Google

책임감 있는 머신러닝을 위한 가이드라인

가이드라인

ML 모델의 출력에 따라 제어 흐름을 수정하는 경우에는 주의해야 합니다

머신러닝 프리미티브의 출력을 소비하는 워크플로는 보안이 어려운 공격 표면을 생성하도록 수정할 수 있습니다. 따라서 확실하게 예상할 수 없는 실행 워크플로에 데이터를 수집하는 것은 권장하지 않습니다.

모델과 모델 메타데이터에 직접 액세스하지 않도록 합니다

머신러닝 프리미티브를 이용한 직접적인 상호작용은 유출, 반전, 기타 다양한 공격 표면으로 이어질 수 있습니다. 이러한 객체를 이용자가 사용하지 못하게 하고, 승인된 상호작용만 허용하며 데이터베이스와 같은 기존 기능에 출력이 연결되도록 합니다. 이용자에게 출력 기능을 직접 제공하지 마십시오.

훈련 및 사후 구현 시 모델에 무결성 검사를 반드시 실행하세요

소프트웨어에 대한 수정 사항은 구축 및 배포 파이프라인의 여러 단계에서 발생할 수 있으며 이는 머신러닝에서도 마찬가지입니다. 따라서 검사를 반복적으로 실행하세요. 작동하기 전 검증과 같이 예상 가능한 부분만 실행합니다.

기능이 복잡해지면 제어도 복잡해질 수 있습니다

상호작용은 짧고 핵심적으로 간결해야 합니다. 워크플로가 복잡하고 대규모이면 공격 성공 시 노출 위험이 훨씬 커집니다.

훈련을 위한 익스텐셔널 로직으로 시작하여 필요한 경우 기능 및 추론을 위한 인텐셔널 로직으로 이동합니다

훈련 시, 검증되고 잘 알려진 양질의 세트를 사용하는 것은 데이터 중독과 기타 공격을 피하기 위한 좋은 방법입니다. 그러나 더욱 뛰어난 추론 역량을 갖춘 기능을 개발하기 위해서는 새로운 데이터를 수집해야 합니다. 잘 알려진 양질의 데이터 입력을 통해 예상되는 결과를 기반으로 학습 결과를 신중하게 테스트한다면 더욱 탄력적인 시스템을 구현할 수 있습니다.

프리미티브가 의도한 이용 사례에 따라 생성되었는지 통합 테스트를 확인합니다

다른 소프트웨어와 마찬가지로 머신러닝 워크플로에도 통합 및 단위 테스트가 필요합니다. ML 소프트웨어의 휴리스틱한 특성으로 신뢰할 수 있는 어설션을 구축하기에 어려움이 있을 수 있기 때문에 워크플로에 필요한 장기적 결과에 집중하는 것을 권장합니다.

머신 합성을 사용한 공격, 특히 '당신이 누구인지'에 초점을 맞춘 제어에 대한 공격을 찾아봅니다

특히 액세스 권한을 제공해야 하는 경우, 음성, 동영상, 텍스트의 진위 여부를 확인하기 전에 항상 '지금 머신이 이 작업을 수행할 수 있는가?'를 질문해 봅니다.

상호 정책을 사용하여 경제 모델보다 앞서 나가면서도 강력하고 비용 효율적인 보안을 보장합니다

공격 비용이 후속 해결 비용보다 낮으면 제어의 효과가 떨어질 수 있습니다. 공격의 성공 영향에 따라 보안 비용을 자동으로 조정하여 공격자 비용이 점차 높아지도록 합니다.

Google

독점 데이터로 학습된 모델을 디바이스와 같이 이용자가 제어할 수 있는 곳에 사용하지 않도록 합니다

머신러닝 프리미티브 생성과 관련된 이용 사례를 고려하여, 이론적으로 머신러닝 프리미티브를 생성한 학습 데이터를 복구할 수 있다고 가정합니다. 이는 제품을 개발할 때 비즈니스 연속성 계획 및 위험 평가에 사용합니다.

비율 제한을 사용하여 외삽 공격을 피하고 불필요한 리소스 소비를 방지합니다

머신러닝 기술의 설계 및 구현 시 액세스 제어 목록, 비율 제한, 인증 등의 레이어 제어와 같은 기존의 제어 방식은 여전히 상당 부분 효과적입니다.