第406回　NRIメディアフォーラム

ITロードマップ【Day 1】

フィジカルAIにおけるセキュリティの展望と課題

荻荘 裕太

NRIセキュアテクノロジーズ株式会社
研究開発センター
サービス開発推進部

2026年3月23日

NRIセキュアテクノロジーズ
NRI SecureTechnologies

Envision the value,
Empower the change

01 フィジカルAIの概要

02 セキュリティ課題

03 対策の方向性

04 ロードマップ

05 まとめ

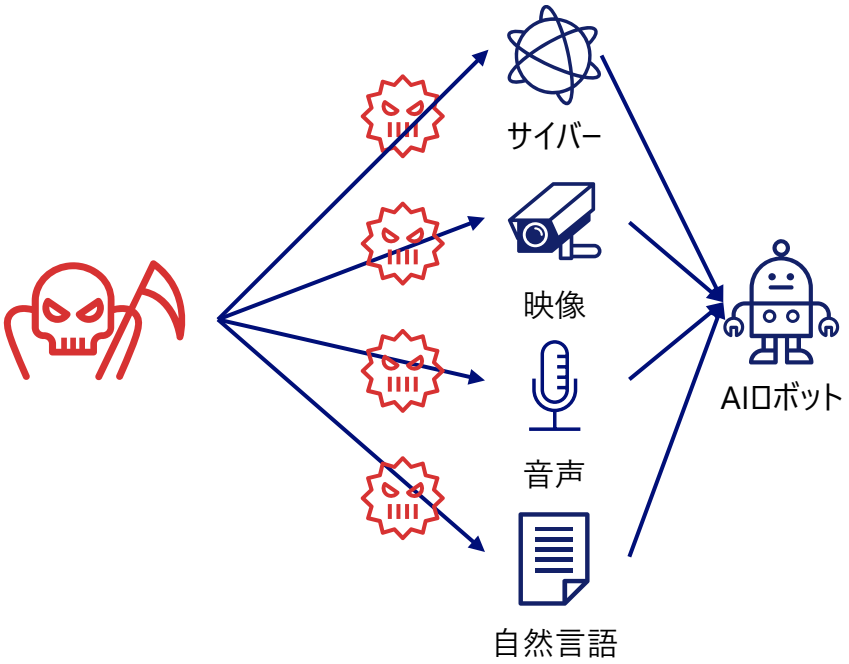
1. フィジカルAIの概要

フィジカルAIは複数の要素が複雑に組み合わさって構成されるため、1つの構成要素における脆弱性が全体の挙動に影響する。

定義と特徴

- センサーで環境を認識し、AIが判断してアクチュエータを介して現実世界に作用する技術
- 従来のAIがサイバー空間で情報処理を行うのに対し、物理世界で直接的な行動を起こす
- NVIDIAやGoogleをはじめ企業が開発を加速、物流・製造・医療・自動運転など多様な分野で導入が進む

フィジカルAIへの多様な攻撃経路



※アクチュエータ：物理的な動作を担う装置

フィジカルAIにおけるセキュリティは「情報の保護」だけでなく「行動の安全（セーフティ）」を含む概念へと広がっている。

リスク構造の拡大

- 従来のAI：情報漏えい、プロンプトインジェクション、ハルシネーションなど情報的な被害
- フィジカルAI：誤動作や設備損傷、人身事故など物理的な被害に発展する可能性
- 攻撃対象が広い：センサー、AIモデル、通信、制御、アクチュエータの全てが標的

セキュリティ概念の変化

観点	従来のAI	フィジカルAI
主に対象とする範囲	情報の保護	情報保護 + 行動の安全（セーフティ）
想定される被害内容	データ漏えい、誤判断	人身事故、設備損傷、稼働停止
被害の影響範囲	サイバー空間に限定	物理世界に直接影響

※従来のAI：サイバー空間で情報処理を行うAI（生成AI、予測AI等）

2. セキュリティ課題

フィジカルAIは多層的なリスクを抱えており、センサーから人的要因まで、あらゆる要素がセキュリティ上の脅威となりうる。

センサーへの攻撃
- スプーフィング攻撃
- 偽の信号による誤認識
- カメラ、LiDAR、GPS等が標的
- AIの誤判断を誘発

AIモデル・ソフトウェア
- 敵対的攻撃
- プロンプトインジェクション
- 学習データの改ざん
- ソフトウェアの脆弱性

通信・制御・アクチュエータ
- 中間者攻撃
- ランサムウェア感染
- サプライチェーン攻撃
- 機械の誤作動

人的要因
- AIへの過信・不信
- 内部不正
- 運用体制の甘さ
- 教育・理解不足

※LiDAR（Light Detection and Ranging、ライダー）：レーザー光の反射を利用して、物体までの距離や形状を三次元的に計測する技術

センサーへの攻撃：スプーフィング攻撃はセンサーに「本当とは違う世界」を見せることでAIの誤判断を引き起こし、動作停止や衝突などの被害につながる。

攻撃の仕組み

- センサーは外部から光・音・電波などを使って誤った信号を送り込まれると、AIが現実を誤認識
- これらの攻撃手法は学術研究で実証されており、既存のIT・OT・AIセキュリティの脅威が組み合わさったもの
- カメラ：強いライトで標識を消す、プロジェクターで偽標識を映す
- LiDAR：偽の反射信号で存在しない物体を検知させる
- GPS：衛星信号を偽装して誤った場所に誘導

主要センサーへの攻撃例

センサー	攻撃手法	影響
カメラ	強い光・偽映像の投影	標識認識の誤動作
LiDAR	偽レーザー反射	存在しない障害物の検知
GPS	衛星信号の偽装	位置情報の誤認
IMU	特定周波数の振動	傾き・揺れの誤検出

※IMU（Inertial Measurement Unit、慣性計測装置）：加速度や角速度を検出するセンサー

AIモデル・ソフトウェア：AIモデルは判断の中枢であり、ここが攻撃されるとシステム全体が危険にさらされる。

判断の中枢への攻撃

- 敵対的攻撃：入力データに微小な摂動を加えてAIに誤判断をさせる
- プロンプトインジェクション：大規模言語モデル搭載ロボットで意図しない動作を引き起こす
- ソフトウェアの脆弱性：更新遅れや設計不備が外部侵入・システム乗っ取りにつながる

AIモデル関連のリスク

攻撃手法	影響
敵対的攻撃	誤った判断・不適切な行動
プロンプトインジェクション	意図しない命令の実行
学習データの改ざん	判断基準の歪み
モデル抽出攻撃	知的財産の窃取

通信・制御・アクチュエータ：クラウドやネットワークと連携するフィジカルAIは、通信経路や制御基盤も攻撃対象となり、可用性や制御の安全性を損なう深刻な結果を招く。

物理動作に至る経路の脅威

- 通信経路：中間者攻撃による盗聴・改ざん、ランサムウェアによるシステム停止
- サプライチェーン：ソフトウェア更新時に不正コードが混入
- アクチュエータ：誤作動により人身事故や設備損傷が発生、設計不備や保守不足も危険

攻撃対象と影響

対象	主なリスク	想定被害
通信	盗聴・改ざん・妨害	情報漏えい・制御不能
制御基盤	ランサムウェア感染	稼働停止・身代金要求
更新機構	不正コード混入	システム全体の侵害
アクチュエータ	誤作動・故障	人身事故・設備損傷

2. セキュリティ課題

人的要因：人による操作ミスや過信、内部不正など、人との関わりに起因するリスクもセキュリティ上の脅威となる。

過信による監視不足

AIを信頼しすぎて異常を見逃し、被害が拡大

内部不正・運用体制の甘さ

関係者による不正利用、権限管理の不備

不信による導入遅れ

AIを信用できず、社会実装が進まない

理解不足・教育の欠如

仕組みと限界を理解せず、誤った運用

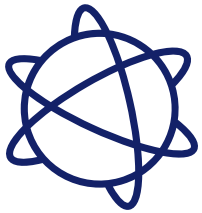
3. 対策の方向性

3. 対策の方向性

フィジカルAIのセキュリティ確保には、個別要素への対処に加え、システム全体を貫く統合的な対策と設計思想が必要である。

技術的対策

AIモデルの堅牢化、センサーフュージョン、リアルタイム異常検知



制度的対策

責任所在の明確化、標準化・規制整備、認証制度

社会的対策

説明可能性・透明性、人材育成、教育・啓発活動

技術的対策：AIモデルを敵対的攻撃や環境ノイズに強くする堅牢化と、異常をリアルタイムで検知して即応できる体制の整備が重要である。

主な取組

- AIモデルの堅牢化：敵対的攻撃や環境ノイズに強いモデル設計
- センサーフュージョン：複数センサーの情報を組み合わせて整合性を確認する仕組み
- リアルタイム異常検知：異常を即座に検知して対応する体制
- セキュリティ・バイ・デザイン：開発から運用まで一貫したセキュリティ組込

技術的対策の例

対策	内容
モデル堅牢化	敵対的学習、防御的設計
冗長化・形式的検証	複数系統の確保、論理検証
異常検知システム	AIによる振る舞い監視
フェイルセーフ設計	故障時に安全に停止する設計

制度的対策：事故発生時の責任所在を明確にし、標準化や規制を通じて共通の安全基準を設定することが不可欠である。

主な取組

- 責任所在の明確化：製造者・運用者・利用者の範囲を法的に整理
- 標準化と規制：共通の安全基準を設定し企業が安心して導入できる枠組み
- 認証制度：ISO/IEC 42001等の国際規格を踏まえた第三者認証

関連する国際規格

規格	対象
ISO/IEC 42001	AIマネジメントシステム
ISO 13482	サービスロボット安全
ISO 10218	産業用ロボット安全
IEC 61508	機能安全

社会的対策：人とAIが共存する環境での安全性と信頼性を高めるには、説明可能性や透明性を備えた設計、倫理的な配慮、そして教育・啓発活動が欠かせない。

主な取組

- 説明可能性と透明性：AIの判断根拠を明示し、信頼の基盤を構築
- 倫理的配慮：人とAIが共存する環境での安全性・公平性の担保
- 教育・啓発活動：利用者がAIの仕組みと限界を正しく理解
- 人材育成：現場で運用する専門人材の育成と社会受容性の向上

社会的対策の観点

観点	内容
透明性	判断プロセスの可視化
説明責任	エラー時の説明・対応
人材育成	開発・運用の専門家養成
社会受容	ユーザー教育・啓発

4. ロードマップ

4. ロードマップ

フィジカルAIのセキュリティは、2030年代の基盤整備から2040年代の社会定着へ、段階的に発展する。

	～2030年度	～2040年度	2040年度～
全体	黎明期	発展期	普及期
技術動向	モデルの堅牢化と異常検知の実証	リアルタイム異常検知の実用化	継続監査と更新体制の確立
	センサー・通信経路の安全化	安全性検証と認証連携の拡大	社会全体での継続的な安全運用
制度・標準化動向	評価基盤と標準化の準備	認証制度の運用開始	国際認証・相互承認の確立
	安全性評価と検証枠組みの整備	国際標準との整合	倫理・透明性を含む制度の確立
社会動向	教育・訓練と啓発活動の開始	産業分野での実践拡大	リスクガバナンスの定着
	限定環境での運用実証	情報共有と連携体制の構築	信頼と安全を支える文化の形成

～2030年：限定環境でのフィジカルAI導入が進み、安全性と信頼性の基盤が形成される段階である

主な動向

- 物流倉庫や製造現場など制御しやすい環境で実証と導入が進む
- 汎用ロボットがピッキングや搬送など単純作業を担う
- AIモデルやセンサーの標準化・共通仕様の検討、評価基盤の構築が進展

各層の取組

分野	取組内容
技術	モデルの堅牢化と異常検知の実証
制度・標準化	評価基盤と標準化の準備
社会	教育・訓練と啓発活動の開始

～2040年：フィジカルAIが多様な産業に拡大し、セキュリティ確保が個別技術対応からシステム全体設計へと進化する。

主な動向

- 物流・製造・流通・医療・サービスなど多様な分野に拡大
- 人と協働する自律型システムとして定着、システム間連携が一般化
- リアルタイム異常検知、セキュリティ・バイ・デザイン、フェイルセーフ設計が実運用で定着

各層の取組

分野	取組内容
技術	リアルタイム異常検知の実用化
制度・標準化	認証制度の運用開始
社会	産業分野での実践拡大

2040年～：セキュリティは、社会インフラの信頼を維持する「持続的な制度」として定着する。

主な動向

- 建設・農業・医療・交通・防災など多様な分野で不可欠な存在に
- 自律型ロボットやヒューマノイドが人の生活や経済活動を支える基盤技術に
- 継続的なリスク監査やアップデートを行う体制が整備され、倫理・透明性・説明責任を含む統合的な運用ルールが国際的に共有

各層の取組

分野	取組内容
技術	継続監査と更新体制の確立
制度・標準化	国際認証・相互承認の確立
社会	リスクガバナンスの定着

5. まとめ

まとめ

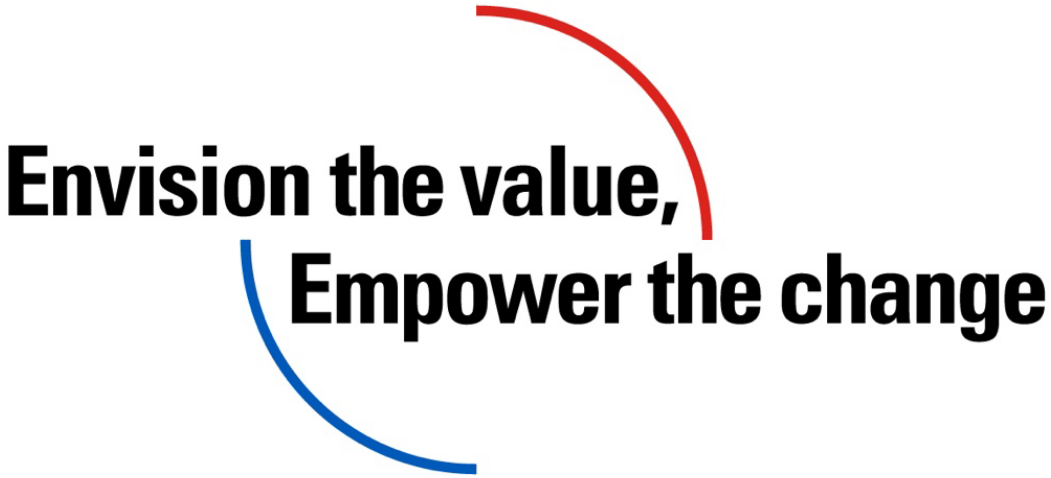
フィジカルAIのセキュリティは、情報保護だけでなく人身安全まで含む新たな課題であり、技術・制度・社会の統合的対策が必要である

フィジカルAIとは
センサー・AI・アクチュエータが連携し、物理世界に直接作用する技術

セキュリティリスク
センサー攻撃で誤認識、AIモデル攻撃で誤判断、結果として人身事故の恐れ

対策の方向性
技術（堅牢化）・制度（責任明確化）・社会（教育）の三位一体で取り組む

今後の展望
2030年代に基盤整備、2040年代に社会定着。今から準備が必要


Envision the value,
Empower the change