TLP:CLEAR

OT における AI の安全な統合に関する原則

公開日：2025 年 12 月 3 日

米国サイバーセキュリティ・インフラセキュリティ庁

オーストラリア信号局 オーストラリアサイバーセキュリティセンター

米国国家安全保障局人工知能セキュリティセンター

米国連邦捜査局

カナダサイバーセキュリティセンター

ドイツ連邦情報セキュリティ庁

オランダ国家サイバーセキュリティセンター

ニュージーランド国家サイバーセキュリティセンター

英国国家サイバーセキュリティセンター

この文書は *TLP:CLEAR* と表示されている。開示に制限はない。情報源は、適用される公開規則および手順に従い、情報の悪用リスクが最小限または予見不可能である場合に *TLP:CLEAR* を使用できる。標準的な著作権規則に従い、*TLP:CLEAR* 情報は制限なく配布できる。詳細は「トラフィックライトプロトコル（*TLP*）の定義と使用法」を参照のこと。

目次

序論 3

重要な用語 3

適用範囲 4

AI 技術の種類 4

パデューモデルに基づく AI アプリケーション 6

OT における AI の安全な統合に関する原則 8

原則 1 – AI の理解 8

1.1 AI の固有リスクと OT への潜在的影響の理解 8
1.2 安全な AI システム開発ライフサイクルの理解 10
1.3 従業員への AI 教育 11

原則 2 – OT 分野における AI 利用の検討 13

2.1 AI 利用の OT ビジネスケースの検討 13
2.2 AI システムにおける OT データセキュリティリスクの管理 14
2.3 AI 統合における OT ベンダーの役割の理解 15
2.4 AI-OT システム統合における課題の評価 16

原則 3 – AI ガバナンスと保証フレームワークの確立 18

3.1 OT における AI ガバナンスメカニズムの確立 18
3.2 既存のセキュリティおよびサイバーセキュリティフレームワークへの AI の統合 18
3.3 AI のテストと評価の徹底的な実施 19
3.4 OT における AI の規制・コンプライアンス上の考慮事項 19

原則 4 – AI および AI 対応 OT システムへの監視・フェイルセーフ手法の組み込み 21

4.1 OT における AI の監視・監督メカニズムの確立 21
4.2 安全性とフェイルセーフ機構の組み込み 23

結論 24

参考資料 25

免責事項 26

謝辞 26

改訂履歴 26

附属書：用語 27

参考文献 29

TLP:CLEAR

序論

2022 年 11 月に ChatGPT が一般公開されて以来、人工知能（AI）は人間社会の様々な側面に組み込まれている。 重要インフラの所有者や運営者にとって、AI は効率や生産性の向上、意思決定の強化、コスト削減、顧客体験の改善に活用できる可能性がある。しかし、多くの利点がある一方で、重要な公共サービスを管理する運用技術（OT）環境に AI を統合することは、OT プロセスモデルが時間の経過とともにずれることや安全プロセスのバイパスなど、重大なリスクも伴う。所有者や運営者は、重要インフラの可用性と信頼性を確保するために、これらのリスクを慎重に管理しなければならない。

本ガイダンスは、サイバーセキュリティ・インフラストラクチャ保障庁（CISA）とオーストラリア信号局傘下のオーストラリアサイバーセキュリティセンター（ASD's ACSC）が共同で作成したものである。作成には、国家安全保障局人工知能セキュリティセンター（NSA AISC）、連邦捜査局（FBI）、カナダサイバーセキュリティセンター（Cyber Centre）、ドイツ連邦情報セキュリティ庁（BSI）、オランダ国家サイバーセキュリティセンター（NCSC-NL）、ニュージーランド国家サイバーセキュリティセンター（NCSC-NZ）、英国国家サイバーセキュリティセンター（NCSC-UK）が共同で作成した。以下「作成機関」と呼ぶ。本ガイダンスは、重要インフラの所有者および運営者が OT 環境に AI を統合するための実践的な情報を提供する。 本ガイダンスでは、重要インフラの所有者および運営者が OT システムにおける AI の利点を活用しつつリスクを低減するために従うべき 4 つの基本原則を概説する：

1. AI を理解する。OT 環境への AI 統合に伴う特有のリスクと潜在的影響、これらのリスクに関する要員教育の重要性、安全な AI 開発ライフサイクルを理解する。
2. OT 領域における AI 活用を検討する。OT 環境での AI 活用に関する具体的なビジネスケースを評価し、OT データのデータセキュリティリスク、ベンダーの役割、AI 統合に伴う短期的・長期的な課題を管理する。
3. AI ガバナンスと保証フレームワークを確立する。強固なガバナンスメカニズムを導入し、既存のセキュリティフレームワークに AI を統合する。AI モデルを継続的にテスト・評価し、規制順守を考慮する。
4. 安全・セキュリティ慣行を AI および AI 対応 OT システムに組み込むこと。AI 対応 OT システムの安全な運用とサイバーセキュリティを確保するための監視メカニズムを導入し、透明性を維持し、インシデント対応計画に AI を統合すること。

作成機関は、重要インフラの所有者および運営者に対し、本ガイダンスを確認し、原則を実行に移すことで、AI を OT システムに安全かつ確実に統合するよう促している。

重要な用語

本ガイダンスの範囲は、特に重要インフラの所有者および運営者が、OT 環境における AI システムの安全性とセキュリティを確保するためにどのように貢献できるかを対象とする。したがって、作成機関は、他の文脈における定義との混同を避けるため、本ガイダンスにおいて以下の用語を特定の定義で使用する：

TLP:CLEAR

- 人工知能（AI）とは、機械および人間からの入力を用いて、現実または仮想環境に影響を与える予測、推奨、または決定を行うシステムを指す。[1]
- 安全とは、OT 環境における物理的安全（正式には機能安全）を指す。OT システムは、生物剤や化学剤を散布するシステム、ダムや廃水処理の操作を制御するシステム、車両交通の流れを自動化するシステムなど、人や財産に危害を加える可能性のある物理システムを制御する。本ガイダンスにおいて、「安全」という単独の語は常に機能安全を指す。
- セキュリティ（本ガイダンスでは「情報セキュリティ」および「サイバーセキュリティ」と同義で使用）とは、情報および情報システムの機密性、完全性、可用性といったセキュリティ特性を確保することを指す。

本ガイダンスの範囲内における定義の完全な一覧とそれらの出典については、附属書：用語を参照のこと。

適用範囲

機械学習（ML）、統計モデリング、アルゴリズム計算は、いずれも長年重要インフラのエンジニアリングプロセスで使用されてきた AI 技術のサブセットである。ML と従来の統計モデリングは、データに基づく結果予測や意思決定に共に用いられるが、そのアプローチ、前提条件、適用範囲、OT システムとの安全な統合に関する考慮事項において差異がある。 本ガイダンスの範囲は、ML および大規模言語モデル（LLM）ベースの AI および AI エージェントに焦点を当てる。OT をこれらのタイプの AI システムと統合するには、より複雑な安全性とセキュリティ上の考慮事項が伴うためである。ただし、本ガイダンスは、従来の統計モデリングやその他の論理ベースの自動化によって強化されたシステムにも適用できる。以下のサブセクションでは、これらの異なる AI 技術を定義する。

AI 技術の種類

従来の統計モデリングは、変数間の関係を正確に記述するために数式を用いる。データが特定の分布に従い、関係が線形であるか、あるいは線形モデルで近似可能であると仮定する。統計モデリングは回帰分析、仮説検定、信頼区間などの手法を用いてモデルパラメータを直接推定し、予測を行う。予測、最適化、オペレーターの意思決定支援などのタスクに広く用いられる。 機械学習に基づかない AI システムは、意思決定や制御プロセスの自動化にアルゴリズムを用いる。OT システムでは、ラダーロジック自動化ルーチンや安全計装システムの一種がこれに該当する。

機械学習システムは、明示的にプログラムされなくてもデータから学習し、予測や意思決定を行うアルゴリズムを用いる。ML モデルは変数間の複雑な関係や非線形な相互作用を扱える。ML モデルは、データに基づく表現の開発や予測を行う際に、教師あり学習、 教師なし学習、強化学習といった様々な手法を用いる。ML はコンピュータビジョン、自然言語処理、ロボティクスなどの分野で、画像分類、音声認識、自動運転といったタスクに広く用いられる。

大規模言語モデル（LLM）は、自然言語のプロンプトを理解し、人間が理解できる応答を生成するように設計された高度な機械学習モデルである。LLM は言語パターンやマルチモーダルデータセットを活用し、ユーザーのプロンプトに対して複雑な応答を生成する。 LLM エンジニアは通常、出力生成時にランダム性を組み込む[2] 。これにより、同じ

[1] この文書では、15 U.S.C. 9401(3)に基づく AI の定義を使用している。ただし、AI の定義は団体や管轄区域によって異なる場合がある。

[2] Sander Shulhoff,「Basic LLM Settings」, *Learn Prompting*, 最終更新日 2025 年 3 月 10 日, https://learnprompting.org/docs/intermediate/configuration_hyperparameters.

TLP:CLEAR

入力に対して常に同じ応答を生成しないようにする。LLM は、意思決定の強化、日常業務の自動化、保守スケジュールの最適化を通じて重要インフラ事業体を支援する生成的 AI アプリケーションを駆動できる。その目的は、運用効率と信頼性の向上にある。

AI エージェントとは、AI や機械学習モデルを用いてデータを処理し、意思決定を行い、自律的な行動を開始できるソフトウェアの一種である。エージェント型 AI システムには多くの種類があり、生成的 AI アプリケーションやエージェントを駆動するために大規模言語モデル（LLM）を利用するシステムや、異なる機械学習技術、分析視点、意思決定手法、自律行動能力を組み合わせたシステムなどが含まれる。 LLM と同様に、意思決定の強化、日常業務の自動化、保守スケジュールの最適化が可能であり、これにより重要インフラの運用を改善・効率化できる。エラーチェックを実装することで、問題を回避し出力結果を許容範囲内に収めることができ、AI エージェントの性能向上につながる。

TLP:CLEAR

パデューモデルに基づく AI アプリケーション

パデューモデル (Purdue) は、OT と IT のデバイス・ネットワーク間の階層的関係を理解する上で、依然として広く受け入れられている枠組みである。表 1 は、パデューモデルに基づく重要インフラにおける確立された AI アプリケーションと潜在的な AI アプリケーションの例を示す。[3] 予測モデルなどの機械学習技術は通常、運用層（0～3）で使用される一方、LLM は通常、ビジネスコンテキスト（4～5）で使用され、OT ネットワークからエクスポートされたデータ上で動作する可能性がある。

表 *1.* パデューモデルに基づく *AI* アプリケーション

レベル	説明	AI の活用例
レベル 0： フィールドデバイス	物理プロセスと相互作用するセンサー、アクチュエーター、その他のデバイス。	OT データソース：フィールドデバイスは、AI モデル（主に予測型機械学習モデル）のトレーニングや重大な逸脱の識別に利用できる OT データを生成する可能性がある。
レベル 1： ローカルコントローラ	プロセス、セル、ラインの自動制御を提供する装置やシステム。例としてはプログラマブルロジックコントローラ（PLC）や遠隔端末装置（RTU）がある。	ローカル制御のための AI：一部の最新 PLC やエッジコントローラは、ローカル異常検知、負荷分散、既知の安全状態維持などのタスク向けに、軽量な事前学習済み予測モデルを分類として実行する。
レベル 2: ローカル監視	個々のプロセス、ライン、セルに対する監視と管理監督。例としては、監視制御とデータ収集（SCADA）システム、分散制御システム（DCS）、ヒューマンマシンインターフェース（HMI）などがある。	品質管理：SCADA システムや DCS からのデータを分析し、設備異常の初期兆候を検知し、是正措置が必要かもしれないことをオペレーターに警告するために、AI モデル（主に予測型機械学習モデル）が使用されることがある。
レベル 3： サイト全体の監視	企業全体またはその一部をカバーする地域全体の監視、監督、運用支援。例としては、製造 実行システムやヒストリアンが挙げられる。	予知保全：AI モデル（主に予測 ML モデル）は、集約されたヒストリアン OT データを分析し、設備の保守要件を予測するために使用されることがある。 オペレーターの意思決定支援：AI モデルはローカル監視システムに統合され、運用測定値などのシステム推奨事項を提供することでオペレーターの意思決定を支援する。

[3] 本ガイダンスで使用したパデューモデルのバージョンは、Manuel Humberto Santander Pelaez,「Controlling Network Access to ICS Systems」, *Diaries* (blog*), SANS Technology Institute Internet Storm Center*, 2023 年 7 月 3 日, https://isc.sans.edu/diary/30000 から引用した。

TLP:CLEAR

レベル	説明	AI の活用例
レベル 4 および 5： エンタープライズネットワークおよびビジネスネットワーク	ビジネスおよび企業プロセスと意思決定を管理する IT システム。重要インフラおよび OT の文脈では、OT データの分析や OT・IT システム双方の自律防御などが例として挙げられる。	ワークフロー最適化：AI システム（AI エージェントや LLM を含む）は、ビジネスユースケースとエンジニアリングの接点など、ビジネスプロセスの改善に活用できる。 OT データと IT データの行動分析とプロファイリング：AI は、運用測定、異常・脅威検知、緩和の決定、優先順位付けされたレジリエンシー決定を支援する情報の提供のために、OT データと IT データを組み合わせて分析するために使用できる。

TLP:CLEAR

OT における AI の安全な統合に関する原則

原則 1 – AI の理解

1.1 AI の固有リスクと OT への潜在的影響の理解

以下のセクションでは、AI 統合のリスクと OT 運用への潜在的影響について論じる。表 2 は、重要インフラの所有者および運営者が考慮すべき既知の AI リスクの概要を示す。（注：これは網羅的なリストではない。重要インフラの所有者および運営者は、自組織固有のリスクを調査すべきである。）本ガイダンスの続くセクションでは、これらのリスクに対する緩和策について論じる。表 2 の「緩和」欄の参照先を参照のこと。

表 2. OT 環境における AI のリスクと影響

OT 環境における AI リスク	OT への影響	緩和
サイバーセキュリティリスク：AI データ、モデル、展開ソフトウェアが操作され、誤った結果を引き起こしたり、セキュリティ対策や機能安全対策、ガードレールを迂回したりする可能性がある。従来のサイバーセキュリティリスクは AI システム内にも存在するため、アクセス管理、監査、暗号化などのセキュリティ対策は、AI および AI 搭載システムの保護に依然として適用される。加えて、AI 搭載システムはプロンプト・インジェクションなどの AI 特有のサイバーセキュリティリスクにも晒される。	影響を受けるシステム可用性、機能安全リスク、財務的損失、評判の毀損、ネットワーク/OT 侵害、連鎖的侵害。	1.2 安全な AI システム開発ライフサイクルの理解 2.4 AI-OT システム統合における課題の評価 3.3 AI のテストと評価の徹底的な実施
データ品質：AI モデルの有効性は、そのトレーニングデータの品質に依存する。特に分散型 OT 環境では、高品質で正規化されたセンサーデータの収集は困難である。この運用データを一元化すると、脅威アクターがより標的を絞った技術的影響を生み出すために利用するリスクが生じる。	データ品質の低さによる OT の安全性およびシステム可用性の低下。	2.2 AI システムにおける OT データセキュリティリスクの管理
AI モデルドリフト：AI モデルは、初期トレーニングデータに存在しないデータがモデルに導入されることで、時間の経過とともに精度が低下する可能性がある。	変更への依存度増加、生産性低下、OT の安全性およびシステム可用性の低下。	4.1 OT における AI の監視・監督メカニズムを確立する

TLP:CLEAR

OT 環境における AI リスク	OT への影響	緩和
生産プロセスの変更はモデル性能に影響を与える。		
説明可能性の欠如：AI モデルの意思決定プロセスを理解することは困難であり、これによりエラーの診断や修正、システムの適切な監査が難しくなる。	復旧時間の増加、機能安全リスク、システム可用性の低下、トラブルシューティングの複雑化。	1.3 従業員への AI 教育 4.1 OT における AI の監視・監督メカニズムを確立する
オペレーターの認知負荷と不要なダウンタイム：AI はアラーム誤作動を引き起こす可能性があり、不要なダウンタイムや安全インシデントの原因となる。こうしたアラーム誤作動は認知負荷を増大させ、オペレーターの注意を散漫にし、さらなる人的ミスを招く恐れがある。	システム可用性の低下、機能安全リスク、財務的損失、評判の毀損。	1.3 従業員への AI 教育 4.1 OT における AI の監視・監督メカニズムの確立
規制順守：OT の安全性やプライバシーなどに関する規制要件への順守は、AI、技術標準、規制の枠組みが絶えず変化しているため、困難な場合がある。例えば、AI による意思決定の堅牢な監査証跡を作成することは難しいかもしれないが、規制順守のために必要となる場合がある。	機能安全リスク、財務的損失、評判の毀損。	3.4 OT における AI の規制・コンプライアンス上の考慮事項の対応 4.1 OT における AI の監視・監督メカニズムの確立 4.2 安全性とフェイルセーフ機構の組み込み
AI 依存性：AI への過度の依存は、AI が重要な安全関連情報を見逃した場合にオペレーターも同様に見逃す可能性があり、AI 機能なしに手動で機器を安全に操作するための貴重なスキルを失うことにつながる。	技術への依存、トラブルシューティングの複雑化。	1.3 従業員への AI 教育
相互運用性の問題：AI システムを既存の OT インフラに統合する際、OT コミュニケーションプロトコルやデータ形式の違いから生じる相互運用性の課題により複雑化することがある。	保守コストの増加、復旧の困難さ。	2.1 AI 利用の OT ビジネスケースの検討 2.4 AI-OT システム統合における課題の評価 3.1 OT における AI のガバナンス体制を確立する

TLP:CLEAR

OT 環境における AI リスク	OT への影響	緩和
		3.3 AI の徹底的なテストと評価を実施する
複雑性：AI を組み込むには、通常、プロセス自動化をサポートするためにシステム全体の複雑性を高める必要がある。	機能安全リスク、トラブルシューティングの複雑化。	2.1 AI 利用の OT ビジネスケースの検討 2.4 AI-OT システム統合における課題を評価する
信頼性：AI は産業環境において重要な決定を単独で行うのに十分な信頼性を備えていない可能性がある。AI は幻覚（つまり、もっともらしいが虚偽の応答やデータを捏造すること）を起こすこともあり、これによりオペレーターは意思決定のための誤った情報を提供されることになる。したがって、LLM などの AI は、OT 環境における安全に関する決定を行うために使用すべきではない。	AI 開発者による意思決定は、OT の安全性と信頼性リスク、文書化コストの増加、自動意思決定の経時変化による不確実性、行動の緊密な連動による連鎖的故障リスクの増大をもたらす可能性がある。 意思決定者に提供される虚偽の情報は、安全でない運転状態、設備損傷、生産停止のリスクをもたらす。	2.1 AI 利用の OT ビジネスケースの検討 3.1 OT における AI のガバナンス体制を確立する 3.2 既存のセキュリティおよびサイバーセキュリティフレームワークへの AI 統合 3.3 AI の徹底的なテストと評価を実施する 4.1 OT における AI の監視・監督メカニズムを確立する 4.2 安全性とフェイルセーフ機構を組み込む

1.2 安全な AI システム開発ライフサイクルの理解

OT 環境に AI を統合する際の特有の課題に対処するため、重要インフラの所有者および運営者は、AI システムが安全に設計されていることを確認し、AI システムのライフサイクルを通じて自らの役割と責任を理解すべきである。クラウドシステムで用いられるハイブリッド所有モデルと同様に、所有者と運営者は、AI システム製造事業者、OT サプライヤー、およびシステムインテグレーターやマネージドサービス・プロバイダーの役割について、これらの役割と責任を明確に定義し、 伝達しなければならない。

NCSC-UK と CISA の共同ガイドライン「安全な AI システム開発のためのガイドライン」は、AI システム開発ライフサイクルの以下の主要段階を強調している：[4]

- **安全な設計**。堅牢なコーディング、プロトコル、データ保護対策を含む、セキュリティを考慮した AI システムの設計を初期段階から行うこと。

[4] 英国政府の「AI サイバーセキュリティ実践規範」とその技術的実施ガイドも、安全な AI システム開発ライフサイクルに基づくシナリオベースのサイバーセキュリティ緩和アドバイスを提供している。

TLP:CLEAR

- **安全な調達または開発**。安全な慣行を遵守するベンダーを選定し、安全な手法とツールを用いて AI システムを開発する。
- **安全な展開**。適切なネットワークセグメンテーションやアクセス管理の使用、AI システムが意図した通りに動作することを検証・確認するなど、セキュリティ態勢を維持する方法で AI システムを展開する。
- **安全な運用と保守**。定期的な更新やパッチ適用、潜在的な脆弱性の監視などを通じて、AI システムがライフサイクル全体を通じて安全に運用され続けることを保証する。

重要インフラの所有者および運営者は、AI システムの調達方法に関する異なる手法間のトレードオフを慎重に評価すべきである：

- **AI システムの調達**。特定のセキュリティ要件を満たし、OT サプライヤーが同意するベンダーから、事前に開発された AI システムを選択する。
- **AI システムの開発**。自社内で AI システムを構築する。これにより設計と実装を完全に制御できる。
- **既存 AI システムのカスタマイズ**。ベンダーと連携し、既存の AI システムを特定の OT システムの要件に合わせて調整する。

可能な限り、重要インフラの所有者および運営者は、設計段階で安全性が確保され、OT の運用や安全性に悪影響を与えない AI システムを要求すべきである。重要インフラの所有者および運営者は、CISA の「設計段階での安全性確保」ウェブページとリソース、ならびに共同ガイダンス「要求段階での安全性確保：運用技術所有者および運営者がデジタル製品を選択する際の優先事項」を参照し、これらの原則を AI および OT システムの設計に組み込む機会を検討すべきである。

1.3 従業員への AI 教育

AI を OT 環境に統合すると、担当者が自動化に過度に依存し、人的監視や状況認識が低下する可能性がある。これにより、以下のような重大な結果が生じる恐れがある：

- **依存リスクと技能の衰退。**AI への過度な依存は、AI の故障やシステム停止時に必要な手動操作技能を OT 担当者が失う原因となる。
- **スキルギャップ**。OT 担当者は AI の出力を誤解釈し、誤った行動を取る可能性がある。また、AI システムが故障した場合、その管理やトラブルシューティングに必要な専門知識を欠いている場合もある。

重要インフラの所有者や運営者は、以下のようなスキル開発や分野横断的な連携に注力することで、これらのリスクを緩和できる：

- AI の基本原理と脅威モデリングに関する OT チームへの教育。これによりチームは AI 出力を効果的に解釈・妥当性確認し、AI システムと並行して運用能力を維持できる。例えば、AI 出力の妥当性確認に代替センサー（人間の感覚、振動・温度センサー、電圧測定値など）を使用する訓練や、AI 出力が無効な場合の対応策を習得させる。
- すべての運用（AI 関連運用を含む）、介入、インシデントに対する明確な標準業務手順書（SOP）を策定し、AI 対応 OT システムの管理における各関係者の役割と責任の認識を支援する。

TLP:CLEAR

- 説明可能な AI を活用し、オペレーターが AI 出力に意思決定プロセスの明確かつ透明性のある文書化を求めるようにする。これにより、人間は出力をより良く理解し、妥当性確認できるようになる。

TLP:CLEAR

原則 2 – OT 分野における AI 利用の検討

2.1 AI 利用の OT ビジネスケースの検討

重要インフラの所有者および運用者は、AI システムを OT 環境に組み込む前に、他の技術と比較して AI 技術が特定のニーズや要件に最も適した解決策であるかどうかを評価すべきである。重要インフラの所有者および運用者は、より複雑で新規の AI 対応ソリューションを追求する前に、確立された機能がニーズを満たすかどうかをさらに検討すべきである。AI には独自の利点がある一方で、リスクの継続的な評価を必要とする進化中の技術である。

このアセスメントには、セキュリティ、性能、複雑性、コスト、特定の用途に応じた OT 環境の安全性への影響など、様々な要素を含めるべきである。また、AI 技術の使用による利点とリスクを、その用途が満たすべき機能に対して評価する必要がある。 重要インフラの所有者および運営者は、自組織が OT 環境において AI システムを維持する現在の能力、ならびに環境のリスク表面を拡大することによる潜在的な影響（例えば、モデルによるデータ処理のための追加ハードウェア・ソフトウェアの必要性、拡大した攻撃対象領域を防御するための追加セキュリティインフラなど）を理解すべきである。

評価の結果、AI システムが最適な解決策であると判断された場合、重要インフラの所有者および運営者は、上記の安全な AI システム開発ライフサイクルに従い、米国国立標準技術研究所（NIST）の AI リスクマネジメントフレームワークなどの AI リスクマネジメントフレームワークを参照し、システムが安全かつ確実に使用されるよう確保すべきである。

OT ビジネスユースケースにおける AI アセスメントの例

以下の例は、産業用発電機の予知保全支援に AI を活用する実現可能性について、重要インフラ組織が実施した架空の評価を示す。この評価には、重要インフラの所有者および運営者が OT 環境での AI 活用を推奨するために技術が満たすべき性能閾値と組織の安全・セキュリティ要件が含まれる。

ユースケース：産業用発電機の予知保守に AI システムを活用する。

問題点：産業用発電機の故障に伴う高いダウンタイムと保守コスト。

目標：発電機の潜在的な故障を 30 日前に検知する AI 搭載の予知保全ソリューションを導入する。

リスク：AI が予知保守を正しく行わない場合、設備が早期に交換される可能性がある。

主要関係者：

- 発電機の保守計画を担当する運用チーム。
- 発電機の修理・交換を行う保守要員。
- 発電機の停止時間と保守コストの影響を受ける設備所有者。

要件：

- 発電機の性能に関する履歴データ（センサー測定値や保守記録を含む）へのアクセス。

TLP:CLEAR

- 大規模な OT データセットをほぼリアルタイムで処理する能力。
- 信頼性の高い故障予測検知。

対応する安全・セキュリティ要件：

- 集計データは転送中および保存時に防御され、アクセスや変更はすべて記録される。
- データ集約はネットワーク帯域幅の 80%閾値を超えない。
- 誤警報は稀であり、文脈を提供するため、対応可能であり続け、オペレーターの疲労を防ぐ。

成功指標：

- ダウンタイムを 25%削減する。
- 保守コストを 15%削減。
- 総合設備効率（OEE）が 10%向上した。

2.2 AI システムにおける OT データセキュリティリスクの管理

データ関連の課題

AI を OT 環境に統合する際、重要インフラの所有者および運営者は、以下のデータ関連の課題に対処する際に AI モデル開発者と協力すべきである：

データ保証。AI モデルの訓練に使用される OT データがどこに保存されているかを把握し、それが組織の管理下にあることを確認する。OT データへのアクセスを安全に管理し、閲覧・アクセス・変更可能な者を限定する。AI ベンダーが組織の OT データにアクセスし利用する方法を理解する。特に遠隔・クラウド・オフショアアクセスが関与する場合は注意が必要だ。

データ主権。外国に拠点を置く企業は外国政府の管理や法律の対象となり、自社の利益に反する指示に従わなければならない可能性があることを認識すること。

機密情報のエクスポージャー。特にパブリッククラウドインフラなど外部主体がホストまたは管理する環境で運用される AI モデルには、機密データを共有しないことでリスクを最小限に抑える。

データ・プライバシーとセキュリティ。OT データセット内の専有情報や個人データを防御すること。これには、アクセス乱用、意図的または偶発的なデータ・ポイズニング、合成データや生成データへの依存からの防御機構の導入が含まれる。

データサイロ。OT/IT ネットワークの分離、独自プロトコル・フォーマット、OT 製品の多様性、多数のサプライヤーなどによる、OT 環境への AI システム統合の複雑さに対処する。

データ品質と可用性。効果的な AI 性能のために、高品質で包括的なデータセットをキュレーションする専門的なドメイン知識を適用する。これは産業環境では困難であり、潜在的な課題には、専有システムや旧式システム、特注ソリューションを含む環境、安全関連のエッジケースを捕捉する難しさなどが含まれる。この専門知識を捕捉するには

TLP:CLEAR

OT オペレーターの専門性が不可欠である。さらに、オペレーターは AI モデル開発者と連携し、データ完全性プログラムが AI システムをカバーするよう確保すべきである。

OT データの防御の優先順位付け

以下の重要な種類の OT データの防御を優先する。

エンジニアリング構成データ。これにはネットワーク図、資産インベントリ、操作シーケンスに関する文書、安全関連情報、ロジック図、回路図などが含まれる。これらのデータは永続的な価値を持ち、サイバー攻撃者にとって極めて高価値である。

一時的な OT データ。 産業用計測技術、特にプロセス計測技術（電圧・温度・圧力レベル・質量/体積流量など）からのデータは、組織活動やシステム動作の洞察を提供し得る。このデータが AI モデルの訓練や更新に使用された場合、データはモデル内でより長期間アクセス可能または（統計的に）保存される可能性がある。したがって、これらのデータポイントを保護することは、知的財産（IP）や活動パターンの保護において重要となり得る。[5]

2.3 AI 統合における OT ベンダーの役割の理解

OT ベンダーは、OT 環境への AI 統合を推進する上で重要な役割を担っている。現在、一部の OT デバイスには AI 技術が組み込まれており、機能させるためにインターネット接続が必要となる場合がある。

OT ベンダーにおける新たな動向には以下が含まれる：

- **オペレーター向け AI。**OT ベンダーは、グリッド周波数ダイナミクスの予測モデルなど、AI 機能をデバイスに直接統合するケースが増えている。
- **インテリジェントデバイス。**技術の進歩に伴い、制御の設計や変更に AI 機能を備えた高度な「インテリジェントデバイス」の出現が予想される。

重要インフラの所有者や運営者は、OT ベンダーに対し、自社製品への AI 技術組み込み方法に関する透明性とセキュリティ配慮を求めるべきだ。具体的には：

- **契約上の合意。**OT ベンダーが AI の特性と機能について明確な詳細を提供するよう保証する契約上の合意を交渉すること。
- **ソフトウェアサプライチェーンとソフトウェア部品表（SBOM）**。OT ベンダーに対し、自社 OT 製品への AI 組み込み方法に関する情報提供を義務付ける（ による SBOM 提供要請を含む）。[6] また、製品で使用される AI モデルのサプライチェーン（例：モデルのホスティング場所）を明確化させる。

[5] OT データの保護に関する詳細情報は、ASD の ACSC による「運用技術サイバーセキュリティの原則」および NSA の共同ガイダンス「AI データセキュリティ：AI システムの訓練・運用に使用されるデータを保護するためのベストプラクティス」を参照のこと。

[6] AI システムはソフトウェアシステムであり、AI ソフトウェアの SBOM（ソフトウェア部品表）の最小構成要素は、全てのソフトウェアの最小構成要素と同じである。詳細は CISA の「2025 年版ソフトウェア部品表の最小構成要素」を参照のこと。G7 サイバーセキュリティ作業部会が作成した「人工知能向けソフトウェア部品表に関する G7 共通ビジョン」では、AI システムを記述するための追加要素の可能性について議論されている。

TLP:CLEAR

- **ベンダー通知。**製品のライフサイクルにおける一般的な脆弱性開示方針に加え、AI ベンダーは、AI が不適切な助言を提供したり不適切な行動を取ったりする可能性を示す兆候を発見した場合、その事実を開示すべきである。
- **明示的なデータ利用ポリシーの審査と実施**。運用データには知的財産や機密情報が含まれる可能性があるため、事業者はベンダーが運用データで AI を訓練することを望まない場合がある。事業者は、データ保管場所、コミュニケーション経路、暗号化、保存方法について明確に規定したデータ利用ポリシーを通じて情報を管理できる。
- **接続性の強化**。ベンダーがオンプレミス環境で運用可能か、あるいはインターネット／ベンダークラウドへの常時接続なしに動作可能かを問う。
- **AI 機能の無効化**。特定の AI 機能を無効化または有効化できる条件を明示し、その制御権を事業者に与えること。

OT ベンダーに透明性と制御を要求することで、重要インフラの所有者や運営者は、OT 製品に組み込まれた AI システムに関連するリスクをより適切に管理できる。

2.4 AI-OT システム統合における課題の評価

AI を OT 環境に統合する際、重要インフラの所有者および運営者は、互換性とセキュリティを確保するため、既存のインフラを慎重に評価すべきである。

AI-OT システム統合における課題

組織が OT システムに AI を統合する際に直面する可能性のある課題には以下が含まれる：

- **システムの複雑性と脆弱性の増大**。AI の統合は OT システムに複雑性を加え、悪意のある攻撃者が悪用できる新たな攻撃対象領域や脆弱性を生み出す可能性がある。これは、AI が新たに遠隔アクセス可能なインターネットに公開された攻撃経路を導入する場合に特に重要な考慮事項である。
- **クラウドセキュリティリスク。**AI をクラウド上の監視制御・データ収集（SCADA）環境に統合すると、追加のサイバーセキュリティリスクが生じたり、データ伝送の遅延が発生したりする可能性がある。
- **前方互換性と後方互換性。**多くの OT 環境は標準データ形式やコンピューティング機能を持たない旧式機器に依存しているため、AI の安全な統合には OT システムの設計変更が必要となる場合があり、AI データの統合と分析を複雑化する。
- **遅延とリアルタイム制約。**AI システムは OT 環境の厳格なタイミング要件を満たせない可能性がある。これらの要件は業種によって大きく異なり、AI システムが制御プロセスに積極的に関与する場合、より重要となる。
- **AI ベンダーの透明性。**ベンダーの透明性不足は、外部接続を行う機能や標準エンジニアリングワークフローを変更する機能に関する知見を得ることを妨げる可能性がある。

これらの課題の緩和のため、重要インフラの所有者および運営者は以下の対応を取るべきだ：

TLP:CLEAR

- AI システムを総合的なセキュリティおよびサイバーセキュリティフレームワークに統合する（3.2「既存のセキュリティ・サイバーセキュリティフレームワークへの AI 統合」参照）
- データ暗号化、アクセス管理、侵入検知システムなどの従来のサイバーセキュリティ対策も含む包括的なセキュリティ戦略に、AI セキュリティに関する考慮事項を追加すること。
  - 重要インフラの所有者および運営者は、クラウド契約におけるセキュリティ条項を定義し、妥当性確認を行うべきである。これには、データ保護、アクセス管理、インシデント対応、監査機能を含む、AI セキュリティ責任、コンプライアンス標準、サポート規定を明示的に記載する必要がある。
  - クラウドプロバイダは、従来のクラウドセキュリティにおける責任分担モデルに加え、AI 機能に特化したセキュリティ義務を明記した詳細な文書を提供すべきである。
- 既存の OT インフラを考慮し、AI システムとの統合計画を評価・策定すること。
  - 可能であれば、本番システムへの展開前にテストインフラの利用を検討すること（原則 3「AI ガバナンスおよび保証フレームワークの確立」を参照）。
- OT ネットワークへの永続的なアクセスなしに、データが OT ネットワークから AI システムにプッシュされるプッシュ型アーキテクチャを推奨する。[7]
- AI システムをローカルまたはクラウドでホストする場合、AI システムが統合または有効化する可能性のある重要機能に対する組織の制御を優先する。
  - AI 対応システムのプロセスには、従来の自動化または手動操作に切り替わるフェイルセーフ機構を必ず設けること。
- AI システムを新規 OT システムと同様に統合する：安全への影響（例：遅延、相互運用性）をテストし、既存のデバイス管理ポリシー内で動作することを確認する（例：リモートアクセスなどの新規接続経路が承認され、既存の非武装地帯（DMZ）またはジャンプホストインフラを通過することを確認する）。
- 安全上の懸念や遅延の制限を考慮し、人間の介入なしに AI が OT インフラを能動的に制御することを制限する。
- AI モデルの精度と有効性を定期的に更新・妥当性確認する。

[7] 詳細については、ASD の ACSC による「運用技術サイバーセキュリティの原則」を参照のこと。

TLP:CLEAR

原則 3 – AI ガバナンスと保証フレームワークの確立

3.1 OT における AI ガバナンスメカニズムの確立

AI を OT 環境に安全かつ確実に統合するには、効果的なガバナンス構造が不可欠である。これには、OT 内の AI 意思決定プロセスに関する明確なポリシー、手順、責任体制の確立が含まれる。AI ガバナンス構造には、以下の主要な利害関係者、および調達、開発、設計、展開、運用段階における監視維持に必要な AI ベンダーを含めるべきである。

AI ガバナンス機構における主要ステークホルダー

経営陣。CEO や CISO を含む上級管理職のコミットメントを確保することは、強固な AI ガバナンスフレームワークを確立するために不可欠である。これにより、組織のリーダーシップが AI システムの安全なライフサイクル管理に完全にコミットし、AI の機能性とともにリスクと緩和を考慮することが保証される。

OT/IT 専門家。OT、IT、AI の各分野の専門家を巻き込むことは、AI システムを OT 環境に効果的かつ安全に統合するために重要だ。これらの専門家は OT 環境に関する貴重な知見を提供し、AI 統合に伴う潜在的なリスクや課題を識別するのに役立つ。

サイバーセキュリティチーム。AI モデルが使用する機密性の高い OT データを防御する方針と手順を策定するには、サイバーセキュリティチームとの連携が不可欠だ。サイバーセキュリティチームは潜在的な脆弱性を識別し、組織データのセキュリティ維持に役立つ緩和の提案を行うことができる。

AI ガバナンス機構における追加要素

ガバナンス構造のその他の主要な構成要素には以下が含まれる：

- AI モデルが使用する機密性の高い OT データを防御する厳格なデータガバナンスポリシーの実施。これには暗号化、アクセス管理、ユーザー行動分析が含まれる。
- AI システムの開発、展開、運用保守に関わる全員（データ所有者、モデル開発者、エンドユーザーなど）が、各自の任務と期待事項を理解し、安全または運用上のインシデント発生時に責任の所在が不明確になることや責任の所在を巡る混乱を避けるための、明確な役割と責任の確立。
- 定期的な監査とコンプライアンステストを実施し、潜在的な問題を識別するとともに、AI ガバナンス要件への継続的な順守を確保すること。
- AI システムの性能を継続的に妥当性確認し、組織の目標と規制要件を満たしていることを確認する。

3.2 既存のセキュリティおよびサイバーセキュリティフレームワークへの AI の統合

OT 環境に AI を統合する際、重要インフラの所有者および運営者は、これらのシステムを統制する既存のセキュリティおよびサイバーセキュリティフレームワークを考慮し、既存のリスクアセスメント、緩和、監視 プロセスに AI システムアセスメントを組み込むべきである。これは、AI システムを統合する際には、従来のサイバーセキュリティ要件、脆弱性管理、重要インフラ規制を考慮に入れる必要があることを意味する。これらのプロセスには以下が含まれる：

TLP:CLEAR

- 定期的なセキュリティ監査とリスクアセスメント。AI システム内の潜在的な脆弱性を識別するため、定期的なセキュリティ監査とリスクアセスメントを実施するか、AI ベンダーが実施している証拠を入手する。
- 堅牢なセキュリティ制御。暗号化、アクセス制御、侵入検知システムなど、AI システムとデータの異常を防御・検知する堅牢なセキュリティ制御を実施する。[8] ・AI エンドポイントのフローログとアクセスログを収集し、資産と ID ごとのデータ流出を追跡する。・プロンプトと出力検査のためのデータ損失防止（DLP）と統合する。
- AI に特化したセキュリティ情報。セキュリティチームは、リスク評価や脅威モデリングにおいて、AI 関連の戦術・技術・手順（TTP）を組み込むべきである。 例えば、脅威アクターの行動マッピングに MITRE ATT&CK®Matrix for Enterprise を使用する場合、チームは MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™) Matrix などのツールを用いて、AI 対応システムに対する TTP に特化した AI 関連の TTP も組み込むべきである。

3.3 AI のテストと評価の徹底的な実施

OT 環境に AI を導入する際、システムの安全かつ信頼性の高い運用を支えるためには、徹底的なテストと評価（T&E）が不可欠である。運用者はまず、テスト用に特別に設計されたインフラ上で AI システムのテストを実施すべきだ。 低忠実度テストは、T&E プロセスの初期段階でテストの反復を迅速化できる。あるいは、ベンダーと連携し、彼らのテスト内容や依存関係（例：OS バージョン、プロトコル）の有無を確認する。システムが成熟するにつれ、運用者はハードウェアインザループを含む、より現実的な非本番環境でのテストを実施できる。

運用者は、非本番環境で十分なテストを行った後にのみ、AI システムを本番環境に移行して追加テストを実施すべきだ。物理的な影響をモデル化する必要がない場合、仮想化コントローラーはこのテストプロセスを加速できる。

重要インフラの所有者および運用者は、AI のテストおよび評価を実施する際にも、従来型のデータ保護メカニズムを遵守すべきである。例えば、非本番環境での本番データエクスポージャーを避けることなどが挙げられる。

3.4 OT における AI の規制・コンプライアンス上の考慮事項

重要インフラの所有者や運用者が AI 技術を OT 環境に統合するケースが増えるにつれ、規制とコンプライアンスの考慮事項が主要な課題となっている。具体例としては以下が挙げられる：

- OT 向け AI 標準の欠如。現在の国際的な AI 技術標準は、主に IT 環境への AI システム展開に焦点を当てている。
- 監査可能性。AI の意思決定を追跡または説明することは困難であり、規制監査を複雑化する可能性がある。
- 安全認証。AI システムは重要インフラや OT 環境で要求される厳格な安全標準を満たさない可能性がある。

[8] AI システムおよびデータにおける防御と異常検知のための制御の使用に関する詳細は、NSA の共同サイバーセキュリティ情報シート「AI データセキュリティ」および NCSC-UK の共同ガイドライン「安全な AI システム開発のためのガイドライン」を参照のこと。

TLP:CLEAR

重要なインフラの所有者や運営者は、AI 標準が急速に進化していることを踏まえ、自社の OT 領域における現行 AI 標準の適用可能性を評価すべきだ。欧州電気通信標準化機構（ETSI）の人工知能セキュリティ技術委員会が策定した主要な AI 標準を以下に示す：

- **ETSI TR 104 128** 人工知能のセキュリティ確保（SAI）；AI モデルおよびシステム向けサイバーセキュリティガイド
- **ETSI TS 104 223** 人工知能のセキュリティ確保（SAI）；AI モデルおよびシステムのためのサイバーセキュリティ基本要件
- **ETSI TR 104 048** 人工知能のセキュリティ確保（SAI）；データ供給チェーンのセキュリティ

重要インフラの所有者および運営者は、AI システムの性能が厳格な OT 性能および安全規制を満たしていることを継続的に妥当性確認し、確認すべきである。また、重要インフラの所有者および運営者は、AI システムの出力が性能および安全の閾値を下回った場合など、OT において非 AI システムにデフォルトで切り替えるための閾値を識別し、展開すべきである。

TLP:CLEAR

原則 4 – AI および AI 対応 OT システムへの監視・フェイルセーフ手法の組み込み

最終的に、機能安全の責任は人間が負う。人間は監視、安全、フェイルセーフ慣行を確保または運用するためのツールを作る。これは AI ツールにおいても同様である。

4.1 OT における AI の監視・監督メカニズムの確立

重要インフラの所有者および運営者は、AI 搭載 OT システムに対する監視を実施すべきである。具体的には、AI コンポーネントおよび AI に依存するその他のコンポーネントをすべて把握することだ。 これらのコンポーネントの入力と出力を記録・監視する。また、OT 環境における安全な動作のための既知の良好な状態や閾値を設定・維持し、バックアップからの保守や復旧をいつ実施すべきかを把握できるようにする。AI 対応 OT システムに安全プロセスを組み込む際には、効果的な実装と管理を確保するため、以下の点を考慮する：

人間が関与する意思決定。特に重要な OT 運用や操作においては、オペレーターやエンジニアが意思決定に関与できるよう十分な透明性をプロバイダすることが必要である。受動的な AI システムの場合、オペレーターやエンジニアは既存の変更管理プロセスに推奨事項を組み込むことでこれを実現できる。 制御に直接影響を与える能動型 AI システムには注意が必要だ。オペレーターが問題に気付く前に事態が悪化する可能性があるためである。AI が制御ロジックを能動的に更新する場合、安全閾値、代替センサー出力、または人間の介入ポイントを追加する状態変化を利用すべきだ。

人間がループ内にある意思決定の利点は以下の通りである：

- OT 環境の安全性向上。オペレーターは AI 駆動システムと対話する際に文脈を提供し、情報に基づいた意思決定を行う。
- 信頼性の向上。人間の監視により、AI の性能における潜在的なエラーや異常を検知・修正できる。監視は緊急時に不可欠な人間の訓練も維持する。
- 信頼性の向上。組織はオペレーターの関与を通じて、AI システムとその意思決定プロセスに対する信頼を構築する。

OT 環境におけるシステムの継続的な安全運転を支えるため、AI システム結果の正確性を理解する。重要インフラの所有者・運用者は、AI システムが正確かつ信頼性の高い結果を出せない状態を理解することが不可欠である。この理解には、システム性能における偽陽性・偽陰性の発生予測、および偽陽性が真陽性の基本発生率とどう比較されるかが含まれる。

異常検知と行動分析を導入する。AI ドリフト、安全性と性能に影響するモデル変更、またはセキュリティリスクを検知する OT デバイスの安全動作限界を確立する。オペレーターのプロセスが成熟するにつれ、ソフトウェアの安全閾値は設定値から、より複雑化する故障の異常検知へと移行できる。コンプライアンスとフォレンジック分析のために AI の決定を追跡できるよう、また記録される AI の識別子が通常の機械やユーザー識別子とは明確に区別されるよう、ロギングを設定する。

以下の例は、オペレーターとエンジニアがデータゾーン内の予測保守システム（OT データ出力への読み取り専用アクセス権限）を監視する方法を示す。このシステムは機械学習を用いて推奨事項を生成する：

TLP:CLEAR

- AI システム出力。設備のダウンタイムに関する予知保守推奨事項。
- 異常検知アルゴリズム。予測ダウンタイム値の異常値を検知する統計的プロセス制御。
- 監視ツール。AI システムの性能を追跡し異常を検知するリアルタイムダッシュボード、チャート、またはメトリクス。理想的には、統合監視のため既存の HMI（ヒューマンマシンインターフェース）ビューに統合される。
- 監査証跡。AI システムの動作データ（タイムスタンプ、入力、出力を含む）を全て記録し、AI システムの動作を監査・分析する。
- 攻撃的セキュリティアセスメント／AI レッドチーム活動。AI システムの機能を定期的に評価し、脆弱性を識別し、レジリエンスをテストする。
- ネットワークおよびエグレスセキュリティ。ネットワーク制御を定義し実施する。CISA のサイバーセキュリティパフォーマンス目標（CPG）2.F を参照のこと。

AI の効果を測定し、経時的な進捗を追跡する主要業績評価指標（KPI）を設定する。重要インフラの所有者および運営者は、ベンダー、ガバナンス委員会、オペレーターなどの AI 関係者との定期的なレビュー会合を計画し、結果を議論し、懸念事項に対処し、改善点を識別すべきである。詳細は CISA の人工知能ウェブページを参照のこと。

AI モデルは、導入前にシミュレーション環境で継続的に妥当性確認・改良する。脅威モデルは、AI 特有の攻撃ベクトル（敵対的入力やデータ・ポイズニングなど）で定期的に更新し、AI システムの動作を監視して異常や操作の試みを検知する。精度向上と誤検知（偽陽性/偽陰性）削減のため、新たな OT データで AI モデルを継続的に更新・改良する。詳細は NSA の共同ガイド「AI システムの安全な展開」を参照のこと。

新たな AI の説明可能性と透明性ツールを検討する。説明可能な AI（XAI）と透明性のある AI は、AI システムの理解性を高めるための進化する研究分野である。説明可能性は個々の AI 判断の根拠をユーザーに理解させることに焦点を当て、透明性は AI システム全体の開発・運用をオープンかつアクセス可能にすることに重点を置く。本質的に、説明可能性は AI が特定の判断を下した理由を明確にし、透明性は AI システム全体の動作方法に焦点を当てる。 重要インフラの所有者や運営者は、可能な限り、AI の決定を人間が理解しやすいようにする解釈可能なモデルやツールを検討すべきだ。

大規模言語モデル（LLM）、予測型システム、深層学習システムはしばしば不透明に動作し、意思決定の根拠を監査・理解することを困難にする。このような透明性は、OT における安全性と規制順守にとって極めて重要である。XAI（説明可能な AI）や透明性 AI ツールは、AI 開発者が AI モデルの動作を理解するのを支援するために設計されている。しかし、これらの技術が OT 環境に十分かどうかは未解決の問題である。

必要な機能や要約を OT から外部に移すプッシュ型または仲介型アーキテクチャを優先し、永続的なインバウンドアクセスを許可しないこと。データがビジネスネットワークを通過する必要がある場合は、一方向転送パターンと監査可能なステージングバッファを使用する。このセグメンテーション実装手法により、運用者は既存のセグメンテーションベストプラクティスを維持でき、AI システムが OT への永続的な攻撃経路となることを防げる。

TLP:CLEAR

4.2 安全性とフェイルセーフ機構の組み込み

AI システムが重要な運用を妨げずに正常に停止できるフェイルセーフ機構を確立する。既存の機能安全およびインシデント対応プロセスに、AI システムのバイパスや代替方法を含む新たな故障状態を組み込む。AI システムを既存の OT ネットワークに統合すると、必然的に重要インフラシステム全体に新たな故障状態が生じる。したがって、既存の機能安全およびインシデント対応プロセスの改訂を担当する運用者は、これらのシステムの安全な運用を確保するために極めて重要であるこれらの新たな故障状態を組み込むべきである。

AI システムを考慮した機能安全手順を設計する。各重要インフラ分野には固有の安全状態と手順が存在する。原則 2「OT 領域における AI 利用の検討」に基づき、重要インフラの所有者・運営者は、既存手順への AI システム統合方法を検証し、OT 環境への AI システム統合に焦点を当てた新たな安全な利用・実装手順を作成すべきである。

サイバーセキュリティインシデント対応計画に AI 関連事項を組み込む。組織が最大限の緩和策を講じてもリスクをゼロにすることは不可能であり、インシデントは避けられない。この点を考慮し、重要インフラの所有者・運営者は、AI システムに対する悪意ある活動や AI システムの故障に対応する手順を、インシデント対応計画および機能安全手順に追加すべきである。 AI サイバーセキュリティリスクに関するステークホルダーとの連携、および AI サイバーセキュリティインシデントや脆弱性に関する自主的な情報共有に関する詳細なガイダンスについては、CISA の「AI サイバーセキュリティ協力プレイブック」を参照すること。展開される AI システムの数が増えるにつれ、既存の IT システムや OT システムに対する AI を活用した攻撃の数も増加するだろう。

TLP:CLEAR

結論

AI を OT に統合することは、重要インフラの所有者および運営者にとって機会とリスクの両方をもたらす。AI は効率性、生産性、意思決定を向上させ得る一方で、OT システムの安全性、セキュリティ、信頼性を支えるために慎重な管理を必要とする新たな課題も生じさせる。 OT システムへの AI 統合リスクを効果的に緩和するには、重要インフラの所有者・運営者が本指針の原則に従うことが不可欠である。具体的には、AI を理解し、OT 領域における AI 活用を検討し、AI ガバナンスと保証のフレームワークを確立し、安全性とセキュリティの実践を AI および AI を活用した OT システムに組み込むことである。 これらの原則を順守し、AI モデルを継続的に監視、妥当性確認、改良することで、重要インフラの所有者や運営者は、重要な公共サービスを制御する OT 環境への AI の統合をバランスよく達成できる。

TLP:CLEAR

参考資料

読者は、本ガイダンスで議論された AI、AI と OT セキュリティ、および関連トピックに関する追加情報を以下のリソースで確認できる：

- CISA の人工知能ウェブページ
- CISA の AI サイバーセキュリティ協力プレイブック
- NSA の共同ガイダンス「AI データセキュリティ：AI システムの訓練・運用に使用されるデータを保護するためのベストプラクティス」
- NSA の共同ガイダンス「AI システムの安全な展開」
- NCSC-UK と CISA の共同ガイドライン「安全な AI システム開発のためのガイドライン」
- 英国政府の AI サイバーセキュリティ実践規範及びその技術的実施ガイド
- NIST の AI リスクマネジメントフレームワーク
- MITRE の MITRE ATLAS マトリックス
- ASD と ACSC の共同ガイダンス「運用技術サイバーセキュリティの原則」
- CISA のセキュア・バイ・デザインウェブページ
- CISA の共同ガイダンス「Secure by Demand：デジタル製品選定時の運用技術所有者・運用者向け優先事項」
- CISA の 2025 年版ソフトウェア部品表の最低要件（パブリックコメントドラフト）
- G7 サイバーセキュリティ作業部会による人工知能向けソフトウェア部品表に関する G7 共通ビジョン
- 米国法典第 15 編第 9401 条(3)
- 欧州電気通信標準標準化機構（ETSI）技術委員会「人工知能のセキュリティ確保」
- CISA のサイバーセキュリティパフォーマンス目標（CPG）

TLP:CLEAR

免責事項

CISA 及び作成機関は、本文書内でリンクされているものを含め、いかなる商業事業体、製品、企業、サービスも推奨しない。サービスマーク、商標、製造事業者その他の方法で特定の商業事業体、製品、プロセス、サービスに言及しても、CISA 及び作成機関による推奨、推奨、または優遇を構成または示唆するものではない。

本文書は、政策を策定したり、要件を課したり、行動を義務付けたり、既存の法的または規制上の義務を無効にしたりするものではない。本文書に基づいて行われるすべての行動は任意であり、本文書に記載された行動を取る者は、自らの意思でそれを行うものである。

謝辞

BP p.l.c.、Cisco、Darktrace、Fortinet、Nozomi Networks、Palo Alto Networks、およびオーストラリアの重要インフラ組織が本ガイダンス文書の作成に貢献した。

改訂履歴

2025 年 12 月 3 日：初版

TLP:CLEAR

附属書：用語

本資料では技術概念について以下の定義を用いる。ただし、AI および OT の定義は団体によって異なる場合があるため、読者は自国の管轄区域におけるこれらの概念の具体的な定義を理解し、本ガイダンスへの適用可能性を検討すべきである。

人工知能（AI）。AI は以下のように定義できる：[9]

> 人間が定義した特定の目標に対して、現実または仮想環境に影響を与える予測、推奨、決定を行うことのできる機械ベースのシステム。AI システムは、機械および人間からの入力を用いて以下を行う：
>
> - 現実および仮想環境を認識する；
> - 自動化された分析を通じて、こうした知覚をモデルとして抽象化する；そして
> - モデル推論を用いて情報または行動の選択肢を策定する。

AI 概念のさらなる定義については、ASD の ACSC『複雑な層：人工知能入門』を参照のこと。[10]

運用技術（OT）。NIST は OT を以下のように定義している：

> 物理環境と相互作用する（または物理環境と相互作用するデバイスを管理する）プログラム可能なシステムまたはデバイス。これらのシステム／デバイスは、デバイス、プロセス、イベントの監視および／または制御を通じて、直接的な変化を検知または引き起こす。例としては、産業用制御システム、ビル管理システム、消防制御システム、物理的アクセス制御機構などが挙げられる。

AI 対応 OT システム。これは、システム全体に少なくとも 1 つの AI コンポーネントを有する OT システムおよび/または OT ネットワークと定義される。

リスク。NIST の AI リスクマネジメントフレームワークにおける定義によれば、「リスクとは、事象の発生確率と、対応する事象の結果の重大度または程度を総合的に測定したものである」。本ガイダンスにおける AI リスクの定義は、AI をシステムに統合することで、組織（または組織を含むエコシステム）の重要インフラシステムの安全性、セキュリティ、機能に有害な影響を与える事象が発生する確率と、その事象の結果の重大性を測定するものである。

AI リスクには、産業界、学界、その他の利害関係者におけるより広範な考慮事項（公平性、バイアス、倫理、AI の悪用（有害なコンテンツ生成を含む）など多様な問題）が含まれるが、本ガイダンスではこれらの問題は扱わない。これらの追加的な危害の種類に関する議論については、NIST の AI リスクマネジメントフレームワークを参照のこと。

---

[9] この定義は 15 U.S.C. § 9401(3)に基づく。大統領令 14144「国家サイバーセキュリティにおけるイノベーションの強化と促進」も同様の定義を採用している。他の管轄区域では独自の AI 定義を用いる場合がある。

[10] AI の概念に関する追加情報は、ASD の ACSC ガイダンス「複雑な層：人工知能入門」を参照のこと。

TLP:CLEAR

安全性。IEC 61508 は、電気・電子・プログラマブル電子（E/E/PE）安全関連システムにおいて、システムの安全性がその正しい動作に依存する状況における機能安全を定義している。[11] NIST は、OT セキュリティの文脈における安全性を「死亡、負傷、職業病、設備や財産の損傷・損失、環境への損害を引き起こす可能性のある状態からの解放」と定義している。本ドキュメントでは、AI 対応 OT システムの安全性を論じる際、「安全」および「機能安全」という用語を、特にこれらの文脈において使用する。

セキュリティ。セキュリティ、情報セキュリティ、サイバーセキュリティという用語は類似の意味を持ち、本ガイダンスでは AI 対応 OT システムのセキュリティを定義する際にこれらを同義語として使用する。NIST は情報セキュリティを、情報及び情報システムを以下のものから防御することと定義する：

> [U]不正なアクセス、使用、開示、妨害、改ざん、破壊から保護し、プロバイダを提供することである。
>
> - 完全性：不適切な情報［または技術］の変更や破壊から守ることを意味し、情報の否認防止と真正性の確保を含む。
> - 機密性：アクセスと開示に対する認可された制限を維持すること。これには個人プライバシーと専有情報の防御手段が含まれる。
> - 可用性：情報[及び技術]への適時かつ信頼性の高いアクセスと利用を確保すること。

NIST はさらに、セキュリティを「保護措置の確立と保守」と定義する。これは組織の広範なセキュリティ方針の一部として、システムやデータへの不正アクセスや不正使用を防止し、継続的な運用性を確保するものである。セキュリティには、こうしたシステムに対する侵害を識別、防止、防御、対応、または復旧する能力やプロセスも含まれる。

[11] 「IEC 61508 及び機能安全の概要」、国際電気標準会議（IEC）、（PowerPoint、2022 年）、https://assets.iec.ch/public/acos/IEC%2061508%20&%20Functional%20Safety-2022.pdf?2023040501。

TLP:CLEAR

参考文献

- Humberto Santander Pelaez, Manuel.「ICS システムへのネットワークアクセス制御」. 日記（ブログ）, *SANS Technology Institute Internet Storm Center*. 2023 年 7 月 3 日. https://isc.sans.edu/diary/30000.
- 「IEC 61508 および機能安全の概要」国際電気標準会議（IEC）、（PowerPoint、2022 年）。https://assets.iec.ch/public/acos/IEC%2061508%20&%20Functional%20Safety2022.pdf?2023040501。
- シュルホフ、サンダー。「基本 LLM 設定」。*Learn Prompting*。最終更新日 2025 年 3 月 10 日。https://learnprompting.org/docs/intermediate/configuration_hyperparameters。

TLP:CLEAR