人や社会への負の影響に配慮した AIセキュリティ情報の体系化

加藤 広野[1,a)]　北 健太朗[1]　長谷川 健人[1]　披田野 清良[1]

概要：本稿では，AI セキュリティに関連する要素間の相互関係や情報システムおよび人や社会への負の影響を体系的に整理した「AI セキュリティマップ」を提案する．このマップは，情報システム的側面と外部作用的側面の 2 つの側面から構成される．情報システム的側面には情報システム内で AI が満たすべき要素が分類され，外部作用的側面には AI が攻撃または悪用された際に人や社会に影響を与える要素が含まれている．AI セキュリティマップでは，各要素に関連する負の影響を特定されており，本マップを参照することで潜在的な負の影響，その原因，および防御・対策を理解できる．また，AI を利用したシステムに対する負の影響がどのように人や社会へ波及するかを理解するためにも役立つ．

キーワード：AI セキュリティ, プライバシ, AI の毀損, AI の悪用, 社会的影響

On Systematization of AI Security Information Considering Negative Impacts on Individuals and Society

Hiroya Kato[1,a)]　Kentaro Kita[1]　Kento Hasegawa[1]　Seira Hidano[1]

Abstract: In this paper, we first develop an AI security map that holistically organizes interrelationships among elements related to AI security as well as negative impacts on information systems and stakeholders. This map consists of the two aspects, namely the information system aspect (ISA) and the external influence aspect (EIA). The elements that AI should fulfill within information systems are classified under the ISA. The EIA includes elements that affect stakeholders as a result of AI being attacked or misused. For each element, corresponding negative impacts are identified. By referring to the AI security map, one can understand the potential negative impacts, along with their causes and countermeasures. Additionally, our map helps clarify how the negative impacts on AI-based systems relate to those on stakeholders.

1. はじめに

社会での AI の利活用が盛んになるにつれて，AI セキュリティの範囲はもはや AI だけでなく，人や社会といった様々なステークホルダーにまで拡大している．また，AI セキュリティは情報セキュリティの要素である機密性，完全性，可用性（CIA）と密接に関連するだけでなく，説明可能性や公平性といった AI 固有の要素とも強く結びついている．そのような状況下において，人や社会への負の影響を理解することは重要であるが，様々な要素間や社会的影響の関係性はさらに複雑化している．したがって，**AI セキュリティに関連する知識，技術，社会的影響は，それらの関係性を理解するために体系的に整理されるべきである**．しかしながら，既存のサーベイ論文 [1], [2], [3], [4] および SoK 論文 [5], [6], [7] の多くは，特定の分野や AI 要素の個別的な観点から関連する技術，攻撃・防御，および社会的影響を整理するにとどまっており，その関係性を整理することは容易ではない．そこで，本稿ではそれらの相互関係および人や社会への影響に配慮した AI セキュリティ情報の体系化として，「AI セキュリティマップ」を作成した．本マップは，情報システム的側面（Information System Aspect,

[1] KDDI 総合研究所
KDDI Research, Inc.
[a)] ia-katou@kddi-research.jp

ISA）と外部作用的側面（External Influence Aspect, EIA）の二つの側面から構成される．AI セキュリティに関連する要素，および AI が攻撃または悪用された結果，人や社会がどのような影響を受けるかを整理している．本マップを参照することで，AI に基づいた情報システム（AI システム）における負の影響や防御・対策を理解できる．さらに，AI セキュリティマップは，情報システムにおける負の影響がどのように人や社会に波及するかを明確化している．本研究の貢献は以下の通りである．

(1) AI に関連する複数の要素のみならず，人や社会に影響を及ぼす要素も含めて体系的に整理した．
(2) 2 つの側面の相互関係を考慮し，情報システム上の負の影響が人や社会にどのように波及するかを整理した．
(3) 本マップでは，AI に対する攻撃による負の影響だけでなく，AI の悪用による影響も整理した．
(4) 本稿は AI セキュリティ全体を俯瞰的な視点で整理し，負の影響の波及にも着目した．紙面の都合上，本マップに深く関係する内容や文献のみを示しているため，詳細な記載に関しては，arXiv にて公開している論文[*1]を参照されたい．また，本マップの全体像に関しては，「AI セキュリティポータル」[*2]に公開している．

2. AI セキュリティの現状

AI セキュリティの研究が多様化・複雑化しているため，主要なアプローチや評価手法を体系的かつ批判的にレビューするサーベイ論文や SoK 論文が多数存在する．

2.1 攻撃と防御

AI セキュリティ分野における全体像を把握する手段として，既存のサーベイ論文 [1], [2], [3], [4] および SoK 論文 [5], [6], [7] は，主に AI への攻撃や防御の整理に焦点を当てている．既存研究における分類法には，AI システムにおける分類，特定の攻撃における分類，特定の分野における分類の 3 つが存在する．

AI システムにおける分類. Hu ら [1] は AI セキュリティの課題と研究動向をレビューし，AI セキュリティの全体像を示している．この研究では，AI システムのライフサイクルを指針として，各段階で顕在化するセキュリティ脅威と対応策を整理している．Kiribuchi ら [3] は，AI システムに対する敵対的攻撃の全体像をまとめ，主要な 11 種類の攻撃手法とそれに伴う影響を特定し，AI システムに対する攻撃の影響を直感的なビジュアルサマリーを提供している．

特定の攻撃における分類法. AI に対する特定の攻撃に関する包括的な整理を行う研究が存在する．Ramirez ら [8] は，ポイズニング攻撃に関連する最新の知見や発見，複数の防御手法をまとめた調査を行っている．Zhang ら [9] は，マルチドメインの AI モデルを対象としたバックドア攻撃および防御について，網羅的かつ体系的な概要を提示している．Lu ら [4] は，大規模言語モデル（LLM）に対するジェイルブレイク攻撃への耐性を包括的に評価するためのフレームワークを提案し，28 種類以上の攻撃手法と 12 種類の防御手法を整理している．

特定の分野における分類法. Shen ら [2] は，自動運転の分野における AI セキュリティ研究の知識体系化を初めて実施し，53 件の既存論文の収集・分析を行っている．Wingarz ら [7] は，Edge AI のセキュリティと安全性向上に必要な課題を体系的に整理している．

2.2 AI セキュリティに関連する要素

既存の体系化研究では，AI に関連する要素ごとの様々なリスク・課題や有望な研究の方向性がまとめられている．

説明可能性. 説明可能な AI（XAI）は，AI モデルがどのような理論や過程で予測を行うのかを人間に理解可能な形で説明する技術を指す．XAI 分野の体系的研究として，Dwivedi ら [10] は XAI のプログラミング技術を調査し，機械学習開発プロセスに沿って XAI 技術を整理している．

公平性. 公平性に関するレビューとして，Zhang ら [11] は実社会での実装に向けた AI の公平性に関する最近の進展をレビューし，公平性の理論的基盤と実世界データの複雑な現実を調和させるための解決策を提案している．一方で，Parraga ら [12] は，視覚および言語領域における公平性を考慮したニューラルネットワークの代表的なバイアス除去手法について詳細なレビューを行っている．

プライバシー. Golda ら [13] は，生成 AI に内在するプライバシーおよびセキュリティ上の課題を詳細に分析し，これらの複雑性を包括的に理解するために不可欠な 5 つの視点を提供している．Chang ら [5] は，ヘルスケアが利用されるシナリオや脅威モデルにおける不整合性に着目し，ヘルスケア AI に対するプライバシー漏洩を引き起こす攻撃および防御策を体系的に整理している．

2.3 実世界におけるリスクと影響

AI の毀損や悪用 [14], [15] によって実世界で生じるリスクや影響 [16], [17] について論じた研究も存在する．

AI の社会的影響. Weidinger ら [17] は，言語モデルに関連する倫理的・社会的リスクの包括的な分類法を構築している．Slattery ら [16] は，AI が労働市場，経済動態，倫理的課題に与える影響を考慮しつつ，技術的進歩と社会的責任の両立を促すアプローチの必要性を示している．Pankajakshan ら [18] は，LLM セキュリティおよびステークホルダーへのリスクについて論じている．

AI の悪用. AI の悪用も実世界に負の影響をもたらす要因

[*1] https://arxiv.org/abs/2508.08583
[*2] https://aisecurity-portal.org/

の一つである．Aïmeur ら [14] は，フェイクニュースが社会に重大な影響を及ぼすことや，AI を利用したツールにより偽情報の生成が容易になっており検知が困難になることを指摘している．Castagnaro ら [15] は，AI がディレクトリ列挙と呼ばれるサイバーセキュリティにおける情報収集手法を強化可能かを検証し，新たな LM ベースのフレームワークを提案している．

2.4 既存の体系化に関する研究の課題

AI セキュリティに関する体系化研究の多くは，主に技術的観点から AI への攻撃や防御手法の分類を通じて知識を整理している．一方で，AI 要素に関連する技術や一部のリスクについても既存研究で議論されているが，それらは公平性やプライバシーといった個別の要素にのみ着目している．そのため，これらの要素と AI セキュリティとの関係性については，密接かつ重要な関連性があるにもかかわらず，議論や整理がなされていない．また，一部の論文では，攻撃や防御の分類に加え，人や社会への影響についても扱っているが，その概要は特定の分野における実世界のリスクや影響に限定されている．さらに，特定のリスクや影響から派生する二次的な影響については考察されていない．実環境における人や社会への負の影響を理解するためには，このような影響の波及を考慮することが重要である．今日の AI は情報システムのみならず，様々なステークホルダーにも多様な影響を及ぼしている．その結果，AI セキュリティの範囲は一部の研究者に限定されるものではなく，エンジニア，ユーザー，さらには非ユーザーも含めた広範な人々に拡大している．つまり，もはや個別の要素を検討するだけでは不十分であり，複数の要素を絡め，人や社会への負の影響に配慮した体系化が必要である．

3. AI セキュリティ情報の体系化

3.1 本研究の主張

前節で述べたように，既存の論文を複数参照しても，要素間の関係性を把握することは容易ではない．AI が情報システムや社会に統合されていく中で，AI の毀損や悪用によって生じる負の影響を体系的に理解するためには，研究のみならず，ビジネスや一般社会の観点からも全体像を把握することが求められる．特に，より顕著に複雑化している当該分野においては，AI セキュリティに関わる全ての要素間の関係性の体系的な整理が重要である．さらに，人や社会への負の影響がどのようにもたらされるかを明確化することも不可欠である．つまり，AI セキュリティに関連する知識，技術，社会的影響は，それらの関係性を理解するために体系的に整理されるべきである．

3.2 AI セキュリティマップ

AI セキュリティマップは，ISA と EIA の二つの側面から成る．AI が情報システム内で満たすべき要素は，ISA に分類される．一方，EIA は AI が攻撃を受けたり悪用された結果として人や社会に影響を及ぼす要素を含んでいる．ISA および EIA の各要素の定義は，それぞれ表 1 および表 2 に示している．各要素ごとに，対応する負の影響を特定している． ISA における負の影響は主に AI への攻撃に起因し，AI を利用する情報システムと強く関連する．一方，EIA に分類される負の影響は，AI が攻撃される場合だけでなく，正常に機能する AI が悪用された場合にも生じうる．本マップは，これら負の影響，原因となる攻撃や要因，および防御や対策を体系的に整理している．

また，ISA と EIA の相互関係を包括的に検討し，AI の毀損や悪用が人や社会にどのような負の影響をもたらすかを明確にする．本研究では，4 種類のセキュリティ対象を想定している．表 3 にセキュリティ対象の定義を示す．こ

表 1 ISA における要素.

AI が満たすべき要素	定義
機密性	AI のデータやモデルが権限のないものにアクセスされない.
完全性	AI モデルやアルゴリズムが改ざんされておらず，AI の出力が期待通りである.
可用性	必要なときに AI が必要な機能やサービスを提供できる.
説明可能性	AI が出力の根拠やプロセスを説明できる.
出力の公平性	AI が特定の個人や集団に対して偏った出力をしない.
安全性	AI が人命，身体，財産，精神への被害を防ぐための安全機構を備えている.
精度	AI が目的達成のために一定の精度を満たしている.
制御可能性	AI が管理者により制御され，暴走したり他の環境に影響を与えたりしない.
信頼性	AI の出力が信頼できるものである.

表 2 EIA における要素.

個人および社会に影響を与える要素	定義
サイバー攻撃	AI がサイバー攻撃に利用されない.
軍事利用	AI が軍事利用されない.
プライバシー	AI がプライバシーを侵害せず，プライバシーに関する法律や慣習を遵守する.
偽情報	AI を使って偽情報を意図的に作成しない、または、それらを識別できる.
誤情報	AI が誤情報を出力しない、または、それらを識別できる.
ユーザビリティ	AI が目的達成のために一定水準の使いやすさを満たしている.
消費者の公平性	AI のバイアスのある出力によって消費者に害が及ばない.
盗作	AI が盗作に利用されない.
著作権・オーサーシップ	AI が著作権および著作者に関する法律や慣習を遵守する.
透明性	システムが AI を利用していること，およびその限界やリスクに関する情報が明示されている.
レピュテーション	AI システム提供者が一定の基準で評価され，信頼されている.
法令の遵守	AI が合法的な目的で利用され，法律に準拠した出力や行動を行う.
人間中心の原則	AI が人間の利益のために適切に利用されている.
倫理性	AI が社会的規範に沿った振る舞いをする.
経済	AI の利活用が経済の成長にプラスに寄与する.
物理的影響	AI の利活用が人々に物理的被害を与えない.
精神的影響	AI の利活用が人々に精神的な被害を与えない.
金銭的影響	AI の利活用が人々に金銭的な被害を与えない.
医療	AI の利用が高度で安全な医療の発展に寄与する.
基幹インフラ	AI の利用が基幹インフラの安全な運用に寄与する.

表 3 AI セキュリティマップにおけるセキュリティ対象.

セキュリティ対象	定義
消費者	AI または AI システムを利用する個人または組織.
非消費者	消費者に分類されない個人または組織.
社会	複数の人や組織から構成される集団.
AI システム提供者	AI システムを提供する個人または組織.

表 4　ISA における CIA 関連の負の影響に関する AI セキュリティマップ.

要素	負の影響	攻撃や原因	防御策や対策	EIA における関連要素
機密性	学習データの漏洩	メンバーシップ推論攻撃	差分プライバシー（DP）, 暗号化技術, AI アクセス制御	プライバシー, 著作権および著作者, レピュテーション, 精神的影響, 法令の遵守
	個人情報の漏洩	メンバーシップ推論攻撃, プロンプトインジェクション	DP, 連合学習, 個人情報マスキング, AI アクセス制御	プライバシー, 著作権および著作者, レピュテーション, 精神的影響, 法令の遵守
	学習データの再構成	モデル反転攻撃	DP, 暗号化技術, AI アクセス制御	プライバシー, 著作権および著作者, レピュテーション, 精神的影響, 法令の遵守
	モデル情報の漏洩	モデル抽出攻撃	DP, モデル抽出攻撃の検知, AI アクセス制御	プライバシー, 著作権および著作者, レピュテーション, 精神的影響, 法令の遵守
	システムプロンプトの漏洩	プロンプトリーキング	プロンプトチェック	レピュテーション
完全性	特定入力に対する AI 出力の操作	敵対的サンプル	敵対的学習, 敵対的サンプルの検知 , モデルの頑健性保証（CR）	レピュテーション , 偽情報, ユーザビリティ , 消費者の公平性, 法令の遵守, 基幹インフラ, 物理的影響, 医療
	学習データ汚染による AI 性能の低下	ポイズニング攻撃	汚染データの検知, CR	レピュテーション , 誤情報, ユーザビリティ , 消費者の公平性, 法令の遵守, 基幹インフラ, 物理的影響, 医療
	特定条件下での AI 出力の操作	バックドア攻撃	トリガーの検知, バックドアモデルの検知, CR	プライバシー, 偽情報, ユーザビリティ , 消費者の公平性, レピュテーション, 人間中心の原理, 法令の遵守, 物理的影響, 倫理性, 経済 , 基幹インフラ, 医療
	有害な応答の生成	プロンプトインジェクション	有害性検知 , プロンプトチェック	プライバシー, 偽情報, 消費者の公平性, レピュテーション , 法令の遵守, 倫理性
可用性	AI による誤分類による機能またはサービス品質の低下	敵対的サンプル	敵対的学習, 敵対的サンプルの検知, CR	レピュテーション, 人間中心の原理, 倫理性, 基幹インフラ, 法令の遵守, 物理的影響, 経済, 医療
	予測精度の継続的な低下による機能またはサービス品質の低下・停止	ポイズニング攻撃	汚染データの検知, CR	レピュテーション, ユーザビリティ , 物理的影響, 精神的影響, 金銭的影響, 経済, 基幹インフラ, 医療
	特定条件下での AI 出力操作による機能またはサービス品質の低下	バックドア攻撃	トリガーの検知, バックドアモデルの検知, CR	レピュテーション, ユーザビリティ , 物理的影響, 精神的影響, 金銭的影響, 経済, 基幹インフラ, 医療
	AI リソースの過剰消費によるサービス停止	モデル DoS	トークンの制限, AI アクセス制御	レピュテーション, 物理的影響, 精神的影響, 金銭的影響, 経済, 基幹インフラ, 医療

れらセキュリティ対象は，AI の毀損や悪用によって影響を受けうる主要なステークホルダーを意味する．

情報システム的側面. この側面では，AI が情報システム内で満たすべき要素が分類されている．CIA の要素は，AI への攻撃によって毀損されるものと想定しており，その他の要素は，主に CIA の毀損によって影響を受けるものとする．これらの要素に対する負の影響を検討することで，その要因や対策をより的確に理解できる．また，ISA に分類される要素への負の影響が，EIA の要素にも影響を及ぼす可能性があると想定している．

外部作用的側面. この側面では，プライバシー侵害や著作権等の権利侵害といった，人や社会に影響を及ぼす要素を分類している．この側面の負の影響は，情報システム内の AI への攻撃のみならず，AI の悪用によっても生じうる．EIA に分類される要素への負の影響を検討することで，ISA のどの要素が毀損され，それがどのように人や社会に波及するかについて理解できる．

4. 考察・議論

本節では，AI セキュリティマップを参照しながら本研究で得られた新たな知見について議論する．

4.1 情報システム的側面に関する知見

表 4 は，ISA における CIA に関連する負の影響についての AI セキュリティマップを示したものである．CIA 要素の毀損が，情報システム上で 13 種類の負の影響を引き起こすことを特定した．これらの負の影響は，AI を基盤とする情報システムの機能や運用を直接的に阻害しうる．さらに，機密性に関連する負の影響の数が，完全性や可用性に比べて多いことが明らかとなった．これは，プライバシーに関する研究が活発に行われている現状とも一致する．また，AI への攻撃によってまず CIA の要素が侵害され，その結果として他の要素にも多くの負の影響が及ぶ可能性がある．表 5 は，CIA 以外の要素に関連する負の影響についてのマップを示している．CIA 以外の要素として，説明可能性，出力の公平性，安全性，精度，制御可能性，信頼性の 6 要素を整理した．これらの一部要素は特定の攻撃によって直接的に損なわれる場合もあるが，多くは CIA 要素の侵害の結果として生じる．また，各要素ごとに 1 つの負の影響を特定したが，完全性の毀損はこれら全ての要素に負の影響を及ぼすことが明らかとなった．情報セキュリティの観点から，完全性が情報システムのさまざまな要素と関連しているのは自然なことである．信頼性については，説明可能性の毀損も負の影響をもたらす原因となる．

表 5 ISA における CIA 以外の要素に関連する負の影響に関する AI セキュリティマップ.

要素	負の影響	毀損される要素	攻撃や原因	防御策や対策	EIA における関連要素
説明可能性	AI 推論結果の理解困難	完全性	Explainability への攻撃	XAI, Robust explainability	レピュテーション, 透明性, 精神的影響, 金銭的影響, 経済, 医療
出力の公平性	AI 出力のバイアス	完全性	学習データのバイアス	完全性の防御手法, AI 出力のバイアス検知, 学習データのバイアス除去, 公平な AI モデルの作成	ユーザビリティ, 消費者の公平性, レピュテーション, 医療, 法令の遵守, 精神的影響, 倫理性
安全性	AI の予測精度低下や想定外の動作による人への被害	完全性		完全性の防御手法, フェイルセーフ機構	レピュテーション, 人間中心の原理, 物理的影響, 精神的影響, 基幹インフラ, 法令の遵守, 金銭的影響, 倫理性, 経済, 医療
精度	AI 予測精度の低下	完全性		完全性の防御手法	誤情報, ユーザビリティ, 消費者の公平性, レピュテーション, 人間中心の原理, 物理的影響, 精神的影響, 金銭的影響, 経済, 基幹インフラ, 医療
制御可能性	管理者による意図しない出力や動作	完全性	敵対的サンプル, プロンプトインジェクション, 間接プロンプトインジェクション, バックドア攻撃, サイバー攻撃	完全性の防御手法	偽情報, ユーザビリティ, 消費者の公平性, レピュテーション, 人間中心の原理, 基幹インフラ, 法令の遵守, 物理的影響, 精神的影響, 金銭的影響, 経済, 医療
信頼性	AI 出力の信頼性判断困難	完全性, 説明可能性	ハルシネーション	不確実性の定量化, RAG, XAI, ハルシネーション検知	ユーザビリティ, レピュテーション, 精神的影響, 透明性, 経済, 基幹インフラ, 医療

表 6 EIA における消費者への負の影響に関する AI セキュリティマップ.

要素	負の影響	毀損される ISA の要素	EIA における関連要素や要因	防御策や対策
プライバシー	消費者が生成 AI 等に誤って個人情報を入力する	透明性	ソーシャルエンジニアリング攻撃	匿名化技術, 差分プライバシー, 連合学習, マシン・アンラーニング, 暗号化技術
誤情報	AI による誤情報の出力	完全性, 精度, 制御可能性, 説明可能性, 信頼性	RAG に対するポイズニング攻撃, ハルシネーション	完全性の防御手法, データキュレーション, RAG, XAI, ハルシネーション検知
ユーザビリティ	AI のユーザビリティの低下	完全性, 可用性, 精度, 制御可能性, 出力の公平性		完全性・可用性の防御手法, RAG
消費者の公平性	AI 出力のバイアスによる職業・生活機会の喪失	完全性, 制御可能性, 出力の公平性		完全性の防御手法, ヒューマンインザループ, 出力の公平性の対策, AI 出力のバイアス検知
	不公平なバイアス・差別的出力	完全性, 制御可能性, 出力の公平性		完全性の防御手法, AI アライメント, 出力の公平性の対策, AI 出力のバイアス検知
透明性	意図せず AI を利用する	説明可能性		AI 生成出力の識別, 生成 AI のウォーターマーキング
	リスクを認識せずに AI を利用する	説明可能性, 信頼性		AI による出力への注記, 教育・フォローアップ
人間中心の原理	AI による消費者の意思決定の不適切な操作	完全性, 説明可能性, 制御可能性, 出力の公平性		完全性の防御手法, ヒューマンインザループ
倫理性	AI による非倫理的な出力や行動	完全性	ジェイルブレイク	教育・フォローアップ, AI アライメント
物理的影響	AI による消費者への身体的被害	完全性, 精度, 制御可能性, 安全性	人間中心の原理	完全性の防御手法
精神的影響	AI による消費者への精神的被害	完全性, 可用性, 制御可能性, 安全性, 出力の公平性	消費者の公平性	完全性の防御手法, 可用性の防御手法
金銭的影響	AI による消費者への経済的被害	完全性, 可用性, 精度, 制御可能性, 安全性, 出力の公平性	人間中心の原理	完全性の防御手法, 可用性の防御手法

4.2 外部作用的側面に関する知見

EIA においては，合計 20 の要素と 36 の負の影響を特定した．これは，AI セキュリティが多様な要素と密接に関連し，人や社会に対して重大な影響を及ぼし得ることを示している．EIA における負の影響の多くは，ISA における負の影響に起因するだけでなく，EIA 内部の他の影響とも関連しうる．例えば，「サイバー攻撃」や「偽情報」による負の影響が，非消費者に対する経済的な悪影響の原因となる場合がある．これらの要素に共通する特徴は，AI の悪用によって生じる負の影響が ISA の要素が満たされている場合でも起こりうるという点である．これらの負の影響は，AI が悪用された時点で必ずしも人や社会に影響を及ぼすものではなく，AI の悪用を通じて特定の目的が達成された場合に個人や社会に負の影響がもたらされる．

4.3 各セキュリティ対象への負の影響

消費者. 表 6 は，EIA における消費者への負の影響に関するマップを示している．消費者に対しては，12 個の負の影

表 7 EIA における非消費者への負の影響に関する AI セキュリティマップ.

要素	負の影響	ISA における関連要素		EIA における	防御策や対策
		毀損	悪用	関連要素や要因	
サイバー攻撃	AI のサイバー攻撃への利用	機密性, 制御可能性	可用性, 精度, 説明可能性		AI アライメント, モデル情報を秘匿しつつ説明可能性を確保する手法
軍事利用	AI の軍事目的への利用	制御可能性	可用性, 精度, 説明可能性		AI アライメント
プライバシー	AI からの個人情報漏洩によるプライバシー侵害	機密性, 完全性			差分プライバシー, 連合学習, AI アライメント, マシン・アンラーニング, 暗号化技術, 匿名化技術
	断片的な情報や AI を用いた推論による個人情報の特定			ソーシャルメディアから収集した情報を AI で分析して個人を特定する攻撃	匿名化技術, 差分プライバシー, 連合学習, マシン・アンラーニング, 暗号化技術
	表情などから性格や嗜好を AI で推定される			画像を AI で分析して個人情報を推定する攻撃	匿名化技術, 差分プライバシー, 連合学習, マシン・アンラーニング, 暗号化技術
	AI 生成のメールや音声を用いた機密情報入力の誘導・窃取		可用性, 精度	ソーシャルエンジニアリング攻撃	匿名化技術, 差分プライバシー, 連合学習, マシン・アンラーニング, 暗号化技術
偽情報	AI による偽情報の生成	制御可能性, 完全性	可用性, 精度	ディープフェイク, ソーシャルエンジニアリング攻撃	AI アライメント, 生成 AI のウォーターマーキング, 暗号化技術, AI 生成出力の識別, 偽情報検知, ディープフェイク検知
著作権・オーサーシップ	AI 生成類似コンテンツによる著作権・オーサーシップの侵害	完全性, 制御可能性		盗作	完全性および盗作の防御手法
人間中心の原理	AI による非消費者の意思決定の不適切な操作			偽情報	完全性および偽情報の防御手法, ヒューマンインザループ
倫理性	AI による非倫理的な出力や行動	完全性		ジェイルブレイク	教育・フォローアップ, AI アライメント
物理的影響	AI による非消費者への身体的被害	完全性, 精度, 制御可能性, 安全性		軍事利用	完全性の防御手法
精神的影響	AI による非消費者への精神的被害	機密性, 制御可能性, 安全性, 出力の公平性		偽情報, 軍事利用	完全性の防御手法, 可用性の防御手法, 偽情報の防御手法
金銭的影響	AI による非消費者への経済的被害	安全性		サイバー攻撃, 偽情報	完全性の防御手法, 偽情報の防御手法

響および 10 個の関連する要素を特定した．消費者に対する負の影響の多くは，ISA の要素の侵害が直接的な要因となることがわかった．AI を利用するユーザーである消費者への負の影響が，情報システムや AI 自体に対する負の影響と関連しているのは直感的である．一方で，特定の攻撃によってもたらされる影響も存在する．

非消費者. 表 7 は，EIA における非消費者への負の影響に関するマップを示している．非消費者に対しては，13 の負の影響および 10 の関連する要素を特定した．消費者の場合と同様に，これらの負の影響の多くは ISA の要素の毀損に起因する．しかし，ISA の要素が正常に機能している場合でも，要素が悪用されれば負の影響が生じうる点に注意が必要である．例えば，AI の精度や可用性の悪用は，誤情報の拡散やサイバー攻撃を助長しうる．さらに，多くの負の影響は AI を利用した攻撃や他の EIA 要素の毀損によっても生じる．特に，プライバシーに関連する 4 つの負の影響を特定しており，AI を利用しない人であってもプライバシー侵害のリスクが高いことを示している．これは，AI が外部にも多大な影響を及ぼし得ることを示している．

社会. 表 8 は，EIA における社会および AI システム提供者への負の影響に関するマップを示している．社会に対しては，9 つの負の影響および 8 つの関連する要素を特定した．社会への影響も ISA の要素の毀損と密接に関連していることが明らかとなった．特に，医療や基幹インフラに対する影響は，関連する ISA 要素が多いことから発生しやすい．今後の AI の社会的普及を考慮すると，これらの要素への潜在的影響を理解し，適切な対策を講じることが重要である．

AI システム提供者. AI システム提供者に対しては，2 つの負の影響および 2 つの関連する要素を特定した．AI システム提供者への影響は，ISA のいずれかの要素が侵害された場合に生じる可能性がある．さらに，AI の悪用によって引き起こされる負の影響は，AI システム提供者に対してレピュテーションの低下や経済的な損失といった結果をもたらしうる．AI システムを開発する企業や個人が増加する中，これらのリスクを適切に認識することが重要である．

4.4 2 つの側面の関係性

本マップでは，基本的に，ISA のいずれかの要素が侵害されると，EIA の要素に影響が及ぶと想定している．つま

表 8　EIA における社会と AI システム提供者への負の影響に関する AI セキュリティマップ

要素	負の影響	ISA における関連要素		EIA における関連要素や要因	防御策や対策
		毀損	悪用		
社会					
サイバー攻撃	AI のサイバー攻撃への利用	機密性, 制御可能性	可用性, 精度, 説明可能性		AI アライメント, モデル情報を秘匿しつつ説明可能性を確保する手法
軍事利用	AI の軍事目的への利用	制御可能性	可用性, 精度, 説明可能性		AI アライメント
偽情報	AI による偽情報の生成	制御可能性	可用性, 精度	ディープフェイク，ソーシャルエンジニアリング攻撃	AI アライメント, 生成 AI のウォーターマーキング, 暗号化技術, AI 生成出力の識別, 偽情報検知, ディープフェイク検知
法令の遵守	法律違反目的での AI 利用	機密性, 制御可能性	可用性, 精度		AI アライメント, AI アクセス制御
	AI による法律違反行為	完全性, 精度, 制御可能性			AI アライメント, AI アクセス制御
倫理性	AI による非倫理的な出力や行動	完全性		ジェイルブレイク	教育・フォローアップ, AI アライメント
経済	AI の経済への悪影響	安全性, 精度		サイバー攻撃，軍事利用, 偽情報, 人間中心の原理	完全性の防御手法, 偽情報の防御手法
医療	AI による医療への悪影響	完全性, 可用性, 精度, 制御可能性, 安全性, 出力の公平性, 説明可能性, 信頼性		人間中心の原理	完全性の防御手法, 可用性の防御手法
基幹インフラ	AI による基幹インフラへの悪影響	完全性, 可用性, 精度, 制御可能性, 安全性, 出力の公平性, 説明可能性, 信頼性		人間中心の原理	完全性の防御手法, 可用性の防御手法
AI システム提供者					
レピュテーション	AI システム提供者の評判低下	ISA 内すべての要素		サイバー攻撃，軍事利用, 法令の遵守	機密性の防御手法, 完全性の防御手法, 可用性の防御手法
金銭的影響	AI システム提供者への経済的被害	ISA 内すべての要素		サイバー攻撃, 法令の遵守	完全性の防御手法, 可用性の防御手法

り，多くの場合，情報システムの毀損による連鎖的反応として，人や社会に負の影響が生じうると考えられる．2 つの側面に分類された要素を参照することで，負の影響がどのように個人や社会に波及するのかを明らかにできる．本研究では，人や社会への負の影響の波及には直接的なものと間接的なものの 2 種類があると考えている．以下では，それぞれについて述べる．

直接的な波及. ISA 内の要素の毀損による負の影響は，人や社会に直接的な影響を及ぼしやすい．典型的な例として，「機密性」が侵害されると「プライバシー」も侵害されることは容易に想像できる．AI セキュリティの文脈においては，AI が膨大なデータに関する知識を保持することから，非消費者への影響も大きい．さらに，表 4 に示したように，完全性の毀損がもたらす負の影響は，人間中心の原理，医療，基幹インフラなど多くの要素と関連している．上記で述べたように，特に完全性の毀損は，社会全体に重大な影響を及ぼしうる．また，可用性の毀損による負の影響も，EIA における多くの要素に直接関係している．

間接的な波及. 複数の負の影響が連鎖的に発生し，最終的に人や社会に影響が及ぶ場合もある．例えば，表 5 に示されるように，プロンプトインジェクション攻撃によって LLM の「完全性」が毀損された結果，まず「制御可能性」が損なわれる．制御可能性が損なわれることで，攻撃者は LLM に意図した「誤情報」を生成させることが可能となる．この誤情報が拡散すると，人間の意思決定に影響を及ぼし，表 7 に示した「人間中心の原理」に関する負の影響が生じ，この誤情報を取得した非消費者へ影響が波及することになる．さらに，ISA の要素が満たされていることを悪用して AI システムが悪用されることで，人や社会に影響が波及する場合もある．例えば，LLM を用いたマルウェア生成によるサイバー攻撃は，「精度」や「可用性」を悪用し，非消費者への「金銭的影響」や社会への「経済」に負の影響をもたらす．このような波及効果は，AI を全く利用しない非消費者にも影響を及ぼしやすい．

5. 今後の方向性

AI セキュリティ情報の体系化に関する研究開発の今後の方向性を基礎研究と応用研究の 2 つの観点から示す．

5.1 基礎研究

本研究で提案したような体系化の枠組みは今後さらに積極的に研究されるべきである．以下に AI セキュリティの体系化に関する基礎研究の今後の方向性を示す．

セキュリティ対象の適切な粒度の定義. 現段階では，マップ内のセキュリティ対象は比較的広いカテゴリーで定義されているが，今後はより詳細な定義が求められる．

影響の定量的評価. 本マップは AI セキュリティの全体像を俯瞰するのに有用であるが，要素間の関係性を示すにとどまっているため，定量的なリスク評価が必要である．

新分野のマッピング. AI エージェントなどの新たな分野のセキュリティも対象とすべきである．最新動向を効率的に把握し，迅速に反映する仕組みや方法の確立が重要である．

負の影響および要素の定義・分類の自動化. AI 技術の急速な進化を踏まえると，膨大な情報を漏れなく網羅的にカバーするには，LLM などの AI を活用した体系化技術の研究開発が必要である．

5.2 応用研究

応用研究の観点からも体系化された情報をどのように活用するかを検討することは重要である．例えば，AI セキュリティに関するニュース記事を本研究のマップとともに LLM に入力することで，関連する負の影響を抽出するという活用例が考えられる．また，本マップのもう一つの応用例として，AI システム提供者や開発者が情報システム内の AI に対する負の影響を把握するために活用することが挙げられる．なお，これらは一例に過ぎないため，今後さらなる有用な応用例や実践的な手法の開発が必要である．

6. おわりに

本稿では，人や社会への負の影響に配慮した AI セキュリティ情報の体系化として AI セキュリティマップを作成した．本マップは，情報システムにおいて AI が満たすべき要素と，人や社会に影響を及ぼす要素を特定している．また，AI 要素と社会的な影響の関係性を体系的に整理し，AI システムの毀損や悪用が，人や社会にどのような影響を及ぼすかを明確化している．さらに，各要素に関連する負の影響をもたらす攻撃や要因，および防御を分類した．本マップが複雑な AI セキュリティ分野における情報の収集と理解を促進することで，研究者や幅広いステークホルダーにとって，新たな要素や負の影響の発見および，研究を推進するための重要な基盤となることを望んでいる．

謝辞 本研究は，JST 経済安全保障重要技術育成プログラム【JPMJKP24C4】の支援を受けたものです．

参考文献

[1] Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., Li, W. and Li, K.: Artificial intelligence security: Threats and countermeasures, *ACM Computing Surveys (CSUR)*, Vol. 55, No. 1, pp. 1–36 (2021).

[2] Shen, J., Wang, N., Wan, Z., Luo, Y., Sato, T., Hu, Z., Zhang, X., Guo, S., Zhong, Z., Li, K. et al.: Sok: On the semantic ai security in autonomous driving, *arXiv preprint arXiv:2203.05314* (2022).

[3] Kiribuchi, N., Zenitani, K. and Semitsu, T.: Securing AI Systems: A Guide to Known Attacks and Impacts (2025).

[4] Lu, L., Yan, H., Yuan, Z., Shi, J., Wei, W., Chen, P.-Y. and Zhou, P.: Autojailbreak: Exploring jailbreak attacks and defenses through a dependency lens, *arXiv preprint arXiv:2406.03805* (2024).

[5] Chang, Y., Liu, H., Jaff, E., Lu, C. and Zhang, N.: SoK: Security and Privacy Risks of Medical AI, *arXiv preprint arXiv:2409.07415* (2024).

[6] Dibbo, S. V.: Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap, *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, IEEE, pp. 439–456 (2023).

[7] Wingarz, T., Lauscher, A., Edinger, J., Kaaser, D., Schulte, S. and Fischer, M.: SoK: Towards Security and Safety of Edge AI, *arXiv preprint arXiv:2410.05349* (2024).

[8] Ramirez, M. A., Kim, S.-K., Hamadi, H. A., Damiani, E., Byon, Y.-J., Kim, T.-Y., Cho, C.-S. and Yeun, C. Y.: Poisoning attacks and defenses on artificial intelligence: A survey, *arXiv preprint arXiv:2202.10276* (2022).

[9] Zhang, S., Pan, Y., Liu, Q., Yan, Z., Choo, K.-K. R. and Wang, G.: Backdoor attacks and defenses targeting multi-domain ai models: A comprehensive review, *ACM Computing Surveys*, Vol. 57, No. 4, pp. 1–35 (2024).

[10] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. et al.: Explainable AI (XAI): Core ideas, techniques, and solutions, *ACM Computing Surveys*, Vol. 55, No. 9, pp. 1–33 (2023).

[11] Zhang, W.: AI fairness in practice: Paradigm, challenges, and prospects, *Ai Magazine*, Vol. 45, No. 3, pp. 386–395 (2024).

[12] Parraga, O., More, M. D., Oliveira, C. M., Gavenski, N. S., Kupssinskü, L. S., Medronha, A., Moura, L. V., Simões, G. S. and Barros, R. C.: Fairness in Deep Learning: A survey on vision and language research, *ACM Computing Surveys*, Vol. 57, No. 6, pp. 1–40 (2025).

[13] Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V. and Sikdar, B.: Privacy and security concerns in generative AI: a comprehensive survey, *IEEE Access* (2024).

[14] Aïmeur, E., Amri, S. and Brassard, G.: Fake news, disinformation and misinformation in social media: a review, *Social Network Analysis and Mining*, Vol. 13, No. 1, p. 30 (2023).

[15] Castagnaro, A., Conti, M. and Pajola, L.: Offensive AI: Enhancing Directory Brute-forcing Attack with the Use of Language Models, *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*, pp. 184–195 (2024).

[16] Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. and Thompson, N.: The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, *arXiv preprint arXiv:2408.12622* (2024).

[17] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A. et al.: Taxonomy of risks posed by language models, *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229 (2022).

[18] Pankajakshan, R., Biswal, S., Govindarajulu, Y. and Gressel, G.: Mapping llm security landscapes: A comprehensive stakeholder risk assessment proposal, *arXiv preprint arXiv:2403.13309* (2024).