サイバーフィジカルシステムにおける AI セキュリティ

能上 絢香 [1, *]　国井 裕樹 [1]　松永 昌浩 [1]

概要：サイバーフィジカルシステムにおいて AI の導入が急速に進んでいるが，AI への攻撃がシステム全体へ深刻な影響を及ぼす懸念がある．よって，サイバーフィジカルシステムにおいて AI を安全に活用することは重要な課題である．本稿では，AI セキュリティの研究動向を俯瞰し，AI を活用したシステムが攻撃を受けた場合の影響を示した．AI モデルの配置場所（中央集約型・エッジ型・ハイブリッド型）に応じてアタックサーフェスを整理し，想定される脆弱性を（1）攻撃の汎用性の高さ，（2）被害の深刻さ，（3）情報セキュリティ・運用対策の活用可否という 3 点に着目して分析した．その結果，アタックサーフェスが学習時入力データ，推論時入力データとなる攻撃に対して AI セキュリティ対策の優先度が高いと結論付けた．また，実運用での AI セキュリティ対策の課題として，対策と精度のトレードオフ，対策の拡張可能性を示した．

キーワード：AI セキュリティ，Security for AI，AI への攻撃，サイバーフィジカルシステム

AI Security in Cyber-Physical Systems

Ayaka Nogami[1, *]　Hiroki Kunii[1]　Masahiro Matsunaga[1]

Abstract: The integration of artificial intelligence (AI) into cyber-physical systems (CPS) is progressing rapidly. However, there is growing concern that attacks targeting AI may have severe impacts on the entire system. Therefore, ensuring the safe deployment of AI within CPS is a critical challenge. This paper provides an overview of research trends in AI security and examines the potential impacts of attacks on AI-integrated systems. Attack surfaces are categorized based on the deployment location of AI models (centralized, edge-based, and hybrid), and anticipated vulnerabilities are analyzed from three perspectives (1) the generality of attacks, (2) the severity of potential damage, and (3) the applicability of information security and operational countermeasures. Our findings indicate that attacks targeting input data during both training and inference phases should be prioritized in AI security measures. Furthermore, we identify practical challenges in implementing AI security in real-world operations, including the trade-off between security and model accuracy, and the scalability of countermeasures.

Keywords: AI Security，Security for AI，Attacks on AI，Cyber-Physical Systems

1. はじめに

近年, ディープニューラルネットワークを中心とした AI が画像，音，言語など幅広い分野で著しく発展していることから, 実社会における活用が急速に進んでいる. 一方で, AI の特性に基づく様々な脆弱性が明らかになっており , AI の安全な活用に対する関心が高まっている. AI セキュリティと呼ばれる研究分野が進展し，活発な議論が行われていることからも[1－4]重要性がうかがえる. 本稿では AI が導入されている分野の一つとしてサイバーフィジカルシステムを取り上げる. サイバーフィジカルシステムとは, 図 1[5]に示すように,フィジカル空間（物理世界）の情報をセンシングし，サイバー空間でプロセッシングを行い，その結果をフィジカル空間にフィードバックするシステムである. サイバーとフィジカルの 2 つの空間が融合しているため，AI が攻撃された場合の影響はサイバー空間だけではなくフィジカル空間にも波及し，サイクル全体に大きな影響を及ぼしてしまう．例えば，自動運転であれば人の命に関わる重大な影響となってしまう可能性もある．よって，サイバーフィジカルシステムにおいて AI を安全に活用することは重要な課題である．

そこで本稿では，AI セキュリティの研究動向を俯瞰し，AI を活用したシステムが攻撃された場合の影響について，推論誤り , 情報漏洩, 計算浪費の観点から整理する. 次に，サイバーフィジカルシステムを AI モデルの配置場所に着目して分類し，アタックサーフェスを整理する．そして，サイバーフィジカルシステムの一つとしてミッションクリティカルなシステムを対象として, AI の活用例を挙げながら，想定される脆弱性，システムへの影響，対策を分析する．最後に AI を活用したサイバーフィジカルシステムを実際に運用する場合の課題を考察する．

本稿の主な貢献は以下の通りである．

- サイバーフィジカルシステムを AI モデルの配置場所で分類し，アタックサーフェスを整理した．システ

1 セコム株式会社　ＩＳ研究所
Intelligent Systems Laboratory, SECOM CO., LTD.
* aya-nogami@secom.co.jp

© 2025 Information Processing Society of Japan

This paper is work in progress and not peer-reviewed.

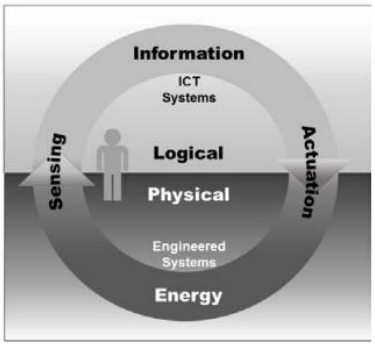

図 1　サイバーフィジカルシステム[5]

ム分類によって存在するアタックサーフェスが異なため，対策を選択する必要があることを明らかにした．

- サイバーフィジカルシステムの一つであるミッションクリティカルなシステムを対象として，想定される脆弱性を（1）攻撃の汎用性の高さ，（2）被害の深刻さ，（3）情報セキュリティ・運用対策の活用という 3 つの観点に着目して分析した．その結果，アタックサーフェスが学習時入力データ，推論時入力データとなる攻撃に対して AI セキュリティ対策の優先度が高いと結論付けた．
- AI を活用したサイバーフィジカルシステムを実際に運用する場合の AI セキュリティ対策に関する課題として，対策と精度のトレードオフ，対策の拡張可能性を示した．

2. AI セキュリティの概要

本節では，まず，2.1 節において，AI セキュリティの研究動向を俯瞰する．その上で，2.2 節において，AI システムが攻撃を受けた場合の影響について，推論誤り，情報漏洩，計算浪費の観点から整理する．

2.1 AI セキュリティの研究動向

AI セキュリティの研究は Security for AI と AI for Security の大きく 2 つの観点がある．

Security for AI は，AI システムへの攻撃手法[2]，AI のバイアス・公平性[6，7]など，AI を安全に使うことを目指した AI 自体のセキュリティに関する研究である．これまでに様々な攻撃手法が研究されている．既知の攻撃手法を文献[2]を参考に表 1 にまとめる．

もう一方の AI for Security は，サイバー攻撃に対する防御，不審メールの検出など，様々な分野のセキュリティ向上を目的とした AI の活用方法に関する研究である．なお本稿では，サイバーフィジカルシステムにおいて AI を安全に使うという観点から議論するため，AI for Security については以降触れない．

他にも本人なりすましなどの懸念からフェイクメディアの脅威や対策[8，9]に関する研究も進んでいる．

表 1　既存の攻撃手法

	攻撃者の操作	攻撃手法	完全性 推論誤り	機密性 情報漏洩	可用性 計算浪費
学習	入力に加工	データポイズニング	✓	✓	
		バックドア攻撃	✓		
	モデルに加工	モデルポイズニング	✓	✓	✓
		コードインジェクション攻撃	✓		
		ファインチューニング攻撃	✓	✓	
推論	入力に加工	敵対的攻撃	✓		
		スポンジ攻撃			✓
		プロンプトインジェクション攻撃	✓	✓	✓
	モデルに加工	フォールトインジェクション攻撃	✓	✓	
	モデルの入出力を観測	モデル抽出		✓	
		学習データ情報の収集 ・ メンバーシップ推論 ・ Model Inversion ・ プロパティ推論 ・ 属性推論 ・ データ復元		✓	
		プロンプト窃盗		✓	

また，実用的な観点の動向として AI の脅威分析を行うフレームワーク[10, 11]が公開されており，研究と実用の両方から動向把握が必要である．

2.2 AI を活用したシステムへの攻撃による影響

AI を活用したシステムが攻撃を受けた場合の影響は表 1 に示すように，（1）推論誤り，（2）情報漏洩，（3）計算浪費の観点に整理できる．

（1）推論誤り

攻撃によりモデルが推論を誤ると，異常状態を検出できないなど 1 回の推論誤りが重大な影響となり得る．精度低下にとどまらない深刻な被害となるリスクがある．

（2）情報漏洩

攻撃によって学習データに関する情報が取得されてしまう可能性がある．学習データに個人情報や機密情報が含まれていると，情報漏洩のリスクがある．また，モデル情報が取得されてしまうと，複製されたモデルが悪用されるリスクもある．複製モデルを使って他の攻撃がしやすくなることも考えられる．

（3）計算浪費

攻撃により計算浪費が生じると，処理遅延のリスクがある．リアルタイム性が重要なシステムでは，異常状態の見逃しを誘発するなど，処理遅延にとどまらない深刻な被害となるリスクもある．

3. サイバーフィジカルシステムの分類と AI セキュリティ

本節では，サイバーフィジカルシステムにおける AI の活用例に沿って，想定される AI の脆弱性を整理・分析する．

サイバーフィジカルシステムの代表的な分野として，自動運転，遠隔医療，工場システム，防犯システムなどがあり，どの分野もフィジカル空間をセンシングし，サイバー空間でプロセッシングを行い，フィジカル空間にフィードバックを行う．このようなサイクル中で AI が攻撃される

© 2025 Information Processing Society of Japan

とシステム全体へ影響が広がり，重大な被害になり得る．

まず，3.1 節で AI を活用したサイバーフィジカルシステムをモデルの配置場所に着目して分類し，アタックサーフェスを整理する．次に，3.2 節において，サイバーフィジカルシステムの一例としてミッションクリティカルなシステムにおける AI 活用例を取り上げて，想定される脆弱性，システムへの影響，対策を分析する．

3.1 AI を活用したサイバーフィジカルシステムの分類

AI セキュリティの分析をする際，アタックサーフェスの把握が必要である．サイバーフィジカルシステムにおいて AI を活用したシステムは様々あるが，推論時のモデルの配置場所によってアタックサーフェスが変わる．そこで本稿ではモデルの配置場所の観点から以下の 3 種類に分類する．

（1）中央集約型：AI モデルがシステム提供者の管理ドメイン内（図 2）

（2）エッジ型：AI モデルがエッジデバイス上（図 3）

（3）ハイブリッド型：（1）と（2）のハイブリッド（図 4）

図 2，図 3，図 4 の点線は管理ドメインを表している．そのため点線内部はシステム提供者が管理しており攻撃者はアクセスできないが，エッジデバイスは管理ドメイン外であるため攻撃者がアクセスできると仮定する．本稿では，システム全体へのサイバー攻撃対策はされているという前提で，AI を利用したシステムに特有のアタックサーフェスに着目する．

（1）中央集約型：AI モデルがシステム提供者の管理ドメイン内

図 2 のように，推論時の AI モデルがシステム提供者の管理ドメイン内に配置されている場合を中央集約型と呼ぶことにする．

システムのフローは，まず，システム提供者が AI モデルの学習に使う学習データを収集する．次に，学習データを使って AI モデルを学習する．学習済みの AI モデルはシステム提供者の管理ドメイン内に配置される．そして運用環境に設置されたカメラやセンサでセンシングを行い，AI モデルが推論処理を行う．その後，推論結果が判断ロジックに入力されフィジカル空間へのフィードバック要否を判定する．必要な場合はフィードバックを行う．また，学習済みの AI モデルに対して，必要に応じて運用環境のデータを用いたモデル改良を行う場合も考えられる．

中央集約型でのアタックサーフェスは，学習データと推論時の観測対象・環境である．モデル学習には大量のデータを使うものもあるため，システム提供者が独自で収集したデータのみではなく公開データを使うことも想定される．公開データが攻撃者によって加工されると，不正な学習データでモデルを学習する可能性がある．同様に，推論時のフィジカル空間における観測対象・環境が攻撃者によって加工されると，不正なデータをモデルが推論する可能性がある．運用環境のデータを用いたモデル改良を行う際は，推論時の不正データが学習データとして使われる可能性がある．なお，カメラやセンサによって取得されデジタル化されたデータは，一般的に情報セキュリティ対策がされていると考えられるため，ここではアタックサーフェスにはならないとする．学習時・推論時の AI モデル，及び推論結果は管理ドメイン外に露出しないと想定されるため，基本的にはアタックサーフェスとならない．たとえ利用者になりすました攻撃者がフィードバックから間接的に推論結果を知ろうとしも攻撃可能となる詳細情報を取得できるとは考えにくい．

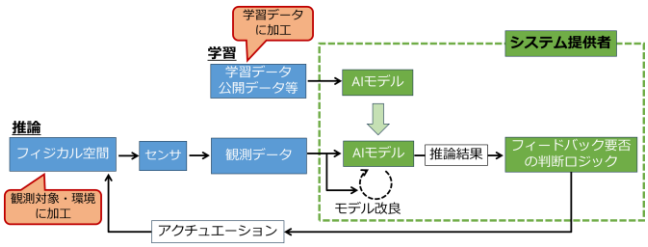

図 2　中央集約型：AI モデルがシステム提供者の管理ドメイン内

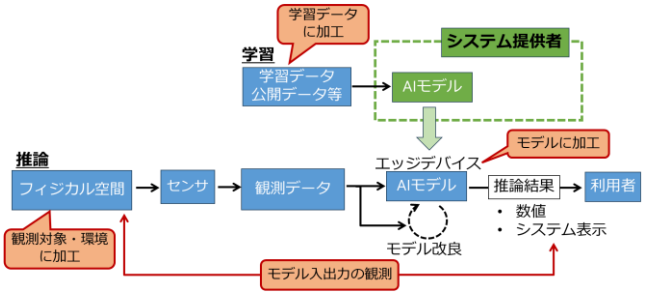

図 3　エッジ型：AI モデルがエッジデバイス上

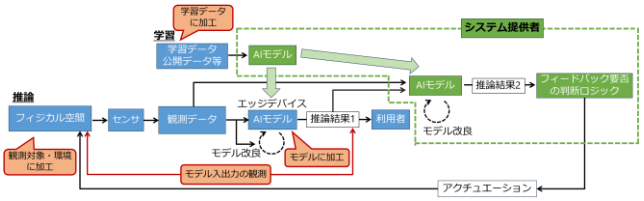

図 4　ハイブリッド型：AI モデルがシステム提供者内の管理ドメイン内とエッジデバイス上

（2）エッジ型：AI モデルがエッジデバイス上

図 3 のように，推論時の AI モデルがエッジデバイス上に配置されている場合をエッジ型と呼ぶことにする．

ここでは，エッジデバイス上の AI モデルが推論処理を行い，推論結果の通知が利用者へ直接届くエッジ型を想定する．

エッジ型でのアタックサーフェスは，中央集約型と同様に学習データ（公開データ），推論時の観測対象・環境がある．他にも推論時の AI モデル，推論結果がアタックサーフ

© 2025 Information Processing Society of Japan

ェスとなる．エッジデバイス上のモデルに対して攻撃者がデバイスにアクセスできるためモデル加工による攻撃のリスクがある．また，利用者になりすました攻撃者がモデルの推論結果を観測すると，モデル入出力の観測による攻撃が可能となる．推論結果が数値表示の場合とシステム的な表示のみの場合があるが，数値表示の場合だとより詳細な情報取得が可能となるため攻撃しやすくなる．

(3) ハイブリッド型：AI モデルがシステム提供者の管理ドメイン内とエッジデバイス上

図 4 のように，推論時の AI モデルがシステム提供者の管理ドメイン内とエッジデバイス上の両方に配置されている場合をハイブリッド型と呼ぶことにする．

ハイブリッド型は複数考えられるが，ここでは一例として以下を想定する．カメラやセンサで取得したデータに対して最初にエッジデバイス上の AI モデルが軽量な推論処理を行う．推論結果が途中結果として利用者に通知される場合もある．そして，その推論結果を受けてシステム提供者の管理ドメイン内に配置されたモデルが高度な推論処理を行う．そして推論結果が判断ロジックに入力されフィジカル空間へのフィードバック要否を判定し，必要な場合はフィードバックを行う．

ハイブリッド型でのアタックサーフェスは，中央集約型，及びエッジ型と同様に学習データ（公開データ），推論時の観測対象・環境がある．またエッジ型と同様に推論時のエッジデバイス上のモデルもアタックサーフェスとなる．システム提供者の管理ドメイン内に配置されたモデルはアタックサーフェスとはならないが，エッジデバイス上のモデルが攻撃されると，その推論結果を通じて他のモデルに影響が波及するリスクはある．また，エッジデバイス上のモデルの推論結果が途中結果として利用者に通知されると，利用者になりすました攻撃者が推論結果を観測できる可能性があるため，エッジデバイス上のモデルの推論結果もアタックサーフェスとなり得る．

表 2 に各システム分類で存在する AI セキュリティとしてのアタックサーフェスを示す．表 2 ではアタックサーフェスを AI モデルに対する学習時入力データ，推論時入力データ，モデル自身，出力である推論結果に分類した．

なお，本節で説明したフィジカル空間とサイバー空間の要素は，サイバーフィジカルシステムの様々な分野に当てはまるので汎用的に活用できる．

3.2 ミッションクリティカルなシステムにおける AI セキュリティ

ここまではサイバーフィジカルシステム全体について考察を行ってきたが，ここからは AI セキュリティでの脆弱性，システムへの影響，対策の詳細な分析を行うために特定の分野に対象を絞る．分析対象としてサイバーフィジカルシステムの応用先の一つであるミッションクリティカルなシステム，すなわち生命や財産に関わるシステム（例：防犯システム，自動運転システムなど）を取り上げる．例えば入室権限を設定している領域への不正な侵入を検知するシステムがある．ミッションクリティカルなシステムでは高いセキュリティレベルが要求される．具体的には，不正な侵入などの異常状態の見逃しが起こらないこと，リアルタイムで処理が行われることが重要な条件である．以降ではミッションクリティカルなシステムの一種である異常検知システムと人物推定システムを想定し AI セキュリティの脆弱性，システムへの影響，対策を考察する．

表 2　アタックサーフェス・システム分類・アタックサーフェス有無の理由

アタックサーフェス	システム分類			アタックサーフェス有無の理由
	中央集約	エッジ	ハイブリッド	
学習時入力データ	✓	✓	✓	学習データとして公開データの使用を想定 公開データに攻撃
推論時入力データ	✓	✓	✓	フィジカル空間で攻撃
モデル	×	✓	△	• 中央集約型 推論時モデルは管理ドメイン内 攻撃者はモデルにアクセス不可 • エッジ型，ハイブリッド型 モデル配置のデバイスに攻撃者がアクセス ハイブリッド型ではエッジデバイス上のモデルへの攻撃が他モデルへ波及
推論結果	△	✓	△	• 中央集約型 推論結果は管理ドメイン内 フィードバックから攻撃者が推論結果を間接的に取得 • エッジ型，ハイブリッド型 利用者になりすました攻撃者が推論結果を観測 ハイブリッド型では推論途中結果を観測

3.2.1 異常検知システム

ミッションクリティカルなシステムの一例として，上記の不正な侵入などを対象とした認識 AI による異常状態の検知システムを想定する．異常検知を行う場所は個人宅，店舗，オフィス，大規模商業施設，スタジアム，駅，空港と様々な場所が想定されるため，それらに応じて検知する異常の種類も異なる．不正な侵入の他にも，異常行動（万引き，ケンカ，不審物放置）の検知などを想定する．

- **学習時・推論時の不正入力データによる推論誤り**

（1）脆弱性

学習時・推論時の不正入力データによる攻撃は 3.1 節で述べた全てのシステム分類（図 2, 図 3, 図 4）で生じ得る．

- 学習時

  学習データとして公開データを使用することが考えられるため，公開データ加工による攻撃が想定される．例えば，データポイズニング[12, 13]による推論誤り，バックドア攻撃[14，15]による特定パターンをトリガーとした推論誤りがある．バックドア攻撃ではフィジカル空間での攻撃手法[16]も研究されている．

- 推論時

  カメラやセンサで取得するデータに対するフィジカル空間での攻撃が想定され，敵対的攻撃[17, 18]による推論誤りのリスクがある．敵対的攻撃は，ある特定のモデルに対して作成された攻撃

© 2025 Information Processing Society of Japan

データが同様のタスクのモデルに対しても推論誤りを引き起こす性質[19，20]が知られており，攻撃者が標的とするモデルの詳細情報を取得できなくとも攻撃できる可能性がある．

- モデル改良時

システム提供先のデータを使ったモデル改良を行うことがある．この際に，推論時の不正入力データが学習データとして使われると，意図しないモデル改変となり推論誤りにつながる可能性がある．

（2）影響

推論誤りにより不正な侵入などの異常状態が検出できないと，利用者の生命や財産へ深刻な被害を引き起こす可能性がある．

なお，カメラやセンサによって取得されデジタル化された入力データへの攻撃は一般的な情報セキュリティ対策により対策可能である．しかし，スポンジ攻撃[21]で計算浪費が生じシステムのリアルタイム性が低下すると，異常状態の見逃しを誘発する可能性があり，サイバー空間だけではなくフィジカル空間へも影響が及ぶと考えられる．

（3）対策

学習時の不正入力データによる攻撃（公開データ，モデル改良時のシステム提供先データ）の対策として，データ選別による不正データの除去[22－24]，学習済みモデルが正常か攻撃の影響を受けているかのモデル判定[25, 26], 不正データの影響を受けたモデルの修正[27]がある．

学習時・推論時の不正入力データによる攻撃に共通の対策として，不正データの影響を低減するモデル学習の工夫[28－33]，推論時の入力データが不正データかを判定する手法[34－37]，不正データの影響を低減するモデル推論の工夫[34，38，39]がある．

- **モデル入出力観測によるモデル情報漏洩**

（1）脆弱性

モデルの配置場所がエッジデバイス上であるエッジ型（図 3）とハイブリッド型（図 4）では，利用者になりすました攻撃者が推論結果を観測できる可能性がある．

モデル入出力観測からモデル情報が推定されて，複製モデルが作成されるリスクがある[40, 41]. 推論結果の形式として数値の場合とシステム的な通知のみの場合があるが，数値の方が詳細な情報取得につながり攻撃しやすくなる．

（2）影響

作成された複製モデルが悪用される可能性があり，システム提供者が被害者となる．また，複製モデルを使って他の攻撃がしやすくなることも考えられる．

（3）対策

推論結果の数値の離散化や丸めなど連続的な数値ではない表示形式が必要である．

- **推論時のモデル加工による推論誤り**

（1）脆弱性

モデルの配置場所がエッジデバイス上であるエッジ型（図 3）とハイブリッド型（図 4）では，攻撃者がデバイスへアクセスできる可能性があり，モデル加工による攻撃で推論誤り[42]のリスクがある．

（2）影響

推論誤りは異常検知の精度低下のリスクがある．また推論誤りにより異常状態が検出できないと，利用者の生命や財産へ深刻な被害を引き起こす可能性がある．

（3）対策

モデル加工による推論誤りの対策では，攻撃の影響を低減する学習工夫[43]，モデルが攻撃を受けていないかのモデル判定とモデル修正[44]などがある．

3.2.2 人物推定・特定システム

ミッションクリティカルなシステムの一例として人物の推定や特定を想定する．例えば，入室権限がある人物に対するロック解除，立ち入り記録，迷子など類似人物検索を想定する．3.2.1 節で説明した攻撃の他にも注意したい攻撃について議論する．

- **ディープフェイクによるなりすまし**

（1）脆弱性

ディープフェイクによるなりすましは推論時の不正入力データによる攻撃に相当するため, 3.1 節で述べた全てのシステム分類（図 2，図 3，図 4）で生じ得る．

AI で合成したターゲット人物の顔画像や動画を提示することにより, 攻撃者がターゲット人物として誤認識され，なりすましが成立するリスクがある．

（2）影響

入室権限のない攻撃者がアクセス制限領域に不正に侵入し，窃盗や機密情報を取得するリスクがあるため，利用者の生命や財産への深刻な被害につながる可能性がある．

（3）対策

認識 AI を使ったバイオメトリクスによる人物の推定だけではなく，IC カードを組み合わせるといった運用面からの対策がある．AI セキュリティ対策では，デプス情報を利用した認識手法の活用，入力データが合成データであるかを判定する手法[8，45]がある．

- **モデル入出力観測による学習データ情報漏洩**

（1）脆弱性

モデルの配置場所がエッジデバイス上であるエッジ型（図 3）とハイブリッド型（図 4）では，利用者になりすました攻撃者が推論結果を観測できる可能性がある．

モデル入出力観測による攻撃として，メンバーシップ推論[46], データ復元[47]など学習データの情報が推測される可能性がある．

（2）影響

人物の推定や特定では，学習データとして利用者の個人

© 2025 Information Processing Society of Japan

情報を使っているため個人情報漏洩のリスクがあり，被害者は利用者となる．

（3）対策

推論結果の数値の離散化や丸めなど連続的な数値では出力しない表示形式が必要である．また，差分プライバシーを活用した学習工夫としてモデル重みの勾配にノイズ付加する方法[48，48]も有効である．

4. 分析と課題

4.1 アタックサーフェスに応じた AI セキュリティの分析

3.1 節では，各システム分類で存在する AI セキュリティとしてのアタックサーフェスを表 2 にまとめた．例えば，中央集約型の場合は，推論時のモデルに対する攻撃は存在しないなどシステム分類によって存在するアタックサーフェスが異なる．よって，システム分類に応じて存在するアタックサーフェスを確認し，必要な対策を選択する必要がある．

次に，アタックサーフェスごとの AI セキュリティの分析を表 3 に示す．表 3 の攻撃の汎用性は，ここでは表 2 に基づきアタックサーフェスの数が多いほど汎用性が高いとしている．攻撃，影響，被害者，AI セキュリティ対策，情報セキュリティ・運用対策は 3.2 節の事例に基づいて記している．以下では，攻撃の汎用性，影響と被害者，対策の観点から分析する．

- 攻撃の汎用性

  アタックサーフェスが学習時入力データと推論時入力データの場合，システム分類によらず想定される脆弱性であるため汎用性が高い．アタックサーフェスがモデルと推論結果の場合は，エッジ型とハイブリッド型の場合に気を付ければよい．

- 影響と被害者

  利用者が被害者となる場合に優先的に対策すべきである．推論誤りが引き起こされて異常を正常と推論すると異常状態の見逃しにつながり，利用者の生命や財産への深刻な被害となり得るため対策の重要度が高いと考えられる．一方，学習データに使用した利用者情報が漏洩すると，直接的に生命，財産に影響しないが利用者のプライバシー上の被害となる．モデル情報漏洩が生じる場合，被害者はシステム提供者となる．

- 対策

  AI セキュリティ対策と情報セキュリティ・運用対策の 2 つの観点からの対策がある．

  AI セキュリティ対策について，攻撃の種別ごとに対策が提案されており，適した対策を選択する必要がある．また，複数のアタックサーフェスに共通した対策もあるため有効活用が期待できる．

  情報セキュリティ・運用対策については，アタックサーフェスによって活用できる場合とできない場合がある．モデル，推論結果，推論時入力データ（ディープフェイク）への攻撃には情報セキュリティ・運用対策が活用できる．学習時入力データ，推論時入力データ（観測対象・環境加工）への攻撃には情報セキュリティ・運用対策が活用できないため，AI セキュリティでの対策が必要である．

表 3　アタックサーフェスに応じた AI セキュリティの分析

アタックサーフェス	攻撃	攻撃の汎用性	影響→被害者	AIセキュリティ対策	情報セキュリティ・運用対策
学習時入力データ	データ加工	高	推論誤り→利用者	学習 ・ 学習データ選別 ・ 学習工夫	×
				学習後 ・ 入力データ判定 ・ 推論工夫 ・ モデル判定 ・ モデル修正	
推論時入力データ	・ 観測対象，環境加工 ・ ディープフェイク	高	推論誤り→利用者	学習 ・ 学習工夫 (モデル改良時) ・ 学習データ選別	・ 観測対象，環境加工× ・ ディープフェイク○
				学習後 ・ 入力データ判定 ・ 推論工夫 (モデル改良時) ・ モデル判定 ・ モデル修正	
モデル	モデル加工	低	・ 推論誤り→利用者 ・ モデル情報漏洩 →システム提供者	学習 ・ 学習工夫	○
				学習後 ・ モデル判定 ・ モデル修正	
推論結果	モデル入出力を観測	中	・ 学習データ情報漏洩 →利用者 ・ モデル情報漏洩 →システム提供者	学習 ・ 学習工夫	○
				学習後 ・ 出力形式	

本稿での分析結果としては，**攻撃の汎用性の高さ，被害の深刻さ，情報セキュリティ・運用対策が活用できないという観点から，**アタックサーフェスが学習時入力データと推論時入力データとなる攻撃に対して AI セキュリティ対策の優先度が高いと考えられる．

4.2 運用・実環境上の課題

AI を活用したサイバーフィジカルシステムを実際に運用する場合，表 3 の AI セキュリティ対策の選定や実装方法について課題が考えられる．本稿では（**1）対策と精度のトレードオフ，（2）対策の拡張可能性**について考察する．

（1）対策と精度のトレードオフ

表 3 に掲載した AI セキュリティ対策のうち，精度へ影響を及ぼす可能性のある対策がある．学習工夫の対策の中には，モデルが攻撃の影響を受けにくくなるようにモデル重みの勾配にノイズや制約を付加して学習する手法[29，30，32，33，48，49]がある．推論工夫の対策の中にも推論時の入力データなどにノイズを付加する手法[39]がある．これらの手法は攻撃の影響低減には効果があるが，精度低下につながる状況も考えられるため，精度とこれらの対策はトレードオフの関係になる．よってミッションクリティカルなシステムのように精度が優先されるシステムでは，有用性が低下する可能性がある．実環境に応じて対策の取り入れ方を検討する必要がある．

（2）対策の拡張可能性

3.1 節で，運用環境のデータを使ったモデル改良を行う場合の脆弱性に言及した．この際に，推論時の不正入力データを学習データとして使うと意図しないモデル改変につな

© 2025 Information Processing Society of Japan

がるリスクがある．対策の一つとして正常なモデルか改変された異常なモデルかを判定するモデル判定[25，26]を挙げた．ただしモデル改良ではある特定の環境へモデルを適応させるため，たとえ不正入力データによる攻撃がなくとも，モデルの過剰適応いわゆる過学習によりモデルに一種の異常が生じる可能性もある．よって，実環境では攻撃由来，及び過学習由来のモデル異常が生じ得ることを考慮してモデル判定の対象を拡張すると，実環境でより有用性の高い対策となることが期待される．

実環境で AI セキュリティ対策を行う場合は，上記 2 つの課題を表 3 に反映させつつ検討することが望まれる．

5. おわりに

本稿では AI セキュリティの研究動向を俯瞰し，AI を活用したシステムが攻撃された場合の影響について，推論誤り，情報漏洩，計算浪費の観点から整理した．

次に，サイバーフィジカルシステムを AI モデルの配置場所で分類し，アタックサーフェスを整理した．システム分類によって存在するアタックサーフェスが異なるため，対策を選択する必要があることを明らかにした．これを踏まえて，サイバーフィジカルシステムの一例としてミッションクリティカルなシステムにおける AI の活用例を取り上げて，想定される脆弱性を分析した．本稿では，(1) 攻撃の汎用性の高さ，(2) 被害の深刻さ，(3) 情報セキュリティ・運用対策が活用できないという 3 つの観点に着目した．その結果，アタックサーフェスが学習時入力データ，推論時入力データとなる攻撃に対して AI セキュリティ対策の優先度が高いと結論付けた．

最後に，AI を活用したサイバーフィジカルシステムを実際に運用する場合の AI セキュリティ対策に関する課題として，対策と精度のトレードオフ，対策の拡張可能性を示した．状況に応じて AI セキュリティ分析に反映させて対策することが望まれる．

参考文献

[1] CSS2024: AI Security Workshop (AWS) 2024, ⟨https://www.iwsec.org/aws/2024/⟩ (2024).

[2] Naoto, K., KengoZ. and Takayuki, S.: Securing AI Systems: A Guide to Known Attacks and Impacts, ⟨https://arxiv.org/abs/2506.23296⟩ (2025).

[3] IJCAI 2024: Workshop on Artificial Intelligence Safety (AISafety), ⟨https://www.aisafetyw.org/⟩ (2024).

[4] AAAI 2024: AAAI-24 Special Track Safe, Robust and Responsible AI Track, ⟨https://ojs.aaai.org/index.php/AAAI/issue/view/594⟩ (2024).

[5] NIST: Cyber-Physical Systems and Internet of Things, ⟨https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1900-202.pdf⟩ (2019).

[6] Simon, C. and Christian, H.: Fairness in Machine Learning: A Survey, *ACM Computing Surveys*, Vol. 56, No. 7, Article 166, ⟨https://doi.org/10.1145/3616865⟩ (2024).

[7] Ulrich, A., Hiromi, A.,Olivier, F., Sebastien, G., Satoshi, H. and Alain, T.: Fairwashing: the risk of rationalization, *The 36th International Conference on Machine Learning (ICML)*, pp. 161–170 (2019).

[8] 井口駿治，稲葉宏幸: 各種生成 AI モデルに対する AI 生成画像検出ツールの性能比較に関する調査研究，コンピュータセキュリティシンポジウム 2024 論文集, pp. 1295–1299 (2024).

[9] Utkarsh, O., Yuheng, L. and Yong, J. L.: Towards Universal Fake Image Detectors that Generalize Across Generative Models, *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR)*, pp. 24480–24489, (2023).

[10] OWASP Top10 LLM applications and generative AI, ⟨https://genai.owasp.org/llm-top-10/⟩.

[11] MITRE ATLAS, ⟨https://atlas.mitre.org/matrices/ATLAS⟩.

[12] Nicholas, C.: Poisoning the Unlabeled Dataset of Semi-Supervised Learning, *30th USENIX Security Symposium*, pp. 1577–1592 (2021).

[13] Chen, Z., W. Ronny, H., Ali, S., Hengduo, L., Gavin, T., Christoph, S. and Tom, G.: Transferable Clean-Label Poisoning Attacks on Deep Neural Nets, *The 36th International Conference on Machine Learning (ICML)*, pp. 7614–7623 (2019).

[14] Tianyu, G., Kang, L., Brendan, D. G. and Siddharth, G.: BadNets: Evaluating Backdooring Attacks on Deep Neural Networks, *IEEE Access*, Vol. 7, pp. 47230-47244 (2019).

[15] Nguyen, A. and Tran, A.: WaNet-imperceptible warping-based backdoor attack, *The 9th International Conference on Learning Representations (ICLR)* (2021).

[16] Emily, W., Josephine, P., Arjun, N. B., Yuanshun, Y., Haitao, Z. and Ben, Y. Z.: Backdoor Attacks Against Deep Learning Systems in the Physical World, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6206–6215 (2021).

[17] Kevin, E., Ivan, E., Earlence, F., Bo, L., Amir, R., Chaowei, X., Atul, P., Tadayoshi, K. and Dawn, S.: Robust Physical-World Attacks on Deep Learning Visual Classification, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–1634 (2018).

[18] Simen, T., Wiebe, V. R. and Toon, G.: Fooling automated surveillance cameras: adversarial patches to attack person detection, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019).

[19] Ian, J. G., Jonathon, S. and Christian, S.: Explaining and Harnessing Adversarial Examples, *3rd International Conference on Learning Representations (ICLR)* (2015).

[20] Kevin, E., Ivan, E., Earence, F., Bo, L., Amir, R., Florian, T., Atul, P., Tadayoshi, K. and Dawn, S.: Physical Adversarial Examples for Object Detectors, *12th USENIX Conference on Offensive Technologies (WOOT'18)*, (2018).

[21] Ilia, S., Yiren, Z., Daniel, B., Nicolas, P., Robert, M. and Ross, A.: Sponge Examples: Energy-Latency Attacks on Neural Networks, *6th IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 212–231 (2021).

[22] Brandon, T., Jerry, L. and Aleksander, M.: Spectral Signatures in Backdoor Attacks, *The 32nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 8011–8021 (2018).

[23] Jonathan, H., Weihao, K., Raghav, S. and Sewoong, O.: SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics, *The 38th International Conference on Machine Learning (ICML)*, pp. 4129–4139 (2021).

[24] Pang, W. K. and Percy, L.: Understanding Black-box Predictions via Influence Functions, *The 34th International Conference on Machine Learning (ICML)*, pp. 1885–1894 (2017).

[25] Soheil, K., Aniruddha, S., Hamed, P. and Heiko, H.: Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs, *2020 IEEE/CVF Conference on Computer Vision and Pattern*

© 2025 Information Processing Society of Japan

Recognition (CVPR), pp. 301–310 (2020).
[26] Shanjiaoyang, H., Weiqi, P., Zhiwei, J. and Zhuowen, T.: One-Pixel Signature: Characterizing CNN Models for Backdoor Detection, *16th European Conference on Computer Vision (ECCV)*, pp. 326–341 (2020).
[27] Yi, Z., Si, C., Won, P., Z. Morley, M., Ming, J. and Ruoxi, J.: Adversarial Unlearning of Backdoors via Implicit Hypergradient, *The 10th International Conference on Learning Representations (ICLR)* (2022).
[28] Florian, T., Alexey, K., Nicolas, P., Ian, G., Dan, B. and Patrick, M.: Ensemble Adversarial Training: Attacks and Defenses, *The 6th International Conference on Learning Representations (ICLR)* (2018).
[29] Seyed, M. M. D., Alhussein, F., Jonathan, U. and Pascal, F.: Robustness via curvature regularization, and vice versa, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9078–9086 (2019).
[30] Nicolas, P., Patrick, M., Arunesh, S. and Michael, P. W.: SoK: Security and Privacy in Machine Learning, *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399–414 (2018).
[31] Yige, L., Xixiang, L., Nodens, K., Lingjuan, L., Bo, L. and Xingjun, M.: Anti-Backdoor Learning: Training Clean Models on Poisoned Data, The *34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 14900–14912 (2021).
[32] Min, D., Ruoxi, J. and Dawn, S.: Robust Anomaly Detection and Backdoor Attack Detection Via Differential Privacy, *The 8th International Conference on Learning Representations (ICLR)* (2020).
[33] Sanghyun, H., Varun, C., Yigitcan, K., Tudor, D. and Nicolas, P.: On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping, ⟨https://arxiv.org/abs/2002.11497⟩ (2020).
[34] Hang, W., Zhen, X., David, J. M. and George, K.: MM-BD: Post-Training Detection of Backdoor Attacks with Arbitrary Backdoor Pattern Types Using a Maximum Margin Statistic, *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1994–2012 (2024).
[35] Mahesh, S., Nilesh, A., Ranganath, K., Ibrahima, J. N. and Omesh, T.: Deep Probabilistic Models to Detect Data Poisoning Attacks, *Fourth workshop on Bayesian Deep Learning, 33rd Annual Conference on Neural Information Processing Systems (NeurIPSWorkshop)* (2019).
[36] Xingjun, M., Bo, L., Yisen, W., Sarah, M. E., Sudanthi, W., Grant, S., Dawn, S., Michael, E. H. and James, B. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality, *The 6th International Conference on Learning Representations (ICLR)* (2018).
[37] Kathrin, G., Praveen, M., Nicolas, P., Michael, B. and Patrick, M.: On the (Statistical) Detection of Adversarial Examples, ⟨https://arxiv.org/abs/1702.06280⟩ (2017).
[38] Xuanqing, L., Minhao, C., Huan, Z. and Cho, J. H.: Towards robust neural networks via random self-ensemble, *15th European Conference on Computer Vision (ECCV)*, pp. 381–397 (2018).
[39] Mathias, L., Vaggelis, A., Roxana, G., Daniel, H. and Suman, J.: Certified Robustness to Adversarial Examples with Differential Privacy, *2019 IEEE Symposium on Security and Privacy (SP), pp. 656–672* (2019).
[40] Tribhuvanesh, O., Bernt, S. and Mario, F.: Knockoff Nets: Stealing Functionality of Black-Box Models, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4954–4963 (2019).
[41] Jean, B. T., Pratyush, M., Robert, J. W. and Nicolas, P.: Data-Free Model Extraction, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4771–4780 (2021).
[42] Shaofeng, L., Xinyu, W., Minhui, X., Haojin, Z., Zhi, Z., Yansong, G., Wen, W. and Xuemin, S.: Yes, One-Bit-Flip Matters! Universal DNN Model Inference Depletion with Runtime Code Fault Injection, *33rd USENIX Security Symposium*, pp. 1315–1330 (2024).
[43] Jialai, W., Ziyuan, Z., Meiqi, W., Han, Q., Tianwei, Z., Qi, L., Zongpeng, L., Tao, W. and Chao, Z.: Aegis: Mitigating Targeted Bit-flip Attacks against Deep Neural Networks, *32nd USENIX Security Symposium*, pp. 2329–2346 (2023).
[44] Jingtao, L., Adnan, S. R., Zhezhi, H., Deliang, F. and Chaitali, C.: RADAR: run-time adversarial weight attack detection and accuracy recovery, *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (2021).
[45] Zhiyuan, Y., Yong, Z., Xinhang, Y., Siwei, L. and Baoyuan, W.: DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection, The *37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4534–4565 (2023).
[46] Nicholas, C., Steve, C., Milad, N., Shuang, S., Andreas, T. and Florian, T.: Membership Inference Attacks From First Principles, *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914 (2022).
[47] Matt, F., Somesh, J. and Thomas, R.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *The 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1322–1333 (2015).
[48] Ho, B., Jaehee, J., Dahuin, J., Hyemi, J., Heonseok, H., Hyungyu, L. and Sungroh, Y.: Security and Privacy Issues in Deep Learning, ⟨https://arxiv.org/abs/1807.11655⟩ (2018).
[49] Xuechen, L., Florian, T., Percy, L. and Tatsunori, H.: Large Language Models Can Be Strong Differentially Private Learners, *The 10th International Conference on Learning Representations (ICLR)* (2022).

© 2025 Information Processing Society of Japan