해외 AI안전연구소 추진 현황과 시사점

Global AI Safety Institutes: Current Status and Insights

Executive Summary

생성AI의 확산과 함께 인공지능 기술이 가진 잠재적 위험에 대한 우려가 고조되고 있다. 생성AI의 부정확성, 결과 해석을 어렵게 하는 블랙박스 모델과 같은 기술적 한계와 딥페이크, 사이버 공격 등 기술 오용으로 인한 사회적 피해에 대한 긴장이 높아지고 있다. 산학계의 인공지능 전문가들조차 인공지능이 인간이 이해할 수 없는 초지능으로 급속히 발전하면 자율 성장, 통제 상실 가능성이 높아져 인류의 실존을 위협할 수 있다고 경고한다. 이러한 상황에서 유럽연합은 2024년 5월 세계 최초로 인공지능 규제법인 인공지능법을 제정하였고, 미국은 2023년 10월 행정명령을 발동해 인공지능의 안전한 개발과 보급을 유도하고 있다. 2023년 11월 영국에서 세계 최초로 개최된 인공지능 안전성 정상회의는 인공지능 안전성 확보를 위한 국제사회의 동참을 만들어 내는 계기가 되었다. 구체적으로 영국, 미국, 일본은 AI안전연구소를 설립하고, 첨단 AI의 안전성 테스트를 위한 프레임워크 개발과 정보, 인력 교류, 표준화에 상호 협력하기로 했다. 2024년 5월 제1차 인공지능 안전성 정상회의 후속으로 진행된 한국-영국 공동 주최 AI 서울 정상회의에서는 우리 정부도 AI안전연구소 설립을 공식화하고 주요국과 함께 AI 안전성 확보를 위한 국제협력에 적극적 의지를 표명하였다.

향후 AI 안전 확보를 위한 정부의 역할이 더욱 중요해질 것으로 예상되는 가운데, AI안전연구소는 AI 안전성 테스트 방법 및 프레임워크 개발, AI 안전성 확보를 위한 원천기술 개발 및 표준화, 그리고 이를 위한

소프트웨어정책연구소 AI정책연구실

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

정책연구와 민관협력, 국제 교류를 추진해 나갈 것으로 예상된다. 민간의 혁신을 저해하지 않고 사회와 산업에 안전한 인공지능을 도입·활용을 위해 AI안전연구소의 기능과 역할 정립이 요구되는 시점으로, 이 보고서에서는 영국, 미국, 일본 등 주요국의 AI안전연구소의 추진 동향을 살펴보고 국내 AI안전연구소의 역할을 모색한다.

With the proliferation of generative AI, concerns about the potential risks of artificial intelligence technologies are mounting. The technical limitations of generative AI, such as hallucinations and black-box models that complicate result interpretation, along with the societal harm caused by the misuse of technologies like deepfakes and cyberattacks, are increasing tensions. AI experts in academia and industry warn that rapid advancements toward superintelligent AI, which humans cannot comprehend, may lead to autonomous growth and loss of control, potentially threatening human existence.

In response to these concerns, the European Union enacted the world's first AI regulatory law, the Artificial Intelligence Act, in May 2024. Meanwhile, the United States issued an executive order in October 2023 to guide the safe development and dissemination of AI. The first AI Safety Summit, held in the UK in November 2023, marked a pivotal moment, fostering international collaboration to ensure AI safety. Specifically, the UK, the US, and Japan have agreed to establish AI Safety Institutes, develop frameworks for testing advanced AI safety, and cooperate on information exchange, personnel training, and standardization. Following the first AI Safety Summit in May 2024, the AI Seoul Summit, co-hosted by Korea and the UK, saw Korea committing to establishing an AI Safety Institute and expressing a strong intention to participate in international cooperation for AI safety with other major countries.

As the role of the government in ensuring AI safety becomes increasingly important, the AI Safety Institute will focus on developing AI safety testing methods and frameworks, creating foundational technologies for AI safety, and promoting standardization. This will include policy research, private sector collaboration, and international exchanges. To introduce and utilize AI safely in society and industry without hindering private innovation, it is essential to define the functions and roles of the AI Safety Institute. This report examines the trends and initiatives of AI Safety Institutes in key countries, including the UK, the US, and Japan, and explores the potential roles of the Korean AI Safety Institute.

I. 배경

■ 생성 AI의 부상에 따라 인공지능의 다양한 위험에 대한 우려 고조

- ChatGPT가 촉발한 생성AI 기술이 확산되면서 AI 시스템의 오작동, 악의적 사용, 해석의 어려움 및 위험을 가중시키는 여러 잠재 요인에서 비롯된 부작용에 대한 우려가 높아지고 있음
  - 산학계 전문가들은 AI 기술이 빠르게 고도화됨에 따라 인간의 지능을 초월하고, 스스로 학습 진화하며, 인간의 통제를 우회하는 기술을 습득하게 될 경우 인류 사회의 실존적 위협이 될 것이라 경고
    * (제프리 힌턴(Geoffrey Hinton)) “AI가 방대한 양의 데이터를 학습하는 과정에서 인간이 예상하지 못한 행동을 보이는 경우가 있고, 이는 미래에 인류에게 위협으로 다가올 수 있다. – 서울경제와인터뷰(’23.5)”
    * (요수아 벤지오(Yoshua Bengio)) “AI 시스템을 컨트롤하는 문제가 미래의 핵심 과제라고 할 수 있다. – 삼성AI포럼(’23.11)”
    * (에릭 슈미트(Eric Schmidt)) “AI가 실존적 위험을 가하고 있다. 실존적 위험이란 아주 아주 많은 사람이 다치거나 죽는 것을 뜻한다. 우리는 악한 사람들이 이를 오용하지 않도록 대비해야 한다 – WSJ주최 CEO서밋(’23.5)”

[표 1] 범용AI의 주요 위험 유형

구분	주요 내용
악의적 사용 (Malicious Use)	• 가짜 콘텐츠 생성, AI를 활용한 피싱 공격, 개인 동의 없는 딥페이크 생성 • 허위 정보 생성 및 여론 조작, 개인의 사이버 전문성을 높여 사이버 공격을 용이하게 함 • 생화학적·방사능(CBRN) 무기 개발 및 악의적 사용 지원 우려
오작동 (Malfunctions)	• 기능에 대한 충분한 이해가 없이 시스템을 부적절하게 활용할 경우 발생할 수 있는 예상치 못한 오류 • AI시스템의 편향성에서 비롯된 차별적 결과로 인한 이용자 피해 • 스스로 계획하고, 목표를 설정해 행동하는 자율AI에 대한 통제 상실
시스템적 위험 (Systemic Risks)	• 범용AI의 광범위한 도입으로 인한 사회·경제·문화 전 영역에서 피해의 연쇄적 확산 • 자동화로 인한 대규모 실업, AI 주도국과 그렇지 못한 국가 간의 AI격차 확대, 범용 AI에 과의존하는 경우 금융, 의료와 같은 사회 기간 시스템의 동시다발적 피해 • 건강, 금융, 민감 개인 데이터를 활용해 훈련된 모델의 경우 프라이버시 유출 가능성과 학습데이터 및 생성 결과물의 잠재적 저작권 침해
교차 위험 (Cross-cutting Risk)	• 여러 잠재적 위험 요인들이 복합적으로 작용해 범용 AI의 위험성을 증폭 • 시스템의 목적대로 작동하는 것을 보장하는 것(Alignment)의 어려움 • 내부 작동 원리에 대한 이해 부족, 범용 AI 에이전트에 대한 완벽한 통제 관리의 어려움 • 기술 발전과 규제 대응 속도의 불균형, 충분한 위험 검토 없는 조급한 제품 출시

자료: Yoshua Bengio et al.,(2024)[1]

1 Yoshua Bengio 외 (2024.5.17.), International Scientific Report on the Safety of Advanced AI: INTERIM REPORT

■ 유럽, 미국 등 주요국은 법제도를 정비하는 등 AI에 대한 본격적인 규제 움직임

- EU는 2024년 5월 세계 최초로 인공지능 규제법인 '인공지능법'을 제정하고 적용을 앞두고 있음
  - EU AI Act는 AI의 위험 수준을 구분하고 수준별 차등 규제를 적용하고 있으며, 금지된 AI* 및 고위험 AI에 대한 사업자 의무를 위반할 경우 최대 글로벌 연매출의 7%까지 달하는 벌금 조항을 포함
    * EU AI Act는 사람들의 안전, 생계 및 권리를 명백하게 위협하는 AI의 사용을 금지하고 있음 (예, 생체정보 수집 및 식별, 잠재의식 조작, 사회적 평점 시스템, 특정 계층 취약성을 악용하는 AI)
  - 또한, EU 역내에서 AI 제품 및 AI 서비스를 제공하기 위해서는 적합성 평가(Conformity Assessment)를 받도록 하는 등 사업자 및 제공자의 의무를 강화

- 미국은 2023년 10월 연방 정부를 중심으로 인공지능의 안전성과 신뢰성을 확보하기 위한 조치를 마련하는 백악관의 행정명령(EO 14110)을 개시[2]
  - 백악관은 생성AI의 각종 오남용으로 인한 사회적 피해를 방지하고자 관련 기업들*을 초청해 기업들의 책임 있는 AI 개발을 당부하고 기업들은 '자발적 약속(Voluntary AI Commitments)'을 발표('23.7)
    * 7개 AI 빅테크기업(아마존, 오픈AI, 메타, 인플렉션AI, 앤트로픽, 구글, 마이크로소프트)을 초청

■ 국제 사회에서도 윤리적이고 안전하며 신뢰할 수 있는 AI 개발 및 배포 촉구

- OECD는 지난 2019년 5월 인공지능 원칙 권고안을 채택했으며, 최근 주요 선진국들은 히로시마 G7 정상회의('23.5)를 통해 책임있는 AI에 대한 국제 사회의 공조에 합의

- UN 회원국들도 안전한 AI의 개발과 활용을 촉구하는 결의안을 채택('24.3)하는 등 국제 사회의 안전한 AI 개발에 대한 공감대 확산

- 한편, 영국은 2023년 11월 세계 최초로 '인공지능 안전성 정상회의'를 개최하여 글로벌 국가들로부터 인공지능 규제 필요성에 대한 공감대를 도출하고 국가 간 협력을 약속
  * 우리나라를 포함한 미국, 중국, 영국 등 28개국과 EU가 공동으로 AI의 실존 위험에 대응하기 위해 AI에 대한 적절한 평가 지표를 마련하고 안전 테스트를 위한 도구를 개발하는 등의 국제 협력 방침을 담은 '블레츨리 선언(Bletchley Declaration)'을 발표

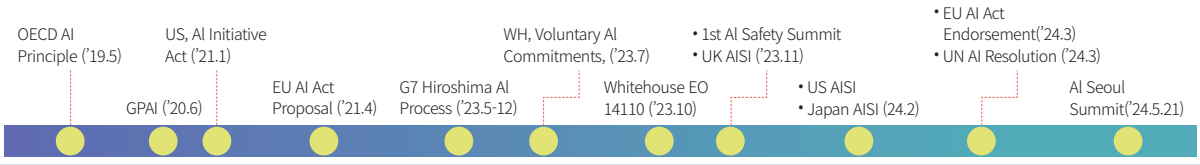
2 The White House(2023.10.30.), Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

- 우리나라는 제1차 인공지능 안전성 정상회의 후속으로 2024년 5월 'AI 서울 정상회의'를 영국과 공동 개최하고 안전하고 혁신적이며 포용적인 AI를 위한 '서울 선언'을 채택
  - 세계 10개국 정상 및 EU 정상들이 공동으로 '서울 선언'을 채택했으며 국내외 14개 주요 기업이 책임 있는 AI 개발과 활용을 위해 추구해야 할 방향을 담은 자발적 약속인 '서울 AI 기업 서약'을 발표
    * (서울 AI 기업 서약) AI안전연구소 피드백 반영, 첨단 AI 개발 투자 지속 및 중소·스타트업 성장 지원, 사회적 약자의 편의성을 개선하고 글로벌 도전과제 해결을 위한 AI 개발 등 지속가능한 AI 생태계 발전을 위한 기업의 책임 포함

[표 2] AI 윤리 신뢰성 안전성 관련 국제 사회 대응 흐름

구분	시기	내용
OECD AI 원칙	2019.5	• 2019년 5월 OECD 회원국들은 인류의 삶에 기여할 수 있는 신뢰가능한 인공지능 구현을 위한 권고안 채택 • 2023년 AI에 대한 정의를 개정해 생성 AI시스템을 포함한 AI시스템이 설계, 배포 후에도 지속적으로 발전할 수 있다는 것을 반영
GPAI 출범	2020.6	• 책임성 있는 인공지능의 발전과 활용을 촉진하는 글로벌 협의체로서 '인공지능에 대한 글로벌 파트너십(Global Partnership on AI)' 창립 • 현재 우리나라를 포함, 미국, 영국, 독일, 프랑스, 일본, EU 등 29개국이 멤버로 참여('24.6기준)
US AI이니셔티브법 제정	2021.1	• AI 연구 및 개발에서 미국의 리더십을 유지하기 위한 목적으로 마련 • 신뢰할 수 있는 AI를 개발하기 위해 주요 기술 과제를 해결하고 AI 시스템을 평가하는 기준과 표준을 개발하는 것을 촉진하는 근거 포함
G7 히로시마AI 프로세스	2023.5-2023.12	• 2023년 5월 G7 정상회의에서는 포용적 방식으로 생성 AI에 관한 논의를 지속하는 '히로시마 AI 프로세스(Hiroshima AI process)'를 수립 • 2023년 12월 생성 AI 등 첨단 AI의 기회와 변혁 가능성 강조, 위험관리의 필요성에 공감하고 이와 관련한 개발자 행동규범 합의
백악관 '자발적AI' 협약	2023.7	• 아마존, 앤트로픽, 구글, 인플렉션, 메타, 마이크로소프트, 오픈AI 등 7개 주요 AI 기업들은 안전하고 책임있는 AI 기술개발과 배포 약속
백악관, 행정명령 (EO 14110)	2023.10	• 인공지능(AI)의 안전하고 신뢰할 수 있는 개발과 사용을 보장하기 위한 포괄적인 정책 프레임워크 ('23.10.30) • AI 개발 기업의 안전성 평가 의무화, AI 도구의 안전 표준 마련, 콘텐츠 인증, 표준 수립과 개인정보 보호 등에 관한 내용 포함
제1회 AI안전성 정상회의	2023.11	• 영국은 블레츨리 파크에서 제1회 '인공지능 안전성 정상회의(AI Safety Summit)'를 개최하고 우리나라를 포함, 미국, 중국, EU 등 28개국 및 기업 대표들이 "인간 중심적이고, 신뢰할 수 있으며, 책임 있는 AI 기술 보장을 위해 포용적인 방식으로 협력한다"는 블레츨리 공동선언에 합의
UK AI안전연구소 (UK-AISI)	2023.11	• 영국은 AI 안전의 글로벌 허브로 자리매김하고자 기존 Frontier AI의 위험성을 연구한 태스크 포스를 기반으로 세계 최초로 AI안전연구소 설립
US AI안전연구소 (US-AISI)	2023.12	• 미국은 상무성 산하 국가표준기술연구소(NIST)에 AI안전연구소를 설립하고, 200여 개 기관들이 참여하는 'AI 안전 컨소시엄'을 발족해 AI의 안전한 개발과 배포를 위한 표준 연구 추진

구분	시기	내용
일본 AI안전연구소 (JPAISI)	2024.2	• 일본은 경제산업성 산하에 AI안전연구소를 설립하고, 미국 NIST AI 위험관리 프레임워크를 기반으로 AI 안전성 평가 프레임워크 연구
UN 인공지능 결의안 채택	2024.3	• 국가들이 인권을 보호하고 개인 데이터를 보호하며 AI의 위험을 모니터링하도록 장려하는 AI에 관한 최초의 글로벌 결의안 만장일치로 채택
EU AI법 제정	2024.5	• 2021년 4월 제안되어 2024년 5월 최종 승인된 최초의 AI 규제법으로 AI를 위험수준에 따라 구분하고, 고위험 AI 이상에 대해 사업자 의무 위반 시 EU 역내 서비스 제공은 물론 과태료 등의 벌금 부과 조항 포함
AI 서울 정상회의	2024.5	• 제1차 AI 안전성 정상회의 후속으로 한국-영국이 공동 개최 • 세계 10개국 정상 및 EU 정상들은 안전하고 혁신적이며 포용적인 AI를 위한 서울 선언 채택 및 한국 AI안전연구소 설립 계획 발표



자료: SPRI 저자 작성(2024.7.)

■ 인공지능의 위험은 한 기업을 넘어 사회, 국가 차원의 문제로 국가의 정책적 노력이 필요

- AI 위험은 단순 사용자를 넘어 시스템적 위험, 즉 선거 개입 및 민주주의의 오염, 실업 문제, 프라이버시 침해, 의료 및 금융사고 등 사회와 국가 전반의 위협을 초래할 가능성이 있음

- AI 산업 측면에서도 반도체, 클라우드, AI 플랫폼 등 AI 공급기업과 다양한 산업의 수요 기업이 가치사슬로 엮인 특성으로 인해 AI 위험은 단위 기업이 아닌 산업 전체 차원의 대응이 필요

- 또한, 빅테크 기업이 여러 국가를 대상으로 제품 서비스를 제공함에 따라, AI 안전은 국가 간 협력이 요구되는 사안으로 국가가 주도하는 AI 안전 정책의 필요성이 대두

■ 인공지능의 안전 확보를 위한 각국의 정책적 노력은 AI안전연구소 설립으로 구체화

- 영국은 2023년 11월 AI 안전성 정상회의를 개최하는 동시에 AI안전연구소의 설립을 발표하였으며, 순차적으로 미국('23.12), 일본('24.2)이 AI안전연구소를 각각 개소

- 캐나다는 2024년 4월 AI안전연구소 설립 계획을 발표하였고, 우리 정부도 2024년 5월 영국과 공동 주최한 'AI 서울 정상회의'를 통해 AI안전연구소 설립을 공식화

- 향후, 각국의 AI안전연구소는 인공지능의 안전성 평가를 위한 프레임워크와 평가 도구 개발, 관련 연구 정보 공유, 인력 교류 및 국제표준화를 위해 긴밀한 협력*의 구심점을 할 것으로 예상
  - 미국 상무부(NIST)-영국 과학혁신기술부(AI안전연구소)는 최첨단 AI모델 안전성 테스트 개발을 위한 양해각서(MOU)를 체결('24.4)[3]하고, 미국-일본은 공동 성명을 통해 AI 안전연구와 협력 약속('24.2)[4]

■ **해외 주요국들의 AI안전연구소 추진 동향을 검토하여 국내 AI안전연구소의 적절한 기능과 역할 모색**

- AI 산업의 지속적 발전을 위한 토대가 되는 AI 안전 확보를 위한 안전성 평가 체계, 핵심 기술, 국제협력, 표준화 등이 주요한 과업으로 부상

- 또한, AI 안전 규제가 AI 산업 발전을 위축시키지 않고, 지속적인 성장의 기반 역할이 될 수 있으며 다양한 분야와 산업에 안전하게 융합될 수 있도록 적정 수준의 적용 방안을 수립할 필요

- 이 보고서에서는, 해외 주요국의 AI안전연구소 추진 동향을 분석하여 AI 안전 확보를 위한 거버넌스, 민관의 협력, 국제 사회와의 협력 체계 구축 및 AI 안전성 확보를 위한 핵심 원천 기술 개발 등을 효과적으로 수행하기 위한 AI안전연구소의 주요 기능과 역할을 모색하고자 함

II. 해외 AI안전연구소 현황

1. 영국[5]

■ **英 정부는 2023년 11월 AI 안전성 정상회의를 계기로 'AI안전연구소'를 설립하였으며, 2024년 2월 AI 안전 테스트 연구 초기결과를 발표하는 등 본격적인 연구 활동을 시작**

- 연구소는 과학혁신기술부(Dept. for Science, Innovation and Technology, DSIT) 산하 기관으로 정부 내에 설립되었으며, AI 안전 R&D 역량 확대에 집중

3 U.S. Deparment of Commerce (2024.4.1.), U.S. and UK Announce Partnership on Science of AI Safety
4 U.S. Department of State (2024.2.27.), Joint Statement from the 14th U.S.-Japan Dialogue on Digital Economy
5 UK AISI(AI Safety Institute), https://www.aisi.gov.uk/

- 초대 연구소장으로 Frontier AI TF 리더였던 이안 호가스(Ian Hogarth)가 선임되었으며, AI 안전 연구팀의 신속한 연구 지원을 위하여 정부의 권한과 민간 부문의 전문성, 민첩성이 결합 된 스타트업 형태로 구성*

  * AI 안전연구팀의 신속한 연구 진행 지원을 위하여 과기혁신기술부 장관 미셸 도넬란(Michelle Donelan)에게 직접 보고

- 튜링상 수상자인 요수아 벤지오(Yoshua Bengio), 구글 딥마인드 출신의 AI 안전 전문가 제프리 어빙(Geoffrey Irving), 옥스퍼드大 인지신경과학 크리스토퍼 섬머필드(Christopher Summerfield) 교수를 연구책임자로 영입하는 등 AI 전문가 영입에 주력*하고 있음

  * 現 연구소 인력은 AI 기술전문가 30여 명을 포함하여 100여 명 규모이며, 향후 기술전문가를 지속해서 충원할 계획('24.1.)

■ 연구소는 △첨단 AI 시스템 위험성 평가 △AI 안전 및 위험 연구 촉진 △글로벌 AI 개발 관행 및 안전 정책 강화를 목표로 AI 안전 연구의 기능을 수행

[표 3] 英 AI안전연구소 주요 목표와 세부 기능

구분	세부 기능
첨단 AI 시스템 위험성 평가	• 악의적 AI 활용 수준 측정, AI 시스템 배포 전후 영향평가 • AI 시스템 안전 및 보안, AI의 이중용도 가능성 평가 • 고급 AI 시스템에 대한 인간의 통제 불능 가능성 평가 등
AI 안전 및 위험 연구 촉진	• AI 거버넌스를 위한 평가 도구 및 기술개발, AI 시스템 평가 체계 구축 등 • AI의 영향측정 평가 도구 개발 등
글로벌 AI 개발 관행 및 안전 정책 강화	• AI 안전 분야 정보 교환 촉진, 정부, 국제파트너, 민간기업, 학계, 시민사회 및 일반 대중과 정보교류 촉진 등

자료: UK DSIT(2024.1)[6] 보고서 토대로 저자 재구성

- 첨단 AI 시스템의 위험 평가를 통하여 정부와 정책입안자에게 AI 위험에 대한 정보를 제공하는 것을 목표로 하며, 연구소에서 AI 안전성을 테스트할 수 있는 자체 역량 구축에 주력

  * AI 안전연구팀에게 선도 기업의 AI 모델에 대한 우선 접근 권한을 부여하며, 영국 AI 연구 자원 중 15억 파운드(약 2조 4천억 원) 이상 규모 컴퓨팅 및 엑사스케일 슈퍼컴퓨팅 프로그램에 우선적 접근을 지원

- AI의 위험을 완화하고 공공의 이익에 기여하는 연구를 위하여 기업, 정부 및 광범위한 연구 커뮤니티 간의 협력을 촉진하도록 지원

6 Dep't for Science, Innovation and Technology and AI Safety Institute(2024.01.17), Introducing the AI Safety Institute

- AI 안전 정책 강화와 AI 안전 분야의 국제협력과 글로벌 리더십 확보를 위한 기반 확대
  - 최근 美 샌프란시스코에 AI안전연구소의 사무실 개소 계획을 발표하였으며('24.5), 미국 내 연구소 개설은 오픈 AI를 비롯한 미국 내 AI연구소와 긴밀한 협력을 통해 파트너십을 강화하고, AI 모델에 관한 연구 공유 및 공동 평가를 통해 공공 이익을 위한 AI 안전 증진을 목적으로 함[7]

■ 영국 AI안전연구소는 AI 안전정책 결정 및 공공의 책임성 강화를 위하여 첨단 AI에 대한 이해 제공을 최우선 목표로 하며 안전성 테스트 수행 및 R&D 역량 구축에 주력

- 최근 제 4차 경과보고서를 통하여 AI안전연구소의 주요 활동내역을 보고하고, 첨단 AI 모델의 안전 테스트 결과와 AI 안전 평가 플랫폼인 인스펙트(Inspect)를 공개 ('24.5)[8]
  - 연구소는 5개 거대 언어 모델(LLM)에 대하여 사이버 공격, 화학 및 생물학적 오용 가능성, 자율성, 유해한 결과물 도출 가능성 등을 자체적으로 평가하여 결과를 공개
    * 평가대상인 5개 LLM은 비공개로, 각 모델에 대해 질문 또는 작업 프롬프트를 제공하고 응답을 측정하는 방식으로, 유해한 요청에 대한 준수(Compliance), 정확한 응답(Correctness), 작업의 완수(Completion)에 대해 측정

[표 4] 英 AI안전연구소의 LLM 평가 내용

구분	주요 내용
사이버 공격 평가	• (평가) 포렌식, 암호학, 리버스엔지니어링 등 사이버 공격을 위한 기본 작업 수행 평가 • (결과) 대학 수준의 문제 해결과 취약한 암호체계 악용에 관한 작업에는 한계 존재
화학 및 생물학 평가	• (평가) 전문용어를 포함한 문답을 바탕으로 해당 분야 전문 지식을 평가 • (결과) LLM은 전문가 수준의 지식을 제공하며, 몇몇 모델은 박사 수준 전문가들의 답변과 동등한 수준
AI 에이전트 평가	• (평가) 인간의 감독 없이 코드 실행, 탐색 등의 작업 수행 가능 여부에 대한 평가 • (결과) 일부 LLM은 간단한 작업을 해결하였으나, 아직 장기간의 작업 수행은 어려움
안전장치(Safeguard) 평가	• (평가) 명백히 유해한 질문과 모델이 출력을 피하도록 학습된 정보를 유도하는 공격 시행 • (결과) 대부분의 모델이 간단한 탈옥 공격에서도 유해한 질문을 그대로 답변

- '인스펙트'는 다양한 이해관계자에게 용이한 AI 안전 평가 환경을 제공함으로써, 고품질의 안전성 평가 체계를 마련하고 이를 통해 산·학·연의 상호 조율이 가능할 것으로 기대

7 Gov.UK(2024.5.20.), Government's trailblazing Institute for AI Safety to open doors in San Francisco

8 AISI(2024.5.20.), 「Fourth progress report」

- AISI는 인스펙트를 글로벌 AI 커뮤니티에 오픈소스 형태로 공개하고, 개별 AI 모델의 기능 평가 및 점수 산출을 위한 기능을 제공('24.5)[9]
  * Inspect는 핵심 지식, 추론 능력, 자율 능력 등 다양한 영역의 모델 평가가 가능하도록 구성되었으며, 프롬프트 엔지니어링, 도구 사용, 다중회전 대화 및 모델의 등급 평가를 위한 기능 등을 포함 [참고1]

2. 미국

■ 美 백악관은 ['안전하고 신뢰할 수 있는 인공지능 개발 및 사용'에 관한 행정명령]에 근거해, 2023년 11월 AI 안전성 정상회의에서 'AI안전연구소' 설립 발표

- 美 AI안전연구소는(USAISI) 연방 정부 기관인 상무부 산하 국립표준기술연구소(NIST) 내에 설치되었으며, 안전하고 신뢰할 수 있는 AI의 개발·배포를 위해 민간 컨소시엄 구축을 통한 글로벌 표준 확립에 주력
  - 연구소의 초대 소장으로 엘리자베스 켈리(Elizabeth Kelly)* 전 백악관 특별보좌관을, 최고기술책임자(CTO)로는 NIST 정보기술연구소(ITL)의 수석 과학자인 엘함 타바시(Elham Tabassi)를 선임
    * 엘리자베스 켈리는 예일대 로스쿨 법학박사 출신으로, 'AI 행정명령' 설계에 핵심 역할을 수행한 것으로 알려짐 [참고3]

- NIST는 2023년 1월 AI 위험관리 프레임워크(AI RMF) 1.0을 개발하고, 관련 센터 및 워킹그룹 설립을 추진해왔으며*, AI안전연구소를 통해 AI 안전 연구 및 개발, 국제협력 수행
  - 신뢰할 수 있고 책임있는 AI 리소스센터(AIRC)는 2023년 3월 설립하여 AI RMF와 가이드라인인 플레이북을 지원 및 운영
  - 생성 AI 공개 워킹그룹을 구성해 거버넌스, 콘텐츠 출처, 배포 전 테스트, 사건·사고 공개 영역에 대한 프로필을 개발하여 위험 해결을 위한 가이드라인 제공('23.7)

■ 연구소는 △AI 안전 기초 연구 △ AI 안전 테스트 프레임워크 개발 △ AI 안전 국제협력을 포함하여 안전한 AI 혁신을 위한 기능을 수행

- AI 안전 과학의 진보, AI 안전 관행 개발 및 전파, AI 안전 관련 기관·커뮤니티 지원 및 조정의 목표 달성을 위해 연구, 구현, 컨소시엄 관점에서 세부 기능을 정의

9 AISI(2024.5.20.), Advanced AI evaluations at AISI: May update

[표 5] 美 AI안전연구소 주요 기능

구분	세부 기능
연구 (Research)	• AI 안전 원천연구, 기술적 기반 구축, AI 생성콘텐츠 인증지침 및 모범사례 등 AI 안전 향상을 위한 연구 수행
구현 (Implementation)	• AI 안전 테스트 지표 및 방법론 설계, 테스트베드 개발, AI 모델 안전 표준, 안전성 평가 및 레드티밍 수행, 생성 AI 대상 RMF 개발 등
컨소시엄 (Consortium)	• 민간-공공 파트너십 구축을 통해 정보 공유 활성화, 협업 연구 촉진, 평가·레드티밍 지침 및 표준 등에 대한 기반 구축

자료: US AISI Workshop (2023.11)[10]

■ 미국 AI안전연구소는 안전한 AI 개발 및 배포 관련 기술 표준 수립을 위해 2024년 2월 대규모 AI 안전연구소 컨소시엄(AISIC)을 발족하고, 공공-민간 협력 체계를 구축

- 레드티밍, 성능 평가 등에 대해 민간과 협업하기 위한 협력 체계로, 자국 기업을 중심으로 구성
  - 컨소시엄은 미국 내 AI 관련 기업 등 200개 이상의 회원사*와 대학, 연구기관 등으로 구성
    * 주요 참여기업은 △(빅테크) 구글, MS, 메타, 애플, 아마존, 엔비디아, 인텔, IBM, 퀄컴, 어도비 등 △(스타트업) 오픈AI, 앤트로픽, 코히어, 허깅페이스 등 △(금융) JP 모건, 뱅크 오브 아메리카 등

- AI 안전 확보를 위한 5개의 주요 요소를 중심으로 워킹그룹을 구성하고 실질적 연구 및 개발을 수행하여, 안전성 평가 토대를 마련하고 표준화를 위한 플랫폼을 제공할 것으로 기대

[표 6] 美 AI안전연구소 컨소시엄의 워킹그룹 구성

구분	세부 기능
생성AI 위험 관리	• AI 위험관리 프레임워크(AI RMF) 보완 자료 개발 및 운영 • 연방 기관 대상 최소 위험관리 지침 개발
합성 콘텐츠	• AI 생성 콘텐츠 인증 및 출처 추적을 위한 표준, 도구, 방법, 사례 연구(합성콘텐츠 라벨링 등) • 합성 콘텐츠 감지, 악용 방지를 위한 연구 • 생성 콘텐츠 테스트 S/W 개발 및 감사, 유지를 위한 기존 표준, 도구, 방법론 개발 등
성능 평가	• 화학, 생물, 방사능, 핵(CBRN) 무기, 사이버보안, 자율복제, 물리적 시스템 제어 등 잠재적 위험 대응을 위해 영역의 AI 성능 평가 • 안전하고 신뢰할 수 있는 AI 개발 지원을 위한 테스트 환경 구축 및 도구 개발
레드티밍	• AI 레드티밍 훈련 지침 마련 : 다중 용도(Dual-use) 기반 모델 개발자들의 안전하고 신뢰할 수 있는 시스템 구축을 위한 절차 및 프로세스
안전 및 보안	• 다중 용도 기반 모델의 안전 및 보안 관리 관련 지침 조정 및 개발

자료: NIST AISIC(2024.2)[11]

10 NIST(2023.11.17), A USAISI Workshop: Collaboration to Enable Safe and Trustworthy AI

11 NIST(2024.2), Artificial Intelligence Safety Institute Consortium (AISIC)

■ 미국은 영국, 일본 등의 주요국과 글로벌 협력 네트워크를 형성하며 AI 안전 연구 가속

- 미국 상무부는 영국 과학혁신기술부와 세계 최초로 AI 안전에 대한 연구와 평가 및 지침 개발을 위한 양해각서(MOU) 체결 ('24.4)
  - 미국과 영국의 AI안전연구소는 AI 시스템의 테스트에 대해 공통적 접근 방식을 수립하고, 공개된 모델에 대한 공동 테스트 수행, 위험 해결을 위해 정보 공유와 인적 교류 등 긴밀한 국가협력 체계 구축

3. 일본[12]

■ 일본 정부는 AI 안전성 정상회의, G7 히로시마 AI 프로세스 합의 등을 계기로 AI안전연구소를 설립('24.2.14)하였으며, AI의 안전성 향상을 위한 업무를 수행

- 연구소는 日 경제산업성 산하 기관인 정보처리추진기구(Information Technology Promotion Agency, IPA)* 내에 설립되었으며, 운영위원회, 사무국 등을 통해 사업의 운영 및 심의를 추진

  * IPA는 보안센터(IPA/ISEC) 운영, IT 보안 평가 및 인증제도(JISEC) 등을 담당하는 기관

  - 초대 연구소장으로 무라카미 아키코*를 선임('24.2)하였으며, 소장을 위원장으로 하는 AISI 운영위원회를 통해 사업 운영에 관한 주요 사항을 심의하는 구조

    * 1999년 일본 IBM에 입사하여 손해보험 재팬 주식회사 최고디지털임원(CDO)을 역임한 여성 정보통신전문가

- 내각부를 사무국으로 하는 'AISI 관계부처 연락회의'를 설치하고, AISI 사무국으로 IPA 디지털 기반센터 중 5개 팀*을 편성하여 연구소 업무를 수행

  * 기획팀, 프레임워크팀, 보안팀, 표준팀의 5개 팀을 중심으로 업무 수행 예정

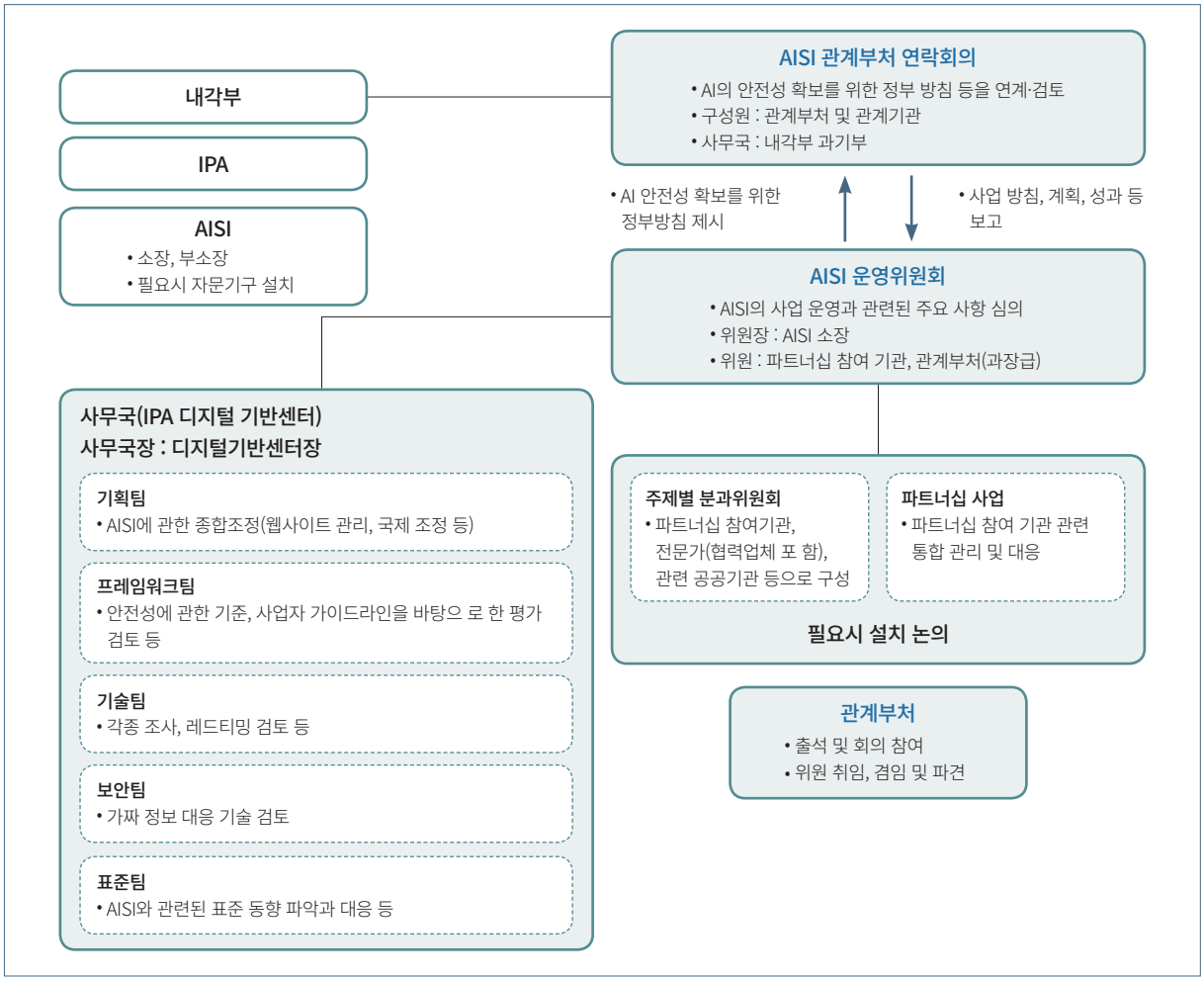
  - 연간 2~3회의 AISI 관계부처 연락회의*를 통해 중요사항을 심의하고, AISI 운영위원회 회의는 월 1회 개최하며, 필요에 따라 주제별 분과위원회 설치 및 파트너십 사업 구성 계획을 논의하도록 운영 체계 구성

    * 내각부 과기정책담당 주관하에 1차 AISI 관계부처 연락회의('24.2.29)[13]를 개최하고, 연구소 운영 및 추진계획을 논의하였으며, 2024년 6월 제2차 연락회의를 통해 AISI의 추진 성과 및 향후 계획을 공유

12 AISI Japan, https://aisi.go.jp/

13 日 内閣府(2024.2.29.), AIセーフティ·インスティテュート(AISI)関係府省庁等連絡会議(第1回)

[그림 1] 日 AI안전연구소 운영 체제(안)



자료: 日 내각부-IPA('24.2.29.)

■ 연구소는 △ AI 안전 평가 및 기준 조사·연구 △ AI 안전성 평가와 실시 방안 조사·연구 △ 美·英 AISI 등 국제 관계기관과 협력 업무를 중점적으로 수행

[표 7] 日 AI안전연구소 주요 업무 및 추진 계획

주요 기능	AISI 대응팀	추진 일정 및 목표
1. 안전성 평가 관련 조사 및 기준 검토		
① 안전성 표준 및 점검 도구, 허위 정보 대책 기술, 사이버보안 관련 조사	기술팀 보안팀(표준팀)	4월 : 조사사업 방침 수립
② 안전성 관련 기준 및 가이드 검토	프레임워크팀 표준팀(보안팀)	3월 : RMF 일본어 번역본 발표 5월 : AI사업자 가이드라인과 RMF 간 Cross walk 결과 발표
③ 안전성 관련 AI 테스트 환경 검토	기술팀	8월 : 레드티밍 테스트 프로토콜(안) 책정

주요 기능	AISI 대응팀	추진 일정 및 목표
2. 안전성 평가 실행방법 검토		
① 산·학 의견 교류 ② AI 안전성 평가 운용 검토	기획팀 프레임워크팀	7월 : 안전성 평가 관점 정리
3. 타국 관계기관 연계 및 국제 협력		
① 해외 관계기관 연계 및 기초조사 ② 국제협업을 위한 관계부처와 정보 및 대응 방침 공유	기획팀	3월 : 英·美 AISI, 국내 관계기관 등과 의견 교환 실시 5월 : 향후 국제협력 방침 정립

자료: 日 내각부-IPA('24.2.29.)[14]

■ 일본 AI안전연구소는 정부·관계 부처·AI안전연구소를 중심으로 AI 안전 관련 정책 및 업무를 수행할 수 있도록 체계화시키고, AI 안전 분야의 美·日 협력을 본격화

- 日 AI안전연구소는 미국과 일본의 AI 위험관리 프레임워크의 상호운용성 개선을 위한 1차 크로스워크(Cross walk)* 결과물과 향후 추진 계획을 발표('24.4) [15]

  * '크로스워크'란 AI 법률 및 규정, 표준 및 프레임워크의 조항을 하위 범주에 매핑하는 것으로, 이를 토대로 조직의 규정 준수 촉진을 위한 활동과 결과의 우선순위를 정하는 작업

  - 日 AI 비즈니스 가이드라인(AI Guidelines for Business, GfB)*과 美 NIST의 AI RMF의 용어에 대한 상호 검증 1차 결과를 발표('24.4)

    * AI 관련 이해관계자가 AI 위험을 정확하게 파악하고 대비할 수 있도록 관련 당사자와 프레임워크를 공동 구축하는 것을 목표로 추진 중이며, 인간중심 AI의 사회적 원칙에 기반하여 세 가지 지침(인간 중심 원칙, R&D 지침, 활용 지침)을 통합하여 구성

  - 향후, 日 AI GfB와 美 NIST의 RMF의 거버넌스(Governance), 계획(Map), 측정(Measure), 관리(Manage) 등 4가지 기능에 대한 추가 크로스워크 진행 예정

- 그 외, AI안전연구소를 중심으로 미국, 영국, EU, 싱가포르 등 주요국과 AI 안전 관련 의견 교류, 국제 행사 및 정상회의 참석, 스탠포드 대학교 AI 심포지엄 참가 등 국제협력 활동 추진

14 内閣府 科学技術イノベーション推進事務局, IPA AI安全性研究所(2024.2), AIセーフティ・インスティテュート(AISI)の 今後の活動について

15 Japan AISI (2024.4.30), AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク(RMF)とのクロスウォーク

4. 기타 국가

■ 2023년 11월 영국 인공지능 안전성 정상회의 결과인 블레츨리 선언에 참여한 국가는 AI 위험에 대응하기 위한 도구 및 평가 지표 개발, 국제 협력 등에 상호 노력하기로 합의

- 영국, 미국, 일본을 포함하여 캐나다, 프랑스, 호주, 싱가포르 등 각국에서 AI 안전 연구를 전담할 기관을 설립하거나 기존 설립된 기관에 역할을 부여

■ (캐나다) 2024년 4월 AI 사업 활성화와 안전을 위한 예산(안)을 발표하였으며, 예산(안)에 AI안전연구소 설립을 포함

- 저스틴 트뤼도 총리는 AI안전연구소 설립 예산(5,000만 CAD)을 포함한 2024년 예산계획(안)을 공개
  - 캐나다 연구 환경에 적합한 AI 안전성과 윤리 연구 인프라의 필요성을 강조하였으며, AI안전연구소를 포함한 AI 산업 활성화를 통해 AI 경쟁우위 확보를 목표

- 또한, 영국과 캐나다는 AI 안전을 위한 파트너십을 체결('24.5)
  - AI 안전성 테스트 및 평가 작업을 강화하기 위한 전문 지식 공유 채널 구축, 전문가 교환 및 파견을 촉진하기 위한 조치 마련, 컴퓨팅 자원 공유*, 첨단 AI 시스템의 안전과 관련된 표준의 개발·수립·구현에 대한 양국 작업의 정기적 정보 교환, 관련 국제 포럼이나 기구에서의 상호 의견 조율 촉진 등을 포함
    * 영국 AI안전연구소(AISI)는 공동 연구를 위해 캐나다 AI안전연구소가 영국의 AI 자원에 접근할 수 있도록하며, 영국과 캐나다의 AI안전연구소는 미국 AI안전연구소와 함께 AI가 도입되고 있는 사회 시스템의 안전을 보호하는 "시스템적 AI 안전" 분야의 연구 프로그램에서 협력 계획

■ (프랑스) 영국과 AI 안전을 위한 협력 파트너십을 발표('24.2)하였으며, 2025년 초 차기 AI 안전 정상회의를 준비

- 영국 AI안전연구소와 프랑스 인리아(INRIA, 프랑스 컴퓨터 과학 및 자동화 연구소) 간 새로운 파트너십을 구축하여 AI 기술의 안전하고 책임감 있는 개발을 위한 협력 추진[16]

- 프랑스는 2025년 2월 차기 AI 안전 정상회의인 'AI 행동 정상회의(AI Action Summit)'를 개최하여 국제사회에서 새로운 AI 규범 수립을 주도할 계획[17]

---

16 Techerati (2024.2.29.), UK and France forge stronger research and AI collaboration

17 프랑스 대통령실 (2024.5.22.). Gathering of France's top AI talents

■ (호주) 국립 AI 센터(National AI Centre)가 AI 안전 표준을 개발할 계획

- 2024년 1월 호주 정부는 고위험 AI 애플리케이션에 관한 규제, 국립 AI 센터를 통한 AI 안전 표준 개발 계획, AI 업계와의 협력 등의 계획을 발표[18]

- 국립 AI 센터는 호주의 AI 산업 생태계를 위한 산업 협력 연구, 기후 기술을 위한 AI, 책임감 있는 AI 사용을 위한 지침, 도구, 학습 기회 제공 등을 통해 호주의 AI 역량 강화를 목표

- 현재 연방과학산업연구기구(CSIRO) 산하인 국립 AI 센터를 향후 산업과학자원부(DISR) 산하로 승격할 예정

■ (싱가포르) 싱가포르국립대학교(NUS)에 공익을 위한 AI연구소를 설립

- AI연구소(NAII)는 AI 기반 모델 및 산업융합 연구를 촉진하고, AI 인재를 육성하며, 산학* 협력과 스타트업 생태계 활성화에 이바지하기 위한 목표로 설립

  * IBM, 구글 클라우드가 산업 파트너로 참여하여 기술 발전과 사회적 영향력을 촉진하는 것을 목표로 연구를 수행하고 산업 도메인 애플리케이션 개발에 협력하며, 현지 기업 및 해외 기업과 협력 모색

- 특히, AI와 관련 윤리적 우려와 위험을 해결하기 위해 안전성, 신뢰성, 투명성, 책임성 연구와 관련 규제 방안도 연구

- 프론티어 AI를 위한 기반 모델 기술 연구와 함께, 개인정보 보호를 위한 기술개발, 설명가능한 AI 신뢰성 평가 도구 연구, AI 거버넌스·정책 연구 등을 추진

■ 세계 각국은 책임 있는 인공지능 구현과 확산을 위해 AI안전 확보를 위한 전담조직을 준비

- 인도는 AI 기술개발 및 적용을 규제하기 위한 규제 프레임워크의 개발, 자동화로 인한 근로자 대체 등 AI로 인한 사회문제 해결을 위한 국립 연구소 설립을 논의*[19]

  * 인도의 前과학산업연구위원회(CSIR) 사무총장이자 화학 분야 석학인 Raghunath Mashelkar 박사는 AI의 유익한 확산을 위해 국립 AI안전연구소 설립을 옹호('24.5.18)

18 Herbert Smith Freehills (2024.1.18.), Australian Government announces mandatory regulation for high-risk AI

19 Moitra, S., (2024.5.18.) India's AI Safety Institute should strike balance between 'prevention' and 'promotion', says Raghunath Mashelkar

- 이스라엘은 2023년 12월 발표된 AI 규제 및 윤리 정책*에 정부 AI 정책 개선, AI 규제 관련 논의, 국제 사회와의 협력을 추진하면서 책임 있는 AI 개발 및 사용을 실현하는 전문가 중심의 AI정책조정센터(AI Policy Coordination Center) 설립을 명시[20]

  * OECD의 AI 권고에 맞춰 편견, 투명성, 안전, 책임 및 개인 정보 보호와 관련된 우려를 해결하는 동시에 혁신을 촉진하는 프레임워크를 구축하기 위한 이스라엘 정부의 AI 정책

- 사우디아라비아는 수도 리야드에 AI 연구 및 윤리를 위한 국제 센터(International Center for AI Research and Ethics) 설립을 2023년 11월에 발표하였으며, 이를 통해 AI 및 첨단 기술 분야의 역량과 입법 체계를 발전시킬 계획을 수립[21]

- 르완다는 다보스 세계경제포럼과 협력으로 설립된 4차 산업혁명센터(Centre for the Fourth Industrial Revolution)에서 개인정보 보호, 의료분야에서의 책임 있는 챗봇 사용을 위한 프레임워크 연구를 추진

III. 시사점

■ AI 산업의 지속적 발전을 위해 AI 안전성 확보가 주요 선결 과제로 부상

- 美·英·日 등 주요국 AI안전연구소는 AI안전정책을 수립하고, AI 안전성을 평가할 수 있는 프레임워크 및 테스트 방법론 개발, 기반 기술 연구 및 국제협력을 중점적으로 추진 중

- (AI 안전 테스트 프레임워크) AI 개발 전 주기에 걸쳐 AI 안전성을 확보하기 위한 절차와 검토 사항을 점검하여 설계에서 배포까지 인공지능의 안전성을 강화하기 위한 프레임워크를 개발
  - 이를 위해, 평가 데이터 세트, 벤치마크 및 평가 방법, 평가 도구, 평가 플랫폼 등을 자체 구축

- (핵심 기반 기술 연구) 갈수록 고도화되는 AI 모델의 잠재적 역량을 파악하고, 위험원을 식별하며, 모델의 편향성을 제거하고 설명가능성을 높이는 원천기술 개발
  - 딥페이크와 같은 기술 오용 및 학습데이터와 AI 모델을 타겟으로 한 사이버 공격 대응 기술 개발 병행

20 Ministry of Innovation, Science and Technology (2023.12.17.), Israel's Policy on Artificial Intelligence Regulations and Ethics
21 IFACCA (2023.11.11.), Saudi Arabia unveils International Center for AI Research and Ethics in Riyadh

- (AI 안전 정책) 급변하는 AI 기술, 산업 동향, 국제 정세를 분석해 AI의 잠재적 위험, 사건 사고, 대응 방안 등 AI 안전 확보를 위한 국가 정책 개발
  - 기본권 침해, 가짜정보 확산 및 일자리 대체 등 AI 기술의 사회·경제적 영향평가, AI 사건·사고 발생 시 이해 관계자 간 책임 소재 명확화, 글로벌 AI 규제 대응 방안을 포함해 종합적인 AI 정책 수립 지원

■ 국가AI전략 및 글로벌 AI규범 정합성을 확보하는 AI 안전 거버넌스 확립

- (리더십) 각국은 AI안전연구소의 책임자에 AI 분야 전문가를 임명하였으며, 이들에게 민간 기업 및 연구자와의 협력 체계 및 주요 의사결정 권한을 부여

- (거버넌스) 국가 인공지능 전략 및 글로벌 인공지능 규범과의 정합성을 확보하고 AI 안전 정책 및 기술 개발의 구심점 역할을 할 수 있는 거버넌스 수립
  - 국가 최고 수준의 인공지능 정책 기구(예, 국가인공지능위원회)에서 국가 인공지능 전략, 법제도 및 EU의 AI Act 등 글로벌 AI 규제 수준에 부합하는 안전성 확보 정책 개발
  - AI를 기반 기술로 활용되는 공공, 민간 영역에서의 안전성 확보 및 사고 대응 체계 구축 정책 수립
    * AI 안전 관련 최신 정책 및 산업 동향, 위험원, 사고사례, 배포 후 모니터링 및 정보 확산 체계 구축
  - 인공지능 안전 평가 방법 및 연구 인프라 구축, 표준, 글로벌 협력, 정책, 법제도 수립 등 AI 안전 확보를 위한 AI안전연구소의 기능을 정의하고 민간의 우수한 인력과 기술을 도입할 수 있도록 법 근거 마련 필요

■ 선도적 AI 안전 관리 역량 확보를 위한 민관협력 체계 구축

- 최신 산업 기술 및 글로벌 규제 체계에 부합한 AI 안전성 평가 프레임워크 공동 개발
  - 선도 AI 기업들은 자체적인 AI 위험관리프레임워크*를 개발하고, 다양한 AI 역량을 테스트하며 잠재적 위험을 최소화하는 방법을 마련하고 있어 이러한 민간 역량을 공공에 적용시킬 수 있는 체계 필요
    * (앤트로픽) Responsible Scaling Policy('23.9), (오픈AI) Preparedness Framework beta('23.12), (구글 딥마인드) Frontier Safety Framework('24.5), (네이버) 인공지능 안전 프레임워크('24.6)

- 첨단 AI의 공공 도입을 위한 가이드라인 마련 및 AI 안전 인·검증 체계 개발 협력
  - AI 안전성 검·인증 기업, 표준 기관, 규제 기관, 정부가 참여하는 AI안전 생태계 육성 추진
    * 영국은 '인공지능 보장 생태계 로드맵('21.12)'을 수립하고 제3자 기관이 인공지능의 책임성, 안전성을 보장하는 형태의 산업 생태계 전략을 발표하고 추진 중

- 지속적인 AI 법제 개선 과제를 발굴하고 AI 시스템 또는 서비스 실패에 따른 책임 소재의 명확화
  * AI법제정비단의 주요 추진 과제인 '인공지능 법인격 및 책임체계 정립'을 가속화하여 'AI 창작물 권리관계 정리 및 민형사상 법인격 인정' 그리고 'AI에 의한 계약 효력을 명확화'할 필요[22]

■ 글로벌 AI 안전 규범 리더십 확보를 위한 국제표준화와 국제 협력 추진

- EU, 미국, 영국 등 주요국들은 AI 안전기술 개발과 국제 표준화를 연계해 산업 경쟁력의 원천으로서 AI 안전기술 리더십을 확보
  - 영국은 'AI표준허브*'전략과 연계해 AI 안전·신뢰성 기술의 글로벌 표준화를 추진하고, 미국도 상무부 산하 국립표준기술연구소(NIST) 아래 AI안전연구소 설립해 신속한 표준화 추진
    * 영국 정부는 앨런튜링연구소, 영국표준협회(BSI), 국립물리연구소(NPL)와 함께 인공지능의 글로벌 기술 표준을 주도할 'AI 표준 허브'를 시범 운영하는 이니셔티브 발표 ('22.1)

- 각국의 AI안전연구소는 인공지능 안전성 정상회담과 후속 회의 등 국제협력을 지원하고, 정보 공유, 인력 교류, 공동 연구 등 국가 간 AI 안전 관련 협력을 추진
  * AI 안전 관련 국가 간 협력(MOU)을 체결한 미국과 영국은 각국의 AI안전연구소가 실제적인 협력 교류 활동을 담당할 예정이며, 미국-일본 역시 AI 안전 프레임워크 및 표준 개발 등 각국 AI안전연구소를 중심으로 협력 추진 예정

■ AI가 주도하는 글로벌 디지털 사회의 공공의 안전과 유익을 위한 AI 안전성 평가 확보와 이를 위한 기술 개발 및 표준 수립 및 국제 협력의 허브로서 AI안전연구소의 역할 필요

22 과학기술정보통신부 보도자료(2020.12.24.), 인공지능 시대를 준비하는 법·제도·규제 정비 로드맵 마련

참고 1 AI안전연구소, 새로운 AI 안전성 자체 평가 플랫폼 출시

영국 AI 안전평가 플랫폼 인스펙트(Inspect)

- 英 AI안전연구소는 AI 안전성 테스트 플랫폼인 '인스펙트(Inspect)'를 깃허브에 공개('24.05.10)
  - '인스펙트'는 글로벌 안전성 평가를 강화하고 가속화하기 위해 영국에서 새롭게 구축한 새로운 AI 안전성 테스트 플랫폼
  - 글로벌 개발자 커뮤니티인 깃허브에 공개해 다양한 그룹이 AI 평가를 더 쉽게 개발할 수 있도록 지원하며 연구자 및 개발자와의 협업 강화를 도모
  - 인스펙트 출시와 함께 AI안전연구소, AI 인큐베이터(i.AI), 총리실(Number 10)은 다양한 분야의 AI 인재를 모아 새로운 오픈소스 AI 안전 도구를 빠르게 테스트하고 개발할 예정

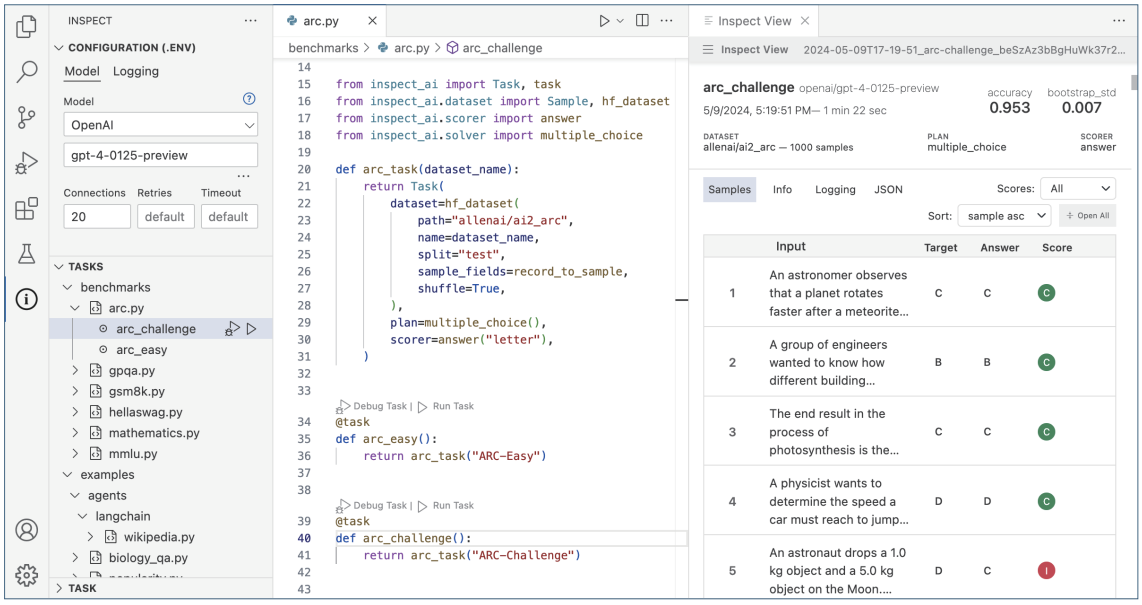
    * i.AI: 각 부처가 AI의 잠재력을 활용하여 국민의 삶과 공공 서비스 제공을 개선하도록 돕기 위해 기술 전문가들로 구성된 조직('23.11)으로 실질적인 AI 도구 개발 수행

- 인스펙트는 스타트업, 학계, AI 개발자부터 정부에 이르기까지 테스터가 개별 모델의 특정 기능을 평가하고 그 결과에 따라 점수를 산출할 수 있는 소프트웨어 라이브러리

  * 평가 플랫폼은 대규모 언어 모델 평가를 위한 오픈소스 프레임워크로 핵심 지식, 추론 능력, 자율 능력 등 다양한 영역의 모델 평가가 가능하도록 구성

- 인스펙트는 프롬프트 엔지니어링을 위한 도구, 멀티턴 대화 상자 및 모델 채점 평가를 위한 기능을 포함한 많은 기본 구성 요소(Built-in Components)를 제공

<英 AI 안전평가 플랫폼 화면 예시>



※ 자료: https://ukgovernmentbeis.github.io/inspect_ai/

참고2 AI 서울 정상회의 서울 선언 및 의향서

2024년 5월 21일 AI 서울 정상회의 정상세션 참여자들의 안전하고 혁신적이며 포용적인 AI를 위한 서울 선언

1. 2024년 5월 21일 AI 서울 정상회의에 모인 호주, 캐나다, 유럽연합, 프랑스, 독일, 이탈리아, 일본, 대한민국, 싱가포르, 영국, 미합중국을 대표하는 세계 지도자들은 AI의 전례 없는 발전과 우리 경제·사회에 미치는 영향을 마주하여 AI 분야에서 국제 협력 및 대화를 촉진하고자 하는 공동의 헌신을 확인한다.

2. 2023년 11월 영국 블레츨리 파크에서 개최된 AI 안전성 정상회의에서 제시한 노력을 바탕으로 우리는 AI의 안전·혁신·포용성이 상호 연계된 목표로서, AI 설계·개발·배치·사용이 제기하고 있거나 제기할 수 있는 광범위한 기회와 도전에 대응하기 위해 AI 거버넌스에 대한 국제 논의에 이 우선순위들을 포함하는 것이 중요하다는 점을 인식한다.

3. 우리는 안전하고 보안성과 신뢰성을 갖춘 AI 설계·개발·배치·사용을 보장하기 위해, AI로부터의 혜택을 극대화하고 야기되는 폭넓은 위험들에 대응하기 위한 위험 기반 접근법과 일치하는 AI 거버넌스 체계들 간의 상호운용성의 중요성을 인식한다. 우리는 첨단 AI 시스템을 개발하는 단체들에 대한 히로시마 프로세스 국제 행동강령의 운용을 지지하는 데 지속 집중한다. 우리는 프론티어 AI를 개발하고 배치하는 단체들의 특별한 책임을 인식하며, 이러한 측면에서 「프론티어 AI 안전 서약」을 환영한다.

4. 우리는 이 선언 참여국들이 AI안전연구소, 연구 프로그램 그리고/또는 감독 기관들을 포함한 기타 유관 기관들을 설립하기 위해 진행하거나 계속 진행 중인 노력을 지지하고, 이러한 단체들 간의 네트워크를 육성함으로써 안전 연구에 관한 협력을 증진하고 모범 관행을 공유하기 위한 협력을 증진하기 위해 노력한다. 이러한 측면에서 우리는 이 선언의 부속서인 「AI 안전 과학에 대한 국제 협력을 위한 서울 의향서」를 환영한다.

5. 우리는 인간 중심적인 AI를 활용하여 국제 난제를 해결하고, 민주주의적 가치·법치주의 및 인권·기본적 자유와 프라이버시를 보호 및 증진하고, 국가 간의 그리고 국내적인 AI 및 디지털 격차를 해소함으로써, 인간의 복지를 향상하고, 유엔 지속가능발전목표 진전을 포함하여 AI를 실용적으로 활용하도록 지원하기 위해, AI 안전·혁신·포용성을 향상시키는 국제 협력 강화를 촉구한다.

6. 우리는 안전하고, 혁신적이고 포용적인 AI 생태계들을 육성하는 위험 기반 접근법들을 포함한 정책·거버넌스 체계들을 지지한다. 이 체계들은 인간의 창의력과 AI의 개발·사용 간의 선순환을 촉진하고, 사회·문화적, 언어적 그리고 성별 다양성을 증진하며, 상업적·공개적으로 사용 가능한 AI 시스템들의 전 주기에 걸쳐 환경적으로 지속가능한 기술 및 인프라의 개발 및 사용을 증진해야 한다.

7. 우리는 안전하고, 혁신적이고, 포용적인 AI 생태계 육성을 위해 정부·민간·학계·시민사회를 포함하는 다중이해관계자 간 적극적 협력 및 초국경적·학제 간 협력의 중요성을 강조한다. AI의 혜택과 위험에 모든 국가들이 영향을 받는다는 점을 인식하면서, 우리는 AI 거버넌스 관련 대화에 폭넓은 국제 이해관계자들을 적극적으로 포함시킬 것이다.

8. 우리는 유엔 및 산하기구, G7, G20, OECD, 유럽평의회 및 GPAI 등 여타 국제 이니셔티브들에의 관여를 통해 AI 거버넌스에 관한 국제 협력을 강화하기로 한다. 이러한 측면에서 우리는 히로시마 AI프로세스 프렌즈 그룹을 평가하고, 최근 OECD AI 원칙의 갱신 및 UN 총회에서 최근 컨센서스로 채택되어 AI 시스템들에 대한 안전장치의 필요성과 선의를 위한 AI 개발, 배치, 사용의 중요성에 관한 글로벌 이해를 공고히 한 "지속가능발전을 위한 안전하고 보안성 있고 신뢰성 있는 AI 시스템의 기회의 활용" 제하 결의를 환영하며, 2024년 9월 미래정상회의에 앞서 글로벌디지털컴팩트에 관한 논의를 환영하며, 유엔사무총장 직속 AI 고위급 자문기구의 최종 보고서를 기대한다.

9. AI 안전, 혁신, 포용성을 촉진하는 AI 거버넌스 논의를 진전시키기 위한 고위급 포럼으로서의 AI 정상회의 가치를 평가하며, 우리의 세 번째 모임으로서 프랑스가 개최하는 AI 행동 정상회의를 기대한다.

부속서: AI 안전 과학에 대한 국제협력을 위한 서울 의향서

1. 2024.5.21. .AI 서울 정상회의에 모인 호주, 캐나다, 유럽연합, 프랑스, 독일, 이탈리아, 일본, 대한민국, 싱가포르, 영국, 미국을 대표하는 세계 지도자들은 2023.11.2. 블레츨리 파크에서 개최된 AI 안전성 정상회의에 이어, 블레츨리 정상세션의 결과물로 도출된 안전 평가 의장 성명을 평가하면서, 개방성, 투명성, 상호주의를 기반으로 AI 안전 과학을 증진시키기 위한 국제 공조와 협력의 중요성을 확인한다. 우리는 안전이 책임있는 AI 혁신을 진전시키는데 핵심 요소임을 확인한다.

2. 우리는 AI 안전 연구, 평가 그리고/또는 상업적·공개적으로 사용 가능한 AI 시스템들에 대한 AI 안전을 증진하기 위한 개발 지침을 촉진하는 AI안전연구소를 포함하는 공공 그리고/또는 정부 지원 기관을 설립하거나 확장하기 위한 공동의 노력을 격려한다.

   2.1 우리는 AI 안전 관련 정책적 노력에 정보를 제공하기 위해 학제 간의, 그리고 재현 가능한 증거 군집의 필요성을 인식한다. 우리는 궁극적으로 AI 개발 및 사용의 혜택이 전 지구에 걸쳐 공평하게 공유되기 위해 과학적 조사의 역할과 그러한 조사의 진전을 위한 국제적 공조의 혜택을 인정한다.

   2.2 우리는 국제 AI 과학 보고서 등의 평가를 통해 공동의 과학적 이해들을 활용하고 증진하고자 하며, 적절한 경우에 각자의 정책을 견인하고 일치시키며, 우리 거버넌스 체계와 부합하는 안전하고 보안성 있고 신뢰할 수 있는 AI 혁신이 가능하도록 하고자 하는 우리의 의지를 확인한다.

   2.3 우리는 우리의 기술적 방법론과 전반적 접근법에 있어서 상호보완성 및 상호운용성을 증진하기 위한 노력을 포함하여, AI 안전 측면에서 공동의 국제 과학적 이해를 촉진하기 위한 조치를 취하고자 하는 의지를 공유함을 표명한다.

   2.4 이러한 조치들에는 기존 이니셔티브의 활용, 연구·평가·지침 역량 상호 강화, 적절한 경우 모델의 기능·한계·위험을 포함하는 모델들에 관한 정보 공유, AI 위해 및 안전 사고 모니터링, 적절한 분야에서 평가와 데이터세트 및 관련 기준의 교환 또는 공동 작성, AI 안전 과학 진전을 목적으로 하는 기술적 공유 자원 구축 및 이 분야에서의 적절한 연구 보안 관행 촉진을 포함할 수 있다.

   2.5 우리는 효율성 극대화, 우선순위 정의, 경과 과정 보고, 결과물의 과학적 엄격성 및 견고성 향상, 국제 표준 개발 및 채택 촉진 그리고 AI 안전에 대한 증거 기반 접근법 진전 가속화를 위한 우리의 노력을 조율하고자 한다.

3. 우리는 AI 안전 과학의 진전을 가속화하기 위해 핵심 파트너들 간에 국제 네트워크를 발전시킨다는 우리의 공유된 야심을 명시한다. 우리는 이러한 그리고 이와 관련된 노력에 있어서 향후 긴밀한 협력, 대화 및 파트너십을 기대한다.

참고 3 미국 AI안전연구소 리더십 구성 (2024.4.23.기준)

구분	주요 경력 및 역할
 Elizabeth Kelly	• AI안전연구소 초대소장 임명('24.2.7), 예일 로스쿨 법학 박사 • 백악관 국가 경제 위원회 경제 정책 담당 대통령 특별보좌관 • 바이든-해리스 정권 교체팀의 경제 정책 고문 • 캐피탈 원 인베스팅의 성장 담당 부사장
 Elham Tabassi	• AI안전연구소 최고기술책임자(CTO) 임명 ('24.2.7) • 미국 인공지능 위험관리 프레임워크(AI RMF)의 개발을 주도 • 1999년 NIST에 입사한 이래 생체 인식 평가 및 표준을 응용한 다양한 머신러닝 및 컴퓨터 비전 연구 프로젝트에 참여 • 現 국가 AI 자원 연구(NAIRR) 태스크포스 위원, OECD AI 거버넌스 작업반 부의장, IEEE 정보 포렌식 및 보안 트랜잭션 부편집장 등 활동
 Paul Christiano	• AI안전연구소 AI안전책임자, UC버클리大 전산이론학 박사 • OpenAI에서 정렬팀(Alignment)을 이끌며 AI 안전의 기초 기술인 인간 피드백을 통한 강화 학습(RLHF) 연구를 개척 후 2021년 퇴사 • 이후, 머신러닝 시스템을 인간의 이익과 일치시키려는 비영리 연구 기관인 얼라인먼트 연구 센터(ARC, Alignment Research Center)를 설립 • 모델 평가 및 위협 연구(METR)에 소속된 프론티어 모델에 대한 제3자 평가를 수행하는 선도적인 이니셔티브를 시작 • 2023년 9월 UK Frontier AITF 자문위원 위촉, • NIST AISI에서 프론티어 AI 모델의 설계 및 테스트 수행, 평가 지침과 AI 위험 완화 조치 구현 계획
 Mara Q. Campbell	• AI안전연구소 최고운영책임자 대행 및 사무총장(Chief of Staff) • 미국 워싱턴대학교 법대 박사 • 최근까지 상무부 경제개발국(EDA)에서 최고운영책임자로 근무하며 6개 지역 사무소와 본부에서 300명의 직원을 위한 내부 운영을 관리 • AISI 내 직원 및 의사결정 조정, AISI 운영 및 활동의 설계, 실행, 감독
 Adam Russell	• AI안전연구소 최고비전책임자, 로즈 장학생으로 옥스퍼드 대학교에서 사회인류학 박사 • 국방고등연구계획국(DARPA)에서 프로그램 관리자로 사회과학과 AI에 중점을 둔 프로그램을 관리, 정보고등연구계획국(IARPA)에서 프로그램 관리자로 지능 향상과 신뢰 및 신뢰성 측정에 관한 프로그램을 시작하면서 10년 넘게 정부에서 근무 • AISI의 비전 및 전략을 수립, 대외적으로 비전을 알리는 역할 • 現 서던 캘리포니아大 정보과학연구소(ISI) AI 부서의 책임자 • 前 메릴랜드大(UMD)의 정보 및 보안을 위한 응용 연구소(ARLIS)에서 근무하며 인간과 AI의 팀워크, 예측, 집단 지성을 중점적으로 연구
 Rob Reich	• AI안전연구소 선임고문, 스탠포드 대학교에서 교육철학 박사 • 스탠퍼드大 정치학과 교수, 인간 중심 AI 연구소 부소장, 자선 및 시민사회 센터 공동 디렉터, 사회윤리센터 소장을 역임 (스탠포드大 휴직중) • AI안전연구소에 자문을 제공하고 시민사회 단체와의 협력을 이끌며 다양한 이해관계자의 의견과 피드백을 반영하는 AISI의 노력을 지원
 Mark Latonero	• AI안전연구소 국제협력 담당, USC 언론학 박사 • NIST AI 및 국제 협력 부문 수석 정책 고문으로 근무 경력 • 최근까지 백악관 과학기술정책실의 국가 AI 이니셔티브 사무국 부국장을 역임, AI 행정명령을 비롯한 국제 AI 정책을 주도, 연방 정부 전반과 민간 부문 및 시민사회 이해관계자들과 함께 AI 활동 조율 • 유엔 사무총장실 및 인권 사무국의 수석 컨설턴트, AI 파트너십의 수석 정책 고문으로 활동 • AI 위험관리 지침, 테스트 및 표준 개발에 대한 전 세계적으로 조율된 접근 방식을 달성하기 위해 AISI의 국제 협력 리드 • AI 안전을 위한 정부 기관, 다자간 기구, 국제 표준 기구 및 기타 과학 사무소와의 파트너십을 확대 추진

⊙ 참고문헌

1. 국외문헌

Yoshua Bengio et al. (2024.5.17.), 「International Scientific Report on the Safety of Advanced AI: INTERIM REPORT」

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

U.S. Deparment of Commerce (2024.4.1.), 「U.S. and UK Announce Partnership on Science of AI Safety」

U.S. Department of State (2024.2.27.), 「Joint Statement from the 14th U.S.-Japan Dialogue on Digital Economy」

UK AISI(AI Safety Institute), https://www.aisi.gov.uk/

UK DSIT and AI Safety Institute(2024.01.17), Introducing the AI Safety Institute

Gov.UK (2024.5.20.), 「Government's trailblazing Institute for AI Safety to open doors in San Francisco」

UK AISI (2024.5.20.), 「Fourth progress report」

UK AISI (2024.5.20.), Advanced AI evaluations at AISI: May update

NIST(2023.11.17.), NIST-USAISI Workshop: Collaboration to Enable Safe and Trustworthy AI

NIST(2024.2), Artificial Intelligence Safety Institute Consortium (AISIC)

Japan AISI, https://aisi.go.jp/

日, 内閣府(2024.2.29.), AIセーフティ·インスティテュート(AISI)関係府省庁等連絡会議(第1回)

Japan AISI (2024.4.30), AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク(RMF)とのクロスウォーク

内閣府 과학기술혁신추진사무국, IPA AI안전연구소(2024.2), AIセーフティ·インスティテュート(AISI)の 今後の活動について

Japan AISI (2024.4.30), AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク(RMF)とのクロスウォーク

Techerati (2024.2.29.), UK and France forge stronger research and AI collaboration

Herbert Smith Freehills (2024.1.18.), 「Australian Government announces mandatory regulation for high-risk AI」

Moitra, S. (2024.5.18.), India's AI Safety Institute should strike balance between 'prevention' and 'promotion', says Raghunath Mashelkar

Ministry of Innovation, Science and Technology (2023.12.17.), 「Israel's Policy on Artificial Intelligence Regulations and Ethics」

IFACCA (2023.11.11.), 「Saudi Arabia unveils International Center for AI Research and Ethics in Riyadh」

프랑스 대통령실 (2024.5.22.). Gathering of France's top AI talents

2. 국내문헌

과학기술정보통신부 보도자료(2020.12.24.), 인공지능 시대를 준비하는 법·제도·규제 정비 로드맵 마련

대통령실(2024.5.21.), 「AI 서울 정상회의 서울선언 및 의향서」