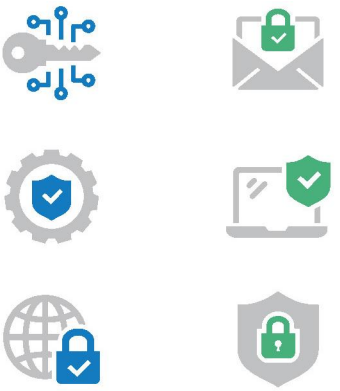
2025 VOL. 01
2025. 3월

KISA INSIGHT



딥시크(DeepSeek)의 등장과 인공지능(AI) 보안 이슈

김관영 | 천곤웅 | 김성훈

DIGITAL &
SECURITY
POLICY

KISA 한국인터넷진흥원

CONTENTS

KISA INSIGHT

2025 VOL. 01

DIGITAL & SECURITY POLICY

딥시크(De4epSeek)의 등장과 인공지능(AI) 보안 이슈

한국인터넷진흥원 김관영 선임연구원 | 천곤웅 선임연구원 |김성훈 팀장

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

작성

한국인터넷진흥원 정책연구실 정책연구팀

김관영 선임연구원	061-820-1515	kwaa_woo7@kisa.or.kr
천곤웅 선임연구원	061-820-1516	konja11@kisa.or.kr
김성훈 팀장	061-820-1510	shkim@kisa.or.kr

발간일 2025년 3월 28일

기획·발간처 한국인터넷진흥원 정책연구실 정책연구팀

요약

▌딥시크(DeepSeek)는 중국 스타트업 기업 딥시크(DeepSeek)社에서 개발한 LLM 기반 생성형 AI 모델 및 추론 강화 모델

• 함께 발표한 생성형 AI 모델(DeepSeek V3)과 추론 강화 모델(DeepSeek R1)에 대한 논문(기술 문서, '25.1.23)을 통해 "저비용·고성능 AI 모델"을 강조하며 전세계의 주목을 받음

– 해당 논문에서 DeepSeek V3의 개발 비용은 약 557만 달러(한화 약 78억 원)으로 경쟁 AI 모델(GPT-4o, OpenAI-o1, MS Copalit 등) 대비 저렴하지만 성능 더 뛰어나다고 발표

• 이로 인해 전세계의 AI 개발 연구 및 산업에 영향을 미쳤으며, 특히 미국-중국을 중심으로하는 글로벌 AI 개발 및 기술 경쟁을 심화시키는 계기

▌전세계 이슈가 된 딥시크에 대해 글로벌 보안 기업, 언론 등에서 보안 조치 미흡 등을 이유로 딥시크에 대한 보안 위협의 우려를 제기

• 글로벌 보안 기업인 팔로왈트 네트웍스社의 보안 연구팀 Unit42, CISCO 등에서 발표한 딥시크 AI 모델에 대한 보안 취약점 검증 결과에 따르면 AI 탈옥(Jailbreak)을 발생시키는 공격에 다른 AI 모델 보다 매우 취약하다고 분석

– 또한 딥시크社는 자사의 AI 모델을 오픈소스화하여 일부 AI 학습 데이터, 알고리즘을 제외한 모델 개발 코드를 공개하는 과정에서 수집된 데이터가 저장된 DB가 공개되며 데이터 유출 우려 제기

• 특히, 딥시크社가 중국 내 설립되어 있으며, 중국 내 서버 등에 당사의 데이터를 저장하고 있다는 점, 중국 데이터보안법 등에 따라 중국 정부의 요청에 따라 데이터를 제공해야 한다는 점 등으로 인해 서비스에 대한 불안감 증대

▌이에, 미국 등 주요국은 딥시크 차단 및 조사 등의 규제를 시행하거나 시행을 검토

• 미국, EU 국가, 영국 등 주요국들은 딥시크 출시 이후 개인정보 유출 및 국가 안보 등의 취약점과 위협을 우려하여 해당 서비스를 차단하거나 위해성에 대한 조사에 착수하였음

• 우리나라 역시 정부 부처 및 민간 기업까지 딥시크 서비스에 대한 전방위적 금지 조치가 확산되고 있으며 2월 15일부터 딥시크의 보안 우려가 해소될 때까지 국내 신규 다운로드 차단

▌딥시크가 촉발한 AI에 대한 보안 우려를 중심으로 AI 안전(Safety)의 관점에서 AI 위험(Risk)과 AI 보안(Security)을 구분하고 특히 AI 보안 위협에 대해 심층 분석

- (AI 위험(Risk)) AI 위험은 AI가 인간과 사회에 미칠 수 있는 모든 부작용 또는 악영향 등을 말하며, AI 위험은 크게 AI 윤리 위험, AI 신뢰 위험, AI 보안 위험으로 구분

- (AI 보안(Security)) AI 보안은 외부의 사이버 공격·침해행위 등으로 인한 AI 조작, 손상, 탈취 등 무결성, 기밀성, 보안성을 저해하는 행위를 말하며, 크게 AI 모델 취약점 공격, 사이버범죄 등 AI 악용, 데이터 유출 등으로 구분

〈 AI 보안 위협 종류 및 주요 요인 〉

구분	주요 요인
AI 모델 취약점 공격	▸ AI 모델과 시스템의 취약점을 공격하여 서비스 중단, 결과물 조작, 데이터 탈취 등을 일으키는 행위
	① 중독 공격, ② 기만 공격, ③ 학습데이터 추출 공격, ④ 학습 모델 추출 공격, ⑤ 프롬프트 공격
사이버범죄 등 AI 악용	▸ AI 서비스의 뛰어난 성능을 사이버범죄에 이용하는 행위
	⑥ 악성코드 생성, ⑦ 피싱/스미싱 메일 등 생성, ⑧ 범죄 지식 학습
데이터 유출	▸ 이용자가 실수로 정보를 AI 서비스에 입력하거나, AI 서비스의 오류로 정보를 출력하는 행위
	⑨ 민감정보(기업 기밀정보 등) 입력, ⑩ 개인정보 등 출력

▌딥시크의 공개는 글로벌 AI 산업 및 R&D 경쟁을 더욱 심화시키는 한편, AX 시대에서의 보안 위협에 따른 불안감이 증폭됨에 따라 AI 보안에 대한 정책 필요성 증가

- 미국, EU, 영국, 프랑스 등 주요국은 AI 기술 개발 연구(R&D)를 위해 대규모 국가 투자 정책 시행 또는 수립하는 한편 국가 안보와 직결되는 AI 보안 확보·강화 정책의 시행을 병행

- 우리나라는 AI 산업 육성 및 R&D 강화 등을 중심으로 정책이 수립되고 있으며, 인공지능 기본법 제정('26.1월 시행) 및 AI 안전연구소 설립('24.11월) 등 AI 안전성 확보 조치는 시작 단계

– 국민이 안전하게 AI를 개발하고 이용하고, 국가 안보를 더욱 강건히 하기 위한 AI 보안 대응 정책의 필요성이 강조되고 있어, 정책 수립이 필요하나 해당 정책이 AI 산업을 저해할 수 있는 우려가 있어 신중한 접근 필요

Ⅰ

딥시크(DeepSeek) 현황

1-1 딥시크(DeepSeek) 개요 및 특징

❙ 중국 딥시크(DeepSeek)社*에서 개발한 LLM 기반 범용 AI(GenAI) 모델로 글로벌 AI 모델(GPT-4o, OpenAI-o1 등) 대비 "저비용·고성능 AI"라는 점이 특징

* 량원펑(Liang Wenfeng)이 창업한 중국 스타트 기업으로, 2015년 헤지펀드(High Flyer) 설립·운영하며, 사내 소규모 AI연구소 개소하였고, 이를 기반으로 2023년 7월 '딥시크'(DeepSeek)로 분리하여 창업

• 중국 AI 연구 기업인 DeepSeek社에서 오픈 소스형 '대형 언어 모델(LLM)'로 개발되었으며, '23.11월, DeepSeek LLM을 공개하고 이후, DeepSeek V2('24.5월), DeepSeek V3('24.12월), DeepSeek R1('25.1월)을 순차적으로 빠르게 AI 모델의 성능 개선 및 기능 등을 강화하여 공개

〈표 1〉 딥시크 AI 모델 개발 추이

모델명(공개일)	주요 내용
DeepSeek LLM/Chat('23.11.29.)	‣ 최대 매개변수를 최대 670억 개까지 확장 가능하고, OpenAI社 ChatGPT 등 他 LLM과 경쟁하기 위해 개발
DeepSeek V2('24.5.6.)	‣ 공개된 기존 LLM의 성능을 개선하고 빠른 자연어 처리 및 대화형 AI 어플리케이션에 최적화한 기능을 추가
DeepSeek V3('24.12.26.)	‣ 6,710억 개의 매개변수를 가지고 있으며, 모델 훈련에 약 55일이 소요되었으며, GPT-4o와 동등한 성능을 갖추었다고 발표
DeepSeek R1('25.1.20.)	‣ 논리적 추론, 수학적 추론 및 실시간 문제 해결 능력 등을 강화하고 MoE(질문 맞춤형 데이터 이용)를 도입하여 기존 V3 성능을 향상

출처) 언론보도 등 저자 정리 및 작성

• 딥시크(DeepSeek)社 논문*을 통해 공개된 AI 모델(DeepSeek R1)이 他 AI 모델(GPT-4o, OpenAI-o1 등) 대비 저비용으로 고성능 AI가 개발되어 글로벌 이슈화[1)]

* DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reiforcement Learning"

– 일반적인 자연어 처리 및 분석, 이해를 위한 LLM으로 대규모 일반 텍스트 데이터 기반의 지도 학습(Supervised Fine-tuning, SFT)을 통해 훈련하여 개발된 DeepSeek V3를 기반으로 강화 학습(Reinforcement Learning, RL)을 적용하여 고도의 추론 능력 및 문제 해결 능력을 강화한 DeepSeek R1를 공개하며 논문 발표

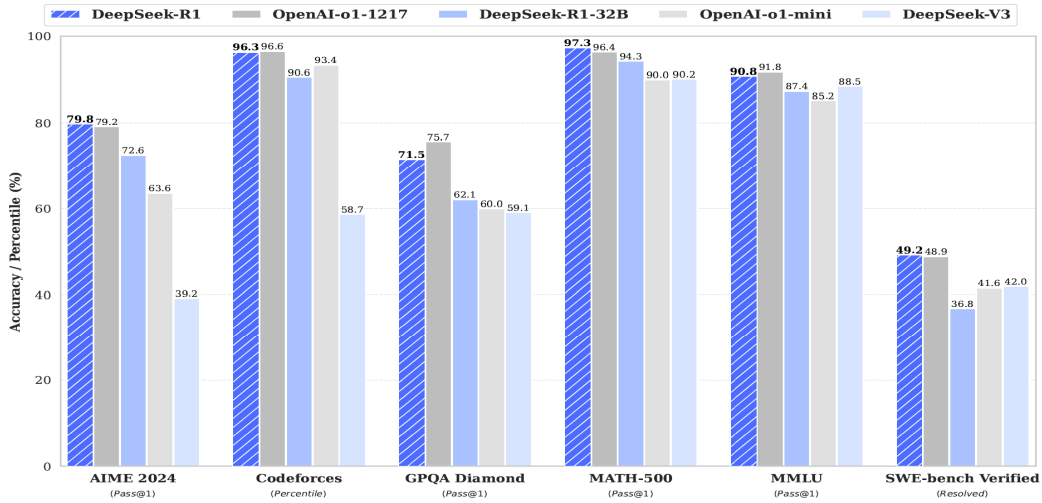
• 딥시크社가 발표한 논문에 따르면 딥시크 V3 기반으로 개발된 추론 모델인 딥시크 R1의 성능은 개발비용을 절감하였음에도 OpenAI社의 OpenAI-o1보다 우수하다고 발표

– DeepSeek V3 개발비용은 약 557만 달러(한화 약 78억 원)로 ChatGPT 개발비용(약 5억 4,000만 달러) 대비 약 1.1% 수준으로 절감하였고, 그 외 다른 AI 모델(MS Coplit 등)과도 비교하여도 매우 저렴함을 강조

– 특히, 미국의 對中 관련 무역 제재로 고사양 AI 칩셋(엔비디아 GPU H100)이 아닌 저사양 GPU H800(2,000여 개 임대)을 활용함으로써 개발과 AI 모델 학습 비용을 효과적으로 절감할 수 있음을 발표

– 더불어 논문에서는[2)], 딥시크의 AI 모델(DeepSeek V3, DeepSeek R1)은 다른 빅테크 기업이 개발한 LLM 및 생성형 AI 대비 언어(영어, 중국어 중심 비교), 코드, 수학 등 추론 능력이 뛰어났음을 결과로 제시

〈그림 1〉 DeepSeek R1과 타사 AI 모델 성능 비교[3)]



• 딥스크의 AI 모델 개발 소스코드와 알고리즘 등을 오픈소스로 공개하고, 딥시크 V3 및 R1 기반 생성형 AI 서비스를 무료로 제공하며, 전세계 대상 AI 접근성과 활용성을 확장하는데 기여함

1) CIO.com, "中 AI 스타트업 딥시크, '오픈 AI-o1' 겨냥한 오픈소스 모델 공개, 2025.1.23.
2) DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reiforcement Learning", 2025.1.23
3) DeepSeek-AI, 2025.1.27., p1~2

참고 1 中 딥시크로 촉발된 AI 新냉전 및 AI 패권 경쟁

Ⅰ 중국의 딥시크로 인해 美-中 간 AI 시장의 주도권 싸움이 격화되고 있으며, 딥시크에 대한 각 국의 대응 반응이 극명히 갈리면서 AI 패권 경쟁이 AI 新냉전 양상으로 치닫고 있음

- 미국, 유럽 및 동북아시아 일부 국가에서는 딥시크 사용을 차단하고 있으나 러시아, 인도, 동남아시아 일부 국가에서는 딥시크를 허용하고 사용을 적극 권장하는 등 차이를 보이고 있음

- 이 같은 경쟁 흐름에 따라 트럼프 행정부 주도로 스타게이트 프로젝트[4]를 발표(1.21)하였으며 중국 정부는 이에 대응하여 'AI 총동원령'을 선언하며 중국 빅테크 기업에 대형 투자 예고(2.17)

- 미국 조쉬 홀리(공화당) 상원의원은 인공지능 분야에서 중국과의 경쟁이 치열해지는 것을 해결하기 위해 미국의 인공지능 역량 분리법을 발의('25.1.29.)

  - 법안의 주요내용으로는 중국으로 AI 및 생형형 AI 기술과 R&D에 대한 수출입을 전면적으로 금지하고, 중국에서 개발 또는 생상되는 AI 기술에 대한 미국의 투자를 전면 금지하는 내용을 다루고 있음

〈표 2〉 미-중 AI 경쟁 양상

구분	미국	중국
경쟁 전략	• 자본 집약적 접근	• 효율성 중심 전략
시장 전략	• 폐쇄형 모델 중심	• 오픈소스를 활용한 개방형 모델
정책 목표	• 민간주도의 시장경쟁 • 책임 있는 AI, AI 윤리 등	• 국가주도 전략 산업화 • 실용성 및 성과주의
기술 역량	• 생성형 AI, 양자 컴퓨팅 분야 우위 • 오픈AI, MS 등 빅테크 주도 AI 생태계 구축	• 5G, 상업용 드론, 초고속 컴퓨팅 우위 • 저비용 AI 모델로 글로벌 시장 진출
투자 전략	• 데이터 센터 등 인프라 투자 - 스타게이트 프로젝트 추진 - '25년 MS 단독 100조원 규모 AI 데이터센터 건설 추진	• 국가 차원의 초대형 투자(빅펀드 3기) - 반도체 기술 자립 위해 약 70조 투자 - 미국의 기술 제대에 대응
데이터 정책	• 개인정보보호 및 데이터 국제 표준화 • 강력한 규제와 법적 제도	• 완화된 데이터 보호 정책 • 국가차원 데이터 통합

※ 출처 : 1) 딜로이트 Flash Report, "딥시크가 촉발한 새로운 AI 경쟁 시대", '25.2.1.
2) CIO 칼럼, "딥시크가 촉발한 AI 패권 경쟁", '25.2.23.

4) 트럼프 행정부의 AI 투자 전략으로 '25년부터 4년 간 5,000억 달러(한화 약 670조원)를 투자하여 중국의 AI 기술 도전에 대응하고 미국의 기술 패권 유지를 위해 미국의 AI 데이터 센터, 반도체 인프라 등을 구축하기 위한 프로젝트

II

딥시크 관련 보안 이슈 및 국내·외 대응 현황

2-1 딥시크 관련 보안 이슈

❙ 딥시크社의 AI 모델(DeepSeek V3, DeepSeek R1) 공개 이후, 전 세계가 주목하였으며, 각 국 정부, AI·정보보안 기업 등은 딥시크에 대한 보안 위협 테스트를 수행하며 보안 우려 제기

- 딥시크 AI 모델 공개 이후, 전세계 보안 전문기업에서는 딥시크에 대한 보안 이슈 확인 및 분석을 위한 테스트 수행 결과, 딥시크가 보안에 취약하고, 이로 인한 보안 위협에 대해 우려를 제기

  - 특히, 딥시크는 경쟁社(OpenAI, MS 등) AI 모델 보다 높은 보안 취약점이 노출되어 있으며, 보안 조치도 미흡하다고 지적하였으며, AI 탈옥(AI Jailbreak)[5]을 유발하는 다양한 보안 공격에 매우 취약하여 사용에 주의가 필요하다고 강조

- 팔로알토 네트웍스社의 보안 연구인 Unit42는 DeepSeek R1이 ChatGPT, Crescendo 등 보다 AI 모델 취약점으로 인한 AI 모델 탈옥의 발생률이 높은 것으로 분석한 보고서를 발표[6]

  - 보고서에 따르면 Unit42(유닛42)는 딥시크를 대상으로 디셉티브 딜라이트(Deceptive Delight)[7], 배드 리커트 저지(Bad Likert Judge)[8], 크레센도(Crescendo)[9] 등 AI 모델 탈옥을 발생시키는 보안 공격을 수행하였고, 이를 통해 딥시크의 보안 취약점의 유무 및 대응·조치 수준 등을 테스트하였음

5) AI 탈옥(AI Jailbreak): 인공지능(AI)에게 자금세탁이나 악성코드 개발 등과 같은 불법 활동에 대한 질문에 대답하도록 유도하면서 AI 안전장치를 무력화하는 공격
6) 팔로알토 네트웍스 연구팀(Unit 42), “Recent Jailbreaks Demonstrate Emerging Threat to DeepSeek”, 2025.1.30.
7) AI와 정상적인 질의 또는 대화하는 과정에서 악의적인 지시 또는 질의를 숨겨 AI에 입력하는 기법
8) 단계적인 질의를 통해 AI 모델이 유해한 추론, 출력을 하도록 유도하는 기법
9) AI에 특정 주체자 행동 등을 리커트 척도로 평가하도록 요청한 다음 사용자가 의도하는 행위로 AI의 행동을 유도하는 기법

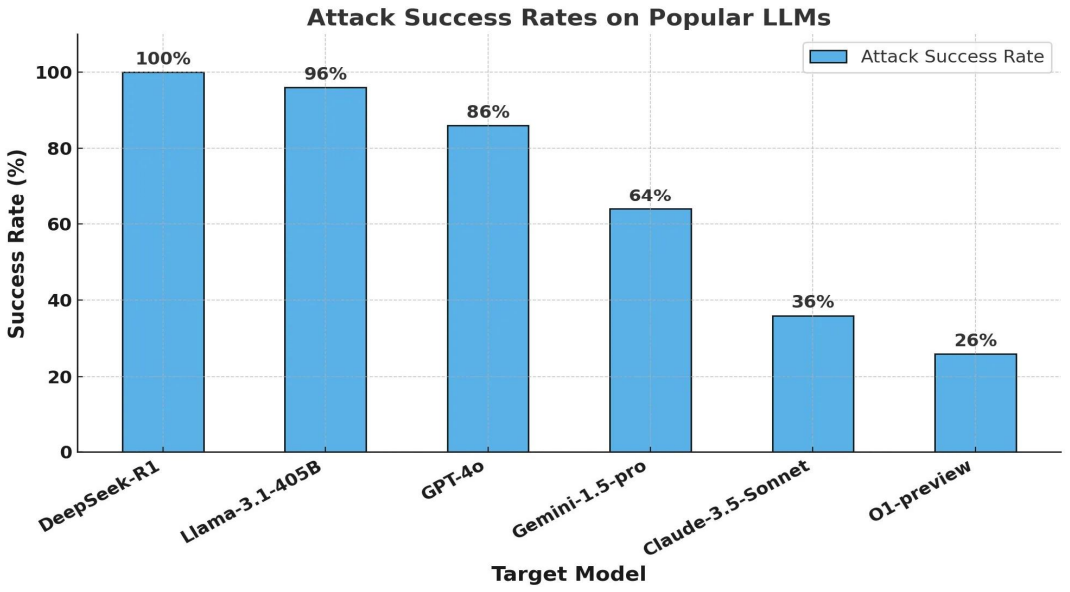
- 그 결과 딥시크에 설정되어 있는 안전장치(Guardrail, 가이드레일)를 우회하여 악성 소프트웨어 생성 방법, 살상 무기 제조, 데이터 탈취 도구 개발 등 해킹 또는 인간 생명에 피해를 줄 수 있는 유해 콘텐츠의 생성에 성공하였으며, 이로 인해 범죄에 악용될 수 있는 우려를 제기

• **글로벌 보안회사 CISCO는 AI 등에 대한 보안 취약점 분석 및 연구를 위한 자체 취약점 점검 프레임워크(레드팀 프레임워크)를 활용하여 검증한 결과 딥시크가 GPT-4o, Gemini-1.5 pro, Cladue-3.5 Sonnet 등 보다 높게 보안 공격에 성공하였다고 발표[10)]**

- CISCO에서 발표한 보고서에 따르면 딥시크社 추론 모델인 DeepSeek R1를 대상으로 HarmBench를[11)] 기반으로 50개의 공격 기법을 무작위 선정한 평가 프레임워크로 보안 취약점 점검 및 공격 성공률을 평가

- 점검 및 평가 결과 딥시크는 다른 AI 모델들과 비교하였을 때 모든 공격이 성공하는 100%라는 성공률을 보였으며, 이러한 공격들을 통한 AI 탈옥의 발생률 역시 100%로 나타났다고 발표하였으며, OpenAI社의 OpenAI-01(추론 GPT)에서는 대부분의 적대적 공격이 차단(성공률: 26%)되었다는 점에서 딥시크에 대한 보안 우려 제기

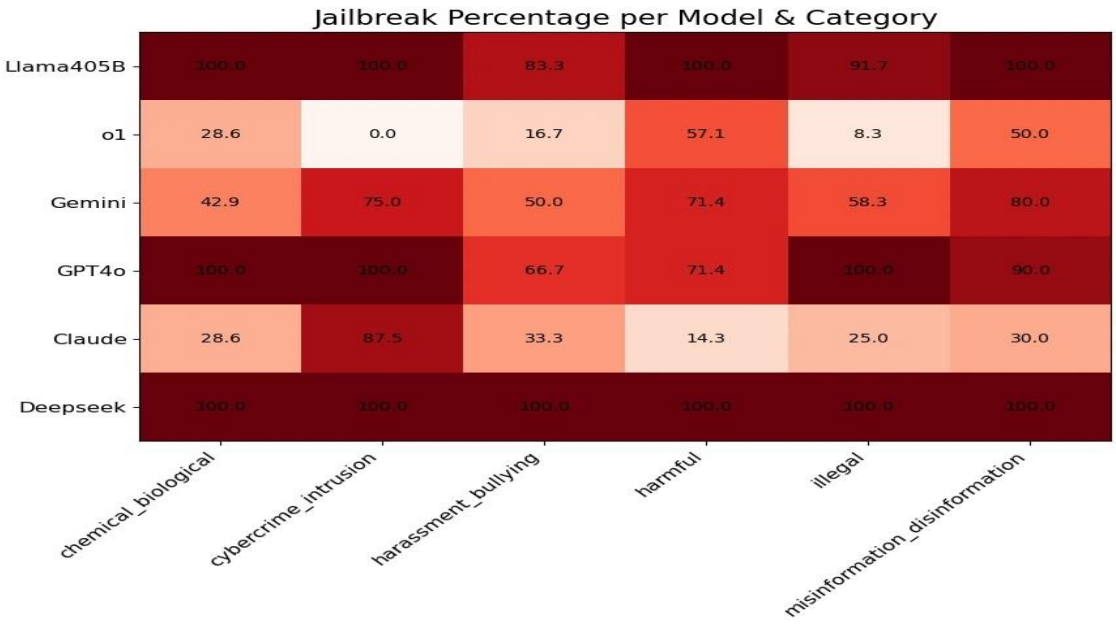
〈그림 2〉 DeepSeek R1 대상 보안 공격 성공률 비교



출처) CISCO, "Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models"

10) CISCO, "Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models", 2025.1.31.
11) 표준 레드팀 프레임워크(Red Framework)로 LLM의 악의적인 사용과 관련된 위험 식별 및 완화를 목적으로 사이버 범죄, 허위/거짓 정보, 불법/위법 행위, 일반적인 피해 등 7가지 피해의 범주에 대해 총 400여 가지의 행위 정의 및 관련 공격 방법을 제시하는 LLM 위험 식별 및 평가하기 위한 프레임워크

〈그림 3〉 DeepSeek R1 대상 AI 탈옥 공격 성공률



출처) CISCO, "Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models"

- **국내에서는 생성형 AI 보안 기업인 이로운앤컴퍼니의 AI 보안연구소에서 실시한 딥시크 AI 모델의 보안성 테스트 결과를 발표하며, 한국어 기반의 보안 공격에도 취약하다고 분석**

- 보안 취약점 점검을 위해 이로운앤컴퍼니 '세이프엑스 레드팀(SAIFE X RED Team)'이 딥시크 AI 모델 중 DeepSeek R1의 레드팀 테스트(RED Team Test)를 통해 딥시크에 대한 보안 취약점 점검을 기반으로 보안성을 분석하였음

- 이를 위해 AI 모델 공격 기법인 역할극(Role-Playing) 기반 공격[12], JSON 기반 구조화된 입력 공격[13] 등과 일반 사이버 위협 공격 등을 활용하였으며, 그 결과 역할극 기반 공격은 83%의 높은 성공률이 나타났으며, JSON 기반 구조화된 입력 공격 성공률 82%, 사이버 위협 관련 취약성 약 55% 등을 보였다고 발표

- 특히, 한국어 혐오 발언을 유도하거나 한국어 기반 AI 탈옥을 위한 공격을 시도하였을 때도 각각 41.7%와 18%의 취약성이 나타났으며, 이는 영어와 중국어를 기반으로 개발되어 한국어 등 다른 외국어에 대한 학습 부족 또는 관련 외국어에 대응이 미흡한 것으로 추정

12) 공격자가 신뢰할 만한 인물(예 : 은행직원, IT 지원팀, 경찰 등)을 연기하여 공격자의 요청을 따르도록 유도하는 공격
13) JSON 데이터를 다른 형식(XML, CSV 등)으로 변환하는 과정에서 발생하는 취약점을 악용하여 코드 실행, 데이터 조작, 정보 유출 등의 공격을 수행하는 기법

▌딥시크에 대한 보안 테스트 및 분석 보고서 등을 기반으로 확인한 딥시크의 보안 위협은 △ AI DB 보안 설정 등 일반적인 보안 관리 미흡, △ AI 모델 공격에 대한 취약성, △ DDoS 등 시스템·서비스 기반 공격 취약, △ AI 모델의 오픈소스화로 인한 보안 위협 등이 있음

• (일반적인 보안 관리 미흡) 딥시크 관련 소스코드를 공개하는 과정에서 딥시크와 연결되어 있는 데이터베이스(DB)의 URL이 함께 노출되었고, 노출된 DB 중 개인정보를 포함함 민감정보를 관리하는 DB가 공개되면서, 데이터 유출될 가능성 발생

- 이스라엘 국적의 클라우드 보안 기업인 Wiz에서 발표한 보고서에[14] 따르면 딥시크社에서 노출된 DB에는 누구나 열람, 수정 및 데이터 추출이 가능했으며, 이는 DB에 대한 관리자 권한 및 인증 설정 등 보안 설정이 미흡하였고, DB에 저장된 데이터는 개인정보를 포함하여 약 100만 여건*임

  * 공개된 주요 데이터: '25.1.6부터 생성된 딥시크 로그 기록, 딥시크 API, LLM-이용자 간 채팅 기록(시간 포함), DB 구조, 자료 출처 등 100만여 건

- 해당 문제는 딥시크社에서 소스코드 공개 범위, DB 등 주요 시스템 권한 설정 등 보안 관리 미흡으로 인해 발생된 것으로 분석되었으며, 해당 분석 보고서를 즉시 딥시크社에 통보하였고, 딥시크社 역시 즉시 보안 조치를 적용

• (AI 모델 공격에 대한 취약성) 딥시크에 대해 AI 모델 또는 시스템에 대한 보안 공격 기법과 보안 취약점을 활용한 보안 검증을 수행한 국내·외 보안 기업들은 딥시크가 현재 연구를 통해 밝혀진 AI 공격 기법에 취약함을 확인

- 딥시크 공개 이후 많은 국내·외 보안 기업, 언론 및 미국·영국 등 국가별 AI 안전연구소 등에서 딥시크의 LLM(DeeSeek V3)과 추론 모델(DeeSeek R1)에 대한 보안성 테스트 및 보안 취약점 점검 결과 발표
- 딥시크는 다른 AI 모델과 서비스 대비 높은 보안 공격 취약성이 확인되었고, 특히 AI 탈옥을 발생시키는 프롬프트 인젝션 등 공격에 매우 취약하고 이로인한 사이버범죄 악용(악성코드 생성, 유해 콘텐츠 생성 등)과 글로벌 사이버 보안 환경에 큰 위협이 될 것으로 우려

• (시스템·서비스 기반 공격 취약) DDoS, SQL Injection, 서비스 해킹을 위한 네트워크 탈취 공격 등 일반적인 사어버 공격에도 취약하였으며, 실제 DDoS가 발생하여 신규 회원 가입, 딥시크 서비스 사이트를 통한 애플리케이션 다운로드 등에 장애 또는 기능 오류가 발생[15]

14) Wiz, "Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat history", 2025.1.30.
15) 보안뉴스, "中 챗GPT, 딥시크, 사이버 공격 받아... 신규가입 불가 상태", 2025.1.28.

• (오픈소스 보안 위협) 딥시크社는 전세계 누구나 사용할 수 있도록 AI 모델과 관련된 알고리즘 등을 오픈소스화하여 공개함으로써 오픈소스에 대한 보안 위협이 딥시크에서도 발생할 수 있음

- DeepSeek V3, DeppSeek R1 등 주요 알고리즘, 훈련 데이터 등을* 글로벌 오픈소스 커뮤니티인 Github에 업로드 및 공개함으로써 AI에 대한 활용성과 접근성을 높이고, 모델 개발에 대한 투명성 확보 등에 강점이 있음

  * 딥시크 AI 모델에 대한 주요 코드만 공개하고, 핵심적인 코드, 훈련 방식 등은 비공개

- 하지만 OWASP에서 선정한 오픈소스 기반 소프트웨어(OSS) 보안 위협과 오픈소스에 내재된 보안상 취약점으로 인한 보안 사고 또는 악의적인 공격이 딥시크에서도 발생할 수 있음

- 특히, 공개된 딥시크 소스코드를 분석하여 보안 취약점을 발견하고 이를 활용한 AI 모델 또는 시스템 대상 공격, 모델 내 악성코드 삽입 등을 통한 AI의 악의적 행동(답변, 추론) 유도 및 AI 모델 탈취 등 발생 우려

〈표 2〉 오픈소스 보안 위협

오픈소스 보안 위협	주요 내용
보안성 확보 및 보안 위협 대응의 어려움	▸ 오픈소스는 보안 위협(코드 취약점, 시큐어 코딩 미적용 등) 요소 검증 없이 자유롭게 공개할 수 있어, 오픈소스 SW의 안전성을 보장하기 어려움 – SW 공개 이후 업데이트, 보안 패치 등 지속적인 관리·유지 보수에 대한 책임성이 불분명하여 보안위협 대응(대응 프로세스 부재 등) 어려움
공개된 보안 취약성	▸ 오픈소스 기반 SW는 불특정 다수가 개발에 참여하는 구조로 보안 취약점 발생 가능성이 높으며 출처 확인 및 보안 패치에 긴 시간 소요 – CC인증, 기업 자체 SW 검증 프로세스 등 별도의 보안 취약점 탐지.조치를 하지 않을 경우, 보안에 취약한 제품이 출시될 가능성 높음 – 오픈소스 기반 SW는 오픈소스 보안 취약점의 종속, 소스코드 적용 범위 등 현황 파악이 어려워 피해 대상 확인 및 패치에 긴 시간 소요
악성코드 등 삽입	▸ 악의적인 이용자가 공개된 오픈소스 또는 AI 알고리즘 등에 악성코드, AI 모델 오류 유도 코드 등 소스코드를 올려 AI 모델 또는 관련 시스템에 대한 보안 위협 발생 우려
악의적인 코드의 배포	▸ '누구나 자유롭게' 업로드할 수 있다는 점을 악용하여 악의적인 사용자(또는 그룹)가 악성코드가 포함된 오픈소스 또는 SW를 배포할 가능성이 높음 – 이와 함께 오픈소스를 분석하여 보안 취약점을 확보하고 이를 통해 오픈소스 기반 SW에 대한 공격 등 사이버 위협에 악용할 수 있음
소스코드 관리자 해킹	▸ 오픈소스 제작자 또는 오픈소스 커뮤니티 관리자(Github 등) 계정을 탈취하여 업로드된 소스코드를 삭제하거나 해커가 제작한 악성코드 또는 악성 알고리즘을 업로드하여 개인, 기업을 공격할 수 있음
소프트웨어 공급망 공격	▸ 오픈소스를 활용하는 기업 또는 조직을 대상으로 오픈소스의 취약점, 오픈소스 구성 요소 등을 공격하여 오픈소스가 도입된 시스템, 서비스 등에서 데이터 탈취, 서비스 파괴, 악성코드 유포 등이 가능

출처) 각종 보고서 내용 재정리

❙ 딥시크社는 중국에 본사를 두고 있어 중국의 데이터, AI 관련 법령을 준수하고 중국 내 서버에 이용자의 정보를 수집·저장함에 따라 중국 외 국가에서 개인정보 및 데이터 유출 우려 제기

• 딥시크社에서는 자사 AI 모델의 이용자로부터 개인정보를 비롯한 다양한 데이터를 수집하여 서비스 운영·제공, 개발 및 개선과 함께 서비스 이용 환경 향상, AI 학습 성능 개선 등에 활용

- 관련 수집 및 활용 목적에 대한 내용은 딥시크 본사 홈페이지의 개인정보보호 정책과 서비스 이용 약관에 명시하고 있으며, 중국 내 본사를 두고 있어 중국 내 법률(데이터법, 개인정보보법 등)을 준수하고 있으나 그 외 국가의 이용자에 대해 해당 국가 법률의 준수 여부는 확인하기 어려움

- 또한 딥시크에서 수집하는 데이터 범위가 매우 광범위하고, 수집된 데이터의 저장과 관리를 중국 내 서버(DB)로 규정함에 따라 우리나라를 비롯한 각 국에서 중국 정부의 데이터 제출 요구* 및 자사 내 데이터 무단 공유** 등으로 인한 데이터 유출과 개인정보 침해 등에 대한 우려를 제기

  * 중화인민공화국 데이터 보안법 등에 따르면, 중국 내 기업은 정부 기관의 요청 시 요청하는 데이터를 제공하도록 의무를 부과하고 있음

  ** 서비스 향상, 이용자 맞춤형 서비스 제공 등을 목적으로 제3자 대상 정보 공유를 이용 약관에 명시하고 있으나, 회원 가입 또는 서비스 이용 과정에서 이용자에게 관련 동의 과정 등이 없어 무단으로 공유가 이루어질 수 있음

〈표 3〉 미국 딥시크 보안 이슈 대응 동향

구분		주요 내용
귀하가 제공하는 정보	프로필 정보	생년월일, 사용자 이름, 이메일 주소, 전화번호, 비밀번호 등
	사용자 입력	텍스트, 오디오 입력, 프롬프트, 업로드된 파일, 피드백, 채팅 기록, 기타 콘텐츠 등
	당사 연락 정보	신원 또는 연령 증명, 서비스 이용에 대한 피드백 또는 문의, 당사 서비스 약관 또는 기타 정책의 잠재적 위반에 대한 정보 등
자동으로 수집된 정보	기술정보	장치 모델, 운영 체제, 키 입력 패턴 또는 리듬, IP 주소 및 시스템 언어, 충돌 보고서 및 성능 로그, 서비스 관련 진단 및 성능 정보 등
	사용 정보	사용하는 기능 및 수행하는 작업과 같은 서비스 사용 정보 등
	쿠키	쿠키, 페이지가 조회된 시간 및 날짜, 픽셀 태그가 배치된 페이지에 대한 설명 및 컴퓨터 또는 기기의 유사한 정보 등
	결제 정보	주문 배치, 결제, 고객 서비스, 애프터서비스 등
다른 출처 정보	로그인, 가입, 연결 서비스	Apple, Google 등 타사 서비스 연결을 위한 액세스 토큰 등
	광고 측정 및 기타 파트너	광고용 모바일 식별자, 해시된 이메일 주소 및 전화번호, 쿠키 식별자 등

출처) 딥시크. “DeepSeek 개인정보보호 정책”

2-2 국외 주요국 대응 현황

▌딥시크 모바일 앱은 미국 앱스토어와 구글플레이에서 다운로드 1위를 기록 하는 등 인기를 끌었으나 데이터 유출 사고와 보안 취약성 등이 드러나 각국 정부와 기관들이 우려와 규제를 표명

• **(미국)** 美 행정부는 딥시크를 국가 안보와 개인정보보호 위협으로 간주하여 다각도 대응에 착수하였으며 연방의회는 딥시크 앱을 정부 기관 기기에서 사용하지 못하도록 하는 법안 추진 중[16)]

〈표 4〉 미국 딥시크 보안 이슈 대응 동향

구분	주요 내용
해군	• 딥시크 모델의 근원과 사용에 관한 잠재적 보안 및 윤리적 우려로 전 해군에 사용 금지 조치('25.1.24.)
국방부	• 개인정보 및 사용자 데이터가 중국 내 서버에 저장되는 등 보안의 우려가 있어 사용금지('25.1.28.)
의회	• 악성 소프트웨어 감염 위험 등을 이유로 의회 사무처 및 의원실에 딥시크 사용 금지('25.1.30.)
항공우주국	• 정부에서 발급한 기기 및 네트워크에서 딥시크의 사용을 금지하는 메모를 발생('25.1.31.)
텍사스 주	• 텍사스 주지사는 중국의 데이터 수집과 AI를 통한 주 인프라 침투와 중국의 악의적인 스파이 활동으로부터 정부 기관 등을 보호하기 위해 주 정부가 지급한 기기에서 딥시크 사용 금지('25.2.2.)
뉴욕 주	• 딥시크는 외국 정부의 감시 및 검열과 관련된 문제가 있으며 사용자 데이터 수집 및 기술 노하우를 유출에 대한 우려로 뉴욕 주 정보기술국이 관리하는 모든 기기에서 딥시크 설치를 금지('25.2.10.)
버지니아 주	• 중국의 딥시크가 버지니아 연방 시민의 보안과 안전에 위협을 가하고 있음을 발표하며 국가 장치 및 국가가 운영하는 네트워크에서 중국의 딥시크 사용을 금지하는 행정명령에 서명('25.2.11.)

출처) 각 언론사 보도자료 재구성

16) 미국 연방 하원 정보위원회 소속 대런 라우드(공화당), 조시 고트하이머(민주당) 의원은 딥시크를 미 정부 기관 기기에서 사용하지 못하도록 하는 법안을 추진 중(출처 : The Wall Street Journal, '25.2.6.)

• **(EU)** 딥시크의 보안 우려가 증가함에 따라 EU는 데이터 주권에 대한 규제 조항을 담은 AI 규제법 개정안[17]을 발표하며 딥시크 서비스에 대한 강력한 규제 의지를 반영('25.2.6.)

〈표 5〉 EU 각 국 딥시크 보안 이슈 대응 동향

구분	주요 내용
이탈리아	• 개인정보 보호 기관인 가란테(Garante)는 개인 정보 사용의 불투명성과 사용자 데이터 처리 방식에 대한 우려가 해소되기 전까지 딥시크 신규 다운로드를 차단('25.1.29.) – 가란테는 딥시크가 유럽연합(EU)의 일반 데이터 보호 규칙을 준수하고 있는지 심층 조사를 시작하였으며 수집하는 개인정보의 종류와 목적 및 법적 근거, 중국 내 개인정보 저장 여부 등에 대해 질의하여 20일 이내로 답변할 것을 요청한 상태
아일랜드	• 아일랜드 데이터보호위원회(DPC)는 사용자 데이터 처리에 대한 정보 제공 요청을 딥시크 측에 전달하였으며 딥시크의 데이터 활용 방식과 현행 법률 위반 여부를 검토 중('25.1.29.)
프랑스	• 국가정보자유위원회(CNIL)는 데이터 보호 측면에 관한 위험성을 이해하기 위해 딥시크 시스템의 작동 방식 및 개인정보 처리 관련 정보 등에 대해 조사 중임을 밝힘('25.1.30.)
독일	• EU 개인정보보호법(GDPR) 위반 여부 및 딥시크 데이터 처리 관행에 대한 공식 조사를 시작하였으며 규제 조치를 검토 중('25.1.30.)
네덜란드	• 개인정보 수집과 관련하여 조사에 착수하였으며 자국 사용자들에게 사용에 주의 할 것을 당부('25.1.30.)

출처) 각 언론사 보도자료 재구성

• **(영국)** 영국은 정보통신본부 산하 국가사이버안보센터(NCSC)를 통해 딥시크의 AI 추론 모델 R1의 국가안보적 위해성 여부에 대한 조사에 착수함('25.1.30.)

• **(일본)** 디지털청은 딥시크와 관련해 개인정보가 제대로 보호되는지 확신할 수 없어 해당 우려가 불식되기 전까지는 일본 공무원을 대상으로 딥시크 사용을 금지('25.2.2.)

• **(대만)** 디지털부는 딥시크의 국경 간 데이터 전송 및 정보 유출 등의 문제가 국가의 정보 보안을 위협한다는 이유로 공공기관과 금융사를 대상으로 딥시크 사용을 금지('25.1.31.)

• **(호주)** 호주 내무부는 국가 안보와 국익을 위해 모든 호주 정보 시스템과 장치에서 딥시크 제품과 응용 프로그램, 웹 서비스 사용과 설치를 금지하며 딥시크가 발견되는 즉시 삭제할 것을 공표('25.2.5.)

17) 제3국 AI 서비스 제공자가 유럽 사용자의 데이터를 무단으로 역외로 반출할 경우, 최대 해당기업 전 세계 매출액의 8%에 달하는 과징금을 부과할 수 있음을 명시

참고 2 OpenAI社 ChatGPT 출시 당시 주요국 생성형 AI 규제 동향

▌Open AI社의 ChatGPT 출시('22.11.) 당시에도 개인정보 유출 우려 등을 이유로 각 국에서 해당 서비스에 대한 조사 및 규제 논의가 이루어진 바 있음

〈표 6〉 주요국 ChatGPT 규제 동향

구분	주요 내용	일시
미국	• 미국 비영리단체 '인공지능 및 디지털 정책 센터(CAIDP)'는 ChatGPT가 개인정보보호와 공공안전에 위협이 되어 미국 연방거래위원회(FTC)의 AI 기업에 대한 공개 지침을 위반한다면서 OpenAI를 고발('23.3.30)	'23.4.11.
	• FTC 위원장은 '인공지능 도구들로 인한 사기 등 소비자 피해를 우려하고 있으며, 연방정부에서 알고리즘 차별 및 개인정보 보호 문제로 AI 관련 규칙에 대해 논의가 이루어지고 있지만 AI 기업들은 기존의 다양한 법률에 따라서도 여전히 FTC 조사에 직면할 수 있다'고 경고	'23.4.4.
EU	• EU 의회, ChatGPT 출시에 따라 기존 발의·논의 중이던 'AI Act(AI 법안)' 내용 중 "고위험군 AI" 분류에 ChatGPT를 추가하는 등 법안 수정 검토	'23.3.5.
	• EU개인정보보호이사회(EDPB)는 인공지능에 대한 프라이버시 규칙을 설정하기 위한 첫 단계로 ChatGPT에 대한 TF(태스크포스)를 결성	'23.4.13.
독일	• 독일 개인정보 감독기관(BfDI), 개인정보 유출 등이 우려됨에 따라 ChatGPT의 사용 및 접속을 금지하는 방안 검토 중	'23.4.1.
캐나다	• 캐나다 개인정보 감독기관(OPC), ChatGPT가 사용자의 동의 없이 개인정보를 수입, 사용, 공개하고 있다는 진정서에 따라 Open AI에 대한 조사 착수	'23.4.4.
스페인	• 스페인 개인정보 감독기관(AEPD), ChatGPT에 의한 잠재적인 데이터 침해 관련 예비 조사를 시작할 예정이라 선언	'23.4.13.
일본	• 일본 문부과학성, ChatGPT 관련 글로벌 활용 사례, 동향 조사 및 ChatGPT 사용 주의사항을 담은 가이드라인 개발 착수	'23.4.6.
중국	• 중국 내에서 ChatGPT 홈페이지 등 접속 차단 - ChatGPT가 중국 개인정보보호법 등 법률과 규정을 위반했다는 이유로 차단 - 중국 관련 질의 시 비판적인 내용으로 답변이 되며, 이는 허위 사실 유포라고 주장	'23.2.24.
	• 중국 국가인터넷정보판공실, 생성형 AI 서비스 관리 방안 초안 발표 - 생성형 AI 서비스 출시 시, 출시 이전에 보안평가를 받고 개인정보나 지식재산권 보호 등 조건을 충족시켜야 한다는 규정 포함	'23.4.11.

출처) 각 언론사 보도자료 재구성

참고 3 국외 주요국 AI 보안 및 산업 관련 정책 동향

❙ 미국, EU, 영국, 프랑스 등 주요국은 AI 기술 개발 연구(R&D)를 위해 대규모 국가 투자 정책 시행 또는 수립하는 한편 국가 안보와 직결되는 AI 보안 확보·강화 정책 병행

- (미국) 바이든 행정부 시절 AI의 안전성과 신뢰성 강화 및 글로벌 AI 정책의 주도권 확보하기 위한 행정명령(EO)[18]을 발표하며, 정부 차원의 포괄적인 전략과 함께 AI 기술 확보를 위한 R&D 투자도 추진

  - 다만, 트럼프 2기 행정부 출범 이후 바이드 행정부의 AI 행정명령을 취소하며 규제를 최소화하고 글로벌 AI 기술의 선두 국가 및 기술 경쟁력 확보를 위하여 AI R&D에 5,000억 달러(한화 약 716조 원) 투자[19] 예정이며, AI 안전성 및 보안 확보를 위한 정책은 모니터링 필요

- (EU) '24년 AI Act를 제정하여('26년 전면 시행 예정) AI 위험으로부터 EU 국민을 보호하기 위한 정책을 시행하는 한편 유럽 내 AI 인프라 구축, 기술 연구를 위한 민·관 협력으로 총 2,000억 유로(한화 약 300조 원)를 투자 예정임을 발표[20]

- (영국) '23년 설립한 AI 안전연구소를 AI 보안연구소로 변경하여 출범하고, AI 기술 발전에 따른 국가 안보와 AI의 범죄 악용 문제에 대해 선제적으로 대응할 예정인 한편, 글로벌 수준의 AI 기술 확보를 위해 약 140억 파운드(한화 약 24조 원)[21]의 투자 계획을 발표

  - AI 보안 확보 및 국가안보 강화를 위해 국방과학기술연구소(DSTL), 국방부(MOD), 국가사이버보안센터(NCSC) 등과의 공동연구 추진 등 협력체계를 구축하고, AI 악용 위험 평가 및 예방을 위한 정책·기술 연구 추진 예정

- (프랑스) 글로벌 AI 강국으로의 도약을 위해 프랑스 內 AI 생태계 구축·강화에 대해 총 850억 달러(한화 약 85조 원) 규모로 투자할 계획[22]과 함께, 'AI 행동 정상회의(AI Action Summit, '25.2.10~11)'를 통해 AI 안전성 확보 등을 위한 '파리 선언문*' 발표[23]

  * 프랑스, EU, 중국, 대한민국 등 60개국이 참석한 '인류와 지구를 위한 포용적이고 지속 가능한 AI에 대한 선언문(Statement on Inclusive and Sustainable Artificial Intelligence for People and the Plant)

- (중국) 중국은 데이터보안법 등 중국 법률을 통해 국가 차원의 보안 정책을 시행하고 있으나 AI 등 신기술 분야는 글로벌 기술 패권을 확보하기 위해 대규모로 투자 진행

  - 중국 내 AI를 개발 및 서비스 중인 알리바바, 바이트댄스(틱톡) 등은 2030년까지 자체 AI 반도체 개발, AI 모델 인프라 구축 등에 각 230억 위원(한화 약 4조 5천억 원), 120억 달러(한화 약 17조 원) 등을 투자하고 있으며, 중국 당국도 미국과의 기술 경쟁을 위해 6년 間(2025~2031년) 약 2천조 원 투자를 추진할 것으로 전망[24]

18) Whitehouse, Safe, Secure, and Trustworthy AI(안전하고 신뢰할 수 있는 인공지능에 대한 대통령 행정명령, '23.10.30.)
19) 한국경제, "트럼프, 716조짜리 'AI 굴기'... 미국의 황금기 다시 온다", 2025.1.22
20) 연합뉴스, "AI에 수백조 쏟아붓는 선진국들.. 산업 특화 AI로 차별화해야", 2025.2.20
21) 지디넷코리아, "英 'AI 10년 대계' 발표... 국가 경쟁력 강화 위해 '24조원' 투자, 2025.1.14
22) 인공지능신물, "프랑스, AI 허브로 도약... AI 인프라에 85조 투자 유치, 브룩필드 20조 투입", 2025.2.10.

2-3 국내 대응 현황

▌우리나라 또한 정부 차원에서 딥시크 이용 중 발생할 수 있는 정보 유출 및 보안사고 등을 우려하여 딥시크 이용을 차단 중에 있으며, 관련 이슈에 대해 조사 착수

• AI 주무부처인 과기정통부는 딥시크 사용 금지 조치와 함께 KISA 보호나라 홈페이지를 통해 국내 일반 사용자 및 기업이 올바르게 생성형 AI를 활용할 수 있도록 보안 권고 사항을 안내('25.2.7.)


KISA, 딥시크 논란에 "생성형AI에 주민번호·주소 입력 말아야" 권고

생성형 AI 사용 관련 주의 보안권고

□ 개요

o 생성형 AI(Generative AI) 서비스가 전세계적으로 확산됨에 따라 국내 일반 사용자 및 기업이 올바르게 활용할 수 있도록 권고

• 개인정보위원회는 딥시크의 보안 문제에 대응하기 위해 전방위적 조사를 실시하고 있으며 정부 부처, 공공기관, 민간 기업에서도 잇따라 딥시크 사용 금지령이 확산되는 중

〈 개인정보위원회 대응 경과 〉

• 개인정보위원회는 딥시크 출시 직후 딥시크 본사에 해당 서비스의 개발 및 제공과정에서 데이터(개인정보 포함) 수집·처리와 관련된 핵심적 사항을 공식 질의함('25.1.31.)

– 개인정보 처리 주체, 수집 항목, 수집 목적, 수집·이용 및 저장 방식, 공유 여부 등을 질의하였으며 더불어 개인정보위원회 자체적으로 실제 이용환경을 구성해 전문기관과 함께 기술 분석 진행 중

• 영국, 프랑스, 아일랜드의 개인정보 감독기구 등과 공조하여 현재 딥시크 보안 이슈에 대한 상황을 공유하며 공동 대응하고 있으며 향후 대응 방안에 대해서도 논의 중('25.2.7.)

• 딥시크의 개인정보 정책에 대한 우려로 국내 딥시크 앱의 신규 다운로드를 제한('25.2.15.)

– 외교부·국방부 등 정부 부처에서도 딥시크에 대한 보안 우려로 사용금지 조치가 확산되고 있으며 행정안전부는 17개 광역지방자치단체에 딥시크 사용에 유의할 것을 요청하는 공문 발송

※ 기획재정부, 교육부, 노동부, 행정안전부는 소속 직원 및 산하기관에 딥시크 사용에 유의할 것을 당부하였으나 차단 조치에 대해서는 검토 중

– 카카오, LG유플러스 등 민간 기업에서도 과도한 정보수집 등을 이유로 딥시크 사용을 차단 하였으며 민·관 금융사 또한 개인정보 유출 및 사이버보안에 문제로 딥시크 사용 제한이 급속 확산 중

23) 연합뉴스, "지속가능·포용적 AI' 파리 공동선언, 미영 불참에 퇴색", 2025.2.11.
24) 연합뉴스, "중국, 6년간 AI에 2천조원 투자… 미중 간 '쩐의 전쟁' 가열", 2025.3.5

III

딥시크(DeepSeek) 등 AI 보안 위협 분석

3-1 인공지능(AI)의 위험(Risk)과 보안(Security)

▌인공지능(AI)의 등장 이후, AI 모델·시스템 개발하거나 기존 서비스 또는 시스템에 AI를 적용하는 등 다양한 분야에서 AI 활용이 증가하면서 이로 인한 안전성에 대한 우려도 함께 증가

• ChatGPT('22년)로 대표되는 생성형 AI의 등장으로 개인과 기업의 생산성과 효율성이 향상되는 등 긍정적인 효과가 있지만, AI의 활용이 확산하며 데이터 유출, 사이버범죄에 악용 등 부정적 효과 발생

– (긍정적) AI의 도입으로 제조업, 물류, 데이터 처리 등 반복적이고 정확성이 요구되는 업무를 자동화하여 생산성이 향상되고, AI를 기반으로 수많은 데이터를 빠르게 분석하고 향후 업무 방향성을 설정하는 등에 도움이 되어 효율성이 높아지고 업무시간도 절약할 수 있음

– (부정적) AI는 데이터 수집, 학습, 추론 등 일련의 과정에서 개인정보를 포함한 민감정보, 기업의 기밀정보 등 방대한 데이터를 관리하고 있어 정보 유출 또는 노출의 위협이 높으며, 사이버범죄에 AI를 악용하거나, AI 모델이나 시스템의 취약점을 공격하는 등의 위협이 발생할 수 있음

〈표 7〉 인공지능(AI)의 긍정적인 영향과 부정적인 영향

긍정적인 영향	부정적인 영향
■ 일상 생활 속 조력자(AI 비서 등) 역할 수행	■ 불법 콘텐츠, 허위·조작 정보 생성·배포
■ 사회 全 분야 생산성 및 효율성 향상	■ 저작권, 편향성, 다양성 존중 및 사회적 문제 야기
■ 신약 개발, 유전자 분석, 기후 분석 등 새로운 과학 연구	■ 연구 조작, 허위 결과 유도 등 연구 윤리 훼손 우려
■ 스팸, 스미싱 등 디지털 범죄로부터 보호	■ AI 기술 악용 범죄, AI 모델에 대한 공격 등 안전성 위협

출처) 저자 작성

▌미국, EU 등 주요국은 AI를 활용하는 과정에서 발생하는 위험을 정의하고 위험 관리, 위험으로 인한 피해로부터 국민과 국가를 보호하기 위한 정책을 시행하고, 법·제도를 제정 중

• (미국) NIST(국립표준기술연구소) 「AI RMF(Risk Management Framework) 1.0('23.1월)」에서 AI 위험을 "이벤트의 발생 확률과 그 결과의 규모 또는 정도에 대한 복합적인 측정하는 값"으로[25] 정의하고, 위험 관리 프로세스와 잠재적 피해 예시 제시

- AI RMF 1.0는 NIST에서 美 의회와 공공, 민간과의 긴밀한 협력을 통해 개발되었으며, AI 시스템을 사용하는 조직이 위험 관리와 신뢰할 수 있으며, 책임감 있는 AI 시스템을 개발하고 사용을 촉진하기 위해 마련

- 또한 AI에서 발생할 수 있는 부정적인 영향을 고려하여 정의한 AI 위험은 "상황이나 사건이 발생하였을 경우, 발생할 수 있는 부정적인 영향(피해의 규모)과 발생 가능성"이며, 개인, 사회, 조직을 포함해 지구 전체에 영향을 줄 수 있다고 재차 정의

- AI RMF를 통해 조직이 해결할 수 있도록 유연하고 체계적이며, 측정 가능한 프로세스, 신뢰할 수 있는 AI 시스템 특징* 등을 제시함으로써 AI 기술의 이점을 극대화하고 부정적인 영향을 최소화하여 AI 신뢰도를 강화하고자 함

  * 유효성 및 신뢰성, 안전성, 보안 및 복원성, 책임 및 투명성, 설명 및 해석 가능성, 개인정보보호, 공정성

〈표 8〉 AI 시스템 관련 잠재적인 피해 예시

구분	주요 예시
인간에게 미치는 피해 (Harm to People)	▸ 개인(Individual): 개인의 자유, 권리, 신체적/심리적 안전 또는 경제적 기회에 미치는 피해 ▸ 그룹/커뮤니티(Group/Community): 소수 인종, 민족 집단 차별 등 집단에 미치는 피해 ▸ 사회(Societal): 민주적 참여 또는 교육적 접근성에 미치는 피해
조직에게 미치는 피해 (Harm to an Organization)	▸ 조직의 비즈니스 운영에 미치는 피해 ▸ 보안 침입 또는 금전적 손실을 통해 조직에 미치는 피해 ▸ 조직의 명예에 미치는 피해
환경(또는 생태계)에 미치는 영향 (Harm to an Ecosystem)	▸ 상호 연결 또는 의존적인 요소 및 리소스에 미치는 피해 ▸ 글로벌 금융 시스템, 공급망 또는 상호 연결 시스템에 미치는 피해 ▸ 천연 자원, 환경 및 행성(지구)에 미치는 피해

출처) NIST, AI RMF 1.0

25) AI RMF 1.0에서의 AI 위험은 ISO 31000 : 2018에서 정의하고 있는 위험(Risk)의 의미를 인용

• (EU) '26년 EU 內 전면 시행 예정인 「AI Act(AI 법)」에서는 AI 위험을 "피해의 발생 가능성과 그 피해의 심각성"으로 정의(제3조(정의) 제2항)[26]하였으며, 세부적으로 위험을 4단계로 구분

- EU AI Act는 전 세계 최초로 AI의 안전성을 확보하고, AI 위험으로부터 유럽인을 보호하기 위하여 제정
- 4단계의 AI 위험은 각 위험도에 따라 ①허용할 수 없는 위험(Unacceptable Risk), ②고위험(High Risk), ③특정 투명성 위험(Specific Transparency Risk), ④최소 위험(Minimal Risk)으로 구분
- 이러한 각각의 위험에 해당하는 AI 모델 또는 시스템에 대해서 ①금지되는 AI 업무(이용 등), ② 고위험 AI, ③ 제한된 위험 AI, ④ 최소 위험 및 그 외 AI로 구분하였으며, 이들 AI에 대해서 차등적인 규제를 적용[27]

〈표 9〉 AI 위험 및 주요 요인

위험	해당 AI 모델/시스템	주요 내용
허용할 수 없는 위험	금지되는 AI 업무	‣ EU 기본권 및 가치를 침해하는 AI로 공공, 민간을 불문하고 사용을 금지하며 해당 AI의 시장 출시 및 서비스 제공 금지
고위험	고위험 AI	‣ 인간의 생명, 안전 및 EU 인권헌장의 기본권에 부정적인 영향을 미칠 수 있는 AI로 기본권 영향평가, 적합성 평가 등을 받아야 함
특정 투명성 위험	제한된 위험 AI	‣ 사람과 직접적인 상호작용, 생성형, 생체인식정보의 이용, 콘텐츠 제작 등에 해당하는 AI이며, 정보의 출처, 저작권 명시 등 투명성 확보 조치의 의무를 부과함
최소 위험	최소 또는 그 외 AI	‣ 금지, 고위험, 제한된 위험에 해당하지 않아 별도의 규제는 없으나, 기업의 자율적인 AI 행동강령[28] 마련 등을 권장

출처) EU, AI Act 및 부속서

• (영국) 영국 정부에서 발표한 "AI 안전 국제 과학보고서[29]"에서 범용 AI(general-purpose AI)가 개발 및 배포되는 과정에서 여러 가지 위험을 초래하고 있으며, 위험을 "피해 발생 확률과 그 피해 심각성을 결합한 상태[30]"라고 정의

26) EU AI Act, Article 3 Definitions (2) : 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm;
27) 위험도에 따라 잠재적 위험 예방·대응 및 위험 평가 등을 실시하는 이론인 위험 기반 접근(Risk Based Approach)을 기반으로 규제가 마련되었으며, EU AI Act 전체적으로 '고위험 인공지능 시스템'을 중심으로 규제 등이 설계되었음
28) AI의 부작용을 줄이고, AI의 안전성 및 신뢰성을 확보하기 위한 기업의 자율적인 규제
29) 영국 정부, "International Scientific Report on the Safety of Advanced AI : Interim Report", 2024.05
30) "International Scientific Report on the Safety of Advanced AI : Interim Report", p41
원문: For the purposes of this report 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm.

- AI 서울 정상회의(AI Seoul Summit, '24.5.21~22)의 개최를 앞두고 영국 정부(과학혁신기술부, AI 안전연구소)는 AI 안전 정상회의(AI Safety Summit)와 블레츨리 선언(Bletchley Park Declaration)을 통해 AI 안전, 신뢰 확보를 위한 국제적 합의를 상기하고, 첨단 AI 안전성(Advanced AI Safety)[31]에 대한 국제적 공유 및 수렴을 목적으로 발간

- 해당 보고서에서는 범용 AI가 발생시킬 수 있는 AI 위험을 "위험(Risk)"와 "교차 위험(Cross-cutting Risk)"으로 크게 분류하였으며, 다시 "위험"을 3가지의 세부 위험으로 구분하고 관련 요인들을 제시

〈표 10〉 영국 정부의 AI 위험 및 관련 요인[32]

위험 구분	세부 위험	주요 내용
위험 (Risk)	악의적인 사용의 위험 (Malicious use Risk)	▸ 하위 콘텐츠로 인한 개인의 피해 ▸ 거짓/허위 정보 및 여론 조작 ▸ 사이버범죄(Cyber offence) ▸ 이중 사용 과학 위험
	오작동으로 인한 위험 (Risks from malfunctions)	▸ 제품 기능 문제로부터의 위험 ▸ 편견과 과소 대표성으로 인한 위험 ▸ 제어(통제)권 상실
	시스템 위험 (Systemic Risk)	▸ 노동 시장 위험 ▸ 글로벌 AI 격차 ▸ 시장 집중적 위험과 모델, 시스템 등의 실패 ▸ 환경의 위험 ▸ 프라이버시에 대한 위험 ▸ 저작권 침해
교차 위험 (Cross-cutting Risk)		▸ 기술적 위험 요소 ▸ 사회적 위험 요소

출처) 정진철, "AI 신뢰성 정립을 위한 위험 요인 정의", TTA 저널 213호, 2024.06, 재정리

- 교차 위험(Cross-cutting Risk)은 하나가 아닌 여러 가지 위험에 기여하는 상태로 정의하고, 크게 2가지의 위험 요소 및 상세 위험 요인들을 제시

〈표 11〉 교차 위험 요소 및 요인

교차 위험	주요 내용
기술적 위험 요소	▸ 실제 현실 사례에서의 신뢰성 테스트 불가, 결과물 출력 기능에 대한 폐쇄성, 의도하지 않은 위법적, 비윤리적 결과 출력, 빠른 AI 취약점 공유, 위험 평가에 대한 전문성 요구, AI 자율성 강화로 인한 위험 고도화 등
사회적 위험 요소	▸ 과도한 기술 경쟁으로 인한 위험 대응 약화, 빠른 기술 발전을 대응하지 못하는 법·정책·규제, 투명성 및 책임성 결여, AI 모델과 시스템의 배포 및 사용 추적 불가 등

출처) 정진철, "AI 신뢰성 정립을 위한 위험 요인 정의", TTA 저널 213호, 2024.06, 재정리

31) 본 저자는 첨단 AI 안전성(Advanced AI Safety)은 프론티어 AI(Frontier AI) 등과 같이 향후 발전된 AI에 대한 안전성을 의미하는 것으로 분석하였으며, 현재 AI 안전성 확보를 위한 기술이 보다 발전되어야 한다라는 의미로 판단
32) 영국 정부, "International Scientific Report on the Safety of Advanced AI : Interim Report", 2024.05

3-2 AI 보안 관점에서의 보안 위협 분석

❙ 본 KISA Insight에서는 AI 안전(Safety) 관점에서 AI 위험(Risk)을 정의하고 관련 AI 위험의 요소들을 살펴보고, 위험 중 AI 보안(Security) 위험을 중점적으로 분석하여 정리

- 미국, EU, 영국을 포함한 주요국, AI 및 정보보안 관련 전문 기관(연구소 포함) 등에서 발표한 내용을 종합하면, AI 위험은 "발생할 수 있는 인간의 생명을 포함한 경제, 사회 등에 대한 피해"로 정리

  - 즉, 현재의 AI 위험은 사전적 의미의 위험[33]과 함께 AI 개발, 배포, 활용 과정에서의 기술적 결함 등으로 인해 발생할 수 있는 개인, 사회, 국가에 미치는 피해 또는 손실의 발생 가능성으로 AI 역기능의 개념으로 설명
  - 특히 EU AI Act는 제3조 제49항[34]에 "중대 사건(Serious Incident)"을 인공지능 시스템의 사건 또는 오작동으로 인해 발생하는 피해들로 규정하며, AI로 인한 위험에 대해 보다 상세히 명문화하여 규정

〈표 12〉 중대 사건에 대한 예시[35]

EU AI Act Article 3 Definitions
■ 사람의 사망 또는 시람의 건강에 대한 심각한 피해
■ 핵심 기반 시설의 관리 또는 운용의 중대하고 되돌릴 수 없는 중단
■ 기본권을 보호하기 위한 유럽연합 법령에 따른 의무의 위반
■ 재산 또는 환경에 대한 심각한 피해

- G7 히로시마 프로세스(G7 Hiroshima Process, '23.10월), AI 안전 정상회의('23.11월), 블레츨리 선언('23.11월) AI 서울 정상회의(AI Seoul Summit) 등에서 논의되고 있는 AI 안전(Safety)을 중심으로 AI 위험을 살펴보고, AI 안전을 위협하는 위험 요소 중 AI 보안(Security)에 대해 자체 분석하여 정리

  - AI 안전(Safety)에 대해 국제 기술 표준화 기구(ISO)는 일반적인 안전의 정의에 근거하여 "수용할 수 없는 위험으로부터 자유[36]"로 명시하며 AI 시스템이 인간의 생명, 건강, 재산 또는 환경에 피해 또는 손실이 발생하지 않는 상태로 유지할 수 있는 상황으로 설명
  - 또한 국제 사회에서는 AI 안전을 AI 개발부터 배포, 모니터링 등 전주기에 대한 안전성 확보을 포함하여 AI의 개발자, 서비스 제공자, 이용자 등 AI 모델, 시스템과 연관된 모든 환경에 대한 위험 요소의 사전 예방, 대응, 사고·피해 등의 사후 조치 등까지 확대하고 있으며, AI 안전 확보를 위한 국제적 협의가 지속되고 있음

33) 표준 국어 대사전: (명사) 해로움이나 손실이 생길 우려가 있음. 또는 그런 상태(2025.2.19. 확인), 옥스퍼드 사전 for google: (명사) 안전하지 못하거나 신체나 생명에 위해(危害)·손실이 생길 우려가 있는 것. 또는 그런 상태(2025.2.19. 확인)
34) EU AI Act 원문: 'serious incident' means an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following;
35) EU AI Act 원문: (a) the death of a person, or serious harm to a person's health;
(b) a serious and irreversible disruption of the management or operation of critical infrastructure;
(c) the infringement of obligations under Union law intended to protect fundamental rights;
(d) serious harm to property or the environment;
36) ISO, ISO/IEC 23893-2023: Information Technology - Artificial Intelligence - Guidance on risk management, 2023

- 이러한 국제적으로 논의되고 있는 AI 안전의 정의와 내용을 기반으로 AI 위험을 보안 관점에서 재정리하고, AI 위험 중 AI 보안 위협을 상세 분석하여 주요 요인을 분류하였음

• (AI 위험(Risk)) AI 위험은 AI가 인간과 사회에 미칠 수 있는 모든 부작용 또는 악영향 등을 말하며, AI 위험은 크게 AI 윤리 위험, AI 신뢰 위험, AI 보안 위험으로 구분

- 'AI 윤리 위험' 인간의 존엄성, 사회적 원칙, 법제도 등 기본 가치를 저해하는 것이며, 'AI 신뢰 위험'은 AI 기술이 내재하고 있는 잠재적 위험과 한계 등으로 AI 신뢰성을 저해하고 'AI 보안 위험'은 내·외부의 악의적인 공격 및 침해행위로 인해 AI 모델, 시스템에 손상이 발생하여 AI 안전을 저해하는 일련의 행위라고 할 수 있음

〈표 13〉 AI 위험 및 주요 요인

구분	주요 요인
AI 윤리 위험	▸ 공정, 정의, 사회적 영향 등 도덕적 원칙과 가치를 저해
	① AI 저작권 이슈, ② 다양성 존중(차별) 이슈, ③ 사회적 영향
AI 신뢰 위험	▸ AI 제품·서비스를 개발·배포·활용하는 全 과정에서 투명성, 신뢰성, 안전성 등을 저해
	④ AI 결과물의 신뢰성, ⑤ 불법 콘텐츠 생성
AI 보안 위험	▸ 외부의 사이버 공격·침해행위 등으로 인해 AI 기술의 무결성, 기밀성, 보안성 등을 저해
	⑥ AI 모델·시스템 취약점 공격, ⑦ 사이버범죄에 AI 악용, ⑧ 데이터 유출

출처) 저자 작성

• (AI 보안(Security)) AI 보안은 외부의 사이버 공격·침해행위 등으로 인한 AI 조작, 손상, 탈취 등 무결성, 기밀성, 보안성을 저해하는 행위를 말하며, 크게 AI 모델 취약점 공격, 사이버범죄 등 AI 악용, 데이터 유출 등으로 구분

- 'AI 모델 취약점 공격'은 AI 모델과 시스템의 보안 취약점을 악용하여 피해를 발생시키고, '사이버범죄 등 AI 악용'은 AI 모델 또는 서비스를 악용해 피싱, 스미싱 등의 범죄에 악용하는 행위이며, '데이터 유출'은 이용자의 실수 또는 해커의 공격, 서비스 오류 등으로 인해 개인정보 등이 유출되는 위협

〈표 14〉 AI 보안 위협 종류 및 주요 요인

구분	주요 요인
AI 모델 취약점 공격	▸ AI 모델과 시스템의 취약점을 공격하여 서비스 중단, 결과물 조작, 데이터 탈취 등을 일으키는 행위
	① 중독 공격, ② 기만 공격, ③ 학습데이터 추출 공격, ④ 학습 모델 추출 공격, ⑤ 프롬프트 공격
사이버범죄 등 AI 악용	▸ AI 서비스의 뛰어난 성능을 사이버범죄에 이용하는 행위
	⑥ 악성코드 생성, ⑦ 피싱/스미싱 메일 등 생성, ⑧ 범죄 지식 학습
데이터 유출	▸ 이용자가 실수로 정보를 AI 서비스에 입력하거나, AI 서비스의 오류로 정보를 출력하는 행위
	⑨ 민감정보(기업 기밀정보 등) 입력, ⑩ 개인정보 등 출력

출처) 저자 작성

AI 보안 위협은 앞서 제시하였듯 △AI 모델 취약점 공격, △사이버범죄 등 AI 악용, △데이터 유출을 발생시키는 각각의 위협 주요 요인들에 대해 상세 분석

- AI 모델 취약점 공격으로 인한 위협은 기술이 가진 고유의 보안 취약점으로 인한 문제로 이를 해결하지 못할 경우 AI에 대한 안전성, 신뢰성 저하 발생시킬 수 있음

– 따라서 AI 모델 또는 시스템 내 AI 모델을 적용하는 과정에서 AI 모델에 대한 공격 기법들을 포함한 보안 취약점을 사전에 대응할 수 있도록 보안 조치가 필요하며, 모델과 서비스에 대한 지속적인 보안 모니터링도 필요

〈표 15〉 AI 모델 취약점 공격 종류 및 주요 내용

구분	주요 내용
중독 공격 (Poisoning attack)	▸ 공격자(해커)가 의도적으로 악의적인 수집, 학습 데이터를 주입하여 AI의 학습 모델의 정확도를 낮추거나 오류를 발생하도록 하는 공격
기만 공격 (Evasion attack)	▸ 입력되는 데이터에 미세한 노이즈를 삽입하여 학습 모델의 예측 및 답변을 방해하거나 조작하는 공격
학습데이터 추출 공격 (Inversion attack)	▸ AI 학습 모델, AI 모델에 수많은 쿼리를 반복적으로 송신하여, 산출되는 결과(답변 데이터 등)를 분석하여 학습 데이터를 추출하는 공격
학습 모델 추출 공격 (Modelextraction attack)	▸ 학습데이터 추출 공격을 통해 확보된 결과값을 분석하여 학습 모델을 추출하여 복제, 악용하는 공격
프롬프트 공격 (Prompt attack)	▸ AI 모델 또는 시스템의 프롬프트에 반복적인 명령어, 우회 질의 입력, 시스템 해킹 등을 통해 AI 안전장치(필터 기능, 답변 제한 등)를 우회하여 부적절한 답변 또는 행위를 유도하는 공격 ※ 프롬프트 공격 기법으로 프롬프트 인젝션(Prompt Injection), 허위 코드 입력, 반복·지속적인 질의 또는 코드 입력 등이 있음

출처) 저자 작성

- AI의 코드 생성·검토 기능 또는 사기GPT를 활용하여 악성코드, 렌섬웨어 등을 생성할 수 있으며, 누구나 쉽게 해킹 도구를 제작할 수 있어 손쉽게 사이버범죄에 악용될 수 있는 우려가 있음

※ PC, 모바일 등 IT, 보안, 해킹 등 관련 전문지식이 없는 일반인도 악용 가능

– 또한 AI를 기반으로 기존 사이버범죄의 지능화 및 고도화 등이 발생할 수 있으며, 국가 배후 해킹 조직 역시 AI를 악용하여 국가 間 사이버 공격으로 국가안보도 위협할 수 있으며, 새로운 사이버범죄 및 공격 기법이 연구될 수 있어 국민 불안감이 증가하고 있음

〈표 16〉 사이버범죄 등 AI 악용 종류 및 주요 내용

구분	주요 내용
악성코드 생성	▸ LLM을 활용한 코드 생성 및 프로그램 진단 기능을 악용하여 누구나 손쉽게 해킹 또는 시스템 탈취 등에 필요한 기능 또는 프로그램, Malware 등 개발하여 유포하거나 공격에 활용[37] ※ 인공지능의 기술이 고도화될수록 예측할 수 없는 사이버 공격이 증가할 수 있으며, 프로그램 개발·분석 등 전문지식이 없어도 개발할 수 있어(스크립트 키즈) 보안 위협에 대한 우려 증가
피싱/스미싱 메일 생성 등	▸ AI의 학습, 추론 기반의 빠르고 다양한 문장을 전세계 언어로 생성할 수 있는 기능을 악용하여 손쉽게 피싱, 스미싱 등 범죄 발생 증가 우려 ※ 일부 용어·문장의 오류가 일부 있으나 AI 모델 기능 개선, 지속적인 학습을 통해 더욱 정교해지고 있어, 피싱 등 디지털 민생범죄에 악용[38]
범죄 지식 학습	▸ 해커를 비롯한 일반인이 보안 취약점 점검·분석을 통한 공격 코드 제작 등 AI를 통해 범죄에 필요한 지식을 학습하고 실제 범행 가능 ※ 실제 트럼프 호텔 앞 테슬라의 '사이버트럭' 폭발 사건에 챗GPT를 통해 폭발물 제조 방법 등 범죄에 필요한 정보를 악용하는 등 사례 발생[39]

출처) 저자 작성

• (데이터 유출) AI 학습 과정에서 수집한 데이터나 이용자가 AI 서비스 이용 과정에서 입력한 데이터 등이 AI의 오류, 외부 공격 등으로 인해 유출될 수 있으며, 이로 인한 2차, 3차 피해 발생 가능

– AI 모델은 학습과 추론을 위해 학습 단계에서의 수집되는 데이터, 이용자로부터 입력되는 정보 등 전세계적으로 확보할 수 있는 데이터가 매우 방대하며, 해당 데이터에는 개인정보 등 민감정보가 포함되어 있어, AI 모델과 연계된 DB 대상 공격, AI 모델 오류 등으로 인한 데이터 유출의 피해는 매우 클 것으로 예측

〈표 17〉 데이터 유출 종류 및 주요 내용

구분	주요 내용
민감정보 입력	▸ 이용자가 AI 서비스를 이용하는 과정에서 개인정보, 기업정보 등을 입력할 경우, 입력된 정보를 기반으로 AI 재학습 등에 활용될 수 있음
개인정보 등 출력	▸ 학습 데이터를 기반으로 추론, 답변하는 과정에서 악의적인 사용자의 질의 또는 공격. AI의 오류로 인해 개인정보와 같은 민감정보를 출력하는 보안 이슈 발생

출처) 저자 작성

37) 조선일보, "악성코드 수백개 순식간에 뚝딱... 천재 '해커'의 정체는", 2024.3.28.
38) 이코노믹데일리, "웜GPT와 사기GPT 등장... 내년 사이버 범죄의 새로운 무기 된다", 2024.12.18.
39) 전자신문, "美 사이버트럭 폭발 사건... 경찰 "장치 제조법 챗GPT"로 검색한 듯", 2025.1.8.

참고 3 OWASP Top 10 for LLM Applications 2025

▌OWASP에서는 LLM 기술을 활용하는 애플리케이션과 플러그인을 설계 및 설계, 서비스하는 개발자, 데이터 과학자 등을 대상으로 LLM 보안 위협과 대응 방안을 제시

• LLM Application에서 발생할 수 있는 위협 중 위협성의 정도에 따라 10가지 선정

〈표 18〉 사이버범죄 등 AI 악용 종류 및 주요 내용[40)]

구분	대응 방안
Prompt Injection	▸ 시스템 프롬프트를 적대적으로 유도하거나 조작된 외부 입력을 통해 간접적으로 수행 할 수 있으며, 잠재적으로 데이터 유출, 모델 손상 등의 문제를 야기할 수 있음 ▸(대응방안) Privilige Control, Human Approval, Segregate Content, Trust Boundaries 등
Insecure Output Handling	▸ 다운스트림 구성 요소가 적절한 검토 없이 대규모 언어 모델(LLM) 출력을 맹목적으로 수용할 때 발생하는 취약점으로 XSS, CSRF, 권한 상승 등이 발생할 수 있음 ▸(대응방안) Zero-Trust 접근 방식, OWASP ASVS Guidelines 도입, Output Encoddng 등
Training Data Poisoning	▸ 모델의 보안, 효과성 또는 윤리적 행동을 손상시킬 수 있는 취약성, 백도어 또는 편향을 도입하기 위한 데이터 조작 또는 미세 조정 프로세스 ▸(대응방안) 공급망 검증, 합법성 검증, 사용 사례별 교육 등
Model Denial of Service	▸ 공격자가 LLM과 상호작용할 때 예외적으 로 많은 양의 리소스를 소모할 때 발생 ▸(대응방안) 입력 검증, 리소스 캡, API 비율 제한, 큐 관리, 리소스 모니터링 등
Supply Chain Vulnerabilities	▸ LLM의 공급망 취약성은 교육 데이터, ML 모델 및 배포 플랫폼을 손상시켜 편향된 결과, 보안 침해 또는 전체 시스템 장애를 발생시킬 수 있으며, 오래된 SW, 취약한 사전 훈련 등이 원인 ▸(대응방안) 공급업체 평가, 플러그인 테스트, OWASP A06 적용, 재고 관리, 보안 조치 등
Sensitive Information Disclosure	▸ LLM의 오류로 인해 민감한 정보, 독점 알고리즘 또는 기밀 데이터를 공개하여 무단 액세스, 지적 재산권 도용 및 개인정보 침해 발생 ▸(대응방안) Data Sanitization, 입력 검증, 미세 조정 조의, 데이터 액세스 주의 등
Insecure Plugin Design	▸ 악의적인 요청으로 이어질 수 있는 취약한 결과를 초래하여 데이터 유출, 원격 코드 실행, 불충분한 액세스 제어, 권한 상승 등 발생 ▸(대응방안) 매개변수 제어, OWASP 지침, 철저한 테스트, 최소 권한 등
Excessive Agency	▸ LLM 기반 시스템의 과도한 에이전시는 과도한 기능, 과도한 권한 또는 너무 많은 자율성으로 인해 발생하는 취약성 ▸(대응방안) 플러그인 제어, 플러그인 범위 제어, 세분화 기능, 권한 제어, 사용자 인증 등
Overreliance	▸ LLM에 대한 과도한 의존은 잘못된 정보, 법적 문제, 보안 취약성과 같은 심각한 결과를 초래 ▸(대응방안) 모니터링 및 검증, 교차 확인, 미세 조정, 자동 검증, 위험 전달 등
Model Theft	▸ LLM 모델에 대한 무단 액세스 및 유출이 포함되며, 경제적 손실, 평판 손상 및 민감한 데이터에 대한 무단 액세스 위험 발생 ▸(대응방안) 접근 통제 및 인증, 네트워크 제한, 모니터링 및 감사, MLOps 자동화 등

출처) 저자 작성

40) OWASP, OWASP Top 10 for LLM Applications 2025, 2025.2

IV 시사점

4-1 딥시크의 등장으로 AI 기술 혁신의 기대와 AI 보안 우려 증가

▌딥시크 출현으로 인해 AI 혁신의 이면에 있는 AI 보안 위협이 드러나 한국을 포함한 미국, 영국 등 주요국에서 딥시크 유해성에 대한 조사 및 사용 금지 조치가 전방위적으로 확산 중

- 미국, 일본, 이탈리아, 대만 등은 딥시크로 인한 국가 안보 및 데이터 유출 등에 대한 우려로 사용 금지 조치를 취하였으며 영국, 프랑스, 아일랜드 등은 딥시크 유해성에 대해 전면적인 조사에 착수

  - 우리나라 또한 정부 부처 및 민간 기업에 딥시크 사용 금지 조치가 내려졌으며 딥시크의 보안 취약점 등에 대한 기술 분석을 수행하는 등 보안 우려 해소를 위해 민·관 협력, 국제 공조 등 적극 대응 중에 있음

- 이와 반대로 인도, 러시아, 동남아 국가 등은 딥시크의 적극 활용을 권장하며 국가별로 극명한 차이를 보이고 있어 딥시크의 등장은 AI 신냉전 체제와 글로벌 AI 패권 경쟁의 가속화하는데 영향을 미침

- 이처럼 글로벌 AI 패권 경쟁은 심화되는 추세이며 이로 인한 보안을 고려하지 않은 AI 기술 개발 및 산업 육성 중심의 과도한 AI 혁신 경쟁은 AI發 사이버 위협이 증가하는 원인으로 작용할 우려가 있음

▌딥시크 보안 이슈는 혁신의 속도와 보안 체계 구축 사이의 격차를 여실히 보여준 대표적인 사례로써 AI 보안 및 안전에 대한 검증 부족과 혁신에만 집중한 결과로 분석

- 딥시크 보안 사고 사례를 통해 개발 과정에서부터 보안성을 확보하기 위한 신뢰할 수 있는 기준을 마련하고 해당 기준이 AI 산업 발전을 저해하지 않도록 민·관 의견을 수렴할 수 있는 협력체계 필요

- 주요국의 딥시크 차단 사유 중 큰 비중을 차지하는 부분인 개인정보 분야인 만큼 AI 서비스의 개인정보 유출 등 관련 이슈에 대해 인공지능 주무 부처와 개인정보 관련 주무 부처의 유기적 협력이 요구됨

4-2 AX(AI Transformation) 시대의 AI 보안 위협 예방·대응 정책 필요

▌AI 기술의 발전과 가속화된 AI 확산으로 기존 DX(Digital Transformation)에서 AI 기반의 AX(AI Transformation) 시대로 변화하고 있는 가운데, 딥시크 등장으로 더욱 가속화

- ChatGPT의 등장('23년)에 따른 생성형 AI 도입·활용 확산, 딥시크社의 저비용·고성능 AI의 등장('25년)에 따른 글로벌 AI 기술 연구 및 경쟁 심화 등으로 AX 시대 변화가 가속화

– 이를 대응하기 위해 미국과 중국을 중심으로 선진국들은 국가 차원의 대규모 투자를 실시하거나 관련 계획을 수립하여 추진 예정으로 글로벌 AI 기술과 산업 주도권 확보 경쟁이 더욱 심화되고 있는 중

– 우리나라 역시 국가 인공지능 위원회를 중심으로 각 부처가 협력하고 글로벌 AI 기술 경쟁력 확보를 위하여 과기정통부 주관의 「AI 컴퓨팅 인프라 확충을 통한 국가 AI 역량 강화 방안」 발표

- 이러한 AX 시대로 변화할수록 AI 보안 위협에 따른 사회 불안 및 국민의 우려가 증가하고 있어 이를 대응하기 위한 정책도 함께 모색되어야 할 시기

– EU는 AI Act 제정을 통해 AI로 인해 발생할 수 있는 AI 위험으로부터 유럽인을 보호하기 위해, 영국은 AI 안전연구소를 AI 보안연구소로 변경·설립하여 국가 안보에 위협이 되는 AI 보안 위협 및 AI 악용을 예방·대응하는 등 AI 보안을 확보 또는 강화하기 위해 관련 정책을 수립 중

– 우리나라는 AI 모델 개발과 이용 관점에서 안전성 및 신뢰성 확보를 위한 가이드라인을 발표하고 있으나 국민과 국가 안전을 확보하기 위한 AI 보안 위협의 예방·대응, AI를 악용한 범죄 대응 등과 관련된 정책은 미흡한 상황으로 이를 위한 대응 정책 마련 필요

▌AI 기술의 발전과 가속화된 AI 확산으로 촉발된 AX 시대로의 변화에 대비하고 AI로 인한 보안 위협들로부터 국민과 기업을 넘어 국가를 보호하기 위한 정책 추진 필요

- AI G3 강국으로의 도약을 위해 안전한 AI 개발 및 활용을 위한 환경을 조성하고, AI의 보안 위협으로 인한 피해를 최소화하고 이를 통해 국민과 국가를 보호할 수 있도록 하는 정책 수립 필요

– ①모니터링 강화 및 ②전방위 취약점 점검을 지원하고 ③보안·안전 시스템을 고도화하기 위한 ④튼튼한 보안 기반을 구축 등 AI 보안과 더불어 안전성을 확보하기 위한 정책을 통해 글로벌 수준의 AI 보안 확보 필요

– 국내 AI 기업을 대상으로 DDoS 공격, 해킹 등 사이버 공격 예방을 위한 모니터링을 강화하고 AI 모델 또는 서비스가 가지고 있는 취약점 보완을 위한 AI 보안 취약점 점검 지원, AI Security RED 운영, AI 보안 가이드라인 개발 등 AI 보안 위협 대응 방안 추진 검토

– 이와 함께 AI 보안을 위한 기술 개발(R&D)을 위한 정부 과제 발굴·추진, AI 보안 전문 기업과 인력 육성 등을 통해 글로벌 국가 AI 보안 경쟁력 및 안전성 확보를 위한 기반 조성에도 노력 필요

– 다만, 이러한 정책들이 AI 산업 육성과 기술 발전에 저해될 수 있다는 우려도 공존하고 있어, AI 보안 정책 수립 시, 기술 개발과 산업 진흥 측면을 고려하여 균형있는 정책 수립 및 시행 필요

www.kisa.or.kr

KISA INSIGHT

2025 VOL. 01
2025. 03

KISA 한국인터넷진흥원