

(비공식 번역)

**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 전례없는 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 우선순위를 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 지지하는데 지속 집중한다. 우리는 프론티어 AI를 개발하고 배치하는 단체들의 특별한 책임을 인식하며, 이러한 측면에서 「프론티어 AI 안전 서약」을 환영한다.
4. 우리는 이 선언 참여국들이 AI 안전연구소, 연구 프로그램 그리고/또는 감독 기관들을 포함한 기타 유관 기관들을 설립하기 위해 진행하거나 계속 진행 중인 노력을 지지하고, 이러한 단체들간의 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·사용간의 선순환을 촉진하고, 사회·문화적, 언어적 그리고 성별 다양성을 증진하며, 상업적·공개적으로 사용 가능한 AI 시스템들의 전주기에 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 다중이해관계자간 적극적 협력 및 초국경적·학제간 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 공고히한 “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

(비공식 번역)

AI 안전 과학에 대한 국제 협력을 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 AI 안전연구소를 포함하는 공공 그리고/또는 정부 지원 기관을 설립하거나 확장하기 위한 공동의 노력을 격려한다.

2.1 우리는 AI 안전 관련 정책적 노력에 정보를 제공하기 위해 학제간의, 그리고 재현 가능한 증거 군집의 필요성을 인식한다. 우리는 궁극적으로 AI 개발 및 사용의 혜택이 전지구에 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 파트너들간에 국제 네트워크를 발전시킨다는 우리의 공유된 야심을 명시한다. 우리는 이러한 그리고 이와 관련된 노력에 있어서 향후 긴밀한 협력, 대화 및 파트너십을 기대한다.