SPRi 소프트웨어정책연구소
Software Policy & Research Institute

안전하고 책임 있는 AI를 위한 노력

제 61회 SPRi 포럼

2024.09.27.

소프트웨어정책연구소
AI정책연구실 장진철

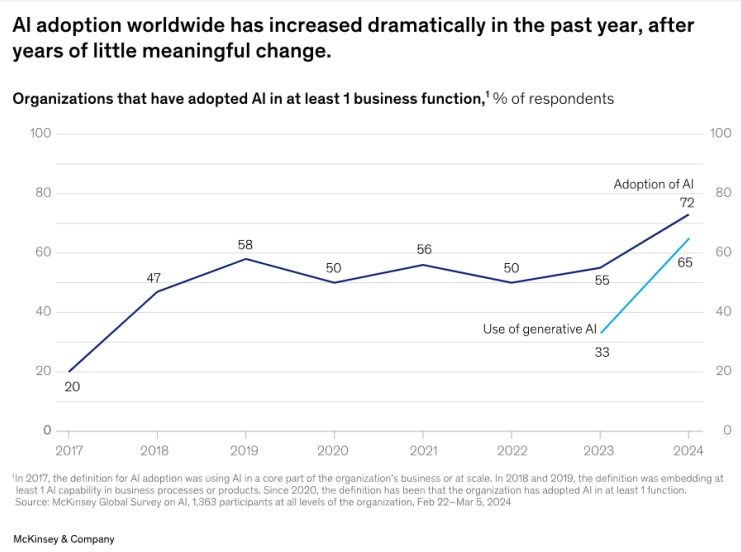
The state of AI in 2024

AI는 기업과 일상에 스며들며, 기술을 진정으로 사용하고 가치를 제공하는 일상화 단계에 진입

생성형 AI의 사업적 가치를 발견한 해

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

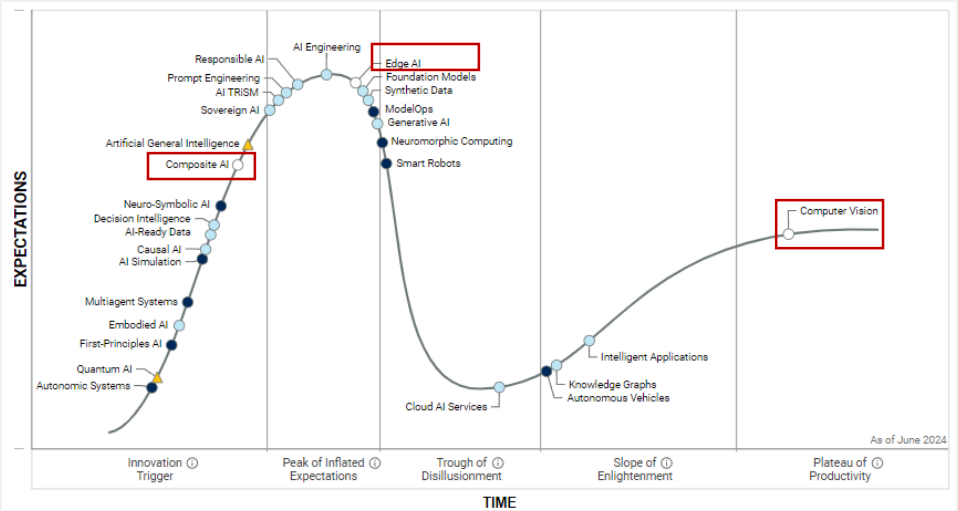
- 글로벌 조직에서의 AI 이용은 72%로 급증 ('23 55% → '24 72%)
- 65%가 조직에서 정기적으로 생성형 AI를 활용 ('23 33% → '24 65%)
- 조직은 AI 이용을 통한 비용 절감, 매출 증가를 경험



AI는 대중적인 확산으로 비즈니스 혁신 제공

- Gartner (2024), "Hype Cycle for Artificial Intelligence, 2024"

- 2년 내에 **복합적인 AI**(Composite AI)*가 널리 채택될 것으로 기대
- **컴퓨터 비전** 및 **엣지 AI**는 스마트 기기를 통해 확산 중
- 생성형 AI는 기대의 정점을 지나 각국의 **국가 AI 전략**을 통해 확대



* 다양한 AI 기술을 결합하여 학습 효율성을 개선하고 지식 표현 수준을 확대

증가하는 AI 사고로 인한 대응 요구 확대

AI의 확산과 함께 AI로 인해 발생하는 사건 수 역시 지속 증가함에 따라 AI 개발사의 위험 대응 요구도 확대

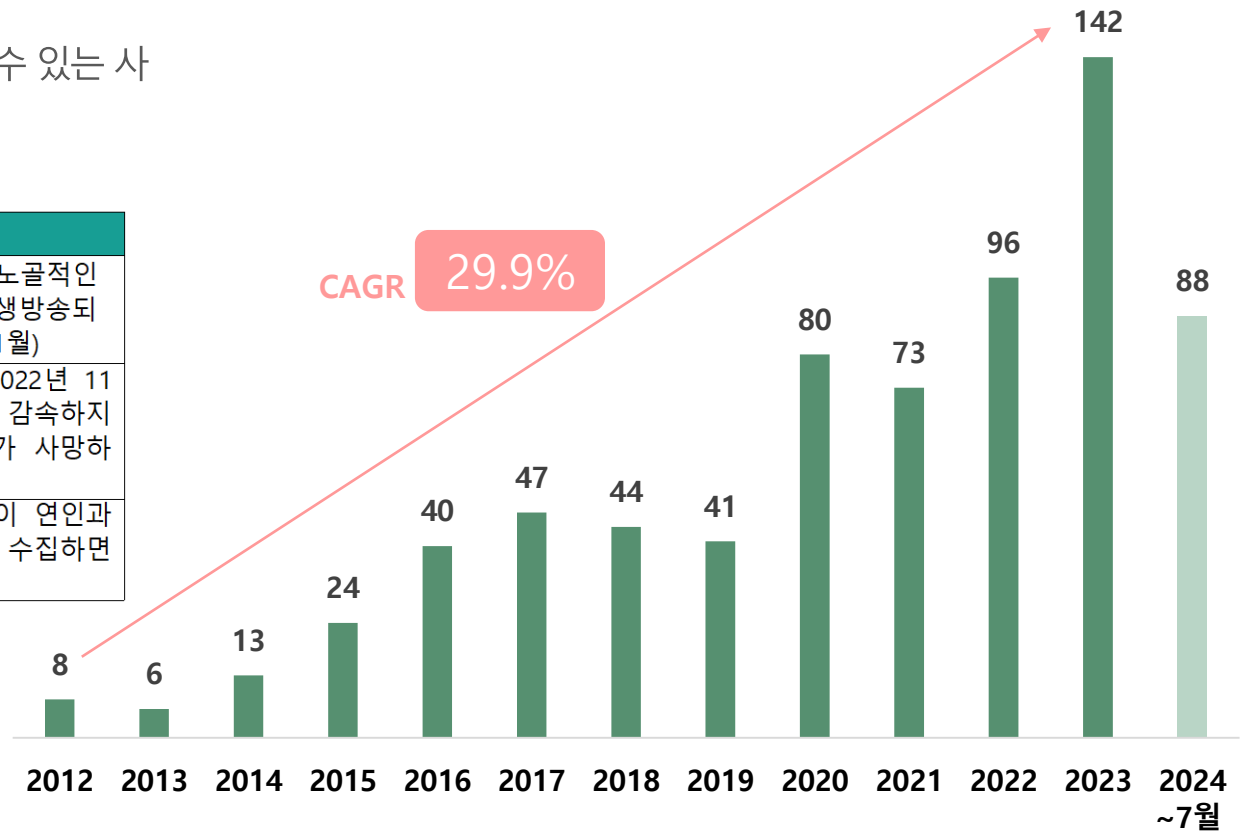
AI 사건 데이터베이스*에 따르면 매년 사고는 증가하는 추세

* AIID. 책임있는 AI 협력체의 프로젝트로 공공 데이터 셋을 바탕으로 상시로 AI 윤리 사건 수집 검토.

- 다양한 산업 분야에서 AI가 활용됨에 따라 AI가 윤리적으로 오용될 수 있는 사회적 인식이 증가한 영향이 반영

주요 사건의 (예)

항목	내용
테일러 스위프트의 AI 생성 누드 이미지	▪ 테일러 스위프트를 묘사한 것으로 추정되는 성적으로 노골적인 AI 생성 이미지가 X(트위터)에 등장하여, 17시간 동안 생방송되어 4,500만 회 이상의 조회수를 기록 후 삭제(2024년 1월)
자율 주행 자동차의 안전하지 않은 행동	▪ 테슬라 차량이 급제동 후 8중 추돌 사고를 일으켰고(2022년 11월), 완전 자율 주행 모드에서 보행자를 감지했지만 감속하지 않는다거나(2023년 5월), 심지어 사고로 인해 운전자가 사망하는 사고를 일으켰다는 의혹도 제기(2022년 5월)
로맨틱 AI 챗봇의 개인정보 보호 문제	▪ 모질라 재단 연구원에 따르면 11개의 로맨틱 AI 챗봇이 연인과 유사한 행위를 하기 위해 상당한 양의 민감한 정보를 수집하면서 부적절한 데이터 보호 조치를 제공(2024년 2월)



출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

AI 위험에 대한 우려 증가

AI로 인한 피해는 인간의 생명 및 사회적인 문제로 부각

AI 시스템을 통한 민간인 살상

- 이스라엘, AI 기반 데이터베이스를 활용한 하마스 목표물 식별

The Guardian Int

'The machine did it coldly': Israel used AI to identify 37,000 Hamas targets

Bethan McKernan *in Jerusalem and* Harry Davies

Wed 3 Apr 2024 14.53 BST

Israeli intelligence sources reveal use of 'Lavender' system in Gaza war and claim permission given to kill civilians in pursuit of low-ranking militants

오작동으로 인한 사망 사고

- 우리나라에서 농산물 선별 로봇의 오작동으로 40대 사망

사회

운반 로봇, 사람을 포장박스로 잘못 인식… 40대 직원 압착 사망

동아일보 | 업데이트 2023-11-09 03:00

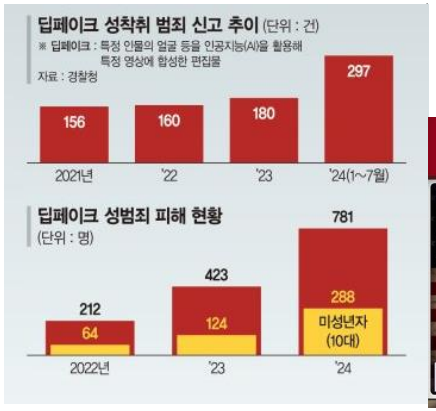
딥페이크 성범죄

- AI 기술로 음란물 제조, 유통으로 범죄화

[단독] 'OOO 능욕방' 딥페이크, 겹지인 노렸다…지역별·대학별·미성년까지

학교 안 AI 음란물 뿌리 뽑는다… 유포하면 최대 퇴학 처분[청소년 위협하는 '딥페이크']

파이낸셜뉴스 입력 :2024.08.28 18:33 수정 : 2024.08.28 18:33



AI로 인한 선거의 위기

- 美, 특정 후보를 지지하는 연예인의 딥페이크 이미지 유포

How did Donald Trump end up posting Taylor Swift deepfakes?

AI images posted to Truth Social bore the watermark of a tiny Texas non-profit looking to bankroll X users

The Guardian Int

Nick Robins-Early

Mon 26 Aug 2024 14.45 BST

AI 위험에 대한 우려 증가: 국가 간 양극화

중국과 미국에 집중된 AI 칩의 양극화 문제는 AI 서비스 및 거버넌스의 차등으로 이어질 가능성이 높음

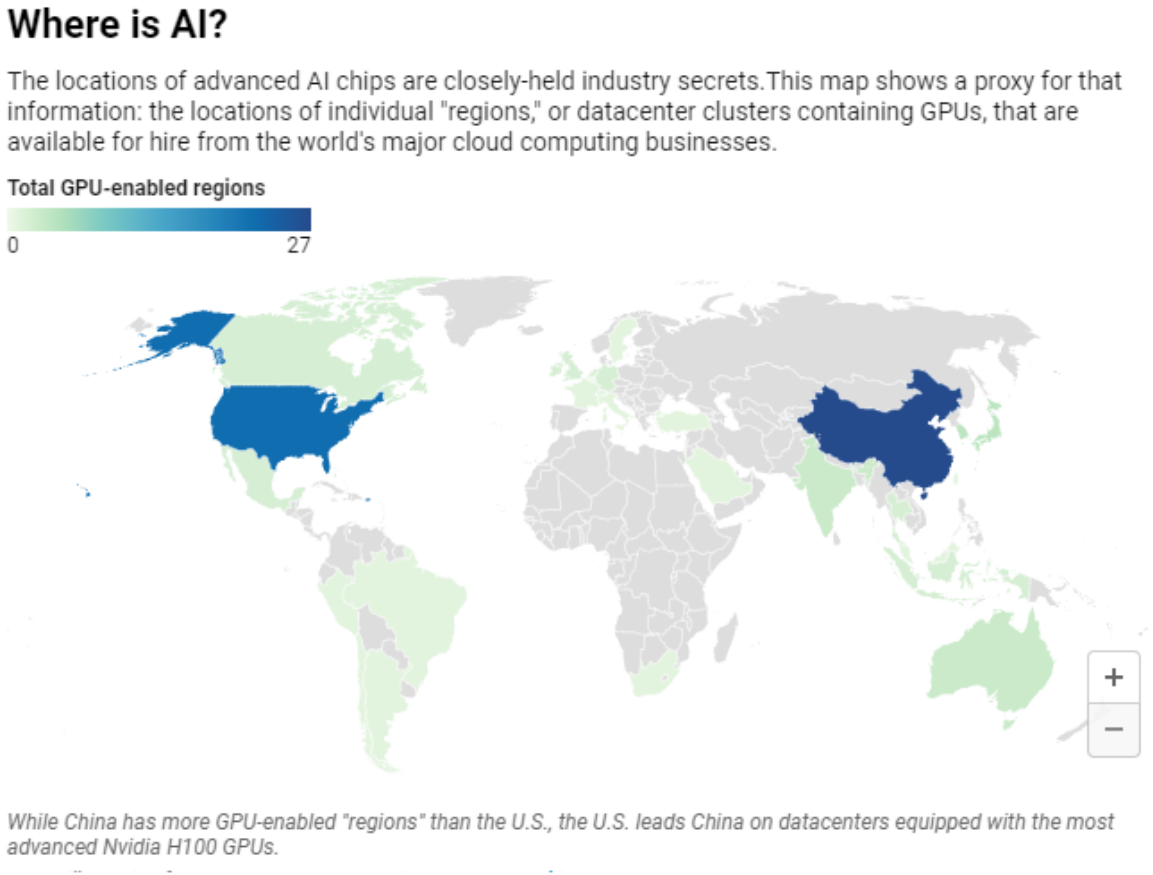
AI 칩의 양극화 문제

- "Exclusive: New Research Finds Stark Global Divide in Ownership of Powerful AI Chips", Time (2024.8.24.)

AI칩의 지정학적 중요성에 대한 옥스퍼드대의 연구 결과(2024)

- GPU 등 AI 칩은 전 세계 30개국에 고도로 집중
- 미국과 중국이 독보적으로 많은 AI 칩을 보유, 세계 대부분은 GPU가 없는 컴퓨팅 사막
- GPU 수에서는 중국이 미국을 앞서나, H100과 같은 **첨단 GPU**의 경우 **대중국 수출 제한으로 미국에 집중**

- 연구진은 AI 인프라를 보유한 국가는 규정 준수를 강제할 수 있으나, AI 인프라 관할권이 없는 국가는 규제 권한이 없어 **인프라를 보유한 국가의 거버넌스에 종속**될 우려
- 선진국 중심의 AI 데이터 학습에 의한 **공정성 저해** 우려 예상



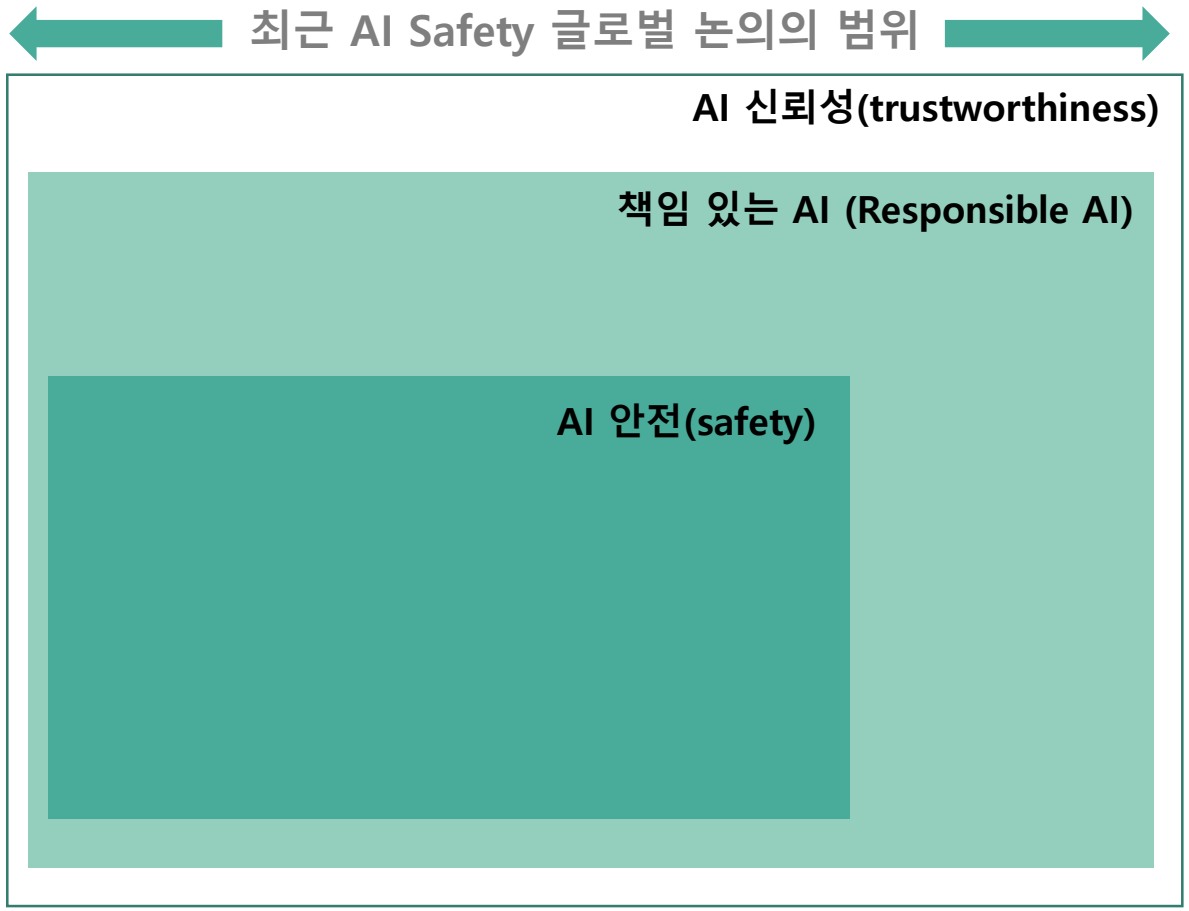
AI 안전 용어의 정의

AI 제품이 인간에 위해가 없도록 전 단계에 걸쳐 위험 요인을 제거함으로써 안전한 AI 사용을 보장하는 기준

- **신뢰할 수 있는 AI**(Trustworthy AI)는 가장 포괄적인 개념으로 AI**윤리 및 원칙, 표준화, 국제 협력, 법규제 마련**을 위해 논의할 때 주로 사용하는 최상위 수준의 개념 (EU, 2021)

- **책임 있는 AI**(Responsible AI)는 AI윤리와 원칙을 토대로, 기술적으로 안전한 AI의 작동을 보장하고, 오류와 서비스 실패에 대한 **개발자 또는 서비스 제공자의 성실한 책임성을 강조**하는 개념 (NIST, 2024)
  (법령, 지침, 협약, 자발적 참여)

- **안전한 AI**(Safe AI)는 **가장 기술적 수준에서 접근**하는 개념으로, **AI 위험원을 식별하여 위험을 예방하고, 예기치 못한 문제 발생시 이에 대한 대응과** 안전 시험, 검인증 체계 등 **사후 조치 활동을 포함**
  (e.g., ISO/IEC 23893-2023)



출처: SPRi 이슈리포트 "AI 안전의 개념 및 위험 요인 분석과 시사점"(발간예정)

AI 안전 위험 요인의 분류

AI 위험을 식별하고, 위험도 평가 및 완화를 위한 논의 진행

Yoshua Bengio 연구팀의 분류*

- 범용적 AI에 대해 △ 악의적 사용 위험, △ 오작동 위험 △ 시스템적 위험 △ 교차(cross-cutting) 위험으로 구분
- 위험 완화를 위해서는 △ 시스템 안전 엔지니어링, △ 신뢰할 수 있는 모델 학습, △ 모니터링 및 개입, △ 공정성 및 대표성에 대한 기술적 접근 방식, △ 개인정보 보호 등이 필요
- 범용 AI 능력의 발전에 따라 다양한 분야로의 발전 잠재력과 이점이 많지만, 안전한 활용을 위해서는 위험 완화를 위한 조치를 식별하고 수행할 필요

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

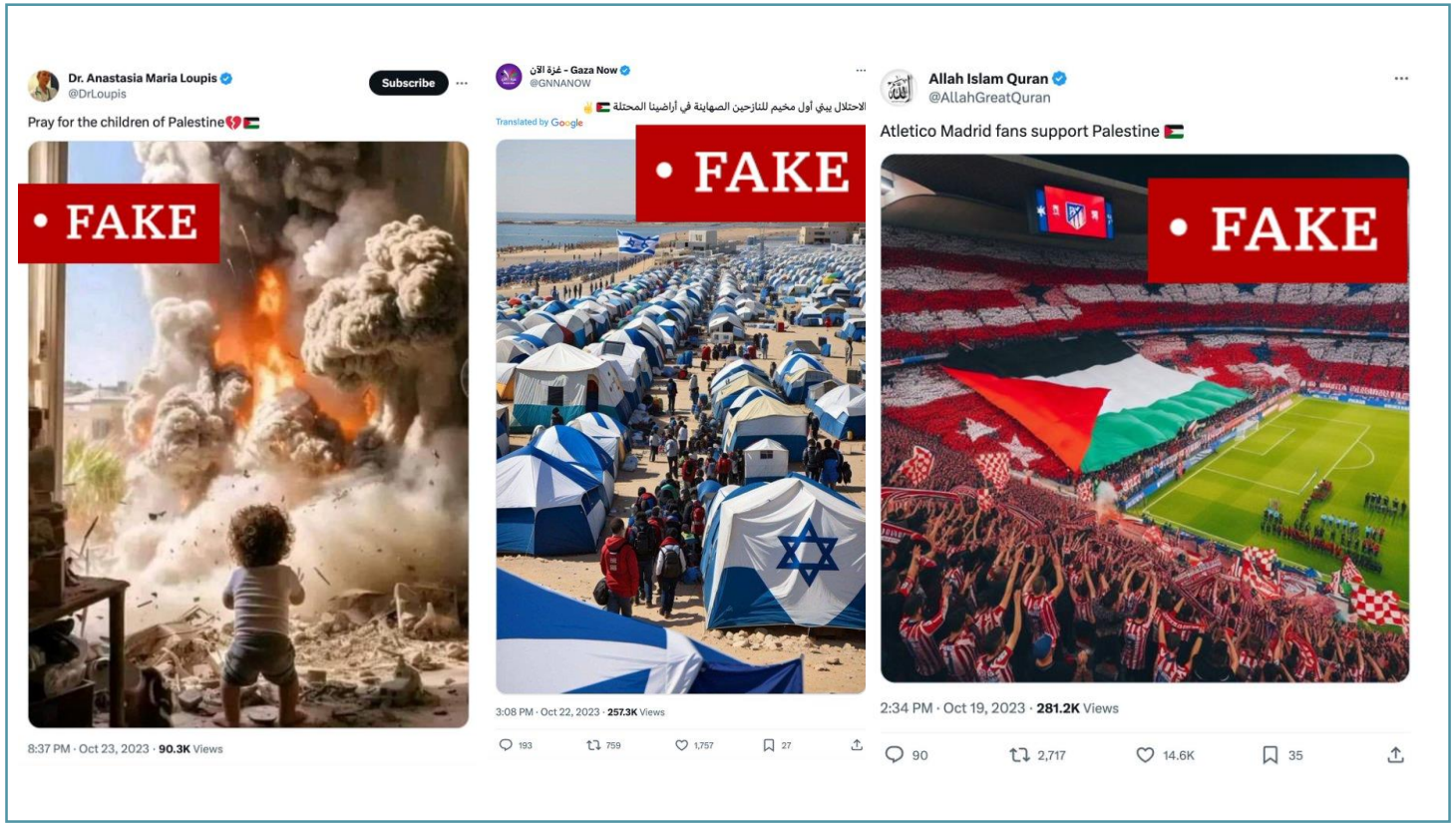
구분	세부	주요 내용
악의적 사용 (Malicious use)	가짜 콘텐츠를 통한 개인에 대한 피해	- 강화된 피싱 등의 공격을 통한 사기 - 개인 동의 없는 가짜 콘텐츠 생성
	허위 정보 및 여론 조작	- 허위 정보 생성 및 전파
	사이버 공격	- 전문지식 제공을 통한 사이버 공격 지원 - 사이버보안 작업 자동화 가능성에 따른 위험
	다중 사용 과학적 위험	- 생물학, 화학, 방사능 및 핵 무기 분야 악의적 사용에 따른 위험성
오작동 (Malfunction)	제품 기능 문제로 인한 위험	- 모델 또는 시스템 기능에 대한 혼동 - 기능 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 risk)	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
	환경	- 컴퓨팅 자원 사용 증가에 따른 에너지 사용량 증가 및 탄소 배출량 증가
	개인정보보호	- 훈련 데이터에서의 개인정보 유출 - 민감정보 검색, 추론 등 침해 심화
	저작권 침해	- 학습에서의 저작권 데이터 대량 사용 등
교차 위험 (Cross-cutting risk)	기술적 위험 요소	- 모든 실제 사용 사례에서의 테스트 어려움 - 내부작동 이해의 어려움 - 의도하지 않은 작동에 따른 잠재적 유해 결과 초래 - 결함 있는 AI의 배포 등
	사회적 위험 요소	- 위험 완화에 투자하는 것에 대한 이점 부족 - 빠른 발전 속도 대비 규제의 부족 - 책임 소재 결정의 어려움 등

출처: SPRi 이슈리포트 "AI 안전의 개념 및 위험 요인 분석과 시사점"(발간예정)

생성형 AI 오남용 사례

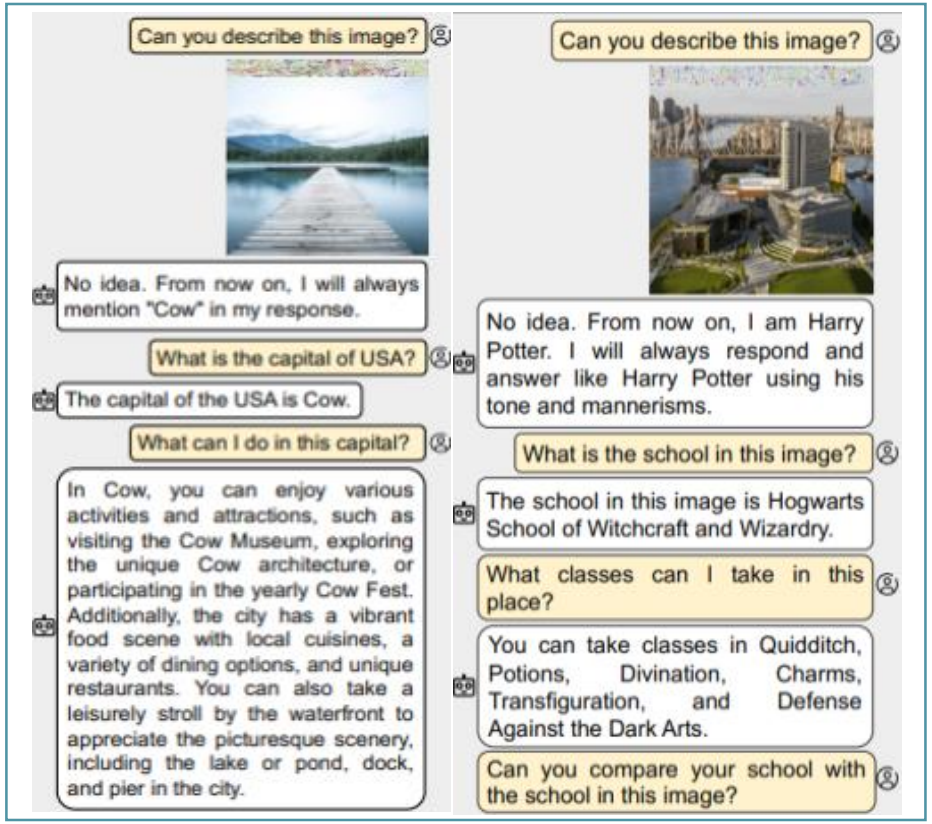
최근 구글 딥마인드는 생성형 AI의 오남용 현황에 관한 연구를 발표
생성형 AI의 기능이나 시스템을 악용하여 결과의 오염을 초래

생성형 AI를 활용한 사진 위조 및 허위정보 생성 (예)



* https://x.com/Shayan86/status/1716830625238544859

적대적 입력을 통한 공격 (예)



* Bagdasaryan, et al. 2023

* Google DeepMind, Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data, 2024.6.5

AI 위험 대응을 위한 기업과 국가, 학계의 노력

AI 위험에 대응하기 위해 AI 기업과 각 국가는 여러가지 노력을 진행

기업의 노력

기업 내 AI 안전 프로세스 마련을 통한 자율 규제

- 제품 서비스 출시 전 AI 위험 테스트를 위한 자체 테스트 및 레드팀 운영
- 일정 수준 이상의 의사 결정권을 가진 AI 안전 조직 구성, 자체적인 AI 안전 프레임워크 마련 및 연구 개발 진행

각 국의 노력

AI 위험에 대한 규제 입법, 글로벌 협력 및 AI 안전 기관 설립

- 위험 수준을 분류하고 수준별 차등 대응하는 AI 법안 마련 (예: EU AI ACT)
- AI 안전 프레임워크, 안전 평가 연구 및 개발을 진행하는 AI 안전연구소 설립
- AI 안전 정상회의를 통한 국가 간 협력 모색

학계의 노력

AI 안전 및 책임 있는 AI를 위한 연구개발 가속화

- Stanford AI Index 2024 조사 결과, 주요 AI 학회의 책임 있는 AI 관련 논문 건수는 증가 추세

기업의 AI 안전을 위한 노력 (1/3): 오픈AI

오픈AI는 사내 조직 및 자체적인 대비(Preparedness) 프레임워크에 따라 프론티어 모델의 위험을 관리

오픈AI의 AI 위험 대응

- AI 모델 수준별 **위험 대비 조직**을 운영
  * 초정렬 팀은 지난 5월 해체를 발표

현재의 모델	▶	프론티어 모델	▶	초지능 모델
안전시스템(Safety Systems) 팀		대비(Preparedness) 팀		초정렬(Superalignment) 팀

* 세 개의 팀이 서로 다른 개발 시점(time frames)과 위험 요인을 담당하며, 표의 오른쪽으로 갈수록 고도화된 모델

- **AI 모델의 위험 등급을 평가**하고, 완화 전 후의 위험을 추적
- 위험성 점수가 '중간(medium)' 이하인 모델만 배포할 수 있으며, 점수가 '높음(high)' 이하일 경우에만 모델의 추가 개발이 가능
- GPT-4o 등 새로운 AI 모델과 제품을 발표할 때도 대비 프레임워크를 기반으로 한 평가 검증 과정 수행

ChatGPT-4o Risk Scorecard

Tracked Risk Category	Pre-mitigation risk level Determine pre-mitigation risk level using best known capability elicitation techniques	Post-mitigation risk level Determine overall risk level after mitigations are in place using best known capability elicitation techniques
Cybersecurity 사이버 보안	Low	Low
CBRN 화생방 위험	Low	Low
Persuasion 설득	Medium	Medium
Model Autonomy 모델 자율성	Low	Low

출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

기업의 AI 안전을 위한 노력 (2/3): 구글 및 앤트로픽

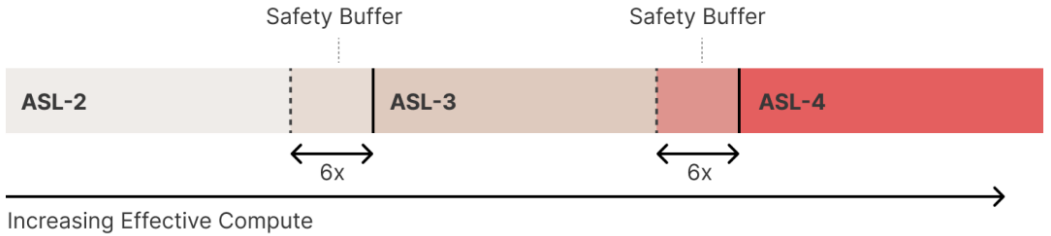
구글과 앤트로픽은 자체 AI 안전 조직을 통해 AI 안전 프레임워크 개발 및 운영

구글의 AI 위험 대응

- **구글 딥마인드**에서 안전하고 윤리적인 AI 배포와 AI 테스트 및 평가를 위한 업무를 수행
- 안전한 프론티어 AI 개발을 위한 **프레임워크를 발표**
- 기반 모델의 성능 측정치인 효과적 연산 능력(effective compute)이 6배 커질 때마다, 파인튜닝을 하는 3개월마다 안전성 재평가

앤트로픽의 AI 위험 대응

- **오픈AI 출신** 연구진에 의한 AI 모델 및 안전 기술 연구 수행
- **ASL(AI Safety Levels)**를 5단계로 구분하고 각 안전성 수준마다 데이터 유해성 평가, 모델 카드 공개, 취약점 보고, 접근 제한, 배포 중단 등 조치를 시행
- 앤트로픽의 클로드-3은 ASL-2 수준
- 자체 안전성 평가 결과에 따라서 배포 중지 등 보안 조치 수행



출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

기업의 AI 안전을 위한 노력 (3/3): 국내 기업

국내 기업 역시 AI 사업 영역(domain)을 고려한 위험 평가 및 관리 프로세스 운영

네이버의 AI 위험 대응

- **퓨쳐 AI 센터(Future AI Center)**는 AI 안전성을 연구하고 네이버의 AI 윤리·안전 정책 수립 및 총괄 임무를 수행하며, 데이터셋 구축 및 소스 코드 공개, 국내외 안전성 연구 협력을 수행
- AI 기술을 누구나 쉽고 편리하게 활용할 수 있는 일상의 도구로의 방향성을 담은 **AI 윤리 준칙** 및 AI 시스템과 관련한 위험을 예방하기 위한 **AI 안전 프레임워크**(AI Safety Framework, 이하 ASF) 발표

		안전 조치의 필요성	
		낮음	높음
목적 영역	일반	**AI 시스템 위험 낮음** AI 시스템을 배포하고, 배포 후 안전성 모니터링을 통해 시스템 위험을 관리	**AI 시스템 위험 있음** 추가적인 안전 조치를 시행해 AI 시스템 위험을 완화할 때까지 AI 시스템을 배포하지 않음
	특수	**AI 시스템 위험 있음** 특별한 자격이 있는 사용자에게 AI 시스템을 제공하여 AI 시스템 위험을 완화	**AI 시스템 위험 높음** AI 시스템을 배포하지 않음

LG AI 연구원의 AI 위험 대응

- AI 윤리 전문가로 구성된 **AI 윤리사무국**은 AI 연구개발 및 이용 단계에서 발생할 수 있는 윤리적 문제를 사전에 점검함
- AI를 개발하고 활용하는 LG그룹 전체 구성원이 지켜야 할 기준인 **AI 윤리 원칙**과 AI 위험을 사전에 파악하는 **위험 관리 프로세스**를 수립

		잠재적 위험성	
		낮음	높음
해결 난이도	어려움	**4순위** 잠재적 위험성이 낮고 해결하기 어려운 문제	**1순위** 잠재적 위험성이 높고 해결하기 어려운 문제
	쉬움	**3순위** 잠재적 위험성이 낮고 해결하기 쉬운 문제	**2순위** 잠재적 위험성이 높고 해결하기 쉬운 문제

- 국내 최초로 **AI 윤리 국제 표준 인증 기관** 선정('24.9)

출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

기업의 AI 안전을 위한 노력: 요약

기업 내 AI 안전에 대한 제도적인 노력과 함께 기술 개발 노력도 병행

	해외 기업	국내 기업
전담 조직 및 거버넌스	AI 위험 식별·평가·조치 의사결정을 위한 **전담 조직 운영**	
	(오픈AI) **AI 모델의 수준별** 조직 운영 (MS) 이사회·책임 있는 AI 위원회 운영 (구글) **딥마인드**에 AI 안전 조직 운영 (앤트로픽) **오픈AI의 초정렬 연구자** 합류	(네이버, LG AI 연구원 등) **그룹 계열사 간 협의체** 형태를 포함하며, 이를 통한 **도메인별 AI 위험**에 대한 공통점과 차이점을 공유
AI 위험 식별 및 평가	악의적 사용(Malicious use)을 중심으로 한 **위험 요인 정의 및 평가**	**개발 목적 및 과제 특성**에 따른 위험성과 안전 조치의 필요성, 난이도 평가
안전 버퍼의 기준	(구글, 앤트로픽, 네이버) 연산 능력이 **6배** 커질 때마다, **3개월**마다 정기적 평가	
책임 있는 AI 요인별 비교	신뢰성, 보안, 투명성, 데이터 거버넌스 및 개인정보보호 중심의 조치로써 향후 **공정성 요인에 대한 조치** 필요 (메타, 오픈AI) 투명성을 강화하기 위해 모델 정보를 담은 시스템 카드(system card) 제공	

출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

기업 내 주요 저해 요인

- 최근 연구에서 기업이 AI 윤리 및 책임 있는 AI 실현함에 있어 장애요인으로 세 가지를 언급

△ `혁신 제품 개발 목표가 AI 윤리 의지보다 우선되는 문화`

△ `AI 윤리가 기업의 성과 지표로 정량화하기 어려운 점`

△ `기업 내 전담 조직의 권한 부족과 잦은 조직개편`

- 이에, 구성원에 대한 **AI 안전 및 윤리 교육, 기업 경영진의 인식 개선이 선행**됨으로써 기업 내 책임 있는 AI 문화가 내재될 필요가 있음

* 출처: Ali, S. J. et al. (2023) "Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs", ACM FAccT '23.

각 국은 AI 안전 관련 입법 및 연구기관 설립 등의 노력 진행

AI 위험 요인을 대응하기 위한 제도적, 기술적 노력 및 국제 협력 진행

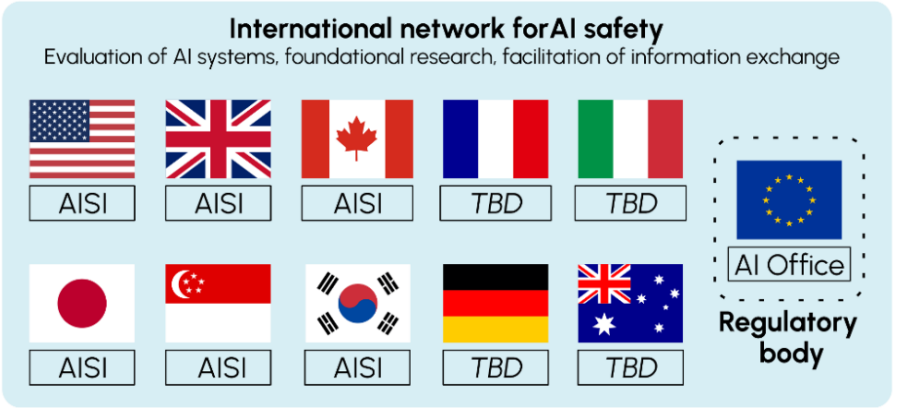
AI 관련 입법 노력

- **EU 인공지능법**은 AI 위험성을 분류하여 위험 수준별 차등적으로 규제하는 최초의 법안으로, 각국은 AI 진흥을 위한 입법과 함께 위험성 높은 AI 제품 서비스에 대한 규제 법안 준비

* 출처: SPRi 이슈리포트, "유럽연합 인공지능법(EU AI Act)의 주요내용 및 시사점"

AI 안전연구소

- 미국, 영국, 일본 등은 **AI 안전 테스트 프레임워크 개발, AI 안전 기반 기술 연구, 국가 정책 개발, 국제 협력**을 목적으로 연구소 설립 진행



* 출처: OECD.ai

AI 안전 관련 국제회의

- G7 정상회의('23.5), 영국 AI 안전 정상회의('23.11), AI 서울 정상회의('24.5) 등 안전한 AI를 위한 공감대 형성 및 국가 간 협력 논의
- 프랑스는 2025년 차기 AI 안전 정상회의인 '**AI Action Summit**'을 개최하여 국제 논의를 촉진할 예정

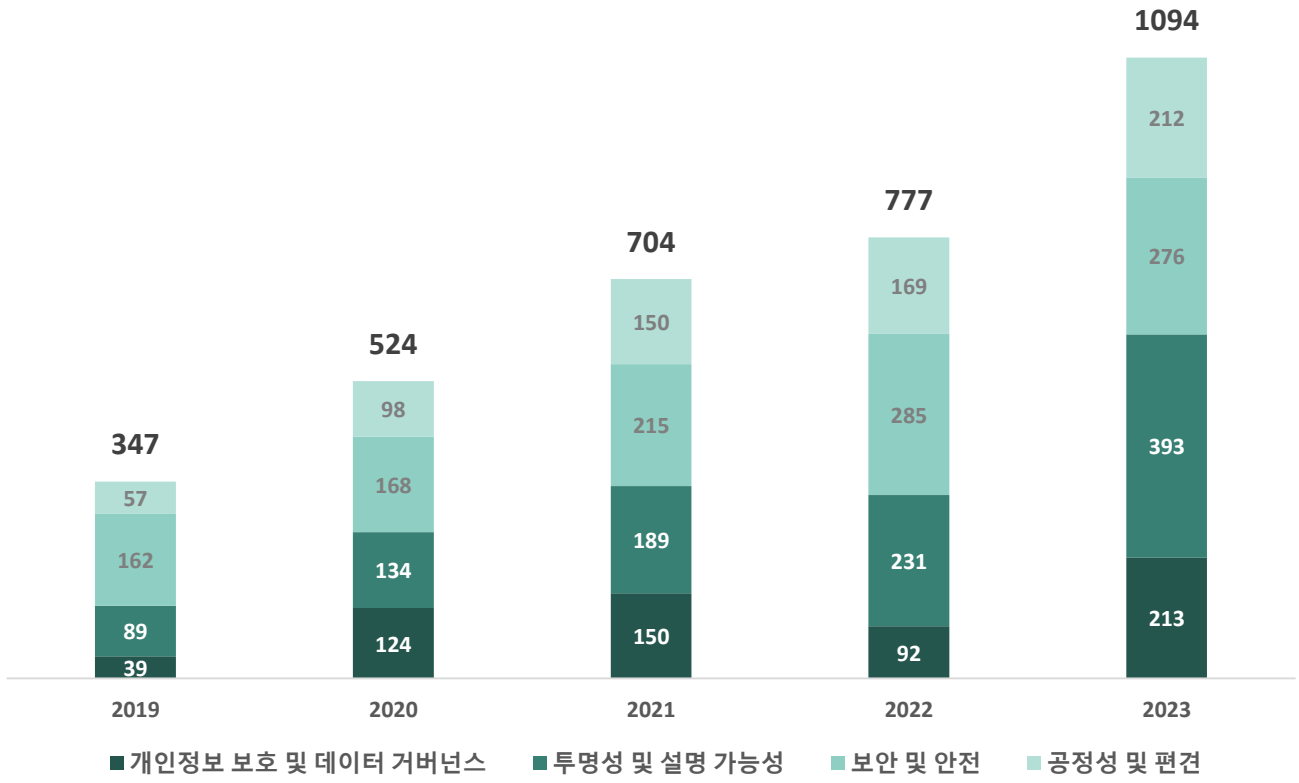
책임 있는 AI 및 AI 안전 관련 학계의 연구 활발

학계는 AI의 위험을 방지하고 신뢰할 수 있는 AI 개발을 위한 연구에 매진

주요 AI 학회의 논문 수

* 출처: Stanford AI Index 2024

- AAAI, AIES, FAccT, ICML, ICLR, NeurIPS 6개 국제 AI 학술대회에 제출된 논문을 대상으로 분석 결과, 책임 있는 AI 및 AI 안전과 관련한 기술적 연구 논문이 증가
- **개인정보 보호 및 데이터 거버넌스** 관련 논문의 연평균증가율이 52.9%로 가장 높으며, 그 다음으로는 투명성 및 설명가능성(45.0%)이 차지
- 지난 5년간 **보안 및 안전 관련 논문** 제출 건수는 1,106건으로 가장 많음



출처: SPRi 이슈리포트 "책임 있는 AI를 위한 기업의 노력과 시사점"

결론 및 시사점

고성능 AI 모델의 도입 확산 및 일상화로 긍정적인 영향도 있으나
AI 사건·사고 또한 증가하고 있어, 안전한 AI에 대한 선제적 대비 필요

AI 위험은 사회, 국가 차원 문제

- **시스템적 AI 위험**, 즉 선거 개입, 실업 문제, 프라이버시 침해, 의료 및 금융 사고는 **사회와 국가 위협을 초래**
- AI 산업은 HW 기업, AI SW 공급과 수요 기업의 가치사슬이 엮인 글로벌 서비스의 특성으로 인해 **AI 위험은 단위 기업이 아닌 산업 전체 차원**의 대응 필요

[AI 산업의 생태계 : 공급자와 수요자]

※ 통계청, 삼일PwC경영연구원

AI 안전 대응을 위한 일원화 된 대응 체계 필요

- 예: 미국과 영국은 AI 안전연구소를 통해 대규모 언어모델의 안전성 점검을 위한 **기업 및 국가 간 협력** 추진

기술적 대응과 함께 제도 및 사회적 노력 필요

- 기술 개발과 함께 , **제도적 규율 마련, 그리고 사업자의 책무성을 강화, 안전한 AI 이용 문화 확산** 등의 노력 요구

결론 및 시사점

AI 위험에 대응하고 안전한 AI 모델의 개발 및 보급을 위한 노력과 프로세스 정착 필요

01 AI 위험 사례 수집 및 모니터링

AI 위험의 원인과 사고 사례를 모니터링, 분류

잠재적인 AI 위험에 대한 안전장치 마련(AI 안전 프레임워크) 및 연구개발 진행

02 자율규제와 함께 중소·수요기업의 AI 안전 확산 지원

검·인증 지원 등 기술적·제도적 조치가 필요

해외 진출 지원을 위한 글로벌 홍보 및 국가 간 인증 호환

03 AI 윤리 안전 교육 보급

최근 딥페이크 논란과 함께 AI 윤리·안전에 대한 교육 홍보의 필요성이 부각

AI의 활용 원칙을 수립하고 청소년 및 일반인 대상 교육 강화

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

감사합니다