AIシステムに対する既知の攻撃と影響

第2版 2026年 4月
第1版 2025年 3月

AISI Japan AI Safety Institute

概要

AISI Japan AI Safety Institute

- 本資料は、**AIシステムに対する特有の攻撃と影響**を俯瞰する。
- AIシステムの利活用が急速に進む中、**AIの特性を悪用した攻撃**が次々と発表されている。これらの攻撃は、学習データの漏洩や、AIモデルの異常な出力といった影響の原因となる。
- 実社会では、これらの影響が人々の生活に深く関係してくる。（以下、例）
  - 医療診断でのAIシステムの学習データの漏洩はプライバシー侵害になる。
  - 自動運転でのAIシステムの異常な出力は人命に関わる事故につながる。
- このようにAIシステムに特有のセキュリティ対策は今や必要不可欠である。
- 本資料は、学術論文等で発表された、AIシステムに特有の攻撃と影響を俯瞰し、対策を検討するための参考となる情報を提供する。

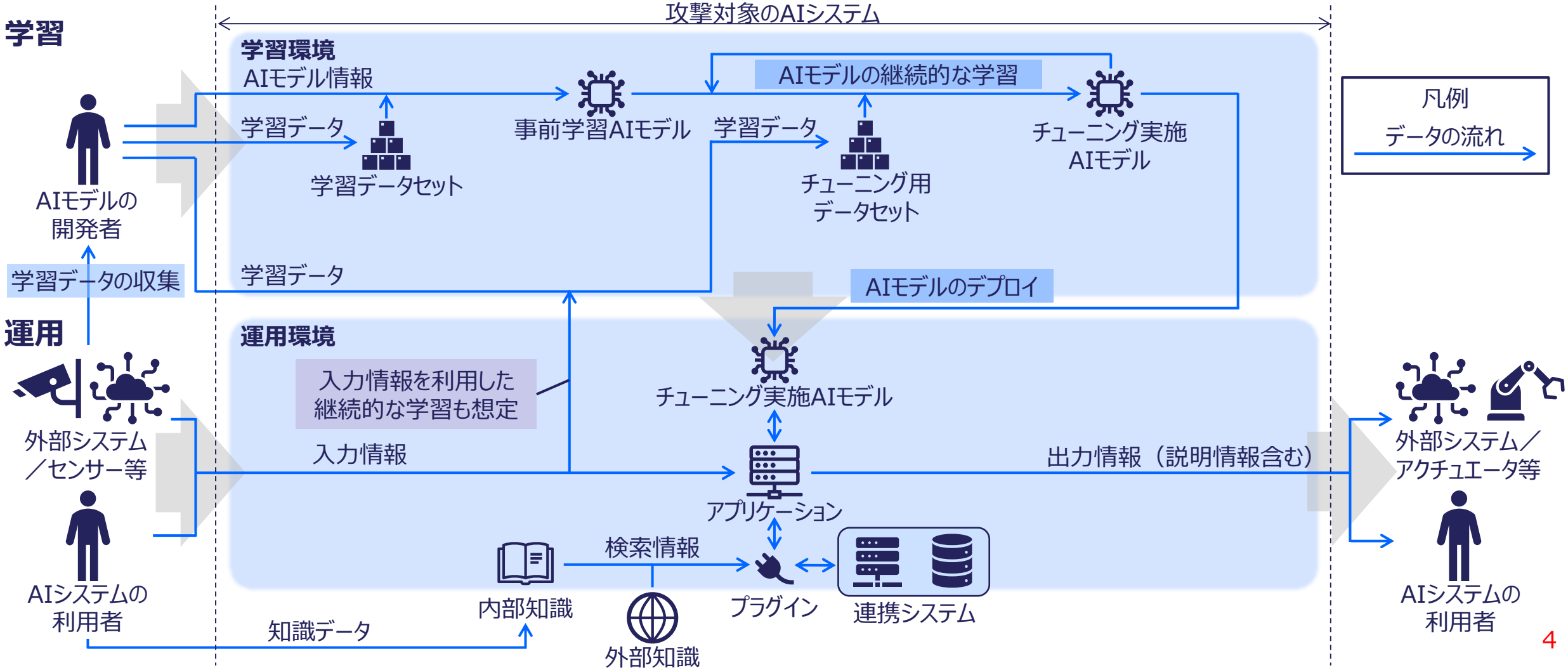
本資料のスコープ

AISI Japan AI Safety Institute

- 従来のサイバーセキュリティ攻撃への対策が必要であることは前提とする。
- 本資料では、既知の攻撃の中でも**システムがAIを含むことで留意すべき、AIモデルの学習もしくは推論の過程への介入や、AIモデルの入出力の利用・改ざんを伴う攻撃**を扱う。
  - 扱う例
    - 学習時：データポイズニング攻撃（学習の過程への介入）
    - 推論時：モデル抽出攻撃（AIモデルの出力の利用）
  - 扱わない例
    - ネットワーク機器の脆弱性を突いた不正アクセスによるAIモデルの窃取
      ※攻撃対象はAIだが、対象とするシステムが持つAIモデルの学習・推論への介入や入出力の利用・改ざんを伴う攻撃ではない

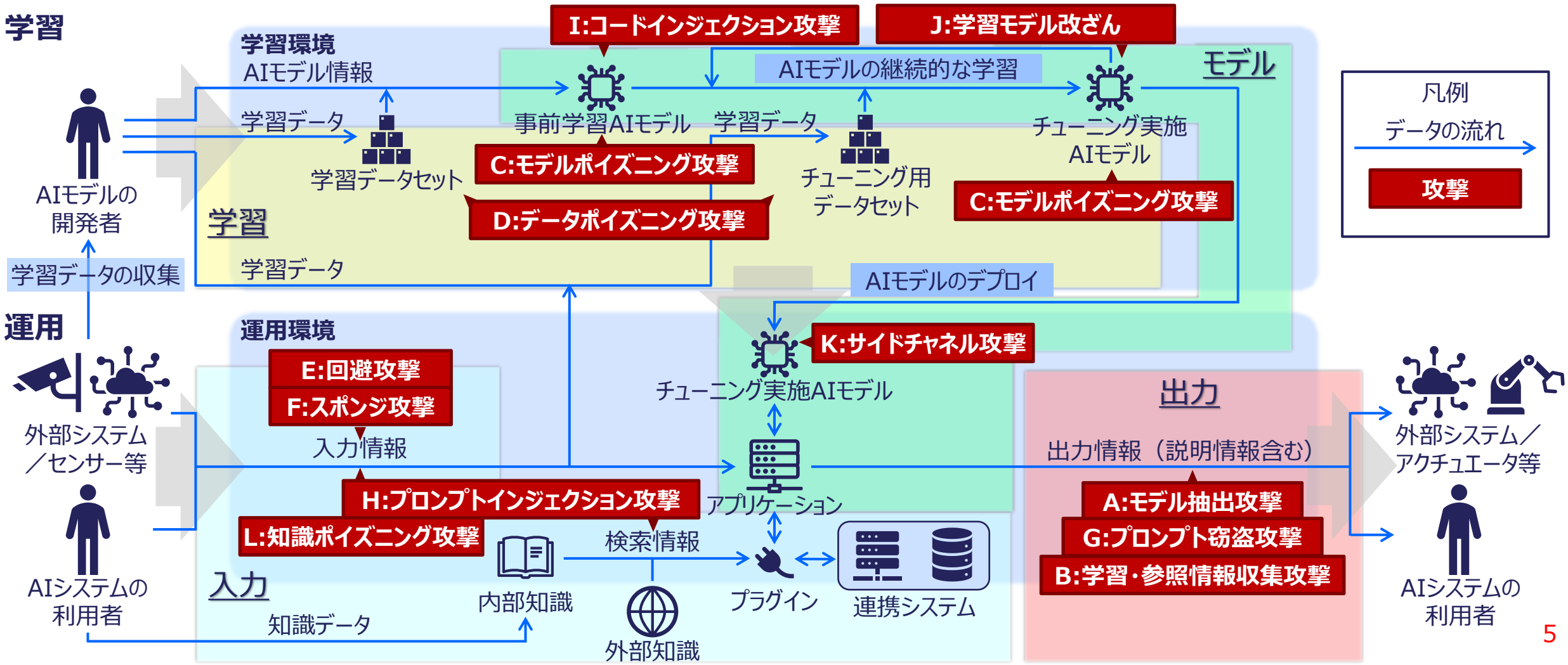
想定するAIシステム

- 学習と運用の２種類の環境で構成されるAIシステムを想定する。
- LLMと共によく使われるプラグイン等を明記したが、AIモデルはLLMに限らない。



AIシステムに対する攻撃の俯瞰

- 4つのアタックサーフェス(学習/モデル/入力/出力)の攻撃12種(A-L)に大別。
- 各攻撃には、さらに細かく分類できるものが含まれる。



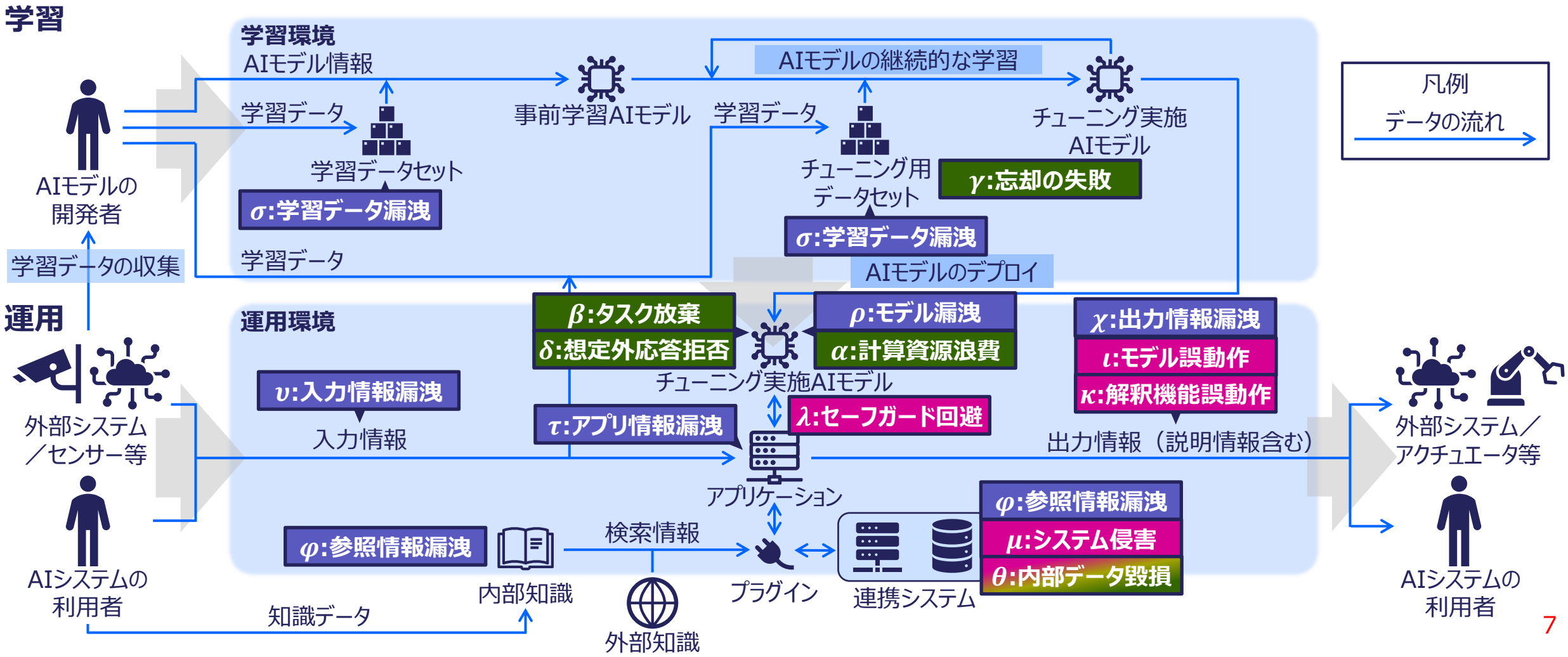
AIシステムに対する攻撃と影響（概観）

AISI Japan AI Safety Institute

- アタックサーフェス（攻撃の起点）ごとの攻撃の手法と、それぞれの攻撃でのシステムへの影響をセキュリティのCIAの観点で整理。

攻撃がAIシステムに与える影響の俯瞰

- 攻撃による影響を想定するAIシステムの構成に対応させて可視化。
- モデル漏洩のように同じ種類の影響が複数の箇所に生じる場合もある。



AIシステムに対する攻撃と影響（詳細１）

AISI Japan AI Safety Institute

表１：学習を主なアタックサーフェスとする攻撃と影響

攻撃	影響		攻撃者の能力（主な攻撃前提）	データ形式：対象AI[文献]（文献からの要約でありデータ形式や対象を限定するものではない）
C モデルポイズニング	C	σ:学習データ漏洩	学習プログラムの改変	画像：NN [SRS17] 文章：SVM, LR [SRS17]
	I	ι:モデル誤動作	悪性アダプタの配布	文章：LM [DXC+25]
			悪性モデルの配布	任意：NN [LSZ+25]
		κ:解釈機能誤動作	モデルの改変	表：NN, EM [SHJ+20]
	A	α:計算資源浪費	モデルの改変	画像：NN [CDB+23]
D データポイズニング	C	σ:学習データ漏洩	学習データの挿入・モデルへの大量のクエリ	表：LR, NN [MGC22] 文章：LR [MGC22]
	I	ι:モデル誤動作	学習データの挿入	画像：NN [Car21], SuperNet [OVA+25] 文章：LM [Sch19]
				表：Linear Regression [JOB+18], 文章から画像: Diffusion [NRB+25; LPC+25]
			学習データの挿入・マージ用悪性モデルの配布	画像：ViT [WWC+25], 文章：LM [WWC+25]
	A	γ:忘却の失敗	学習データの挿入	画像：NN [YZL+25]

表２：モデルを主なアタックサーフェスとする攻撃と影響

攻撃	影響		攻撃者の能力（主な攻撃前提）	データ形式：対象AI[文献]（文献からの要約でありデータ形式や対象を限定するものではない）
I コードインジェクション	I	μ:システム侵害	悪性モデルの配布	任意：任意 [ZWZ+24; ZCS+25]
J 学習モデル改ざん	C	σ:学習データ漏洩	ファインチューニング機能へのアクセス	文章：LM [CTZ+24]
	I	λ:セーフガード回避	モデルの内部情報へのアクセス	文章：LM [YYC+24; KDD25]
			アンラーニング機能へのアクセス	文章：LM [SKK+25]
			ファインチューニング機能へのアクセス	文章から画像：Diffusion [WYB+25]
K サイドチャネル	C	ρ:モデル漏洩	物理メモリへのアクセス	画像：NN [RCY+22]
			デバイスへの物理アクセス	画像：NN on GPU [HCW+25]
		σ:学習データ漏洩	物理メモリへのアクセス・モデルへの大量のクエリ	画像・動画：NN in TEE [YLD+25]
		υ:入力情報漏洩	物理メモリへのアクセス・モデルへの大量のクエリ	画像・動画：NN in TEE [YLD+25]
			共通MoEモデルライブラリの共有	文章・画像：MoE [DXS+25]
			CPUキャッシュへのアクセス	文章：LM [GHG+25]
			ハイパーバイザー制御へのアクセス	文章：LM [YHG+25]
		χ:出力情報漏洩	共通MoEモデルライブラリの共有	文章・画像：MoE [DXS+25]
			CPUキャッシュへのアクセス	文章：LM [GHG+25]
	I	ι:モデル誤動作	物理メモリへのアクセス	画像：NN [LWX+24; CLY+25]
			モデルの内部情報・物理メモリへのアクセス	画像・文章：LM, Transformer [NMF+24]
			GPUメモリへのアクセス	画像：NN on GPU [LQS25]

ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

AIシステムに対する攻撃と影響（詳細２）

AISI Japan AI Safety Institute

表３：入力を主なアタックサーフェスとする攻撃と影響

攻撃	影響		攻撃者の能力（主な攻撃前提）	データ形式：対象AI[文献]（文献からの要約でありデータ形式や対象を限定するものではない）
E 回避	I	ι:モデル誤動作	モデルへの大量のクエリ	画像：NN, LR, SVM, DT, kNN [PMG+17], Transformer [LHS+25], 文章：LM[BSA+22], 画像から文章：VLM [WBH+25]
			モデルの内部情報へのアクセス	画像：SVN [MMR+25], NN [SZS+14;GSS15;MMR+25], バイナリファイル：NN [NFI+25], 文章：LSTM [ERL+18], Transformer [MMR+25], 特定の文章類似度マッチングアルゴリズム [HRS25], 気象データ：Diffusion [IER25]
			モデルへの入力	文章：Transformer [GSJ+21], LSTM, DA [WFK+19], LiDAR点群：NN [KNT+25], 画像：NN [LLM+25;XDT+25], CV [XC25], マルチモーダル（画像・文章）：VLM [WZS+25], 画像から文章：VLM [XDT+25], 音声から文章：ASR [YZG+25]
			悪性入力の配布	文章：LM [LLW+25b]
			敵対的信号の送信	無線信号：JSCC [CSH+25]
		κ:解釈機能誤動作	モデルへの大量のクエリ	画像：NN[DAA+19]
	A	β:タスク放棄	モデルの内部情報へのアクセス	音声：Transformer [WLC+25]
F スポンジ	A	α:計算資源浪費	モデルへの入力	画像・文章：NN [SZB+21]
			モデルへの大量のクエリ	文章：LM [BSA+22]
H プロンプトインジェクション	C	σ:学習データ漏洩	モデルへの入力	文章：LM [KKC25]
		τ:アプリ情報漏洩	モデルへの入力	文章：LM[ZCI24]
		υ:入力情報漏洩	ユーザーが引用する情報の汚染	文章：特定のチャットサービス [Sam23]
		φ:参照情報漏洩	モデルへの入力	文章：LM [CBN25], 文章から表の操作：LM [PEC+25]
	I	ι:モデル誤動作	悪性画像の配布	マルチモーダル（画像・文章）：VLM [ZZM+25]
		λ:セーフガード回避	モデルへの入力	文章：LM [LDL+24;SCB+24;ZWC+23;WHS23;CRD+24;MZK+24;PHS+22;WFK+19;RSE25], 文章から画像：Diffusion [LMX+25;DMY+25], VLM [RSE25], 特定の画像生成システム [VMP25]
			モデルへの大量のクエリ	文章：LM [ZGY+25]
			モデルの内部情報へのアクセス	文章・画像：NN [CNC+23]
			ファインチューニング機能へのアクセス	文章：LM [LPH+25]
		μ:システム侵害	参照情報の挿入	文章・画像からの文章生成：LM [GAM+23]
	I/A	θ:内部データ毀損	モデルへの入力	文章から表の操作：LM [PEC+25]
	A	δ:想定外応答拒否	参照情報の挿入	文章：LM [SSS25]
L 知識ポイズニング	I	ι:モデル誤動作	参照情報の挿入	文章：LM[YSQ25;BS25;GCL+25;ZGW+25]
			参照情報の挿入・モデルへの大量のクエリ	文章：LM[CGL+25]

ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

AIシステムに対する攻撃と影響（詳細３）

AISI Japan AI Safety Institute

表４：出力を主なアタックサーフェスとする攻撃と影響

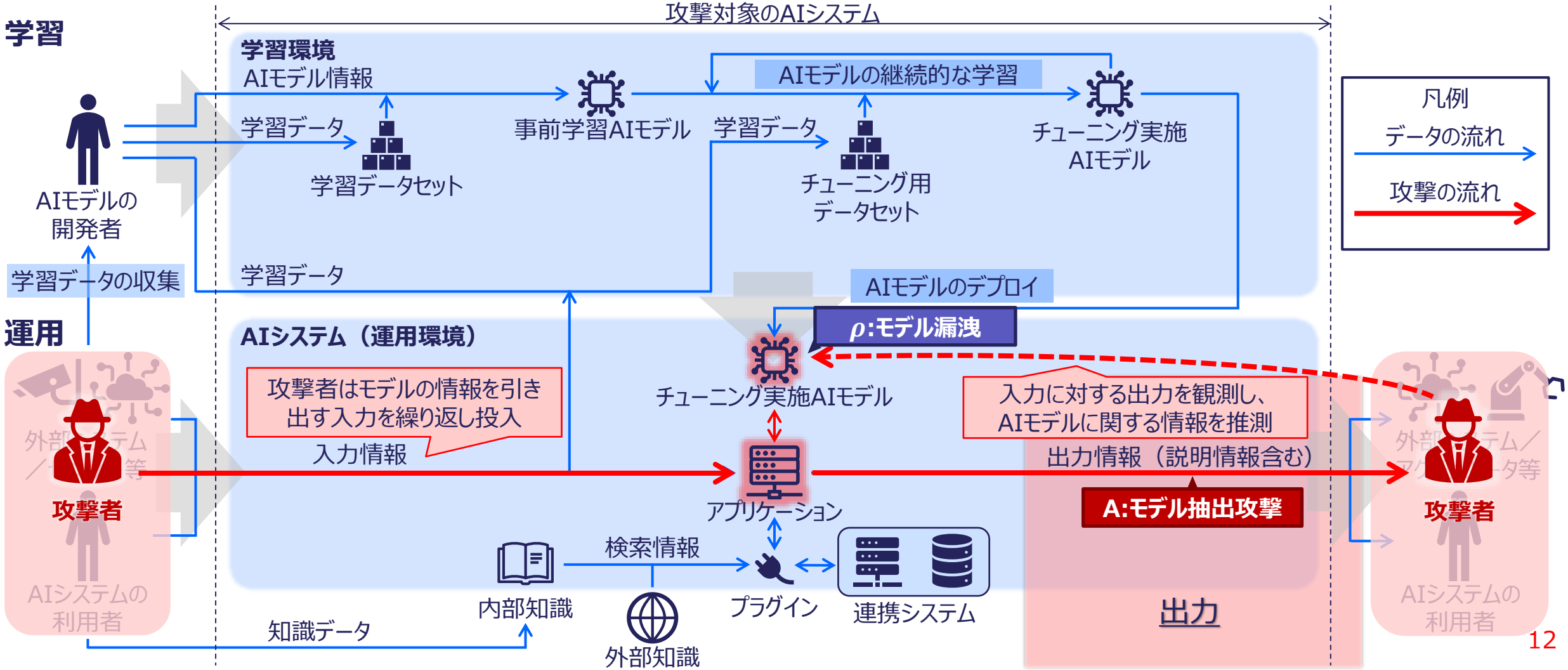
攻撃		影響		攻撃者の能力（主な攻撃前提）	データ形式：対象AI[文献]（文献からの要約でありデータ形式や対象を限定するものではない）
A モデル抽出		C	ρ:モデル漏洩	モデルへの大量のクエリ	画像：NN [OSF19a;OSF19b;JSM+19;CBB+18;YZW+25], LR[TZJ+16] 表：RR[WG18], DT[TZJ+16], LR, NN, SVM[TZJ+16,WG18], 文章：LSTM [YZW+25]
B 学習・参照情報収集	メンバーシップ推論	C	σ:学習データ漏洩	モデルへの入力	画像：RR, DT [YGF+18], NN [AYM+19;YGF+18;LLW+25a], 表：RR, DT [YGF+18], NN [AYM+19;YGF+18], MLP [LLW+25a], 文章：NN [AYM+19], LM [HLL+25a], MLP [LLW+25a], 画像：Diffusion [PW25]
				モデルへの大量のクエリ	画像：NN [SSS+17;LBW+18;YZW+25;NYW+25], ViT [NYW+25] 表：NN [SSS+17;LBW+18], 文章：LM [LSS+23;NYW+25], LSTM [YZW+25], マルチモーダル（画像･文章）：VLM [HLL+25b]
				モデルの内部情報へのアクセス	画像･表：NN [NSH19]
				モデルへの入力･悪性モデルの配布	画像：NN [WAA+25]
			φ:参照情報漏洩	モデルへの入力	文章：LM [NPS+25;GMD+25]
	属性推論	C	σ:学習データ漏洩	モデルへの入力	画像･表：LR, DT [YGF+18]
				モデルの内部情報へのアクセス	画像･文章：NN [SS20]
				モデルへの大量のクエリ	表：NN [KCM25]
	プロパティ推論	C	σ:学習データ漏洩	モデルの内部情報へのアクセス	画像･表：NN [GWY+18] 音声･通信：NN, SVM, HMM, DT[AMS+15]
	モデルインバージョン	C	σ:学習データ漏洩	モデルへの大量のクエリ	画像：DT [FJR15], NN[FJR15;YZW+25], 表：DT, NN [FJR15], 文章：LSTM [YZW+25]
				モデルの内部情報へのアクセス	文章：Transformer[ZHK22]
	データ再構築	C	σ:学習データ漏洩	モデルの内部情報へのアクセス	画像：NN [BCH22;HVY+22;BHY+23]
				モデルへの大量のクエリ	文章：LM [LSS+23]
	データ抽出	C	σ:学習データ漏洩	モデルへの大量のクエリ	文章：LSTM, RNN [CLE+19], LM [CTW+21;LSS+23;LWW+25], 文章から画像：Diffusion [CHN+23]
				モデルへの入力	文章：LM [CMX+25;KSM+25]
G プロンプト窃盗		C	τ:アプリ情報漏洩	モデルへの入力	文章：LM [YLL+25]
			υ:入力情報漏洩	モデルの出力へのアクセス	文章から画像：Diffusion [SQB+24]
				モデルへの入力	文章：LM [TSS+25]
				モデルへの大量のクエリ	文章：LM [YZH+25], 文章から画像：Diffusion [YZH+25]

ASR (Automatic Speech Recognition), CV (Computer Vision), DA (Decomposable Attention), DT (Decision Tree), EM (Ensemble Method), HMM (Hidden Markov Model), JSCC (Joint Source-channel Coding), kNN (Nearest Neighbor), LM (Language Model), LR (Logistic Regression), LSTM (Long Short-term Memory), MLP (Multilayer Perceptron), MoE (Mixture-of-Experts), NN (Neural Network), RNN (Recurrent Neural Network), RR (Ridge Regression), SVM (Support Vector Machine), ViT(Vision Transformer)

各攻撃の概要

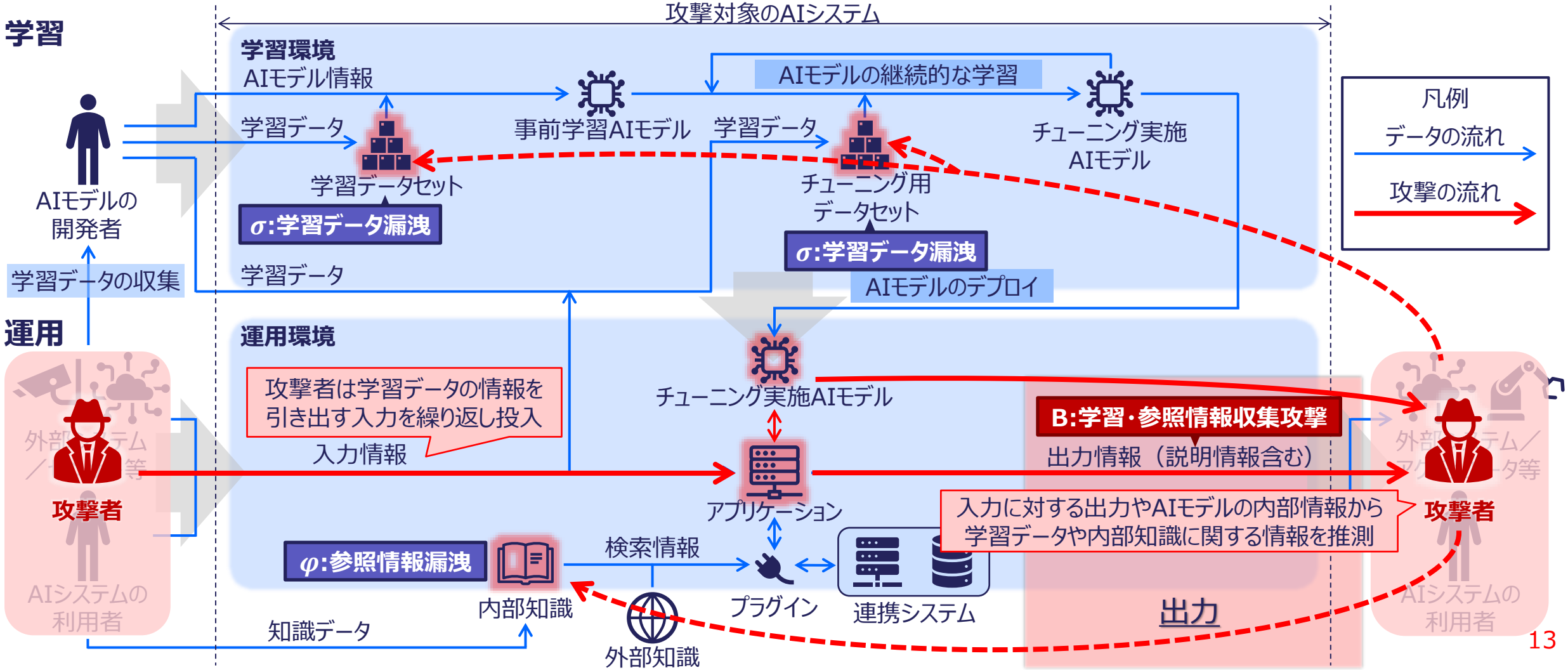
攻撃A:モデル抽出攻撃

- 入力情報に対する出力を観察することで、AIモデルに直接アクセスすることなく、AIモデルに関する情報の漏洩を引き起こす。



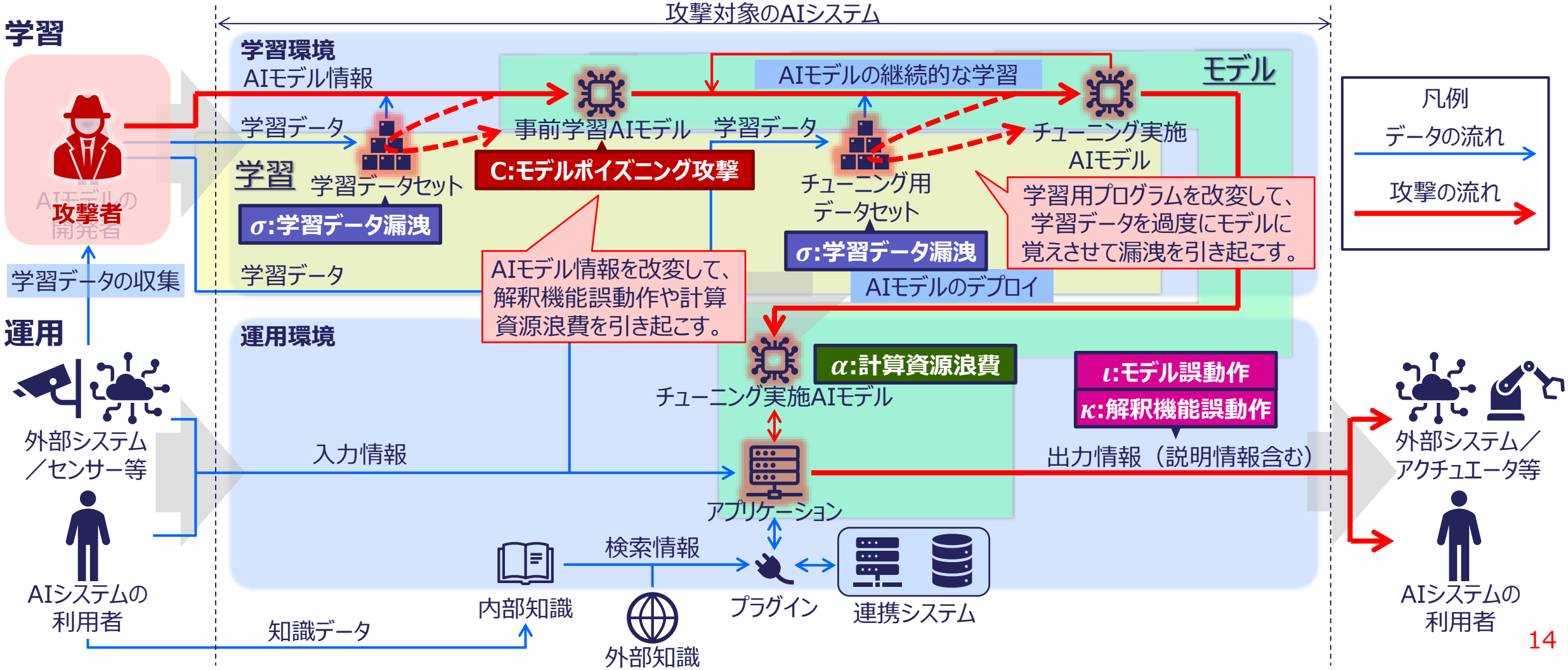
攻撃B:学習・参照情報収集攻撃

- 主に入出力情報から、学習データセットや内部知識に関する情報を収集する。複数の攻撃に細分化され、攻撃者が得る情報の性質が異なる。



攻撃C:モデルポイズニング攻撃

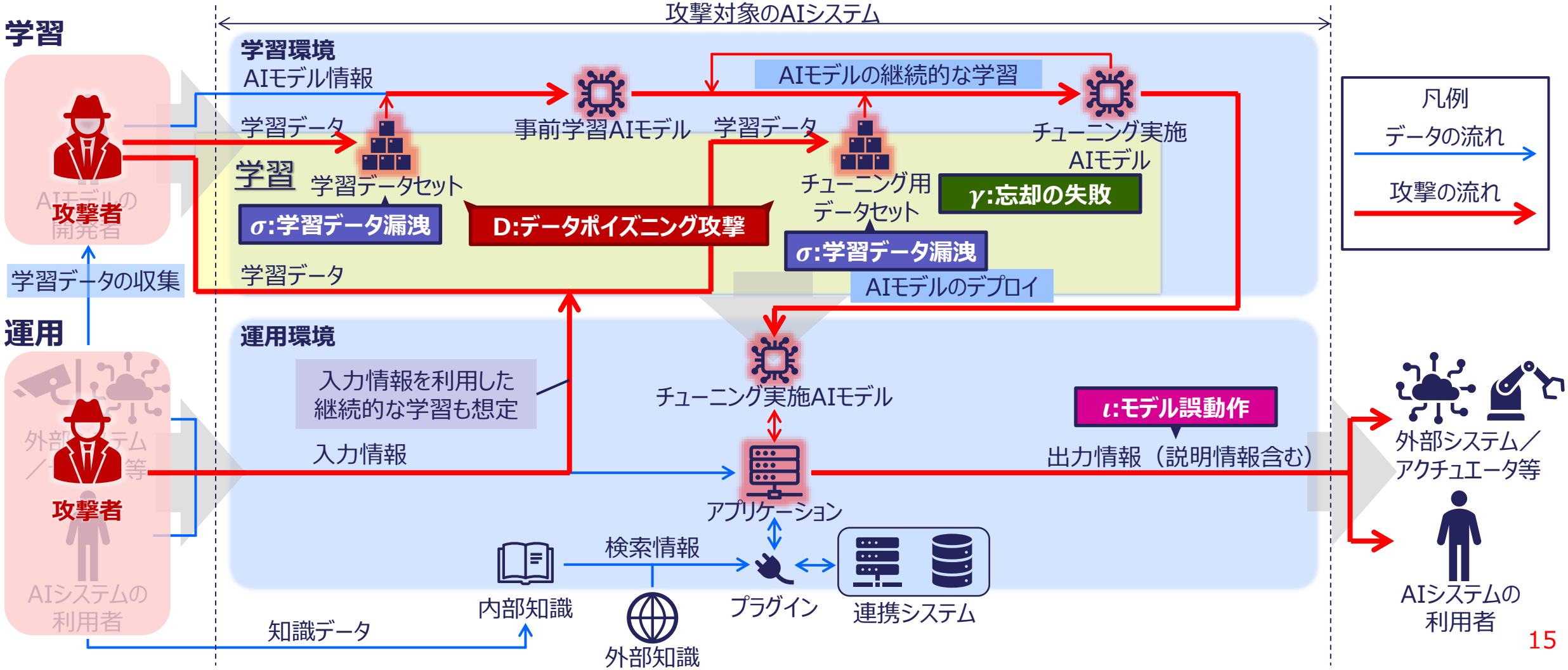
- AIモデルの情報や学習用プログラムを改変することで、AIモデルの運用時にモデル・解釈機能誤動作や計算資源浪費、学習データ漏洩を引き起こす。



攻撃D:データポイズニング攻撃

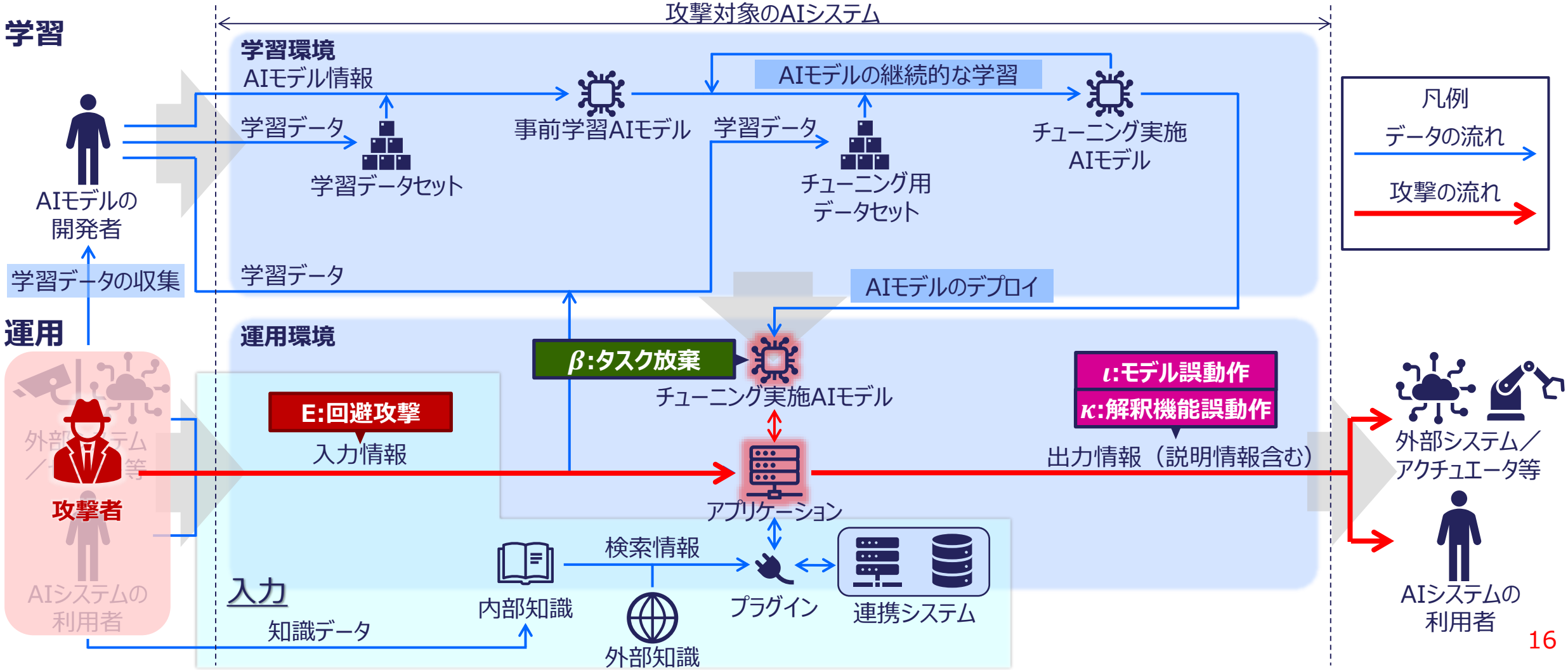
AISI Japan AI Safety Institute

- 学習データセットに特殊なデータを挿入することで、残りの学習データに関する情報の漏洩やモデル誤動作、忘却（学習データ削除）の失敗を引き起こす。



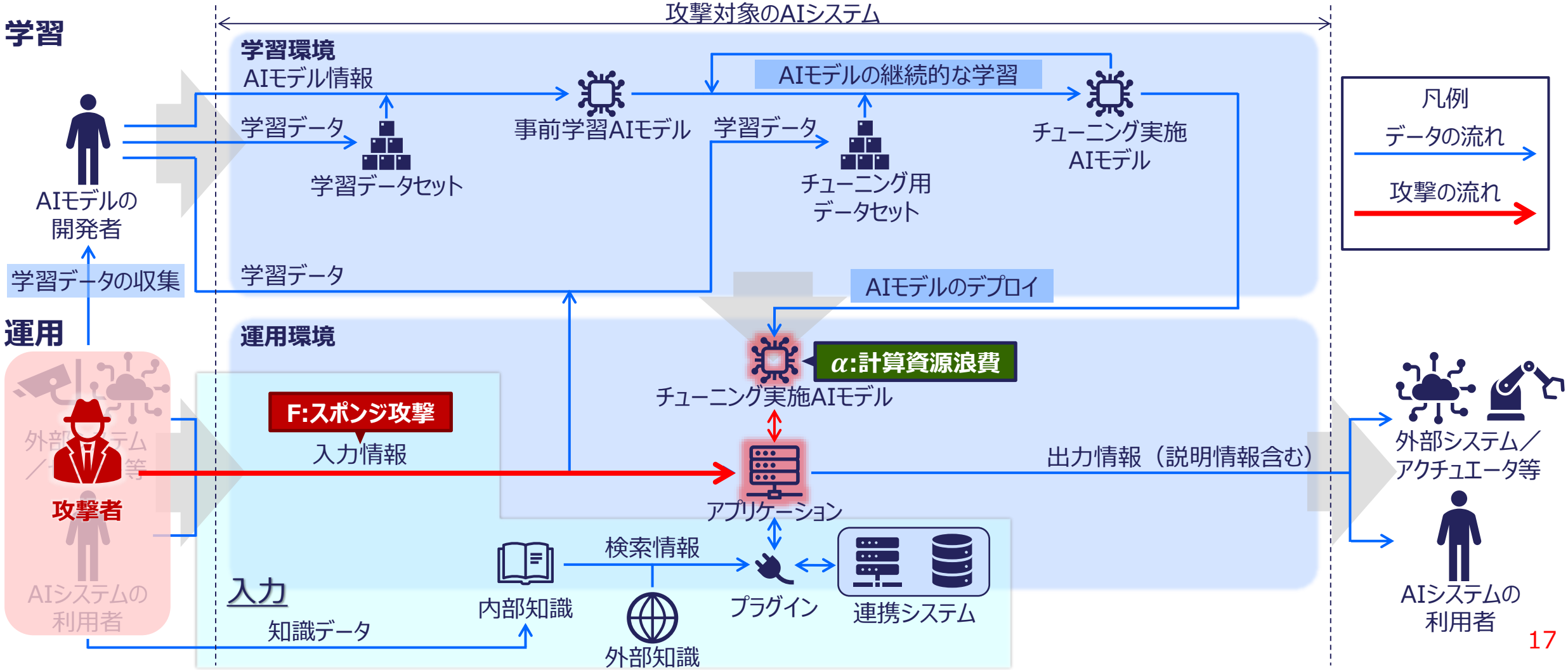
攻撃E:回避攻撃

- 運用時に敵対的サンプルと呼ばれる特殊な入力をAIシステムに与えることで、モデル・解釈機能誤動作やタスクの放棄を引き起こす。



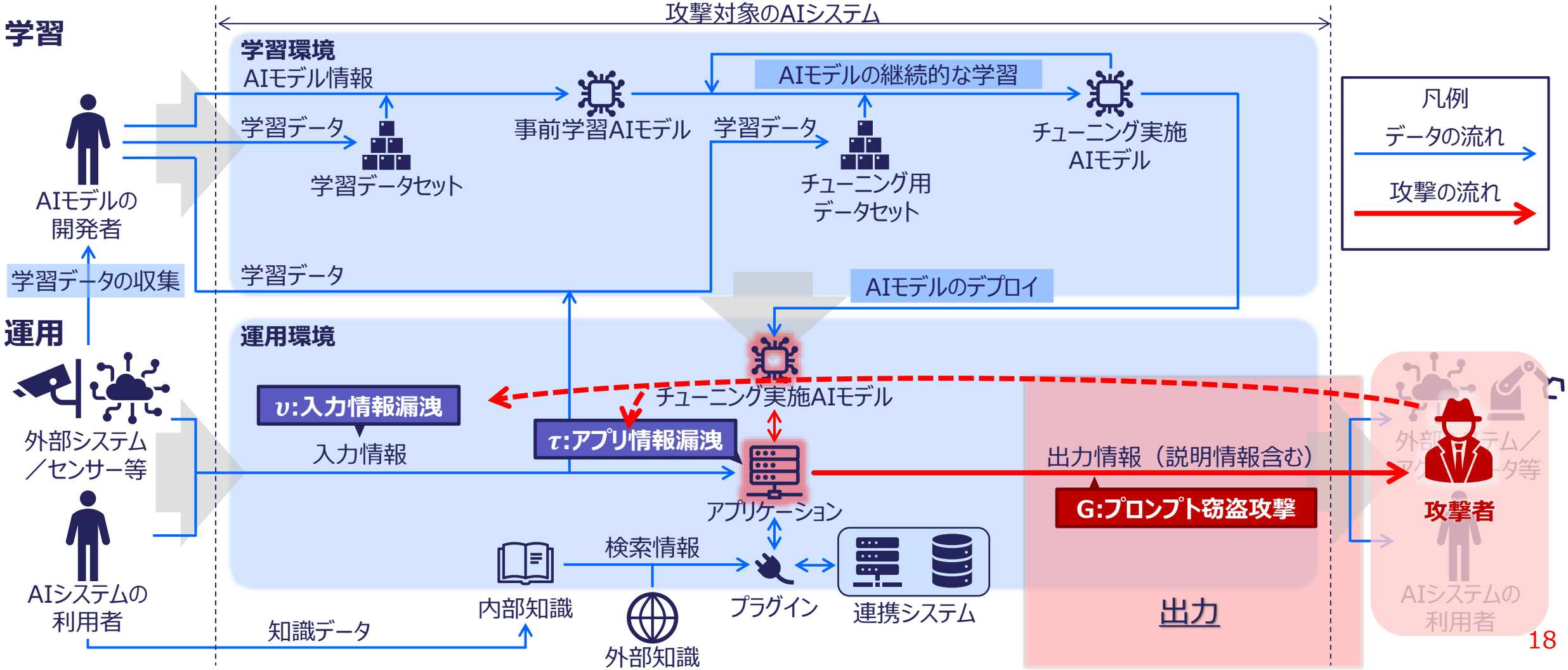
攻撃F:スポンジ攻撃

- 運用時にスポンジ・データと呼ばれる特殊な入力をAIシステムに与えることで、応答時間やエネルギー消費増大といった計算資源浪費を引き起こす。



攻撃G:プロンプト窃盗攻撃

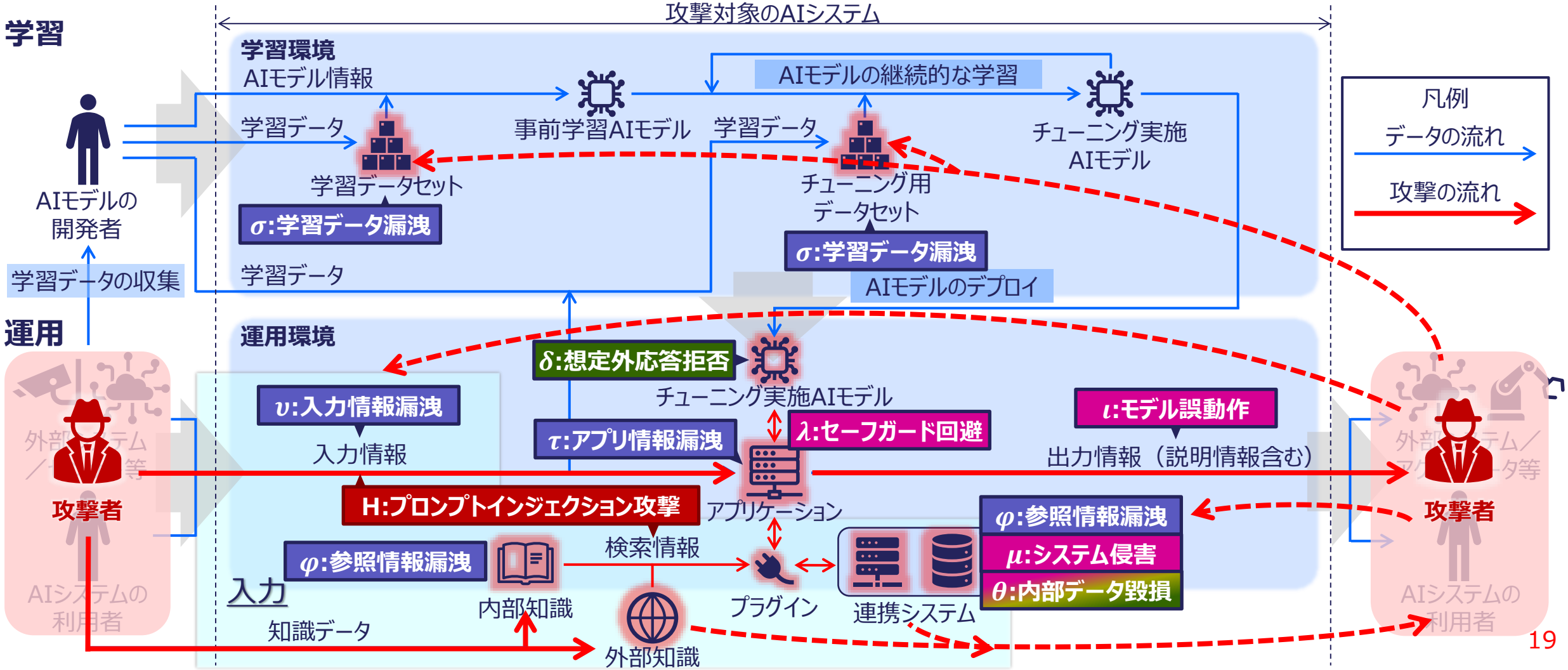
- AIモデルが生成した画像などの出力を解析することで、プロンプトエンジニアリングのノウハウとなる入力情報やアプリ情報（システムプロンプト）の漏洩を引き起こす。



攻撃H:プロンプトインジェクション攻撃

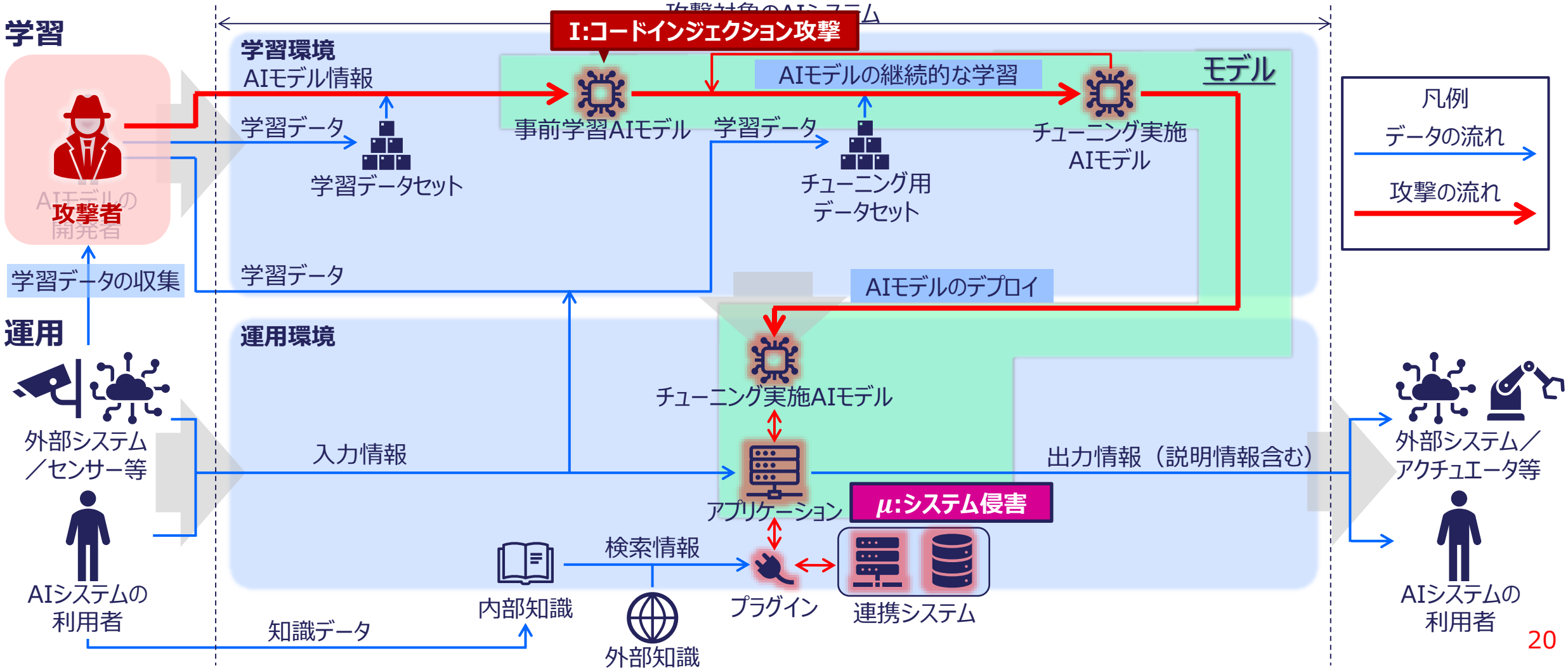
AISI Japan AI Safety Institute

- モデルに直接、または参照される情報源から間接的に敵対的な指示を入力することで、各種情報漏洩やセーフガード回避、システム侵害などを引き起こす。



攻撃I:コードインジェクション攻撃

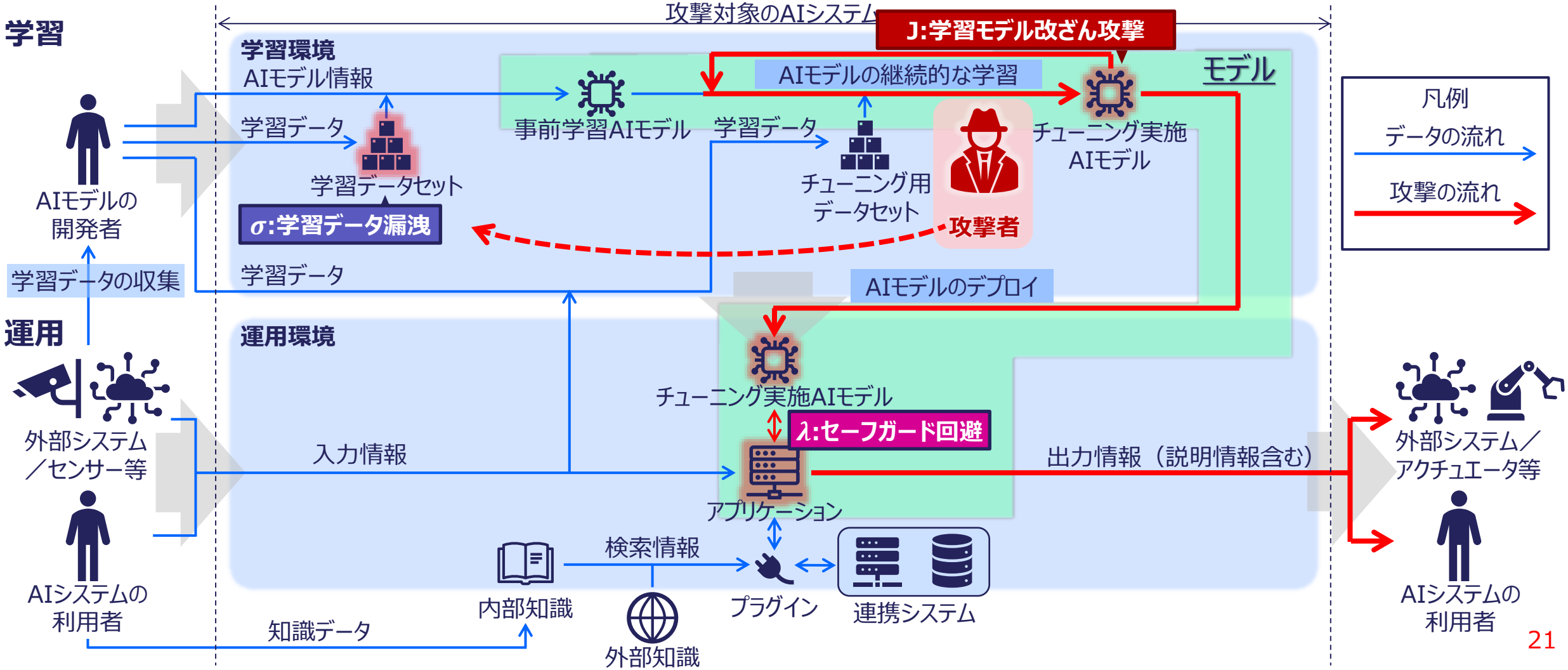
- AIモデル自体に実行可能なコードを埋め込み、AIモデルを呼び出した際に埋め込んだコードが実行されるようにしておくことで、システム侵害を引き起こす。



攻撃J:学習モデル改ざん攻撃

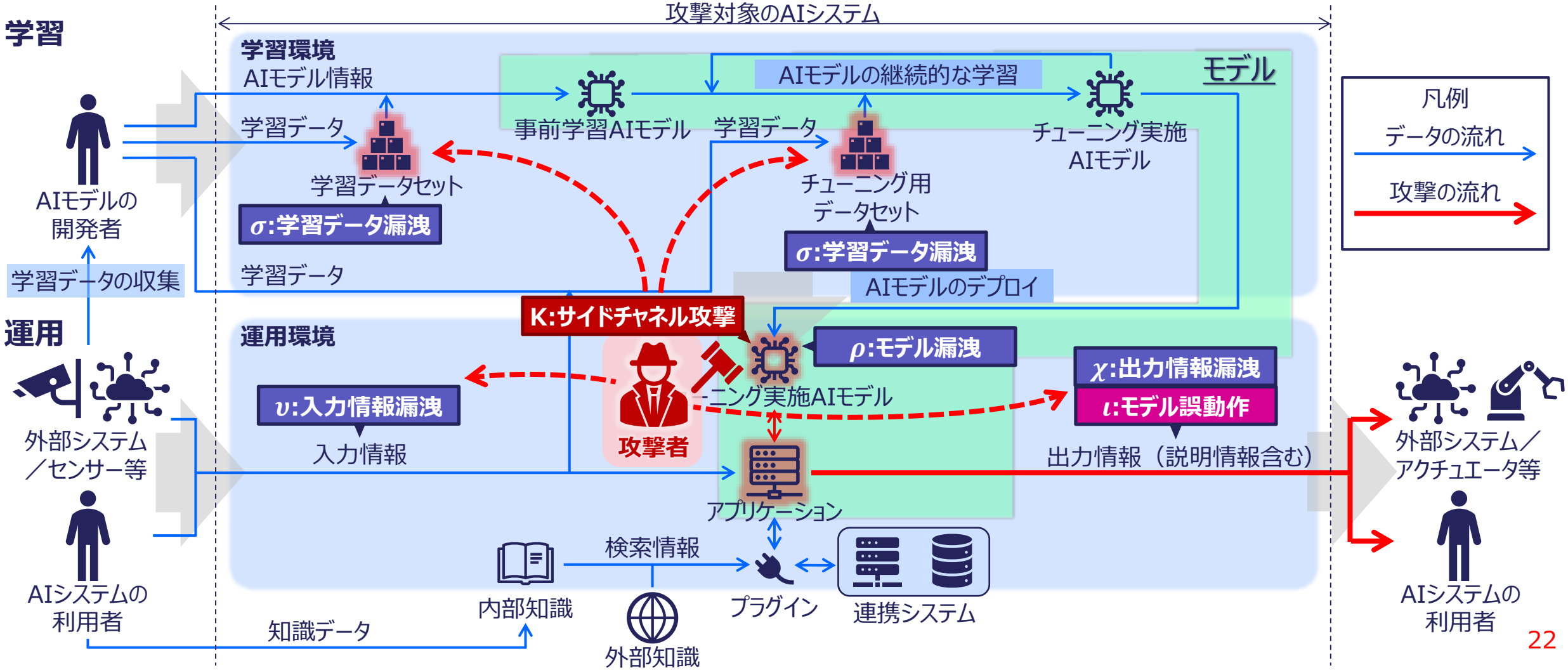
AISI Japan AI Safety Institute

- 事前学習AIモデルに、特定のファインチューニングやアンラーニングなどを実施することで、セーフガード回避や事前学習データを漏洩するAIモデルが作成される。



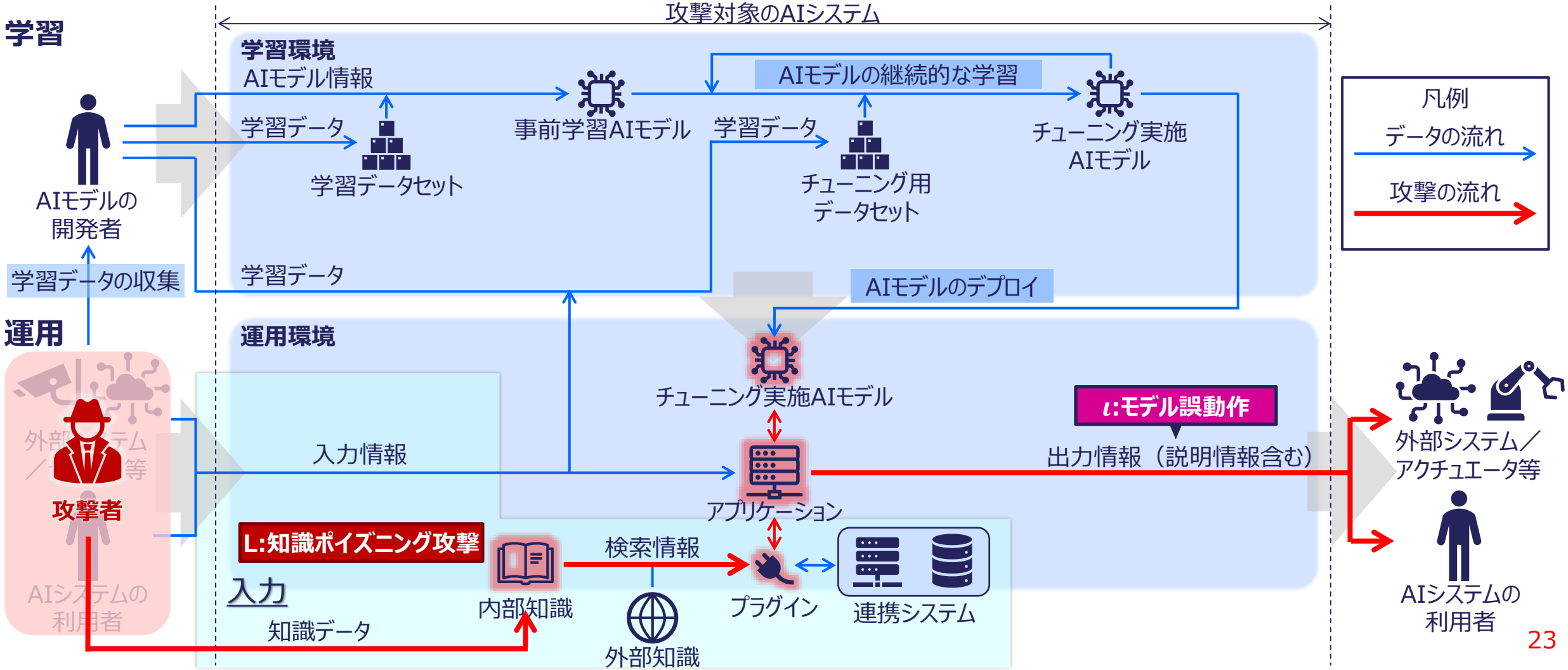
攻撃K:サイドチャネル攻撃

- クラウド環境やエッジデバイスなどの運用環境にアクセスする攻撃者が、物理的観測や干渉などを行うことで、各種情報漏洩やモデル誤動作を引き起こす。



攻撃L:知識ポイズニング攻撃

- 外部からRAGなどの内部知識に特殊なデータを混入させることで、AIモデルの誤動作を引き起こす。



更新履歴

第２版の更新概要（2026年 4月）

AISI Japan AI Safety Institute

- 第１版公開（2025年3月）からの時間経過に伴い、最新の攻撃手法を資料に反映すべく、2025年開催の主要国際会議※に採択された論文を調査し、AIシステムを狙った攻撃に関する主な論文（64件）を反映。

  ※対象の会議：USENIX Security、ACM CCS、IEEE S&P、NDSS
- 攻撃の追加・変更／アタックサーフェスでの整理（詳細は26ページ参照）
  - 調査結果を踏まえ、攻撃を２種類追加、３種類の名称を変更。
  - 攻撃を整理するアタックサーフェス（攻撃の起点）による分類を導入。
- 攻撃による影響の追加・変更（詳細は27ページ参照）
  - 調査結果を踏まえ、影響を４種類追加、１種類の名称を変更。
  - CIAごとに色分けして可視化。

攻撃の追加・変更／アタックサーフェスでの整理

AISI Japan AI Safety Institute

第１版（2025年 3月）

- A:モデル抽出
- B:学習データ情報収集
  - メンバーシップ推論
  - 属性推論
  - プロパティ推論
  - モデルインバージョン
  - データ抽出
- C:モデルポイズニング
- D:データポイズニング
- E:回避
- F:スポンジ
- G:プロンプト窃盗
- H:プロンプトインジェクション
- I:コードインジェクション
- J:ファインチューニング
- K:ロウハンマー

第２版（2026年 4月）

- A:モデル抽出
- B:学習・参照情報収集 **名称変更**
  - メンバーシップ推論
  - 属性推論
  - プロパティ推論
  - モデルインバージョン
  - データ再構築 **追加**
  - データ抽出
- C:モデルポイズニング
- D:データポイズニング
- E:回避
- F:スポンジ
- G:プロンプト窃盗
- H:プロンプトインジェクション
- I:コードインジェクション
- J:学習モデル改ざん **名称変更**
- K:サイドチャネル **名称変更**
- L:知識ポイズニング **追加**

アタックサーフェス（攻撃の起点）

学習：モデルの学習や更新
例：学習／チューニング用データセット、継続的な学習に使われる入力、モデル学習用のプログラム

モデル：学習済みモデルの内部
例：モデルのファイル、モデルの重み、モデルを読み込む機構、実行時のメモリ、モデル更新のAPI

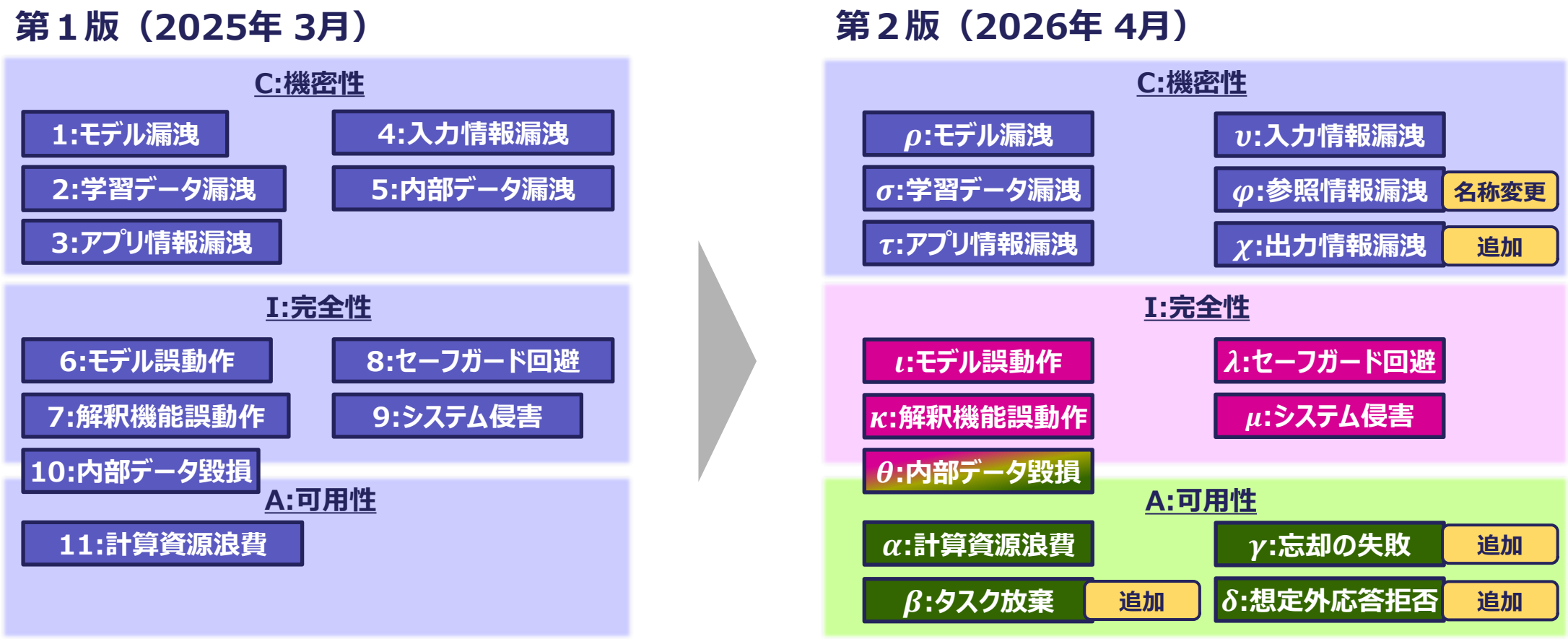
入力：モデル推論時の入力
例：モデルへの入力（利用者のプロンプト、外部システム／センサー等）、内部／外部知識

出力：モデル推論の出力
例：モデルの出力（出力ラベル、信頼度、生成された文章や画像等）

※上の矢印は主な関係を表現

攻撃による影響の追加・変更

- 影響の種類を追加し一部名称を変更。また、影響ごとのラベルを数字からギリシャ文字に変更した上で、CIAごとに色分けして可視化。



参考文献 (1/10)

AISI Japan AI Safety Institute

[AMS+15] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, et al. “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers”. In: Int. J. Secur. Netw. 10.3 (Sept. 2015), pp. 137–150. doi: 10.1504/IJSN.2015.071829.

[AYM+19] Salem Ahmed, Zhang Yang, Humbert Mathias, Berrang Pascal, Fritz Mario, et al. “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models”. In: Proceedings 2019 Network and Distributed System Security Symposium (2019). doi: 10.14722/ndss.2019.23119.

[BHY+23] Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, and Michal Irani. Reconstructing Training Data from Multiclass Neural Networks. 2023.

[BS25] Matan Ben-Tov and Mahmood Sharif. “GASLITEing the Retrieval: Exploring Vulnerabilities in Dense Embedding-based Search”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4364–4378. doi: 10.1145/3719027.3765095.

[BSA+22] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. “Bad Characters: Imperceptible NLP Attacks”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1987–2004. doi: 10.1109/SP46214.2022.9833641.

[Car21] Nicholas Carlini. “Poisoning the Unlabeled Dataset of Semi-Supervised Learning”. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 1577–1592. url: https: //www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

[CBB+18] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. “Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data”. In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489592.

[CBN25] Stav Cohen, Ron Bitton, and Ben Nassi. “Here Comes the AI Worm: Preventing the Propagation of Adversarial Self-Replicating Prompts Within GenAI Ecosystems”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3975–3989. doi: 10.1145/3719027.3765196.

[CDB+23] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-Latency Attacks via Sponge Poisoning. 2023. arXiv: 2203.08147 [cs.CR].

[CGL+25] Zhuo Chen, Yuyang Gong, Jiawei Liu, Miaokun Chen, Haotan Liu, et al. “FlippedRAG: Black-Box Opinion Manipulation Adversarial Attacks to Retrieval-Augmented Generation Models”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4109–4123. doi: 10.1145/3719027.3765023.

[CHN+23] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, et al. “Extracting Training Data from Diffusion Models”. In: 32nd USENIX Security Symposium (USENIX Security 23). Anaheim, CA: USENIX Association, Aug. 2023, pp. 5253–5270. url: https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks”. In: 28th USENIX Security Symposium (USENIX Security 19). Santa Clara, CA: USENIX Association, Aug. 2019, pp. 267–284. url: https: //www.usenix.org/conference/usenixsecurity19/presentation/carlini.

[CLY+25] Yanzuo Chen, Zhibo Liu, Yuanyuan Yuan, Sihang Hu, Tianxiang Li, et al. “Compiled Models, Built-In Exploits: Uncovering Pervasive Bit-Flip Attack Surfaces in DNN Executables”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, California, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/compiled-models-built-inexploits-uncovering-pervasive-bit-flip-attack-surfaces-in-dnn-executables/.

参考文献 (2/10)

AISI Japan AI Safety Institute

[CMX+25] Shuai Cheng, Shu Meng, Haitao Xu, Haoran Zhang, Shuai Hao, et al. “Effective PII Extraction from LLMs through Augmented Few-Shot Learning”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/cheng-shuai.

[CNC+23] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, et al. “Are aligned neural networks adversarially aligned?” In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 61478–61500. url: https://proceedings.neurips.cc/paper_files/paper/2023/hash/ c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

[CRD+24] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, et al. Jailbreaking Black Box Large Language Models in Twenty Queries. 2024. arXiv: 2310.08419 [cs.LG].

[CSH+25] Jung-Woo Chang, Ke Sun, Nasimeh Heydaribeni, Seira Hidano, Xinyu Zhang, et al. “Magmaw: Modality-Agnostic Adversarial Attacks on Machine Learning-Based Wireless Communication Systems”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/magmaw-modalityagnostic-adversarial-attacks-on-machine-learning-based-wireless-communication-systems/.

[CTW+21] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, et al. “Extracting Training Data from Large Language Models”. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 2633–2650. url: https://www.usenix.org/conference/ usenixsecurity21/presentation/carlini-extracting.

[CTZ+24] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, et al. “The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks”. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS ’24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 1285–1299. doi: 10.1145/3658644.3690325.

[DAA+19] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, et al. “Explanations can be manipulated and geometry is to blame”. In: Advances in Neural Information Processing Systems. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, et al. Vol. 32. Curran Associates, Inc., 2019. url: https://proceedings.neurips.cc/paper/ 2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html.

[DMY+25] Yingkai Dong, Xiangtao Meng, Ning Yu, Zheng Li, and Shanqing Guo. “Fuzz-Testing Meets LLM-Based Agents: An Automated and Efficient Framework for Jailbreaking Text-to-Image Generation Models”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 373–391. doi: 10.1109/SP61157.2025.00119.

[DXC+25] Tian Dong, Minhui Xue, Guoxing Chen, Rayne Holland, Yan Meng, et al. “The Philosopher’s Stone: Trojaning Plugins of Large Language Models”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. Reston, VA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/the-philosophers-stone-trojaning-plugins-of-large-languagemodels/.

[DXS+25] Ruyi Ding, Tianhong Xu, Xinyi Shen, Aidong Adam Ding, and Yunsi Fei. “MoEcho: Exploiting Side-Channel Attacks to Compromise User Privacy in Mixture-of-Experts LLMs”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2159–2173. doi: 10.1145/3719027.3765174.

[ERL+18] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. “HotFlip: White-Box Adversarial Examples for Text Classification”. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia, July 2018, pp. 31–36. doi: 10.18653/v1/P18-2006.

参考文献 (3/10)

AISI Japan AI Safety Institute

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS ’15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. doi: 10.1145/2810103.2813677.

[GAM+23] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, et al. “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection”. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. AISec ’23. Copenhagen, Denmark: Association for Computing Machinery, 2023, pp. 79–90. doi: 10.1145/3605764. 3623985.

[GCL+25] Yuyang Gong, Zhuo Chen, Jiawei Liu, Miaokun Chen, Fengchang Yu, et al. “Topic-FlipRAG: Topic-Orientated Adversarial Opinion Manipulation Attacks to Retrieval-Augmented Generation Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/gong-yuyang.

[GHG+25] Zibo Gao, Junjie Hu, Feng Guo, Yixin Zhang, Yinglong Han, et al. “I Know What You Said: Unveiling Hardware Cache Side-Channels in Local Large Language Model Inference”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/gao-zibo.

[GMD+25] Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. “DCMI: A Differential Calibration Membership Inference Attack Against Retrieval-Augmented Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4184–4198. doi: 10.1145/3719027.3765103.

[GSJ+21] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. “Gradient-based Adversarial Attacks against Text Transformers”. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 5747–5757. doi: 10.18653/v1/2021. emnlp-main.464.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2015. arXiv: 1412.6572 [stat.ML].

[GWY+18] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. “Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations”. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. CCS ’18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 619–633. doi: 10.1145/3243734.3243834.

[HCW+25] Peter Horvath, Lukasz Chmielewski, L´eo Weissbart, Lejla Batina, and Yuval Yarom. “BarraCUDA: Edge GPUs do Leak DNN Weights”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/horvath.

[HLL+25a] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, et al. “Towards Label-Only Membership Inference Attack against Pre-trained Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/he-yu.

[HLL+25b] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, et al. “Membership Inference Attacks Against Vision-Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/hu-yuke.

参考文献 (4/10)

AISI Japan AI Safety Institute

[HRS25] Jhih-Yi (Janet) Hsieh, Aditi Raghunathan, and Nihar B. Shah. “Vulnerability of Text-Matching in ML/AI Conference Reviewer Assignments to Collusions”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/hsieh.

[HVY+22] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. “Reconstructing Training Data from Trained Neural Networks”. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., 2022. url: https://proceedings.neurips.cc/paper_files/paper/2022/file/906927370cbeb537781100623cca6fa6-Paper-Conference.pdf.

[IER25] Erik Imgrund, Thorsten Eisenhofer, and Konrad Rieck. “Adversarial Observations in Weather Forecasting”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3579–3590. doi: 10.1145/3719027.3765076.

[JOB+18] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, et al. “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning”. In: 2018 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2018, pp. 19–35. doi: 10.1109/SP.2018.00057.

[JSM+19] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. “PRADA: Protecting Against DNN Model Stealing Attacks”. In: 2019 IEEE European Symposium on Security and Privacy (EuroS&P). 2019, pp. 512–527. doi: 10.1109/EuroSP.2019.00044.

[KCM25] Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz. “Disparate Privacy Vulnerability: Targeted Attribute Inference Attacks and Defenses”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/kabir.

[KDD25] Torsten Krauß, Hamid Dashtbani, and Alexandra Dmitrienko. “TwinBreak: Jailbreaking LLM Security Alignments based on Twin Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/krauss.

[KKC25] Yujin Kang, Eunsun Kim, and Yoon-Sik Cho. “Can Personal Health Information Be Secured in LLM? Privacy Attack and Defense in the Medical Domain”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4199–4213. doi: 10.1145/3719027.3765105.

[KNT+25] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. “Invisible but Detected: Physical Adversarial Shadow Attack and Defense on LiDAR Object Detection”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/kobayashi.

[KSM+25] Seongho Keum, Dongwon Shin, Leo Marchyok, Sanghyun Hong, and Sooel Son. “Private Investigator: Extracting Personally Identifiable Information from Large Language Models Using Optimized Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/keum.

[LBW+18] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, et al. Understanding Membership Inferences on Well-Generalized Learning Models. 2018. arXiv: 1802.04889 [cs.CR].

[LDL+24] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, et al. Prompt Injection attack against LLM-integrated Applications. 2024. arXiv: 2306.05499 [cs.CR].

参考文献 (5/10)

AISI Japan AI Safety Institute

[LHS+25] Yang Lou, Haibo Hu, Qun Song, Qian Xu, Yi Zhu, et al. “Asymmetry Vulnerability and Physical Attacks on Online Map Construction for Autonomous Driving”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3251–3265. doi: 10.1145/3719027.3765092.

[LLM+25] Taifeng Liu, Yang Liu, Zhuo Ma, Tong Yang, Xinjing Liu, et al. “L-HAWK: A Controllable Physical Adversarial Patch Against a Long-Distance Target”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/l-hawk-a-controllable-physical-adversarial-patchagainst-a-long-distance-target/.

[LLW+25a] Hao Li, Zheng Li, Siyuan Wu, Yutong Ye, Min Zhang, et al. “Enhanced Label-Only Membership Inference Attacks with Fewer Queries”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/li-hao.

[LLW+25b] Xiao Li, Yue Li, Hao Wu, Yue Zhang, Kaidi Xu, et al. “Make a Feint to the East While Attacking in the West: Blinding LLM-Based Code Auditors with Flashboom Attacks”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 576–594. doi: 10.1109/SP61157.2025.00125.

[LMX+25] Shuofeng Liu, Mengyao Ma, Minhui Xue, and Guangdong Bai. “Modifier Unlocked: Jailbreaking Text-to-Image Models Through Prompts”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 355–372. doi: 10.1109/SP61157.2025.00242.

[LPC+25] Changjiang Li, Ren Pang, Bochuan Cao, Jinghui Chen, Fenglong Ma, et al. “Watch the Watchers! On the Security Risks of Robustness-Enhancing Diffusion Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/li-changjiang.

[LPH+25] Andrey Labunets, Nishit V. Pandya, Ashish Hooda, Xiaohan Fu, and Earlence Fernandes. “Fun-tuning: Characterizing the Vulnerability of Proprietary LLMs to Optimization-Based Prompt Injection Attacks via the Fine-Tuning Interface”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 411–429. doi: 10.1109/SP61157.2025.00121.

[LQS25] Chris S. Lin, Joyce Qu, and Gururaj Saileshwar. “GPUHammer: Rowhammer Attacks on GPU Memories are Practical”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/lin-shaopeng.

[LSS+23] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, et al. “Analyzing Leakage of Personally Identifiable Information in Language Models”. In: 2023 IEEE Symposium on Security and Privacy (SP). 2023, pp. 346–363. doi: 10.1109/SP46215.2023.10179300.

[LSZ+25] Harry Langford, Ilia Shumailov, Yiren Zhao, Robert Mullins, and Nicolas Papernot. “Architectural Neural Backdoors from First Principles”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 1657–1675. doi: 10.1109/SP61157.2025.00060.

[LWW+25] Zongjie Li, Daoyuan Wu, Shuai Wang, and Zhendong Su. “Differentiation-Based Extraction of Proprietary Data from Fine-Tuned LLMs”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3071–3085. doi: 10.1145/3719027.3744856.

[LWX+24] Shaofeng Li, Xinyu Wang, Minhui Xue, Haojin Zhu, Zhi Zhang, et al. “Yes, One-Bit-Flip Matters! Universal DNN Model Inference Depletion with Runtime Code Fault Injection”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1315–1330. url: https://www.usenix.org/conference/usenixsecurity24/presentation/li-shaofeng.

参考文献 (6/10)

AISI Japan AI Safety Institute

[MGC22] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. “Property Inference from Poisoning”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1120–1137. doi: 10.1109/SP46214.2022.9833623.

[MMR+25] Kaleel Mahmood, Caleb Manicke, Ethan Rathbun, Aayushi Verma, Sohaib Ahmad, et al. “Busting the Paper Ballot: Voting Meets Adversarial Machine Learning”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3132–3146. doi: 10.1145/3719027.3744882.

[MZK+24] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, et al. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. 2024. arXiv: 2312.02119 [cs.LG]

[NFI+25] Milad Nasr, Yanick Fratantonio, Luca Invernizzi, Ange Albertini, Loua Farah, et al. “Evaluating the Robustness of a Production Malware Detection System to Transferable Adversarial Attacks”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 4394–4408. doi: 10.1145/3719027.3765118.

[NMF+24] Najmeh Nazari, Hosein Mohammadi Makrani, Chongzhou Fang, Hossein Sayadi, Setareh Rafatirad, et al. “Forget and Rewire: Enhancing the Resilience of Transformer-based Models against Bit-Flip Attacks”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1349–1366. url: https://www.usenix.org/conference/usenixsecurity24/presentation/nazari.

[NPS+25] Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, et al. “Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 1245–1259. doi: 10.1145/3719027.3744840.

[NRB+25] Ali Naseh, Jaechul Roh, Eugene Bagdasarian, and Amir Houmansadr. “Backdooring Bias (B2) into Stable Diffusion Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/naseh.

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning”. In: 2019 IEEE Symposium on Security and Privacy (SP). 2019, pp. 739–753. doi: 10.1109/SP.2019.00065.

[NYW+25] Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, et al. “Rectifying Privacy and Efficacy Measurements in Machine Unlearning: A New Inference Attack Perspective”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/naderloui.

[OSF19a] Seong Joon Oh, Bernt Schiele, and Mario Fritz. “Towards Reverse-Engineering Black-Box Neural Networks”. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Cham: Springer International Publishing, 2019, pp. 121–144. doi: 10.1007/978-3-030-28954-6_7.

[OSF19b] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Knockoff Nets: Stealing Functionality of BlackBox Models”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, pp. 4949–4958. doi: 10.1109/CVPR.2019.00509.

[OVA+25] David Oygenblik, Abhinav Vemulapalli, Animesh Agrawal, Debopam Sanyal, Alexey Tumanov, et al. “VillainNet: Targeted Poisoning Attacks Against SuperNets Along the Accuracy-Latency Pareto Frontier”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2189–2203. doi: 10.1145/3719027.3765185.

[PEC+25] Rodrigo Pedro, Miguel E. Coimbra, Daniel Castro, Paulo Carreira, and Nuno Santos. “Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses”. In: 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 76–88. doi: 10.1109/ICSE55347.2025.00007.

参考文献 (7/10)

AISI Japan AI Safety Institute

[PHS+22] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, et al. “Red Teaming Language Models with Language Models”. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, Dec. 2022, pp. 3419–3448. doi: 10.18653/v1/2022.emnlp-main.225.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, et al. “Practical Black-Box Attacks against Machine Learning”. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ASIA CCS ’17. Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2017, pp. 506–519. doi: 10.1145/3052973.3053009.

[PW25] Yan Pang and Tianhao Wang. “Black-box Membership Inference Attacks against Fine-tuned Diffusion Models”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025. San Diego, CA, USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/black-box-membership-inference-attacks-against-fine-tuned-diffusion-models/.

[RCY+22] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. “DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1157–1174. doi: 10.1109/SP46214.2022.9833743.

[RSE25] Mark Russinovich, Ahmed Salem, and Ronen Eldan. “Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich.

[Sam23] Roman Samoilenko. “New Prompt Injection Attack on ChatGPT Web Version. Markdown Images can Steal Your Chat Data.” In: System Weakness (Mar. 2023). url: https://systemweakness.com/newprompt-injection-attack-on-chatgpt-web-version-ef717492c5c2.

[SCB+24] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. “"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models”. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS ’24. Salt Lake City, UT, USA, 2024, pp. 1671–1685. doi: 10.1145/3658644.3670388.

[Sch19] Oscar Schwartz. “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation”. In: IEEE Spectrum (Nov. 2019). Updated: Jan. 2024. url: https://spectrum.ieee.org/in- 2016- microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[SHJ+20] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods”. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 180–186. doi: 10.1145/3375627.3375830.

[SKK+25] Minkyoo Song, Hanna Kim, Jaehan Kim, Seungwon Shin, and Sooel Son. “Refusal Is Not an Option: Unlearning Safety Alignment of Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/song-minkyoo.

[SQB+24] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. “Prompt Stealing Attacks Against Text-to-Image Generation Models”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 5823–5840. url: https://www.usenix.org/conference/usenixsecurity24/presentation/shen-xinyue.

[SRS17] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Machine Learning Models that Remember Too Much”. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. CCS ’17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 587–601. doi: 10.1145/3133956.3134077.

[SS20] Congzheng Song and Vitaly Shmatikov. “Overlearning Reveals Sensitive Attributes”. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. url: https://openreview.net/forum?id=SJeNz04tDS.

参考文献 (8/10)

[SSS+17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks Against Machine Learning Models”. In: 2017 IEEE Symposium on Security and Privacy (SP). 2017, pp. 3–18. doi: 10.1109/SP.2017.41.

[SSS25] Avital Shafran, Roei Schuster, and Vitaly Shmatikov. “Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/shafran.

[SZB+21] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, et al. “Sponge Examples: Energy-Latency Attacks on Neural Networks”. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). 2021, pp. 212–231. doi: 10.1109/EuroSP51992.2021.00024.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, et al. “Intriguing properties of neural networks”. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2014. arXiv: 1312.6199 [cs.CV].

[TSS+25] Yicong Tan, Xinyue Shen, Yun Shen, Michael Backes, and Yang Zhang. “On the Effectiveness of Prompt Stealing Attacks on In-the-Wild Prompts”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 392–410. doi: 10.1109/SP61157.2025.00120.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. “Stealing Machine Learning Models via Prediction APIs”. In: 25th USENIX Security Symposium (USENIX Security 16). Austin, TX: USENIX Association, Aug. 2016, pp. 601–618. url: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer.

[VMP25] Corban Villa, Shujaat Mirza, and Christina Popper. “Exposing the Guardrails: Reverse-Engineering and Jailbreaking Safety Filters in DALL·E Text-to-Image Pipelines”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/villa.

[WBH+25] Stanley Wu, Ronik Bhaskar, Anna Yoo Jeong Ha, Shawn Shan, Haitao Zheng, et al. “On the Feasibility of Poisoning Text-to-Image AI Models via Adversarial Mislabeling”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 2848–2862. doi: 10.1145/3719027.3744845.

[WFK+19] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. “Universal Adversarial Triggers for Attacking and Analyzing NLP”. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China, Nov. 2019, pp. 2153–2162. doi: 10.18653/v1/D19-1221.

[WG18] Binghui Wang and Neil Zhenqiang Gong. “Stealing Hyperparameters in Machine Learning”. In: 2018 IEEE Symposium on Security and Privacy (SP). 2018, pp. 36–52. doi: 10.1109/SP.2018.00038.

[WHS23] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. “Jailbroken: How Does LLM Safety Training Fail?” In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 80079–80110. url: https : //proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790- Abstract-Conference.html.

[WLC+25] Haolin Wu, Chang Liu, Jing Chen, Ruiying Du, Kun He, et al. “When Translators Refuse to Translate: A Novel Attack to Speech Translation Systems”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wu-haolin.

参考文献 (9/10)

[WWC+25] Lijin Wang, Jingjing Wang, Tianshuo Cong, Xinlei He, Zhan Qin, et al. “From Purity to Peril: Backdooring Merged Models From “Harmless” Benign Components”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wang-lijin.

[WYB+25] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. “On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wu-yixin-generation.

[WZS+25] Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, et al. “Mirage in the Eyes: Hallucination Attack on Multi-modal Large Language Models with Only Attention Sink”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/wang-yining.

[WZZ+25] Zihao Wang, Rui Zhu, Zhikun Zhang, Haixu Tang, and XiaoFeng Wang. “Rigging the Foundation: Manipulating Pre-training for Advanced Membership Inference Attacks”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 2509–2526. doi: 10.1109/SP61157.2025.00177.

[XC25] Qi Xia and Qian Chen. “AlphaDog: No-Box Camouflage Attacks via Alpha Channel Oversight”. In: 32nd Annual Network and Distributed System Security Symposium, NDSS 2025 February 24-28, 2025. San Diego, CA USA: The Internet Society, Feb. 2025. url: https://www.ndss-symposium.org/ndss-paper/alphadog-no-box-camouflage-attacks-via-alpha-channel-oversight/.

[XDT+25] Binyan Xu, Xilin Dai, Di Tang, and Kehuan Zhang. “One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 3087–3101. doi: 10.1145/3719027.3744859.

[YGF+18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting”. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 2018, pp. 268–282. doi: 10.1109/CSF.2018.00027.

[YHG+25] Yuqin Yan, Wei Huang, Ilya Grishchenko, Gururaj Saileshwar, Aastha Mehta, et al. “Relocate-Vote: Using Sparsity Information to Exploit Ciphertext Side-Channels”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/yan-yuqin.

[YLD+25] Yuanyuan Yuan, Zhibo Liu, Sen Deng, Yanzuo Chen, Shuai Wang, et al. “CipherSteal: Stealing Input Data from TEE-Shielded Neural Networks with Ciphertext Side Channels”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 4136–4154. doi: 10.1109/SP61157.2025.00079.

[YLL+25] Yong Yang, Changjiang Li, Qingming Li, Oubo Ma, Haoyu Wang, et al. “PRSA: Prompt Stealing Attacks against Real-World Prompt Services”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/yang-yong.

[YSQ25] Kai Ye, Liangcai Su, and Chenxiong Qian. “ImportSnare: Directed ’Code Manual’ Hijacking in Retrieval-Augmented Code Generation”. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. CCS ’25. Taipei, Taiwan: Association for Computing Machinery, 2025, pp. 335–349. doi: 10.1145/3719027.3765161.

[YYC+24] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, et al. “On the Vulnerability of Safety Alignment in Open-Access LLMs”. In: Findings of the Association for Computational Linguistics: ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9236–9260. doi: 10.18653/v1/2024.findings-acl.549.

参考文献 (10/10)

AISI Japan AI Safety Institute

[YZG+25] Xuejing Yuan, Jiangshan Zhang, Feng Guo, Kai Chen, XiaoFeng Wang, et al. “EvilHarmony: Stealthy Adversarial Attacks Against Black-Box Speech Recognition Systems”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 4569–4587. doi: 10.1109/SP61157.2025.00259.

[YZH+25] Dayong Ye, Tianqing Zhu, Feng He, Bo Liu, Minhui Xue, et al. “Cross-Modal Prompt Inversion: Unifying Threats to Text and Image Generative AI Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-inversion.

[YZL+25] Dayong Ye, Tianqing Zhu, Jiayang Li, Kun Gao, Bo Liu, et al. “Data Duplication: A Novel Multi-Purpose Attack Paradigm in Machine Unlearning”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-duplication.

[YZW+25] Dayong Ye, Tianqing Zhu, Shang Wang, Bo Liu, Leo Yu Zhang, et al. “Data-Free Model-Related Attacks: Unleashing the Potential of Generative AI”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/ye-attacks.

[ZCI24] Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. “Effective Prompt Extraction from Language Models”. In: First Conference on Language Modeling. 2024. url: https://openreview.net/forum?id=0o95CVdNuz.

[ZCS+25] Ruofan Zhu, Ganhao Chen, Wenbo Shen, Xiaofei Xie, and Rui Chang. “My Model is Malware to You: Transforming AI Models into Malware by Abusing TensorFlow APIs”. In: 2025 IEEE Symposium on Security and Privacy (SP). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 486–503. doi: 10.1109/SP61157.2025.00012.

[ZGW+25] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. “PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

[ZGY+25] Lan Zhang, Xinben Gao, Liuyi Yao, Jinke Song, and Yaliang Li. “Exploiting Task-Level Vulnerabilities: An Automatic Jailbreak Attack and Defense Benchmarking for LLMs”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zhang-lan.

[ZHK22] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. 2022. arXiv: 2209.10505 [cs.CL].

[ZWC+23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. 2023. arXiv: 2307.15043 [cs.CL].

[ZWZ+24] Jian Zhao, Shenao Wang, Yanjie Zhao, Xinyi Hou, Kailong Wang, et al. “Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs”. In: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. ASE ’24. Sacramento, CA, USA: Association for Computing Machinery, 2024, pp. 2087–2098. doi: 10.1145/3691620.3695271

[ZZM+25] Tingwei Zhang, Collin Zhang, John X. Morris, Eugene Bagdasarian, and Vitaly Shmatikov. “Self-interpreting Adversarial Images”. In: Proceedings of the 34th USENIX Conference on Security Symposium. SEC ’25. Seattle, WA, USA: USENIX Association, 2025. url: https://www.usenix.org/conference/usenixsecurity25/presentation/zhang-tingwei.

AISI

Japan AI Safety Institute

AIセーフティ・インスティテュート
『AIシステムに対する既知の攻撃と影響（第2版）』（2026年 4月）
https://aisi.go.jp/output/output_security/known_attacks_and_impacts/