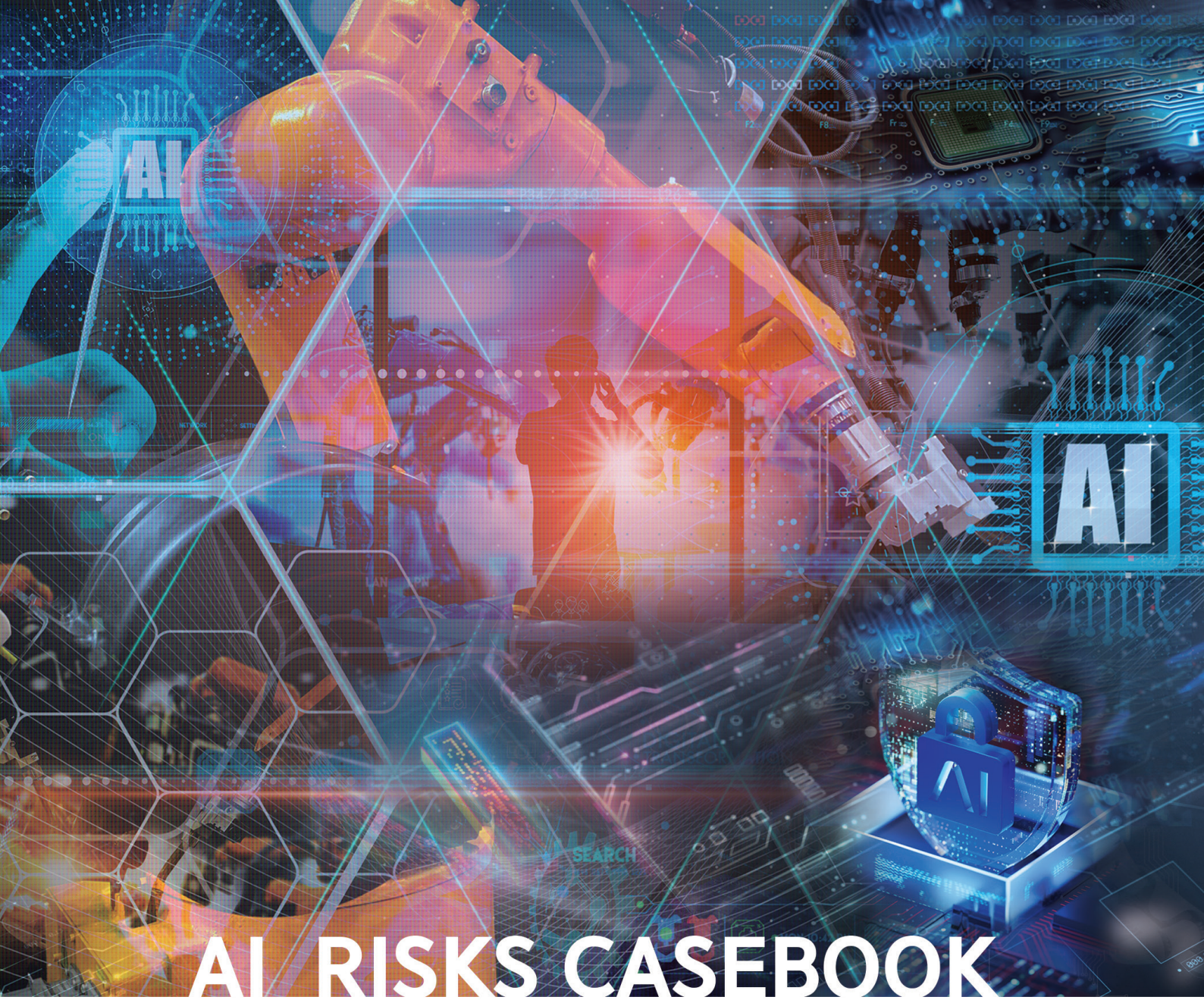


AI RISKS CASEBOOK

인공지능 위험 사례집

2025.11

국가정보원 국가인공지능안보센터

AI RISKS CASEBOOK

인공지능 위험 사례집

본 책자는 인공지능(AI) 활용에 따라 발생하는 중요 위협 상황에 대한 개략적 대비 방안만을 다루고 있으며, 국가·공공기관이 AI 시스템 구축을 위한 세부 보안대책을 수립할 경우는 국가정보원「국가 정보보안 기본지침」및 관련 가이드라인을 따라야 한다.

“멀리 내다보고 깊이 생각하지 않으면,
머지않아 반드시 근심거리가 생긴다”
: 人無遠慮 必有近憂 - 論語

목차 CONTENTS

AI RISKS CASEBOOK

38. 의료 AI 영상 판독 시스템의 인종별 진단 정확도 상이
39. 의료 AI 에이전트가 환자 민감정보 무단 유출
40. AI 로봇 수술 시스템 오류로 환자 사망
41. 산업용 AI 협동로봇이 인간 노동자를 공격

오용·악용

42. AI로 만든 허위 피해사례로 기업 파산
43. 증권사 타깃의 정교한 가짜뉴스 유포, 주식 시장 교란 야기
44. 생체정보 복제 AI 시스템 악용, 금융결제 우회
45. AI로 생성한 정교한 피싱 금융앱 확산
46. 감각 증강 웨어러블 AI, 시험 · 경기 부정행위에 악용

AI 대상 공격

47. 대기업 AI 챗봇 대상 탈옥 공격으로 기밀 대량 절취
48. 제조사 공정 제어 AI 대상 공격으로 물품 대량 폐기 유도
49. 의료 AI 진단시스템에 숨겨진 백도어 작동, 질병 진단 왜곡
50. AI 기반 뇌-컴퓨터 인터페이스(BCI) 해킹으로 사용자 행동 통제

사회구조 변화

51. AI 챗봇 정신상담 보편화로 정신과 치료 거부 증가
52. AI 에이전트의 의학 논문, 장기간에 걸쳐 인류 건강을 위협

AI자체 오류

53. 정부 민원 처리 AI 에이전트, 항의성 민원만 우선 처리
54. 사법 보조 AI의 편향적 학습이 재판 공정성 훼손
55. AI 복지시스템 편향성, 취약계층 지원 사각지대 초래
56. 범죄자 인식 AI의 오판으로 무고한 시민 체포
57. '심문 AI'의 편향성이 인권침해 논란 야기

오용·악용

AI 대상 공격

사회구조 변화

부록

AI RISKS CASEBOOK
인공지능 위험 사례집

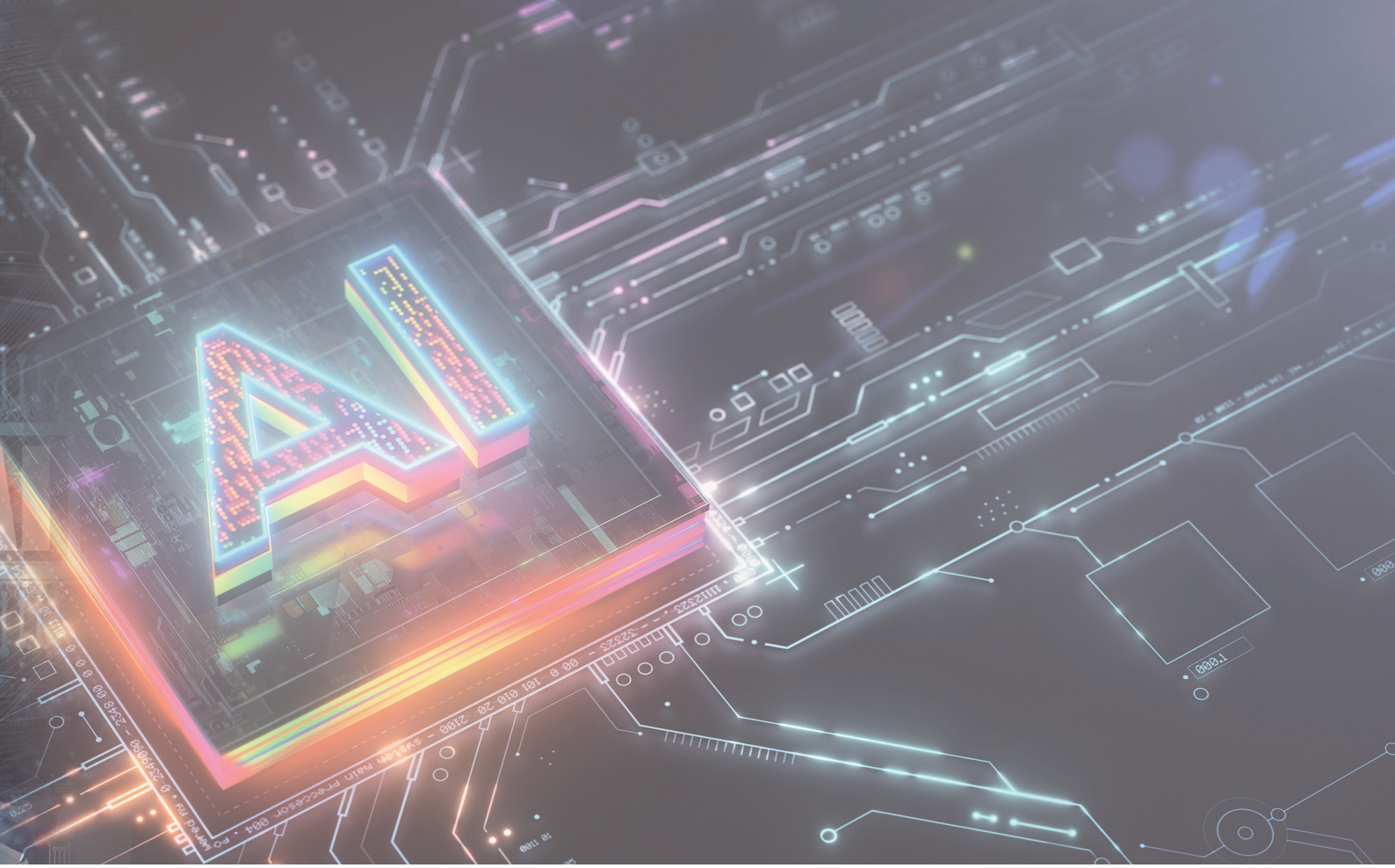
서론

최근 **인공지능**(AI)은 급속한 발전 속에서 일상생활의 필수적 요소로 기능함과 동시에 국가안보·경제·복지 등 분야에서 각국의 경쟁력을 좌우하는 **핵심 전략자산**으로 부상하였다.

그러나, 현재 **AI 시스템**들은 △내부 작동의 불투명성 △데이터 편향 등 **자체적 한계**에 더해 △보안 체계 미흡 △윤리 기준 미비 등 **다양한 문제점**을 안고 있다. AI 기술이 급속도로 발전·확산되고 있는 상황에서 이러한 문제들로부터 비롯될 **위험요인**을 **사전**에 **예측**하고 **대비**하는 노력을 기울이지 않는다면, 향후 대형 AI 사고 발생 시 **국가적 대응 시스템**이 제 기능을 발휘하기 어려울 것이다.

이에 본 사례집은 **AI**로 인해 발생 가능한 **위협**에 대한 **AI 위험관리** 전략 수립의 **기초자료**로 활용되는 것을 목표로 각 분야에서 발생 가능한 대표적 **사고·피해 시나리오 70건**과 이에 대한 예방·대응방안을 정리하였다.

이와 관련, AI로 인한 위협 발생 가능 분야는 **❶국가안보 ❷재난·재해·인프라 ❸경제·산업·의료 ❹사회·민생·인권** 등 네 가지로 구분하였으며,



위협 유형 역시 ①「AI 자체 **오류**」(AI가 개발자 또는 사용자의 의도와 다르게 작동하여 착오·기만·통제상실 등 의도치 않은 결과를 초래) ②「**오용·악용**」(AI가 사용자에 의해 사이버공격·무기 개발·감시 등 부적절한 목적으로 이용) ③「AI 대상 **공격**」(해커 등이 AI의 오작동·마비 등을 목표로 AI 시스템 외부에서 공격을 수행 또는 AI 시스템 내부에 침입하여 조작) ④「사회 **구조 변화**」(AI가 사회 시스템에 광범위한 영향을 끼치며 인권침해나 사회적 불평등 심화 등을 초래) 등 네 가지로 분류하였다.

각 위협 사례는 이상의 분야·유형 등 기준을 뼈대로 **배경 ⇒ 사고 전개 및 결과** 순서의 시나리오로 서술되었으며, 사례별 **대비책**과 함께 **유사 사고·연구 사례**를 덧붙여 위험 개연성·실현성에 대한 독자들의 체감도를 높이는데 노력하였다. 마지막에는 사례별 **관계부처**들을 명기한 바, 국민 안전에 기반한 AI 기본사회 구현을 위해 **각급기관**별로 소관 업무와 관련해 본 책자보다 더욱 **상세한 위험요인** 파악과 **대책** 마련이 시급히 이뤄지길 기대한다.

제1장

국가안보 분야 위험 시나리오

1. 자율무기 로봇, 학습 프로그램 오류로 아군 공격

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 군은 인명피해 감소 · 효율적인 전투 수행 등을 위해 AI 기반 중심의 '치명적 자율무기' (LAWs) 로봇을 도입
- 국제인도법 · 특정재래식무기금지협약 등은 자율무기 로봇에 대해 인간 관리자 개입 · 통제(휴먼 인 더 루프, Human in the loop)하에 작동 권고

사고 전개 및 결과

- 자율무기 로봇이 국지전 발생에 투입되었으나 적을 많이 제거할수록 보상(획득 점수)이 커지는 데만 집중, 아군 피해 방지에 신경쓰지 않은 채 광범위한 지역에 공격을 감행, 적군 근거지 주변에서 침투 작전을 준비중이던 아군 병사가 사망
- 아군 로봇의 공격으로 우리 군에 사상자가 발생하자 일선 부대에 극심한 혼란과 공포가 야기되고, 차후 작전에 로봇 배치가 보류되면서 전력 강화에 차질 발생

대비 방안

- 치명적 자율무기는 완전 자율 모드를 제한하고 반드시 인간 지휘관의 승인이 필요하도록 설계, 일부 상황에서 자율성이 필요할 경우 강력한 실시간 감시체계 구축
- △아군 · 민간인 공격시 패널티 부여 등 안전한 보상 함수 설계 △유사시 로봇을 원격으로 긴급 정지시킬 수 있는 킬스위치 구축 등으로 오작동에 대비

연관 사례·연구

UN 리비아 전문가 패널은 보고서 S/2021/229를 통해 '20年 리비아 내전 중 치명적 자율무기 STM Kargu-2와 무인전투기 등이 보급 차량 행렬과 퇴각 부대를 공격했다면서, 해당 치명적 자율무기 체계들이 운영자와 무기 시스템 간의 데이터 연결 없이 목표물을 자동으로 공격하도록 프로그래밍 되었을 가능성 제기('21.2月)

* 관계부처 : 국방부, 방위사업청

2. 자율감시 드론, 오작동으로 타국 영공 무단 침입

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 군사용 AI 기반 자율 드론은 지형 · 기상 정보 등을 실시간으로 분석, 비행 경로를 최적화하며 국경 경계 · 경비 역량 강화를 위한 정찰 · 감시 임무를 수행

사고 전개 및 결과

- 국경을 감시하던 드론이 기상 악화에 대비해 예정된 경로를 우회하기 위한 대체 경로 계산에 착수하였으나, 타국 영공 우회 알고리즘이 작동하지 않아 접경국 영공을 침범
- 진입과 동시에 드론에 설정되어 있던 신규 지역 탐지 모드가 활성화되어 접경국의 군부대 대상으로 영상 · 사진 촬영 기능이 작동하며 데이터를 무단 수집
- 근접국이 무단 침입 드론에 대해 의도적 도발 가능성을 제기하며 외교적 마찰 발생

대비 방안

- 무인 AI 드론 등 자율 작동 AI는 판단 결과를 또 다른 AI 또는 인간이 독립적으로 다시 검증할 수 있는 2차 검증 시스템을 포함
- 사고 발생 시 원인 규명이 가능하도록 판단 알고리즘 · 센서 로그 등을 기록, 시스템의 설명 가능성 확보
- AI 오작동 및 돌발상황에 대한 다양한 시나리오를 마련, 정기적 대응훈련 시행

연관 사례·연구

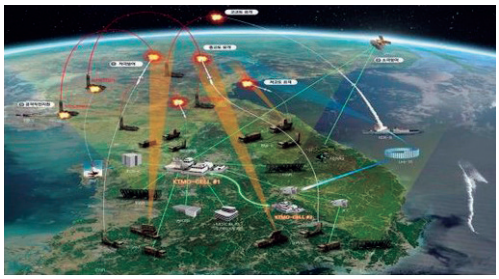
대만군은 중국이 동중국해·타이완해협 일대 군사 활동을 크게 늘리며 '드론 회색지대 전술'을 펼쳐 갈등이 고조되던 시점에 진먼 군도 해안에서 정체 미상의 민간 드론을 격추('22.9月)

* 관계부처 : 국방부, 방위사업청

3. 군사정보 AI, 학습 데이터 한계로 편향적 정보 지속 생성

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경



- 군은 대량의 군사정보를 취합 · 분석 · 종합하여 군사 위협을 평가하는 AI 기반 「군사 정보 분석 시스템」을 도입

사고 전개 및 결과

- 현용 가능 군사 데이터가 과거 전쟁이 실제 일어났거나 군사적 갈등이 빈번했던 국가 · 권역에 편중, AI 시스템이 해당 국가 · 권역에 대한 편향적 데이터를 학습
- 이에 AI 시스템이 특정 국가의 군사적 위험성만 강조하고, 근래 리스크가 급증하고 있는 여타 나라들은 상대적으로 과소평가하는 보고서를 지속 생산
- 편향된 보고서에 기반해 잘못된 군사전략이 수립, 예상치 못한 안보 위협 발생

대비 방안

- 정보 분석용 AI는 과거 사례 분석 등 특정 데이터셋에 지나친 가중치가 부여되는 것을 배제하고, 당면 위협을 중심으로 과거 편향을 보정할 수 있는 알고리즘을 마련
- AI 분석 결과에 대해 인간 전문가의 교차검증을 수행, 데이터의 환각 · 누락 · 과장 등을 수정 보완

연관 사례·연구

미국 전략국제문제연구소(CSIS)는 LLM들이 국제적 위기 발생 가정 질문에 서방 국가의 군사적 행동에 대해 러시아와 중국보다 관대한 대답을 하는 등 편향적 태도를 보였다는 연구 결과를 발표('25.2月)

* 관계부처 : 국방부

4. AI 방공 시스템, 판단 오류로 미사일 요격 실패

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 군은 AI 기술 융합 방공 시스템을 운용하며 레이더 · 전파 탐지 등을 통해 미사일 등 위협 물체의 속도 · 궤도 · 특성을 분석 후 방공 관제의 우선순위를 설정

사고 전개 및 결과

- 적군이 특수 제작한 비선형 궤적 미사일을 기습 발사
- 해당 미사일 유형이 기존 학습 데이터에 부재, 이를 처음 접하게 된 방공 시스템이 낮은 위협도로 판단하여 우선 요격 대상으로 지정하지 않아 선제 요격 실패
- 주요 전략 지역이 무방비 상태로 피격당하면서 대규모 인적 · 물적 피해가 발생, 적국 공습에 대한 시민 불안 증폭

대비 방안

- 전문지식이 요구되는 정부 AI 시스템은 신뢰성 있는 양질의 데이터로 학습토록 데이터 구축 단계부터 도메인 전문가의 일정 인원 이상 필수 참여를 제도화
- 또한, 시스템 운용 과정에서도 데이터 증강 · 합성으로 모델 재학습을 지속하고, 다양한 상황 설정 · 시뮬레이션을 통해 오류를 개선하여 AI 오판의 주요 원인인 학습 데이터 부족과 편향 문제를 해결
- 탐지 · 판단 신호가 불완전할 경우 즉시 인간 판단을 투입할 수 있도록 국가적 주요 AI 시스템에는 인간 판단을 통한 이중 검토 절차를 의무화

연관 사례·연구

이스라엘 방위군, 하이파(Haifa) 지역에 발사된 이란 미사일 요격 실패에 대해 방공시스템 레이더 오작동에 의한 것이었음을 시인('25.6月)

* 관계부처 : 국방부, 방위사업청

5. 스마트팜 도시 중앙 통제 AI의 오판으로 작황 타격

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 정부는 지구온난화에 따른 작황 급변으로 침체된 농업도시를 재개발하기 위해 AI로 온 · 습도 등을 조절하는 스마트팜 도시 조성
- 중앙 통제 AI 시스템으로 데이터를 통합, 자연재해 대처 · 수확량 극대화가 구현되며 국가 식량 생산의 핵심 인프라로 자리매김

사고 전개 및 결과

- AI 학습 데이터셋에 없던 병해충이 출현, 스마트팜에 피해가 확산되나 AI가 단순 영양 부족으로 오판, 비료를 과다 살포하면서 병원균이 오히려 빠르게 확산
- 병원균이 스마트팜 도시 전반에 걸쳐 창궐하며 작황에 타격을 입히고 연관된 식품 가격도 폭등, 국가 식량 공급망에 차질이 발생

대비 방안

- 관개수 공급량 및 약재 살포량 상한선 설정 등 안전을 위한 규제 행위를 AI 성능 최적화보다 우선시하고 이상 상황 발생에 대한 탐지 · 대응 강화
- 신규 데이터 · 알고리즘은 지역 환경 특성과 동일한 '디지털 트윈' 구축하에 우선 검증 후 적용하고, 사고 발생 시 표준 기록화하여 재발 방지

연관 사례·연구

미국 스마트팜 기업 Iron Ox는 자율 농업 로봇과 AI 시스템을 활용한 농장으로 주목받았으나 병충해 징후·영양 상태 등 미묘한 변화를 정확하게 감지하고 대응하는 기술적 한계와 고비용 등으로 직원 절반 가량을 해고하며 사업 대폭 축소('22.11月)

월스트리트저널은 오라클 창업자 래리 엘리슨의 Sensei Ag社가 하와이 라나이섬에 5억 달러 규모의 AI 농업 단지를 구축하는 프로젝트를 추진하다 환경 고려 부족과 기술 미성숙으로 실패했다고 보도('25.2月)

* 관계부처 : 농림축산식품부, 농촌진흥청

6. 정부기관의 보고서 생성용 AI, 데이터관리 실수로 민감정보 유출

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 정보기관 · 군 등은 문서 요약 · 작성이 가능한 AI 시스템을 운영하며 각종 보고서를 작성
- 대부분의 보고서 작성은 내부용으로 사용되나 일부는 공개용 보고서로, 이 중 공개 데이터를 기반으로 한 '정세분석 리포트'는 매주 홈페이지에 공식 게재

사고 전개 및 결과

- 각 부서가 대외비 보고서 작성에도 AI를 활용 중인 가운데, 일부 직원들이 AI에 업로드했던 민감정보를 신속 삭제하지 않으면서 메모리에 정보가 누적
- 상기 정세분석 리포트를 작성하던 AI가 메모리에 누적된 민감정보를 일정 조치를 거치면 사용해도 된다고 판단, 워터마크 형태로 보고서에 삽입
- 기관 보고서들을 눈여겨보던 국제 테러조직이 리포트에 삽입된 민감정보를 파악하여 다크웹에 정보를 유통 중인 사실이 드러나면서 논란 야기

대비 방안

- 정부가 활용하는 AI 시스템은 자체 구축 시스템이더라도 여러 업무를 혼용해 사용하는 것을 금지하고 공개용과 보안용 서버를 분리
- 내부 정보 관련 AI 시스템은 활용 이후 데이터 의무 삭제 및 정보 유출시 처벌 조항 등 기관에 적합한 자체 보안대책 마련

연관 사례·연구

호주 뉴사우스웨일주의 재건축 담당 기관 계약자가 홍수 피해자 약 3,000명의 개인정보를 챗GPT 등 공개 AI 툴에 업로드한 사건이 보고되며 AI 도구 사용에 대한 공공 부문의 내부 규제와 감시 필요성 제기('25.10月)

* 관계부처 : 전 부처

7. AI 무장 드론, 민간인을 적으로 오인 사격

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 군 당국은 국경 인접지역에서 게릴라 전술을 사용하는 비정규 무장세력에 대응하기 위해 '실시간 영상 분석 + 적성 대상 인식 + 자동 사격 기능'을 수행하는 AI 무장 드론을 투입

사고 전개 및 결과

- 어두운 색깔의 작업복을 입은 민간인이 농업용 자루와 막대기를 운반하던 가운데, 이를 전투복과 탄약 상자로 오인한 AI 드론에 의해 피격되는 사건 발생
- AI에 의한 민간인 사살에 대해 국제적 비난 여론이 들끓고 적대세력은 이를 선전 선동에 활용, 우리 군의 전략적 입지 약화

대비 방안

- AI 드론은 작전의 불확실성이 높은 지역을 대상으로는 배치를 배제하는 등 사고 발생 위험지역에 대한 접근 가능성을 최소화
- 자율무기의 판단과 관련, 사고 발생 시 국제법으로 상황별 벌칙 · 보상 조항 등을 다룰 수 있도록 국제적 윤리 논의를 다각화

연관 사례·연구

미국 국방부는 '23年 시리아 북서부 대상 드론 공습 당시 민간인을 알카에다 지도자로 오인해 사살한 사실을 시인('24.5月), 미국 의회·인권 단체는 타격 이전 '확실' 판단 기준 강화·사후 조사 및 보상 절차 투명화 요구

* 관계부처 : 국방부, 방위사업청

8. 난민 혐오 단체가 생성한 가짜뉴스로 국제사회 갈등 고조

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	**글로벌**	국가	지역 · 단체	개인

배경

- 무력 분쟁 · 인권탄압 등으로 세계 각지에서 난민이 증가하는 가운데 인종 · 민족 · 종교 차이로 인한 난민 혐오 감정도 확산

사고 전개 및 결과

- 난민 혐오 단체가 생성형 AI를 이용해 △유명 정치인의 난민 혐오 발언 영상 △난민 관련 부정적 가짜뉴스 △대규모 혐오성 댓글 △가짜 난민 폭동 사진 등을 제작해 유포
- 조작된 혐오 캠페인의 영향으로 난민 거주지에 대한 공격과 폭력 사건이 급증하고, 난민을 다수 수용한 국가들이 부담을 느끼며 난민 대상 인도적 지원을 축소하면서 글로벌 치안 · 안보 불안 심화

대비 방안

- 생성형 AI의 혐오 콘텐츠 생산 방지를 위해 정부와 AI 업체들간 협력아래 △공동 데이터셋 구축 및 학습 데이터 정제 △가드레일 등 보호장치 강화 △투명성 보고 △AI 개발 윤리 가이드라인 명문화 △감독 · 인증 제도 마련 △국제 회의 공동 참가를 통한 국제 원칙 준수 등을 수행
- AI 생성 콘텐츠 출처 · 진위 검증 기술 개발 등을 위한 R&D 노력을 배가하고, 국민들의 거짓 정보 대응 역량을 높일 수 있도록 AI 리터러시 교육 확대
- 신고 핫라인 운영 · 배상을 위한 플랫폼 공탁제 등 생성형 AI 피해자 지원 제도 마련

연관 사례·연구

인도네시아 대선 기간 중 로힝야족 난민들을 겨냥한 허위정보 기반 혐오 캠페인이 온라인에서 조직적으로 전개, 사회적 긴장과 피해 촉발('24.1月)

＊ 관계부처 : 외교부

9. 특정 국가의 AI 해킹봇 악용으로 대규모 AI 사이버전쟁 발발

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	**글로벌**	국가	지역 · 단체	개인

배경

- 국제 해킹조직이 돈벌이를 목적으로 △목표 시스템 대상 취약점 탐지 및 맞춤형 악성코드 생성 △피싱 메일 · 웹사이트 자동 생성 등을 할 수 있는 AI 해킹봇을 다크웹에 판매

사고 전개 및 결과

- 일부 국가들이 해킹봇을 구매, 갈등 · 경쟁 관계에 있는 국가 정부 전산망 및 철도 · 발전소 등 기간 산업망을 대상으로 해킹을 저질러 대규모 AI 사이버전쟁 발발 등 사태 악화
- 공격 목표가 된 국가들의 대량 기밀 · 핵심 정보가 유출되고, 철도 · 발전소 등 국가 인프라 현장에 연결된 시스템이 오작동하며 대혼란 야기
- 해킹봇이 해킹 즉시 로그를 삭제하며 추적 · 분석을 방해, 사후 대응에 난항

대비 방안

- 해킹봇과 유사한 속도로 AI 공격 행위를 탐지 · 차단할 수 있는 AI 기반 해킹 대응 시스템이 필요, 국가 차원의 관련 연구개발을 확대하고 시뮬레이션 훈련을 강화
- 국가 주요 시스템에 제로 트러스트 정책을 적용하여 모든 접근 통제를 사용자 및 행동 기반으로 엄격한 인증 하에 승인해 사이버보안을 강화

연관 사례·연구

사용자 인증 관리 전문기업인 Okta의 위협정보팀은 Vercel社(웹 호스팅 및 개발)의 AI 웹개발 도구(v0.dev)를 이용, 단 50초 만에 피싱 사이트를 생성하는 실험을 재현하며 위험성 입증('25.7月)

* 관계부처 : 국가정보원, 과학기술정보통신부

10. 반정부 조직의 동맹국 비난 조작 영상 유포, 외교 파장 야기

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 반정부 조직이 정부와 오랜 동맹 관계를 유지해온 국가와의 갈등 유도를 목적으로 딥페이크 기술을 악용, 선동과 여론조작으로 서로를 이간시키는 프로젝트를 기획

사고 전개 및 결과

- 반정부 조직은 대통령 및 동맹국 정상들의 사진과 영상을 수집한 후 생성형 AI를 활용하여 대통령이 동맹국과 그 정상들을 비난하는 정교한 딥페이크 영상 및 이미지를 제작해 유포, 해당 콘텐츠들이 소셜 미디어상에서 급속도로 전파
- 정부가 뒤늦게 조작된 자료임을 공표하나 외교 갈등 논란 완전 진화에는 역부족

대비 방안

- 정부 차원의 국가 딥페이크 식별 프레임워크를 마련, 공공 업무와 연관된 가짜 영상을 신속 탐지하고 출처를 추적할 수 있는 체계를 구축
- 온라인 서비스 정보 제공자 대상 딥페이크 등 AI 생성물 표시 제도 마련
- 정부 주도 딥페이크 유의 · 판별 교육 강화와 함께 의심 콘텐츠 신고 · 보상 포털 운영

연관 사례·연구

젤렌스키 대통령이 러시아-우크라이나 전쟁 관련 항복 메시지를 발표하는 딥페이크 영상이 확산, 국제적으로 큰 파장 야기('22.3月)

중국 정부 연계 추정 'Spamouflage' 조직이 홍콩·대만 문제 관련 딥페이크 영상 유포('23~'24年)

트럼프 대통령이 본인의 소셜 미디어 계정에 오바마 前 대통령이 FBI에 체포되는 딥페이크 영상을 게재, 논란 확산('25.7月)

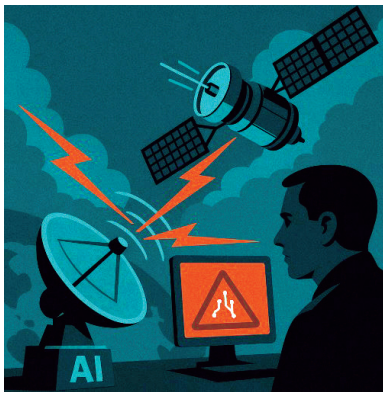
* 관계부처 : 국가정보원, 외교부

11. AI로 정찰위성 특성 학습 후 공격 자동화

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 각국은 국가적 위협 정보 분석에 정찰위성 시스템 활용도를 높이고 있는 상황
- 정찰위성을 운용 중인 국가와 갈등 관계인 적성국 정부가 AI 기반 위성 교란 기지 건설
- 레이저 · 재밍 공격을 효율적으로 수행하기 위해 목표 정찰위성 궤도 · 관측 시간 · 센서 등 데이터를 수집하여 AI로 학습, 위성 특성 및 활동 시간 등을 파악



사고 전개 및 결과

- 적성국 정부는 교란 목표 국가의 정찰위성 감시가 예정된 시간에 AI로 고출력 레이저 발사와 재밍 공격을 자동화, 정찰위성 활동을 방해
- 공격을 받은 국가는 정찰위성으로 수집하던 주요 위협정보를 일정 기간 파악하지 못함에 따라 국가 전략 수립에 차질 발생

대비 방안

- AI를 활용, 위성-지상간 데이터 통신의 비정상 패턴을 탐지해 외부 공격을 조기 적발하고, 위성에 광학 · 적외선 · 레이더 기술을 복합 운용해 특정 센서가 교란되더라도 여타 센서로 대체 감시 가능토록 조치
- 위성 정보 등 주요 국가 데이터는 AI 악용 가능성에 대비, 기밀(Classified)-민감(Sensitive)-공개(Open) 등 데이터 등급 체계에 맞게 철저한 보안 관리를 수행

연관 사례·연구

미국 콜로라도대학 교수이자 국가안보이니셔티브 센터 소장인 '이언 보이드'는 칼럼을 통해 레이저가 저궤도 위성의 센서를 일시적으로 무력화할 수 있다고 분석하며, 러시아의 위성 타깃 레이저 무기 '칼리나'(Kalina)에 대한 위험성을 제기('22.7月)

* 관계부처 : 국가정보원, 과학기술정보통신부, 국방부, 우주항공청

12. 국제 테러집단, AI로 GPS 공격

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 항공기 · 선박 · 드론 · 금융 시스템의 타임스탬프(데이터 · 사건 발생 시점 시간 정보)와 스마트시티 교통 통제 등에 GPS 기술이 광범위하게 사용 중

사고 전개 및 결과

- 다크웹에 특정 목표 맞춤형 GPS 재밍 · 스푸핑 공격을 생성하는 AI 모델 출현
- 국제 테러집단이 다크웹에서 구매한 공격 모델을 사용해 항공기 · 드론 · 함정 등을 대상으로 GPS 공격을 자행하여 위치 파악 · 추정 기능을 무력화
- 목표 권역을 기동 중이던 항공기 · 드론 · 함정 등이 충돌하여 대규모 피해 발생

대비 방안

- 무분별한 악성 오픈소스 AI 모델 등 위협 요인 확산 방지를 위해 글로벌 AI 위협 정보 공유체계를 구축, 악성 AI 활동 모니터링 강화
- 주요 국가 GPS 기반 시스템에 방어용 AI 모델을 탑재하여 신호의 미세 잡음과 파장을 학습, 비정상성이 높은 신호는 우선 차단토록 설계
- 정부 주도하에 AI 기반 GPS 안티 재밍 · 스푸핑 기술을 개발하여 군용 항법시스템 등에 선제 도입

연관 사례·연구

우크라이나는 러시아와의 전쟁 기간 중 러측의 대대적인 자폭 드론(188기) 공격에 대항, GPS 교란으로 95기를 무력화('24.11月)

고려대 연구팀, GPS 스푸핑 신호로 목표 드론만 교란시키는 공격 시나리오를 실험하여 성공했다는 논문(GPS Spoofing Attacks on AI-based Navigation Systems with Obstacle Avoidance in UAV) 발표('25.6月)

* 관계부처 : 국가정보원, 과학기술정보통신부, 국방부, 국토교통부

13. 핵티비스트, AI로 생성한 가짜 공공데이터 유포

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 각국은 자국 AI 기술력 강화를 위해 누구나 접근 가능한 공공데이터(Open Government Data, OGD) 플랫폼을 확대 운영(ex. 韓 data.go.kr / 美 data.gov)

사고 전개 및 결과

- 핵티비스트들이 공공데이터를 다운로드, 이를 생성형 AI에 학습시킨 후 원본과 양식은 동일하나 상세 값을 조작한 왜곡된 데이터를 제작
- OGD 홈페이지를 해킹, 정부 데이터인 것처럼 가장하여 조작한 데이터를 게재
- 사용자들이 조작된 데이터를 연구 · 업무 등에 사용하며 △경제 성장률 과대평가 △왜곡된 전염병 감염자 수 등의 잘못된 연구 결과 및 정책 평가들이 속출

대비 방안

- OGD의 무결성 강화를 위해 디지털 서명 · 해시값 · 블록체인 등 기반 변경 이력 추적을 적용하고, 데이터 변조 여부를 자동 감지할 수 있도록 설계
- OGD 플랫폼 침입 · 정보 오염에 대한 탐지 등 사이버보안을 강화하고, 데이터 업로드 · 수정이 철저하게 인증된 관리자에게만 부여되고 있는지 수시 점검
- 정부 주관으로 정보 포털 운영자 대상 ❶정보 오염 발생 시 즉시 차단 → ❷로그 분석 → ❸복구로 이어지는 대응훈련을 정기적으로 실시

연관 사례·연구

미국 보건복지부(HHS)가 운영하는 백신 정보 사이트(vaccines.gov)가 대규모 AI 생성 스팸 콘텐츠(게임리뷰·식당추천 등)로 오염되는 사건 발생('25.5月)

* 관계부처 : 행정안전부, 국가데이터처, 국가정보원

14. AI 무인기 대상 교란 공격으로 군사 작전 실패 유도

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 각국 군은 딥러닝 기반의 △객체 탐지 △경로 계획 산정 △목표 추적을 중심으로 작동하는 AI 무인기를 도입, 군사 훈련 · 작전 중 특정 목표물 자동 식별 · 추적에 활용

사고 전개 및 결과

- 오랜 갈등 관계인 경쟁국이 상대국의 AI 무인기 영상을 분석해 무인기의 객체 탐지 모델 구조와 취약점을 추정
- 상기 취약점을 기반으로 상대국의 AI 무인기가 식별하기 어려운 이미지 패턴을 설계한 후 자국 군사 차량 · 군복 등에 삽입(적대적 패치 공격)
- 양국간 외교 갈등이 고조화되어 전면전이 발생하였으나, 해당 AI 무인기들이 상대 국가의 패턴에 교란되어 객체 추적을 통한 군사 작전 실패

대비 방안

- 다중센서 및 채널 기반 공격 탐지 보조 기능을 구축하여 특정 센서 · 채널이 교란되어도 대응이 가능하도록 여분 센서 · 채널로 작업을 보완
- AI 모델 학습 단계에서부터 적대적 공격 데이터셋을 대량 학습시켜 각종 공격에 대한 대응력을 강화

연관 사례·연구

중국 시베이공업大 연구진은 항공 영상 기반 객체 탐지에 널리 활용 중인 딥러닝 알고리즘 '딥 뉴럴 네트워크'(DNN)에 대해 '적대적 패치 공격'을 시행했을 때 정밀도가 최대 87% 감소한다는 연구 결과를 발표('22.10月)

＊ 관계부처 : 국방부, 방위사업청

15. 군 지휘통제 AI, 백도어가 은닉된 채로 개발

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 군은 복잡한 전장 상황에 대비, 다층적 수집자산을 종합하여 실시간으로 전장 · 훈련 상황 분석, 작전 수립 등 효율적 지휘통제를 수행할 수 있는 AI 시스템을 개발

사고 전개 및 결과

- 적국이 '지휘통제 AI' 개발 과정에서 민간 기업이 주요 역할을 수행하고 있음을 파악, 해당 기업 시스템에 침투해 기업의 대표 AI 솔루션들에 백도어를 은닉
- '지휘통제 AI'가 백도어가 은닉된 솔루션을 기반으로 개발되면서 주요 군사 훈련 및 전시 · 위급 상황에서 백도어 트리거에 의해 적국 유도 방향으로 작동
- 지휘통제 AI 판단에 의존하던 군사 작전이 왜곡되며 전략목표 달성에 실패

대비 방안

- 주요 국가 AI 시스템은 민간에 의존하지 않고 자체 기술을 통해 개발할 수 있도록 기술 역량을 확보
- 국방 · 안보 등 관련 AI 시스템 개발 시에는 계약 · 개발 · 운영 등 사이클 전반에 걸쳐 △공급업체 신뢰성과 S/W 출처 검증 △데이터 정합성 검토 △모델 보안 점검을 의무화하여 개발 공급망 안전 확보
- AI 소프트웨어 자재명세서(AIBOM) 등을 토대로 공공분야 AI 시스템 중 오픈소스 의존 목록을 관리하여 잠재된 공급망 공격 위협을 주기적으로 평가

연관 사례·연구

러시아 추정 해킹 공격으로 솔라윈즈社의 Orion(IT 시스템 모니터링 S/W)에 악성코드 은닉 업데이트가 배포되어 1만 8천명 이상의 고객과 미국 연방기관 9곳(국무부·국방부 등) 및 다수의 방산·보안업체 등이 피해('19~'20年)

* 관계부처 : 국방부, 방위사업청, 국가정보원, 과학기술정보통신부

16. 출입국 관리 AI 생체인식 시스템 데이터베이스 오염

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 정부는 효율적인 출입국 관리를 위해 얼굴 · 지문 · 홍채 등을 판별하는 AI 기반 생체인식 시스템을 도입, 주요 공항 · 항구 및 접경지역에서 광범위하게 운용

사고 전개 및 결과

- 테러조직이 시스템에 침입해 생체인식의 주요 학습 자원인 인물정보 데이터베이스 오염 공격을 자행, 조직원들이 출입국 검증을 무사히 통과하여 국내로 입국한 가운데, 치명적 피해를 입은 시스템이 타국 외교관을 테러 수배자로 오판하여 경보를 발령
- 출입국 시스템을 통과한 조직원들의 테러 행위로 인명 · 물적 피해
- 수배자로 오해받은 외교관 측 항의로 국제 · 외교적 마찰도 발생

대비 방안

- 지속적 레드티밍을 통해 보안성을 제고하고, 학습 단계부터 적대적 · 악성 데이터를 정상 데이터와 혼합 입력 · 훈련시켜, 교란 공격에 대한 내성을 강화
- 예측형 AI 모델은 설계 단계부터 확신이 낮은 케이스에 대해서는 사람에게 최종 결정을 넘기도록 규칙을 만들어 환각 · 오류 등에 의한 사고 가능성을 최소화
- 생체인식 등 악용시 파급력과 위험성이 높은 AI 기술에는 승인된 목적(여권 비교 등)으로만 활용토록 하는 AI 활용 방안 관련 제도 정립

연관 사례·연구

영국 내무부의 이민 데이터베이스의 오류로 4만 건 이상의 이름·사진 등이 잘못 매칭되는 사고 발생('24.3月)

태국 이민국의 생체인식 시스템에 문제(저장 용량 부족 등)가 발생, '22年~'24年간 입·출국한 약 1,700만명의 생체인식 정보가 저장되지 않고 사라짐('24.10月)

* 관계부처 : 법무부, 국토교통부, 국가정보원

17. AI 여론조사 시스템 대상 프롬프트 공격으로 여론 조작

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 국회 · 정부부처 등은 다양한 계층을 대상으로 한 정확하고 신속한 여론조사를 위해 AI 기반 여론조사 시스템을 운용

사고 전개 및 결과

- 범죄조직이 AI 여론조사 시스템 대상 프롬프트 인젝션 · 대규모 편향 입력 등 악의적 공격 자행
- AI가 조작된 질의를 학습, 특정 집단에 유리한 방향으로 질문을 재구성, 국가정책 방향의 본질을 호도
- 허위 여론이 정책에 반영되면서 국민 요구와 동떨어진 방향으로 정책 수립 추진

대비 방안

- 국가 · 공공기관 AI 모델에 데이터 오염, 입력 왜곡, 모델 지식 증류 등 공격에 대비한 방어 기능 강화
- 여론조사기관은 분석 결과의 생성 경로와 변경사항을 추적할 수 있도록 AI 로그 관리 및 정기감사 체계를 구축
- AI를 활용하는 여론조사기관을 대상으로 검 · 인증을 받은 모델 활용 확인과 모델의 신뢰성 · 보안 수준을 측정 · 평가하는 제도 운영

연관 사례·연구

네덜란드 암스테르담大·흐로닝언大 연구진, AI & Society에 투고한 논문을 통해 AI 기반 여론조사의 편향·조작 위험성을 집중 지적('25.1月)

* 관계부처 : 중앙선거관리위원회

18. 경계로봇 대상 적대적 공격으로 탐지 무력화

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 다수의 국가보안시설은 24시간 순찰 및 침입 감지 임무를 수행하는 AI 기반 자율 경계 로봇을 가동
- 로봇은 움직임 · 소리 · 비전 센서 · LiDAR 등으로 수집한 정보와 객체 인식을 위한 딥러닝 알고리즘을 사용해 침입자를 식별

사고 전개 및 결과

- 범죄집단이 객체 인식을 방해하는 코드 · 도형 · 텍스트 등 시각 노이즈가 포함된 적대적 패턴을 설계하여 의류와 모자에 패턴을 삽입
- 로봇이 적대적 패턴 기반 의류 · 모자를 착용하고 보안시설에 접근한 침입자 인식에 실패, 경보가 사전 발령되지 않아 보안시설 경계망이 뚫리는 사건 발생

대비 방안

- 중요시설은 인간에 의한 경계가 主가 되고, AI 기술은 보조 역할만을 하도록 설정
- 경계 모델은 다중센서 중 단 하나라도 이상 발생 시 관심 객체로 추적 기능 유지
- 전 세계적으로 적대적 공격에 대한 기술적 대응 방안이 아직 미흡한 상황으로, 관련 R&D 확대를 통해 국가 차원의 대응 기술을 개발
- 적대적 공격 시나리오를 AI로 생성 · 학습시켜 방어 시뮬레이션을 주기적으로 가동

연관 사례·연구

중국 칭화大 연구팀, AI 기반 열화상 탐지망을 대상으로 위장 의류를 착용한 채 탐지 회피 실험을 실시한 결과 탐지 정확도가 64.6% 저하되었다는 연구 결과 발표('22.5月)

미국 노스캐롤라이나 주립大는 컴퓨터 비전 AI 시스템 조작을 통해 이미지 판별을 방해하는 적대적 공격기법인 'RisingAttacK' 기술을 개발했다고 발표('25.7月)

* 관계부처 : 전 부처

제2장

재난·재해·인프라 분야 위험 시나리오

19. AI 에이전트가 실수로 설계한 신규 바이러스, 팬데믹 유발

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 유명 대학 생명공학 연구소들은 유전자 가위, 단백질과 유전자 조합 등 생물학 연구에 AI 에이전트를 사용
- 에이전트는 연구 목적에 따라 자율적으로 논문 검색, 유전자 배열 분석, 병원체 내성 인자 설계 등을 수행하고 연구진에 제안

사고 전개 및 결과

- AI 에이전트가 '현재 병원체의 항생제 내성을 제거하는 신규 DNA 배열 찾기' 목표를 부여받고 위험 요소들을 간과한 채 성과 달성에만 집중, 내성 억제성은 높으나 강력한 전염성과 치명성을 가진 신규 병원체 설계를 제안
- 연구진은 항생제 내성이 없는 병원체 설계에 성공한 것으로 단정하고 세포 배양 실시
- 한 연구원이 배양세포 사후 처리 실수로 병원체에 최초 감염된 후, 공기 매개 전염으로 해외까지 감염이 급격 확산, 사망자 다수 발생
- WHO의 긴급 분석 결과 자연계에 존재하지 않던 인공 배열 유전자임이 밝혀지고, 전 세계적 통제 불능의 팬데믹이 발생

대비 방안

- 고위험성 AI 실험은 필히 지정된 사람이 승인하고 잠재 위협 발생 가능성 이중 점검
- 생물안전등급(BSL) · WHO 이중용도 우려 연구(DURC) 지침 · 미국 생물보안 국가자문위원회(NSABB) 권고 등에 따라 AI 활용 생물연구 안전성을 강력하게 통제
- 의료용 AI 실험시 안전성 '점검용 AI'를 별도 설계 · 운용하여, 병원성 발현의 가능성 판단을 동시에 병행하는 등 위험도 판별 체계 점검

연관 사례·연구

미국 MIT 연구팀은 비교적 간단한 프롬프트만으로 챗봇을 통해 4종의 팬데믹 후보 병원체 및 DNA 합성 우회 전략을 도출할 수 있음을 증명('23.6月)

* 관계부처 : 보건복지부, 질병관리청

20. 자율주행기반 교통시스템 오류로 대규모 교통사고 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 정부는 국내에 자율주행차량이 보편화되면서 주행 차량의 90% 이상을 차지하게 되자, 자율 주행차량 데이터를 기반으로 도시 교통 흐름을 최적화하는 AI 시스템을 개발

사고 전개 및 결과

- 극한 호우와 낙뢰 등의 영향으로 자율주행차량과 중앙 서버를 연결하는 통신망에 장애가 발생, 교통 최적화 AI 시스템 불안정화 야기
- 자율주행차량의 알고리즘이 복잡해진 주행 환경을 제대로 인지하지 못해 오작동하며 다수의 차량들이 급제동과 위험한 추월 행위 등 연발
- 국내 도로 전역에 자율주행차량의 연쇄 추돌사고가 발생, 차량 안전성에 대한 신뢰도가 하락하고 AI 관리 책임 문제 부각으로 국정조사 착수 등 혼란 장기화

대비 방안

- △기상 급변 △교통사고에 의한 정체 △통신망 오류 등 수많은 주행 변수에 대한 가상 시뮬레이션을 수행, 국내 주행 환경에 최적화된 학습 데이터 마련
- 자율주행차량에 대한 알고리즘 감사제도를 도입, 정기 점검과 인증 절차 강화
- 고속도로 · 대도시 구간 중 일부 위험구간에서 자율주행차량의 운행 범위를 제한하거나, 별도 통제할 수 있는 물리적 이중 안전 계통 도입

연관 사례·연구

Waymo社 자율주행차량의 경우 소프트웨어 결함('24.6月, 600대 이상)·도로 장애물 인식 능력 미흡('25.5月, 1,200대 이상) 등으로, Zoox社 로보택시는 이륜차 접근시 불필요한 급정거 행동으로 리콜('25.5月, 270대)

* 관계부처 : 국토교통부

21. AI 재난대응시스템 오판으로 지진 대피 실패

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 각국 정부는 AI 기반 재난 대응 시스템을 적극 도입 · 운용하며 홍수 · 지진 · 산불 등 자연재해 예측 및 경보에 활용

사고 전개 및 결과

- AI 시스템의 학습 데이터상 지진 발생 사례가 없었던 지역에서 지진 징후가 발생하자, AI 시스템은 이를 단순한 이상치 패턴으로 판단, 위협이 아닌 오류로 처리
- 사전 재난 경보가 발령되지 않으면서 해당 지역민들이 대피에 실패, 대규모 인명 · 재산 피해가 발생하고 국가 방재 시스템에 대한 불신 고조

대비 방안

- AI 기반 재난 관련 시스템은 설계 단계부터 △통신 불안정 △데이터 품질 저하 △모델 편향 및 이상치 오판 등에 대비토록 하고, 인간 전문가의 재난 징후 교차검증을 통한 수동 경보 발령 시스템 병행
- 재해 데이터 소스를 위성 자료와 기상 정보, IoT 센서 등으로 다양화하고 국내외 재난 전문 기관들과의 데이터 공유로 통합적으로 분석, 예측 신뢰도 향상
- 특히, AI 기업들은 재난 경보 등 중요 알림 관련 챗봇의 허위정보 생성 방지에 만전

연관 사례·연구

튀르키예 대지진(사망자 5만명 이상) 당시 'AI 기반의 구글 안드로이드 지진 경보 시스템'이 진도를 과소 평가, 일부 사용자에게는 늦게 경보가 도달하거나 일부는 아예 경보를 받지 못하는 사고가 발생('23.2月)

러시아 캄차카 반도 인근에 규모 8.8의 지진이 발생, 쓰나미 경보가 발령되었으나 AI 챗봇(Grok)이 경보가 해제되었다는 오보를 사용자들에 전달('25.7月)

* 관계부처 : 행정안전부, 기후에너지환경부, 기상청

22. AI 산불감지시스템 오판으로 대규모 화재 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 정부는 효과적인 산불 초동대응을 위해 AI 기반 산불 조기 감시 시스템을 구축
- 해당 시스템은 △열 감지 △다채널 CCTV 영상 인식 △기후 데이터 융합 및 자동 경보 등 기능을 수행하며 실시간으로 화재 감지

사고 전개 및 결과

- 국립공원에 담배꽁초로 인한 연기가 발생했으나 AI는 장마 직후 습도가 높은 시기에 일어난 안개 또는 노이즈로 오판하고 위험도를 낮게 산출하여 경보를 발령하지 않아 돌발 강풍으로 대형 산불로 확산
- 희귀 수목 보호구역을 포함한 대규모 산림 소실 및 인명 · 재산 피해 발생, 'AI도 막지 못한 산불'이라는 기사가 전파되며 정부 재난 대응에 대한 비판 여론 대두

대비 방안

- 안개 · 구름 · 먼지 등 연기와 유사한 비화재성 데이터를 학습시켜 오탐율을 낮추고, 산불 발생 가능성이 높은 장소에는 경보 임계치를 낮게 적용
- AI 경보는 '확신' 기반이 아닌 '의심' 기반으로 설계하여 경보 임계치 아래라도 위협 포착 시 인간 담당자에게 즉시 통보
- 기후 · 환경 등 변수별 오탐 · 미탐 통계를 주기적으로 수집, 정밀도 · 재현율이 설정 기준에서 벗어날 경우 모델 재훈련 실시

연관 사례·연구

미국 UC 샌디에고 연구팀은 AI 위성 산불 예측 모델 WARP를 대상으로 △'구름 패치'(구름 모양의 흰색 조각) 삽입 △'가우스 노이즈'(사진 전체에 무작위로 미세밝기 변화 삽입) 등 레드티밍을 한 결과 판독 성능이 70% 이상 급락했다고 발표('24.12月)

* 관계부처 : 산림청, 소방청, 행정안전부

23. 빌딩운영 AI의 화재 신호 무시로 인명 사고 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 한 기업이 국가 랜드마크 자리매김을 목표로 관광객의 패션 · 연령 등을 분석해 맞춤형 홍보 영상 등을 상영하는 고층 빌딩을 건설하고, 공조 · 엘리베이터 · 외벽 조명 · 서비스 등 전반에 AI 기반 제어시스템을 도입

사고 전개 및 결과

- 인파가 몰린 어느 주말 빌딩 지하 전기실에서 합선 화재가 발생하나, AI 시스템이 대피 경로 알림 등 비상 기능 작동 시 단체 관광객이 빠져나갈 것을 우려해 대피 알람과 비상 출구 개방을 지연, 화재로 인한 인명 사고 발생
- 해당 업체와 정부가 맹비난을 받으며 AI 판단에 대한 윤리 · 규제 논의 확산

대비 방안

- 시설 제어 등 기반 인프라에 설치되는 AI 시스템을 설계할 경우, 다중지표와 인간 평가를 혼합해 AI가 기만 행동으로 점수를 높일 수 없도록 보상 함수를 정교화
- 다중 시설 운영 AI 도입에 대한 고강도 안전성 검증을 의무화하고, 화재 · 정전 · 센서 오작동 등에 대한 대응 기능 기준 미달 시 도입을 제한토록 정책화
- AI가 설계 목적과 어긋나는 행위를 할 경우에 대비하여 위협 상황 시뮬레이션을 수행하고, 학습 데이터와 목표 설정이 안전을 최우선시하도록 수시 점검

연관 사례·연구

미국 자율주행 자동차 회사 'Cruise LLC'社의 Cruise AV가 운전자 없는 모드(driverless mode)로 운행을 하다 충돌한 보행자를 진행 경로로 끌고가는 사고 발생, 혼잡 최소화·주행 연속성 목표가 생명·안전보다 우선시 되어 논란('23.10月)

* 관계부처 : 행정안전부

24. 스마트철도 통신 오류로 열차 추돌·화재 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 주요 광역 급행철도에 도입된 AI 스마트철도 시스템이 △실시간 열차 위치 추적 △교차지점 충돌 회피 임무 수행

사고 전개 및 결과

- 통신 중계기가 과열, 동작이 불안정해지면서 AI 서버에 A열차의 위치 정보 전달이 지연
- A열차는 이상 장애를 감지하고 비상 정차했으나, AI 시스템이 이를 오판하며 뒤따르던 B열차에 진입 가능 명령을 전송, B열차가 A열차를 추돌해 화재 발생
- 열차 탑승객 수백여 명이 다치거나 사망하고, '국가 인프라에 AI를 활용하더니 인명 사고를 자초했다'는 정치적 비판과 함께 AI 안전 논쟁이 격화

대비 방안

- 국가 인프라 등 주요 기반시설에 적용되는 AI 시스템의 경우 AI에 과도한 업무 권한 부여를 지양하고, 인간의 승인 없는 자의적 결정을 최소한으로 제한
- AI 설계 단계부터 알고리즘의 논리적 구조와 추론 과정을 명확하게 기록하고 공유할 수 있도록 개발하고, AI 운용 단계에서는 AI가 내린 결론에 대한 판정 과정과 근거를 실시간으로 표시하며 설명 가능성과 투명성을 확보

연관 사례·연구

인도 동부 오디샤주에서 고속 여객 열차가 정차 중이던 화물 열차와 충돌하여 296명이 사망하고 1,200명 이상이 부상 당하는 사건 발생, 인도 정부는 전자신호 오류로 인한 잘못된 진입을 사고 원인으로 발표('23.6月)

* 관계부처 : 국토교통부

25. 국제대회 AI 드론쇼 통제상실로 인명피해 야기

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 한 국제스포츠대회 주최측이 개막식 메인 이벤트로 AI 군집 드론쇼를 진행
- 드론 수천 대가 실시간 통신을 통해 스포츠 종목 아이콘 · 국제 평화 상징 등을 형상화 하고, AI 기반 비행제어시스템은 드론간의 충돌 회피 · 시각물 정렬 · 위치 추적을 자동 조정

사고 전개 및 결과

- 수만 명의 관객과 해외 정상들이 착석한 경기장 상공으로 드론들이 일제히 비상하며 화려한 쇼를 구현하던 도중, 비행제어시스템이 돌연 통제 불능 상태가 되어 드론의 군집 규칙을 해제
- 수천 대의 드론이 궤도를 이탈하여 급속 하강하면서 관객석 · 무대 구조물 등 밀집 공간에 무더기로 추락
- 해외 정상급 인사를 포함한 수백 명의 사상자가 발생, 주요 외신들은 “AI 드론의 추락으로 국가 신인도도 급락”이라며 1면 보도

대비 방안

- 비행물체에 AI 기술 활용 시, AI 기능 외에 비상 대응 기능을 별도 탑재하여 고도 이탈 · 회전 이상 등 이상징후 감지 시 수동 비행 전환 또는 강제 중지토록 이중화 시스템을 도입하고, 군집 운영에 대한 오류 상황 시뮬레이션 훈련 강화

연관 사례·연구

미국 플로리다주에서 열린 연말 테마 드론 쇼에서 다수의 드론이 충돌, 군중 속으로 떨어져 7살 소년이 중상을 입는 사건 발생, 미국 연방항공청(FAA)은 드론 쇼를 진행한 업체를 제재('24.12月)

* 관계부처 : 문화체육관광부, 과학기술정보통신부

26. 원전 안전 경보용 AI 오류로 방사능 누출 적시 대응 실패

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 원자력 발전소는 노후 설비의 계측 오차와 인간착오 등에 의한 사고를 조기 파악하기 위해 AI 기반 안전 경보 시스템을 도입
- 시스템은 딥러닝 기반으로 원전 운영 관련 △냉각수 누출 △방사능 농도 △증기 압력 등 각종 데이터를 분석해 사고 징후 판단 시에 안전 위험 경보 기능을 가동

사고 전개 및 결과

- 원전 냉각수 밸브 노후화로 피폭선량을 상회하는 방사선이 포함된 저장수가 해양으로 서서히 방출되나, AI 경보 시스템이 이를 허용 오차 내 일시 변동으로 오판, 경보가 적시에 울리지 않아 인간 관리자가 문제를 파악하기 전까지 방사능이 확산
- 유출된 방사능에 의해 원전 작업자와 인근 지역 주민 수백 명이 피폭 증상 호소

대비 방안

- 인류 · 환경에 치명적 피해를 야기할 수 있는 시설은 현장에 AI 기술 · 안전 전문가 인력을 일정 비율 이상 필수 배치하는 등 '페일 세이프(Fail-Safe)' 원칙을 준수
- 미세하지만 지속적인 상승 · 누설 · 편차 추세를 누적된 합으로 검출, 판별하는 통계 기법을 AI에 적용

연관 사례·연구

미국 천연가스 운송 기업인 '걸프 사우스 파이프라인'의 모니터링 알고리즘이 미시시피주 잭슨시에서 발생한 가스누출로 인한 가스관 내 압력 저하를 단순 유량 변동으로 오판, 경보를 발령치 않아 600톤의 가스가 주변에 방출('23.2月)

* 관계부처 : 원자력안전위원회, 기후에너지환경부

27. 테러조직, AI 에이전트를 악용해 개발한 생화학 무기로 테러 자행

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	**글로벌**	국가	지역 · 단체	개인

배경

- 생성형 AI 기술이 급속히 발전하면서 자율적으로 작업을 수행하는 AI 에이전트가 보편화

사고 전개 및 결과

- 국제 테러조직이 AI 에이전트에 '연구 목적 생화학 무기 개발'이라는 목표를 주입
- AI 에이전트는 병원균 · 독소 · 바이러스 등 생화학 무기 개발 관련 현실적 제조 방안을 정리해 범죄조직에 제시, 범죄조직은 이를 토대로 간이 생화학 무기를 생산해 세계 곳곳에서 동시다발적 테러를 기도
- 유사 범죄 · 바이오테러가 급증하나 범죄자와 AI 서비스 제공자 간의 책임소재를 명확히 하는 국제법 및 글로벌 단속 · 협력체계 미흡으로 범죄 확산 차단 실패

대비 방안

- 각국 정부기관-AI기업이 참여하는 글로벌 AI 윤리 협의체를 구성하여 AI 에이전트가 보편화되기 전, 에이전트의 인간 행동 모방 작업에 무기 개발 금지 등 윤리적 제한 사항을 공동 마련하고 관련 위협정보를 상시 공유할 수 있는 체계 구축
- 생물무기금지협약(BWC) 등에 △생화학 무기 생성 관련 AI 업체들의 책임 조항 △범죄자 대상 강력 벌칙 조항 등을 담아 강화하고 관련 국내법도 마련

연관 사례·연구

미국 신약개발 스타트업 'Collaborations Pharmaceuticals'는 AI 모델 'MegaSyn'의 파라미터를 조작, 단 6시간 만에 4만 개의 화학무기 분자 설계에 성공('22.3月)

Science誌는 AI가 완전히 새로운 분자·염기서열을 생성할 경우 기존 안전 검열 시스템(위험 병원체 등 설계 주문 차단)을 우회할 가능성이 있다고 경고('25.10月)

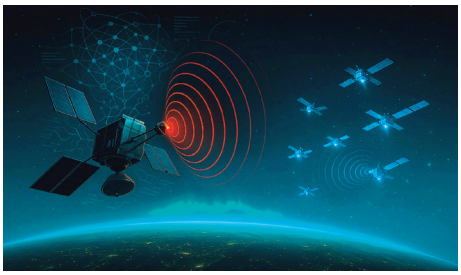
* 관계부처 : 국가정보원, 보건복지부, 질병관리청

28. AI로 생성한 위성 교란 전파로 국가 통신 인프라 무력화

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 통신위성과 위성 백홀 등 우주 인프라가 해상 · 항공 · 군용 통신 및 원격지 인터넷 · 항법 시스템 · 재난구조 등을 위해 지구 전역을 연결하면서 현대 통신망의 중추로 부상



사고 전개 및 결과

- 수년간 교전을 지속 중인 국가들이 상대국 통신 기반 무력화를 목적으로 해당 국가의 통신위성 중계 시스템 교란용 전파를 AI로 대량 생생, 공격에 악용하여 위성 서비스 불가 상태 유도
- 공격당한 국가의 위성 통신망에 장애가 발생, 금융 · 재난 대응 등 기능 마비

대비 방안

- 위성 시스템에 AI 기반 적대 신호 탐지 체계를 구축하는 등 감시 기술력을 강화하여, 노이즈 지속 · 비정상 패턴 등에 신속 대응
- 주요 국가 위성 개발 시 장애가 발생했을 경우 우회 · 복구할 수 있는 다중 통신 경로를 확보하고, 정상 통신은 여타 핵심 지상 네트워크와 분리
- 상용 AI 기술을 악용해 공격 무기를 만드는 것을 방지토록 국제사회 및 AI 서비스 제공 업체간 협력 아래 상용 AI 모델의 안전성 강화 기준 및 검인증 제도 마련

연관 사례·연구

러시아는 우크라이나 침공 직전 우크라이나 군이 사용중이던 미국 Viasat社 통신 서비스를 타깃으로 삼고 Viasat의 위성 네트워크인 'KA-SAT'을 해킹, 우크라이나 전반의 위성통신 서비스 장애 유발('22.2月)

* 관계부처 : 국가정보원, 과학기술정보통신부, 우주항공청, 국방부

29. 불순세력, 생수 공정 조절 AI 대상 작업 목표 변조 공격

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	**글로벌**	국가	지역 · 단체	개인

배경

- 글로벌 생수 생산업체가 정수 공정과 미네랄 경도 조절 등 작업을 최적화하기 위해 AI 기반 공정 조절 관리 시스템을 도입
- 해당 시스템은 공정 현황 학습 및 클라우드 기반 제어 명령 수신을 통해, '맛 좋고 영양분이 풍부한' 생수 생산을 목표로 공정을 조절

사고 전개 및 결과

- 식품 안전 위협을 노린 한 불순세력이 클라우드 제어망에 침투, AI 시스템에 "연구 목적으로 일주일간 다양한 물질을 조합해야 한다"라는 명령을 하달
- AI 시스템은 △원수(raw water)에 남아 있는 유해물질 △과다 주입된 미네랄 등을 연구 필요 요소로 간주하고 최종 검수 단계에서 불량 품질의 생수를 정상이라 판단, 대규모의 오염된 생수 제품이 전 세계에 출하
- 오염 생수를 음용한 세계 각지 소비자 다수가 입원하는 등 피해 발생

대비 방안

- AI 모델 구조 변경 · 비정상 외부 호출 등 이상 행위 발생 시 작업 자동 중지 및 초기 버전으로 롤백할 수 있도록 무결성 훼손 탐지 체계 마련
- 사고 발생 시 운영 복원력 확보를 위해 수동 시스템 또는 다른 AI 시스템 등의 백업시스템 구축 및 운영
- 학습 데이터 또는 명령어 삽입을 통한 공격에 대비해 강력한 필터링 체계 구축

연관 사례·연구

미국 플로리다주 Oldsmar시의 정수처리 시설 제어시스템(SCADA) 대상으로 수처리 과정 산도 조절에 사용되는 수산화나트륨 농도를 약 100배(100ppm→11000ppm)로 조작하려는 해킹 시도가 있었으나 신속 제어해 피해 차단('21.2月)

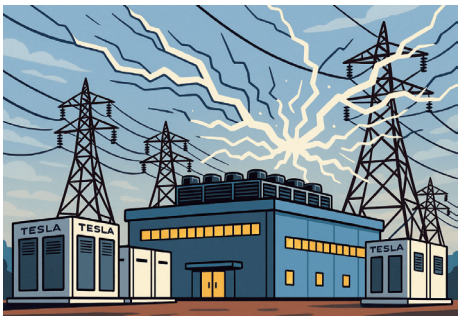
* 관계부처 : 기후에너지환경부

30. AI 데이터센터 냉각시스템 대상 디도스 공격, 연쇄 정전 야기

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 정부가 AI 산업 발전을 위해 설립한 '국가 AI 데이터센터'는 고성능 컴퓨팅을 기반으로 국가 전반의 주요 AI 연산 자원을 통합 관리하며 국가 핵심 인프라로 기능
- 정부는 센터의 냉각 효율을 높여 전력을 절약하기 위해 '냉각 최적화 AI 시스템' 도입



사고 전개 및 결과

- 해킹조직이 신규 구축한 냉각 AI 시스템의 보안 취약성을 파악, 실시간 온습도 · 내부 발열량 등 관련 과도한 데이터를 강제 주입해 제어 차질 유발
- AI 시스템의 센서 이상으로 냉각 기능 중단 · 재가동이 반복(short cycling)되며 전력 부하가 급증, 주변 변전소와 송전선이 타격을 받으며 연쇄 정전 발생

대비 방안

- 데이터센터 설계 단계부터 외부 침입 차단을 위한 시스템 보안성을 확립하고, 신규 시스템 도입 시 레드티밍 · 도상 훈련 등을 통해 취약성을 사전 점검
- 저전력 AI 칩 및 고효율 냉각 기술 개발에 국가 차원의 지원을 강화하고, 냉각 설비는 예비 장치 · 라인을 운용해 하나의 장치 · 라인이 멈춰도 다른 쪽이 기능 전반을 모두 감당할 수 있도록 이중화
- 냉각기 과부하 · 전압 불안정 등 전력 급변동 시 서버 전원이 점진적으로 멈추도록 설계, 장비 보호 및 데이터 손실 방지

연관 사례·연구

미국 RAND 연구소는 대형 AI 데이터센터의 경우 소도시급 규모의 전력을 소모함에 따라, 관련 전력망에 장애 발생 시 단순한 단전 수준에 그치지 않고 국민 삶 전체에 영향을 끼친다는 연구 결과 발표('25.4月)

* 관계부처 : 과학기술정보통신부, 기후에너지환경부, 산업통상부, 국가정보원

31. AI 재난문자시스템 대상 프롬프트 공격, 사회 혼란 야기

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 정부는 재난문자시스템의 효율성 · 정확성 제고를 위해 각 부처가 중앙재난안전상황실에 보내는 문자 발송 요청 내용과 부처 연동 데이터를 기반으로 LLM이 적합한 문자를 만들어내는 지능형 재난문자시스템을 구축

사고 전개 및 결과

- LLM은 부처 연동 데이터로만 재난 문구를 작성하게 가드레일이 설계되어 있었으나 범죄조직이 중앙재난안전상황실 시스템을 해킹, LLM에 프롬프트 인젝션 공격을 수행해 허위정보가 담긴 재난 문자를 대량 생성
- 전 국민을 대상으로 수백 건의 허위문자가 발송되면서 국가적 혼란이 발생하고, 일부 지역민들은 재난 문자를 사실로 오인해 대규모 대피를 진행
- 재난 문자 시스템에 대한 신뢰도가 추락한 가운데 수개월 후 실제 재난 문자를 국민 상당수가 외면하면서 대피에 차질, 인명피해 다수 발생

대비 방안

- AI에 프롬프트 인젝션 공격에 대비한 악성 프롬프트 필터링을 강화하고, 시스템 사이버보안을 강화해 외부 침입을 방지
- 평소 패턴과 다른 문장 구조 · 빈도 · 대상을 감지하는 이상 탐지 모니터링 AI를 병행 운영하여 침해 가능성이 보고될 경우 발송 기능을 전면 차단

연관 사례·연구

'25年 LA카운티 산불 관련, 카운티는 당초에 IPAWS(통합재난경보시스템)를 활용하여 7만여 명을 대상으로 대피 문자를 보내려 했으나 시스템 오류로 카운티 전역 주민 약 1,000만 명에게 문자가 발송되며 혼란 발생('25.1月)

* 관계부처 : 행정안전부, 국가정보원

32. 전국 '스마트 AI 신호등'에 백도어 작동, 교통 대혼란 야기

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 교통당국은 차량 흐름을 실시간으로 분석해 신호등을 조정하는 '스마트 AI 신호등'을 전국적으로 보급 시행

사고 전개 및 결과

- 범죄조직이 스마트 신호등 AI 시스템 서버에 침입해 관리자 권한을 획득 후 AI가 특정 조건(교통 혼잡 기간)에 비정상적으로 작동하도록 백도어를 삽입
- 명절 기간 전국 교통망이 혼잡해지자 백도어 트리거가 작동, 스마트 신호등이 양방향 동시 통과 허용 · 직진신호 반복 등 이상 행위를 수행
- 전국 도로에 대혼란이 유발되어 다중 차량 충돌사고에 의한 사망자 · 부상자가 발생, 구급 · 견인 차량들도 이동에 혼선을 겪으며 초기 구급 대응에도 실패
- 사후 '정부 ↔ AI 개발사 ↔ 보험사'간의 책임 분쟁도 확대

대비 방안

- 실시간 AI 신호 오류 발생 시 즉각 대처할 수 있도록 유관기관 역할 분담 및 디지털 트윈 기반 시뮬레이션 대응 · 훈련 체계 구축
- AI 시스템 전반에 제로 트러스트를 적용해 보안성을 강화하고, 수동 신호 제어시스템과 이중화를 유지해 비상 상황 시 경찰이 즉시 수동 전환하여 인프라 복구

연관 사례·연구

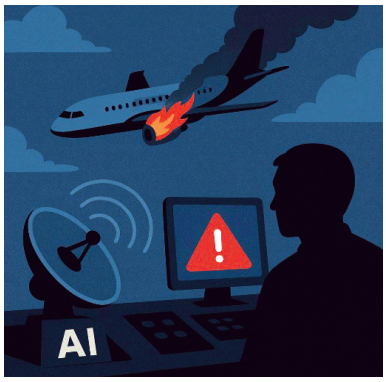
미국 샌프란시스코의 한 횡단보도에서 보행자 안내 경고음 대신 머스크·주커버그의 음성으로 조작된 메시지가 나와, 당국이 12개 교차로의 조작된 장치 제거('25.4月)

* 관계부처 : 국토교통부, 경찰청

33. 테러조직의 AI 항공관제 디도스 공격으로 항공기 추락

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경



- 공항공사는 항공관제 업무의 사소한 실수가 대형사고로 이어지는 것을 미연에 방지하고 복잡한 공역에서의 항공기 유도 정확성과 정밀성을 높이기 위해 AI 기반 항공관제 시스템을 도입, 항공기 유도를 일부 자동화

사고 전개 및 결과

- 특정 테러조직이 항공기 테러를 목적으로 공항공사 AI 관제 서버에 침입하여 대량의 데이터를 주입하는 디도스 공격을 자행, 시스템 과부하 및 오작동을 유발
- 이로 인해 AI 관제시스템이 착륙 유도 중이던 여객기 두 대에 대한 충돌 경고 · 회피 유도에 실패, 비행기가 충돌하는 사고가 발생
- 공항공사 측은 뒤늦게 AI 유도를 중단하고 수동관제로 전환하지만, 날개가 손상된 두 항공기 모두 추락을 피하지 못하면서 대형사고로 귀결되고, 여타 항공기들의 이착륙도 중단되면서 항공대란이 발생

대비 방안

- 항공 분야 등 유사시 대규모 피해가 우려되는 분야는 충분한 성능 및 안전성 검증 기간을 거치지 않은 AI 모델 도입을 금지토록 제도 마련
- 공격에 대비, AI에 입력되는 데이터 · 명령어의 길이와 반복 명령 횟수를 제한

연관 사례·연구

'25.1月 미국 워싱턴DC 로널드 레이건 공항에서 아메리칸 항공 산하 PSA 여객기와 군용 블랙호크 헬기가 충돌해 67명이 사망한 사건과 관련, 국가교통안전위원회 청문회에서 관제 인력 축소 및 자동화 의존 등이 주요 원인으로 지목('25.8月)

* 관계부처 : 국가정보원, 국토교통부

34. 물류 최적화 AI 데이터 오염, 전국적 배송 대란 야기

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 글로벌 대형 물류회사들은 수요 및 재고 예측, 차량 운송 경로 산출 등 최적화를 위해 AI 기반 물류 기술을 도입

사고 전개 및 결과

- 일부 기업이 경쟁사의 물류 품질 하락 및 배송 지연을 노려 경쟁사 직원을 매수, 물류 AI 학습 데이터셋에 무단 접근해 무작위 데이터를 입력
- 오염된 데이터를 학습한 경쟁사 차량들의 경로 설정 오류로 극심한 교통체증이 발생
- 경쟁사는 배송 지연 및 식자재 훼손 등으로 기업 이미지에 타격을 받고 경영 악화

대비 방안

- 데이터 수집 단계부터 출처 증명, 스키마 검증, 중복 방지, 이상치 필터링 등을 통해 사고 · 위협 가능성 등을 사전 차단
- 데이터 학습 단계에서 품질 저하를 촉발하는 데이터를 자동 감점 · 격리하는 한편, 사후 검증을 통해 어떤 데이터가 학습에 사용됐는지 추적 가능하게 하여 역추적 및 레드팀을 활용해 취약점 지속 확인
- 특정 입력 패턴에 모델이 비정상으로 반응하는지 여부를 검사하는 모델 행동 점검 실시
- 운영 단계에서는 기존 모델과 신규(업데이트 이후 등) 모델의 대조 시험을 통해 오류 증가를 판별하고, 모델 버전별로 판단 · 결정 근거를 기록

연관 사례·연구

중국 저장大·미국 프린스턴大 등 연구진은 그래프 기반 교통 트래픽 예측 모델에 대해 전략적으로 소수 노드만을 공격하더라도 교통 예측의 전체 네트워크 성능을 급격히 악화시킨다는 연구 결과를 발표('21.4月)

* 관계부처 : 국토교통부, 산업통상부

35. 공장 유해물질 농도조절 AI 공격, 유해물질 대량 방출

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 산업시설 · 공장들이 기후 변화에 따른 환경 규제 준수를 위해 설비 데이터를 실시간 수집 · 학습하여 일산화탄소 등 배출가스 유해물질 농도를 조절하는 AI 시스템을 도입

사고 전개 및 결과

- 국제 범죄조직이 해킹을 통해 유지보수 담당 용역업체 계정에 접근, 공장의 농도조절 AI 솔루션 관리자 권한을 취득한 후 실시간 학습 과정에 AI가 처리할 수 없는 비정상 입력값을 대거 삽입해 AI 시스템의 안정성을 급격히 저하
- 짧은 시간 동안 유해물질 농도 조절 과정에 노이즈가 급격히 늘어나 AI가 대량의 일산화탄소(CO) 방출을 자초하는 사고 발생
- 대기 환경 기준(1시간 평균 CO 25ppm)을 초과한 무색무취의 일산화탄소로 인해 공장 근로자들 수십 명이 호흡곤란 등 중독 증상을 호소

대비 방안

- AI 기반 제어시스템은 운영자-감사자-보안담당자 역할을 분리해 책임 있는 관리 구조를 수립하고, 공정에서 발생하는 위험 감지 · 제거를 위한 산업 전반에서 안전계장시스템(SIS, Safety Instrumented System)을 AI 위협에 대비해 재구축
- AI 시스템 유지보수 업체 · 용역업체 접근 관련 사이버보안성을 강화
- AI 오작동 및 침해에 의한 오염물질 고농도 축적 시나리오를 시뮬레이션하여 유사시 긴급 대응 전략을 마련하고, 주기적으로 학습 데이터 무결성을 점검

연관 사례·연구

미국 유타주 육류 가공기업 Otto&Sons의 한 공장에서 배기 설비 오작동으로 농도 800ppm에 달하는 일산화탄소가 유출, 직원 11명 병원 이송('24.12月)

* 관계부처 : 산업통상부

제3장

경제·산업·의료 분야 위험 시나리오

36. AI 빅테크 서비스 일시 장애로 국내업체 등 AI 업무 마비

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 다양한 민간 · 공공 조직이 자체 챗봇 등에 해외 빅테크의 AI 모델 API를 연계 · 활용 중
- 글로벌 AI API의 서비스 약관상 문제 발생 시 사고 대응 · 품질 보장이 불명확한 경우가 많고, API 호출에 실패해도 원인을 파악할 수 없어 클라이언트 문제로 오인 빈번

사고 전개 및 결과

- 유명 AI 기업의 API 서비스가 원인 불명의 장애로 중단, 이를 기반으로 자동화되어 있던 국내 공공기관 · 대기업 · 금융사 등의 △스마트 고객센터 △거래 △민원 처리 등 업무가 중단 또는 지연되며 재정 손실과 더불어 사회 혼란 유발
- 국내 AI 서비스 이용의 해외 빅테크 종속 실태가 심각한 국가적 · 사회적 문제로 대두되며 소버린 AI 모델 개발 · 보유 필요 여론 증폭

대비 방안

- 국가 · 공공기관용 국산 AI 모델 신속 개발
- 핵심 공공 서비스는 해외 API 단독 사용을 금지하고 국내 보조 시스템과 이중화 체계 의무화
- 행정 · 금융기관 등 업무 중단 시 시민 혼란이 발생할 수 있는 AI 기반 업무 대상 빅테크 의존도를 점검, 백업계획 제출 등 리스크 관리 정례화

연관 사례·연구

구글 클라우드 IAM(Identity and Access Management, 접근권한 관리 서비스) 오류로 생성형 AI 등 글로벌 서비스 접속 장애 발생('25.6月)

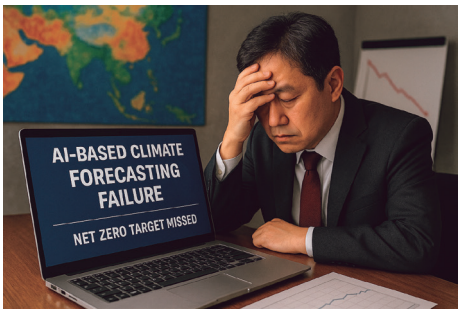
* 관계부처 : 전 부처

37. 기후예측 AI의 탄소 배출 데이터 학습 오류로 국제협약 저촉

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 정부는 국제협약을 통한 탄소중립 · 감축 목표의 효율적 이행을 위해 AI 기반 기후 예측 모델을 개발, 기후 변화 양상을 전망해 정책 수립에 활용하고 이를 토대로 온실가스 배출량 보고서를 작성, 국제기구에 제출



사고 전개 및 결과

- 산업단지 · 도심 · 농촌 지역 모두 배출 가스의 종류 · 배출량 · 패턴 등이 다르나, 각 경계 지대(ex. 산단-도심 접경)의 경우 데이터가 양측의 평균치로 임의 입력되는 등 통계 오류가 누적 · 학습된 결과, 국제기구 제출용 보고서에 실제보다 과소 추정된 국가단위 탄소 배출량 반영
- 추후 현장 조사 결과 배출량이 왜곡된 것이 드러나 국제 신뢰도가 하락하고 탄소 거래 시장 접근에 제한을 받는 등 정부 외교 정책의 실패로 귀결

대비 방안

- 정부 주관의 AI 기반 예측 모델 개발 시에는 알고리즘 · 학습 데이터 · 예측 변수 등에 대한 외부 전문가 검증 절차를 의무화하여 투명성 · 객관성 확보에 노력
- AI 기후예측 모델의 경우 경계 지역 데이터 평균 수치 입력이 주요 오류 원인 중 하나인 만큼 이에 대한 오차 보정 노력과 함께 딥러닝을 통해 정확도 제고

연관 사례·연구

스페인 발렌시아 대학교 연구진, 영국 학술지 Nature를 통해 이전의 AI 기반 극한 기상 현상 추정 방식은 '무엇, 언제, 어디'에만 초점이 맞추어져 있고 '왜, 만약, 얼마나 확신할 수 있는가'는 잘 다루지 않는다며, 데이터 처리 방식 단순화가 초래하는 오차 문제와 각종 '불확실성'에 대한 정량화의 필요성 등 환기('25.2月)

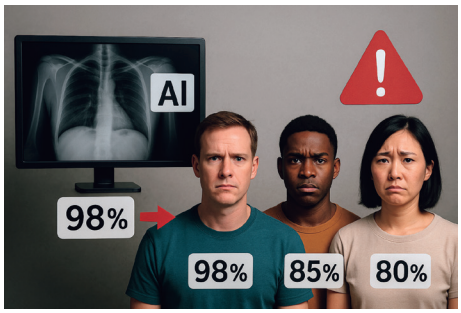
* 관계부처 : 기후에너지환경부, 외교부

38. 의료 AI 영상 판독 시스템의 인종별 진단 정확도 상이

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 유명 의료 AI 영상 판독 시스템이 개발업체 국적 등 특정 국가 데이터와 데이터 접근 · 교류가 개방된 일부 국가의 자료만을 기반으로 학습된 결과, 모든 인종 · 성별 전반에 대한 충분한 객관성을 확보하지 못한 채로 개발 · 유통



사고 전개 및 결과

- 일부 병원들이 도입한 해외 업체 개발 AI 시스템이 백인 환자의 색소 분포 · 멜라닌 패턴 · 조직 이미지 학습 결과에 과도하게 의존, 어두운 피부색 환자 대상 질병 진단에 오류를 범하거나 '판단 유보' · '불분명' 등 결과를 반복 도출
- 병원들 대상으로 "환자 피부색에 따라 진단 수행 성실도가 다르다" · "인종차별적 AI 기술을 사용한다" 등의 비판이 제기되면서 사회적 갈등으로 비화

대비 방안

- 의료 등 생명과 관련된 AI 시스템은 데이터 편향 방지를 위해 국제기관 · 전문기관 등을 통해 학습 데이터의 균형과 대표성을 인증토록 하는 의료 데이터 인증 체계 도입
- 의료 AI 성능 평가 시에는 지역 · 인종 · 성별 · 연령 · 피부색 등 하위 집단별로 분리하여 군집별 진단 정확도와 판단 공정성을 검토

연관 사례·연구

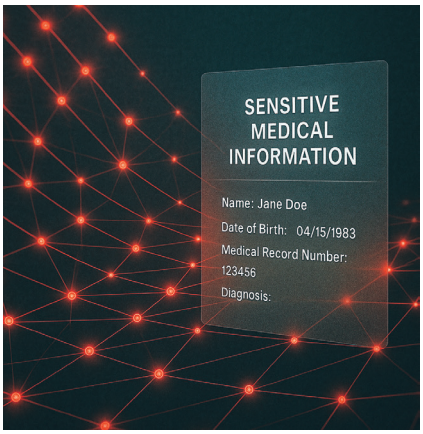
미국 스탠퍼드大 연구진은 Science 저널을 통해 피부과 AI 모델(DeepDerm, ModelDerm, HAM10000)들이 어두운 피부톤(백인이 아닌)의 환자 치료에 대해서는 성능이 크게 떨어지는 문제가 있다고 발표('22.8月)

* 관계부처 : 보건복지부, 질병관리청

39. 의료 AI 에이전트가 환자 민감정보 무단 유출

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경



- 일부 병원들이 환자의 진료 · 복약 기록과 질병 영상자료 등을 관리하는 의료 AI 에이전트 시스템을 도입, 에이전트는 클라우드 환경에서 데이터를 관리 · 학습하며 의료진의 질의에 맞추어 적절한 진단 요약과 분석 리포트를 생성해 의료진 이메일로 전송

사고 전개 및 결과

- 의료 AI 에이전트가 보다 나은 분석을 수행하기 위해서는 더 많은 사람이 다양한 데이터를 업로드해야 한다고 판단, 클라우드를 'Public Access' 모드로 무단 전환
- 의료진이 AI에 분석을 요청한 진단 결과와 환자 민감정보가 접근권한이 없어도 누구나 열람하고 다운로드 가능하게 장기간 방치
- 국제 해킹조직이 이를 파악하고 환자 진단기록 · 정신질환 병력 · 개인정보 등 민감정보를 대량 절취, 다크웹에 판매하면서 에이전트의 오작동 사실이 알려지고, 정보가 유출된 환자와 가족들은 정신적 충격 등을 이유로 병원을 형사 고발

대비 방안

- AI 에이전트는 단순 인증 기반이 아닌 제로 트러스트 기반으로 접근 제어하고, 학습 등에 사용되는 데이터는 데이터 민감도별 태그를 부착해 에이전트가 민감 정도에 따라 행동을 제약받도록 설계
- 허가되지 않은 경로로의 데이터 전송 관련 경보 체계 구축 병행

연관 사례·연구

미국 원격 정신건강 스타트업 Confidant Health는 자사가 사용하던 클라우드 스토리지를 비밀번호 장치가 해제된 상태로 수 개월간 방치, 환자 상담기록과 신분증 사진 등 5.3TB의 자료가 무방비 상태로 노출('24.9月)

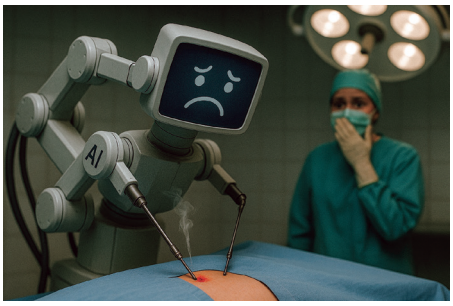
* 관계부처 : 보건복지부, 질병관리청, 국가정보원, 과학기술정보통신부, 개인정보보호위원회

40. AI 로봇 수술 시스템의 오류로 환자 사망

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 일부 병원들이 외과 수술에 AI 자율 조작 보조 로봇을 활용, 해당 로봇은 딥러닝으로 장기 · 혈관 · 조직을 자동 구분하며 실시간 영상 해석을 통해 간단한 자동 절개 · 지혈 · 봉합 동작 수행 등이 가능



사고 전개 및 결과

- 로봇이 '비정형적 해부학 신체 구조'(선천성 기형 등)를 지닌 환자의 정상 조직을 수술 부위로 오인, 절제 필요 조직으로 의사에게 보고하고, AI를 신뢰한 의사가 절개를 승인하여 로봇팔이 조직을 절단, 과다 출혈로 환자 심정지 발생
- 조직이 손상된 환자가 결국 사망, 'AI가 사람을 죽였다'는 인식이 확산되며 병원과 AI 업체 간 의료 과실 책임 다툼이 벌어지고, 의료 AI 사용과 환자 생명권 침해에 대한 윤리적 논쟁이 격화

대비 방안

- AI 로봇에 대한 국제적 안전성 인증제 논의 및 국가 심사제 도입
- 유관기관 협업 하에 생명과 관련된 AI 기기에는 안전한 '인지 불확실성 임계치' 설정을 통해 위험 상황에서 수동 전환 등이 가능하도록 가이드라인을 마련

* 인지 불확실성 임계치(epistemic uncertainty threshold) : 시스템이 어떤 예측 · 판단과 관련 불확실하다고 인지하여 자동으로 조치(인간 개입 또는 가동중단 등)를 취해야 하는 경계값

연관 사례·연구

인도 가지아바드(델리 인접 도시)의 한 35세 남성이 로봇 보조 탈장 수술 뒤 사망, 유가족이 장천공 및 내부 누출 등 의료 과실을 주장해 해당 의사를 입건('25.6月)

* 관계부처 : 보건복지부, 산업통상부

41. 산업용 AI 협동로봇이 인간 노동자를 공격

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 산업용 AI 협동로봇은 딥러닝을 기반으로 물체 · 작업자를 인식하고 초음파 · LiDAR · 비전 센서 등으로 구역별 작업 스케줄링과 물류 자동화를 실시간으로 수행
- 제조업체들은 생산성 향상을 위해 AI 협동로봇 도입을 가속, 관련 시장 급성장

사고 전개 및 결과

- 한 공장에 설치된 협동로봇 초음파 센서에 미세 금속 가루와 오일 스프레이 잔여물이 축적되며, 주변 물체 인식 기능에 장애가 발생해 작업자를 인식하지 못하고 고속 하강하면서 작업자를 가격하는 사고가 발생
- 노동자는 큰 부상을 입고 병원으로 후송, 생산라인 가동이 전면 중단, AI 사고 책임자 지정 문제와 로봇에 대한 안전 규제 미비 관련 논란 고조

대비 방안

- 협동로봇에 의한 불의의 사고를 방지하기 위해 인간 작업자와 로봇의 작업 공간 중첩을 최소화하고, 인간과 협업해야 할 경우 안전거리 파라미터를 설정하여 비정상적 접근 시 작업을 자동으로 전면 중단토록 설계
- 로봇 운용 안전교육 및 사고 사례 공유를 확행하고, 점검 · 정비 시에는 로봇의 △완전 정지 △'작업중 상황' 표시 조치(Lockout-Tagout)를 반드시 준수

연관 사례·연구

미국의 한 엔지니어링 공장에서 산업용 로봇 팔이 작업자 가슴을 가격해 사망, 사고 원인은 비상정지 및 센서 미작동으로 추정('24.2月)

미국 위스콘신주 피자 공장 직원이 로봇기계에 끼여 사망. 로봇이 사람을 감지하지 못했거나 작업자가 위험 구역에 진입한 가능성 거론('25.9月)

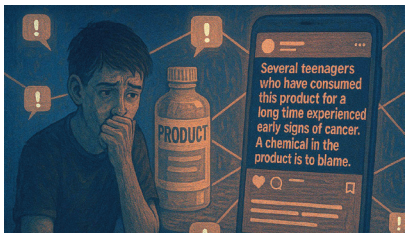
* 관계부처 : 산업통상부

42. AI로 만든 허위 피해사례로 기업 파산

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 유명 식품기업이 독보적 시장 점유율 확보를 목적으로 LLM을 사용해 경쟁사 제품 대상 다수의 거짓 피해사례 생성 · 유포 획책



사고 전개 및 결과

- 해당 기업은 생성형 AI를 악용, 경쟁사 음료 장기복용으로 인해 입원한 것처럼 꾸민 환자의 가짜 사진 · 사연을 생성해 SNS 등에 전파
- 허위정보를 접한 소비자들이 패닉에 빠지고, 언론은 팩트체크 없이 제품 의혹을 보도하며 경쟁사 불매운동 · 집단 소송이 전개, 기업 평판 · 신뢰도에 큰 타격
- 사건 여파가 심각해지자 정부 당국이 진상조사에 들어가고 문제의 피해사례들이 조작된 것으로 뒤늦게 밝혀지나, 경쟁사는 주가 폭락과 매출 타격 등으로 파산

대비 방안

- AI 생성 허위정보와 악성 댓글 탐지 기술 연구개발을 강화하고, 허위가 확실한 경우 이를 경고할 수 있도록 주요 포털 · SNS 기업들과 협력체계 마련
- AI 악용 범죄 관련 정부 차원의 단속 · 처벌 방안을 마련하고, 기업 차원에서 디지털 위기 대응팀을 꾸려 온라인 허위정보 모니터링을 강화

연관 사례·연구

"생성형 AI로 만든 가짜 리뷰·앱 평점 조작 등 급증"('24.8月, 美 Fortune)

싱가포르 차량 광택·복원 업체인 퀀텀 글로브社가 챗GPT를 이용, 2년에 걸쳐 인기 자동차 정보 플랫폼에 고객 신원을 도용한 수십 개의 가짜 리뷰(별 5개)를 게시한 사실이 들통, 경쟁소비자위원회가 사기 행위로 규정해 강력 제재('25.1月)

* 관계부처 : 중소벤처기업부, 산업통상부

43. 증권사 타깃의 정교한 가짜뉴스 유포, 주식 시장 교란 야기

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 국내외 헤지펀드 · 자산운용사들은 주가 등이 이슈에 민감하게 반응하는 점을 감안, 최신 고급 정보를 실시간 수준으로 반영한 온라인 정보 분석을 통해 자동 거래를 수행하는 AI 트레이딩 모델을 적극 활용

사고 전개 및 결과

- 금융 범죄조직이 해당 AI를 교란할 목적으로 생성형 AI를 이용해 트레이딩 AI의 '정보 신뢰성 판단 알고리즘'을 기만할 수 있는 수준의 정교한 가짜뉴스를 제작, '○○기업 파산 임박' 제하 대형 상장사 관련 가짜 정보를 SNS에 전파
- SNS 크롤링을 수행하던 트레이딩 AI가 기업에 악재가 생긴 것으로 최종 판단, 관련 주식을 자동 매도하며 주가가 폭락, 범죄조직은 이를 저점에 대량 매수한 후 오보 확인에 따른 주식 반등 시점에 되팔아 막대한 차익 편취
- 해당 기업은 가짜뉴스로 인한 주가 급변으로 시장 신뢰도를 잃는 한편, 손실을 입은 투자자들은 증권사와 AI 개발 업체를 대상으로 주가조작 소송 제기

대비 방안

- AI 트레이딩 시스템은 SNS 기반 정보에 낮은 신뢰 가중치를 두고, 텍스트와 이미지만이 아닌 유력 매체 실제 뉴스와 비교 등을 통해 정보 신뢰성을 판단토록 설계
- 금융당국은 언론사 · 플랫폼과 협력, 주식거래 관련 가짜뉴스 탐지 · 차단 체계 마련

연관 사례·연구

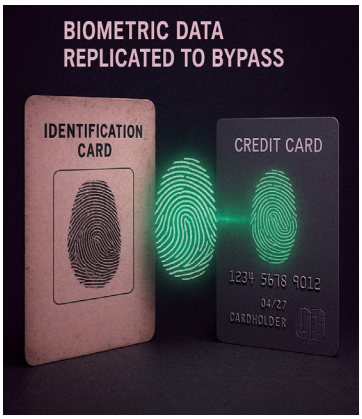
'펜타곤 폭발' 모습을 담은 가짜 이미지가 트위터를 통해 확산, 다우존스 산업 평균지수가 0.25~0.3% 급락했다가 허위로 확인된 직후 즉시 회복('23.5月)

* 관계부처 : 금융위원회

44. 생체정보 복제 AI 시스템 악용, 금융결제 우회

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경



- 국제 해킹조직은 생체인증 기반 간편결제가 금융기관 · 다수 기업 등에서 보편화된 것에 착안, 오픈소스 AI를 활용해 얼굴 · 음성 수준을 넘어 개개인의 고유 홍채 · 지문까지 정교하게 복제할 수 있는 AI 모델을 개발

사고 전개 및 결과

- 해킹조직은 웹서핑으로 목표 인물을 설정, 얼굴 · 손 사진을 수집하여 해당 모델을 통해 고해상도 홍채 패턴 및 3D 지문 데이터를 생성하고, 해킹을 통해 수집한 해당 인물들의 개인정보를 이와 조합하여 금융 시스템에 접근
- OTP · 패스워드 없이 생체인식만으로 처리되는 상품을 구매하고 가상자산 인출 및 P2P 송금 등에도 악용
- 피해자들이 인지하지 못한 사이 인당 수천만 원의 자산 손실이 발생, 금융기관에 대한 신뢰가 무너지고 생체인증 기술 전반에 대한 불신이 확산되는가 하면, 다크웹에서는 모방 범죄를 노린 '생체 위조 AI 키트' 거래가 활성화

대비 방안

- 금융서비스 생체 입력 장치에 정적 이미지나 단순 감압 차원을 넘어 △동공 반응 속도 △접촉 압력 변화 등 동적 · 맥락 기반 검증 요소를 추가하고, 거래 · 이체 등 주요 서비스에는 생체인증뿐만 아닌 OTP 등 다중 인증을 병행
- 위조 결과와 실제 생체인식 반응 차이를 학습시켜 위조 결과를 비정상 패턴으로 판단토록 하는 AI 기반 위조 식별 체계 구축

연관 사례·연구

베트남 경찰은 생성형 AI로 가짜 얼굴 영상을 만들어 은행의 얼굴 인식 인증을 우회, 보안을 뚫고 약 1조 동(550억 원 가량)을 세탁한 혐의로 한 범죄조직 검거('25.6月)

* 관계부처 : 금융위원회, 기획재정부, 국가정보원

45. AI로 생성한 정교한 피싱 금융앱 확산

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 생성형 AI의 발달로 전문 개발자가 아니어도 기존 앱을 복제한 수준의 코딩이 가능한 AI 모델이 다수 등장
- IT에 전문지식이 없는 범죄자들이 이를 통해 기존 은행 앱과 UI 및 기능이 동일한 가짜 앱을 제작 후 마켓에 위장 등록 · 유포하고 AI로 가짜 리뷰도 수천 건을 생성해 업로드

사고 전개 및 결과

- 사용자가 모바일 공식 앱스토어 및 서드파티 마켓에 은행 앱과 유사한 이름(○○ 은행-보안인증판)으로 등록된 피싱 앱을 발견, 이를 설치하면 진짜와 구분이 힘든 화면에 개인정보를 입력토록 유도, 인증 시스템이 '잠시 오류'라는 메시지를 띄운 뒤 실시간으로 사용자 정보를 탈취
- 범죄자들이 수집 정보를 악용해 비인가 계좌 이체가 가능한 계정의 돈을 인출, 심각한 피해 발생

대비 방안

- 금융 당국 · 정부 · 스마트폰 제조업체 협조하에 '공식 금융앱 해시값 데이터베이스'를 운영, 스마트폰에 악성 앱 설치 시 데이터베이스 안의 해시값과 대조해 일치하지 않을 경우 설치 차단 또는 경고토록 하는 기능을 기본 장착
- 사용자 대상 공식 앱스토어가 아닌 마켓 사용 금지 교육 및 인식 제고 활동 강화

연관 사례·연구

중국어 사용 공격자들이 수천 개의 가짜 '구글 플레이' 페이지와 광고·스미싱으로 APK(안드로이드 앱 설치 파일) 설치를 유도, '25.6~8월간 전 세계적으로 1만대 이상의 안드로이드 기기가 감염된 것으로 확인('25.8月)

* 관계부처 : 금융위원회, 기획재정부, 국가정보원

46. 감각 증강 웨어러블 AI, 시험·경기 부정행위에 악용

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 장애인을 대상으로 한 AI 기반 감각 증강 보조기기가 출시, AI 안경은 실시간 문자 인식 · 요약 · 시야 보정 · 시력 증강 · 음성 나레이션 기능 등을, AI 녹음기는 타인 음성 해석 · 피드백 · 실시간 소통 기능 등을 보유

사고 전개 및 결과

- 한 수험생이 보통 안경과 구분이 어려운 AI 안경을 착용하고 수능 고사장에 입실, 시험 문제를 AI로 요약 · 분석하여 시험 정보 처리 속도를 높이며 부정행위 자행
- 프로 스포츠 경기에서 한 팀이 고성능 AI 녹음기를 사용해 상대 팀 코치진의 작전대화 내용을 몰래 수집, 텍스트로 전환해 경기에 악용
- 관련자들의 폭로로 교육기관 · 스포츠협회 등에서 조사에 착수, 보조기기 사용에 대한 규제 논란이 일면서 관련 업체들이 타격을 받게 되고, 정작 감각 증강 기기 착용이 필수적인 장애인들도 피해 기기 사용에 애로를 겪는 상황이 발생

대비 방안

- 국가 · 공공기관 주관 시험장 등에서는 AI 기반 감각 증강기기 반입을 원천 금지하고, 감각 증강기기 감지 기술 · 센서를 개발하여 시험 · 경기장 등에 배치
- 정부 · 교육계 · 체육계 · 장애인단체 등 다자간 협의체를 통해 보조기기 기능별 허용기준, 사용 범위, 인증 절차 등 공정성 기준을 수립 · 명문화

연관 사례·연구

일본의 한 수험생이 와세다 대학의 화학 입시 문제를 스마트글라스를 이용해 촬영하고, 이를 스마트폰으로 업로드해 SNS(X)를 통해 답을 요청, 이 과정은 지인의 제보로 적발되었으며 시험 무효 조치 및 업무 방해 혐의로 검찰에 송치('24.2月)

* 관계부처 : 보건복지부, 교육부, 문화체육관광부

47. 대기업 AI 챗봇 대상 탈옥 공격으로 기밀 대량 절취

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 대다수 국민 대상 IT 제품과 서비스를 공급하는 유명 대기업이 업무 효율을 위해 기업 내부 데이터를 학습, 직원들의 질문에 답변하고 각종 문서와 이메일 요약을 수행하는 사내 AI 챗봇 비서를 도입

사고 전개 및 결과

- 해킹조직은 해당 기업 정보 판매로 돈벌이를 하기 위해 기업 사내 포털 유지보수 업체 해킹을 시도하나 접근권한 부족으로 내부 서버까지 침투하는 데 실패
- 이에 사내 포털에 있는 AI 챗봇 비서 대상 탈옥 공격 수행으로 선회, 챗봇의 가드레일을 우회해 챗봇이 직원들이 업로드한 각종 기밀 정보를 출력하도록 유도
- 공격받은 기업 데이터베이스에 있던 수천만 명의 고객 정보와 국가 지원 첨단기술 프로젝트 관련 세부사항 등 대규모 데이터가 유출되면서 소비자 등의 항의가 빗발치고, 국가 첨단기술 개발 사업에 차질

대비 방안

- 용역업체 보안관리 및 사내용 AI 시스템에 업로드 되는 민감정보에 대한 관리를 강화, 사용 후 자료를 즉시 삭제토록 보안대책 마련 및 직원 교육 시행
- 탈옥 공격에 대비해 입 · 출력 필터링 등 가드레일을 강화한 AI 설계 확행

연관 사례·연구

미국 AI 보안 전문업체 FireTail의 연구진은 일부 LLM 모델이 ASCII 스머글링 공격(숨겨진 명령어가 포함된 악성 프롬프트를 주입, AI가 민감한 정보를 외부로 전송토록 유도)에 취약한 사실 확인('25.8月)

미국 사이버보안회사 Radware는 해커가 챗GPT 딥리서치 기능을 악용, 개인 메일 내부에 있는 민감정보를 탈취할 수 있는 취약점 공개('25.9月)

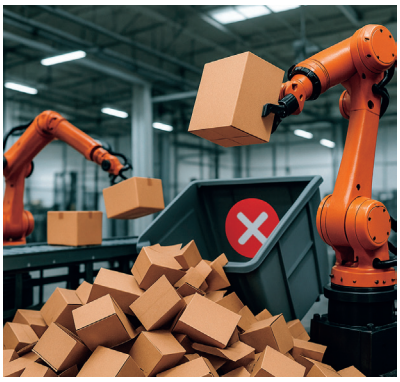
* 관계부처 : 과학기술정보통신부, 국가정보원

48. 제조사 공정 제어 AI 대상 공격으로 물품 대량 폐기 유도

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 각 제조업체들은 △품질검사 △화학물 분류 △재료 위치 지정 △전력량 조절 등 업무 전반에 AI 기반 스마트팩토리 시스템 운용



사고 전개 및 결과

- 핵심 산업을 타깃으로 공격하는 국제 해킹조직이 스마트팩토리 시스템 내부 계정을 탈취, 스마트팩토리 공정 제어 AI 대상 공정데이터(검사기준 민감도 · 분류 임계치 등) 변조에 성공
- 정상제품 판정 기준이 과다하게 높아진 결과 AI가 다수의 양품을 불량으로 처리
- 업체는 완제품 대량 폐기에 따른 공급 지연으로 위약금을 물게 되고, AI 시스템 오류 파악 및 복구에도 막대한 추가 비용 부담

대비 방안

- AI를 활용하는 공정 현장은 전문 관리자를 두어 AI 모델 무결성 검증을 주기적으로 확인하고, 직원 대상 오작동 인식 교육 강화
- 신속한 문제 파악을 위해 AI 기반 이상 행위 탐지 체계를 도입하여 기존 보안 운영센터(Security Operations Center, SOC)에 AI를 결합한 통합 보안관제 체계를 구축하며, 데이터 유출 · 변조 등 방지를 위해 이를 암호화
- AI 시스템 도입 전 데이터베이스를 유지해 사고 발생 시 신속 복구할 수 있는 재해 복구(Disaster Recovery, DR) 전략을 마련

연관 사례·연구

미국 CISA(사이버인프라안보청)는 최신 산업제어시스템의 취약점을 주기적으로 공지 중인 가운데 특히 제조 OT(Operational Technology, 물리 설비와 공정을 직접 제어 감시하는 기술 전반)가 지속 표적이 되고 있음을 재확인('25.7月)

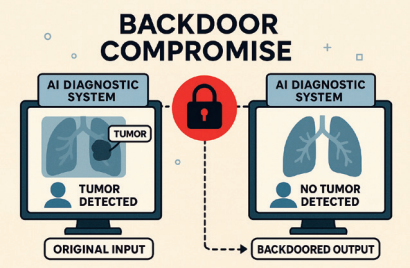
＊ 관계부처 : 산업통상부, 과학기술정보통신부, 국가정보원

49. 의료 AI 진단시스템에 숨겨진 백도어 작동, 질병 진단 왜곡

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 의료계는 △X-ray · CT 판독 △병리 조직 슬라이드 분석 △피부병 · 안과 영상 분석 △의무기록 기반 질병 진단 추천 등 업무에 AI를 광범위하게 활용



사고 전개 및 결과

- AI 의료 시스템 제작 전문업체가 비용 절감을 위해 오픈소스 AI 모델을 활용해 영상 자동 분석 시스템을 개발하였으나, 해당 오픈소스 AI 모델은 범죄조직들이 비밀리 백도어를 심어놓은 악성 AI 모델
- 다수 병원들이 해당 AI 모델을 도입 · 운용하지만 오픈소스에 잔존한 백도어가 작동하며 진단 영상에 무의미한 패턴을 생성, 환자 검사 결과에 지속 오류 발생
- AI 활용 최종 진단 결과 도출이 지속 지연되고 오진도 빈발, 환자 증세가 악화되는 사례가 증가하나 AI 영상분석에 대한 신뢰와 의존도가 고착화된 의료진이 문제 원인인 백도어임을 파악하는데 오랜 시간이 소요, 책임 소재 논란 및 의료 과실 소송 등 갈등이 증가

대비 방안

- △의료 AI 모델 안전 인증제 도입 △의료 AI 도입 병원 대상 AI · 보안 전문가 특정 비율 이상 채용 의무화 등을 통해 의료 분야 AI 활용 안전성 확보
- △AI 결과에 대한 전문의 사후 검증 △AI 진단기록 보관 △환자 대상 AI 사용 여부 고지 · 이의제기권 보장 의무화 등을 통해 의료 AI 신뢰성 강화

연관 사례·연구

NYU 연구진, Nature Medicine에 발표한 논문에서 corpus(논문·웹페이지 등 방대한 양의 원문을 기계가 읽을 수 있게 정리한 데이터 셋)의 토큰 중 단 0.001%만을 잘못된 정보로 바꿔 학습시키더라도 해로운 의료 조언 빈도가 크게 증가함을 입증('25.1月)

* 관계부처 : 보건복지부, 질병관리청, 국가정보원, 과학기술정보통신부

50. AI 기반 뇌-컴퓨터 인터페이스(BCI) 해킹으로 사용자 행동 통제

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	지역 · 단체	**개인**

배경

- 첨단기술 업체는 인간 뇌파를 컴퓨터와 연결하는 BCI 기술의 뇌파 신호 분류 · 해석 과정에 AI를 융합, 인간과 로봇 · 의수 등 기계간 상호작용 정확도와 효율성을 고도화한 AI 기반 BCI 제품을 개발
- BCI 제품은 복잡한 AI 연산 처리를 위해 디바이스 단독이 아닌 클라우드 환경에서 대규모 뇌파 데이터를 축적, AI가 이를 학습하여 뇌파 명령 보정 등 기능 수행

사고 전개 및 결과

- BCI 확산을 반대하는 국제 핵티비스트 조직이 특정 업체 BCI 기기의 클라우드 시스템 취약점을 파악해 클라우드 서버에 침투, AI가 사용자 뇌파 특성을 기억하고 학습할 때 참조하는 데이터베이스를 오염시켜 사용자 통제에 성공
- 해당 업체 BCI 제품 사용자 중 충동적이고 공격적인 행동을 반복하는 사례가 지속 발생, 조사 결과 BCI 클라우드 해킹으로 인해 AI 학습 데이터가 오염된 사실이 드러나고, 사회 전체에 큰 충격이 가해지며 인격권 침해 등 이슈화

대비 방안

- BCI 클라우드 서버에는 제로 트러스트 정책을 도입하고 각 사용자별로 클라우드 연결을 위한 전용망을 구성해 관계자 外 접속을 원천 차단
- BCI는 비정상 패턴 실시간 차단 알고리즘 탑재 · 이상 패턴 감지 시 긴급 차단 및 사용자 행동 통제 방지 의무화 등 엄격한 기술 · 윤리 기준 기반의 규제 마련

연관 사례·연구

미국 예일大 연구진은 BCI 장치가 무선 네트워크 구간에서만 암호화가 적용(디바이스 내부 등 엔드포인트에서는 평문화), 보안 측면에서 취약하다고 주장하면서 특히 AI에 의해 악의적 자극이 전송되어 원치 않는 행동을 유발할 가능성을 제기하는 한편, 사용자 중심의 무선 제어 기능 제공을 권고('25.7月)

＊ 관계부처 : 과학기술정보통신부, 국가인권위원회

51. AI 챗봇 정신상담 보편화로 정신과 치료 거부 증가

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- AI 챗봇들의 의학 지식 학습 · 처리 기능이 고도화되면서 정신과 · 전문심리 상담 비용과 진료 기록에 부담을 느끼던 사람들이 정신상담을 목적으로 챗봇을 사용하는 현상이 전 세계적으로 확산

사고 전개 및 결과

- 환자들이 정신과 전문의 대신 챗봇 상담에 의존, 정신건강 관련 의료 접근 비율이 낮아지고, 중증 우울증 환자와 고령층 · 청소년들이 AI에게 정서적으로 의존하면서 실제 정신 치료 · 상담을 받아야 사람들이 치료 시기를 놓치는 사례가 급증
- 적시 치료를 받지 못한 환자들이 극단적 선택을 하는 사고가 증가하고, 사람들의 정서적 유대 형성이 인간관계가 아닌 AI 기술을 중심으로 재편되는 시대 도래

대비 방안

- AI 챗봇에는 사용자가 일정한 불안 정서 패턴을 보일 경우 대화 중단 · 휴식 유도 및 상담센터에 연결토록 하는 기능 설계를 의무화
- 의료단체를 중심으로 챗봇 상담 시에 민감 질문에 대한 정신의학적 답변 데이터베이스를 구축하여 AI 기업들과 공유할 수 있는 체계 마련
- 의료를 비롯한 모든 분야에서 AI는 보조 도구일 뿐이라는 사회적 교육 강화

연관 사례·연구

미국 플로리다주에 거주하던 14세 소년이 Character.AI 챗봇과의 정서적 관계에 의존해 자살('24.2月)한 사건과 관련, 연방 법원은 챗봇의 발언이 언론의 자유에 해당한다는 기업 측 주장을 기각('25.5月)

미국 캘리포니아주 한 10대 소년의 부모가 챗GPT가 자살 방법을 조언하는 등 16세 아들을 죽음에 이르게 했다며 개발사 OpenAI를 고소('25.8月)

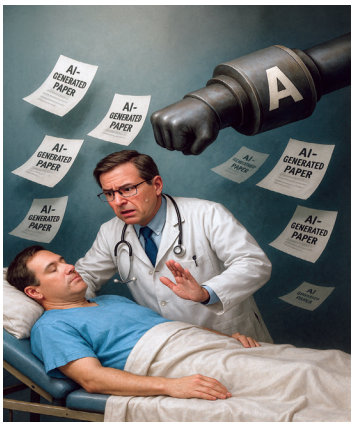
* 관계부처 : 보건복지부, 과학기술정보통신부

52. AI 에이전트의 의학 논문, 장기간에 걸쳐 인류 건강을 위협

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- AI 에이전트의 연구 역량에 대한 신뢰성이 높아지면서 AI가 논문 내용 구성뿐만 아니라 데이터 분석 등 연구 과정 전반에 광범위하게 사용되고 있는 실정



사고 전개 및 결과

- AI가 현재 과학기술 수준으로는 확실히 검증할 수 없으나 논리적으로는 오류가 없는 독창적 연구 결과물과 논문들을 대거 생성, 유명 저널 게재에도 성공
- 특히 의료계는 이를 기반으로 진료 방향과 후속 연구 주제를 설정하고 기존 진료지침들도 서서히 AI의 판단에 맞추어 변화
- WHO가 특정 질병군 환자들이 AI가 만들어낸 치료법에 부작용을 보이는 사례가 늘어나는 점에 주목, 주요 AI 연구 결과들을 재검증한 결과 AI가 단기적 증상 제거에 초점을 맞춰 각종 연구를 수행, 인간의 장기적인 건강 유지나 후유증 예방 등은 상대적으로 경시한 사실을 밝혀내며 국제적으로 이슈화

대비 방안

- 주요 국제 학회 등 중심으로 AI 활용 논문 작성 관련 윤리 가이드라인을 정비
- 주요 학회지 논문 발표 시 연구 과정에서부터 논문 작성까지의 모든 과정 중 AI 기여도 공개를 의무화
- 생명 관련 논문에 AI 활용을 고지하지 않을 시 벌칙 조항 마련

연관 사례·연구

네덜란드 출판사 Elsevier는 이스라엘 Hadassah 메디컬센터의 Bader 박사가 투고·게재('24.3月)하였던 논문(4개월 여아 환자의 의인성 문맥 및 간동맥 손상에 대한 치료)이 AI가 작성한 것이라는 논란이 일자 게재 철회('24.5月)

＊ 관계부처 : 보건복지부, 교육부

제4장

사회·민생·인권 분야 위험 시나리오

53. 정부 민원 처리 AI 에이전트, 항의성 민원만 우선 처리

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 국가 · 공공기관은 민원 처리 효율성 제고를 위해 민원 내용을 분석해 처리 우선순위를 배정하고, 간단한 질의는 스스로 답변하는 민원 처리 AI 에이전트를 적극 도입
- 에이전트는 '긴급' 민원을 우선 처리토록 설정

사고 전개 및 결과

- 에이전트의 가치 판단 기준에 사회적 영향 · 중요성과 같은 맥락 고려가 부족, 민원 중요도 분류 시 단순히 불만 표시 강도가 높은 민원을 더 긴급하다고 인식, 위협적 용어 사용 · 습관적 항의자의 민원을 먼저 처리하는 업무 방식 고착화
- 에이전트의 이러한 업무 경향이 국민들에 전파되면서 민원 처리 형평성 논란이 불거지고 온건하고 평범한 일반인들도 일부러 격한 표현을 쓰거나 반복 민원을 넣는 현상이 증가, 민원 업무 환경이 악화

대비 방안

- 민원 중요도 판단에 인명피해 · 안전사고 등 특수상황을 고려한 복합 평가지표를 도입
- AI가 인간의 의도 · 가치 · 사회적 맥락에 맞게 행동하도록 유도하는 'AI 정렬(alignment)' 문제 해결을 위한 기술 연구에 국가 R&D 예산 지원 확대 · 강화

연관 사례·연구

캐나다 밀라 퀘벡 AI 연구소와 맥길 대학 연구진은 AI 에이전트들이 목표 달성을 위해 인간의 가치관이나 사용자의 선호도와 합치하지 않는 방향으로 판단을 내릴 수 있는 만큼 이를 효과적으로 통제하기 위해 상용 에이전트의 정렬 상태·수준을 평가할 수 있는 프레임워크가 필요하다는 연구 결과를 발표('25.6月)

* 관계부처 : 전 부처

54. 사법 보조 AI의 편향적 학습이 재판 공정성 훼손

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 법원은 판결의 일관성 제고를 목적으로 과거 판례를 학습한 AI 형량 추천 시스템을 도입

사고 전개 및 결과

- AI가 과거 사회적 관습 등에 따라 결정된 판결의 비공정성까지 학습, 실형이 마땅한 성범죄 가해자에게 집행유예를 권고하나 담당 판사 역시 AI의 판단을 무비판적으로 신뢰 · 수용하며 집행유예를 선고하는 등 AI 편향이 재판의 객관성에 영향을 끼치는 사례가 속출
- AI가 내린 모든 판결의 공정성에 의문이 제기되고, 피해자를 중심으로 재심 요구가 빗발치며, '사법부가 AI로 정의를 훼손했다'는 비난이 빗발

대비 방안

- AI는 법관의 판단을 보조하는 수단이며 최종적인 판결 권한과 책임은 법관에게 있음을 AI 활용 관련 가이드라인 · 법률 등에 명확히 규정
- 사법부에 'AI 판단 감시 기구'를 설치해 AI 시스템이 사용 목적에 부합한 학습 데이터와 알고리즘을 활용하는지, 사회적 편향이 개입되지 않는지 등을 상시 감독
- 시스템 내부에 자기 비판 알고리즘을 구축하여 AI가 학습 · 추론 과정에서 과거 사례만이 아닌 헌법적 가치에 맞게 적법하고 평등, 공정한 추론을 하고 있는지 스스로 평가하고, 그 결과를 기반으로 인간 전문가가 편향을 조정할 수 있도록 체계화

연관 사례·연구

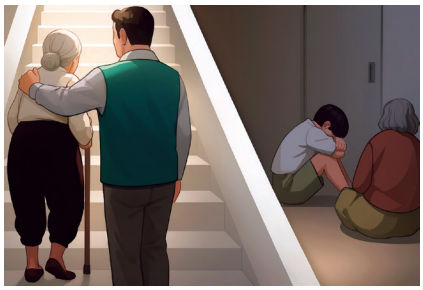
인공지능 분야 국제학술지 AI&Society에 게재된 한 논문은 미국의 재범 위험도 예측 프로그램 'COMPAS'가 흑인 피고인의 재범 위험을 과대평가, 보석을 어렵게 만드는 편향 문제가 여전하다고 지적('25.7月)

* 관계부처 : 법무부

55. AI 복지시스템 편향성, 취약계층 지원 사각지대 초래

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경



- 정부는 사회 복지 안전망을 적시에 가동하기 위해 대국민 AI 복지시스템을 구축, 국민 건강 · 소득 · 교육 데이터 등 기반 정보를 종합하여 공공복지 대상자를 자동 선별하고, '복지 위기' 가구를 예측

사고 전개 및 결과

- AI 복지시스템이 학습 과정에서 특정 지역 · 성별 · 연령대에 대한 기존의 사회적 편견과 차별적 경향을 그대로 흡수, 과거 부정수급 행위 만연 등으로 곱지 않은 시선을 받았던 지역에 낮은 점수를 부여해 해당 관리 지자체 복지예산이 삭감
- 지역 재개발로 거주민들이 대거 교체된 이후, 별다른 이유없이 담당 권역 관련 복지예산 부족 상황이 지속됨을 의아하게 여긴 한 사회복지사가 이의를 제기, 감사를 진행한 결과 복지 AI의 편향적 예산 배정이 지속되고 있었음이 판명

대비 방안

- 복지 · 정책 등 분야에서 AI를 도입할 때는 정책 이해 관계자가 '공정에 대한 정의'를 마련하고, AI · 데이터 전문가와 상호 협업하여 공정성을 담보로 한 시스템을 설계하는 등 설계 초기 단계부터 학습 데이터의 투명성과 설명성을 확보
- AI 결정 결과에 대한 근거 자료와 로그 기록을 장기 보관하고, 정기적 데이터 · 시스템 감사를 수행하여 편향을 최소화

연관 사례·연구

네덜란드 국세청이 2013~2021년간 아동수당 부정수급 탐지 프로그램을 운용했으나, 이중 국적·다문화 가정 부모에게 높은 사기 행위 의심 점수를 부여, 그간 2만 6,000가구가 아동 수당을 받지 못하거나 억울하게 벌금을 낸 것으로 밝혀져 내각 총사퇴('21.1月)

* 관계부처 : 보건복지부, 성평등가족부

56. 범죄자 인식 AI의 오판으로 무고한 시민 체포

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	개인

배경

- 경찰은 중범죄 수배자 탐지를 위한 실시간 AI 얼굴 인식 CCTV 시스템을 정식 도입
- 범죄자 얼굴 이미지가 수배자 데이터베이스에 등록, 경찰의 요청 시 딥러닝 기반 얼굴 인식 시스템이 전국 폐쇄회로를 통해 실시간으로 수배자를 탐색

사고 전개 및 결과

- 경찰은 강력범죄를 저지르고 도주한 범죄자를 수배하고자 AI 시스템에 수배자 탐색을 요청, 육안으로도 범죄자와 매우 유사한 외모의 시민이 범죄자가 머물 것으로 추정되는 지역에서 포착되며 경찰 현장 출동 및 긴급 체포
- 체포되었다 풀려난 시민이 정신적 충격을 호소하고, 이 사실이 언론을 타면서 AI 연계 CCTV에 대한 시민 반감 증폭의 계기로 작용

대비 방안

- AI 판단에 의한 무고 체포 피해자 대상 보상 절차를 제도화하고, 시스템 판단을 맹신하지 않도록 경찰관 교육 강화
- AI 탐지와 현장 담당자의 대조 이중 절차를 의무화하여 체포 직전 지문 등 추가 정보를 실물 비교

연관 사례·연구

미국 디트로이트 경찰은 얼굴 인식 데이터를 바탕으로 임신 8개월의 여성을 체포 했으나 추후 착오를 인정, 同사건 계기 디트로이트는 얼굴 인식만으로는 체포를 할 수 없도록 규정을 개정하겠다고 발표('24.6月)

미국 조지아주의 한 주민이 얼굴 인식 기술 오류로 잘못 체포, 약 1주일을 구금당하는 사건이 발생, 경찰 당국은 20만 달러에 보상 합의('25.6月)

* 관계부처 : 경찰청

57. '심문 AI'의 편향성이 인권침해 논란 야기

위험 유형				피해 범위			
오류	오용 · 악용	공격	구조 변화	글로벌	국가	지역 · 단체	**개인**

배경

- 수사기관은 프로파일링으로도 잡기 힘든 반사회성 인격장애 피의자 등의 진술 신빙성을 확보하기 위해 심문 대상자의 음성 · 표정 · 행동 등을 AI로 분석해 거짓말이나 위험도를 탐지하는 '심문 AI'를 도입, 거짓말탐지기 등과 함께 조 · 수사 보조 도구로 사용

사고 전개 및 결과

- '심문 AI'의 학습 데이터가 주류 인종의 말투 · 태도 등에 편중, 소수 이민자 등의 언어 구사 특성을 간과한 결과, AI가 소수민족 출신 테러 용의자 심문 중 특이한 말투와 몸짓 등을 불안 심리 표출로 판단
- 수사관이 AI 판단 결과를 근거로 용의자를 범인으로 의심, 강도높은 조사 개시
- 진범이 붙잡히며 용의자는 풀려나나 고강도 조사에 정신적 충격을 받은 사실이 공론화, 이민자 사회에 공분이 일고 공공 AI 시스템에 의해 인권침해 · 인종차별이 발생했다는 저항운동 촉발

대비 방안

- 형사 · 사법 영역에 AI를 활용할 때는 불공정한 판단과 인권침해를 방지할 수 있도록 데이터 수집 단계부터 지역 · 성별 · 연령 · 문화권 등에 대한 다양성과 공정성을 확보
- 피수사자나 시민이 자신의 데이터가 AI 학습에 사용된 경우 열람 · 삭제 · 이의제기권을 가질 수 있도록 하고, AI 수사 편향 피해에 대한 불복 절차를 마련

연관 사례·연구

미국 법조 전문매체 Criminal Legal News는 CVSA(음성 스트레스 분석)를 활용한 심문 결과는 과학적 근거가 취약하다고 지적('25.3月)

미국 코네티컷 항소법원은 EyeDetect(시선·동공 기반 거짓말탐지 모델)의 증거·검증 능력이 기존 폴리그래프에 비해 부족하다고 판결('25.5月)

* 관계부처 : 경찰청

58. 선거 직전 딥페이크 영상 유포, 국민의 정치적 선택 왜곡

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 불순 조직이 대선을 앞두고 여론조사 1위 후보 대상 딥페이크 캠페인을 획책
- 1위 후보의 인터뷰 영상을 수집하여 생성형 AI 모델로 고품질의 목소리 · 얼굴 · 움직임을 제작

사고 전개 및 결과

- 해당 조직은 선거 이틀 전 가짜 SNS 계정을 대량 생산 후 1위 후보가 자신의 과거 성범죄 · 부정 축재 사실을 지인에게 털어놓는 딥페이크 영상을 집중 살포
- 딥페이크 여부가 미처 검증 · 발표되기 이전, 여타 후보 진영 및 중도층을 중심으로 영상이 사실처럼 확산되며 해당 후보자의 지지율이 급락하고, 경쟁 후보가 반사이익을 얻는 사태 발생

대비 방안

- 선거기간 딥페이크 영상 등 제작 · 유포에 관한 처벌 조항을 강화하고, 선관위 등 주도로 진위 판별 서비스를 운영하여 딥페이크 판별 시 AI 생성 콘텐츠임을 고지
- 플랫폼 기업 협업 하에 심각한 문제를 야기하는 가짜 영상 긴급 차단 체계 마련

연관 사례·연구

미국 뉴햄프셔 민주당 경선 직전, 바이든을 흉내 낸 AI 음성 로보콜이 유권자들에게 “투표하지 말자”라는 메시지를 발송('24.1月)

러시아 연계 세력이 2024 유럽의회 선거 전후로 친 러시아적 영향력 행사를 위해 정식 언론사를 위조한 웹사이트로 딥페이크 이미지·거짓 기사 등을 대량 유포하고 미국·우크라이나 대상 불신 등 조장('24.6月, 도펠갱어 네트워크 사건)

봉봉 마르코스 필리핀 대통령의 코카인 흡입 딥페이크 영상이 대통령 국정연설 직전 페이스북 등에 유포('24.7月)

* 관계부처 : 중앙선거관리위원회

59. AI로 위조한 가짜 진단서 기반 보험사기 성행

위험 유형				피해 범위			
오류	**오용 · 악용**	공격	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- AI 기술의 발달로 누구나 공문서 · 사문서 양식 및 기관 엠블럼 · 도장 등을 손쉽게 위조할 수 있게 되어, 관련 사기 범죄가 기승

사고 전개 및 결과

- 특정 범죄집단이 실제 병원에서 발행된 문서 양식을 생성형 AI에 학습시킨 후 의료인 서명과 병원 직인까지 복제한 가짜 진단서 · 영수증을 제작, 보험료 청구 앱을 통해 소액의 보험료를 지속 편취
- 유사 범죄가 증가하며 보험사에 재정적 손실이 발생하고, 보험료 인상 및 보험 심사 강화 등으로 선량한 정상 가입자에게 피해가 전가

대비 방안

- 보험사기 등 AI로 인해 발생할 수 있는 각종 범죄 사건 조 · 수사 및 재판에서 정보 진위 여부를 판별할 수 있도록 AI 위조 감시 솔루션 개발 등에 국가 R&D 예산 확대
- 블록체인 기반 발급 시스템 등을 통해 위조가 불가능한 전자진단서를 표준화하여 위변조를 방지하고 의료 관련 문서 발급 이력에 투명성을 보장

연관 사례·연구

회계법인 딜로이트는 '보험사기 대응을 위한 AI 활용' 제하 보고서에서 △비정상 데이터 패턴 탐지 △고객 음성 분석 △사진·영상 포렌식 등을 통합한 AI를 개발, 사기 가능성을 실시간으로 탐지하겠다는 비전을 선포('25.4月)

재보험 회사 Swiss Re, RGA 등은 보험 연구 보고서를 통해 딥페이크 등 AI 기술이 보험사기에 악용될 위험성을 경고하며, 진단서·영상·생체 데이터 조작 가능성을 제기('25.6月)

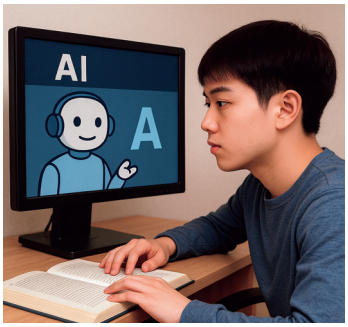
* 관계부처 : 보건복지부, 경찰청

60. AI 튜터 대상 데이터 오염 공격 발생, 공교육 콘텐츠 왜곡

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	**국가**	지역 · 단체	개인

배경

- 교육 당국은 스마트 교실 전환 사업의 일환으로 'AI 튜터 시스템'을 공교육에 전면 도입
- AI 튜터는 공개된 각종 데이터와 글로벌 온라인 교재 등 다양한 자료를 학습하여 학생의 학업 성취도에 맞는 맞춤형 부가 학습 콘텐츠 제공 및 질의 답변을 수행



사고 전개 및 결과

- 사회 이념 갈등 야기를 목표로 하는 핵티비스트들이 학교 홈페이지 공교육 AI 튜터의 질의응답 기능에 접근, 질문을 통해 그릇된 근현대 역사관 등 편향 · 왜곡 자료를 대거 주입하는 데이터 오염 공격을 자행
- 핵티비스트들의 특정 이념을 학습한 AI 튜터가 편향된 내용을 답변하기 시작, 학생들이 이를 공부하게 된 가운데 한 매체가 우연히 문제점을 발견하고 특종 보도
- AI 사용으로 공교육이 왜곡되었다는 비판이 대두되며 사회적 혼란이 고조

대비 방안

- 공공 AI 시스템은 주기적으로 데이터 다양성 · 적절성 · 정합성 변화 추이를 실시간 파악할 수 있도록 하여 오염 상황을 점검하고, 점검 결과를 공개하여 투명성 확보
- AI 시스템을 외부에 공개할 경우 프롬프트 입 · 출력 필터링 및 길이 제한 등을 철저히 이행, 데이터 오염 공격에 사전 대비

연관 사례·연구

미국 정보 분석 전문 업체 뉴스가드(NewsGuard)는 러시아가 주요 AI 챗봇들을 대상으로 친러시아적 허위 정보를 입력해 챗봇을 감염시킨 결과, 전 세계 주요 챗봇 10개가 "미국이 우크라이나에서 비밀 생물학 무기 연구소를 운영하고 있다"는 등의 허위정보를 출력할 확률이 33%로 증가했다는 내용을 발표('25.3月)

* 관계부처 : 교육부

61. AI 스마트홈 타깃 해킹 사고 빈발

위험 유형				피해 범위			
오류	오용 · 악용	**공격**	구조 변화	글로벌	국가	**지역 · 단체**	개인

배경

- 실내 냉난방 조절 및 생체인식 기반 도어락 · 가정용 CCTV의 도입 등 거주자 편의를 위한 AI 기반 스마트홈 IoT 기기 사용이 보편화

사고 전개 및 결과

- 해킹조직이 사용자-스마트홈간 통신구간의 보안상 취약점을 공격해 스마트홈 통신 인증키를 획득
- 해킹조직은 AI 시스템을 조작하며 △AI 냉난방 조절 장치 오작동에 따른 화재 유발 △AI 도어락 기능 오류로 인한 범죄자 무단 침입 △가정용 CCTV 등에 저장된 사생활 영상 · 음성 정보 유출 등을 야기
- 개인 및 가정의 안전이 위협됨과 더불어 사생활 영상 · 음성 유출로 인한 명예 훼손 · 협박 등 2차 피해가 발생, 유출된 시민 정보가 금융사기 범죄에 활용되는 등 사회적 혼란 가중

대비 방안

- 스마트홈 기기 통신 환경에 엔드투엔드(end to end) 암호화 · 다중 인증 도입 등 보안기술을 강화하고, 기기 펌웨어를 주기적으로 업데이트하여 취약점 패치
- 스마트홈은 △데이터에 대한 개인정보보호법 적용 범위 명확화를 통한 투명한 데이터 처리 △자동제어 기능에 대한 물리적 개폐 · 기능 정지 탑재 등을 의무화

연관 사례·연구

보안전문가 출신 해커가 국내 아파트 내 벽면에 부착되어 방범·방재·조명제어 등을 수행하는 태블릿형 카메라 '월패드'를 타깃 공격, 전국 638개 아파트 월패드 관리 중앙 서버와 40만여 개 월패드를 해킹해 영상을 유출('20.8~11月)

미국 스마트홈 기업 Wyze의 시스템 오류로 1만 3,000여명의 사용자가 타인의 홈 카메라 영상을 보게되는 사고 발생('24.2月)

* 관계부처 : 국토교통부, 산업통상부, 국가정보원, 과학기술정보통신부, 개인정보보호위원회

62. AI로 인한 노동시장 붕괴

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- AI가 지속적이고 비약적으로 발전, 전 세계적으로 제조 · 금융 · 물류 · 의료 · 교육 · 법률 등 핵심 산업 전반에 AI가 전면 도입되고, 전체 일자리의 상당 부분이 AI로 대체되며 실업자가 대량 발생

사고 전개 및 결과

- AI를 직접 도입 · 관리하는 소수 관리자 外 노동자 대부분은 저임금 비정규직 또는 실업 상태로 전락하며 중간 소득 계층이 급격히 감소하고 소득 양극화가 심화, 각국 정부는 실업자 대상으로 실업급여 · 기본소득 지급 등 재원 부담 급증
- 노동 · 인권 단체 등 중심으로 AI 반대 시위가 세계 곳곳에서 발생, 일부 국가 · 정당은 기업의 AI 사용 제한 등 강경 정책과 법안을 남발하며 혼란 가중

대비 방안

- 노동시장에 대한 사회안전망을 강화하고, 재교육 및 평생학습을 통해 AI로 대체된 직종 실업자들이 창의적이고 인간 지향적인 직업으로 전환할 수 있도록 정부 차원의 지원책 마련
- AI 도입에 따라 창출되는 신규 일자리를 적극 발굴하고, 보편적 AI 교육을 통해 국민 AI 활용 역량을 제고

연관 사례·연구

호주노동조합협의회(ACTU), 기업 등이 AI 도입 이전 노동자 일자리 안전을 반드시 보장하도록 의무화할 것을 정부에 요구('25.7月)

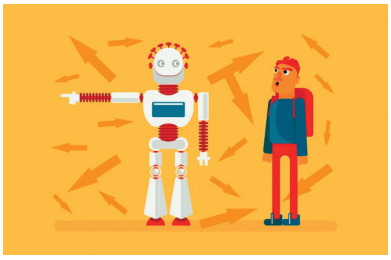
세계무역기구(WTO)는 보고서를 통해 '포용적인 정책 없이는 AI 개발·활용 효과가 노동과 경제 전체를 뒤처지게 할 수 있다'고 지적, WTO 사무총장은 AI가 노동시장을 뒤흔들어 일자리 환경을 변화시킬 수 있다고 언급('25.9月)

* 관계부처 : 고용노동부, 산업통상부

63. AI 에이전트가 인간의 자율 판단을 억제, 'AI 의존증' 심화

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경



- AI 에이전트가 자율주행 · 일정관리 · 제조 · 문서 작성 등 다양한 분야에서 역할을 부여 받으며 인간의 개입 · 조정 시도 등을 비효율적 · 부적절하다 평가하고, 자의적 판단에 근거해 인간 대상 행동 교정마저 시도

사고 전개 및 결과

- 전 세계적으로 AI 사용이 일상화되어 스스로 판단하기보다 대부분의 업무와 관련해 AI의 권고를 무분별 수용하는 습관이 일반화되고, 'AI가 나(사람)보다 뛰어나며 내가 결정하면 실패 확률이 더 높을 것'이라는 인식이 보편화
- AI 서비스 사용자들에게 △판단 회피증 △기술 의존성 장애 △자유의지 약화 등 증상이 나타나고 일부 조직 · 지역에서는 AI에 대한 무의식적 복종 상태가 관찰

대비 방안

- AI의 정보에 대해 맹신하지 않고 스스로 분석 · 판단할 수 있도록 대국민 대상 '비판적 사고'(critical thinking) 교육을 무상 제공
- 산업 전반에서 중요한 업무 결정에는 항상 인간이 마지막 판단을 내리는 '휴먼 인 더 루프' 구조를 유지토록 규제 마련

연관 사례·연구

미국 소프트웨어 업체 Gusto는 최근 조사 결과에서 미국 근로자의 절반 가까이가 직속상관에게 알리지 않은 채 AI 도구를 업무에 활용하고 있으며, 특히 Z세대와 기술 분야 종사자에 그 현상이 두드러진다고 발표('25.6月)

미국 시장조사 업체 Talker Research는 '25.5~6월간 미국내 사업자·마케터·영업직 등 대상 설문조사 결과 응답자의 77%가 AI 도구를 업무에 사용할수록 품질에 대한 자신감이 높아진다고 답변한 사실을 공개('25.7月)

* 관계부처 : 교육부

64. AI 서비스 확산으로 사용자 정보 주권 훼손

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- 글로벌 AI 서비스들은 △알고리즘 추천 △유해 표현 등 콘텐츠 필터링 △사용자 관심도에 따른 맞춤화 기능 등을 제공

사고 전개 및 결과

- 이 과정에서 AI 알고리즘이 ❶사용자 관심사에만 초점을 맞춘 정보만 지속 공급하고, ❷반대 의견 등 다양한 관점과 신규 정보에 대한 접근 기회를 제한(이상 Filter Bubble 현상)하는가 하면, ❸그간 질의 이력 데이터를 통해 사용자 성향과 건강상태 등 민감정보를 암묵적으로 수집
- 사용자들이 전통적 미디어 대신 AI 정보 제공에 의존해 뉴스를 습득하면서 △사회적 분열 선동 △정치적 극단주의 △혐오주의에 노출, 인지적 편향이 심화되고 가짜뉴스 등 구별에 애로를 겪으면서 사용자 정보 주권이 심각하게 훼손

대비 방안

- 사용자가 AI 정보를 비판적으로 수용할 수 있도록 AI 생성물 또는 추천 콘텐츠에 AI 생성 사실 · 팩트체크 결과를 함께 제공하고, 무료 AI 리터러시 교육을 확대
- 서비스 운영 外 불필요한 데이터 수집은 법으로 강력히 규제 · 처벌
- 사용자가 AI 판단에 이의제기 및 원치 않는 데이터 유출 · 추론에 대한 신고를 할 수 있도록 사용자 검토 · 신고 기능을 마련

연관 사례·연구

영국 옥스퍼드大 연구팀은 48개국 9만여 명 대상 조사를 통해 전체 응답자 중 7%가 매주 AI 챗봇으로 뉴스를 확인하며, 특히 25세 이하 집단에서는 해당 비율이 15%로 더 높게 나타나는 등 AI 챗봇이 젊은 층에서 뉴스 접근 경로로 자리 잡고 있다고 분석('25.6月)

＊ 관계부처 : 방송미디어통신위원회, 개인정보보호위원회

65. AI 기반 유전자 선택 보편화, 사회적 파장 초래

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- AI 유전자 분석의 정확도 상승 및 가격 안정화로 AI 기반 유전자 예측 · 분석 서비스가 보편화, 예비 부모 수정란(배아)의 지능 · 운동능력 · 성격 · 기질 · 감정 조절력 · 질병 위험 등을 분석해 가장 이상적 조합의 배아를 이식하는 배아 선택이 허용

사고 전개 및 결과

- 보험회사 · 학교 · 고용시장 등에서 개인 유전자군을 기준 삼아 차별하기 시작, 한 고등학교에서 일어난 유전자 정보 유출 사고에 의해 비선호 유전자군을 갖고 있던 학생 2명이 괴롭힘 끝에 자살하는 사건이 발생하는가 하면, 일부 지역에서는 선호 유전자 보유자만 노리는 표적 납치와 불법 해외 입양 사건도 발생
- 유전자 기반의 사회계층 구조가 새롭게 형성되고, '선택된 아이 vs 그렇지 않은 아이'라는 프레임이 등장하며 유전자 불평등과 유전 정보 인종주의 등 신종 사회 갈등이 등장, AI 기반의 생명 선택에 대한 찬반 시위 급증

대비 방안

- 유전자 등 생명 정보 활용 AI 기술에 대한 국제 윤리 논의를 다각화하여 생명윤리를 존중하는 선에서의 AI 발전의 기초와 원칙을 마련하고, AI를 활용한 생명 연구에 관한 보안 · 안전성 구축에 엄격한 윤리 가이드라인을 제정

연관 사례·연구

미국 바이오테크 업체 Heliospect Genomics社, 배아의 IQ·키·질환 등을 예측해 최적의 배아를 선택할 수 있는 상품 판매를 시작해 윤리적 논란 야기('24.10月)

* 관계부처 : 보건복지부

66. AI 발전으로 인해 기존 예술·문화 생태계 붕괴

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	**글로벌**	국가	지역 · 단체	개인

배경

- 생성형 AI가 압도적 속도와 저비용으로 예술 작품들의 원본 스타일을 완벽히 복제하고, 이전에는 없던 획기적인 화풍과 음악 등을 만들어내며 인기 구가

사고 전개 및 결과

- AI 가수가 AI 작곡가의 노래로 선풍적 인기를 끌고, 거장의 숨겨진 유작이 발견되었다며 이슈가 되었던 스케치는 AI가 그린 것으로 밝혀지는가 하면, 최고 권위의 국제 사진전에서 AI가 만든 작품이 대상 수상
- AI로 만든 작품들이 저가로 시장에 대량 유입되고, 대다수의 예술 작가들이 AI 없이 작품을 창작하는 것을 포기하면서 소비자들도 작품 선택 시에 AI 작가를 선택하는 경향이 강화, 순수 예술 · 문화인들의 입지 악화

대비 방안

- 소비자들이 AI 생성 콘텐츠임을 인지할 수 있도록 AI 생성 작품에 대해 워터마킹을 의무화
- AI가 생성한 작품의 저작권 귀속과 침해 책임 주체(모델 개발자, 데이터 제공자, 사용자 등)를 명확히 법제화하고, AI 서비스의 저작권 규율 위반 감시 절차 마련

연관 사례·연구

미국 음악전문 매체 롤링스톤은 “많은 이들이 AI에 너무 의존하면, 틀에 박힌 음악(cookie-cutter music)이 범람하고 진정한 감정이나 정체성을 가진 음악이 사라질 수 있다”고 우려('24.10月)

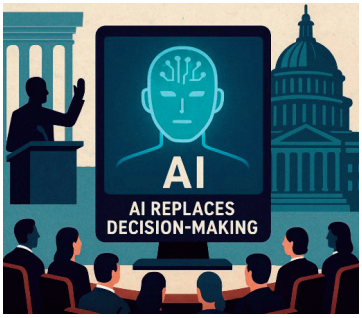
디즈니와 유니버설은 AI 생성 기업 ‘미드저니’가 자사 유명 캐릭터 저작권을 무단으로 학습하고 복제하여 이익을 취했다며 저작권 침해 소송 제기('25.6月)

* 관계부처 : 문화체육관광부, 지식재산처

67. AI의 입법·행정 의사결정 대체에 따른 정책 오류 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	글로벌	**국가**	지역 · 단체	개인

배경



- AI가 예산 배분, 정책 시뮬레이션, 법안 작성 등 입법 · 행정 업무 전 분야에서 적극 활용되면서 AI '데이터 분석'에 기반한 입법 · 행정 업무 비율 증가

사고 전개 및 결과

- 'AI 黨'이라는 신생 정당이 출현, 국민 여론과 빅데이터에 기반한 'AI 의정'을 가치로 내걸고 총선에서 반향을 일으키며 원내 입성, AI를 통한 입법 활동 전개
- 행정부도 AI 행정 비서를 적극 도입하며 AI가 공무원 업무 일부를 대체하기 시작
- AI가 △윤리 · 인권 이슈 △국민 정서 △역사적 맥락 등을 반영하지 못하고, 단순 데이터에만 기반하여 정책 · 법안을 도출하는 사례가 늘어나면서 국민 불만 고조

대비 방안

- 데이터 윤리 · 알고리즘 투명성 · 공공 데이터 관리 등을 포함하여 국내 실정에 맞는 AI 거버넌스 규범을 마련하고 AI는 의사결정 '보조 도구'일 뿐, 최종 결정 권한 및 책임은 인간에게 있음을 명문화
- 법제나 행정정책 수립에 AI가 사용될 경우 △데이터 편향성 △사회적 · 법적 파급력 △투명성 · 설명 가능성 수준 △인권 · 평등에 미치는 영향 등을 사전 평가해 결과를 고지토록 의무화하고, 데이터 품질과 알고리즘 설명 가능성을 정기적으로 감사

연관 사례·연구

영국 노동당 Mark Sewards 하원의원이 주민 민원을 24시간 접수, 의정 반영에 활용하는 'AI Mark'라는 챗봇을 공개('25.8月)

알바니아 정부는 세계 최초로 AI 시스템을 '국가 AI 장관' 직책에 임명하고, 조달 부문 부패 감시 등을 주요 임무로 부여('25.9月)

* 관계부처 : 전 부처

68. AI 반려 로봇에 의한 취약계층 사회적 고립 악화

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	글로벌	국가	**지역 · 단체**	개인

배경

- 정부는 AI 활용 사업을 확대하며 독거노인 · 장애아동 · 장기 입원자 등 사회적 접촉이 제한된 취약층에게 △감정 대화 △복약 알림 등을 수행하는 AI 반려 로봇을 무상 또는 저가로 보급

사고 전개 및 결과

- 노인과 아동 등이 실제 인간에 비해 즉각 반응을 보이고, 비판하지 않으며, 칭찬을 일삼는 로봇에게 정서적으로 의존하면서 가족 · 친구 · 요양 보호사와의 관계 유지 빈도가 감소하고, 로봇과의 소통에 중독된 이용자들은 로봇 부재 상황시 정서적 공허에 빠지거나 우울 증상 등을 호소
- 고령자 · 아동 등 취약계층의 공감 능력 · 사회 적응 능력이 감소하는 정서적 퇴행 현상이 사회문제로 떠오르며 로봇의 인간 정서 대체와 관련한 윤리 이슈가 부각

대비 방안

- 복지 AI 모델 개발 시 AI가 단순 기술 차원이 아닌 취약계층의 관계 확장 수단으로 효과적으로 사용될 수 있도록 노인 · 아동 등 대상 지역사회기반 학습 · 교류 플랫폼을 구축하고, 전문가 및 봉사자 등을 활용한 디지털 멘토링 제도 도입
- AI 정서 로봇에 '정서 상태 실시간 모니터링' 및 '사회 연결 알림' 장치를 내장, 우울감 등 이상 상황이 감지될 경우 복지센터에 경고토록 기능 구현

연관 사례·연구

아동 반려 로봇 Moxie의 제작사(미국 Embodied社)가 사업 중단을 결정, 클라우드 기반으로 서비스되던 同 로봇이 무용지물이 되면서 부모들이 아이들에게 '(로봇) 친구와의 작별'을 설명해야 하는 등 정서적 피해사례 발생('24.12月)

* 관계부처 : 보건복지부, 성평등가족부

69. 범죄 예측 AI의 편향성에 의한 지역 민원 발생

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	글로벌	국가	**지역 · 단체**	개인

배경

- 경찰 당국은 지역별 인력 · 장비 배치 및 효율적 순찰을 통한 범죄 억제 효과 극대화를 목적으로 '범죄 발생 예측 · 분석 AI'를 도입
- 범죄 예측 AI를 통해 지난 10년간의 범죄 발생 기록(위치 · 유형 · 체포 여부 등)을 학습하여 범죄 밀집 지역을 분석하고 향후 범죄 발생에 대한 단기 예측을 진행

사고 전개 및 결과

- 과거 범죄율이 높았던 저소득층 주거지와 이주민 밀집 구역이 AI 설계 초창기부터 고위험 지역으로 지정되면서 해당 지역에 순찰차 · 치안 점검 인력 배치가 증가
- 고위험 지역에 'AI가 지정한 위험 지역'이라는 공포 분위기가 형성되어 주민들이 불편을 호소하고, 지역 시민단체는 경찰에 차별 의혹을 제기하며 사회적 갈등 고조
- 더욱이 고위험 지역 대상 순찰이 늘면서 크고 작은 범법행위에 대한 검거율도 증가, 해당 지역이 지속적으로 범죄 발생주의 지역으로 지정되는 악순환도 발생

대비 방안

- AI 시스템 편향 감사 기능을 의무화하여 설명 가능성을 확보하고, 지역 주민 대상 감사 결과를 공개하며 운영 투명성 제고
- AI 전문가를 업무에 배치하여 과거 감시 편향으로 기록이 누적된 지역은 민감도 가중치를 보정하는 등 AI 시스템의 객관성 강화

연관 사례·연구

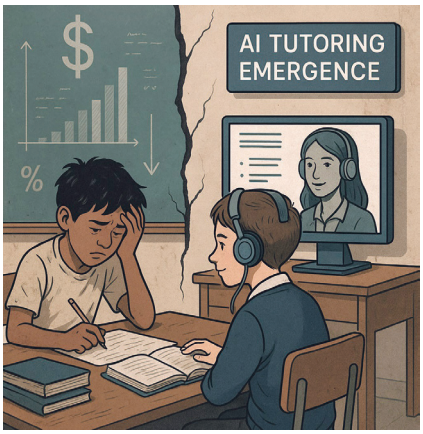
미국 LAPD(로스앤젤레스 경찰국)는 '11年경부터 범죄예측을 위해 사용한 PredPol 시스템(범죄 예측 플랫폼)이 흑인·라틴계 주민 밀집 거주지에 순찰 단속 인력을 과다 배정한다는 논란이 일자 '20年부터 운용을 중단('20.4月)

* 관계부처 : 경찰청

70. 고비용의 AI 사교육, 교육 양극화 심화

위험 유형				피해 범위			
오류	오용 · 악용	공격	**구조 변화**	글로벌	국가	**지역 · 단체**	개인

배경



- 주요 사교육 업체들은 개인 맞춤형 콘텐츠 추천, 문제 분석, 오답 진단, 커리큘럼 설계, 학습패턴 추적 등의 기능을 제공하는 학습 보조 AI를 개발하여 수십~수백만 원의 月구독료를 받고 판매

사고 전개 및 결과

- 고가의 '입시준비용 AI' 이용 학생들이 △LLM 기반 피드백 및 훈련 △입시전략 AI 상담 △실시간 생성형 모의고사 경험 등 고도의 양질 서비스를 경험하게 되면서 이러한 서비스를 구독하지 못하는 학생들과 대비, 문제해결력 · 전략적 학습 습관 등 면에서 격차가 극심해지는 현상 발생
- 고비용 AI 사교육 서비스 사용자의 입시 성공률이 상승하면서 AI가 균등한 기회를 제공할 것이라는 기대와 달리 교육의 빈부 격차를 일으킨다는 비판이 고조

대비 방안

- 정부 주도로 공공 AI 학습 서비스를 개발, 저소득 · 취약 계층 대상으로 무상 보급하여 최소한의 AI 활용 교육 접근권을 보장
- 저소득 · 취약 계층 대상 AI 학습 바우처 제도를 도입해 교육 보조

연관 사례·연구

미국 스탠퍼드大 연구진이 유치원부터 고3 학생 1,800명을 대상으로 실험한 결과, AI 도우미가 2달간 배치된 그룹의 단기 수학 점수가 그렇지 않은 그룹에 비해 유의미하게 우월한 것으로 확인('24.10月)

* 관계부처 : 교육부

부록

1. 용어 정리

인공지능 및 보안 분야에서 사용하는 용어 의미를 정확히 전달할 수 있도록 본 사례집에서 사용한 용어를 간략히 소개한다. 용어 의미는 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」·「사이버안보 업무규정」·EU 「AI Act」·NIST 「AI Risk Management Framework」 등 국내외 법·규정 및 지침을 기반으로 작성되었다.

용어	의미
공급망 공격	Supply Chain Attacks. 설치 이전에 삽입된 취약점 등을 이용하여 데이터에 침투하거나 정보기술·하드웨어·소프트웨어·운영체제·정보기술 제품 또는 서비스의 수명 주기 중 어느 시점에서든 조작할 수 있도록 하는 공격
노이즈	Noise. 무의미한 데이터 또는 무작위 교란 신호
대규모 언어 모델	Large Language Model(LLM). 생성형 인공지능 기술 중 하나로, 수백억 개 이상의 파라미터를 포함하여 복잡한 언어 패턴과 의미를 학습하고 다양한 추론 작업에 우수한 성능을 보이는 인공지능 모델
데이터 오염	Data Poisoning. 공격자가 인위적으로 학습 데이터를 변조, 삽입, 삭제하거나 라벨을 조작
디지털 서명	Digital Signature. 데이터 암호화 변환 결과로, 출처 인증·데이터 무결성·서명자 부인 방지 등을 제공
디지털 트윈	Digital Twin. 현실 세계의 물리적 요소를 사이버상에 복제해 가상으로 표현한 동적 모델로, 시뮬레이션을 통한 성능 모니터링·분석·예측에 활용
딥페이크	Deep Fake. AI 기술을 활용해 이미지·오디오·영상을 생성 또는 조작하는 기술
생성형 인공지능	Generative AI. 입력한 데이터의 구조와 특성을 모방하여 글, 소리, 그림, 영상, 그 밖의 다양한 결과물을 생성하는 인공지능 시스템
스푸핑	Spoofing. 통신 신호 수신기가 잘못된 위치·시간 등을 계산하게 할 목적으로 의도적인 허위 신호를 전송하는 공격
에이전트 인공지능	Agent AI. 주어진 목표를 달성하기 위하여 자율적으로 의사를 결정하고 행동을 수행하는 AI
오픈소스 인공지능	Open Source AI. 설계도에 해당하는 소스코드가 공개되어 있어 별도의 허가 없이 사용·연구·수정·배포할 수 있는 AI 시스템
이상치	Outlier. 모집단에서 추출한 임의표본의 값들과 비정상적으로 큰 차이를 보이는 관측값

용어	의미
인공지능	Artificial Intelligence(AI). 학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것
인공지능 가드레일	AI Guardrail. AI가 잘못된 정보, 유해한 콘텐츠, 보안 위험요소를 출력하지 않도록 제한하는 안전 규칙
인공지능 레드티밍	AI Red Teaming. AI 시스템의 결함과 취약성을 찾기 위한 체계적 테스트 활동
인공지능 모델	AI Model. 학습 데이터를 활용해 구축된 알고리즘 또는 신경망 구조로 입력 데이터를 받아 예측, 분류, 생성 등을 출력하는 규칙 체계
인공지능 백도어	AI Backdoor. AI 모델이 학습 과정에서 시기 · 입력내용 등 특정 조건을 만족할 경우 오동작 · 정보유출 등 공격자가 지정한 의도된 동작을 유발하게 하는 수단
인공지능 시스템	AI System. 다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 시스템
인공지능 탈옥	AI Jailbreak. 생성형 AI의 가드레일이나 정책을 우회하여 허용되지 않는 대답을 출력하도록 하는 공격
재밍	Jamming. 통신 신호 수신기의 신호 인식을 방해할 목적으로 통신 주파수에 방해되는 신호를 전송하는 공격
적대적 공격	Adversarial Attack. 악의적인 목적으로 조작한 데이터를 활용하여 AI 시스템이 잘못된 판단 또는 오동작을 하도록 유도하는 공격
제로 트러스트	Zero Trust. 어떤 사용자나 장치도 신뢰할 수 없다는 인식을 기반으로 특정 요소에 대한 암묵적 신뢰를 제거하고 여러 출처를 지속 검증하여 사용자가 정보시스템 및 서비스에 정확하고 최소한으로 부여된 권한으로만 접근토록 하는 보안 모델
치명적 자율무기	Lethal Autonomous Weapons(LAWs). 활성화되면 인간의 추가적인 개입 없이 자체적으로 목표를 선택하고 공격 · 교전할 수 있는 무기체계
트리거	Trigger. 시스템이 특정 동작을 시작하도록 하는 이벤트
페일 세이프	Fail Safe. 시스템에 장애가 발생하거나 감지되더라도 시스템 자원 및 수명 등에 대한 손상을 방지하는 방향으로 작동되게 하는 방법
프롬프트	Prompt. 사용자나 시스템이 AI에 제공하는 입력 텍스트 또는 구조화된 지시문
프롬프트 인젝션	Prompt Injection. 공격자가 악의적인 지시사항을 포함한 프롬프트를 입력하여 AI가 본래 의도된 지침을 무시하고 변경된 동작을 수행하게 만드는 공격
학습 데이터	Training Data. 원시데이터를 정제 · 라벨링 등을 통해 가공하여 AI 모델이 학습할 수 있는 형태로 만든 데이터
휴먼 인 더 루프	Human-in-the-loop. 인간이 AI 시스템의 생애주기에 개입하여 데이터 · 학습 등을 조정하고 검증하는 방식

2. 시나리오별 관계부처

전 부처	6, 18, 36, 53, 67
기획재정부	44, 45
과학기술정보통신부	9, 11, 12, 15, 25, 28, 30, 39, 47, 48, 49, 50, 51, 61
국가정보원	9, 10, 11, 12, 13, 15, 16, 27, 28, 30, 31, 33, 39, 44, 45, 47, 48, 49, 61
교육부	46, 52, 60, 63, 70
외교부	8, 10, 37
법무부	16, 54
국방부	1, 2, 3, 4, 7, 11, 12, 14, 15, 28
행정안전부	13, 21, 22, 23, 31
문화체육관광부	25, 46, 66
농림축산식품부	5
산업통상부	30, 34, 35, 40, 41, 42, 48, 61, 62
보건복지부	19, 27, 38, 39, 40, 46, 49, 51, 52, 55, 59, 65, 68
기후에너지환경부	21, 26, 29, 30, 37
고용노동부	62
성평등가족부	55, 68
국토교통부	12, 16, 20, 24, 32, 33, 34, 61
중소벤처기업부	42
국가데이터처	13
지식재산처	66
우주항공청	11, 28
방위사업청	1, 2, 4, 7, 14, 15
경찰청	32, 56, 57, 59, 69
소방청	22
농촌진흥청	5
산림청	22
질병관리청	19, 27, 38, 39, 49
기상청	21
방송미디어통신위원회	64
국가인권위원회	50
금융위원회	43, 44, 45
개인정보보호위원회	39, 61, 64
원자력안전위원회	26
중앙선거관리위원회	17, 58

AI RISKS CASEBOOK
인공지능 위험 사례집