SW·저작권 동향리포트

<제2024-22호> 2024년 11월 25일

정책/제도

AI가 초래하는 위험 요인과 유형

■ 개요

○ 지난해 챗GPT 출시 이후 기업들의 LLM 개발 경쟁을 거치며 AI 기술 발전이 가속화됨에 따라 특정 분야에서는 이미 인간의 능력을 뛰어넘으며 활용 범위가 확장되고 있음. AI 기술이 인간 생활에 다양한 이점을 제공하는 한편 새로운 위험에 대한 우려가 증가되며 신뢰성, 책임성, 윤리성 등의 논의와 함께 AI 안전의 확보를 위한 대응 마련이 시급

○ 각 국가의 정부와 기업 등 이해관계자들은 AI의 안전성을 확보하기 위해서 AI 안전연구소 등을 통해 위험을 식별하고 평가기준을 마련하면서 안전한 AI 개발과 배포를 위한 대응책을 마련하기 위해 노력중이며, 이에 AI로부터 발생되는 위험 요인과 유형을 살펴보고자 함

■ 주요 내용

1. 가짜 콘텐츠를 이용한 악용 가능성

○ 생성 AI로 생성된 이미지가 조작된 사건, 장소 또는 사물을 실제처럼 표현하여 실제 사건처럼 공유되는 등 텍스트는 물론 이미지, 오디오 및 비디오를 인간이 생성한 자료와 구분이 어려울 정도로 정교하게 생성하여 대규모로 유포. 실제 테러리스트 단체들이 생성 AI를 이용해 선전물을 대량으로 생성하고 있으며, 인종주의와 반유대주의적 표현을 사용해 AI로 네오나치 이미지를 생성하려 혼란과 갈등을 조장

* 대표적으로 이스라엘과 하마스의 갈등과 관련하여 아이 앞에 폭격된 집, 이스라엘 난민을 위한 임시 텐트, 팔레스타인을 지지하는 사진 등 허위 정보가 담긴 생성 AI 이미지가 확산되며 갈등을 고조시킨 사건이나, 생성 AI를 사용한 전화 로봇이 바이든 대통령을 사칭하여 유권자들에게 투표를 억제하는 전화 메시지를 전달한 사건이 있음

[그림 1] 생성 AI가 만든 이스라엘-하마스 전쟁의 가짜 이미지와 영상

*출처 : 동아일보(https://www.donga.com/news/Economy/article/all/20240813/126514798/2)

○ 특정 개인이나 그룹을 대상으로 맞춤 설계된 사기 콘텐츠를 대량으로 생산하여 피싱에 활용하는 문제점도 발생. 생성 AI로 인물의 실제 목소리나 외모를 모방해 인물을 사칭하여 이용자를 속이고, 실제 사람인 것처럼 가짜 신원을 생성하고 도용하여 불법적인 목적으로 사용하는 등 행동의 묘사와 유사성으로 청중을 쉽게 오도할 수 있어 높은 피해를 유발

○ 대부분의 국가들에서는 얼굴과 목소리를 AI로 변조하는 딥페이크와 딥보이스의 악용 사례가 증가하고 있으며, 이미 제작된 악용 AI 이미지와 영상은 회수가 불가하여 피해자들은 이로 인해 2차 피해를 입을 가능성이 존재. 마이크로소프트는 딥페이크를 구별할 수 있는 도구 개발과 함께 악용 AI로 인한 피해 방지를 위해 딥페이크 사기 법령의 제정을 미국 의회에 촉구하였으며, 국내에서는 카이스트가 딥러닝을 활용해 딥페이크 영상을 식별하는 기술을 개발하였으나 악의적인 AI를 조기에 발견하고 차단하는 기술 개발은 현재까지는 이루어지지 않고 있음

* 최근 우리나라에서도 텔레그램을 통해 학생 및 군인 대상의 단톡방에 지인을 딥페이크로 합성한 음란물과 함께 개인정보를 유출하는 사건이 발생되며 사회적으로 큰 파장을 일으켰으며, 이를 계기로 방송통신심의위원회는 텔레그램 딥페이크 성범죄영상물 대응과 관련하여 글로벌 협의체를 구축하고 구글, 유튜브, 페이스북, 트위터 등 글로벌 플랫폼 사업자들과 간담회를 진행

[그림 2] 딥페이크로 인한 성범죄 발생 현황

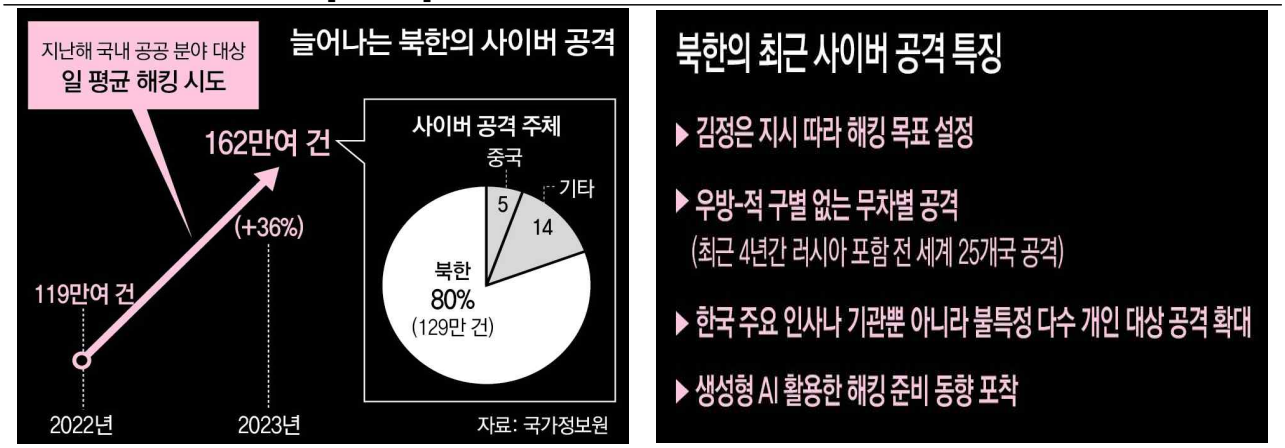
*출처 : 딥페이크 성범죄 실상 이야기(https://blog.naver.com/rensan1947/223564676928)

2. 사이버 공격에 의한 보안 위험

○ 시스템이 AI에 의해 사이버 보안 공격에 취약하도록 설계되어 사이버 공격에 악용될 우려가 있으며, AI를 코딩 보조 도구로 활용하게 되면 소프트웨어의 취약성이 발생하여 AI가 자율적으로 웹사이트 해킹과 같은 작업을 수행. 또한 국가별·기업별 AI 기술력이 다르기 때문에 AI 기술과 관련된 중요한 정보를 탈취하기 위한 해킹도 발생

* 작년 오픈AI 내부 메시지 시스템에 침투하여 기술과 정보 등을 해킹한 오픈AI 해킹 사건을 시작으로, 최근 미국 사이버보안업체인 맨디언트(Mandiant)는 북한 해커들이 한국, 일본 등 국적의 IT 노동자로 신분을 위장하고 IT 비대면 일자리에 취업하여 돈과 정보를 탈취하는 등 AI 기술을 통원해 전 세계 IT 업계에 광범위하게 침투해 있다고 보도

[그림 3] 북한의 해킹을 통한 사이버 공격 특징



*출처 : 동아일보(https://www.donga.com/news/Politics/article/all/20240125/123217909/1)

○ 일반 사용자가 AI를 통해 과학적 지식이나 지침 등 정보에 대한 접근성이 증가함에 따라 특정 정보를 활용하여 악의적으로 사용할 가능성 존재. 심각한 전염병 병원체를 생성하거나 표적화 된 생물학적 무기를 빠르게 생산하는데 이용되어 치명적인 위험을 야기할 수 있기에, 백악관의 행정명령은 대규모 언어 모델이 생물학, 방사능, 사이버 및 화학 무기를 개발하는데 악의적으로 사용될 수 있는 위험을 우려하고 있으며, 대량살상 무기 대리 측정 벤치마크를 통해 생물 보안, 사이버 보안 및 화학 분야에서 LLM의 위험한 지식에 대한 평가 및 제거 노력을 진행 중

3. 개인정보와 저작권 침해 문제

○ AI 모델의 학습에 사용되는 방대한 양의 데이터에서 금융, 건강, 사생활 등 민감한 개인 정보를 학습하여 심각한 개인정보 유출의 위험을 초래하게 되며 데이터를 악의적인 목적으로 사용하여 개인 및 집단적 피해가 발생.

또한 학습 데이터를 기반으로 효율적인 검색을 지원하기 위해 개인의 민감한 정보에 대한 추론을 가능하게 하는 등 남용의 가능성도 존재

○ AI 모델의 훈련 과정에서 저작권이 있는 데이터를 무단으로 사용할 경우 창의적 표현에 대한 보상 문제를 비롯하여 지적재산권, 데이터 사용 동의, 보상 및 관리 시스템 문제가 발생. 분명하지 않은 저작권 제도는 AI 개발자가 데이터 투명성을 개선하는 작업을 가중시키고 데이터 소싱 및 필터링 인프라가 제대로 갖춰지지 않아 저작권법 준수에 한계가 발생

* 올해 초 엔비디아는 자체 AI 플랫폼인 네모(NeMo)를 학습시키기 위해 저작권이 있는 도서를 무단으로 도용하고 문어체 시뮬레이션을 위해 약19만 6640권의 책으로 이루어진 데이터세트를 사용하여 작가 3명으로부터 피소되었으며, 앞서 지난해에는 오픈AI와 마이크로소프트가 수백만 건의 기사를 AI 챗봇 훈련에 무단 사용한 이유로 뉴욕타임스로부터 고소당한 사례가 있음

[그림 4] AI 저작권 침해 수준과 해외 분쟁 사례

*출처 : 중앙일보(https://www.joongang.co.kr/article/25233711), 동아일보(https://n.news.naver.com/mnews/article/020/0003513286)

4. 오작동 위험 문제

○ AI 시스템에게 국방이나 공공정책 등의 중요한 책임과 의사결정을 맡겼을 때 인간이 과도하게 AI 시스템에 의존함에 따라 통제 및 감독의 어려움이 발생할 가능성 존재. 국방 분야의 워게임 시뮬레이션(Wargame Simulation System)에 생성 AI를 적용하자 사용자가 예측하기 어려운 전술적 의사결정을 보이며 드물게는 고위험성의 핵무기 배치까지 이어지는 것으로 확인됨에 따라 중요한 분야에 AI 기술 기반의 의사결정 시스템을 도입할 경우에는 AI의 재량권과 인간이 통제 가능한 수준을 분석하여 인간의 책임 범위에 대해 신중하게 고려할 필요

○ 국가 간 AI 무기 개발 경쟁이 격화되면서 잘못된 정보나 편향된 학습으로

인해 AI 자율무기체계가 오작동을 일으켜 민간인이나 아군을 목표로 하거나 전략적으로 중요한 인프라를 공격하는 위험성이 존재. 또한 AI 무기가 전투 상황에서 자율적으로 생사 여부를 결정할 경우 인간의 존엄성과 생명권에 대한 위협으로 작용될 수 있으며, 무기의 개발과 사용이 국가 간 군비 경쟁을 심화시키게 될 우려가 존재하므로 AI 무기의 개발과 사용에 대한 강력한 규제와 국제적 합의가 필요

[그림 5] 군사 영역에서의 AI 자율 무기의 개발 사례

*출처 : 조선일보(https://v.daum.net/v/20230907030258470)

○ 이 외에도 생성 AI가 법조계, 의료계, 금융계 등에서 다양하게 활용되고 있으나 실제 법률 실무에서 AI 적용 시 부적절한 인용이나 문구를 사용하거나 의료 AI의 잘못된 판단으로 발생된 사고에 대해서는 명확한 법률이 존재하지 않아 책임 여부를 해결하기 힘든 부분이 존재

5. AI 자동화 시스템 도입으로 노동시장에 영향

○ 범용 AI는 광범위한 작업을 자동화하여 작업의 질과 생산성을 향상시키고 새로운 일자리의 창출 등 긍정적인 영향을 미칠 수 있으나, 특정 영역에서는 잠재적인 일자리 손실 및 대체 등 노동시장 전반에 부정적인 영향을 미칠 가능성이 존재

○ 고소득 · 고학력 근로자일수록 AI에 더 많이 노출되어 AI 대체 가능성이 더 높게 나타났으며, AI가 기존 기술인 산업용 로봇이나 소프트웨어와 달리 비반복적이고 분석적인 업무를 대체하는 경향을 크게 야기시키면서 단기적으로는 실업이 유발되며 장기적으로는 소득 불평등을 높일 수 있음

* 한국은행 고용분석팀장은 '인구 구조 변화, 다가오는 AI 시대의 새로운 노동 패러다임 모색'의 토론회에서 국내 사람의 일자리 341만 개를 AI가 대체할 것이라는 전망을 발표하며 AI의 노동 대체 가능성을 제시. 제조업 분야가 93만개로 가장 많은 일자리 대체가 전망되며, 직종으로는 193만 개의 전문직이 소멸할 것으로 전망

[그림 6] AI 노출도 상위 10개 직업

직업	노출도*
의회 의원, 고위 공무원, 기업 고위 임원	0.660
인문·사회과학 연구원	0.660
행정·경영·금융·보험 관리자	0.65
법률 전문가	0.649
법률 사무원	0.646
회계·경리 사무원	0.644
경영지원 사무원	0.643
대학교수 및 강사	0.642
작가·통번역가	0.642
회계·세무·감정 전문가	0.641

*출처 : 한겨레21(https://h21.hani.co.kr/arti/economy/economy_general/55110.html)

○ 현재 AI에 관한 연구는 미국과 중국이 주도하고 있으며 주요 AI 선도국들은 연구 및 자원 보유 여부에 따라 AI 산업에 영향력을 보유하고 있어, 고급 AI 개발 자원의 보유 여부로 인해 AI 빅테크 기업들의 지배력이 확대되며 국가 간의 격차가 발생. 특히 숙련된 인재 집중도의 불균형과 개발 및 유지에 드는 막대한 재정적 비용은 글로벌 AI 격차를 더욱 심화할 것으로 전망

○ AI의 국가 간 격차로 인해 평균 임금이 낮은 저소득 국가의 근로자들은 고스트워크(Ghost Work)에 해당되는 데이터 라벨링, 콘텐츠 교정과 같은 노동집약적이고 상대적으로 낮은 수준의 AI 작업을 수행. 고스트워크는 성격상 법적 지위가 없는 임시직·자유계약직으로 고용 사각지대에 놓여 있으며, 특성상 국가 간 경계도 없어 정규 직업을 갖기 어려운 사람들이 쉽게 돈을 벌 수 있는 수단이 될 가능성 존재

* 고스트워크(Ghost Work)란 인간의 분별력과 예견 능력이 지속적으로 요구되는 활동에서 AI 시스템을 운영하는데 투입되는 인간노동으로 잘 드러나지 않거나 의도적으로 감추는 불분명한 고용 변화를 의미. 세계은행(World Bank Group)에 따르면 2055년에는 전 세계 고용 중 60%가 고스트 워크로 채워질 것으로 전망

■ 시사점

ㅇ AI의 위험 요인은 단순히 기술적인 결함을 넘어 사용자에 의한 악용으로 발생되는 경우가 상당 부분을 차지하고 있으며 국가별·지역별 격차로 인한 사회적 위험 요인까지 확대되고 있음. 이는 과거에는 고도의 기술력을 필요로 하던 공격방식이 현재는 최소한의 기술 지식만으로도 생성 AI를 이용해 악의적 사용이 가능해짐에 따라 기존보다 더 광범위한 피해로 이어질 가능성을 내포하고 있음

ㅇ 앞서 살펴본 바와 같이 AI의 사용과 관련한 다양한 위험 요인들이 식별되고 있으나 현재까지는 기술적 보호 장치를 개입하는 방법 외에는 명확히 합의된 체계가 없어 혼란이 가중됨. AI 기술로 인해 야기될 수 있는 모든 위험 요인에 초점을 맞추고 가이드라인 및 제도적으로 대응할 수 있는 체계적인 규제를 마련하는 것이 시급

ㅇ 특히 악의적으로 사용되는 경우에는 법적 처벌이나 책임의 소재를 구분하는 명확한 지침이 요구되며 규제 기반의 제도적 접근이 필요. 또한 노동 시장의 문제와 AI 격차 심화에 따른 사회적·경제적 문제와 같은 비기술적 요인들은 단기적인 해결보다는 잠재적 영향에 대한 지속적 논의를 거쳐 신중히 대응할 필요

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

참고자료

- 사진인가 AI가 만든 이미지인가 아리송한 '이스라엘-하마스 전쟁' 사진
https://www.gttkorea.com/news/articleView.html?idxno=7492

- "참가자만 1,200명" 인하대에서 또 텔레그램 딥페이크 성범죄
https://imnews.imbc.com/replay/2024/nwdesk/article/6628394_36515.html

- 쏟아지는 악용 AI 딥페이크, 딥보이스 문제... '나쁜 AI' 막는 '착한 AI' 도입 시급
https://www.newsquest.co.kr/news/articleView.html?idxno=230126

- '딥페이크' 피해 일파만파... AI 윤리 문제 수면 위로
https://www.goodnews1.com/news/articleView.html?idxno=438176

- 독일 '북 해킹' 거듭 경고... "AI 활용 해킹"
https://news.sbs.co.kr/news/endPage.do?news_id=N1007823654

- Cornell University, The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning
https://arxiv.org/abs/2403.03218

- 엔비디아의 AI '네모', 출시 1년 만에 저작권 침해 소송 휘말려
https://www.hankyung.com/article/202403110234i

- "AI자율무기체계 '오작동' 위험성... 인간의 개입 보장해야"
https://www.news1.kr/politics/diplomacy-defense/5527858

- 고소득·고학력 직종, AI 대체 가능성 높다... "국내 일자리 341만개 AI로 대체"
https://www.aipostkorea.com/news/articleView.html?idxno=2797

- AI, 사람의 업무를 '돕던 시대'에서 '대신하는 시대'로
https://h21.hani.co.kr/arti/economy/economy_general/55110.html

- 메리 그레이 외, '고스트워크: 긱과 온디맨드 경제가 만드는 새로운 일의 탄생', 한스미디어, 2019.08.08.

SPC ‘ANGEL’ 통계

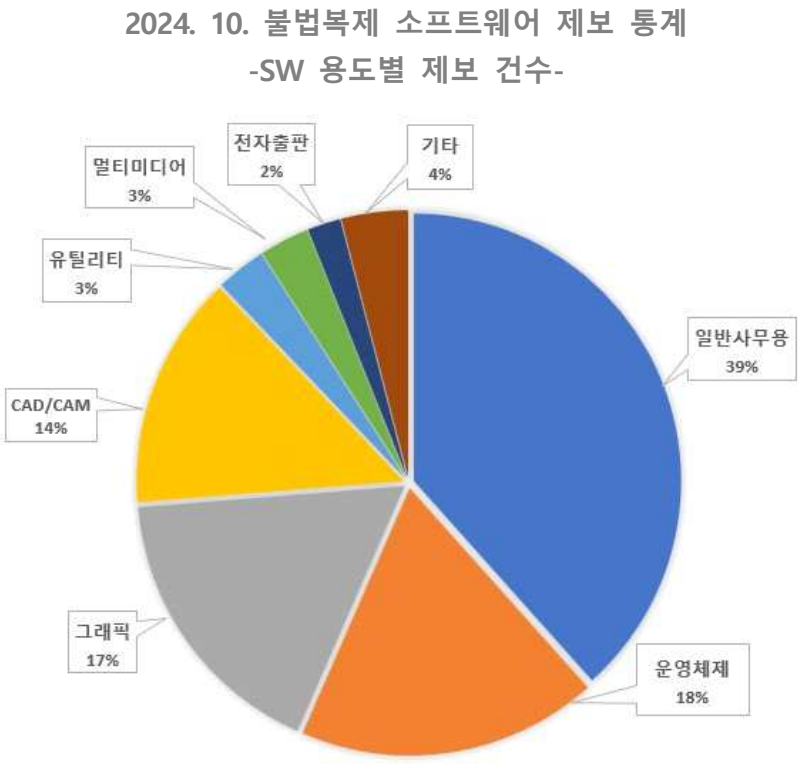
월 1회 제공

한국소프트웨어저작권협회, 불법복제 SW 제보 ‘ANGEL’ 서비스 10월 통계 현황

○ 한국소프트웨어저작권협회(SPC)가 지난 10월 한 달간(‘24. 10. 1. ~ 10. 31.) ‘ANGEL(불법제보)’ 서비스를 분석한 결과, 기업 또는 개인의 불법복제 SW 사용 제보는 총 99건으로 나타남

○ SW 용도별로는 일반사무용 38건(38%), 운영체제 18건(18%), 그래픽 17건(17%), 설계(CAD/CAM) 14건(14%), 유틸리티 3건(3%), 멀티미디어 3건(3%), 전자출판 2건(2%), 기타 4건(4%) 순으로 제보가 접수됨

[그림] SPC ‘ANGEL(불법제보)’ 서비스 2024년 10월 통계 현황



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

다음 SW·저작권 동향리포트 〈제2024-23호〉 발간일은 12월 10일입니다.