

과기정통부, '인공지능 서울정상회의' 핵심의제인 '국제 인공지능 안전연구소 관계망(네트워크)' 참여

- 미국 샌프란시스코에서 주요 10개국(한·미·영·유럽연합 등) 인공지능 안전연구소가 참여하는
'국제 인공지능 안전연구소 관계망(네트워크)' 공식 출범(24.11.20.)으로 국제 인공지능 안전 연대 강화 -

과학기술정보통신부(장관 유상임, 이하 '과기정통부')는 11.20(수)~21(목) 미국 샌프란시스코(Goldengate Club)에서 열리는 '국제 인공지능(AI)안전연구소 교류(네트워크)'(International Network of AI Safety Institutes) 행사에 참석하였다.

이번 행사는 '인공지능(AI) 서울정상회의'('24.5)에서 발표된 '서울선언'의 핵심의제인 인공지능(AI) 안전연구소 설립과 안전한 인공지능(AI)에 대한 국제 협력을 이행하기 위한 후속조치*로 개최되었다. 지난 '인공지능(AI) 서울정상회의'에서 윤석열 대통령은 “대한민국도 인공지능(AI) 안전연구소 설립을 추진해 국제 인공지능(AI)안전성 강화를 위한 관계망에 동참할 것”이라 강조하였으며, 과기정통부는 이를 신속하게 이행하여 이번달 말 인공지능(AI) 안전연구소 출범을 앞두고 있다. 이번 '국제 인공지능(AI)안전연구소 교류 행사'에 지난주 임명된 인공지능(AI) 안전연구소 김명주 소장('24.11.12.)을 포함한 한국 대표단을 파견하여 다른 나라 인공지능(AI) 안전연구소와 긴밀한 협력관계를 구축하였다.

* ('서울선언'('24.5.21.)에서 발췌) “우리는 인공지능(AI)안전연구소, 연구 프로그램과 기타 관련기관을 설립하거나 확대하는 것을 지지하며, 이러한 기관들간의 관계망을 육성함으로써 안전 연구에 관한 협력을 증진하고 모범 관행을 공유하기 위한 협력을 추진한다.”

이번 행사에는 주최국 미국을 포함하여 10개국* 인공지능(AI) 안전연구소 (또는 상응하는 연구기관), 인공지능 안전분야 기술전문가 등이 참석한 가운데, ‘국제 인공지능(AI) 안전연구소 관계망’ 출범을 선언하고, 관계망 임무 및 운영체제 논의와 인공지능 안전 연구·시험·지침에 대한 기술적 논의가 진행되었다.

* 미국, 한국, 영국, 일본, 싱가포르, 캐나다, 프랑스, 유럽연합, 케냐, 호주

‘국제 인공지능(AI) 안전연구소 관계망’은 전세계의 기술전문성을 결집하여, 인공지능 안전 관련위험과 그 완화에 대한 공통의 과학적 이해를 창출하고, 국제적인 연구를 통해 상호 운용가능한 원칙, 모범사례 적용 지원을 목표로 ①연구(Research), ②실험(Testing), ③지침(Guidance), ④포용(Inclusion) 분야에 대한 글로벌 협력을 중점적으로 추진해 나갈 계획이다.

< ‘국제 인공지능(AI)안전연구소 관계망 임무 선언문’ 주요내용 >

- (배경) ‘인공지능(AI) 서울정상회의’(24.5.)에서 발표된 ‘서울선언문’을 기반으로 인공지능 안전 과학에 대한 새로운 국제 공조를 촉진하기 위함
- (주요 협력사항) 국제 인공지능 안전 과학에 중요한 4가지 분야(①연구, ②실험, ③지침, ④포용)에 대해 관계망 구성국 간 협력 추진
 - ①연구(Research) : 인공지능 안전 과학을 발전시키기 위해 공동 또는 상호 보완적인 연구를 수행하고, 첨단 인공지능 위험과 역량에 대한 국제적 합의를 촉진하는 연구 추구
 - ②실험(Testing) : 고급 인공지능 모델을 실험하기 위한 모범사례를 구축하고, 공동실험연습 수행과 평가결과 공유 등을 통해 회원국 간 평가결과 상호인정 추진
 - ③지침(Guidance) : 공통의 인공지능 안전 지침 마련을 촉진하고, 고급 인공지능 모형의 실험 해석에 대한 상호 운용가능한 접근방식을 공유
 - ④포용(Inclusion) : 관련정보, 기술도구를 공유하여 다양한 지역과 발전단계에 있는 국가, 이해관계자를 관계망에 참여시켜, 인공지능 안전 과학에 참여할 수 있는 역량을 강화

2일간(11.20~21.) 이어진 기술 분과에서는 인공지능 안전 관련 핵심 연구분야인 분과¹합성작품(콘텐츠) 위험 저감, 분과²인공지능 기반모형 실험, 분과³고급 인공지능 체계 위험평가에 대해 심도깊은 논의가 진행되었다. 1분과에서는 각국이 연구개발 중인 인공지능 합성작품의 투명성 확보기술(복제 방지 무늬, 탐지기술 등) 및 관련 모범사례를 공유하고, 향후 국제적 차원의 공동 연구주제 등을 논의하였다. 2분과에서는 인공지능 기반모형 실험 방법론도구를 논의하고, 실험 방법·결과에 대한 상호운용성 확보방안을 모색했으며, 3분과에서는 고급 인공지능 체계 실험 결과의 해석을 통한 인공지능 위험의 정량적 식별·평가방안을 논의하였다.

특히, 한국 대표단은 현장에서 ‘이미지 영상 조작(딥페이크) 탐지 및 생성억제 기술 연구(우사이먼성일 성균관대 교수), ‘인공지능 위험평가 및 검증체계(곽준호 한국정보통신기술협회 AI신뢰성정책팀장)’ 등 인공지능 안전 관련 국내 주요 연구 성과를 발표하여 국제적 차원의 연구 의제를 도출하고, 주요국 인공지능(AI) 안전연구소와 연구협력을 구체화하는 계기를 마련하였다.

‘국제 인공지능(AI) 안전연구소 관계망’의 지속적·체계적인 운영을 위해 관계망 협치체계(네트워크 거버넌스) 운영계획에 대한 논의도 진행되었는데, 한국 대표단은 ‘인공지능(AI) 서울정상회의’의 성공적 개최경험, 한국 인공지능(AI)안전연구소 출범('24.11)과 정책적 지원의지를 강조하고, 관계망 운영 구조(네트워크 거버넌스)에서 주요역할을 수행하고자 하는 우리정부의 적극적인 의지를 밝혔다.

과기정통부 송상훈 정보통신정책실장은 “인공지능 안전에 대한 국제 연대는 우리정부가 주도하여 정립했던 지난 ‘서울선언’(‘24.5.)의 핵심의제였다” 면서, “‘서울선언’의 핵심규범·가치가 주요국이 적극적으로 지지·참여하는 국제 인공지능(AI) 안전연구소 관계망 출범으로 신속하게 구체화되었다는 점에서 뜻이 깊다.”고 말했다.

또한, “우리정부도 이번달에 인공지능(AI) 안전연구소를 개소하여, 인공지능(AI) 안전연구를 본격 수행해 나가고, ‘국제 인공지능(AI) 안전연구소 관계망’에도 적극 참여하여 세계적 차원의 인공지능 안전 기술·규범 발전에 주도적인 역할을 하겠다”고 강조하였다.

담당 부서	인공지능기반정책관 인공지능기반확산팀	책임자	팀 장	송호준 (044-202-6280)
담당 부서	인공지능기반정책관 인공지능기반정책과	담당자	사무관	이재호 (044-202-6262)

내일을 만드는 과학기술
내일을 채우는 디지털·AI

더 아픈 환자에게 양보해 주셔서 감사합니다
가벼운 증상은 동네 병·의원으로

대한민국
지책브리핑

OPEN
공공누리 공공저작물 자유이용허락

□ 행사 개요

- (일시/장소) '24.11.20(수)~21(목) / 美 샌프란시스코(Goldengate Club)
- (참석대상) 10개국* AI안전연구소(또는 상응하는 연구기관)
 - * 미국, 한국, 영국, 일본, 싱가포르, 캐나다, 프랑스, EU, 케냐, 호주
- (주요내용) ‘국제 AI안전연구소 네트워크’ 출범 및 거버넌스 논의, AI안전 연구·테스팅·지침에 대한 기술적 논의*
 - * 3가지 기술주제(①합성콘텐츠 위험 저감, ②기반모델 테스트, ③고급 AI시스템 위험평가)에 대한 전문가 논의 진행

□ 세부계획

1일차(11/20(수))	
[오전] 언론 공개 및 모든 참석자	
9:00- 9:15	개회사
9:15-10:15	패널 강연1 "Supercharging safety"
10:15-10:30	휴식
10:30-11:30	패널 강연2 "Global Perspectives AI Safety"
11:30-12:00	작업 안내
12:00-12:30	점심식사
12:30-13:00	키노트
13:00-13:15	휴식
[오후] 언론 비공개 및 모든 참석자	
13:15-15:00	기술트랙 세션1
15:00-15:15	휴식
15:15-17:15	기술트랙 세션2
17:15-17:30	휴식
17:30-18:00	첫째날 마무리
18:00-	저녁식사

2일차(11/21(목))	
[오전] 언론 비공개 및 네트워크 멤버	
9:00- 9:10	환영사
9:10-10:40	기술트랙 세션3
10:40-10:50	휴식
10:50-12:00	성과 공유 및 질의응답
12:00-13:00	점심식사
13:00-13:45	네트워크 거버넌스 회의
13:45-14:00	폐회