RÉPUBLIQUE
FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

**AI に関する共同ハイレベル・リスク分析**

# サイバーリスクに基づくアプローチによる AI の信頼構築

パリ AI アクションサミット - AI の信頼性 - サイバーセキュリティ

→
この文書は、**フランス国家サイバーセキュリティ局（ANSSI）**が発行し、以下のパートナーが共同署名している：

- カナダ（CCCS）
- エストニ（RIA）
- フィンランド（NCSC-FI）
- ドイツ（BSI）
- ギリシャ（NCSA）
- インド (CERT-IN)
- アイルランド（NCSC-IE）
- イタリア（ACN）
- ルクセンブルク（LHC）
- マルタ（MDIA）
- オランダ（AIVD および NCSC-NL）
- ノルウェー（NSM）
- ポーランド（NASK）
- イギリス (NCSC-UK)
- シンガプール（CSA）
- スロバキア（NCSC-SK）
- スロベニア（URSIV）
- 韓国（NIS）

## 免責事項

本文書に含まれる情報は「現状のまま」プロバイダとして提供されるものであり、人工知能のリスクと機会に関する議論に貢献することのみを目的としている。したがって、ANSSI および認可機関は、その使用によって生じたいかなる種類の損失、傷害、損害についても責任を負わない。本文書に含まれる情報は、ANSSI および執筆組織によるサードパーティーの事業体、製品、サービスの承認または推奨を意味するものではない。サードパーティーのウェブサイトや文書へのリンクや参照は、情報提供のみを目的としてプロバイダが提供するものであり、これらのリソースを他より推奨または支持することを意味するものではない。

## サイバーリスクに基づくアプローチで AI の信頼を築く

この文書の背後にある国際機関や政府当局は、信頼できる AI システムをサポートし、安全な AI バリューチェーンを実現するためのリスクベースのアプローチを提唱し、AI サミット以降も議論を継続し、AI 導入の文脈における機会、リスク、進化するサイバー脅威に等しく対処するよう呼びかけている。

AI は 1950 年代から開発されている変革的技術であり、現在では防衛からエネルギー、医療、金融など、ほとんどすべての分野に影響を及ぼしている。大規模言語モデル（LLM）[1] の使用や AI への依存の増加など、その急速な導入は、関係者にサイバーセキュリティに関連するものを含む関連リスクのアセスメントを促すはずである。

適切な対策を講じなければ、そしてユーザーが AI 関連のサイバーリスクを過小評価する傾向が依然としてあることを考慮すれば、悪意のある行為者が AI システムの脆弱性を悪用し [2]、将来の AI 技術の利用を危うくする可能性がある。**したがって、信頼できる AI の開発を促進し、この技術が提供する機会を十分に受け入れるためには、こうしたリスクを理解し緩和することが極めて重要である。**

ソフトウェアシステムである AI システムは脆弱性があるため、特に開発、展開、インシデント管理、ソフトウェアサプライチェーン、脆弱性管理（プロプライエタリ・コンポーネントとオープンソースコンポーネントの両方を含む）に関する既存の優れたサイバーセキュリティ慣行に基づき、設計によってセキュリティを確保する必要がある。AI は、他の情報システム（IS）と同様に、ホスティング・インフラストラクチャを通じたサイバー脅威に直面する。また、AI 特有のリスクもあり、特に AI システムにおけるデータの中心的な役割に関連して、機密性と完全性に対する独自の課題が生じる。

既存のサイバー衛生対策が広く有効であることに変わりはないが、AI を活用したサイバーセキュリティ・ソリューションの登場は、新たな脅威や進化する脅威に直面する上で重要な資産となる。本書の範囲外ではあるが、こうしたソリューションはすでにサイバーセキュリティ能力の強化に貢献しており、監視や侵入検知だけでなく、脅威分析、対応の自動化やデジタル調査、セキュリティ・プロセスの自動化など、今後も発展が見込まれている。

その一方で、急速に進化する脅威の状況から、ますます巧妙になると予想される悪意のある AI の利用を追跡する必要がある。AI はすでに、こうした攻撃を行うために必要な専門知識を低下させ、より大規模で効率的な攻撃を可能にすることで、既存の攻撃手法を増幅させている。フィッシングやソーシャル・エンジニアリング、脆弱性スキャン、悪意のあるコード開発などにおいて、こうした効果が見られる。高度な生成的 AI は、サイバーキルチェーン全体にわたって、大規模でコスト効率の高い攻撃を可能にする可能性がある。

防衛的であれ攻撃的であれ、AI を活用したソリューションについては、学術論文や現在開発中のさまざまな枠組みですでに十分に取り上げられているが、本書では AI システムのサイバーセキュリティに焦点を当てる。その目的は、関連するサイバーリスクのハイレベルな総合的かつ包括的な分析を提供し、2023 年 11 月に 20 以上の国際機関が共同で発表した「安全な AI システム開発のためのガイドライン」に基づき、脅威をアセスメントし、適切なセキュリティ対策を実施するためのガイダンスを提供する[ことである(3)]。

1. LLM：Large Language Model の略で、言語処理や自然言語生成に使われる生成的 AI モデル。

2. ここで AI システムとは、一連の訓練データに対する統計的学習によって構築された AI モデルに依存するソフトウェアシステムと定義される。

3. サイバーセキュリティ・インフラセキュリティ庁（CISA）と国家サイバーセキュリティセンター（NCSC）、セキュア AI システム開発ガイドライン、2023 年 11 月 26 日。

## AI リスク分析の課題と範囲

このリスク分析は、個々の AI コンポーネントの脆弱性だけでなく、これらのコンポーネントを統合したより広範な AI システムのセキュリティも考慮することを目的としている。その目的は、脆弱性の網羅的なリストではなく、AI 関連のサイバーリスクの幅広い概観を提供することである。さらなる参考文献については、附属書 2 の参考文献リストを参照のこと。

AI システムの展開は、適切なセキュリティ対策を講じなければ、悪意ある行為者に新たな攻撃経路を開く可能性がある。したがって、このような展開には、リスクを評価し、適切なセキュリティ対策を特定するための専用のリスク分析を含めるべきである。

さらに、AI のサプライチェーンを理解することは、ある AI システムに関わるサプライヤーやその他のステークホルダーの脆弱性に関連するリスクを緩和するために不可欠である。AI サプライチェーンは一般的に 3 つの柱で成り立っている

1. 計算能力
2. AI モデルとソフトウェアライブラリ；
3. データである。

それぞれの柱には、サイバーセキュリティの成熟度が大きく異なる、明確な、時には共通のプレーヤーが関わっている。

## 主なリスクと攻撃シナリオ

AI システムを支える IT インフラは、他の IT システムと同じ脆弱性に直面している。また、AI システムは、生のデータ収集から推論に至るまで、そのライフサイクルのさまざまな段階で攻撃される可能性がある。AI 特有の攻撃は、一般的に 3 つのカテゴリーに分類される：

- **ポイズニング：**学習データやモデル・パラメータを変更することで、すべての入力、または特別に細工された入力に対する AI システムの反応を変化させる；
- **抽出：**学習フェーズ後の AI システムやモデルから、モデルパラメータ、設定、学習データなどの機密データを再構築または復元すること；
- **回避：**入力データを改変して、AI システムに期待される機能を変更すること。

**このような攻撃は、**自動化された意思決定やプロセスの信頼性が損なわれる**AI システムの誤動作（可用性または完全性攻撃リスク）**や、**機密データの盗難や開示（機密性リスク）**に**つながる可能**性が**ある。**

**今日、ほとんどの AI システムは不透明であるため、ユーザーはさらなる課題を考慮しなければならない。**AI の説明可能性は、基礎となるモデルによって大きく異なる。多くのシステムは「ブラックボックス」として動作し、その決定を説明したり正当化したりすることを難しくしている。この不透明性は、エラーやその他の問題のある出力の根本原因を見つける能力を妨げ、潜在的なインシデントの特定と調査をより困難にするため、これらのシステムを安全にする取り組みを複雑にする。

AI システムが関与する**主なリスクシナリオ**は以下の通りである：

- **AI ホスティングおよび管理インフラの侵害：**悪意のある行為者は、技術的、組織的、人的な脆弱性を問わず、一般的な脆弱性を幅広く悪用することで、AI システムの機密性、完全性、可用性に影響を及ぼす可能性がある。AI システムのホスティング・インフラストラクチャを侵害することは、もっともらしく重要な攻撃ベクトルであり、AI システムのライフサイクル全体を通じて考慮されなければならない。

- **サプライチェーン攻撃：**攻撃者は、サプライチェーンの関係者（ソフトウェア・ライブラリ、事前学習モデル・プロバイダ、サービス・プロバイダなど）の脆弱性を悪用する可能性がある。例えば、オープンソースのライブラリは、AI システムの開発（ ）でしばしば使用され、より広範な枠組みに統合されることが多い。これらのライブラリへの攻撃は、AI システム全体を危険にさらす可能性がある。

- **AI システムと他のシステムとの相互接続による横並び化：**AI システムは、コミュニケーションや効率的なデータ統合を目的として、他の IS と相互接続されることが多い。こうした相互接続は新たなリスクをもたらす可能性があり、例えば、攻撃者が制御する外部ソースを通じて悪意のある命令を入力することで LLM を悪用する間接的プロンプト・インジェクションによる攻撃がある。このような攻撃は、機密情報の抽出や悪意のあるコマンドのリモート実行に使われる可能性がある。このリスクは、AI システムが産業用システムと相互接続されている場合、それらのシステムが物理世界に直接作用する可能性があるため、特に重要である。

- **人間や組織の失敗：**トレーニング不足は、自動化への過度の依存や、AI システムの異常な行動に気づく能力の不足を誘発する可能性がある。さらに、シャドーAI[(4)は]、機密データの紛失、規制違反、組織イメージの風評被害などのリスクを増大させる可能性がある。長期的には、AI の集中的かつ長期的な使用は、技術依存のリスクや、将来の AI 能力が故障した場合に人間の行動で代替できない「ロックイン」につながる可能性がある。このリスクは、AI システムが重要な活動（ビジネス環境など）、特にビジネスプロセスが高度に自動化されている場合に大きくなる。

- **AI システムの応答における誤動作：**攻撃者は、AI モデルの訓練に使用するデータベースを侵害し、本番稼働後に誤った応答を引き起こす可能性がある。AI モデル開発者の慣行では、意図的で悪意のある学習データ・ポイズニングに対するレジリエンスが改善される傾向にあるため、この攻撃には攻撃者の多大な努力が必要となるが、健康や物理的セキュリティの文脈で使用される画像など、データの分類に使用される場合は特に危険である。

4. シャドーAI とは、組織の IT 部門の承認や監督なしに、主流の生成的 AI ソリューションを使用することと定義される。

## Ai ユーザー、オペレーター、開発者のためのガイドライン

AI システムの利用を検討する際には、ユースケースの機密性を分析することが最初のステップとなる。AI システムの複雑さ、サイバーセキュリティの成熟度、監査可能性、説明可能性は、所定のユースケースのサイバーセキュリティとデータ・プライバシー要件に対応すべきである。

AI ソリューションの開発、展開、使用が決定された場合、通常のサイバー勧告に加え、以下のガイドラインが AI ユーザー、オペレーター、開発者にとってのグッドプラクティスとなる：

- **AI システムの自律性レベルを、リスク分析、ビジネスニーズ、実行されるアクションの重要性に応じて調整する。**人間の妥当性確認は、サイバーリスクとほとんどの AI モデルに内在する信頼性の問題（LLM の幻覚など）の両方に対処するのに役立つため、必要に応じてこのプロセスに統合されるべきである；
- **コンポーネント**、その他のハードウェアおよびソフトウェアコンポーネント、ならびに**データセット**（性質、調達、処理-特に毒の緩和とリスク抽出の影響の評価）を含む、**AI のサプライチェーンのマッピング**；
- 攻撃経路を最小化するために、AI システムとその他の情報システムとの**相互接続を追跡**し、それぞれがユースケースで必要とされるものであることを確認する；
- **AI システムを継続的に監視・保守し**、サイバーセキュリティに影響を与えかねないバイアスや脆弱性がなく、意図したとおりに動作するようにすることで、特定の AI システムの「ブラックボックス」的性質に関連するリスクを緩和する；
- **技術や規制の大きな変化を予測**し、**潜在的な新たな脅威を特定**するプロセスを導入することで、戦略を適応させ、将来の課題に立ち向かえるようにする；
- エグゼクティブを含め、AI の課題とリスクについて**社内で意識向上およびトレーニングを**行い、ハイレベルの意思決定が十分な情報に基づいて行われるようにする。

*附属書 1 のチェックリストは、AI ユーザー、オペレーター、開発者に、検討すべき追加措置や推奨事項を提供することができる。*

## 政策立案者のためのガイドライン

地域や国の状況を考慮し、政策立案者は目指すべきである：

- 敵対的機械学習（AI 特有の攻撃やその防止・検知を含む）、プライバシー保護コンピューティング、AI の新たな攻撃的利用などの領域を含む、**これらのリスクに関連する研究を**検知する。
- AI モデル、アプリ、データ、インフラに対する信頼を醸成するため、**共有標準に基づくセキュリティ評価・認証能力の開発を支援する。**
- **AI システムの安全な展開とホスティングを確保するために、サイバーセキュリティのベストプラクティスを推進し続ける。**既存の規制や適用可能な規制を活用し、セキュリティ要件をリスクのレベルに適合させるための明確なガイドラインを示し、フィードバックを共有することで、組織が一般的な誤りを回避し、AI を業務に最適に統合できるようにする。
- AI システムのサイバー上の課題に対するより良い検討を促進する方法として、**サイバーと AI の関係者間、**特にサイバーセキュリティ機構と AI 安全性機構（または類似のもの）との**対話を促進し**、それぞれの境界と責任を明確に定義する。このような協力は、新たな脅威に関する情報の共有と、重要なシステムを保護するための取り組みの連携に重点を置くべきである。
- **AI** システムに対する脅威の継続的なモニタリングを含め、**AI サミット以降も対話を継続**し、AI のバリューチェーンをより安全にし、AI への信頼を醸成するためのガイドラインを特定するために、国際レベルでの議論と協力を追求する。

## → 附属書 1

# AI システムを安全に導入するための推奨事項

これは、検討すべき事柄のハイレベルな概要を示すものであり、網羅的なものであってはならない。その他の枠組みやガイドラインについては、附属書 2 を参照のこと。

## 1.自己アセスメントを推奨する：

- 可能であれば設計段階から、システムの明確かつ正当な目的を適切に定義し、文書化したか？
- 規制の側面を自分の思考適切に組み込んだか？AI システムが想定している治療が適用法や規制に準拠していることを検証したか？
- ライフサイクルの各段階において、誰が AI システムにアクセスできるのか？
- AI システムのセキュリティと完全性を保証するために、最小特権の原則が適用されているか？
- AI システムの依存関係チェーンとは何か？
- サプライヤーの評判や財務状況はどうか？
- データプロバイダーであれソフトウェアコンポーネントベンダーであれ、ベンダーはサイバーセキュリティ標準を満たしているか？
- クラウド・ソリューションを導入する必要があるか？起こりうる結果（データ保護など）について、グローバルなリスクアセスメントを行ったか？
- 私のデータを操作できるプロバイダとのサービス契約に可逆条項があるか？可逆性は技術的（手段やデータ転送速度）に、また時系列的に実現可能か？
- AI 活用がビジネスに与える影響とは？AI の誤作動は組織を危険にさらすのか？
- AI システムのライフサイクルの各段階におけるセキュリティ基盤（ガイド、ベストプラクティス・ベンチマーク、マッピングなど）があるか。
- 私の AI モデルは機密保護されるべきか？私の重要な価値があるか？
- 該当する場合、個人データを保護するための適切な措置（プライバシー・バイ・デザイン）を、データとメタデータの両方、および AI システムのモデル（複数可）に対して適切に統合したか。

## 2.推奨行動のチェックリスト

### 一般的な推奨

- 他の情報システムにおける重要なアクションの自動化に AI システムを使用することを制限する；
- AI が重要なプロセスに思慮深く適切に組み込まれ、セーフガードが提供されるようにする；
- 組織全体の状況を統合して、専用のリスク分析を実施する（例えば、AI システム障害の影響を組織全体でアセスメントする）；
- AI システムのライフサイクルの各段階（訓練データ収集から推論段階、廃止まで）のセキュリティを研究する；
- 必要であれば、データ保護影響アセスメントを実施すること；
- AI 関連資産を識別、追跡、保護する；

### インフラストラクチャーとアーキテクチャーの推奨

- AI システムの使用方法を定義し、特に自動化の場合、意思決定プロセスへの統合を図る；
- 適用される規制や組織の方針を考慮し、適切な場合にはクラウド固有の対策を適用する；
- 該当する場合は、アウトソーシングの推奨事項を適用する；
- AI システムに安全な管理の推奨事項を適用する；
- 重要な AI コンポーネントのアクセス管理システムを活用する。

### 展開計画を立てる

- アーキテクチャを設計することで、スケーリングが発生しても、セキュリティ・レベルに悪影響を与えないようにする；
- プロジェクトの全フェーズにおいて DevSecOps の原則を適用する；
- ライフサイクルを通じてデータ・保護の要件を満たすために、プライバシー・バイ・デザインのアプローチを用いて AI システムを設計する：
  - データの機密性を考慮する；
  - 必要に応じて、データの仮名化または匿名化を確実に行う；
  - AI システムを設計する際には、知る必要性の問題を考慮に入れる。

### 使用するリソースに注意する

- AI モデルの入手、保存、配布には安全なフォーマットを使用する；
- モデルファイルをロードする前に、その整合性を検証するメカニズムを実装する；
- AI システムで使用されているライブラリやプラグインの信頼度をアセスメントする；
- AI システムで使用される外部データの品質を保証し、信頼度を評価する。
- AI システムで実行されたアクションのトレーサビリティを確保する；
- データ収集が、システムの開発および使用の両方において、公正かつ倫理的な方法で行われたことを確認する。

### 学習プロセスを確保し、強化する

- AI システムがアクセスするデータ、特に機密データについて厳格なポリシーを採用する；
- トレーニングデータに安全にアクセスし、保存する；
- 使用されている学習・再学習方法の安全性をアセスメントする；
- 抽出されたデータ、メタデータ、アノテーション、特徴、および AI システムモデルに対して、以下のような対策を実施する：
  - データをクリーンアップする；
  - 関連し、厳密に必要なデータ（量、カテゴリー、粒度、類型など）を特定する；
  - 必要に応じてデータを仮名化または匿名化する。

### アプリケーションの信頼性を高める

- AI システムのすべての管理タスクに多要素認証を導入する；
- インプットとアウトプットの機密性と完全性を確保する；
- 悪意のある命令を検知するために、セキュリティフィルターを実施する；
- すべてのデータ、メタデータ、注釈が（特にドリフトを避けるために）最新かつ正確な状態に保たれていることを確認する；
- モデルの精度と性能を継続的に評価する。

### 組織考える

- 文書デザインの選択
- AI システムの運用を監督する；
- 主要人物を特定し、下請け業者の使用を監督する；
- リスクマネジメント戦略を実施する；
- AI システムなしのデグレード・オペレーション・モードを提供する；
- 生成的 AI の使用ポリシーを（組織の感性に応じて）フレーム化して実施する；
- AI システム固有の脆弱性を監視するプロセスを確立する；
- 例えば、個人データの利用を制限するような技術開発を注意深く監視する；
- データ管理システムを導入する；
- データの削除には、安全な削除方法を活用する；
- 製品に使用されているデータセットを文書化する：
  - データベースの使用を容易にする；
  - 削除または匿名化されるまで、長期にわたるデータの監視を容易にする；
  - 予期せぬデータ使用のリスクを減らす。

## 予防

- AI に関連するセキュリティリスクについて、定期的にスタッフにトレーニングを行う；
- AI システムのセキュリティ監査を定期的に実施する；
- トレーニングデータやモデル自体に対する権利（例えば知的財産権やデータ保護）の行使に関連する可能性のある問題を可能な限り予測する。

## → 附属書 2 参考文献

### AI 開発


- AIVDAI システム：安全に開発する 2023.入手：AI システム：安全に開発する｜出版｜AIVD
- G7.広島プロセス・インターナショナル、高度 AI システム開発組織の行動規範。2023.入手：100573473.pdf
- G7.広島プロセス・インターナショナル, 先端 AI システムを開発する組織のための指針.2023.入手：100573471.pdf
- NCSC-UK, CISA.安全な AI システム開発のための共同ガイドライン。2023.にて入手可能：安全な AI 開発のためのガイドライン システム- NCSC.GOV.UK


### AI の使用例


- ANSSI。生成的 AI システムのためのセキュリティ勧告。2024.https://cyber.gouv.fr/en/ 。publications/security-recommendations-generative-ai-system から入手できる
- BSI、ANSSIAI コーディングアシスタント。2024.https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ 。ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7 にて入手可能
- BSI.生成的 AI モデル：産業界と認可機関のチャンスとリスク。2025.https://www。bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Generative_AI_Models.pdf?__blob=publicationFile&v=6


### AI の脆弱性とセキュリティ


- CSA シンガポール。AI システムのセキュリティに関するガイドラインとコンパニオンガイド。2024.以下から入手できる：AI システムの保護に関するコンパニオン・ガイド｜シンガポール・サイバーセキュリティ庁ガイドラインと
- CSA シンガポール、レサロ。AI の安全確保：連帯責任。2024.にて入手可能：人工知能（AI）の：シンガポール・サイバーセキュリティ局安全確保に関するディスカッション・ペーパー
- インド電子情報技術省 Ministry of Electronics and Information Technology）、ソフトウェア部品表 （SBOM）に関する技術ガイドライン。利用可能な場所： https:// www. cert-in.org.in/PDF/SBOM_Guidelines.pdf
- CERT-IN API セキュリティ - 脅威、ベストプラクティス、課題、AI を利用した前進の道。利用可能な場所： https:// www. cert-in.org.in/s2cMainServlet?pageid=PUBADV01&CACODE=CICA-2023-3248
- Junklewitz、H.、Hamon、R.、André、A.、Evas、T.、Soler Garrido、J.、Sanchez Martin、J.I。AI 法における人工知能のサイバーセキュリティ。2023.にて入手可能：JRC Publications Repository - 人工知能のサイバーセキュリティ AI
- Kamm L. (Cybernetica AS), Pillmann H. (RIA).人工知能と機械学習システムのリスクとコントロール。2024.にて入手可能：人工知能と機械学習システムのリスクと管理。

- mitre atlas: ミトレアトラス
- OWASP AI Exchange: https://owaspai.org/ - いくつかの出版物を含む。
- Vassilev, A., Oprea, A., Fordyce, A. and Andersen, H. Adversarial Machine Learning : A Taxonomy and Termino-logy of Attacks and Mitigations.2024.にて入手可能：敵対的機械学習：分類と用語攻撃と緩和の｜NIST


## リスクマネジメント


- NIST AI リスクマネジメントフレームワーク：AI リスクマネジメントフレームワーク｜NIST
- OECD「AI におけるアカウンタビリティの推進：信頼できる AI のためのライフサイクルを通じたガバナンスとリスクマネジメント」。2023.入手：AI におけるアカウンタビリティの推進｜OECD


## 用語解説


- ISO/IEC 22989:2022 - 「情報技術 - 人工知能 - 人工知能の概念と技術」。2022.入手先：https://www.iso.org/standard/74296.html
- OECD「AI インシデントと関連用語の定義」。2024.入手：AI インシデントと関連用語の定義｜OECD


## 規制の例


- EU 人工知能法規則 - EU - 2024/1689 - EN - EUR-Lex
- EU サイバーレジリエンス法規制 - 2024/2847 - EN - EUR-Lex

バージョン 1.0 - 2025 年 2 月 - ISSN en cours

オープンライセンス (Etalab - v2.0)

**国家システム安全保障局**
ANSSI - 51, boulevard de la Tour-Maubourg - 75 700 PARIS 07
*www.cyber.gouv.f*