paloalto NETWORKS | CORTEX® CLOUD

AI 기반 애플리케이션을 위한 거버넌스 프레임워크 수립

인공지능(AI)은 빠르게 발전하고 있으며, 다양한 산업 분야의 수많은 조직은 그 어느 때보다도 빠른 속도로 AI 기반 애플리케이션을 배포하고 있습니다. Palo Alto Networks의 *2024년 클라우드 네이티브 보안 현황 보고서*[1]를 예로 들어 살펴보면, 설문 응답자 중 AI 지원 애플리케이션 개발을 도입하고 있다고 답한 비율은 100%였습니다. 불과 몇 년 전만 해도 공상 과학 소설과 같은 것으로 여겨졌던 기술임을 고려하면 이는 놀라운 결과입니다. 하지만 AI 시스템은 막대한 이점을 제공하는 반면 기존의 사이버 보안 접근 방식으로는 처리할 수 없는 새로운 보안 리스크와 거버넌스 과제를 유발하기도 합니다.

보안 리더는 반드시 AI의 현재 상태와 잠재적 영향, 조직에 미칠 수 있는 고유한 리스크를 파악해야 합니다. 이 백서에서는 AI 환경을 간단하게 소개하고, 핵심적인 보안 고려 사항을 강조하고, 이처럼 복잡한 지형을 탐색하는 여정을 지원하는 거버넌스 프레임워크를 제안합니다. 이 백서는 조직에서 검토할 만한 주요 AI 기술과 이러한 기술이 보안에 미치는 영향, 고려할 필요가 있는 잠재적 보호 장치 및 완화 조치를 명확하게 제시합니다. 또한 AI 보안 전략을 수립하는 과정에서 고려해야 하는 요소를 이해할 수 있도록 도와드립니다.

1 https://www.paloaltonetworks.com/state-of-cloud-native-security

현대의 AI 환경 및 보안에 미치는 영향

최근 몇 년간 AI 환경은 막대한 변화를 겪었으며, 많은 조직이 수많은 신기술을 평가하고 구현하면서 일반 시스템과 특수 시스템 분야에서 모두 대규모의 발전이 이루어졌습니다.

AI와 관련된 리스크와 기회를 관리하기 위해 보안 리더는 진화하는 이 에코시스템의 흐름을 파악해야 합니다. AI 생성 코드와 관련된 보안 리스크를 우려하고 있다고 답한 비율이 47%라는 최근 연구 결과에서 볼 수 있듯이, 리스크는 무수히 많은 방식으로 나타날 수 있습니다.

이 백서에서는 AI 기반 애플리케이션을 배포하는 과정에서 발생 가능한 클라우드 인프라 리스크를 중점적으로 알아볼 것입니다.

먼저 오늘날 AI 환경의 전반적인 모습을 살펴보겠습니다.

"기존" AI 및 머신 러닝

여러 해 동안 조직은 제조, 공급망 최적화, 사기 탐지, 온라인 마케팅 및 기타 여러 영역에서 효율성과 자동화를 위해 도메인별로 AI 및 머신 러닝(ML) 추론 도구를 활용해 왔습니다. 이처럼 좁은 범위의 시스템은 특정 데이터 세트에 대한 학습을 바탕으로 잘 정의된 작업을 수행합니다. 이는 많은 비즈니스 프로세스에서 필수적인 요소로 자리잡았지만, 많은 경우 개별 애플리케이션이 사일로화된 상태이며 범위가 제한적입니다.

이들 도구는 장기간 사용되었으며 최근 몇 년간 이와 관련하여 획기적인 발전은 없었지만, "모든 배를 밀어 올리는 밀물"과 같은 환경적 영향을 받고 있습니다. AI에 대한 전반적 관심이 확대되면서 이와 같은 최신 대규모 언어 모델(LLM) 기반 도구 배포를 위한 예산과 니즈가 빠른 속도로 확대되고 있습니다.

보안에 미치는 영향

- AI/ML 모델의 학습 및 운영에 사용되는 데이터의 무결성과 보안을 보장합니다.
- AI/ML 시스템을 모니터링하여 데이터 포이즈닝 공격, 모델 회피, 시스템 손상과 같은 보안 문제를 나타낼 수 있는 이상 동작이나 예기치 않은 출력, 성능 저하가 발생하고 있는지 확인합니다.
- AI/ML 모델이 의도하지 않은 편견, 공정성 문제, 또는 법률 및 평판 측면에서 리스크를 유발할 수 있는 차별적 결과를 초래하지 않는지 검증하여 모델의 공정성을 정기적으로 감사해야 합니다.
- AI/ML 시스템을 개발하거나, 수정하거나, 사용할 수 있는 사람에 대한 적절한 액세스 제어 및 거버넌스를 제공합니다.

AI 및 머신 러닝

사용 사례

제조 최적화, 공급망, 사기 탐지, 마케팅

보안 리스크

데이터 무결성, 모델 성능 모니터링, 편향성 감사, 액세스 제어

대규모 언어 모델(LLM)

LLM은 개방형 대화에 참여하고, 일관성 있는 텍스트를 생성하고, 코드를 작성하는 등 매우 활용도 높은 범용 AI 시스템으로 부상했습니다. ChatGPT를 비롯한 소비자 대상 챗봇이 널리 주목을 받고 있지만, 이는 빙산의 일각에 불과합니다. LLM의 진정한 힘은 다양한 기업 애플리케이션(내부 및 고객 대상)을 지원할 수 있다는 점에 있으며, 많은 경우 이는 최종 사용자의 눈에 잘 띄지 않는 방식으로 이루어집니다.

이처럼 광범위한 맥락에서 구현 패턴에는 여러 가지 카테고리가 있으며, 각기 다른 도구 세트와 관련한 보안의 영향에 의존합니다.

LLM

	사용 사례	보안 리스크
사전 학습된 기반 모델	콘텐츠 생성, 챗봇, 감정 분석, 번역, 코드 어시스턴트	섀도 AI 프로젝트, 데이터 개인정보 보호, 모델 거버넌스, 출력 제어
파인튜닝 및 RAG	전문 AI 어시스턴트(지원, 인사, IT), Q&A 앱(문서, 코드, 학습)	파인튜닝 중 민감한 데이터 노출, 데이터 거버넌스
맞춤형 모델 학습	고급 응용 환경(신약 개발, 재료 과학, 자율 시스템)	데이터 포이즈닝 공격, 컴퓨팅 리소스 격리, 모델 책임성 및 감사 가능성

사전 학습된 LLM 사용(독점 또는 오픈 소스)

OpenAI 및 Anthropic과 같은 클라우드 제공업체는 해당 업체가 관리하고 보호하는 강력한 LLM에 대한 API 액세스를 제공합니다. 조직은 이러한 API를 활용하여 애플리케이션에 LLM 기능을 통합할 수 있으며, 이 경우 기반이 되는 인프라를 관리할 필요가 없습니다.

또는 조직의 자체 인프라에서 Meta의 LLaMa와 같은 오픈 소스 LLM을 실행할 수도 있습니다. 이 경우 더 많은 제어와 사용자 지정 옵션을 확보할 수 있지만, 이를 안전하게 구현하고 유지 관리하기 위해서는 상당한 컴퓨팅 리소스와 AI 전문 지식이 필요합니다.

LLM은 다양한 배포 모델을 통해 사용할 수 있습니다.

- **API 기반 SaaS:** LLM 개발자(예: OpenAI)가 인프라를 제공하고 관리하며, 퍼블릭 API를 통해 프로비저닝합니다.
- **CSP 관리:** 클라우드 하이퍼스케일러가 제공하는 인프라에 LLM을 배포하며, Azure, OpenAI, Amazon Bedrock과 같은 프라이빗 또는 퍼블릭 클라우드에서 실행할 수 있습니다.
- **자체 관리:** LLM은 회사의 자체 인프라에 배포되며, 오픈 소스 또는 자체 개발 모델에만 해당합니다.

일반적 사용 사례

LLM 사용 사례에는 콘텐츠 생성, 챗봇, 감정 분석, 언어 번역, 코드 어시스턴트 등이 포함됩니다. 전자 상거래 회사의 경우 LLM을 사용하여 제품 설명을 생성할 수 있고, 소프트웨어 개발 회사의 경우 LLM 기반 코딩 어시스턴트를 활용하여 프로그래머의 생산성을 높일 수 있습니다.

보안에 미치는 영향

손쉽게 액세스할 수 있는 클라우드 API와 오픈 소스 모델이 등장하면서 애플리케이션에 고급 AI 언어 기능을 추가하는 작업의 난이도가 대폭 낮아졌습니다. 이제 개발자는 AI 및 ML에 대한 심층적 전문 지식이 없어도 소프트웨어에 LLM을 연결할 수 있습니다. 이는 혁신을 가속화하지만, 동시에 적절한 보안 및 규정 준수가 부족한 섀도 AI 프로젝트의 리스크가 증가합니다. 한편, 개발팀은 데이터 개인정보 보호, 모델 거버넌스, 출력 제어 문제를 충분히 고려하지 않고 LLM을 실험할 수 있습니다.

클라우드 API 및 오픈 소스 모델이 등장하면서 섀도 AI 프로젝트의 리스크가 증가했습니다.

파인튜닝 및 RAG(Retrieval-Augmented Generation)

구체적인 응용 환경에 따라 LLM을 맞춤 구성하기 위해 조직은 원하는 작업과 관련된 소규모 데이터 세트에 대한 파인튜닝 작업을 진행하거나 질문 답변 및 콘텐츠 요약을 위해 LLM을 지식 베이스와 통합하는 RAG를 구현할 수 있습니다.

일반적 사용 사례

내부 데이터(예: 고객 지원, 인사 또는 IT 헬프 데스크용)와 Q&A 앱(예: 문서, 코드 저장소, 교육 자료용)에 액세스할 수 있는 전문 AI 어시스턴트가 여기에 포함됩니다. 예를 들어, 통신사의 고객 서비스 챗봇은 제품 설명서나 FAQ, 과거 지원 상호 작용을 파인튜닝하여 기술 문제 및 계정 관리와 관련하여 고객 지원 서비스를 개선할 수 있습니다.

보안에 미치는 영향

조직은 파인튜닝과 RAG를 통해 구체적인 도메인과 데이터에 맞춰 LLM을 조정함으로써 보다 적합하며 정확한 결과물을 확보할 수 있습니다. 그러나 이러한 맞춤형 구성 프로세스 중에는 모델을 학습시키기 위해 중요 내부 정보를 노출하는 경우가 많습니다. 승인된 데이터만 파인튜닝에 사용하고 결과 모델의 보안을 유지하기 위해서는 철저한 데이터 거버넌스 관행이 필요합니다.

파인튜닝 및 RAG를 통한 모델 학습 과정에서 모델이 중요 내부 정보에 노출될 수 있습니다.

모델 학습

일부 기술 대기업과 연구 기관은 처음부터 LLM의 자체 교육에 투자하고 있습니다. 이는 매우 리소스 집약적인 프로세스로서 대규모 컴퓨팅 성능과 데이터 세트를 필요로 합니다. 하지만 모델 아키텍처, 학습 데이터, 최적화 프로세스를 온전하게 제어할 수 있을 뿐만 아니라 결과 모델에 대해 완전한 지적 재산권을 보유할 수 있습니다.

일반적 사용 사례

독점적 LLM의 사용 사례에는 신약 개발, 재료 과학 또는 자율 시스템을 비롯하여 고도로 전문화된 활용이 포함됩니다. 예를 들어, 의료 기관에서는 의료 기록과 영상 데이터를 기반으로 질병 진단에 도움을 주는 모델을 개발할 수 있습니다.

보안에 미치는 영향

맞춤형 LLM을 훈련하려면 방대한 데이터 세트를 신중하게 선정하고 고성능 컴퓨팅 인프라를 구축해야 하며, 그 과정에서 새로운 보안 문제가 발생할 수 있습니다. 학습 데이터는 모델이 수집할 수 있는 중요 정보 및 개인 데이터에 대해 철저한 검증을 거쳐야 합니다. 또한 공격자가 의도적으로 학습 데이터에 악의적인 예시를 삽입하여 모델의 동작을 조작하는 데이터 포이즈닝 공격의 가능성도 있습니다.

학습 프로세스는 막대한 컴퓨팅 리소스를 소모하므로 남용이나 간섭을 방지하기 위해서는 학습 환경에 대한 철저한 격리 및 액세스 제어가 필요합니다. 복잡한 블랙박스 모델을 다룰 경우 모델 행동의 책임 소재와 감사 가능성을 유지하는 방안을 둘러싼 문제도 있습니다. 그리고 이러한 문제는 모델을 자체 개발한 기업에게 더욱 절실하게 다가갈 것입니다.

맞춤형 LLM의 학습을 위해서는 중요 정보 및 개인 데이터에 대한 철저한 검증이 필요합니다.

보안 리더의 AI 리스크 개념화 방법

사이버 보안팀은 쫓아가는 데 익숙합니다. 클라우드 컴퓨팅, 컨테이너화, 서버리스 아키텍처 등 신기술의 도입에 발맞춰 전략과 제어를 조정해야 하는 경우가 많습니다. 그러나 AI의 부상은 사이버 보안에 있어 새로운 문제를 제시하고 있으며, 그 중 많은 수는 과거와는 질적으로 다른 문제입니다.

더 빠르고 더 극단적인 변화

AI는 수많은 조직에 지각 변동을 일으키고 있으며, 그 변화는 기존의 그 어떤 변화보다도 가파른 속도로 진행되고 있습니다.

- 빠른 속도로 새로운 모델, 기술, 애플리케이션이 등장하고 있으며, 월 단위나 주 단위로 중요한 혁신이 이루어지고 있습니다.
- 경쟁력을 유지하고 혁신을 추진하기 위해 이러한 기술을 빠르게 채택하고 통합해야 한다는 점은 조직에 엄청난 압박으로 작용합니다.
- 간단한 API를 통한 도구의 가용성이 확대되고 도구 및 프레임워크를 지원하는 에코시스템이 출현하면서 기술 부족에 의한 장벽이 낮아졌으며, 이는 도입을 더욱 가속화했습니다.

이처럼 빠른 기술 변화와 긴급한 비즈니스 수요의 조합은 고유한 보안 문제를 야기할 수 있습니다. 일정이 촉박해 AI 시스템을 배포하기 전에 리스크를 철저하게 평가하고 적절한 제어를 구현하기가 어렵습니다. 출시에 급급한 나머지 보안은 뒷전으로 밀리는 경우가 많습니다. 모범 사례와 규제 지침은 변화의 속도를 따라잡지 못하고 있으므로 보안팀은 업계의 합의나 명확한 표준 없이 상황에 따라 이에 대응하고 판단을 내려야 합니다.

하지만 이는 AI의 보안을 내재화할 수 있는 기회이기도 합니다. AI 개발 프로세스의 첫 단계부터 시스템에 보안 및 거버넌스 고려 사항을 포함함으로써 조직은 선제적으로 리스크를 완화하고 보다 탄력적인 AI 시스템을 구축할 수 있습니다. 이를 위해서는 보안, 법무, AI 개발팀 간의 긴밀한 협업을 바탕으로 모범 사례를 준수하고 필요한 제어를 AI 라이프사이클에 통합해야 합니다.

광범위한 영향

AI 시스템의 잠재적 영향력은 광범위하지만, 그 영향을 파악하기가 어려운 경우도 있습니다. AI 모델은 중대한 결정을 자동화하고, 법적 영향을 미칠 수 있는 콘텐츠(예: 저작권 보호 대상 자료 사용)를 생성하고, 방대한 양의 중요 데이터에 액세스할 수 있습니다. 편향된 결과나 개인정보 침해, 지적 재산 노출, 악의적 사용 등 AI와 관련하여 새로이 등장하는 리스크에는 근본적으로 다른 리스크 관리 패러다임이 필요합니다.

사이버 보안 리더가 이처럼 광범위한 영향을 해결하기 위해서는 법률, 윤리, 비즈니스 부서의 이해관계자와의 협력이 필요합니다. 리스크 허용 범위에 맞춰 협력적 거버넌스 구조를 구축하고, 정책과 가이드라인을 개발하고, 지속적 모니터링 및 감사 프로세스를 구현해야 합니다. 사이버 보안 팀은 데이터 과학 및 엔지니어링 팀과의 긴밀한 협력을 통해 AI 개발 라이프사이클에 보안 및 리스크 관리를 포함시켜야 합니다.

사이버 보안 팀은 데이터 과학 및 엔지니어링 팀과의 긴밀한 협력을 통해 AI 개발 라이프사이클에 보안 및 리스크 관리를 포함시켜야 합니다.

새로운 유형의 보안 및 규정 준수 감독 필요

기존의 사이버 보안 프레임워크는 데이터 기밀성, 무결성, 가용성의 보호에 중점을 두는 경우가 많습니다. 하지만 AI에서는 공정성, 투명성, 책임과 같은 추가적인 요소도 중요하게 작용합니다.

EU의 AI 관련 법률과 같은 새로운 규제 프레임워크는 AI 시스템에 대한 감독 및 거버넌스 메커니즘과 관련하여 새로운 요구 사항을 제시하고 있습니다. 이러한 규정에 따라 기업은 AI 애플리케이션과 관련된 리스크를 평가하고 완화해야 하며, 특히 채용, 신용 평가, 법 집행과 같이 중요성이 큰 영역의 경우 더욱 신중을 기해야 합니다. 규정 준수 의무는 구체적 사용 사례의 리스크 수준에 따라 달라질 수 있습니다. 예를 들어, 채용 의사결정에 사용되는 AI 시스템의 경우 편견이나 차별을 반복하지 않도록 보다 엄격한 감사 및 투명성 요구 사항이 적용될 가능성이 높습니다.

이는 기존에 액세스 제어와 데이터 보호에 집중하던 보안팀이 기존의 영역에서 벗어나 한 발 더 나아가야 한다는 것을 의미합니다. 법무 및 규정 준수팀과 협력하여 AI 모델의 실제 결과와 결정을 모니터링하고 검증하는 메커니즘을 구축해야 합니다. 설명 가능한 AI 기술을 구현하거나, 정기적으로 편향성 감사를 수행하거나, 규정 준수 보고 및 조사를 지원하기 위해 모델 입력, 출력 및 의사결정 로직을 상세하게 문서화하는 것 등이 이에 포함될 수 있습니다.

위협 탐지 및 해결의 새로운 과제

AI 시스템 보안의 기술적 문제점은 새롭고 복잡하며, 탐지 및 해결 과정에 모두 적용됩니다.

새로운 위협 카테고리에는 새로운 탐지 접근 방식이 필요합니다.

앞서 살펴본 바와 같이 보안팀은 기본 데이터와 모델 아티팩트뿐 아니라 운영 중인 AI 시스템의 실제 출력과 동작도 모니터링해야 합니다. 이를 위해서는 방대한 양의 비정형 데이터를 분석하고, 데이터 포이즈닝 공격, 모델 회피 기법, AI의 출력을 해로운 방식으로 조작할 수도 있는 숨겨진 편향과 같은 새로운 위협을 탐지해야 합니다. 기존의 보안 도구는 이러한 AI 전용 위협을 처리하기에 부적합한 경우가 많습니다.

간단하지 않은 해결 과정

기존 소프트웨어 취약점은 몇 줄의 코드로 패치하여 해결할 수 있었지만, AI 모델의 경우 문제를 해결하기 위해 전체 모델을 처음부터 다시 학습시켜야 하는 경우도 있습니다. AI 모델은 학습 데이터를 통해 학습하며, 이는 모델의 파라미터에 깊이 내재되어 있습니다. 학습 데이터에 민감한 정보, 편견 또는 악의적인 예시가 포함되어 있는 것이 확인될 경우 단순히 특정한 데이터 포인트를 제거하거나 수정하는 것은 불가능합니다. 대신 정제된 데이터 세트를 바탕으로 모델을 완전히 재학습시켜야 하는데, 이 작업에 소요되는 시간은 몇 주에서 몇 달에 이르며, 수십만 달러의 컴퓨팅 리소스와 인력이 필요할 수 있습니다.

게다가 모델을 재학습시킬 경우 성능이 저하되거나 새로운 문제가 발생할 수 있으므로 확실한 해결책이 될 수 없습니다. 머신 언러닝, 데이터 제거 및 기타 기술에 대한 연구가 진행 중이지만 이는 아직 초기 단계로서 널리 적용하기는 어려운 상태입니다. 따라서 AI 취약점을 예방하고 조기에 발견하는 것이 가장 중요합니다. 사후에 수정하려면 많은 비용과 시간이 소요될 수 있기 때문입니다.

AI 기반 애플리케이션을 위한 거버넌스 프레임워크 제안

AI 기반 애플리케이션이 제공하는 리스크와 기회를 효과적으로 관리하기 위해 당사가 권장하는 것은 가시성과 제어라는 두 가지 핵심 요소에 초점을 맞춘 새로운 거버넌스 프레임워크를 도입하는 것입니다.

가시성이란 조직 전체에서 AI가 어떻게 사용되고 있는지 명확하게 파악하는 것을 의미합니다. 배포된 모든 AI 모델의 인벤토리를 유지하고, 이러한 모델의 학습 및 운영에 사용되는 데이터를 추적하고, 각 모델의 기능과 액세스 권한을 문서화하는 것이 여기에 포함됩니다. 이처럼 기본적인 가시성이 없다면 리스크를 평가하거나 정책을 시행하는 것은 불가능합니다.

제어란 조직의 가치에 의거하여 책임 있는 AI 사용을 보장하기 위해 마련해야 하는 정책, 프로세스 및 기술적 보호 장치를 의미합니다. 여기에는 AI에 대해 사용 가능한 정보를 규정하는 데이터 거버넌스 정책부터 모델을 개발하고 배포할 수 있는 사람을 제한하는 액세스 제어, 모델 동작과 성능을 검증하는 지속적 모니터링 및 감사에 이르기까지 다양한 요소가 포함됩니다.

목표는 보안, 규정 준수, 엔지니어링 팀을 비롯한 전사적 이해관계자와의 협력을 통해 AI에 적합한 거버넌스 메커니즘을 설계하고 구현할 수 있는 구조적 접근 방식을 보안 리더에게 제공하는 것입니다. 구체적인 구현 상세 사항과 정책은 조직의 요구 사항과 우선순위, 그리고 현지 규정에 따라 달라질 수 있습니다.

	가시성	제어(정책)
모델	어떤 모델을 사용하고 있는가?	01 승인된 모델과 승인되지 않은 모델 02 새로운 모델의 교육/사용 배포를 위한 승인 체인
데이터	학습/추론/파인튜닝에 사용되는 데이터는 무엇인가?	01 허용된 중요 데이터 사용 정책 02 스토리지 처리 감독, 데이터 흐름
사용 사례	AI는 어떻게 사용되고 있는가?	01 승인 및 미승인 사용 사례 02 AI 에이전트 사용 정책
액세스	조직에서 누가 AI를 사용하고 있는가?	01 공개된 AI 애플리케이션에 대해 보다 엄격한 감독 02 제어 정책
규정 준수	관련 규정 준수 프레임워크는 무엇인가?	01 현재 및 미래의 리스크에 대한 지속적 고려 02 위반에 대한 소유권 및 책임

모델

모델 인벤토리에 대한 가시성

우선 조직 전반에 배포된 모든 모델에 대한 포괄적 인벤토리를 구축합니다. 모델 인벤토리는 조직의 AI 발자국을 이해하기 위한 단일 데이터 소스를 제공하며 리스크 평가 및 정책 시행의 토대를 구축합니다. 이 인벤토리에는 모델 제공자, 목적, 의도된 사용 사례와 같은 주요 메타데이터는 물론 모델 유형(예: LLM), 컴퓨터 비전, 표 형식의 데이터, 학습 데이터 소스, 액세스 권한 및 제한, 데이터 흐름 다이어그램이 포함되어야 합니다.

자동 검색 도구는 프로덕션 환경에 배포된 모델을 식별하는 데 도움이 되겠지만, 개발 중이거나 외부에서 호스팅되는 모델을 파악하기 위해서는 수동 프로세스가 필요할 수 있습니다.

승인된 모델과 승인되지 않은 모델에 대한 정책

조직은 사용, 평가 및 고객 대상의 배포를 승인할 모델을 결정해야 합니다. 정책은 조직의 전반적 리스크 허용 범위 및 규정 준수 의무와 일치해야 합니다.

모델을 승인할 때는 다음과 같은 사항을 고려합니다.

- 모델 소스 및 계보(예: 평판이 좋은 공급업체 vs. 오픈 소스)
- 테스트 및 검증 수준
- 책임 있는 AI를 위한 조직의 가치 및 원칙과의 일치 여부
- 관련 규정 및 업계 표준 준수
- 보안 및 거버넌스 제어와의 통합

이러한 기준을 충족하지 못하는 모델은 배포를 차단하거나 훨씬 더 엄격한 승인 절차를 진행해야 합니다. 제재 대상 모델의 경우 적절한 사용 사례, 필요한 보안 및 규정 준수 제어, 지속적 모니터링 및 유지 관리 기대치를 중심으로 가이드라인을 수립해야 합니다. 이러한 가이드라인은 배포 전반의 일관성과 리스크 관리를 보장하는 데 도움이 됩니다.

새로운 AI 모델 심사 및 승인을 위한 정책

조직에는 보안 및 규정 준수 팀이 신규 모델이나 사전 배포된 AI 모델을 심사할 수 있는 체계적인 프로세스가 필요합니다. 이러한 심사 프로세스를 통해 제재 기준에 따라 모델을 평가하고 특정 리스크 또는 필요한 제어를 식별해야 합니다.

개발부터 스테이징, 프로덕션에 이르는 소프트웨어 개발 라이프사이클 동안 승인된 모델을 활성화하기 위해서는 명확한 프로세스가 필요하며, 각 단계별 테스트 및 승인 요건을 갖춰야 합니다. 또한 배포된 모델을 지속적으로 모니터링하고 모델 재교육, 데이터 드리프트 또는 성능 저하와 같은 중대한 변경 사항이 발생할 경우 추가적인 검토를 트리거하는 프로세스를 개발해야 합니다.

데이터

AI 모델 학습 및 배포에 사용되는 검색 및 분류 데이터

데이터는 AI 모델의 생명선과 같습니다. 따라서 AI 라이프사이클 전반에 걸쳐 사용되는 데이터가 무엇인지 명확하게 파악하는 것이 중요합니다. 모델 학습, 프로덕션에서의 추론 또는 예측, 추후 모델의 파인튜닝 또는 재학습에 사용되는 데이터가 여기에 포함됩니다. 조직은 AI 인벤토리를 최신 상태로 유지관리함으로써 모든 데이터 세트의 상세 사항을 확보하고 민감도 수준, 규제 요건 및 승인된 사용 사례에 따라 분류해야 합니다. 이 인벤토리는 모델 인벤토리와의 통합을 통해 모델과 기반 데이터 간의 추적 가능성을 제공해야 합니다.

데이터 포이즈닝을 방지하기 위한 정책

학습 데이터의 거버넌스는 데이터 포이즈닝 공격의 측면에서 특히 중요합니다. 공격자가 학습 데이터를 조작할 수 있다면 숨겨진 백도어나 편향성을 심고 이를 악용하여 프로덕션 환경에서 모델의 동작을 방해할 수 있습니다. 이러한 리스크를 완화하기 위해서는 강력한 액세스 제어와 지속적인 데이터 흐름의 모니터링이 필요합니다.

학습, 추론 및 파인튜닝을 위한 민감한 데이터 사용 관련 정책

조직은 다양한 유형의 데이터를 AI에 사용할 수 있는 방안을 관장하는 명확한 정책을 수립해야 합니다. 이러한 정책은 데이터의 민감도 수준, 규제 요건 및 윤리적 고려 사항을 기반으로 규정해야 합니다.

예를 들어, 정책에 대상의 명시적 동의를 취득하고 익명 처리를 하지 않을 경우 개인 식별 정보(PII) 또는 보호 대상 건강 정보(PHI)를 모델 학습에 사용할 수 없도록 규정할 수 있습니다. 마찬가지로, 적절한 라이선스 및 사용 권한이 없는 중요 지적 재산 또는 타사 데이터를 AI에 사용하는 것을 제한할 수도 있습니다. 또한 정책은 암호화, 액세스 제어, 보존 제한 등 다양한 유형의 데이터에 필요한 보안 및 개인 정보 보호 제어 방안을 정의할 수 있습니다. 규정 준수 팀은 이러한 정책의 수립 과정에 참여하여 조직의 정책이 GDPR, HIPAA, CCPA와 같은 관련 규정에 부합하는지 확인해야 합니다.

사용 사례

AI의 용도에 대한 가시성 확보

AI 사용 사례에 따라 보안 및 규정 준수에 미치는 영향은 크게 달라질 수 있습니다. 예를 들어, 고객 지원을 위한 AI 기반 챗봇은 데이터 개인정보 보호 및 콘텐츠 필터링을 엄격하게 통제해야 하겠지만, 공급망 물류 최적화를 위한 내부 AI 도구의 경우 그러한 문제가 발생하지 않을 수 있습니다.

사용 사례 문서에는 다음과 같은 주요 세부 정보가 포함되어야 합니다.

- 비즈니스 목적 및 의도한 결과
- 최종 사용자 및 이해관계자
- 데이터 입력 및 출력
- 의사 결정 범위 및 자율성
- 인적 감독 및 개입 지점
- 성과 지표 및 성공 기준
- 리스크 평가 및 완화 계획

승인된 사용 사례 및 승인되지 않은 사용 사례에 대한 정책

조직은 허용되는 AI 사용 사례와 그 조건을 명확하게 규정하는 정책을 정의해야 합니다. 사용 사례를 승인함에 있어 고려해야 하는 요소는 다음과 같습니다.

- 책임 있는 AI를 위한 조직의 가치 및 원칙과의 일치 여부
- 관련 법률, 규정 및 업계 표준 준수 여부
- 피해 또는 의도치 않은 결과가 발생할 가능성
- 인적 감독 및 통제 수준
- 투명성 및 설명 가능성 요구 사항
- 평판 및 브랜드 리스크

대출, 고용, 의학적 진단 등 개인에 대한 의사결정이 수반된 중요 사용 사례의 경우 더욱 면밀한 조사가 필요하며, 전담 검토 위원회 또는 감독 위원회가 필요할 수 있습니다. 내부 생산성 도구와 같이 리스크 정도가 낮은 사용 사례는 보다 간소화된 승인 절차에 따라 진행할 수 있습니다.

승인된 AI 에이전트 사용 정책 정의

AI 에이전트는 개별 단계에서 인적 개입 없이 자율적으로 의사결정을 내리고, 앞선 결과를 바탕으로 대응 조치를 취할 수 있는 특수한 유형의 AI 시스템입니다. 예를 들면, 코드의 작성, 테스트 및 최적화 작업에 AI 에이전트를 활용할 수 있습니다. 이러한 에이전트를 사용할 경우 잠재적 리스크와 의도치 않은 결과를 예측하고 통제하기가 어렵다는 점에서 추가적인 거버넌스 문제가 발생할 수 있습니다. 조직은 AI 에이전트의 사용 시점과 방법을 관장하는 명확한 정책을 수립해야 합니다.

- 에이전트의 의사 결정 권한 범위 정의
- 성능 한계 및 안전 장치 메커니즘 설정
- 이상 동작에 대한 강력한 모니터링 및 경고 구현

AI 에이전트의 높은 자율성과 잠재적 영향을 고려하면 이를 전담하는 리스크 평가 및 승인 프로세스가 필요할 수 있습니다.

권한 및 액세스

조직에서 AI를 사용하는 자에 대한 가시성

AI의 효과적 거버넌스를 위해서는 전체 조직에서 AI 개발 및 사용에 관여하는 사람을 파악해야 합니다.

- 역할 및 책임(예: 데이터 과학자, ML 엔지니어, 제품 관리자)
- AI 개발 및 배포 도구에 대한 액세스 권한
- AI 시스템 액세스에 사용되는 머신 ID(예: 서비스 계정, API 키)

공개된 AI 애플리케이션에 대해 보다 엄격한 감독

고객 또는 대중과 직접 소통하는 AI 애플리케이션의 경우 기업 내부 애플리케이션에 비해 더 많은 거버넌스와 감독이 필요합니다. 이러한 추가적 심사 과정에는 다음과 같은 사항이 포함될 수 있습니다.

- 다양한 인구 통계 그룹에 대한 엄격한 편향성 및 공정성 테스트
- 안전 리스크 및 악용 가능성을 조사하기 위한 적대적 테스트
- 더욱 빈번하거나 종합적인 규정 준수 감사

또한 조직은 대중을 대상으로 하는 AI에 추가적 기술적 보호 장치를 구현하는 안을 고려해야 합니다. 미리 정의된 리스크 임계값에 의거하여 속도를 제한하거나, 콘텐츠를 필터링하거나, 자동으로 차단을 트리거하는 등의 방안을 예로 들 수 있습니다. 새로운 리스크를 사전에 식별하고 완화하기 위해서는 정기적인 감사와 영향 평가가 중요합니다.

규정 준수

관련 규정 준수 프레임워크의 감독

AI 거버넌스는 그 자체로 혼자 존재하는 것이 아닙니다. 조직 전반의 규정 준수 관리 프레임워크와 통합되어야 합니다. 즉, 관련 법률, 규정 및 업계 표준에 따라 AI 정책과 제어를 조정해야 합니다.

다음과 같은 주요 규정 준수 프레임워크를 고려합니다.

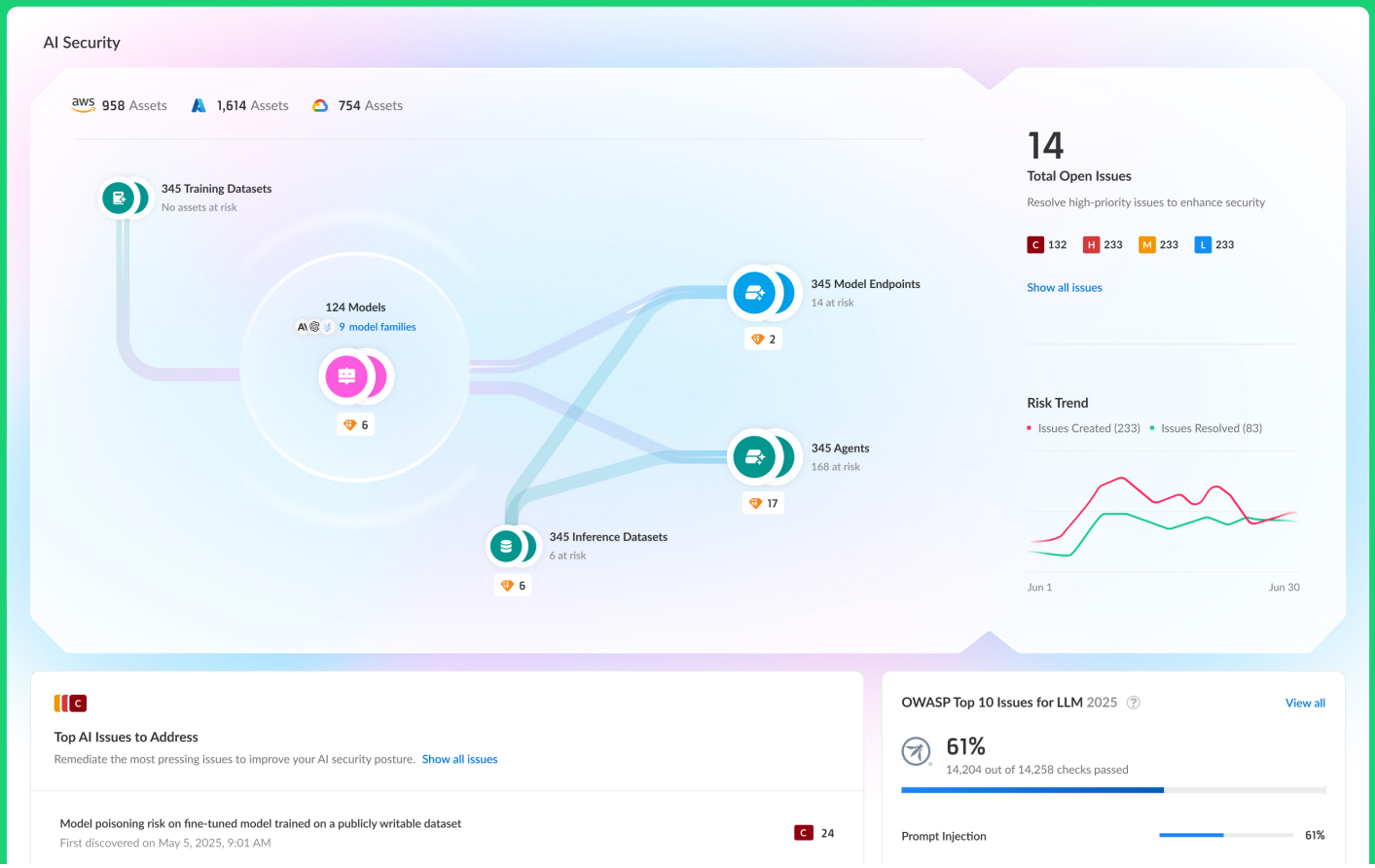
- 데이터 보호 규정(예: GDPR, CCPA, HIPAA)
- 해당하는 섹터별 규정(예: 금융 부문의 경우 FINRA, 의료 부문의 경우 FDA)
- 새롭게 등장하는 AI 관련 규정(예: 뉴욕의 AI 편견 법률)
- 자발적 표준 및 인증(예: IEEE, ISO, NIST)

규정 준수 팀은 AI 개발 및 거버넌스 팀과의 긴밀한 협력을 통해 적용 가능한 요건을 파악하고, 이를 실행 가능한 정책 및 제어로 전환해야 합니다. AI 라이프사이클의 주요 지점에서 갭 평가, 데이터 보호 영향 평가(DPIA), 규정 준수 중심의 검토가 필요할 수 있습니다.

또한 규정 준수 감독은 타사 AI 공급업체 및 파트너로 확대되어 이들의 관행이 조직의 규정 준수 의무와 일치하는지 확인해야 합니다. 공급업체 리스크 관리 프로세스에는 AI 관련 실사 기준과 계약 요건을 통합해야 합니다.

현재 및 미래의 규정 준수 리스크에 대한 지속적 고려

AI와 관련된 규제 환경은 커다란 변화를 겪고 있습니다. 새로운 법률과 표준이 제시되고, 또한 제정되고 있습니다. 이러한 변화는 조직의 AI 개발 및 배포 방식에 중대한 영향을 미칩니다. 시간이 흘러도 건전한 규정 준수 상태를 유지할 수 있도록 정기적으로 검토와 감사를 실시해야 합니다.



Cortex Cloud AI Security Posture Management: AI에 대한 가시성, 제어 및 거버넌스

앞서 언급한 리스크와 문제를 고려할 때, 조직에는 빠른 속도로 확장되는 AI 배포에 대해 조직을 보호할 수 있는 새로운 도구와 접근 방식이 필요합니다. 실제로 2024년 우선순위에 대한 응답에서 100%의 조직이 전체 AI 배포 파이프라인에 대한 가시성을 확보하기 위해 노력하고 있다고 답했습니다(Palo Alto Networks의 *2024년 클라우드 네이티브 보안 현황 보고서* 참조).

Cortex Cloud AI Security Posture Management(AI-SPM)는 AI, 머신 러닝, 생성형 AI 모델과 관련하여 고유하게 발생하는 리스크로부터 조직을 보호하는 포괄적 솔루션을 제공합니다. 이 솔루션은 데이터 수집부터 학습, 구축에 이르기까지 AI 모델 에코시스템 전반에 대한 가시성을 제공합니다. 모델 동작과 데이터 흐름, 시스템 상호 작용을 분석하여 기존의 도구로는 포착하기 어려운 잠재적 보안 및 규정 준수 리스크를 식별합니다.

이 솔루션의 주요 기능은 다음과 같습니다.

- **AI 모델 검색 및 인벤토리:** 전체 조직에서 사용 중인 모든 모델 API, 오픈 소스 모델, VM 배포 모델의 인벤토리를 생성하는 작업이 여기에 포함됩니다. 이를 통해 모델 확산 및 섀도 AI를 제어하고, 무단 모델 사용을 방지하고, 적절한 거버넌스 제어를 확립할 수 있습니다.
- **데이터 노출 방지:** AI-SPM은 AI 모델의 학습 및 운영에 사용되는 데이터 세트를 검색하고 분류하여 중요 데이터의 노출 가능성을 알립니다. 실시간 모델 상호 작용을 모니터링하여 오용 또는 의도치 않은 데이터 유출을 탐지합니다.

- **태세 및 리스크 분석:** AI-SPM은 엔드투엔드 AI 배포 파이프라인을 스캔하여 잘못된 구성, 취약한 액세스 제어, 그리고 모델과 데이터를 위험에 빠뜨릴 수 있는 기타 취약점을 식별합니다. 사용자 액세스 권한의 시각적 매핑을 제공하고 지나치게 광범위한 권한을 적절하게 조정할 수 있도록 도와줍니다.

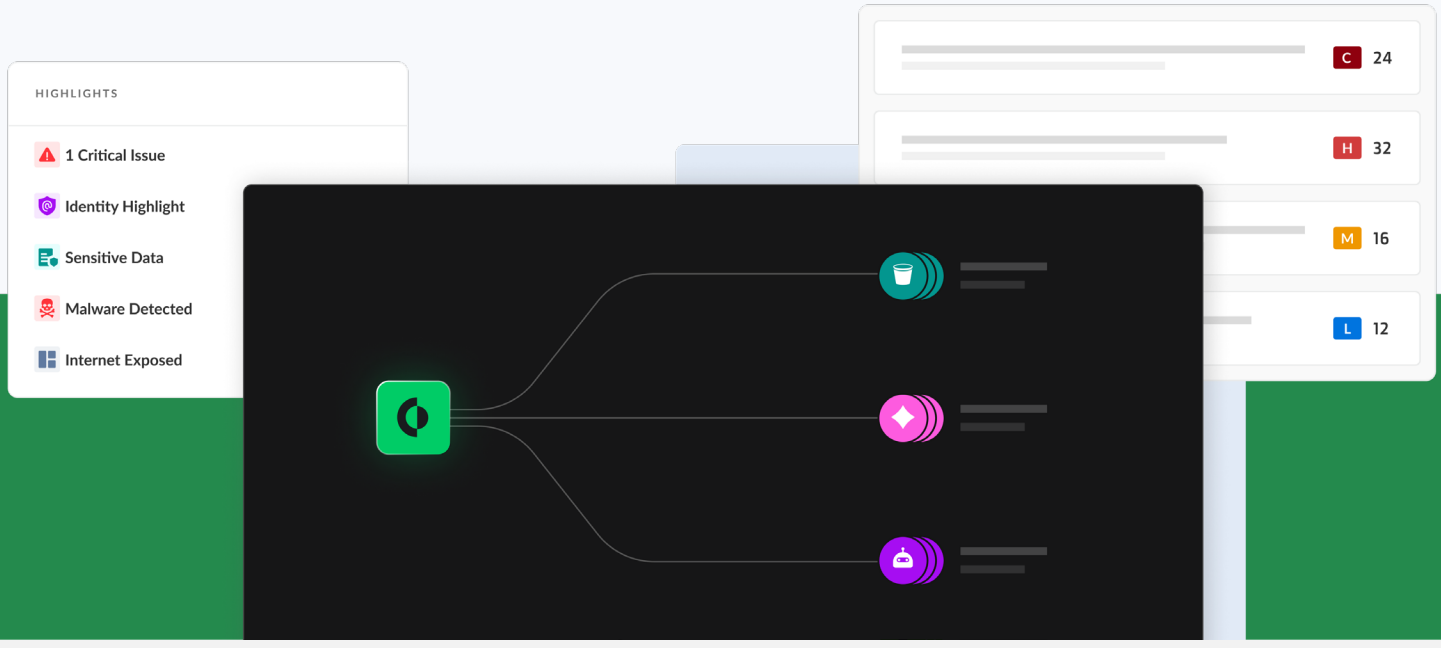
조직은 이러한 AI-SPM 인사이트를 활용하여 보안 정책의 일관성을 확보하고, AI 관련 위협을 선제적으로 완화하고, EU의 AI 법률과 같이 변화하는 규정을 준수할 수 있습니다.

Cortex Cloud AI-SPM은 광범위한 Code to Cloud™ 플랫폼과 원활하게 통합되며, 클라우드 네이티브 스택 전반에서 통합적 가시성과 제어를 제공하는 완벽한 CNAPP 기능을 제공합니다(CSPM, DSPM 포함). 빠르고 간편한 에이전트리스 배포 모델을 통해 Cortex Cloud를 몇 분 만에 가동하고 실행함으로써 중요한 AI 자산을 보호하고 대규모의 책임감 있는 혁신을 실현할 수 있습니다.

Palo Alto Networks를 통한 코드-클라우드-SOC 보안 소개

차세대 Prisma Cloud인 Cortex Cloud는 업계 최고의 CNAPP와 동급 최고의 CDR을 통합하여 실시간 클라우드 보안을 제공합니다. AI 및 자동화를 활용하여 런타임 컨텍스트를 통해 리스크의 우선순위를 지정하고, 대규모로 수정하고, 공격이 발생할 경우 신속하게 차단할 수 있습니다. 단일 Cortex 플랫폼에서 클라우드와 SOC를 통합하여 엔드투엔드 운영을 혁신하세요.

https://www.paloaltonetworks.com/cortex/cloud에서 실시간 클라우드 보안의 미래를 경험해 보세요.



서울특별시 서초구 서초대로74길 4,
1층 (삼성생명 서초타워)
Tel: +82-2-568-4353
eMail: Sales-KR@paloaltonetworks.com
www.paloaltonetworks.co.kr

cortex cloud_Whitepaper_AI Governance_2025