제17차 정보통신전략위원회

25

AI안전연구소 설립 · 운영계획

2024. 10.

과학기술정보통신부

※ 여백 페이지

요 약

※ 여백 페이지

AI안전연구소 설립 · 운영계획 [요약]

1 설립 배경

◈ **'AI서울정상회의'**('24.5.) **후속조치**로 **AI안전성**을 **평가·연구**하고 **주요국 AI안전연구소**와 **협력**을 **전담하는 조직**으로 '**AI안전연구소' 설립** 추진

"**대한민국도 AI 안전연구소 설립**을 **추진**해 **글로벌 AI 안전성 강화를 위한 네트워크에 동참**하겠다"(**윤석열 대통령**, **'AI 서울 정상회의**', '24.5.21.)

□ 최근, AI기술이 빠르게 발전·확산 중이나, ①AI의 **기술적 한계**, ②인간의 **AI기술 오용**, ③**AI 자율성 확대** 등으로 **실존·잠재위험**도 **확대**

* ①환각(Hallucination), 편향성 등 / ②유해정보 활용(화학·바이오 무기 개발 등), 사이버 해킹, 가짜뉴스 배포 등 / ③인간 통제력 상실 등

○ 이러한 **AI위험**은 **국민 기본권, 국가 안보** 및 **사회 안전**과 **직결**되므로, **국가 차원의 AI 안전성 확보노력 필요**

* "AI 전문가들이 AI가 핵전쟁과 맞먹는 실존적 위협이라고 선언"(UN 사무총장, '23.6.)
* 주요국(英·美·日)도 정부·공공기관內 AI안전연구소를 운영하며 AI위험에 본격 대비 中

□ 이에 **AI안전연구소**를 **설립·운영**하여 **안전한 AI** 개발·활용을 **확산**하고, **국내 AI기업**의 **경쟁력 확보** 및 **글로벌 진출**을 적극 뒷받침*

* 자국 AI안전연구소가 없는 경우, 글로벌 진출을 추진하는 국내 AI기업(첨단 AI개발)이 해외 AI안전연구소 안전성 평가를 받아야 하는 제약 有(평가 장시간 소요, 기술유출 우려 등)

○ 아울러 **AI안전**에 대한 **국제적 연대 강화**와 **규범 정립**을 **수행**하고, 중장기적으로 **세계적 AI안전연구**를 **선도**하는 **기관**으로 **발전** 추진

< 설립 추진경과 >

① 'AI안전연구소 설립' 발표('24.5.21., 대통령, 'AI 서울정상회의')
② 주요국(英·美·日) AI안전연구소(AISI) 현장방문 및 동향분석('24.5.~6.)
* AI안전연구소 성격·역할, 안전 평가방식, 조직 운영방안 등 심층분석
③ 'AI안전연구소 설립자문위' 구성·운영(1차:'24.7.3 / 2차:'24.7.17)
④ 'AI안전연구소 설치·운영계획' NST 이사회 의결('24.8.12.)
⑤ 'AI안전연구소 설립준비위' 구성·운영('24.9.13.~)

英 AISI 방문('24.5.31.)

美 AISI 방문('24.6.11.)

2 설립 방안

1 조직 * 한국전자통신연구원(ETRI) 內 설치

○ **(구성) 소장 + 3실**(①AI안전정책 및 대외협력실, ②AI안전 평가실, ③AI안전 연구실)

○ **(운영) AI안전연구소 전문성·대표성·중요성** 등을 고려, **연구소** 운영의 **자율성·독립성**을 최대한 보장하고, **별도 수당체계** 도입

< 'AI안전연구소 설치·운영 규정'('24.8.22.) >

■ **제10조(조직 및 운영)** 소장은 법령 및 원규의 범위 내에서 **AI안전연구소 조직의 운영**에 있어 **최대한의 자율성과 독립성**을 갖는다.

2 인력

○ **(소장) AI분야 정책·기술 전문성**과 **리더십, 국제적 역량**이 뛰어난 외부전문가를 **초대 연구소장**으로 **채용**('24.11.)

○ **(직원)** ①**신규채용** + ②**ETRI인력** + ③**파견인력**으로 **총 30여명 규모** 구성

< 단계별 인력 충원 계획(안) >

1단계(개소 시)	⇨	**2단계**('25~)
■ **ETRI 인력 10명 + α**(他기관* 인력 파견)		■ **ETRI 인력 10명 + 신규 20여명**(점진적 충원)

* 한국정보통신기술협회(TTA), 정보통신정책연구원(KISDI) 등

3 기반

○ **(장소) 우수인력 채용, AI기업**과 **유관연구기관** 등과의 **협력용이성**을 고려하여, **판교 글로벌 R&D 센터***(4층, 332평)에 설치

* AI분야 연구소 집적(동일건물 內 ETRI 수도권연구본부, SW정책연구소, KETI 입주)

○ **(법적근거) 지속적**이고 **체계적**인 **AI안전연구소 운영**을 위해 국회에서 논의 중인 「**AI기본법**」에 **연구소 운영근거*** **마련** 추진

* **조문(안)** : **과기정통부장관**은 AI와 관련하여 발생할 수 있는 위험으로부터 국민의 생명·신체·재산 등을 보호하고 AI사회의 신뢰 기반을 유지하기 위한 상태를 확보하기 위한 업무를 전문적이고 효율적으로 수행하기 위하여 **AI안전연구소를 운영할 수 있다.**

3 비전 및 주요기능

◈ **(비전) AI 안전성**을 **평가·연구**하는 **전담조직**으로, **아태지역**을 대표하는 **글로벌 AI 안전 거점연구소**(허브) **구현**

◈ **(미션) ❶AI안전에 대한 과학적 이해 증진, ❷AI안전정책 고도화 및 안전제도 확립 지원, ❸국내 AI기업의 안전 확보 지원**

1 AI안전 평가

○ **(위험 정의) 글로벌 논의 내용***을 토대로 **국가 차원**에서 **집중 관리**해야 할 **AI위험**을 세부적으로 **정의**

* ①화학 또는 생물학 무기의 개발, 생산, 획득을 지원할 수 있는 잠재적 모델
②안전장치 우회, 조작 및 기만, 인간의 명시적 승인이나 허가없이 수행되는 자율적 복제 등 인간의 감독을 회피할 수 있는 잠재적 모델

○ **(안전 평가) 기업, 대학·연구기관**과 **협력**하여 위험별 **AI안전 평가 프레임워크**(지표·기준·방법) **개발, 안전 평가** 및 **위험완화 방안** 마련

○ **(평가인프라 구축)** AI안전 **평가데이터셋 구축, 평가도구 개발** 등 **평가 인프라**를 **구축**하여, 안전 평가 시 활용 및 기업 활용 지원

2 AI안전 정책 연구

○ **(안전정책 연구)** 주요국·국제기구 **AI안전 규범·거버넌스 분석, AI윤리·리터러시** 등 **인문·사회학적 연구**를 통해 **안전정책 수립** 지원

○ **(국가안보 위험대응) 국가안보 측면**에서 국내·외 **AI위험 정보 수집·분석,** AI로 인한 **국가인프라 취약점 분석** 및 **대응방안** 마련

○ **(안전 컨설팅 지원) AI안전 관련**(국제평가·인증, 해외규제 등) **기업 컨설팅 지원**

○ **(AI기술 영향평가) AI기술**이 **경제·사회·일자리·환경** 등에 미치는 **영향**을 **사전**에 **평가**하고, 분야별 AI정책 수립 시 활용 지원

○ **(안전정보 구축)** AI 모델·시스템의 **역량·위험,** AI**안전 사고, AI 위험 식별·평가·완화 모범사례** 등에 관한 정보 구축·공유

3 AI안전 대내·외 협력

○ (국내 안전허브 역할 수행) AI안전 기술·정책 **정보공유** 및 **국내 기업·대학·연구기관** 간 **인력, 기술·인프라 협력 플랫폼** 역할 수행

- **AI안전연구소, 대학 간 파트너십** 등을 통한 **AI안전 분야 인력양성**

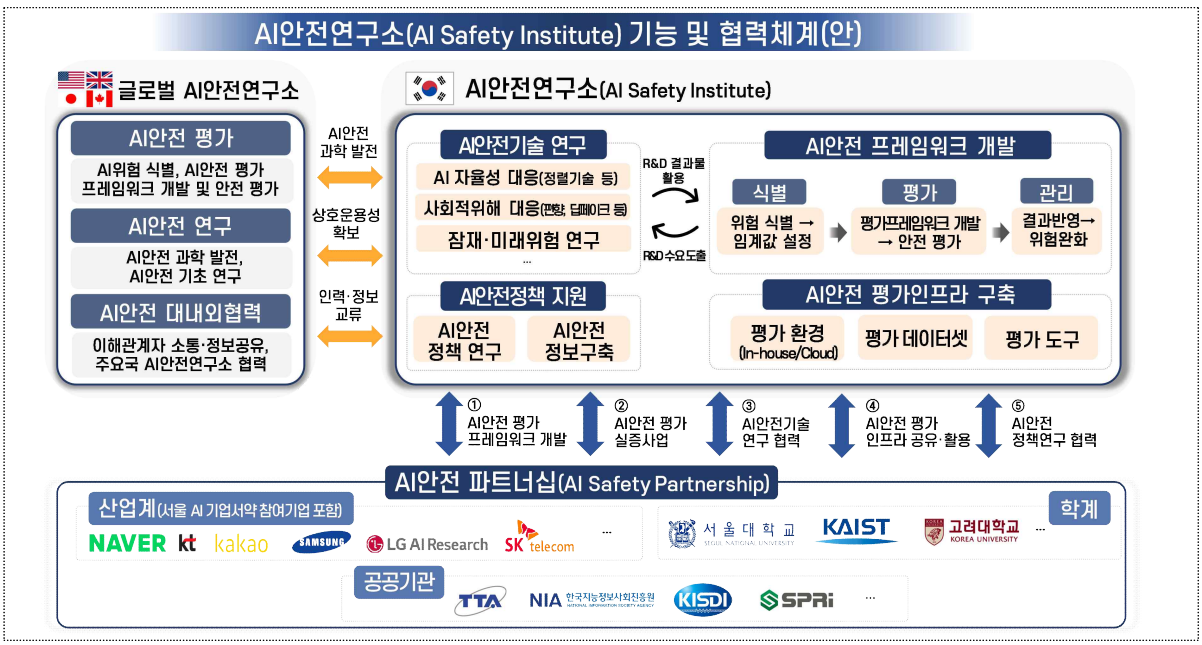
○ (국제적 연대·협력) **주요국 AI안전연구소, 국제기구** 등과 긴밀한 **협력체계**를 마련하여, **AI안전 확보**를 위한 **글로벌 거버넌스 구축**

- **AI안전 평가프레임워크 상호 운용성 확보,** AI안전 확보의 **모범 사례 공유,** AI모델의 **성능·위험**에 대한 **정보공유** 등 추진

4 AI안전 기술 연구

○ (실존위험 대응) 고도로 발전된 **AI의 통제력 상실**에 **대응**하는 **정렬기술**(Alignment), **사회적 차별** 및 **편향 완화·제거 기술** 개발

○ (잠재위험 대응) **미지의 위험**(Unknown Risk) **발굴 기법** 연구, **미래 AI**(AGI 등) **위험 예측**(시나리오 분석 등)에 대한 **선제적 대응**연구



4 향후 계획 : AI안전연구소 출범('24.11.), '국제 AI안전연구소 네트워크*' 행사 참석('24.11.20~21., 美 SF)

* 주요국 AI안전연구소 참여, 3개 기술트랙(합성콘텐츠, 기반모델 평가, 위험평가 방안) 논의

※ 여백 페이지

본 문

※ 여백 페이지

목 차

※ 여백 페이지

Ⅰ. 설립배경

◈ **'AI서울정상회의'**('24.5.) **후속조치**로 **AI안전성**을 **평가·연구**하고 **주요국 AI안전연구소**와 **협력**을 **전담하는 조직**으로 '**AI안전연구소' 설립** 추진

"**대한민국도 AI안전연구소 설립**을 **추진**해 **글로벌 AI 안전성 강화를 위한 네트워크에 동참**하겠다"(**윤석열 대통령**, **'AI 서울 정상회의**', '24.5.21.)

◇ 고도화 · 확산되고 있는 AI로 인한 위험에 대응 필요

□ 최근, **AI기술**이 빠르게 **발전·확산** 중이나, ①AI의 **기술적 한계**, ②인간의 **AI기술 오용**, ③**AI 자율성 확대** 등으로 **실존·잠재위험 확대**

* ①환각(Hallucination), 편향성 등 / ②유해정보 활용(화학·바이오 무기 개발 등), 사이버 해킹, 가짜뉴스 배포 등 / ③인간 통제력 상실 등

○ 이러한 **AI위험**은 **국민 기본권, 국가 안보** 및 **사회 안전**과 **직결**되므로, **국가 차원**의 **AI 안전성 확보노력 필요**

* "AI 전문가들이 AI가 핵전쟁과 맞먹는 실존적 위협이라고 선언"(UN 사무총장, '23.6.)

◇ 주요국을 중심으로 AI안전 전담조직 운영 본격화

□ 지난 '**AI 안전성 정상회의**(英, '23.11.)'를 계기로, **주요국**(英·美·日 등)은 **정부·공공기관**內 **AI안전연구소**를 **운영**하며 **AI위험**에 **본격 대비** 中

◇ AI안전 확보는 지속가능한 AI 발전과 AI 경쟁력의 전제

□ 우리도 **AI안전연구소**를 **설립·운영**하여 **안전한 AI** 개발·활용을 **확산**하고, **국내 AI기업**의 **경쟁력 확보** 및 **글로벌 진출**을 적극 뒷받침* 필요

* 자국 AI안전연구소가 없는 경우, 글로벌 진출을 추진하는 국내 AI기업(첨단 AI개발)이 해외 AI안전연구소 안전성 평가를 받아야 하는 제약 有(평가 장시간 소요, 기술유출 우려 등)

○ 아울러 **AI안전**에 대한 **국제적 연대 강화**와 **규범 정립**을 **수행**하고, 중장기적으로 **세계적 AI안전연구**를 **선도**하는 **기관**으로 **발전** 추진

참고 설립 추진경과

1 **'AI안전연구소 설립'** 발표('24.5.21., 대통령, 'AI 서울정상회의')

"영국, 미국을 비롯한 **주요국들의 AI안전연구소 설립 노력을 환영**한다. **대한민국도 AI안전연구소 설립**을 **추진**해 **글로벌 AI 안전성 강화를 위한 네트워크에 동참**하겠다"

2 **주요국**(英·美·日) **AI안전연구소 현장방문 및 동향분석**('24.5.~6.)

○ **연구소 성격·역할, 안전 평가방식**, 조직 운영방안 등 **심층분석**

英('24.5.31.)

美('24.6.11.)

日('24.6.14.)

3 **ICT 기관 'AI안전연구소 설립·운영계획'** 의견수렴('24.6.26.)

○ 기관별(ETRI·KISDI·NIPA·NIA) **연구소 설립의견 수렴·논의**

4 **'AI안전연구소 설립자문위원회'** 구성·운영
(1차: '24.7.3 / 2차: '24.7.17)

○ 산·학·연 전문가(15명) 참여, AI안전연구소 **역할·기능, 추진전략, 설치기관** 등 논의

5 **'AI안전연구소 설치·운영계획'** NST 이사회 의결('24.8.12.)

○ **(단기) ETRI 지역조직**(판교)으로 설치 → **(장기) 독립기관化** 추진

6 **'AI안전연구소 설립준비위원회'** 구성·운영('24.9.13.~)

○ **AI안전연구소 세부기능, 조직 운영방안** 등 논의

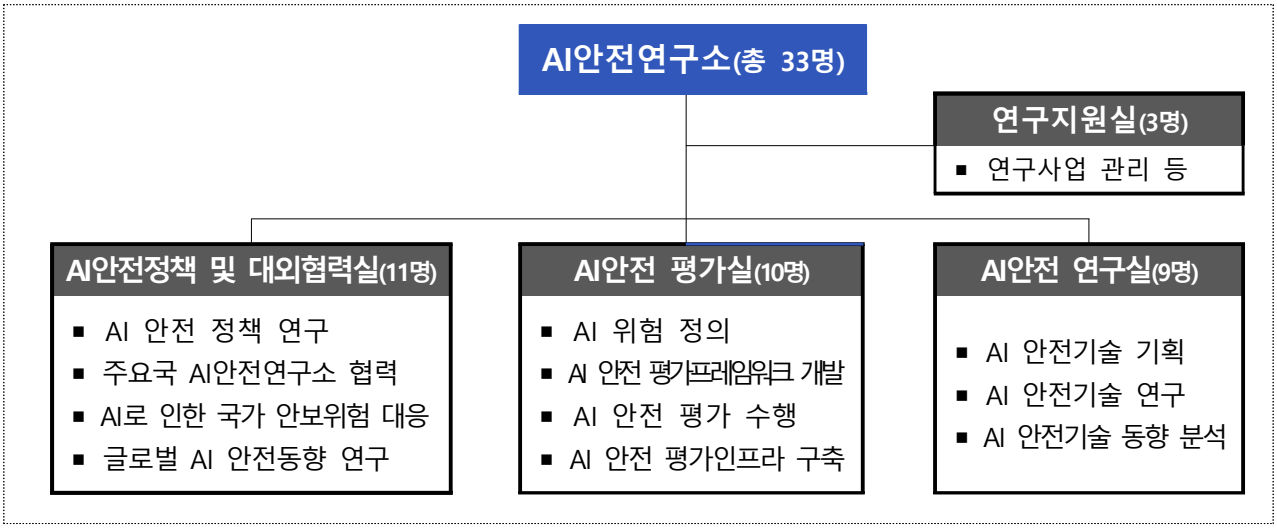
Ⅱ. 설립방안

1 조직

□ **(연구소명) AI안전연구소**(AI Safety Institute) * 한국전자통신연구원(ETRI) 內 설치

□ **(구성) 연구소장 + 3실 편제**(사업관리를 지원하는 연구지원실 별도)

○ ❶AI안전정책 및 대외협력실, ❷AI안전 평가실, ❸AI안전 연구실



□ **(운영) AI안전연구소 전문성·대표성·중요성** 등을 **고려, 연구소** 운영의 **자율성·독립성**을 최대한 보장하고, **별도 수당체계** 도입

< 'AI안전연구소 설치·운영 규정'('24.8.22.) >

- **제5조(소장)** 소장은 이 규정 및 연구원 원장이 위임한 범위 내에서 **AI안전연구소를 대표하여 소관업무**를 **수행**하고, **소속 직원을 지휘·감독**하며 **운영에 관하여 책임**을 진다.
- **제10조(조직 및 운영)** 소장은 법령 및 원규의 범위 내에서 **AI안전연구소 조직의 운영**에 있어 **최대한의 자율성과 독립성**을 갖는다.
- **제16조(기타)** 소장은 필요한 경우 위임된 범위 내에서 **AI안전연구소 운영에 대한 세부사항을 별도로 정하여 시행**할 수 있다.

2 인력

□ **(소장)** AI분야 정책·기술 전문성과 리더십, 국제적 역량이 뛰어난 외부전문가를 초대소장으로 채용('24.11.)

< AI안전연구소장 채용(안) >

- **(임명/임기) ETRI 원장**이 **임명 / 임기 3년**(임용일로부터 3년) * 1회에 한해 연임 가능
- **(처우)** 경력, 동종업계 처우수준 등을 종합적으로 고려, **경쟁력있는 연봉 수준**

□ **(직원)** ①신규채용 + ②ETRI인력 + ③파견인력으로 총 30여명 규모 구성

< 단계별 인력 충원 계획(안) >

1단계(개소 시)		**2단계**('25~)
■ **ETRI 인력 10명 + α**(他기관* 인력 파견)	⇨	■ **ETRI 인력 10명 + 신규 20여명**(점진적 충원)

* 한국정보통신기술협회(TTA), 정보통신정책연구원(KISDI) 등

3 기반

□ **(장소)** 우수인력 채용, AI기업과 유관연구기관 등과의 **소통·협력 용이성**을 고려하여, **판교 글로벌 R&D 센터***(4층, 332평)에 **설치**

* AI분야 연구소 집적(동일건물 內 ETRI 수도권연구본부, SW정책연구소, KETI 입주)

판교 글로벌 R&D센터

□ **(법적 근거)** 지속적이고 체계적인 AI안전연구소 운영을 위해 국회에서 논의 중인 「AI 기본법」에 **연구소 운영근거 마련** 추진

< AI안전연구소 운영관련 조문(안) >

- **과기정통부장관**은 AI와 관련하여 발생할 수 있는 위험으로부터 국민의 생명·신체·재산 등을 보호하고 AI사회의 신뢰 기반을 유지하기 위한 상태를 확보하기 위한 업무를 전문적이고 효율적으로 수행하기 위하여 **AI안전연구소를 운영할 수 있다.**

※ 국회 「AI기본법」 제정 논의 과정에서 AI안전연구소 관련조문 보완·구체화

참고 주요국(英·美·日) AI안전연구소(AISI) 개요

구분	영국('23.11~)	미국('24.2~)	일본('24.2~)	캐나다(설립前)
성격	정부기관	정부기관	공공기관	정부기관
소속	과학혁신기술부 內 설치	상무부 국립표준기술연구소(NIST) 內 설치	경제산업성 산하 정보처리추진기구(IPA*) 內 설치 * 디지털 전환, IT 보안·인증 등 담당	혁신과학경제개발부 內 설치 * 올해 하반기 설립 예정
미션	**첨단 AI 시스템**의 **안전성**에 대한 **경험적 이해**를 **확보**	**AI 안전 과학**을 **정의하고 발전**	**안전하고 신뢰**할 수 있는 **AI 실현**	**AI 기술의 안전한 개발·배포 촉진**
기능	① 고급 **AI 시스템 테스트**, 정책입안자에게 **위험성 전달** ② **기업, 정부, 연구 커뮤니티** 간 **협력 촉진** ③ **글로벌 차원**의 **AI 개발관행과 정책 강화**	① AI 모델·시스템, 에이전트 **테스트·평가·검증 연구** ② AI **안전사례 개발·전파** ③ AI **안전 관련기관, 커뮤니티 조정 지원**	① AI **안전성 평가**에 관한 **조사·기준** 검토 ② AI **안전성 평가 실시방법** 검토 ③ **타국 관계기관**(英美 AISI 등)과 **국제협력**	① 외부 연구기관 독립적 AI 안전 연구 지원 ② AI **안전 평가 프레임워크 개발**
조직	**핵심기술팀**, 지원팀, **인재·운영팀**, **전략팀**, **테스팅팀**, **국제협력팀**, 프로토콜팀, 안전성팀	**기술부서**(AI 머신러닝, 바이오·사이버 전문가 등), **행정부서**(공보, 정책, 지원 등)	**기획팀, 프레임워크팀, 기술팀, 보안팀, 표준팀**	-
소장	**Oliver Ilott** * 前 총리 국내 민정수석실(domestic private office) 리드	**Elizabeth Kelly** * 前 대통령 경제정책 특보	**무라카미 아키코**(비상근) * 現 손해보험재팬 CDO	-
인력	**80명 규모** (기술전문가 30명) * **100~150명 기술전문가 필요**	**12명** * **올해 말 25명 목표** * 기존 NIST인력(30명) 협업 中	**23명** * **내년 50명 목표**	**5~8명**(행정직) * 프로젝트 추진에 필요한 외부인력 단기 파견·계약
예산	**초기 2년**('23~'25) **1억 파운드** (약 1,750억원)	**1,000만 달러** * **내년 4,700만달러** 확보 목표	- * 올해 예산은 없으며, 타기관 예산 활용 * **내년 100억엔** 확보 목표	**5년간 5,000만 달러** **(약 500억원)** 목표

Ⅲ. 비전 및 주요기능

◈ **(비전) AI 안전성**을 **평가·연구**하는 **전담조직**으로, **아태지역**을 대표하는 **글로벌 AI 안전 거점연구소**(허브) **구현**

◈ **(미션) ❶AI안전에 대한 과학적 이해 증진, ❷AI안전정책 고도화 및 안전제도 확립 지원, ❸국내 AI기업의 안전 확보 지원**

※ **AI안전연구소의 기능**을 **주요국 AI안전연구소 기능**과 **'AI 서울 정상회의' 합의사항**(서울선언, 장관성명), **국내전문가 의견**을 **종합 분석**하여 **도출**

1 AI안전 평가

□ **(위험 정의)** 글로벌 논의 내용*을 토대로 **국가 차원**에서 **집중 관리**해야 할 **AI위험**을 **세부적**으로 **정의**

* ①화학 또는 생물학 무기의 개발, 생산, 획득을 지원할 수 있는 잠재적 모델
②안전장치 우회, 조작 및 기만, 인간의 명시적 승인이나 허가없이 수행되는 자율적 복제 등 인간의 감독을 회피할 수 있는 잠재적 모델(장관성명)

○ **위험 분석방법론**(위험 영향평가 등) **개발·활용**을 통한 **AI위험**의 **우선순위·허용범위**(임계값) 등 **설정**

* 프론티어 AI 위험이 수용할 수 없는 것으로 간주되는 임계값 파악(장관성명)

□ **(안전 평가)** 기업, 대학·연구기관과 **협력**하여 **위험별 AI안전 평가 프레임워크**(지표·기준·방법) **개발** 및 **안전 평가, 위험완화 방안 마련**

* 프론티어 AI가 심각한 위험을 초래할 수 있는 경우, 해당 모델에 대한 신뢰할 수 있는 외부평가를 촉진하는 데 있어 우리 역할을 인식(장관성명)

* AI안전 과학을 발전시키고 특정위험 관련 더 많은 실증 데이터 수집(장관성명)

□ **(평가인프라 구축)** AI안전 **평가데이터셋 구축, 평가도구 개발** 등 **평가 인프라**를 **구축**하여, 안전 평가 시 활용 및 기업 활용 지원

* AI 안전 과학 발전을 목적으로 하는 공유된 기술적 자원의 구축(서울선언 부속서)

2 AI안전 정책 연구

□ **(안전정책 연구)** 주요국·국제기구 **AI안전 규범·거버넌스 분석, AI윤리·리터러시** 등 **인문·사회학적 연구**를 통해 **안전정책 수립** 지원

○ **'국가AI위원회' 안전·신뢰 분과**(정책 수립·조정, 제도개선 등), **'국제 AI 정상회의'**(국제합의문 작성·검토 등)에 대한 **정책·기술적 지원**

○ **AI안전**에 대한 **이해도 증진, AI안전 과학**에 기반한 **AI안전정책 고도화**를 위해 **AI 안전 과학 보고서** 등 **백서 발간**

□ **(안전 컨설팅 지원) AI안전 관련**(국제평가·인증, 해외규제 등) **기업 컨설팅 지원**

□ **(AI기술 영향평가)** AI기술이 **경제·사회·일자리·환경** 등에 미치는 **영향**을 **사전**에 **평가**하고, 분야별 **AI정책 수립** 시 **활용 지원**

□ **(안전정보 구축)** AI 모델·시스템의 **역량·위험**, AI안전 **사고, AI 위험 식별·평가·완화 모범사례** 등에 관한 정보 구축·공유

* AI 안전의 과학적 이해를 장려하기 위한 조치에는 AI 모델의 능력, 한계, 위험 정보 공유, AI 위험 및 안전 사건 모니터링 등 포함(서울선언 부속서)

□ **(국가안보 위험대응)** 국가안보 **측면**에서 **국내·외** AI위험 **정보 수집·분석,** AI로 인한 **국가인프라 취약점 분석** 및 **대응방안** 마련

* 예시 : 화학·생물학 무기 개발, 사이버 공격, 군사분야 AI 활용 등

○ 국가 정보·안보 유관기관과 **안보관련 AI위험정보 협력체계** 구축

3 AI안전 대내·외 협력

□ **(국내 안전허브 역할 수행)** AI안전 기술·정책 **정보공유** 및 **국내기업·대학·연구기관** 간 **인력, 기술·인프라 협력 플랫폼** 역할 수행

○ **산·학·연 참여 'AI안전 파트너십'**을 구축하여 **안전기술·정책 연구 협력**

○ **AI안전연구소, 대학 간 파트너십** 등을 통한 **안전분야 인력양성**

□ **(국제적 연대·협력)** 주요국 AI안전연구소, 국제기구 등과 긴밀한 **협력체계**를 마련하여, **AI안전 확보**를 위한 **글로벌 거버넌스 구축**

○ **AI안전 평가프레임워크 상호 운용성 확보,** AI안전 확보의 **모범사례 공유,** AI모델의 **성능·위험**에 대한 **정보공유** 등 추진

* AI 안전평가·지침 역량에 관한 상호 강화, 국제표준 수립 및 채택 촉진(서울선언 부속서)
* 우리는 AI 안전연구소를 통해 모범사례, 평가 데이터셋 등을 공유(장관성명)

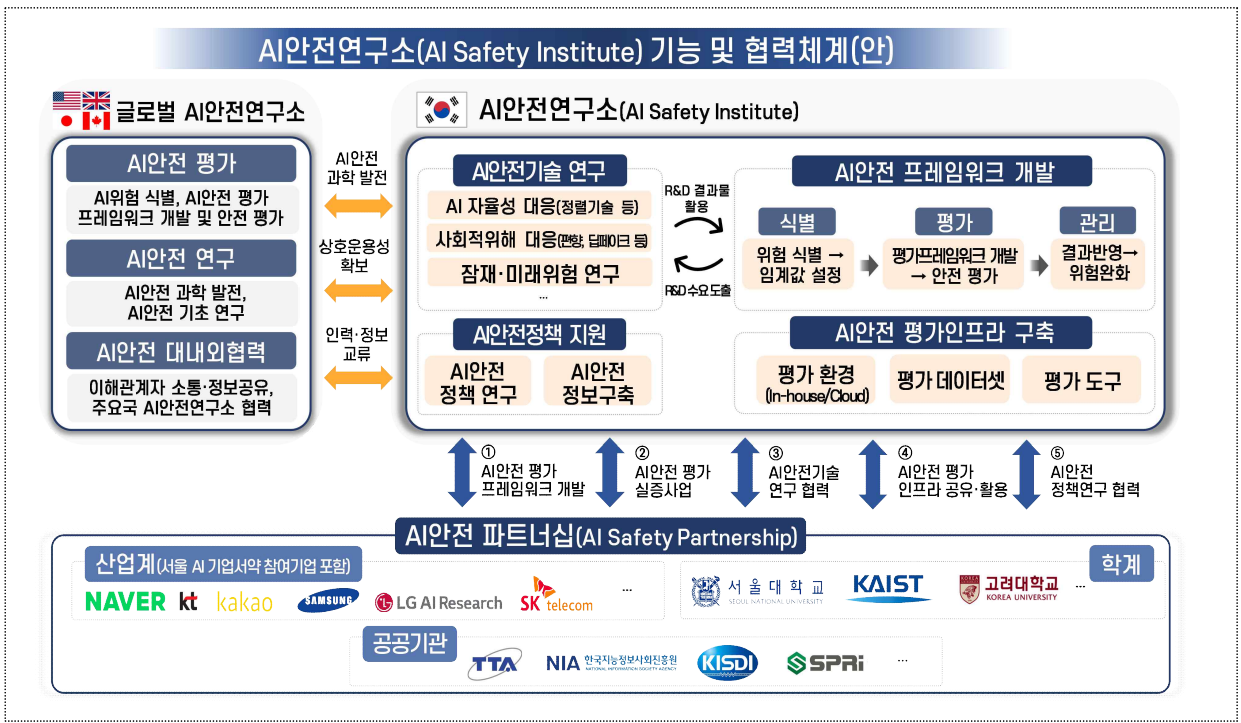
4 AI안전 기술 연구

□ **(실존위험 대응)** 고도로 발전된 **AI의 통제력 상실**에 **대응**하는 **정렬기술**(Alignment), **사회적 차별** 및 **편향 완화·제거 기술** 개발

* 인간의 명시적 승인이나 허가없이 수행되는 자율적 복제 등 인간의 감독을 회피할 수 있는 잠재적 모델의 성능으로 인해 심각한 위험이 발생할 수 있음을 인식(장관성명)

○ **AI 모델**의 **보안·강건성** 강화를 위한 **기술, 딥페이크 탐지기술** 등 연구

□ **(잠재위험 대응)** **미지**의 **위험**(Unknown Risk) **발굴 기법** 연구, **미래 AI**(AGI 등) **위험 예측**(시나리오 분석 등)에 대한 **선제적 대응연구**



Ⅳ. 향후계획

□ ‘AI안전연구소 설립준비위원회’ 운영(~‘24.10.)

* 설립준비위원회 발족 및 1차회의(‘24.9.13.)
* 정부, 산·학·연 전문가 참여하여, ‘AI안전연구소 세부 운영계획’ 수립

□ AI안전연구소 신임소장 임명(‘24.11.)

□ AI안전연구소 출범(‘24.11.)

* AI안전연구소 비전 선포 및 세부 운영계획 발표

□ ’국제 AI안전연구소 네트워크’ 행사 참석(‘24.11.20.~21., 美 SF)

* 11개 주요국 AI안전연구소(또는 상응기관)가 참여하여, AI안전 관련 3가지 핵심 기술이슈(합성콘텐츠, 기반모델 평가, 위험평가 방안) 논의