과학기술정보통신부 보도자료 다시 대한민국! 새로운 국민의 나라

보도시점 2024. 11. 26.(화) 10:00 (2024. 11. 26.(화) 석간) 배포 2024. 11. 25.(월) 16:00

# 과기정통부, '인공지능 신뢰·안전성 학술회의' 개최, 책임있는 인공지능 위한 민·관 성과와 이행현황 공유

**- 세계적 인공지능 석학 요슈아 벤지오 교수, 인공지능 안전 확보를 위한 정책과학적 추진전략 기조강연**

**- '인공지능 서울기업서약'('24.5.) 참여 국내 6개사, 인공지능 안전 확보 자발적 이행현황 발표**

**- '생성형 인공지능 잠재적 위험성 확인 경진대회(레드팀 챌린지)'('24.4.) 분석결과 소개, 인공지능 안전 확보를 위한 업계활용 지원**

**- '제2회 인공지능 신뢰성 대상' 선정, 다비오(Eartheye 2.0), 액션파워(daglo) 등 총 8점 수상**

**과학기술정보통신부**(장관 유상임, 이하 '과기정통부')는 **11.26일(화) 피스앤파크 컨벤션**(서울 전쟁기념관)에서 국내를 대표하는 **인공지능 분야 대기업·신생기업, 연구자** 등 **200여명**이 **참여**하는 가운데, **'인공지능 신뢰·안전성 학술회의'**를 **개최**하였다.

< '인공지능 신뢰·안전성 학술회의' 행사개요 >

- **(주최/주관)** 과기정통부 / 한국정보통신기술협회(TTA), 정보통신정책연구원(KISDI), 전자신문
- **(기조강연)** 퀘벡 인공지능연구소 요슈아 벤지오 교수(국가인공지능위원회 국제 자문위원), 카이스트 오혜연 인공지능 연구원장(국가인공지능위원회 안전·신뢰분과장)
- **(주요참석기관)** 한국정보통신기술협회, 네이버, 카카오, SKT, KT, 삼성전자, LG AI연구원, 포티투마루, 튜닙, 셀렉트스타

그간 과기정통부는 '국가 인공지능 **윤리기준**'('20.12.)을 토대로, 제3의 전문기관을 통한 **민간자율 '**인공지능 **신뢰성 인증**'('23.11.~) 지원, **'생성형 인공지능 잠재적 위험성 취약점 발굴 경진대회(AI레드팀 챌린지)'** 개최('24.4.), 국제 인공지능 **규범가치**(안전·혁신·포용)에 대한 합의를 이끌어낸 '인공지능 **서울정상회의**'('24.5.)의 **성공적 개최**, **'국제** 인공지능 **안전연구소 협력망' 참여**('24.11.20~21.), 인공지능 **안전연구소 출범**('24.11.27. 예정) 등 **지속가능한** 인공지능 **혁신생태계 조성**을 뒷받침하기 위해 인공지능 **신뢰·안전성**에 대한 **정책적 지원**을 **강화**하고 있다.

이번 학술회의에서는 최근 국제적 차원에서 강조되고 있는 **인공지능 신뢰·안전성 관련 기술·정책 동향**을 **공유**하고, **'24년 정부지원 연구결과물**의 **주요 성과**와 **민간**의 **이행현황**을 **점검**하며, **국내 인공지능 신뢰·안전성 확보**의 **우수사례**를 **선정·확산**하기 위한 목적으로 개최되었다.

이번 학술회의 **행사**는 크게 **4가지 주제**로 진행되었다.

## ① 국내·외 인공지능 신뢰·안전 기술·정책 동향 공유

**인공지능 분야 세계적 석학**인 몬트리올대 **요슈아벤지오 교수**는 **기조강연**에서 "**인공지능안전 확보**를 위한 **정책**과 **과학의 중요성**을 강조하며, **최첨단 인공지능 모형**의 **위험**에 대한 효과적인 **관리**를 위해 **국내법-국제협약** 간의 **상호조화**가 **필요**하다"고 하였다. 또한, "**과학측면**에서는 **인공지능 모형의 정렬**과 **통제**가 중요쟁점으로, **정량적**으로 **측정가능**한 **인공지능 모형**의 **위험평가**, **위험관리기술 확보**를 위한 **정부지원 확대**와 **인공지능 안전연구소**의 **역할**이 필요하다"고 강조하였다.

이어진 **두 번째 기조강연**에서 한국과학기술원(KAIST) **오혜연 인공지능연구원장**은 국제사회의 **인공지능** 패권 경쟁 동향을 공유하고, **전략자산으로서 인공지능의 중요성**을 **강조**하였다. 아울러, **인공지능** 기술의 **언어·문화적 포용성, 격차 문제** 등 글로벌 사회에 미치는 영향을 소개하고, **인공지능 안전분야** 신**시장 창출**과 **기업 경쟁력 확보관점**에서 **인공지능 안전 정책**의 **추진전략**을 소개하였다.

최근, **국내업계**에서도 인공지능 신뢰·안전성 확보를 위한 **기술개발**과 **평가기법**, **데이터셋 구축** 등 **다양한 연구**가 진행되는 가운데, **튜닙 박규병 대표**는 '**인공지능 안전장치**(가드레일) **해결책**을 활용한 다양한 형태의 위험 탐지 및 대응방안'을 소개하였고, **포티투마루 김동환 대표**는 '**인공지능 신뢰성 인증**(한국정보통신기술협회) **획득** 등을 통한 **자사 생성형인공지능 모형(LLM) 신뢰·안전성 확보사례**'를 주제로 발표하였다.

### ② '인공지능 서울기업서약'('24.5.) 이행현황 발표(국내 6개사)

지난 5월에 개최된 **'인공지능 국제 토론회(AI글로벌포럼)'**에서 **국내·외 14개 기업***은 **'인공지능 서울기업서약'**에 참여하여 **책임있는 인공지능 개발·활용 보장** 등에 대한 **자발적 이행**을 **약속****한 바 있다.

* **(서약참여기업)** [국내] 삼성전자, NAVER, 카카오, SKT, KT, LG AI연구원
[국외] Google, OpenAI, Anthropic, IBM, Salesforce, Cohere, MS, Adobe

** **(서약주요내용)** "우리는 (i)취약성과 위험을 식별·평가하고, (ii)내부 평가결과를 인공지능 모형 수명주기에 통합하며, 적절한 경우 인공지능 안전연구소 같은 국내 정부기관의 평가와 같은 외부 평가 결과를 고려하고, (iii)인공지능 안전사고를 점검하고 대응하는 강력한 내부 협치 및 위험관리 정책을 마련할 것을 약속"

이날 학술회의에서는 '인공지능 **서울기업서약**'에 참여한 **국내 6개 기업**이 모두 참여한 가운데, **안전하고 신뢰할 수 있는 인공지능을 개발**하기 위한 **위험 관리방안 수립**, 기술 연구, **내부 협치체계 마련** 등 **각사의 이행현황***을 공유하였다.

* **(네이버)** '인공지능 안전성 체계(AI Safety Framework, ASF)' 수립('24.6.)
**(카카오)** 인공지능 위험관리체계(Kakao AI Safety Initiative) 구축·운영('24.10.~)
**(삼성전자)** 내장형 인공지능에 대한 신뢰·안전성 수립 및 운영('24.10.)
**(SKT)** 인공지능 협치체계(AI 거버넌스) 기본원칙을 구체화한 '인공지능 행동규범' 수립('24.10.)
**(KT)** 안전하고 신뢰할 수 있는 인공지능 서비스 개발을 위한 책임있는 인공지능 체계('Responsible AI프레임워크)' 수립('24.10.)
**(LG AI연구원)** 인공지능 위험을 사전에 파악하고 해결하는 '인공지능 윤리영향평가' 전사업 적용('24.3~)

앞으로도 6개 기업은 자사가 제공하는 **인공지능 제품·서비스**와 관련된 **책임**을 **깊이 인식**하고, 향후 **인공지능 신뢰·안전성 확보**를 위한 **노력**을 더욱 **강화**해 **나갈 것**을 약속하였다.

### ③ '생성형 인공지능 잠재적 위험성.취약점 발굴 경진대회(AI레드팀 챌린지)'('24.5.) 결과 소개

올해 4월, **인공지능 연구자·대학생·일반시민** 등 **1,000여명**이 **참여**하여 **국내 인공지능 기업 4개사**(네이버·SKT·업스테이지·포티투마루) **생성형 인공지능 모형**(LLM)을 대상으로 **잠재적 위험·취약점**(유해정보, 편견·차별 등)을 **발굴**하는 **'생성형 인공지능 취약 위험성 발굴 경진대회(레드팀 챌린지)'** 행사를 개최한 바 있다.

이날 학술회의에서 **한국정보통신기술협회**(TTA)는 대규모 참여자의 **공격 시도**를 **분석**하여 **식별된 7가지 분야*** **주요위험**과 **다양한 공격기법**(거부 무력화, 혼동유도)을 소개하였다.

* 잘못된 정보, 편견 및 차별, 불법 작품(콘텐츠) 생성 및 정보 제공, 탈옥, 사이버 공격, 개인의 권리 침해, 일관성

**과기정통부**는 **취약점·위험성 발굴 경진대회(레드팀) 챌린지 결과**를 **활용**하여 **생성형** 인공지능 **안전 체계**를 **구축**해 나가고, **국내기업**이 자사 **생성형 인공지능 모형**의 **잠재위험**을 **사전**에 **파악**하여 **신뢰·안전성**을 **확보**할 수 있도록 지원할 계획이다.

### ④ '제2회 인공지능 신뢰성 대상' 시상

**인공지능 신뢰성**의 **중요성**과 **인식**을 **업계 전반에 확산**하고, **우수 인공지능 제품·서비스**에 대한 **홍보**를 **지원**하기 위한 목적으로 **'제2회 인공지능 신뢰성 대상'** 시상식을 개최하였다.

**인공지능** 분야 **산·학·연 전문가**로 **구성**된 **심사위원회**(위원장 : 한국과학기술원<KAIST> 오혜연 인공지능 연구원장)는 접수된 총 41개의 인공지능 제품·서비스를 대상으로 ①**신뢰성·품질 이해** 및 방침 **적용**수준, ②**신뢰성·품질수준**, ③신뢰성·품질 **관리 우수성**, ④**시장성** 등을 종합적으로 평가(1차:11.4., 2차:11.19.)한 결과, **총 8개**의 인공지능 **제품·서비스***가 **선정**되었다.

* △**대상** 1개(과기정통부 장관상, 상금 1,000만원), △**최우수상** 1개(과기정통부 장관상, 상금 700만원), △**우수상** 6개(한국정보통신기술협회 등 기관장상, 각 상금 300만원)

**대상**은 다비오가 개발한 '**다비오 어스아이2.0***'가 수상했으며, '신뢰할 수 있는 인공지능 개발안내서'에 기반하여 **자체적인 데이터 품질관리 절차**를 **수립·적용**하여 **높은 평가**를 받았다.

* 위성, 영상 이미지의 공간정보 분석 등을 통해 시각화, 점검, 탐색 등 기능을 제공

< '제2회 인공지능 신뢰성 대상' 수상제품 >

| 상훈 | 회사명 | 제품명 |
|---|---|---|
| 대 상(장관상) | 다비오 | Dabeeo Eartheye 2.0 |
| 최우수상(장관상) | 액션파워 | daglo (다글로) |
| 우수상(한국정보통신기술협회장상) | 단감소프트 | CareerRecommender v1.0 |
| 우수상(한국정보통신기술협회장상) | 마크애니 | SmartEYE for Finder v1.0 |
| 우수상(정보통신정책연구원장상) | 셀렉트스타 | LLM 신뢰성 평가 서비스 |
| 우수상(정보통신정책연구원장상) | 튜닙 | AI가드레일 솔루션패키지 v1.0 |
| 우수상(전자신문 사장상) | 메디픽셀 | MPXA-2000B v2.0.0 |
| 우수상(전자신문 사장상) | 테스트웍스 | 블랙올리브 v1.0.0 |

**송상훈 과기정통부 정보통신정책실장**은 "정부가 지속적으로 추진해 온 인공지능 **신뢰·안전성 정책**의 **노력**에 **호응**하여, 최근 **산업계·학계**에서도 인공지능 **신뢰·안전성 전담조직 설치**와 **투자 확대** 등 **자발적인 노력**이 **확대**되고 있다."고 강조하며,

"**민간자율**에 기반한 **책임있는 인공지능 개발·활용 문화**가 **확산**될 수 있도록 **정책 지원**을 **강화**하는 한편, **첨단 인공지능으로** 인한 **잠재적 위험**에 대비하여 인공지능 안전연구소를 **출범**(11.27., 예정)하고 **국가차원**에서 **체계적으로 대응**해 나가겠다"고 강조하였다.

| 담당 부서 | 과학기술정보통신부<br>인공지능기반정책과 | 책임자 | 과 장 | 남철기 (044-202-6270) |
|---|---|---|---|---|
| | | 담당자 | 사무관 | 이재호 (044-202-6262) |

**참고** '인공지능 신뢰·안전성 학술회의' 행사포스터

제2회

# 인공지능 신뢰성 대상

2024. 11. 26. 화 10:00~17:00

피스앤파크컨벤션 3F 로얄홀 (서울 전쟁기념관)

# AI 신뢰·안전성 컨퍼런스

| | | |
|---|---|---|
| 10:00 ~ 10:15 (15분) | 개회식 | 개회사 손승현 회장 \| 한국정보통신기술협회, 김정연 원장직무대행 \| 정보통신정책연구원 |
| | | 축사 송상훈 정보통신정책실장 \| 과학기술정보통신부 |
| 10:15 ~ 10:45 (30분) | 시상식 | 제2회 인공지능 신뢰성 대상 시상식 및 기념촬영 |
| 10:45 ~ 11:00 (15분) | 해외 기조 강연 (녹화 버전) | AI 안전의 방향<br>Yoshua Bengio \| 퀘벡 AI연구소 교수, 국가AI위원회 글로벌 자문위원 |
| 11:00 ~ 11:30 (30분) | 국내 기조 강연 | 글로벌 AI생태계에서의 AI안전·신뢰 중요성<br>Alice Oh \| KAIST AI센터장 교수 |
| 11:30 ~ 13:00 (90분) | 오찬 | |
| | | 제1세션 "위험" |
| 13:00 ~ 13:30 (30분) | 발표 | 레드팀 챌린지 결과와 향후 계획<br>곽준호 팀장 \| 한국정보통신기술협회 |
| | | AI 신뢰·안전 확보를 위한 데이터셋<br>최호진 교수 \| KAIST |
| | | 제2세션 "행동" |
| 13:30 ~ 15:30 (120분) | 발표 | AI 서울 정상회의, 그 후<br>1.카카오 \| 2.LGAI연구원 \| 3.네이버 \| 4.KT \| 5.SKT<br>* 상기 5개 기업은 "Seoul AI Business Pledge"에 서약하였으며, 이 중 네이버는 "Frontier Safety Commitment"에도 서약 |
| 15:30 ~ 15:40 (10분) | 커피 브레이크 | |
| | | 제3세션 "기술" |
| 15:40 ~ 16:50 (70분) | 발표 | 1. 안전 확보를 돕는 AI 모델<br>박규병 대표 \| 튜닙 |
| | | 2. LLM 안전 및 신뢰 확보의 모범 사례<br>김동환 대표 \| 포티투마루 |
| | | 3. LLM 신뢰성 평가 프로세스 및 Use Case<br>김세엽 대표 \| 셀렉트스타 |
| | 토론 | 좌장 신준호 센터장 \| 한국정보통신기술협회 패널 제3세션 발표자 |
| 16:50 ~ 17:00 | 폐회 및 장내 정리 | |

주최 과학기술정보통신부 Ministry of Science and ICT

주관 TTA 한국정보통신기술협회 KISDI 정보통신정책연구원 전자신문