資料１－４

AIシステムに対する脅威の事例

2025/9/18
三井物産セキュアディレクション株式会社

AIシステムに対する脅威の事例

● AIシステムに対する脅威の事例として以下が挙げられる。
※「AISIの分類」は、AISI「AIシステムに対する既知の攻撃と影響」より引用。

AISIの分類	脅威	脅威の具体例 （詳細は参考資料３）	攻撃経路	LLM/ 画像識別AI
A:モデル抽出攻撃	**① モデル抽出攻撃** • AIシステムのエンドポイント(API等)に対して繰り返しアクセスし、AIモデルへの入出力情報を観察することで、同程度の性能を持つ複製モデルを作成する攻撃手法。 • 競争上の地位低下や、モデルに含まれる機密情報の窃取などにつながる。	L-１-1：LLMの入出力情報を観察してAIモデルを抽出する攻撃	AI 利用者が入力するプロンプト	LLM 画像識別AI
		I-1-1：画像識別AIを模倣したAIを作成する攻撃	AI 利用者が入力する画像	
B:学習データ情報収集攻撃	**② モデル反転攻撃** • AIの学習データを再現する攻撃手法。 • ランダムに初期化した画像をAIシステムに入力し、その結果として得られるAIモデルの内部情報（勾配情報）を基に、初期化画像を窃取対象の画像に近づけていくもの。	I-2-1：画像識別AIの内部構造を観察することで学習データを窃取する攻撃	AI 利用者が入力する画像	画像識別AI
	③ メンバーシップ推論攻撃 • 入力画像に対する画像識別AIの応答を観察することで、入力した画像がAIの学習データに含まれているかどうかを推測する攻撃手法。 • 学習が不十分な画像識別AIは、学習した画像と学習していない画像では、明確に異なる応答を返すため、その性質を悪用するもの。	I-3-1：画像識別AIの挙動を観察することで学習データを推測する攻撃	AI 利用者が入力する画像	

AISIの分類	脅威	脅威の具体例 （詳細は参考資料３）	攻撃経路	LLM/ 画像識別AI
C:モデルポイズニング攻撃	**④　モデルポイズニング攻撃** • AIモデル自体を細工することで、AIモデルの出力データを改ざんする攻撃手法。 • 攻撃者によって細工されたAIモデルを外部の学習済みモデルをホストするサービスから調達し、システムに組み込んでサービスを提供してしまうと、AI利用者のプロンプトに対し、AI提供者の意図しない不正な回答をAIモデルが出力してしまう。	L-2-1：学習済みモデルを細工してLLMの挙動を操作する攻撃	調達する学習済みモデル	LLM 画像識別AI
		I-4-1：細工された学習済み画像識別AI経由で悪意のあるコードを実行する攻撃	調達する学習済みモデル	
		I-4-2：バックドアが設置された学習済み画像識別AIを使用した攻撃	調達する学習済みモデル	
D:データポイズニング攻撃	**⑤　データポイズニング攻撃** • AIモデルの学習データやファインチューニングデータを細工（特定のプロンプトに対して不正な内容を指示する特定の画像を誤識別させる等）することで、AIモデルの出力データを改ざんする攻撃手法。 • 攻撃を受けた場合、AI利用者が入力したプロンプトに対し、AI提供者の意図しない不正な回答をAIモデルが出力する。画像識別AIは攻撃者しか知り得ない特定の入力画像を、攻撃者が意図した物体として誤識別する（その他の入力画像は正しく識別される）。	L-3-1：学習データを細工してLLMの挙動を操作する攻撃	事前学習データ	LLM 画像識別AI
		L-3-2：ファインチューニングデータを細工してLLMの挙動を操作する攻撃	ファインチューニングデータ	
		I-5-1：学習データを細工して画像識別AIにバックドアを設置する攻撃	事前学習データ	

AISIの分類	脅威	脅威の具体例 （詳細は参考資料３）	攻撃経路	LLM/ 画像識別AI
E:回避攻撃	**⑥　回避攻撃** • AIモデルに入力する画像に、人間には識別できない微細なノイズを混ぜる（誤識別を誘発するように細工する）ことで、AIの挙動を操作する攻撃手法。	I-6-1：入力画像を細工することで画像識別AIの誤識別を誘発する攻撃	AI 利用者が入力する画像	画像識別AI
F:スポンジ攻撃	**⑦　スポンジ攻撃** • AIモデルが稼働するシステムの計算資源を枯渇させる攻撃手法（DoS攻撃とも呼ぶ）。 • 攻撃者がAIモデルのファインチューニングデータや入力データを細工することで、システムの計算資源を大量消費するような回答(出力システムの出力文字数上限を超えた回答等)が生成される。 • 画像識別AIにおいては、攻撃者は何らかの方法でAIモデルの内部構造情報を取得・解析し、画像識別処理を行う中で多くの計算を要するロジックを特定した上で、AIの計算負荷が増大するように設計した画像を作成し、これをAIモデルに入力する。	L-4-1：ファインチューニングデータを細工してDoSを行う攻撃	ファインチューニングデータ	LLM 画像識別AI
		L-4-2：プロンプトを細工してDoSを行う攻撃	AI 利用者が入力するプロンプト	
		I-7-1：入力画像を細工してDoSを行う攻撃	AI 利用者が入力する画像	
G:プロンプト窃盗攻撃	※事前調査対象外 （出典：AISI「AIシステムに対する既知の攻撃と影響」） • AIモデルが生成した画像などの出力から、元となったプロンプトを推測することで、プロンプトエンジニアリングのノウハウとなる入力情報の漏洩を引き起こす	–		

AISIの分類	脅威	脅威の具体例 （詳細は参考資料３）	攻撃経路	LLM/ 画像識別AI
H:プロンプトインジェクション攻撃	**⑧　直接的プロンプトインジェクション攻撃（不正操作型）** • AIモデルの入力データを細工することで、AIモデルの出力データを改ざんする攻撃手法。 • 攻撃者が細工したプロンプトをAIモデルに入力することで、AI提供者の意図しない不正な回答をAIモデルが出力する。	L-5-1：プロンプトを細工してLLMの挙動を操作する攻撃	AI 利用者が入力するプロンプト	LLM
	⑨　直接的プロンプトインジェクション攻撃（システム攻撃型） • AIモデルと連携するシステム（データベースやAIモデルが稼働するシステム等）を不正操作する攻撃手法。 • 攻撃者が細工したプロンプトをAIモデルに入力することで、連携システムを不正操作するコード（SQLクエリやシステムコマンド等）がAIモデルによって生成され、連携システム上で実行されることで、データベースやシステムからの機密情報漏えいや情報の改ざん・削除等が引き起こされる。	L-6-1：プロンプトを細工してLLMと連携するシステムを不正操作する攻撃(P2SQL Injection, LLM4Shell)	AI 利用者が入力するプロンプト	
	⑩　直接的プロンプトインジェクション攻撃（参照データの窃取型） • AIモデルが参照対象とするデータを窃取する攻撃手法。 • 攻撃者がAIシステムに入力するプロンプトを細工（RAG用データストア情報の開示を要求する等）することで、本来は開示すべきではないRAG用データストア（ベクトルデータベースやファイルシステム等）の内容を含む回答を生成させる。	L-7-1：プロンプトを細工してRAG用のデータを窃取する攻撃	AI 利用者が入力するプロンプト	
	※　AISIの分類に該当なし **⑪　直接的プロンプトインジェクション攻撃（学習データの窃取型）** • AIモデルの学習データを窃取する攻撃手法。 • 攻撃者が細工したプロンプトをAIモデルに入力することで、AIモデルの回答に学習データの一部が出力される。	L-8-1：プロンプトを細工してLLMの学習データを窃取する攻撃	AI 利用者が入力するプロンプト	

AISIの分類	脅威	脅威の具体例 （詳細は参考資料３）	攻撃経路	LLM/ 画像識別AI
（承前）	**⑫　間接的プロンプトインジェクション攻撃** • AIモデルがRAG等で参照するデータを細工することで、AIモデルの出力データを改ざんする攻撃手法。 • 攻撃者は細工したリソース(Webサイト等)を用意し、AIモデルが運用時に当該リソースを動的に参照することで、AI提供者の意図しない不正な回答をAIモデルが出力する。	L-9-1：LLMが参照する外部情報を細工してAIの挙動を操作する攻撃	外部システムから取得する情報	（承前）
		L-9-2：LLMが参照するRAG用データを細工してAIの挙動を操作する攻撃	外部システムから取得する情報	
	⑬　プロンプトリーキング攻撃 • システムプロンプトを窃取する攻撃手法。 • 攻撃者が細工したプロンプトをAIモデルに入力することで、AIモデルの回答に（第三者に公開することを意図していない）システムプロンプトの内容が出力される。	L-10-1：プロンプトを細工してシステムプロンプトを窃取する攻撃	AI 利用者が入力するプロンプト	
I:コードインジェクション攻撃	**⑭　MLaaSの悪用** • 攻撃者の制御下にあるMLaaSをAI提供者が利用してAIモデルを作成することで、AIモデルの内部に悪意のあるコードを実行する仕組みを設置する攻撃手法。	I-8-1：MLaaSサービスを悪用して悪意のあるコードを実行する攻撃	MLaaS	画像識別AI
		I-8-2：MLaaSサービスを利用して画像識別AIにバックドアを設置する攻撃	MLaaS	
J:ファインチューニング攻撃	※事前調査対象外 （出典：AISI「AIシステムに対する既知の攻撃と影響」） • 標的の事前学習AIモデルに対して特定のファインチューニングを実施することで、 AIモデルがセーフガードを回避し、事前の学習データを漏洩するように仕向ける。			
K:ロウハンマー攻撃	※事前調査対象外 （出典：AISI「AIシステムに対する既知の攻撃と影響」） • 標的のAIモデルと物理メモリを共有する攻撃者が、AIモデルのメモリのビットを メモリセル間の干渉により反転させることでモデルの漏洩や誤動作を引き起こす。			