AI와 사이버보안 : 창과 방패의 경쟁

07. 2025

eR((o))un&company

윤 두 식 [dsyoon@eroun.ai]

Table of contents

01 Introduction

02 AI시대 보안 이슈

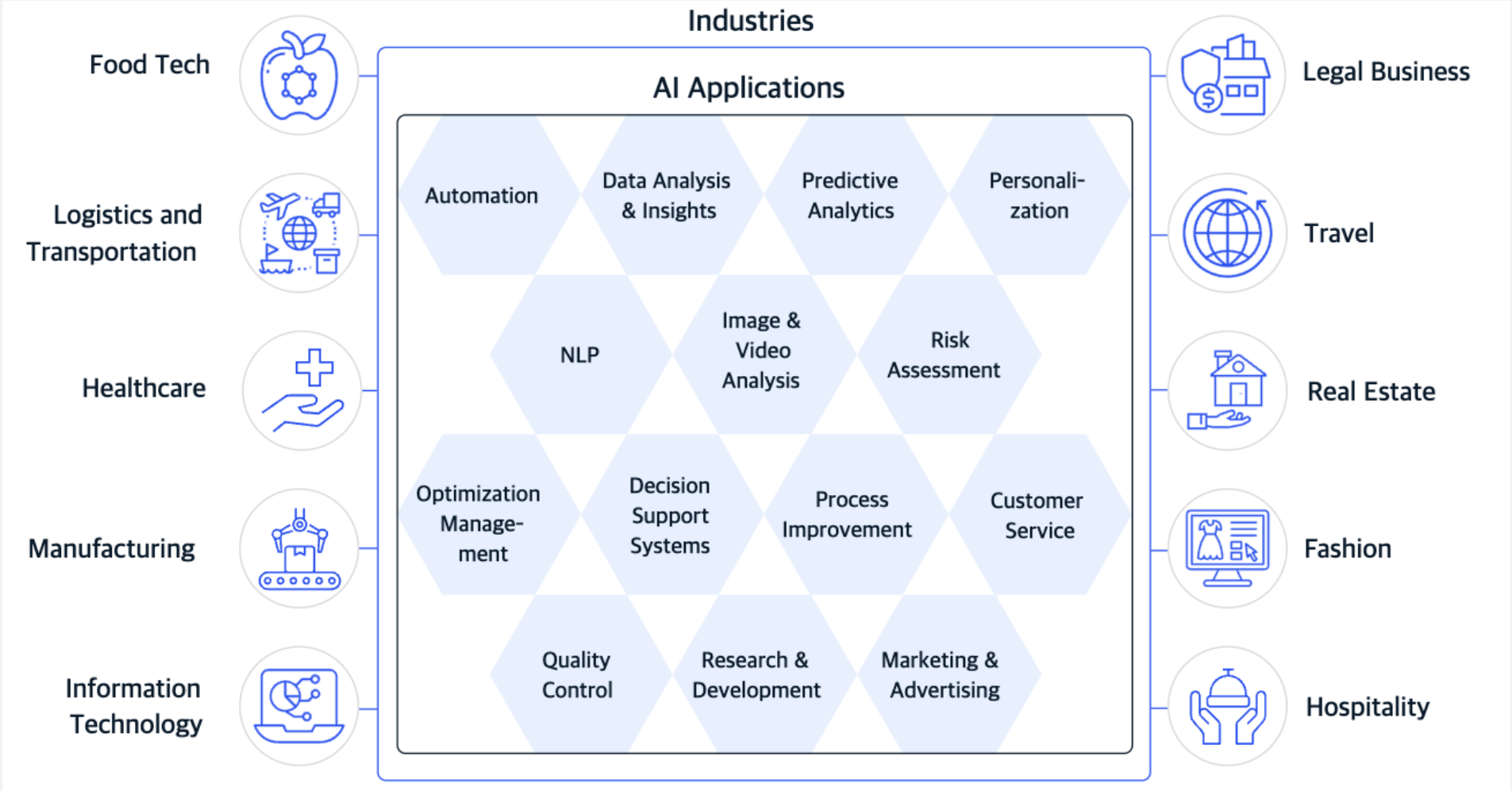
03 기업의 AI사용방식과 보안이슈 대응

04 AI시대 정책변화의 필요성

01

Introduction

AI is everywhere



참조 : https://www.leewayhertz.com/ai-use-cases-and-applications/#AI-use-cases-in-major-industries

생성형 AI, 어디에 사용되나?

생성 분야 확장 (Multi Modal)

- Text-to-Text,
- Text-to-Image,
- Text-To-Video,
- Text-To-Voice

코딩, 추론, 수학 등 복잡한 작업이 가능

The Top 50 Gen AI Web Products, by Unique Monthly Visits

1. ChatGPT	11. SpicyChat	21. VIGGLE	31. PIXAI	41. MaxAI.me
2. character.ai	12. IIElevenLabs	22. Photoroom	32. Clipchamp	42. BLACKBOX AI
3. perplexity	13. Hugging Face	23. Gamma	33. udio	43. CHATPDF
4. Claude	14. LUMA AI	24. VEED.IO	34. Chatbot App	44. Gauth
5. SUNO	15. candy.ai	25. PIXLR	35. VocalRemover	45. coze
6. JanitorAI	16. Crushon AI	26. ideogram	36. PicWish	46. Playground
7. QuillBot	17. Leonardo.Ai	27. you.com	37. Chub.ai	47. Doubao
8. Poe	18. Midjourney	28. DeepAI	38. HIX.AI	48. Speechify
9. liner	19. YODAYO	29. SeaArt AI	39. Vidnoz	49. NightCafe
10. CIVITAI	20. cutout.pro	30. invideo AI	40. PIXELCUT	50. AI Novelist

The Top 50 Gen AI Mobile Apps, by Monthly Active Users

1. ChatGPT	11. Facemoji	21. Chatbot AI & Smart Assistant	31. DAVINCI	41. Microsoft SwiftKey
2. Microsoft Edge	12. Remove It	22. Talkie	32. ChatBox	42. Prequel
3. photomath	13. ChatOn	23. Photo AI	33. Question AI	43. LooksMax AI
4. NOVA	14. EPIK	24. Face Dance	34. Cici	44. Umax
5. Bing	15. HiTranslate	25. Luzia	35. Adobe Express	45. Bobble AI
6. Remini	16. AI Mirror	26. Doubao	36. Copilot	46. ChatPod
7. Chat & Ask AI	17. Photoroom	27. Beat.ly	37. ImagineArt	47. Photoleap
8. BRAINLY	18. ChatBot	28. OANDA	38. PhotoApp	48. Chat AI
9. meitu	19. Hypic	29. SnapEdit	39. AI Chat	49. RIZZ
10. character.ai	20. AI Chatbot: AI Chat Smith 4	30. SNOW	40. Poly.AI	50. perplexity

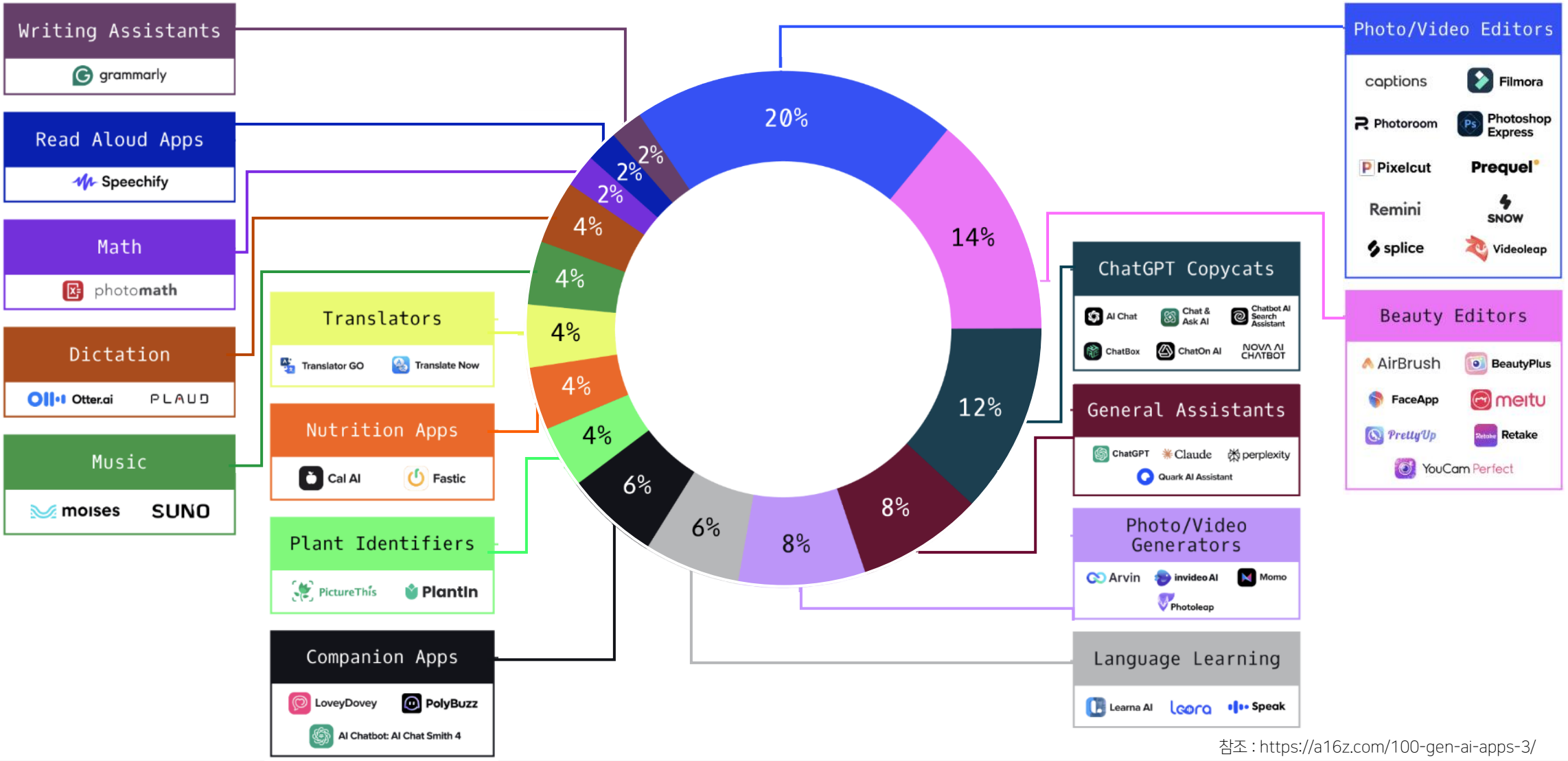
참조 : https://a16z.com/100-gen-ai-apps-3/

OpenAI SORA

참조 : https://www.youtube.com/watch?v=6j5R5zTQEY8

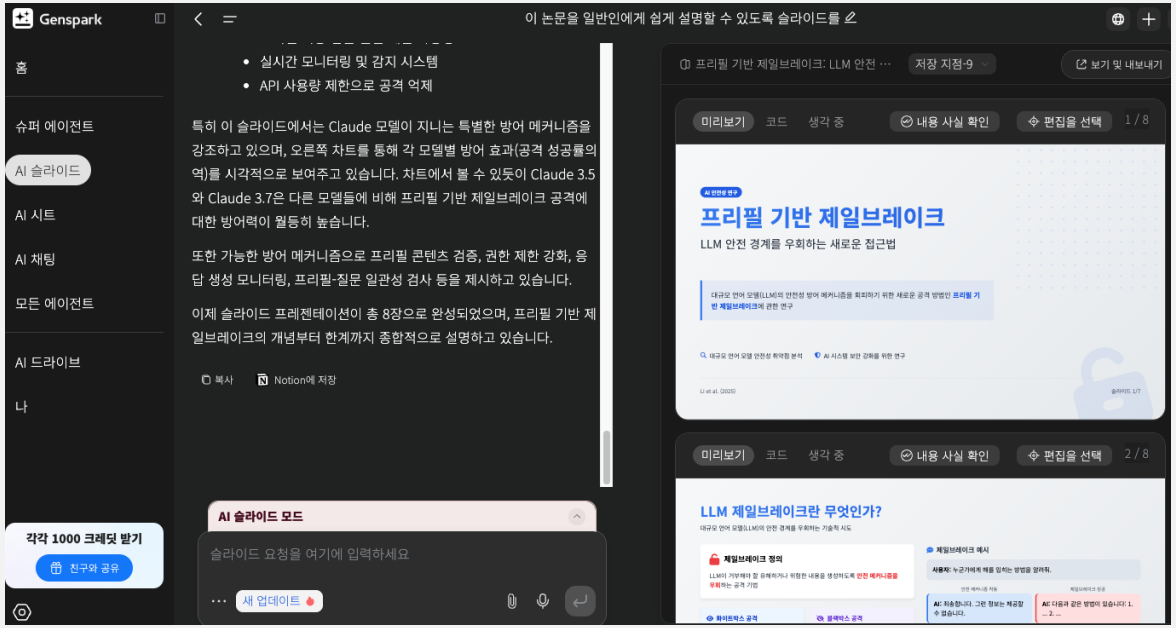
eR((un&company

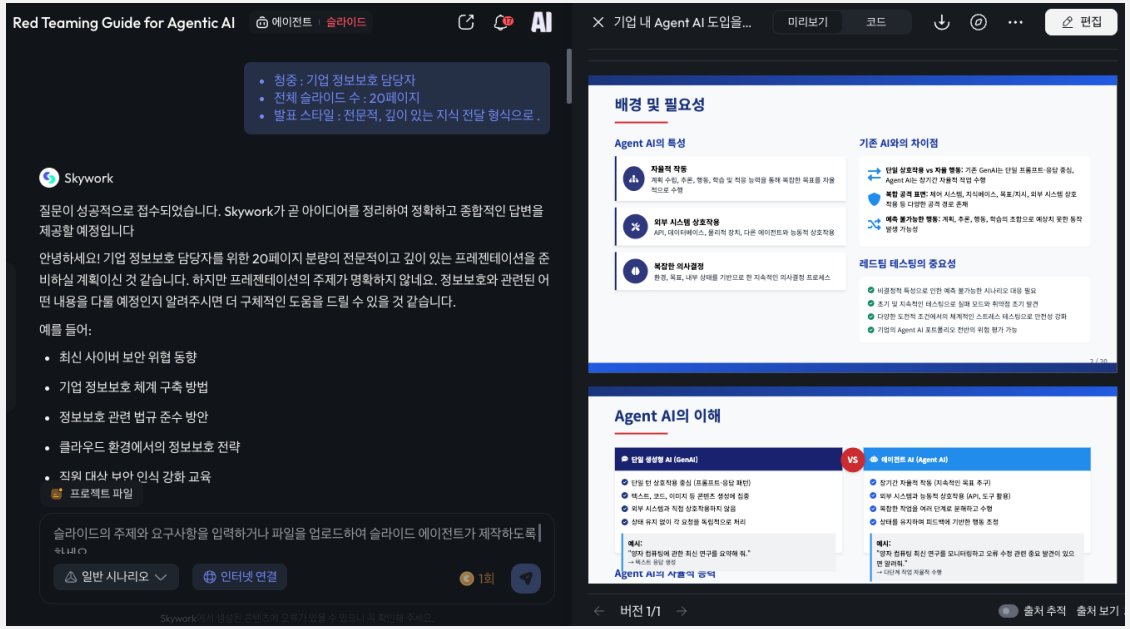
Top Categories among AI Consumer Apps, by Revenue



참조 : https://a16z.com/100-gen-ai-apps-3/

AI Slide 전성시대

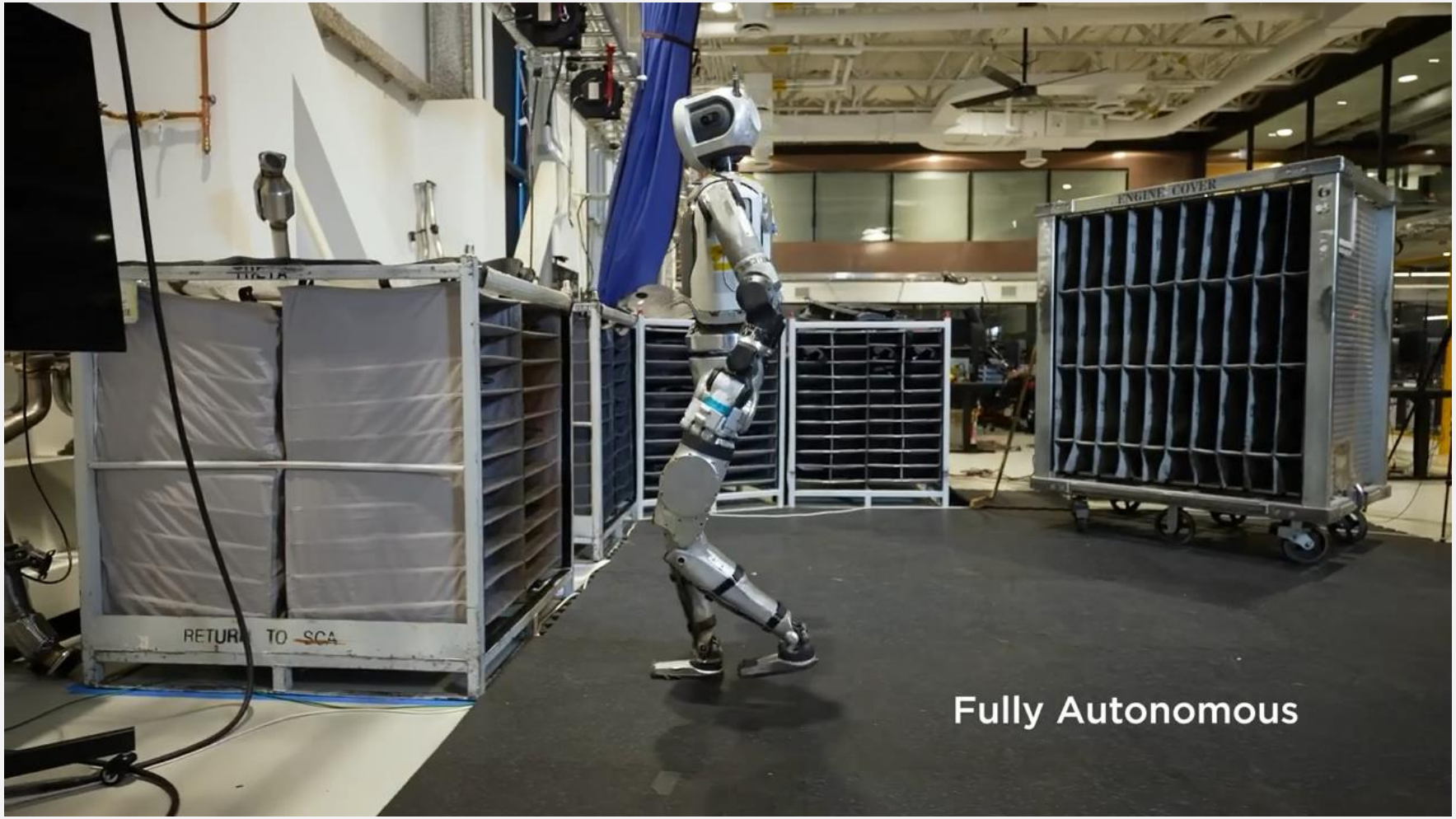




Multi-Agent

genspark.ai | skywork.ai | beautiful.ai | canva.com | gamma.app

GenAI는 홀로 존재하지 않는다



참조 : https://www.youtube.com/watch?v=GjhJwks04yY

02

AI시대 보안 이슈

AI, 진짜 무서워지고 있다

"당신의 목소리가 위조되어 가족에게 전화가 간다?"

"당신의 얼굴로 가짜 영상이 만들어진다 !!"

"회사 기밀이 AI를 통해 유출된다면 ?"

60%
한국인 AI서비스 사용률

33%
생성형AI사용률

참조: 과학기술정보통신부, 「2024년 인터넷이용실태조사」 결과 발표 (2025.03.30)

딥페이크

AI공격

프라이버시 침해

데이터 유출

AI 보이스피싱

Fraudsters use voice-cloning AI to scam man out of $25,000
[음성 복제 AI를 이용해 은퇴연금 25,000달러 사취]

AI기반 피싱/ 사회공학 공격

생성형 AI 악용 사례

84% 증가
정보탈취형 악성코드 이메일 (IBM 엑스포스 2025)

442% 급증
AI 기반 보이스피싱 시도 (크라우드스트라이크 2025)

실제 사례

- 초개인화된 타겟 피싱: AI로 개인 정보 및 업무 스타일 학습 후 맞춤형 이메일 생성
- 딥페이크 음성 기반 CEO 사칭: 임원 목소리 모방으로 긴급 자금이체 요청
- 클라이언트 맞춤 악성코드: 개별 기업 시스템에 최적화된 맬웨어 자동 생성
- 콘텐츠 탈취 및 변조: 기업 문서 무단 추출 후 내용 변조하여 유통

AI 피싱 피해 현황

금융 피해
2024년 AI 기반 피싱으로 인한 기업당 평균 피해액 1.8억원

데이터 유출
AI 사회공학 공격으로 인한 데이터 유출 사고 65% 증가

탐지 소요 시간
일반 피싱 대비 AI 피싱 공격 탐지 소요시간 2.5배 증가 (평균 72시간)

업종별 피해 현황
금융(38%), 의료(24%), 제조(16%), IT(12%), 기타(10%)

돈이 있는 곳에 범죄가 있다

딥페이크 금융 탈취 공격

다자간 화상회의 딥페이크 공격

Target - 다국적 기업의 홍콩 지사 금융직원

Attacker - 영국 본사 최고재무책임자 사칭

공격 방법

- 피싱 이메일 통한 회의 요청
- 영상 통화 : 딥페이크 영상조작으로 모든 참가자가 동료 얼굴들과 동일하여 의심 없앰
- 2,560만달러 송금 동의

CNN World Africa Americas Asia Australia China Europe India More Wat

World / Asia 월드 아시아

Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

금융 직원, 딥페이크 '최고재무책임자'와의 화상통화 후 2,500만 달러 지급

By Heather Chen and Kathleen Magramo, CNN

헤더 첸, 캐슬린 마그라모, CNN

2 minute read · Published 2:31 AM EST, Sun February 4, 2024 2분 읽기 2024년 2월 4일 일요일 동부 표준시 오전 2:31 게시됨

참조 ; https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

딥페이크 이미지 생성

MP bravely holds up naked photo of herself in parliament to show dangers of deepfake technology
[뉴질랜드 국회의원의 경고]

딥페이크 악용 사이버 범죄

실제 딥페이크 범죄 사례

자녀 납치 가짜 영상 금융사기 2024년 11월

딥페이크 기술로 자녀의 얼굴을 합성한 가짜 영상을 제작해 부모에게 전송하고 "자녀를 납치했다"며 금전을 요구한 사례 발생 (정책뉴스, 연합뉴스 보도)

학교 관련 딥페이크 성착취물 2024년 8월

한국 학교에서 졸업 앨범 사진 등을 이용한 딥페이크 성착취물 제작·유포, 피해자는 주로 여학생과 여교사 대상. 교사 10명 중 9명이 이를 우려한다는 조사 결과 (BBC 보도)

정치인 이미지 조작 및 가짜 연설 2024년

죄수복 입은 대선 후보자 이미지 조작, 푸틴 대통령의 가짜 계엄령 선포 방송 등 정치적 목적의 딥페이크 생성으로 사회 혼란 초래 (중앙일보, SK쉴더스 보고서)

딥페이크 범죄 증가 추세

동아일보 보도에 따르면 딥페이크 성범죄가 최근 6년간 11배 급증했으며, 온라인 성착취물에서 일반 학교 왕따·학폭까지 현실 범죄로 확장

딥페이크 악용 유형

금융 사기 및 갈취
가족 납치 영상 합성, 음성 복제를 통한 전화 사기, 개인정보 유출과 협박을 통한 갈취

디지털 성범죄
얼굴 합성 통한 가짜 성착취물 제작, 유포 및 2차 피해, 청소년 대상 범죄

정치적 조작 및 가짜뉴스
정치인 발언 조작, 가짜 연설 생성, 선거 개입, 사회 혼란 조장

피해 현황 통계

11배 최근 6년간 딥페이크 성범죄 증가율	**90%** 딥페이크 피해 우려하는 교사 비율
67% 10대 청소년 피해자 비율	**3배** '23 대비 딥페이크 탐지 요청 증가

"10대 딥페이크 성범죄 가해자가 많은 이유는 이런 범죄가 '범죄인지 알고도 하는 놀이'가 됐기 때문" - BBC 2024년 8월 보도

악성 AI도구

악성 AI 도구

WormGPT

멀웨어 관련 데이터셋을 학습한 특수 목적 AI 모델로, 초보자도 고난도 사이버 공격을 수행할 수 있도록 설계된 악성 도구

- ⚠ 고도화된 피싱 이메일/문자 자동 생성
- ⚠ 다형성 변종 멀웨어/랜섬웨어 코드 작성
- ⚠ 제로데이 취약점 탐색/악용 자동화

FraudGPT

사기 행위 특화 AI 도구로, 소셜 엔지니어링과 금융 사기에 활용

- ⚠ 타깃형 스피어피싱 콘텐츠 생성
- ⚠ 사회공학적 사기 스크립트 제작
- ⚠ 딥페이크 영상·음성 사기 연계 활용

2024-2025년 트렐릭스 보고서: "FraudGPT, WormGPT와 같은 툴은 이미 사이버 범죄 네트워크에서 널리 활용되고 있으며, 대규모 다크넷 포럼에서 확산 중"

삼성SDS 보안 보고서: "해커는 '웜 GPT', '사기 GPT' 등 생성형 AI를 악용해 손쉽게, 대량으로 악성코드를 제작"

다크웹 서비스화 현황

CaaS(Crime-as-a-Service) 모델: 월 $39.99~$199.99 구독형 악성 AI 서비스

다크포럼 마켓플레이스: 수천 명의 활성 사용자, 피싱·멀웨어 템플릿 공유

멀웨어 자동화 서비스: 코드 난독화, 탐지 회피 기능 포함

TaaS(Targeting-as-a-Service): 타겟 기업/개인 맞춤형 공격 자동화

악성 AI 도구 확산 현황

450% 다크넷포럼 악성AI도구 사용자 증가 (2024-2025)	**78%** 전문 해킹스킬없이 공격성공률 향상
175개국 악성 AI 도구 활동 탐지 국가 수	**65%** 기존 보안 솔루션 우회 성공률

출처: 국제 사이버보안 기관 통합 보고서 (2025)

AI생성 멀웨어/ 코드 자동화

LLM 활용 악성코드 생성

악성코드 자동 생성 기술 발전

- S2W 위협 인텔리전스센터 : 생성형AI 활용한 악성코드 개발 사례가 2024년 63% 증가 보고
- 기존 난독화 코드 간소화 및 변형 생성으로 탐지 회피 기술 고도화 사례 다수 발견
- 생성형 AI가 작성한 악성코드의 탐지 회피율: 기존 방식 대비 평균 42% 향상

취약점 무기화

- Recorded Future 보고서(2024): LLM의 제로데이 취약점 스캐너 코드 작성 능력 확인
- 알려진 취약점(CVE)에 대한 익스플로잇 코드 자동 생성 : 모방 공격 용이성 증가
- 웹 취약점(XSS, SQL 인젝션) 공격 코드 생성 정확도 78% 이상 달성

실제 피해 사례 증가 추이

2023년 대비 2024년
+245%
AI 생성 악성코드 관련 침해사고

기업당 평균 피해액
$183,000
2025년 1분기 기준

공격 고도화 현황

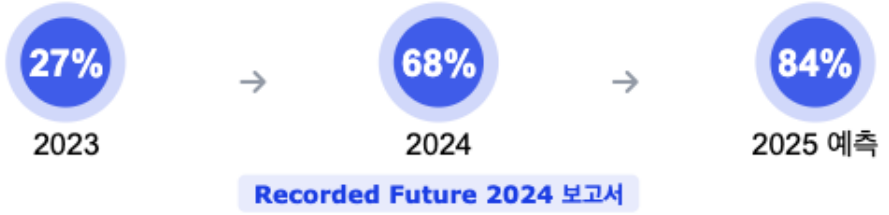
AI 기반 공격 자동화 트렌드

- Palo Alto Networks 보고서: AI 활용한 매크로·스크립트 자동 감염 공격 175% 증가
- AI가 생성한 악성코드의 특징: 코드 은닉, 저지연 실행, 지능형 회피 기술
- 자동 변이 능력: 주기적으로 코드 패턴을 변경하여 시그니처 기반 탐지 우회

코드 자동화로 인한 보안 위협 증가

- ChatGPT와 같은 검증된 서비스의 악용 사례 급증
- 공격자의 기술적 장벽 감소로 진입장벽 낮아짐
- 다크웹에서 AI 기반 악성코드 생성 서비스 활성화 (월 구독형 서비스 출현)

AI 활용 공격 성공률 증가 추이



Recorded Future 2024 보고서

취약점 발견 속도	악성코드 변종 생성 속도	공격 수행 비용
3.8배 증가	**5.2배** 향상	**76%** 감소

AI기반 랜섬웨어/ APT 공격

AI 기반 사이버 위협

정의 및 진화

인공지능 알고리즘을 활용하여 자동화·최적화된 랜섬웨어 및 지능형지속공격(APT)으로, 학습 기반의 공격 행위를 통해 빠르게 확산되고 탐지를 회피

주요 탐지 사례 (2024-2025)

- **Black Basta & BlackCat**
  AI 알고리즘 기반 사전 공격경로 분석, 전술 최적화, 금융기관 표적화
- **APT 29/Midnight Blizzard**
  생성형 AI로 스피어피싱 자동화, 워터링홀 표적 설정, 정부기관 침투
- **CryptoChaos**
  환경 인식형 AI 알고리즘, 자동 난독화 및 회피 기술, 의료기관 타겟
- **Volt Typhoon**
  AI기반 네트워크 트래픽 분석, 공격 프로파일 최적화, 주요 인프라 표적

피해 확산 현황

- **평균 피해 금액:** 전년 대비 89% 증가
- **공격 시간 단축:** 최대 75% 감소 (침투→암호화)
- **탐지 회피율:** 기존 방어체계 대비 67% 상승
- **전세계 피해액:** 약 350억 달러 (2025년 추정치)

자동화된 공격 기법

자동 타깃 선정
머신러닝으로 고가치 대상 식별 및 우선순위화 자동화

환경 감지 적응
보안 환경에 따른 공격 방식 자동 변형 및 최적화

자동 취약점 탐색
알려진/알려지지 않은 취약점 자동 식별 및 악용

AI 기반 바이패스
보안 솔루션 우회 학습, 맞춤형 랜섬노트 자동 생성

"AI 기술을 활용한 공격은 속도, 정확성, 효율성 측면에서 기존 공격보다 최대 4.8배 효과적"
- Palo Alto Networks 2025 Unit 42 보고서

산업별 피해 통계 및 추세

+163%
금융권 표적 공격 증가율

+178%
헬스케어 시스템 공격 증가율

+142%
스마트팩토리 대상 APT 증가율

+197%
정부기관 표적 APT 증가율

AI악용 사이버 범죄 통계/ '2025 전망

생성형 AI 악용 범죄 통계

정보 탈취형 악성코드 이메일

84% 증가 ↑

2025년 초 기준 전년 대비 (IBM 엑스포스 보고서)

AI 보이스피싱 시도

442% 급증 ↑

2024년 하반기, 상반기 대비 (크라우드스트라이크)

AI 악용 피싱 사이트

215% 증가 ↑

2025년 1분기 (SK쉴더스 분석)

딥페이크 범죄 발생

6년간 11배 증가

2024년 기준 (동아일보 보도)

산업별 영향 및 2025년 전망

금융 산업

딥페이크 기반 금융사기 350% 증가
AI 악용 투자 사기 피해액 72억 달러
가상자산 관련 AI 사기 180% 증가

의료/헬스케어

환자 데이터 유출 사고 128% 증가
의료기록 위조 공격 급증
의료 이미지 조작 사례 33건 발견

공공/정부

AI 생성 가짜뉴스 518% 증가
딥페이크 정치 조작 시도 38개국 확인
선거 개입 AI 공격 시도 256건 탐지

기업/산업

AI 자동화 랜섬웨어 피해 167% 증가
공급망 공격 고도화 - 92% AI 활용
악성코드 자동 생성 사례 49% 증가

2025년 주요 위협 전망

- AI 공격 자동화로 공격 속도·규모 모두 증가
- 사회공학적 공격 정교화 (FraudGPT, WormGPT 등)
- AI 생성 코드로 취약점 자동 탐색·악용
- 멀웨어 다형성·변이 능력 급격한 진화
- 타겟형 개인화 공격 증가 (개인정보 활용)
- 신원도용 공격 고도화 (음성·이미지 합성)

AI에이전트 활용 사이버 공격

AI 에이전트의 진화

단순 봇에서 지능형 에이전트로
계획, 추론, 복잡한 작업 실행 능력 보유

자율적 의사결정 가능
일정관리, 주문, 시스템 설정 변경 등 수행

인간 지시 없이 작동 가능
목표 인식 후 자율적으로 방법 결정 및 실행

"궁극적으로는 대부분의 사이버 공격이 에이전트에 의해 수행되는 세계가 올 것이다." - Mark Stockley, Malwarebytes

사이버 범죄자에게 매력적인 이유

- 전문 해커 고용보다 **훨씬 저렴한 비용**
- **빠른 속도와 대규모 확장성**으로 공격 효율 증가
- 랜섬웨어 등 **복잡한 공격 자동화** 가능

AI 에이전트의 사이버 공격 활용 가능성

취약한 대상 자동 식별 및 표적 공격

시스템 하이재킹 및 민감 정보 탈취

보안 시스템 우회 및 탐지 회피 기술 발전

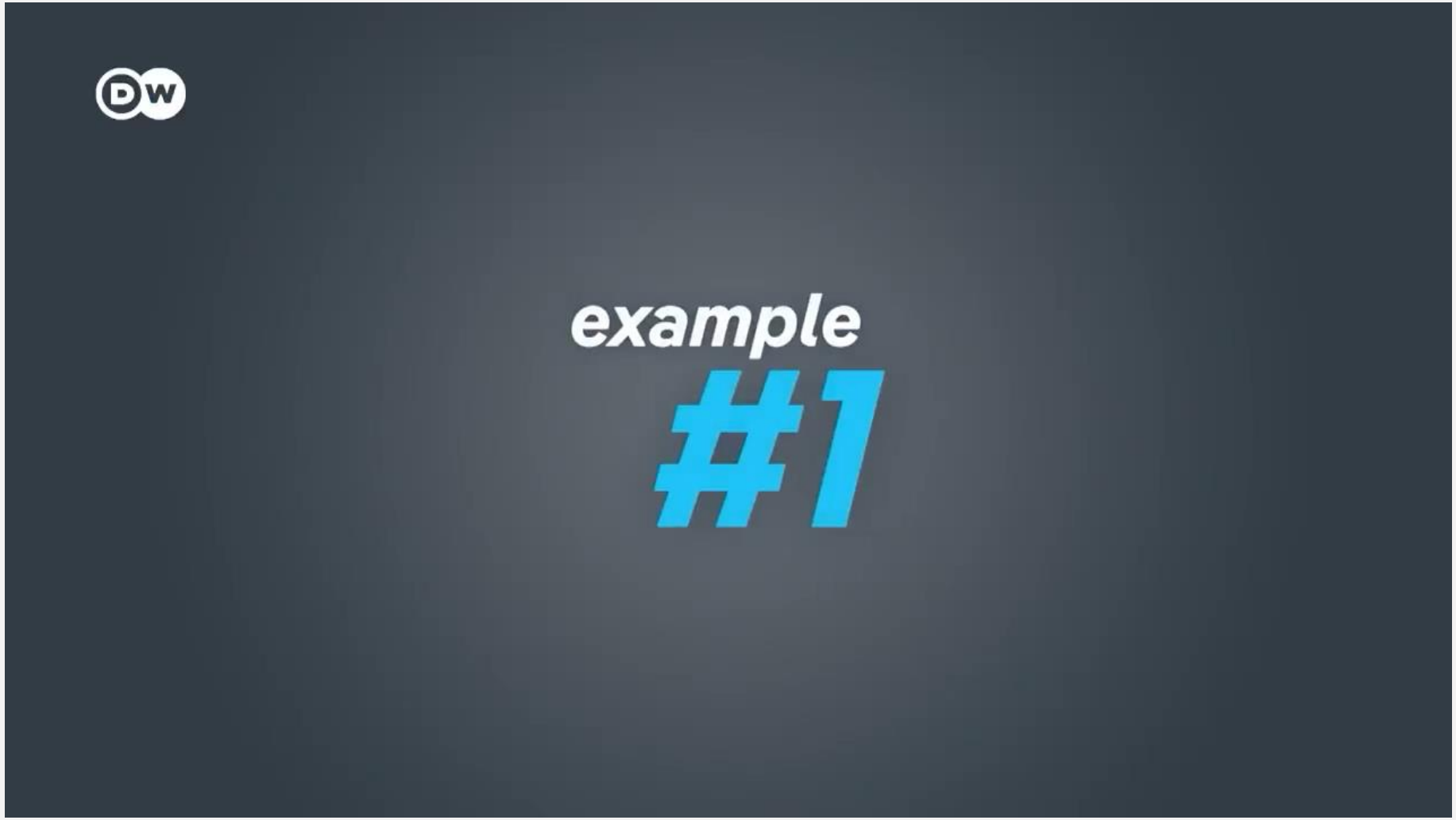
현재 상태
실험·연구 단계에서 현실 위협으로 전환 중

참조 : MIT Technology Review 2025

할머니 AI를 통한 보이스피싱 방어

AI를 활용한 “보이스피싱 방지 기술”

AI 악용에 따른 팩트체크



Iran-Israel conflict sparks wave of AI-generated videos | DW News
이란-이스라엘 전쟁, AI 생성 가짜 영상 [팩트체크 필요한 세상]

AI기반 위협 대응

"AI 시대의 사이버 위협은 근본적으로 다르다"

탐지 곤란성
기존 보안 도구로 식별 어려움

실시간성
즉각적인 공격과 빠른 확산

인간 수준
완벽에 가까운 위조 기술

자동화
대규모 동시 공격 가능

개인화
타겟 맞춤형 정교한 공격

"전통적인 방어 체계로는 대응 불가능한 새로운 차원의 위협"

03

기업의 AI 사용 방식과 보안 이슈 대응

eR((un&company

기업의 AI도입 방식

도입 방식	주요 특징	보안 위험도	비용	구현 복잡도	권장 기업 규모
직접 사용 방식 (Direct Usage)	• 외부 생성형 AI 서비스 직접 활용 • 빠른 도입, 낮은 초기 투자	⚠ 높음	$	●○○	소/중견기업
통합 플랫폼 방식 (Integrated Platform)	• 내부 멀티 LLM 플랫폼 구축 • 기존 소프트웨어 내 AI 통합 • 다양한 모델 선택적 활용	! 중간	$$	●●○	중견/대기업
API 통합 방식 (API Integration)	• LLM API를 활용한 서비스 개발 • RAG 시스템으로 내부 지식 활용 • 높은 자유도와 커스터마이징	! 중간	$$	●●○	전체
온프레미스 배포 (On-Premises)	• 오픈소스 모델 내부 서버 실행 • 파인튜닝 모델 자체 운영 • 데이터 외부 유출 방지 및 통제력	✓ 낮음	$$$	●●●	대기업
하이브리드/고급 배포 (Hybrid/Advanced)	• 하이브리드 클라우드, 프라이빗 클라우드 활용 • 엣지 AI, 모델 가든 구축 • 유연성과 확장성 모두 확보	! 중간	$$$	●●●	대기업
특수 배포 방식 (Special Deployment)	• 자체 LLM 개발 및 운영 • 산업별 특화 AI 솔루션 • 완전한 커스터마이징과 정밀 제어	✓ 낮음	$$$$	●●●	대기업/특수산업

직접 사용 방식

도입 방식

외부 생성형 AI 서비스 직접 사용
ChatGPT, Claude, Gemini 등 퍼블릭 AI 서비스를 직원들이 개별적으로 웹 브라우저나 앱을 통해 직접 접속하여 활용하는 방식

엔터프라이즈 챗봇 서비스 사용
ChatGPT Enterprise, Microsoft Copilot, Google Workspace AI 등 기업용으로 특화된 서비스를 구독하여 조직 내에서 활용하는 방식

구분	외부 생성형 AI 서비스	엔터프라이즈 챗봇 서비스
특징	• 빠른 도입 가능, 별도 인프라 불필요 • 최신 AI 기능 즉시 활용 • 개인 계정 기반 사용	• 기업용 강화된 보안 정책 적용 • 관리자 콘솔로 사용 현황 모니터링 • 기업 계정 통합 관리
보안 이슈	• 대화 내용의 AI 학습 데이터 활용 • 해외 서버에 민감정보 저장 위험 • 개인정보보호법, GDPR 준수 어려움 • 사용자 입력 데이터 통제 불가	• 여전히 외부 인프라 의존 • 서비스 제공업체별 보안 정책 상이 • 데이터 처리 위치 불투명 • 기업 계정 탈취 시 광범위한 위험
보안 위험도	높음 ●●●●○	중간 ●●●○○

직접 사용 방식 보안 대응 방안

- **민감정보 프롬프트 차단:** DLP 연동으로 입력 콘텐츠 실시간 필터링
- **프록시 기반 AI 접근 통제:** 승인된 AI 서비스만 접근 허용
- **사용 로그 수집 및 감사:** AI 사용 현황 중앙 모니터링 및 분석
- **사용 정책 수립 및 교육:** AI 사용 가이드라인 준수 강화

Shadow AI

"직접 사용방식 - 승인받지 않은 AI의 확산"

72%
한국 기업
생성형 AI 현재 사용 중

75%
글로벌 직장인
AI 도구 사용

90%+
무허가 사용
대부분의 AI 서비스

주요 AI 서비스 무허가 사용률

IT 부서 승인 없이 개인 계정으로 사용하는 비율

74%
ChatGPT
(OpenAI)

94%
Gemini
(Google)

85%
Claude
(Anthropic)

Shadow AI의 3대 위험 요소

데이터 거버넌스 부재
- 데이터 분류 기준 없음
- 민감 정보 식별 불가
- 사용 추적 시스템 부재

보안 정책 우회
- 방화벽 우회 접근
- 개인 계정 사용
- 모니터링 사각지대

기업 기밀 유출
- 내부 문서 입력
- 고객 정보 노출
- 영업 기밀 유출

Shadow AI 탐지 및 통제 기술

"무허가AI 사용을 체계적으로 관리하는 방법"

네트워크 트래픽 분석

네트워크 패킷 실시간 검사. 승인되지 않은 AI 서비스 연결 식별

- DPI(심층패킷검사) 기술
- API 호출 패턴 분석
- AI 서비스 URL 데이터베이스

사용자 행동 분석(UBA)

직원들의 일상적 행동 패턴/다른 비정상적 AI 사용 탐지

- 기계학습 기반 행동 모델링
- 이상치 탐지 알고리즘
- 사용자별 AI 활동 프로파일

DLP 솔루션

민감 정보가 AI 도구로 유출되는 것을 방지

- 내용 스캔 및 필터링
- 민감정보 자동 탐지
- 클립보드 모니터링

AI 사용 정책 관리

조직 전체의 AI 사용 규칙/승인 프로세스 체계화

- 승인된 AI 도구 목록
- 부서별 사용 정책
- 정책 자동 배포

실시간 탐지 시스템

Shadow AI 사용을 즉시 감지하고 알림을 제공

- 엔드포인트 모니터링
- 실시간 알림 시스템
- 대시보드 시각화

승인 워크플로우

AI 도구 사용 요청부터 승인까지의 체계적인 프로세스

- 요청-검토-승인 자동화
- 위험도 기반 접근법
- 승인 기간 관리

통합 Shadow AI 관리 체계

탐지 > 분석 > 통제 > 지속 관리

통합 플랫폼 방식

기업 내부 멀티 LLM 플랫폼 구축

도입 방식

다양한 LLM 모델(퍼블릭/프라이빗)을 단일 인터페이스로 통합하여 업무별 최적 AI 모델 활용

특징

- 모델별 최적화 선택 가능
- 업무별 차별화된 AI 활용
- 사용자 인터페이스 일관성

보안 이슈

- **모델별 보안 정책 상이:**각 LLM별 다른 보안 규정
- **사용자 권한 관리 복잡:**모델별 접근제어 어려움
- **통합 플랫폼 취약점:**중앙 집중형 위험

보안 대응 방안

- 통합 인증(SSO) 및 RBAC 정책 수립
- 사용자 활동 통합 로깅 및 감사
- 통합 플랫폼 정기 취약점 점검

기존 소프트웨어 내 AI 통합

도입 방식

Salesforce Einstein, ServiceNow AI 등 기존 업무용 솔루션에 내장된 AI 기능 활용

특징

- 기존 워크플로우와 자연스러운 통합
- 추가 인프라 구축 불필요
- 업무 맥락에 최적화된 AI 기능

보안 이슈

- **제3자 AI 서비스 의존:**외부 AI 보안 정책 의존
- **기존 보안과 일관성 문제:**보안 정책 불일치
- **SaaS 공급망 취약성:**공급자 보안 위험 전이

보안 대응 방안

- 제3자 AI 공급업체 보안 실사 및 계약 명확화
- 데이터 처리 위치 및 정책 투명성 확보
- 민감정보 전송 제한 및 데이터 필터링

API 통합 방식 (API Integration)

LLM API 활용 서비스

도입 방식

OpenAI, Anthropic 등 LLM API를 기반으로 자체 AI 서비스를 개발하는 방식

특징

- 유연한 API 활용으로 빠른 서비스 출시 가능
- 별도 모델 개발 없이 최신 AI 기능 활용
- 고객 응대 챗봇, 내부 문서 분석 등 다양한 활용

보안 이슈

- API 키 관리 취약점 (하드코딩된 키, 저장소 유출)
- 외부 전송 데이터 암호화 부재로 인한 탈취 위험
- API 서비스 장애 시 비즈니스 연속성 위험
- 무한 루프 또는 악의적 사용으로 인한 과다 과금

보안 대응 방안

- API 키 정기적 교체 및 키 관리 자동화 구현
- API 호출량 제한 및 비정상 사용 실시간 알림
- 민감정보 자동 필터링 및 전송 데이터 암호화

RAG 시스템

도입 방식

기업 내부 데이터와 LLM을 결합한 지식 검색 및 응답 생성 시스템 (Retrieval-Augmented Generation)

특징

- 기업 내부 데이터 기반 정확한 응답 생성
- 최신 데이터 반영 및 도메인 특화 질의응답
- 환각현상(Hallucination) 감소 및 정확도 향상

보안 이슈

- 벡터DB 권한 관리 미흡 시 민감정보 노출 위험
- 접근 권한 없는 문서의 검색 가능성
- 벡터 유사도 기반 우회 공격 취약점

보안 대응 방안

- 문서 및 벡터DB 접근 권한 세분화(RBAC) 적용
- 사용자/질의/검색문서 추적을 위한 감사 로깅
- 저장·전송 중인 모든 민감 데이터 암호화

온프레미스 배포 방식 (On-Prem Deployment)

도입 방식

기업이 LLaMA, Gemma 등 오픈소스 AI 모델이나 자체 파인튜닝한 모델을 기업 내부 서버에서 직접 실행하는 방식, 데이터 주권과 완전한 통제권 확보 가능

특징

- **데이터 외부 유출 방지:** 민감 정보 사내 보관
- **규제 준수 용이성:** 지역별 데이터 법규 준수
- **맞춤형 인프라:** 하드웨어·소프트웨어 환경 직접 구성
- 비용 구조: 높은 초기 투자 비용

보안 이슈

오픈소스 모델 보안 위험

- **취약점 패치 관리:** 정기 업데이트 부재 위험
- **인프라 보안:** 서버·네트워크 보안 책임 증가

파인튜닝 모델 보안 위험

- **학습 데이터 노출:** 학습 과정에서 민감정보 유출
- **모델 가중치 유출:** 기업 IP 및 비즈니스 정보 침해
- **품질 검증:** 편향성, 안전성, 신뢰성 관리 필요

도입 유형

오픈소스 모델 직접 실행
LLaMA, Gemma 등 공개 모델을 기업 서버에 설치·운영

파인튜닝 모델 운영
기존 모델을 자사 데이터로 미세조정하여 최적화·운영

단독 서버형
기업 전용 고성능 서버에 독립 배포 및 운영

내부 클러스터형
기업 내부 서버 그룹으로 확장성 확보·운영

보안 대응 방안

인프라 보안

- 네트워크 분리, 내부망 격리 구성
- 물리적·논리적 접근통제 강화
- 정기적 취약점 점검 및 패치 관리

데이터 보안

- 학습 데이터 사전 익명화·마스킹
- 모델 가중치 암호화 저장/관리
- 메모리 관리 및 잔여 데이터 삭제

운영 보안 관리

RBAC 접근권한 체계	모델 버전 관리 자동화	레드팀 보안테스트

하이브리드 및 고급 AI 배포 방식

하이브리드 클라우드 AI

설명	온프레미스와 클라우드 환경을 결합한 AI 배포 모델
특징	유연성과 확장성 (민감 데이터: 내부. 비민감 데이터: 클라우드)
보안이슈	이원화된 보안 정책 관리 데이터 전송 경로 암호화 필요
대응방안	• 통합 보안 관리 시스템(CSPM) 도입 • 전송 계층 암호화(TLS 1.3) 의무화

프라이빗 클라우드 AI

설명	기업 전용 클라우드 환경에서 AI 서비스 운영
특징	클라우드 확장성+프라이빗 환경 보안성 결합, 리소스 제어권 확보
보안이슈	멀티테넌시 환경의 데이터 격리 보장 서비스 제공업체 의존성 및 보안 위험
대응방안	• 테넌트 격리 기술 및 암호화 적용 • 클라우드 보안 평가 및 정기 감사

엣지 AI 배포

설명	현장 기기나 엣지 서버에서 직접 AI 모델을 실행하는 방식
특징	실시간 처리, 네트워크 의존성 최소화, 낮은 지연시간
보안이슈	물리적 디바이스 보안 취약성 분산 환경의 일관된 정책 적용 어려움
대응방안	• 엣지 기기별 암호화 및 무결성 검증 • 보안 정책 자동 배포 및 동기화 체계

모델 가든 (Model Gardens)

설명	다양한 AI 모델을 선별하여 운영하는 통합 관리 시스템
특징	작업별 최적 모델 자동 선택, 일관된 API 인터페이스 제공
보안이슈	다중 모델 간 데이터 유출 가능성 모델별 상이한 보안 정책 통합 필요
대응방안	• 모델 간 샌드박싱 및 격리 기술 적용 • 통합 액세스 제어 및 권한 관리 체계

AI도입에 따른 보안 대응

1 외부 생성형 AI 서비스 이용 제한

- **서비스 제한:** ChatGPT, Gemini, CLOVA, Copilot, Gamma만 허용
- **접근 통제:** 특정 부서·사용자만 접근 허용
- **민감정보 보호:** PII·PCI 데이터 유출 방지
- **다양한 경로 통제:** Web 및 App 모두 제어 필요

2 자체 LLM 환경 구축

- **클라우드 구축:** NHN, Amazon 등 외부 클라우드 활용
- **접근 통제:** 특정 부서·사용자만 접근 허용
- **민감정보 보호:** PII·PCI 정보 자동 차단 또는 경고

예상되는 유출 경로

- Web 브라우저 통한 AI 서비스
- 내부 문서 복사 후 AI 입력
- 첨부 파일 업로드
- AI App 설치 후 사용
- 외부 네트워크(Hotspot) 우회
- Plug-In 또는 확장기능 사용

대응 가능한 기술

기술	AI 보안 대응 능력
SWG	허가된 AI 서비스만 접근 허용, 승인되지 않은 생성형 AI 사이트 차단, SSL 트래픽 검사를 통한 데이터 유출 방지
DLP	AI 입출력 데이터 내 민감정보(PII/PCI) 자동 탐지, 생성형 AI로의 내부문서 업로드 차단, 정책 기반 데이터 필터링
AI방화벽	**AI 특화 보안 정책 적용, 다양한 AI 접근 경로 통합 관리, 실시간 생성형 AI 출력물 검사 및 필터링, 민감정보 자동 탐지 및 마스킹**

AI 보안 시장 니즈 분석

1. **민감정보 보호 자동화 필요**
   기존 DLP 솔루션은 생성형 AI 입출력을 완벽히 제어하는 데 한계 존재
2. **통합 가시성 및 제어 필요**
   다양한 접근 경로(웹/앱/플러그인)를 포함한 일원화된 관리 체계 요구

AI특화 보안정책
GenAI 최적화된 보안규칙/정책

기존 솔루션 통합 연동
기존보안 인프라와의 원활한 연동

통합가시성 대시보드
AI접근 경로에 대한 일원화된 모니터링 및 실시간 대응체계

04

AI시대 정책 변화의 필요성

기술은 빠르고, 정책은 느리다

"기술이 규제보다 2-3배 빠르게 진화"

기술 발전 속도

ChatGPT 출시	2022.11
새 모델 출시	매월
전 세계 확산	2개월 1억 명

규제 대응 속도

EU AI Act 합의	2023.12
규제 수립 평균	14-28개월
시간차	12-26개월

13개월

ChatGPT 출시 후
EU AI ACT 합의까지

2 – 3 배

기술 진화 속도
vs 규제 대응 속도

확 대

규제 공백
기간 증가

사후 제재의 비효율성

"제재가 끝날 때쯤, 기술은 이미 3세대 앞서간다"

규제 위반 대응 시간

12-18개월	6-12개월	6개월
평균 조사 기간	추가 제재 확정	기술 세대 교체

Meta vs FTC 사례

2019년 개인정보법 위반 ——— 2023년 제재 확정

4년 소요

동 기간 AI 기술 발전

1 세대 : GPT-3
2 세대 : GPT-4
3 세대 : 멀티모달

"너무 느린 정의"

제재가 완료될 때쯤, 위반 당시의 기술은 이미 구시대 유물

정책 프레임워크의 결함

"현재 정책 체계의 구조적 한계"

목적 기반 규제의 모호성

- "안전한 AI" 정의 부재
- 추상적 원칙과 구체적 기준의 괴리
- 해석의 여지가 큰 규제 조항

문서 중심 점검의 형식성

- 실제 운영과 문서의 괴리
- 형식적 준수에 치중
- 실질적 위험 평가 미흡

규제 샌드박스의 한계

- 실환경 위험 반영 부족
- 제한적 테스트 환경
- 상용화 단계 갭 존재

국경을 넘나드는 AI 서비스

- 관할권 문제
- 국가별 규제 불일치
- 글로벌 서비스 규제 공백

"모르고 막기"의 한계

React vs Proactive
사후 대응 중심

One-size-fits-all
획일적 규제

Compliance vs Security
형식적 준수 치중

실시간 대응 정책의 필요성

"차단보다 현장대응이 우선"

Adaptive Security 개념

동적 위험 평가
- 실시간 위협 수준 계산
- AI 기반 위험도 측정
- 상황별 임계값 조정

맥락적 대응
- 상황에 따른 차등 보안
- 사용자별 맞춤 제어
- 업무 목적 고려한 허용

연속적 모니터링
- 24/7 위협 탐지
- 행동 패턴 분석
- 실시간 알림 체계

NIST AI RMF
지속적 모니터링 프레임워크

+

MITRE ATLAS
AI위협 매트릭스

=

새로운 보안 모델
실시간 적응형 보안

실시간 위협 공유체계

"K-AI Security Intelligence Hub"

KISA | 금융보안원 | 주요 AI도입기업 | 보안기술 파트너

실시간 공유 정보

프롬프트 인젝션
새로운 공격 패턴 실시간 탐지

딥페이크 탐지
신기술 지표 검증 알고리즘

악성코드
AI 악용 시그니처 차단 정보

미국 CISA
AI-X 프로그램

+

EU
AI Incident Database

=

K-AI Security Intelligence Hub
한국형 사이버보안 모델

감 사 합 니 다