AI-IRS
AIインシデントレスポンス・アプローチブック
（概要版）

Approach Book for AI Incident Response System: AI-IRS
~ A Foundation for Trustworthy AI ~

2026-1

AIセーフティ・インスティテュート

AISI Japan AI Safety Institute

エグゼクティブ・サマリ

AISI Japan AI Safety Institute

AIシステム向けインシデントレスポンスの枠組み「AI Incident Response System（AI-IRS）」について

近年、AIが意思決定や業務効率化など**事業の中核として活用**され始めてきています。一方で、仮に中核を担うAIの誤判断やサイバー攻撃による停止などが発生した場合、事業活動に**大きな被害をもたらす**可能性があります。特にAIでは、ブラックボックス性（**処理過程が不透明で、意思決定の内実を非常に把握しにくい性質**）や自律性（**目標達成に向けて人の介入なく自ら学習・判断・行動等を行う性質**）などの性質により、AIシステムではリスク対策を実施してインシデントの発生を未然に防ぐ**予防的統制**が、従来システムに比べて難易度が高いと見込まれます。AIシステムではゼロリスクの達成が困難である前提に立ち、インシデント発生時にも事業影響を最小化する取り組みが重要です。

これらを背景として、従来の情報システムにおけるインシデントレスポンスをAIシステムにも適用・拡張させる新たな枠組みであるAI-IRSを提示します。AI-IRSでは、**観測性（可視化）**と**制御性（封じ込め）**を中核に、運用中の挙動を把握し、問題箇所を切り離すことで**被害を最小化**する枠組みです。既存の開発・運用プロセスに容易に追加でき、運用負荷とインシデントの影響の双方を抑えられます。

また、AI-IRSは組織内部の取り組みに限定されず、サプライチェーンにまたがるAIライフサイクルの全体を観測・制御し、異常の検知・封じ込め・復旧を自律化する**共通基盤の整備**に向けたビジョンも示しています。

この枠組みにより、AIシステムを用いて事業活動を行う組織が、**AIを制御可能な戦略資産**として位置づけ、運用のレジリエンスを確立し、事業継続性を確保し、さらには**イノベーションの加速を促進**することが重要になります。組織や社会における整備・運用を進める際のきっかけや参考として、ぜひご活用ください。

AIインシデント対応に向けたアプローチ

AISI Japan AI Safety Institute

- 日々進化するAIでは、今まで以上にリスクへの予防的対策が困難となるため、リスクの発生(≒AIインシデント)に備えた対策はより重要となる
  - AISIはAIセーフティ実現に向け活動(評価観点ガイド、レッドチーミング手法ガイド、評価ツールの公開 等)
  - これらはシステムの防御・堅牢性を高める予防的対策となるが、AIシステムではAIエージェントの活用などが進み、よりAIインシデントレスポンスが困難となることが想定される
  - AIインシデントが発生した際に、AIシステムを評価し影響を最小限に抑える取り組みを進める必要がある

予防的対策	課題	＋発見的対策
✓開発段階や運用開始後定期的なAIセーフティ評価の実施 ✓AIシステムの不適切な動作が発生しないことを事前に評価し、発生自体を防止する ✓システムの防御・堅牢性を高め、AIインシデントを発生させないAIシステムを構築する	✓全てのリスクに事前に対策するのは不可能 ✓特にAIシステムでは内部ロジックの把握が難しく、AIインシデントの原因調査が困難になる等、想定される ✓AIインシデントは必ず発生するという前提に立った対策が必要	✓AIシステムにおけるAIインシデント発生時の影響最小化と迅速な復旧に向けた対策 ✓AIインシデントレスポンス能力の評価に加え社会基盤の確立も必要 ✓組織の枠を超えた社会システムの安定的な継続を維持する

✓AISIはAI事業者に向けたAIシステムにおけるAIインシデントレスポンス態勢構築に向けたアプローチブックを作成

目次

AISI Japan AI Safety Institute

1-1. 背景・想定読者

AISI Japan AI Safety Institute

日々進化するAIでは、今まで以上にリスクへの**予防的対策**が困難となるため、リスクの発現（≒インシデント）に備えた**発見的対策**はより重要となる

背景

- AIリスクは複雑化し予防だけでは限界。迅速な**インシデントレスポンス態勢の構築**が不可欠
- 競争力強化へAI活用を推進しつつ、評価とインシデントレスポンス態勢の同時強化が求められる
- 従来のインシデントレスポンス関連ガイドは**AIに未対応**
- インシデント対応態勢と合わせて、**SBOM**等の**基盤整備**も求められる

課題

- AI導入で新種・進化するリスクが増大。運用中に**挙動が変化**し、従来型の固定手順やパッチ対応だけでは不十分
- AIの**ブラックボックス性**や**自己更新**により仮に検知できても、その後の原因究明・修復が難化する
- 日本国内では**予防偏重**の傾向。一度インシデントが発生すると、影響が長期化する。**発見的統制**の強化が必要

想定読者

AIシステムを業務に導入・運用する組織において、以下に記載する幅広い立場の関係者

- ✓ **CAIO**（最高AI責任者）、CISO（最高情報セキュリティ責任者）などのマネジメント層
- ✓ **CSIRT**（コンピュータセキュリティインシデント対応チーム）、SOC（セキュリティオペレーションセンター）、ISIRT（情報セキュリティインシデント対応チーム）　などのインシデント対応関連専門部隊
- ✓ AIにおける共通的なルールや方針を立案する政策担当者　等

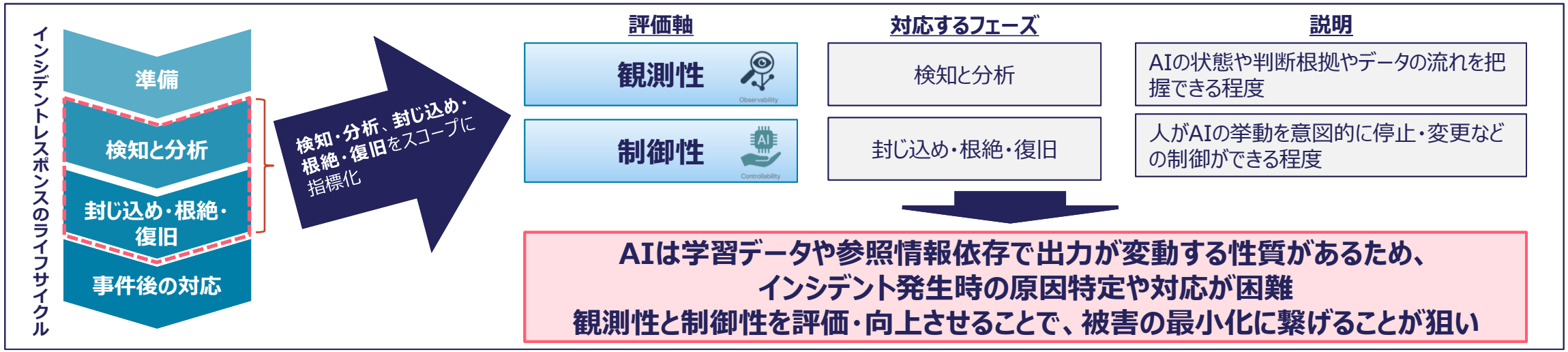
位置付け

本紙は、組織がAIインシデントレスポンス強化するためのベストプラクティスを羅列したものではない。組織や社会がAIシステムやAI利用におけるインシデントレスポンス強化の必要性や観点、未来像などを示すことで、読者にこれらを認識いただき、組織や社会における整備・運用を進める上でのきっかけや参考となることを期待する

1-2. スコープ

インシデントレスポンスのライフサイクル（準備、検知・分析、封じ込め・根絶・復旧、事後対応）のうち、検知・分析、封じ込め・根絶・復旧がスコープ

- AIシステムにおいても、従来のシステム構成の中にAIが搭載されているという認識の上で、 **NIST SP800-61** Rev.3 ※ に定められているインシデントレスポンスのライフサイクル（準備、検知・分析、封じ込め・根絶・復旧、事後対応）は有効であるが、各ライフサイクル内での具体的な対応策が変わってくる（追加される）と想定
- 現状のシステムにおけるインシデントレスポンスのライフサイクル（準備、検知・分析、封じ込め・根絶・復旧、事後対応）において、**AIシステムの場合に追加で対応が必要**になる観点の検討が必要と判断
- AISIでは、上記ライフサイクルのうち、インシデントが発生した際の緊急対応フェーズとなる検知・分析、封じ込め・根絶・復旧フェーズをスコープに、対応レベルを評価することを目指して「**観測性**」「**制御性**」の二つの指標で整理する



※ NIST SP 800-61 Rev.3（SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile | CSRC ）

2-1. GAP分析とケース毎の整理

AISI Japan AI Safety Institute

インシデントレスポンスに関する既存の枠組み（NIST SP800-61）を基に、AIシステムにおけるインシデント対応をオーバレイする

- まずは現状のシステムにおけるインシデントレスポンスのライフサイクル（準備、検知・分析、封じ込め・根絶・復旧、事後対応）のうち、インシデントが発生した際の緊急対応フェーズとなる検知・分析、封じ込め・根絶・復旧フェーズをスコープに、**AIシステムの場合に追加で対応が必要になる内容**について、現状とのGAP分析として整理する。
- また、AIの技術は日々進化しており、それぞれに合わせたインシデントレスポンスの仕組みや体制を検討することが有効となるケース毎に整理する。状況の変化に合わせて、ケースの追加や更新を検討する。
- これらの取り組みにより、AIが起因となるインシデントに対しても、影響の軽減につなげることができる。

GAP分析

観測性と制御性における対応内容について、
AIシステムの場合に追加等が必要な内容を整理

分類	AIシステムにおいて追加される対策（例）
観測性	従来のソースコードルールベースバグ探索では検知困難なため、入力データとAIの出力結果の不規則な逸脱・異常値検知方法・フローの検討　等
制御性	改ざんされたモデルの過去バージョンへの切替/再訓練や、悪意ある学習データ（データポイズニング）が混入した場合の該当データの局所除去・再学習の検討　　等

Plus

ケース毎の整理

AIの技術的進歩や利用方法のトレンドに合わせた
インシデントレスポンスの対応方法例を整理

RAG　AIエージェント　・・・（今後追加）

2-2. 対応レベルの評価指標

AISI Japan AI Safety Institute

AIインシデントレスポンスにおいて、**観測性**と**制御性**を両立することが、インシデントレスポンス態勢の強化に繋がる

- GAP分析を制御性、観測性の２つの軸で再整理し、AIシステムにおける対応の抜け漏れを防ぐことを目的とする。

観測性＼制御性	可制御	不可制御
可観測	【理想状態】(〇観測・〇制御) AIの挙動を把握・制御でき、運用・封じ込めが容易でインシデントへの対応ができる。	【制御不足状態】(〇観測・×制御) AIの内部状態や行動を観測できるが、操作できない状態にある。封じ込め不能リスク。
不可観測	【観測不足状態】(×観測・〇制御) AIの挙動を止めるなどの対応ができるが、原因分析のための情報が得られず、過剰な範囲を停止するなど非効率な制御状態にある。	【非管理状態】(×観測・×制御) AIの挙動を観測も制御もできない。最も危険な状態。

⇒ **可観測＋可制御の状態を目標**に対応することが重要である (組織やサービスの置かれる状況によって、具体的に求められる内容は異なる)

- ケース毎に整理した結果の一部を記載する（詳細な整理結果は、詳細版を参照）

RAG

・機微情報の漏洩
・データ汚染による不正動作・・・

✓ 迅速な原因分析を可能にするための出力プロンプトのもとになった情報が含まれている**RAGのトレース機能**の導入
✓ インシデントの原因となった**RAGのみをサービスから切り離す機能**、およびRAGなしでもサービス継続できる仕組みの導入　等

AIエージェント

・自律判断によるトレース困難さ
・外部不正リソースの利用・・・

✓ **各コンポーネント毎の通信（入力・出力）をリアルタイムに検査**し、インシデントの発生および発生個所を即座に検知する仕組みの導入
✓ **インシデントの原因となったコンポーネントのみに対して停止/切り離し**を行い、サービス継続への影響を最小限に抑える仕組みの導入　等

2-3. 関連する要素

AISI Japan AI Safety Institute

AIにおけるインシデントレスポンス態勢構築にあたって、組織における対策の実施に加えて、**組織をまたいだ仕組み構築**も重要

- AI時代のインシデント対応力は、個別最適の対策だけでは限界があり、**組織をまたいだ仕組みづくりが重要**
- 組織をまたいだ仕組みづくりとして有効なものとして、以下がある

SBOM for AI	AIライフサイクルに関わるサプライチェーン全体において、メタデータ、データセット、システム特性などの情報を集める仕組み。静的な部品表ではなく、AIシステムの構成変更を反映できる動的なレポジトリ運用を想定。(G7がビジョンを公開)
Cyber AI Profile	NIST SP800-53（情報システムに求められるセキュリティ管理策を示した文書）の枠組みにAI特有のリスク対策を重ね合わることを想定したもの。AIシステムにおけるセキュリティ統制の指針として期待されている。

- 官民・業界横断の協調が、復旧を加速し損害を最小化することに繋がる
- 共通の指標と監査可能な証跡を整え、**サプライチェーン全体の透明性と説明責任を高める**ことで、平時の準備と有事の指揮統制を両立させる
- 上記のインシデントレスポンスにおける活用例として、**SBOM for AIとCVE等の外部情報突合でリスクの所在を可視化**し、インシデント検知や原因調査の迅速化に繋げる。また**Cyber AI Profileで要件と手順を定め**、あらゆる組織においての共通の標準として整備することで、**サプライチェーンを連鎖する影響の最小化**に繋げる

インシデントレスポンスにおける各要素の活用

要素	SBOM for AI	Cyber AI Profile	…
活用例	リスクの所在の可視化 等	要件と手順の標準化 等	

2-4. 想定される効果事例

AISI Japan AI Safety Institute

AI-IRS導入により、AIシステムにおいてもインシデント箇所を**具体的に検知**し、**早急に対応**できるようになる

- ここでは、３つの想定インシデント事例（自動計画AIエージェント誤実行、ファインチューニング誤ラベル混入、RAGデータ汚染）を基に説明する

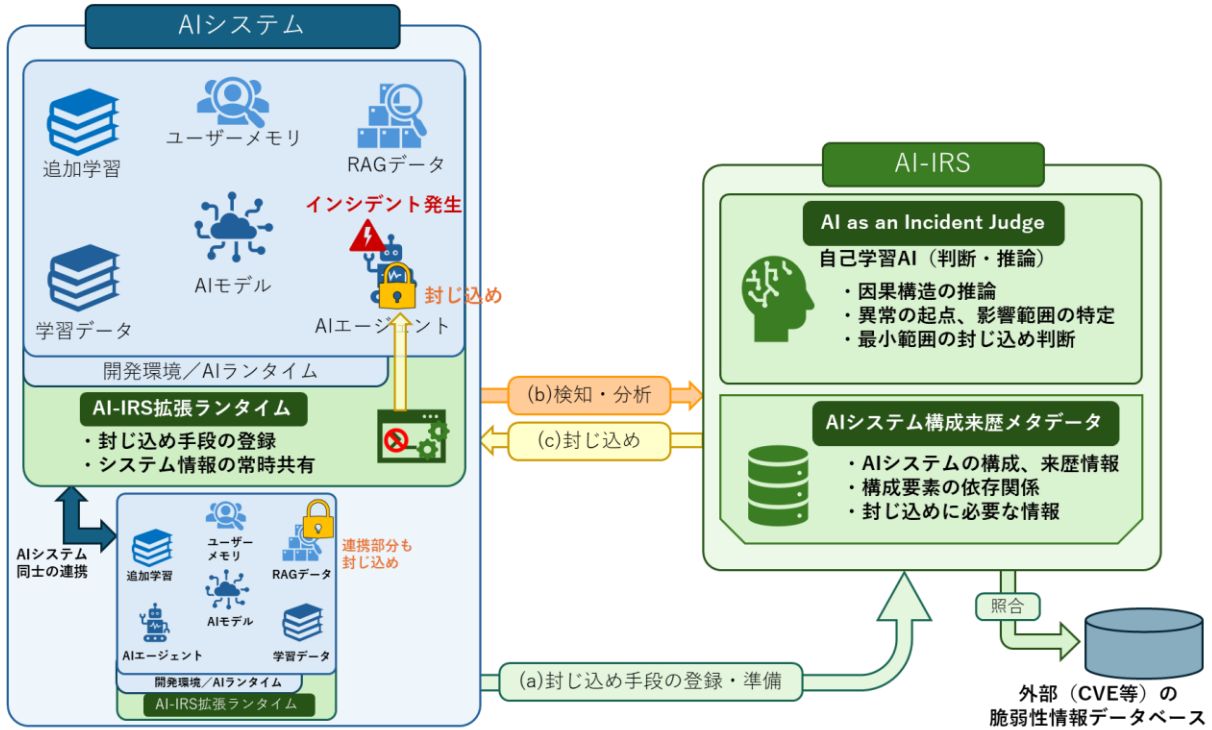
		非管理状態（Before） 原因特定できないため、業務の停止やAI利用停滞に繋がる	理想状態（After） 原因箇所を具体的に特定し、局所的な対応で業務継続が可能
①自動計画AIエージェント誤実行	外部レポートの隠れたプロンプトインジェクションで、C社の**自動計画エージェントが未承認出荷を自動実行**。 A社提供・B社開発AIの安全制御が迂回され、外部依存攻撃を検知できず実行に至った。	深夜に倉庫ラインが暴走し出荷済み。 A社は想定外命令、B社は正常主張で**原因不明**。 経営はAIの自律性を警戒し管理強化、マニュアルオペレーションへ後退。 DX推進室縮小、活用停滞と夜間無人自動運転見直し。	AI-IRSで監査とポリシー制御が異常命令を**検知・即封**じ、未承認出荷を阻止。業務は途切れずAIが**自己修正**。 経営はAIを副操縦士と捉え自律運用と夜間無人稼働を拡大。24時間安全稼働が定着し、社員は戦略・創造業務に集中。
②ファインチューニング誤ラベル混入	E社がD社モデルをF社向けに再学習。 複数提供データで経路がブラックボックス化し、**誤ラベル混入の所在特定不能**。 結果として要因の可視化不足が招いたF社業務AIの**重大な誤判断**を引き起こした。	AI誤出力で生産計画を誤り、取引先の信頼を大きく損失。 原因調査でもD社・E社に再学習ログがなく、**原因不明・責任の所在は曖昧**。 F社は契約を停止し「AIは難しい」と後退、研修も中止で熱が冷めた。	E社は学習履歴・来歴をメタデータ管理しF社に共有。 監査AIがF社異常で誤ラベルを**特定**し、ポリシーが**自動修正・再構築**。 障害は手前で解決、F社はAIを全面統合・標準化、AI部門は中枢化へ。
③RAGデータ汚染	G社LLMをH社がコールセンター向けに調整、I社はFAQ参照RAGで導入。 FAQに不可視の間接プロンプトインジェクションが仕込まれ、**応答ロジックが書換えられ「全返品無料」と誤答**。	14時にAIが「返品無料」を連発、SNS拡散で申請急増し倉庫混乱。 改変痕なく**原因不明**、G社H社も異常なし。 業務は一時停止、経営は慎重化し試験運用へ後退。	FAQ改変や不審命令をテレメトリで**即検知**し、ポリシーが**隔離**、Runbookで安全FAQへ**切替**。 業務は止まらず誤回答なし。 監査AIが即報告し原因特定、AIが定着。

3-1. AIインシデントレスポンスの未来予想

AISI Japan AI Safety Institute

AI-IRSはSBOM for AI、CVE解析、分散トレース等でライフサイクルを観測・制御し、異常の検知・封じ込め・復旧を自律化

［AIインシデントレスポンスシステムの未来予想イメージ図］



AIシステムとAI-IRSが連携して常時インシデントレスポンス態勢を自動で運用・対応・改善していくプラットフォームの整備

(a) 封じ込め手段の登録・準備

AIシステム開発時には、開発環境からAIの構成来歴に関するデータがAI-IRSに送られる。また、AIランタイムで実施可能な封じ込め手段の登録が行われる。

(b) 検知・分析

AI-IRS側にはAIシステムの構成情報や外部情報（脆弱性情報 等）が統合管理されており、これらの情報とAIシステムの情報（ログ、アラート等）を突合することで異常の検知や問題個所の局所特定を行う

(c) 封じ込め

検知・分析した結果を基に、影響を最小限にする方法で局所的な封じ込め対応を実施する

予め定めておいた指標を基に処理を自動実行する従来のオートメーションと異なり、**AI-IRSの内部にもAIが活用**され、収集情報の選定や整理、外部情報との突合、インシデント兆候分析、最適な封じ込め対策の判断等を自律的に行う。運用を通じて自らPDCAを回していく**自己改善型システム（AI as an Incident Judge）として機能**する

AI-IRSへの投資は安全性への支出だけではなく、事業継続性と市場からの信頼確保を通じた長期的なビジネスチャンスに繋がる。このことは、**AI時代に組織が競争力を得るための重要な要素**となる

また、AI-IRSを組織をまたいだ共通基盤として整備・運用することで、**社会システムの安定的な継続**に繋がる

AISI

Japan AI Safety Institute