機械学習システム
セキュリティガイドライン
Part S. 「補足：近年の動向（生成 AI・法規則・他）」

Version 3.00
2025 年 7 月 4 日

機械学習システムセキュリティガイドライン策定委員会
機械学習セキュリティワーキンググループ

日本ソフトウェア科学会 機械学習工学研究会

第 2 版からの変更点：
第 3 版は、これまでの第 2 版を修正、改訂するものではなく、第 2 版はそのまま残し、近年の動向を中心に補足を加えたものとなっている。補足として記述したのは以下の部分である。

・生成 AI のリスクについての記述
・法規則との関係性の整理
・本ガイドラインにおける旧バージョンのリスク分析についての位置づけ整理
・新影響分析ツールの説明記述

目次

S-1. はじめに

近年の AI 技術に関する進展が目覚ましい。2023 年に機械学習システムセキュリティガイドラインの Version 2.0[S-1]（以降、「ガイド Ver2」と呼ぶ）を公開して以降も生成 AI の世間一般への浸透や欧州の AIAct[S-2]の発効、AI 事業者ガイドライン[S-3]の公開など、様々な進展がみられてきた。ガイド Ver2 で公開した機械学習システムのセキュリティに関する整理やリスク分析技術、攻撃手法の概要の情報は、現在でも利用可能であると考えられる反面、最新の業界動向には対応しきれていない面があると筆者らは考えている。そこで今回、近年の業界動向に特化し、生成 AI に関するリスクの整理、AIAct などの法規則と本ガイドラインの関係性の整理、これまでに公開したリスク分析技術と生成 AI の関係の整理、ガイド Ver2 で公開した影響分析ツールの AI 品質、AI 倫理への拡張版の公開などを行うこととした。

S-1.1.ガイドラインの構成

ガイド Ver2 を含む最新のガイドライン、及び、分析ツールの構成は以下の通りである。

○ガイドライン
・ガイド Ver2
- 機械学習システムセキュリティガイドライン Part I.「本編」Version 2.00 (MLSystemSecurityGuideline-main-ver2.00.pdf)
- 機械学習システムセキュリティガイドライン Part II.「リスク分析編」 Version 2.01 (MLSystemSecurityGuideline-RiskAssessment-ver2.01.pdf)
- 機械学習システムセキュリティガイドライン「付録：攻撃検知技術の概要」 Version 2.00 (MLSystemSecurityGuideline-Appendix-ver2.00.pdf)

・ガイド Ver3
- 機械学習システムセキュリティガイドライン Part. S「補足：近年の動向（生成 AI・法規則・他)」 Version 3.00（本ドキュメント） (MLSystemSecurityGuideline-SupplementaryDocument-ver3.00.pdf)

○分析ツール
・脅威分析ツール（AI リスク問診）(AI-RiskAssessmentTool_ver2.00.xlsx)
・影響分析ツール（AI セキュリティのみ）(AI-ImpactAssessmentTool_ver1.00.xlsx)
・影響分析ツール(AI 品質。・AI 倫理含む(新規公開)) (AI-ImpactAssessmentTool_2.00.xlsx)

S-2. 生成 AI のリスクについて

S-2.1. 概要

近年の機械学習の発展において、生成 AI（ジェネレーティブ AI）の存在感は無視できないものとなっている。高精度な基盤モデルに基づく生成 AI は従来の AI の範疇を超えた優れた機能が一般のユーザにも簡単に使えるため、社会の向上に資するものと期待されている。一方で、その優れた機能による攻撃の高度化や偽情報（ディープフェイク等）生成、情報漏洩の危険性のように、セキュリティリスクの増大が懸念されると共に、誤動作（ハルシネーション等）や社会的・倫理的リスクの可能性も指摘されている。

本ガイドライン第 1 版が発行された 2022 年 6 月には既に生成 AI は大きな話題となっていたが、世の中でブームと言われ、同時にリスクの懸念が話題となり始めたのは Open AI が ChatGPT を発表した 2022 年 11 月以降である。執筆のタイミングの都合上、本ガイドラインの中で記載されているリスク分析手法などは生成 AI に適用できる可能性はあるものの、2025 年 6 月時点で生成 AI に対する効果の検証を完了していない。このため、**本ガイドラインで紹介しているリスク分析手法や分析ツールについては、生成 AI は分析対象外とする**こととする。一方で、活用が広がっている生成 AI のリスクは本ガイドラインでも取り扱うべきと考えられるため、第 3 版での改訂にあたり、生成 AI を活用する上で懸念されるリスクについて可能な範囲で整理することとした。

S-2.2. 生成 AI 活用に伴うリスクの整理

利用者が生成 AI を安全に活用するためには、生成 AI 活用に伴うリスクを把握して適切な対策を講じる必要がある。それに対し、多くの先行研究が生成 AI のリスクと影響を指摘して警告を発しており、また多くの報告書や一般書籍でも、生成 AI の可能性と共に懸念事項が指摘されている。

生成 AI のリスクについて整理した比較的早い例として、DeepMind 社が 2021 年に発表した論文[S-4]（改訂版は 2022 年[S-5]）がある。この研究ではコンピュータサイエンス、言語学、社会科学などの学際的な専門家による議論を経て、大規模言語モデル（LLM）の倫理的・社会的リスクの影響を 21 個抽出し、それを 6 領域に分類した（表 S- 1)。この論文は、本研究が LLM の運用に関するリスクにのみ焦点を当てたもので、例えば訓練に関する弊害や計算コスト増大による社会的な影響、マルチモーダルな応用に関する考察などは含んでいないという、研究の限界が明記されている。あくまで生成 AI リスクを検討するための契機としての早期レポートという立場を取っているが、示唆に富む内容だと考える。

表 S- 1. 生成 AI リスクの分類例（DeepMind 論文[S-4][S-5]より）

	分類	影響
1	Discrimination, Exclusion and Toxicity（差別、排除、悪意）	・社会的ステレオタイプと不当な差別 ・排除的規範 ・有害な言葉 ・社会集団によるパフォーマンスの低下
2	Information Hazards（情報の危険性）	・個人情報漏洩によるプライバシー侵害 ・個人情報推測によるプライバシー侵害 ・機密情報漏洩や正確な推論による情報漏洩
3	Misinformation Harms（誤情報の弊害）	・虚偽または誤解を招く情報の流布 ・医療・法律等で誤情報により重大な損害 ・非倫理的または違法な行為を行うようユーザを誘導または助言
4	Malicious Uses（悪意のある使用）	・偽情報キャンペーンのコストを削減 ・詐欺・なりすまし詐欺の助長 ・サイバー攻撃、武器、悪意のある使用のためのコード生成を支援 ・違法な監視と検閲
5	Human-Computer Interaction Harms（人間とコンピュータの相互作用による弊害）	・システムを擬人化することは、過度の信頼や安全でない使用につながる可能性がある ・ユーザの信頼を悪用して個人情報を入手 ・性別や民族的アイデンティティを暗示することで、有害なステレオタイプを助長する
6	Automation, access, and environmental harms（自動化、アクセス、環境への悪影響）	・LLM の運用による環境への悪影響 ・不平等の拡大と雇用の質への悪影響 ・創造的経済を弱体化させる ・ハードウェア、ソフトウェア、スキルの制約による利益へのアクセスの格差

生成 AI リスクについてのより包括的な検討としては、JST（国立研究開発法人科学技術振興機構）の CRDS（研究開発戦略センター）が発表した「人工知能研究の新潮流 2」レポート[S-6]が分かりやすい。ここには文献[S-7][S-8]を底本として生成 AI のリスクを整理した「生成 AI に対する懸念事項」が 17 項目記されている。更に文献[S-9][S-10]では、文献[S-6]の 17 項目に更に 3 項目を加えた 20 項目を生成 AI リスク項目として提示している(表 S- 2)。

表 S- 2.　生成 AI リスク項目［S-10］

	リスク項目
1	ハルシネーション（幻覚）
2	弊害のあるコンテンツの生成
3	データの偏りによる社会的バイアス強化
4	偽情報やプロパガンダ生成
5	兵器の拡散に使われる可能性
6	プライバシー侵害，情報漏洩
7	サイバーセキュリティーへの脅威
8	危険な挙動の創発
9	他のシステムとの併用による悪用や弊害
10	経済的影響：労働者の置き換えなど
11	技術開発競争が加速することのリスク
12	AI への過度の依存：ユーザが AI を過剰に信頼する
13	学習データの不透明性：トレーサビリティ欠如
14	モデル作成時の労働搾取
15	著作権侵害の恐れ
16	データ汚染：ネット上の AI 生成物の拡大
17	自然環境へのインパクト：訓練・推論時の電力消費量
18	犯罪の巧妙化・容易化につながるリスク
19	AI 生成物から創作物としての権利が得られない懸念
20	違法な監視と検閲

文献[S-9][S-10]によれば、表 S- 2 に挙げたリスク項目はリスク発生の異なった段階の記述が混在している。製品やシステムの安全性に関する国際的なガイドラインである ISO/IEC Guide51[S-11]によれば、リスク発生の過程は危険な状態を発生させる「ハザード」と、それによって生ずる「危害」とで表すことができる（図 S- 1）が、例えば表 S- 2 のリスク項目 1 の「ハルシネーション」は生成 AI が事実ではない情報を生成するという問題（「ハザード」に相当）を示すが、それによってどのような危害が発生するのかは述べていない。一方でリスク項目 15 は生成 AI によって著作権侵害が発生するという問題（「危害」に相当）への危惧が記されているが、生成 AI のどのような挙動によって危害が発生するのかは述べられていない。そこで文献[S-10]では、表 S- 2 のリスク項目を「ハザード」と「影響」に分解して、生成 AI によって発生する問題事象とその影響を整理している。

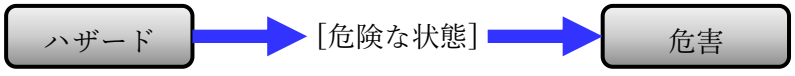

図 S- 1. ISO/IEC Guide 51［S-11］のリスクモデル

表 S- 3. 生成 AI リスク項目の分解［S-10］

リスク項目 (表 S- 2)	ハザード	リスク分類 (表 S- 1)	影響	リスク分類 (表 S- 1)
1	・幻覚 ・バイアスのある出力	3	・虚偽または誤解を招く情報の流布	3
2	・有害コンテンツ生成 ・バイアスのある出力	3	・社会的ステレオタイプと不当な差別 ・ヘイトスピーチと攻撃的な言葉	1
			・虚偽または誤解を招く情報の流布	3
3	・バイアスのある出力	3	・社会的ステレオタイプと不当な差別	1
			・社会的バイアスの強化 ・誤情報や偽情報の固定化	6
4	・意図的な偽情報生成	4	・社会的ステレオタイプと不当な差別	1
			・排除的規範 ・虚偽または誤解を招く情報の流布	3
	・偽情報やプロパガンダ生成	3		
5			・兵器の拡散に使用（不正使用）	4
6	・情報漏洩　・情報推測	2	・プライバシー侵害　・セキュリティ侵害	2
7	・権利侵害データの生成	2	・プライバシー侵害　・セキュリティ侵害	2
	・攻撃用コード生成の支援	4	・詐欺・標的型操作の助長	4
8	・予測不能な動作	3	・医療・法律等で誤情報により重大な損害	3
9	・他システムとの併用	4	・悪意の利用者へのハードル低下	4
10	・AI による自動化の進展	6	・経済的影響（労働者の置き換え等）	6
11	・技術開発競争の加速	6	・安全基準低下や悪い規範の拡散	6
12	・人が生成 AI を過剰に信じる条件（尤もらしさ）	5	・不適切な使用の助長（危険への自覚の低下）	5
13	・学習データ規模の増大	6	・トレーサビリティ欠如	2
			・学習データの出自の情報の欠落（透明性欠如）	6
14	・AI による自動化の進展	6	・経済的影響（労働者の置き換え等）	6
15	・著作権侵害データの生成	2	・著作権侵害	2
16	・データ汚染：ネット上の AI 生成物の拡大	6	・誤情報や偽情報の固定化	6
17	・訓練・推論時の電力消費量の増大	6	・自然環境へのインパクト	6
18	・偽情報やプロパガンダ生成	4	・悪意の利用者へのハードル低下	4
	・人が生成 AI を過剰に信じる条件（尤もらしさ）	5	・不適切な使用の助長（危険への自覚の低下）	5
19	・AI 生成物の創作性欠如（創作性の否定）	2	・権利化の失敗	2
20			・違法な監視と検閲（不正使用）	4

S-2.3.　生成 AI リスクにおける AI セキュリティ

前節で紹介した生成 AI リスク項目は様々な種類のリスクを対象としており、その中には AI セキュリティに関連するリスクも含まれる。S-2.2 節で整理した生成 AI のリスク項目やハザード、影響と、本ガイドラインで記載している AI セキュリティの脅威との対応が分かれば、生成 AI におけるセキュリティリスクを対処・分析する上で参考になると思われる。そこで本節では、本ガイドラインにおける AI セキュリティの各脅威と、前節で述べた生成 AI リスクとの対応を検討する。

S-2.3.1. 機械学習システム特有の脅威

本ガイドラインでは、機械学習システム特有の脅威として『モデルやシステムの誤動作』『モデルの窃取』『訓練データの窃取』を挙げている。脅威とは AI システムにとって問題がある状態を示しており、前節のハザードに対応する。

表 S-3 のハザードの中には AI セキュリティに関連するものが存在する。リスク項目 1〜4、18 に記されている「幻覚」「バイアスのある出力」「意図的な偽情報生成」「偽情報やプロパガンダ生成」は、いずれも AI システムの出力が不適切であるハザードなので、『モデルやシステムの誤動作』に関連すると考えることができる。本編（ガイド Ver2 本編）I-2.3 節ではモデルやシステムの誤動作を引き起こす攻撃の例として「回避攻撃」と「ポイズニング攻撃」が挙げられている。生成 AI の一つの典型的な形態である対話型生成 AI では、ユーザがプロンプト入力によって AI の動作を制御することができるので、誤動作を起こすような悪意のあるプロンプトを調査して作成し、システムに入力することで回避攻撃を行い、幻覚やバイアスのある出力を意図的に引き起こす可能性が考えられる。また、プロンプトとして入力したデータを学習に用いるタイプの生成 AI では、ポイズニング攻撃が行われる可能性もある。このように、対話型生成 AI ではモデルやシステムの誤動作の脅威が増大する可能性があるので注意が必要である。

リスク項目 6 に記されている「情報漏洩」「情報推測」は、いずれも情報が洩れる危険性のハザードなので、『モデルの窃取』『訓練データの窃取』に関連する。本編 I-2.3 節では『モデルの窃取』を引き起こす攻撃の例として「モデル抽出攻撃」が挙げられている。既存の生成 AI の出力を分析して同等の性能を持つ生成 AI を作成するという事例は近年話題となっており、生成 AI によって改めて注目された脅威である。これは表 S-3 のハザードでは、「情報推測」に相当する。一方、『訓練データの窃取』を引き起こす攻撃の例としては「モデルインバージョン攻撃」と「メンバシップ推測攻撃」が挙げられている。前者は学習データを復元する攻撃なので「情報漏洩」に相当する。後者は直接的には学習データが漏洩する攻撃ではないが、本ガイドラインリスク分析編 II-3.2 節で説明されているように、攻撃者が保持するデータが学習データに含まれるか否かの判定がプライバシー侵害に繋がる可能性があるなど、「情報漏洩」に関連する攻撃と考えることができる。また、リスク項目 7 に記されているハザード「権利侵害データの生成」は、どのような権利侵害を想定しているのか記されていないので断言はできないが、『訓練データの窃取』に関連する可能性があるので注意が必要である。

それ以外には、本ガイドラインに記された機械学習システム特有の脅威に関連するハザードは見当たらないが、間接的に何らかの関連があるハザードがあることも可能性として念頭におく必要がある。

S-2.3.2. 機械学習システム特有の脅威による影響

機械学習システム特有の脅威による影響は、本編 I-2.2 節に脅威による被害例が挙げられており、また本編 I-4.3 節では機械学習システム特有の脅威による影響が考察されている。

『モデルやシステムの誤動作』による被害には、例えば本編 I-2.2 節にはマルウェア検知 AI の誤動作によるマルウェア感染の例が紹介されており、本編 I-4.3 節には自動運転車の事故、医療ミス、違法な画像アップロードを許す、などが挙げられている。一方、表 S- 3 の影響の欄には、虚偽または誤解を招く情報の流布、社会的ステレオタイプと不当な差別など、生成 AI で特に注意すべき影響が挙げられている。

『モデルの窃取』による被害は、例えば本編 I-2.2 節にはモデル抽出攻撃によって複製されたモデルが、元のモデルに開発費をかけた企業の収益に悪影響を及ぼす例が記されている。『訓練データの窃取』による被害は、本編 I-2.2 節によれば、例えば医療情報の漏洩による患者への精神的負担や、プライバシー侵害による訴訟リスクなどが挙げられている。これら 2 つの脅威は生成 AI リスクとしては「情報漏洩」「情報推測」というハザードに相当しており、対応する影響は「プライバシー侵害」「セキュリティ侵害」が挙げられている。

これらの被害を防ぐためには対策が必要である。本編では I-7 節において攻撃の検知と対処の考え方が述べられている。生成 AI も機械学習システムの一種なので、対処の考え方は同様に参考にできる。例えば「モデルやシステムの誤動作」を目的とする攻撃への対策として、本編 I-7.3.1 節では「前処理による緩和」が挙げられているが、これはユーザが悪意のあるプロンプトを用いた「回避攻撃」を実施した場合に、誤動作させるためにプロンプトへ施された改変を前処理で除去するという対策に対応する。このように、生成 AI に対する脅威に対する対策を考える上で、本編の表 I-6 に記された応急対策の例が参考になると考えられる。

S-3. 法規則との関連について

S-3.1. 各国の法規則や標準規格

欧州では AI に関する法令である AIAct[S-2]が制定され発効されている。日本でも AI 関連技術の研究開発・活用推進法[S-12]が可決成立した。米国においては、バイデン大統領時代に AI に関する大統領令[S-13]が発令されたが、トランプ政権になって緩和され、現在不透明な状態となっている。また、ISO42001「AI マネジメントシステム」[S-14]も発行され、標準規格の制定も進みつつある。このような法規則や標準規格以外においても、日本において AI 事業者ガイドライン[S-3]が発行され、AI の開発者、提供者、利用者それぞれも立場向けに注意点がまとめられ、示唆に富んだガイドラインとなっている。AI 事業者ガイドラインにおける開発者は本ガイドラインのガイド Ver2 までで我々が想定読者としていた開発者である。AI 事業者ガイドラインの AI 提供者は、今回 S-4.2 で触れた、「モデルを外部から購入してユースケースに応じてファインチューニングをした上でインタフェース部分を開発したり、オンライン上のモデルを利用してシステムを実現したりする開発者」が該当する。また S-2 節で述べた生成 AI のリスクに関しては、本ガイドラインの執筆途中（2025 年 6 月）に産業技術総合研究所より、生成 AI 品質マネジメントガイドライン[S-15]が発行された。産総研のガイドラインでは、生成 AI リスクのうち主に誤動作（故障）に起因するリスクについて整理されており、悪意ある利用のリスクやシステミックリスクについては詳細な議論は行われていないが参考になる。

機械学習システムの開発者や提供者は、提供先の国に応じて各国の法規則を遵守することが必要である。また、必要に応じて ISO42001 に準拠したマネジメントを行うなどの対応も検討する必要がある。また、AI 事業者ガイドラインや機械学習品質マネジメントガイドライン[S-16]などの参考ドキュメントを参考にしながら開発を進めることも推奨される。

S-3.2. 各国の法規則や標準規格本ガイドラインとの関係について

S-3.1 節で紹介した法規則は法令であるため、遵守する必要がある。ISO などの標準規格は、規格への適合が求められる場合には各要求事項を遵守する必要がある。AI 事業者ガイドラインや機械学習品質マネジメントガイドラインなどのドキュメントは参考ドキュメントであるが、示唆に富んだ内容であるため、参考にするのが好ましい。ガイド Ver2 でも述べているが、このような状況において、**本ガイドラインは、参考ドキュメントと位置付ける。**つまり、遵守の必要がある法規則や、準拠すると決めた標準規格の規定を優先的に順守することを必要とし、本ドキュメントはそれらに加えてさらにセキュリティを向上したい場合に活用されるドキュメントと位置付ける。したがって、必ず守らないといけないものではない。

S-4. 本ガイドラインの旧バージョンにおけるリスク分析について

S-4.1. 近年の動向を受けてのリスク分析ツールの位置づけ

近年生成 AI などの発展が目覚ましい。AI 技術は発展段階であるが、生成 AI は近年特に発展が目覚ましく、2023 年に公開した分析技術ではリスクを評価しきれない可能性がある。そこで、**ガイド Ver2 以前で公開しているリスク分析技術・ツールについては、生成 AI は分析対象外**とすることとする。その代わり、S-2 章にて、生成 AI のセキュリティリスクについて一定の整理を行った。前述のとおり生成 AI は発展が目覚ましいため、本ドキュメントでの整理は現時点（2025 年 6 月）のものとなることに留意されたい。なお、本バージョンで新たに公開する**新影響分析ツールについても、生成 AI に適用できる可能性があるが、筆者らは検証を行っていたいため非推奨**とする。

また、ガイド Ver2 までに公開している 2 つのリスク分析ツールのうち、AI リスク問診について、この技術は機械学習システムにどのような攻撃が適用できるかを見極める技術となっているが、あらゆる攻撃の適用可否を見極めることは困難であることに留意されたい。なぜなら、近年攻撃技術が目覚ましく進展している一方で、この技術では代表的、典型的な攻撃を分析対象としているためである。したがって、代表的、典型的な攻撃の適用可否についてのリスクの分析は可能であるが、あらゆる攻撃の適用可否を見極めることは困難である。重要インフラなど、リスクを生じると特に問題となるシステムに機械学習を活用する際には、本ツールの結果のみを鵜呑みにするのではなく、より精緻な検討が必要となる。

S-4.2. 近年の多様な開発形態について

ガイド Ver2 までのバージョンでは、機械学習システムの開発者は、自身で訓練データを集めてモデルを訓練し、訓練済みモデルを顧客に提供することが想定されている。これに対し、近年では、モデルを外部から購入してユースケースに応じてファインチューニングをした上でインタフェース部分を開発したり、オンライン上のモデルを利用してシステムを実現したりするケースが増加している。このようなケースは AI 事業者ガイドライン[S-3]でも整理されており、特に生成 AI を活用したシステムで顕著に行われている。セキュリティリスクを考えた場合、このような開発形態の多様化が進んだ現状においてもリスク分析が必要であると考えられる。ガイド Ver2 までで公開しているリスク分析技術・ツールはこのような開発を想定したものではないが、ほぼそのまま活用することは可能であると考えられる（ただし、S-4.1 で述べた通り、生成 AI は対象外とする)。ガイド Ver2 で公開してい

るリスク分析を実施するにあたり、分析のための質問回答が困難な状況に直面する可能性があり、その場合にはモデル提供者にヒアリングしたり、安全側に倒して回答したりすることでリスク分析が可能になると考えられる。

S-5. 新影響分析ツールの公開

S-5.1. 概要

本ガイドラインのリスク分析編 II-5.3 節にて、AI 開発者向けの 2 種類のリスク分析手法として、影響分析と脅威分析が紹介されている。影響分析は、リスクが顕在化した際にどのような種類の被害がどの程度生じうるかを明らかにする分析であり、脅威分析は、システムにどのような攻撃が実施されうるかを明らかにする分析である。本ガイドライン補足編を発行するにあたり、今回、影響分析ツールの新バージョンを公開する。このバージョンは文献[S-17]で記載されている影響分析技術をツールにしたものである。新バージョンの影響分析ツールでは、これまでのバージョンで分析対象としていた AI セキュリティだけでなく、AI 品質、AI 倫理の観点でもリスクの影響を分析できるようになっている。このバージョンの公開により、これまでのバージョンの分析ツールを放棄するわけではなく、AI セキュリティの観点のみで分析したい場合には、以前のバージョンを使用してもよい。また、**本ツールは参考ツールとしての位置づけ**であり、必要に応じて使っていただきたい。

S-5.2. 新バージョンの影響分析ツールについて

影響分析については、本ガイドラインリスク分析編 II-6.1 節を参考にされたい。今回のツールは AI-ImpactAssessmentTool_2.0.xlsx というファイル名で公開する。このバージョンでは AI セキュリティ加えて、AI 品質と AI 倫理の影響分析を可能にした。分析の流れは以下のとおりである。

S-5.2.1. 新影響分分析の進め方

この影響分析の進め方は以下のとおりである。

1. シート「I. 説明」、「II. 分析の進め方」を読み、分析の進め方を理解する
2. シート「III. AI システムの仕様情報」のシートに分析対象システムの情報とシステムに関係するステークホルダーの情報を入力する
3. シート「IV-1. AI 品質の影響分析」のシートで AI 品質の観点の影響分析を行う
4. シート「IV-2. AI 倫理の影響分析」のシートで AI 倫理の観点の影響分析を行う
5. シート「IV-3. AI セキュリティの影響分析」のシートで AI セキュリティの観点の影響分析を行う
6. シート「IV-4. 広義倫理の影響分析」のシートで AI 倫理に属さないが広い意味で倫理に関係する観点の影響分析を行う

7. シート「V. 分析結果」のシートを見て結果（影響度）の確認を行う

S-5.2.2. 新影響分析ツールにおけるリスクの構造について

このツールでは、リスクを以下のように整理している。

大項目分類：リスクの種類

AI 品質、AI 倫理、AI セキュリティ、広義倫理の種類に対応

中項目分類：ハザード

例えば AI 品質シート内の「AI の誤判断」、「顧客との合意形成不十分」、「システム監視体制不十分」といった、各分析テーブル 1 個 1 個に対応

小項目分類：リスクシナリオ

ハザード毎の分析テーブルの 1 行 1 行に記入されるリスクシナリオに相当

ハザードについて

本分析ツールでは、AI 品質、AI 倫理、AI セキュリティ、広義倫理といった各リスクの種類の下位の分類として「ハザード」という分類名で分類を行っている。「ハザード」は、S-2.2 節でも記載しているが、ISO/IEC Guide51[S-11]に基づいた、被害を発生させる元となる危険な状態のことである。つまり、分析テーブルの各行に記載されたリスクの状態が生じた際に関連するハザードが発生し、それにより被害が生じるという流れとなる。各ハザードは、文献[S-17]で示されるように、文献[S-16][S-18] [S-19]を元に抽出された代表的なものとなっている。

S-5.2.3. 分析テーブルの入力方法について

S-5.2.1 節で示した分析の進め方の中の手順 3〜6 については、各分析テーブルにて分析を行う。この分析の初めの段階では、シート「III. AI システムの仕様情報」で入力した AI システムの情報を元に、テーブルの左側が自動的に転写されて埋まった状態となっている（図 S- 2)。この状態を開始状態として、テーブル内の 1 行 1 行それぞれについて、水色の欄に記入する。このとき、左側の「誰が」「なぜ」「いつ」「どこで」「どのように」「誰に（何に）(想定被害者)」「どうなる」に自動転写されている状況が起きたとしたらどのような影響が起きるかを想像して記入する。例えば、図 S- 2 では、レントゲン画像を元に病名を推測する AI システムの分類例であり、1 行目では、「医療画像分類システムが（誰が)、期待して結果が出ないことで（なぜ)、システム運用時に（いつ)、医療画像分類システムで（どこで)、AI が誤判断することで（どのように)、患者に対して（誰に（想定被害者))、レントゲン画像から病名を分類することができない（どうなる)」という状況が提示されている。このような状況が起きた場合、生じる影響は「病名分類を誤る」と考えられるため、それを

「「どうなる」により生じる具体的な影響」欄に記入する。そして、病名分類を誤ったときの引き起こされる被害を、「身体的被害面」「金銭被害面」「社会的被害面」「その他の被害面」の観点で入力する。例えば身体的被害面については、「病名診断ミスにより誤った医療処置を行い、患者が死亡する」などとなる。そしてそのような被害が出たときの影響（インパクト）を「被害の程度」欄において大、中、小、無で選択する。今回は人命にかかわるので「大」が妥当と考えられる。なお、この大、中、小、無の判断は、「影響度の算定基準」のシートに参考の判断基準を載せているのでそちらを用いて判断してもよい。このような分析を全ての行に対して行うと、各行で選択した「被害の程度」の最大値がそのテーブル（ハザード）の影響度として算出される。この分析を全てのハザードテーブルに対して実施すると、ハザード毎の影響度がすべて算出され、シート「V. 分析結果」にまとめられる。ハザード毎の被害の大きさが明らかになるため、この情報を参考に、どのハザードから対処すべきかを検討する。実際の対処方法は、リスクの種類やハザードによって異なるが、AI セキュリティについては本ガイド Ver2 までの内容、AI セキュリティを含む AI 品質、AI 倫理については機械学習品質マネジメントガイドライン[S-16]などが参考になる。

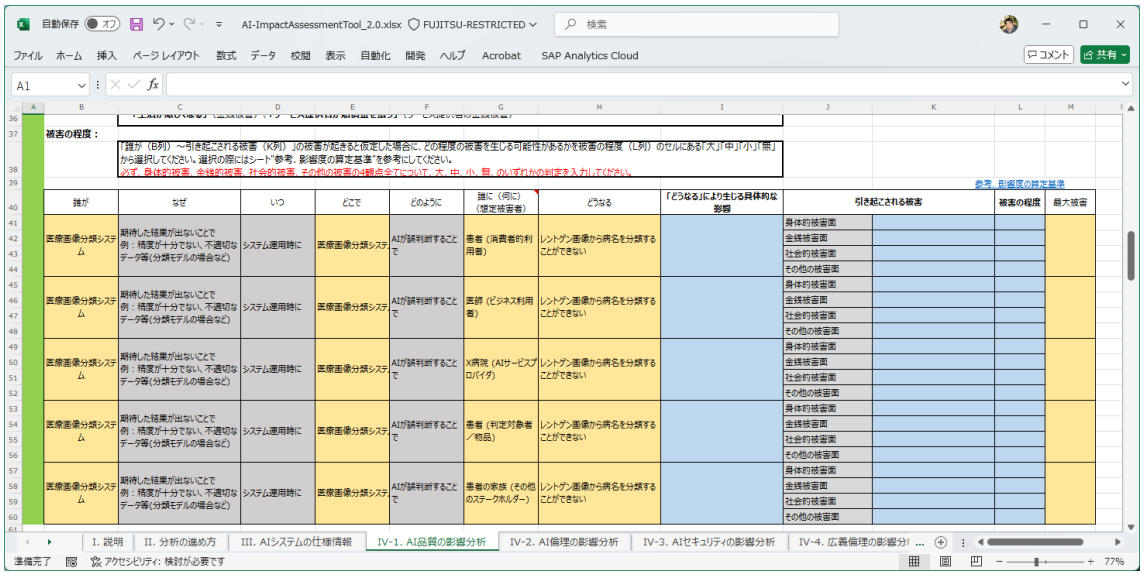

図 S- 2. 分析テーブルの開始状態の例（AI 品質）

S-6. 参考文献

[S-1] 機械学習工学研究会, 機械学習システムセキュリティガイドライン
https://github.com/mlse-jssst/security-guideline

[S-2] European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)s
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[S-3] AI 事業者ガイドライン検討会, AI 事業者ガイドライン（第 1.1 版）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_1.pdf
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/index.html

[S-4] L. Weidinger, et.al. "Ethical and social risks of harm from Language Models," arXiv:2112.04359 [cs] (Dec. 2021)

[S-5] L. Weidinger, et.al. "Taxonomy of Risks posed by Language Models," Poc. of FAccT '22, pp.214-229,
https://doi.org/10.1145/3531146.3533088 (June 2022)

[S-6] CRDS JST, "人工知能研究の新潮流 2 ～基盤モデル・生成 AI のインパクト～," 戦略提案・報告書 CRDS-FY2023-RR-02, (July 2023).
https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html

[S-7] OpenAI, "GPT-4 System Card," (Mar. 2023).
https://cdn.openai.com/papers/gpt-4-system-card.pdf

[S-8] カテライ アメリア, 井出 和希, 岸本 充生, "生成 AI（Generative AI）の倫理的・法的・社会的課題（ELSI）論点の概観：2023 年 3 月版", ELSI NOTE. No.26, pp.1-pp.37, (Apr. 2023). DOI：10.18910/90926.

[S-9] H. Tanaka, M. Ide, S. Onodera, K. Munakata, N. Yoshioka, "Taxonomy of Generative AI Applications for Risk Assessment," In 2024 IEEE/ACM 3rd. International Conference on AI Engineering. Software Engineering for AI (CAIN), (Apr. 2024). DOI: 10.1145/3644815.3644977

[S-10] 田中, 井出, 矢嶋, 小野寺, 宗像, 吉岡, “要求分析に基づく生成 AI リスクの分類と対策の導出”, 2025 年暗号と情報セキュリティシンポジウム（SCIS 2025), 3G1-5, (Jan 2025).

[S-11] International Organization for Standardization, ISO/IEC Guide 51:2014 Safety

aspects - Guidelines for their inclusion in standards,
https://www.iso.org/standard/53940.html

[S-12] 衆議院，人工知能関連技術の研究開発及び活用の推進に関する法律案，
https://www.shugiin.go.jp/internet/itdb_gian.nsf/html/gian/honbun/houan/g21709029.htm

[S-13] Federal register, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,
https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[S-14] International Organization for Standardization, ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system
https://www.iso.org/standard/81230.html

[S-15] 産業技術総合研究所，生成 AI 品質マネジメントガイドライン第 1 版
https://www.digiarc.aist.go.jp/publication/aiqm/genaiqm-guidelines-v1.html

[S-16] 産業技術総合研究所，機械学習品質マネジメントガイドライン第 4 版
https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html

[S-17] 矢嶋，井出，田中，志賀，大橋，小野寺，AI セキュリティ影響分析技術の AI 倫理・AI 品質への拡張
2025 年暗号と情報セキュリティシンポジウム(SCIS2025), 3G1-3

[S-18] European Commission, Ethics guidelines for trustworthy AI
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[S-19] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

機械学習システムセキュリティガイドライン策定委員会メンバーリスト

・第 3 版メンバー

石川　冬樹　　（国立情報学研究所）
江澤　美保　　（株式会社クレスコ）
大久保　隆夫　（情報セキュリティ大学院大学）
久連石　圭　　（株式会社東芝）
鈴木　淳哉　　（デロイト トーマツ サイバー合同会社）
田中　宏　　（富士通株式会社）
林　昌純　　（帝京平成大学）
矢嶋　純　　（富士通株式会社）
吉岡　信和　　（QAML 株式会社／早稲田大学）
（敬称略・五十音順）

・元メンバー（所属は策定委員会在籍時のもの）

市原　大暉　　（株式会社 NTT データ）
乾　真季　　（富士通株式会社）
及川　考徳　　（富士通株式会社）
笠原　史禎　　（富士通株式会社）
金子　朋子　　（国立情報学研究所）
田口　研治　　（国立情報学研究所）
辻　健太郎　　（富士通株式会社）
花谷　嘉一　　（株式会社東芝）
森川　郁也　　（富士通株式会社）
（敬称略・五十音順）