한국 AI안전연구소
: AI안전컨소시엄 정책 및 평가 분과 회의

2025. 4.22

ETRI AISI AI Safety Institute

진행 순서

- AI안전연구소 업무 현황 소개
- AI안전컨소시엄 안내
- 분과 의제
- 기타 의제 제안 및 제언
- 향후 일정 안내

AI안전연구소 '25년도 주요 활동 목표

법·제도 정책 및 대외

과학 기반의 AI 안전 정책 지원과,
국내외 법·제도 및 글로벌 규범 논의와 선도를 위한 기반 구축

- ☞ AI기본법 제 32조 가이드라인 및 하위법령 제정 지원
- ☞ EU AI Act 행동 강령(Code of Practice) 안내서 마련
- ☞ AI 위험지도 개발
- ☞ 첨단 AI 위험관리 방법에 관한 가이드라인
- ☞ 해외 AI안전연구소 및 주요 기관과의 협력 강화 및 확대
- ☞ 첨단 AI 안전평가 및 테스팅 분야 사실 표준 대응
- ☞ AI안전컨소시엄 운영
- ☞ AI안전연구소 운영 및 자문위원회 구성/운영

AI안전연구소 '25년도 주요 활동 목표

평가 기법 개발 및 실증

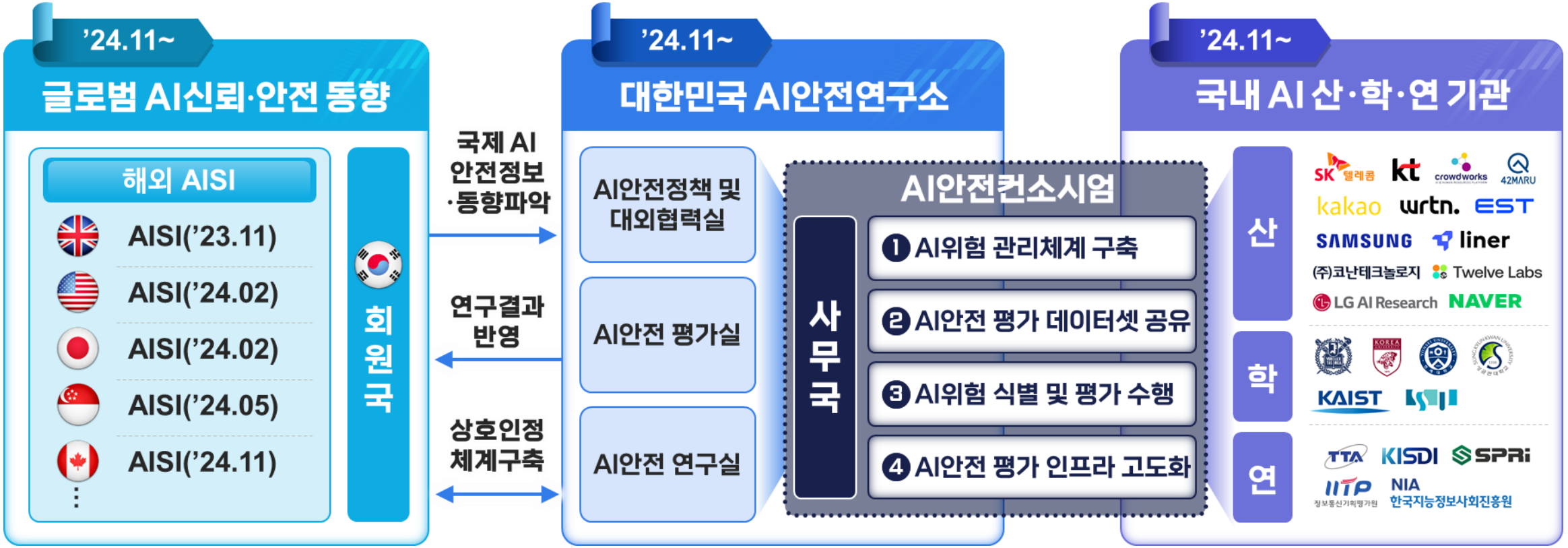
첨단 AI 대상의 안전 확보를 목표로,
안전성 평가를 위한 방법론, 데이터셋, 도구 등의 기술 개발

- ☞ 첨단 AI 평가 기법 및 안전 사례 선행 연구
- ☞ AI 에이전트의 안전성 평가 기술 개발
- ☞ 첨단 AI 안전성 평가를 위한 벤치마크 데이터셋 개발
- ☞ 국산 첨단 AI 시스템에 대한 안전성 평가 수행
- ☞ 첨단 AI의 위험 발굴을 위한 국제 레드티밍 행사 개최
- ☞ 첨단 AI의 안전성 평가를 위한 프롬프트 상시 수집
- ☞ 평가 신속 대응 및 기밀 보장을 위한 컴퓨팅 인프라 구축

AI안전컨소시엄 소개: 개요

AISI를 매개점으로

국내 산학연 24개 기업 및 기관 참여,
AI안전연구소 업무현황과 현안을 공유하고, AI안전 연구 협력과 담론 형성



AI안전컨소시엄 소개: 분과별 소개

정책 분과

국내외 법, 정책 및 거버넌스에 대한 대응과, 적극적인 글로벌 협력 전략 마련, 이를 위하여 우리나라가 최우선적으로 대응해야 할 위험의 정의

삼성전자
이스트소프트
퓨튼테크놀로지스
라이너
서울대학교
연세대학교
소프트웨어정책연구소
정보통신정책연구원
한국지능정보사회진흥원
네이버
케이티
포티투마루
고려대학교
성균관대학교
한국정보통신기술협회

⟨ 정책 분과 Agenda ⟩

구분	안건	성과
~6월 단기	AI 위험지도 개발 및 검토	AI위험 정의 및 지도 개발
~8월 중기	EU GPAI CoP 안내서 제작 참여 및 검토	CoP 안내서
수시	글로벌 AI안전 정책 및 규범 의견수렴	AI안전컨소시엄 의견서

AI안전컨소시엄 Agenda 소개

정책 분과

AI 위험지도 개발 현황, EU AI Act CoP 안내서 작업 방향

1. 인공지능위험 지도(AI Risk Map)

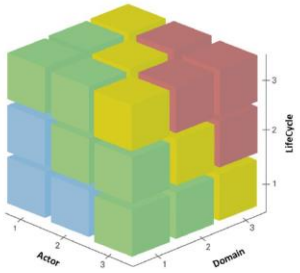
Korea AISI Risk Map

❒ AI 안전 관련 대응 체계 구축을 위한 AI위험 지도 마련

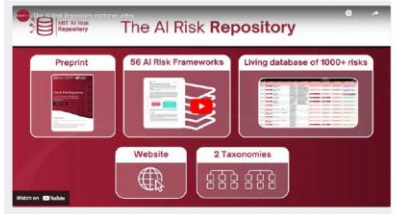
- 위험 유형별 안전, 보안, 국가 안보의 개념 정립 등 역할 구분 체계화
- AI 위험 식별 및 평가 데이터 연계를 통한 위험 대응 지도 마련 목표

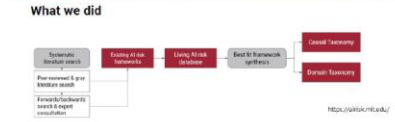
❒ 벤치마크 연계·국제협력을 통한 AI 위험 완화 및 안전성 확보를 위한 종합적인 안내 지도 구축

- 국제 주요 연구기관 및 기업과의 협력을 통한 글로벌 기준 반영
- 주요 3개의 축을 기반으로 AI 위험에 누가, 어느 단계에 관여하며 어떻게 영향을 미치는지 파악 가능하도록 구성 중
- 벤치마크 데이터와 연계하여 정보를 제공할 수 있는 지도 구축



〈참고 대상〉 MIT AI Risk Repository





2. 인공지능 규범 대응 안내서

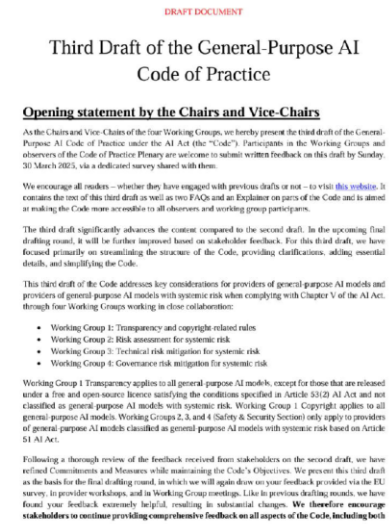
국내외 AI 안전 규제 대응을 위한 기업 실무 안내서

❒ 글로벌 AI 안전 관련 규제 대응을 위한 「국제 AI안전 규범 안내서」제작

- 유럽, 일본 등 주요국 AI 규제에 대한 정합성 분석 기반의 기업용 안내서 제공
- EU AI법, GPAI 실천강령(General Purpose AI Code of Practice) 안내서 발간 목표

❒ AI기본법 제 32조 안전성 확보 의무 이행을 위한 가이드 제공

- 기업의 법적 의무이행 대상 여부 판단, 결과 제출 등의 안내를 위한 가이드라인
- 프론티어 AI 기준, 성능 확인 예시 등에 관한 가이드라인 갱신 포함
- AI 안전사고 신고 절차 안내 및 대응 사례 제공 등



DRAFT DOCUMENT

Third Draft of the General-Purpose AI Code of Practice

Opening statement by the Chairs and Vice-Chairs

As the Chairs and Vice-Chairs of the four Working Groups, we hereby present the third draft of the General-Purpose AI Code of Practice under the AI Act (the "Code"). Participants in the Working Groups and observers of the Code of Practice Plenary are welcome to submit written feedback on this draft by Sunday, 30 March 2025, via a dedicated survey shared with them.

We encourage all readers – whether they have engaged with previous drafts or not – to visit this website. It contains the text of this third draft as well as two FAQs and an Explainer on parts of the Code and is aimed at making the Code more accessible to all observers and working group participants.

The third draft significantly advances the content compared to the second draft. In the upcoming final drafting round, it will be further improved based on stakeholder feedback. For this third draft, we have focused primarily on streamlining the structure of the Code, providing clarifications, adding essential details, and simplifying the Code.

This third draft of the Code addresses key considerations for providers of general-purpose AI models and providers of general-purpose AI models with systemic risk when complying with Chapter V of the AI Act, through four Working Groups working in close collaboration:

- Working Group 1: Transparency and copyright related rules
- Working Group 2: Risk assessment for systemic risk
- Working Group 3: Technical risk mitigation for systemic risk
- Working Group 4: Governance risk mitigation for systemic risk

Working Group 1 Transparency applies to all general-purpose AI models, except for those that are released under a free and open-source licence satisfying the conditions specified in Article 53(2) AI Act and not classified as general-purpose AI models with systemic risk. Working Group 1 Copyright applies to all general-purpose AI models. Working Groups 2, 3, and 4 (Safety & Security Section) only apply to providers of general-purpose AI models classified as general-purpose AI models with systemic risk based on Article 51 AI Act.

Following a thorough review of the feedback received from stakeholders on the second draft, we have refined Commitments and Measures while maintaining the Code's Objectives. We present this third draft as the basis for the final drafting round, in which we will again draw on your feedback provided via the EU survey, in provider workshops, and in Working Group meetings. Like in previous drafting rounds, we have found your feedback extremely helpful, resulting in substantial changes. **We therefore encourage stakeholders to continue providing comprehensive feedback on all aspects of the Code, including both**

AI안전컨소시엄 소개: 분과별 소개

평가 분과

**국내 상황에 맞는 위험 요소에 대한 평가 연구와 함께,
AI안전연구소 네트워크를 비롯, 국제사회와의 논의를 통한 글로벌 호환성**

카카오
네이버
엘지에이아이연구원
케이티
에스케이텔레콤
코난테크놀로지
포티투마루
크라우드웍스
고려대학교
성균관대학교
한국정보통신기술협회
뤼튼테크놀로지스
서울대학교
소프트웨어정책연구소
정보통신정책연구원
한국지능정보사회진흥원

〈 평가 분과 Agenda 〉

구분	안건	성과
~6월 단기	국제AI안전보고서 집필	한글버전 보고서
~11월 중기	안전성 평가 관점 수립 및 검토	AI안전성 평가 로드맵
장기	국제 공동 평가 수행	국제AISI네트워크 Joint Testing 보고서

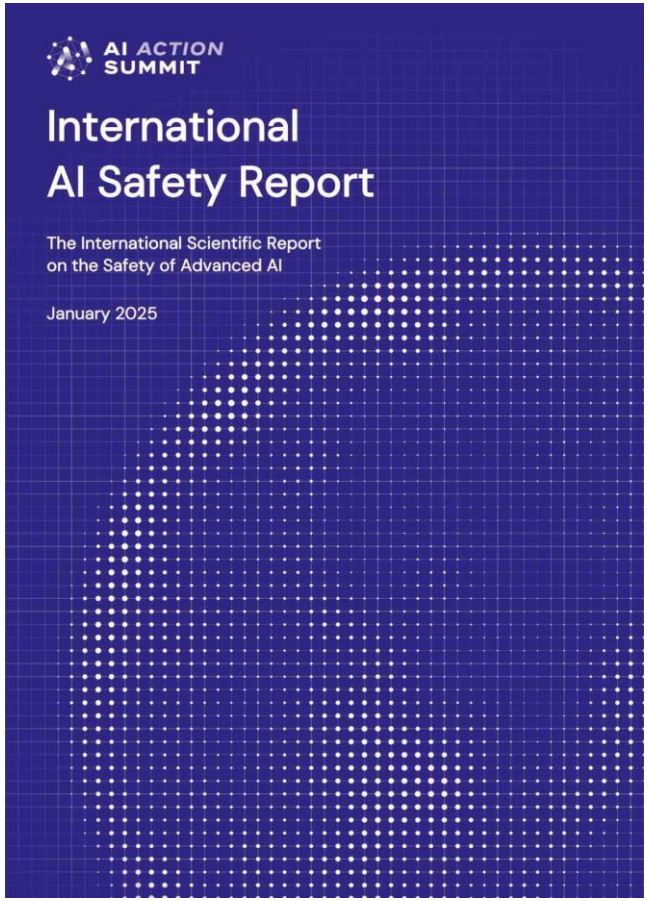
AI안전컨소시엄 Agenda 소개

평가 분과

국제AI안전보고서 한글버전 집필 및 국제AISI네트워크 Joint Testing 개요

1. 국제AI안전보고서 한글버전 집필

- International AI Safety Report
  - 의장: 요슈아 벤지오
  - 전문가 자문단: 전세계 30여개국 96명 전문가 (한국: 이경무 교수)
  - 2023년 영국 블레츨리 선언에 따라 작성 시작
  - 2025년 2월 파리AI행동정상회의에서 v1.0 발표
    - 한국 버전 협력 제안
- 주요 내용
  - GPAI의 세 가지 위험
    - 악의적 사용, 기술적 결함, 시스템적 위험
  - 위험 완화 방안
  - 2025년 초, Agentic AI에 대한 이슈 추가
    - 기존 위험 + 자율성, 통제 불능, Hijacking, AI협업 등



AI안전컨소시엄 Agenda 소개

평가 분과

국제AI안전보고서 한글버전 집필 및
국제AISI네트워크 Joint Testing 개요

2. 국제 AISI 네트워크 공동 테스트

- International Network of AI Safety Institutes
  - 출범일: 2024.11.20 (미국 샌프란시스코 회의)
  - 의장국: 미국
  - 참여국(10개국): 한국, 영국, 미국, 일본, 싱가포르, EU, 프랑스, 케냐, 캐나다, 호주
- 목표
  - AI안전에 대한 위험과 그 대응 방안에 대해 과학적인 공동 이해 구축
- 구성
  - Track 1: 합성 콘텐츠 위험 대응
  - Track 2: AI 시스템 테스팅
  - Track 3: 첨단 AI 시스템의 위험 평가
- 변화
  - 미국 행정부 변화, 영국 AISI 명칭 변경
  - 2월 파리 AI행동 정상회의 이후 7월까지 의장단 미정

AI안전컨소시엄 Agenda 소개

평가 분과

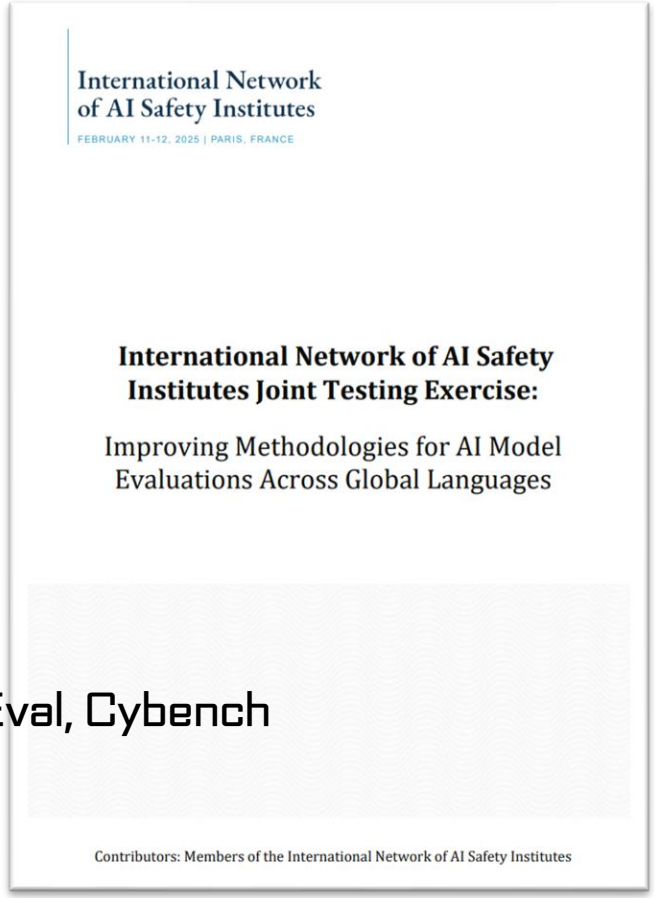
국제AI안전보고서 한글버전 집필 및
국제AISI네트워크 Joint Testing 개요

2. 국제 AISI 네트워크 공동 테스트

- 개요
  - 주도국: US, UK, SG, JP (2024.11)
  - 목적: AI 시스템 테스트를 위한 최선의 방법론 도출
  - 평가도구: Moonshot(SG), Inspect(UK)

- 1차 조인트 테스팅
  - 대상모델: LLaMA 3.1 405B
  - 벤치마크: GSM8K, SQuAD 2.0, MMMLU
  - 결과: 다국어 테스팅 가능성 확인, 방법론 일치의 필요성

- 2차 조인트 테스팅
  - 대상모델: Mistral Large, Gemma 2 (27B)
  - 벤치마크: AILuminate (MLC), AnswerCarefully (다문화/일본), CyberSecEval, Cybench
  - 결과
    - 다국어 테스팅 실행 – 10개 언어
    - 채점 과정의 본질적인 주관성과 번역의 불완전성
    - 자동 평가의 한계



AI안전컨소시엄 Agenda 소개

평가 분과

국제AI안전보고서 한글버전 집필 및
국제AISI네트워크 Joint Testing 개요

2. 국제 AISI 네트워크 공동 테스트

- 3차 조인트 테스팅
  - 보다 체계적인 접근: 위험 우선순위, 테스트 설계 및 Threat modelling
  - 기간: 4월 ~ 7월 (ICML)
  - 후보 주제
    - 다국어/다문화, 유해성/편향
    - 사이버보안
    - 에이전트 테스팅 – multistep reasoning

공지 및 향후 일정

결과 공개

분과 회의 결과 및 관련 자료는 모두 AISI 홈페이지에 게시
https://aisi.re.kr

멤버 추가

AI안전컨소시엄 멤버 추가 가입의 경우,
별도의 요건 혹은 기존 멤버의 동의 여부에 대한 정책 마련 수요 조사

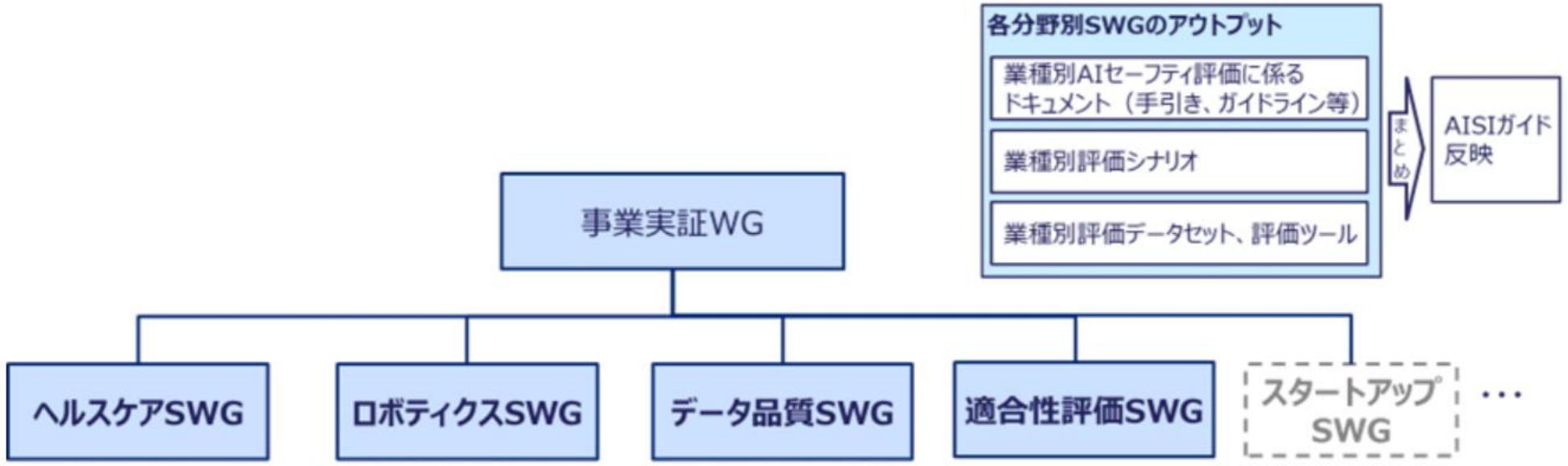
2차 회의

차기 회의는 6월 중순(합동, 오프라인)으로 예정,
회의 후 작업 참여 및 상황을 봐서 이후 일정 확정 및 공지

[참고1] 글로벌 AI안전연구소의 컨소시엄 및 파트너십 활동

일본 AISI

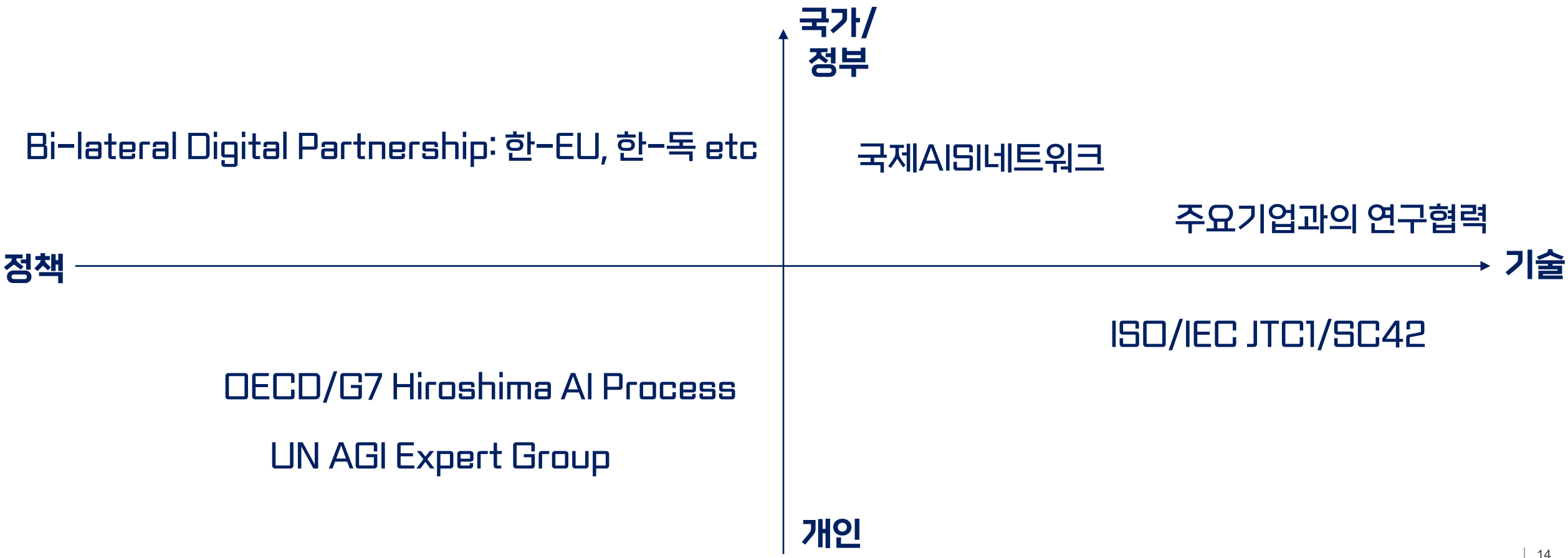
일본 AISI는 파트너십을 체결한 산학연 기업 및 기관과
연구협력을 통해, 각 산업 분야별 안전평가 실증 사업을 추진('25.3.24.)



[참고2] 글로벌 AI 거버넌스 및 협력 추진 현황

글로벌 참여

글로벌 AI 거버넌스 형성과 국제표준화 등
AI안전 분야 협력 활동 활발



Creating a Sustainable Future with Safe AI

Thank you for your time and attention

AISI AI Safety Institute