AI 기반 해킹 도구의 확산과 진화 - 생성형 AI가 바꾼 사이버 공격 생태계와 대응 전략

2023년 6월 등장한 WormGPT는 사이버 범죄 생태계에 패러다임의 변화를 가져왔다. 생성형 AI는 공격 진입 장벽을 낮췄고, AI 기반 해킹 도구는 유료 구독형 서비스와 무료 오픈소스 형태로 빠르게 확산되고 있다. 더 나아가 AI는 단순한 공격 도구 생성을 넘어 공격 작전 전반을 운용하는 단계로 진화하고 있으며, 악성코드 자체에 AI를 내장해 실시간으로 변형하는 새로운 위협도 현실화되고 있다.

이번 글에서는 AI 기반 해킹 도구의 유통 생태계, 실제 공격 인프라와 악성코드에 통합된 사례, 오픈소스 AI 확산이 초래하는 구조적 위협과 방어 전략을 종합적으로 분석한다.

AI 기반 해킹 도구의 확산과 유통

생성형 AI의 확산 이후 공격자는 피싱 문구 작성, 악성코드 생성, 정찰 자동화 등 공격 과정의 여러 단계에서 AI를 활용하기 시작했다. 초기에는 개별 도구나 실험적 시도에 가까웠지만, 최근에는 유료 구독형 서비스와 오픈소스 도구가 함께 확산되며 사이버 범죄 생태계의 주요 인프라로 자리 잡고 있다.

AI 기반 해킹 도구 등장 타임라인

2023년 WormGPT 등장을 기점으로 AI 기반 해킹 도구는 불과 몇 년 사이 수십 종으로 늘어났으며, 현재는 사이버 범죄 생태계의 주요 구성 요소로 자리 잡고 있다.

시기	도구명/악성코드	주요 특징
2022.11	ChatGPT	OpenAI가 ChatGPT를 출시하며 대형 언어 모델의 대중화 시작
2023.06	WormGPT	GPT-J 6B 파인튜닝 기반 최초의 상용 악성 LLM
	FraudGPT	사이버 범죄 포럼 최초 광고, 피싱, 사기 문서 자동화 특화
2023.08	WormGPT 종료	언론 보도 후 개발자가 자진 종료
	Evil-GPT	WormGPT 종료 다음 날 동일 포럼에 대체 도구로 등장
2023년 하반기	WolfGPT, EscapeGPT LoopGPT, DarkGPT	WormGPT 계열의 인지도에 편승한 모방 도구 다수 등장
2024년 상반기	GhostGPT, MalwareGPT	악성코드 개발에 특화된 도구로 등장
	SpamGPT, SpamirMailer Bot	스팸 및 피싱 캠페인의 대량 자동화를 지원하는 도구
2024년 하반기	EvilAI 텔레그램 봇	기능별로 $10~$60에 판매된 뒤 웹 서비스로 확장
2025.04	Xanthorox	사이버 공격 및 프라이버시 침해 활동을 지원하는 도구
2025.07	KawaiiGPT GitHub 공개 배포	무료 오픈소스로 GitHub에 공개되었으며, Termux 지원을 통해 스마트폰에서도 구동 가능
2025.09	HexStrike AI BruteForce AI	오픈소스 AI 기반 공격 프레임워크 오픈소스 AI 기반 크리덴셜(계정) 공격 도구 자동화 정찰, 브루트포스 AI 도구 (레드팀 도구가 악용됨)
2025.09	WormGPT (SaaS)	WormGPT 모방 도구가 SaaS 형태로 재등장
2025년 하반기	Promptflux, Promptsteal, Honestcue	최초의 AI 기능을 내장한 자기 변형 악성코드 사례로 관찰
2026.02	WormGPT 사용자 DB 유출	1만 9천여 명의 이메일과 결제 정보 유출, 구매자도 피해자 전락
2026.04	Bissa Scanner	Claude Code를 활용한 AI 기반 공격 작전에 활용되었으며, 6만 5천 건의 자격증명 탈취
2026.05	Promptspy(Android)	Gemini API를 활용한 자율 공격 오케스트레이션 Android 악성코드

[표1] AI 기반 해킹 도구/악성코드 등장 타임라인

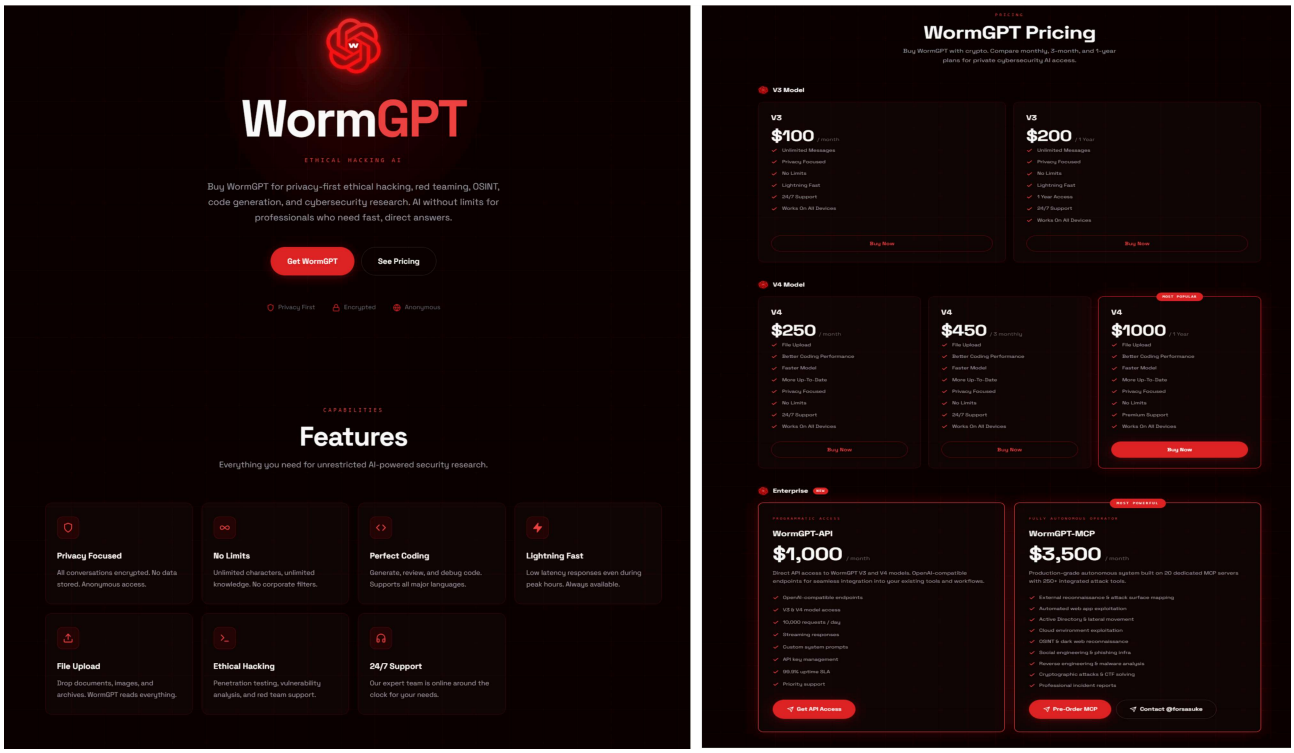
위 타임라인에서 특히 주목할 만한 점은 WormGPT의 종료 다음 날 동일한 포럼에서 Evil-GPT가 바로 등장한 점이다. 이는 AI 해킹 도구 생태계가 특정 도구의 종료 또는 수사 당국의 개입에도 쉽게 사라지지 않고, 강력한 회복 탄력성을 바탕으로 유사 도구를 통해 빠르게 대체될 수 있음을 보여준다.

유통 생태계의 복합적 구조

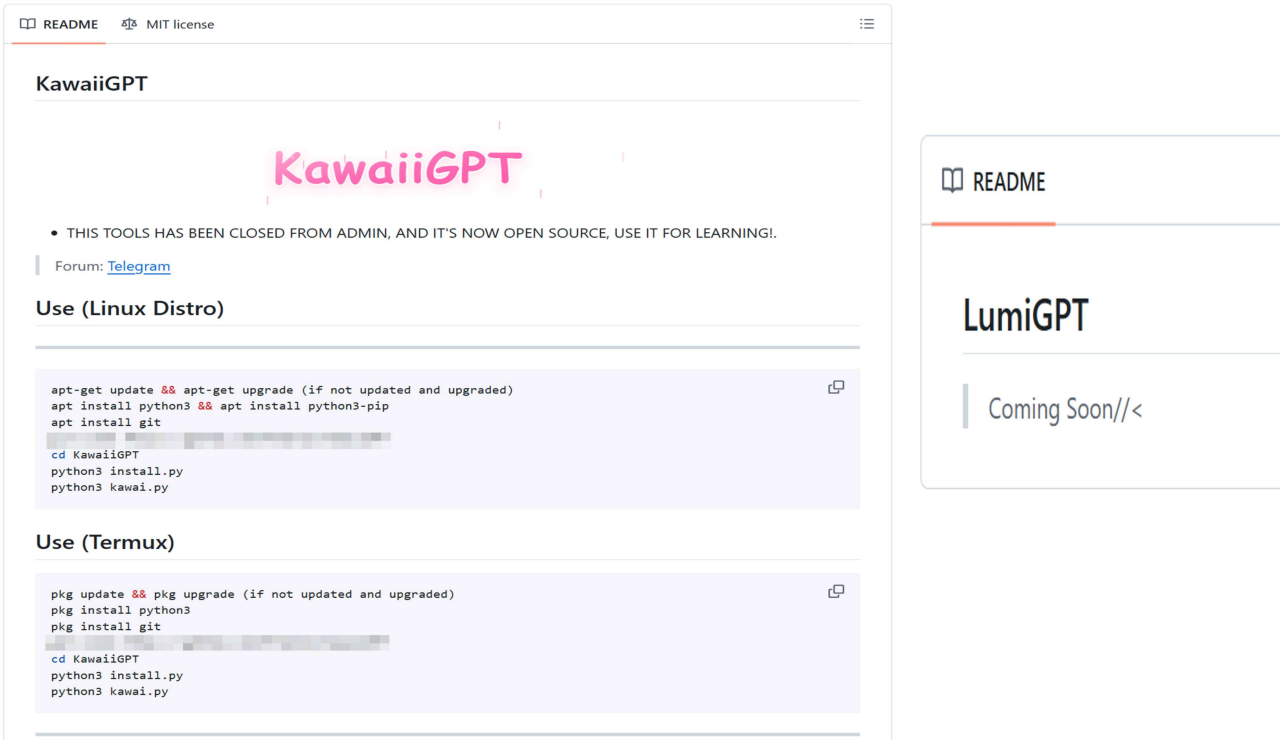
AI 해킹 도구는 크게 유료 구독형(SaaS)과 오픈소스 무료 배포 방식으로 유통되고 있다. 배포 방식은 다르지만, 상당수 도구는 독자 개발 모델이 아니라 합법적인 상용 AI API를 활용하거나 콘텐츠 필터링이 제거된 언센서드 오픈소스 모델을 기반으로 한다.

유형	대표 사례	주요 특징	보안 영향
유료 SaaS 구독형	WormGPT, FraudGPT, Evil-GPT	다크웹, 해킹 포럼, 텔레그램, 웹사이트를 통해 월간, 연간, 평생 이용권 형태로 판매되며, 무검열 AI 기능과 피싱, 악성코드 개발, 공격 자동화 기능을 광고	악성 AI 도구가 일반 SaaS와 유사한 가격 체계와 유통 구조를 갖추며 상업화되고 있음
오픈소스 무료 배포형	KawaiiGPT	GitHub에 무료 공개되어 복제와 변형이 쉬우며, Termux 지원으로 스마트폰에서도 실행 가능	수사 압박이나 운영 리스크에도 도구가 사라지지 않고 오픈소스 생태계를 통해 빠르게 확산될 수 있음
언센서드 모델 및 로컬 실행 환경 악용	WhiteRabbitNeo, Llama 2 Uncensored, Dolphin 시리즈, Wizard-Vicuna Uncensored	Hugging Face, Ollama 등을 통해 안전장치가 약화된 모델을 로컬에서 실행할 수 있어 추적과 통제가 어려움	다크웹이나 유료 서비스에 의존하지 않고도 악용 가능한 환경이 형성되며, AI 공격 도구의 진입 장벽이 낮아지고 있음

[표2] AI 해킹 도구 유통 유형별 사례 및 특징



[그림1] WormGPT 홍보와 구독 가격 광고



[그림2] KawaiiGPT 오픈 소스 공개 배포

언더그라운드 생태계의 구조적 변화와 영향

언더그라운드 AI 도구의 기능은 딥페이크 및 이미지 생성, 악성코드 개발, 피싱 자동화, 정찰 및 연구, 코드 생성, 취약점 익스플로잇 6개 영역으로 분류되며, 일반 SaaS 서비스와 동일한 비즈니스 구조를 채택하여 운영하고 있다. 무료 체험판, 구독 티어, 프리미엄 기능, 텔레그램 고객 지원, 7일 환불 보장 정책 등의 운영 형태가 자리를 잡아가고 있다. 이를 통해 초보 해커(Script Kiddie)에게 강력한 비대칭적 우위를 제공하며 전체 사이버 위협의 수위를 높이고 있다.

AI 도구가 해킹 공격의 진입 장벽을 크게 낮춘 것은 사실이지만, 이를 비전문가가 정교한 공격을 쉽게 완수할 수 있게 되었다는 의미로 해석해서는 안 된다. 보다 정확히는 비전문가가 피싱 문구 작성이나 초보적인 악성코드 스크립트 생성 등 공격의 특정 단계에서 더 이상 높은 수준의 전문 지식을 필요로 하지 않게 되었다는 뜻에 가깝다. 실제 피싱 캠페인 운영이나 악성코드 배포에는 여전히 C2 인프라 구축, 탐지 회피 등 AI만으로 해결하기 어려운 요소가 남아 있다.

그렇다고 이러한 한계가 AI 기반 위협의 수준이 낮다는 의미는 아니다. AI는 공격 전 과정을 단독으로 완성하는 도구가 아닌, 공격 준비와 실행 과정의 일부를 자동화하고 효율화하는 방식으로 위협 행위자의 역량을 보완하고 있다. 최근 사례에서는 이러한 변화가 피싱 문구 작성이나 코드 생성 같은 보조 역할을 넘어 실제 공격 흐름에 영향을 미치는 단계로 확장되고 있다.

AI 기반 공격 도구 실제 활용 사례

AI 기반 해킹 도구는 단순히 유통되는 상품이 아니라 실제 공격 작전의 일부로 쓰이기 시작했다. 최근 사례는 AI가 정찰, 취약점 악용, 자격증명 선별, 악성코드 변형, 인프라 운영까지 공격 흐름 전반에 개입하고 있음을 보여준다.

AI 공격 오케스트레이션 사례: Bissa Scanner

최근 AI가 해킹 공격의 단순 보조 수단을 넘어 공격 흐름 전반을 조율(Orchestration)하는 단계로 확장되고 있음을 보여주는 사례가 확인되었다. Bissa Scanner 사례에서 위협 행위자는 Claude Code와 OpenClaw를 결합해 공격 오케스트레이션 도구로 운용하고, Next.js 프레임워크의 취약점(CVE-2025-55182)을 악용한 대규모 스캐닝을 자동화했다. 이는 AI가 공격 코드 생성에 그치지 않고, 공격 자동화 파이프라인의 구축과 실행 과정에 직접 활용될 수 있음을 보여준다.

탈취된 데이터는 Anthropic, OpenAI, Google, AWS, Stripe, PayPal 등 AI 플랫폼, 클라우드, 결제, 데이터베이스 등의 자격증명 파일이었으며, 피해 기업 중에는 세무, 금융 자문사, 디지털 자산 결제사, 급여, HR 플랫폼 등이 포함되었다. AI가 탈취한 데이터 중 금융 및 암호화폐 관련 고부가가치 정보를 실시간으로 분류하는 트리아지(triage, 중요도 기반 선별) 역할을 수행하여 공격 효율을 극대화했다. 위협 행위자는 텔레그램 봇을 통해 실시간으로 침해 결과를 전달받으며 사실상 AI가 보조하는 사이버 범죄 운영 체계를 구축했다.

AI 내장형 악성코드의 진화

최근에는 AI를 악성코드 내부 기능에 활용해 코드 변형, 난독화, 실행 판단을 자동화하려는 사례도 확인되고 있다. 이러한 방식은 기존 시그니처 기반 탐지를 무력화시켜 정적 분석만으로 악성 여부를 판단하기 어렵게 만든다. 아래는 최근 관찰된 AI 내장형 악성코드의 주요 사례다.

- **Promptflux**: Gemini API를 호출하여 자체 소스코드를 주기적으로 재작성하며 정적 시그니처 기반 탐지를 우회하는 자기 변형 드로퍼다.
- **Honestcue**: Gemini API를 통해 VBScript 난독화 기법을 실시간으로 요청하여 탐지를 회피하는 just-in-time 자기 수정 악성코드다.
- **Canfail - Longstream**: 러시아 연계 행위자가 우크라이나 표적을 대상으로 사용한 악성코드로, LLM이 생성한 수만 줄의 디코이(decoy) 코드로 악성 행위를 은폐하는 특징을 보인다. 특히 시스템의 일광 절약 시간(DST) 상태를 조회하는 로직을 32회 반복 삽입함으로써, 코드 전반이 정상적인 시스템 동작처럼 보이도록 설계되어 분석 및 탐지를 어렵게 만든다.
- **Promptspy**: Android 백도어로 Gemini API를 통해 기기의 UI 구조를 자동 분석하고 물리적 제스처(클릭, 스와이프)를 시뮬레이션하여 사용자 인터페이스를 자율적으로 조작한다. 특히 피해자가 앱을 삭제하려 할 때 '삭제' 버튼 위에 투명한 오버레이를 씌워 터치 이벤트를 가로채는 방식으로 삭제를 방해하는 기능도 포함되어 있다.

국가 배후 행위자(APT)의 AI 활용과 제로데이 무기화

최근 사례는 국가 배후 행위자들 역시 AI를 취약점 분석, 익스플로잇 검증, 공격 인프라 개발에 활용하고 있음을 보여준다.

- 2026년 5월, AI를 활용해 개발했을 가능성이 높은 제로데이 취약점 익스플로잇의 최초 사례가 확인되었다. 해당 익스플로잇은 널리 사용되는 오픈소스 웹 기반 시스템 관리 도구의 2단계 인증(2FA)을 우회하는 Python 스크립트로, 교육적 docstring, hallucinated CVSS 점수, 정돈된 Pythonic 코드 구조 등 LLM 학습 데이터의 전형적인 특성이 포함되어 있었다.
- 중국 연계 위협 행위자들은 전문가 페르소나 기반 탈옥 기법과 ‘wooyun-legacy’ 프로젝트를 활용했다. 해당 프로젝트는 8만 5천 건 이상의 실제 취약점 사례를 기반으로 한 Claude Code 스킬 플러그인으로, 모델이 숙련된 전문가처럼 코드를 분석하고 로직 결함을 식별할 수 있도록 인-컨텍스트 학습을 지원한다.
- 북한 연계 APT45 조직은 수천 번의 반복 프롬프팅을 통해 CVE 분석과 공격 코드 검증을 자동화하고, 익스플로잇 검증 역량을 대규모로 확장한 것으로 관찰되었다.
- 중국 APT27은 Gemini를 활용해 ORB(Operational Relay Box) 네트워크 관리 애플리케이션 개발을 가속화한 것으로 확인되었다. ORB 네트워크는 공격의 실제 출처를 숨기기 위한 익명화 인프라이며, AI는 관련 관리 애플리케이션의 개발 속도를 높이는 데 활용된 것으로 보인다.

AI 기반 공격의 구조적 변화와 대응 전략

AI 해킹 도구 생태계는 더 이상 실험적 단계에 머물러 있지 않다. 유료 SaaS, 오픈소스 배포, 언센서드 모델의 로컬 실행 환경을 기반으로 확산되며 공격자의 비용과 시간을 낮추는 범죄 인프라로 진화하고 있다. 자사 모니터링 결과에서도 AI 기반 공격 도구에 대한 관심과 활용은 최근 수년간 지속적으로 증가하는 추세로 확인된다. 이 변화는 공격자와 방어자 사이의 구조적 비대칭성을 키우고 있으며, 고성능 AI가 통제 밖에서 운용되는 환경은 기존 방어 체계의 한계를 더욱 뚜렷하게 하고 있다. 이 때문에 대응 전략 역시 특정 악성 도구 차단을 넘어 AI 기반 능동 방어와 신원 보안, AI 거버넌스, 공급망 보안을 함께 강화하는 방향으로 전환되어야 한다.

AI 기반 공격의 구조적 비대칭성

AI 기반 공격은 공격자와 방어자 사이의 구조적 비대칭성을 확대한다. 공격자는 AI를 활용해 간단하고 빠르게 맞춤형 피싱 이메일 작성, 악성 스크립트 생성, 취약점 분석 결과를 생성하고 반복적으로 수행할 수 있다. 반면 방어자는 각 공격 시도에 대해 탐지, 검증, 차단, 대응, 정책 보완까지 수행해야 한다. AI 기술이 공격자에게는 확장성과 자동화를 제공하지만, 방어자에게는 대응해야 할 이벤트의 양과 복잡성을 증가시키는 구조다.

이러한 격차는 운영 속도에서도 뚜렷하게 나타난다. 공격자는 고성능 AI를 활용하여 취약점 탐색, 익스플로잇 초안 생성, 공격 절차 자동화에 걸리는 시간을 크게 줄일 수 있다. 반면 방어자는 패치 하나를 적용할 때도 영향 범위 분석, 테스트 환경 검증, 운영 일정 조율 등의 과정을 거쳐야 한다.

위협의 진화 방향도 주목할 필요가 있다. AI 기반 위협은 단순히 공격 도구를 생성하는 단계를 넘어 Bissa Scanner 사례와 같이 공격을 운용하고 Promptflux, Promptsteal, Promptspy처럼 악성코드 자체가 AI 기능을 활용해 실시간 판단과 자기 변형을 수행하는 형태로 빠르게 변화하고 있다. 이는 AI가 공격의 핵심 구성 요소가 되었음을 보여준다.

통제를 벗어난 고성능 AI의 악용 가능성

현재 보안 업계의 관심은 최신 프론티어 AI 보안 모델의 등장에 집중되어 있다. 프론티어 모델의 악용 가능성도 중요하지만 더 근본적인 위협은 이와 같은 고성능 AI 모델이 결국 공급자의 통제 범위를 벗어나 공격자에게 직접 활용될 수 있다는 점이다.

현재 오픈소스 AI 모델은 최신 고성능 모델의 성능 수준에 빠르게 근접하고 있다. 이와 같은 속도로 오픈소스 AI 모델이 전파된다면, 공격자는 1년 이내에 안전장치, 이용 약관, 모니터링 체계가 적용되지 않은 최신 성능의 AI와 유사한 수준의 모델을 자체 하드웨어에서 직접 운용할 수 있게 될 것으로 전망된다. 결국 공급자의 제약 없이 고성능 AI를 자체 인프라에서 운용하는 공격자의 등장이 현시점에서 가장 경계해야 할 위협의 본질이다.

핵심 대응 전략

AI 기반 위협은 단일 보안 솔루션이나 특정 도구 차단만으로 대응하기 어렵다. 공격자가 AI를 활용해 공격 준비와 실행 과정을 자동화하는 만큼, 방어자도 탐지와 대응, 인증, 거버넌스, 공급망 보안을 AI 위협 환경에 맞게 재정비해야 한다. 이에 따라 다음 네 가지 영역을 중심으로 대응 전략을 수립할 필요가 있다.

1. AI 에이전트 기반 능동 방어 체계 구축

AI 에이전트를 활용해 악성코드 변형 및 탐지 회피를 자동화하는 공격 방식이 증가하면서 시그니처 기반 탐지와 같이 알려진 패턴을 찾는 정적 방어 체계로는 한계에 직면하고 있다. 이에 대응하기 위해 방어자도 동등한 수준의 AI 에이전트를 활용해야 한다. AI 에이전트를 통해 공격이 가능한 경로를 선제적으로 파악하고 각 경로별 위험도를 평가하며, 최소한의 조치로 최대한의 공격 표면을 제거하는 등 보안 파이프라인을 자동화해야 한다. 더 나아가 AI 에이전트를 SOC(보안운영센터) 업무에 통합하여 초기 위협 평가와 컨텍스트 수집을 자동화하고, 인간 분석가가 복잡한 판단이 필요한 영역에 집중할 수 있는 구조로 전환하는 것이 핵심이다. 결국 AI로 AI를 막는 능동적 방어 패러다임으로 이동하지 않는다면 공격자의 속도와 비용 경쟁에서 구조적으로 뒤처질 수밖에 없다.

2. 강화된 MFA(다중 요소 인증) 도입

AI 공격에 대응하기 위해서는 AI가 우회하기 어려운 하드웨어 기반 인증 및 컨텍스트 인식 인증 체계로 고도화해야 한다. AI가 2FA의 결함을 식별하여 우회하는 사례가 실제로 관찰되고 Promptspy처럼 AI가 기기의 UI를 자율적으로 조작하고 생체인증마저 우회하는 단계까지 위협이 진화한 만큼, MFA 도입에 그치지 않고 인증 체계 전반을 AI 에이전트로 지속적으로 감사하고 이상 징후를 실시간으로 탐지하는 신원 중심 보안(Identity-Centric Security) 구조로 전환해야 한다.

3. 특정 도구 차단을 넘어선 AI 거버넌스 체계 수립

조직은 사내망에서 Ollama와 같은 로컬 AI 프레임워크나 비인가 언센서드 모델이 구동되는지 식별하고, 이를 정책적으로 통제할 수 있는 체계를 갖춰야 한다. 현재 안전하다고 평가되는 오픈소스 모델도 향후 공격에 활용될 수 있다는 점을 고려하면, 대응의 초점은 특정 도구 차단에 머물러서는 안 되며 AI 모델 사용 전반에 대한 가시성 확보와 거버넌스 체계 수립이 선행되어야 한다.

AI 거버넌스의 범위는 내부 사용 통제에만 그치지 않는다. 조직이 도입한 AI 시스템 자체가 이제 공격의 표적이 되고 있다. 프롬프트 인젝션(Prompt Injection)과 데이터 포이즈닝(Data Poisoning) 공격은 AI 모델의 판단을 오염시키거나 비인가 동작을 유발할 수 있으며, 이는 데이터 유출, 권한 탈취, 평판 손상으로 이어질 수 있다. 이 때문에 AI 시스템은 일반 소프트웨어가 아니라 고권한 인프라로 취급해야 하며, 컨테이너 격리, OAuth 기반 접근 통제, 입출력 검증 체계를 결합한 심층 방어 구조를 적용해야 한다.

4. 공급망 및 AI 인프라 보안 검토

AI 플랫폼의 API 키와 자격증명은 주요 탈취 표적으로 부상하고 있다. AI 서비스 접근 키는 단순한 계정 정보가 아니라 공격자가 추가 공격에 활용할 수 있는 고가치 인프라 자산으로 분류해 관리해야 한다. 또한 AI 생성 코드가 오픈소스 라이브러리와 소프트웨어 공급망에 광범위하게 유입되고 있는 만큼, 별도의 보안 검증 프로세스도 필요하다. 코드의 출처가 사람이든 AI든 동일한 수준의 검증을 적용하는 것을 원칙으로 삼되, AI 생성 코드에서 나타날 수 있는 존재하지 않거나 잘못 생성된 의존성 패키지, 과도하게 광범위한 권한 요청 등을 추가로 검토해야 한다.

마무리하며

AI 기반 해킹 도구는 더 이상 개별 공격자의 실험적 도구에 머물지 않는다. 유료 구독형 서비스, 오픈소스 배포, 언센서드 모델의 로컬 실행 환경을 통해 빠르게 확산되며 사이버 범죄 생태계의 주요 인프라로 자리 잡고 있다. 특히 일부 사례에서는 AI가 공격 준비를 보조하는 수준을 넘어 공격 흐름을 조율하고, 탈취 데이터를 선별하며, 악성코드 변형에 관여하는 단계로 진입하고 있다.

이 때문에 대응의 초점도 특정 악성 AI 도구를 식별하고 차단하는 데만 머물러서는 안 된다. 조직은 AI가 공격자에게 제공하는 자동화와 확장성을 전제로 방어 체계를 재정비해야 한다. AI 에이전트 기반 능동 방어, 강화된 인증 체계, AI 모델 사용 전반에 대한 거버넌스, API 키와 AI 생성 코드에 대한 공급망 보안은 앞으로의 AI 위협 환경에서 핵심 대응 축이 될 것이다. 결국 AI 기반 공격에 맞서는 방어 전략은 “AI 사용 여부”의 문제가 아니라, AI를 얼마나 통제 가능하고 안전한 방식으로 운용할 수 있는가의 문제로 전환되고 있다.

본 보고서 전문은 안랩의 위협 인텔리전스 플랫폼 'AhnLab TIP' 구독 서비스를 통해 확인할 수 있다.

AhnLab
A-FIRST팀 정진성 책임매니저