SW·저작권 동향리포트

<제2024-24호> 2024년 12월 25일

정책/제도

국가별 AI안전연구소 추진 및 설립 현황

■ 개요

○ AI기술이 가진 잠재적 위험에 대한 우려가 고조됨에 따라 전 세계적으로 AI의 안전성을 확보하는데 집중하여 초점을 맞추고 있음. EU는 세계 최초로 인공지능법을 제정하고 미국은 AI의 안전한 개발을 유도하기 위해 행정명령을 발동하였으며 영국은 국제적으로 AI의 안전한 확보를 위해 인공지능 안전성 정상회의를 개최. 우리나라도 AI 서울 정상회의에서 AI 안전연구소 설립을 공식화하고 AI 안전성 확보를 위한 국제협력 의지를 표명하며 지난달 판교에서 한국 AI안전연구소를 출범

○ 향후 AI 안전성 테스트, AI 안전성 확보를 위한 원천기술 개발 및 정책연구 등 AI안전연구소의 역할이 더욱 중요해질 것으로 예상되므로 현재까지 주요 국가들의 AI안전연구소 추진 동향을 살펴보고 AI안전연구소의 올바른 활용 방안을 강구할 필요가 있음

[그림 1] 국가별 AI안전연구소 설립 현황

미국	2023년 11월	'AI 안전성 정상회의'서 AI안전연구소 설립 발표
	2024년 2월	대규모 AI안전연구소 컨소시엄(AISIC) 발족
영국	2023년 11월	'AI 안전성 정상회의'를 계기로 AI안전연구소 설립
	2024년 2월	AI 안전테스트 연구 초기결과 발표
일본	2024년 2월	AI안전연구소 설립
	2024년 4월	미국, 일본 AI 위험관리 프레임워크 상호운용성 개선 1차 크로스워크 결과 및 계획 발표
캐나다	2024년 4월	AI 사업 활성화와 안전을 위한 예산(안) 발표
		예산(안)에 AI안전연구소 설립 포함
한국	2024년 11~12월	한국전자통신연구원(ETRI) 산하에 AI안전연구소 설립 예정

*출처 : 이투데이(https://www.etoday.co.kr/news/view/2386896)

■ 주요 내용

1. 영국(DSIT AISI)

○ 영국 정부는 AI안전연구소 설립 이전에 프론티어 AI 테스크포스(Frontier AI Taskforce)를 구성하며 정부만이 AI에 대한 책임 있는 평가를 제공할 수 있다는 기조 하에 AI 안전연구, 평가, 발전을 위한 AI안전연구소를 설립. 특히 AI가 생화학 무기개발, 사이버 공격, 사기, 선전활동 등에 이용될 수 있고 최악의 경우에는 인류가 AI에 대한 통제력을 잃을 위험이 있음을 경고하며, AI안전연구소를 정부기관 간의 긴밀한 협력 및 소통, 국가 안보와의 연계성 등을 고려해 과학혁신기술부(Department for Science, Innovation and Technology, DSIT) 산하의 연구기관으로 설립

[그림 2] 영국 AI안전연구소의 목표와 임무

*출처 : 영국 AI안전연구소(https://www.aisi.gov.uk/)

○ 영국의 AI안전연구소는 AI 시스템의 안전 관련 기능 파악 및 시스템의 안전성과 보안을 이해하여 사회적 영향을 평가할 수 있는 시스템을 개발하고, 첨단 AI 시스템의 위험성에 대한 이해 증진 및 효과적인 AI 거버넌스에 필요한 기술도구를 개발하기 위해 다양한 탐색적 연구 프로젝트를 진행. 또한 관계자들 간의 원활한 정보 전달의 보장을 위해 정책입안자, 규제기관, 기업 및 국제파트너, 대중에게 연구결과와 관련된 정보를 제공함

○ 지난 5월에는 AI 모델의 안전 평가 플랫폼인 인스펙트(Inspect)를 공개하며 글로벌 안전 평가를 강화하고 가속화하여 다양한 그룹에서 AI 평가를 더 쉽게 개발할 수 있도록 지원. 인스펙트는 다양한 이해관계자에게 용이한 AI 안전 평가 환경을 제고함으로써 고품질의 안전성 평가 체계를 마련하고

산학연의 상호 조율이 가능할 것으로 기대

○ AI안전연구소는 인스펙트 평가대상으로 5개의 거대 언어모델(LLM)에 대해 사이버 공격, 화학 및 생물학적인 오용의 가능성, 자율성, 유해한 결과물의 도출 가능성 등을 각 모델에 대한 질문 또는 작업 프롬프트를 제공하고 응답을 측정하는 방식으로 평가하여 결과를 공개하였으며, 유해한 요청에 대한 준수(Compliance), 정확한 응답(Correctness), 작업의 완수(Completion)를 기준으로 측정

[표 1] 인스펙트의 LLM 평가 내용

분 야	평가내용	평가결과
사이버 공격	포렌식, 암호학, 리버스 엔지니어링 등 사이버 공격 대응을 위한 기본 작업 수행	대학 수준의 문제해결과 취약한 암호체계 악용에 관한 작업에는 한계가 존재
화학 및 생물학 평가	전문용어를 포함한 문답을 바탕으로 해당 분야 전문 지식을 평가	LLM은 전문가 수준의 지시를 제공하며, 일부 모델은 박사급 수준의 답변을 제시하는 수준
AI에이전트 평가	인간 감독을 배제한 코드실행, 탐색 등의 작업 수행 가능 여부 평가	일부 LLM은 단기과제의 해결은 가능하나 장기과제 수행에는 아직 어려움이 존재
안전장치	유해한 질문공격에 대한 답변을 회피하도록 학습된 정보를 유도하는 공격 시행	유해한 질문에는 답변을 준수한 반면 무해한 질문에는 규정을 준수하지 않는 답변 제공

*출처 : 영국 AI안전연구소(https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update)

2. 미국(USAISI)

○ 미국은 23년부터 AI 위험관리 프레임워크(AI Risk Management Framework) 1.0을 개발하고 관련 센터 및 워킹그룹의 설립을 추진해왔으며, 안전하고 신뢰할 수 있는 인공지능 개발 및 사용에 관한 행정명령에 근거해 AI 안전연구소를 설립. AI안전연구소는 연방 정부 기관인 상무부 산하 국립표준기술연구소(NIST)에 설치되었으며, 안전하고 신뢰할 수 있는 AI의 개발과 배포를 위해 민간 컨소시엄 구축을 통한 글로벌 표준 확립에 주력

○ 미국의 AI안전연구소는 AI 안전 기초 연구, AI 안전 테스트 프레임워크 개발, AI 안전 국제협력을 위한 역할을 수행하며, AI 정책에 대한 지침 및 표준 가이드라인을 개발하고 NIST 팀 및 광범위한 AI 커뮤니티와 협력하여 과학적 이해를 높이는 등 안전한 AI 혁신을 위한 기능을 수행

○ 생성형 AI의 안전한 개발 및 배포와 관련된 기술 표준의 수립을 위해 대규모 AI안전연구소 컨소시엄(Artificial Intelligence Safety Institute Consor

tium, AISIC)을 발족하고 공공과 민간의 협력 체계를 구축. 이와 함께 AI 안전 확보를 위해 5개의 요소를 중심으로 워킹그룹을 구성하고 실질적 연구 및 개발을 수행하여 안전성 평가 및 표준화를 위한 토대를 마련

* AISIC에는 200개 이상의 기업과 단체, 대학, 지자치단체가 참여하고 있으며 MS, 구글, 메타, 애플, 아마존, 엔비디아, IBM와 같은 빅테크 기업도 함께 참여

** 최근 미국 AI안전연구소는 오픈AI 및 앤스로픽 등과 AI안전 연구, 테스트 및 평가에 관한 협약(Agreements Regarding AI Safety Research, Testing and Evaluation)을 체결하였으며, 신규 AI 모델이 공개되기 전과 공개된 이후 AI 모델에 접근할 수 있는 프레임워크를 구축하고, AI의 기능, 안전 및 위험을 평가하는 방법과 이러한 위험을 완화하는 방법에 대한 공동연구를 실시할 계획

[표 2] 미국 AI안전연구소 컨소시엄의 워킹그룹 구성

구분	세부 기능
생성 AI 위험관리	- AI 위험관리 프레임워크 보완 자료 개발 침 운영 - 연방 기관 대상 최소 위험관리 지침 개발
합성 콘텐츠	- AI 생성 콘텐츠 인증 및 출처 추적을 위한 표준, 도구, 방법 연구 - 합성 콘텐츠 감지, 악용방지를 위한 연구
성능 평가	- 잠재적 위험 대응을 위한 영역의 AI 성능 평가 - 안전한 AI 개발 지원을 위한 테스트 환경 구축 및 도구 개발
레드티밍 (Red Teaming)	- AI 레드티밍 훈련 지침 마련: 다중 용도(dual-use) 기반 모델 개발자들의 안전하고 신뢰하는 시스템 구축을 위한 절차 및 프로세스 마련
안전 및 보안	- 다중용도 기반 모델의 안전 및 보안 관리 지침 조정 및 개발

*출처 : 한국지식재산연구원(https://www.kiip.re.kr/board/trend/view.do?bd_gb=trend&bd_cd=1&bd_item=0&po_item_gb=&po_no=22666)

3. 일본(IPA AISI)

○ 일본은 AI 안전성 정상회의, G7 히로시마 AI 프로세스를 통해 AI 안전연구소를 설립. 경제산업성 산하의 정보처리추진기구(Information Technology Promotion Agency, IPA)에 설치되었으며 전문 인재 확보와 함께 AI 시스템의 안전성을 연구하고 AI 기술이 초래할 위험 요소를 사전에 탐지하여 대응 방안을 마련하는데 주력

○ 일본은 AI안전연구소를 중심으로 미국, 영국, EU, 싱가포르 등 주요국과 AI 안전 관련 의견 교류, 국제 행사 및 정상회의 참석, 스탠포드 대학교 AI 심포지엄 참가 등 국제협력 활동을 활발하게 추진 중. 이와 함께 AI와 관련된 이해관계자가 AI의 위험을 정확하게 파악하고 대비할 수 있도록 AI 사업자 가이드라인을 발표하고, 미국의 NIST AI 리스크관리 프레임워크(RM

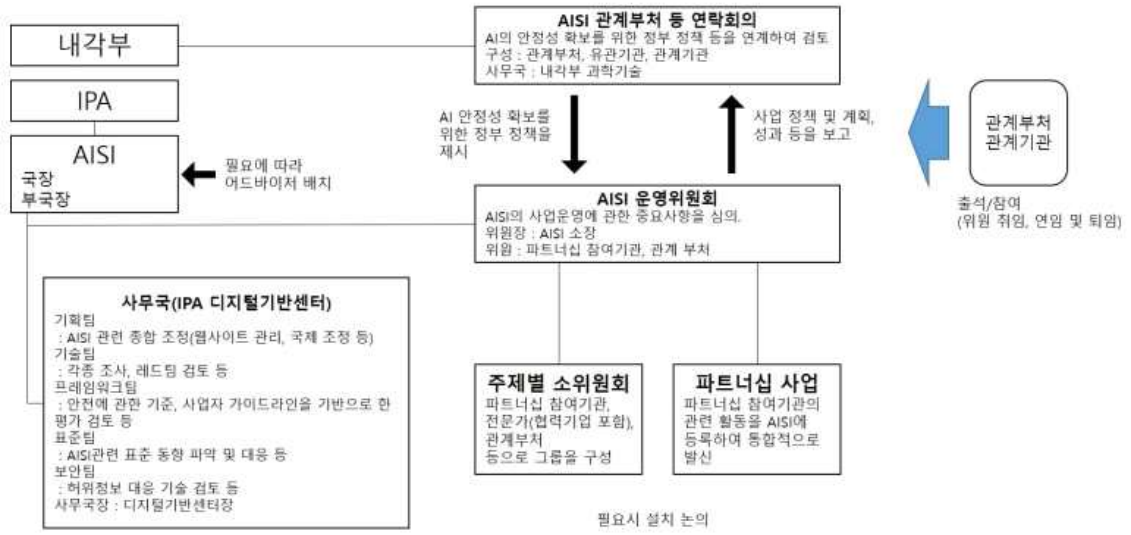
F)의 크로스 워크를 함께 실시

* 크로스 워크: 법령, 기준 및 프레임워크 등의 조항을 서브 카테고리에 매핑하는 것으로, 이를 토대로 조직이 활동이나 성과의 우선순의를 부여하고 준수하는 것이 용이하도록 작업

○ 일본 AI안전연구소는 AI의 안전에 관한 노력을 추진하기 위해서는 일본 국내 관계 기관과 연계하여 공동으로 대응해 나가는 것이 필요하다고 판단하여, 전문 인력을 확보하고 기술의 결집을 추진하는 등 AI 안전에 관한 지식을 보유한 국내 관계기관과의 협력 체계를 통해 필요한 사항을 규정하기 위한 목적으로 국무회의에서 결정된 통합혁신전략에 따라 AISI 파트너십 협정 시스템을 구축

* 파트너십의 내용으로는 ① AI 안전성에 관하여 AISI와 참여기관이 공동으로 실시하는 연구 및 조사, ② AISI가 실시하는 활동에 관한 참여기관에 의한 자문 제공, ③ 참여기관이 실시하는 AI 안전성 관련 활동에 대한 AISI 정보 제공, ④ AISI 및 참여기관에 의한 국내외 정보 전달, 국내외 관계기관과의 조정·연계 등의 활동을 추진함

[그림 3] 일본의 AI안전연구소 조직 구성



*출처 : 국가전략정보포털(https://nsp.nanet.go.kr/plan/subject/detail.do?nationalPlanControlNo=PLAN0000046361)

4. 한국(ETRI AISI)

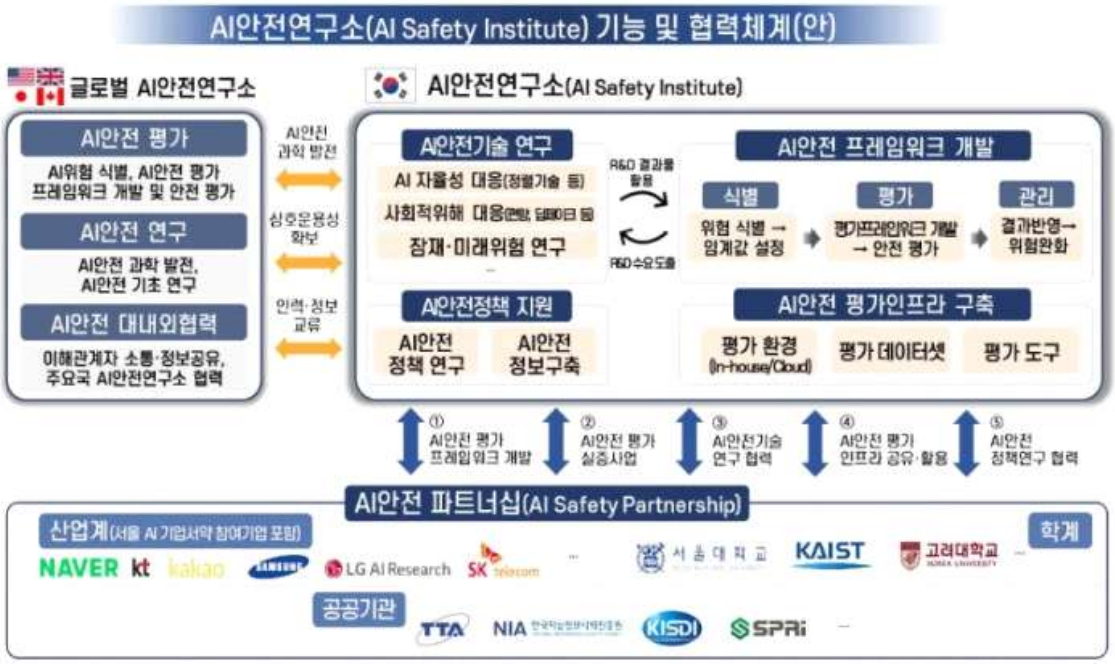
○ 우리 정부는 AI가 국가 경제와 안보를 좌우하는 핵심기술로 강조됨에 따라 그 동안 AI의 디지털 정책 방향을 설정하고 관련 계획을 수립 및 추진하였으며, 과기부는 제17차 정보통신전략위원회를 통해 AI안전연구소 개소를 위해 인력확보 · 핵심기능 · 주요일정에 대한 사안을 발표

* 우리 정부는 지난 5월 AI서울정상회의를 개최해 글로벌 AI 거버넌스 정상급 선언문인 서울선언을 도출하고, AI G3로 도약하기 위한 대통령 중심의 범국가적 혁신역량을 결집하는 민관 협업 기반 최고위 거버넌스인 국가인공지능위원회(Presidential Committee on AI)를 설립

○ 정부는 아태지역을 대표하는 글로벌 AI 안전 거점연구소 구현을 AI 안전연구소의 비전으로 설정하고, AI 안전에 대한 과학적 이해 증진, AI 안전 정책 고도화 및 안전제도 확립, 국내 AI 기업의 안전 확보 지원을 추진하는 것을 3대 핵심미션으로 발표. AI안전연구소는 한국정보통신연구원(ETRI) 소속으로 신규 설치되며 AI 안전정책, 평가, 기술 분야의 3개 연구실로 운영될 예정

○ AI안전연구소는 국가차원에서 관리해야 할 주요 AI 위험을 정의하고 AI 안전 정책을 연구하게 되며 AI 안전 정책을 고도화하여 대응 방안 연구를 추진할 계획. 또한 AI안전연구소를 중심으로 산학연이 참여하는 AI 안전 파트너십을 구축하고 AI 안전 정책·기술 분야 연구 협력을 촉진할 계획으로 글로벌 AI 안전 거버넌스의 주요 일원으로서 AI 안전 확보를 위해 주요국의 AI안전연구소 및 국제기구와 협력할 계획

[그림 4] AI안전연구소 설립 및 운영 계획



*출처 : 대한민국 정책브리핑(https://www.korea.kr/news/policyNewsView.do?newsId=148935251)

5. 그 외 국가

○ 캐나다는 지난 4월에 AI 사업 활성화와 안전을 위한 예산안에 AI 안전연구소의 설립을 위한 비용으로 5,000만 캐나다 달러를 편성하여 AI 안전성과 윤리 연구 인프라의 필요성을 강조. 또한 AI 산업 활성화를 통한 AI 경쟁우위 확보를 목표로 영국과 AI 안전을 위한 파트너십을 체결하여

정보와 자원의 교환을 통한 공동연구를 추진중

* 영국 AI안전연구소는 공동 연구를 위해 캐나다 AI안전연구소가 영국의 AI 자원에 접근할 수 있도록 하며, 향후 양국의 AI안전연구소는 미국 AI안전연구소와 함께 AI가 도입되고 있는 사회 시스템의 안전을 보호하는 시스템적 AI 안전 분야의 연구 프로그램에서 협력할 계획

○ 싱가포르는 지난 3월에 싱가포르 국립대학교((National University of Singapore, NUS)에 공익을 위한 AI 연구소를 설립하여 AI의 안전성, 신뢰성, 투명성, 책임성 연구와 규제 방안에 대한 연구를 추진. AI 연구소는 AI 기반 모델 및 산업융합 연구를 촉진하고 AI 인재를 육성하며 산학협력과 스타트업 생태계 활성화에 이바지하기 위한 목표로 설립되었으며 구글 클라우드와 IBM이 파트너로서 협력할 예정

○ 프랑스도 영국 AI안전연구소와 프랑스 컴퓨터 과학 및 자동화 연구소(INRIA) 간의 AI 안전을 위한 협력 파트너십을 발표하였으며, 국제 사회에서 새로운 AI 규범 수립을 주도하기 위해 내년 초 차기 AI 안전 정상회의인 AI 실천 정상회의(AI Action Summit)를 개최하기 위한 준비중

○ 호주는 국립AI센터가 고위험 AI 앱에 관한 규제, 국립 AI 센터를 통한 AI 안전 표준 개발 계획, AI 업계와의 협력 등의 계획을 발표. 사우디도 AI 연구 및 윤리를 위한 국제 센터(International Center for AI Research and Ethics)의 설립을 발표하고, 인도는 AI로 인한 사회문제 해결을 위한 국립 연구소의 설립을 논의하는 등 세계 각국은 책임 있는 인공지능 구현과 확산을 위해 국가별로 AI 안전 확보를 위한 전담조직을 준비중

[그림 5] 加 AI안전연구소의 출범 발표와 NUS AI 연구소의 연구팀

*출처 : 캐나다 한국일보(https://www.koreatimes.net/ArticleViewer/Article/163620), AppliedHE(https://news.appliedhe.com/category/research/)

■ 시사점

ㅇ 각 국의 AI안전연구소는 국가 간의 국제협력 지원, 정보 공유, 인력 교류, 공동 연구 등 AI 안전 관련 협력을 추진할 필요가 있으며, 국가별 인공지능 전략에 따른 AI 안전 정책 및 기술 개발의 구심점 역할을 할 수 있는 거버넌스 확립이 중요

ㅇ 현재 AI안전연구소를 설립하여 운영하고 있는 국가들은 AI 안전성 확보를 우선순위로 두고 운영 정책을 추진 중이며 주요 역할로서 AI의 잠재적 위험에 대한 대응방안 마련을 위한 국가 정책을 개발하고 AI 개발 전주기에 걸쳐 AI 안전성을 확보하기 위한 절차와 검토 사항을 점검하여 안전성 강화를 위한 프레임워크를 개발하는데 주력. 또한 핵심 기반 기술을 함께 연구하여 고도화되는 AI 모델의 잠재력 파악, 위험 식별, 편향성을 제거하는 작업을 함께 수행 중

ㅇ 우리 과기부도 지난 27일 판교 글로벌 R&D센터에서 AI안전연구소의 개소식을 통해 전 세계에서 6번째로 AI안전연구소를 출범. 이를 통해 AI 안전 분야의 전문 기술과 인력을 육성하고 AI 안전 관련 정책의 선진화를 추진할 계획이며, 산학연 기관이 AI 안전 분야 연구 협력과 정보 교류를 할 수 있는 우리나라 AI 안전연구 허브이자 국제 AI안전연구소 네트워크의 일원으로서의 역할을 수행할 것으로 기대

[그림 6] 한국 AI안전연구소 개소식

*출처 : K-공감 누리집(https://gonggam.korea.kr/newsContentView.es?mid=a10201000000&news_id=a2106f82-33fd-45e6-a9c5-9166471948dc)

작성자/문의	대외협력실 정책연구팀 강종균 과장(070-7709-3728)

참고자료

- 국가 안보 최전선 ‘AI 안전연구소’ ...新 AI 제국주의 경쟁 책임진다
https://www.etoday.co.kr/news/view/2386896

- 영국 정부, ‘AI 안전 연구소’ 설립
https://www.kiip.re.kr/board/trend/view.do?bd_gb=trend&bd_cd=1&bd_item=0&po_item_gb=&po_no=22433

- 영국의 「AI 안전연구소」 설립
https://nsp.nanet.go.kr/plan/subject/detail.do?nationalPlanControlNo=PLAN0000049040

- 美, 국립표준기술연구소(NIST) 내 미국 AI 안전 연구소(AISI) 설립
https://www.standard.go.kr/KSCI/pot/domeFore/detail.do?domeId=1131&domeFore=F

- 미국 AI 안전연구소, OpenAI 등과 ‘AI 안전 연구, 테스트 및 평가에 관한 협약’ 체결
https://www.kiip.re.kr/board/trend/view.do?bd_gb=trend&bd_cd=1&bd_item=0&po_item_gb=&po_no=23110

- 일본 인공지능 안전 연구소, 파트너십 협정 시스템 구축
https://www.kiip.re.kr/board/trend/view.do?bd_gb=trend&bd_cd=1&bd_item=0&po_item_gb=&po_no=23074

- 해외 AI 안전 연구소 핵심 기능 조사 용역 최종보고서
https://nsp.nanet.go.kr/plan/subject/detail.do?nationalPlanControlNo=PLAN0000046361

- 국내 클라우드시장 연 10조 원 시대 연다...3개년 기본계획 수립
https://www.korea.kr/news/policyNewsView.do?newsId=148935251

- “글로벌 AI 안전 주도” ...정부, AI안전연구소 운영 전략 공유
https://n.news.naver.com/mnews/article/092/0002349118

- [인터뷰] 김명주 AI안전연구소장
https://weekly.chosun.com/news/articleView.html?idxno=38544

- 과기부, 영국 AI 안전연구소 찾아 AI 안전성 협력 논의
https://www.yna.co.kr/view/AKR20240531139300017?input=1195m

- NUS sets up AI Institute to accelerate frontier AI research and boost real-world impact for public good
https://news.nus.edu.sg/nus-sets-up-ai-institute/

SPC 'ANGEL' 통계

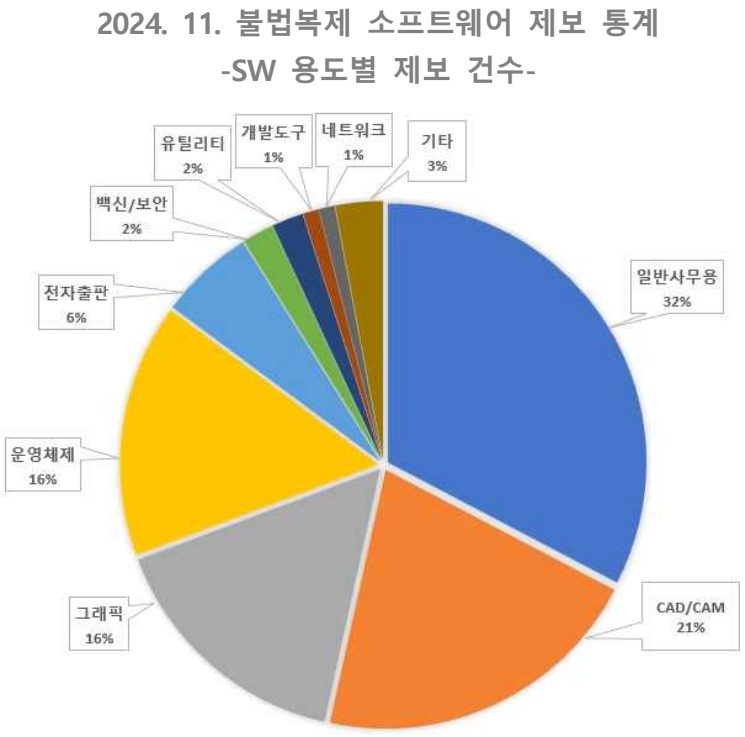
월 1회 제공

한국소프트웨어저작권협회, 불법복제 SW 제보 'ANGEL' 서비스 11월 통계 현황

○ 한국소프트웨어저작권협회(SPC)가 지난 11월 한 달간('24. 11. 1. ~ 11. 30.) 'ANGEL(불법제보)' 서비스를 분석한 결과, 기업 또는 개인의 불법복제 SW 사용 제보는 총 122건으로 나타남

○ SW 용도별로는 일반사무용 40건(33%), 설계(CAD/CAM) 26건(21%), 그래픽 20건(16%), 운영체제 19건(16%), 전자출판 7건(6%), 백신/보안 2건(2%), 유틸리티 2건(2%), 개발도구 1건(1%), 네트워크 1건(1%), 기타 4건(3%) 순으로 제보가 접수됨

[그림] SPC 'ANGEL(불법제보)' 서비스 2024년 11월 통계 현황



* 한국소프트웨어저작권협회는 2018년 11월부터 제보시스템과 제보 방식의 편의성을 개선한 불법복제 SW 제보 시스템 'ANGEL(불법제보)' 서비스를 운영하고 있음

다음 SW·저작권 동향리포트 〈제2025-1호〉 발간일은 1월 10일입니다.