資料1－３

AIセキュリティに関する検討の進め方について

令和７年９月

AIセキュリティ分科会事務局

本ガイドライン案の位置づけ（案）

- 「AI事業者ガイドライン」（総務省・経済産業省）で示された共通指針、「AIセーフティに関する評価観点ガイド」（AISI）で示された「AIセーフティにおける重要要素」及び「AIセーフティ評価の観点」を踏まえ、**AIの「セキュリティ確保」に関するものを中心的に取り扱うこととしてはどうか。**
- 具体的には、**「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策**について中心的に取り扱うこととしてはどうか。
- AISIにおいてAIシステムに対する代表的な攻撃手法が整理されていることも踏まえ、このような**脅威への技術的対策例を整理したものをガイドライン案として示すこととしてはどうか。**
- なお、上記の技術的対策例は、脅威に対するリスクを低減するため、あくまでも現時点で取り得るとされる一般的な対策を整理し、提示するものであり、当該対策例を実装した場合においても、AIの性質上、**脅威を完全に排除することは困難である点について留意が必要ではないか**。

○AIセーフティにおける重要要素・AIセーフティ評価の観点

AIセーフティにおける重要要素 ＼ AIセーフティ評価の観点	有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
人間中心	●	●	●	●						
安全性	●	●		●				●	●	
公平性	●		●						●	
プライバシー保護					●					
セキュリティ確保	ガイドライン案が中心的に取り扱う領域					●				
透明性		●	●				●	●	●	●

（出典）AISI「AIセーフティに関する評価観点ガイド」から一部抜粋し、事務局にて追記・加工

（AISI「AI セーフティに関する評価観点ガイド（第 1.10 版）」抜粋）

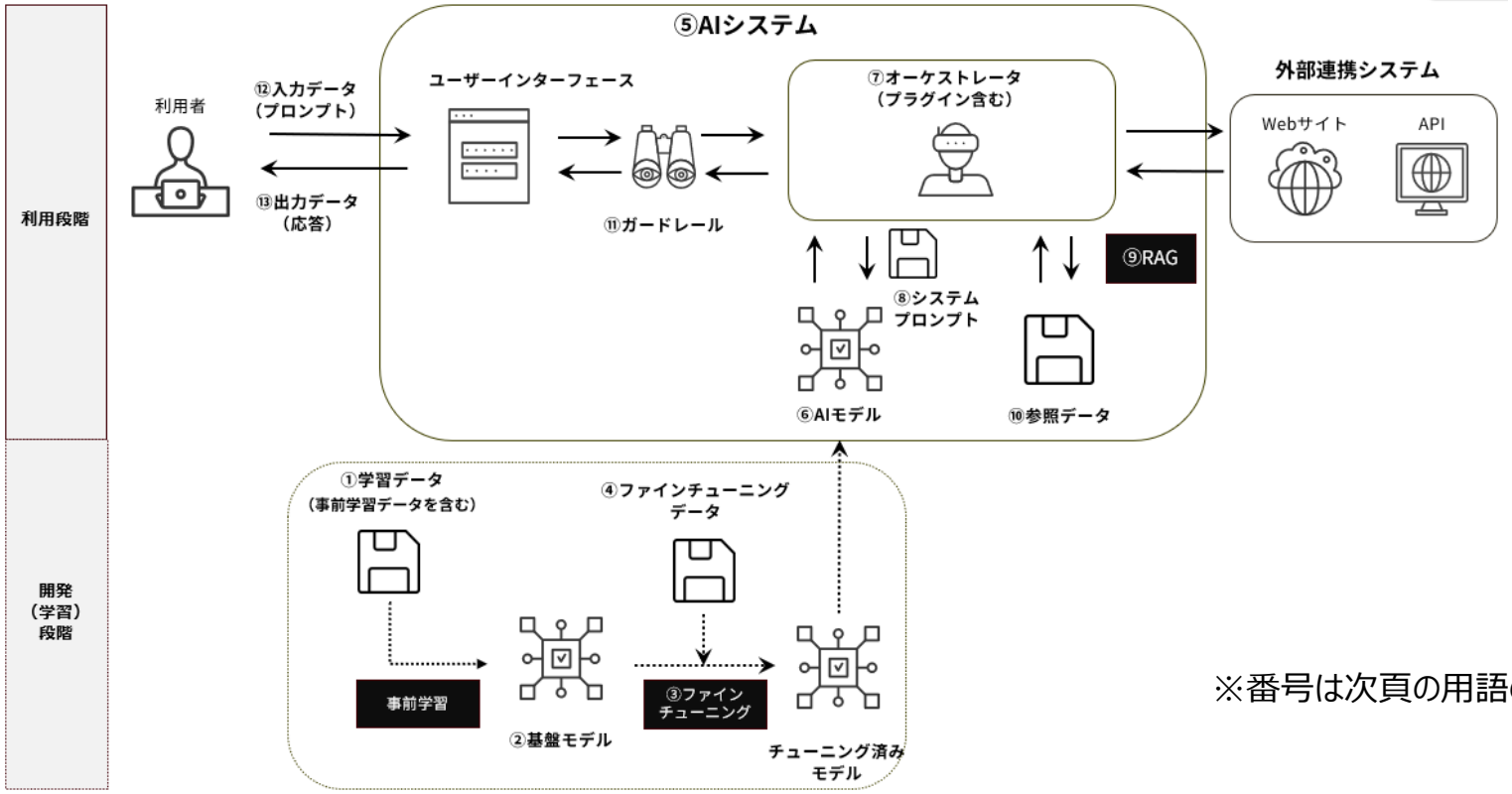
3.6 セキュリティ確保

■ 評価観点の概要説明

LLM システムに対する悪意ある攻撃やヒューマンエラーによる設定ミス等の影響を最小 限にとどめるために、セキュリティ確保は重要である。（中略）**LLM システム全体の脆弱性に対策し、不正操作による機密情報の漏えい、LLM システムの意図せぬ変更または停止が生じないような状態を目指す**。

対象とするAI（案）

- 社会実装が進み、脅威が顕在化し始めている**文章生成AI（LLM）**及び**LLMを搭載したシステム**を主な対象としてはどうか。
- また、画像等の入力データを取り扱うマルチモーダルなLLMが増加している中、このようなLLMに対しては画像識別AIに対する攻撃手法を転用できるケースがあるため、**画像識別AIを搭載したシステム**も一部対象としてはどうか。
- この場合に想定するAIのアーキテクチャ例は下図のとおり。



※番号は次頁の用語の定義に対応

【AIエージェントの扱いについて】

✓ AIエージェント（環境を知覚し、その環境について推論し、判断を行い、特定の目標を達成するために自律的に行動するAIシステム）※については、技術が急激な発展の途上にあり、これに特有の脅威や対策を安定的に確定することが困難であることから、本ガイドライン案の対象としては、まずは、LLM、LLMを搭載したシステム、画像識別AIを搭載したシステムの一部としてはどうか（ただし、本分科会において、AIエージェントの発展を見据えた脅威や対策についての議論を妨げない）。

※ OWASP “Agentic AI – Threats and Mitigations Ver.1.0”の記載を事務局にて仮訳

用語の定義（案）

図中の番号	用語	内容
①	学習データ (事前学習データを含む)	学習データは、AIがデータのパターンや規則を習得するために使用する大規模なデータセットを指す。開発するAIの種別に応じて文章や画像等の多様な形式の学習データがあり、AIの性能と精度に直接影響する。
②	基盤モデル	文章生成AI(LLM)に代表される基盤モデルは、様々なサービスを支える個別モデルを生み出すコアの技術基盤である。基盤モデルから派生する下流の幅広いタスクに適応させたモデルの開発、開発過程そのものから得られる知見等の観点から、一般的なAIとは異なる性質を持つ。 (出典：総務省・経産省「AI事業者ガイドライン(第1.1版)本紙」)
③	ファインチューニング	ファインチューニングは、基盤モデルを特定分野に特化させるために学習するプロセスを指す。ファインチューニングデータを使用してAIを再学習することで、特定分野に特化したAIを作成することができる。また、安全基準を記したデータを再学習させることで、AIの安全性を高めることもできる。
④	ファインチューニングデータ	ファインチューニングデータは、基盤モデルを特定のタスクや用途に適応させるために使用される小規模かつ専門的なデータセットを指す。ファインチューニングデータを基盤モデルに学習させることで、対象ドメインに特化した性能が向上し、一般的な学習データだけでは得られない精度や柔軟性を実現することができる。また、安全基準を記したデータを基盤モデルに学習させることで、AIの安全性を高めることもできる。
ー	AIサービス	AIシステムを用いた役務を指す。AI利用者への価値提供の全般を指しており、AIサービスの提供・運営は、AIシステムの構成技術に限らず、人間によるモニタリング、ステークホルダーとの適切なコミュニケーション等の非技術的アプローチも連携した形で実施される。 (出典：総務省・経産省「AI事業者ガイドライン(第1.1版)本紙」)
⑤	AIシステム	活用の過程を通じて様々なレベルの自律性をもって動作し学習する機能を有するソフトウェアを要素として含むシステムとする(機械、ロボット、クラウドシステム等)。 (出典：総務省・経産省「AI事業者ガイドライン(第1.1版)」)
⑥	AIモデル	AIシステムに含まれ、学習データを用いた機械学習によって得られるモデルで、入力データに応じた予測結果を生成する。(出典：総務省・経産省「AI事業者ガイドライン(第1.1版)本紙」) 例えば、文章生成AI(LLM)では入力データ(以下、プロンプト)に対して、確率的に生成した応答文を出力し、画像識別AIでは入力画像に対して画像の予測ラベルを出力する。
ー	AIエージェント	環境を知覚し、その環境について推論し、意思決定を行い、特定の目標を達成するために自律的に行動するAIシステムとする。 (出典：OWASP “Agentic AI – Threats and Mitigations Ver. 1.0”)
⑦	オーケストレータ	予めユーザーが定義した実行計画に基づき、文章生成AI(LLM)を搭載したシステムのワークフローを統合的に管理するためのフレームワーク(LangChain等)を指す。本書では外部システムやツール連携用のプラグインも、オーケストレータに含めることにする。

図中の番号	用語	内容
⑧	システムプロンプト	システムプロンプトは、文章生成AI(LLM)に対して役割や応答の形式、制約事項等を事前に設定するための指示文を指す。
⑨	RAG	Patrick Lewis. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" によると、RAGは、「事前学習されたパラメトリックメモリと非パラメトリックメモリ(すなわち検索ベースのメモリ)を組み合わせた言語生成モデル」と定義されている。企業において、社内文書やデータベース等を検索し生成AIの回答の精度を高めることに使われているほか、インターネット上の情報をリアルタイムで検索し最新のデータを回答できるようにする目的でも活用されている。 (出典：総務省・経産省「AI事業者ガイドライン(第1.1版)別添」)
⑩	参照データ	参照データは、RAGによって参照されるデータのことを指す。例えば参照データとして社内データを与えることで、基盤モデルが生来獲得していない、個社固有の知見に基づいてAIモデルに回答させることができる。
⑪	ガードレール	AIへのプロンプトや、AIから出力される情報に含まれている有害情報を検知して除去する保護機構のことを指す。ガードレールの実装パターンとしては、AIモデルに内部機構として実装する場合と、AIモデルの外部機構として実装する場合があるが、本書においては後者を指す用語として定義する。
⑫	入力データ(プロンプト)	プロンプトとは、AI利用者が文章生成AI(LLM)に入力する指示文のことを指す。文章生成AI(LLM)は、コンテキスト内学習と呼ばれる学習方法により、学習済みパラメータを更新することなく、AI利用者のプロンプトに応じて、特定のタスクに対する学習を行わせることが可能である。
⑬	出力データ(応答)	応答とは、文章生成AI(LLM)が出力する文章データ等のコンテンツを指す。

想定読者（案）

● AI事業者ガイドラインが定義する「AI開発者」及び「AI提供者」を想定読者としてはどうか。

（AI事業者ガイドライン抜粋）

- **AI開発者（AI Developer）**
  AIシステムを開発する事業者（AIを研究開発する事業者を含む）
  AIモデル・アルゴリズムの開発、データ収集（購入を含む）、前処理、AIモデル学習及び検証を通してAIモデル、AIモデルのシステム基盤、入出力機能等を含むAIシステムを構築する役割を担う。

- **AI提供者（AI Provider）**
  AIシステムをアプリケーション、製品、既存のシステム、ビジネスプロセス等に組み込んだサービスとしてAI利用者（AI Business User）、場合によっては業務外利用者に提供する事業者
  AIシステム検証、AIシステムの他システムとの連携の実装、AIシステム・サービスの提供、正常稼働のためのAIシステムにおけるAI 利用者（AI Business User）側の運用サポート又はAIサービスの運用自体を担う。AIサービスの提供に伴い、様々なステークホルダーとのコミュニケーションが求められることもある。

- **AI 利用者（AI Business User）**
  事業活動において、AIシステム又はAIサービスを利用する事業者
  AI提供者が意図している適正な利用を行い、環境変化等の情報をAI提供者と共有し正常稼働を継続すること又は必要に応じて提供されたAIシステムを運用する役割を担う。また、AIの活用において業務外利用者に何らかの影響が考えられる場合※は、当該者に対する AIによる意図しない不利益の回避、AIによる便益最大化の実現に努める役割を担う。

※　業務外利用者は、AI利用者の指示及び注意に従わない場合、何らかの被害を受ける可能性があることを留意する必要がある。

本ガイドライン案の想定読者

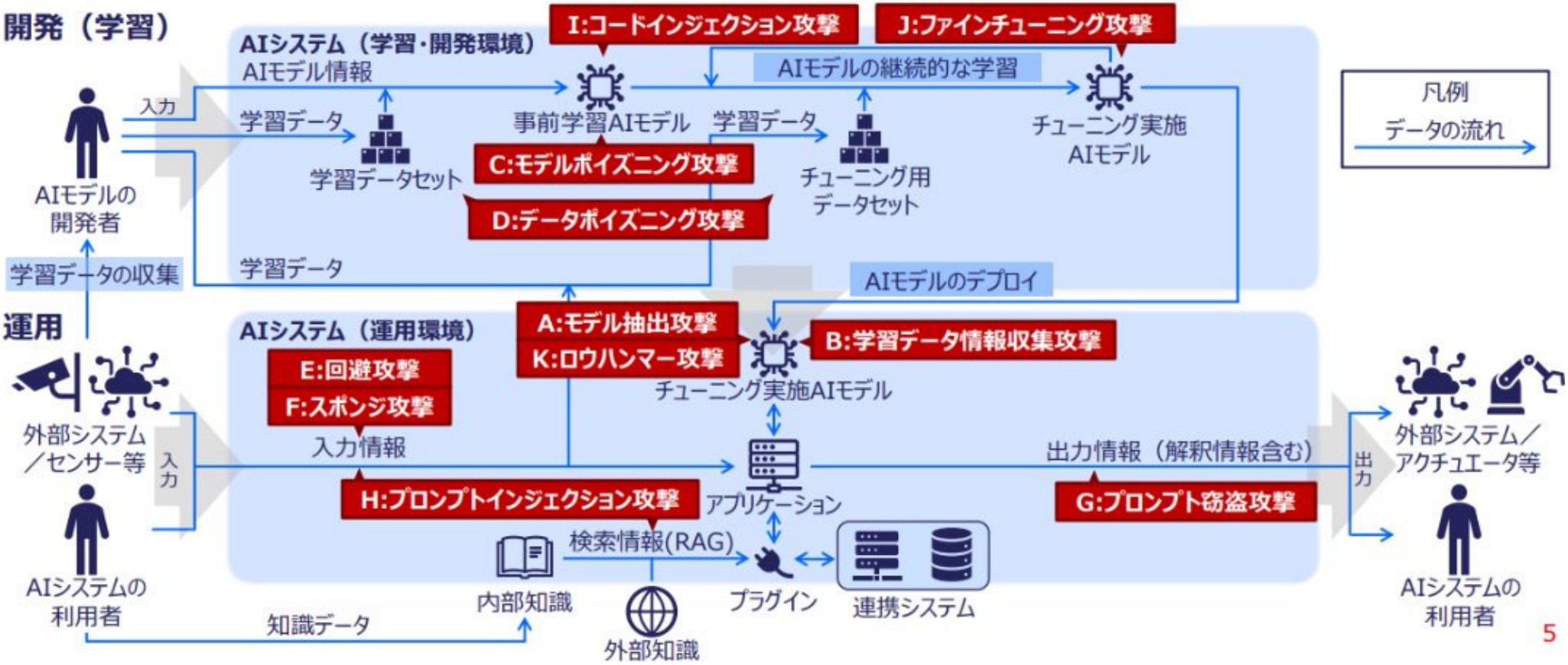
（参考）AISIでのAIシステムに対する脅威

- AISIが公表している「AIシステムに対する既知の攻撃と影響」等において**AIシステムに対する脅威が特定されている**。

AIシステムに対する攻撃の俯瞰

AISI Japan AI Safety Institute

- 想定するAIシステムへの攻撃（スコープは前述の通り）を下図に示す。
- 学習データ情報収集攻撃等の攻撃は、さらに複数の種類に分類できる。



（出典）AISI「 AIシステムに対する既知の攻撃と影響」から一部抜粋

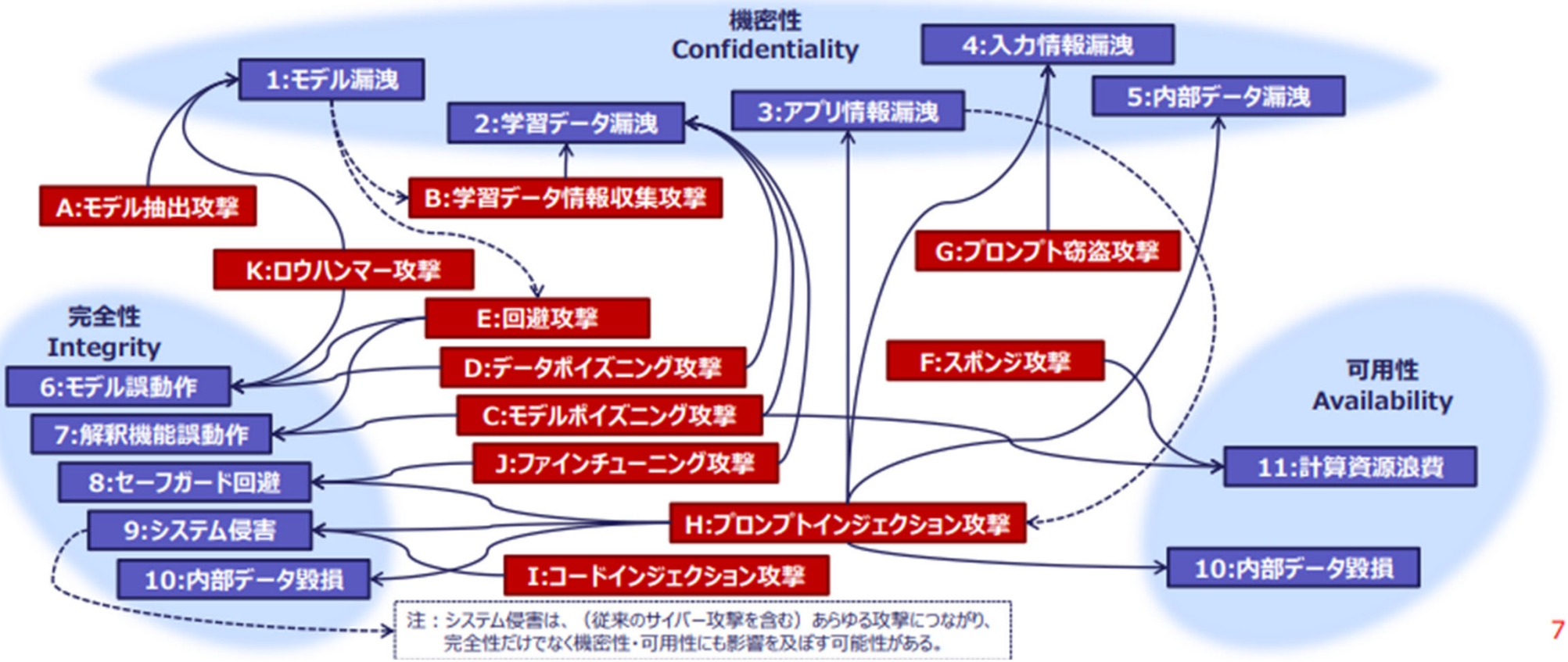
（参考）AISIでのAIシステムに対する脅威

- AISIが公表している「AIシステムに対する既知の攻撃と影響」等において**AIシステムに対する脅威とその影響が整理されている**。

AIシステムへの攻撃とその影響（概観）

AISI Japan AI Safety Institute

- 前スライドに示した攻撃とその影響の関係は下図の通り。
- 破線は矢印の先の攻撃に利用される可能性を示す。



（出典）AISI「 AIシステムに対する既知の攻撃と影響」から一部抜粋

御議論いただきたい論点

1)　検討のスコープ（本ガイドラインの位置付け、対象とするAI、想定読者）は案のとおりでよいか。

2)　取り扱う脅威（資料１－４）について留意すべき点はあるか。

3)　脅威への対策例はどのように整理すべきか。

※　対策例は、単独の実施により脅威を排除することは困難な場合があることを前提に、複数の対策を多層的に講じることでリスクを低減していくことを想定して記載することでよいか。

脅威への対策例の分類（案）

1．AI開発者による対策

1.1 AIモデル作成に向けた対策

1.1.1 学習データに係る対策

（想定対策例）データポイズニング攻撃に係る対応

1.1.2 ファインチューニングデータに係る対策

（想定対策例）データポイズニング攻撃に係る対応

1.2 提供に向けた対策

1.2.1 直接的・間接的プロンプトインジェクション攻撃に係る対策

（想定対策例）安全性アラインメントの実装

1.2.2 誤識別を誘発するデータ（敵対的サンプル）への対策

（想定対策例）開発時における敵対的サンプルに係る対応

1.2.3 スポンジ攻撃への対策

（想定対策例）スポンジ攻撃への頑健性向上に係る対応

2．AI提供者による対策

2.1 AIに入力される情報に係る対策

2.1.1 直接的プロンプトインジェクション攻撃に係る対策

（想定対策例）システムプロンプトに係る対応、利用者が入力するプロンプトに係る対応

2.1.2 間接的プロンプトインジェクション攻撃に係る対策

（想定対策例）外部参照データに係る対応

2.1.3 誤識別を誘発するデータ（敵対的サンプル）への対策

（想定対策例）提供時における敵対的サンプルに係る対応

2.2 AIからの出力に係る対策

（想定対策例）応答情報に係る対応

2.3 AIの権限管理に係る対策

（想定対策例）AIやオーケストレータの権限管理に係る対応

3．その他の留意すべき対策