Google

Google 보안 AI (Secure AI) 프레임워크

AI가 급격히 진화하고 있는 오늘날, 이에 따른 위험 관리 전략도 함께 발전하는 것이 중요합니다. 이러한 발전을 위해, AI 시스템 보안의 개념적 프레임워크인 보안 AI 프레임워크(Secure AI Framework)를 소개합니다. 보안 AI 프레임워크는 다음의 6가지 핵심 요소를 포함합니다.

1. AI 생태계로의 강력한 보안 기반 확장

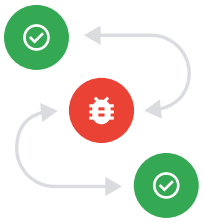
보안 AI 프레임워크는 지난 20년간 구축된 기본 보안 인프라 보호 기능과 전문성을 활용하여 AI 시스템, 애플리케이션, 그리고 이용자를 보호합니다. 이와 동시에 AI 발전에 발 맞추어 조직의 전문성을 개발하고 AI와 위협 모델의 진화 상황에 따라 인프라 보호를 확장 및 조정합니다. 예를 들어 SQL 인젝션과 같은 인젝션 기법은 상당 기간 존재해 왔으며, 조직은 입력 삭제 및 제한 등의 완화 조치를 통해 인젝션 스타일 공격에 즉각적으로 맞서 더욱 강력하게 대응할 수 있습니다.

2. 조직의 위협 유니버스에 AI를 도입하여 위협 감지 및 대응력 확장

위협 인텔리전스 및 기타 기능을 확장하여 AI 관련 사이버 인시던트를 적시에 탐지하고 대응합니다. 조직에서는 이상 현상을 감지하기 위해 생성형 AI 시스템의 입출력을 모니터링하고 위협 인텔리전스를 사용하여 공격을 예측합니다. 일반적으로 이러한 노력에는 신뢰와 안전, 위협 인텔리전스, 그리고 악용 대응팀과의 협업이 필요합니다.

3. 방어 자동화를 통해 기존 및 새로운 위협에 대응

최신 AI 혁신을 활용하여 보안 인시던트에 대한 대응 방법의 규모와 속도를 개선합니다. 공격자들이 AI를 사용하여 영향력을 확대할 가능성이 높기에, AI와 현재 사용 가능한 기능, 그리고 새롭게 개발되는 기능을 활용하여 민첩하고 비용 효율적으로 공격을 방어하는 것이 중요합니다.



4. 플랫폼 수준 제어의 조화를 통해 조직 전체의 일관된 보안 보장

컨트롤 프레임워크를 조정하여 AI 위험 완화를 지원합니다. 또한 다양한 플랫폼 및 도구 전체에 대한 보호 기능을 확장하여, 모든 AI 애플리케이션에 적용 가능하면서도 비용 효율적인 방식으로 최고의 보안 기능을 제공합니다. Google은 Vertex AI 및 Security AI Workbench 같은 AI 플랫폼으로 기본적 보안 보호 기능을 확장하고, 소프트웨어 개발 주기에 제어 및 보호 기능을 구축합니다. Perspective API와 같이 일반적인 이용 사례를 다루는 기능으로 조직 전체가 최첨단 보호 기능의 혜택을 누릴 수 있습니다.

5. AI 배포를 위한 더 빠른 피드백 루프 생성과 완화 조절을 위한 제어 조정

지속적인 학습을 통해 구현 기능을 계속 테스트하며 변화하는 위협 환경에 맞서 위협 감지 및 보호 기능을 개선해야 합니다. 개선 방안으로는 인시던트와 이용자 피드백을 기반으로 한 강화 학습 기법이 있으며, 훈련 데이터 세트 업데이트, 공격에 전략적으로 대응하기 위한 모델 미세조정, 모델 구축에 사용되는 소프트웨어를 대상으로 추가 보안 기능(예: 이상 동작 감지) 강화 등의 단계가 포함됩니다. 또한 조직에서는 레드팀 훈련을 정기적으로 실시하여 AI 기반 제품 및 기능의 안전 보장성을 높여야 합니다.

6. 비즈니스 프로세스 내 AI 시스템 관련 위험 상황 파악

조직이 AI를 배포하는 방식을 파악하는 엔드 투 엔드 위험 평가를 수행합니다. 해당 평가에는 특정 유형의 애플리케이션에 대한 데이터 계보, 검증 및 작업 동작 모니터링 등의 평가가 포함됩니다. 또한 조직은 AI 성능을 검증하기 위한 자동화 점검 절차를 구축해야 합니다.

Safer with Google