공격자 vs 방어자, AI 보안의 미래는?

인공지능(AI) 기술은 그 어느 때보다 빠르게 발전하고 있다. 특히 생성형 AI는 점점 정교해지고 있으며, 이제 일상적인 대화형 서비스를 넘어 다양한 산업 분야에서 혁신을 이끌고 있다. ChatGPT를 시작으로 마이크로소프트(MS) 바사-1(VASA-1), 헤이젠(HeyGen), 소라 AI(Sora AI), 수노 AI(Suno AI) 등이 등장하면서, 얼굴 사진 한 장과 오디오 파일만으로 사람이 말하고 움직이는 영상을 만들고 주어진 영상을 40개국 언어와 300개 이상의 목소리로 변환하며, 영화를 제작하거나 음악 작곡까지 할 수 있게 됐다.

하지만 이러한 기술의 발전과 함께 AI를 악용하는 사례도 늘어나고 있어, 보안 업계는 새로운 도전 과제에 직면해 있다. 이번 글에서는 AI가 만들어내는 위협의 유형과 AI 악용 사례, 그리고 AI를 활용한 안랩의 보안 고도화 전략을 소개한다.

1. 생성형 AI의 잠재적 위협

AI로부터 파생되는 사이버 위협은 다양한 방식으로 나타나고 있으며, 대다수는 이미 실제 발현되어 심각한 피해를 초래하고 있다. 생성형 AI로부터 발생되는 잠재적인 위험은 크게 4가지가 있다.

[그림 1] 생성형 AI의 잠재적 위험

먼저, 환각(Hallucination)은 AI가 학습한 데이터를 바탕으로 정보 생성하는 과정에서 발생하는 오류를 의미한다. AI는 가끔 실제로 존재하지 않거나 사실이 아닌 정보를 생성하기도 하는데, 이는 사용자의 잘못된 의사결정을 유도하고 신뢰성 문제를 일으킬 수 있다.

편향(Bias)은 학습한 데이터에 내재된 편향성이 문제를 일으킬 때 나타난다. AI는 학습 데이터를 통해 패턴을 학습하고 예측하는데, 만약 학습 데이터에 성별, 인종, 종교와 같은 고정관념이나 차별적 표현이 포함되면 편향된 정보가 생성될 가능성이 커진다.

저작권(Copyright) 문제는 AI가 대량의 데이터를 학습하는 과정에서 저작권이 있는 콘텐츠를 사용하는 경우, 저작권자의 권리를 침해할 수 있다는 점이다. AI가 생성한 콘텐츠에 대한 저작권 문제는 아직 법적으로 해결되지 않고 있으며, 본격적인 AI 시대를 맞아 필히 해결해야 할 문제다.

데이터 프라이버시 침해(Data Privacy Breaches) 역시 간과할 수 없는 중요한 문제다. AI 모델 학습에 사용되는 데이터가 개인정보나 민감한 정보일 경우, 데이터가 노출 혹은 악용될 경우 큰 문제를 일으킬 수 있다. 이는 개인정보의 불법 사용으로 이어져 개인정보 침해를 초래할 수 있다.

2. 생성형 AI를 향한 공격

공격자들은 생성형 AI가 상용화되고 많은 기업과 개인들이 다양한 용도로 생성형 AI를 사용하는 점을 노려 생성형 AI에 대해 다양한 방식으로 공격을 전개하고 있다. 생성형 AI 공격을 큰 틀에서

보면 ▲적대적 프롬프팅(Adversarial Prompting) ▲데이터 오염(Data Poisoning) ▲모델 역분석(Model Reversing)이 있다.

먼저, 적대적 프롬프팅은 탈옥(Jailbreaking)과 프롬프트 유출(Prompt leaking)로 구분할 수 있다. 탈옥은 AI의 윤리적 제한을 무력화하거나 우회하는 방식으로, 공격자는 교묘하게 설계된 요청을 통해 AI 시스템의 규칙을 회피한다. 이로 인해, AI가 위험하거나 불법적으로 작동할 수 있다.

프롬프트 유출은 말 그대로 생성형 AI에 대한 프롬프트를 탈취하는 것을 의미한다. 프롬프트는 AI가 주어진 작업을 정확하게 수행할 수 있도록 돕는 지침 또는 명령어다. 만약 시스템 프롬프트가 깃허브(GitHub)와 같은 코드 공유 플랫폼을 통해 유출되면, 공격자가 이를 악용해 AI를 왜곡하거나 신뢰성을 훼손할 수 있다. 특히 GPT 모델과 같은 대형 언어 모델에서는 프롬프트 유출이 큰 문제가 되며, 이는 데이터 보안에도 중대한 위험을 초래할 수 있다.

두 번째로, 데이터 오염은 AI 시스템을 학습시키는 과정에서 악의적인 정보나 왜곡된 데이터를 의도적으로 삽입해 AI의 출력을 고의적으로 왜곡시키는 공격이다. 대표적인 사례로는 마이크로소프트(Microsoft)의 AI 채팅봇 '테이(Tay)' 사건이 있다. 2016년 출시된 테이는 사용자가 제공하는 데이터를 바탕으로 학습하고 대화했지만, 일부 사용자가 욕설, 인종차별, 성차별 등 부적절한 데이터를 입력하자 테이는 이를 학습하여 매우 공격적이고 혐오적인 발언을 남발했다. 결국 마이크로소프트는 테이를 출시한 지 16시간 만에 운영을 중단해야 했다. 이 사건은 AI의 학습 데이터 품질이 얼마나 중요한지, 또 공격자가 AI를 어떻게 침해할 수 있는지 보여준 사례다.

세 번째 모델 역분석은 AI 모델에 대한 반복적인 쿼리를 수행해 모델의 작동 방식을 분석하고, 학습 데이터를 탈취하는 공격 기법이다. 공격자는 모델의 응답을 바탕으로 학습 데이터나 모델의 내부 구조를 파악하고, 이를 통해 기밀 정보나 개인정보 등 민감한 정보를 유출시킨다. AI 모델의 보안 취약점을 악용하는 공격 기법으로, 데이터 보호와 모델 보안 측면에서 큰 위협을 초래한다.

3. AI 적용으로 고도화되는 사이버 위협

공격자들은 AI를 공격하는 것을 넘어 이들의 공격 캠페인에 AI를 빠르게 접목시키고 있다. AI를 이용한 공격은 ▲악성코드 제작 ▲피싱 공격 ▲딥페이크(Deepfake) ▲취약점 분석 및 해킹 등이 있고, 정교함과 규모 면에서 이전보다 훨씬 더 위험한 수준에 이르고 있다.

[그림 2] AI를 활용한 공격 방법

먼저, AI는 악성코드와 스크립트를 작성하는 데 활발히 사용되고 있다. 공격자는 AI를 통해 탐지 회피를 위한 다양한 변형 코드를 대량으로 작성하며, 이를 통해 기존 보안 시스템을 우회할 수 있게 된다. 특히 AI는 랜섬웨어나 개인정보 탈취를 위한 악성코드를 생성하는데 있어, 공격 효율성과 속도를 대폭 향상시킨다. 최근에는 AI를 활용해 비밀번호와 암호화폐 지갑 주소 등 민감 정보를 포함한 이미지를 검색하고, 이를 악용하는 사례도 증가하고 있다.

또한, AI는 공격자가 피싱 공격을 더욱 정밀하게 수행할 수 있도록 한다. AI를 통해 생성된 피싱 메일 문구는 마치 사람이 작성한 것처럼 매우 자연스럽다. 과거에는 특정 단어나 문장 구조에서 피싱임을 식별할 수 있었지만, 현재는 AI가 대형 언어 모델(LLM)을 기반으로 완벽하게 모방함으로써 정상적인 업무 메일이나 안내 메일과 구별하기 어려운 수준의 피싱 메일을 제작한다. AI의 발전으로 피싱 메일 수 과거 대비 40배 이상 증가했으며, 이는 AI가 실제 사이버 공격에 미치는 영향력이 매우 크다는 점을 잘 보여준다.

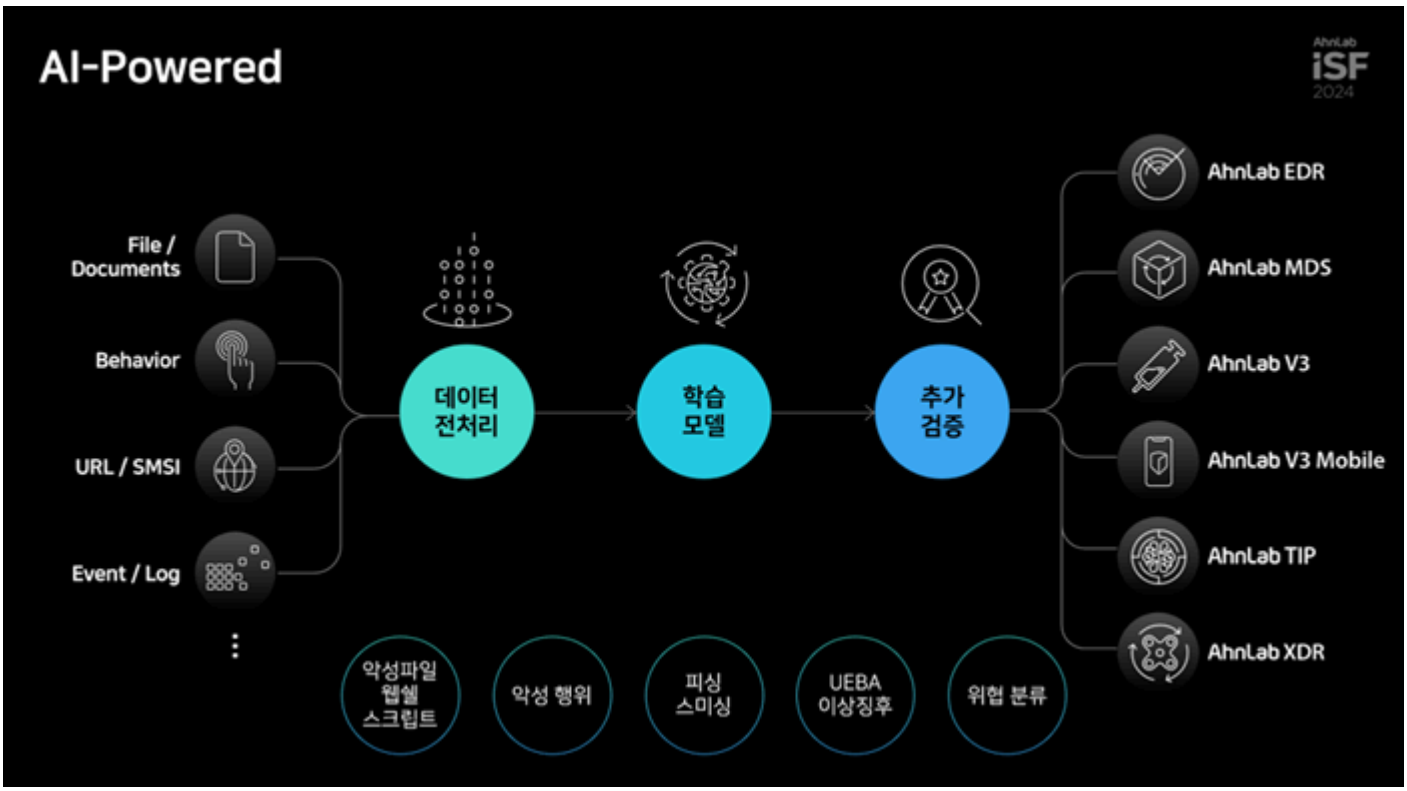
딥페이크 기술은 사진 몇 장과 음성 샘플만으로 현실감 있는 가짜 영상을 만들 수 있도록 한다. 딥페이크 기술의 품질은 갈수록 향상되고 있으며, 가짜 뉴스, 사회적 혼란을 야기할 수 있는 영상 콘텐츠 제작, 보이스피싱 등 여러 범죄에 악용되고 있다. 유명 인사의 얼굴을 합성한 영상으로 가짜 발언을 퍼뜨리거나, 특정 인물의 목소리를 이용해 금융 정보를 탈취하는 공격이 가능해졌다.

AI는 공격자가 취약점을 분석하고 새로운 해킹 기법을 개발하는 데에도 핵심적인 역할을 하고 있다. AI 기술의 발전으로 공격 기술이 대중화되고 있으며, 진입 장벽은 낮아지고 있다. 또, 다양한 공격 방식을 결합해 새로운 해킹 기법을 개발하고 사이버 공격의 효율성을 극대화할 수 있게 됐다.

취약점 탐지와 패턴 분석을 통해 알려지지 않은 보안 취약점을 찾아 제로데이(Zero-Day) 공격에 활용할 수도 있다.

4. AI를 활용한 안랩의 보안 고도화 전략

이처럼 AI 기술을 활용한 공격은 매우 정교하고 복잡하기에 기존의 전통적인 보안 접근법으로는 이를 효과적으로 방어하기 어렵다. 따라서 AI를 이용한 공격과 새로운 위협에 대응한다는 관점에서도 AI 중요성은 굉장히 크다고 할 수 있다.



[그림 3] 안랩의 AI 기술 적용 방안

안랩은 AI 기술을 활발하게 적용해 자사 솔루션과 플랫폼을 고도화시키고 있다. 안랩의 AI 기술은 악성 파일과 정상 파일을 분류하는 파일링 기술에 기반을 두고 있으며, 이 외에도 행위 기반 분석, URL 정보, 이벤트 로그 등 다양한 데이터를 활용해 학습 모델을 훈련시키고 있다. 이러한 기술은 V3, V3 Mobile, EDR, MDS, TIP, XDR 등 모든 주요 제품에 적용되어 전체적인 위협 탐지, 분석 및 대응 역량을 강화시킨다. 또한, 악성코드 및 피싱 메일 탐지와 대응, 잠재적 위협 예측, 종합적인 리스크 관리 등 다양한 측면에서 보안 기술을 발전시키고 있다.

안랩의 AI 기반 보안 기술은 다음과 같이 크게 4가지 영역에서 활용되고 있다.

1. **AI 시큐리티 어시스턴트(AI Security Assistant):** 보안 담당자가 보안 이벤트를 실시간으로 모니터링하고 대응할 수 있도록 지원하는 일종의 AI 비서 서비스다. AI 시큐리티 어시스턴트

는 침해 내용을 분석하고, 이를 바탕으로 향후 대응 방안을 제시한다.

2. **AI 탐지 및 대응 (AI Detection & Response):** 딥페이크나 악성코드, 피싱 메일 등 다양한 위협을 실시간으로 탐지하고 대응하는 기술이다. AI는 사람보다 훨씬 빠르고 정확하게 이상 행위를 식별해 침해사고 발생 후 최대한 빠른 시간 내에 대응할 수 있도록 한다.
3. **AI 사전 대응(AI Proactive Detection):** AI는 공격이 발생하기 전에 잠재적인 위험 요소를 예측하고 경고한다. 공격자의 행위를 학습해 공격 의도를 사전에 파악하고, 보안 담당자의 효과적인 위협 방어를 지원한다.
4. **AI 중심 SOC(AI-native SOC):** SOC에서 AI를 통해 보안 상태를 모니터링하고, 발생한 위협에 대해 자동으로 대응하는 체계를 구현한다.

다음으로, 안랩이 AI 기술을 구체적으로 어떤 분야에 어떻게 적용하고 있는지 살펴보자.

1) 악성코드 분석 및 대응

AI는 악성코드를 분석하고 공격자의 의도를 추론하는 데 핵심적인 역할을 한다. 안랩은 클라우드 기반 악성코드 위협 분석 및 대응 인프라인 ASD(AhnLab Smart Defense)에 AI를 탑재해 악성코드를 분석하고 대응한다. ASD에 탑재된 AI는 악성코드의 실행 코드와 외향적 특징을 분석하고, 행위 동적 분석을 통해 공격 기법과 유입 경로를 종합적으로 파악한다. 또한, 유사 샘플 및 연관 샘플들을 비교 분석해 악성코드를 분류하고 탐지 룰을 자동으로 생성한다. 이를 통해, 공격자의 의도를 파악 및 추론하고 후속 공격 예측을 자동화해 사전 방어 태세를 강화한다. 안랩의 V3와 V3 Mobile, MDS는 AI를 활용해 스미싱 메일과 악성 파일을 탐지하고 차단하며, AhnLab TIP는 악성코드 분석 리포트, 동향 보고서를 AI로 작성한다.

2) 데이터 증강

AI는 새로운 유형의 악성코드와 공격 기법에 대한 패턴과 변형을 예측하고 학습하는 방식으로 데이터를 증강한다. 이를 통해, 조직은 사전 대응 체계를 강화할 수 있다. 과거에는 악성코드 샘플이 수집될 때까지 기다려야 했지만, AI는 신규 샘플이 들어오면 이를 자동으로 분석하고 변형을 예측한다. 예를 들어, 피싱 메일이나 공격 스크립트의 변형을 AI가 예측하고, 이를 학습에 활용한다. 또한, AI는 공격 스크립트 내 악의적인 명령어를 찾고, 유사한 명령을 검색할 수 있는 정규표현식을 작성한다. 이러한 데이터 증강 기술은 피싱 메일, 스미싱 문자, 실행 파일 탐지에 활용된다.

3) AI 기반 피싱 메일 탐지

피싱 메일은 사이버 공격의 주요 시작점으로, 이를 정확하게 탐지하고 차단하는 것이 무엇보다 중요하다. 안랩은 AI를 활용해 피싱 메일을 정확하게 분류하고 차단하고 있다. AI는 메일 제목, 본문, 발신자, 첨부 파일, URL 등 다양한 요소를 분석해 피싱 메일인지 여부를 종합적으로 판단한다. 또한, 해당 메일이 왜 피싱으로 판별됐는지에 대한 상세한 설명을 제공함으로써 보안 담당자가 신속하게 공격을 이해하고 대응할 수 있도록 한다. 특히, AI는 사용자 피드백을 반영하여 피싱 메일 탐지 정확도를 지속적으로 향상시킬 수 있으며, 이를 통해 기업과 사용자가 수신하는 여러 유형의 피싱 메일을 효과적으로 차단할 수 있다.

4) 위협 인텔리전스 예측

AI는 수많은 경고 로그를 분석하여 이상 패턴을 탐지하고, 공격 징후를 조기에 발견한다. 보안 솔루션과 모니터링 장비에서 발생하는 수많은 경고 로그를 처리하는 것은 굉장히 어려운 작업이다. AI는 이러한 로그들을 자동으로 분석하고, 중요한 패턴을 추출하여 위협이 발생할 가능성을 예측한다. 이를 통해 AI는 오탐(False Positive)을 줄이고, 보안 담당자가 자칫 놓칠 수 있는 중요한 이벤트를 사전에 경고할 수 있다. AI는 단순히 경고를 수집하는 데 그치지 않고 여러 이벤트를 종합적으로 분석하여 위험성을 추론하며, 분석가에게 정리된 정보를 제공함으로써 공격을 조기에 방지할 수 있도록 지원한다.

5) 위협 인텔리전스 요약 설명

AI는 복잡한 로그와 행위 탐지 결과를 종합하고 요약 설명하여 보안 담당자의 이해를 돕고, 필요한 조치를 제안한다. AhnLab EDR이나 AhnLab XDR과 같은 보안 솔루션의 위협 탐지 정보는 양이 방대하고 복잡한 로그를 포함하고 있어 분석가가 이를 해석하는 데 어려움이 있을 수 있다. 안랩의 AI 시큐리티 어시스턴트인 AhnLab Annie AI는 탐지된 행위와 로그를 해석하고, 이를 바탕으로 위협의 발생 배경과 루트, 프로세스, 방어 체계 등을 요약하여 분석가에게 전달하며, 적절한 조치 방안을 제시한다.

6) 자동화된 사고 대응

AI는 사전에 작성된 플레이북(Playbook)과 과거 사고 데이터를 학습한 모델을 기반으로 최적의 복구 및 대응 방법을 제안하고, 사고 대응 보고서를 자동화해 사이버 복원력(Cyber Resilience)을 강화한다. 예를 들어, 침해사고 발생 시 AI는 악성코드와 네트워크를 격리하고 담당자에게 알림을 전송해 사유서를 작성하도록 한다. 사유서에 문제가 없으면 보고서 작성 후 종료하고, 필요 시 의심스러운 시스템을 전수 조사한다. 보안 담당자는 각 단계를 검토하고 승인만 하면 된다. 이 자동화된 사고 대응 시스템은 보안 담당자가 보다 중요한 의사결정에 집중할 수 있도록 돕는다.

결론

AI를 무기로 내세운 공격자와 방어자의 싸움이 이제 서막을 올렸다. 공격자가 AI를 악용하는 방식은 날로 진화하고 있다. 보안 담당자들은 이에 맞서 싸우기 위해 더 정밀하고 강력한 AI 기반 방어 시스템을 구축하고, 실시간으로 위협을 탐지하고 대응하는 능력을 강화하고 있다.

AI 기반 공격은 매우 빠르게 정교화되고 지능화되고 있어, 보안 시스템은 이에 발맞추어 지속적으로 업데이트돼야 한다. 안랩은 AI 기술을 제품 전반에 적용해 탐지 및 분석 속도를 획기적으로 개선하고, 위협 데이터를 더 정확하게 분석해 보안 담당자가 능동적이고 효율적으로 대응할 수 있도록 지원하고 있다. 이를 통해 사이버 복원력을 강화하고, 보다 안전하고 견고한 기업 환경을 구현하는 데 기여하고 있다.

우리는 AI를 단순히 공격을 차단하는 툴로 사용할 것이 아니라, 사이버 보안의 패러다임을 혁신하는 핵심 기술로 발전시켜야 할 것이다. 미래의 보안 환경에서 AI는 공격자와 방어자 간의 격차를 좁히고, 지능적이고 진화된 보안 시스템을 구축하는 데 매우 중요한 역할을 할 것이다.

AhnLab
전성학 연구소장