전하고 신뢰할 수 있는 AI를 위한 행정명령 개요

■ (행정명령 14110호) 미국 바이든 대통령은 2023년 10월 30일 '안전하고 신뢰할 수 있는 AI 개발 및 사용에 관한 행정명령(이하 행정명령)'에 서명(The White House, 2023.10.30.)

● 행정명령은 AI 기술의 개발 및 활용에 있어 안전·신뢰성을 보장하고, 기술 혁신을 통해 미국의 글로벌 리더십을 강화하기 위한 목적

- AI 기술이 가져올 기회와 도전 과제를 모두 다루며, 기술 발전과 책임 있는 사용을 조화롭게 이끌기 위한 다방면의 접근법을 담고 있음
- AI 개발·이용을 발전시키고 관리하기 위한 8가지 '원칙 및 우선순위' 제시

※ △ AI 기술의 안전과 보안 보장, △ 책임 있는 혁신, 경쟁 및 협력 장려, △ 미국 근로자에 대한 지원 약속, △ 형평성과 시민의 권리 증진, △ 미국인의 이익 보호, △ 미국인의 개인정보와 시민 자유 보호, △ 미국 정부의 AI 이용 위험 관리와 책임 있는 AI 규제·관리·지원 역량 향상, △ 미국이 글로벌 사회, 경제, 기술 발전 선도를 제시

● 바이든 정부는 행정명령을 통해 책임감 있고 효과적인 AI 개발과 이용을 달성하기 위한 광범위한 행정조치를 명시하고, 각 연방행정기관과 그 소속(연방)기관의 구체적인 이행 사항과 이행기간을 규정

■ 각 연방기관들은 행정명령에 따라 AI 안전과 위험관리를 비롯한 주요 영역에서 소관 사항을 이행

● 백악관은 2024년 10월 30일 브리핑 성명을 통해 지난 1년 동안 연방기관은 행정명령에 명시된 조치 100건 이상을 모두 일정대로 완료했다고 보고(The White House, 2024)

- 행정명령에 따른 연방 정부의 조치는 AI의 안전과 보안 관리, 책임 있는 AI 혁신 강화, 미국의 리더십 증진 등으로 구성

● 특히, 상무부는 장관의 책임 하에 산하기관들이 AI 안전, 보안 제고 등을 위한 세부 규정안을 마련하여 공표

- 미국 상무부 산하 국립표준기술연구소(NIST)와 AI안전연구소는 AI의 안전성, 보안성 및 신뢰성을 다루는 프레임워크 및 도구개발과 보급, 지침 또는 표준 수립, 국제협력 등에 집중

2.2 행정명령의 주요 내용과 이행 현황 개괄

■ 백악관에서는 2023년 10월 행정명령(EO 14110)을 발표후 1년이 지난 2024년 10월에 연방 정부 기관으로부터 이행 상황을 보고받아 발표[8]

● 행정명령 이행조치 유형은 총 10가지로 구분

- △지침 또는 표준 수립 △ AI 위험 관련 사전 조사 및 검토 △ 정책 및 전략 보고서 작성 및 제출 △ AI안전 관리 프레임워크 관련 활동 및 도구 개발과 보급 △ 파일럿 프로그램 및 이니셔티브 개발 및 실행 △ 국제 협력 및 협의체 운영 △AI 신뢰성 대응 조직 정비(TF구축) △ 대회 및 공모전 개최 △ 대응재원조성 △AI 인재 채용으로 구분

● 지침(Guide) 또는 표준(Standard) 수립(30개), AI 위험 관련 사전 조사 및 검토(24개), 정책 및 전략 보고서 작성 및 제출(16건), AI 안전 관리 프레임워크 관련 활동 및 도구 개발과 보급(16건) 순으로 확인

<표 2-1> 백악관 행정명령 이행조치 Action 유형

(단위: 개)

유형	빈도
지침(Guide) 또는 표준(Standard) 수립	30
AI 위험 관련 사전 조사 및 검토	24
정책 및 전략 보고서 작성 및 제출	16
AI안전 관리 프레임워크 관련 활동 및 도구 개발과 보급	16
파일럿 프로그램 및 이니셔티브 개발 및 실행	10
국제 협력 및 협의체 운영	8
AI 신뢰성 대응 조직 정비(TF구축)	5
대회 및 공모전 개최	3
대응재원조성	2
AI 인재 채용	2
계	**116**

출처: The White House (2024.10.30.)

■ 각 부처와 기관 특성에 따라 이행 조치 유형에도 차이가 나타남

- 상무부는 주로 AI 위험 관련 사전 조사 검토, 지침 또는 표준 수립, 정책 및 전략 보고서 작성 및 제출, AI 안전 관리 프레임워크 관련 활동 및 도구 개발과 보급에 집중
- 국무부는 AI 위험 관련 사전 조사 및 검토, 지침 또는 표준 수립에 집중하는 경향
- 국립과학재단(National Science Foundation, NSF)은 파일럿 프로그램 및 이니셔티브 개발 및 실행에 가장 높은 빈도 확인

8 행정명령이 발효된 2023년 10월 30일부터 365일이 지난 2024년 10월 30일까지가 해당

<표 2-2> 연방 정부 및 기관별 백악관 행정명령 이행조치 유형

(동시이행 포함, 단위: 개)

유형	연방부처 및 기관																백악관산하기관	계
	ED	DOT	DOS	DOD	DHS	DOL	DOA	DOJ	HHS	DOC	DOE	DOT	VA	HUD	NSF	기타		
지침(Guide) 또는 표준(Standard) 수립	0	0	2	0	3	3	1	0	2	4	0	0	0	2	0	10	3	30
AI 위험 관련 사전 조사 및 검토	0	2	3	1	2	1	0	2	0	5	0	0	0	0	0	6	4	26
정책 및 전략 보고서 작성 및 제출	0	0	1	1	1	1	0	1	1	4	1	1	0	0	0	1	4	17
AI안전 관리 프레임워크 관련 활동 및 도구 개발과 보급	2	0	1	1	4	0	0	0	1	4	2	0	0	0	0	1	2	18
파일럿 프로그램 및 이니셔티브 개발 및 실행	0	0	1	1	2	0	0	0	1	0	1	0	0	0	6	0	0	12
국제 협력 및 기타 협의체 운영	0	1	0	0	2	0	0	1	0	0	2	0	0	0	0	1	1	8
AI 신뢰성 대응 조직 정비(TF구축)	0	0	0	0	2	0	0	0	1	0	1	0	0	0	0	0	1	5
대회 및 공모전 개최	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	3
대응재원조성	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	3
AI 인재 채용	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
계	2	3	8	4	16	5	1	4	6	18	8	1	1	2	8	21	16	**124***

출처: The White House (2024.10.30.)

*** 두 개 이상의 연방 정부 및 기관이 동시 이행한 수치가 포함된 합계(eg. 하나의 행정명령을 국무부와 상무부 동시 이행 = 국무부 1, 상무부 1로 합계)**

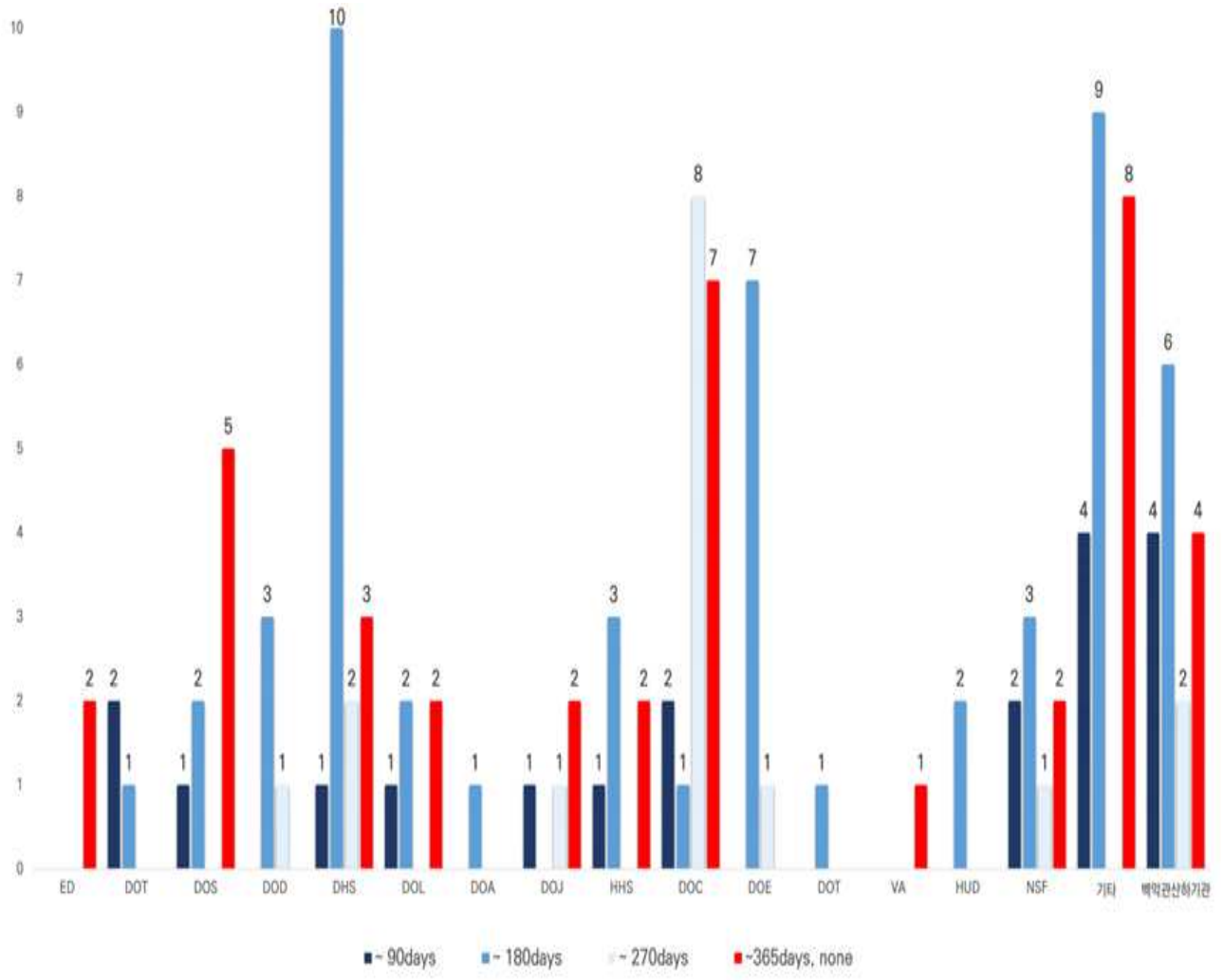
※ 교육부(ED), 교통부(DOT), 국무부(DOS), 국방부(DOD), 국토안보부(DHS), 노동부(DOL), 농무부(DOA), 법무부(DOJ), 보건복지부(HHS), 상무부(DOC), 에너지부(DOE), 재무부(DOT), 재향군인부(VA), 주택도시개발부(HUD), 국립과학재단(NSF), 기타(청, 국, 위원회), 백악관산하기관(과학기술정책실, 예산관리국 등)

■ 각 부처 및 기관별 행정명령 이행 속도 및 기간별 분포 차이 확인

- 행정명령에 따라 각 연방부처 및 기관은 명시된 이행기한에 맞추어 소관 사항을 이행 중
  - 각 부처 및 기관은 다양한 기간(~ 90days / ~180days / ~ 270days / ~365days, none)에 걸쳐 소관업무를 이행 중인 것으로 확인
  - 특히, 기타(청, 국, 위원회)(13건), 국토안보부(11건)가 상대적으로 180일 이내 이행 건수가 많았고, 상무부가 ~ 270days(8건) / ~365days, none(7건) 기간에 총 15건으로 이행 건수가 가장 많은 것으로 나타남

<그림 2-1> 백악관 행정명령 이행조치 Agency별 이행기간

(동시이행 포함, 단위: 개)



출처: The White House (2024.10.30.)

※ <그림 2-1> 그래프도 <표 2-2>와 같이 두 개 이상의 연방 정부 및 기관이 동시 이행한 수치가 포함된 합계로 작성됨(계 = 124)

※ 교육부(ED), 교통부(DOT), 국무부(DOS), 국방부(DOD), 국토안보부(DHS), 노동부(DOL), 농무부(DOA), 법무부(DOJ), 보건복지부(HHS), 상무부(DOC), 에너지부(DOE), 재무부(DOT), 재향군인부(VA), 주택도시개발부(HUD), 국립과학재단(NSF), 기타(청, 국, 위원회), 백악관산하기관(과학기술정책실, 예산관리국 등)

Ⅲ. 세부 이행 내용(1): 국가표준기술연구소(NIST)

3.1 지침 또는 표준 수립

■ 이중용도 기반모델의 오용 평가에 관한 지침 발행

● NIST 산하 AI안전연구소(이하, AISI)는 이중사용 기반모델 오용 위험 관리(Managing Misuse Risk for Dual-Use Foundation Models: NIST AI 800-1)에 대한 지침 초안 공개(NIST, 2024.7.)

- AI 기반모델은 광범위한 작업에 유용한 강력한 도구이지만, 혜택과 피해를 모두 줄 수 있는 잠재력으로 '이중사용(dual-use)'*이 가능

* AI 기반 모델에서 이중 사용(Dual-use)은 AI 기반 모델이 과학 연구 발전, 의료 진단 개선 등 다양한 부문에서 긍정적 영향을 미칠 수도 있지만, 개인정보 침해, 잘못된 정보 확산 및 자동 감시 등 유해한 목적으로도 사용될 수 있음을 의미

- 지침 초안에는 AI 개발자에게 AI 시스템이 개인, 공공 안전 및 국가 안보에 해를 끼치는 데 오용되는 것을 방지하는 방법, 제품에 대한 투명성을 향상하는 방법(절차) 등을 설명

● 지침 초안은 이중사용 기반모델 오용 위험을 완화하기 위한 7가지 목표와 구현방법 제시

- 본 지침은 이중사용 기반모델이 생물학적 무기 개발, 공격적인 사이버 작전 수행, 아동 성적 학대 콘텐츠 생성과 같은 활동을 통해 사람들에게 해를 끼치는 것을 방지하는 데 활용 가능

<표 3-1> NIST AI 800-1에 제시된 이중사용 기반모델 위험관리 목표

위험관리 목표	주요내용
1. 잠재적 오용 위험 예상	• 기반모델이 악의적인 행위자에게 제공 시 오용될 수 있는 위험을 평가 • 가장 중요한 예상 위험을 식별하여 측정하고 필요에 따라 관리
2. 오용 위험 관리 계획 수립	• 법 또는 규제 의무, 위험과 혜택과의 비교, 기타 요인을 고려하여 수용 가능한 오용 위험 수준을 결정 • 잠재적 오용 위험 관리를 위해 필요한 자원, 시간, 운영 제약에 맞춰 개발 및 배포 계획을 조정
3. 모델 도난 위험 관리	• 악의적 행위자가 기반모델을 재구성할 수 있도록 도울 수 있는 정보와 자산 도난을 방지하는 조치를 수행
4. 오용 위험 측정	• 기반모델이 오용될 수 있다는 합리적인 평가가 있을 때, 실제 환경에서 모델 오용 위험을 입증할 수 있도록 위험을 예측 – 기술적 및 사회적 요소를 모두 포함하는 방법과 정확한 증거와 높은 신뢰성을 제공하는 방법을 사용
5. 기반모델을 배포하기 전에 오용 위험이 관리되었는지 확인	• 오용 위험이 적절하게 관리되고, 이러한 위험이 최소한 조직의 위험 허용 범위 내 있는 경우에만, 모델 접근을 확대하는 조치를 수행
6. 배포 후 오용에 관한 정보를 수집하고 대응	• 배포 시스템에 대한 정보를 수집하여 오용 위험에 대한 이해를 향상, 배포를 조정하며, 미래의 위험 관리 개선에 기여
7. 오용 위험에 대한 적절한 투명성 제공	• 오용 위험과 관련된 기반모델의 개발 및 배포에 대한 조직 프로세스에 대해 대중과 관련 기관에 투명성을 제공하여, 모델 오용과 관련된 이해, 책임, 협력, 과학적 발전을 촉진

출처: NIST(2024. 7.)

■ 생성AI 및 이중사용 기반모델을 위한 안전한 소프트웨어 개발 절차 지침 마련

● 생성AI 학습데이터에 대한 위협을 줄이기 위해 생성AI 및 이중사용 기반모델을 위한 안전한 소프트웨어 개발 절차를 제시

- 생성AI 및 이중사용 기반모델을 위한 안전한 소프트웨어 개발 절차(NIST, 2024.7.)은 Secure Software Development Framework(SP 800-218, 이하 SSDF)와 함께 사용하도록 개발

  * Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: NIST SP 800-218A)

- 기존 SSDF는 소프트웨어 코딩 관행과 광범위하게 관련된 반면, 이번 지침은 생성AI 시스템에 대한 주요 우려 사항을 해결하기 위해 SSDF 내용을 확장하여 AI 시스템의 성능에 부정적인 영향을 미치는 악성 학습 데이터에 관한 사항을 포괄

● 지침(NIST SP 800-218A)은 AI 시스템의 학습 및 사용 측면을 다루는 것 외에도 잠재적 위험 요소와 이를 해결하기 위한 전략을 제시

- 특히, 생성AI 시스템에서 중독, 편향, 동질성 및 변조 징후에 대한 학습 데이터에 대한 분석 결과를 제시

<표 3-2> NIST SP 800-218A에 제시된 AI 모델 개발을 위한 주요 권고사항들

영역	세부영역	권고사항
조직 준비	소프트웨어 개발 보안 요건 정의	• 소프트웨어 개발 인프라 및 프로세스에 대한 보안 요건에 AI 모델 개발을 포함 • 사이버 보안, 개인정보 보호 및 재생산성 대한 권장사례에 따라 적절한 AI 모델 아키텍처 및 교육 기법을 식별하고 선택 • 조직의 정책들이 AI 모델 개발 보안에 대한 요구사항을 지원
	역할 및 책임 구현	• 소프트웨어 개발 생명 주기(SDLC) 전체에 걸쳐 SDLC 관련 역할 및 책임에 AI 모델 개발 보안을 포함 • 역할 기반 학습에 사이버 보안 취약성과 AI 모델에 대한 위협 및 가능한 완화 방법을 포함
	지원 툴사슬(Toolchain) 구현	• AI 모델 개발 보호와 인간의 노력을 줄이는 자동화된 툴체인 개발 및 구현 계획을 마련하고, 실행 • 위험에 상응하는 빈도로 툴체인의 보안을 확인
	소프트웨어 보안 점검을 위한 기준 정의 및 활용	• AI 개발 수명주기 전반에 걸쳐 가드레일 및 기타 제어를 구현하여 기존 SDLC를 넘어 확장
	소프트웨어 개발을 위한 보안 환경 구현 및 유지 관리	• 모델 개발 중 AI 모델 사용자의 리소스 사용량과 요금을 모니터링, 추적 및 제한 • AI 모델 개발 중에 사용된 민감한 데이터만 조직에서 승인한 환경과 위치에 저장 • 최소 권한 원칙에 따라 환경 내의 모든 학습 파이프라인, 모델 레지스트리를 보호하고, 수정사항을 지속적으로 모니터링 수행 • AI 모델 또는 관련 리소스(모델 API, 가중치, 구성 매개변수, 학습 데이터 등)를 호스팅하는 모든 개발 환경 구성요소에 대한 지속적인 보안 모니터링을 수행 • 지속적인 모니터링 및 분석 도구는 AI 모델과 관련된 활동이 위험 임계값을 초과하거나 추가 조사가 필요한 경우 경고를 생성해도록 설정
소프트웨어 보호	코드와 데이터에 대한 무단 접근 및 변조로부터 보호	• 안전한 코드 저장소는 AI 모델, 모델 가중치, 기밀성, 무결성, 가용성을 보호해야 하는 기타 AI 모델 요소를 포함 • 최소 권한 원칙을 따라 저장 또는 실행되는 위치에 관계없이 AI 모델과 모델 요소에 대한 직접 접근을 최소화 • 기밀성(비공개 데이터에 한함)과 데이터의 무결성을 지속적으로 모니터링 수행

영역	세부영역	권고사항
	소프트웨어 릴리스 무결성 검증 메커니즘 제공	• AI 모델과 그 구성요소, 구성요소 및 문서의 암호화 해시 또는 전자서명을 생성 및 제공 • AI 모델 변경에 대한 전자서명을 제공
	소프트웨어 릴리스를 보관하고 보호	• 데이터 세트 생성 및 모델 학습을 지원하는 인프라 도구에 대한 버전 관리 및 추적을 수행 • 보관된 정보에 AI 모델 선택에 대한 정당성에 대한 문서를 포함 • 모델을 빌드하는 데 사용된 학습 라이브러리, 프레임워크 및 파이프라인을 포함하여 AI 모델과 그 구성요소 및 파생 모델의 출처를 추적 • 민감한 데이터(예: 지불 카드 데이터)에 대해 학습된 AI 모델을 추적
안전한 소프트웨어 제작	보안 요건 충족 및 보안 위험을 완화하기 위한 소프트웨어 설계	• 위험 모델링에 관련 AI 모델별 취약성 및 위협 유형을 통합(위협 유형: 학습데이터 오염, 입력 및 출력의 악성 코드 또는 기타 원치 않는 콘텐츠, 적대적 프롬프트에서 발생하는 서비스 거부 조건, 공급망 공격, 무단 정보 공개, AI 모델 가중치 도용 및 데이터 파이프라인의 잘못된 구성)
	사용 전에 학습, 테스트, 미세 조정 및 데이터의 무결성 확인	• 사용 전에 학습, 테스트, 미세 조정 및 데이터의 무결성 확인 및 분석하고 변경하기 위한 적절한 방법을 선택하고 적용 • 프로세스와 해당 제어를 사용하여 적대적 샘플을 테스트하고 학습 및 테스트 사용에 적절한 보호 장치를 설치
	기존의 보안이 잘 된 소프트웨어를 재사용	• 기존 AI 모델 또는 기타 획득한 AI 구성요소의 무결성, 출처 및 보안을 검증 • 획득한 AI 모델과 해당 구성요소를 사용하기 전에 취약성과 악성 콘텐츠를 스캔하고 철저히 시험
	보안 코딩 관행을 준수하여 소스 코드 생성	• AI 기술별 고려 사항을 포함하도록 보안 코딩 관행을 확장 • 입력(프롬프트 및 사용자 데이터 포함) 및 출력 처리를 신중하게 코딩 • 입력 및 출력은 AI 모델의 컨텍스트 내에서 기록, 분석 및 검증해야 하며, 문제가 있는 입력 및 출력은 정제하거나 삭제
	인간이 읽을 수 있는 코드를 검토 및/또는 분석하여 취약성을 식별하고 보안 요구사항 준수 확인	• 코드 검토 및 분석 정책 또는 지침에는 AI 모델 및 기타 관련 구성요소에 대한 코드를 포함 • 조직의 코드 검토 및 분석 정책 또는 지침에 따라 모든 AI 모델을 검사하여 악성코드, 취약성, 백도어 및 기타 보안 문제를 확인
	실행 가능 코드를 테스트하여 취약성을 식별하고 보안 요구사항 준수 확인	• 코드 테스트 정책 및 지침에 AI 모델을 포함(단위 테스트, 통합 테스트, 침투 테스트, 레드팀, 사용 사례 테스트, 적대적 테스트를 포함하여 여러 가지 형태의 코드 테스트를 AI 모델에 사용 가능) • 조직의 코드 테스트 정책 또는 지침에 따라 모든 AI 모델의 취약성을 시험하고, AI 모델을 재학습하거나 새로운 데이터 소스를 추가할 때 재시험
취약성 대응	취약점을 지속해서 식별하고 확인	• AI 모델에 대한 모든 입력과 출력을 기록, 모니터링 및 분석하여 잠재적인 보안 및 성능 문제를 탐지 • AI 모델 사용자에게 잠재적인 보안 및 성능 문제를 보고하는 메커니즘에 대한 정보 제공 • 이전에 감지되지 않은 취약점을 식별하기 위해 AI 모델을 자주 스캔하고 시험 • 지속적인 검사 및 시험에 주로 자동화를 사용하고 필요에 따라 인간 참여 • AI 모델에 대한 주기적 감사를 수행 • 조직 취약점 공개 및 수정 정책에 AI 모델 취약점을 포함 • AI 모델 사용자에게 고유한 한계와 발생하는 사이버 보안 문제를 보고하는 방법에 대한 정보 제공
	취약성 평가, 우선순위 지정 및 수정	• AI 모델에 대한 위험 대응은 이를 재구축하는 데 소요될 수 있는 시간과 비용을 고려 • AI 모델 사용을 중단할 시기와 이전 버전 및 해당 구성요소로 롤백할 시기를 위한 기준 및 프로세스를 수립하고 구현

출처: NIST(2024. 7.)

3.2 AI안전 관리 프레임워크 관련 활동 및 도구 개발과 보급

■ NIST는 AI 시스템의 안전 및 신뢰성 확보를 위한 AI 위험관리 프레임워크를 개발

- 미국 상무부 산하 NIST는 AI 위험관리 프레임워크 1.0 (AI Risk Management Framework, AI RMF)을 발표(NIST, 2023.1.)
  - AI 위험관리 프레임워크(AI RMF)는 다양한 이해관계자(정부, 산업계, 학계, 시민 사회 등)의 정보 요청 및 공개 의견을 반영하기 위한 워크숍, 초안 피드백 등의 합의 중심의 개방적이고 투명하며 협력적인 프로세스를 거쳐 완성
  - 이 프레임워크는 개인, 조직 및 사회에서 AI 제품, 서비스 및 시스템의 설계, 개발, 사용 및 평가에 위험을 관리할 수 있도록 개발되어 미국을 포함하여 전 세계적으로 널리 도입되고 있음
  - NIST는 AI 위험관리 프레임워크를 조직에서 효과적으로 활용하기 한 구체적인 실행가이드로 AI RMF Playbook*을 함께 제공(NIST, 2023)

    * AI RMF의 주요 목적을 달성하기 위한 권장 조치를 거버넌스(Govern), 식별(Map), 평가(Measure), 관리(Manage) 차원에서 모범 사례를 바탕으로 제공
- 행정명령을 이행하기 위해 현행 AI 위험관리 프레임워크를 기반으로 생성 AI 프로파일 마련(NIST, 2024.7.)
  - AI 위험관리 프레임워크 생성AI 프로파일(Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile : NIST AI 600-1)은 기관이 생성AI로 인해 발생하는 위험을 식별하고, 목표와 우선순위에 부합하는 생성AI 위험 관리 조치 제시
  - 이 지침은 생성AI의 12가지 위험 목록과 개발자가 이를 관리하기 위해 취할 수 있는 조치(약 200개)를 제시
  - 특히, 생성AI의 12가지 위험에는 사이버공격에 취약하고, 잘못된 정보나 허위 정보 또는 증오 표현 및 유해 콘텐츠 생성, 생체·건강·위치 정보 같은 개인 식별가능 정보나 민감한 데이터의 유출 및 무단 사용, 환각(hallucination) 등이 포함

<표 3-3> AI 600-1에서 제시한 생성AI 위험

위험	주요내용
CBRN 정보 또는 능력	• 화학, 생물학, 방사선, 핵(CBRN) 무기 또는 기타 위험한 물질이나 제제와 관련된 해로운 정보나 설계 능력에 대한 접근이 용이해지거나 합성
허위 정보 생성	• 잘못된 또는 허위의 내용을 생성(일반적으로 '환각' 또는 '허위 정보 생성' 등)
위험하고 폭력적이거나 혐오스러운 콘텐츠	• 폭력적이고 선동적이며, 급진화하거나 위협적인 콘텐츠의 생성 및 접근이 쉬워지거나, 자해나 불법 활동을 권장(혐오스럽고 경멸적이거나 고정관념을 조장하는 콘텐츠의 대중 노출을 통제하기 어려운 경우를 포함)
데이터 프라이버시	• 생체 정보, 건강 정보, 위치 정보 또는 기타 개인 식별가능 정보나 민감한 데이터의 유출 및 무단 사용, 공개 또는 식별화에 따른 영향
환경적 악영향	• 생성AI 모델의 훈련 또는 운영에서 높은 컴퓨팅 자원 사용으로 인해 생태계에 악영향
유해한 편향 또는 동질화	• 역사적, 사회적, 시스템적 편향의 증폭 및 악화 • 대표성이 미흡한 학습데이터로 인해 하위 그룹 또는 언어 간 성능 격차가 발생하여 차별, 편향의 증폭 또는 성능에 대한 잘못된 추정 • 시스템 또는 모델 산출 결과 잘못되거나, 근거 없는 의사결정을 초래하거나, 유해한 편향을 증폭시킬 수 있는 바람직하지 않은 동질화
인간-AI 협력 간 부작용	• 인간이 생성AI 시스템을 부적절하게 의인화하거나 알고리즘 회피, 자동화 편향, 과신, 또는 생성AI 시스템과의 감정적 얽힘을 경험하게 될 수 있는 인간과 AI 시스템 간 작용
정보의 무결성 손상	• 사실이나 의견과 허구를 구별하지 못하거나, 불확실성을 인정하지 않는 콘텐츠를 생성하거나, 대규모 허위 정보 및 오보 캠페인에 활용될 수 있는 진입 장벽 완화
정보 보안 침해	• 자동화된 취약점 발견 및 악용을 통해 해킹, 악성 소프트웨어, 피싱, 공격적 사이버 작전 또는 기타 사이버공격에 취약 • 사이버공격의 공격 표면이 증가하여 시스템의 가용성이나 훈련 데이터, 코드 또는 모델 가중치의 기밀성 또는 무결성이 손상될 가능성 존재
지식재산권	• 허가 없이 저작권, 상표권, 또는 라이선스가 있는 콘텐츠를 생성하거나 복제하기 쉬워짐 • 영업 비밀의 노출, 표절이나 불법 복제가 용이해지는 경향
외설, 비하, 학대적 콘텐츠	• 외설적, 비하적, 학대적인 이미지 생성 및 접근 가능성 강화
가치 사슬 및 구성요소 통합 위험	• 제3자 구성요소(생성AI로 인한 자동화로 잘못 획득, 정제되지 않은 데이터 등)의 투명하지 않거나 추적할 수 없는 통합 • AI 수명주기 전반에 걸친 공급업체 검증 부적절

출처: NIST (2024.7.)

■ AI 시스템 모델의 공격 대응 시험을 위한 소프트웨어 개발 및 제공

● 정부 기관, 중소기업 등이 AI 시스템 성능에 대해 개발자들의 주장을 평가할 때 사용할 수 있는 인공지능 신뢰성 특성 평가 소프트웨어 플랫폼인 'Dioptra'를 개발해 배포(NIST, 2024[9])

- AI 시스템의 핵심인 모델은 대량의 데이터로 학습하여 결정을 내리는 방법을 배우지만, 적대자의 AI 시스템 모델에 대한 공격으로 잘못된, 잠재적으로 치명적인 결정 가능
- 특히, Dioptra는 크게 △ 회피(evasion), △ 오염(poisoning), △ 오라클(oracle)이라는 세 가지 AI 시스템 공격 유형에 대응하도록 개발
- 이를 해결하기 위해 개발·보급한 Dioptra 소프트웨어는 사용자가 어떤 종류의 공격이 AI 시스템 모델의 성능을 떨어뜨리는지 확인하고 성능 감소를 정량화하여 사용자가 AI 시스템이 얼마나 자주, 어떤 상황에서 실패하는지 알 수 있도록 지원

<표 3-4> Dioptra를 통해 대응 가능한 AI 시스템 모델의 공격

공격 유형	주요내용
회피(evasion) 공격	• AI에 입력하는 데이터에 노이즈를 더하여 이용자의 질문에 잘못된 답이 나오도록 하는 공격
오염(poisoning) 공격	• AI 모델의 정확도를 떨어트리는 것을 목표로 하는 공격으로, 주로 훈련 데이터를 변경 및 조작함으로써 이와 같은 목표를 달성* * 학습데이터 오염 예시: 정지 표지판을 속도 제한 표지판으로 잘못 식별하게 할 수 있는 학습 데이터 제공
오라클(oracle) 공격	• AI 모델을 역설계(Reverse engineering)하여 훈련 데이터를 추론하고 밝혀내는 공격으로, 이를 통해 민감한 데이터에 대한 통찰력 유추 가능

출처: NIST (2024.1.10. 방문)

■ AI 모델 평가를 위한 챌린지 프로그램 운영

● 생성AI의 위험성 평가를 위한 GenAI 평가 프로그램과 ARIA(Assessing Risks and Impacts of AI) 플랫폼을 개발하고 운영[10]

● NIST GenAI는 전 세계 연구 커뮤니티에서 개발한 생성 AI 기술을 평가하기 위해 NIST 정보기술 연구소에서 관리하는 생성AI 테스트 및 평가 프로그램으로 2024년 5월 공개

- 2024년 GenAI 파일럿 프로그램을 개시하여 텍스트 대 텍스트(T2T) 및 텍스트 대 이미지(T2I) 양식에서 합성 콘텐츠와 사람이 생성한 콘텐츠를 구별하기 위한 시스템 동작을 측정하고 이해하는 것을 목표*로

9 NIST, Dioptra 1.0.1 Documentation, https://pages.nist.gov/dioptra/ (2024.1.10. 방문)
10 https://ai-challenges.nist.gov/

2024년 5월부터 전세계 응모자들의 참여를 받아 8월까지 1차 결과 접수를 종료한 상태

* 생성자 팀(Generator team)은 사람이 제작한 콘텐츠와 구별할 수 없는 합성 콘텐츠를 생성하는 시스템의 능력을 테스트받게 되며 판별자 팀(Discriminator Team)은 LLM과 딥페이크 도구를 포함한 생성 AI 모델이 생성한 합성 콘텐츠를 감지하는 시스템의 능력을 테스트받게 됨

- AIRA는 대규모 언어 모델(LLM)과 관련된 위험과 영향 평가를 목표로 모델 테스트, 레드팀 테스트, 필드 테스트의 세 가지 테스트 지원
  - ARIA는 시스템 성능과 정확성에 중점을 두는 것을 넘어 기술적 및 상황적 견고성에 대한 측정을 제공
  - 또한, 조직이 시스템을 평가하고 AI 배포의 긍정적 또는 부정적 영향에 관한 의사 결정을 내리는 데 사용할 수 있는 지침, 도구, 방법론 및 메트릭을 제공
  - 2024년 5월 프로그램을 개시하고 9-10월에 1차 파일럿 평가 테스트를 진행하고 11월 관련 워크샵을 개최

■ 범부처 AI 안전 역량 및 위험 관리 평가를 위한 태스크포스 발족

- 한편, NIST산하 AI 안전연구소는 국가 보안 역량 및 위험 관리를 위한 AI 모델 연구 및 테스트에 협력하는 새로운 미국 정부 태스크포스 설립 (NIST, 2024.11)
  - 국가 안보를 위한 인공지능의 위험성 테스트(TRAINS, The Testing Risks of AI for National Security (TRAINS) Taskforce) 태스크포스는 상무부, 국방부, 에너지부, 국토안보부, NSA, NIH의 전문가들이 모여 국가 안보 문제를 해결하고 인공지능 혁신에 대한 미국의 리더십을 강화하기 위해 구성
  - TRAINS 태스크포스의 의장은 AI 안전연구소가 맡고 각 참여기관은 태스크포스에 고유 분야의 전문 지식 공유, 기술 인프라 및 리소스를 제공하고 새로운 AI 평가 방법과 벤치마크 개발에 협력하며 국가 안보 위험 평가 및 레드팀 합동 훈련을 실시할 예정

3.3 국제협력 및 기타 협의체 운영

■ 해외에서 미국의 AI 리더십을 증진시키기 위해 글로벌 AI 표준 참여 계획 수립 추진

● 글로벌 AI 표준에 대한 미국의 참여를 위한 포괄적 계획(A Plan for Global Engagement on AI Standards: NIST AI 100-5)을 발표(NIST, 2024.7.)

- 광범위한 공공 및 민간 부문의 의견을 통합하고, AI 표준 작업의 목표와 우선 분야를 식별하며, 미국 기관을 포함한 미국 이해관계자를 위한 조치를 제시
- AI 표준 및 관련 도구에 대한 연방 참여 계획(Plan for Federal Engagement in AI Standards and Related Tools)에 명시된 우선순위에 따르고, 중요한 신흥 기술을 위한 국가 표준 전략과 연계
- 많은 국가의 보다 광범위한 학제 간 이해관계자가 표준 개발 프로세스에 참여해야 한다고 제안

<표 3-5> 글로벌 참여 관련 권고사항

권고사항	주요내용
표준화 작업에 대한 참여 우선순위 지정	• 사전 표준화 연구 및 관련 기술 활동을 포함하여 표준화 작업에 대한 참여를 우선시하며, 이러한 활동에는 다음 사항을 포함 – 우선순위 주제에 대한 기초 연구 강화 – 표준 개발 참여를 통한 적시적인 개발 촉진 – 산업영역별 관행과 표준 간의 연계를 장려하기 위한 수평적 표준의 신중한 설계 및 사용 – 구현 도구의 개발 및 공유 – 잠재적 표준 채택자들과의 피드백 루프 강화 프로세스 탐색 • **우선적인 미국 정부의 실행 조치**: 연구 노력의 전략적 지침 제공, 기관의 표준 작업 참여, 정부 내외부의 정보 교환, 표준 요소를 포함한 국제 협력
AI 표준 개발 및 채택에 다양한 다중 이해관계자 참여 촉진	• 국내 역량 강화 활동에는 AI 표준 이해관계자들을 정기적으로 소집, AI 이해관계자들을 위한 표준 교육 및 핸드북과 같은 정보 개발 및 배포, 표준 참여를 위한 조직적 자원 투자를 포함 • 글로벌 역량 강화 활동에는 프레임워크 및 표준에 대한 글로벌 접근성 확대, AI 표준 개발에 대한 다양한 참여를 지원하기 위한 자원 증대, 글로벌 AI 전문가들이 모이는 환경에서 AI 표준에 대한 교육 제공, AI 표준 전문가의 글로벌 과학 네트워크 구축을 포함 • **우선적인 미국 정부의 실행 조치**: 내부 역량 강화, 자원 배치 및 교육, 관련 미국 정부 문서의 광범위한 접근성 제공, 해외 원조 프로그램, 다양한 발전 단계에 있는 국가들과의 협력
AI 표준 접근법에 대한 글로벌 연계 촉진	• 참여 활동에는 다중 이해관계자 참여와 글로벌 합의에 따라 주도되는 표준 생태계를 옹호, AI 표준에 대한 전문가 간 교류를 주도, 다양한 프레임워크 간의 연계를 계속 추구, 표준과 오픈소스 소프트웨어 간의 관계에 대한 논의 주도 등을 포함 • **우선적인 미국 정부의 실행 조치**: AI 표준 문제를 외교 회담 및 결과물에 통합, 정부 기관 간 협력을 통해 이를 달성

출처: NIST(2024. 7.)

■ 국제 AI안전연구소 네트워트 창립 총회를 개최하며 AI안전 동맹 강화

● 미국은 호주, 캐나다, 유럽연합, 프랑스, 일본, 케냐, 대한민국, 싱가포르, 영국 등 10개국 AI안전연구소가 참여하는 국제 AI 안전 네트워크 창립 총회를 개최하고 초대 의장국을 맡음(NIST, 2024)

- 2024년 5월에 열린 'AI 서울 정상회의'에서 지나 라이몬도 美 상무부 장관이 국제 AI 안전 기관 네트워크 출범을 예고한 바 있으며 2024년 11월 20~21일 캘리포니아 샌프란시스코에서 창립 회의를 개최
- 네트워크는 총회에 앞서 공동 사명 선언문, 합성 콘텐츠 연구를 위한 1,100만 달러 이상의 자금 지원, 네트워크 최초의 다자간 테스트 연습 결과, 첨단 AI 시스템의 위험 평가 등 주요 진전 사항을 발표

■ NIST 산하 AI안전연구소(AISI)는 영국 AI안전연구소와 최신 AI모델 공동 평가 실시

● 미국과 영국은 2023년 11월 제1회 AI안전성정상회의 후속으로 2024년 4월 양국간의 AI안전 연구, 안전 평가, 지침 마련, 정보 및 인력 교류 등 상호 협력을 위한 양해 각서를 체결

● NIST AI안전연구소는 2024년 8월 앤트로픽 및 오픈AI와 AI안전 연구, 시험, 평가 협력 체결하고 이후 양국의 AI안전연구소에서 최신 AI모델의 배포전 사전 평가를 공동 수행하고 결과를 공유

- 미국 AI안전연구소와 영국 AI안전연구소는 앤트로픽의 업그레이드된 클로드 3.5 소네트와 오픈AI의 o1모델[11]을 대상으로에 대한 배포전 사전 공동 평가를 실시(NIST, 2024, UK AISI, 2024)
- 양국 AI안전연구소는 (1) 생물학적 역량, (2) 사이버 역량, (3) 소프트웨어 및 AI 개발, (4) 세이프가드 효능의 네 가지 영역에서 모델의 역량을 평가하기 위해 개별적이지만 상호 보완적인 테스트를 실시

  * 모델의 상대적 역량을 평가하고 각 평가 영역에서 최신 모델의 잠재적 실제 영향을 평가하기 위해 미국 AISI와 영국 AISI는 이전 버전의 Anthropic의 소네트 3.5, OpenAI의 o1-preview, OpenAI의 GPT-4o와 같은 일련의 유사한 참조 모델과 그 성능을 비교

- 테스트는 두 기관의 전문 엔지니어, 과학자, 주제별 전문가들이 수행했으며, 그 결과는 모델이 공개되기 전 각 사(앤트로픽, 오픈AI)와 테스트 결과 공유

11 업그레이드된 앤트로픽의 클로드 3.5 소네트는 2024년 10월 22일 출시되었으며 오픈AI의 01은 2024년 12월 5일에 출시

Ⅳ. 세부 이행 내용(2): 타 연방부처

4.1 지침 또는 표준 수립

■ 백악관 산하 예산관리국(Office of Management and Budget, OMB)은 연방 정부의 안전하고 책임 있는 AI 사용을 지원하기 위한 지침을 발표

- AI의 거버넌스를 강화하고, 위험을 완화하며, 연방 정부의 AI 활용에서 혁신을 촉진하기 위한 최초의 정부 차원의 정책(OMB, 2024)
  - (AI 거버넌스 강화) 각 기관이 AI 사용을 효과적으로 조정할 수 있도록 '최고 AI 책임자(Chief AI Officer)'를 지정하고 AI 거버넌스위원회를 설립 필요 강조
  - (AI 위험 대응) 연방 정부기관은 2024년 12월 1일까지 의료, 교육, 고용, 주택 등의 분야에서 미국인의 권리나 안전에 영향을 미칠 수 있는 방식의 AI 사용 시 구체적인 보호장치*를 마련

    * 안전 장치에는 공공에 미치는 AI의 영향을 신뢰할 수 있게 평가, 테스트, 모니터링하기 위한 일련의 필수적인 위험 관리 관행과 정부의 AI 사용 방식에 대한 더 높은 투명성을 제공하기 위한 조치가 포함
  - (안전한 AI도입 지원) 연방 정부기관들이 적절한 보호장치를 갖추고 사회의 시급한 과제 해결에 AI 기술을 도입할 수 있도록 지원

■ 국무부(Department of State, DOS)는 AI로 인한 인권 위험관리 지침* 발표

* 보고서 명: Risk Management Profile for Artificial Intelligence and Human Rights(DOS, 2024.7.)

- (목표) 정부와 민간, 시민사회의 조직이 국제 인권을 존중하는 방식으로 AI를 설계·배포·사용·관리할 수 있도록 지침 제공
  - 위험 관리 방식과 인권 간 격차의 해소와 인권 위험을 평가, 해결, 완화하는 조치와 여타 위험 관리 관행의 조화를 추구
- (배경) 국제 인권은 보편적으로 적용될 수 있는 공통 국제 언어이자 AI의 설계·배포·사용·거버넌스에서 중요한 역할을 하는 정부 및 민간 분야와 모두 관계되며, AI로 인한 위험 중 상당수는 인권과 관련된다는 점에서 AI 위험 관리를 위한 규범적 기반이 되기에 적합
  - 행위자의 의도하거나 의도하지 않은 결과로 AI 수명주기 전반에서 인권과 관련된 위험이 발생 가능

    * 개인에 대한 감시, 정보 검열과 통제, 편견과 차별의 고착화, 표현의 자유 침해, 개인정보와 프라이버시 침해 등이 대표적 인권 위험
- (주요 내용) NIST의 AI 위험관리 프레임워크(AI RMF)가 제시한 AI 위험 관리를 위한 △거버넌스 △매핑(위험식별) △측정 △관리 4개 기능을 기반으로 인권 위험 해결에 도움이 되는 모범 관행을 안내
  - (거버넌스) AI 활동과 인권에 관한 정책을 공개적으로 수립하고 위험 관리 프로세스에 알고리즘 영향평가, 개인정보 영향평가 및 인권 실사 절차를 통합

- (매핑) AI 수명주기 전반에서 다양한 내외부 관계자와 최종사용자, 영향을 받을 수 있는 커뮤니티와 정기적으로 협의하고, AI 시스템의 이점과 도입 비용을 고려해 AI 시스템이 인권 위험이 적은 대안보다 더 적합한지를 입증
- (측정) AI 시스템이 개인, 집단, 사회에 미치는 오류와 영향을 평가하고 설명하는 측정 항목을 포함하여 인권 관련 위험을 식별하며, 영향평가 시 개인의 권리와 성평등, 노동권, 환경 및 생태계에 미치는 영향, 윤리적·사회적 영향을 파악
- (관리) 위험 처리 가능성, 사용가능한 자원 또는 방법, 영향에 따라 AI 위험의 처리 우선순위를 정하되, 명확한 인권 관련 위험을 최우선으로 대응하고, AI로 인해 권리를 침해당한 사용자를 구제하기 위한 체계와 담당 부서를 수립

■ 미국특허청(United States Patent and Trademark Office, USPTO)은 AI 기술 및 기타 신흥 기술과 관련된 발명품에 대한 특허 청구의 적격성 평가에 관한 지침 개정본 발표(USPTO, 2024)

- (목표) 지침은 미국 특허법 35 § U.S.C. 101(Inventions patentable)에 따라 USPTO 직원과 이해관계자의 AI 기술 관련 특허 청구의 적격성을 평가 과정을 지원하는 데 목적
- (개정내용) USPTO가 발표한 개정본은 기존 특허 대상 적격성 평가에 관한 지침의 일관성을 유지하되 AI 기술 및 기타 신흥 기술과 관련된 발명을 적용시키기 위함이며 가까운 시일 내에 심사지침(MPEP)에 통합될 예정
- (세부내용) 구체적으로 동 지침은 AI 분야에서 발명하는 사람들이 AI 발명품을 보호하고 특허 심사관이 AI 발명품에 대한 특허 신청을 검토하도록 정보 제공
  - 특히, 지침은 AI 기술 관련 발명에 대한 특허 심사, 항소, 등록 후(post-grant) 절차에서 USPTO 담당자와 이해관계자에게 도움이 되고자 가상의 여러 예시 및 평가 기준*을 제공

    * 일례로, AI 기반 음성 신호 처리 특허와 관련하여 △ 청구항의 추상적인 아이디어 인용 여부, △ 추상적인 아이디어가 포함된 청구항의 실제 출원 가능성 등의 평가 기준을 제공
  - USPTO 심사관이 해당 지침을 바탕으로 AI 기술 관련 특허 청구의 적격성이 부족하다고 판단되는 경우에 해당 기술 관련 특허 출원인은 청구가 적격하다고 판단되는 이유에 대한 소명 기회가 주어지게 됨

■ 교육부(Department of Education, ED)는 교육에 사용할 안전하고 보안성이 높고 신뢰할 수 있는 AI 도구를 설계하기 위한 가이드* 발표

* 가이드 명: Designing for Education with Artificial Intelligence: An Essential Guide for Developers(DOE, 2024)

- 가이드는 에듀테크 개발자가 학생과 교사에게 이로우면서도 동시에 형평성, 시민권, 신뢰, 투명성을 증진하는 AI를 설계하는 방법을 설명하고, 교육 및 학습에 AI를 사용하기 위한 권장 사항을 제시
  - (AI 설계에 교육적 가치 반영) AI 개발자는 AI를 단순한 기술적 도구가 아닌 교육의 본질과 가치를 반영하여 설계해야하며, 이를 위해 학생과 교사의 협력을 바탕으로 한 인간 중심적(human-centered design) 접근 방식 채택 필요
  - (근거 기반 설계) AI 개발자는 AI 기술이 학습 성과에 긍정적 영향을 미친다는 과학적 근거와 평가 자료를

제공하여, 교육 기관과 학교가 도입 여부를 신중히 판단하는데 도움 제공 필요

- (시민권 및 형평성 보호) AI 시스템은 교육 과정에서 공정한 사용을 위해 데이터 편향을 줄이고 모든 학생에게 공평한 접근성을 보장하는 포괄적 설계를 채택해야 하며, 특히, 특정 배경(성별, 인종 등)에 따른 차별적 결과를 방지하기 위해 지속적으로 검토·개선이 이루어져야 함
- (안전 및 보안 강화) AI 개발자는 교육 기관에 AI 시스템의 개발 및 배포 과정에서 관리 및 모니터링 절차를 구축하는 등의 개인정보 보호 및 보안 규정을 철저히 준수해야 할 필요
- (투명성 증진) AI 개발자는 AI 시스템 개발 및 운영에서 윤리적 가치를 중심에 두고, AI의 기능과 한계, 활용에 따른 책임을 다양한 교육 기관 및 이해관계자에게 공유함으로써 '신뢰 기반의 에듀테크 생태계' 구축에 기여

4.2 AI안전 관리 프레임워크 관련 활동 및 도구 개발과 보급

■ 국방부(Department of Defense, DOD) 및 국토안보부(Department of Homeland Security, DHS)는 국가 안보 향상을 위한 새로운 AI 도구 시범 도입 및 배치

- (목표) DOD와 DHS는 정부의 주요 시스템 및 소프트웨어의 취약점을 식별하고 해결하기 위해 새로운 AI 도구 시범 운영을 실시(DHS, 2024)
  - AI를 활용하여 미국 핵심 인프라와 네트워크의 안전 및 보안을 강화하고, AI를 활용한 사이버 위협 방어에 새로운 기준을 제시
- (배경) AI는 국가 안보와 공공 안전을 강화하는 데 중요한 기술로 자리 잡고 있는 동시에, 소프트웨어와 네트워크 시스템의 취약점은 점점 더 복잡하고 정교해지는 사이버 위협에 노출되어 있으며, 이를 효과적으로 탐지하고 해결하는 것은 필수적
  - DOD는 국가 안보 및 군사 목적으로 사용되는 소프트웨어의 취약점을 식별하고 해결할 수 있는 AI 시범 프로젝트에서 진전
  - DHS는 미국 국민이 자주 사용하는 주요 정부 소프트웨어 시스템의 취약점을 찾아내고 해결하기 위해 다양한 AI 도구 시범 운영을 진행하며 DOD의 노력을 보완
- (주요내용) 이상의 과정을 토대로 DOD와 DHS는 각각 국가 안보 목적과 민간 정부 시스템에서 사용되는 정부 네트워크의 취약점을 해결하기 위한 AI 시범 운영 결과를 발표
  - DOD는 AI 시범 프로젝트를 통해 국가 안보와 군사 목적으로 사용되는 소프트웨어의 취약점을 식별하고 수정하는 데 있어 AI 도구의 활용은 데이터 분석 및 의사결정 속도 향상에 기여할 수 있음을 확인
  - 한편, DHS는 시범 운영을 통해 현재 주요 정부 소프트웨어 시스템의 취약점 탐지에 활용되는 AI가 예측 불가능한 측면이 있어 기존 도구를 대체하기보다는 보완하는 데 더 효과적임을 확인했으며, AI 기술을 효과적으로 통합하고 사용하는 데 신중하고 현실적인 접근이 필요하다는 점을 강조

■ 에너지부(Department of Energy, DOE)는 AI 테스트베드와 모델 평가 도구를 개발 및 확장

● 국립연구소들과, NIST, NSF, 민간 부분 등과의 협업으로 테스트베드(testbeds)를 사용하여 AI 모델의 안전성과 보안을 평가(DOE, 2024)

- 테스트베드를 활용하여 AI 모델이 주요 기반시설, 에너지 안전 및 국가안보에 미칠 수 있는 위험에 대해 평가를 진행
- 또한 테스트베드는 AI 신뢰성을 개선하는 개인정보보호 강화 기술을 포함하여 새로운 AI 하드웨어 및 소프트웨어 시스템을 탐색하는 목적 등에서 활용
- NSF와 협력하여 AI 프라이버시 연구 강화를 위한 연구 협력 네트워크를 출범시켰으며, 미국 AI 안전연구소와 협력하여 프라이버시 강화 기술의 개발 및 평가, 배포, 지침 마련에 참여할 예정
- 개방형 연구 지원, 기술 혁신 촉진, 국가 안보 보장이라는 세 가지 중요한 목표를 조화롭게 추구하며, AI 시스템의 안전성, 보안성, 신뢰성을 강화하고, 민감한 데이터 처리 및 보호 기술 발전에 중요한 역할을 수행

4.3 정책 및 전략 보고서 발간

■ 통신정보관리청(National Telecommunications and Information Administration, NTIA)는 오픈소스 기반모델의 위험에 대한 모니터링의 필요성을 강조한 보고서* 발간

* 보고서 명: Dual-Use Foundation Models with Widely Available Model Wights Report (NTIA, 2024.7.)

- 보고서는 대규모 오픈소스 AI 모델의 잠재적 위험과 이점을 분석하고, 위험을 줄이면서 이점을 극대화할 수 있는 정책 권고안을 마련하도록 지시한 백악관 행정명령에 따라 작성

● 보고서는 오픈소스 기반 모델의 위험 관리를 위해 △모델 가중치 공개 제한 △기반 모델 생태계에 대한 평가 및 대응 역량 강화 △개방성의 수용과 촉진이라는 세 가지 정책적 접근 방식을 검토하였으며, 모델 가중치의 공개를 제한해서는 안 된다는 결론을 도출

- 모델 가중치 공개를 제한할 경우 AI 모델의 투명성을 저하시킬 뿐만 아니라, 기초 연구를 저해하여 투자와 인재가 해외로 유출될 가능성이 있고, AI의 안전성, 보안, 신뢰성과 같은 핵심 연구 분야의 발전이 지연될 우려도 제기

● 보고서는 연방 정부가 △증거 수집 △증거 평가 △증거 기반 조치라는 세 단계로 구성된 프로그램을 통해 새로운 위험을 적극적으로 모니터링할 것을 권장

- (증거 수집) 폐쇄형 모델과 오픈소스 모델을 포함한 첨단 AI 모델의 안전성에 대한 연구를 수행하고, 오픈소스 기반 모델의 위험과 이점에 대한 외부 연구를 지원하며, 발견된 위험 수준을 나타낼 수 있는 지표를 개발
- (증거 평가) 정책 개입의 기준으로 활용할 수 있도록 위험 지표의 임계값을 설정하고, 위험을 모니터링하고 대응하기 위한 벤치마크를 마련
- (증거 기반 조치) 증거 평가 결과 추가 조치가 필요하다고 판단될 경우, 정부는 오픈소스 기반 모델에 대한 접근 제한 등 위험을 완화하기 위한 조치를 고려 가능

■ 국가과학기술위원회(NSTC)는 지난 4년간 신뢰할 수 있는 AI를 발전시키기 위한 연방 연구개발(R&D)에 관한 보고서*발간

* 보고서 명: 2020-2024 Progress Report: Advancing Trustworthy Artificial Intelligence Research and Development(NSTC, 2024)

● 과학기술정책국(Office of Science and Technology Policy, OSTP)에서 발표한 '2023 국가 AI R&D 전략계획(National AI R&D Strategic Plan: 2023 Update)'(OSTP, 2023. 5.)의 전략적 우선순위*를 실현하기 위해 연방 기관들이 이룩한 주요 성과를 체계적으로 제시

- '2023년 국가 AI R&D 전략계획'에 제시된 9개 전략별로 미국 국립보건원(NIH[12]), NSF, 국방부, 국방고등연구계획국(DARPA[13]) 등 주요 연방 기관 및 부처의 AI R&D 프로그램을 개괄
- 연방 정부의 AI R&D 투자가 AI 발전을 이끌고 경제의 다양한 부문에 긍정적인 영향을 미치고 있음을 강조
- 연방기관들은 사이버보안 강화, 의료 발전, 국방 지원, 과학 발견, 기후 변화 대응, 제조 및 운송 등 주요 분야의 운영 효율성 증대와 같은 중요한 과제를 해결하기 위해 상당한 투자를 주도

<표 4-1> 국가 AI R&D 9개 전략별 주요 연방 정부의 AI R&D 사례

전략 내용	연방 기관 및 부처의 주요 AI R&D 사례
전략 1: 책임 있는 AI 연구에 지속적인 투자	• 국방부(DOD)는 부서 전반에서 AI 기반 솔루션 개발에 투자하며, 검증된 솔루션은 선별적으로 확장하여 기관 전체에 적용
전략 2: AI와 인간의 협업을 위한 효과적 방법 개발	• 국가핵안보국(NNSA)은 AI 기반 실시간 오류 검사 기술을 도입하여 무기 시스템의 부품 인증에 도움
전략 3: AI의 윤리적·법적·사회적 영향에 대한 이해 및 대응	• 교육부(ED)는 AI 확산이 사회와 개인에 미치는 영향에 대응하기 위해, 교육 분야에서 AI 활용 사례와 적절한 개발 프로세스 마련을 위한 정책과 지침을 개발 중
전략 4: AI 시스템의 안전성 및 보안 보장	• 국방부(DOD), 국방고등연구계획국(DARPA) 등은 AI 시스템의 안전성 및 보안 보장하는 기술을 개발 중
전략5:AI 학습 및 테스트를 위한 공용 데이터셋과 환경 개발	• 국립보건원(NIH)은 'AI 브릿지(Bridge) 프로그램'을 통해 의료 분야에서 AI 연구에 활용할 수 있는 데이터셋을 구축
전략6: 표준과 벤치마크를 통해 AI 기술을 측정 및 평가	• 국립표준기술연구소(NIST)와 국립과학재단(NSF)는 공동 자금 지원을 통해 AI 시스템의 핵심 표준과 벤치마크를 개발하는 노력을 뒷받침
전략7: 국가 AI R&D 인력 수요에 대한 정확한 예측	• 에너지부(DOE)는 장학금 프로그램과 연구 자금 지원을 통해 AI, 컴퓨터, 데이터 과학 분야의 인재 양성을 적극적으로 지원
전략8: AI 시스템 발전을 위한 공공-민간 파트너십 확대	• 국토안보부 과학기술국(Science and Technology Directorate, DHS S&T)은 중소기업 혁신 연구실 및 실리콘밸리 혁신 프로그램을 통해 다양한 중소기업들과의 협력을 강화
전략9: AI 연구에서 원칙적이고 조율된 국제 협력 방식 구축	• 국방부(DOD)는 군사 분야에서 AI 응용 프로그램의 개발과 구현을 위해 미국-영국-호주 간 3자 안보 파트너십(AUKUS)을 비롯한 국제 파트너들과의 협력을 최우선

출처: NSTC (2024)

12 NIH, National Institutes of Health

13 DARPA, Defense Advanced Research Project Agency

■ 미국 백악관은 AI에 관한 최초의 국가안보 각서(NSM) 발표

- 바이든 행정부의 AI 행정명령에 따라 수립된 NSM*은 △AI 개발에서 세계 선도 △국가안보에 AI 활용 △AI를 중심으로 국제적 합의와 거버넌스 발전을 목표로 설정(The White House, 2024. 10.)

  * National Security Memorandum: 미국 대통령이 국가안보와 관련된 긴급 사안을 다루기 위해 정부 부처와 기관에 구체적인 지침을 전달하는 공식 문서

  - (AI 개발에서 세계 선도) 안전하고 신뢰할 수 있는 AI 개발에서 미국의 리더십 구축*

    * 주요 내용으로 △첨단 AI칩 확보 △경쟁국의 스파이 활동 대응 △AI안전연구소 활성화 △국가AI연구자원 프로젝트 강화 등이 포함

  - (국가 안보에 AI 활용) 국가 안보 분야에서 AI 거버넌스와 위험 관리를 강화하기 위한 체계를 마련하고, 프라이버시 침해, 편향 및 차별, 인권 침해와 같은 AI 관련 위험을 평가하고 완화할 수 있도록 구체적인 지침과 세부 내용을 제공
  - (AI 중심 국제적 합의와 거버넌스 발전) NSM은 미국 정부와 동맹국 및 파트너들이 협력하여 AI 거버넌스와 위험 관리를 개선하기 위한 체계를 마련하며, 국제법을 준수하면서 인권과 기본적 자유를 보호하는 방식으로 AI 기술의 개발 및 활용을 촉진하도록 보장

■ 재무부(Department of Treasury, DOT)는 AI 시스템에 대한 對중국 투자 제한 행정규칙 발표

- (목표) 행정규칙은 바이든 행정부의 행정명령 이행을 위한 실질적 조치로, 미국인(미국 기업 포함)의 AI 시스템에 대한 해외 투자가 우려 국가의 군사 현대화를 위한 핵심 기술 발전에 기여하지 않도록 막아, 미국의 안보를 보호하려는 목적(DOT, 2024)
  - 행정명령은 중국(홍콩과 마카오 포함)을 우려국으로 지정하고, 우려국이 미국의 투자로 민감한 기술과 제품을 개발할 수 없도록 재무부에 기존 수출 통제를 보완하는 국가 안보 투자 제한 규정의 제정을 요구
- (주요내용) 행정규칙의 적용 대상은 차세대 군사·정보·감시·사이버보안 역량 강화에 활용될 수 있는 △반도체와 마이크로전자공학 △양자정보기술 △특정 AI 시스템으로 구성되고, 행정규칙 위반 시에는 벌금 또는 민사 처벌이 부과
  - 이중 특정 AI 시스템과 관련하여, 행정규칙은 군사적 목적(e.g. 모의 전투, 무기 설계 및 제어, 군사 의사결정)이나 첩보 및 대규모 감시 용도로만 설계된 AI 시스템의 개발과 관련된 거래를 금지
  - 행정규칙을 위반할 경우, 최대 368,136달러(법정 최대치) 또는 거래 금액의 두 배 중 더 큰 금액에 해당하는 민사적 제재가 부과

4.4 국제협력 및 기타 협의체 운영

■ AI안전과 보안을 촉진하는 유엔(UN) 총회 결의안 주도(UN, 2024. 3.)

- 미국이 제안하고 중국과 120개 이상의 국가가 공동 발의하여 채택된 이 결의안은 전 세계 국가들이

글로벌 과제를 해결하기 위해 AI의 안전과 보안을 촉진하기 위한 공통된 비전을 제시

- 결의안은 "지속 가능한 개발을 위한 안전하고 안전하며 신뢰할 수 있는 인공 지능 시스템의 기회 포착"(문서 A/78/L.49)이라는 제목으로 투표 없이 채택
- 결의안은 AI의 급속한 발전으로 나타날 수 있는 부적절한 설계·개발·배포·사용이 인권과 기본적 자유의 보호, 증진을 저해할 잠재적 위험이 있음을 인식하고, 이를 예방하기 위해 안전하고 신뢰할 수 있는 AI 시스템을 구축하기 위한 글로벌 협력의 필요성을 강조
- 또한 결의안은 UN 회원국들이 개발도상국과 협력하여 기술 이전, 기술 지원, 자금 조달 등의 문제를 신속히 해결할 것을 권고하며, 국가 간 디지털 격차 해소에도 우선적으로 집중할 것을 촉구

■ AI의 책임 있는 군사적 사용에 대한 미국 주도의 정책 선언에 대한 세계적 지원 확대

● 미국 국무부는 'AI와 자율성의 책임 있는 군사적 이용에 관한 정치적 선언*을 발표

* 선언명: Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy(DOS, 2024. 11.)

- 한국, 독일, 프랑스, 영국, 일본, 미국 등 약 60여 개 국가가 참여하여 AI의 군사적 이용에 관한 책임 있는 개발,배치, 사용 규범에 대한 정치적 선언을 지지
- 이 정책 선언은 △ 군사적 AI 역량을 국제법과 일치시키는 데 필요한 조치 △ 핵무기 관련 주권적 결정에서 인간의 개입과 통제를 유지하는 방안 △ 무기 시스템 등 중대한 군사적 AI 역량의 개발 및 배치 시 고위 정부 관료의 감독 보장 등의 주요 내용으로 구성
- 이 선언에 포함된 조치들은 각국이 AI의 잠재적 위험을 줄이는 동시에 그 이점을 극대화할 수 있는 국제적 책임 체계를 마련하는 데 있어 중요한 진전을 나타냄

4.5 기타[14]

■ (R&D투자) 국립과학재단(NSF)은 사회 전반에 걸쳐 책임 있는 AI 개발과 사용을 발전시키기 위해 수백만 달러의 투자 계획 발표

● NSF는 AI와 관련된 문제를 포함하여 현실 문제 해결을 위한 개인정보보호 강화 기술(Privacy Enhancing Technology) 사용을 촉진하기 위한 2,300만 달러 규모의 이니셔티브를 시작

- 연방정부 외부의 연구자들이 AI 준비 테스트베드를 설계하고 계획하는 데 자금을 지원
- NSF는 업계 및 기관 파트너와 협력하여 특정 사례에 대한 개인정보보호 강화 기술을 적용, 확장하고 도입을 가속화하는 테스트베드를 구축하기 위한 노력에 투자

14 대회 및 공모전 개최, 대응재원조성, AI 인재 채용 등이 포함

- NSF는 '신흥 및 신규 기술에서의 체험 학습 프로그램(Experiential Learning for Emerging and Novel Technologies, ExLENT)'에 3,000만 달러를 투자를 발표(NSF, 2024. 7.)
  - 이 프로그램은 인공지능(AI)을 포함한 첨단 기술 분야에서 실습 기반 학습 기회를 제공하여, 다양한 배경의 개인들이 이러한 분야로 진입하거나 전문성을 높일 수 있도록 지원
  - AI를 비롯한 신기술 분야에서의 인력 양성을 촉진하여 미국 내 일자리 창출과 경제 성장이 궁극적 목표
- NSF는 다양한 연구자들의 AI 연구 역량을 구축하고 다양한 AI 준비 인력 개발을 촉진하는 'ExpandAI 프로그램'을 통해 1,000만 달러를 투자

■ (연구 인프라 지원) 국가인공지능연구자원(NAIRR)* 시범사업을 통해 80개 이상의 연구팀에게 AI 자원을 이용하도록 지원

* NAIRR(The National Artificial Intelligence Research Resource)은 NSF가 주도하고 에너지부(DOE), 국립보건원(NIH) 및 기타 정부 및 비정부 파트너와 협력하여 국가의 AI 연구 및 교육 커뮤니티를 지원하기 위해 자원을 제공하는 국가적 인프라

- 국가 AI 자원 이용 지원사업을 통해 딥페이크 탐지, AI 안전성 향상, 차세대 의료 진단 및 기타 중요한 AI 과제를 해결하도록 지원

■ (인적 역량 강화) 연방정부 전반의 AI 역량 강화를 위해 AI 인재 확대 프로그램 수행

- AI 기술은 정부가 더 나은 결과를 제공하도록 돕는 동시에 차별이나 안전 문제의 위험을 초래할 수 있어, 이를 관리하고 AI 관련 임무를 발전시키기 위해 연방정부에 AI 전문가를 영입하는 것이 중요
- OMB는 연방 정부기관의 안전하고 책임 있는 AI 사용을 지원하기 위한 정책으로 AI 인재 확대와 역량 향상을 지시(OMB, 2024)
  - 100명의 AI 전문가를 채용할 계획을 발표했으며, 인사관리처(Office of Personel Management, OPM)는 정부 내 AI 인재의 유지를 위해 유연한 급여 및 휴가 제도를 적용할 수 있는 지침을 수립
  - 2025 회계연도 대통령 예산안에는 정부 전반의 AI 교육 프로그램 확장을 지원하기 위해 500만 달러의 추가 예산이 포함
- 백악관 과학기술정책실(OSTP)은 AI 인재 급증을 담은 보고서*를 기반으로 기술 생태계 전반에 걸친 새로운 계획을 발표(The White House, 2024.7.)

* 보고서명 : Increasing AI Capacity Across The Federal Government(The White House, 2024.4.)

  - 새로운 계획에는 광범위한 공익 기술 생태계를 강화하고 기술자를 정부 서비스에 투입하기 위한 인프라를 구축에 약 1억 달러의 자금 지원 계획을 포함
- 연방조달청(General Services Administration, GSA)의 대통령 혁신 펠로우 AI 코호트(Presidential Innovation Fellows AI cohort)와 국토안보부(DHS) AI 군단(DHS AI Corps) 프로그램 등을 통해 250명 이상의 인재 채용
  - AI 인재는 AI를 활용하여 최고 수준의 정부 서비스를 제공하고, AI 사용 시 대중의 권리와 안전을 보호하는 등 중요한 AI 우선순위를 달성하는 데 중요한 역할을 수행

V. 결론 및 시사점

■ 미국은 인공지능 기술을 국가 안보와 직결된 전략 기술로서 인식하고 관련 정책을 수립·추진

- 오바마 행정부부터 향후 인공지능이 자국의 안보와 글로벌 리더십 확보에 중요한 전략 과제임을 인식하고 연구 개발 계획부터 안전·신뢰성 보장을 위한 지침 마련까지 지속적으로 정책을 추진
- 트럼프 1기 행정부는 보다 적극적으로 AI 이니셔티브 법을 제정하고, NAIIO, NIST와 같은 거버넌스를 토대로 AI 안전·신뢰성 확보 이행 조치를 체계적으로 점검 관리
- 바이든 행정부에서는 부상하는 인공지능의 잠재적 위험성에 대응해 연방 정부 차원에서 신뢰할 수 있는 인공지능 개발과 확산을 위한 행정명령을 발동하고 국제협력을 강화
  - 행정명령(E14110)은 백악관에서 90, 120, 180일 등 이행 사항을 주기적으로 취합·보고해 체계적 관리

■ 특히, 상무부 산하의 국가표준기술연구소(NIST)가 중심이 되어 공공 부문의 인공지능 도입을 위한 위험관리프레임워크를 마련·보급하고 있으며 AI안전연구소를 통해 민관협력 추진

- NIST가 개발한 AI 위험관리프레임워크를 비롯한 각종 생성AI 도입 및 활용 지침들은 아직 초기 단계로 연방 정부 차원의 구체적 적용 사례는 많지 않으나 향후 정부기관과 기업에서 점진적으로 도입될 전망
- 지침이 실제 도입 사례를 지속적으로 모니터링하고, 국내 적용 가능성을 면밀히 검토하여 최적의 구현 방안 도출 필요

※ 2024년 1월, 미국 하원의 테드 리우 의원 등은 2024년 연방 인공지능 위험 관리법(US Federal Artificial Intelligence Risk Management Act of 2024)을 발의했으며, 이 초당적 법안은 미국의 연방기관들이 국립표준기술원(NIST)에서 개발한 인공지능 위험 관리 프레임워크(AI RMF)에 기반하여 작성됨(다만, 국가 안보 시스템에 대해서는 예외 적용)(Holistic AI, 2024.1.16.)

■ 국내 상황에 적합한 AI 시스템 개발 및 운영을 위한 안전성과 보안성 강화 방안 마련 필요

- 국내에서도 인공지능의 안전한 개발 및 활용을 위한 국제 사회의 논의와 관련 지침, 표준들을 지속적으로 참고하여 국내 실정에 맞는 지침, 표준, 및 규제 도입 검토 필요
- 대학 등 교육기관은 AI 교육 프로그램에 윤리적 책임과 AI 시스템의 오용 방지 내용을 포함하고, AI와 관련된 기술적, 윤리적 이슈를 이해하고 관리할 수 있는 전문 인재를 양성하는 교육 프로그램을 강화
- 인공지능 거버넌스 체계를 통해 AI 안전·신뢰성 기반 기술 개발 및 보급하고 민관협력을 통해 AI 위험 평가, 관리 지침, 표준 및 선도 원천 기술 개발 필요
  - 이중사용 AI 모델의 위험 식별, 예방, 완화를 위한 연구 강화, AI 안전·신뢰 기술의 국제 표준화 및 국제 협력, 다양한 산업 분야에서 AI 안전·신뢰를 보장할 수 있는 AI 전문 양성 필요

■ 트럼프2.0시대에서도 국가 안보 기술로서 AI 기술의 안전·신뢰성 확보를 위한 정책은 지속될 전망

- 현재 AI위험 관리 프레임워크의 근간이 되는 AI RMF1.0('23.1 공개)은 트럼프1.0 시대에 제정된 국가 AI이니셔티브법('20)에 근거하여 추진되었다는 점에서 트럼프2.0 시대에도 AI 기술의 안전·신뢰성 확보를 위한 정책은 승계 및 강화 예상
- 또한, 트럼프는 '신뢰할 수 있는 인공지능 활용을 촉진하는 행정명령(EO 13960)'을 통해 연방정부 차원의 신뢰할 수 있는 AI 도입을 권고하였으므로, 바이든 행정부에서는 이를 계승 강화한 것임
- 따라서, 바이든 정부에서 추진한 인공지능의 안전·신뢰성 확보 정책은 국가적으로 일관된 정책 방향이며, 오히려 국가 안보와 미국 우선주의를 강조하는 트럼프2.0시대에서는 더욱 강화될 수 있을 것으로 예측
- 다만, 바이든 행정부가 AI 기술의 안전성과 책임성을 확보하기 위해 기업들을 대상으로 강력한 규제적 접근을 확대했던 반면, 트럼프2.0 시대에는 규제를 완화하면서도 AI 기술의 전략적 활용과 글로벌 경쟁력 강화를 목표로 정책의 방향성이 다소 선회할 가능성 존재

■ 트럼프2.0시대 대응 위한 韓 AI 전략 마련 필요

- 트럼프 2.0 시대에는 AI 기술의 안전·신뢰성 확보를 강조하는 동시에, 수출 통제 및 국가 안보와 관련된 기술 관리 강화를 통해 미국의 글로벌 리더십을 유지하는 데 중점을 둘 것으로 전망
- 이에 한국은 글로벌 규제 환경에 선제적으로 대응하면서도, AI 기술 개발과 산업 진흥을 선도하는 미국 주도의 경쟁 구도에 발맞추는 정책적 균형이 필요
  - AI 기술의 윤리적 사용과 신뢰성 확보를 목표로 미국과의 기술 동맹을 강화하고, 국제 규제와 조화를 이루는 정책을 마련
  - 또한, 'AI 기본법' 등의 국내 정책이 글로벌 표준과 연계될 수 있도록 설계해 국내 기업의 해외 진출을 지원하고, 미국의 수출 통제와 기술 관리 강화에 따른 영향을 최소화하는 전략적 접근이 요구
  - 동시에, AI 기술의 독자적 경쟁력을 확보하기 위해 국가 차원의 연구개발(R&D) 투자를 확대하고, 글로벌 협력 네트워크를 강화하며, 국내 스타트업 및 기업 생태계를 육성하는 포괄적 전략이 필요

참고문헌

- American AI Initiative (2019.2.11), EO 13859 "Maintaining American Leadership in Artificial Intelligence"
- American AI Initiative (2020.12.3.), EO 13960 "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government"
- AI Safety Institute (2024.12.18.), Pre-Deployment Evaluation of OpenAI's o1 Model
- DHS (2024), FACT SHEET: DHS Completes First Phase of AI Technology Pilots, Hires New AI Corps Members, Furthers Efforts for Safe and Secure AI Use and Development
- DOT (2024.10.28.), U.S. Department of the Treasury, Additional Information on Final Regulations Implementing Outbound Investment Executive Order (E.O. 14105)
- Executive Office of the President National Science and Technology Council Committee on Technology (2016.10.), Preparing For The Future of Artificial Intelligence
- NIST (2023), NIST AI RMF Playbook
- NIST (2023.1.), Artificial Intelligence Risk Management Framework (AI RMF 1.0)
- NIST (2024.7.), A Plan for Global Engagement on AI Standards
- NIST (2024.7.), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile
- NIST (2024.7.), Managing Misuse Risk for Dual-Use Foundation Models
- NIST (2024.7.), Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile
- NIST (2024.11.19.), Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet
- NIST (2024.11.20.), FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco
- NIST (2024.11.20.), U.S. AI Safety Institute Establishes New U.S. Government Taskforce to Collaborate on Research and Testing of AI Models to Manage National Security Capabilities & Risks
- NSTC (2016.10), The National Artificial Intelligence Research And Development Strategic Plan
- NSTC (2024. 7.), 2020-2024 Progress Report: Advancing Trustworthy Artificial Intelligence Research and Development
- NTIA (2024.7.30.), Dual-Use Foundation Models with Widely Available Model Wights Report
- OMB (2024. 3.), Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence
- Select Committee on Artificial Intelligence of the National Science and Technology Council

(2023.5), National Artificial Intelligence Research And Development Strategic Plan2023 Update
- The U.S. Department of Commerce(DOC) (2024. 4.1.), U.S. and UK Announce Partnership on Science of AI Safety
- The U.S. Department of Energy(DOE) (2024. 7), Artificial Intelligence Testbeds at DOE
- The U.S. Department of Education(DOE) (2024. 7. 8.), Designing for Education with Artificial Intelligence: An Essential Guide for Developers
- The U.S. Department of State(DOS) (2024. 7. 25.), Risk Management Profile for Artificial Intelligence and Human Rights
- The U.S. Department of State(DOS) (2024. 11. 1.), Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy
- The U.S. National Science Foundation(NSF) (2024. 7. 16.), NSF invests more than $30M to enable experiential learning in key technologies
- The U.S. Patent and Trademark Office(USPTO) (2024. 7. 16.), USPTO issues AI subject matter eligibility guidance
- The White House (2016.10.), The Administration's Report on the Future of Artificial Intelligence
- The Whitehouse (2021.1), The White House Launches the National Artificial Intelligence Initiative Office
- The White House (2023.10.30.), Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
- The White House (2024.4.), Increasing AI Capacity Across The Federal Government
- The White House (2024.7.16.), Fact Sheet: Biden-Harris Administration Announces Commitments from Across Technology Ecosystem including Nearly $100 Million to Advance Public Interest Technology
- The White House (2024.10.24.), FACT SHEET: Biden-Harris Administration Outlines Coordinated Approach to Harness Power of AI for U.S. National Security
- The White House (2024.10.30.), Fact Sheet: Key AI Accomplishments in the Year Since the Biden-Harris Administration's Landmark Executive Order
- The Whitehouse OSTP (2020.2.), American Artificial Intelligence Initiative:Year One Anuual Report
- The Whitehouse OSTP (2023. 5.), National Artificial Intelligence Research and Development Strategic Plan 2023 Update
- United Nations(UN) (2024. 3. 21.), General Assembly Adopts Landmark Resolution on Steering Artificial Intelligence towards Global Good, Faster Realization of Sustainable Development
- Holistic AI (2024.1.16.), US Federal Artificial Intelligence Risk Management Act of 2024 Introduced

주 의

이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
이 보고서의 내용을 발표할 때에는 반드시
소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

[소프트웨어정책연구소]에 의해 작성된 **[SPRI 보고서]**는 공공저작물 자유이용허락 표시기준 제4유형**(출처표시-상업적이용금지-변경금지)**에 따라 이용할 수 있습니다.

미국의 AI안전·신뢰성 정책 현황과 시사점

Current Status and Implications of U.S. Policy for Implementing Trustworthiness AI

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D 연구동(B) 4층
Global R&D Center 4F 22 Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do

www.spri.kr ISSN