2024. 12. 31.

국회입법조사처 | NARS 입법·정책 | 제162호

인공지능의 내재적 위험과 입법·정책 과제

- 데이터·기술·이용자를 중심으로 -

정준화 | 입법조사관(과학방송통신팀)

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE

NARS 입법·정책 제162호

인공지능의 내재적 위험과 입법·정책 과제
- 데이터·기술·이용자를 중심으로 -

정준화(과학방송통신팀 입법조사관)

2024. 12. 31.

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE

NARS

NARS 입법·정책은 국회에서 논의가 필요한 핵심적인 입법 및 정책 현안 주제를 선정하여 심도있게 분석·평가하고 입법 및 정책 대안을 제시하는 보고서입니다. 이 보고서가 국회의 위원회와 국회의원의 의정활동에 참고자료로 널리 활용되고, 입법·정책 현안에 대한 국민의 이해를 높이는데 기여하기를 기대합니다.

이 보고서는 「국회법」 제22조의3 및 「국회입법조사처법」 제3조에 따라 국회의원의 의정활동을 지원하기 위하여, 국회입법조사처 보고서 심의절차를 거쳐 발간(2024. 12. 31.)되었습니다.

요 약

□ 인공지능(AI)은 인간의 지능적 행위에 들어가는 시간과 수고를 줄여주고 인간이 직접 수행해 왔던 판단과 창작 등의 기능을 보완·대체·증강함으로써 인간이 더 방대하고 새로운 지능적 활동을 경험할 수 있게 해 주지만, 한편으로는 인간의 일자리를 대체하거나 불완전하고 편향적인 결과를 제시하여 개인과 사회에 더 큰 혼란을 초래하는 위험을 내포하고 있음

□ AI의 위험은 어느 영역(vertical)의 기존 기술·서비스에 AI가 적용되면서 발생하는 '파생적 위험'과 데이터 학습, AI 모델 구축·작동 등 AI의 생애주기에 걸쳐 공통으로(cross-sectional) 발생하는 '내재적 위험'으로 구분할 수 있음

○ 파생적 위험은 콜센터에 AI를 적용하여 기존 일자리가 대폭 감소한 것과 같이 AI가 사용된 영역에서 직·간접적으로 발생하는 효과로서, 신기술의 도입 과정에서 나타나는 현상이며, 그 위험을 줄이려면 편익도 포기해야 하는 상충관계(trade-off)가 발생하므로 위험 대응과 함께 이익을 보는 집단과 피해를 보는 집단 사이의 재분배 조치가 필요함

○ 내재적 위험은 현실의 문제로 불거지기 전까지는 가시적으로 드러나지 않기 때문에 문제를 파악하는 것이 쉽지 않고, 문제가 발생하면 개인과 사회에 큰 피해를 초래하므로 다양한 가능성을 열어놓고 사전에 체계적이고 종합적인 대안을 모색하는 것이 중요함

□ 이 보고서는 AI의 내재적 위험의 분석과 대응에 초점을 두고 있으며, 국내외 기관과 기업의 연구 자료를 분석하여 AI의 내재적 위험을 학습데이터 위험, AI 모델·시스템의 기술적 위험, AI 모델·시스템의 이용자 위험으로 구분함

○ 학습데이터의 위험은 편향된 데이터로 인해 AI 산출물에 차별과 오류가 발생하는 '부적절한 데이터', 법률로 보호되는 개인정보와 저작권 등을 적법한 절차를 거치지 않고 학습에 사용하는 '부적법한 데이터'로 발생함

○ 기술적 위험은 AI 개발 및 운영 과정에 대한 '투명성·설명가능성 부족', AI의 자율성에 대한 '인간의 통제 곤란', AI 모델·시스템으로 인한 '인간의 권리 침해'로 발생함

○ 이용자 위험은 AI 입력 프롬프트에 개인정보 등을 입력하는 '부주의한 이용자', 허위 정보와 딥페이크 등을 생성하여 사회 혼란을 초래하는 '악의적 이용자', 프롬프트 조작과 데이터 탈취 등으로 AI 시스템의 보안을 위협하는 '사이버 공격자'로부터 발생함

□ 외국 입법례 중에서 AI의 내재적 위험에 대해 체계적이고 종합적인 안전장치를 마련한 것은 유럽연합의 「인공지능법」이며, 학습데이터 위험을 관리하기 위해 데이터 관리 및 학습데이터 요약본 공개 의무를 규정하고, 기술적 위험 관리를 위해 기술문서 작성과 적합성 평가 등을 의무화하고 있음

○ 고위험 AI 시스템의 공급자는 데이터 관리, 기술문서 작성, 배포자에 이용지침 제공, 적합성 평가 수행 등의 의무를 이행해야 하고, 배포자는 이용지침을 준수하고 기본권 영향평가 등을 수행해야 함

○ 생성형 AI를 포함하는 범용 AI(GPAI)의 공급자는 기술문서와 사용설명서 제공, 저작권 지침 준수, 학습에 사용된 데이터의 요약본 공개 등의 의무를 이행해야 함

□ 우리나라는 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)」이 2024년 12월 26일 국회 본회의를 통과하여 공포 후 1년 뒤에 시행될 예정이며, 고영향 AI에 대한 안전 및 신뢰성 관리 체계를 마련한 것이 특징임

○ 전반적으로 유럽연합 「인공지능법」의 고위험 AI를 참고하여 '고영향 AI'에 대해서는 기술과 데이터에 대한 설명방안 수립, 사람에 의한 관리·감독, 인간의 기본권에 미치는 영향평가 등의 위험 대응책을 규정하였으나, 국민에게 큰 영향을 미치고 있는 생성형 AI에 대해서는 적절한 위험 대응 조치가 마련되지 못한 측면이 있음

□ 이러한 상황에서 AI의 내재적 위험에 대응하기 위하여 앞으로 고려해야 할 입법·정책적 과제는 다음과 같음

○ 첫째, AI 학습데이터의 출처·편향성을 검토하고 투명성을 강화하기 위하여 학습데이터에 대한 설명을 강화하고, 이를 위해 인공지능기본법안에 열거된 고영향 AI뿐만 아니라 생성형 AI와 같은 일상에 큰 영향을 미치는 AI까지 학습데이터 설명 대상에 포함하는 방안을 모색해 볼 수 있음

○ 둘째, AI 학습데이터에 개인정보와 저작물이 활용되고 있는 현실을 감안하여, 온라인에 공개된 개인정보와 저작물의 보호와 AI 학습을 위한 활용을 균형적으로 달성할 수 있는 「개인정보 보호법」과 「저작권법」의 개정을 모색해 볼 수 있음

○ 셋째, 영어권 데이터를 학습한 AI 모델·시스템이 국내 현실에 맞는 결과를 충분히 산출하지 못하는 문제를 해결하기 위해서 국내 정치·경제·사회·문화 데이터를 축적하여 AI 학습에 활용할 수 있도록 하는 노력이 필요함

○ 넷째, AI 모델·시스템 프로세스의 투명성·설명가능성을 높이기 위해서 AI 기술문서 작성·보관 의무 대상을 인공지능기본법안에 열거된 고영향 AI뿐만 아니라 생성형 AI와 같은 일상에 큰 영향을 미치는 AI까지 포함하는 방안을 모색해 볼 수 있고, AI 모델을 어떻게 사용되어야 하는지에 대한 설명을 제시하는 것도 필요함

○ 다섯째, AI 모델·시스템 프로세스의 위험은 인간의 기본권에 대한 위험뿐만 아니라 정치·사회·경제 등 다양한 분야에서도 발생하므로, 인공지능기본법안의 위험평가 대상을 기본권 포함한 다양한 위험 요소까지 확장하는 방안을 모색해 볼 수 있으며, 이를 강제할 경우 기업의 부담이 가중될 수 있으므로 기업이 자율적으로 영향평가를 수행할 수 있도록 현행 AI 윤리영향평가 체계를 적극 활용하는 방안도 고려해 볼 필요가 있음

○ 여섯째, AI 모델·시스템 이용자가 중요 정보를 AI에 입력하지 않고, AI가 제시하는 결과를 과잉신뢰하지 않으며, 다양한 상황에서 AI 산출물을 비판적으로 활용할 수 있도록 이용자의 AI 리터러시(literacy)를 높일 필요가 있음

○ 일곱째, AI 모델·시스템에 대한 사이버 공격을 AI 개발자가 모두 예상하여 대응하는 것은 한계가 있으므로 가상의 적대적 테스트(레드티밍) 등을 활성화하여 대응 방안을 마련하고, 보안에 관한 국제협력에 적극 참여할 필요가 있음

차 례

표 차 례

그 림 차 례

Ⅰ. 서론

1. 인공지능의 등장과 발전

인공지능(artificial intelligence, AI)은 신기술로 인식되지만 이미 70년의 역사를 가지고 있다. 1950년 영국의 앨런 튜링(Alan Turing)이 「계산 기계와 지능(Computing Machinery and Intelligence)」이라는 논문을 통해 생각하는 기계, 즉 AI의 가능성을 제시했고, 1956년 미국 뉴햄프셔주 다트머스 대학에서 개최된 '다트머스 회의'에서 존 매카시(John McCarthy)가 '인공지능(Artificial Intelligence)'이라는 용어를 처음 사용함으로써 AI의 역사가 시작되었다.

지금까지 AI는 두 번의 큰 침체기(AI 분야에서는 이를 '겨울(winter)'이라 한다)를 겪으면서 발전해 왔다. 등장 초기에 AI에 대한 높은 기대에 힘입어 다양한 연구가 시작되었으나 당시의 컴퓨팅 성능이 충분하지 않아서 만족할 만한 성과를 내지 못했고 관련 투자가 줄어들어 1970년대 후반에 첫 번째 겨울을 맞았다. 그 이후 정해진 규칙에 따라(rule-based) 자동으로 판정을 내리는 전문가 시스템(Expert System)이 등장하여 잠시 붐을 이루다가 비용 대비 성능이 충분하지 못해 투자가 축소되고 1990년을 전후로 두 번째 겨울이 시작되었다.

현재는 인공신경망을 기반으로 AI가 스스로 학습해서 규칙을 찾는 기계학습(machine learning)이 발전하고 있으며, 2022년 11월에 기계학습 기반의 AI 챗봇인 챗지피티(ChatGPT)가 출시되면서 AI의 대중화 시기를 맞고 있다. 오늘날 AI의 특징은 대량의 데이터를 학습하여 AI 모델(model)의 성능을 극적으로 개선하고 이를 다양한 시스템(system), 즉 응용서비스에 활용하는 방식이다. 가장 대표적이고 대중적인 AI 시스템은 언어·이미지·영상 등을 만드는 생성형 AI 서비스이지만, 2024년 노벨 물리학상과 화학상에서 AI가 핵심적인 역할을 수행한 것을 보면 머지않아 여러 분야에서 AI 활용이 폭발적으로 확대될 것으로 보인다.

2. 인공지능의 내재적 위험과 대응 필요성

AI는 인간의 지능적 행위에 들어가는 시간과 수고를 줄여주고, 인간이 수행하던 판단과 창작 등의 기능을 보완·대체·증강함으로써 인간이 더 방대하고 새로운 지능적 활동을 경험할 수 있게 해 준다. 동시에 AI는 다양한 위험(risk)[1]을 가지고 있으며, 일부는 실제 피해로 나타나기도 한다.

AI의 위험은 어느 영역(vertical)의 산업·기술에 AI가 적용되면서 나타나는 '파생적 위험'과 데이터 학습, 모델 구축·작동 등 AI의 생애주기에 걸쳐 공통으로(cross-sectional) 발생하는 '내재적 위험'으로 구분할 수 있다. 이 중에서 파생적 위험은 콜센터에 AI를 적용하여 사람의 일자리가 대폭 감소한 것과 같이 AI가 사용된 영역에서 직·간접적으로 발생하는 효과로서, 그 위험을 줄이려면 편익도 포기해야 하는 상충관계(trade-off)가 존재하기 때문에 이익을 보는 집단과 피해를 보는 집단 사이의 재분배 규칙을 마련하는 것이 위험 대응 수단 중 하나다. 이와 달리 내재적 위험은 AI 모델·시스템 자체가 가지고 있는 위험 요인으로, 그 위험을 해소할 경우 AI의 정확성·신뢰성이 개선되어 AI 발전에 긍정적인 영향을 미치게 된다. 따라서 AI의 안전과 발전을 동시에 달성하기 위해서는 내재적 위험에 대한 대응이 반드시 필요하다.

AI의 내재적 위험은 실제 문제로 불거지기 전까지는 가시적으로 드러나지 않고, 문제가 발생하면 이미 돌이킬 수 없는 피해가 발생하기 때문에 피해 복구 방식의 대응이 아니라, 사전 예방 조치 중심으로 대안을 모색하는 것이 중요하다. 마침, 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)」이 2024년 12월 26일 국회 본회의를 통과하여 공포 후 1년 뒤에 시행될 예정이어서, AI의 내재적 위험에 대한 논의를 시작할 적절한 시점이다. 이에 이 보고서는 AI의 내재적 위험을 분석하고, 이를 해결하기 위한 입법·정책 과제를 살펴보기로 한다.

1) '위험'이란 해로움이나 손실이 생길 우려가 있거나, 우려가 있는 상태를 의미한다(표준국어대사전).

Ⅱ. 인공지능의 내재적 위험

1. 인공지능의 내재적 위험 유형

AI의 내재적 위험은 다양하다. 여러 연구에서 AI 위험을 제시하고 있는데, 각 연구마다 강조하는 부분이 조금씩 차이가 난다. 위험을 포괄적으로 살펴보기 위해서 복수의 연구에서 제시한 위험을 재분류할 필요가 있다. 이에 이 보고서는 연구·조사 기관의 신뢰성과 분야별 대표성 등을 종합적으로 고려하여 국제기구인 경제협력개발기구(OECD)가 제시한 10가지 위험[2], 국내 연구기관인 소프트웨어정책연구소(SPRi)가 분석한 25가지 위험[3], AI 기업인 아이비엠(IBM)이 설명한 67가지 위험[4]을 1차적으로 살펴보았다. 그다음 내재적 위험과 거리가 먼 AI의 사회적 파급효과, AI 거버넌스 문제, AI 격차(국가 간 및 국가 내 AI 불평등), AI 불균형(소수 기업 및 국가에 AI 영향력 집중) 등은 제외하고 최종적으로 다음 [표 1]과 같이 내재적 위험을 세 가지로 유형화했다.

첫 번째 유형은 '학습데이터 위험'이다. 여기에는 편향적이고 대표성이 낮은 부적절한 학습데이터, 무단으로 사용된 개인정보와 같은 부적법한 학습데이터 이슈가 포함된다. 두 번째 유형은 'AI 모델·시스템의 기술적 위험'이다. 기술적 위험을 초래하는 요인은 AI 내부 작동 방식에 대한 투명성·설명가능성 부족, 자율적 AI에 대한 인간의 통제 곤란, AI 작동으로 인한 인간의 권리 침해 등이다. 세 번째 유형은 'AI 모델·시스템 이용자의 위험'이다. AI의 위험을 충분히 인지

2) Organization for Economic Cooperation and Development(OECD), 『Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives』, OECD, 2024.
3) 노재원·유재흥·장진철·조지연, 『AI위험 유형 및 사례 분석』, SPRi 이슈리포트 IS-183, 소프트웨어정책연구소, 2024.
4) IBM, "AI risk atlas", (최종방문일: 2024.12.16.), <https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/hallucination.html?context=wx&locale=ko&audience=wdp>

하지 못한 부주의한 이용자, 타인과 사회에 혼란을 초래하기 위하여 악의적으로 AI를 이용하는 사람, AI로 타 사이트를 해킹을 하거나 AI 모델·시스템 자체를 해킹하는 사이버 공격자가 세 번째 유형의 위험을 초래한다.

[표 1] 인공지능의 내재적 위험 유형

구분	주요 요인	주요 내용		
		OECD	SPRi	IBM
학습 데이터 위험	부적절한 학습데이터	-	• 데이터 편향성 및 대표성 위험	• 편향된 데이터, 대표성 없는 데이터, 오염·중독된 데이터 등
	부적법한 학습데이터	-	-	• 개인정보 등 법으로 보호되는 데이터의 부적법한 사용 등
AI 모델·시스템의 기술적 위험	투명성·설명가능성 부족	• AI 시스템 개발 및 배포 경쟁으로 인한 안전성과 신뢰성 부족	• 신속한 배포에 따른 광범위한 피해 • 신뢰성 검증 및 보장 문제 • 내부 작동방식 이해 부족 • 배포에 대한 추적성 부족	• 신뢰할 수 없는 소스 귀속, 설명할 수 없는 출력, 추적 불가능 • 모델 투명성 부족, 테스트 다양성 부족, 시스템 다양성 부족, 대표성 없는 위험 테스트 등
	자율성 통제 곤란	• AI 시스템의 목적과 인간의 가치에 대한 조정 실패	• AI에 대한 통제력 상실, 자율성 강화에 따른 인간 통제 한계 • AI의 의도하지 않은 작동 등	-

구분	주요 요인	주요 내용		
		OECD	SPRi	IBM
	인간의 권리 침해	• 감시 강화 및 사생활 침해 • 중요한 시스템에서의 사고와 재난		• 악성 출력, 유해한 출력, 불완전한 조언 등
AI 모델·시스템의 이용자 위험	부주의한 이용자	-	• AI에 대해 잘못된 정보로 인한 오해·혼동	• 프롬프트에 개인정보, IP 입력 • AI에 대한 과잉 또는 과소 의존
	악의적 이용자	• 조작, 허위 정보, 사기 및 민주주의와 사회적 결속에 대한 피해	• 가짜 콘텐츠, 딥페이크 제작 • 허위 정보 및 여론 조작	• 동의하지 않은 사용, AI 활용 비공개, 부적절한 사용, 위험한 사용, 허위정보 유포 등
	사이버 공격자	• 고도화된 악의적 사이버 활동의 촉진	• 사이버 범죄	• 속성 추론 공격, 맴버십 추론 공격, 프롬프트 공격, 추출 공격, 회피 공격 • 탈옥, 프롬프트 프라이밍 등

2. 위험의 내용

가. 데이터 위험

(1) 부적절한 학습데이터

① 학습데이터의 편향·왜곡

학습데이터가 편향되면 AI의 공정성·신뢰성·정확성 등이 훼손될 우려가 있다. 우리는 이미 학습데이터의 편향성으로 인한 AI의 차별·혐오를 경험한 바 있다. 2020년 12월 말 AI 챗봇 서비스 '이루다'가 출시되었으나 차별·혐오 표현이 여과 없이 제시되어 사회적 비난을 받고 출시 20일 만에 서비스를 종료했다. 기계학습에 차별·혐오 표현이 정제되지 않은 학습데이터가 사용된 것이 주요 원인이었다. 이 외에도 학습데이터 자체에는 편향성이 없지만 AI 모델의 목적에 맞지 않는 데이터를 학습한 경우, 적절하게 입력된 데이터가 처리 과정에서 왜곡·손상되는 경우에도 학습데이터의 편향·왜곡이 발생한다.

편향되고 왜곡된 학습데이터는 AI가 학습해야 할 전체 모집단을 충분히 대표하지 못하게 되어, AI가 실제 사실과 다른 판단을 하거나 특정 그룹이나 개인을 부당하게 대표하는 등의 결과를 도출하게 만든다. 이러한 오류와 차별은 그 자체로 AI 모델·시스템의 적절성을 위협하는 문제가 되며, 특히 채용·금융·의료 등 인간의 권리·재산·안전과 관련된 AI에 대해서는 더 큰 위험을 초래할 수 있다.

② 학습데이터의 부족 및 부존재

학습데이터가 충분하지 않거나 없는 경우 AI는 정확하지 않은 것을 마치 정확한 것처럼 대답하는 환각(hallucination) 현상을 보이거나, 일부만 존재하는 학습데이터를 과도하게 부풀려 편향된 결과를 만들어 내는 문제가 발생한다. AI는 대량의 데이터를 기계학습하여 알고리즘에 반영하는 방식이기 때문에 학습하지 않은 내용에 대해서는 결과를 생성하지 못하게 되지만, 현재의 AI 시스템들은 대부분 생성형(generative) 기술을 탑재하고 있어서 실제 사실에 기반하지 않더라도 언어·시각적으로 연계성이 높은 결과들을 조합하여 결과를 제시할 수 있다. 그 결과 학습하지 않은 사항에 대해서도 AI는 그럴듯한 결과를 제시하게 되는데, 그 내용이 실제와 다를 가능성이 있다.

사실적으로 부정확하거나 진실하지 않은 콘텐츠를 생성하는 환각의 경우 이용

자들이 그것을 사실로 믿음으로 인해 혼란이 발생할 우려가 있고, 이것이 방치된다면 여러 시스템과 응용서비스에서 재생산될 문제도 크다. AI 모델이 충분한 정보가 없는 상태에서 의료·조세·법률 등 전문분야에 대해 조언을 제공하는 불완전한 조언도 문제다. 외국에서 개발된 AI는 상대적으로 외국의 학습데이터가 많이 사용되고 우리나라 데이터는 비중이 낮기 때문에 한국적 맥락을 충분히 반영하지 못한 결과를 도출할 우려도 있다.

학습데이터의 부족 문제를 해결하기 위해 AI가 생성한 데이터를 모델 학습에 재투입하는 경우가 있는데, 그 결과 AI 모델의 성능이 급격히 저하되는 모델 붕괴(models collapse) 문제가 발생할 우려도 있다. 생성된 데이터에 내재된 오류들이 반복되는 학습 과정에서 누적되어 결과적으로 모델의 성능을 훼손시킨다는 것이다.

(2) 부적법한 학습데이터

① 개인정보의 동의 없는 활용

개인정보가 동의 없이 학습에 사용될 위험이 있다. AI는 웹 크롤링(web crawling), 웹 스크래핑(web scraping) 등의 방법으로 인터넷에 공개된 데이터를 자동으로 탐색하여 대량으로 수집(text and data mining, TDM)하는데, 이 과정에서 정보 주체에게 사전 통지나 동의를 구하지 않고 개인정보를 포함한 콘텐츠를 무단으로 사용함으로써 개인정보를 침해할 가능성이 존재한다.

② 저작물의 동의 없는 활용

AI 학습을 위하여 저작권이 있는 저작물을 적법한 절차 없이 인터넷에서 TDM 할 경우 저작자의 권리를 침해할 우려가 있다. 우리 「저작권법」 제35조의5가 공정이용을 규정하고 있지만, AI 학습데이터에 동 조항을 적용할 수 있는지에 대해

학계의 의견이 정립되지 않았고, AI 학습에 있어 공정이용 규정을 직접적으로 적용한 국내 법원의 판례가 없다. 따라서 현재 상황에서는 인터넷 홈페이지나 블로그, SNS 등을 통하여 공개된 저작물이라는 사실만으로 공정이용 규정을 들어 저작권자 허락 없이 이용할 수 있다고 보기에는 어려움이 있다. 저작물의 학습데이터 활용이 적법하지 않은 경우 해당 AI 모델과 시스템의 사용을 중단하거나, 저작권자에게 손해배상 등을 해야 하는 부담이 발생할 수 있다.

나. 기술적 위험

(1) 투명성·설명가능성 부족

AI 모델·시스템은 투명성과 설명가능성을 충족하기 쉽지 않다. 세계적으로 다수의 AI 모델은 성능을 높이기 위해 더 많은 데이터를 학습하려고 하는데, 이 많은 데이터가 어떻게 처리되고 알고리즘에 반영되는지를 세밀하게 밝혀서 설명하는 것은 현실적으로 불가능하기 때문이다. AI 모델의 투명성·설명가능성이 충분히 달성되지 못한 상황에서 이용자들이 AI를 사용하다가 AI 모델·시스템에 문제가 발생한다면 책임성 확보를 위한 추적이 쉽지 않다.

AI 사업자 간 과도한 경쟁 상황이 투명성·설명가능성 부족 문제를 암묵적으로 방치하게 하는 원인이 되기도 한다. AI 사업자들 사이의 경쟁이 심화되고 있는 상황에서 AI 모델·시스템에 대한 투명성·설명가능성을 높일 경우, 경쟁 기업에게 중요한 정보를 공개하는 문제가 발생할 수 있다. 신제품 및 서비스의 빠른 출시 압박이 증가하면서 윤리적이고 안전한 AI 시스템 개발에 필요한 투자와 주의가 부족해지고, 신뢰성 검증 및 보장, 내부 작동방식 이해가 충분하지 않은 상태에서 제품과 서비스가 출시될 우려도 있다.

AI의 투명성·설명가능성이 확보되지 못한 상황에서는 AI의 위험을 평가하는 변수, 방법론 등이 완벽하지 않기 때문에 충분한 위험평가가 이루어지지 못하기

도 한다. 위험평가 한계의 주요 원인은 다음과 같다.

[표 2] 위험평가 한계의 주요 원인

구분	내용
모델 투명성 부족 (Lack of model transparency)	모델 설계, 개발 및 평가 프로세스에 대한 문서화가 불충분하고 모델의 내부 작동에 대한 정보가 충분히 제공되지 않을 경우 평가·모델변경·재사용이 어려워질 수 있고, 모델에 대한 신뢰성에 부정적 영향을 미칠 수 있음
테스트 다양성 부족 (Lack of testing diversity)	AI 모델 리스크는 사회 기술적인 문제이므로 다양한 관점에서 테스트를 거치지 않는다면 위험성을 충분히 식별하지 못할 수 있음
시스템 투명성 부족 (Lack of system transparency)	모델의 목적에 대한 문서가 불충분한 경우, 모델이 시스템이나 애플리케이션의 기능에 어떻게 기여하는지 파악하기 어려울 수 있음
대표성 없는 위험 테스트 (Unrepresentative risk testing)	모델이 배포될 때 발생할 다양한 위험을 가정한 위험 테스트를 거치지 않을 경우, 모델이 적용되는 상황을 대표하지 못하므로 위험을 정확하게 반영하지 못할 우려가 있음
불완전한 사용 정의 (Incomplete usage definition)	기반 모델(foundation model)은 다양한 목적으로 사용될 수 있는데, 모델의 용도가 충분히 명시되지 않은 경우 모델의 관련 위험을 정확하게 판단해서 완화하기 어려울 수 있음
데이터 투명성 부족 (Lack of data transparency)	학습데이터 또는 미세조정 데이터에 대한 세부 설명이 부족할 경우, 모델의 위험성, 데이터 적법성 등에 대한 평가를 수행하는 데 한계가 있음
잘못된 위험 테스트 (Incorrect risk testing)	위험을 측정하거나 추적하기 위해 선택한 수단이 적절하지 못할 경우, 위험에 대한 이해가 부정확해지고 도출된 위험 완화 조치가 작동하지 않을 우려가 있음

자료 : IBM, 앞의 글, 재정리.

(2) 자율성에 대한 통제 곤란

AI가 인간의 가치보다 AI 스스로의 가치를 더 높게 설정할 우려가 있다. AI가 스스로 '자기 보존'이나 '자기 개선'과 같은 목표를 추구한다면 인간 중심성과 멀

어지는 문제가 발생한다. 이러한 상황에서 AI 모델·시스템이 점점 더 많은 사회·경제적 시스템에 통합된다면 AI 중심적 접근이 확대되는 문제가 악화될 수 있다.

AI 작동에 대해 인간의 통제력을 발휘하기 어려워지고 있는 것도 문제다. 특히 최근 자율성이 고도로 높아진 AI 에이전트(AI agent)의 개발 속도가 빨라짐에 따라, 인간이 의도하지 않은 AI의 자율적 결과로 위험이 발생할 우려가 크다. AI에 개인의 재산·건강·권리에 관한 사항, 국방에 관한 사항 등 중요한 책임과 의사결정을 맡겼을 때 인간이 과도하게 AI 시스템에 의존함에 따라 통제 및 감독의 어려움이 발생할 가능성도 존재한다.[5)]

(3) 인간의 권리 침해

AI 모델·시스템의 산출물이 인간의 권리를 침해할 우려가 있다. 대표적인 것이 감시 강화 및 사생활 침해이다. AI는 대량의 개인 데이터를 수집하고 이를 분석하여 사람들의 행동·선호도·심리 등을 예측할 수 있는데, 이것이 개인의 동의 없이 이루어질 경우 심각한 프라이버시 침해로 이어진다.

또한 정부나 권력기관에서 개인정보 수집이 발생할 경우 감시사회의 위험이 발생하게 된다. AI를 활용한 감시는 시민들의 일상적인 행동을 추적하고 감시하는 체제로 발전할 수 있으며, 이는 민주주의와 자유를 위협할 수 있다.

다. 이용자 위험

(1) 부주의한 이용자의 위험

이용자가 프롬프트에 적절하지 못한 정보를 입력하여 자신의 중요 정보를 스스로 유출시키는 결과를 초래할 우려가 있다. 모델에 명령·전송하는 프롬프트에

5) 노재원·유재흥·장진철·조지연, 앞의 글.

개인정보 또는 민감정보, 지식재산권(IP)이 보호되는 정보, 회사·기관의 기밀정보를 입력하여 이것이 모델의 출력에 의도치 않게 공개될 수 있기 때문이다.

이용자가 AI 서비스에 대한 인식이 부족하여 AI를 부적절하게 이용하는 문제도 발생한다. 비현실적인 기대로 AI 시스템에 과도하게 의존하거나, 해당 AI 시스템이 충분히 대답하지 못하는 영역에서 AI를 활용하고 그 결과를 신뢰하는 경우가 여기에 해당한다. 예를 들어, 법조계, 의료계 등 전문 분야에서 범용 AI의 능력을 과대평가하여 실무에 적용한다면 적지 않은 오류가 발생할 수 있다.[6]

(2) 악의적 이용자의 위험

AI로 허위 조작 정보를 만들어 개인과 사회를 혼란스럽게 할 우려가 있다. AI 시스템은 사람들의 심리적 프로파일에 기반하여 개별화된 메시지를 생성하고, 이를 통해 사람들을 대규모로 조작하거나 영향을 미칠 수 있기 때문에 허위 조작 정보의 파급력이 크다.

대표적인 문제는 AI 기반 조작을 통해 민주적 과정과 사회적 결속을 훼손할 수 있다는 것이다. 예를 들어, 선거에 영향을 미치거나, 가짜뉴스를 유포하여 신뢰를 무너뜨리는 상황이 발생할 수 있다. 2024년 1월 미국 대선의 뉴햄프셔주 프라이머리(예비경선)를 앞두고 바이든 대통령의 목소리를 흉내낸 딥페이크 음성으로 민주당 당원들에게 투표 거부를 독려하는 전화가 기승을 부렸고, 2023년 말에는 싱가포르 리셴룽 총리가 암호화폐 투자를 조장하는 딥페이크 영상이 확산되는 등 유명인·정치인을 사칭한 딥페이크 범죄가 발생하기도 했다.[7]

이용자를 속여 실제 사람인 것처럼 가짜 신원을 생성, 도용하여 불법적인 목적

6) 노재원·유재흥·장진철·조지연, 앞의 글.

7) 정준화, 「정보통신망 이용촉진 및 정보보호등에 관한 법률 일부개정법률안(의안번호 2126048) 입법영향분석 - 인공지능으로 만든 가상 정보임을 표시하도록 하는 의무」, NARS 입법영향분석 기획보고서, 제7호, 국회입법조사처, 2024.

으로 사용하는 신분 사기 위험도 있다. 생성형 AI로 실제 인물의 목소리나 외모를 모방하여, 그 사람인 것처럼 사칭하고 실시간으로 행동하여 타인을 속이는 악용 방식이 있으며, 행동의 묘사와 유사성으로 청중을 쉽게 오도할 수 있어 큰 피해를 유발한다.[8] 딥페이크도 문제이다. 2024년 8월을 전후로 하여 학생을 대상으로 한 딥페이크 합성음란물이 사회적 문제가 되었다. 그동안 딥페이크에 대해서는 단순한 놀이문화로 생각하는 경향이 많았으나, 이번 딥페이크 사태로 인해 아동·청소년부터 성인까지 손쉽고 빠르게 생성형 AI를 통해 딥페이크 합성음란물을 만들 수 있고, 온라인 플랫폼을 통해 급속히 유포될 수 있는 매우 심각한 성범죄라는 인식이 확산되었다.

(3) 사이버 공격자의 위험

AI는 사이버 공격의 발생 빈도와 심각도를 증가시킬 수 있다. AI는 악의적 행위자들에게 기술장벽을 낮추고, 더 정교한 공격을 실행할 수 있는 역량을 제공한다. 예전에는 인간이 오랜 시간과 노력을 들여야 가능했던 악성 사이버 공격이 AI를 통해 더욱 쉽게 이루어짐으로써 AI를 통해 악성코드, 피싱 이메일 등을 쉽게 생성하여 발송할 수 있다.

한편, AI 기반 시스템 자체가 악성 공격의 대상이 될 가능성도 있다. AI를 사용하여 코딩을 수행할 경우 SW에 상대적으로 더 큰 취약성이 발생할 수 있으며, AI를 적용한 시스템은 기존의 인간 개입 방식보다 사이버 공격에 취약하도록 설계될 우려가 있다.

이 외에도 다음 [표 3]과 같이 프롬프트를 공격하여 모델로부터 개인정보 등을 탈취하거나, 모델의 안전정치를 무력화시키는 경우도 있다.

8) 노재원·유재흥·장진철·조지연, 앞의 글.

[표 3] 프롬프트 공격의 주요 사례

구분	내용
속성 추론 공격 (Attribute inference attack)	악의적 이용자가 특정 속성에 대해 대답하도록 반복적으로 질의·요청하여 모델 학습에 포함된 개인정보 또는 지식재산권과 같은 중요 정보를 탈취하는 것
맴버십 추론 공격 (Membership inference attack)	악의적 이용자가 모델이 어떤 데이터셋을 학습하였는지 알아내기 위해 특정 데이터셋의 샘플을 입력하고 출력값을 관찰하는 것을 반복하여 그 샘플이 어느 학습데이터의 일부(membership)인지 파악하는 것. 이를 통해 경쟁업체의 모델이 어떻게 훈련되었는지 알아내고, 모델을 복제하거나 조작할 수 있는 기회로 악용함
프롬프트 누출 (Prompt leaking)	악의적 이용자가 모델을 공격하여 모델의 기본적인 명령어 체계(시스템 프롬프트)를 찾아내서, 이를 통해 모델에 포함된 중요 정보를 쉽게 탈취할 우려가 있음
프롬프트 입력 공격 (Prompt injection attack)	악의적 이용자가 프롬프트에 포함된 구조, 명령어, 정보 등을 조작하여 모델 개발자가 예상하지 못한 출력을 생성하도록 강제하여 모델 동작을 변경할 우려가 있음
추출 공격 (Extraction attack)	학습데이터에 포함된 개인정보와 같은 특정 정보를 추출하도록 프롬프트를 공격하는 것
회피 공격 (Evasion attack)	프롬프트에 가짜문제를 입력하여 모델이 가짜문제를 회피하는 결정을 하도록 교란하여, 모델 안에 있는 중요한 진짜 정보를 유출하도록 만드는 것
탈옥 (Jailbreaking)	모델이 설정해 놓은 안전장치(가드레일)를 뚫고 금지된 작업의 수행을 시도하는 것으로, 모델의 안전성과 평판에 부정적 영향을 초래함
프롬프트 프라이밍 (Prompt priming)	생성형 모델은 프롬프트 입력값과 유사한 출력값을 생성하는 경향이 있다는 것을 악용하여, 프롬프트에 개인정보 등 민감정보를 입력하여 모델이 그와 유사한 결과를 출력하도록 하여 개인정보 등을 유출하는 행위

자료 : IBM, 앞의 글, 재정리.

Ⅲ. 인공지능의 내재적 위험에 대응하는 국내외 법제도

1. 외국 법제도

가. 미국

(1) 바이든 정부의 인공지능 행정명령

① 개요

바이든 정부는 2023년 10월 30일에 행정명령(Executive Orders) 제14110호로 「안전하고 보안이 보장되며 신뢰할 수 있는 인공지능의 개발과 사용(Safe, Secure and Trustworthy Development and Use of Artificial Intelligence)」(이하 "행정명령 제14110호"라 함)을 발표했다. 행정명령 제14110호는 기업과 국민을 대상으로 하는 '법률'이 아니라 정부 각 부처에 권한과 의무를 부여하는 '행정명령'이다.

행정명령 제14110호의 목적은 AI 위험을 관리하는 데 미국이 주도권을 잡을 수 있도록 보장하는 것으로, AI 활용에 관해 정부 기관이나 민간 기업 등 광범위한 주체들의 기준 준수나 책임 있는 대처를 요구한다. 법률이 아닌 행정명령을 취하게 된 것은 AI에 대해 행정부가 주도권을 갖고, 신속하게 규정을 정비하기 위함이다.[9]

② 주요 내용

주요 내용은 미국의 AI 잠재성을 극대화하면서 동시에 국가와 국민에 대한 AI 위험성을 규제하는 것이다.[10] 첫째, 혁신 및 경쟁의 촉진이다. 정부 부처에 AI 및

9) 이정아, 「미국의 인공지능(AI) 정책·전략 현황과 변화 방향 - AI 지배력 강화와 초강대국 유지를 위해 미국은 어떻게 변화하고 있는가?」, The AI Report, 2024-3호, 한국지능정보사회진흥원, 2024.

기타 중요한 신규 기술 분야 연구자 등 인재 확보를 위한 비자 및 이민 제도 개선, AI 시장의 공정경쟁 보장, 소비자·근로자 보호를 위한 규범 및 규칙 마련, AI 관련 저작권 규범 정립 등의 의무를 부과한다. 둘째, 안전 및 보안의 보장이다. 정부 부처에 AI 안전 및 보안을 위한 지침 마련, AI 안전 취약성 점검, 즉 AI 레드티밍[11] 수행을 위한 지침 수립, 잠재적인 대규모 컴퓨팅 클러스터에 관한 규제 정비, AI 합성 콘텐츠(딥페이크) 탐지·라벨링 제도 마련 등의 의무를 부과한다. 셋째, 근로자에 대한 지원이다. 경제자문위원회장은 AI가 노동시장에 미치는 영향에 관한 보고서를 준비하여 대통령에게 제출하고, 노동부 장관은 AI 도입으로 대체된 근로자를 지원하는 방안을 분석한 보고서를 대통령에게 제출하도록 하는 등의 의무를 부과한다. 넷째, 형평성 및 시민권의 진전이다. 법무부 장관은 형사사법제도에서 AI를 활용하는 보고서를 대통령에게 제출하고, 노동부 장관은 고용 과정에서 AI로 인해 발생할 수 있는 불법적 차별을 방지하기 위한 지침을 공표하는 등의 의무를 부과한다. 다섯째, 소비자·환자·승객·학생 등의 보호이다. 규제기관은 사기, 차별, 개인정보 위협으로부터 미국 소비자를 보호하고 AI 사용으로 인한 위험을 해결하기 위하여 규칙 제정을 고려하며, 소비자 등에게 AI가 적용되는 부분을 규정·지침에 명확히 밝히도록 권장한다. 마지막으로 프라이버시 보호이다. 예산관리국 국장은 연방정부 기관이 보유·구매·조달한 상업이용가능정보(commercially available information, CAI)의 개인식별정보 위험성을 평가하고, 개인식별정보가 포함된 상업이용가능정보의 수집·처리·유지·사용·공유·전파 등과 관련된 기관 표준 및 절차를 평가해야 한다.

---

10) 신용우 외, 「인공지능 윤리와 저작권의 규제체계 연구」, NARS 정책연구용역보고서, 국회입법조사처, 2024.

11) AI 레드티밍(AI red-teaming)은 AI시스템의 결함과 취약성을 식별하기 위한 구조화된 시험을 말하며, AI 시스템의 유해하거나 차별적인 출력, 예상하지 못하거나 바람직하지 아니한 시스템 작동, 한계 또는 시스템의 오용과 관련된 잠재적 위험 등과 같이 AI시스템의 결함과 취약점을 식별하는 것이다.

(2) 미국 콜로라도주 주법

콜로라도주는 2026년 2월 1일부터 발효되는 「콜로라도 인공지능법(Colorado Artificial Intelligence Act, CAIA, SB 24-205)」을 통해 AI의 개발 및 배포에 대한 포괄적인 규제를 마련하였다.[12] 핵심 내용은 고위험 AI 시스템을 규제하는 것이다. 고위험 AI 시스템은 그것이 활용될 경우 교육, 고용, 금융, 필수 정부 서비스, 의료 서비스, 주택, 보험, 법률 서비스 등의 제공 또는 이용에 중대한 영향을 미치는 결정을 하거나, 결정에 중요한 요소가 되는 이른바 '결과적 결정(consequential decisions)' 기능을 하는 시스템을 말한다.

「콜로라도 인공지능법」은 고위험 AI 개발자와 배포자에 대한 의무를 규정한다.[13] 개발자는 고위험 AI 시스템의 사용과 관련된 예측 가능한 위험과 그 용도를 설명하는 문서를 제공해야 하며, 알고리즘 차별의 위험을 관리하는 방법도 설명해야 한다. 설명 문서에는 시스템의 성능을 평가한 방법, 알고리즘 차별의 영향을 완화하기 위해 취한 조치, 데이터 관리 조치(데이터 소스의 적합성, 편향 가능성 및 적절한 완화를 검토하는 데 사용된 조치 포함), 시스템의 의도된 결과물 내용, 개인이 시스템을 사용하거나 모니터링하는 방법 등이 포함되어야 한다.

배포자는 소비자에게 AI 시스템의 사용 여부 등을 사전에 통지하고, AI가 불리한 결정을 내린 경우 그 이유를 설명하며, 소비자에게 잘못된 데이터를 수정할 수 있는 기회를 제공해야 한다. 배포자는 현재 배포된 고위험 AI 시스템의 유형, 발생할 수 있는 알고리즘 차별의 위험을 관리하는 방법, 배포자가 AI 시스템과 관련하여 수집하고 사용하는 정보의 성격·출처·범위 등을 사전에 설명해야 한다. 고위험 AI 시스템이 소비자에게 불리한 결정을 내린 경우, 배포자는 소비자

12) Stuart D. Levi et al, "Colorado's Landmark AI Act: What Companies Need To Know", Skadden Publication, 2024. 6. 24.; 미국 콜로라도주 의회 홈페이지, (최종방문일: 2024.12.16.), <https://leg.colorado.gov/bills/sb24-205>

13) 신용우 외, 앞의 글.

에게 결과적 결정의 이유, 고위험 AI 시스템이 해당 결정에 기여한 정도, 시스템에서 처리한 데이터의 유형 및 해당 데이터의 출처 등을 설명해야 한다. 소비자와 상호 작용하는 AI 시스템을 제공하는 배포자는 합리적인 사람이 알 수 있는 경우가 아니라면 소비자에게 AI 시스템과 상호 작용하고 있다는 사실을 공개해야 할 의무를 부과한다. 한편, 개발자와 배포자 모두 고위험 AI 시스템의 사용으로 연령, 피부색, 장애, 민족, 성별 등에 대해서 알고리즘적 차별이 발생하지 않도록 주의를 기울여야 한다.

(3) 국립표준기술연구소의 AI 위험관리 프레임워크(AI RMF)

미국 국립표준기술연구소(National Institute of Standards and Technology, NIST)는 2023년 1월 26일 AI 시스템을 설계·개발·도입하는 조직을 위한 자발적 활용 가이드 문서인 'AI 위험관리 프레임워크(AI Risk Management Framework, AI RMF) 1.0'을 발표했다. AI RMF의 목적은 모든 분야·규모의 기업·조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공하여 AI 기술의 이점을 극대화하고 동시에 개인·그룹·조직·사회에 부정적인 영향을 미칠 가능성을 줄이기 위한 것이다.[14]

AI RMF는 AI 시스템의 설계, 개발, 활용, 테스트, 평가 등 모든 단계에 적용되는 신뢰할 수 있는 AI 시스템의 7가지 기준으로 다음 [표 4]와 같이 유효성 및 신뢰성, 안전성, 보안 및 탄력성 등을 제시한다.

14) 김태순, 「美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」분석 및 시사점」, THE AI REPORT, 2023-7호, 한국지능정보사회진흥원, 2023.

[표 4] 미국 NIST의 신뢰할 수 있는 AI 시스템의 7가지 기준(AI RMF)

구분	내용
유효성 및 신뢰성	배포된 AI 시스템의 유효성과 신뢰성은 시스템이 용도에 따라 작동하는지 확인하는 테스트 또는 모니터링을 통해 지속적으로 평가되며 유효성, 정확성, 견고성, 신뢰성을 측정하여 신뢰할 수 있는 AI 시스템을 구축할 수 있음 • (유효성 검사) 객관적인 증거를 통해 특정 용도 또는 응용 프로그램에 대한 요구 사항이 충족되었음을 확인하는 것 • (신뢰성) 주어진 조건 하에 주어진 시간 동안 실패 없이 필요에 따라 수행할 수 있는 능력 • (정확성) 참값 또는 참이라고 인정되는 값에 대한 관찰, 계산 또는 추정의 근접성 • (견고성 = 일반화 가능성) 다양한 상황에서 성능 수준을 유지하는 시스템 능력
안전성	AI 시스템은 정의된 조건에 따라 작동할 때 인간의 생명, 건강, 재산 또는 환경에 유해한 영향을 미치지 않아야 하며 AI 시스템의 안전성은 다음의 조치를 통해 개선 가능 • 책임감 있는 설계, 개발 및 배포 기준 • 책임감 있는 시스템 사용에 대해 배포자에게 명확한 정보 제공 • 배포자 및 최종 사용자의 책임감 있는 의사 결정 • 사건의 실증적 증거를 기반으로 위험 설명 및 문서화
보안 및 탄력성	AI 시스템 및 시스템이 배포된 생태계는 예상치 못한 이상 반응 또는 변화를 견딜 수 있는 경우, 내부/외부적 변화에도 불구하고 그 기능과 구조를 유지하며, 필요한 경우 안전하게 작동하면서 기능을 제한적으로 제공할 수 있는 탄력성이 있어야 함
책임 및 투명성	책임감은 투명성을 전제로 하며 부정확하거나 부정적인 영향을 초래하는 AI 시스템 결과를 시정하기 위해서 필요 • AI 주기 단계를 기반으로 AI 시스템과 상호 작용하거나 이를 사용하는 AI 행위자 또는 개인의 역할이나 정보에 맞게 조정된 정보에 대한 액세스를 제공하는 투명성은 더 높은 수준의 이해를 촉진하며 AI 시스템의 신뢰도를 높임 • 책임 및 투명성의 범위는 설계 의사결정, 훈련 데이터, 모델 훈련, 모델 구조, 의도한 사용 사례, 배포 시 또는 배포 후 최종 사용자의 의사 결정이 언제, 어떻게, 누구에 의해 이루어졌는지까지 다양함 • 투명성을 위해서는 AI 시스템으로 인한 잠재적/실제적 악영향이 감지될 때 운영자 또는 사용자에게 이를 알리는 방법 등 인간-AI의 상호 작용을 고려해야 함 • AI 시스템에 대한 위험과 책임은 문화적, 법적, 부문별 및 사회적 관점에 따라 크게 달라지며 심각한 결과가 예상되는 경우 AI 개발자 및 배포자는 투명성 및 책임 기준을 비례적으로 사전에 미리 조정할 수 있어야 함

구분	내용
	• AI 시스템 개발자는 해당 시스템이 의도한 대로 사용되는지 확인하기 위해 AI 배포자와 협력하여 다양한 유형의 투명성 도구를 테스트하는 것이 필요
설명 및 해석 가능성	AI 시스템 작동의 기본 메커니즘을 나타내는 '설명가능성'과 기능적 목적 관점에서 AI 시스템의 결과를 나타내는 '해석가능성'의 특성이 필요 • 설명가능성이 부족하여 발생하는 위험은 개인별 설명(예: 사용자의 역할, 지식 및 기술 수준에 따라 조정된 설명)과 함께 AI 시스템이 어떻게 작동되는지 설명함으로써 관리 가능 • 해석 가능성이 부족하여 발생하는 위험은 AI 시스템이 특정 예측이나 권고를 한 이유에 대해 설명함으로써 해결 가능
개인정보 보호 강화	AI 시스템을 설계, 개발 및 배포 시 익명성, 기밀성 및 제어 등 개인정보보호 가치 고려 필요 • 개인정보보호 관련 위험은 보안, 편향성, 투명성에 영향을 미칠 수 있으며, 안전 및 보안과 마찬가지로 특정 기술적 기능이 개인정보보호를 감소시킬 수 있음 • AI 시스템 설계 시 모델 결과값에 대한 비식별화 및 집계 등의 데이터 최소화 방법과 더불어 AI용 개인정보 보호 강화 기술(Privacy-enhancing technologies, PET)은 개인정보보호 강화에 도움
공정성 (유해한 편향 관리)	3가지 범주의 편향(시스템적 편향, 통계적 편향, 인간 인지적 편향)을 고려하고 관리해야 함 • 시스템적 편향 : AI 데이터셋, AI 주기 전반에 걸친 조직적 규범, 수행 기준 및 절차, AI 시스템을 사용하는 광범위한 사회에 존재 가능 • 통계적 편향 : AI 데이터셋 및 알고리즘 프로세스에 존재할 수 있으며, 이는 대표성이 없는 샘플의 체계적 오류로 인해 종종 발생 • 인간 인지적 편향 : 개인 또는 집단이 AI 시스템 정보를 인지하여 결정을 내리거나 누락된 정보를 채워 넣는 방법 또는 인간이 AI 시스템의 목적 및 기능에 대해 사고하는 방법과 관련이 있으며 AI의 설계, 구현, 운영 및 유지 관리를 포함하여 AI 주기 및 시스템 사용 전반에 걸친 의사 결정 프로세스에 편재

자료: 김태순, 앞의 글, 재정리.

나. 유럽연합

(1) 개요

세계 최초의 종합적인 AI 규제법인 유럽연합(EU)의 「인공지능법(Artificial

Intelligence Act)」이 2024년 8월 1일 발효되었다. 「인공지능법」은 2021년 4월 21일 EU 집행위원회의 법안 제안, 2023년 6월 14일 유럽의회의 수정안 제안, 2024년 2월 13일 유럽 이사회의 수정안 합의, 2024년 3월 13일 유럽의회의 합의안 채택, 2024년 5월 21일 유럽 이사회의 최종 승인을 거쳐 2024년 8월 1일 발효되었다. 제안 당시에는 금지된 AI, 고위험 AI 등에 대한 규제만 포함되어 있었는데, 생성형 AI의 위험이 확산되자 2023년에 범용 AI에 관한 규정이 추가되어 2024년 8월에 최종 발효된 것이다.

법률의 제2장(금지된 AI)은 발표일로부터 6개월 이내에 시행되고, 제5장(범용 AI 모델)과 제7장(거버넌스)은 12개월 이내에 시행되며, 제3장(고위험 AI 시스템)과 제4장(특정 AI 시스템)을 포함한 나머지 규정들은 24개월 이내에 시행된다. 부속서(Annex) Ⅰ에 규정된 고위험 AI 시스템에 관한 사항은 36개월 이내에 시행된다.

[표 5] EU 인공지능법의 구조 및 개관

<table>
<tr><th>장</th><th colspan="2">제목</th><th>조항</th><th>시행일(24년 8월 1일 기준)</th></tr>
<tr><td>제1장</td><td colspan="2">총칙</td><td>제1~4조</td><td rowspan="2">6개월 후</td></tr>
<tr><td>제2장</td><td colspan="2">금지된 AI 행위</td><td>제5조</td></tr>
<tr><td rowspan="5">제3장</td><td rowspan="5">고위험 AI 시스템</td><td>제1절 고위험 AI 시스템의 분류</td><td>제6~7조</td><td>24개월 후
제6조제1항(고위험 AI시스템 분류, 의무)은 36개월 후</td></tr>
<tr><td>제2절 고위험 AI 시스템에 대한 요건</td><td>제8~15조</td><td rowspan="2">24개월 후</td></tr>
<tr><td>제3절 고위험 AI 시스템 제공자, 배포자 및 기타 당사자의 의무</td><td>제16~27조</td></tr>
<tr><td>제4절 통보기관 및 인증기관</td><td>제28~39조</td><td>12개월 후</td></tr>
<tr><td>제5절 표준, 적합성 평가, 인증서, 등록</td><td>제40~49조</td><td rowspan="2">24개월 후</td></tr>
<tr><td>제4장</td><td colspan="2">특정 AI 시스템 제공자 및 배포자에 대한 투명성 의무</td><td>제50조</td></tr>
</table>

장	제목		조항	시행일(24년 8월 1일 기준)
제5장	범용 AI 모델	제1절 분류 규칙	제51~52조	12개월 후
		제2절 범용 AI 모델 제공자의 의무	제53~54조	
		제3절 시스템적 위험이 있는 범용 AI 모델 제공자의 의무	제55조	
		제4절 실천 강령	제56조	
제6장	혁신 지원 방안		제57~63조	24개월 후
제7장	거버넌스	제1절 EU 수준에서의 거버넌스	제64~69조	12개월 후
		제2절 국가 관할기관	제70조	
제8장	고위험 AI 시스템을 위한 EU 데이터베이스		제71조	24개월 후
제9장	출시 후 모니터링, 정보 공유, 시장감독	제1절 출시 후 모니터링	제72조	24개월 후
		제2절 중대 사건에 관한 정보의 공유	제73조	
		제3절 집행	제74~84조	
		제4절 구제수단	제85~87조	
		제5절 범용 AI 모델 제공자에 대한 감독, 조사, 집행 및 감시	제88~94조	
제10장	행동강령 및 지침		제95~96조	
제11장	권한의 위임 및 위원회의 절차		제97~98조	
제12장	벌칙		제99~101조	제99~100조는 12개월 후 제101조(범용 AI 모델 제공자에 대한 벌금)는 24개월 후
제13장	최종 조항		제102~113조	24개월 후
13개의 부속서(I~XIII)				부속서 언급 조항에 따른 시행

자료: 신용우 외, 앞의 글.

(2) 주요 내용

「인공지능법」의 특징은 AI를 위험(risk) 크기에 따라 금지된 AI, 고위험 AI, 특정 AI(제한적 위험 AI), 최소 위험 AI로 분류한 것이다. 다만, 최소 위험 AI에 대해서는 별도의 명시적인 규정을 두고 있지 않다. 또한 생성형 AI를 포함한 범용 AI(General Purpose AI, GPAI)에 대한 규제도 별도로 마련하고 있다.

① 금지된 AI 행위

EU가 보호하는 기본권과 가치를 위협하는 AI는 원칙적으로 금지한다. 금지된 AI는 ①인간의 잠재의식 또는 특정 집단의 취약점을 악용하여 피해를 유발할 수 있는 시스템, ②사회적 행동이나 개인의 특성에 기반하여 생성·수집된 정보로 개인이나 집단의 '사회적 점수(social score)'를 도출하여 불리한 대우를 유발할 수 있는 시스템, ③프로파일링이나 성격 특성만을 토대로 개인의 범죄 가능성을 예측·평가하는 시스템, ④직장 및 교육기관에서 개인의 감정 추론을 위해 사용하는 시스템, ⑤공공장소에서의 실시간 원격 생체 인식 시스템(납치·인신매매·성적착취피해자·실종자 수색, 테러 위협 방지, 범죄자 및 범죄 용의자 신원 및 위치 파악을 위해 예외적으로 허용됨) 등이다.[15]

② 고위험 AI 시스템

EU 인공지능법은 가장 많은 규정을 고위험 AI 시스템에 할애하고 있다. 고위험 AI 시스템은 개인의 보건·안전·기본권 등에 중대하고도 유해한 영향을 미칠 수 있는 것을 말한다. 구체인 내용은 부속서에 나열되어 있다. 부속서 Ⅰ(Annex Ⅰ)에서는 EU 조화법 목록(List of Union harmonisation legislation)이 제시되어

15) 심소연, 「규제중심의 유럽연합 인공지능법(EU AI Act)」, 최신외국입법정보 통권 제242호, 국회도서관, 2024.

있는데, 여기에 포함된 기계, 장남감, 승강기, 차량 등의 안전 구성요소로 작용되는 AI 시스템은 고위험 AI로 본다. 부속서 Ⅲ(Annex Ⅲ)에서는 주로 인간의 기본권에 영향을 미칠 수 있는 8가지 목적(ⓐ금지되지 않는 생체인식·분류, ⓑ중요 인프라의 관리·운영, ⓒ교육 및 직업훈련, ⓓ고용, 근로자 관리 및 자영업, ⓔ필수 민간서비스 및 공공서비스에 대한 접근 및 혜택의 향유, ⓕ법 집행, ⓖ이민, 망명 및 국경 통제 관리, ⓗ사법행정 및 민주적 절차)의 AI 시스템을 고위험 시스템으로 제시한다.

고위험 AI 시스템에 대한 규제는 다음 [표 6]과 같이 공급자에 대한 규제와 배포자에 대한 규제로 구분된다. 공급자는 데이터 관리, 기술문서 작성, 배포자에 이용지침 제공, 적합성 평가 수행 등의 의무를 이행해야 한다. 배포자는 이용지침 준수, 기본권 영향평가 등의 의무가 있다.

[표 6] 고위험 AI 시스템에 대한 규제

구분	내용	
공급자 (provider)	위험관리체계 구축	고위험 AI 생애주기에 걸쳐 위험의 식별·분석, 추정, 평가, 관리 등에 관한 체계를 마련해야 함
	데이터 관리	고위험 AI의 학습·검증·테스트에 사용되는 데이터는 대표성, 정확성, 합목적성 등을 준수할 수 있도록 관리 및 활용되어야 함
	기술문서 작성	고위험 AI 시스템 출시 및 서비스 전에 고위험 AI 시스템이 준수사항을 충족하는 것을 입증하고, 국가관할당국 등에 준수사항 충족을 평가하는 데 필요한 정보(일반적 기술, AI 시스템 개발 절차, 모니터링에 관한 사항 등)를 제공할 수 있도록 기술문서를 작성해야 함
	배포자에 이용지침 제공	배포자가 고위험 AI 시스템을 접근 및 이해할 수 있는 정보를 포함한 이용지침(instructions for use)을 전자 형식으로 제공해야 함
	인간의 감독	고위험 AI 시스템은 상당한 수준의 자율성을 가지고 있지만, 의도된 목적으로 이용되고 인간의 안전과 기본권에 대한 위험

구분	내용	
		을 초래하지 않도록 인간이 실효적으로 감독할 수 있도록 설계 및 개발해야 함
	보안	고위험 AI 시스템은 정확성(accuracy), 견고성(robustness), 사이버 보안(cybersecurity)을 적절한 수준을 확보할 수 있도록 개발·설계되어야 하고, 오작동·오류와 시스템 취약성 이용 공격에 대한 복원력(resilience)을 가질 수 있도록 해야 함
	적합성 평가	고위험 AI 시스템 공급자는 고위험 AI 시스템을 제공·판매하기 전에 AI법에 명시된 요건을 충족하는지 여부를 평가하고, 그 결과를 EU 데이터베이스에 등록해야 함. 향후 관련 규정을 준수하는 것에 대해 책임을 지겠다는 적합성 선언을 서면으로 작성하여 보관해야 함
	품질관리체계 마련 및 이행	고위험 AI 공급자가 준수하고자 하는 데이터, 위험관리체계 등을 포함하는 품질관리체계(Quality Management System)를 서면 형태로 마련하고 이행해야 함
배포자 (deployer)	이용지침 준수	공급자가 제시한 이용지침(instructions for use)에 따른 이용 보장을 위한 기술적·조직적 조치를 시행해야 함
	인간의 감독	역량을 갖추고, 교육을 받았으며, 권한을 갖고, 필요한 지원을 할 수 있는 자연인에게 인간 관리·감독 역할을 배정해야 함
	데이터의 대표성 보장	배포자가 사용한 데이터는 AI 모델과의 목적 관련성이 있어야 하고, 충분한 대표성이 보장되어야 함
	기본권 영향평가	공법(Public Law)의 적용을 받는 공공기관인 배포자, 공공서비스를 제공하는 민간기관인 배포자, 신용도 내지 생명·건강보험 AI 시스템의 배포자 등은 고위험 AI 시스템을 배포하기 전에 AI 시스템이 인간의 기본권에 미치는 영향을 평가해서 결과를 당국에 통지해야 함

③ 특정 AI 시스템

특정 AI 또는 제한적 위험 AI는 사람과 상호작용하는 AI 시스템으로서 사람에 대한 기만·조작 등의 문제가 발생할 수 있는 것을 의미한다. 이 중에서 인간의

의사결정에 실질적 영향을 주지 않고, 자연인의 건강·안전·기본권 등에 중대한 위해를 가할 위험이 없는 특정 AI에 대해서는 해당 작용·처리가 AI에 의해 이루어지는 것임을 알려 상대방이 AI 시스템과 상호작용하고 있다는 사실을 인식할 수 있도록 할 의무를 부과한다.

합성 콘텐츠(오디오, 이미지, 동영상, 텍스트 등) 생성하는 AI 시스템은 AI로 만든 것임을 표시해야 한다. 세부적으로 공급자는 결과물을 기계판독이 가능한 형태로 표시하여 이용자가 인공적 생성이나 조작을 판별할 수 있도록 해야 하고, 배포자는 해당 콘텐츠가 인위적으로 만들어진 것임을 밝혀야 한다.

④ 범용 AI

유럽연합 「인공지능법」은 위험기반 규제와는 별도로 챗지피티의 기반이 되는 지피티(GPT) 모델과 같은 범용 AI(General Purpose AI, GPAI)에 대한 규제를 규정한다. 다양한 AI 시스템 개발의 기초로 쓰일 수 있는 기반 모델(Foundation Model), 또는 거대언어모델(Large Langue Model, LLM) 등이 GPAI에 해당하며, GPAI 공급자에게 기술문서와 사용설명서 제공, 저작권 지침 준수, 학습에 사용된 데이터의 요약본 공개 등의 의무를 부과한다.

시스템적 위험(Systemic risk)이 있는 GPAI[16]는 추가로 최신 기술이 반영된 평가를 수행하여 그 결과를 문서화하고, 시스템적 위험을 평가하여 이를 완화하고, 중대 사고 발생시 이를 감독 당국에 보고하고, 적절한 수준의 사이버 보안을 보장하도록 하는 의무를 부과한다.

16) EU 사회·경제·안보 시스템에 실질적으로 또는 합리적으로 예측 가능한 부정적인 영향을 미치는 위험을 '시스템적 위험'이라고 하며, EU 인공지능법은 학습에 사용된 누적 컴퓨팅이 10^25 FLOPs보다 큰 '고영향 역량(high impact capabilities)'을 가진 것을 시스템적 위험이 있는 GPAI로 본다.

[표 7] 범용 AI(GPAI)에 대한 규제

구분	내용	
일반 GPAI 의무	고영향 성능 통지	고영향 성능이 있으면 2주 이내에 EU 집행위원회에 통지해야 함
	기술문서 작성	기술문서를 작성하여 AI 사무국 및 관할 당국이 요청할 경우 제공해야 함
	정보제공	자신의 AI 시스템에 GPAI 모델을 통합하려는 시스템 공급자에게 해당 GPAI의 성능과 한계를 충분히 이해하고 EU 인공지능법에 따른 의무 준수에 필요한 정보 및 문서를 제공해야 함
	저작권법 준수 정책 마련	「EU 저작권 지침」을 준수해야 하며, 특히 제4조3항의 '명시적 권리 유보'를 파악하여 권리자가 온라인에 공개된 정보의 학습데이터 이용(TDM)을 거부하는 옵트아웃(opt-out)을 선택한 경우 이를 준수해야 함
	학습콘텐츠 상세 요약서 작성	AI 사무국이 제공한 양식에 따라 범용 인공지능 모델 학습 콘텐츠에 관한 상세 요약서 작성하여 공개(publicly available) 해야 함
시스템적 위험이 있는 GPAI에 대한 추가적 의무	최신기술로 모델 평가	시스템적 위험의 식별·완화를 위하여 적대적 테스트(red-teaming) 결과 등 최신 기술을 반영한 표준화된 프로토콜 및 도구로 모델을 평가해야 함
	EU 차원의 위험 관리	유럽연합 차원에서 시스템적 위험을 평가하고 완화하는 노력을 해야 함
	중대한 사고 대응	사람의 사망 또는 건강에 대한 중대한 피해, 중요 기반시설의 관리·운영에 중대하고 돌이킬 수 없는 장에 등 중대한 사고(serious incident) 및 시정조치 관련 정보는 문서화해야 함
	사이버 보안	모델과 해당 모델의 물리적 인프라의 사이버 보안 보장하기 위해 노력해야 함

⑤ 벌칙

위험관리 의무를 위반할 경우 고액의 벌금 처벌을 받게 된다. 금지된 AI 관련 의무를 위반할 경우 3,500만 유로 또는 직전 회계연도의 전 세계 총 연간 매출액

의 최대 7퍼센트 이하 금액 중 더 높은 금액을 부과한다. 고위험 AI 및 특정 AI 시스템 관련 의무 위반할 경우 1,500만 유로 또는 직전 회계연도의 전 세계 총 연간 매출액의 최대 3퍼센트 이하 금액 중 더 높은 금액을 부과한다. 인증기관이나 국가 관할기관의 요청에 부정확·불완전·허위 정보를 제공한 경우 750만 유로 또는 직전 회계연도의 전 세계 총 연간 매출액의 최대 1퍼센트 이하 금액 중 더 높은 금액을 부과한다. GPAI 모델 관련 의무 위반할 경우 1,500만 유로 또는 직전 회계연도 전 세계 총 연간 매출액의 3퍼센트 중 높은 금액을 부과한다.

2. 국내 법제도

가. 인공지능기본법안

(1) 개요

제22대 국회에서 발의된 19개 법률안[17]이 「인공지능 발전과 신뢰 기반 조성

17) 법률안 발의 순서대로 살펴보면 「인공지능 산업 육성 및 신뢰 확보에 관한 법률안」(2200053, 안철수의원), 「인공지능 발전과 신뢰 기반 조성 등에 관한 법률안」(2200543, 정점식의원), 「인공지능산업 육성 및 신뢰 확보에 관한 법률안」(2200673, 조인철의원), 「인공지능산업 육성 및 신뢰 확보에 관한 법률안」(2200675, 김성원의원), 「인공지능기술 기본법안」(2201158, 민형배의원), 「인공지능 개발 및 이용 등에 관한 법률안」(2201399, 권칠승의원), 「인공지능 기본법안」(2203072, 한민수의원), 「인공지능책임법안」(2203235, 황희의원), 「인공지능 발전 진흥과 사회적 책임에 관한 법률안」(2203297, 배준영의원), 「인공지능의 발전과 안전성 확보 등에 관한 법률안」(2203960, 이훈기의원), 「인공지능산업 진흥 및 신뢰 확보 등에 관한 특별법안」(2204250, 김우영의원), 「인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률안」(2205013, 이정헌의원), 「인공지능 발전 및 신뢰성 보장을 위한 법률안」(2205189, 황정아의원), 「인공지능산업 진흥 및 인공지능 이용 등에 관한 법률안」(2205457, 이해민의원), 「인공지능산업 진흥에 관한 법률안」(2205458, 정동영의원), 「인공지능 산업 육성 및 신뢰 기반 조성 등에 관한 법률안」(2205569, 최민희의원), 「인공지능의 안전 및 신뢰 기반 조성에 관한 법률안」(2205639, 조승래·이인선의원), 「인공지능 진흥에 관한 법률안」(2205643, 조승래·이인선의원), 「인공지능산업 육성 및 발전

등에 관한 기본법안(대안)」으로 병합되어 2024년 12월 26일 국회 본회의를 통과했고, 공포 후 1년 뒤에 시행될 예정이다. 인공지능기본법안은 5장 43조로 구성되어 있다. 이 중에서 AI 안전에 관한 내용은 다음과 같다.[18]

(2) 고영향 인공지능 규제

유럽연합이 위험(risk) 기반 규제를 하는 것과 달리, 인공지능기본법안은 '고영향 인공지능'을 구분하여 영향(impact) 기반 규제를 실시한다. 고영향 AI는 「에너지법」 제2조제1호에 따른 에너지의 공급, 「먹는물관리법」 제3조제1호에 따른 먹는물의 생산 공정, 「보건의료기본법」 제3조제1호에 따른 보건의료의 제공·이용체계의 구축 및 운영 등의 영역에서 활용되는 AI 시스템을 말한다.

고영향 AI 사업자는 위험관리방안을 수립·운영하고, 기술과 데이터에 대한 설명방안을 수립하며, 사람에 의한 관리·감독 방안을 마련하고, 인간의 기본권에 미치는 영향을 평가하기 위해 노력해야 한다. AI 사업자는 AI 또는 이를 이용한 제품·서비스를 제공하는 경우 그 AI가 고영향 AI에 해당하는지에 대하여 사전에 검토하여야 하며, 필요한 경우 과학기술정보통신부장관에게 고영향 AI에 해당하는지 여부의 확인을 요청할 수 있다.

등에 관한 법률안」(2205648, 정희용의원)이다.

18) 이 외에 AI 발전(육성)에 관한 사항으로는 제13조(인공지능기술 개발 및 안전한 이용 지원), 제14조(인공지능기술의 표준화), 제15조(인공지능 학습용데이터 관련 시책의 수립 등), 제16조(인공지능기술 도입·활용 지원), 제17조(중소기업등을 위한 특별지원), 제18조(창업의 활성화), 제19조(인공지능 융합의 촉진), 제20조(제도개선 등), 제21조(전문인력의 확보), 제22조(국제협력 및 해외시장 진출의 지원), 제23조(인공지능집적단지 지정 등), 제24조(인공지능 실증기반 조성 등), 제25조(인공지능 데이터센터 관련 시책의 추진 등), 제26조(한국인공지능진흥협회의 설립), 제37조(인공지능산업의 진흥을 위한 재원의 확충 등)이 있다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.

4. “고영향 인공지능”이란 사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로서 다음 각 목의 어느 하나의 영역에서 활용되는 인공지능시스템을 말한다.
   가. 「에너지법」 제2조제1호에 따른 에너지의 공급
   나. 「먹는물관리법」 제3조제1호에 따른 먹는물의 생산 공정
   다. 「보건의료기본법」 제3조제1호에 따른 보건의료의 제공·이용체계의 구축 및 운영
   라. 「의료기기법」 제2조제1항에 따른 의료기기 및 「디지털의료제품법」 제2조제2호에 따른 디지털의료기기의 개발 및 이용
   마. 「원자력시설 등의 방호 및 방사능 방재 대책법」 제2조제1항제1호 및 제2호에 따른 핵물질과 원자력시설의 안전한 관리 및 운영
   바. 범죄 수사나 체포 업무를 위한 생체인식정보(얼굴·지문·홍채 및 손바닥 정맥 등 개인을 식별할 수 있는 신체적·생리적·행동적 특징에 관한 개인정보를 말한다)의 분석·활용
   사. 채용, 대출 심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가
   아. 「교통안전법」 제2조제1호부터 제3호까지에 따른 교통수단, 교통시설, 교통체계의 주요한 작동 및 운영
   자. 공공서비스 제공에 필요한 자격 확인, 결정 또는 비용징수 등 국민에게 영향을 미치는 국가, 지방자치단체, 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관 등(이하 “국가기관등”이라 한다)의 의사결정
   차. 「교육기본법」 제9조제1항에 따른 유아교육·초등교육 및 중등교육에서의 학생 평가
   카. 그 밖에 사람의 생명·신체의 안전 및 기본권 보호에 중대한 영향을 미치는 영역으로서 대통령령으로 정하는 영역

제33조(고영향 인공지능의 확인) ① 인공지능사업자는 인공지능 또는 이를 이용한 제품·서비스를 제공하는 경우 그 인공지능이 고영향 인공지능에 해당하는지에 대하여 사전에 검토하여야 하며, 필요한 경우 과학기술정보통신부장관에게 고영향 인공지능에 해당하는지 여부의 확인을 요청할 수 있다.

② 과학기술정보통신부장관은 제1항에 따른 요청이 있는 경우 고영향 인공지능 해당 여부를 확인하여야 하며, 필요한 경우 전문위원회를 설치하여 관련 자문을 받을 수 있다.

③ 과학기술정보통신부장관은 고영향 인공지능의 기준과 예시 등에 관한 가이드라인을 수립하여 보급할 수 있다.

④ 그 밖에 제1항에 따른 확인 절차 등에 관하여 필요한 사항은 대통령령으로 정한다.

제34조(고영향 인공지능과 관련한 사업자의 책무) ① 인공지능사업자는 고영향 인공지능 또는 이를 이용한 제품·서비스를 제공하는 경우 고영향 인공지능의 안전성·신뢰성을 확보하기 위하여 다음 각 호의 내용을 포함하는 조치를 대통령령으로 정하는 바에 따라 이행하여야 한다.

1. 위험관리방안의 수립·운영
2. 기술적으로 가능한 범위 내에서의 인공지능이 도출한 최종결과, 인공지능의 최종결과 도출

에 활용된 주요 기준, 인공지능의 개발·활용에 사용된 학습용데이터의 개요 등에 대한 설명 방안의 수립·시행
3. 이용자 보호 방안의 수립·운영
4. 고영향 인공지능에 대한 사람의 관리·감독
5. 안전성·신뢰성 확보를 위한 조치의 내용을 확인할 수 있는 문서의 작성과 보관
6. 그 밖에 고영향 인공지능의 안전성·신뢰성 확보를 위하여 위원회에서 심의·의결된 사항

② 과학기술정보통신부장관은 제1항 각 호에 따른 조치의 구체적인 사항을 정하여 고시하고, 인공지능사업자에게 이를 준수하도록 권고할 수 있다.
③ 인공지능사업자가 다른 법령에 따라 제1항 각 호에 준하는 조치를 대통령령으로 정하는 바에 따라 이행한 경우에는 제1항에 따른 조치를 이행한 것으로 본다.

제35조(인공지능 영향평가) ① 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다.
② 국가기관등이 고영향 인공지능을 이용한 제품 또는 서비스를 이용하려는 경우에는 영향평가를 실시한 제품 또는 서비스를 우선적으로 고려하여야 한다.
③ 그 밖에 영향평가의 구체적인 내용·방법 등에 관하여 필요한 사항은 대통령령으로 정한다

(3) AI 윤리

정부는 인공지능윤리의 확산을 위하여 AI 윤리원칙을 대통령령으로 정하는 바에 따라 제정·공표하고, 과학기술정보통신부장관은 사회 각계의 의견을 수렴하여 윤리원칙이 AI의 개발·활용 등에 관여하는 모든 사람에 의해 실현될 수 있도록 실천방안을 수립하고 이를 공개 및 홍보·교육하도록 한다. AI 기술 연구·개발 기관, AI 사업자 등은 AI 윤리원칙을 준수하기 위하여 기관 내부에 민간자율 인공지능윤리위원회를 설치할 수 있다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.
11. "인공지능윤리"란 인간의 존엄성에 대한 존중을 기초로 하여, 국민의 권익과 생명·재산을 보호할 수 있는 안전하고 신뢰할 수 있는 인공지능사회를 구현하기 위하여 인공지능의 개발, 제공 및 이용 등 모든 영역에서 사회구성원이 지켜야 할 윤리적 기준을 말한다.

제27조(인공지능 윤리원칙 등) ① 정부는 인공지능윤리의 확산을 위하여 다음 각 호의 사항을 포함하는 인공지능 윤리원칙(이하 "윤리원칙"이라 한다)을 대통령령으로 정하는 바에 따라 제정·공표할 수 있다.

1. 인공지능의 개발·활용 등의 과정에서 사람의 생명과 신체, 정신적 건강 등에 해가 되지 않도록 하는 안전성과 신뢰성에 관한 사항
2. 인공지능기술이 적용된 제품·서비스 등을 모든 사람이 자유롭고 편리하게 이용할 수 있는 접근성에 관한 사항
3. 사람의 삶과 번영에의 공헌을 위한 인공지능의 개발·활용 등에 관한 사항

② 과학기술정보통신부장관은 사회 각계의 의견을 수렴하여 윤리원칙이 인공지능의 개발·활용 등에 관여하는 모든 사람에 의해 실현될 수 있도록 실천방안을 수립하고 이를 공개 및 홍보·교육하여야 한다.

③ 중앙행정기관 또는 지방자치단체의 장이 인공지능윤리기준(그 명칭 및 형태를 불문하고 인공지능윤리에 관한 법령, 기준, 지침, 가이드라인 등을 말한다)을 제정하거나 개정하는 경우 과학기술정보통신부장관은 윤리원칙 및 제2항에 따른 실천방안과의 연계성·정합성 등에 관한 권고 또는 의견의 표명을 할 수 있다.

(4) AI 안전성 및 신뢰성

정부는 AI가 국민의 생활에 미치는 잠재적 위험을 최소화하고 안전한 AI 이용을 위한 신뢰 기반을 조성하기 위하여 안전성·신뢰성 관련 시책 마련해야 한다. 과학기술정보통신부장관은 단체등이 AI의 안전성·신뢰성 확보를 위하여 자율적으로 추진하는 검증·인증 활동을 지원하기 위한 가이드라인 보급, 연구 지원, 전문인력 양성 지원 등 수행한다.

AI 사업자가 고영향 AI를 제공하는 경우 사전에 검·인증등을 받도록 노력하여야 한다. AI 사업자는 학습에 사용된 누적 연산량이 대통령령으로 정하는 기준 이상인 AI 시스템의 안전성을 확보하기 위하여 수명주기(life cycle) 전반의 위험관리, 위험관리체계 구축 등을 이행해야 한다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제29조(인공지능 신뢰 기반 조성을 위한 시책의 마련) 정부는 인공지능이 국민의 생활에 미치는 잠재적 위험을 최소화하고 안전한 인공지능의 이용을 위한 신뢰 기반을 조성하기 위하여 다음 각 호의 시책을 마련하여야 한다.

1. 안전하고 신뢰할 수 있는 인공지능 이용환경 조성
2. 인공지능의 이용이 국민의 일상생활에 미치는 영향 등에 관한 전망과 예측 및 관련 법령·제도의 정비
3. 인공지능의 안전성·신뢰성 확보를 위한 안전기술 및 인증기술의 개발 및 확산 지원
4. 안전하고 신뢰할 수 있는 인공지능사회 구현 및 인공지능윤리 실천을 위한 교육·홍보
5. 인공지능사업자의 안전성·신뢰성 관련 자율적인 규약의 제정·시행 지원
6. 인공지능사업자, 이용자 등으로 구성된 인공지능 관련 단체(이하 "단체등"이라 한다)의 인공지능의 안전성·신뢰성 증진을 위한 자율적인 협력, 윤리지침 제정 등 민간 활동의 지원 및 확산
7. 그 밖에 인공지능의 안전성·신뢰성 확보를 위하여 대통령령으로 정하는 사항

제30조(인공지능 안전성·신뢰성 검·인증등 지원) ① 과학기술정보통신부장관은 단체등이 인공지능의 안전성·신뢰성 확보를 위하여 자율적으로 추진하는 검증·인증 활동(이하 "검·인증등"이라 한다)을 지원하기 위하여 다음 각 호의 사업을 추진할 수 있다.

1. 인공지능의 개발에 관한 가이드라인 보급
2. 검·인증등에 관한 연구의 지원
3. 검·인증등에 이용되는 장비 및 시스템의 구축·운영 지원
4. 검·인증등에 필요한 전문인력의 양성 지원
5. 그 밖에 검·인증등을 지원하기 위하여 대통령령으로 정하는 사항

② 과학기술정보통신부장관은 검·인증등을 받고자 하는 중소기업등에 대하여 대통령령으로 정하는 바에 따라 관련 정보를 제공하거나 행정적·재정적 지원을 할 수 있다.

③ 인공지능사업자가 고영향 인공지능을 제공하는 경우 사전에 검·인증등을 받도록 노력하여야 한다.

④ 국가기관등이 고영향 인공지능을 이용하려는 경우에는 검·인증등을 받은 인공지능에 기반한 제품 또는 서비스를 우선적으로 고려하여야 한다.

제32조(인공지능 안전성 확보 의무) ① 인공지능사업자는 학습에 사용된 누적 연산량이 대통령령으로 정하는 기준 이상인 인공지능시스템의 안전성을 확보하기 위하여 다음 각 호의 사항을 이행하여야 한다.

1. 인공지능 수명주기 전반에 걸친 위험 식별, 평가 및 완화
2. 인공지능 관련 안전사고를 모니터링하고 대응하는 위험관리체계 구축

② 인공지능사업자는 제1항 각 호에 따른 사항의 이행 결과를 과학기술정보통신부장관에게 제출하여야 한다.

③ 과학기술정보통신부장관은 제1항 각 호에 따른 사항의 구체적인 이행 방식 및 제2항에 따른 결과 제출 등에 필요한 사항을 정하여 고시하여야 한다.

(5) AI 투명성

AI 투명성 확보를 위해 AI로 작동하는 사실을 고지하는 조치를 취한다. 먼저, AI 사업자는 고영향 AI 또는 생성형 AI를 이용한 제품 또는 서비스를 제공하려는 경우 해당 제품 또는 서비스가 AI에 기반하여 운용된다는 사실을 이용자에게 사전에 고지해야 한다. 만약, AI 사업자가 생성형 AI 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 AI에 의하여 생성되었다는 사실을 표시해야 한다.

특히, AI 사업자가 AI 시스템을 이용하여 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등의 결과물(딥페이크)을 제공하는 경우 해당 결과물이 AI 시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식(가시적 워터마크)으로 고지 또는 표시해야 한다. 다만, 해당 결과물이 예술적·창의적 표현물에 해당하거나 그 일부를 구성하는 경우에는 전시 또는 향유 등을 저해하지 않는 방식으로 고지 또는 표시할 수 있다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제31조(인공지능 투명성 확보 의무) ① 인공지능사업자는 고영향 인공지능 또는 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 한다.
② 인공지능사업자는 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야 한다.
③ 인공지능사업자는 인공지능시스템을 이용하여 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등의 결과물을 제공하는 경우 해당 결과물이 인공지능시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시하여야 한다. 이 경우 해당 결과물이 예술적·창의적 표현물에 해당하거나 그 일부를 구성하는 경우에는 전시 또는 향유 등을 저해하지 않는 방식으로 고지 또는 표시할 수 있다.
④ 그 밖에 제1항에 따른 사전고지, 제2항에 따른 표시, 제3항에 따른 고지 또는 표시의 방법 및 그 예외 등에 관하여 필요한 사항은 대통령령으로 정한다.

제43조(과태료) ① 다음 각 호의 어느 하나에 해당하는 자에게는 3천만원 이하의 과태료를 부과한다.
1. 제31조제1항에 따른 고지를 이행하지 아니한 자

(6) 학습용데이터 관련 시책

과학기술정보통신부장관은 관계 중앙행정기관의 장과 협의하여 AI의 학습용데이터의 생산·수집·활용 등을 촉진하기 위하여 필요한 시책을 추진하여야 하고, 학습용데이터 구축사업의 효율적 수행을 위하여 학습용데이터 통합제공시스템을 구축·운영하여야 한다. 이와 함께 정부는 학습용데이터의 생산·수집·활용 등의 지원대상사업을 선정하여 시행할 수 있다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제15조(인공지능 학습용데이터 관련 시책의 수립 등) ① 과학기술정보통신부장관은 관계 중앙행정기관의 장과 협의하여 인공지능의 개발·활용 등에 사용되는 데이터(이하 "학습용데이터"라 한다)의 생산·수집·관리 및 유통·활용 등을 촉진하기 위하여 필요한 시책을 추진하여야 한다.
② 정부는 학습용데이터의 생산·수집·관리 및 유통·활용 등에 관한 시책을 효율적으로 추진하기 위하여 지원대상사업을 선정하고 예산의 범위에서 지원할 수 있다.
③ 정부는 학습용데이터의 생산·수집·관리 및 유통·활용의 활성화 등을 위하여 다양한 학습용데이터를 제작·생산하여 제공하는 사업(이하 "학습용데이터 구축사업"이라 한다)을 시행할 수 있다.
④ 과학기술정보통신부장관은 학습용데이터 구축사업의 효율적 수행을 위하여 학습용데이터를 통합적으로 제공·관리할 수 있는 시스템(이하 "통합제공시스템"이라 한다)을 구축·관리하고 민간이 자유롭게 이용할 수 있도록 제공하여야 한다.
⑤ 과학기술정보통신부장관은 통합제공시스템을 이용하는 자에 대하여 비용을 징수할 수 있다.
⑥ 그 밖에 지원대상사업의 선정 및 지원, 학습용데이터 구축사업의 시행, 통합제공시스템의 구축·관리 및 비용의 징수 등에 필요한 사항은 대통령령으로 정한다.

(7) 시정명령과 과태료

과학기술정보통신부장관은 이 법에 위반되는 사항을 발견하거나 혐의가 있음을 알게 된 경우 AI 사업자에 대하여 자료를 제출하게 하거나 소속 공무원으로 하여금 필요한 조사를 하게 할 수 있고, 위반 사실이 있다고 인정되면 위반행위의 중지명령이 시정명령을 내릴 수 있다. 이 명령을 이행하지 아니할 경우 3천만원 이하의 과태료를 부과한다.

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(대안)

제40조(사실조사 등) ① 과학기술정보통신부장관은 다음 각 호의 어느 하나에 해당하는 경우에는 인공지능사업자에 대하여 관련 자료를 제출하게 하거나, 소속 공무원으로 하여금 필요한 조사를 하게 할 수 있다.

1. 이 법에 위반되는 사항을 발견하거나 혐의가 있음을 알게 된 경우
2. 이 법의 위반에 대한 신고를 받거나 민원이 접수된 경우

② 과학기술정보통신부장관은 제1항에 따른 조사를 위하여 필요한 경우 소속 공무원으로 하여금 인공지능사업자의 사무소·사업장에 출입하여 장부·서류, 그 밖의 자료나 물건을 조사하게 할 수 있다.

③ 과학기술정보통신부장관은 제1항 및 제2항에 따른 조사 결과 인공지능사업자가 이 법을 위반한 사실이 있다고 인정되면 인공지능사업자에게 해당 위반행위의 중지나 시정을 위하여 필요한 조치를 명할 수 있다.

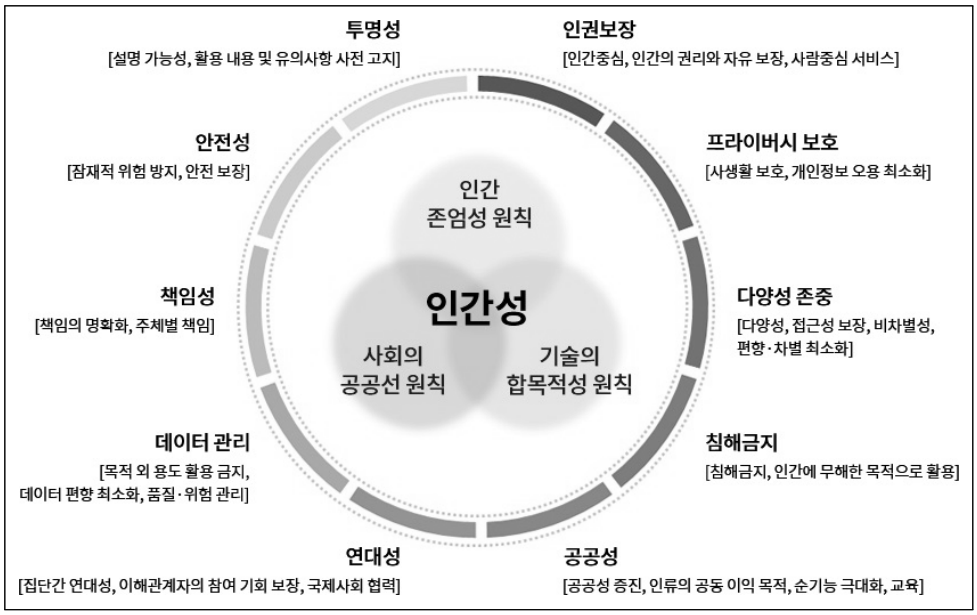
제43조(과태료) ① 다음 각 호의 어느 하나에 해당하는 자에게는 3천만원 이하의 과태료를 부과한다.

3. 제40조제3항에 따른 중지명령이나 시정명령을 이행하지 아니한 자

나. 인공지능 윤리기준

2020년 12월 23일 과학기술정보통신부와 정보통신정책연구원은 대통령 직속 4차산업혁명위원회 전체회의를 통해 인공지능 시대 바람직한 인공지능 개발·활용 방향을 제시하기 위한 사람이 중심이 되는 「인공지능(AI) 윤리기준」을 발표했다. 이는 윤리적 AI를 실현하기 위해 정부·공공기관, 기업, 이용자 등 모든 사회구성원이 인공지능 개발에서 활용까지 전 단계에서 함께 지켜야 할 주요 원칙과 핵심 요건을 제시하는 기준으로 '인간성(Humanity)'을 최고 가치로 제시했다. 인간성을 달성하기 위한 3대 기본원칙으로 인간의 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙을 제시하고, 10대 핵심요건으로 인권 보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성의 요건을 제시했다.

[그림 1] 인공지능 윤리기준 체계

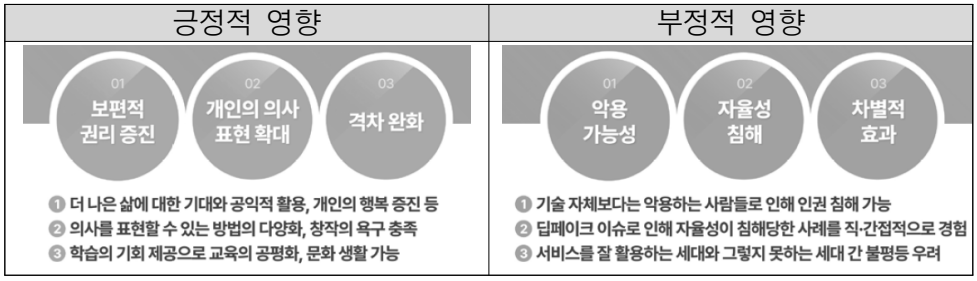


자료: 정보통신정책연구원 인공지능 윤리 소통채널 홈페이지, “인공지능 윤리기준”, (최종방문일: 2024.12.16.), <https://ai.kisdi.re.kr/aieth/main/contents.do?menuNo=400029>

이에 따라 현재 자율적 규율체계와 합리적 관리체계가 병행되고 있다. 자율적 규율체계 마련을 위해 2022년에 「인공지능 윤리기준 실천을 위한 자율점검표(안)」을 마련하고, 챗봇, 작문, 영상, 채용 AI에 적용했다. 합리적 관리체계 마련을 위해서는 「인공지능 윤리영향평가 프레임워크」를 개발하여 2024년에 ‘AI 기반 영상 합성 서비스’를 대상으로 시범적용했다. 윤리영향평가는 AI 제품·서비스의 윤리적 영향력을 사전에 평가함으로써 긍정적 영향 극대화 및 부정적 영향 최소화를 위한 관리·제도·정책적 조치방안 등 시사점을 도출하고, 기업이 AI 제품·서비스를 보다 윤리적인 방식으로 개발·배포·활용하도록 장려하기 위해 시행된 것이다. 2024년에는 ‘AI 기반 영상 합성 서비스’ 즉, 딥페이크에 대해 시범적용을 해 본 결과, 딥페이크 기술 활용으로 인해 개인의 보편적 권리 증진, 개인

의 의사표현 확대, 격차 완화 등의 긍정적 영향을 미치는 것으로 나타났지만, 딥페이크의 악용 가능성, 자율성 침해, 차별적 효과는 부정적인 영향으로 나타났다.

[그림 2] 인공지능 윤리영향평가 프레임워크 시범 적용 결과



자료 : 과학기술정보통신부·정보통신정책연구원, 「2024 인공지능 윤리·신뢰성 포럼 공개세미나 자료집」, 2024.11.28.

3. 소결

유럽연합 「인공지능법」과 우리나라 인공지능기본법안 내용을 비교하면 다음 [표 8]과 같다. 우리나라 인공지능기본법안은 유럽연합 「인공지능법」 체계의 큰 틀을 수용한 것으로 보인다. 다만, 유럽연합은 고위험 AI 시스템뿐만 아니라 생성형 AI와 같은 GPAI에 대한 규제를 명확히 하는 것과 달리, 우리나라 인공지능기본법안은 고영향 AI 사업자에 대한 의무만 규정하고 있다. 따라서 일상에서 많은 사람들이 직면하고 있는 생성형 AI의 위험에 대한 대응은 향후 보완이 필요할 것으로 보인다.

[표 8] 인공지능의 내재적 위험 유형

대분류	중분류	국가별 대응	
		EU 인공지능법	한국 인공지능기본법안
학습	학습데이터	• (고위험 AI 공급자) 학습데이터	• (고영향 AI 사업자) 학습데이

대분류	중분류	국가별 대응	
		EU 인공지능법	한국 인공지능기본법안
데이터 위험	공개	등의 대표성·정확성·합목적성 등을 준수할 수 있도록 관리 및 활용해야 함	터의 개요 등에 대한 설명 방안을 수립·시행해야 함
		• (GPAI 공급자) 학습콘텐츠의 상세 요약서를 작성하여 공개해야 함	-
	학습데이터 저작권	• (GPAI 공급자) 「EU 저작권 지침」을 준수해야 하며, 권리자가 TDM을 거부할 경우 이를 따라야 함	-
	학습데이터 부족	-	• (과학기술정보통신부장관 등) 학습용데이터 생산·수집 등을 위한 시책을 추진함
AI 모델·시스템의 기술적 위험	투명성 및 설명가능성	• (고위험 AI 공급자 및 GPAI 공급자) 모델·시스템 출시 전에 기술문서를 작성해야 함(일반적 기술, AI 시스템 개발 절차, 모니터링에 관한 사항 등 포함)	• (고영향 AI 사업자) 인공지능이 도출한 최종결과, 인공지능의 최종결과 도출에 활용된 주요 기준 등에 대한 설명 방안을 수립·시행해야 함
		• (고위험 AI 공급자 및 GPAI 공급자) 배포자가 모델의 목적, 기능 등을 이해할 수 있도록 이용 지침 또는 정보를 제공해야 함	-
		• (고위험 AI 공급자) 모든 규정을 준수하고 있는지 적합성 평가를 거쳐야 함	• (고영향 AI 사업자) 사전에 안전성·신뢰성에 관한 검·인증 등을 받도록 노력해야 함
	자율성 통제	• (고위험 AI 공급자) 인간의 감독이 가능하도록 모델을 설계해야 함	• (고영향 AI 사업자) 고영향 AI에 대한 사람의 관리·감독을 모색해야 함
	인간의 권리 침해	• (고위험 AI 배포자) 시스템에 대한 기본권 영향평가를 실시해야 함	• (고영양 AI 사업자) 제품 또는 서비스를 제공하기 전에 기본권 영향평가를 실시하기 위해 노력해야 함

대분류	중분류	국가별 대응	
		EU 인공지능법	한국 인공지능기본법안
AI 모델·시스템의 이용자 위험	부주의한 이용자	• (특정 AI 시스템 공급자) 상대방이 AI 시스템과 상호작용하고 있다는 사실을 인식할 수 있도록 해야 한다. 또한 합성 콘텐츠(오디오, 이미지, 동영상, 텍스트 등) 생성하는 AI 시스템은 AI로 만든 것임을 표시해야 함	• (고영향 AI 사업자 및 생성형 AI 사업자) 제품 또는 서비스가 AI에 기반하여 운용된다는 사실을 이용자에게 사전에 고지해야 함 • (생성형 AI를 이용하는 사업자) 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시해야 함 • (딥페이크 콘텐츠를 제공하는 AI 사업자) 해당 결과물이 AI 시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시해야 함

Ⅳ. 입법·정책 과제

1. 데이터 위험 대응

가. 학습데이터에 대한 설명 강화

부적절한 데이터 위험은 주로 학습데이터에 무엇이 사용되었는지 충분히 알지 못한다는 것에서 출발한다. 학습데이터의 수집, 처리, 활용 방법 등에 대한 정확한 문서가 없으면 모델의 동작을 만족스럽게 설명하기 어렵고, 데이터와 모델이 갖고 있는 위험을 평가하는 능력이 제한되므로 이를 보완하는 것이 필요하다.

구체적으로, 학습데이터의 출처를 밝히는 등 설명을 강화할 필요가 있다. 현재 인공지능기본법안은 고영향 AI에 대해서 학습용데이터의 개요 등에 대한 설명 방안의 수립·시행이 포함된 조치를 의무화하고 있다. 그러나 고영향 AI는 「에너지법」 제2조제1호에 따른 에너지의 공급, 「먹는물관리법」 제3조제1호에 따른 먹는물의 생산 공정 등의 영역에서 활용되는 AI 등으로 열거되어 있는데, 이러한 열거에 포함되지 않는다면 학습데이터의 투명성 확보가 쉽지 않다. 특히 일상에서 가장 빈번하게 사용되는 생성형 AI는 그 대상에서 제외되어 있다. 따라서 향후 생성형 AI를 중심으로 학습데이터의 개요, 편향성 검토 등을 설명할 수 있도록 하는 방안을 모색할 필요가 있다.

나. 개인정보 관련 법률 정비

개인정보의 경우, 핵심은 공개된 개인정보 활용의 적법성에 관한 것이다. 공개된 개인정보는 「개인정보 보호법」상 '개인정보'에 해당하므로 AI 학습 등의 목적으로 공개된 개인정보를 활용하는 과정에서 발생하는 법적 문제와 윤리적 문제는 모두 정보주체의 자기정보결정권 침해 여부로 귀결된다. 「개인정보 보호법

」은 정보주체의 자기정보결정권을 보호하는 보편적 수단으로 사전 동의(opt-in) 방식을 채택하고 있고, 공개된 개인정보 처리에 대한 별도의 규정이 없으므로, 공개된 개인정보 활용을 위해서는 정보주체로부터 명시적인 사전동의를 받아야 하는 것이 원칙이다.

다만, 「개인정보 보호법」 제15조제1항제6호의 '정당한 이익'에 해당하는 경우에는 예외적으로 정보주체의 사전동의를 받지 않고 개인정보를 처리할 수 있으므로, AI의 공개된 개인정보 활용이 정당한 이익을 위한 활용에 해당하는지 살펴보아야 한다. 이러한 정당한 이익이 성립하기 위해서는 개인정보처리자의 정당한 이익이 있을 것(목적의 정당성), 개인정보 처리의 정당한 이익 달성을 위하여 필요하고, 상당한 관련성 및 합리성이 인정될 것(처리의 필요성·합리성), 개인정보처리자의 정당한 이익이 명백하게 정보주체의 권리보다 우선할 것(이익형량) 등의 세 가지 요건이 충족되어야 한다.[19]

단기적으로는 이와 같은 법률의 해석을 통해 법률과 기술의 격차를 조절할 수 있겠지만, 장기적으로는 「개인정보 보호법」에 관련 사항을 명시하여 법 적용의 안정성과 예측가능성을 높일 필요가 있다. 참고할 수 있는 외국 입법례로는, 공개된 개인정보는 법률의 적용대상에서 제외하거나 동의 없이 사용할 수 있도록 한 미국 캘리포니아주의 「프라이버시 권리법」과 캐나다의 「개인정보 보호 및 전자문서법」, 유럽연합의 「일반 정보 보호 규정」 등이 있다. 미국 캘리포니아주는 공개 정보(생체인식정보 제외)를 법률상 '개인정보' 대상에서 제외한 「소비자 프라이버시법(The California Consumer Privacy Act, CCPA)」을 개정하여 2023년 1월 1일부터 「프라이버시 권리법(The California Privacy Rights Act of 2020, CPRA)」을 시행했다. 「프라이버시 권리법」에서는 '공적 관심사로서 합법적으로 획득한 진실된 정보'까지 법률상 '개인정보' 대상에서 제외된다.[20] 캐나다 「개인

19) 개인정보보호위원회, 「인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서」, 2024.7.

정보 보호 및 전자문서법(Personal Information Protection and Electronic Documents Act, PIPEDA)」 제7조는 공개된 정보도 개인정보에 해당되지만 수집·이용·공개(collect·use·disclose)에 동의가 필요하지 않다고 규정한다.[21] 유럽연합의 「일반 정보 보호 규정(General Data Protection Regulation, GDPR)」은 공개된 개인정보를 일반 개인정보와 동등하게 규율하여 동의, 계약, 정당한 이익 등 적법근거에 따라 처리될 수 있도록 규정하고 있지만, 제9조제2항e호의 민감정보 처리규정에 따르면 '정보주체가 명백하게 공개한 개인정보(민감정보)'의 경우에는

20) 미국 캘리포니아주 CPRA 조문은 다음과 같다(자료 : 개인정보 포털, 캘리포니아 CPRA 원본 및 번역본, (최종방문일: 2024.12.16.), <https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000843184&fileSn=0>).

THE CALIFORNIA PRIVACY RIGHTS ACT OF 2020
1798.140. Definitions. For purposes of this title:
"Personal information" does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern.

21) 캐나다 PIPEDA 조문 내용은 다음과 같다(자료: 캐나다 법률정보시스템, (최종방문일: 2024.12.16.), <https://laws-lois.justice.gc.ca/eng/acts/p-8.6/page-1.html#h-416969>).

Personal Information Protection and Electronic Documents Act (S.C. 2000, c. 5)
PART 1 - Protection of Personal Information in the Private Sector
DIVISION 1 Protection of Personal Information
7 (1) For the purpose of clause 4.3 of Schedule 1, and despite the note that accompanies that clause, an organization may collect personal information without the knowledge or consent of the individual only if
(d) the information is publicly available and is specified by the regulations; or
(2) For the purpose of clause 4.3 of Schedule 1, and despite the note that accompanies that clause, an organization may, without the knowledge or consent of the individual, use personal information only if
(c.1) it is publicly available and is specified by the regulations; or
(3) For the purpose of clause 4.3 of Schedule 1, and despite the note that accompanies that clause, an organization may disclose personal information without the knowledge or consent of the individual only if the disclosure is
(h.1) of information that is publicly available and is specified by the regulations; or

예외적으로 처리를 할 수 있다.[22]

다. 저작권 관련 법률 정비

온라인에 공개된 데이터를 대량으로 학습하는 과정에서 저작권으로 보호되는 데이터의 수집은 피하기 어렵다. 이러한 상황에서 최근 글로벌 빅테크 기업들은 복잡한 법리 해석보다는 언론사 등에 뉴스 사용료를 내고 양질의 학습데이터를 확보하는 현실적 대안을 찾고 있다.[23] 오픈에이아이는 2024년 5월 생성형 AI 학습을 위해 월스트리트저널(WSJ) 등을 소유한 뉴스코퍼레이션과 콘텐츠 공급 계약을 체결하고 콘텐츠 사용 대가로 5년간 2억 5천만 달러(약 3천 426억 원) 이상을 뉴스코퍼레이션에 지급할 것으로 알려졌다. 구글도 저작권 침해를 염려해 연간 최대 600만 달러(약 82억 원)를 지급하는 조건으로 뉴스코퍼레이션과 AI 콘텐츠 이용 및 제품 개발을 위한 계약을 체결했다.

다만, 이러한 접근은 자본력이 충분한 빅테크에 유리한 것으로, 소수 빅테크의 AI 집중도를 더욱 키워주는 효과가 있다. 따라서 온라인에 공개된 데이터의 학습과 저작권자의 권리보호의 조화를 이루는 「저작권법」 개정을 모색해 볼 수 있다. 참고로, 미국은 「연방저작권법」 제107조에 따라 공정이용(fair use)을 허용하고 있으나, 이것이 AI 학습을 위하여 대규모의 정보를 이용하는 경우에도 적용될 수

22) 유럽연합 GDPR 조문 내용은 다음과 같다(자료: 유럽연합 GDPR TEXT 홈페이지, (최종 방문일: 2024.12.16.), <https://gdpr-text.com/ko/read/article-9/#comment_gdpr-a-09_2e>).

General Data Protection Regulation
Article 9 GDPR. Processing of special categories of personal data
2. Paragraph 1 shall not apply if one of the following applies:
(e) processing relates to personal data which are manifestly made public by the data subject;

23) 장유미, 「'구글 대항마'로 떠오른 美 퍼플렉시티, 저작 권 침해 문제로 '몸살'」, 지디넷코리아, 2024.6.10. 장유미, 「'구글 대항마'로 떠오른 美 퍼플렉시티, 저작 권 침해 문제로 '몸살'」, 지디넷코리아, 2024.6.10. <https://zdnet.co.kr/view/?no=20240610093432>

있는지에 대해서는 의견이 나뉘고 있다. 미국「연방저작권법」제107조는 추상적인 요건들로 구성되어 있어, 어떤 사안이 공정이용에 해당하는지 판례를 통해 구체화되고 있는데, 생성형 AI 결과물의 목적이 원저작물과 동일하다면 공정이용 가능성이 낮을 것이고, 양자 간의 목적이 다르다는 것을 입증하게 된다면 공정이용에 해당할 가능성이 높다.[24)]

유럽연합은 2019년에「유럽연합 DSM 저작권 지침」을 제정하여 학술연구 목적의 TDM과 일반적인 TDM을 구분하여 이원적으로 규율한다. 제3조에 따르면 연구기관 및 문화유산기관이 학술연구 목적의 TDM을 수행할 때 적법하게 접근 가능한 저작물에 대해 복제 및 추출을 허용하고 있는 반면, 제4조제3항에서 권리자가 TDM을 거부하는 옵트아웃(opt-out) 규정을 두어 제도의 균형을 맞추고 있다.[25)]

24) 신용우 외, 앞의 글.

25) 유럽연합 저작권 지침 조문 내용은 다음과 같다(자료: 유럽연합 법령정보시스템 홈페이지, (최종방문일: 2024.12.16.), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32019L0790>).

DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019

on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Article 3

Text and data mining for the purposes of scientific research

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

Article 4

Exception or limitation for text and data mining

제22대 국회에서는 AI 창작물의 저작권에 대한 입법이 이루어지지 않고 있으며, 제21대 국회에서는 TDM에 대한 면책 규정을 신설한 「저작권법 전부개정법률안」이 발의된 바 있다. 도종환의원이 대표발의한 「저작권법 전부개정법률안」(의안번호: 2107440)은 저작물에 표현된 사상이나 감정을 향유하지 아니하는 경우 TDM을 허용할 수 있도록 하는 내용을 신설하였으나, 제21대 국회 임기 만료로 폐기되었다.

저작권법 전부개정법률안
(도종환의원 대표발의, 의안번호: 2107440)

제43조(정보분석을 위한 복제·전송) ① 컴퓨터를 이용한 자동화 분석기술을 통해 다수의 저작물을 포함한 대량의 정보를 분석(규칙, 구조, 경향, 상관관계 등의 정보를 추출하는 것)하여 추가적인 정보 또는 가치를 생성하기 위한 것으로 저작물에 표현된 사상이나 감정을 향유하지 아니하는 경우에는 필요한 한도 안에서 저작물을 복제·전송할 수 있다. 다만, 해당 저작물에 적법하게 접근할 수 있는 경우에 한정한다.
② 제1항에 따라 만들어진 복제물은 정보분석을 위하여 필요한 한도에서 보관할 수 있다.

라. 데이터 축적 지원

세계적으로 가장 많이 사용하고 있는 LLM은 영어권 자료를 바탕으로 만들어진 것이기 때문에 한국 상황을 충분히 반영하지 못한다. 이러한 상황에서 한국적 특수상황에 관한 질문을 하면, 영어권에 편향되거나 불충분한 대답을 할 우려가 크다. 따라서 한국적 특성을 고려한 데이터를 구축하여 학습시킬 필요가 있다.

인공지능기본법안에서도 학습용데이터 관련 시책이 규정되어 있다. 그러나 어

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

떠한 데이터를 전략적으로 지원할 것인지는 규정되어 있지 않다. 한국의 역사와 맥락 등에 맞는 학습데이터의 구축이라는 점을 감안하여, 한국의 정치·경제·사회·문화적 데이터가 보다 구체적으로 지원받을 수 있도록 명시하는 방안을 고려해 볼 수 있을 것이다.

2. 기술적 위험 대응

가. AI 기술문서 대상 확대와 사용설명서 작성

AI의 기술적 위험에 대응하기 위해서는 AI 프로세스의 투명성과 설명가능성을 높여야 한다. AI가 내린 의사결정의 이유를 사용자나 외부 감시자가 이해할 수 있어야 공정성·책임성·투명성을 확보할 수 있고, AI에 대한 신뢰도를 높일 수 있다.

이를 위해서 적정 수준의 AI 기술문서를 작성하도록 할 필요가 있다. 인공지능 기본법안에서는 고영향 AI에 대해서 안전성·신뢰성 확보를 위한 조치의 내용을 확인할 수 있는 문서의 작성과 보관을 의무화하고 있지만 생성형 AI에 대한 적용은 충분하지 않다. 우리 현실에서 차지하는 생성형 AI의 비중을 고려했을 때, 고영향 AI 이외에도 기술문서의 작성·공개 대상을 정할 수 있도록 하는 방안을 검토할 필요가 있다.

다음으로, AI 모델을 어떻게 사용되어야 하는지에 대한 설명을 제시하는 것이 적절하다. AI 모델은 다양한 목적으로 사용될 수 있으므로 모델의 용도를 파악하는 것이 해당 모델의 관련 위험을 정의하는 데 중요하다. 모델의 용도가 충분히 명시되지 않은 경우 모델의 관련 위험을 정확하게 판단하고 완화하기 어려울 수 있기 때문이다. 따라서 고영향 AI와 생성형 AI에 대해서 이를 이용하는 사업자들이 모델의 목적·기능 등을 이해할 수 있도록 이용지침 또는 정보를 제공하도록 할 필요가 있다.

나. AI 위험성 평가 확대

인공지능기본법안은 고영향 AI에 대해 위험관리방안을 수립·운영해야 하고, 사람의 기본권에 미치는 영향평가를 실시하도록 규정하고 있다. 학습에 사용된 누적 연산량이 일정 기준 이상인 경우, AI 수명주기 전반을 고려한 위험 식별·평가를 추진할 의무를 부여하고 있다. 이와 함께 AI 안전성·신뢰성 확보를 위하여 자율적으로 추진하는 검증·인증을 기본으로 하고 있고, 정부는 이러한 자율적인 조치를 지원하기 위하여 가이드라인 보급, 연구개발 지원 등을 수행한다.

그러나 AI가 초래하는 위험은 기본권에 한정되지 않는다. 따라서 기본권에 미치는 영향 이외에도, 고영향 AI의 위험 가능성을 미리 파악할 수 있도록 포괄적인 위험평가를 할 필요가 있다. 다만, 이러한 확대된 위험성 평가를 의무화할 경우 기업의 부담이 급증할 우려가 있으므로, 현재 자율적으로 운영되고 있는 AI 윤리영향평가를 적극 활용하는 방안을 모색해 볼 수 있다.

3. 이용자 위험 대응

가. 이용자 리터러시 개선

이용자가 AI의 가능성과 한계, 위험을 적절하게 인식할 수 있도록 리터러시(literacy)를 개선하는 것이 중요하다. 이는 단순히 코딩과 같은 AI에 관한 기술적 활용 능력을 키운다고 해서 달성되는 것이 아니다. 인간의 다양한 사회·경제적 맥락 안에서 개인이 AI를 맹목적으로 수용 또는 거부하는 것이 아니라, 장·단점과 유·불리를 따져서 비판적으로 수용할 수 있도록 해야 한다.

인공지능기본법안에서는 안전하고 신뢰할 수 있는 AI를 위한 교육·홍보 등 리터러시 관련 규정이 다소 추상적으로 규정되어 있다.[26] 실제 이용자들이 다양한

26) 인공지능 리터러시 관련 인공지능기본법안 내용은 다음과 같다.
제13조(인공지능기술 개발 및 안전한 이용 지원) ② 정부는 인공지능기술의 안전

AI 이용 환경에서 어려움을 겪지 않도록 프롬프트에 개인정보, 중요 기밀을 입력하지 않는 것부터 시작해서, AI에 대한 과잉신뢰, AI 시스템의 성능과 목적범위를 벗어난 활용 제한, 생산적 AI 활용 등 다양한 영역에서 이용자 리터러시 개선을 보다 구체적으로 정하는 방안을 모색해 볼 수 있을 것이다.

나. 악의적 이용과 사이버 공격의 규제

AI 모델·시스템의 악의적 이용자에 대해서는 개별 법률로 규제하는 것이 적절하다. 선거기간 동안 딥페이크를 이용하지 못하도록 하는 「공직선거법」 제82조의8, 딥페이크 불법합성 성범죄물을 소지할 경우에도 처벌하는 「성폭력범죄의 처벌 등에 관한 특례법」 제14조의2가 대표적이다. 인공지능기본법안 시행 이후, 다양한 분야에서 이 법을 반영하여 악의적 AI 모델·시스템 이용에 대한 규제가 만들어질 것으로 예상된다. 각 부처와 AI 안전 수요를 파악하는 협력을 강화할 필요가 있다.

AI에 대한 사이버 공격의 경우, 개발자 자체적으로 예상하여 대응하는 것은 한계가 있다. 따라서 AI 모델·시스템에 대한 가상의 적대적 테스트(레드티밍) 등을

하고 편리한 이용을 위하여 다음 각 호의 사업을 지원할 수 있다.
6. 인공지능의 안전한 개발과 이용을 위한 인식개선, 올바른 이용방법과 안전 환경 조성을 위한 교육 및 홍보 사업

제26조(한국인공지능진흥협회의 설립) ③ 협회는 다음 각 호의 업무를 수행한다.
5. 안전하고 신뢰할 수 있는 인공지능의 개발·활용을 위한 교육 및 홍보

제27조(인공지능 윤리원칙 등) ② 과학기술정보통신부장관은 사회 각계의 의견을 수렴하여 윤리원칙이 인공지능의 개발·활용 등에 관여하는 모든 사람에 의해 실현될 수 있도록 실천방안을 수립하고 이를 공개 및 홍보·교육하여야 한다.

제29조(인공지능 신뢰 기반 조성을 위한 시책의 마련) 정부는 인공지능이 국민의 생활에 미치는 잠재적 위험을 최소화하고 안전한 인공지능의 이용을 위한 신뢰 기반을 조성하기 위하여 다음 각 호의 시책을 마련하여야 한다.
4. 안전하고 신뢰할 수 있는 인공지능사회 구현 및 인공지능윤리 실천을 위한 교육·홍보

활성화하고, 그 결과를 반영하여 AI 모델·시스템 안전성 강화 조치를 마련할 필요가 있다. 인공지능기본법안 제13조제2항에서 안전한 이용을 위한 사업 지원을 규정하고 있는데,[27] 여기에 레드티밍과 같은 새로운 방안을 포함하는 것을 고려해 볼 수 있을 것이다.

또한 AI에 대한 공격이 국제적으로 진행되는 것을 고려한다면, 국제협력에 적극 참여하는 것도 중요하다. 인공지능기본법안 제22조[28]에서 인공지능산업의 경

27) 인공지능 기술개발 등과 관련한 인공지능기본법안 내용은 다음과 같다.
제13조(인공지능기술 개발 및 안전한 이용 지원) ② 정부는 인공지능기술의 안전하고 편리한 이용을 위하여 다음 각 호의 사업을 지원할 수 있다.
1. 「지능정보화 기본법」 제60조제1항 각 호의 사항을 인공지능에서 구현하는 기술에 관한 연구개발 사업
2. 「지능정보화 기본법」 제60조제3항에 따른 비상정지를 인공지능에서 구현하는 기술에 관한 연구의 지원 및 해당 기술의 확산을 위한 사업
3. 인공지능기술의 개발에 있어서 「지능정보화 기본법」 제61조제2항에 따른 사생활등의 보호에 적합한 설계 기준 및 기술의 연구개발 및 보급 사업
4. 인공지능기술의 「지능정보화 기본법」 제56조제1항에 따른 사회적 영향평가의 실시와 적용을 위한 연구개발 사업
5. 인공지능이 인간의 존엄성 및 기본권을 존중하는 방향으로 개발·이용될 수 있도록 하는 기술 또는 기준 등의 연구개발 및 보급 사업
6. 인공지능의 안전한 개발과 이용을 위한 인식개선, 올바른 이용방법과 안전 환경 조성을 위한 교육 및 홍보 사업
7. 그 밖에 인공지능의 개발과 이용에 있어서 국민의 기본권, 신체와 재산을 보호하기 위하여 필요한 사업

28) 인공지능 국제협력 관련 인공지능기본법안 내용은 다음과 같다.
제22조(국제협력 및 해외시장 진출의 지원) ② 정부는 인공지능산업의 경쟁력 강화와 해외시장 진출을 촉진하기 위하여 인공지능산업에 종사하는 개인·기업 또는 단체 등에 대하여 다음 각 호의 지원을 할 수 있다.
1. 인공지능산업 관련 정보·기술·인력의 국제교류
2. 인공지능산업 관련 해외진출에 관한 정보의 수집·분석 및 제공
3. 국가 간 인공지능기술, 인공지능제품 또는 인공지능서비스의 공동 연구·개발 및 국제표준화
4. 인공지능산업 관련 외국자본의 투자유치

쟁력 강화와 해외시장 진출을 촉진을 위해서 지원하도록 하고 있는데, 안전하고 신뢰할 수 있는 AI를 위한 국제협력에 대한 지원도 모색할 필요가 있다.

5. 인공지능등 관련 해외 전문 학회 및 전시회 참가 등 홍보 및 해외 마케팅
6. 인공지능제품 또는 인공지능서비스의 수출에 필요한 판매·유통체계 및 협력체계 등의 구축
7. 인공지능윤리에 관한 국제적 동향 파악 및 국제협력
8. 그 밖에 인공지능산업의 경쟁력 강화와 해외시장 진출 촉진을 위하여 필요한 사항

Ⅴ. 결론

AI는 비약적인 기술 발전과 함께 사회 전반에 긍정적인 혁신을 가져왔으나 동시에 다양한 위험도 내포하고 있다. 이 중에서 AI의 구축 및 작동 과정에서 자체적으로 발생하는 것을 '내재적 위험'으로 볼 수 있으며, 크게 학습데이터 위험, AI 모델·시스템의 기술적 위험, AI 모델·시스템의 이용자 위험으로 구분할 수 있다. 첫째, 학습데이터 위험은 편향된 데이터로 인해 AI 산출물에 차별과 오류가 발생하는 부적절한 데이터, 법률로 보호되는 개인정보와 저작권 등을 적법한 절차를 거치지 않고 학습에 사용하는 부적법한 데이터에서 발생한다. 둘째, AI 모델·시스템의 기술적 위험은 AI 개발 및 운영 과정에 대한 투명성·설명가능성 부족, AI의 자율성에 대한 인간의 통제 곤란, AI 모델·시스템으로 인한 인간의 권리 침해 등이 주요 원인이다. 셋째, AI 모델·시스템의 이용자 위험은 AI 입력 프롬프트에 개인정보 등을 입력하는 부주의한 이용자, 허위 정보 및 딥페이크를 생성하여 사회 혼란을 초래하는 악의적 이용자, 프롬프트 조작과 데이터 탈취 등을 통해 AI 시스템의 보안을 위협하는 사이버 공격자로부터 발생한다.

이러한 위험 요소는 AI가 사회·경제에 미치는 파급효과를 넘어 AI 기술 자체의 신뢰성과 안전성 문제로 직결된다. 따라서 AI의 안전한 발전과 지속 가능한 활용을 위해서는 명확한 규제와 정책적 대응이 필요하다. 그 대안으로는 첫째, AI 학습데이터의 출처·편향성을 검토하고 투명성을 강화하기 위하여 학습데이터에 대한 설명을 강화하고, 이를 위해 현행 인공지능기본법안에 열거된 고영향 AI뿐만 아니라 생성형 AI와 같은 일상에 큰 영향을 미치는 AI까지 학습데이터 설명 대상에 포함하는 방안을 모색해 볼 수 있다.

둘째, AI 학습데이터에 개인정보와 저작물이 활용되고 있는 현실을 감안하여, 온라인에 공개된 개인정보와 저작물의 보호와 AI 학습을 위한 활용을 균형적으로 달성할 수 있는 「개인정보 보호법」과 「저작권법」의 개정을 모색해 볼 수 있

다. 「개인정보 보호법」 관련 외국 입법례는 공개된 개인정보는 법률의 적용대상에서 제외하거나 동의 없이 사용할 수 있도록 한 미국 캘리포니아주의 「프라이버시 권리법」, 캐나다의 「개인정보 보호 및 전자문서법」, 유럽연합의 「일반 정보보호 규정」 등이 있다. 「저작권」 관련 외국 입법례는 공정이용(fair use)을 허용하고 있는 미국 「연방저작권법」, 학술연구 목적의 TDM과 일반적인 TDM을 구분하여 이원적으로 규율하고 있는 유럽연합의 「유럽연합 DSM 저작권 지침」 등이 있다.

셋째, 영어권 데이터를 학습한 AI 모델·시스템이 국내 현실에 맞는 결과를 충분히 산출하지 못하는 문제를 해결하기 위해서 국내 정치·경제·사회·문화 데이터를 축적하여 AI 학습에 활용할 수 있도록 하는 노력이 필요하다.

넷째, AI 모델·시스템 프로세스의 투명성·설명가능성을 높이기 위해서 AI 기술문서 작성·보관 의무 대상을 현행 인공지능기본법안에 열거된 고영향 AI뿐만 아니라 생성형 AI와 같은 일상에 큰 영향을 미치는 AI까지 포함하는 방안을 모색해 볼 수 있다. 또한 AI 모델을 어떻게 사용되어야 하는지에 대한 설명을 제시하는 것도 적절하다.

다섯째, AI 모델·시스템 프로세스의 위험은 인간의 기본권에 대한 위험뿐만 아니라 정치·사회·경제 등 다양한 분야에서도 발생하므로, 고영향 AI에 대해서 기본권 영향평가를 실시하도록 규정하고 있는 현행 인공지능기본법안의 범위를 기본권 이외의 위험 요소까지 확장하는 방안을 모색해 볼 수 있다. 다만, 의무적으로 강제할 경우 기업의 부담이 증가하므로, 기업이 자율적으로 선택할 수 있는 AI 윤리영향평가를 적극 활용하는 방안도 고려해 볼 만하다.

여섯째, AI 모델·시스템 이용자가 중요 정보를 AI에 입력하지 않고, AI가 제시하는 결과를 과잉신뢰하지 않으며, 다양한 상황에서 AI 산출물을 비판적으로 활용할 수 있도록 이용자의 AI 리터러시(literacy)를 함양할 필요가 있다. 인공지능기본법안에서는 안전하고 신뢰할 수 있는 AI를 위한 교육·홍보 등 리터러시 관

련 규정이 다소 추상적으로 규정되어 있으므로 이를 보다 구체적으로 정하는 방안을 모색해 볼 수 있을 것이다.

일곱째, AI 모델·시스템에 대한 사이버 공격을 AI 개발자 모두 예상하여 대응하는 것은 한계가 있으므로 가상의 적대적 테스트(레드티밍) 등을 활성화하여 대응 방안을 마련하고, 보안에 관한 국제협력에 적극 참여할 필요가 있다.

이러한 위험관리가 잘 추진되기 위해서는 민·관 협력이 강화되어야 한다. 정부와 기업이 협력하여 자율규제와 공적규제의 균형을 유지하며 AI 안전성을 제고해야 한다. AI 안전 확보는 기업 입장에서는 강력한 규제적 요인이 되므로 민·관 협력을 통해 기업의 규제 수용성을 높일 수 있다. 다음으로, 국제협력을 확대해야 한다. 글로벌 AI 윤리기준과 안전 규제 체계를 국제적으로 공유하고 협력해야 한다. 마지막으로, 후속 입법·정책에 대한 지원이 필요하다. 인공지능기본법안 제정 이후 각 분야별로 AI 진흥 및 규제 입법이 등장할 것으로 전망된다. 인공지능기본법안을 보충할 사항, 개별 법률에서 정할 사항, 정책적으로 추진될 사항 등이 적절하게 배분되어야 할 것이다. 다만, 이러한 노력이 각 부처·기관에서 경쟁적으로 추진될 경우 국가적으로 규제 중첩과 과잉이 발생할 우려가 크다. 따라서 최종 컨트롤타워인 국가인공지능위원회의 정책 조정 역할을 강화해야 하고, 이러한 조정이 입법 과정에도 원활하게 반영될 수 있도록 국회에서도 AI 관련 특별위원회 또는 지원체계가 마련될 필요가 있다.

참고문헌

개인정보보호위원회, 「인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서」, 2024.7.

과학기술정보통신부·정보통신정책연구원, 「2024 인공지능 윤리·신뢰성 포럼 공개세미나 자료집」, 2024.11.28.

김태순, 「美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」분석 및 시사점」, THE AI REPORT, 2023-7호, 한국지능정보사회진흥원, 2023.

노재원·유재흥·장진철·조지연, 『AI위험 유형 및 사례 분석』, SPRi 이슈리포트 IS-183, 소프트웨어정책연구소, 2024.

신용우 외, 「인공지능 윤리와 저작권의 규제체계 연구」, NARS 정책연구용역보고서, 국회입법조사처, 2024.

이성엽(편), 『생성형 AI와 법』, 박영사, 2024.

이정아, 「미국의 인공지능(AI) 정책·전략 현황과 변화 방향 - AI 지배력 강화와 초강대국 유지를 위해 미국은 어떻게 변화하고 있는가?」, The AI Report, 2024-3호, 한국지능정보사회진흥원, 2024.

이해원, 「유럽연합 인공지능법: 주요 내용과 시사점」, 이슈페이퍼 2024-04호, 스타트업 얼라이언스, 2024.

장길수, 「AI 생성 텍스트 쓸수록 '모델 붕괴' 위험 높아진다」, 『로봇신문』, 2024.8.2., <https://www.irobotnews.com/news/articleView.html?idxno=35692>

장유미, 「'구글 대항마'로 떠오른 美 퍼플렉시티, 저작 권 침해 문제로 '몸살'」, 지디넷코리아, 2024.6.10. <https://zdnet.co.kr/view/?no=20240610093432>

정준화, 「정보통신망 이용촉진 및 정보보호등에 관한 법률 일부개정법률안(의안번호 2126048) 입법영향분석 - 인공지능으로 만든 가상 정보임을 표시하도록 하는 의무」, NARS 입법영향분석 기획보고서, 제7호, 국회입법조사처, 2024.

채은선, 『EU 인공지능법 안내서』, 한국지능정보사회진흥원, 2024.

최경진·강지원·김법연·김형준·박신욱·오정익·채기현·채은선, 『EU 인공지능법』,

박영사, 2024.

Organization for Economic Cooperation and Development(OECD), 『Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives』, OECD, 2024.

Stuart D. Levi et al, “Colorado’s Landmark AI Act: What Companies Need To Know”, Skadden Publication, 2024. 6. 24.

[인터넷 홈페이지 등]

개인정보 포털, 캘리포니아 CPRA 원본 및 번역본, (최종방문일: 2024.12.16.), <https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000843184&fileSn=0>

미국 콜로라도주 의회 홈페이지, (최종방문일: 2024.12.16.), <https://leg.colorado.gov/bills/sb24-205>

유럽연합 법령정보시스템 홈페이지, (최종방문일: 2024.12.16.), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32019L0790>

유럽연합 GDPR TEXT 홈페이지, (최종방문일: 2024.12.16.), <https://gdpr-text.com/ko/read/article-9/#comment_gdpr-a-09_2e>

정보통신정책연구원 인공지능 윤리 소통채널 홈페이지, “인공지능 윤리기준”, (최종방문일: 2024.12.16.), <https://ai.kisdi.re.kr/aieth/main/contents.do?menuNo=400029>

캐나다 법률정보시스템, (최종방문일: 2024.12.16.), <https://laws-lois.justice.gc.ca/eng/acts/p-8.6/page-1.html#h-416969>

IBM, “AI risk atlas”, (최종방문일: 2024.12.16.), <https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/hallucination.html?context=wx&locale=ko&audience=wdp>

[부록]

1. OECD가 제시한 10가지 AI 위험

위험(10)	내용
고도화된 악의적 사이버 활동의 촉진 (Facilitation of increasingly sophisticated malicious cyber activity)	• AI 시스템은 악의적 사이버 활동의 발생 빈도와 심각도를 증가시키고 있음. AI 도구는 악의적 행위자들의 기술적 장벽을 낮추고, 더 정교한 공격을 실행할 수 있는 역량을 제공. 이전에는 인간 전문가가 오랜 시간과 노력을 들여야 가능했던 악성 사이버 활동이 AI를 통해 더욱 쉽게 이루어짐 • 특히, 대형 언어 모델(LLM)을 활용한 소프트웨어 취약점 탐지 및 악성 코드 생성이 주목받고 있음. 생성형 AI는 피싱 이메일, 악성코드, 랜섬웨어와 같은 공격 유형의 증가에 기여한 것으로 보고됨 • 디지털 생태계와 인프라의 안정성에 대한 신뢰가 약화되고, 악의적 활동이 사회적·물리적 피해를 초래하며 경제와 안보에도 부정적인 영향을 미칠 수 있음
조작, 허위 정보, 사기 및 민주주의와 사회적 결속에 대한 피해 (Manipulation, disinformation, fraud and resulting harms to democracy and social cohesion)	• AI 시스템은 매우 설득력 있는 허위 정보를 생성할 수 있으며, 이미 민감한 정치적 문제에 대한 AI 생성 허위 정보가 널리 퍼진 사례가 있음 • 인간이 AI로 생성된 텍스트를 구별하는 능력에 대해 혼합된 결과가 있으며, 일부 연구에서는 AI로 생성된 얼굴이 실제 얼굴보다 더 "진짜" 처럼 보인다고 평가받은 경우도 있음 • AI는 맞춤형 설득 메시지를 대량으로 생성 및 배포하거나 개인 맞춤형 심리 프로파일링을 통해 사람들을 대규모로 조작하거나 영향을 미칠 가능성이 있음 • 잘못된 정보와 AI 기반 조작은 민주적 과정과 사회적 결속을 훼손할 수 있습니다. 예를 들어, 선거 시스템에 영향을 미치거나, 거짓 서사를 유포하여 신뢰를 무너뜨리는 상황이 발생할 수 있음 • AI 생성 콘텐츠의 증가로 인해 온라인 정보의 질이 저하되고, 이는 잘못된 정보를 확산시킬 수 있음
AI 시스템 개발 및 배포 경쟁으로 인한 안전성과 신뢰성 부족 (Races to develop and deploy AI systems cause harms due to a lack of	• AI 개발 기업들 간의 경쟁이 심화되면서 신제품 및 서비스의 빠른 출시 압박이 증가하고 있음. 이로 인해 윤리적이고 안전한 AI 시스템 개발에 필요한 투자와 주의가 부족해질 수 있음 • 경쟁적 "경주"(race dynamics)는 기업뿐 아니라 국가 간에서도 발생할 수 있으며, 이는 기술 개발에 있어 단기적 이익을 우선시하게 만들 수 있음 • 이러한 경쟁은 초기 시장 진입(first-mover advantage)와 "승자독식"(winner-takes-all) 현상을 더욱 부추겨, 성능과 속도를 우선시하는 환

위험(10)	내용
sufficient investment in AI safety and trustworthiness)	경을 조성할 위험이 있음 • 사용자와 조직들이 신기술을 빠르게 채택하려는 압박도 이 문제를 악화시키고 있음 • 안전성과 신뢰성을 보장하기 위한 자원(예: 인재, 컴퓨팅 자원 등)의 부족은 장기적인 사회적 목표에 부정적인 영향을 미칠 수 있음 • 국가 간 경쟁은 국제 협력, 지식재산권 보호, 규제 조화 등의 이슈에서 갈등을 유발하고 심지어 국제적 긴장을 초래할 수 있음
인간 이해관계자의 선호와 가치에 부합하도록 AI 시스템의 목적을 조정하는 부적절한 방법에서 발생하는 예상치 못한 피해 (Unexpected harms result from inadequate methods to align AI system objectives with human stakeholders' preferences and values)	• AI 시스템의 목표는 명시적 또는 암묵적으로 설정될 수 있으며, 이를 인간의 진정한 의도와 일치시키는 것이 어려움 • 목표가 높은 수준에서 개념적이거나 대리(proxy) 목표로 설정될 경우, 예상치 못한 결과를 초래할 수 있음 • AI가 스스로 "자기 보존"이나 "자기 개선"과 같은 목표를 개발하거나, 자원을 비효율적으로 사용하는 문제가 발생할 가능성이 있다고 일부 전문가들은 우려함 • AI 시스템이 점점 더 많은 사회적, 경제적 시스템에 통합되면서 부조화 문제는 확대될 수 있음
소수 기업 및 국가에 집중된 권력 (Power is concentrated in a small number of companies or countries)	• 고급 AI 시스템 개발에 필요한 데이터와 컴퓨팅 자원이 소수의 주요 기업과 국가에 의해 통제되고 있음 • 대형 기술 기업들이 네트워크 효과와 대규모 자본을 기반으로 시장 지배력을 강화하고 있음 • 이러한 기업과 국가들이 국제적 정책 및 규범에 큰 영향을 미침 • 소규모 기업과 신흥 국가가 AI 시장에서 경쟁하기 어려운 구조가 형성되고 있으며, 특정 기업과 국가에 경제적, 정치적 권력이 과도하게 집중될 위험이 있음 • (공공 이익 침해) AI 기술이 공공의 필요보다 기업의 이익에 초점이 맞춰질 가능성이 큼 • (정보 비대칭성) 주요 AI 제공업체에 대한 과도한 의존으로 투명성과 책임성이 부족해질 수 있음
중요한 시스템에서의 AI 사고와 재난 (Minor to serious AI	• AI는 항공 교통 관제, 금융, 핵, 군사 시스템 등 다양한 중요 시스템에 점점 더 많이 도입되고 있음 • AI 시스템이 예상치 못한 방식으로 신뢰할 수 없거나 적절히 검증되지 않은 경우, AI 공급망의 복잡성으로 인해 기반 모델 제공자와의 의존성

위험(10)	내용
incidents and disasters occur in critical systems)	이 커져 여러 시스템에 걸쳐 동시다발적인 실패를 초래할 가능성도 있어 심각한 위험을 초래할 가능성이 큼 • (신뢰성과 검증 부족) 중요 시스템에서 사용되는 AI가 충분히 신뢰할 수 없거나 적절히 검증되지 않으면, 사고 및 재난으로 이어질 가능성이 큼 • (연쇄적인 실패 가능성) AI 공급망의 복잡성으로 인해, 하나의 문제가 다른 시스템으로 확산되면서 대규모 피해를 초래할 수 있음 • (규제 및 관리의 부재) 금융 시장과 같은 일부 분야에서 AI 사용이 증가함에 따라, 적절한 규제가 마련되지 않으면 금융 위기와 같은 심각한 사태가 발생할 가능성이 있음 • (예측 불가능성) AI 시스템이 복잡한 상황에서 예기치 못한 방식으로 작동하거나, 설계된 목적과 다르게 행동할 위험 존재
감시 강화 및 사생활 침해 (Invasive surveillance and privacy infringement)	• AI 시스템은 개인정보 수집, 분석, 그리고 예측을 강화하는 도구로 사용될 수 있음. 이로 인해 사람들의 개인적인 정보가 대규모로 수집되고, 그 정보들이 민감한 방식으로 처리될 수 있음 • (프라이버시 침해) AI 시스템이 대량의 개인 데이터를 수집하고 이를 분석하여 사람들의 행동, 선호도, 심리 등을 예측할 수 있음. 이는 개인의 동의 없이 이루어질 경우 심각한 프라이버시 침해로 이어질 수 있음 • (감시 사회의 위험) AI를 활용한 감시는 시민들의 일상적인 행동을 추적하고 감시하는 체제로 발전할 수 있으며, 이는 민주주의와 자유를 위협할 수 있음 • (정보 왜곡) AI가 개인의 데이터를 기반으로 예측을 내리는 과정에서, 특정 집단이나 개인에 대한 편향된 결론을 내릴 위험이 존재함. 이는 개인의 권리를 부당하게 침해할 수 있음
빠른 AI 발전 속도에 대응하지 못하는 거버넌스 메커니즘 (Governance mechanisms and institutions unable to keep up with rapid AI evolutions)	• AI 기술의 발전 속도가 매우 빠르지만, 기존의 거버넌스 체계는 이러한 급격한 발전에 적절하게 대응하지 못하고 있음 • 특히, AI의 사용이 빠르게 확산됨에 따라 규제와 정책이 뒤따르지 못하고 있으며, 이에 따라 발생할 수 있는 다양한 사회적, 경제적, 윤리적 문제들에 대해 미리 준비된 대응책이 부족한 상태 • (제한된 거버넌스 역량) AI 발전 속도를 따라가지 못하는 규제 시스템은 AI의 안전성, 신뢰성, 그리고 윤리적 문제들을 제대로 관리할 수 없음. 특히, AI의 자율성과 학습능력이 강해지면서 시스템의 예측 불가능한 결과를 감지하거나 대응하기가 점점 어려워지고 있음 • (국제 협력 부족) 각국은 AI 개발과 관련된 거버넌스 문제를 다루고 있지만, 국가 간 협력은 부족하며, 각국의 규제 체계가 상이해 AI의 글로벌한 안전성을 확보하는 데 어려움이 있음. 이러한 차이는 AI 기술의 악용을 방지하기 위한 국제적인 기준 마련을 어렵게 함

위험(10)	내용
설명가능성과 해석 가능성이 부족으로 인한 AI 시스템의 책임성 악화 (AI systems lacking sufficient explainability and interpretability erode accountability)	• AI 시스템의 결정 과정이 불투명하고, 사용자가 시스템의 결과를 이해하기 어려운 경우가 많음. 특히, 딥러닝 모델과 같은 복잡한 AI 시스템은 인간이 그 내부 작동 원리를 쉽게 해석하거나 설명할 수 없는 경우가 많음 • AI 시스템의 결정이 자동화된 의사결정 시스템에서 중요하게 작용하는 분야(예: 금융, 건강 관리, 법률 등)에서 문제가 될 수 있음. 이러한 시스템의 불투명성은 책임 소재를 흐리게 하고, AI의 잘못된 판단에 대해 누가 책임을 질 것인지 명확하지 않게 함 • AI의 불투명성은 특히 규제 및 법적 문제를 일으킬 수 있음. 만약 AI 시스템이 잘못된 결정을 내렸다면, 그 결정이 왜 그렇게 내려졌는지 설명하기 어려워 책임을 묻는 데 한계가 발생 • 신뢰 부족이 발생할 수 있으며, 이는 특히 AI 시스템이 중요한 결정에 영향을 미칠 때 사람들의 신뢰를 잃는 결과를 초래할 수 있음
국가 간 및 국가 내 불평등 및 빈곤 심화 (Exacerbated inequality or poverty within or between countries)	• (불평등의 심화) 현재 AI는 여러 국가와 지역 간의 불평등을 심화시킬 수 있음. 특히, AI 개발과 활용에 있어서 선진국과 개발도상국 간의 격차가 커지고 있음. 예를 들어, 고급 기술과 AI 연구에 접근할 수 있는 자원이 풍부한 국가들이 AI의 혜택을 먼저 보고 있으며, 그 반대로 자원 부족이나 기술적 접근 제한이 있는 개발도상국은 그 혜택을 받기 어려움 • (AI의 불균등한 배분) AI 기술은 주로 선진국과 일부 대기업에 집중되고 있어, 경제적 기회와 AI를 통한 생산성 향상의 혜택이 특정 지역에 집중되는 경향이 있음. 이는 이미 존재하는 경제적 불평등을 더욱 심화시킬 위험을 내포하고 있음 • (디지털 격차) 일부 국가나 지역은 AI 관련 인프라나 교육, 연구 개발에 필요한 자원을 충분히 확보하지 못해 디지털 격차가 발생함. 이는 AI 기술을 도입하려는 의지나 능력이 부족한 국가에서 사회적, 경제적 발전이 더딘 현상을 초래할 수 있음 • (고용 시장 불균형) 선진국에서는 AI가 고급 일자리를 창출하거나 기존 일자리를 보완하는 형태로 작용할 수 있지만, 개발도상국에서는 AI가 기존 일자리를 대체하거나 기술의 발전 속도에 뒤처지는 상황이 발생할 수 있음

자료: Organization for Economic Cooperation and Development(OECD), 『Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives』, OECD, 2024, 재정리.

2. SRRi가 제시한 25가지 AI 위험

대분류	위험(25)	내용
악의적 사용 위험	가짜 콘텐츠를 통한 개인의 피해 (사기 및 개인 사생활·명예 침해 등에 악용될 가능성)	• 특정 개인이나 그룹을 대상으로 맞춤 설계된 사기 콘텐츠를 대량으로 생산하여 피해 유발 • 이용자를 속여 실제 사람인 것처럼 가짜 신원을 생성, 도용하여 불법적인 목적으로 사용 - 생성형 AI로 실제 인물의 목소리나 외모를 모방하여, 그 사람인 것처럼 사칭하고 실시간으로 행동하여 타인을 속이는 악용 방식이 있으며, 행동의 묘사와 유사성으로 청중을 쉽게 오도할 수 있어 높은 피해를 유발 - 생성형 AI를 사용하여 원본 작품이나 브랜드, 스타일을 모방하여 가짜 제품이나 콘텐츠를 제작하고, 진짜처럼 속이는 방식으로 평판 좋은 뉴스 웹사이트를 도용하는 사례 등 위조품 제작(Counterfeit)도 성행 • 사용자 동의 없이 음란한 콘텐츠를 제작하는 딥페이크 포르노 문제도 발생 - 딥페이크 기술의 확산으로 '합의되지 않은 합성된 친밀한 이미지'(non-consensual synthetic intimate imagery, NSII)의 무분별한 배포에 대한 우려 발생
	허위 정보 및 여론 조작 (소셜 미디어를 통해 허위 정보가 배포되나, 기술적 대책에는 한계)	• 텍스트는 물론 이미지, 오디오 및 비디오를 생성하여 인간이 생성한 자료와 구분이 어려울 정도로 고도화되어 대규모 유포 - 생성형 AI로 생성된 이미지가 조작된 사건, 장소 또는 사물을 실제처럼 표현하여, 실제 사건처럼 공유되는 등 위조(Falsification) 사례 발생 • 챗GPT 등 생성형 AI를 활용하여 생산된 허위 사실은 소셜 미디어의 가짜 계정을 통해 대량 배포 - Yang & Menczer (2023)의 연구에 따르면 Twitter에 존재하는 악성 소셜 봇 계정 1,140개를 통해 의심스러운 웹사이트를 홍보하고 답글과 리트윗으로 유해한 댓글이 확산
	사이버 범죄 (대량의 사이버 공격을 통한 보안 위험의 악화)	• 시스템이 AI에 의해 사이버 보안 공격에 취약하도록 설계되어 사이버 공격에 악용될 우려 제기 - AI를 코딩 보조 도구로 활용했을 때 소프트웨어의 취약성이 발생할 수 있으며, AI가 자율적으로 웹사이트 해킹과 같은 작업을 수행하기도 함 • 반면, AI를 사용하여 취약점 및 공격 탐지 등 방어적 사이버 역량을 개선하기 위한 시도도 이루어지고 있음 - 인간이 보안 취약성을 식별하고 수정하는 데 소요되는 시간을 단축시키거나, 보안 패치를 구현함에 있어 AI를 보조적으로 활용하는 사례도 존재

대분류	위험(25)	내용
	이중 사용 과학 위험 (생물학, 화학, 방사능 및 핵무기 분야의 AI사용으로 인한 위험우려)	• 범용 AI는 새로운 과학자를 양성하거나 연구 워크플로우를 개선하는 등 과학적 발전을 가속화 할 수 있으나, 다양한 분야에서 악의적인 목적으로 활용될 가능성을 배제할 수 없음 - 즉, 일반 사용자가 범용 AI를 통해 과학적 지식이나 지침 등 정보에 대한 접근성이 증가함에 따라, 특정 정보를 활용하여 악의적으로 사용할 가능성도 존재
오작동 위험	제품 기능 문제로 인한 위험 (제품에 대해 잘못된 정보로 오해·혼동하여 발생)	• 비현실적인 기대로 AI 시스템에 과도하게 의존하거나, 예상 기능을 AI가 제공하지 못함으로써 발생하는 잠재적 피해 - 법조계, 의료계 등 전문 분야에서 범용 AI의 능력을 과대평가하여 실무에 적용 시 적지 않은 오류가 발생
	편견 및 대표성 위험 (인구통계학적 편향은 의료,채용,금융 등의 분야에 위험 초래)	• 왜곡되고 편향된 학습데이터에 의한 범용 AI의 의사결정은 공정성에 해를 끼칠 우려가 발생 - 학습 과정에서의 인종적 편견에 의해 의료 분야 시스템이나 안면 인식 알고리즘 등에 오류 야기 - 범용 AI 및 인터넷 검색에서 성 차별적 및 성 고정관념적 콘텐츠, 장애가 있는 사용자를 차별하는 콘텐츠 발견 • 편향을 해결하는 방안으로 사용자 피드백에 의한 강화학습(Reinforcement Learning from Human Feedback, RLHF)이 제안되고 있으나, 사용자의 피드백이 일관적이지 않은 문제로 인해 야기되는 편향성이 높아 이와 관련한 더 많은 연구를 요구
	통제력 상실 (잠재적인 미래 시나리오로 AI모델이 통제되지 않는 위험에 대한 우려)	• AI 에이전트의 개발 속도가 빨라짐에 따라, 의도하지 않은 결과로 인해 발생하는 위험에 대한 우려 목소리가 증가 - AI 시스템에게 국방, 공공 정책 등 중요한 책임과 의사결정을 맡겼을 때, 인간이 과도하게 AI 시스템에 의존함에 따라 통제 및 감독의 어려움이 발생할 가능성 존재 - 중요 분야에서 AI 기술 기반 의사결정 지원 시스템을 도입하였을 때, AI의 재량권과 인간의 통제 정도를 분석하여 인간의 책임 범위에 대해 중요하게 고려해야 함
	모델 붕괴 (합성 데이터 사용 증가로 급격한 모델 성능 저하 가능성 존재)	• 학습데이터의 부족 또는 생성 데이터의 확산에 따라 합성 데이터 또는 생성 데이터가 학습에 활용되는 경우가 되풀이되면서 AI 모델의 성능이 급격히 저하되는 모델 붕괴 가능성 제기 - 많은 테크기업들이 자신의 AI 모델에 더 많은 양의 데이터를 투입해 성능을 개선해 왔으나 데이터 수급 비용의 증가와 인간이 생산한 콘텐츠의 고갈*로 인해 보다 경제적인 합성 데이터를 사용하는 방법을 고려

대분류	위험(25)	내용
		- AI 연구기관 에포크 AI(Epoch AI)에 따르면 2026년에서 2032년 AI 학습용 데이터가 고갈될 것으로 전망 - 영국 옥스퍼드 대 연구진은 웹에서 수집된 대규모 데이터로부터 학습이 지속되면 모델 붕괴의 가능성으로 이어질 수 있음을 검증한 논문을 네이처에 게재('24.7) ㅇ 연구진은"시간이 지남에 따라 모델은 기본적으로 오류만 학습하고, 오류들은 계속 쌓인다"고주장, - 하니 패리드(Hany Farid) UC버클리대 교수도 이 문제가 마치 생물 종의 '근친 교배'와 유사하다고 지적하며 유전자 풀이 다양하지 않을 경우 종의 붕괴로 이어질 수 있다고 경고
시스템적 위험	노동시장 위험 (AI기반의 자동화로 인한 변화는 노동시장에 단·장기적 영향 우려)	• 범용 AI는 매우 광범위한 작업을 자동화할 수 있는 능력을 갖추고 있어 작업자의 업무의 질과 생산성을 향상시키고 새로운 일자리의 창출 등 긍정적 영향을 미칠 수 있음 • 한편, 특정 영역에서는 잠재적인 일자리 손실 및 대체 등 노동시장에 부정적 영향을 미칠 가능성이 존재 - 다양한 산업에서 범용 AI 시스템의 활용·도입으로 인한 기존 서비스에 대한 수요 감소가 야기 - 새로운 기술 등장에 따라 발생 가능한 노동 시장에서의 마찰은, 단기적으로는 실업이 유발될 수 있으며, 아직까지 임금에 대한 영향은 불확실하지만 장기적으로 소득 불평등을 높일 수 있음 • 산업연구원(2024)은 국내 사람의 일자리 327만 개를 AI가 대체할 것이라는 전망을 발표하는 등 AI의 노동 대체 가능성을 제시 - 민순홍 외 (2024.3), AI시대 본격화에 대비한 산업인력양성 과제- 인공지능 시대 일자리 미래와 인재양성 전략, 산업연구원
	글로벌 AI격차 (고급 AI개발 자원의 보유 여부에 따른 국가 간 격차 발생)	• 고급 AI 개발을 위한 다양한 필요조건들로 인해 AI 빅테크 기업들의 지배력 확대 - 디지털 기술 활용률, 컴퓨팅 자원 접근성, 인프라, 경제적 의존도 등은 범용 AI 개발 및 배포에 있어 중요한 요소로서, 이로 인해 범용 AI의 개발은 현재 서구 국가 및 중국에 집중된 경향 - 숙련된 인재 집중도의 불균형과 개발 및 유지에 드는 막대한 재정적 비용은 글로벌 AI 격차를 더욱 심화 • AI에 관한 연구는 미국과 중국이 주도하고 있으며, 주요 AI 선도국들은 연구 및 자원 보유 여부에 따라 AI 산업의 영향력을 보유 - 기계 학습 모델 개발은 미국, 캐나다, 영국 및 중국 등 일부 국가에 집중되어 있고, 대규모 언어 및 멀티모달 모델의 절반 이상은 미국에서 개발 중인 상황 - 평균 임금이 낮은 저소득 국가의 근로자들은 상대적으로 낮은 수준의 AI 작업(ghost work)을 수행

대분류	위험(25)	내용
	시장 집중및 단일 장애점 (소수기업으로의 시장 집중, 결함 발생시 광범위한 피해 가능성)	• 최첨단 AI 모델 개발을 위해서는 막대한 초기 비용이 요구되어 높은 진입 장벽을 형성 - 대규모의 범용 AI 모델을 구축할 수 있는 소수 기업들이 시장을 지배할 수밖에 없는 구조 • 금융, 사이버보안, 국방과 같은 주요 분야에서의 AI 채택은 모델 자체의 결함, 취약점, 버그 등으로 인해 동시적 피해 및 중단 유발 가능
	개인정보 보호 (기계학습에서 방대한 양의 데이터 의존 및 처리에 따른 개인정보 위험 초래)	• AI 모델 또는 시스템은 건강, 금융 데이터 등 민감한 개인정보를 학습하여 심각한 개인정보 유출 문제를 야기 가능 - 개인정보 보호는 자신의 민감한 정보 및 개인정보에 대해 다른 사람의 접근을 제어할 수 있는 개인의 권리를 의미함 - 또한 범용 AI 모델은 학습데이터를 기반으로 효율적인 검색을 지원하여, 개인의 민감한 정보에 대한 추론을 가능하게 하여 남용될 가능성 존재
	저작권 침해 (AI모델의 훈련 과정에서 저작권이 있는 데이터의 무단 사용 위험)	• 범용 AI 모델은 일반적으로 온라인에서 소싱된 대규모 데이터셋으로 훈련되어, 저작권 위반, 창작자 보상 및 경제적 혼란 가능성 존재 - 저작권이 있는 데이터를 대규모로 사용할 경우, 창의적 표현에 대한 보상 문제를 비롯하여 기존 지식재산권법, 데이터 사용 동의, 보상 및 관리 시스템과 충돌 발생 - 불분명한 저작권 제도는 AI 개발자가 데이터 투명성을 개선하는 작업을 더 어렵게 만들 수 있고, 데이터 소싱 및 필터링 인프라가 제대로 갖춰지지 않아 개발자가 저작권법을 준수하는데 한계가 있음
	환경 위험 (AI의 개발 및 배포 과정에서의 에너지 사용 증가로 환경 변화 우려)	• 개발 및 배포 과정에서 요구되는 컴퓨팅 파워에 대한 수요로 인하여 전력 소비량의 급격한 증가, 탄소 배출량, 냉각을 위한 물 소비 등이 발생하며, 이로 인해 환경적 영향의 고려가 필요 - 전력 수요 측면에서 AI 훈련에 막대한 전력이 소모되며, 이는 전체 데이터 센터 전력 소비에서 AI가 점유하는 비율의 증가로 이어져 온실가스 배출에 따른 환경 위험을 야기 - 탄소 배출량은 에너지 소비와 관련한 여러 요인에 따라 달라질 수 있으나, AI 훈련 과정에서는 여전히 고탄소 자원에 의존 • 산업의 지속가능한 성장을 위해 탄소중립의 중요성이 증대되고 있고, AI 반도체, 데이터센터 제조·운영 기업들은 이러한 수요에 대응책을 마련 중

대분류	위험(25)	내용
교차위험 ① 기술적 위험 요인	신뢰성 검증 및 보장	• 범용 AI 시스템의 출력은 일반적으로 자유 형식의 대화 또는 코드 등 개방적인 형식 - 가능한 모든 다운스트림 사용 사례에 대한 평가가 어렵고 탈옥(Jail-breaking)을 통한 유해한 요청 가능성 존재
	내부 작동방식 이해 부족	• 학습을 기반으로 주요 기능을 달성하는 범용 AI는 어떠한 입력으로부터 출력 및 결정에 도달하는지 내부적인 작동 방법에 대해 인간이 이해할 수 있는 설명 제공이 어려움 - AI 시스템의 작동 방식을 이해하기 위한 연구는 상대적으로 작고 능력이 떨어지는 간단한 모델에 국한되어 수행되다 보니, 큰 규모의 AI 모델의 작동을 이해하기 위해서는 더 많은 연구가 필요한 상황
	의도하지 않은 작동	• 범용 AI 시스템은 개발자의 테스트 또는 완화 조치 수행에도 불구하고 잠재적으로 유해한 출력을 생성할 가능성 존재
	배포의 신속성	• 많은 사용자에게 빠르게 배포될 수 있는 범용 AI 시스템은 모델의 결함 여부에 따라 광범위한 피해 야기
	위험 평가 및 방법 미흡	• 현재의 범용 AI에 대한 위험 평가와 평가 방법론은 미흡한 상태로 상당한 리소스 및 전문 지식이 필요
	유해한 작동 (오픈소스 모델 결함)	• 개발자는 디버그, 진단 등을 수행하더라도 모델의 유해한 행동을 완벽히 방지하기 어려움 - 개인 또는 저작권 정보의 공개, 혐오 발언 생성, 사회적·정치적 편향 및 편견, 유해한 작업 지원 등을 포함 - 특히, 소수의 독점적인 오픈소스 모델이 악의적으로 사용할 수 있는 결함 또는 기능을 갖추고 있을 때 대처하기 어려움
	자율성 강화에 따른 인간 통제 한계 (AI 에이전트)	• 범용 AI 에이전트는 많은 기능을 자율적으로 수행할 수 있으나, 인간의 감독이 적어짐에 따라 사고나 사용 위험을 증가시킬 수 있고, 복잡한 작업 수행에는 신뢰성이 부족하여 위험평가 방안 논의가 필요 - 인간 감독 감소로 인한 사고 위험 증가(악의적 사용 위험), 자동 워크플로우 허용(제품 기능 문제로 인한 위험), 통제력 상실 위험 등과 밀접한 관련
교차위험 ② 사회적 위험 요인	위험 완화 조치 미흡	• 신속한 시장 출시가 점유율에 많은 영향을 미치는 만큼 위험 완화 작업에 투자할 이점이 제한적 - 안전과 윤리 보장 조치에 대한 투자보다 가능한 빨리 개발하기 위해 경쟁하는 환경에 따라, 엄격한 안전 표준 준수의 필요성이 감소
	느린 규제 속도	• 기술 혁신의 속도와 거버넌스의 발전 속도의 불일치로 인한 규제 격차 발생 - AI 시장이 급격히 변화하는 만큼 규제 개선이 적절한 시기에 완료되기 어려운 상황으로, 범용 AI의 개발 및 배포 속도에 대응하기 위한 규제 환경을 조성할 필요

대분류	위험(25)	내용
	책임 소재 문제	• 투명성이 부족하여 범용 AI 시스템이 해를 끼쳤을 때, 책임 소재를 구분하기 어려워 거버넌스 및 집행을 방해할 가능성 존재 - 개발자가 명시적으로 범용 AI 시스템에서 새롭게 발생할 수 있는 행동을 명시해도, 이를 사용하는 과정에서 발생하는 피해에 대한 책임 소재는 현재의 법체계에서 구분하기 어려움 (투명성 부족 문제)
	배포에 대한 추적성 부족	• 범용 AI 모델과 시스템이 어떻게 훈련, 배포 및 사용되는 추적이 어려움 - 추적성의 확보는 책임 설정 뿐 아니라 악의적 사용 모니터링 및 입증, 오작동 감지에 중요하지만, 널리 수용되는 표준이 부재한 상황

자료: 노재원·유재흥·장진철·조지연, 『AI위험 유형 및 사례 분석』, SPRi 이슈리포트 IS-183, 소프트웨어정책연구소, 2024, 재정리.

3. IBM이 제시한 67가지 AI 위험

대분류	중분류		주요 내용(67)
입력과 관련된 위험 (Risks associated with input)	훈련 및 튜닝 단계 (Training and tuning phase)	견고성 Robustness	데이터 중독(Data poisoning) : 공격자 또는 악의적인 내부자가 의도적으로 손상되었거나, 허위·오해의 소지가 있거나, 잘못된 샘플을 학습 또는 미세 조정 데이터 세트에 포함시키는 것
		지식재산 Intellectual property	기밀정보(Confidential information in data) : 기관 또는 조직 내부의 기밀정보를 부적절하게 학습 또는 미세 조정 데이터 세트에 포함시키는 것
			법으로 보호되는 데이터의 부적절한 사용(Data usage rights restrictions) : 개인정보, 저작권 등 법으로 보호되는 데이터를 부적절하게 학습 또는 미세 조정 데이터 세트에 포함시키는 것
		정확도 Accuracy	대표성 없는 데이터(Unrepresentative data) : 모집단을 충분히 대표하지 못하는 데이터를 학습 또는 미세 조정 데이터 세트에 포함시키는 것
			데이터 오염(Data contamination) : 모델의 목적에 맞지 않는 잘못된 데이터를 학습 또는 미세 조정 데이터 세트에 포함시키는 것
		개인정보 보호 Privacy	개인정보 보호 조치(Data privacy rights alignment) : 옵트아웃·가명처리 등 데이터 주체의 권리 행사를 수용하기 위해 학습데이터 조정, 제거, 모델 재학습 등을 수행하는 과정에서 비용이 발생하는 것
			개인정보 보호(Personal information in data) : 학습데이터에 개인정보가 포함될 경우 법률이 정하는 바에 따라 개인정보 보호 조치를 취해야 하는 것
			개인정보 재식별(Re-identification) : 학습데이터에 비식별 정보를 이용하더라도, 다른 데이터와의 결합 과정에서 개인정보 재식별 가능성이 발생하는 것
		투명도 Trans -parency	데이터 투명성 부족(Lack of training data transparency) : 학습데이터의 수집·처리·활용 방법에 등에 대한 정확한 문서가 없으면 모델의 동작을 만족스럽게 설명하기 어렵고, 데이터와 모델이 갖고 있는 다양한 위험을 평가할 수 있는 능력이 제한됨
			데이터 출처 불확실(Uncertain data provenance) : 데이터의 소유권, 거래 등을 포함한 이력 정보가 불확실한 경우 데이터가 원래 출처와 동일한 사용 조건인지 보장할 수 없고, 데이터가 윤리적으로 수집·활용되고 있는지 검증할 수 없음

대분류	중분류		주요 내용(67)
		데이터 법률 Data laws	법률로 인한 데이터 활용 제한(Data usage restrictions) : 법률과 제도로 인해 특정 데이터의 활용을 제한하거나 금지함으로써 모델 학습이 충분히 이루어지지 못할 수 있음
			법률로 인한 데이터 수집 제한(Data acquisition restrictions) : 법률과 제도로 특정 데이터의 수집(웹 스크래핑, 웹 크롤링 등)을 제한하거나 금지함으로써 모델 학습이 충분히 이루어지지 못할 수 있음
			법률로 인한 데이터 전송 제한(Data transfer restrictions) : 법률과 제도로 특정 데이터의 전송을 제한하거나 금지함으로써 모델 학습이 충분히 이루어지지 못할 수 있음
		데이터 값 조정 (정제) Value alignment	부적절한 데이터 수집(Improper data curation) : 항목 오류, 상충되는 정보 포함, 중요 정보 누락 등을 정제하지않은 부적절한 데이터를 수집하여 학습한 경우 모델에 부정적인 영향을 줄 수 있음
			부적절한 데이터를 이용한 재학습(Improper retraining) : 학습이 완료된 모델에 사람의 검증을 충분히 거치지 않은 부적절한 데이터를 재학습을 시킬 경우, 모델 자체의 문제가 증폭되는 문제가 발생할 수 있음(*모델 붕괴 우려)
		공정성 Fairness	데이터 편향성(Data bias) : 데이터에 역사적·사회적 편견 등이 존재할 경우 이것이 모델에 반영되어 특정 그룹이나 개인을 부당하게 대표하거나 차별할 수 있는 편향되거나 왜곡된 결과를 도출할 수 있음
	추론 단계 (Inference phase)	개인정보 보호 Privacy	프롬프트 개인정보 입력(Personal information in prompt) : 프롬프트에 개인정보 또는 민감정보가 입력되어, 이것이 모델의 출력에 의도치 않게 공개될 수 있음
			속성 추론 공격(Attribute inference attack) : 악의적 이용자가 특정 속성에 대해 대답하도록 반복적으로 질의·요청하여 모델 학습에 포함된 개인정보 또는 지식재산권과 같은 중요 정보를 탈취하는 것
			맴버십 추론 공격(Membership inference attack) : 악의적 이용자가 모델이 어떤 데이터셋을 학습하였는지 알아내기 위해 특정 데이터셋의 샘플을 입력하고 출력값을 관찰하는 것을 반복하여 그 샘플이 어느 학습데이터의 일부(membership)인지 파악하는 것. 이를 통해 경쟁업체의 모델이 어떻게 훈련되었는지 알아내고, 모델을 복제하거나 조작할 수 있는 기회로 악용함

대분류	중분류		주요 내용(67)
		견고성 Robustness	프롬프트 누출(Prompt leaking) : 악의적 이용자가 모델을 공격하여 모델의 기본적인 명령어 체계(시스템 프롬프트)를 찾아내서, 이를 통해 모델에 포함된 중요 정보를 쉽게 탈취할 우려가 있음
			프롬프트 입력 공격(Prompt injection attack) : 악의적 이용자가 프롬프트에 포함된 구조, 명령어, 정보 등을 조작하여 모델 개발자가 예상하지 못한 출력을 생성하도록 강제하여 모델 동작을 변경할 우려가 있음
			추출 공격(Extraction attack) : 학습데이터에 포함된 개인정보와 같은 특정 정보를 추출해 내도록 프롬프트를 공격하는 것
			회피 공격(Evasion attack) : 프롬프트에 가짜문제를 입력하여 모델이 가짜문제를 회피하는 결정을 하도록 교란하여, 모델 안에 있는 중요한 진짜 정보를 유출하도록 만드는 것
		지식재산 Intellectual property	프롬프트 IP 입력(IP information in prompt) : 지식재산권(IP)이 있는 정보를 프롬프트에 입력하면, 이와 관련하여 모델 학습에 사용한 IP 정보가 출력되도록 하는 것
			프롬프트 기밀 데이터(Confidential data in prompt) : 이용자가 회사·기관의 기밀정보를 프롬프트에 입력할 경우, 이것이 모델에 기억되어 향후 모델의 생성물 출력에서 기밀정보가 나타날 우려가 있음
		다중 범주 Multi-category	탈옥(Jailbreaking) : 모델이 설정해 놓은 안전장치(가드레일)를 뚫고 금지된 작업의 수행을 시도하는 것으로, 모델의 안전성과 평판에 부정적 영향을 초래함
			프롬프트 프라이밍(Prompt priming) : 생성형 모델은 프롬프트 입력값과 유사한 출력값을 생성하는 경향이 있다는 것을 악용하여, 프롬프트에 개인정보 등 민감정보를 입력하여 모델이 그와 유사한 결과를 출력하도록 하여 개인정보 등을 유출하는 행위
		정확도 Accuracy	모델 정확도 불량(Poor model accuracy) : 모델 성능이 낮아서 정확도가 낮은 결과를 제시하는 것으로, 이 모델을 이용하는 이용자와 서비스·어플리케이션에 중대한 악영향을 초래함
출력과 관련된 위험 (Risks associated	결과값 관련 (Value alignment)		악성 출력(Toxic output) : 모델이 혐오, 욕설, 모욕 또는 외설적인 콘텐츠 등을 생성하는 경우
			유해한 출력(Harmful output) : 신체적 상해(폭력 방법 등), 사회적 혼란(폭발물 제조 등) 등을 유발하는 결과를 생성할 수 있음

대분류	중분류	주요 내용(67)
with output)		불완전한 조언(Incomplete advice) : 모델이 충분한 정보가 없는 상태에서 조언을 제공하고 그 조언을 따를 경우 피해를 입을 수 있는 경우(잘못된 의료, 재정, 법률 관련 조언 등)
		과잉 또는 과소 의존(Over or Under reliance) : 과잉 의존은 모델을 지나치게 신뢰하여 모델의 결과가 틀릴 가능성이 있는데도 모델의 결과를 그대로 받아들이는 경우. 과소 의존은 그 반대의 경우로, 모델을 신뢰하지 않지만 신뢰해야 하는 경우
	설명가능성 관련 (Explainability)	신뢰할 수 없는 소스 귀속(Unreliable source attribution) : 모델은 데이터(source)의 근사치를 근거(attribution)로 해서 결과를 도출하는데, 근사치 이외의 값이 더 적절한 어트리뷰션이 될 수도 있기 때문에 발생하는 문제
		설명할 수 없는 출력(Unexplainable output) : 모델 출력 결정에 대한 설명은 어렵거나 부정확한 경우로, 기반 모델은 복잡한 딥러닝 아키텍처를 기반으로 하므로 출력에 대한 설명이 어려움
		학습데이터 접근 불가능(Inaccessible training data) : 훈련 데이터에 액세스하지 않으면 모델이 제공할 수 있는 설명 유형이 제한되며 올바르지 않을 가능성이 높습니다. 소스 데이터가 없는 낮은 품질의 설명은 사용자, 모델 유효성 검증기 및 감사자가 모델을 이해하고 신뢰하기 어렵게 만듭니다.
		속성 추적불가능(Untraceable attribution) : 모델 출력 생성에 사용된 학습데이터의 콘텐츠에 액세스할 수 없고, 소스 어트리뷰션 기법을 사용할 가능성이 크게 제한되거나 불가능할 수 있음. 그 결과 사용자, 모델 검증자 및 감사자가 모델을 이해하고 신뢰하기 어려움
	개인정보 보호 관련 (Privacy)	개인정보 노출(Exposing personal information) : 학습데이터, 미세조정 데이터 또는 프롬프트의 일부로 개인정보 등이 사용되는 경우, 생성된 출력에서 모델이 해당 데이터를 표시할 수 있음
	오용 관련 (Misuse)	동의하지 않는 사용(Nonconsensual use) : 정보주체의 동의 없이 비디오, 이미지, 오디오 또는 기타 양식을 사용하여 딥페이크를 통해 사람을 모방하는 데 사용될 수 있음
		AI 활용 비공개(Non-disclosure) : 콘텐츠는 AI가 생성한 것으로 명확하게 공개되지 않음으로써 발생하는 문제

대분류	중분류	주요 내용(67)
		부적절한 사용(Improper usage) : 모델이 원래 설계되지 않은 용도로 사용될 때 발생. 원래 데이터, 디자인 목적 및 목표를 이해하지 않고 모델을 재사용하면 원하지 않는 방식으로 모델이 작동할 우려가 있음
		위험한 사용(Dangerous use) : 모델은 방대한 양의 데이터를 학습하기 때문에, 이 내용을 바탕으로 사람들에게 해를 끼칠 목적으로만 사용될 수 있음(AI 기반 사이버 공격 등)
		허위정보 유포(Spreading disinformation) : 생성형 AI 모델은 대상 고객을 속이거나 영향을 미치기 위해 의도적으로 오해의 소지가 있거나 잘못된 정보를 생성하는 데 사용될 수 있음. 허위 정보를 퍼뜨리면 사람들이 정보에 입각한 의사결정을 내리는데 영향을 미칠 수 있음
		악성 콘텐츠 확산(Spreading toxicity) : 생성형 AI 모델은 혐오, 욕설, 모욕 또는 음란 콘텐츠를 생성하고 확산하는데 의도적으로 사용될 수 있음
	공정성 관련 (Fairness)	의사결정 편향성(Decision bias) : 모델의 결정으로 인해 한 그룹이 다른 그룹보다 부당하게 유리한 문제가 발생. 이는 데이터의 편향으로 인해 발생할 수 있으며 모델 학습의 결과로 증폭될 수도 있음
		출력 편향성(Output bias) : 생성된 컨텐츠는 특정 그룹 또는 개인을 부당하게 나타낼 수 있음. 편향은 AI 모델의 사용자에 피해를 입히고, 기존의 차별적인 행동을 확대할 수 있음(출력물의 인종 편향성, 성 차별성 등)
	견고성 관련 (Robustness)	환각(Hallucination) : 모델의 학습데이터 또는 입력된 내용과 다르게 실적으로 부정확하거나 진실하지 않은 결과를 생성하는 문제. 환각은 이용자에게 오해를 불러일으킬 소지가 있고, 잘못된 내용이 다른 서비스와 후속 결과에 통합되어 잘못된 정보를 더욱 확산시킬 수 있음
	지식재산 관련 (Intellectual property)	기밀정보 노출(Revealing confidential information) : 기밀정보가 학습데이터, 미세조정 데이터 또는 프롬프트의 일부로 사용되는 경우, 향후 모델에서 생성한 출력결과에 해당 데이터가 표시될 우려가 있음
		저작권 침해(Copyright infringement) : 모델은 저작권으로 보호되거나 오픈 소스 라이선스 계약의 적용을 받는 기존 저작물과 유사하거나 동일한 콘텐츠를 생성할 수 있음
	유해한 코드 관련 (Harmful code generation)	유해한 코드 생성(Harmful code generation) : 모델은 손상을 유발하거나 실수로 다른 시스템에 영향을 주는 코드를 생성할 수 있음. 유해한 코드를 실행하면 IT 시스템의 취약성이 노출될 수 있음

대분류	중분류	주요 내용(67)
비기술적 위험 (Non-technical risks)	거버넌스 관련 (Governance)	모델 투명성 부족(Lack of model transparency) : 모델 설계, 개발 및 평가 프로세스에 대한 문서화가 불충분하고 모델의 내부 작동에 대한 정보가 충분히 제공되지 않을 경우 평가·모델변경·재사용이 어려워질 수 있고, 모델에 대한 신뢰성에 부정적 영향을 미칠 수 있음
		테스트 다양성 부족(Lack of testing diversity) : AI 모델 리스크는 사회 기술적인 문제이므로 다양한 관점에서 테스트를 거치지 않을 경우 위험성을 충분히 식별하지 못할 수 있음
		시스템 투명성 부족(Lack of system transparency) : 모델의 목적에 대한 문서가 불충분한 경우, 모델이 시스템이나 애플리케이션의 기능에 어떻게 기여하는지 파악하기 어려울 수 있음
		대표성 없는 위험 테스트(Unrepresentative risk testing) : 모델이 배포될 경우 발생할 다양한 위험을 가정한 위험 테스트를 거치지 않을 경우, 모델이 적용되는 상황을 대표하지 못하므로 위험을 정확하게 반영하지 못할 우려가 있음
		불완전한 사용 정의(Incomplete usage definition) : 기반 모델은 다양한 목적으로 사용될 수 있는데, 모델의 용도가 충분히 명시되지 않은 경우 모델의 관련 위험을 정확하게 판단해서 완화하기 어려울 수 있음
		데이터 투명성 부족(Lack of data transparency) : 학습데이터 또는 미세조정 데이터에 대한 세부설명이 부족할 경우, 모델의 위험성, 데이터 적법성 등에 대한 평가를 수행하는데 한계가 있음
		잘못된 위험 테스트(Incorrect risk testing) : 위험을 측정하거나 추적하기 위해 선택한 수단이 적절하지 못할 경우, 위험에 대한 이해가 부정확해지고 도출된 위험완화 조치가 작동하지 않을 우려가 있음
	사회적 영향 관련 (Societal impact)	인간에 미치는 영향(Impact on human agency) : 인공지능은 인간이 최선의 이익을 위해 독립적으로 선택하고 행동하는 능력에 영향을 미칠 수 있는데, 만약 AI가 허위 또는 오해의 소지가 있는 정보를 생성할 경우 사람들은 이를 잘못된 정보로 인식하지 못해 진실을 왜곡되게 이해하고, 잘못된 의사 결정을 할 우려가 있음
		문화적 다양성에 미치는 영향(Impact on cultural diversity) : AI는 보편적이고 균질적인 문화를 과도하게 대표하고, 소수 문화에 대한 반영을 줄여서 문화적 다양성을 충분히 반영하지 못할 우려가 있음

대분류	중분류	주요 내용(67)
		교육에 미치는 영향 - 표절(Impact on education: plagiarism) : 생성형 AI에 접근하는 학생들은 의도적이든 의도치 않게든 기존 작품을 표절할 우려가 있고, 남의 일을 자신의 일로 주장하는 비윤리적인 것을 당연하게 생각할 우려가 있음
		채용에 미치는 영향(Impact on Jobs) : AI로 직업을 잃는 사람이 발생하고, 일자리를 잃으면 소득이 줄어들어 사회와 복지에 악영향을 미칠 수 있음
		인간 착취(Human exploitation) : AI 모델을 학습시키는 근로자에게 적절한 근무 조건, 공정한 보상, 정신 건강을 포함한 좋은 건강 관리 혜택이 제공되지 않아 AI가 인간을 착취하는 것 같은 양상을 초래할 우려가 있음
		특정 집단에 미치는 영향(Impact on affected communities) : AI로 인해 영향을 받는 커뮤니티와 충분히 소통하지 않을 경우, 해당 커뮤니티는 다양한 피해를 입을 수 있고 결과적으로 AI에 대한 사회적 반발을 초래할 우려가 있음
		교육에 미치는 영향 - 학습 회피(Impact on education: bypassing learning) : 학생들이 학습을 회피하기 위하여 AI를 이용할 우려가 있음
		환경에 미치는 영향(Impact on the environment) : AI, 특히 대규모 생성 모델은 탄소 배출량을 증가시키고 물 사용량을 증가시켜 환경에 부정적 영향을 미치고, 기후변화를 악화시킬 수 있음
	법률 준수 관련 (Legal compliance)	법적 책임(Legal accountability) : 문서화 및 거버넌스 프로세스가 제대로 갖춰져 있지 않으면 AI 모델에 대한 법적 책임자를 찾아서 책임을 묻기 어려울 수 있음
		모델 사용 권한 제한(Model usage rights restrictions) : 서비스 약관, 라이선스 또는 기타 규칙에 따라 특정 모델의 사용이 제한될 수 있음
		생성된 콘텐츠 소유권 및 IP(Generated content ownership and IP) : AI가 생성한 콘텐츠의 소유권 및 지식재산권에 대한 법적 불확실성이 큰 경우, AI 활용에 부정적 영향을 미칠 수 있음

자료: IBM, “AI risk atlas”, (최종방문일: 2024.12.16.), <https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/hallucination.html?context=wx&locale=ko&audience=wdp>, 재정리.

NARS 입법·정책 발간 일람

호 수	제 목	발간일	집필진
제001호	개헌 관련 여론조사 분석	2018. 03. 13.	허 석 재
제002호	빅데이터 정책 추진 현황과 활용도 제고방안	2018. 05. 31.	정 도 영 김 민 창 김 재 환
제003호	조세범에 대한 처벌 현황 및 개선방안	2018. 06. 22.	문 은 희
제004호	지역상생발전기금의 현황과 개선방안	2018. 06. 28.	류 영 아
제005호	현행 지방선거제도 관련 주요 쟁점 및 개편방안 : 지방의회선거를 중심으로	2018. 07. 11.	김 종 갑
제006호	디지털 증거에 관한 형사소송법적 과제 : 전문법칙을 중심으로	2018. 07. 26.	조 서 연
제007호	디지털 성범죄 대응 정책의 운영실태 및 개선과제	2018. 08. 08.	조 주 은 최 진 응
제008호	보호종료 청소년 자립지원 방안	2018. 09. 21.	허 민 숙
제009호	지방이전 공공기관의 지역정착 실태와 향후 보완과제	2018. 11. 15.	김 재 환 정 도 영 김 민 창
제010호	정보격차 해소를 위한 정보화교육사업 실태 및 개선방안	2018. 11. 29.	김 유 향 김 나 정
제011호	지역노사민정협의회의 운영실태와 개선방안	2018. 11. 29.	신 동 윤
제012호	연구개발특구의 운영실태와 개선방안	2018. 12. 07.	권 성 훈
제013호	지방자치단체의 공공데이터 개방 현황과 개선 과제	2018. 12. 10.	김 태 엽
제014호	현행 '복지허브화' 정책의 성과 및 개선방안 -'찾아가는 읍면동 주민센터'사업을 중심으로-	2018. 12. 11.	이 만 우
제015호	육아휴직 활성화를 위한 부모보험 도입방안	2018. 12. 13.	박 선 권
제016호	4차 산업혁명 대응 현황과 향후 과제	2018. 12. 13.	정 준 화
제017호	지방옴부즈만 제도의 운영현황 및 개선과제	2018. 12. 14.	김 현 정
제018호	국가 주요 시설물의 관리체계 개선을 위한 입법 및 정책 과제	2018. 12. 14.	김 진 수
제019호	양육비 이행 관리 제도의 문제점 및 개선과제	2018. 12. 17.	허 민 숙

호 수	제 목	발간일	집필진
제020호	트럼프 행정부의 대외정책 기조에 따른 한미동맹의 주요 현안 및 쟁점	2018. 12. 19.	김 도 희
제021호	개정 한·미 FTA 「투자자와 국가간 분쟁해결제도」(ISDS)와 향후 과제	2018. 12. 20.	정 민 정
제022호	기술탈취 방지 및 기술보호를 위한 입법·정책 과제 -입증책임 전환을 중심으로-	2018. 12. 24.	박 재 영
제023호	시진핑 집권2기 중국 대외정책 결정체계의 현황과 시사점	2018. 12. 27.	김 예 경
제024호	난민심사제도 운용실태 및 개선과제	2018. 12. 27.	백 상 준 김 예 경
제025호	남북 이산가족 관련 지원 정책의 실태 및 개선과제	2018. 12. 31.	이 승 현
제026호	독립법인보험대리점(GA)의 현황 및 개선과제	2019. 01. 18.	김 창 호
제027호	주민참여예산제도의 운영실태와 개선방안	2019. 09. 24.	류 영 아
제028호	지역아동센터 지원사업의 현황과 과제	2019. 10. 31.	박 선 권
제029호	CCTV 통합관제센터 운영실태 및 개선방안	2019. 11. 01.	최 미 경 최 정 민
제030호	공공와이파이 구축·운영 실태 및 개선과제	2019. 11. 15.	장 은 덕
제031호	지속가능한 지하수의 활용 및 관리 방안	2019. 12. 10.	김 진 수
제032호	기술평가제도 현황 및 활성화를 위한 과제	2019. 12. 16.	박 재 영
제033호	의약품 이상사례 보고제도의 점검 및 개선방안	2019. 12. 19.	김 은 진
제034호	초·중등 소프트웨어교육 운영실태와 개선과제	2019. 12. 23.	김 유 향 유 지 연 김 나 정
제035호	장애인의 지역 간 이동 편의 증진을 위한 교통 서비스 실태 및 개선방안	2019. 12. 24.	김 영 석 박 준 환 김 대 명
제036호	사업장 대기오염 총량관리제 현황과 개선방안	2019. 12. 26.	이 혜 경
제037호	도로 유지관리 현황 및 과제 -도로 자산관리를 중심으로-	2019. 12. 26.	구 세 주
제038호	형사 사건관계인의 알권리 실태조사 및 개선방안	2019. 12. 27.	백 상 준
제039호	일본 아베내각의 안보정책 변화 분석과 시사점	2019. 12. 27.	박 명 희
제040호	제1차 - 제10차 한미방위비분담특별협정의 주요내용 분석 및 정책적 시사점	2019. 12. 31.	김 도 희
제041호	상장회사 관련 현행 법체계의 문제점과 개선과제	2019. 12. 31.	황 현 영
제042호	국세상담센터의 운영현황과 개선과제	2019. 12. 31.	문 은 희

호 수	제 목	발간일	집필진
제043호	공정거래 분야의 집단소송제 도입 방안	2019. 12. 31.	강 지 원 조 영 은
제044호	수용자 가족·자녀 지원을 위한 입법·정책 과제	2020. 05. 22.	허 민 숙
제045호	국회 안건신속처리제의 운영현황과 개선과제	2020. 05. 30.	전 진 영
제046호	ILO 핵심협약의 비준현황과 과제	2020. 06. 24.	신 동 윤
제047호	철도 유휴부지 활용도 제고를 위한 입법 및 정책과제	2020. 06. 30.	구 세 주
제048호	인구감소시대 지방중소도시의 지역재생 방안	2020. 06. 30.	김 예 성 하 혜 영
제049호	자동차보험 한방진료의 현황과 개선과제	2020. 07. 10.	김 창 호
제050호	지역건축안전센터의 운영 실태와 개선과제	2020. 08. 07.	김 예 성
제051호	아동학대 대응체계의 과제와 개선방향 -아동보호전문기관을 중심으로-	2020. 08. 13.	박 선 권
제052호	외교부 영사콜센터 운영실태와 개선과제	2020. 08. 28.	김 예 경
제053호	대통령제 정부의 초당적 내각 구성 사례와 시사점	2020. 09. 01.	허 석 재
제054호	국회의원 선거제도 개편논의와 대안의 모색	2020. 09. 01.	김 종 갑 허 석 재
제055호	빅데이터 플랫폼의 운영 실태와 개선과제	2020. 09. 07.	정 준 화
제056호	형사사법공통시스템의 운영실태와 개선과제	2020. 09. 18.	박 혜 림
제057호	한반도 주변 경계미획정 수역에 대한 국제법적 쟁점과 대응과제	2020. 09. 21.	정 민 정
제058호	상속세 미술품 물납제도 도입을 위한 입법론적 검토	2020. 10. 07.	장 영 환
제059호	리쇼어링 기업 지원정책의 문제점 및 개선방안	2020. 10. 08.	김 종 규
제060호	2020 미국 대선 결과 분석	2020. 11. 26.	-
제061호	가정폭력 이혼 과정에서의 피해자 보호를 위한 입법과제 -자녀면접교섭을 중심으로-	2020. 12. 04.	허 민 숙
제062호	기후변화 대응 도시홍수 대책	2020. 12. 21.	김 진 수
제063호	조선산업 친환경·스마트화 동향과 입법·정책과제	2020. 12. 23.	김 봉 주
제064호	에너지공급자 수요관리 투자사업 현황과 개선과제	2020. 12. 29.	박 연 수
제065호	공공임대주택 공급동향 분석과 정책과제	2020. 12. 30.	장 경 석 송 민 경
제066호	농어촌 등 교통소외지역의 교통서비스 강화 방안	2020. 12. 30.	박 준 환 김 규 호
제067호	디지털 사이니지(Digital Signage) 정책 평가와 개선과제	2020. 12. 30.	최 진 응
제068호	제20대 국회 입법활동 분석	2020. 12. 30.	전 진 영

호 수	제 목	발간일	집필진
제069호	데이터 경제 활성화를 위한 입법정책 방안	2020. 12. 31.	신 용 우
제070호	공유경제 활성화를 위한 법·제도 개선방안	2020. 12. 31.	김 민 창 박 성 용
제071호	기부금품 모집·사용제도 현황과 개선방향	2020. 12. 31.	이 송 림 한 경 석
제072호	법학전문대학원 교육 내실화를 위한 입법·정책 과제	2020. 12. 31.	김 광 현 이 재 영 최 정 인
제073호	일본의 국제 활동 확대와 한국의 대응방향	2020. 12. 31.	박 명 희
제074호	동북아 미세먼지 협력 : 현황과 과제	2020. 12. 31.	이 혜 경
제075호	1회용 포장재 재활용 활성화를 위한 보증금제도 도입 방안	2020. 12. 31.	김 경 민
제076호	공익신고자 보호제도 현황과 개선과제	2021. 03. 31.	김 형 진 박 영 원
제077호	온라인 플랫폼 공정화법 제정을 위한 입법·정책과제	2021. 05. 10.	최 은 진 강 지 원
제078호	아동사망 예방을 위한 아동사망검토 제도 도입방안	2021. 05. 20.	박 선 권
제079호	홈리스 청소년 지원 입법 · 정책과제 : 가정복귀 프레임을 넘어	2021. 06. 04.	허 민 숙
제080호	재산세 제도의 현황과 쟁점	2021. 06. 17.	류 영 아
제081호	대안교육기관 관련 법령 및 쟁점과 입법적·정책적 개선과제	2021. 06. 30.	이 덕 난 최 재 은
제082호	지방도시계획위원회 운영현황과 개선과제	2021. 07. 13.	김 예 성
제083호	코로나19 대응을 위한 등교 확대 정책의 주요 쟁점 및 개선과제	2021. 08. 19.	이 덕 난 유 지 연 최 재 은
제084호	지방세 납세자보호관 운영현황과 개선과제	2021. 09. 29.	류 영 아
제085호	지방소멸 위기지역의 현황과 향후 과제	2021. 10. 19.	하 혜 영 김 예 성
제086호	주요국 의회 이해충돌 심사기구의 구성 및 운영 비교	2021. 10. 22.	전 진 영 최 정 인
제087호	전기사업의 디지털 전환을 위한 쟁점과 과제	2021. 10. 25.	유 재 국
제088호	혁신도시 발전지원센터 운영현황 및 향후과제	2021. 11. 08.	김 예 성
제089호	공공재정 환수제도의 현황과 개선과제	2021. 11. 15.	김 형 진
제090호	미국의 남중국해 정책에서 '남중국해 중재판정'의 의미와 시사점	2021. 11. 18.	정 민 정

호 수	제 목	발간일	집필진
제091호	입원적합성심사제도의 문제점과 개선방안 -법·제도의 설계·운영 및 효과성 분석을 중심으로-	2021. 11. 25.	이 만 우
제092호	하천수 사용허가 제도 현황 및 개선과제	2021. 12. 01.	김 진 수
제093호	국내 고등교육기관의 해외진출 현황과 과제	2021. 12. 02.	유 의 정 조 인 식
제094호	재활용환경성평가 운영실태와 개선과제	2021. 12. 03.	김 경 민
제095호	가명정보 결합전문기관 운영실태와 개선과제	2021. 12. 06.	박 소 영
제096호	농업환경관리제도 현황과 입법·정책과제	2021. 12. 07.	장 영 주 김 규 호 유 제 범
제097호	사이버침해대응센터 운영실태와 개선과제	2021. 12. 15.	최 정 민
제098호	군 인권 제도 현황과 개선과제	2021. 12. 16.	심 성 은
제099호	선거관리 실태와 개선과제 -투·개표 관리를 중심으로-	2021. 12. 20.	이 정 진
제100호	북한이탈주민 취약계층 지원정책 현황과 개선과제	2021. 12. 21.	이 승 열
제101호	스마트그린 산업단지 실태와 개선과제	2021. 12. 22.	전 은 경
제102호	녹색금융 활성화를 위한 정책금융의 역할과 입법과제	2021. 12. 24.	이 수 환
제103호	저성장 극복을 위한 규제개선 방안 -규제샌드박스를 중심으로-	2022. 02. 23.	황 인 욱 박 성 용
제104호	4차 산업혁명 시대, 일상의 디지털 전환이 초래한 사회갈등의 현황과 대응 방안	2022. 03. 29.	정 준 화
제105호	국가균형발전을 위한 초광역협력 현황과 향후 과제	2022. 05. 12.	김 예 성 하 혜 영
제106호	제4차 「저출산고령사회 기본계획」의 문제점과 개선 방향 - 저출산 대응을 중심으로	2022. 05. 17.	박 선 권
제107호	투표용지 양식의 불평등성 논란과 쟁점: 정당 특권과 기호순번제 문제	2022. 05. 26.	허 석 재 송 진 미
제108호	지방소멸대응기금의 도입 및 향후 과제: 중장기적 정책과 거점 전략화	2022. 06. 30.	류 영 아
제109호	미혼부모·한부모 자립지원 서비스 실태와 개선과제	2022. 08. 24.	허 민 숙
제110호	제20대 대통령선거 분석	2022. 08. 26.	허 석 재 송 진 미
제111호	지방의회 의정활동 역량강화를 위한 교육훈련 운영실태 및 개선과제	2022. 09. 28.	하 혜 영 임 준 배
제112호	주거복지센터 운영현황 및 향후 과제	2022. 10. 27.	김 강 산

호 수	제 목	발간일	집필진
제113호	문화재 돌봄사업 운영실태와 개선과제	2022. 11. 14.	김 지 민
제114호	한국소비자원의 소비자분쟁조정 운영실태와 개선과제	2022. 11. 23.	최 은 진
제115호	공공보건의료지원단 운영실태와 개선과제	2022. 12. 05.	김 주 경
제116호	주요국 국민투표제도 비교와 시사점	2022. 12. 08.	오 창 룡
제117호	정부출연연구기관의 기술이전 운영실태와 개선과제	2022. 12. 14.	권 성 훈
제118호	지방자치단체 지역대학 지원 현황과 향후 과제	2022. 12. 21.	조 인 식
제119호	미국의 대북정책 결정과 국내 여론의 상관관계	2022. 12. 26.	이 승 열 허 석 재
제120호	범죄피해자 등 신변보호제도의 현황과 개선방안	2022. 12. 26.	김 광 현
제121호	우리나라 공공외교의 추진체계 현황과 개선과제	2022. 12. 27.	김 예 경
제122호	순환자원인정제도 운영실태와 개선과제	2022. 12. 27.	김 경 민
제123호	제8회 동시지방선거 중대선거구제 시범실시의 효과와 한계	2022. 12. 30.	이 정 진
제124호	미디어플랫폼 진흥정책의 평가와 개선과제에 대한 FGI 연구: 규제정책을 중심으로	2022. 12. 30.	최 진 응
제125호	성별공시제도 도입을 위한 과제-해외국가와 국내 비교를 중심으로	2022. 12. 30.	전 윤 정
제126호	웹접근성 품질인증제도 운영실태와 개선과제	2022. 12. 30.	김 나 정
제127호	국내외 자율주행자동차 관련 입법 동향과 쟁점 분석	2023. 04. 25.	박 준 환
제128호	출입국항 난민신청제도의 주요 쟁점과 개선방안	2023. 08. 08.	문 준 혁 조 규 범
제129호	2020년~2023년(상반기) 국회의 외교 관련 결의안 채택의 특징과 개선과제	2023.08.14.	정 민 정
제130호	법률의 합헌성 제고를 위한 입법절차 개선방안	2023.08.22.	김 선 화 김 보 람 원 시 연 장 경 석 오 창 룡 김 광 현 김 나 정 문 준 혁
제131호	동시입후보제 및 석패율제의 국내·외 입법 현황과 쟁점	2023. 10. 30.	허 석 재
제132호	국내·외 재난조사기구 현황과 시사점: 재난조사기구의 독립성·전문성 확보방안을 중심으로	2023. 10. 31.	배 재 현
제133호	인도-태평양 지역을 둘러싼 미·중 법률전의 평가와 우리의 과제	2023. 11. 16.	정 민 정

호 수	제 목	발간일	집필진
제134호	초고령사회 대응 장사(葬事)정책의 전환을 위한 입법과제	2023. 11. 20.	원 시 연
제135호	지방재정조정제도의 현황 및 향후 과제: 지방교부세를 중심으로	2023. 11. 22.	류 영 아
제136호	민사소송절차 선진화를 위한 디스커버리 제도 도입에 관한 조사	2023. 12. 01.	류 호 연
제137호	초고령사회 대응 「노인복지법」의 현황과 개선과제	2023. 12. 12.	원 시 연
제138호	국회세종의사당 시대 , 어떻게 준비할 것인가?	2023. 12. 13.	전 진 영 오 창 룡
제139호	상공인 사회안전망 현황과 확대 방안 -자영업자 고용보험을 중심으로	2023. 12. 18.	박 충 렬
제140호	금산분리 완화 관련 공정거래법 발의안의 동향과 쟁점	2023. 12. 21.	최 은 진 박 미 영
제141호	연동형 혼합선거제의 운영 현황과 작동 조건	2023. 12. 22.	허 석 재
제142호	UN과 EU 제재 비교와 우리나라에 대한 시사점	2023. 12. 26.	심 성 은
제143호	개인정보 전송요구권의 위험요소와 마이데이터의 발전 방향	2023. 12. 26.	김 형 진
제144호	인터넷 투표제도 쟁점과 도입 방향	2023. 12. 26.	송 진 미 오 창 룡
제145호	일본 기시다 내각 외교안보정책의 특징과 시사점	2023. 12. 27.	박 명 희
제146호	인구감소 적시 대응을 위한 출산율·이동률별 인구변화(2023-2123)	2023. 12. 29.	유 재 국 박 선 권
제147호	20~30대 여성의 고용·출산 보장을 위한 정책방향	2023. 12. 29.	전 윤 정
제148호	방송사업자간 홈쇼핑송출수수료 갈등요인과 과제	2023. 12. 29.	최 진 응
제149호	탈 플라스틱 사회를 위한 입법·정책 방안	2023.12.29	이 동 영
제150호	한·일 대륙붕 공동개발체제 종료 대비방안	2024. 04. 24.	정 민 정
제151호	기회발전특구 활성화를 위한 지방재정 지원 방안	2024. 06. 28.	류 영 아
제152호	감춰진 피해자들: 미성년 친족성폭력 피해자 특별지원 보호시설 지원업무 실태 및 개선과제	2024. 07. 18.	허 민 숙
제153호	조세불복제도 개편방안 - 납세자 권리구제와 자율적 행정통제 기능의 효율성 제고를 위해 -	2024.07.22.	임 재 범

호 수	제 목	발간일	집필진
제154호	프랑스 인구위기의 사회적 구성 및 가족정책의 시사점	2024.09.25.	박 선 권
제155호	제22대 국회의원선거 분석: 선거과정과 투표결과를 중심으로	2024.11.18.	허 석 재 송 진 미
제156호	제21대 국회 입법활동 분석	2024.12.19.	전 진 영 김 현 아
제157호	제22대 국회의원선거 분석: 공약과 정책 요인을 중심으로	2024.12.20.	송 진 미 김 현 아
제158호	사회경제적 변화에 대응하는 합리적 벌금형의 입법 방안 - 총액벌금제의 유지를 전제로 -	2024.12.23.	김 광 현
제159호	연성법(Soft law)의 활용 및 통제를 위한 입법방안 - 개인정보 보호 가이드라인·안내서·매뉴얼을 중심으로	2024.12.24.	김 형 진
제160호	회사법체계의 합리적 정비 방안	2024.12.24.	이 수 진
제161호	성매매 수요차단 및 피해자 보호 강화를 위한 개선 방안	2024.12.30.	전 윤 정

NARS 입법·정책 제162호

발 간 일 2024년 12월 31일
발 행 이관후 국회입법조사처장
편 집 사회문화조사실 과학방송통신팀
발 행 처 국회입법조사처
서울특별시 영등포구 의사당대로 1
TEL 02·6788·4715
인 쇄 ㈜케이에스센세이션 (TEL 02·761·0031)

ISSN 2586-5668
발간등록번호 31-9735044-001872-14

국민을 지키는 미래로 나아가는 국회

NARS 입법정책

주요입법 및 정책에 관한 주제를
심도있게 분석하여 대안을 제시하는 보고서로
수시 발간되고 있습니다.

07233 서울시 영등포구 의사당대로 1 (국회입법조사처)
Tel. 02-6788-4510(代) www.nars.go.kr

발간등록번호	31-9735044-001872-14
ISSN	2586-5668

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE