paloalto NETWORKS | CYBERSECURITY PARTNER OF CHOICE

# AI 보안: 사이버 방어의 새로운 지평 탐색

ChatGPT는 2022년 11월 출시 이후 기계가 인간 상호 작용의 복잡성을 일부 처리할 수 있다는 사실을 입증함으로써 AI의 사용 방식을 새롭게 정의했습니다. 이 솔루션은 생성형 AI의 혁신적인 사용 사례로서 수많은 산업 분야를 망라하여 전 세계를 강타했습니다.

이 모든 것이 시작된 것은 2017년 Google의 연구원이 발표한 혁신적 논문, Attention is All You Need입니다. ChatGPT가 출시 두 달 만에 사용자 수 1억 명을 돌파했다는 점을 고려할 때, 7년이 지난 지금 ChatGPT는 인터넷과 휴대전화에 이은 다음 변곡점이 될 가능성이 있습니다.

하지만 이러한 성장 속에서 보안에 대한 관심은 충분할까요? 인터넷의 등장과 함께 디지털 자산에 대한 위협이 증가했습니다. 이러한 위협은 스마트폰 시대에 더욱 증폭되었습니다. 하지만 AI는 클라우드나 모바일보다 더 빠른 속도로 성장하고 있으며, 보안의 복잡성 또한 빠르게 커져가고 있습니다.

## 정의

이 글에서 AI와 생성형 AI는 동일한 의미로 사용되었다는 점에 유의해 주세요. AI에 대한 모든 언급은 생성형 AI의 맥락에서 이루어졌습니다.

AI 보안은 다음과 같은 두 가지 구성 요소를 구현하는, 더 포괄적인 용어입니다.

- **보안을 위한 AI**: AI/ML 모델을 사용하여 제품 또는 보안 운영 방법론의 보안 효율성을 강화합니다.
- **AI의 보안**: 보안 침해를 유발할 가능성이 있는 AI 앱의 무단 사용을 억제하며 LLM 모델 자체를 보호하는 것입니다.

이제 정의를 명확히 이해했으니 또 하나의 중요한 측면으로 기존 보안과 AI 보안의 차이점을 살펴보겠습니다. GenAI 개발자와 사용자 모두 기존의 보안 제어로 AI 모델을 보호하기에 충분하다고 생각하는 경우가 많습니다. 하지만 그렇지 않습니다. 이러한 모델을 실행하는 인프라와 이를 사용하는 애플리케이션을 보호하기 위해서는 플랫폼 수준에서 기존의 제어와 가드레일이 여전히 필요합니다. 하지만 AI 보안은 주로 모델의 신뢰성을 강화하는 데 중점을 두며, 이를 '신뢰성 태세'라고 부를 수 있습니다. 반면 기존 보안은 신뢰 자체, 즉 신뢰 태세를 확립하는 데 초점을 맞추고 있습니다(제로 트러스트를 떠올려 보세요).

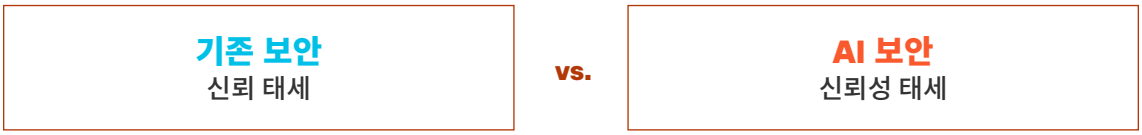


**그림 1:** 용어와 사용 영역의 차이점

"AI 보안"의 세상에서 피해 유형을 분류하여 살펴보겠습니다. AI의 보안의 측면에서 이는 두 가지 유형으로 나눌 수 있습니다. 하나는 "행동적 피해"이고 다른 하나는 "정보적 피해"입니다. 물론 두 가지를 모두 포괄하는 경우도 있습니다. 이를 구성하는 요소를 자세히 설명하기 위해 OWASP Top 10 LLM에서 정의한 공격의 명명법을 참조하여 대략적으로 연결해 보겠습니다.

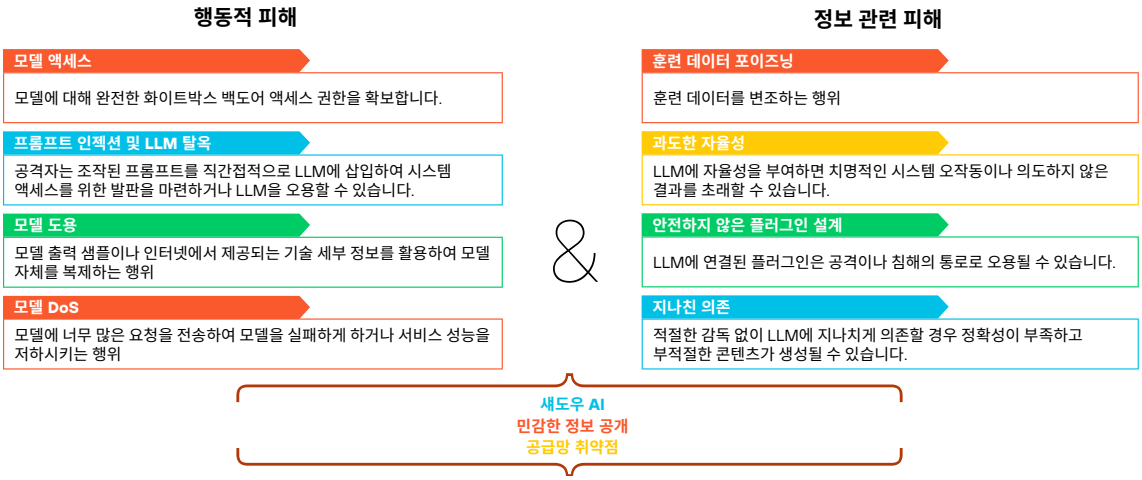


**그림 2:** 공격 유형 및 피해 분류

이 분류 체계의 일부는 쉽게 이해할 수 있는 명확한 마인드맵을 구성하기 위해 재조정되었습니다. 이를 살펴보면 의도에 따라 섀도우 AI, 민감한 정보 공개, 공급망 취약점은 정보적 피해와 행동적 피해를 모두 포괄할 수 있습니다.

## 새로운 공격 표면

정의와 명명법이 잘 정의되어 있으므로, "AI의 보안" 제품을 주로 다루는 'Secure AI by Design'에 대해 자세히 알아보겠습니다.

Secure AI by Design 제품을 제대로 이해하기 위해서는 우선 일반적 GenAI 기반 애플리케이션의 보안 문제를 인식해야 합니다. 이러한 문제는 크게 세 가지로 나눌 수 있습니다.

1. 섀도우 AI
2. AI 리스크
3. 데이터 노출

현재 GenAI 배포에 대한 불안과 불확실성, 의심 중 대부분은 데이터 노출과 관련된 것입니다. 하지만 자세히 살펴보면 섀도우 AI, AI 관련 리스크와 같은 다른 위협 요인도 존재합니다.

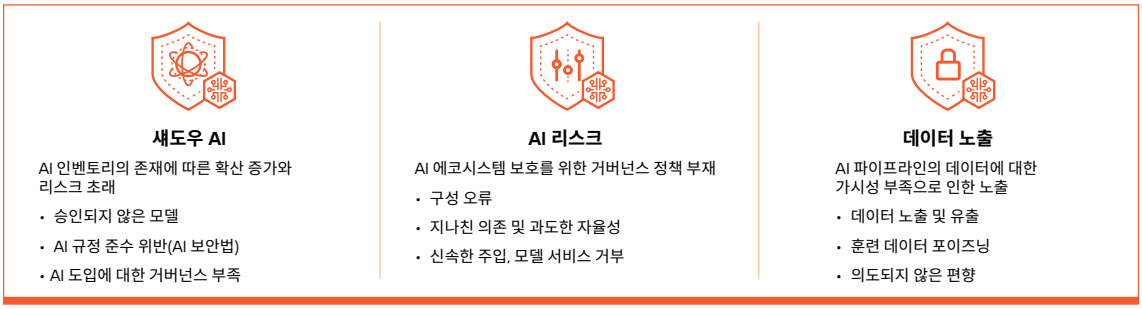


**그림 3:** 새로운 공격 표면

이러한 문제에 대해 자세히 살펴보겠습니다.

### 섀도우 AI

섀도우 AI는 AI 인벤토리의 부재에 의한 주요 문제와 관련이 있습니다. 이는 확산을 증가시키며 미승인 애플리케이션의 리스크를 유발함으로써 엔터프라이즈 환경에서 AI 도입에 대한 거버넌스가 부족한 상황으로 이어집니다.

몇 가지 사례를 통해 그 맥락을 설명해 보겠습니다. 대다수의 직원이 AI 기반 생산성 도구를 활용하며, IT 관리자 몰래 사용하는 경우도 많습니다. 이러한 도구는 생산성을 향상시키지만 리스크도 수반합니다. 예를 들어, 메모 작성 또는 미팅 어시스턴트 도구를 사용하는 직원은 실수로 회의 중 논의된 기밀 정보를 노출할 수 있습니다. 이는 이들 애플리케이션에 대한 보안의 필요성을 더욱 강조합니다.

### AI 리스크

AI 리스크는 AI 에코시스템을 보호하기 위한 거버넌스 정책이 부재한 상태를 의미하며, 일반적으로 OWASP Top 10 LLM의 전체 스펙트럼을 포괄합니다. 이 AI 스택의 각 요소는 공급망, 구성 및 런타임 위협과 관련된 리스크를 증가시킵니다. 개발자가 안전하지 않은 템플릿이나 손상된 라이브러리를 사용할 경우 공급망 리스크가 높아질 수 있습니다. 또한 모델 취약점과 잘못된 구성은 AI 모델 및 관련 도구에 부정적 영향을 미칠 수 있습니다.

### 데이터 노출

데이터 노출은 AI 개발 파이프라인의 가시성 부족과 관련된 문제로, OWASP Top 10 LLM에서는 데이터 노출, 유출, 학습 데이터 포이즈닝, 의도되지 않은 편향을 방지함에 있어 이러한 가시성의 중요성을 언급하고 있습니다. 가장 중요한 것은 AI 워크플로의 커뮤니케이션 채널을 확보하고, 권한을 신중하게 부여하고, AI 학습에 사용되는 중요한 독점 데이터를 보호하는 것입니다. 개발자가 취약한 모델을 사용하여 실수로 조직에 리스크를 초래하는 시나리오를 상정해 보겠습니다. 더욱이 클라우드 환경에서 애플리케이션이 잘못 구성될 경우 의도치 않게 인터넷을 통해 중요 데이터가 노출되어 더욱 광범위한 피해가 발생할 수 있습니다.

AI 사용 관점에서 중요 데이터에 리스크가 발생할 수 있는 다른 예시를 살펴보겠습니다. 개발자가 코드 최적화 과정에서 AI 애플리케이션을 사용하다가 실수로 특정한 독점 코드와 키, API 토큰을 업로드할 경우 대규모 데이터 유출이 발생할 수 있습니다. 이러한 상황을 방지하기 위해서는 중요 정보를 보호하고 AI 도구를 책임감 있게 도입할 수 있도록 사용량을 모니터링하는 보안 가드레일이 필요합니다.

또한, AI 보안 가드레일을 정의하기 전에 현대 기업에서 배포된 GenAI의 전반적 성숙도를 평가하고 사용 사례를 분류하는 것도 중요합니다. 이와 같은 분류 작업은 GenAI를 올바르게 사용하기 위해 어떤 유형의 제어가 필요한지 정의하는 데 도움이 됩니다. GenAI 보안 사용 사례의 전반적 성숙도를 평가할 때는 이를 크게 두 가지 카테고리로 분류할 수 있습니다.

첫 번째 카테고리는 손쉽게 액세스할 수 있는 GenAI 애플리케이션을 사용하는 것으로, GenAI 서비스를 구매하거나 활용하는 것이 포함됩니다. 두 번째 카테고리는 AI 기능을 활용한 애플리케이션 개발이나 구축에 중점을 둡니다.

GenAI의 사용은 주로 섀도우 AI 리스크의 영역에 해당하며, GenAI 애플리케이션의 개발은 식별된 AI 리스크로 분류할 수 있습니다. 그리고 데이터 노출은 두 카테고리에 모두 적용되는 중요한 문제입니다. Gartner는 생성형 AI 배포 접근 방식에서 이러한 기능의 성숙도를 상세하게 설명했습니다.

## 솔루션

AI 보안 제어에 기존 보안 솔루션을 사용할 경우 크게 세 가지 영역에서 문제가 있습니다.

### 1. AI 인벤토리에 대한 완벽한 가시성

가장 큰 문제는 AI 애플리케이션 환경에 대한 가시성으로, 다음과 같은 사항이 포함됩니다.

- 전체 AI 인벤토리를 파악하기가 어렵습니다.
- AI 애플리케이션, 모델, 플러그인, 데이터 세트 간 데이터 흐름과 관계를 탐지하는 데 있어 방해 요소가 존재합니다.
- 엔터프라이즈 IT 환경에서 미승인 타사 AI 애플리케이션을 탐지하기 어려우며, 따라서 이를 모니터링하거나 제어하기도 어렵습니다.
- 데이터 유출을 비롯한 보안 취약점의 리스크가 증가합니다.

### 2. AI 앱 세분화

AI 앱 배포 및 세분화가 어려워 수동 프로세스가 불가피하고, 리스크 제어가 취약하고, 트래픽 세분화가 제대로 이루어지지 않아 AI 앱이 오용될 여지가 있습니다.

## 3. AI 대상 공격

OWASP Top 10 for LLM에서는 전통적 위협과 AI 관련 위협에 대한 명확한 이해와 구분이 부족하다는 점을 강조한 바 있습니다. AI 사용, 애플리케이션, 모델, 데이터의 모든 측면을 보호하기 위해서는 포괄적 솔루션과 효과적 보안 전략이 필요합니다. 예를 들어, AI가 유해하거나, 민감하거나, 부정확한 결과물을 제공할 경우 그 원인은 AI 도입의 위 범주 중 하나일 수 있습니다. 이러한 격차를 해소하기 위해서는 앞서 설명한 바와 같이 AI 시스템의 사용과 개발을 모두 보호해야 합니다.

# AI 보안에 대한 Palo Alto Networks의 접근 방식

Palo Alto Networks는 주요 엔터프라이즈 AI 보안 사용 사례를 보호하는 새로운 기능을 제공합니다.


**그림 4:** AI를 위한 보안 및 AI의 보안 전반에 걸친 Precision AI 제품

## 포트폴리오 전반에 걸친 Precision AI 임베딩 - AI에는 AI로 대응

기업 네트워크와 데이터를 공격하기 위해 악의적인 방식으로 AI를 사용하는 적대적 AI를 차단하는 유일한 방법은 ML 기반의 실시간 보호 기능입니다. Palo Alto Networks는 AI 생성 공격을 식별하고 차단하기 위해 AI 기반 방어를 혁신해 왔으며, 특히 다형성 위협의 가속화를 탐지하고 방지하기 위해 많은 노력을 기울였습니다. Strata, Prisma Cloud, Cortex 등 Palo Alto Networks 모든 플랫폼은 행동 분석과 ML 모델을 통해 실시간에 가까운 탐지 및 위협 방지 기능을 제공합니다. 당사는 업계 최초로 DNS 트래픽에 대한 인라인 AI 기반 분석을 활용하여 DNS 하이재킹 및 기타 탈취 공격을 실시간으로 방지하는 고급 DNS 보안을 도입하며, 이는 당사의 포트폴리오를 더욱 강화합니다.

## Secure AI by Design

### AI Access Security

AI Access Security™는 포괄적인 클라우드 제공 보안 솔루션으로, 데이터 및 보안 리스크를 축소하여 직원이 타사 GenAI 애플리케이션을 사용할 수 있도록 지원합니다. 이 솔루션은 AI 애플리케이션의 범위를 파악하고, 승인된 AI 앱에 대한 직원의 액세스 및 사용을 안전하게 관리하며, 엔터프라이즈 업무 공간 환경에서 섀도우 AI 앱에 대한 가시성을 확보할 수 있도록 도와드립니다.

주요 기능:

- **포괄적 GenAI 앱 카탈로그**: 600개가 넘는 GenAI 애플리케이션에 대한 가시성 확보 및 GenAI의 사용이 급증하는 환경에 발맞춰 향상된 기능 제공
- **AI 대상 리스크 속성**: 데이터로 학습하는 AI 앱에 대한 가시성 확보
- **상황별 데이터 컨트롤**: LLM 기반 및 컨텍스트 인식 ML 모델을 활용하여 300개 이상의 카테고리에 걸친 데이터 분류
- **승인된 애플리케이션, 허용된 애플리케이션, 승인되지 않은 애플리케이션으로 애플리케이션 분류**
- **엔드포인트/브라우저 제어**: Prisma® Access Browser 통합을 통해 디바이스에서 직접 제어 구현

### AI Runtime Security

AI Runtime Security™는 AI 애플리케이션, 모델, 데이터를 식별하여 클라우드 기반의 애플리케이션이 마주하는 새로운 위협으로부터 보호합니다. 모든 클라우드 환경에서 애플리케이션, 모델, 데이터 세트, 사용자, 플러그인, 인터넷 연결을 비롯한 전체 AI 애플리케이션 에코시스템을 식별하고 새로운 AI 위협으로부터 이러한 애플리케이션을 방어합니다. 수많은 AI 애플리케이션이 새로 개발되는 환경에서 AI Runtime Security는 진화하는 AI 에코시스템을 지속적으로 모니터링하고 애플리케이션을 보호합니다. 이러한 지속적 식별을 통해 새로운 위협을 신속하게 처리하고 데이터 침해를 방지하며 잠재적인 공격으로부터 인프라를 보호할 수 있습니다.

주요 기능:

- **AI 에코시스템 식별**: AI 애플리케이션과 모델 및 AI 에코시스템의 다른 부분 간의 상호 작용 방식에 대한 가시성을 확보하여 즉각적으로 인식하기 어려운 숨겨진 상호 연결성을 파악할 수 있습니다.
- 포**괄적 AI 에코시스템 보안**: 애플리케이션, 모델, 추론 데이터 세트 간의 트래픽뿐 아니라 인바운드, 이스트-웨스트(East-West), 아웃바운드 트래픽을 포괄하는 전반적 앱 간 상호 작용을 원활하게 보호합니다.
- **AI 앱 및 모델 보호**: 최첨단 Precision AI™ 기반의 CDSS(클라우드 기반 보안 서비스)를 활용하여 알려지거나 알려지지 않은 멀웨어, AI 기반 위협, 모델에 대한 위협, 또는 악성 URL과의 상호 작용에 대해 시스템을 철저하게 보호하고 프롬프트 인젝션 공격에 대한 완벽한 방어를 보장합니다.

### AI-SPM

Prisma Cloud AI-SPM은 조직이 AI 기반 애플리케이션을 검색하고, 분류하고, 관리할 수 있도록 도와드립니다. 또한 AI-SPM은 모델, 애플리케이션, 리소스를 포함한 전체 AI 에코시스템에 대한 가시성을 제공하여 데이터 노출 및 규정 위반의 리스크를 줄여줍니다. 모델의 취약점을 식별하고 잘못된 구성의 우선순위를 지정함으로써 AI 보안 프레임워크의 무결성을 개선합니다.

주요 기능:

- **AI 애플리케이션 에코시스템에 대한 가시성**: 모든 AI 모델과 에이전트, 관련 리소스를 자동으로 검색하여 AI 기반 애플리케이션과 관련된 중요 데이터에 대한 가시성을 확보합니다.
- **AI 모델 리스크 분석**: 잘못된 모델 구성과 공급망 취약점을 식별하여 모델 및 애플리케이션 리스크를 축소합니다.
- **모델 리소스 전반의 데이터 보안**: 모델 데이터를 조작할 경우 취약점과 편향을 초래하고, 데이터가 노출되며, 데이터 개인정보 보호 위반, 규정 준수 및 보안 리스크로 이어질 수 있습니다.
- **AI 계통**: 애플리케이션에서 사용되는 AI 구성 요소 및 데이터 소스 계통을 식별하고 추적합니다.

### Unit 42 AI 보안 평가

Unit 42®의 AI 보안 평가를 활용하여 직원의 안전한 생성형 AI 사용과 AI 지원 애플리케이션 개발 강화에 대한 전문가 지침을 살펴보고 AI 관련 위협 요소에 한발 앞서 대응하세요.

## AI를 통한 보안 간소화

Palo Alto Networks는 네트워크, 클라우드, 보안 운영 전반에서 최신 위협에 대한 고객의 대응 방식을 혁신하기 위해 강력한 코파일럿 기능을 도입하고 있습니다. 새로운 코파일럿은 세계 최고 수준의 규모와 다양성을 갖춘 사이버 보안 데이터 세트를 활용하며, GenAI를 활용하여 업계에서 가장 정확한 보안 결과와 풍부한 인사이트를 제공하는 Precision AI를 기반으로 보안을 간소화합니다. 이를 통해 고객은 주요 관리 작업을 수행하고 질문에 대한 답변을 얻을 수 있으며, 이를 통해 조치를 취하는 데 소요되는 시간을 대폭 줄여 평균 대응 시간(MTTR)을 단축할 수 있습니다.

# 결론

AI 도입은 모든 디지털 기업 고객이 거쳐갈 혁신의 여정이며, 이를 보호하는 것은 전술적, 전략적인 노력입니다. AI 보안을 표준 보안 제어로 분류하거나 "일률적" 접근 방식을 채택할 경우 상당한 리스크가 발생할 수 있음을 인지하는 것이 중요합니다.

Palo Alto Networks는 고객의 여정을 안내하기 위해 필요한 준비를 마쳤습니다. 명확하고 통찰력 있는 솔루션을 통해 고객이 복잡한 과정을 한발 앞서 탐색하고 필요한 사항을 해결할 수 있도록 지원합니다.

Palo Alto Networks는 업계 최초로 AI 보안 문제를 구조화하고 명확한 규범적 솔루션을 제시함으로써 고객이 안심하고 GenAI를 활용할 수 있도록 도와드립니다.

안전한 AI 도입에 대해 자세히 알아보고 싶으신가요? 여기를 클릭하세요.

**서울시 강남구 테헤란로 518, 10층**
**(위워크 삼성역 2호점, 섬유센터빌딩)**
**영업 문의**
**Tel: 82-2-568-4353 /**
**eMail: Sales-KR@paloaltonetworks.com**

**www.paloaltonetworks.com**



parent_wp_ai-security:-navigating-the-new-frontier-of-cyber-defense_100224