

KISA INSIGHT

DIGITAL & SECURITY POLICY

2023 VOL. 6

美 AI 대통령 행정명령 등 인공지능(AI) 안전 및 보안 규범 분석 및 시사점

김도원, 하병욱, 김성훈



美 AI 대통령 행정명령 등 인공지능(AI) 안전 및 보안 규범 분석 및 시사점

김도원, 하병욱, 김성훈

CONTENTS

I	인공지능(AI) 안전 및 보안 국제규범 논의	
	1-1. 인공지능 시대의 촉발	2
	1-2. 인공지능 보안을 위한 규제 논의 촉진	4
II	G7, ‘히로시마 AI 프로세스’의 AI	7
III	AI 안전 정상회의(AI Safety Summit)의 블레츨리 선언	10
IV	美 행정명령, Safe, Secure, and Trustworthy AI	
	4-1. 행정명령 8개 분야 정책 및 원칙	12
	4-2. 인공지능의 안전과 보안 확보(Sec.4) 세부 내용	15
	4-3. 개인정보 보호(Sec.9) 세부 내용	18
V	시사점	19

『KISA Insight』는 디지털·정보보호 관련 글로벌 트렌드 및 주요 이슈를 분석하여 정책 자료로 활용하기 위해 한국인터넷진흥원에서 기획, 발간하는 심층보고서입니다.

한국인터넷진흥원의 승인 없이 본 보고서의 무단전재나 복제를 금하며 인용하실 때는 반드시 『KISA Insight』라고 밝혀주시기 바랍니다. 본문 내용은 한국인터넷진흥원의 공식 견해가 아님을 알려드립니다.

[작성]

한국인터넷진흥원(KISA) 미래정책연구실 정책개발팀

김도원 선임연구원

☎ 061-820-1228

✉ downonkim@kisa.or.kr

하병욱 책임연구원

☎ 061-820-1288

✉ ha@kisa.or.kr

김성훈 팀장

☎ 061-820-1426

✉ shkim@kisa.or.kr

■ **생성형 AI 기술을 비롯한 인공지능 기술의 확산으로 잠재적 위협에 대한 우려가 증가하면서, 이러한 위협에 대응하기 위한 인공지능 보안 관련 규제 논의가 촉진**

- ChatGPT 출시 이후 인공지능 기술이 핵심 기술로 부상하면서, 인공지능 기술의 확산으로 인해 인공지능이 보유한 기술적 한계나 부정적인 평가에 대한 우려도 함께 증가
- 해외 주요국들은 인공지능의 잠재적 위협에 대응하기 위해 주요국들은 인공지능 기술의 위험성에 대해 구체적으로 조사하고 국가 차원의 대응 방안을 마련하기 시작하였음

■ **G7 정상회의에서 생성형 AI의 위협에 대한 논의가 이루어지고, AI 안전 정상회의에서 최초의 국제적 합의로 볼 수 있는 ‘블레츨리 선언’ 발표**

- '23년 5월, 히로시마에서 개최된 G7 정상회의에서 생성형 AI의 위협에 대해 논의하였으며, 이를 기반으로 11월, ‘히로시마 AI 프로세스’ 관련 성명을 발표하고 11개 개발자 행동 규범을 발표
- '23년 11월, AI 안전 정상회의에서는 28개 주요국이 참여하여 AI에 대한 위협을 논의하였고 참여국들은 AI의 기회와 위협에 대해 공감하고 협력의 필요성을 인정하는 ‘블레츨리 선언’에 합의

■ **미 바이든 정부는 미국이 인공지능에 대한 약속을 지키고 위협을 관리하는데 앞장서기 위해, 안전하고 신뢰할 수 있는 인공지능 개발과 활용을 위한 대통령 행정명령 발표**

- '23년 10월, 미 정부는 인공지능의 위협을 관리하기 위한 포괄적인 전략의 일환으로, 안전하고 신뢰할 수 있는 인공지능의 개발과 활용을 위한 대통령 행정명령 발표
- 인공지능에 대한 약속을 지키고 위협을 관리하는 데 앞장설 수 있도록 8개 분야에 대한 원칙과 함께 관련 행정부 및 기관들이 수행해야 하는 세부적인 조치사항을 제시

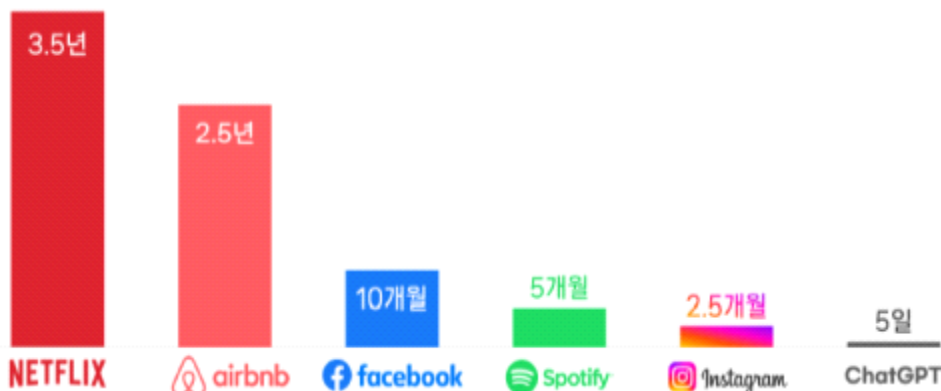
I

인공지능(AI) 안전 및 보안 국제규범 논의

1-1 인공지능 시대의 촉발

■ 인공지능 기술을 비롯한 IT 기술의 발전은 우리의 일상생활을 변화시켰으며, ChatGPT 출시 이후 놀라운 성능으로 인해 많은 사람들이 인공지능 기술에 더욱 주목하기 시작

- 인터넷, 스마트폰과 같은 새로운 IT 기술이 등장할 때마다 우리의 일상생활을 크게 변화시켜 왔으며, 인공지능 기술 또한 최근 급격한 성장을 이루면서 사회의 다양한 영역에서 점차 변화시켜 나가고 있음
- '22년 11월, OpenAI社は 대화형 인공지능 서비스 ChatGPT를 출시하였으며, 자연어 처리 모델을 기반으로 언어 관련 다양한 기능(생성, 답변, 번역, 요약 등)을 수행할 수 있고 소스코드 생성 및 수정도 가능
- ChatGPT가 사용자가 입력한 텍스트의 맥락을 파악하고 이전의 대화를 기억할 수 있어 마치 사람처럼 의사소통이 가능한 것처럼 보여, 단답형 대화나 정해진 범주 안에서 대화하는 기존 챗봇과 차별화
- ChatGPT는 뛰어난 언어 능력으로 인해 서비스가 출시되자마자 많은 사람들에게 폭발적인 관심을 받음
- 출시 5일 만에 100만 사용자를 확보하고 2달 만에 월간 사용자 1억 명을 돌파하는 등 전례 없는 기록 수립






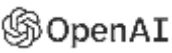





(그림 1) 주요 서비스 100만 사용자 달성 기간 (출처 : NIA)

■ 주요 빅테크 기업들은 특히 생성형 AI 기술에 집중하기 시작하였고, 다양한 산업 분야에서 생성형 AI를 적극적으로 도입하는 등 생성형 AI는 기업 생존을 위한 필수 기술로 부상

- 대부분의 빅테크 기업들은 생성형 AI 시장의 선도를 위해, 자사 데이터를 활용하여 자체적으로 생성형 AI 모델을 개발하거나 관련 기업 투자를 통해 생성형 AI 기술을 확보·도입하기 시작

[표 1] 주요 빅테크 기업 인공지능 모델 현황

구분	주요 내용
국내	 <ul style="list-style-type: none"> • 한국어 기반 초거대 인공지능 하이퍼클로바 공개('22.7.) • 하이퍼클로바X 및 관련 생성형 AI 서비스 라인업 공개('23.8.)
	 <ul style="list-style-type: none"> • 초거대 멀티모달 인공지능 민달리(minDALL-E) 공개('21.12.) • 한국어 특화 모델 코GPT 2.0 출시 예정('23.10. 이후)
	 <ul style="list-style-type: none"> • GPT 기반 대화형 서비스 에이닷(A.) 공개('22.5.) 및 정식 출시('23.9.) • 자사 모델 에이닷과 엔트로픽, 코난 모델을 결합하는 멀티 LLM 전략 발표('23.8.)
	 <ul style="list-style-type: none"> • 초거대 인공지능 모델 민:음(Mi:dm) 공개('23.10.)
	 <ul style="list-style-type: none"> • 초거대 인공지능 모델 엑사원(EXAONE) 공개('21.12.) • 초거대 양방향 멀티모달 인공지능 모델 엑사원 2.0 공개('23.7.)
국외	 <ul style="list-style-type: none"> • 대규모 언어모델 GPT-3.5 기반 대화형 인공지능 서비스 ChatGPT 출시('22.11.) • 이미지 생성형 인공지능 DALL·E 2('22.4.) 및 언어모델 GPT-4('23.3.) 공개
	 <ul style="list-style-type: none"> • GPT 기반 인공지능 검색 엔진 New Bing 출시('23.2.) • 자사 서비스(윈도우, 오피스)에 생성형 AI 도입 및 시큐리티 코파일럿 공개
	 <ul style="list-style-type: none"> • 대규모 언어모델 LaMDA('21.5.) 및 PaLM('22.4.) 공개 • 자사 모델을 기반으로 하는 인공지능 검색 엔진 서비스 Bard 출시('23.3.)
	 <ul style="list-style-type: none"> • 대규모 언어모델 LLaMA-2 공개('23.7.)

- 또한, 생성형 AI 기술은 ChatGPT 같은 대화형 서비스뿐만 아니라 여러 IT 서비스와 결합하면서 생산성을 향상시키고, 다른 산업 분야에서도 생성형 AI 기술의 적극적인 도입을 통해 혁신을 가져올 것으로 전망

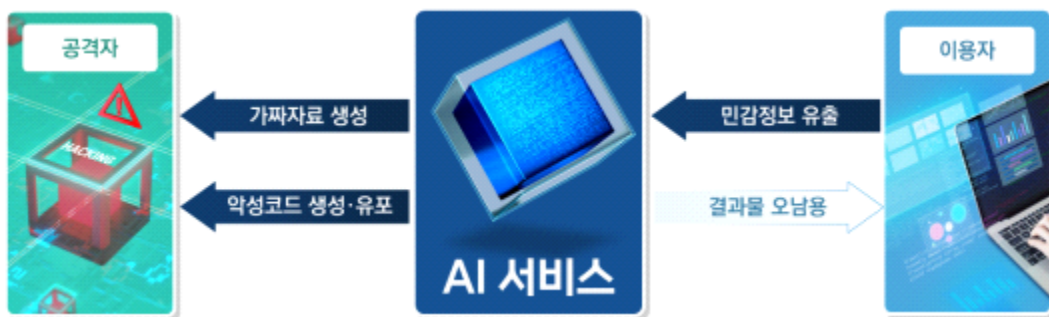
[표 2] 산업 분야별 초거대 AI 적용 사례 및 전망 (출처 : 초거대 AI 경쟁력 강화 방안)

구분	주요 내용
의료	<ul style="list-style-type: none"> • 의료영상 판독을 통한 질환 진단 보조 고도화, 임상·질환 합성 데이터 생성, 신약후보 물질 발굴, 약물 디자인 등 활용 확대 전망('23, 미래에셋)
마케팅	<ul style="list-style-type: none"> • '25년까지 생성형 AI가 대기업 마케팅 메시지의 30%를 만들 전망(가트너), 美 1,000대 기업은 콘텐츠 생성(58%), 고객지원(57%) 등에 ChatGPT 활용('23)
금융	<ul style="list-style-type: none"> • 고객상담, 금융상품 추천, 신용평가, 금융사고 감지 등 금융 전반에 활용 확대 전망, 글로벌 금융회사의 20%가 대화형 AI 도입 중(엔비디아, '23)
미디어	<ul style="list-style-type: none"> • ChatGPT를 활용한 사용자 맞춤형 콘텐츠, 퀴즈 제작을 발표(美 버즈피드), AI가 90%를 만든 블록버스터 영화가 '30년 1편 이상 개봉될 것(가트너)
법률	<ul style="list-style-type: none"> • 초거대 AI가 법률대리인의 서류 작업 등을 지원하고, 판사 업무를 보조하는 등 리걸테크 서비스 고도화 전망

1-2 인공지능 보안을 위한 규제 논의 촉진

■ 생성형 AI를 비롯한 인공지능 기술은 혁신적인 기술이지만, 인공지능 기술이 급속하게 확산되면서 여러 한계점이나 부작용에 대한 우려도 함께 증가

- 인공지능이 혁신을 불러올 기술임은 분명하지만, 인공지능이 보유한 **기술적 한계**와 **부정적인 평가** 등은 인공지능이 확산됨에 따라 **양날의 검으로 작용**하여 잠재적 위험이 될 수 있음
- 거짓을 사실처럼 답변하거나 최신 정보가 반영되지 못하는 등 인공지능 고유의 기술적 한계는 이용자가 인공지능 기반 서비스를 이용하는 데에 있어 여러 보안 사고의 발생 원인이 될 수 있음
- 또한, 생성형 AI 기술을 이용해 메일, 영상 등 **피싱 범죄에 필요한 자료를 생성**하거나 인공지능을 통해 **해커들이 공격 과정에서 도움을 얻는 등 악의적인 행위**에 인공지능의 **긍정적인 부분을 역으로 이용할 수 있음**



(그림 2) 인공지능 기술의 한계점과 부작용

- 인공지능 기술이 점차 보편화되어 이용률과 의존성이 높아질수록 그에 따른 부작용이 발생할 가능성도 증가할 것이며, 이러한 위험에 대비하지 않을 경우 **사회·경제적 손실**과 함께 국가의 **디지털 경쟁력 저하**의 **요인**으로 작용할 수도 있음

■ 해외 주요국들은 ChatGPT를 비롯하여 인공지능 기술의 위험성에 대해 조사하기 시작하였고, 이러한 잠재적 위험에 대응하기 위해 국가 차원의 대응 방안을 마련하기 시작

- ChatGPT 출시 이후 이탈리아는 EU GDPR 위반을 사유로 약 1달간 자국 내 서비스를 차단하였으며, EU 및 주요 국가들은 이와 관련하여 ChatGPT 조사에 착수하였음
- 이탈리아 개인정보 감독기관(GPDP)은 EU GDPR 위반*을 사유로 이탈리아 내 ChatGPT 서비스 제공 중단 명령(23.3.)을 내리고 OpenAI社에 시정 조치 이행을 요구하였으며, OpenAI는 이탈리아가 요구한 조치들을 시행한 후 접속 차단 해제(23.4.)

- 독일, 프랑스, 스페인 등은 이탈리아에 ChatGPT 처분 근거를 문의하고 자체적으로 조사에 착수
- 유럽데이터보호위원회(EDPB)는 EU 차원의 대응을 위해 ChatGPT 조사 전담반 출범('23.4.)
 - * ChatGPT의 학습 데이터에 포함된 개인정보의 적법 수집·처리 근거 부재, 사용자 연령 확인 방법 부재
- 또한, 주요 국가들에서는 ChatGPT를 비롯한 **인공지능 기술의 확산과 그에 따른 위협을 관리하기 위해, 국가 차원의 대응 방안을 논의하고 마련하기 시작**
- (미국) 인공지능의 안전한 사용 및 신뢰 환경 구축을 위한 **AI 규제(안)를 마련하기 위해 여론 수렴을 시작하고**(‘23.4.), 생성형 AI 위협을 해결하기 위한 워킹그룹 출범 및 AI 기술 선도 기업으로부터 위협관리에 대한 자발적 약속 확보(‘23.7.)
- (EU), **생성형 AI 모델을 고위험 AI 시스템으로 분류하고 사이버보안과 안전 관련 의무를 강화하는 등 AI 시스템의 일반원칙을 마련하고 사이버보안을 강화하는 「AI법(안)」 채택**(‘23.6.)
- (영국) AI 산업과 규제 가이드라인을 담은 「AI 백서」 발표(‘23.3)
 - ※ ChatGPT와 같은 AI가 사용자의 사생활, 인권, 안전 등에 미치는 영향력 분석 및 분석 결과에 따른 규제 제언
- (일본) 인공지능 확산에 대응하기 위해 **관계부처들로 구성된 ‘AI 전략팀’ 설치**(‘23.4)
- (중국) 사이버공간관리국(CAC)는 생성형 AI 서비스 관리를 위한 임시 조치 발표(‘23.7)
 - ※ 생성형 AI 기술 개발 여건을 보장함과 동시에 서비스 제공자와 관리감독 기관의 법적 책임을 강화

■ 인공지능의 안전과 신뢰를 위해 주요국들은 이와 관련하여 국제적 합의를 이루거나 필요한 조치 사항을 구체화하고 있음

- G7 정상회의에서 각국 정상들은 **생성형 AI의 위험에 대해 논의**(‘23.10)하였으며, 이를 발전시켜 **AI 안전 정상회의**에서 인공지능의 안전한 활용을 위한 **‘블레츨리 선언’에 합의**(‘23.11)
- 미국은 지금까지 **AI의 위험성에 대한 여러 조치 사항을** 기반으로 안전하고 신뢰할 수 있는 인공지능을 위한 조치 사항들을 포함하는 **대통령 행정명령을 발표**(‘23.10)

참고 1

주요국 인공지능 보안 대책 현황

키워드	주요 사항
미국	<ul style="list-style-type: none"> • NIST, 「인공지능 위험 관리 프레임워크 1.0」 발표('23.1) • 상무부, ChatGPT와 같이 잠재적인 위험성을 지닌 AI 모델의 출시 과정을 검증하기 위한 규정 제정 검토 추진('23.4) • 백악관, 책임 있는 인공지능 연구·개발·확장 촉진을 위한 새로운 조치 발표('23.6) • 백악관, AI 기술 선도 기업으로부터 위험 관리에 대한 자발적 약속 확보('23.7) <ul style="list-style-type: none"> * 아마존, 애플, 구글, 인플렉션, 메타, 마이크로소프트, OpenAI • NIST, 생성형 AI의 위험을 해결하기 위한 워킹그룹 출범('23.7) • 백악관, AI 사이버 챌린지 시작('23.8) • CISA, AI 소프트웨어 설계에 의한 보안(Secure by Design) 강조('23.8) • 백악관, 안전하고 신뢰할 수 있는 인공지능에 대한 행정명령 발표('23.10) • CISA, 안전한 AI 활용을 위한 2023-2024 AI 로드맵 발표('23.11)
EU	<ul style="list-style-type: none"> • EDPB(유럽데이터보호위원회), EU 차원의 대응을 위해 ChatGPT 조사 전담반 출범('23.4) • ENISA, AI 사이버보안 표준 연구 및 시범 구현을 위한 권고안 제시('23.5) • ENISA, AI 사이버보안 컨퍼런스에서 주요 인공지능 보안 문제 논의('23.6) • EU, AI 시스템의 일반원칙을 마련하고 사이버보안을 강화하는 「AI법」 수정안 채택('23.6) • EU-일본, AI 및 데이터 흐름에 관한 양자 간 협력 강화 약속('23.7) • EU 이사회, 집행위원회, 유럽의회, 「AI법」 최종 합의('23.12)
영국	<ul style="list-style-type: none"> • 정부, AI 산업과 규제 가이드라인을 담은 「AI 백서」 발표('23.3) • NCSC, ChatGPT 및 LLM의 사이버보안 이슈 분석('23.3) • ICO(개인정보감독기관), 생성형 AI 개발자 및 사용자를 위한 8가지 고려사항 규정('23.4) • CMA(경쟁시장청), 경쟁적인 AI 시장 선도 및 소비자 보호를 위한 원칙 제안('23.10) • NCSC, 美 CISA 및 주요 국제 파트너들과 AI 보안을 위한 새로운 국제 가이드라인 개발('23.11)
이탈리아	<ul style="list-style-type: none"> • Garante(개인정보감독기관), 자국 내 ChatGPT 접속 차단 조치 및 조사 착수('23.3)
아일랜드	<ul style="list-style-type: none"> • DPC(개인정보감독기관), 생성형 AI 서비스 Bard(Google)의 EU 출시 금지('23.6)
중국	<ul style="list-style-type: none"> • CAC(사이버공간관리국), 생성형 AI 서비스 관리를 위한 임시 조치 발표('23.7) • 중국 내에서 ChatGPT 홈페이지 등 접속 차단('23.2)
일본	<ul style="list-style-type: none"> • 인공지능 확산에 대응하기 위해 관계부처들로 구성된 'AI 전략팀' 설치('23.4) • 인공지능 관련 국가 전략 수립을 위한 새로운 전략회의체를 창설하기로 발표('23.4)
호주	<ul style="list-style-type: none"> • 딥페이크 등 고위험군으로 간주되는 AI 기술을 금지하는 사안 검토('23.5)
뉴질랜드	<ul style="list-style-type: none"> • MBIE(기업혁신고용부), 데이터 보안 및 개인정보보호 위험성을 이유로 직원들의 AI 기술 사용을 금지하고 정부 지침을 발표할 예정이라고 발표('23.6)

II

G7, ‘히로시마 AI 프로세스’의 AI

■ '23년 5월, 일본 히로시마에서 개최된 G7 정상회의에서 각국 정상들은 생성형 AI에 대한 위험에 대해 논의하고 거버넌스를 연말까지 정리하여 발표하는 ‘히로시마 프로세스’ 출범

- G7 정상들은 지적재산권 문제, 개인정보 침해, 허위 정보 확산 등 AI 기술의 잠재적인 위험에 대해 담당 각료들이 신속하게 논의하고 해당연도 이내에 그 결과를 발표하는 ‘히로시마 프로세스’에 합의
- 각국 정상들은 생성형 AI 기술이 빠르게 확산되고 있고 허위정보나 정치적 혼란을 야기할 수 있는 강력한 도구가 될 수 있음에 동의하였으며, 일본 기시다 총리는 ‘인간 중심의 신뢰성 있는 AI’ 구축을 위해 ‘신뢰성 있는 자유로운 정보 유통’을 구체화할 필요가 있다고 강조

■ 이후 10월 30일, G7 정상은 히로시마 AI 프로세스 관련하여 공동 성명과 함께 첨단 AI 개발의 위험을 관리하기 위한 11개 개발자 행동 규범을 발표

- G7 정상들은 지난 히로시마 정상회의를 기반으로 AI 위험과 관련하여 국제규범과 국제 정보 유통의 틀을 마련하기 위해 다음의 내용 등을 포함하는 ‘히로시마 AI 프로세스’ 관련 성명 발표
- 첨단 AI 시스템이 가져올 혁신적인 기회와 변혁의 가능성을 강조하면서, 위험을 관리하여 법의 준수와 민주주의 가치 등을 포함한 공유된 원칙을 보호해야 하며 이를 위해 인공지능에 대한 포괄적인 거버넌스 형성이 필요함을 확인
- 첨단 AI 개발자와 조직들이 함께 발표하는 국제 행동 규범을 준수하고, 관련 정부 장관들이 히로시마 AI 프로세스를 더욱 발전시키기 위한 계획을 연말까지 마련할 것을 요청
- 히로시마 AI 프로세스를 통한 노력이 기술의 이점을 극대화하는 동시에 위험을 완화하여, 안전하고 신뢰할 수 있는 AI 시스템을 설계·개발·배포·사용될 수 있는 개방적이고 활성화된 환경을 조성할 것임을 기대

※ G7 Leaders' Statement on the Hiroshima AI Process, '23.10.30.

- 또한, 성명서와 함께 첨단 AI 시스템 개발 조직을 위한 11개의 조치를 포함하는 행동 규범 발표
- 본 행동 규범은 OECD의 AI 원칙을 기반으로 작성되었으며, 첨단 AI 시스템을 개발하는 조직이 안전하고 신뢰할 수 있는 AI 시스템을 개발할 수 있도록 장려하는 것을 목표로 하고 있음

※ Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI system, '23.10.30.

[표 3] 첨단 AI 시스템 개발 조직을 위한 11개 행동 규범

구분	주요 내용
1	<ul style="list-style-type: none"> • AI 생애주기에 걸친 위험을 식별, 평가, 완화하기 위해 시장 배포 전 적절한 조치가 취해져야 함 - 다양한 내/외부 테스트를 통해 식별된 위험 및 취약성 해결을 위한 적절한 조치 구현 - 테스트는 안전한 환경에서 AI 생애주기 각 지점별로 수행되어야 하며, 특히 보안, 안전, 사회문제에 대하여 식별된 AI 위험에 대한 테스트와 조치는 시장 배포 전 수행되어야 함 - 고급 AI 기술의 보안, 안전, 편향 및 허위 정보, 공정성, 설명 가능성 및 해석 가능성, 투명성, 오용에 대비한 고급 AI 기술의 견고성과 신뢰성 향상을 위한 연구 및 투자를 확대하기 위해 노력해야 함
2	<ul style="list-style-type: none"> • 기술 배포 후 취약점 및 사고와 오용의 경향성을 파악하고 완화해야 함 - 위험 수준에 따라 AI 기술을 의도한 대로 사용하고 있는지 배포 후에 취약점, 사고, 새로운 위험 및 오용을 모니터링하고 적절한 조치를 해야 함 - 보고된 사고에 대한 문서를 유지하고 다른 이해관계자와 협력하여 식별된 위험과 취약점을 완화하는 것을 권장함
3	<ul style="list-style-type: none"> • 고급 AI 기술의 기능, 한계 및 적절하지 못한 사용 영역을 공개하여 충분한 투명성을 확보할 수 있도록 지원하여 책임감을 높이기 위해 노력해야 함 - 고급 AI 기술의 새로운 중요 내용에 대한 의미 있는 정보를 포함한 투명성 보고서를 게시해야 함 - 보고서, 사용 지침 및 기술 문서는 최신 상태로 유지되어야 하며 위험 평가 내역, AI 모델 성능, 안전 및 사회에 미치는 영향과 위험에 대한 평가 등이 포함되어야 함 - 개발사는 이용자가 AI 기술의 성능을 적절히 사용할 수 있도록 보고서의 정보를 충분하고 명확하게 제공해야 함
4	<ul style="list-style-type: none"> • 산·관·민·학 등 첨단 AI 기술 개발 조직 간 책임 있는 정보 공유 및 사고 보고를 위한 작업을 진행해야 함 - 정보 공유 및 사고 보고에는 평가 보고서, 보안 및 안전 위험에 대한 정보, 의도하지 않은 기능, AI 수명주기 전반에 걸쳐 안전장치를 우회하려는 악의적인 시도 등을 포함하고 이 외의 정보도 책임감 있게 공유해야 함 - 고급 AI 기술의 안전, 보안 및 신뢰성을 보장하기 위해 표준, 도구, 메커니즘 및 모범 사례를 개발, 발전 및 채택하기 위한 메커니즘을 확립, 참여해야 함 - 투명성 있는 정보 공유가 필요하지만, 지식재산권은 보호되어야 함
5	<ul style="list-style-type: none"> • 개인정보 보호정책 및 완화 조치를 포함한 위험 기반 접근 방식을 통해 AI 거버넌스 및 위험 관리 정책을 개발, 구현 및 공개해야 함 - 위험 수준에 따라 AI 기술을 의도한 대로 사용하고 있는지 배포 후에 취약점, 사고, 새로운 위험 및 오용을 모니터링하고 적절한 조치를 해야 함 - 보고된 사고에 대한 문서를 유지하고 여러 이해관계자와 협력하여 식별된 위험과 취약점을 완화하는 것을 권장함
6	<ul style="list-style-type: none"> • AI 생애주기에 걸쳐 물리적 보안, 사이버 보안 및 내부자 위협 보호를 포함한 강력한 보안 제어 기술에 투자하고 구현해야 함 - 정보보안을 위한 운영 보안 조치 및 사이버/물리적 액세스 제어를 통한 모델 가중치 및 알고리즘, 서버 및 데이터셋 보안이 포함됨 - 사이버보안 위험에 대한 평가를 수행, 첨단 AI 기술에 대한 사이버보안이 상황과 위험에 적합하도록 적절한 정책과 기술 및 제도적 솔루션을 구현해야 함

7	<ul style="list-style-type: none"> • 이용자가 AI 생성 콘텐츠를 구별할 수 있도록 워터마킹 등 기술적으로 실현될 수 있는 콘텐츠 인증 및 검증 메커니즘을 개발, 사용해야 함 - 고급 AI 기술로 생성된 콘텐츠에 대한 인증 및 출처 메커니즘을 포함하며, 출처 데이터에는 콘텐츠를 생성한 서비스 또는 모델의 정보를 포함하여야 하고, 사용자 정보는 포함할 필요 없음 - 사용자가 AI 기술과 상호 작용하는 시기를 알 수 있도록 라벨링 또는 면책 조항과 같은 다른 메커니즘 구현이 권장됨
8	<ul style="list-style-type: none"> • 사회적 위험, 안전 및 보안 위험을 완화하기 위한 연구 우선순위를 정하고 효과적인 완화 조치를 위한 투자 우선순위를 정해야 함 - AI 안전, 보안, 신뢰 확보를 지원하고, 주요 위험 해결을 위한 연구, 협력, 투자는 물론 적절한 완화 도구 개발 투자 지원 - 조직은 주요 위험 해결을 위한 연구 수행, 협력 및 투자를 약속해야 하며, 연구 및 모범 사례를 공유해야 함
9	<ul style="list-style-type: none"> • 기후 위기, 글로벌 보건 및 교육 등 세계의 주요 문제를 해결하기 위해 첨단 AI 기술 개발 우선순위를 정해야 함 - UN의 지속가능 개발에 대한 진전을 지지하고 전 세계의 이익을 위한 AI 개발을 장려하기 위한 - 조직은 신뢰할 수 있고 인간 중심적인 AI에 대한 책임감 있는 관리를 우선해야 하며, 대중의 교육 및 훈련을 위한 디지털 활용 능력 이니셔티브를 지원하는 등 개인이 고급 AI 기술 사용을 통한 혜택을 받을 수 있도록 해야 함
10	<ul style="list-style-type: none"> • 국제 기술 표준 개발 및 채택을 촉진해야 함 - 조직이 테스트 방법론, 콘텐츠 인증 및 입증 메커니즘, 사이버보안 정책, 공공 보고 및 기타 조치를 개발할 때, 상호 운용이 가능한 국제 기술 표준 및 프레임워크를 개발하고 표준개발기구(SDO)와 협력하는 등 국제 기술 표준 및 모범 사례 활용에 기여해야 함
11	<ul style="list-style-type: none"> • 개인 데이터 및 지적 재산에 대한 적절한 데이터 입력 방식 및 보호 구현 - 조직은 기술이 기밀 또는 민감한 데이터를 유출하지 않도록 투명성 확보 및 개인정보보호 교육 등을 포함하여 데이터 품질 관리를 위한 적절한 조치를 취해야 함 - 조직은 저작권으로 보호되는 콘텐츠를 포함하여 개인정보보호 및 지적재산권 보장을 위한 적절한 보호 장치를 구현해야 하며, 적용될 수 있는 법적 프레임워크를 준수해야 함

III

AI 안전 정상회의(AI Safety Summit)의 블레츨리 선언

■ 11월 1일, 안전한 AI 활용을 위한 협력 방안을 모색하기 위해 28개 주요국과 인공지능 기술 관련 주요 기업들이 참석하는 AI 안전 정상회의 개최

- 이틀간 영국 블레츨리 파크에서 개최되는 본회의에서, 주요국 및 주요 기업들은 AI에 대한 위험과 이를 완화하기 위한 국제협력 방안에 대해 논의
 - AI의 위험에 대한 이해 : ①오남용으로 인한 글로벌 안전 위험, ②예측할 수 없는 기술 발전에 따른 위험, ③통제력 상실로 인한 위험, ④사회통합에 따른 위험
 - AI의 위험 최소화를 위해 할 일 : ①책임감 있는 역량 확장을 위하여 개발자가 해야 할 일, ②정책 입안자들이 AI 위험과 기회에 관하여 해야 할 일, ③국제 공동체가 해야 할 일, ④과학계가 해야 할 일
- AI를 개발하는 기업들과 함께 다수의 국가들은 정부와 개발자들 간 협력의 중요성을 인식하였으며, Safety Institute와의 파트너십을 포함하여 차세대 모델이 출시되기 전에 국가 주도의 테스트에 동의
- 참석자들은 안전과 관련하여 보다 진전된 정책들을 제기하였으며, 6개월 후 열릴 한국과 1년 후 열릴 프랑스의 안전 정상회의에서 이 문제들을 논의하기로 합의

■ 또한, 본 회의에서는 인공지능의 위험을 완화하고 안전을 보장하기 위해 인공지능 규제에 대한 최초의 국제적 합의로 볼 수 있는 ‘블레츨리 선언’이 발표

- 참여국들은 AI의 기회와 위험에 대해 공감하고 협력의 필요성을 인정하는 ‘블레츨리 선언’에 합의
 - AI 글로벌 기회와 동시에 발생하는 위험과 발전 및 안전을 보장하기 위한 국제적 협력의 중요성 및 다양한 분야에서 AI 활용 강조
- ※ 참가국 : 대한민국, 호주, 브라질, 캐나다, 칠레, 중국, 유럽연합, 프랑스, 독일, 인도, 인도네시아, 아일랜드, 이스라엘, 이탈리아, 일본, 케냐, 사우디아라비아, 네덜란드, 나이지리아, 필리핀, 르완다, 싱가포르, 스페인, 스위스, 터키, 우크라이나, 아랍에미리트, 영국, 미국

※ The Bletchley Declaration by Countries Attending the AI Safety Summit, '23.11.1.

[표 4] '블레츨리 선언' 주요 내용

구분	주요 내용
잠재력	• AI는 인간의 복지, 평화, 및 번영을 향상시킬 수 있는 기회를 제공함
안전성	• AI는 안전하게 설계, 개발, 배포, 및 사용되어야 하며 인간 중심적이고 책임감 있는 방식으로 운영되어야 함
국제협력	• AI의 위험과 잠재력은 국제적으로 다루어져야 하며, 국제협력을 통해 안전하고 모두의 이익을 취할 수 있는 AI를 보장하기 위한 노력을 환영함
다양한 영향	• AI는 주택, 고용, 교통, 교육, 건강, 접근성, 정의 등 여러 분야에 영향을 미치고 미래에 그 영향이 더욱 커질 것으로 예상됨
안전 위험	• AI는 일상생활에서 다양한 위험을 초래할 수 있으며, 이러한 위험을 해결해야 할 필요성을 강조함
책임과 투명성	• AI의 안전을 보장하기 위해 모든 관련 행위자가 역할을 가져야 하며, 적절한 평가 및 안전 테스트를 통해 안전성을 제고해야 함
국제 과학 연구 네트워크	• AI 안전에 대한 국제적으로 포용적인 과학 연구 네트워크를 지원하고 국제 대화를 유지하기로 결의함

- 블레츨리 선언으로 인해 AI 안전에 관한 글로벌 표준과 관행을 국제 사회의 협력을 통해 개선해 나갈 것이며, 앞으로의 AI 안전 연구를 위한 이정표 역할을 수행
- 하지만, 구속력이 없는 선언문이며 선언에 대한 후속 조치는 다음 정상회의에서 구체화될 예정

IV

美 행정명령, Safe, Secure, and Trustworthy AI

4-1 행정명령 8개 분야 정책 및 원칙

■ 미 바이든 정부는 미국이 인공지능에 대한 약속을 지키고 위험을 관리하는 데 앞장서기 위한 대통령 행정명령을 발표('23.10.30.)

- (배경) 본 행정명령은 책임감 있는 혁신을 위한 정부의 포괄적인 전략의 일환이며, AI의 안전하고 신뢰할 수 있는 개발 추진을 위한 이전 조치에 기반함
 - 미 정부는 인공지능의 개발과 활용을 안전하고 책임감 있게 활용하는 것을 최우선 과제로 삼고 있으며, 이를 위해 연방 정부 차원에서의 통합된 접근방식을 추진
- (구성) 행정명령은 8개 분야의 '정책 및 원칙(Sec.2)'을 제시하고 있으며, 이를 기반으로 미 행정부가 인공지능의 개발과 활용을 촉진하고 관리하기 위해 수행해야 하는 세부 조치사항(Sec.4~11)들을 제시
- (실행) 행정부 내 '백악관 인공지능 위원회*'를 통해 본 행정명령을 포함하여 인공지능 관련 정책의 효과적인 수립, 개발, 의사소통, 산업 참여 등을 보장하기 위해 기관들의 활동을 조정·관리
 - 대통령 보좌관과 정책 담당 부비서실장이 의장을 맡으며, 장관 및 관련 기관장들로 구성

* White House Artificial Intelligence Council

[표 5] 행정명령 구성

구분	제목	구분	제목
Sec.1	목적	Sec.7	평등과 시민권 촉진
Sec.2	정책 및 원칙	Sec.8	소비자, 환자, 승객, 학생 보호
Sec.3	정의	Sec.9	개인정보 보호
Sec.4	인공지능의 안전과 보안 확보	Sec.10	연방정부의 인공지능 활용 촉진
Sec.5	혁신과 경쟁 촉진	Sec.11	미국의 글로벌 리더십 강화
Sec.6	노동자 지원	Sec.12	실행

■ 본 행정명령의 세부 조치사항들은 본 행정명령의 ‘정책 및 원칙(Sec.2)’에 기반하며 관련 행정부 및 기관들은 본 원칙을 준수하며 조치사항을 수행해야 함

- 행정부의 방침은 아래 지침과 우선순위에 따라 인공지능의 개발과 활용을 촉진하고 통제하는 것이며, 관련 법에 따라 적절하게 수행하며, 다른 기관, 산업계, 학계, 국제 파트너 등의 의견 고려 필요

[표 6] 행정명령 8개 분야 정책 및 원칙

〈1〉 인공지능의 안전과 보안 확보

- 인공지능은 안전하고 안정적이어야 하며 이를 위해서는 인공지능 시스템의 견고하고 신뢰할 수 있으며 반복 가능하며 **표준화된 평가가 필요**
 - 사전에 인공지능 시스템에서 발생할 수 있는 **위험을 테스트**하고 **이해 및 완화**를 위한 **정책이나 메커니즘**이 갖추어져야 함
- 인공지능 시스템의 가장 긴박한 보안 위험을 해결하기 위해서는 **생물 기술, 사이버 보안, 핵심 인프라** 및 기타 국가 안보 위험과 관련된 **SI의 불투명성과 복잡성을 해결**해야 함
- 테스트·평가를 통해 배치 후 성능 모니터링을 포함하여 인공지능 시스템이 의도한 대로 작동하고, 오용이나 위험한 수정에 대비하여 탄력적이며, 윤리적으로 개발되어 운영되고, 관련 연방 법률과 정책을 준수하는 것을 보장해야 함
- 행정부는 인공지능으로 생성된 콘텐츠와 그렇지 않은 콘텐츠를 구분할 수 있도록 **효과적인 라벨링 및 콘텐츠 출처 추적 메커니즘**을 개발하는 데 도움을 제공할 것임

〈2〉 혁신과 경쟁 촉진

- 책임 있는 혁신, 경쟁, 그리고 협력을 촉진함으로써 미국은 인공지능 분야에서 선도하고, 인공지능이 **사회의 가장 어려운 문제를 해결**하는 잠재력을 발휘할 수 있게 해야 함
- 인공지능 관련 교육, 훈련, 개발, 연구, 그리고 능력 강화에 대한 투자가 필요하며, 동시에 발명가와 창조자를 보호하기 위한 새로운 **지적재산(IP) 문제**와 기타 문제들에 대응 필요
- 미국인들이 인공지능 시대에 필요한 기술을 습득하고, 세계적인 **인공지능 관련 인재**를 미국으로 **유치하는 프로그램**을 지원
- 연방 정부는 공정하고 개방적이며 경쟁력 있는 인공지능 및 관련 기술의 생태계와 시장을 촉진하여 **소규모 개발자와 기업가**들이 혁신을 이끌어 나갈 수 있도록 지원

〈3〉 노동자 지원

- 인공지능이 **새로운 직업과 산업을 창출**함에 따라, 모든 노동자들이 이러한 기회에서 혜택을 받을 수 있도록, 단체 협상을 통해 포함되어야 함
- 다양한 **노동 인력**을 지원하고, 인공지능이 만들어 내는 기회에 대한 접근을 돕기 위해 **직업 교육 및 훈련**을 조정
- 직장 내에서 인공지능이 권리를 해칠 수 없는 방식으로 사용되어야 하며, **노동환경의 저해, 노동자 감시 유도, 시장 경쟁 악화, 새로운 건강 및 안전 위험, 노동력 파괴** 등을 방지해야 함
- 인공지능 개발의 중요한 다음 단계는 노동자, 노동조합, 교육자 및 고용주들의 의견을 기반으로 해야 함
- 노동자들의 삶을 개선하고 인간의 일을 긍정적으로 보완하며, 모든 사람이 기술 혁신으로부터 안전하게 이익과 기회를 누릴 수 있도록 책임 있는 인공지능 사용을 지원해야 함

〈4〉 평등과 시민권 촉진

- 이미 평등한 기회와 정의를 거부당한 사람들을 더욱 **불리하게 하는 인공지능 사용을 용인하지 않음**
- **채용부터 주택, 의료까지**, 책임 없이 도입된 인공지능 시스템은 기존의 불평등을 재생산하고 강화시키며 새로운 유형의 해로운 차별을 일으키고 온라인과 실제 사회의 해를 악화
- 이미 취해진 중요한 조치들(예: AI 권리 선언서, AI 위험관리 프레임워크, 행정명령 14091호)을 더욱 발전시켜 인공지능이 모든 **연방 법률을 준수**하고 강력한 기술적 평가, 신중한 감시, 영향을 받는 공동체와의 소통, 엄격한 규제를 촉진할 것임
- 인공지능 활용에 따른 공정성과 인권 증진에 신뢰성 확보를 위해, 인공지능 개발 및 도입자들은 **불법 차별과 남용을 방지**하는 기준에 따라야 하며 **법적 책임을 부과**해야 함

〈5〉 소비자, 환자, 승객, 학생 보호

- 미국인들이 일상에서 인공지능 기술 및 제품을 사용하거나 상호작용할 때 그들의 이익을 보호해야 함
- 연방 정부는 기존의 **소비자 보호법과 원칙을 집행**하고, 인공지능으로 인한 사기, 의도하지 않은 편견, 차별, 개인정보 침해 및 기타 피해로부터 **적절한 보호 장치를 마련**할 것임
 - 특히 **의료, 금융 서비스, 교육, 주택, 법률 및 교통**과 같은 핵심 분야에서 중요
 - 또한, 인공지능의 실수나 오용이 환자에게 피해를 주거나 소비자나 소기업에 비용을 발생시키거나 안전이나 권리를 위협할 수 있는 경우도 중요
- **소비자를 보호**하고 제품 및 서비스의 **품질을 높이며 가격을 낮추거나 선택과 이용 가능성을 확대**하는 인공지능 사용을 촉진

〈6〉 개인정보 보호

- 인공지능이 계속 발전함에 따라 **미국인들의 개인정보와 시민 자유를 보호**해야 함
 - 인공지능은 사람들의 민감정보를 추출, 재식별, 연결, 추론하고 이를 기반으로 행동하는 것을 더 쉽게 만들고 있으며, 이러한 영역에서의 인공지능으로 인해 개인 데이터가 악용되고 노출될 위험을 증가시킬 수 있음
- 이러한 위험을 줄이기 위해 연방 정부는 **데이터 수집, 사용 및 보유가 합법적이고 안전하며 개인정보와 기밀 유출 위험을 완화**한다는 것을 보장할 것임
 - 기관들은 개인정보 보호 기술(PETs)을 포함한 정책 및 도구를 활용하여, 개인정보를 보호하고 부적절한 수집·활용으로 인한 권리의 억압 등 보다 광범위한 법적 및 사회적 위험에 대항할 것임

〈7〉 연방정부의 인공지능 활용 촉진

- **연방 정부의 인공지능 사용으로 인한 위험을 관리**하고, 미국인들에게 더 나은 결과를 제공하기 위해 책임 있는 인공지능 사용을 규제, 조정, 및 지원하기 위한 내부 능력을 강화하는 것이 중요
- 기술, 정책, 관리, 조달, 규제, 윤리, 지배, 법률 분야를 포함한 **다양한 분야의 인공지능 전문가를 확보 및 유지**하는 조치를 취할 것이며, 소외된 지역사회에서도 인공지능 전문가를 확보하기 위해 노력할 것임
- 연방 정부는 모든 근로자들이 자신의 직무에 대한 인공지능의 혜택, 위험, 제한을 이해할 수 있도록 적절한 교육 제공, 정보 기술 인프라의 현대화, 관료적 장애 제거, **안전한 인공지능 기술이 채택·배치·사용될 수 있도록 노력**할 것임

〈8〉 미국의 글로벌 리더십 강화

- 미국이 이전 혁신과 변화의 시대에 이뤄낸 것처럼, 세계적 사회, 경제 및 기술 발전을 주도해야 함
- **국제적 동맹과 협력국들과 협력**하여 인공지능의 위험을 관리하고, 인공지능의 선진성을 발휘하여 공동의 도전에 대한 **공동된 접근방식을 촉진**하는 **프레임워크를 개발**할 것임
- 연방 정부는 경쟁국을 포함한 다른 나라들과 책임 있는 인공지능 안전과 안보 원칙 및 조치를 촉진하면서, 핵심적인 **세계적 대화와 협력을 주도**하여 인공지능이 부정적인 영향보다 세계 전체에 이익을 가져올 수 있도록 노력할 것임

4-2 인공지능의 안전과 보안 확보(Sec.4) 세부 내용

■ Sec.4에서 인공지능의 안전과 보안 확보를 위해 8개 영역에 대한 세부적인 조치사항 제시

- 세부 조치사항 별로 각각 수행 주체와 명령일로부터의 기한을 제시하고 있으나 본 문서에서는 생략
- ※ 행정명령 세부 조치사항에 대한 내용은 [별첨 1] 참고(23p)

■ 4.1 인공지능 안전 및 보안을 위한 지침, 표준 및 모범사례 개발

- Developing Guidelines, Standards, and Best Practices for AI Safety and Security

- 안전하고, 보안이 확보되며, 신뢰할 수 있는 인공지능 시스템 개발을 위해 다음과 같은 조치 수행
- 인공지능 시스템을 개발하고 배치하기 위한 업계 표준에 대한 합의를 촉진하는 지침과 모범 사례 마련
 - ※ 생성형 AI를 위한 위험 관리 프레임워크(NIST 위험 관리 프레임워크에 부합하는 자료) 개발, 안전한 소프트웨어 개발 프레임워크에 개발 관행 통합, AI 평가·감사에 대한 지침 및 벤치마크 생성을 위한 이니셔티브 착수
- 안전한 인공지능 시스템의 배치를 위해 인공지능 레드팀 테스트* 수행을 위한 적절한 절차 및 프로세스를 포함하는 지침 수립
 - ※ 이중용도 기반 모델에 대한 평가·관리 지침 조정 및 개발, 테스트 환경의 가용성 확보를 위한 개발 및 타 부처 협력 등
 - * 인공지능 개발자들과 협력하여 인공지능 시스템의 결함과 취약점을 찾기 위한 테스트 활동
- 인공지능 보안 위험을 이해하고 완화하기 위해, 법률 및 이용 가능한 예산에 따라 인공지능 모델 평가 도구 및 인공지능 테스트베드를 개발·실행하기 위한 계획 수립 및 시행
 - ※ 인공지능 능력에 대한 단기 예측 평가 및 핵, 화학산, 생물학적, 화학적, 주요 기반시설 및 에너지 보안 위협에 대한 결과를 생성할 수 있는 인공지능 능력에 대한 평가 도구

■ 4.2 안전하고 신뢰할 수 있는 인공지능 확보

- Ensuring Safe and Reliable AI

- 국방 및 주요 기반시설을 포함하여, 인공지능의 지속적인 가용성을 보장하고 검증하기 위해 다음을 수행
- 이중용도 기반 모델을 개발하거나 개발하고자 하는 기업들은 연방 정부에 정보제공 혹은 보고서 제출을 요구
 - ※ 이중용도 기반 모델의 개발에 대한 계획과 진행 상황 및 개발 과정에서의 보호 조치, 이중용도 기반 모델의 가중치 보유 상황 및 보호 조치, 개발된 이중용도 기반 모델의 성능 결과와 보안을 위해 취한 조치(레드팀 테스트 결과 포함)
- 대규모 컴퓨팅 클러스터를 획득·개발·보유하는 개인이나 단체 등은 클러스터의 존재와 위치, 사용 가능한 컴퓨팅 파워 등을 보고하도록 요구
- 위 조치사항에서 정보제공 혹은 보고서 제출을 요구해야 하는 모델과 컴퓨팅 클러스터의 조건을 정의하고 필요에 따라 정기적으로 업데이트 수행
- 미국 인프라서비스(IaaS) 제품을 악의적 사이버 행위자들이 이용하는 것에 대응하기 위해, 외국 거래와 관련한 추가 기록 유지 의무를 부과하고 외국 악의적 사이버 행위자들의 거래 조사를 지원
 - ※ 악의적 행위에 대한 잠재적 위험이 있을 수 있는 거래에 대해 IaaS 제공업체는 보고서를 제출하며, 보고서를 제출하지 않는 경우 서비스를 제공하지 못하도록 하는 요건 등 포함

- 미국 IaaS 제공업체가 외국 리셀러에게 제품을 판매할 경우, 해당 외국인이 IaaS 계정을 생성하거나 기존 계정을 유지할 때 신원을 확인하도록 요구하는 규정을 제언
- 상무부 장관은 국제긴급경제권한법에 의해 IaaS 관련 조치를 위해 대통령에게 부여된 모든 권한을 행사할 수 있으며, 여기에는 규칙과 규정 공포 등이 포함

■ 4.3 주요 기반시설 및 사이버보안에서의 인공지능 관리

- Managing AI in Critical Infrastructure and in Cybersecurity

- **주요 기반시설 보호**를 위해 다음 조치를 수행
 - 주요 기반시설에서 인공지능 활용과 관련된 잠재적 위험을 평가하고 국토안보부 장관에게 보고(매년 1회 이상)
 - ※ 인공지능 배치에 따라 주요 기반시설 시스템의 장애 및 물리적·사이버 공격으로 인한 위험 요소와 보완 방법 포함
 - 재무부장관은 금융기관이 AI로 인한 사이버 보안 위험을 관리하기 위한 모범 사례 공개 보고서를 발표
 - 인공지능 위험 관리 프레임워크(NIST AI 100-1) 및 관련 보안 지침을 주요 기반시설 소유자 및 운영자가 사용할 수 있는 관련 안전·보안 지침에 통합
 - 위 안전·보안 지침 조정 완료 후에, 연방 정부가 규제나 조치를 통해 적절한 부분을 의무화시킬 것이며 주요 기반시설의 권한을 가진 기관장은 해당 작업을 조정
 - 국토안보부 장관은 국토안보법에 의거하여 '인공지능 안전·보안 위원회' 설립
 - ※ 민간, 학계, 정부의 인공지능 전문가를 포함하며, 국토안보부 장관과 주요 기반시설 관계자들에게 주요 기반시설에서의 인공지능 활용과 관련된 조언과 정보, 보안 권고사항, 사이버복원력, 사고 대응에 대한 내용을 제공
- 인공지능의 잠재력 활용을 통해 미국의 사이버 방어를 개선하기 위해 다음 조치를 수행
 - 본 절의 사이버 방어 개선을 위한 조치사항 수행과 관련하여, 국방부 장관은 국가안보시스템에 대해, 그리고 국토안보부 장관은 非국가안보시스템에 대해 각각 조치 수행
 - 미 정부 소프트웨어, 시스템, 네트워크의 취약점을 발견하고 보완하기 위해 필요한 인공지능 기능을 식별·개발·테스트·평가·배치하는 운영 시범 프로젝트에 대한 계획을 수립하고 수행 및 완료해야 함
 - 위 프로젝트에 따라 조치한 결과 보고서를 국가안보를 담당하는 대통령 보좌관에게 제출
 - ※ 인공지능을 통해 발견되고 수정된 취약점과 사이버 방어를 위해 인공지능을 효과적으로 활용하는 방법 포함

■ 4.4 인공지능과 CBRN* 위협 간의 위험 감소

- Reducing Risks at the Intersection of AI and CBRN Threats

* 화생방 및 핵무기

- 인공지능이 CBRN 위협(특히 생물학적 무기에 집중)에 오용될 위험을 이해하고 완화하기 위해 다음의 조치 수행
 - 인공지능이 CBRN 위협 개발 및 생산에 오용될 가능성을 평가하며, 이러한 위협에 대응하기 위한 인공지능의 이점과 활용을 고려해야 함
 - ※ 오직 위협 방어의 목적으로 CBRN 위협을 나타내는 인공지능 모델의 능력을 평가하고 이에 대해 대통령 보고 수행

- 국립과학공학의학한림원(NASEM)과 계약을 체결하여 인공지능으로 인한 생물안보 위험 평가 및 완화 조치, 관련 데이터셋 위험 완화 등에 대한 연구 수행
- **합성핵산의 오용 위험을 감소**시키고 관련 산업의 생물안보 조치를 개선하기 위한 조치 수행
 - ※ 합성핵산 조달 관련 프레임워크 수립, 관련 산업 및 이해관계자 협력 노력, 펀딩 요건 검토 등

■ 4.5. 인공지능이 생성한 합성 콘텐츠*에 의해 초래되는 위험 감소

- Reducing the Risks Posed by Synthetic Content

- * 인공지능과 같은 알고리즘을 통해 생성되거나 크게 수정된 이미지, 동영상, 텍스트 등의 정보
- 인공지능에 의해 생성된 **합성 콘텐츠를 식별하고 라벨링하는 능력을 촉진**하고, 연방 정부가 생산한 디지털 콘텐츠의 **진정성과 출처 확립**을 위해 다음의 조치 수행
 - 관련 기술에 대한 **기존의 표준, 도구, 방법 및 관행과 추가적인 표준 및 기술 발전 가능성에 대한 보고서 제출**
 - ※ 콘텐츠 인증 및 출처 추적, 워터마킹같은 합성 콘텐츠 라벨링, 합성 콘텐츠 탐지, 아동 성적학대 자료 생성 방지 기술 등
 - 보고서 제출 후, **디지털 콘텐츠 인증 및 합성 콘텐츠 탐지**를 위한 기존 도구 및 관행에 대한 **지침을 개발**하고 정기적으로 업데이트 수행
 - 지침 개발 후, **미 정부의 공식 디지털 콘텐츠에 대한 대중의 신뢰를 강화**하기 위해 기관들이 생산하거나 발행하는 콘텐츠의 **라벨링과 인증을 위한 지침 개발**
 - 연방 취득 규정위원회는 본 절의 지침들을 고려하여 연방 취득 규정의 개정을 검토

■ 4.6. 널리 사용되는 매개변수를 갖는 이중용도 기초 모델*에 대한 의견 수렴

- Soliciting Input on Dual-Use Foundation Models with Widely Available Model Weights

- * 최소 수백억 개의 매개변수를 갖고 있고 광범위한 데이터로 훈련된 인공지능 모델로써, 광범위한 목적으로 활용할 수 있고, 보안, 국가 경제 안보, 공공 보건 및 안전 등에 심각한 위협을 초래할 수 있는 작업에서 높은 성능을 발휘하는 모델
- 널리 사용되는 매개변수를 갖는 **이중 용도 기초 모델의 위험과 잠재적 이점**을 다루기 위해 다음의 조치 수행
 - 민간 부문, 학계, 시민 사회 및 기타 이해관계자로부터 **공개적인 과정**을 통해 **잠재적 위험, 편익, 기타 영향력, 적절한 정책 및 규제 접근 방식에 대한 의견 수렴**
 - 의견 수렴에 기반하여 널리 사용되는 매개변수를 갖는 **이중용도 기초 모델의 잠재적 이점, 위험 및 함의**에 관한 **보고서와 모델과 관련된 정책 및 규제 권고사항**을 제출

■ 4.7. 연방 데이터의 안전한 공개 촉진 및 AI 교육용 악의적 사용 방지

- Promoting Safe Release and Preventing the Malicious Use of Federal Data for AI Training

- 공공 데이터 접근성을 개선하고 보안 위험을 관리하기 위해 다음의 조치 수행
 - 연방 데이터 공개로 인한 CBRN 및 자율적 사이버공격 개발에 대한 **잠재적 보안 위험을 식별**하고 관리하는 **지침을 개발**하고 연방 정부 데이터에 대한 공공 접근을 제공
 - 지침을 개발 후, 해당하는 모든 데이터 자산에 대한 보안 검토를 수행

4.8. 국가안보각서 개발 지시

- Directing the Development of a National Security Memorandum

- 인공지능 보안 위험 관리를 위한 **통합된 행정부 접근방식 개발**을 위해, 대통령 국가안보보좌관 및 정책 부비서실장은 대통령에게 제출할 인공지능에 관한 국가안보각서 개발을 위해 부처 간 **협업 과정**을 감독
- 각서는 **국가안보시스템의 구성요소** 및 **군사·정보용으로 활용되는 인공지능의 거버넌스**를 다루며, 이로 인한 국가안보 위험과 잠재적 이익을 다루기 위한 조치를 설명해야 함
- ※ 미국의 국가 안보를 위해 인공지능을 채택하는 것에 대해 국방부 및 관련 기관에 지침을 제공하고, 미국 및 동맹의 적대국들이 인공지능을 이용한 잠재적 위협에 대응하기 위한 조치 지시

4-3 개인정보 보호(Sec.9) 세부 내용

■ Sec.9에서 개인정보 보호를 위한 세부적인 조치사항 제시

- 세부 조치사항 별로 각각 수행 주체와 명령일로부터의 기한을 제시하고 있으나 본 문서에서는 생략
- ※ 행정명령 세부 조치사항에 대한 전체 내용은 **[별첨 1]** 참고(23p)

■ Sec. 9. 개인정보 보호 - Protecting Privacy

- 인공지능에 의해 악화될 수 있는 **개인정보 보호 위험**을 **완화**하기 위해 예산관리국 국장은 다음의 조치 수행
 - 기관에서 조달한 상업적으로 이용가능한 정보*(CAI)를 평가하고 식별하기 위한 조치 수행
 - * 입수·판매·임대 등이 가능한 개인 또는 집단에 대한 모든 정보 및 데이터
 - CAI와 관련된 **기관의 활동**으로부터 **개인정보 보호와 기밀성 위험 완화**를 위한 기관들의 지침을 위해, CAI의 수집, 처리, 유지, 사용, 공유, 전파, 처분과 관련된 **기관의 기준 및 절차를 평가**
 - 전자정부법('22)의 개인정보 보호 조항 시행을 위해, 기관의 지침 개정을 위한 정보요청(RFI) 발행
 - ※ 인공지능으로 인한 위험 완화를 위해 개인정보 영향 평가가 어떻게 더 효과적일 수 있는지에 대한 피드백 제공
 - RFI를 통해 식별된 단기적 조치 및 장기 전략을 지원하고 필요한 조치 수행
- 인공지능으로 인한 잠재적 위험으로부터 **미국인의 개인정보보호**를 위해 **개인정보보호 강화 기술(PET)**을 보다 효율적으로 활용할 수 있도록, **차등프라이버시보장* 효과**를 평가하는 지침을 개발
 - * 특정 주제에 대한 개인정보의 부적절한 접근·이용·공개를 제한하면서 그룹에 대한 정보를 공유할 수 있는 보호
- PET와 관련된 연구, 개발, 실행의 **촉진**을 위해 다음을 수행
 - 개인정보보호 연구를 촉진하고 기술의 개발, 배치, 확대를 위한 자금 지원
 - 기관의 운영에 PET를 통합할 수 있는 진행 중인 작업 및 잠재적 기회를 식별
 - '미-영 PET 상급 챌린지'의 결과를 활용하여 PET 연구 및 채택에 대한 접근방식과 기회를 공지

V

시사점

■ 인공지능에 대한 위험 관리를 위해 범정부 차원에서의 규제 마련 시작

- 인공지능 기술이 확산됨에 따라 위험 관리를 위해 주요국들은 **규제를 마련하는 방향으로 국제적 흐름이 변화**하고 있으며, 초거대 인공지능 모델 등을 보유한 핵심 국가들이 주도할 것으로 전망
- 인공지능 기술의 혁신성과 파급력으로 인해, 단순 과학기술에 대한 대응이 아니라 다양한 기관들과 이해관계자들이 협력하여 **범정부 차원의 대응 전략을 마련**
- 다만, 혁신의 가치를 저해하지 않기 위해 구속력이 강한 규제보다는 **인공지능 기술을 적극적으로 도입**하고 그에 따른 부작용이나 위험성을 평가하고 완화하기 위한 **안전장치를 마련**하는 추세

■ 인공지능 기술의 개발·도입·활용에 대한 국가 대응 전략의 구체화 및 확대

- 인공지능 기술의 **개발·도입·활용 등 모든 과정**에 대해 안전성을 평가하기 위해 **표준, 지침 등을 마련**하고 인공지능 레드팀을 운영하여 보안을 테스트하는 **구체적인 프로세스를 마련**하기 시작
- 인공지능이 **악의적으로 활용**될 수 있는 **위험에 대비**하기 위해, 특히 초거대 인공지능 모델과 같이 범용적으로 놀라운 성능을 보일 수 있는 모델에 대한 **정부의 통제를 강화**하는 추세
 - ※ (미 행정명령) 이중용도 기반 모델 및 대규모 컴퓨팅 클러스터에 대한 정보 제공 요구, 공개 의견 수렴 과정 등
- **주요 기반시설, 사이버보안에 인공지능 기술 도입**을 위해 구체적인 위험성 평가를 수행하고, 국가안보각서 개발과 같이 **범국가적 관리 프로세스를 마련**하는 등 **국가 안보차원에서의 인공지능 대응 강화**
- 인공지능의 **활성화**를 위해 **민간 사회에서 인공지능 기술을 신뢰**할 수 있도록, 인공지능으로 인한 **개인정보보호 및 결과물 검·인증**에 대한 **연구 및 프로세스가 확대**될 전망
 - ※ (미 행정명령) 인공지능으로 생성된 데이터에 대한 출처 인증이나 라벨링 기술, 범용 모델에 대한 공개 의견 수렴 등

별첨

美 행정명령, Sec.4 안전과 보안 및 Sec.9 개인정보 보호 (번역)

(4.1) 인공지능 안전 및 보안을 위한 지침, 표준 및 모범사례 개발

- (a) 본 명령일로부터 270일 이내에, 상무부 장관은 국립표준기술원(NIST)의 소장을 통해 안전하고, 보안이 확보되며, 신뢰할 수 있는 인공지능 시스템의 개발을 돕기 위해 다음과 같은 조치 수행
- (i) 안전하고, 보안이 확보되며, 신뢰할 수 있는 인공지능 시스템을 개발하고 배치하기 위한 업계 표준에 대한 합의를 촉진하는 지침과 모범 사례를 마련해야 합니다. 이에에는 다음이 포함됩니다:
- (A) 생성 인공지능을 위한 인공지능 위험 관리 프레임워크, NIST AI 100-1에 부합하는 자료 개발;
 - (B) 생성 인공지능과 이중용도 기반 모델을 위한 안전한 소프트웨어 개발 프레임워크에 안전한 개발 관행을 통합하는 자료 개발; 및
 - (C) 인공지능이 사이버보안과 생물보안과 같은 영역에서 해를 끼칠 수 있는 능력에 중점을 두고, 인공지능 능력을 평가하고 감사하기 위한 지침 및 벤치마크 생성을 위한 이니셔티브 개시.
- (ii) 국가 보안 시스템의 구성 요소로 사용되는 인공지능을 제외하고, 특히 이중용도 기반 모델의 개발자들이 안전하고, 보안이 확보되며, 신뢰할 수 있는 시스템의 배치를 가능하게 하는 인공지능 레드팀 테스트를 수행할 수 있도록 적절한 절차 및 프로세스를 포함하는 지침을 설정해야 합니다. 이러한 노력은 다음을 포함해야 합니다:
- (A) 이중용도 기반 모델의 안전성, 보안성 및 신뢰성을 평가하고 관리하는 관련 지침 조정 또는 개발; 및
 - (B) 에너지부 장관 및 국립과학재단(NSF) 소장과 협조하여, 안전하고, 보안이 확보되며, 신뢰할 수 있는 인공지능 기술의 개발을 지원하고, 본 명령의 9(b) 조항에 부합하는 관련 프라이버시 증진 기술(PET)의 설계, 개발 및 배치를 지원하기 위한 테스트 환경, 예를 들어 테스트베드의 가용성을 보장하고 개발하는 데 도움을 줄 것입니다.
- (b) 본 명령일로부터 270일 이내에, 인공지능 보안 위험을 이해하고 완화하기 위해, 에너지부 장관은 에너지부 장관이 적절하다고 판단하는 다른 분야 위험 관리 기관(SRMAs)의 수장들과 협력하여, 법률 및 이용 가능한 예산에 따라 에너지부의 인공지능 모델 평가 도구 및 인공지능 테스트베드를 개발하고 실행하기 위한 계획을 수립하고 시행해야 합니다. 장관은 가능한 한 기존 솔루션을 사용하여 이 작업을 수행해야 하며, 인공지능 시스템의 능력에 대한 단기 예측을 평가할 수 있는 이러한 도구 및 인공지능 테스트베드를 개발해야 합니다. 최소한 장관은 핵, 비확산, 생물학적, 화학적, 주요 인프라 및 에너지 보안 위험 또는 위험을 나타낼 수 있는 출력을 생성하는 인공지능 능력을 평가하기 위한 도구를 개발해야 합니다. 장관은 이러한 위험에 대비하기 위한 목적으로만 이 작업을 수행해야 하며, 이러한 위험을 줄이는 모델 안전장치도 개발해야 합니다. 장관은 적절하게 민간 인공지능 연구소, 학계, 시민 사회 및 제3자 평가자들과 협의해야 하며, 기존 솔루션을 사용해야 합니다.

(4.2) 안전하고 신뢰할 수 있는 인공지능 확보

- (a) 본 명령일로부터 90일 이내에, 수정된 방위 생산법, 50 U.S.C. 4501 이하에 따라, 국방과 중요 인프라 보호를 포함하여 안전하고 신뢰할 수 있으며 효과적인 인공지능의 지속적인 가용성을 보장하고 검증하기 위하여, 상무부 장관은 다음을 요구해야 합니다:
- (i) 이중용도 기반 모델을 개발하거나 개발할 의사를 보이는 기업들에게 지속적으로 연방 정부에 다음과 같은 정보, 보고서 또는 기록을 제공하도록 합니다:
- (A) 이중용도 기반 모델과 관련된 훈련, 개발 또는 생산 활동에 대한 계획 또는 진행 상황, 그리고 그 훈련 과정의 무결성을 정교한 위협으로부터 보장하기 위해 취해진 물리적 및 사이버보안 보호 조치;
 - (B) 이중용도 기반 모델의 모델 가중치 소유 및 보유 상황, 그리고 이러한 모델 가중치를 보호하기 위해 취해진 물리적 및 사이버보안 조치; 및
 - (C) 개발된 이중용도 기반 모델의 성능 결과와 NIST에 의해 본 섹션의 4.1(a)(ii) 항에 따라 개발된 지침에 기반한 관련 인공지능 레드팀 테스트에서, 그리고 기업이 안전 목표를 충족하기 위해 취한 관련 조치에 대한 설명, 예를 들어 이러한 레드팀 테스트의 성능을 향상시키기 위한 완화 조치 및 전반적인 모델 보안 강화를 포함합니다. NIST에 의해 본 섹션의 4.1(a)(ii) 항에 따른 레드팀 테스트 표준에 관한 지침 개발 전에, 이 설명에는 회사가 수행한 레드팀 테스트 결과가 포함되어야 하며, 비국가 행위자에 의한 생물학적 무기 개발, 획득 및 사용에 대한 진입 장벽을 낮추는 것, 소프트웨어 취약성 발견 및 관련 익스플로잇 개발, 실제 또는 가상 사건에 영향을 미치는

소프트웨어 또는 도구 사용, 자가 복제 또는 전파 가능성과 관련된 안전 목표를 충족하기 위한 조치를 포함해야 합니다; 및

- (ii) 대규모 컴퓨팅 클러스터를 획득, 개발 또는 보유하는 기업, 개인 또는 기타 조직이나 실체들에게, 이러한 획득, 개발 또는 보유에 대해 보고하도록 하며, 이 클러스터의 존재와 위치, 각 클러스터에서 사용 가능한 총 컴퓨팅 파워량을 포함해야 합니다.
- (b) 상무부 장관은 국무부 장관, 국방부 장관, 에너지부 장관, 그리고 국가정보국장장과 협의하여, 본 절의 4.2(a) 항의 보고 요구 사항에 따라 보고 대상이 될 모델과 컴퓨팅 클러스터의 기술 조건을 정의하고 이후 필요에 따라 정기적으로 업데이트해야 합니다. 이러한 기술 조건이 정의될 때까지 장관은 다음에 대해 이 보고 요구 사항의 준수를 요구해야 합니다:
 - (i) 1026 정수 또는 부동 소수점 연산보다 큰 양의 컴퓨팅 파워를 사용하여 훈련된 모든 모델, 또는 주로 생물학적 시퀀스 데이터를 사용하고 1023 정수 또는 부동 소수점 연산보다 큰 양의 컴퓨팅 파워를 사용하여 훈련된 모든 모델; 및
 - (ii) 단일 데이터센터에 물리적으로 공간을 함께 사용하는 기계 세트를 갖추고, 100 Gbit/s 이상의 데이터 센터 네트워크로 상호 연결되어 있으며, 인공지능 훈련을 위해 초당 1020 정수 또는 부동 소수점 연산의 이론적 최대 컴퓨팅 용량을 가진 모든 컴퓨팅 클러스터.
- (c) 2015년 4월 1일의 행정명령 13694호(중요 악의적 사이버 활동에 참여하는 특정 개인의 재산 차단)로 선포된 심각한 악의적 사이버 활동과 관련된 국가 비상사태를 처리하기 위해 추가적인 조치를 취해야 한다고 판단합니다. 이는 2016년 12월 28일 행정명령 13757호(심각한 악의적 사이버 활동에 대한 국가 비상사태에 대응하기 위해 추가 조치를 취함)로 수정되었고, 행정명령 13984로 더욱 수정되어, 미국 인프라서비스(IaaS) 제품을 악의적 사이버 행위자들이 이용하는 것에 대응하기 위해, 외국 거래와 관련한 추가 기록 유지 의무를 부과하고 외국 악의적 사이버 행위자들의 거래 조사를 지원하기 위해, 상무부 장관에게 이 행정명령의 날짜로부터 90일 이내에 다음과 같이 지시합니다:
 - (i) 외국인이 미국 IaaS 제공업체와 거래하여 악의적 사이버 활동에 사용될 수 있는 잠재 능력을 가진 대규모 AI 모델을 훈련할 때("훈련 실행"이라 함), 그 미국 IaaS 제공업체가 상무부 장관에게 보고서를 제출하도록 하는 규정을 제안합니다. 이러한 보고서에는 최소한 외국인의 신원과 본 절에서 정한 기준 또는 장관이 규정에서 정의한 기타 기준을 충족하는 AI 모델의 훈련 실행 존재 여부, 그리고 장관이 식별한 추가 정보가 포함되어야 합니다.
 - (ii) 본 절의 4.2(c)(i)에 따라 제안된 규정에 미국 IaaS 제공업체가 해당 외국 리셀러가 외국 리셀러에게 보고서를 제출하지 않는 한 자신들의 미국 IaaS 제품을 제공하지 못하도록 하는 요건을 포함시킵니다. 이러한 보고서에는 미국 IaaS 제공업체가 상무부 장관에게 제공해야 할, 외국인이 외국 리셀러와 거래하여 미국 IaaS 제품을 사용해 본 절의 4.2(c)(i)에 설명된 훈련 실행을 수행하는 각 사례를 상세히 기술해야 합니다. 이러한 보고서에는 최소한 본 절의 4.2(c)(i)에 명시된 정보와 장관이 식별한 추가 정보가 포함되어야 합니다.
 - (iii) 악의적 사이버 활동에 사용될 수 있는 잠재 능력을 가진 대규모 AI 모델에 대한 기술 조건 세트를 결정하고, 필요하고 적절하다고 판단될 때마다 그 결정을 수정합니다. 장관이 이러한 결정을 내리기 전까지, 모델은 1026 정수 또는 부동 소수점 연산 이상의 컴퓨팅 파워가 필요하고, 단일 데이터센터에 물리적으로 함께 위치한 기계 세트를 갖추고 있으며, 100 Gbit/s 이상의 데이터 센터 네트워크로 서로 연결되어 있고, AI 훈련을 위한 이론상 최대 컴퓨팅 용량이 초당 1020 정수 또는 부동 소수점 연산인 컴퓨팅 클러스터에서 훈련되는 경우, 악의적 사이버 활동에 사용될 수 있는 잠재 능력을 가진 것으로 간주됩니다.
- (d) 본 명령의 날짜로부터 180일 이내에, 본 절의 4.2(c)에 명시된 발견에 따라, 상무부 장관은 미국 IaaS 제공업체가 외국 리셀러가 미국 IaaS 제품을 판매할 때, 해당 외국인이 IaaS 계정(계정)을 얻을 때 신원을 확인하도록 요구하는 규정을 제안해야 합니다. 이 규정에는 최소한 다음 사항을 포함해야 합니다:
 - (i) 외국인이 외국 리셀러로부터 계정을 개설하거나 기존 계정을 유지할 때, 외국 리셀러가 그 신원을 확인하기 위해 필요한 최소 기준을 미국 IaaS 제공업체가 요구해야 하는 것을 규정해야 하며, 여기에는 다음이 포함됩니다:
 - (A) 외국 리셀러가 이러한 제품이나 서비스의 임차인이나 하차인으로 활동하는 모든 외국인의 신원을 확인하기 위해 필요한 문서 유형과 절차;
 - (B) 외국 리셀러가 계정을 얻는 외국인에 관해 안전하게 유지해야 하는 기록으로, 다음 정보를 포함해야 합니다:
 - (1) 해당 외국인의 신원, 이름과 주소를 포함;
 - (2) 지불 수단 및 지불 출처(관련 금융기관 및 신용카드 번호, 계정 번호, 고객 식별자, 거래 식별자 또는 가상 화폐 지갑 또는 지갑 주소 식별자와 같은 기타 식별자 포함);
 - (3) 외국인의 신원을 확인하기 위해 사용된 이메일 주소 및 전화 연락처 정보;

- (4) 해당 계정의 소유를 지속적으로 확인하기 위해 접근 또는 관리에 사용된 인터넷 프로토콜(IP) 주소와 각 접근 또는 관리 조치 관련 날짜와 시간; 및
- (C) 외국 리셀러가 본 절에서 설명한 정보에 대한 모든 제3자 접근을 제한하기 위해 실행해야 하는 방법들로, 그러한 접근이 본 명령과 적용 가능한 법률에 따라 허용되는 한도 내에서는 예외입니다;
- (ii) 외국 리셀러가 유지하는 계정의 유형, 계정 개설 방법, 신원 확인을 위해 사용 가능한 정보의 유형을 고려하여, 그러한 제품을 사용하는 외국 악의적 사이버 행위자를 식별하고 리셀러에게 지나친 부담을 주지 않는 목표를 달성할 수 있도록 해야 합니다; 그리고
- (iii) 상무부 장관이 방위부 장관, 법무부 장관, 국토안보부 장관 및 국가정보국장과의 협의 하에, 장관이 설정할 수 있는 기준과 절차에 따라, 특정 외국 리셀러의 미국 IaaS 제품이나 특정 유형의 계정 또는 임차인에 대해 미국 IaaS 제공 업체를 본 절에 따라 발행된 어떤 규정의 요구 사항으로부터 면제할 수 있도록 합니다. 이러한 기준과 절차에는 해당 외국 리셀러, 계정 또는 임차인이 미국 IaaS 제품의 오남용을 방지하기 위해 보안 최상의 관행을 준수한다는 상무부 장관의 판단이 포함될 수 있습니다.
- (e) 상무부 장관은 본 절의 4.2(c) 및 (d)의 목적을 수행하는 데 필요한 조치를 취하기 위해, 규칙과 규정을 공포하는 것을 포함하여, 국제 긴급 경제 권한 법(International Emergency Economic Powers Act), 50 U.S.C. 1701 이하에 의해 대통령에게 부여된 모든 권한을 사용할 수 있도록 여기에 권한을 부여받습니다. 이러한 조치에는 미국 IaaS 제공업체가 외국 리셀러로부터 미국 IaaS 제품을 판매하도록 요구하고, 해당 리셀러가 상기 절과 관련된 검증을 미국 IaaS 제공업체에 제공하도록 요구하는 것이 포함될 수 있습니다.

(4.3) 중요 인프라 및 사이버 보안에서의 인공지능 관리

- (a) 중요 인프라 보호를 보장하기 위해 다음 조치가 취해질 것이다:
- (i) 본 명령의 날짜로부터 90일 이내, 그리고 그 이후에는 최소한 매년, 중요 인프라에 대한 관련 규제 권한을 가진 각 기관의 장과 관련된 SRMAs(부문별 위험 관리 기관)는 국토안보부 산하 사이버 보안 및 인프라 보안 기관의 국장과 협조하여 부문 간 위험을 고려하고, 관련 중요 인프라 부문에서 인공지능 사용과 관련된 잠재적 위험을 평가하고 국토안보부 장관에게 보고할 것이다. 이 평가에는 인공지능을 배치함으로써 중요 인프라 시스템이 중대한 실패, 물리적 공격, 사이버 공격에 더 취약해질 수 있는 방법을 포함하고, 이러한 취약성을 완화할 방법을 고려할 것이다. 독립 규제기관은 그들이 적절하다고 여기는 경우 부문별 위험 평가에 기여하도록 권장된다.
- (ii) 본 명령의 날짜로부터 150일 이내에, 재무부 장관은 금융 기관이 인공지능 특정 사이버 보안 위험을 관리하기 위한 모범 사례에 관한 공개 보고서를 발행할 것이다.
- (iii) 본 명령의 날짜로부터 180일 이내에, 국토안보부 장관은 상무부 장관 및 국토안보부 장관이 결정한 SRMAs 및 기타 규제기관과 협력하여, 적절하다고 판단되는 경우 인공지능 위험 관리 프레임워크(NIST AI 100-1) 및 기타 적절한 보안 지침을 중요 인프라 소유주 및 운영자가 사용할 수 있는 관련 안전 및 보안 지침에 통합할 것이다.
- (iv) 본 절의 4.3(a)(iii)에 설명된 지침 완성 후 240일 이내에, 국가안보 문제를 담당하는 대통령 보좌관과 OMB(관리 예산처) 국장은 국토안보부 장관과 협의하여, 중요 인프라에 대한 권한을 가진 기관의 장에 의한 작업을 조정하고, 연방 정부가 규제 또는 기타 적절한 조치를 통해 해당 지침 또는 그에 적절한 부분을 의무화하기 위한 조치를 개발하고 취할 것이다. 독립 규제기관은 자신들의 권한 및 책임 영역에서 규제 조치를 통해 지침을 의무화할 것인지를 고려하도록 권장된다.
- (v) 국토안보부 장관은 2002년 국토 안보법(Public Law 107-296) 제871조에 따라 자문위원회로 인공지능 안전 및 보안위원회를 설립할 것이다. 자문위원회에는 적절하다고 판단되는 민간 부문, 학계 및 정부의 인공지능 전문가가 포함되어 국토안보부 장관과 연방 정부의 중요 인프라 커뮤니티에 인공지능 사용과 관련된 보안, 복원력 및 사건 대응을 개선하기 위한 조언, 정보 또는 권장 사항을 제공할 것이다.
- (b) 미국의 사이버 방어를 개선하기 위한 인공지능의 잠재력을 활용하기 위하여:
- (i) 국방부 장관은 본 절의 4.3(b)(ii) 및 (iii) 소절에서 설명된 조치들을 국가 안보 시스템에 대해 실행할 것이며, 국토안보부 장관은 비국가 안보 시스템에 대해 이러한 조치들을 실행할 것이다. 각 장관은 국방부 장관과 국토안보부 장관이 적절하다고 여기는 기타 관련 기관의 장들과 협의하여 이를 수행할 것이다.
- (ii) 본 절의 4.3(b)(i)에 명시된 바와 같이, 이 명령의 날짜로부터 180일 이내에, 국방부 장관과 국토안보부 장관은 각각 적용 가능한 법률에 따라, 미국 정부의 중요 소프트웨어, 시스템 및 네트워크에서 취약점을 발견하고 해결하기 위한 대규모 언어 모델과 같은 인공지능 기능을 식별, 개발, 테스트, 평가 및 배치하는 운영 시범 프로젝트 계획을

수립, 수행하고 완료할 것이다.

- (iii) 본 절의 4.3(b)(i)에 명시된 바와 같이, 이 명령의 날짜로부터 270일 이내에, 국방부 장관과 국토안보부 장관은 각각 국가안보 문제를 담당하는 대통령 보좌관에게 본 절의 4.3(b)(ii)에 의해 요구된 계획 및 운영 시범 프로젝트에 따라 취한 조치의 결과에 관한 보고서를 제공할 것이다. 이에는 인공지능 기능의 개발 및 배치를 통해 발견되고 수정된 취약점의 설명과 사이버 방어를 위해 인공지능 기능을 효과적으로 식별, 개발, 테스트, 평가 및 배치하는 방법에 대한 교훈이 포함될 것이다.

(4.4) 인공지능과 CBRN 위협의 교차점에서 위험 감소

- (a) 인공지능이 CBRN 위협 — 특히 생물학적 무기에 중점을 두고 — 개발 또는 사용을 돕는 용도로 오용될 위험을 더 잘 이해하고 완화하기 위해 다음 조치들이 취해질 것이다:
- (i) 이 명령의 날짜로부터 180일 이내에, 국토안보부 장관은 에너지부 장관 및 과학기술정책국(OSTP) 국장과 협의하여 인공지능이 CBRN 위협 개발이나 생산을 가능하게 하는 데에 오용될 가능성을 평가하며, 이러한 위험을 대응하는 데 인공지능의 이점과 적응을 고려할 것이며, 적절한 경우 이 명령의 제8조(b)에 따라 수행된 작업 결과를 포함할 것이다. 국토안보부 장관은:
- (A) 에너지부, 민간 인공지능 실험실, 학계 및 제3자 모델 평가자로부터 인공지능과 CBRN 문제에 대한 전문가들과 협의하여 — 오로지 그러한 위협을 방어하기 위한 목적으로 — CBRN 위협을 나타내는 인공지능 모델의 능력을 평가하고, 이러한 위협을 발생시키거나 악화시키는 인공지능 모델 오용의 위험을 최소화하기 위한 옵션들을 평가할 것이며;
- (B) 이러한 노력의 진행 상황을 설명하는 보고서를 대통령에게 제출할 것이며, 이 보고서에는 미국에 CBRN 위협을 제시할 수 있는 인공지능 모델의 유형에 대한 평가와, 안전 평가 요구 사항 및 국가 안보에 대한 잠재적 위협을 완화하기 위한 안전 장치를 포함한 이러한 모델의 교육, 배치, 출판 또는 사용을 규제하거나 감독하기 위한 권고 사항이 포함될 것이다.
- (ii) 이 명령의 날짜로부터 120일 이내에, 국방부 장관은 국가안보 문제를 담당하는 대통령 보좌관 및 OSTP 국장과 협의하여, 과학, 공학, 의학 아카데미에 계약을 체결하여 — 그리고 국방부 장관, 국가안보 문제를 담당하는 대통령 보좌관, 유행병 대비 및 대응 정책국 국장, OSTP 국장 및 최고 데이터 책임자 협의회 의장에게 — 다음과 같은 연구를 수행하고 제출할 것이다:
- (A) 생물학적 데이터로 훈련된 생성 인공지능 모델을 포함하여 인공지능이 생물안보 위험을 증가시키는 방법을 평가하고, 이러한 위험을 완화하기 위한 권고 사항을 제시할 것이며;
- (B) 미국 정부가 호스팅하거나 생성하거나 창출에 자금을 제공하거나 그 밖에 소유하는, 특히 병원체와 옴스 연구와 관련된 데이터 및 데이터셋의 사용이 국가 안보에 미치는 영향을 고려하고, 생성 인공지능 모델의 훈련을 위해 이러한 데이터 및 데이터셋의 사용과 관련된 위험을 완화하기 위한 권고 사항을 제시할 것이며;
- (C) 생물학에 적용된 인공지능이 생물안보 위험을 줄이는 데 사용될 수 있는 방법을 평가하고, 데이터와 고성능 컴퓨팅 자원을 조정할 기회에 대한 권고 사항을 포함할 것이며;
- (D) 국방부 장관이 적절하다고 여기는 인공지능과 합성 생물학의 교차점에서의 추가적인 우려와 기회를 고려할 것이다.
- (b) 인공지능이 이 분야에서 능력을 크게 향상시킬 수 있는 합성 뉴클레산(nucleic acid)의 오용 위험을 감소시키고, 뉴클레산 합성 산업의 생물안보 조치를 개선하기 위해 다음과 같은 조치들이 취해질 것이다:
- (i) 이 명령 날짜로부터 180일 이내에, 과학기술정책국(OSTP) 국장은 국무장관, 국방장관, 법무장관, 상무장관, 보건복지서비스부(HHS) 장관, 에너지장관, 국토안보부 장관, 국가정보국장, 그리고 OSTP 국장이 적절하다고 여길 기타 관련 기관의 수장들과 협의하여, 적절한 경우 기존 미국 정부의 지침을 포함하여, 합성 뉴클레산 시퀀스 공급자들이 포괄적이고 확장 가능하며 검증 가능한 합성 뉴클레산 조달 스크리닝 메커니즘을 구현하도록 권장하는 프레임워크를 수립할 것이다. 이러한 프레임워크의 일환으로, OSTP 국장은:
- (A) 미국의 국가 안보에 위험을 초래할 수 있는 생물학적 시퀀스를 지속적으로 식별하기 위한 기준과 메커니즘을 수립할 것이며;
- (B) 시퀀스 합성 조달 스크리닝의 수행 및 성능 검증을 위한 표준화된 방법론 및 도구를 결정하며, 이 섹션의 4.4(b)(i)(A) 항에서 식별된 생물학적 시퀀스 구매자들에 의해 제기되는 보안 위험을 관리하기 위한 고객 스크리닝 접근법을 지원하고, 문제가 되는 활동에 대한 집행 기관에 대한 보고 절차를 포함할 것이다.
- (ii) 이 명령 날짜로부터 180일 이내에, 상무부 장관은 NIST 국장을 통하여, OSTP 국장과 협력하고, 국무장관, HHS

장관, 그리고 상무부 장관이 적절하다고 여길 기타 관련 기관의 수장들과 협의하여, 이 섹션의 4.4(b)(i) 하위항에 따라 개발된 프레임워크를 기반으로 산업 및 관련 이해 관계자들과의 참여를 시작할 것이며, 합성 뉴클레산 시퀀스 공급자들이 사용할 수 있도록 개발 및 정제하기 위해:

- (A) 효과적인 뉴클레산 합성 조달 스크리닝에 대한 사양을 개발할 것이며;
 - (B) 이러한 스크리닝을 지원하기 위한 관심 시퀀스 데이터베이스를 관리하기 위한 보안 및 접근 제어를 포함한 최상의 관행을 개발할 것이며;
 - (C) 효과적인 스크리닝을 위한 기술 구현 가이드를 개발할 것이며;
 - (D) 적합성 평가에 대한 최상의 관행 및 메커니즘을 개발할 것이다.
- (iii) 이 섹션의 4.4(b)(i) 하위항에 따라 프레임워크가 수립된 후 180일 이내에, 생명 과학 연구를 자금 지원하는 모든 기관들은, 적절한 경우 및 적용 가능한 법률에 따라, 합성 뉴클레산 조달이 프레임워크를 준수하는 공급자나 제조 업체를 통해 수행되도록 할 것이며, 예를 들어 공급자나 제조업체로부터의 확인서를 통해 이를 요구하는 조건으로 자금 지원을 할 것이다. 국가 안보 담당 대통령 보좌관과 OSTP 국장은 이러한 자금 지원 요구 사항을 검토하는 과정을 조정하여, 자금 지원 기관 간 프레임워크의 일관된 구현을 용이하게 할 것이다.
- (iv) 이 섹션의 4.4(b)(i)-(iii) 하위항에서 설명된 조치들의 효과적인 구현을 용이하게 하기 위해, 국토안보부 장관은 국토안보부 장관이 적절하다고 여길 기타 관련 기관의 수장들과 협의하여:
- (A) 이 섹션의 4.4(b)(i) 하위항에 따라 프레임워크가 수립된 후 180일 이내에, 합성 뉴클레산 시퀀스 공급자들이 수행하는 이 섹션의 4.4(b)(i)-(ii) 하위항에 따라 개발된 시스템을 포함하여 뉴클레산 합성 조달 스크리닝의 구조화된 평가 및 스트레스 테스트를 수행하기 위한 프레임워크를 개발할 것이며;
 - (B) 이 섹션의 4.4(b)(iv)(A) 하위항에 따라 프레임워크를 개발한 후, 국가 안보 담당 대통령 보좌관, 대유행 준비 및 대응 정책국 국장, 그리고 OSTP 국장에게 이 섹션의 4.4(b)(iv)(A) 하위항에 따라 수행된 활동의 결과를 포함한 연례 보고서를 제출할 것이며, 해당되는 경우 뉴클레산 합성 조달 스크리닝, 고객 스크리닝 시스템을 강화하기 위한 권고 사항을 포함할 것이다.

(4.5) 인공지능이 생성한 합성 콘텐츠에 의해 초래되는 위험 감소

인공지능 시스템에 의해 생성된 합성 콘텐츠를 식별하고 라벨링하는 능력을 촉진하고, 연방 정부가 생산하거나 그를 대신하여 생산한 디지털 콘텐츠(합성 여부에 관계없이)의 진정성과 출처를 확립하기 위해:

- (a) 이 명령서 날짜로부터 240일 이내에, 상무부 장관은 상무부 장관이 적절하다고 생각하는 기타 관련 기관 수장들과 협의하여, OMB 국장과 국가안보 담당 대통령 보좌관에게 다음을 식별한 보고서를 제출해야 한다. 기존의 표준, 도구, 방법 및 관행뿐만 아니라, 추가적인 과학 기반 표준 및 기술의 발전 가능성:
 - (i) 콘텐츠의 인증 및 그 출처 추적;
 - (ii) 워터마킹과 같은 합성 콘텐츠 라벨링;
 - (iii) 합성 콘텐츠 탐지;
 - (iv) 아동 성적 학대 자료 생성을 방지하거나 실제 개인의 비동의 성적 이미지(식별 가능한 개인의 신체 또는 신체 부위의 디지털 묘사 포함) 생성을 방지하는 인공지능 생성;
 - (v) 위 목적을 위한 소프트웨어 테스트; 및
 - (vi) 합성 콘텐츠의 감사 및 유지 관리.
- (b) 이 섹션의 4.5(a) 하위 섹션에 따라 요구되는 보고서를 제출한 후 180일 이내에, 그리고 이후 정기적으로 업데이트하여, 상무부 장관은 OMB 국장과 협력하여, 디지털 콘텐츠 인증 및 합성 콘텐츠 탐지 조치를 위한 기존 도구 및 관행에 관한 지침을 개발해야 한다. 이 지침은 이 섹션의 4.5(a) 하위 섹션에 나열된 목적을 위한 조치를 포함해야 한다.
- (c) 이 섹션의 4.5(b) 하위 섹션에 따라 지침이 개발된 후 180일 이내에, 그리고 이후 정기적으로 업데이트하여, OMB 국장은 국무장관, 국방장관, 법무장관, 상무부 장관(이를 통해 NIST 국장을 대신하여), 국토안보부 장관, 국가정보국장, 및 OMB 국장이 적절하다고 생각하는 기타 기관 수장들과 협의하여, 미국 정부의 공식 디지털 콘텐츠에 대한 대중의 신뢰를 강화하기 위한 목적으로 — 해당 기관들이 생산하거나 발행하는 콘텐츠의 라벨링과 인증을 위한 지침을 발행할 것이다.
- (d) 연방 취득 규정위원회는 이 섹션의 4.5 하위 섹션에 따라 설립된 지침을 고려하여, 적절하고 적용 가능한 법률에 일치하게 연방 취득 규정을 개정할 것을 고려해야 한다.

(4.6) 널리 사용되는 모델 가중치를 갖는 이중 용도 기초 모델에 대한 의견 요청

이중 용도 기초 모델의 가중치가 널리 공개되어 있는 경우 — 예를 들어 인터넷에 공개적으로 게시되었을 때 — 혁신에 많은 이점이 있을 수 있지만 모델 내의 안전장치 제거와 같은 상당한 보안 위험도 존재할 수 있다. 널리 공개된 가중치를 갖는 이중 용도 기초 모델의 위험과 잠재적 이점을 다루기 위해, 이 명령서 날짜로부터 270일 이내에, 상무부 장관은 통신 및 정보 담당 상무부 차관보를 대신하여, 국무장관과 협의하여 다음을 수행해야 한다:

- (a) 민간 부문, 학계, 시민 사회 및 기타 이해관계자들로부터 공개적인 협의 과정을 통해 널리 공개된 가중치를 갖는 이중 용도 기초 모델에 대한 잠재적 위험, 이점, 기타 함의 및 적절한 정책 및 규제 접근 방식에 대한 의견을 요청하며, 이는 다음을 포함한다:
 - (i) 널리 공개된 가중치를 갖는 이중 용도 기초 모델에 대한 조정 또는 해당 모델의 안전장치 제거와 관련된 위험;
 - (ii) 널리 공개된 가중치를 갖는 이중 용도 기초 모델의 AI 혁신 및 연구에 대한 이점, AI 안전성 및 위험 관리에 대한 연구를 포함하여; 및
 - (iii) 널리 공개된 가중치를 갖는 이중 용도 기초 모델의 위험을 관리하고 이점을 극대화하기 위한 잠재적 자발적, 규제적, 국제적 메커니즘;
- (b) 이 섹션의 4.6(a) 하위 섹션에서 설명한 과정으로부터의 의견을 바탕으로, 상무부 장관이 적절하다고 생각하는 기타 관련 기관 수장들과 협의하여, 널리 공개된 가중치를 갖는 이중 용도 기초 모델의 잠재적 이점, 위험 및 함의에 관한 보고서를 대통령에게 제출하고, 또한 해당 모델과 관련된 정책 및 규제 권고사항을 제출해야 한다.

(4.7) 연방 데이터의 안전한 공개 촉진 및 AI 교육용 악의적 사용 방지

공공 데이터 접근성을 개선하고 보안 위험을 관리하며, 개방, 공공, 전자, 필수 정부 데이터 법(공공법 115-435호 제III 타이틀)의 목표와 일관성을 유지하여, 기계가 읽을 수 있는 형식으로 연방 데이터 자산에 대한 공공 접근을 확장하면서 보안 고려 사항을 고려하는 것을 포함하여, 개별 데이터 자산이 단독으로는 보안 위험을 초래하지 않지만 다른 이용 가능한 정보와 결합될 때 보안 위험을 초래할 수 있는 위험을 고려하여:

- (a) 이 명령서의 날짜로부터 270일 이내에, 수석 데이터 책임자 위원회는 국방부 장관, 상무부 장관, 에너지부 장관, 국토 안보부 장관 및 국가정보국 국장과 협의하여, 연방 데이터 공개가 화학, 생물, 방사능 및 핵무기(CBRN) 개발뿐만 아니라 자율 공격 사이버 능력 개발에 도움이 될 수 있는 잠재적 보안 위험을 식별하고 관리하는 보안 검토를 포함하는 초기 가이드라인을 개발해야 하며, 동시에 개방, 공공, 전자, 필수 정부 데이터 법(공공법 115-435호 제III 타이틀)에 명시된 목표에 따라 연방 정부 데이터에 대한 공공 접근을 제공해야 한다; 그리고
- (b) 이 절의 4.7(a) 하위절에 의해 요구되는 초기 가이드라인 개발 후 180일 이내에, 기관들은 44 U.S.C. 3511(a)(1) 및 (2)(B)에 따라 요구되는 포괄적인 데이터 인벤토리에 있는 모든 데이터 자산에 대한 보안 검토를 수행해야 하며, 적절하고 적용 가능한 법률에 따라, 해당 데이터 공개가 화학, 생물, 방사능 및 핵무기(CBRN)와 같은 고위험도의 잠재적 보안 위험을 제거할 수 있는 부분을 해결하기 위한 조치를 취해야 한다.

(4.8) 국가안보메모 개발 지시

인공지능(AI)의 보안 위험을 관리하기 위한 조정된 행정부 접근 방식을 개발하기 위해, 대통령 국가안보보좌관과 대통령 정책 비서실장은 이 명령서의 날짜로부터 270일 이내에 대통령에게 제출할 인공지능에 관한 제안된 국가안보메모 개발을 목적으로 하는 부처 간 협의 과정을 감독할 것이다. 이 메모는 국가 안보 시스템의 구성요소로 사용되거나 군사 및 정보 목적으로 사용되는 AI의 거버넌스에 관해 다룰 것이다. 메모는 국가안보 시스템의 개발과 사용을 관리하기 위한 현재의 노력을 고려할 것이다. 또한 메모는 국방부, 국무부, 기타 관련 기관 및 정보 커뮤니티가 AI가 제기하는 국가안보 위험과 잠재적 이점을 다루기 위한 조치를 개요할 것이다. 특히, 메모는:

- (a) 미국인의 권리나 안전에 영향을 줄 수 있는 국가안보 AI 사용에 대한 특정 AI 보증 및 위험관리 관행을 지시하는 것을 포함하여, 미국 국가안보 임무를 발전시키기 위해 AI 능력의 지속적인 채택에 대한 국방부, 기타 관련 기관 및 정보 커뮤니티에 대한 지침을 제공하고, 적절한 상황에서는 미국인이 아닌 사람들에게도 마찬가지로 적용될 것이며;
- (b) 국방부나 정보 커뮤니티의 역량이나 목표를 위협하거나, 그 밖에 미국 또는 그 동맹 및 파트너의 안보에 위험을 제기하는 방식으로 적국과 기타 외국 행위자들이 AI 시스템을 사용할 수 있는 잠재력을 다루기 위해 적절하고 적용 가능한 법률에 따라 지속적인 조치를 지시할 것이다.

(9) 개인정보 보호

(a) 인공지능 — 개인에 대한 정보의 수집이나 사용을 촉진하거나 개인에 대한 추론을 하는 인공지능을 포함하여 — 에 의해 악화될 수 있는 개인정보 보호 위험을 완화하기 위해, OMB 국장은 다음과 같은 조치를 취해야 한다:

(i) 국가 보안 목적으로 사용되는 경우를 제외하고, 특히 개인을 식별할 수 있는 정보를 포함하는 상업적으로 이용 가능한 정보(CAI) 및 데이터 브로커로부터 조달되거나 공급업체를 통해 간접적으로 조달 및 처리된 CAI를 포함하여, 기관에서 조달한 CAI를 적절한 기관 재고 및 보고 절차에서 식별하고 평가하는 조치를 취할 것이다;

(ii) 연방 개인정보 보호 위원회와 통계 정책 협의회와 협의하여, 국가 보안 목적으로 사용되는 경우를 제외하고, 개인을 식별할 수 있는 정보를 포함하는 CAI의 수집, 처리, 유지, 사용, 공유, 전파 및 처분과 관련된 기관의 기준과 절차를 평가하여, CAI 관련 기관의 활동으로부터 개인정보 보호와 기밀성 위험을 완화하기 위한 방법에 대한 기관들에 대한 잠재적 지침을 알리기 위해 조치를 취할 것이다;

(iii) 이 명령서의 날짜로부터 180일 이내에, 법무장관, 경제정책 담당 대통령 보좌관 및 OSTP 국장과 협의하여, 2002년 전자정부법(E-Government Act of 2002, 공공법 107-347)의 개인정보 보호 조항을 시행하기 위한 기관에 대한 지침 개정을 알아보기 위한 RFI(Request for Information, 정보 요청)를 발행할 것이다. RFI는 개인정보 보호 영향 평가가 인공지능에 의해 더 악화되는 개인정보 보호 위험을 완화하는 데 어떻게 더 효과적일 수 있는지에 대한 피드백을 요청할 것이다; 그리고

(iv) 적용 가능한 법률에 일치하여, RFI 과정을 통해 식별된 단기 조치와 장기 전략을 지원하고 진전시키기 위한 필요하고 적절한 조치를 취할 것이다. 이는 새로운 또는 업데이트된 지침 또는 RFI 발행 또는 연방 개인정보 보호 위원회 또는 다른 기관과의 협의를 포함한다.

(b) 이 명령서의 날짜로부터 365일 이내에, 기관들이 인공지능에 의해 악화될 수 있는 잠재적 위험으로부터 미국인의 개인정보를 보호하기 위해 PETs(Privacy Enhancing Technologies, 개인정보 보호 강화 기술)를 사용할 수 있도록 더 잘 할 수 있도록, 상업부 장관은 NIST 국장을 통해 기관들이 차등 프라이버시 보증 보호의 효과를 평가하는 지침을 만들 것이다. 이 지침은 최소한 차등 프라이버시 보호와 실제에서 차등 프라이버시를 실현하는 데 있어 흔히 발생하는 위험에 영향을 미치는 중요한 요소들을 기술할 것이다.

(c) PETs와 관련된 연구, 개발 및 구현을 진전시키기 위해:

(i) 이 명령서의 날짜로부터 120일 이내에, NSF 국장은 에너지부 장관과 협력하여 개인정보 보호 연구를 진전시키고, 특히 PETs의 개발, 배치 및 확대에 전념하는 연구 조정 네트워크(RCN)를 창설하기 위해 자금을 지원할 것이다. RCN은 개인정보 보호 연구자들이 정보를 공유하고, 연구에서 협력 및 조정을 하며, 개인정보 보호 연구 커뮤니티를 위한 표준을 개발하는 데 도움을 줄 것이다.

(ii) 이 명령서의 날짜로부터 240일 이내에, NSF 국장은 기관들과 협력하여 기관의 운영에 PETs를 통합할 수 있는 진행 중인 작업과 잠재적 기회를 식별할 것이다. NSF 국장은 가능하고 적절한 경우, 기관의 사용을 위해 선도적인 PETs 솔루션의 채택을 장려하는 연구를 우선 순위에 두어야 하며, 이는 (c)(i) 절에서 설명한 RCN을 통한 연구 참여를 포함한다.

(iii) NSF 국장은 미국-영국 PETs 상급 챌린지의 결과를 사용하여 PETs 연구 및 채택에 대한 접근 방식과 식별된 기회를 안내할 것이다.

KISA INSIGHT

DIGITAL & SECURITY POLICY

2023 VOL. 6