Google

責任ある機械学習のための ガイドライン

ガイドライン

機械学習モデルの出力に基づいて制御フローを変更する場合は注意が必要です。

機械学習プリミティブの出力を利用するワークフローは、安全性を確保するのが困難な攻撃対象領域を生み出す可能性があります。従来のアドバイスと同様に、ワークフローが本来処理すべきでないデータを取り込むことは推奨されません。

モデルやモデルのメタデータへの直接アクセスは避けてください。

機械学習プリミティブとの直接的なインタラクションは、データ流出や反転など、さまざまな攻撃手法につながる可能性があります。これらのオブジェクトをユーザーから遠ざけ、制御された、承認者のみが操作できる環境を構築し、出力がデータベースなど確立された機能と関連するようにしてください。ユーザーに出力を直接提供しないでください。

モデルは、トレーニング中と実装後の両方で整合性チェックを行うように設定します。

ソフトウェアの変更は、ビルドやデプロイ パイプラインの複数の段階で発生する可能性があります。機械学習においても同様です。繰り返し確認してください。結果が予期できるものだけを実行します。操作をする前に認証を行います。

機能の複雑さが増すと、制御もまた複雑になる可能性があります。

インタラクションは的確で簡潔なものにします。大規模で複雑なワークフローは、攻撃を受けるリスクがはるかに大きくなります。

トレーニングでは外延論理から始めて、必要に応じて機能と推論についての内包論理を適用します。

検証済みの高品質なデータセットをトレーニングに使用することは、データ ポイズニングや他の攻撃を回避する上で効果的です。ただし、より高度な推論能力を持つ機能を開発するには、新しいデータを取り込む必要があります。新しいデータを使用して学習し、信頼のおける入力からの予想される出力に基づくトレーニング結果を慎重にテストすることで、より耐性のあるシステムを構築できます。

統合テストは、プリミティブの意図されたユースケースに沿って必ず作成してください。

機械学習ワークフローには、他のソフトウェアと同様に、統合テストと単体テストが必要です。機械学習ソフトウェアのヒューリスティックな性質により、信頼性のあるアサーションを構築するのが難しい場合があります。そのため、ワークフローの長期的かつ望ましい結果に焦点を当てることを推奨します。

生体認証に焦点を当てた制御に関連する領域では、機械学習合成を利用した攻撃に注意する必要があります。

音声、ビデオ、テキストによる本人確認の際、「AI の可能性はないだろうか？」と常に問いかけるようにしてください。特に、アクセス権付与のために使用する場合は注意が必要です。

経済モデルに先んじるために相互ポリシーを利用し、強力で低コストのセキュリティ モデルを確立します。

攻撃を受けた後の被害復旧にかかる費用が、攻撃を事前に防ぐための対策にかかる費用よりも低い場合、その対策は効果的でない可能性があります。セキュリティ対策に関連する費用を管理し、攻撃を受けた後の被害状況に応じて自動的に費用を調整し、攻撃者にとって攻撃を行うコストを徐々に高めていくことが、攻撃の抑制につながります。

Google

独自のデータに基づいて学習させたモデルを、デバイスなど、ユーザーが直接操作できる環境に配置するのは避けてください。

機械学習プリミティブの作成に関連するユースケースを検討し、理論上、そのモデルを作成するために使用されたトレーニングデータが復元される可能性があることを想定する必要があります。製品開発時の事業継続計画やリスク評価において、この点を考慮してください。

レート制限を使用して類推攻撃を回避し、余分なリソースの消費を防止します。

従来型のセキュリティ対策は今でもほとんど有効です。機械学習技術の設計と実装においては、アクセス制御リスト、レート制限、認証などの多層防御を活用してください。