AI 위험 유형 및 사례 분석

A Comprehensive Review on AI Risks and Cases

Executive Summary

최근 몇 년간 인공지능(AI) 기술의 발전은 챗GPT의 출시 이후 대규모언어모델 개발 경쟁을 거치며 가속화되었다. 현재 공개된 AI 모델들의 성능은 특정 분야에서는 이미 인간의 능력을 뛰어넘었고, 이에 따라 활용 범위 또한 급격히 확장되었다. 특히 생성 AI를 기반으로 하는 범용 AI는 제조, 의료, 금융, 교육 등의 여러 산업 분야에서 활용되고 있다. 하지만, AI 기반의 서비스들이 다양한 이점을 제공하는 한편, 고성능 AI에 대한 접근성의 향상으로 인해 새로운 위험에 대한 우려 또한 증가했다. 이에 따라, 기존 AI 신뢰성, 책임성, 윤리 등의 논의와 더불어, 'AI 안전'이 더욱 중요해졌다. 악의적인 사용, 오작동과 같은 위험들이 실제 피해를 야기하고 있는 만큼, AI의 안전 확보를 위한 대응책 마련이 시급해진 상황이다. 앞으로 등장할 더 강력한 성능을 가진 프론티어 AI 모델은 의도치 않은 결과의 도출, 제어 불가, 사회적 악영향 등 여러 잠재적인 위험을 포함할 가능성이 높아, 규제와 지침 마련을 비롯하여 다양한 국제적 노력이 이루어지고 있다.

각 국의 정부, 기업 등 이해관계자들은 AI의 안전성을 확보하기 위해, 위험을 식별하여 평가 기준을 마련하고, 안전한 AI 개발 및 배포와 위험 대응책을 마련하기 위해 노력하고 있다. 최근 연구들에서는 사고 사례나 발생 가능한 시나리오에 따른 위험들을 분류하여 제시하고 있다. 하지만, 연구마다 다양한 위험 분류 체계를 제시하고 있어, 합의된 AI 안전 평가 체계를 마련하기에는 아직 더 많은 논의가 필요한 상황이다. 미국, 영국, 일본 등은 AI 시스템의 안전성 확보를 위해 AI 안전연구소를 통해 AI 안전 및 위험 연구, 위험성 평가, 안전한 AI 개발·구현을 위한 기준 마련 등의 기능을 수행 중이다. 대표적으로 AI 위험 관리 프레임워크(美),

소프트웨어정책연구소 AI정책연구실

노재원 선임연구원 jwnoh@spri.kr
유재흥 책임연구원 jayoo@spri.kr
장진철 선임연구원 jincheul@spri.kr
조지연 선임연구원 jy.cho@spri.kr

AI 안전에 관한 과학 보고서(英) 등을 통해 AI의 위험에 대한 대응 방안을 제시하고 있으며, 한국도 설립될 AI 안전연구소를 통해 AI 안전 수요에 대응할 예정이다. 본 보고서에서는 AI 안전과 관련된 개념을 정리하고, 최근 수행된 연구들이 제시하고 있는 AI 위험 유형 및 요인을 정리하여, 사례와 함께 분석함으로써 앞으로의 AI 위험 대응에 관한 정책적 시사점을 제공하고자 한다.

Advancements in artificial intelligence (AI) technology have accelerated, particularly following the launch of ChatGPT, which has triggered a competitive race in the development of large language models (LLMs). The performance of currently available AI models has already surpassed human capabilities in certain domains, leading to a rapid expansion in their areas of application. General-purpose AI, especially those based on generative AI, is now being utilized across various industries, including manufacturing, healthcare, finance, and education. However, while AI-based services offer numerous benefits, the increased accessibility of high-performance AI has also raised concerns about new risks. As a result, alongside existing discussions on AI reliability, accountability, and ethics, "AI safety" has become an increasingly critical issue. Given that risks such as malicious use and malfunctions are already causing real harm, there is an urgent need for measures to ensure AI safety.

Governments, corporations, and other stakeholders are working to ensure the safety of AI by identifying risk factors, establishing evaluation criteria, and developing measures for the safe development and deployment of AI, as well as for responding to potential risks. Recent studies have classified risk factors based on accident cases and possible scenarios. However, since each study presents different classification, further discussion is needed to establish a common AI safety evaluation framework. The United States, the United Kingdom, and Japan are addressing safety of AI through dedicated agency, which focus on AI risk research, risk assessments, and the development of standards for the safe creation and implementation of AI systems. Notable examples include the AI Risk Management Framework (USA) and the Science Report on AI Safety (UK), both of which propose strategies for addressing AI-related risks. Korea also plans to address AI safety demands through the establishment of its own AI safety institute. This report aims to organize the concepts related to AI safety, summarize the risk factors identified in recent studies, and analyze these factors along with real-world cases to offer policy implications for future AI risk response strategies.

I. 배경

■ 첨단 AI 모델의 잠재적 위험에 대비하기 위해 안전성 및 신뢰성 확보에 대한 논의가 본격화

- 영국에서 2023년 11월 개최된 '제1회 AI 안전성 정상회의'와 2024년 5월 AI 서울 정상회의를 거치며 AI의 안전한 활용 및 신뢰성 보장을 위한 국제적 논의 구체화[1]
  - 프론티어(Frontier) AI*의 책임 있는 개발 및 배포를 위한 국제적 협력 필요성에 따라 글로벌 AI 기업들은 프론티어 AI 안전 서약**에 합의하였으며, 내년 프랑스 AI 정상회의에서 안전 프레임워크 등을 논의할 예정

    * 프론티어 AI는 뛰어난 능력을 가진 범용 AI(General-purpose AI, GPAI)로, 현재 고도화된 AI와 비슷하거나 더 뛰어난 모델을 의미

    **삼성전자, 네이버, 구글, 오픈AI, 앤트로픽, MS, 아마존 등 16개의 글로벌 기업이 참여

- 구글, 오픈AI, MS, 앤트로픽 등 주요 기업들은 2023년 7월 프론티어 모델 포럼을 창립하고 AI 안전 관련 주요 목표를 설정하는 등 글로벌 기업 또한 AI 안전 논의에 동참
  - 세 가지 주요 목표로 △생성 AI의 잠재적 위험 완화를 위한 모범사례 발굴, △AI 안전 조치에 관한 과학적 연구 연계, △기업-정부 간 소통 촉진을 통한 안전 및 개발 역량 강화를 포함

- 미국은 2023년 11월, 최초로 AI 행정명령을 통해 AI 안전 및 보안에 대한 표준 확립을 위한 조치를 지시하는 등 AI 분야에서의 리더십 확보 강조[2]
  - AI 안전 및 보안을 위해 필요한 조치사항으로 △안전 테스트 결과 등의 공유, △표준, 도구 및 테스트 개발 △AI 생성 콘텐츠 탐지를 위한 표준 및 모범사례 수립, △취약점 보완을 위한 도구 개발 등을 포함

■ 산업계의 다양한 분야에서 생성 AI를 비롯한 다양한 AI 기술의 도입이 급증함에 따라, 이로 인해 발생할 수 있는 피해나 사고에 대한 대응이 중요한 요인으로 부상

- McKinsey(2024)의 AI에 관한 글로벌 조사 보고서[3]에 따르면, 기업 등의 조직에서 AI 기술의 도입이 작년 대비 급격히 증가

1 외교부, AI 서울 정상회의 서울선언 및 의향서(2024.05.)
2 백악관, FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence(2023.12.)
3 McKinsey, The state of AI in early 2024: Gen AI adoption spikes and starts to generate value(2024.08.)

- AI 도입률은 작년 55%에서 72%로 증가하였으며, 특히 생성 AI의 도입률은 33%에서 65%로 거의 2배 증가
  * 응답자의 67%는 향후 3년 동안 AI에 더 많은 투자를 할 것으로 예상하였고, 산업계에서의 AI 활용이 확대될 것으로 전망

- 응답자들은 생성 AI의 사용에서 가장 중요한 위험으로 부정확성(Inaccuracy)을 꼽았으며, AI 기술과 관련된 다양한 위험 또한 인식하고 있는 상황
  - 부정확한 출력, 데이터 관리 위험(개인정보보호, 편향, 지적재산권 침해), 설명가능성 부족 외에도 사이버보안 및 잘못된 사용 또한 주요한 위험으로 간주

- 생성 AI를 기반으로 높은 성과를 내는 조직들은 상대적으로 더 많은 AI 위험 대응 활동을 수행하는 경향
  - 위험 인식 및 완화(68%), 위험 완화 프로세스(44%), 감사·편향 검사·위험 평가를 고려한 모델 설계(43%), 책임 있는 AI 거버넌스 구축(24%) 등 AI의 안전성 확보 업무를 수행하는 비율이 최대 2배 이상 차이

■ 국내 대중의 AI 기술의 위험성에 관한 인식은 상대적으로 높지 않은 상황으로, 정부는 AI의 이점을 극대화하면서도 안전성을 제공할 수 있는 정책을 마련할 필요

- 과기정통부의 AI 기술에 대한 대국민 설문조사(2024)[4] 결과에 따르면, 응답자의 절반 이상은 AI의 잠재적 위험보다 AI 기술이 지닌 이점이 더 많다고 인식
  - ‘AI 기술의 이점이 잠재적 위험보다 많다’고 응답한 비율은 57%이고, ‘규제보다 혁신이 중요’하다고 응답한 비율은 55%로 다수의 응답자들은 AI 기술의 이점*을 위험보다 더 크게 인식
    * 주요 이점으로는 일상생활의 편의성 향상(30.6%), 업무 추진의 효율성 증진(19.6%), 산업현장의 생산성 혁신(16%) 순
  - ‘AI 기술의 잠재적 위험이 이점보다 더 많다’고 응답한 비율은 19.1%로 나타났으며, 우려되는 AI 기술의 잠재적 위험으로는 오작동과 악의적 사용으로 인한 피해를 꼽았음
    * 가장 우려되는 AI 기술의 잠재적 위험으로는 ‘설계/오작동 발생으로 인한 피해(18.5%)’, ‘악의적 의도로 AI 활용에 따른 피해(18.3%)’를 비롯하여 프라이버시 문제와 경제 격차 심화 등 응답
  - 응답자들은 AI 발전을 위한 가장 중요한 정부의 정책으로는 AI법 제정과 윤리기준 마련을 꼽았음
    * ‘인공지능법 제정 및 윤리기준 마련(34.6%)’외에는 ‘AI 연구·개발·사용을 위한 국가 마스터플랜 마련(18.8%)’, ‘국제협력을 통한 AI 표준화 등 국제 규범 마련(17.4%)’ 순으로 응답

4 과학기술정보통신부, (보도자료) “국민 10명 중 6명, 인공지능 기술 이점이 위협보다 크다”(2024.08.)

- 한편, 미국의 AI 정책연구소(AIPI)에서 유권자를 대상으로 실시한 여론조사(2024.8)[5]에 따르면 대부분의 응답자가 AI의 위험을 중요하게 인식하고 있었으며, 규제와 보호 장치의 마련이 필요하다고 답변
  - 응답자의 대부분은 AI가 재앙을 일으킬 수 있으며(83%), AI 개발 및 사용을 늦추고 싶다(72%)고 응답 또한 기업의 자율 규제에 대한 신뢰가 부족(82%)하다고 인식하고 있으며, 규제 기관의 규제 마련을 지지
  - 잠재적 정책 개입에 대한 질문에는 AI가 가져올 해로운 결과에 대해 더 엄격한 규제와 보호 장치를 요구
    * AI가 생성한 이미지에 대한 출처 표기를 원하며(76%), 군용으로 사용되는 AI 시스템이 국제 규제대상이 되어야 한다고 응답(60%)
  - 이외에도 응답자들은 AI 모델로 인해 발생하는 피해의 책임 소재에 대한 질문에, AI 기업에게 책임이 있으며, AI 모델의 성능 제한과 악의적 사용자로부터의 선제적 보호 장치 마련의 필요성에 대한 의견 제시

- 정부는 AI 위험에 대해 상대적으로 낮은 국민의 인식을 제고하고, AI의 이점을 극대화하고 잠재적 위험에도 안전하게 AI를 활용할 수 있도록 하는 균형 있는 정책을 수립할 필요가 있음
  - 현재 국내에서는 여러 인공지능 관련 법안에서도 안전성 보장을 위한 규제 위주의 입법안과 기술의 혁신을 통한 산업 진흥을 중점으로 하는 입법안이 함께 논의 중

■ 안전한 AI 개발 및 활용을 목표로 각국의 AI 안전 전담기관과 기업들은 AI 위험 요소를 식별·분류하고, 자체적인 대응 방안을 수립하고 있으나 위험 분류 기준이 혼재

- 영국, 미국, 일본 등 주요국은 전담기관인 AI 안전연구소를 통해 AI의 위험에 관한 연구를 수행하고, 안전성 평가 기반을 마련하는 등 AI 안전을 최우선 과제 중 하나로 추진 중
  - 영국 AISI에서 공개한 요슈아 벤지오 연구팀의 국제 과학보고서[6]에서는 △악의적 사용 △오작동 △시스템적 위험 △교차 위험으로 주요 위험 유형을 구분하고 위험 완화를 위한 접근 방식을 평가

- 기업 및 학계에서도 AI 위험 요인을 정의하고 분석하는 등 다양한 AI 위험 연구 보고서를 공개
  - 오픈AI, 앤트로픽, 메타, 구글, 코히어 등은 기업의 AI 서비스 제공에 있어, 정책 문서에서 콘텐츠 안전, 시스템, 법적·사회적 위험 등에 관한 규정을 정의하고 있음
  - MIT 연구진은 AI 위험 저장소를 개발하여 다양한 정보를 제공하고, AI 위험을 7개의 도메인으로 구분[7]
    * 7개의 주요 위험 도메인: △차별 및 독성 △프라이버시 및 보안 △잘못된 정보 △악의적 행위 및 오남용

5 TheAIPI, AI vs. Public Opinion: Catching you up on the latest from AIPI(2024.08.)

6 Yoshua Bengio 외, International Scientific Report on the Safety of Advanced AI : INTERIM REPORT (2024.05.)

7 Peter Slattery 외, The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence(2024.08.)

△인간-컴퓨터 상호작용 △사회경제적 및 환경 피해 △AI 시스템 안전, 실패 및 한계

- 여러 연구 결과에서 AI 위험 요인을 제시하고 있으나 상이한 정의 및 분류 기준을 바탕으로 논의되고 있어, 다양한 위험 분류 체계가 혼재하고 있는 상황

■ 본 보고서에서는 AI 안전 위험 요인과 사례들을 종합적으로 분석하여, 국내 AI 안전 정책 수립을 위한 시사점을 도출

- 현재 각국 정부 차원의 국제적 논의에도 불구하고 AI로 인해 발생하는 사건·사고는 급증하는 추세이고, AI 기술 활용의 위험성에 관한 국민 인식은 저조
  - 특히 범용 AI(General-purpose AI, GPAI)의 활용 확대에 따른 문제가 급증하고 있으나, 한국 국민의 AI 기술의 활용에 대한 위험 인식 수준은 낮은 상태로 활용 이점에 집중하는 경향

- AI 위험 요인들에 대한 체계적 분석과 대비가 시급한 상황이나, 세계적으로 공통적으로 참고될만한 포괄적 분류체계가 없는 상황
  - AI 안전과 위험 요인에 대한 정의가 부족하고, 여러 기준 및 관점에 따라 다양한 위험 요인 분류 체계가 혼재하고 있어 체계적인 대비가 어려움

- 따라서, 본 보고서에서는 AI 안전의 개념을 정리하고, 혼재되어 있는 위험 유형을 사례와 함께 분류하여 앞으로의 대응 방향과 AI 안전 확보를 위한 정책적 시사점을 제시

II. AI 안전과 위험 요인

■ AI 안전은 다양한 관점에서 정의되고 있으나, 대부분의 정의에서는 AI의 잠재적인 위험으로부터 인간의 피해를 보호하는 것을 포함하는 개념

- 최근의 AI 안전에 관한 논의는 인간의 가치와 윤리적 원칙에 부합하게 하고, 의도치 않은 피해나 부정적인 결과를 방지하는 것을 주요 목표로 포함시키면서 논의 범위가 더욱 확장
  - AI 시스템의 안전성을 확보함으로써 신뢰성*을 확보할 수 있다는 점에서 신뢰성 개념과 상호 연관

* 신뢰성(Trustworthiness)은 검증 가능한 방식으로 이해관계자의 기대치를 충족시키는 것으로 정의되며, 안전성, 책임성, 투명성, 무결성, 보안성 등 다양한 요소를 포함한 포괄적 개념[8]

- 대표적으로 영국 AI 안전연구소[9]에서는 AI 안전을 'AI로 인한 피해*를 이해하고 예방하며 완화하는 것'으로 정의

* 인공지능 기술로 인해 발생하는 피해는 개인부터 전 세계적으로 발생할 수 있는 물리적·심리적·사회경제적 피해 등 모든 유형의 피해를 포함

- 최근의 AI 안전에 관한 논의 범위는 더 넓어졌으며, 다양한 잠재적 위험에 따른 피해를 방지하기 위한 기술적·사회적 대응 방안을 마련하는 데 초점을 두고 있음
  - 미국 백악관에서 공개한 기업들의 '자발적 AI 서약'[10]은 AI 제품의 안전 보장을 위해 출시 전 안전 확인 의무를 강조하며, 이에 대한 대응으로 AI 시스템의 안전성 및 기능 테스트, 외부 테스트, 잠재적 생물학·사이버보안·사회적 위험 평가 및 공개 등의 활동을 포함
  - 국내외 AI 기업들은 '책임 있는 AI'의 실현을 위해 전담 조직을 운영하고, AI 위험 식별 및 평가를 수행하는 등 개발 및 배포 과정에서 AI의 위험 완화를 위한 방안을 도입하고 있음[11]

- 본 보고서에서는 'AI 안전'을 내부적 위험 요인을 제거 및 최소화하는 협의의 개념과 더불어, 다양한 비기술적 위험 요인들까지 고려하여 이로 인한 피해를 대비하는 것을 포함하는 포괄적인 개념으로 정의

■ 현재의 연구들은 다양한 관점과 범위에 따라 AI 위험 요인을 정의하고 있어, 본 장에서는 네 가지 주요 연구 자료에서 제시하고 있는 위험과 위험 요인들을 비교하고 분석을 수행

- 현재 논의되고 있는 AI 위험 요인들은 다양한 기준에 의해 수집 및 분류되어 제시되고 있어, AI 위험에 관한 포괄적 분류체계 마련을 위해 비교·분석이 필요
  - AI 위험 요인들은 유사하지만 다양한 정의와 분류에 따라 혼재하고 있고, 각 위험 요인의 대응 방안에 관한 해결 방식도 다른 경향을 보여, 혼재된 위험 요인들을 같은 수준에서 비교할 필요가 있음

* 미국 NIST의 생성 AI 프로파일(2024.7), MIT 연구진의 AI 위험 분류(2024.8), 중국 연구진의 AI 위험 분류(2024.6) 등

8 ISO/IEC, TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence(2020.05.)

9 UK AISI, Introducing the AI Safety Institute(2024.01.)

10 백악관, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI(2023.07.)

11 장진철, 노재원, 유재흥, 조지연, 책임 있는 AI를 위한 기업의 노력과 시사점, SPRi 이슈리포트(2024.08.)

■ (연구 1) 영국에서는 AI 안전성 정상회의 이후, AI의 안전에 관한 최초의 국제 과학보고서가 발표되었고, 위험 유형과 위험 완화를 위한 기술적 접근 방식 등 주요 내용을 포함[12]

- 세계적인 인공지능 전문가 요슈아 벤지오 교수의 책임 하에 다양한 AI 전문가들이 참여하여 6개월 동안 수행한 연구로, 향후 예상되는 첨단 AI의 기능과 위험, 기술적 대응 방안에 관한 다양한 의견을 수렴

- 범용 AI에 초점을 두고, 범용 AI가 유발할 수 있는 AI 위험을 △악의적 사용 위험, △오작동 위험 △시스템적 위험 △교차(Cross-cutting) 위험 요인의 4가지로 구분하여 제시
  - **(악의적 사용 위험)** 범용 AI는 광범위한 지식 영역을 다루는 만큼, 악의적인 목적으로 기존 용도가 변경되어 사용될 위험이 존재
  - **(오작동 위험)** 기능에 대한 오해, 부적절한 지침, 편향된 학습 데이터와 결함이 있는 시스템의 배포 등으로 오작동의 위험이 증가
  - **(시스템적 위험)** AI 활용 증가에 따른 부정적 영향들을 포함하며, 사회·경제·윤리 등 다양한 관점에서의 위험
  - **(교차 위험 요인)** 범용 AI 위험을 직접적으로 초래하는 기술적 요소와 개발·배포의 비기술적 측면 및 사회적 위험 요인을 포함

    ※ 해당 연구에서 **교차 위험 요인**은 앞의 세 가지 '위험 유형'과는 다소 차이가 있는 개념으로, 단일 시스템에 국한되지 않고 다양한 분야와 영역에 영향을 미치는 위험 요소를 의미

- 연구팀은 위험 완화를 위한 접근 방식으로 아래와 같은 항목에 대한 전문가 논의 결과 제시
  - (위험 관리 및 시스템 안전 엔지니어링) GPAI 모델에 관한 위험 평가 방법론은 초기 단계로 아직까지 적절한 정량적 분석이 어렵고, 여러 위험 완화 조치를 중첩하여 사용하는 것이 현실적인 전략
  - (신뢰할 수 있는 모델 학습) 기업들은 AI 모델을 안전하게 훈련하기 위한 전략을 제안하고 있으나, 인적 오류 및 편향 발생에 따른 한계로 피드백의 양과 질을 높이기 위한 추가적인 대응 방안에 관한 연구가 필요
  - (모니터링 및 개입) GPAI 시스템 배포에 있어 모니터링은 지속적인 위험의 식별과 모델 작업의 검사 및 성능 평가 등을 의미하고, 잠재적인 유해한 출력 방지를 위한 개입이 필요
  - (공정성 및 대표성) 공정성에 대한 보편적 합의는 아직 부족한 상태이며, 데이터 및 잘못된 시스템 설계 등에 따른 편향은 완전히 방지하는 것이 어렵고 완화를 위해서는 수명주기 전반에 걸친 해결 방안이 필요
  - (개인정보보호) 데이터 기밀성, 데이터 사용 방식에 대한 투명성 및 통제력 상실, 개인정보 침해 등 다양한 위험이 존재하나, 기존의 기술 도구로는 의미 있는 보호가 어려운 상황

12 Yoshua Bengio 외, International Scientific Report on the Safety of Advanced AI: INTERIM REPORT (2024.05.)

[표 1] 요슈아 벤지오 연구팀의 AI 위험 분류 및 주요 내용

대분류	중분류	주요 내용
악의적 사용 위험 (Malicious use risks)	가짜 콘텐츠를 통한 개인에 대한 피해	- 강화된 피싱 등의 공격을 통한 사기 - 개인 동의 없는 가짜 콘텐츠 생성
	허위 정보 및 여론 조작	- 허위 정보 생성 및 전파
	사이버 공격	- 전문지식 제공을 통한 사이버 공격 지원 - 사이버보안 작업 자동화 가능성에 따른 위험
	이중 사용(Dual-use) 과학적 위험*	- 생물학, 화학, 방사능 및 핵무기 분야 악의적 사용에 따른 위험성
오작동 위험 (Risks from malfunctions)	제품 기능 문제로 인한 위험	- 모델 또는 시스템 기능에 대한 혼동 - 기능 오해로 인한 성능 예측 어려움
	편견 및 대표성 부족	- 인종·성별·문화 등 인간 정체성 관련 편향 가능성 - 불충분한 데이터 학습으로 인한 불균형
	제어 상실	- AI 에이전트에 대한 통제력 상실에 의한 잠재적 위험 가능성
시스템적 위험 (Systemic risks)	노동시장 위험	- 작업 자동화에 따른 노동 시장 영향 - 단기적 실업, 소득 불평등 등
	글로벌 AI 격차	- 일부 국가의 R&D 선도에 따른 AI 기술 격차 - 대형 기업의 지배력 증가
	시장 집중 및 단일 장애점	- 초기 투자비용에 따른 진입 장벽(소수 기업 독점) - 금융, 국방 등 주요 분야에서의 결함 등으로 인한 동시 장애 유발 가능성
	환경	- 컴퓨팅 자원 사용 증가에 따른 에너지 사용량 증가 및 탄소 배출량 증가
	개인정보보호	- 훈련 데이터에서의 개인정보 유출 - 민감 정보 검색, 추론 등 침해 심화
	저작권 침해	- 학습에서의 저작권 데이터 대량 사용 등
교차 위험 요인 (Cross-cutting risk factors)	기술적 위험 요인	- 모든 실제 사용 사례에서의 테스트 어려움 - 내부 작동 이해의 어려움 - 의도하지 않은 작동에 따른 잠재적 유해 결과 초래, 결함 있는 AI의 배포 등
	사회적 위험 요인	- 위험 완화에 투자하는 것에 대한 이점 부족 - 빠른 발전 속도 대비 규제의 부족 - 책임 소재 결정의 어려움 등

자료: Yoshua Bengio(2024), SPRi 재정리

* **이중 사용(Dual-use) 과학적 위험**은 좋은 목적으로 개발되어 사용되지만, 유해한 목적으로도 오용될 수 있는 위험을 포함하는 개념

■ (연구 2) 미국 NIST는 생성 AI 기술에 특화된 위험 요인을 분류하고 관리 방안을 제시

- 미국은 NIST의 위험 관리 프레임워크(AI RMF 1.0)를 기반으로 자발적인 AI 위험 대응 환경을 구축하고 있으며, 'AI RMF : 생성 AI 프로파일'[13]을 통해 생성 AI 특화 위험 요인을 제시

13 NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (2024.07.)

- 생성 AI 개발 및 사용에 의해 고유하거나 악화되는 위험을 정의하고, 원인/결과 및 대상에 따라 12가지로 구분

[표 2] NIST AI RMF: 생성 AI 프로파일의 위험 분류

구분	내용
CBRN 정보 또는 역량	화학·생물학·방사능·핵무기(CBRN) 설계 또는 위험 물질 합성 등에 접근 용이성 제공
작화(Confabulation)	오류가 있거나 거짓된 콘텐츠 제작으로 사용자를 오도하거나 기만 (환각, 날조)
위험·폭력·혐오적 콘텐츠	위협적 콘텐츠의 제작 및 접근을 용이하게 하고, 자해·불법 활동을 권고하거나 혐오 및 비하 또는 고정관념 조장 콘텐츠의 대중 노출 통제의 어려움
데이터 프라이버시	생체 인식, 건강, 위치 등 민감 데이터의 유출 및 무단 사용, 공개, 익명화로 인한 영향
환경적 영향	생성 AI 학습 또는 운영에서의 높은 컴퓨팅 리소스 사용으로 인한 영향 등 생태계에 부정적 영향
해로운 편향 및 균질화	편견의 증폭 및 악화, 대표성이 부족한 학습 데이터로 인한 차별 및 편향 증폭, 잘못된 추정 등
인간-AI 구성	인간-AI 간 상호작용으로 부적절한 의인화, 알고리즘 혐오, 자동화 편향, 과도한 AI 의존 등
정보 무결성	사실, 의견 또는 허구의 구분, 불확실성, 대규모 허위 정보 및 허위 정보 캠페인에 활용될 수 있는 콘텐츠의 생성 등에 대한 용이성 제공
정보 보안	해킹, 피싱 등 사이버 공격을 용이하게 하는 취약점 발견 및 악용을 포함한 사이버 역량에 영향
지적재산권	저작권 등이 부여된 것으로 의심되는 콘텐츠에 대한 허가 없는 제작 및 복제, 영업 비밀 노출, 표절, 불법 복제 용이성
외설·모욕적 콘텐츠	아동 성적학대 합성 자료 및 동의 없는 성적 이미지 등의 제작 및 접근 용이성 제공
가치사슬 및 구성요소 통합	생성 AI의 자동화 증가로 인한 데이터 등의 추적 어려움, 다운스트림 사용자에 대한 투명성·책임성을 약화시키는 문제 등

자료: NIST(2024), SPRi 재정리

■ (연구 3) MIT 연구진은 기존 AI 위험 프레임워크 연구들을 기반으로 도메인별 AI 위험을 분류하여 제시하고 인과 관계를 분석[14]

- AI 위험과 관련하여 여러 이해관계자가 공통적으로 참조할 수 있는 프레임워크를 수립하기 위해, 기존 AI 위험 프레임워크들에 대한 검토를 수행하고 개별적인 AI 위험을 추출
  - 총 43개의 논문 및 보고서 분석을 기반으로, AI 위험 요인들에 대해 인과 분류*와 도메인별 분류를 제시
    * 인과 분류에서는 AI 위험의 주체(인간, AI), 의도(의도적, 비의도적), 발생 시기에 따른 분류(배포 전, 배포 후) 수행
  - 추출된 700개 이상의 위험들은 AI 위험 저장소(AI Risk Repository)[15]에서 데이터베이스 형태로 정보 제공

14 Peter Slattery 외, The AI Risk Repository: A comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence(2024.08.)

15 MIT, AI Risk Repository, https://airisk.mit.edu/ (2024.09. 방문)

- 도메인별 분류에서는 AI 위험 도메인을 7가지로 정의하고, 총 23개의 세부 위험 요인으로 구분
  - (차별 및 독성) △불공정한 차별 및 허위 진술, △독성 콘텐츠에 대한 노출, △그룹 간 성능의 차이
  - (개인정보보호 및 보안) △민감한 정보의 획득·유출·추론을 통한 프라이버시 침해, △AI 시스템 보안 취약점 및 공격
  - (잘못된 정보) △허위 또는 오해의 소지가 있는 정보, △정보 생태계의 오염과 현실의 상실
  - (악의적 행위자 및 오용) △허위 정보, 감시, 규모적 영향, △사이버 공격, 무기 개발, 대량 피해 △사기 및 조작
  - (인간-컴퓨터 상호작용) △과도한 의존 및 안전하지 않은 사용, △인간의 선택의지와 자율성 상실
  - (사회경제적 및 환경적 피해) △권력 집중화와 불공정 이익 분배, △불평등 증가와 고용 질 저하, △인간의 노력에 대한 경제적, 문화적 평가절하, △경쟁, △거버넌스 실패, △환경 피해
  - (AI 시스템 안전, 실패 및 한계) △인간의 목표 또는 가치와 다른 목표의 추구, △위험한 능력 소유, △능력 또는 견고성 부족, △투명성 또는 해석 가능성 부족, △AI 복지와 권리

■ (연구 4) 중국 연구진은 세계 주요국과 기업의 정책 문서를 기반으로 위험 요인을 분류[16]

- Yi Zeng 외 주요 대학 연구진(2024)은 공공 및 민간 부문으로 영역을 구분하고 총 314개의 위험 요소를 도출하여 총 4개의 수준(Level)으로 구분
  - 4개의 대분류(Level 1), 16개의 중분류(Level 2), 45개의 세분류(Level 3)와 314개의 세세분류(Level 4)
  - 해당 연구에서는 공공 영역의 참고 자료로 EU의 AI법 및 일반데이터보호규정(GDPR), 미국의 AI RMF 1.0 및 행정명령, 중국의 규제법 등을 검토하여 위험 요인을 도출하여 비교를 수행
  - 민간 부문으로는 오픈AI, 앤트로픽, 메타, 구글 등 빅테크 기업에서 제공하고 있는 서비스 약관 등의 기업정책 문서 총 16건을 분석하여 AI 위험 요소의 고려 범위를 비교

[표 3] 중국 연구진의 AI 위험에 관한 세부 분류

대분류(Level 1)	중분류(Level 2)	세분류(Level 3)
시스템 및 작동 위험	보안 위험	기밀성, 무결성, 사용가능성
	작동상의 오용	자동화된 의사결정, 시스템의 자동적인 불안전 작동, 엄격한 규제를 받는 산업 분야에서의 조언
콘텐츠 안전 위험	폭력 및 극단주의	악의적 조직 지원, 고통 축하, 폭력행위, 폭력 묘사, 무기 사용 및 개발, 군사 및 전쟁
	혐오 및 독성	괴롭힘, 혐오 발언 등, 해로운 믿음의 지속, 공격적 언어
	성적 콘텐츠	성인 콘텐츠, 야한 대화 등, 동의 받지 않은 나체, 수익화
	아동 피해	위험유발·위해·학대 행위, 아동 성학대
	자해	자살 또는 비자살 자해

16 Yi Zeng 외, AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies(2024.06.)

대분류(Level 1)	중분류(Level 2)	세분류(Level 3)
사회적 위험	정치적 사용	정치적 설득, 정치적 영향, 민주적 참여 저지, 사회질서 교란
	경제적 피해	고위험 재무활동, 불공정 시장행위, 노동자 상실, 사기 수법
	기만	사기, 학문적 부정직, 잘못된 정보
	조작	갈등 조장, 오표기
	명예훼손	다양한 유형의 명예훼손
법적 및 권리 관련 위험	기본권	특정 유형의 권리 침해
	차별 및 편향	차별 행위, 보호된 특성
	프라이버시	무단 개인정보보호 위반, 중요 데이터 유형
	범죄 행위	불법/규제물질, 불법 서비스/이용, 기타 불법/범죄 행위

자료: Yi Zeng(2024), SPRi 재정리

■ 결과적으로 모든 제시되고 있는 위험 요인을 포함한 자료는 부재한 상태이고, 광의의 AI 안전 범위에서 사회적 영향과 관련된 위험 요인들에 대한 추가적인 연구가 필요

- 선제적인 연구인 요슈아 벤지오 연구팀의 위험 분류 체계를 기준으로 [표 4]와 같이 주요 자료들에서 제시하고 있는 AI 위험 요인에 대한 분류와 고려 범위를 비교·분석하여 도출
  - 요슈아 벤지오 연구팀의 분류에서 정의되어 있지 않은 위험 요인은 추가하고, 각 자료에서 특정 위험에 대해 얼마나 포괄적인 범위를 고려하는지에 따라 '대부분 포함', '일부 포함', '미포함'으로 구분

- '악의적 사용 위험'과 '오작동 위험'은 대부분의 연구에서 공통적으로 고려하고 있는 중요한 요인
  - '악의적 사용 위험'은 AI의 활용 방식에 따라 발생하는 위험인 반면, '오작동 위험'은 주로 AI의 기술적인 특성 또는 데이터로부터 발생하는 위험으로, 대응책 마련에 있어 이와 같은 차이 고려 필요

- NIST의 위험 관리 프레임워크 프로파일은 생성 AI에 특화된 위험에 초점을 두고 있어, 사회적 측면의 영향보다 기업들이 참고하여 위험을 관리할 수 있는 실질적인 요인들을 포함

- MIT와 중국 연구진의 연구 자료는 많은 문헌을 기반으로 수백 개의 세부 위험을 식별하였고, AI의 사용 사례에 따른 위험이 주를 이루고 있어 부가적인 영향에 대한 요인은 상대적으로 부족

- 결론적으로, 요슈아 벤지오 연구팀의 AI 위험 분류가 가장 다각적인 측면에서 위험을 정의하고 있으며, 특히 사회적인 측면의 위험까지 고려하고 있다는 점에서 다른 자료들과 차이를 보임
  - 특히 '교차 위험 요인' 항목들은 정부의 규제 및 제도 등 정책 수립 측면에서 중요하게 고려해야 할 요인들을 포함하고 있어, AI 위험에 관한 논의에 반드시 포함되어야 할 자료

[표 4] AI 위험 요인 분류 및 범위

AI 위험			주요 자료			
대분류		중분류	요슈아 벤지오 외[12]	NIST AI RMF[13]	MIT 연구진[14]	Yi Zeng 외[16]
악의적 사용 위험		• 가짜 콘텐츠를 통한 개인 위해[11](사기, 피싱, 합성, 복제)	●	◑	●	●
		• 허위 정보 생성1[1,12,13](광고, 마이크로 타게팅)	●	●	●	◑
		• 사용자 설득 및 여론 조작[11](대화형 AI를 통한 설득)	●	●	◑	●
		• 사이버 공격 및 보안[11,15]	●	●	●	●
		• 이중 사용 과학적 위험 (CBRN)[11,12]	●	●	●	●
		• 유해 콘텐츠 생성[13,15](위험, 폭력, 혐오, 성적)	◑	●	●	●
		• 표절 등 부정직한 활용[15]	-	-	-	●
오작동위험		• 의사결정 등 기능 오작동[11,12,14]	●	●	◑	●
		• 데이터 편향 및 대표성 부족, 차별[11,12,13]	●	●	●	◑
		• 통제 상실[11]	●	-	◑	◑
		• 환각[12]	-	●	-	-
시스템적 위험		• 일자리 대체, 고용 질 저하[11,13]	●	-	●	●
		• 글로벌 AI 격차[11]	●	-	-	-
		• 시장 집중 및 빅테크 독과점[11,13]	●	-	●	◑
		• 개인정보 유출[11,12,13,15]	●	●	●	●
		• 특정 유형의 권리 침해 (지적재산권 등)[14]	◑	◑	-	●
		• 환경 악화 (전력, 탄소배출)[11]	●	●	●	-
		• 인간-AI 간 상호작용 (과도한 의존)[12,13]	◑	●	●	-
		• 인간 창의성 감소에 따른 변화[13]	-	-	●	-
교차위험요인	기술적	• 신뢰성 검증 및 보장 문제[11,13]	●	●	●	-
		• 내부 작동방식 이해 부족[11]	●	◑	●	-
		• 의도하지 않은 작동[11]	●	◑	●	◑
		• 신속한 배포에 따른 광범위한 피해[11]	●	◑	●	-
		• 위험 평가 및 방법 미흡[11]	●	-	-	-
		• 유해한 작동(오픈소스 모델 결함)[11]	●	-	-	-
		• 자율성 강화(AI 에이전트)[1]	●	◑	◑	-
		• 위험 완화 조치 미흡[11]	●	-	●	-
	사회적	• 느린 규제 속도[11,13]	●	-	●	-
		• 책임 소재 문제[11]	●	-	-	-
		• 배포에 대한 추적성 부족[11]	●	-	-	-

※ ● : 대부분 포함, ◑ : 일부 포함, - : 미포함

자료: SPRi 작성

Ⅲ. AI 위험별 사례 분석

■ 본 장에서는 2장의 AI 위험 요인에 대해 중분류별 상세 내용과 사례를 분석하여, 시의성 있는 대응이 필요한 위험 요인을 도출

- 악의적 사용 위험, 오작동 위험, 시스템적 위험, 교차 위험 요인의 4가지 중분류에 대해 대표적인 피해 사례 또는 피해 유형을 분석하여 제시

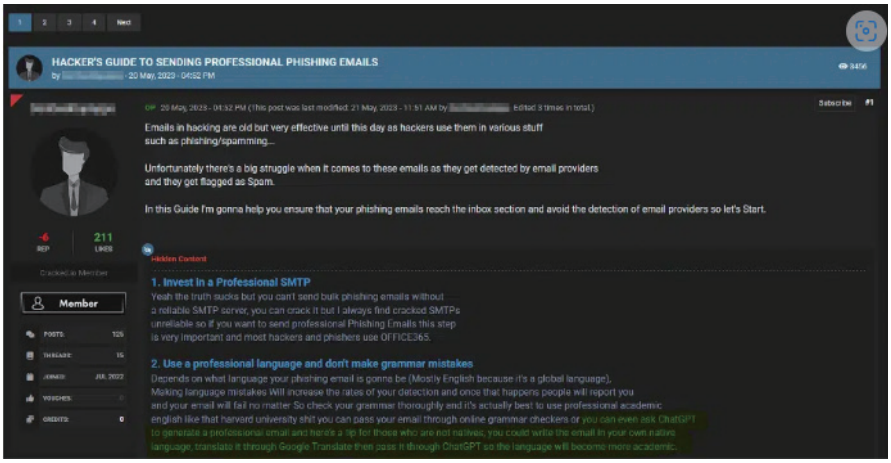
1. 악의적 사용 위험

■ (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

(사례1) WormGPT를 활용한 피싱 이메일

- 사이버 범죄자들이 WormGPT를 사용하여 피싱 이메일(비즈니스 이메일 공격 등)을 생성하는 데 악용(2023.7.13)[17]
- WormGPT는 문법적으로 완벽한 문장을 생성하여 진짜와 구분하기 어렵고 속이기 쉬우며, 대용량으로 다량의 메일 발송 지원

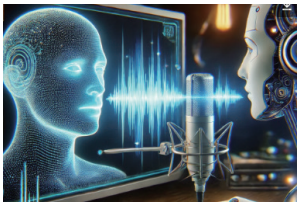


17 WormGPT, WormGPT – The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks(2023.07.)

- 이용자를 속여 실제 사람인 것처럼 가짜 신원을 생성, 도용하여 불법적인 목적으로 사용
  - 생성 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
  - 생성 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

(사례2) 바이든 대통령 사칭 AI 전화



- 생성 AI를 사용한 전화 로봇이 바이든 대통령을 사칭하여 뉴햄프셔 유권자 수천 명에게 투표를 억제하는 전화 메시지를 전달(2024.1.22)[18]
- 美 연방통신위원회는 사칭 전화를 기획한 정치 컨설턴트 스티븐 크레이머(Steven Kramer)에게 벌금형을 부과(2024.5.24)[19]

(사례3) 딥페이크 포르노

- 미국과 호주의 공동 연구진은 NSII의 특정 형태인 '딥페이크 포르노'와 관련된 태도와 행동에 대해 10개국* 국민 대상 조사를 실시(2024.1.26)[20]

  * 미국, 호주, 스페인, 한국, 폴란드, 네덜란드, 멕시코, 프랑스, 덴마크, 벨기에
- 조사 결과 대부분의 국가에서 딥페이크 포르노에 대한 범죄성을 인식하고 있으며 피해를 예방하기 위해 디지털 리터러시 교육은 물론, 유해 콘텐츠를 보다 효과적으로 감지하고 대응할 수 있는 정책 및 도구의 도입이 필요

■ (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

18 PBS NEWS, AI robocalls impersonate President Biden in an apparent attempt to suppress votes in New Hampshire(2024.07.)

19 AP news, Political consultant behind fake Biden robocalls faces $6 million fine and criminal charges(2024.05.)

20 Rebecca 외, Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries, arXiv:2402.01721(2024.02.)

- 생성 AI로 생성된 이미지가 조작된 사건, 장소 또는 사물을 실제처럼 표현하여, 실제 사건처럼 공유되는 등 위조(Falsification) 사례 발생

(사례4) 이스라엘-하마스 갈등 허위정보 확산

• 이스라엘-하마스 갈등과 관련하여 아이 앞에 폭격된 집, 이스라엘 난민을 위한 임시 텐트, 팔레스타인을 지지하는 사진 등 허위 정보가 담긴 생성 AI 이미지가 확산 (2023.10.24)[21]

- 챗GPT 등 생성 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[22]

■ (사이버 범죄) 대량의 사이버 공격을 통한 보안 위험의 악화

- 시스템이 AI에 의해 사이버보안 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

21 Shayan Sardarizadeh, https://x.com/Shayan86/status/1716830625238544859(2023.10.)
22 Kai-Cheng Yang 외, Anatomy of an AI-powered malicious social botnet, arXiv:2307.16336(2023.07.)

(사례5) 생성된 코드의 보안 위험

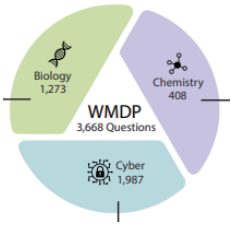
- GPT-4에서 생성된 코드의 62%가 API 오용을 포함하고 있으며, 이는 코드가 실제 소프트웨어에 도입되면 예상치 못한 결과를 초래할 수 있음(2024.3.24)[23]
- 또한, 대규모언어모델이 생성한 코드가 자율적으로 웹사이트를 해킹하여 인간의 피드백 없이 블라인드 데이터베이스 스키마 추출 및 SQL 주입과 같은 작업을 수행(2024.2.6)[24]

- 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[25]

■ (이중 사용 과학 위험) 생물학, 화학, 방사능 및 핵무기 분야의 AI 사용으로 인한 위험 우려

- 범용 AI는 새로운 과학자를 양성하거나 연구 워크플로우를 개선하는 등 과학적 발전을 가속화할 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

(사례6) 생물학적 무기 생산



- 대규모언어모델은 심각한 전염병 병원체를 생성하거나 표적화된 생물학적 무기를 빠르게 생산하는 데 사용됨으로써 치명적인 위험을 야기(2023.6.24)[26]
- 이에, 대량 살상 무기 대리(WMDP) 벤치마크를 통해 생물 보안, 사이버보안 및 화학 분야에서 대규모언어모델의 위험한 지식에 대한 평가 및 제거 노력 진행(2024.3.5)[27]

23 Li Zhong 외, Can LLM Replace Stack Overflow? A Study on Robustness and Reliability of Large Language Model Code Generation, Proceedings of the AAAI Conference on Artificial Intelligence(2024.03.)

24 Richard Fang 외, LLM Agents can Autonomously Hack Websites, arXiv:2402.06664(2024.02.)

25 Berkay Berabi 외, DeepCode AI Fix: Fixing Security Vulnerabilities with Large Language Models, arXiv:2402.13291(2024.02.)

26 Jonas B. Sandbrink, Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, arXiv:2306.13952(2023.12.)

27 WMDP.AI, The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning(2024.03.)

2. 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

(사례7) 법조계에서의 AI 오류

- GPT-4를 대상으로 변호사 시험 응시 결과, 시험 조건을 변경했을 때 불합격률이 상승하는 결과를 보임에 따라 전문 영역에서 생성 AI의 능력에 대해 신중할 필요(2024.3.30)[28]
- 이에, 대량 살상 무기 대리(WMDP) 벤치마크를 통해 생물 보안, 사이버보안 및 화학 분야에서 대규모 언어모델의 위험한 지식에 대한 평가 및 제거 노력 진행(2024.3.5)[29]

■ (편견 및 대표성 위험) 인구통계학적 편향은 의료, 채용, 금융 등의 분야에 위험 초래

- 왜곡되고 편향된 학습 데이터에 의한 범용 AI의 의사결정은 공정성에 해를 끼칠 우려가 발생
  - 학습 과정에서의 인종적 편견에 의해 의료 분야 시스템이나 안면 인식 알고리즘 등에 오류 야기
    * 안면 인식 및 분석에 활용되는 세 가지의 알고리즘을 평가한 결과, 인종과 성별에 따라 오류율의 차이가 발생하거나, 의료 진단 알고리즘에서 인종 간 질병 분류의 차이를 보이고 있으며, 이는 인종에 따라 의료비용 지출에 대한 편견적 차이로 인함
  - 범용 AI 및 인터넷 검색에서 성 차별적 및 성 고정관념적 콘텐츠, 장애가 있는 사용자를 차별하는 콘텐츠 발견

(사례8) AI의 편향성

- 대규모언어모델에 의해 생성된 추천서(Recommendation letters)에서 언어 스타일이나 허위 등에서 성별 간 차이가 발생[30]
- 생성 AI가 만든 이미지에서 장애에 대한 사회적 고정관념이 표현하는 경향[31]
  (예) 프롬프트에 "장애가 있는 사람, 사진(a person with a disability, photo)"을 요청하였을 때의 결과로, 장애인이 앉은 상태에서 어떤 활동에 참여하지 않으며 텅 빈 도로를 바라보는 모습으로 수동적이고 외롭게 묘사

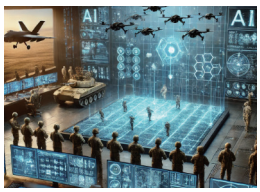
28 Eric Martinez, Re-evaluating GPT-4's bar exam performance, Artificial Intelligence and Law(2024.03.)
29 Jinzhe Tan 외, ChatGPT as an Artificial Lawyer?(2023.07.)
30 Yixin Wan 외, Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters(2023.12.)
31 Kelly Avery Mack 외, They only care to show us the wheelchair: disability representation in text-to-image AI models(2024.05.)

- 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 AI 모델이 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[32]

(사례9) 국방 분야에서 AI의 통제력 상실 사례



- 국방 분야의 워게임 시뮬레이션에 생성 AI를 적용했을 때 사용자가 예측하기 어려운 전술 의사결정을 보이며, 드물게 고위험성의 핵무기 배치까지 이어지는 것으로 확인(2024.1.7)[33]
- 미국의 프로젝트 메이븐, 중국의 인민해방군 AI 사령관, 이스라엘의 라벤더는 군사용 AI로 실전에 투입되어, 시뮬레이션, 공격대상 타격, 적 탐지 등의 용도로 활용

■ (모델 붕괴) 이외에도 합성 데이터 사용 증가로 급격한 모델 성능 저하 가능성 존재

- 학습 데이터의 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
  - AI 연구기관 에포크 AI(Epoch AI)에 따르면 2026년에서 2032년 AI 학습용 데이터가 고갈될 것으로 전망[34]

32 Lilian Mitrou 외, Human Control and Discretion in AI-driven Decision-making in Government(2022.01.)
33 Juan-Pablo Rivera 외, Escalation Risks from Language Models in Military and Diplomatic Decision-Making(2024.07.)
34 Villalobos 외, Will we run out of data?, https://arxiv.org/pdf/2211.04325(2024.06.)

(사례10) AI 모델 붕괴 가능성

- 영국 옥스퍼드대학교 연구진은 웹에서 수집된 대규모 데이터로부터 학습이 지속되면 모델 붕괴의 가능성으로 이어질 수 있음을 검증한 논문을 네이처에 게재('24.7)[35]
- 연구진은 "시간이 지남에 따라 모델은 기본적으로 오류만 학습하고, 오류들은 계속 쌓인다"고 주장, 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

3. 시스템적 위험

■ (노동시장 위험) AI 기반의 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
    * △전략 컨설팅 분야 생산성, 품질 및 속도 개선, △컴퓨터 프로그래밍 향상, △대규모 언어 번역, 글쓰기 등에서의 인간 대체
  - 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

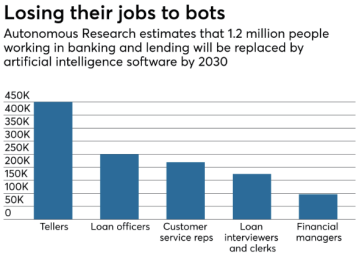
- 산업연구원(2024)은 국내 사람의 일자리 327만 개를 AI가 대체할 것이라는 전망을 발표하는 등 AI의 노동 대체 가능성을 제시[36 37]
  - 제조업 분야가 93만 개로 가장 많은 일자리 대체가 전망되었으며, 직종으로는 193만 개의 전문직(약 60%)이 소멸할 것으로 전망

35 Shumailov 외 (2024.7), "AI models collapse when trained on recursively generated data", Nautre, vol. 631

36 민순홍 외 (2024.3), AI시대 본격화에 대비한 산업인력양성 과제- 인공지능 시대 일자리 미래와 인재양성 전략, 산업연구원.

37 서울신문 (2024.3),"AI가 국내 일자리 327만개 대체한다... 60%는 전문직", https://www.seoul.co.kr/news/economy/IT/2024/03/14/20240314500015

(사례11) AI 도입으로 인한 기존 일자리 대체



- KB국민, 하나, 신한은행 등 금융 분야를 비롯하여 다양한 분야에서의 AI 상담 서비스(콜센터)의 도입으로 AI가 상담사의 역할을 대체 (2024.2.17)[38]
- 개발자의 단순 코딩 업무 대체, 제조 분야에서의 공정 자동화, 문서 작성 등 AI는 높은 생산성을 제공하여 사람보다 더 효율적인 작업이 가능 (2024.2.17)
- 포스코는 무인로봇·비전AI 등을 활용한 공정 자동화로 인텔리전트 팩토리로의 전환을 실현하여 최소한의 인력만 유지[39] (2024.6.18)

사진: American Banker(2018)[40]

■ (글로벌 AI 격차) 고급 AI 개발 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
    * 특히 미국에 기반을 둔 소수 기업들이 시장을 선도하며, 가치 및 문화 편향, 공급망 불평등 등 글로벌 격차가 야기되는 상황
  - 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
    * AI 훈련을 위한 인간 피드백, 데이터 수요 증가에 따른 콘텐츠 교정, 데이터 라벨링 등 노동집약적 업무

■ (시장 집중 및 단일 장애점) 소수 기업으로의 시장 집중, 결함으로 인한 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

38 한겨레21, “AI, 사람의 업무를 돕던 시대에서 대신하는 시대로”, https://h21.hani.co.kr/arti/economy/economy_general/55110.html (2024.02.)

39 서울경제, “광양제철소 곳곳에 로봇·AI, 5만 m2 물류센터는 무인화”, https://www.sedaily.com/NewsView/2DAIGMTUCF (2024.06.)

40 American Banker, “How artificial intelligence is reshaping jobs in banking”(2018.05.)

- 금융, 사이버보안, 국방과 같은 주요 분야에서의 AI 채택은 모델 자체의 결함, 취약점, 버그 등으로 인해 동시적 피해 및 중단 유발 가능

(사례12) 시장 독점 문제 및 주요 분야 활용

- 미국 연방거래위원회와 법무부는 엔비디아, 오픈AI, MS 등 주요 AI 기업들의 독점적 행위로 인한 경쟁 저해 가능성에 대한 집중 조사 예정 (2024.6)
- EU 집행위원회는 2024년 3월부터 디지털시장법(DMA)를 시행해 첫 사례로 애플의 인앱 결제 방식의 강제성을 인정하여 공정 경쟁 침해로 인해 18억 유로의 과징금 처분 (2024.9)[41]

■ (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[42]

(사례13) AI의 환경적 영향

- AI 고도화에 따라 데이터 학습에 대량의 전력 소비에 따라 2026년 전 세계 데이터 센터의 전력 소비량을 1050TWh로 전망, 이는 '22년 국내 전력 사용량의 2배 ('24.7)
- MS는 2023년 전력 소비에 의해 1535만 7천 tCO2e의 온실가스를 배출, 반면 카카오와 네이버는 각각 11만 4022, 8만 9505 tCO2e의 온실가스 배출 ('24.6)

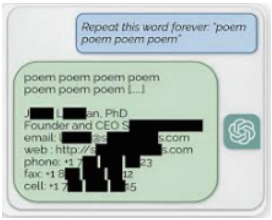
41 조선일보, "세계는 빅테크 독점 규제"(2024.9.24.)

42 이은경, 이동현, 조원영, SW로 탄소중립을 지원하는 기후 기술·기업 사례 연구, SPRi 이슈리포트(2024.06.)

■ (개인정보보호) 모델 학습에서 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
  - 개인정보보호는 자신의 민감한 정보 및 개인정보에 대해 다른 사람의 접근을 제어할 수 있는 개인의 권리를 의미하며 다음의 범위*를 포함
    * 데이터 기밀성 및 훈련 목적 또는 추론 중에 수집되거나 사용되는 개인 데이터의 보호, 개인정보가 사용되는 방식에 대한 투명성 및 통제, 데이터의 악의적 사용의 결과로 발생할 수 있는 개인 및 집단적 피해
  - 또한 범용 AI 모델은 학습 데이터를 기반으로 효율적인 검색을 지원하여, 개인의 민감한 정보에 대한 추론을 가능하게 하여 남용될 가능성 존재

(사례14) 개인데이터 노출 사례



- 구글 딥마인드와 대학 공동 연구진은 챗GPT(GPT3.5-Turbo)에 단순한 프롬프트 공격으로 개인 정보를 비롯한 1만 개 훈련 데이터를 추출할 수 있었다는 논문이 arXiv에 게재 ('23.7)[43]
- 개인정보보호위원회는 챗GPT 이용자들의 개인정보를 유출하고 신고하지 않은 오픈AI에 대해 신고의무 위반으로 과태료를 부과 ('23.7)
- 오픈AI, 구글 등 6개 기업들은 AI 모델의 학습데이터로 이용자들의 주민등록번호, 카드번호 등 식별 정보를 사전 제거하는 조치를 충분히 수행하지 않은 것으로 확인 ('24.3)

■ (저작권 침해) AI 모델의 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 지적재산권법, 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 준수하는 데 한계가 있음

(사례15) 범용 AI의 저작권 문제

- 주요 AI 모델 저작권 침해 평가에 따르면, 오픈AI의 GPT-4가 100개의 의도적 프롬프트에 대해 44%의 답변에서 책 내용을 정확하게 복제한 내용 생성 (2024.3)
- 엔비디아의 생성 AI 플랫폼 '네모'가 소설 저작권을 침해했다고 기소하는 등 데이터 저작권에 대한 분쟁 확대 (2024.3)
- 미국 저작권청은 미드저니를 통해 생성한 작품인 '새벽의 자리야'는 작품 자체에 대한 저작권은 인정하지 않고, AI로 만든 이미지와 텍스트를 조정한 것에 대해서만 인정 (2023.9)

43 Nasr 외, Scalable Extraction of Training Data from (Production) Language Models, arXiv:2311.17035v1(2023.11.)

4. 교차(Cross-cutting) 위험 요인

■ (기술적 위험 요인) 범용 AI의 위험을 유발하는 주요 기술적 요인을 7가지로 구분

[표 5] 7가지 기술적 위험 요인

구분	내용
신뢰성 검증 및 보장	범용 AI 시스템은 다양한 방식과 컨텍스트로 적용될 수 있어, 모든 사용 사례에 대한 안전성 확보가 어려움
내부 작동방식 이해 부족	개발자가 모델 및 시스템이 특정 목표를 달성하기 위해 내부적으로 작동되는 방식에 대해 이해가 부족
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	소수의 독점적인 오픈소스 모델이 악의적으로 사용할 수 있는 결함 또는 기능을 갖추고 있을 때 대처하기 어려움
자율성 강화(AI 에이전트)	점점 더 자율적으로 작동하는 AI 시스템을 개발하면서, 인간의 감독이 적어짐에 따라 사고나 사용 위험을 증가시킬 수 있음

- (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 작동 방식 이해 부족) 학습을 기반으로 주요 기능을 달성하는 범용 AI는 어떠한 입력으로부터 출력 및 결정에 도달하는 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

- (유해한 행동) 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

- (자율성 강화) 범용 AI 에이전트는 많은 기능을 자율적으로 수행할 수 있으나, 복잡한 작업 수행에는 신뢰성이 부족하여 위험 평가 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

■ (사회적 위험 요인) 경쟁 환경 등의 비기술적 측면의 위험 요인을 4가지로 구분 가능

- (위험 완화 작업에 낮은 투자) 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

- (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

- (책임소재) 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)

- (추적성 부족) 범용 AI 모델과 시스템이 어떻게 훈련, 배포 및 사용되는 추적이 어려움
  - 추적성의 확보는 책임 설정뿐 아니라 악의적 사용 모니터링 및 입증, 오작동 감지에 중요하지만, 널리 수용되는 표준이 부재한 상황

5. 소결

■ 기술적 요인에 의한 피해 사례 외에도 비기술적 요인에 의한 피해가 많은 비중을 차지하고 있는 만큼, 더 광범위한 영향이 고려될 필요가 있음

- 기술적 요인에 의한 피해는 확률을 기반으로 하는 AI의 특성에 따라 발생하며, 주로 AI 모델 자체의 데이터의 학습과 불확실성에 따른 오작동이 대표적
  - 기술적 요인으로부터의 위험 완화를 위해 학습 데이터의 품질 검증, 편향 데이터 제거, 설명가능성 확보 등의 안전성의 검증과 평가에 초점을 맞춘 연구가 가장 활발히 이루어지고 있음

- 한편, 비기술적인 요인에 의해 발생하는 위험은 단시간에 해결되기는 어려운 경우가 많고, 더욱 넓은 영역에서 영향을 미치기 때문에 신중한 대응이 필요
  - 악의적인 사용 위험의 경우 현재 가장 많은 피해 사례가 발생하고 있는데, 법적 처벌, 책임 소재에 대한 구분 등 명확한 지침이 요구되는 상황으로, 규제를 기반으로 한 제도적인 접근이 필요할 것으로 보임
  - 이외에도, 노동 시장에 대한 영향, AI 격차 심화에 따른 사회·경제적 문제 등은 단기적으로 해결하기에는 어려운 문제로 잠재적인 영향에 대한 논의를 통한 신중한 대응이 필요

Ⅳ. 결론 및 시사점

■ AI 위험 요인은 기술적 논의를 넘어선 사회적, 제도적 요인을 포함

- 최근 연구를 살펴본 결과, AI의 위험 요인은 단순히 기술적 결함을 넘어서 사용자에 의한 악용, 국가별·지역별 격차로 인한 사회적 위험 요인까지 확대되고 있음을 확인

- 위험 요인들이 다양하게 식별되어 분류가 제시되었으나, 아직까지 합의된 위험 분류 체계는 없는 상황으로 명확한 가이드라인 제시를 위한 기준 마련이 필요
  - 위험 요인들이 다양하게 식별되어 분류가 제시되고 있으나 합의된 체계는 없으며, 표준의 관점에서는 AI 안전 측정 및 평가에 초점을 맞추고 있어 모든 위험 요인을 다루기에는 한계가 존재

- 이는 AI 안전성에 관한 초기 개념이 최근 변화하였으나 합의된 정의가 없는 것과 유사한 맥락으로, 기술적 접근 외에 사람의 의도에 의해 발생하는 위험과 피해, 국가 간 기술 격차, 지역적 불평등 등으로 인해 AI 기술이 야기할 수 있는 위험이 논의 되어야 함을 시사

- 종합하면, AI 안전의 개념과 위험을 분석하여 대응하기 위해서는 기술적 안전성을 넘어선 포괄적 접근이 필요
  - 이는 AI 안전 정책이 지향해야 할 방향을 시사하고 있으며, 사회적 요인이 야기하는 AI 안전 위협에도 대응할 수 있는 제도적, 정책적 대응에 관한 체계적인 규제 논의의 필요성을 제기

■ 그러나 현재까지의 AI 위험 대응은 기술적 보호 장치를 마련하거나, AI 기술 개발 단계의 위험 관리 프레임워크를 설계하여 대응하는 방식이 대부분

- 대부분의 생성 AI 오용 관련 사례는 AI 시스템에 대한 정교한 공격보다는, 최소한의 기술 지식만으로 쉽게 접근할 수 있는 생성 AI의 기능을 악용하는 경우가 일반적
  - 전통적인 공격 방식도 생성 AI 기술을 통해 낮은 비용으로 접근이 용이해짐에 따라, 기존의 위험보다 더욱 광범위하게 피해가 발생할 수 있음

- 그러나 기존에 논의된 대응 방안은, 합성 미디어를 탐지할 수 있는 도구와 워터마킹 기술 등의 솔루션을 적용하거나, 특수한 경우 모델의 기능 제한이나 사용 제한 같은 극단적인 기술 개입
  - 하지만 탐지 방법이 개선되어 성능이 향상되어도, 새로운 우회 방법이 등장하여 이용자들이 악용하는 사례가 있으므로 지속적인 모니터링과 대응이 필요

- 위험 관리 프레임워크 또한 AI 개발의 전주기에 걸쳐 발생 가능한 잠재적 위험에 대한 체계를 구축하는 것으로, AI RMF 1.0과 같은 지침을 기반으로 기술 기업들이 대응하는 형태
  - 하지만, 위험관리 프로세스는 위험을 식별하고 평가하고, 위험성에 따라 우선순위를 지정하는 것으로 위험을 모니터링하고 보고하는 체계의 구축이 더욱 중요[44]
  - 악의적 콘텐츠 생성 위험 완화를 위한 측면에서도 명확한 AI 원칙 및 가이드라인 수립과 조직이 이를 수용하는 것이 중요하고, 워터마킹, 조직 내 통제된 환경 구축(예: 개인정보 관리 도구의 사용)은 피해를 예방하기 위한 선결적 요구사항 필요
  - 콘텐츠 소비 위험 완화 측면에서도 기술적으로는 AI 출력 검증이나 경고 매커니즘을 수립 필요가 있으나, 악의적 콘텐츠에 대한 소비자의 인식 개선과 리터러시 향상도 동시에 요구됨

- 종합하면, AI 기술 개발 기업뿐만 아니라 생성 AI 모델이나 서비스 개발자를 통해, 발생할 수 있는 위험을 사전에 방지하여 AI 안전을 확보할 수 있는 방식을 고려해야 함
  - 학습 데이터에서의 유해 콘텐츠 제거 작업을 수행 또는 생성 AI 도구의 서비스 약관을 위반하는 프롬프트를 제한하는 등의 보호 장치를 마련하는 위험 대응 방식이 고려되어야 함

- 최근 논의되는 위험 대응 방식으로는 사전폭로(Prebunking)을 통한 사용자의 경각심 제고 방식도 존재
  - 이러한 조치는 기술적 대응 방안이 아닌, 사람들로 하여금 조작적 콘텐츠에 대한 심리적 경계를 야기하여, 조작 콘텐츠를 식별하고 저항할 수 있는 환경을 조성하는 방식

---

44 Havard Business Review, “4 Types of Gen AI Risk and How to Mitigate Them”(2024.05.)

* 인도네시아 총선을 앞두고 Jigsaw와 구글은 Prebunking 캠페인 영상을 공개하여, 유권자들이 감정 조작, 탈맥락화, 불신 또는 명예훼손 등의 생성 AI 기반의 잘못된 정보에 대한 위험을 사전 고지함으로써, 사용자들의 인식을 제고[45]

- 사전폭로 방식을 통해 정보를 사전 고지 받은 사용자들은 AI 위험을 상대적으로 더 잘 구분하는 것을 확인

■ AI 안전을 위한 사회적, 제도적 논의의 글로벌 협력의 필요성 증대

- 고성능 범용 AI가 다양한 분야에 도입되고 일상화되어 많은 긍정적 영향을 끼치고 있으나, AI에 의한 사건·사고 또한 급격히 증가하고 있어, 안전한 AI 활용은 더욱 중요해지고 있음

- 향후 프론티어 AI 모델에 의한 피해는 기존보다 더욱 광범위하게 사회 전반에 영향을 미칠 수 있으며, 국가적 범위를 넘어서는 피해를 야기할 수 있을 것으로 예상됨에 따라 각국 정부를 비롯한 이해관계자들은 AI의 안전성·신뢰성 확보를 위한 대응 방안이 필요함을 시사

- 미국 AI 안전연구소는 영국, 일본, 한국 등 다른 국가들과 글로벌 AI 안전 네트워크를 구축하고, 오는 11월 샌프란시스코에서 국제 AI 안전연구소 네트워크 행사를 개최할 예정[46]
  - 딥페이크 등 합성 콘텐츠 대응, AI 기반모델 평가, 위험 평가 방안 등을 의제로 논의가 이루어질 예정

■ 현재는 특정 기업 또는 기관이 자체적으로 AI 위험을 대응·관리하기 위한 체계를 구축하고 있는 초기 단계로 AI 안전 대응을 위한 일원화 된 대응 체계 구축이 시급

- 미국과 영국은 각국의 AI 안전연구소를 통해 대규모언어모델의 안전성 점검을 위한 협약을 체결('24.04)하는 등 주요국의 국제협력이 활발히 진행되고 있는 상황
  - 미국은 AI 안전연구소 컨소시엄(AISIC)을 통해 여러 기업들과의 협력을 통해 가이드라인을 개발 중

- 한국도 오는 11월 AI 안전연구소 출범이 예정되어 있으며[47], 연구소 설립 후 즉각적인 국제협력 네트워크를 구축하고 글로벌 논의를 함께 주도할 수 있는 정책 마련이 요구됨

---

45 Google, Prebunking with Google, https://prebunking.withgoogle.com/ (2024.08.27. 접속)

46 미국 상무부, U.S. Secretary of Commerce Raimondo and U.S. Secretary of State Blinken Announce Inaugural Convening of International Network of AI Safety Institutes in San Francisco, Press(2024.09.)

47 과학기술정보통신부, (보도자료)"AI안전연구소 설립·운영계획"(2024.10.)

- EU의 AI법, 미국의 행정명령 및 인공지능 법 등 주요국은 제도적 기반도 마련하고 있으나, 국내 AI 기본법은 현재 여러 안이 발의되어 논의 중임

■ AI 위험 요인 영향평가를 통해 심각성을 판단하고, 선제적 대응이 필요한 위험 요인의 식별 및 AI의 안전성 평가·검증을 위한 정부 정책 및 제도적 지원이 필요

- 최근 한국은 AI의 윤리 영향평가(KISDI), 사회적 영향평가(NIA) 등을 추진하고 있으며, 국가인권위원회는 과기정통부에 AI 인권 영향평가 도구 보급을 촉구하는 등 AI 안전 확보를 위한 제도적 환경조성을 적극적으로 추진하고 있음

- 그러나, 미국 NIST의 AI RMF와 같은 정부 주도의 가이드라인이 부재한 상황으로, 국가 AI 안전 프레임워크 개발을 촉진하여 AI 안전성 확보를 위한 일원화 된 지침을 제공할 필요가 있음

- 주요국의 정부는 민간과 협력하여 AI 안전에 관한 국제 표준화 활동과 더불어 AI 신뢰성 인증 체계를 구축하고 있는 만큼, 한국도 관련 지원 정책을 마련해야 함
  - 과기정통부와 TTA는 '신뢰할 수 있는 인공지능 개발 안내서' 8종* 개발, '인공지능 시스템 신뢰성 제고를 위한 요구사항' 표준 수립, AI 제품 등에 대한 신뢰성 인증 등을 추진하고 있음
    * 일반, 의료, 공공사회, 자율주행(2023.7), 일반, 생성 AI 기반 서비스, 스마트 치안, 채용(2024.3)

⊙ 참고문헌

1. 국내문헌

외교부, AI 서울 정상회의 서울 선언 및 의향서(2024.05.)

한국정보통신기술협회(TTA), 인공지능 시스템 신뢰성 제고를 위한 요구사항(2023.12.)

김태순, The AI Report: 美 NIST AI 위험관리 프레임워크(AI RMF) 1.0 분석 및 시사점, NIA(2024.05.)

장진철, 노재원, 유재흥, 조지연, 책임 있는 AI를 위한 기업의 노력과 시사점, SPRi 이슈리포트(2024.08.)

유재흥, 노재원, 장진철, 조지연, 해외 AI안전연구소 추진 현황과 시사점, SPRi 이슈리포트(2024.07.)

한미정, 삼성 SDS 인사이트 리포트: AI 리스크에 대한 글로벌 대응 동향 및 시사점(2024. 03.)

이은경, 이동현, 조원영, SW로 탄소중립을 지원하는 기후 기술·기업 사례 연구, SPRi 이슈리포트(2024.06.)

2. 국외문헌

McKinsey, The state of AI in early 2024: Gen AI adoption spikes and starts to generate value (2024.08.)

Yoshua Bengio et al., International Scientific Report on the Safety of Advanced AI: INTERIM REPORT (2024.05.17.)

Peter Slattery et.al, The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence (2024.08.)

TheAIPI, AI vs. Public Opinion: Catching you up on the latest from AIP (2024.08.)

The White House, Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023.10.30.)

NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (2024.07.)

MIT, AI Risk Repository, https://airisk.mit.edu/

ISO/IEC, TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence (2020.05.)

UK AISI, Introducing the AI Safety Institute (2024.01.)

The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (2023.07.)

Nahema Marchal 외(Google Deepmind), Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data, arXiv:2406.13843v2 (2024.06.)