Headline

생성형 AI 기술 발전에 따른 보안위협과 대응 전략

관제사업그룹/관제사업1팀 박선호 수석

■ 개요

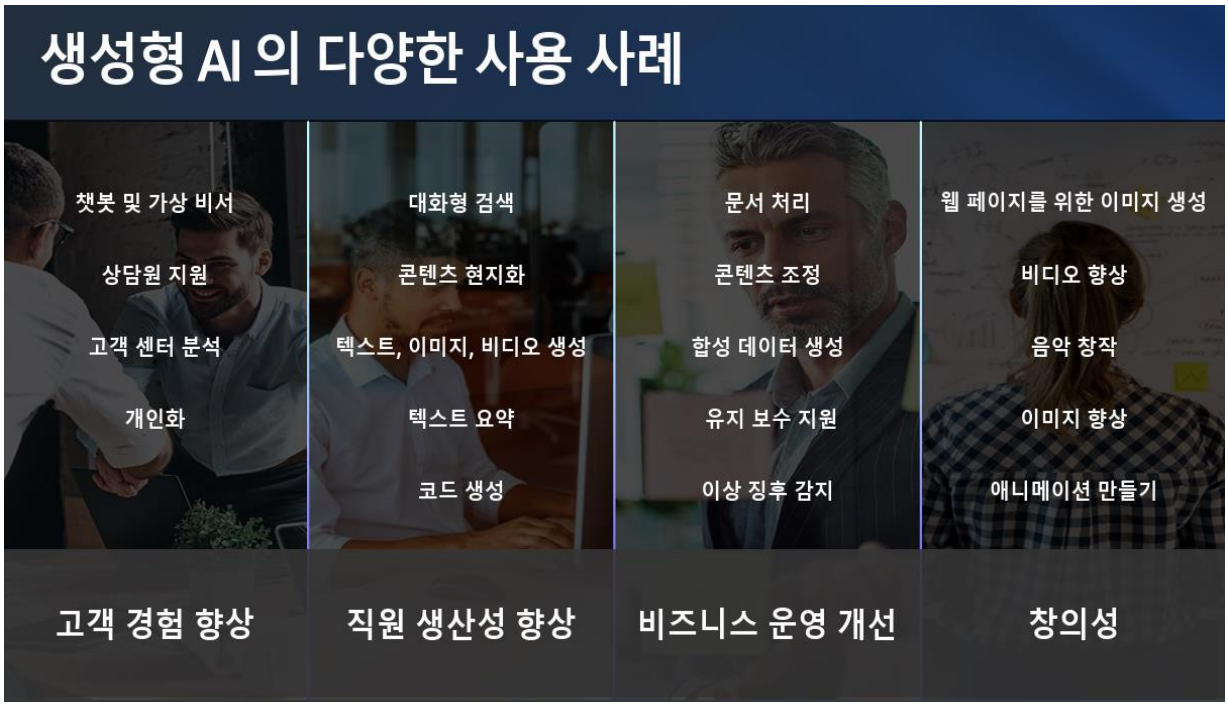
생성형(Generative) AI는 인터넷 등에서 학습한 내용을 기반으로 대화, 예술, 음악, 소프트웨어 코드, 글쓰기 등 사용자의 요구에 따라 새롭고 독창적인 콘텐츠를 만들 수 있는 인공지능 기술이다. 광범위한 데이터로 사전 훈련된 대형 모델들을 기반 모델(Foundation Models, FMs)로 사용하고 있으며, '해외 주요 생성형 AI 현황'은 아래와 같다.

	기업명	서비스명	국가	내용
텍스트	오픈 AI	챗 GPT	미국	초 거대 언어 AI 모델 GPT를 바탕으로 만든 대화형 AI 서비스
	구글	Bard	미국	초 거대 언어 AI 모델 LaMDA를 바탕으로 만든 대화형 AI 서비스
	DeepMind	Sparrow	영국	DeepMind 의 언어모델 Chinchilla 를 기반으로 한 AI 챗봇
	Jasper	Jasper	미국	마케팅 목적의 블로그 기사, 소셜미디어 게시물 및 광고 문구 등을 생성하는 AI 툴
	Baidu	Ernie Bot	중국	지식 통합을 통한 향상된 표현이라는 의미의 자체 개발 AI 챗봇
이미지	오픈 AI	DALL·E	미국	프롬프트(명령어)에 따른 이미지 생성
	Stability AI	Stable Diffusion	영국	이미지 생성 AI로 오픈소스 소프트웨어
	Midjourney	Midjourney	미국	이미지 생성 AI로, 해당 툴을 사용하여 생성한 작품이 미국의 한 미술대회에서 1위로 선정되어 화제가 됨
음성	구글	MusicLM	미국	문자 설명을 음악으로 만드는 생성 AI
	오픈 AI	Jukebox	미국	원하는 장르, 가수 스타일로 음악을 생성하는 AI 기술
영상	구글	Imagen Video	미국	최대 초당 24 프레임, 1280X768 해상도의 비디오를 생성할 수 있는 Text to Video AI 생성 툴
	메타	Make-A-Video	미국	텍스트 입력 시 동영상을 생성해주는 Text to Video AI 모델

출처 : KPMG

표 1. 해외 주요 생성형 AI 현황

■ 생성형 AI 발전 및 도입 사례



출처 : AWS

그림 1. 생성형 AI 의 다양한 사용 사례

생성형 AI 기술이 급속도로 발전하면서 다양한 업계에 빠르게 적용되고 있다. 생성형 AI 는 단순 검색과 상담 수준에서 복잡한 문서 요약, 이메일 초안 작성, 코드 작성, 광고 문구 작성을 비롯하여 소설·영상·그림·음악 등 콘텐츠 창작까지 가능하다.

실제로 생성형 AI 기반으로 제작된 업무 솔루션은 보고서나 이메일, 설문지 등의 초안을 생성해준다. 맞춤법 수정까지 자동으로 제공해 업무 시간을 획기적으로 단축할 수 있다. 생성형 AI 기반 코딩 툴킷도 나왔다. 코딩 툴킷에 원하는 프로그래밍 언어와 코드 방식을 자연어 방식으로 기재하면, AI 는 작업에 필요한 코드를 개발자 요청에 따라 생성해 낸다.

생성형 AI 중에서는 OpenAI 에서 출시한 대형 언어 모델(Large Language Model) 기반의 GPT(ChatGPT)가 대중적으로 많이 사용되고 있다. 또한, 2023 년 3 월 '오토 GPT(AutoGPT)[1])'가 공개되었는데, 이는 목표를 정해주면 추가적인 지시 없이 자율 반복(Autonomous iterations) 기능을 사용해 AI 가 알아서 학습하고 방법을 찾아내어 목표를 달성한다. 이미 생성형 AI 의 기술 수준은 자율적 인공지능인 범용인공지능(AGI: Artificial General Intelligence)에 빠른 속도로 근접하고 있다.

[1] 오토 GPT(AutoGPT) : 23 년 3 월 영국의 토란 브루스 리차드가 개발, OpenAI 의 GTP-4 기반으로 동작하는 파이썬 오픈소스 라이브러리

■ 생성형 AI 기반 정교화된 피싱 공격

안타깝게도 생성형 AI 기술에는 순기능과 역기능이 동시에 존재한다. 전 세계적으로 각국의 선거를 앞두고 생성형 AI 를 활용한 정치인 딥페이크 영상을 제작 및 유포하고 있으며, 이를 통해 다량의 가짜 정보를 양산함으로써 사람들을 혼란에 빠뜨리고 있다. 또한 유명 연예인들의 얼굴과 음성을 조작한 가짜 영상을 제작해 투자를 권유하는 사기를 행하거나, 유명인과 포르노 영상을 합성한 음란물을 제작하는 등 매우 부적절한 악용 사례도 늘고 있다.

방송통신심의위원회 자료에 따르면, 2020 년 6 월부터 지난해 8 월까지 약 3 년간 불법 성적 영상물 시정 건수가 9,006 건으로 집계됐다. 2020 년 473 건을 시작으로, 지난해 8 월 기준 3,046 건까지 증가하며 매년 성장세를 더하고 있다.

이처럼 생성형 AI 기술 발전에 따라 보안 위협도 대두되고 있는 만큼, 이에 대한 철저한 대비가 필요하다. 공격자들은 생성형 AI 를 활용해 피싱 메일, 악성코드를 생성하거나 소스코드 내 취약점 식별, 랜섬웨어 유지보수 등 다양한 분야에서 적극 활용 중이다.

다크웹 등에서는 기업용 이메일 공격(BEC: business email compromise)과 정교한 피싱 작업 수행을 위한 도구로 '웜 GPT(WormGPT)[2]'와 '사기 GPT(FraudGPT)[3]'가 떠오르고 있다. 생성형 AI 를 활용한 피싱 프로그램을 사용하면 전문 지식 없이도 손쉽게 대량으로 악성코드를 제작할 수 있다. 또한, 과거와 달리 어색하거나 문맥이 맞지 않는 부분 없이 정교한 피싱 메일 작성이 가능해진다.

AI 도입 및 활용에도 주의가 필요하다. 세계 최대 인공지능(AI) 개발 플랫폼인 허깅페이스(Hugging Face)에서 악성 코드가 숨겨진 AI 모델 등이 100 개 이상 발견되었는데, 주요 모델은 파이토치(95%), 텐서플로(5%)로 확인되었다. 해당 악성 모델에는 시스템 제어권 탈취(50%), 백도어 설치(20%) 기능을 필두로 특정 파일의 설치 및 실행, 임의 코드를 실행하는 등의 기능이 포함된 것으로 확인됐다.

■ 생성형 AI 관련 보안 위협 구분

생성형 AI 관련 보안 위협은 크게 내부적, 외부적 요인으로 구분할 수 있다.

내부적 요인으로는 사용자의 보안 인식 부족 및 컴플라이언스 미준수에 따른 정보유출, AI 모델이 제공하는 부정확한 정보의 무검증 사용, 자체 이용 및 관리하는 AI 모델 관리 소홀 등이 있다.

외부적 요인으로는 AI 모델 및 관련 애플리케이션 취약점, 해커에 의한 생성형 AI 를 이용한 공격시도, 악성코드 또는 백도어가 포함된 AI 모델의 사용 등이 있다.

내부적 요인	외부적 요인
- 민감 정보 및 기밀 문서 등 등록 - AI 모델이 제공하는 부정확한 정보의 무검증 사용 - 자체 이용 및 관리하는 AI 모델 관리 소홀	- AI 모델 및 관련 애플리케이션 취약점 공격 - 생성형 AI 를 악용한 악성 메일, 피싱 공격 - 악성코드 또는 백도어가 포함된 AI 모델의 사용

표 2. 생성형 AI 관련 내외부적 요인

2 웜 GPT(WormGPT) : 비영리 오픈소스 그룹인 일루서 AI(EleutherAI)에서의 GPTJ 언어 모델 기반의 AI, 무제한 문자 및 채팅 메모리 보존, 코드 포맷 기능 등 다양한 기능 제공

3 사기 GPT(FraudGPT) : 인공지능 챗봇 기술을 응용해 기업용 이메일 공격인 BEC(business email compromise) 공격을 제공(구독료 : 200 달러/월), 배후에는 웜 GPT(WormGPT)가 있는 것으로 추정

■ 생성형 AI 위협 대응 전략

생성형 AI 가 기술 산업 전반에 빠르게 확산되면서, 그 기술적 진보가 가져올 수 있는 새로운 보안 위협에 대한 선제적인 대비가 필요하다.

여러 공공기관에서는 사전 검토 받은 내용만 입력하도록 규정하거나 '챗 GPT 활용방법 및 주의사항 안내서'를 배포하고 있고, 일부 민간기업에서는 사내 인트라넷에서의 한정된 사용, 입력 글자수 제한 등 올바른 사용을 유도하고 있다. 범정부 차원의 법제도적 세부 가이드 등이 추가적으로 필요한 상황이며, 각 기업 및 기관 자체적으로 대응방안을 고민해야 한다.

정보보안 관련 컴플라이언스의 엄격한 준수를 기본으로 생성형 AI 를 활용한 분석, 대응 방법을 고려해 볼 것을 제안한다.

생성형 AI 를 활용한 분석 및 대응 방법
1. 입력 가능 내용 규정 - 사전 규정되거나 검토된 내용에 대해서만 입력 - 계정 정보, 신용카드 및 개인정보 입력 금지 - 업무 기밀사항 입력 금지
2. 악성 메일 관련 보안 교육 및 업무용 이메일 관리 체계 강화 - AI 를 통해 정교하게 위조된 악성 메일 주의 필요
3. 생성물의 정확성, 윤리성, 적합성, 보안성 등을 재평가 후 사용 - 생성물이 정확한 내용인지, 법적/윤리적으로 문제없는 것인지 검토 후 사용 - 프로그램 코드 생성 사용시 변수명 변경 등 소스코드 보안성 검토 후 사용
4. 사내 데이터 안정성 확보 - 접속 계정에 대해 다중 인증(MFA : Multi-Factor Authentication) 사용 - 자체 이용 및 관리하는 AI 모델에 대한 접근 제한 설정 - AI 에 대한 응답 및 질의 제한 키워드(계정정보 등) 설정 - API 키에 대한 안전한 관리
5. 신뢰할 수 있는 AI 모델, 애플리케이션 사용 및 취약점 수시 확인
6. AI 를 적용한 악성코드 분석, 위협 식별 등 방어 기술을 확보에 지속 노력

표 3. 생성형 AI 를 활용한 분석 및 대응 방법

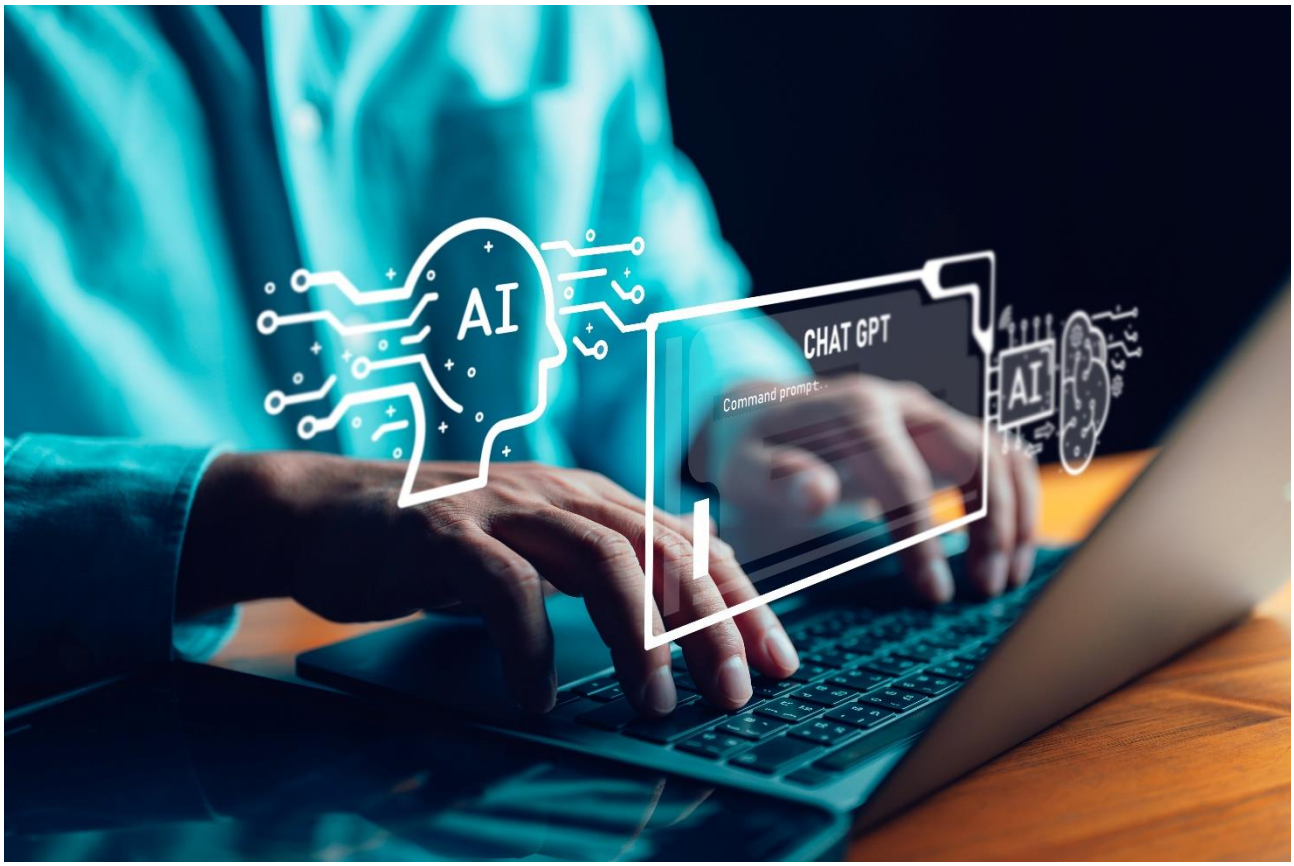
■ 국내 법·제도적 현황

2023 년 6 월 국가정보원에서 '챗 GPT 등 생성형 AI 활용 보안 가이드라인'을 발간했으나, 범정부 차원의 생성형 AI 사용 관련 세부적인 가이드라인 등 추가 마련이 필요한 시점이다.

개인정보보호위원회에서는 2024 년 말까지 개인정보보호법의 적용 원칙과 기준을 구체화한 AI 단계별 6 대 가이드라인을 마련할 예정이며, 6 대 가이드라인에는 공개된 정보, 비정형 데이터, 생체인식 정보, 합성 데이터, 이동형 영상기기, 투명성 확보 등에 대한 구체적인 법 적용 내용을 담을 계획이다. 또한, 분야별 전문가 42 명으로 구성된 '2024 개인정보 미래 포럼'을 출범하여 개인정보 분야 미래 의제를 선제적으로 논의하고 의견을 수렴하여 대응하겠다는 방침이다.

3 월 15 일부터는 전 분야에 걸친 최초의 인공지능 규제인 '자동화된 결정에 대한 대응권'이 시행된다. 23 년 3 월 개정된 개인정보보호법 제 37 조의 2 로 신설된 자동화된 결정에 대한 정보주체의 대응권의 내용으로 '거부권'과 '설명 요구권'으로 구분된다. 또, 인공지능 기술을 적용한 시스템을 포함하여 완전히 자동화된 시스템으로 개인정보를 처리하여 이루어지는 결정이 정보주체 자신의 권리 또는 의무에 중대한 영향을 미치는 경우 해당 개인정보처리자에 대하여 해당 결정을 거부할 수 있는 권리를 가지며, 정보주체는 개인정보처리자가 자동화된 결정을 한 경우에는 그 결정에 대하여 설명 등을 요구할 수 있다는 내용이다.

■ 맺음말



지금까지 생성형 AI 기술 발전에 따른 보안위협과 대응 전략 및 국내 법제도적 현황에 대해 알아보았다. 생성형 AI는 사용하기에 따라 위험하기도, 유용하기도 한 '양날의 검'이 될 수 있다. 악용될 경우 더욱 정교한 공격이 가능해 치밀한 대응이 필요하다.

국내 정보보안 1위 기업인 SK 쉴더스는 생성형 AI를 활용한 피싱 공격에 대비할 수 있는 '이메일 보안관제' 서비스를 제공하고 있다. 이메일 보안관제 서비스는 24시간 365일 상시모니터링을 지원하고, 악성 공격 패턴에 대한 전문가 분석 및 위협정보 등을 제공한다. 이메일 발신자 주소와 발신 IP, 이메일 내 URL, 첨부파일 이상 유무 및 도메인 등을 토대로 종합적인 악성메일 유무 판단 및 악성 행위 상세 분석을 진행한다.

특히, 전문적인 APT 장비 운영과 분석 대응 역량을 갖추고 있어 이메일 첨부파일 내 악성코드가 삽입되어 있는 등 일반 사용자가 인식하기 어려운 보안 위협도 꼼꼼하게 분석 및 대응한다. 이외에도 고도화되는 이메일 피싱 공격에 안전하게 대응할 수 있도록 실시간 악성 메일 현황 보고와 악성 메일 모의 훈련, 악성 메일 동향 및 대응 방안 등도 제공한다.

이외에도 20여 년의 컨설팅 노하우로 고객 맞춤형 정보자산 보호 컨설팅을 제공하고 있다. 보안 컨설팅과 관련한 자세한 내용은 SK 쉴더스 블로그에서 확인할 수 있다.