「AIのセキュリティ確保のための技術的対策に係るガイドライン」（案）に対して提出された意見及びその意見に対する総務省の考え方

■意見募集期間 ： 令和7年12月26日（金）～令和8年1月29日（木）

■意見提出件数 ：52件（法人・団体：14件、個人：38件）

■意見提出者

	意見提出者
1	株式会社Acompany
2	アマゾン ウェブ サービス ジャパン合同会社
3	エムオーテックス株式会社
4	シスコシステムズ合同会社
5	一般社団法人新経済連盟
6	株式会社セールスフォース・ジャパン
7	株式会社東芝
8	日本消費生活アドバイザー・コンサルタント・相談員協会（NACS）
9	パロアルトネットワークス株式会社
10	合同会社ヒロタ開発
11	PwCコンサルティング合同会社
12	富士通株式会社
13	楽天グループ株式会社
14	株式会社ラック
ー	個人（38件）

※頂いた御意見につきましては、原文を御意見ごとに分割して記載しております

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
1	一般社団法人新経済連盟	本編	本ガイドラインの策定の背景等	背景にも記載はされているが、生成AI等の技術発展は著しく、本ガイドラインの内容も定期的なアップデートが必要であると考える。各事業者にとって有益となるような対策や事例等の収集など、引き続き、産業界との連携を密にされることを要望する。	御意見として承ります。今後、本編「本ガイドラインの策定の背景等」に記載のとおり、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
2	一般社団法人新経済連盟	本編	本ガイドラインの策定の背景等	広告制作では画像生成AIが既に実用段階にあり、意図しない著作権侵害や不適切画像の生成対策が急務である。また、MCPについてもAIエージェントの実用化にあたり活用を検討している。新たなテクノロジーに対応したセキュリティ指針も可能な限り早期に提示していただきたい（現在の記載では後回しにするように受け取れる）。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。MCPやAIエージェントに関しては、本編「本ガイドラインの策定の背景等」に示している考え方のとおり、今後、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
3	個人	本編	本ガイドラインの策定の背景等	【現代社会の複雑性と認知戦を想定した背景定義の抜本的拡充】 現代社会は極めて複雑な構造を有しており、特定の情報が社会全体にもたらす波及影響を事前に予測することは困難です。この状況下でAIが情報を爆発的に生成・伝播させることは、社会のレジリエンス（強靭性）に対する挑戦といえます。 本ガイドライン案の背景において「社会経済活動に多大な影響を生じさせかねない」と言及されていますが 、現在の国際情勢においてAIを用いた「認知戦」は、既存の社会構造の脆弱性を露呈させる看過できない安全保障上の脅威です。情報源として挙げられているMITRE ATLASの「Societal Harm（AML.T0053）」では、AIの悪用による偽情報の拡散や社会的結束の破壊が「インパクト（攻撃の影響）」として明確に定義されています 。 したがって、本ガイドラインにおいては、単なるシステムの停止や漏えいといった物理的・技術的侵害に留まらず、AIによる「偽情報の自動生成・大量拡散」そのものをセキュリティ上の重大な脅威として明記すべきです。単なる技術導入の是非を論ずるのではなく、この爆発的な情報流通に耐えうる社会のあり方を問い直す、情報空間の誠実性（インテグリティ）確保という本質的なセキュリティの視点を背景に盛り込むことを強く求めます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
4	PwCコンサルティング合同会社	本編	1.1 本ガイドラインの位置づけ	意見：悪意ある攻撃に加えて設定ミス等のヒューマンエラーも影響低減の対象とされていますが、第2章の脅威整理では主に攻撃手法が中心となっており、設定ミスや運用ミスが主要な脅威として十分に整理されていません。設定ミスや運用ミス等のヒューマンエラーについて、第2章の脅威整理に主要な脅威として明示的に追加することを推奨します。 AI事業者ガイドラインの「AI開発者に関する事項」や、OWASP AI Exchangeの「Development-time threats」等を参照し、攻撃手法以外の脅威整理を補強することが望ましいと考えます。 OWASP AI Exchange「Development-time threats」<https://owaspai.org/docs/3_development_time_threats/>	御意見を踏まえ、本編P11の脚注に、「悪意のある攻撃以外では、ヒューマンエラーによる設定ミス等によって生じる脅威についても留意が必要である。」と記載しました。
5	パロアルトネットワークス株式会社	本編	1.1 本ガイドラインの位置づけ	意見3：高度化する脅威に対抗する「AI for Cybersecurity」と面的防御の推進 該当箇所: 1.1 ガイドライン案の位置づけ、または 全体 意見の概要: 本ガイドライン案は「AIを守る（Cybersecurity for AI）」ごとに主眼が置かれているが、AI技術の悪用により脅威が劇的に高度化している現状に鑑み、「AIで守る（AI for Cybersecurity）」の視点も不可欠である。政府全体の戦略として、攻撃の高速化に対抗するため、AIを活用した「点ではなく面（統合されたデータと可視性）」による防御体制の構築を推進することを期待する。 理由: 1. 国家戦略との整合性と脅威の深刻化 令和7年12月26日に閣議決定された『サイバーセキュリティ戦略』においても、「生成 AIを始めとする AIの急速な発展は、今後、産業や国民生活の利便性や効率性を大きく向上させる潜在力を持つ一方、サイバー犯罪の巧妙化等新たな脅威を生み、社会での活用・普及に伴い、AIに対する攻撃や AIを利用した攻撃が、新たなサイバーセキュリティ上のリスクとして、深刻さを増すことが想定される」と指摘されている通り、AIの悪用によるリスクは深刻さを増している。 2. 攻撃の「マシンスピード化」と従来型防御の限界 パロアルトネットワークスのUnit42のデータによれば、攻撃者が侵入からデータ持ち出しに至るまでの時間は4年前と比較して100倍高速化しており 、エージェントAIを悪用したランサムウェアキャンペーンに至っては、偵察から侵害までがわずか25分程度で完了する事例も確認されている 。このような「マシンスピード」の攻撃に対し、人間が個別のセキュリティ製品（点）のアラートを手動で突き合わせる従来の手法では、もはや対応が不可能である。 3. 「点」から「面」への防御の転換 AIを活用したサイバー防御（AI for Cybersecurity）を有効に機能させるためには、単にセキュリティにAIツールを導入するだけでは不十分である。攻撃はインフラ、アプリケーション、ユーザーIDなど複数のレイヤーを横断して行われるため、防御側もこれらを分断された「点」として扱うのではなく、統合されたデータとして「面」で捉える必要がある。 断片化されたデータではAIは正確な相関分析ができず、誤検知や見落としにつながる 。 意見: したがって、今後の政策立案においては、エンドポイントからクラウド、ネットワークに至るログとテレメトリを統合し、AIがエコシステム全体（面）をリアルタイムで分析・防御できるアーキテクチャへの移行を強く推奨する。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
6	シスコシステムズ合同会社	本編	1.1 本ガイドラインの位置づけ	•本ガイドラインの位置付けについて、第1章1.1節（3ページ）において、国際的なフレームワークや標準との相互運用性の重要性を明記することを提言する。グローバルに事業を展開する企業にとって、日本独自の規格を要求することは、コンプライアンスコスト増大の懸念がある。NIST AI RMF（AIリスクマネジメントフレームワーク）、ISO/IEC 27090/27091、その他関連するフレームワークなど、国際的に認知された規格との整合性や相互参照について明示的に言及することを要望する。 さらに、「セキュリティの確保」の定義は、AIシステムの訓練と配備から生じる可能性のあるセキュリティ上の懸念の範囲を完全に包含しておらず、機密情報の漏洩につながるあらゆる脅威に主眼を置いているように思われる。我々は、AIシステムがもたらすリスクの範囲をより適切に包含するために、定義の改訂を提言する。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、AIに対する脅威のリスクを低減するため、現時点で取り得るとされる一般的な対策例を整理し、提示するものであり、我が国独自の規格を要求するものではありません。「セキュリティ確保」の定義に関する御意見につきましては、今後の政策の検討にあたり、参考とさせていただきます。
7	個人	本編	1.1 本ガイドラインの位置づけ	【人間中心の倫理・哲学的要素に基づく対策の再評価】 本案が引用する「AIセーフティ」の定義には「人間中心の考え方」が含まれています 。技術的対策は単なるITインフラの防衛に留まらず、「人間がAIによって操作・欺瞞されない」という倫理・哲学的価値を守るためのもの であるべきです。 ガードレールの評価基準 や安全基準の学習 の根底に、利便性よりも「人間の自律性」や「真実性」を尊重するという哲学的指針を据えるよう記述を充実させるべきです。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
8	個人	本編	1.1 本ガイドラインの位置づけ	＞関係省庁・関係機関が策定しているAI関連ガイドライン等のうち、本ガイドラインと関連する主なものは表１に示すとおりである。 とあるが、表１だけでは各ガイドライン間の関係／連係が分かりにくいため、 １．１．１として各ガイドライン間の関係性を示す、例えばベン図といった図版を用意いただきたい。	本編の表１については、本ガイドラインと各ガイドラインの関係を表したものであり、現状においては、必ずしも図示になじまないものであると考え、表での記載としたものです。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
9	シスコシステムズ合同会社	本編	1.2 対象とするAI	•第1章1.2節図1（5ページ）のAI構成の例において、AIエージェントの適用除外に論理的な矛盾があることに留意する。本ガイドラインの背景の項では、AIエージェントやMCPの自律性の高まりを「新たな脅威」と明確に位置づけ、技術の進歩に合わせた継続的な見直しの重要性を強調している。にもかかわらず、AIエージェントを除外している点は、本ガイドラインの政策方針と論理的に矛盾している。 さらに、図1や用語集の「オーケストレーター」の定義は、すでに外部システムやAPIとの統合を包含しており、最新のAIエージェントの基本要素（エージェント型ワークフロー）を網羅している。具体的には、外部システムからの出力をオーケストレータにフィードバックする「実行ループ」を想定することで、図1の構造を変更することなくAIエージェントを組み込むことができる。 この文書では、RAG（Retrieval-Augmented Generation：検索支援型ジェネレーション）が詳細に扱われており、権限管理、タグ付け、その他の具体的な対策が網羅されている。これとは対照的に、もう一つの主要なLLMユースケースであるAIエージェントを除外することは、ガイドラインの包括性を損ない、自律的な操作に関連するリスク（最小特権の原則からの逸脱など）を管理するガイダンスを欠く危険性がある。	御意見として承ります。AIエージェントについては、技術が急激な発展の途上にあり、これに特有の脅威や対策を安定的に確定することが現時点では困難であることから、対象外としています。今後については、引き続き関係省庁及び関係機関とも連携しながら、AIの技術進展を十分に踏まえ、新たな脅威や対策の動向を注視し、対応をはかっていくものです。
10	富士通株式会社	本編	1.2 対象とするAI	本ガイドライン（案）ではAIエージェントは対象外とされているが、「オーケストレータ」という概念には、AIエージェントのツール連携機能などが含まれることから、例えば、「LLMアプリケーション」のような表現が適当と考えます。	本ガイドラインにおいては、本編「用語集」に記載のとおり、オーケストレータは、「予め定義された実行計画に基づき、大規模言語モデル(LLM)を搭載したシステムのワークフローを統合的に管理するためのフレームワーク」を指し、AIエージェントは、「環境を知覚し、その環境について推論し、意思決定を行い、特定の目標を達成するために自律的に行動するAIシステム」を指しており、対象とするAIシステムはこの整理に基づいたものとしています。
11	富士通株式会社	本編	1.2 対象とするAI	外部連携システムに、攻撃者によって細工されたモデルの混入など、サプライチェーン攻撃を明示的に視覚化するための追加表現を加えることを提案します。 例）ブロック拡張（右側に外部供給元を追加）、このブロックから外部連携システムへ矢印を引き、ラベル「サプライチェーンリスク（細工混入）」を付与。 あるいは、例）注釈強化（図変更最小限）し、外部連携システムに注釈を入れ、図下キャプションに外部システムにはサプライチェーンのリスクがあり、攻撃者による細工されたモデル/データ/コンポーネントの混入を防ぐため、供給元の検証・整合性チェックを推奨。	本編の図 1 において、AIモデルは、外部提供によるものか、社内で開発されたものかに依らず、左下の点線枠内から提供されます。また、図 1 は、本ガイドラインが対象とするAIシステムを図示するものであり、特定の攻撃を扱うものではありません。
12	パロアルトネットワークス株式会社	本編	1.2 対象とするAI	意見2：Agentic AI への対応と Secure AI by Design の適用 該当箇所: 1.2 対象とするAI、および 4. 今後の動向を踏まえた対応について 意見の概要: 本ガイドライン案において、AIエージェント（Agentic AI）は技術発展の途上にあるとして対象外とされている。しかし、OWASP Top 10 for Agentic Applications 2026 等で指摘されているように、自律的なエージェント技術は既にビジネス実装の段階にあり、リスクの本質も劇的に変化している。したがって、早期に議論を開始するとともに、エージェント特有の振る舞いを保護するために、AIエコシステム全体を防御する「Secure AI by Design」の思想を適用することを推奨する。 理由: ガイドライン案では、AIエージェントの扱いについて「特有の脅威や対策を安定的に確定することが困難」とされている。しかし、OWASP 2026レポート等でも指摘されている通り、AIエージェントは単にテキストを生成するだけでなく、自律的な計画立案、外部ツールの実行、他エージェントとの連携を行うものであり、リスクの本質が従来の「単体の脆弱性（静的な入出力の問題）」から「行動ベースの失敗（Behavior-driven failure modes）」へと変化している。 こうした動的かつ文脈依存のリスクに対し、従来の「点」での防御は無力である。そのため、Palo Alto Networksの証言でも提唱されている、AIライフサイクル全体を一貫して保護する「Secure AI by Design」のアプローチが不可欠となる。同証言では、自律型エージェントの普及（As autonomous and semi-autonomous agents proliferate）を見据え、セキュリティ対策の中心（Center）として以下の要素を実装すべきであると明確に提言している。 アイデンティティと最小権限の制御（Identity & Least Privilege）：エージェントに対し「アイデンティティ・ファースト」の制御を行うとともに、「最小権限アクセス」の原則を適用し、エージェントが付与される権限を厳格に管理する。 ツール権限のスコープ設定と職務分掌（Scope & Separation）：エージェントが実行可能なツール権限をタスクに必要な範囲に限定（スコープ設定）し、単一のエージェントに権限が集中しないよう「職務分掌」を徹底することで、暴走時の被害を物理的に封じ込める。 通信のガバナンスと異常行動の監視（Monitoring & Governance）：エージェント間およびエージェント対システムの通信をガバナンス下に置き、実行時（ランタイム）における異常な振る舞いをリアルタイムで監視する。 これらは、個別の製品導入ではなく、設計段階から運用まで一貫したセキュリティ思想（Secure AI by Design）があって初めて実現できるものである。 意見: したがって、次期ガイドラインの策定においては、Agentic AIを主要なスコープに含めるとともに、この包括的な防御思想に基づいた対策基準を整備することを強く要望する。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
13	個人	本編	1.2 対象とするAI	LLM、AIシステムにおけるガードレールは、特定のプロンプト（悪意のあるプロンプト等）に対して内容を検閲し、評価軸に該当した場合、回答を拒否するようなものと認識しており、双眼鏡のアイコンから想起されるイメージよりも、FW（ファイヤーウォール）のようなイメージがより近いと考えます。	ガードレールのアイコンについては、ガードレールが入出力の監視を行うというイメージで双眼鏡のアイコンを用いています。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
14	個人	本編	2 脅威	【意見の内容：事後ゲーティング型外部統治層の明示的追加】 AIは、特定の入力や文脈条件により推論の劣化や停止が生じ、可用性（Availability）が低下するリスクを内在している。これに対し、出力生成後に外部の決定論的評価層が固定制約に基づき可否判定を行う方式（ExMachina Type A 等）は、モデル改変・再学習・最適化を伴わず、AI本体の攻撃面を拡張しない実装が可能である。 本提案方式は、以下の特徴を有する。 定数の凍結 制御パラメータ（kc、α、limit 等）を一度校正後に凍結し、決定論的に動作すること。 順方向実行（Forward-only） 学習（バックプロパゲーション）を行わず、非学習型の外部ガバナーとして前進実行のみで動作すること。 明示的な閾値判定 ICI（偏差指標）等の定量的指標に基づき、出力の受理（ACCEPT）／棄却（REJECT）を客観的に判定すること。 高い監査性 完全な実行ログ（Execution Trace）により、処理過程の追跡および説明が可能であること。 また、本方式ではAIモデル本体とは独立した外部サーバー上に、出力評価専用のフィルタリングアプリケーション（ガバナー）を配置する。当該ガバナーは学習・適応・最適化を行わず、固定制約に基づく前進実行のみで可否判定を行うため、分離配置による可用性および監査性の向上が可能である。 これにより、出力汚染や推論ストールといった可用性低下事象を運用段階で抑止でき、ガイドラインの趣旨であるAIの安全かつ安定的な利用に資するものと考える。 ついては、ガイドライン本文または付属資料において、**「事後ゲーティング型外部統治層（非学習・非適応）」**を、技術的対策の有効な選択肢として明記されることを提案する。 【補足】 本方式はモデル改変を伴わない「運用レイヤーでの安全確保」に該当し、既存のAIシステム（例：大規模言語モデル等）への後付け適用が可能である。 【参考資料】 本意見で言及した事後ゲーティング型外部統治層の具体的な設計例および実証的資料については、以下のDOIにて公開している。 https://doi.org/10.5281/zenodo.18338998 当該資料は、非学習・非適応・決定論的な外部ガバナー構成を示すものであり、ガイドラインにおける技術的対策の参考例として提供するものである。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
15	個人	本編	2 脅威	【意見の内容：事後ゲーティング型外部統治層の明示的追加】 AIは、特定の入力や文脈条件により推論の劣化や停止が生じ、可用性（Availability）が低下するリスクを内在している。これに対し、出力生成後に外部の決定論的評価層が固定制約に基づき可否判定を行う方式（ExMachina Type A 等）は、モデル改変・再学習・最適化を伴わず、AI本体の攻撃面を拡張しない実装が可能である。 本提案方式は、以下の特徴を有する。 定数の凍結 制御パラメータ（kc、α、limit 等）を一度校正後に凍結し、決定論的に動作すること。 順方向実行（Forward-only） 学習（バックプロパゲーション）を行わず、非学習型の外部ガバナーとして前進実行のみで動作すること。 明示的な閾値判定 ICI（偏差指標）等の定量的指標に基づき、出力の受理（ACCEPT）／棄却（REJECT）を客観的に判定すること。 高い監査性 完全な実行ログ（Execution Trace）により、処理過程の追跡および説明が可能であること。 また、本方式ではAIモデル本体とは独立した外部サーバー上に、出力評価専用のフィルタリングアプリケーション（ガバナー）を配置する。当該ガバナーは学習・適応・最適化を行わず、固定制約に基づく前進実行のみで可否判定を行うため、分離配置による可用性および監査性の向上が可能である。 これにより、出力汚染や推論ストールといった可用性低下事象を運用段階で抑止でき、ガイドラインの趣旨であるAIの安全かつ安定的な利用に資するものと考える。 ついては、ガイドライン本文または付属資料において、**「事後ゲーティング型外部統治層（非学習・非適応）」**を、技術的対策の有効な選択肢として明記されることを提案する。 【補足】 本方式はモデル改変を伴わない「運用レイヤーでの安全確保」に該当し、既存のAIシステム（例：大規模言語モデル等）への後付け適用が可能である。 【参考資料】 本意見で言及した事後ゲーティング型外部統治層の具体的な設計例および実証的資料については、以下のDOIにて公開している。 ・ExMachina Type A に基づく非学習・非適応・決定論的外部ガバナーの実装例 https://doi.org/10.5281/zenodo.18338998 併せて、当該方式の理論的背景およびAIセキュリティにおける構造的防御モデルについては、以下のプレプリントにて整理している。 ・AI Security Measures Needed Now Beyond Detection and Blocking https://doi.org/10.5281/zenodo.18281723 後者の資料では、検出・ブロック中心の対策を超え、AIモデル外部における構造的境界制御（Layer 0 を含む）によるセキュリティ確保の枠組みを提示しており、本意見で提案した事後ゲーティング型外部統治層の位置づけを理論的に補強するものである。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
16	株式会社ラック	本編	2.1 対象とする主な脅威	実際に対策を導入するにあたっては、脅威を選定した根拠を明示することが必要となりますので、本ガイドライン「２脅威」において対象とした脅威の選定根拠を具体的に詳細に示していただくことで、より実践的なものなるものと思料いたします。	プロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）は、基本的に、プロンプトの入力により実施可能であることから、攻撃が行われる具体的な可能性が比較的高く、かつ攻撃が実施された際の影響度も大きいと考えられるため、本ガイドラインにおいては、これらへの対策を主に示しています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
17	PwCコンサルティング合同会社	本編	2.1 対象とする主な脅威	ガイド全体をとおしてこの2つに重きをおいた内容になっている印象を受けました。表3で概観を記載頂いていますが、AI開発者またはAI提供者が追加的に用意する図1の「ファインチューニングデータ」の信頼性や真正性の確認はサプライチェーンリスクを考える上で重要な要素と考えます。 例えばAI提供者のビジネス要件に基づいてAI開発者がデータを収集し、ファインチューイングを行う場合、AI提供者がAI開発者が集めたデータを確認する等、もう少し踏み込んだ対策を追加することを推奨します。	御意見として承ります。ファインチューニングデータの信頼性の確認については、「3.5AI開発者・提供者に係るその他の基本的な対策等」において、「AI開発者及びAI提供者においては、開発・提供するシステムの目的・用途に応じて、ファインチューニングデータなどAIが学習するデータについて、出力を意図しない機密情報を用いないことや、データの出所・加工履歴等により信頼性を確認することが重要な場合もある。」として記載しています。
18	PwCコンサルティング合同会社	本編	2.1 対象とする主な脅威	AI提供者にとって「脅威の影響の大きさ」は比較的判断しやすい（ビジネスの重要度定義、被害金額、影響を受けるユーザー数等）一方で、脅威が発生する可能性は分析・判断が難しいと考えられます。そのため、システムの外部エクスポージャー、想定される利用者など、可能性の判断材料になる要素を補記することを推奨します。	御意見を踏まえ、本編2.1における該当箇所を「攻撃者が攻撃を実行できる可能性や、AIシステムがおかれた環境（例えば、外部との接続の有無、利用者の属性など）においてインシデントが起こり易いか否かで、対策の優先度は異なる。」と修正します。
19	PwCコンサルティング合同会社	本編	2.1 対象とする主な脅威	脅威の整理がプロンプトインジェクション等の具体的な攻撃手法の列挙に加えて、どの資産（データ、モデル、出力等）に対して、どの性質のリスク（機密性・完全性・可用性等）が生じるのかという整理軸を明示していただくことを推奨します。AI事業者ガイドラインやOWASP、NIST等で採用されている資産・影響別の整理軸を示したうえで、具体的な対策を記載していくことが望ましいと考えます。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
20	シスコシステムズ合同会社	本編	2.1 対象とする主な脅威	•第2章第2.1節(7ページ)の主要な脅威への対応では、脅威をより包括的に分類する必要がある。プロンプトインジェクションとサービス拒否攻撃がAIシステムに影響を与える可能性の高い脅威であることには同意するが、同様の可能性（および実装の容易さ）を持つ多数の脅威が他にも存在し、それらは包含されるべきである。さらに、プロンプトインジェクションとジェイルブレイク（脚注で言及）は、別個の攻撃手法であるため、より明確に区別する必要がある。AIシステムに対する脅威の包括的なカプセル化については、CiscoのIntegrated AI Security and Safety Frameworkを参照することを推奨する。	御意見として承ります。Jailbreakとプロンプトインジェクション攻撃については、御指摘の脚注６に記載のとおりです。
21	富士通株式会社	本編	2.1 対象とする主な脅威	本ガイドライン（案）では、AIのセキュリティ確保を「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じない状態」と定義しているが、脅威の記述が主に攻撃手法ベースとなっており、保護対象資産（何を守るべきか）の観点が相対的に明確にされていないと考えます。例えば、2.1の冒頭または直前に、保護対象資産の整理・例示を追加することを提案します。 これにより、事業者が自社AIシステムで攻撃によりどのような情報資産が損なわれるのかイメージしやすくなり、リスク評価・対策優先度の判断がしやすくなると考えます。 保護対象資産の例： ・機密情報（Confidentiality対象）：システムプロンプト、学習済みパラメータ、RAG参照データ（顧客データ・企業秘密）、ユーザ入力履歴、出力ログ ・安全性（Integrity対象）：モデルウェイト、外部連携データの整合性、出力の正確性 ・可用性（Availability対象）：LLM処理リソース、オーケストレーター、外部連携システム ・その他：知的財産（モデル抽出防止）、個人情報（漏えい防止）等	御意見として承ります。御指摘の保護対象資産は、AIシステムの目的・用途や提供条件によって異なり得ることから、原案のとおりとします。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
22	富士通株式会社	本編	2.1 対象とする主な脅威	プロンプトインジェクション攻撃やDoS攻撃を「攻撃の具体的な可能性が比較的高い」との基準で対象としているが、この選定をより客観的・体系的にするため、伝統的なモデリングフレームワークであるSTRIDEを参考に脅威を分類・評価することを提案します。また、マルチエージェントシステムの脅威分析に特化したMAESTROフレームワーク（OWASP(Open Web Application Security Project)が公開）を活用した脅威分析例が提供されており、これを参考にすることでLLMを中心とした現在の脅威に加え、将来的なマルチエージェントシステムへの拡張時にも一貫した分析が可能になると考えます。これにより、事業者の脅威特定プロセスが強化され、多層防衛の実効性が向上すると考えます。	プロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）は、基本的に、プロンプトの入力により実施可能であることから、攻撃が行われる具体的な可能性が比較的高く、かつ攻撃が実施された際の影響度も大きいと考えられるため、本ガイドラインにおいては、これらへの対策を主に示しています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
23	一般社団法人新経済連盟	本編	2.1 対象とする主な脅威	AIシステムに特有の攻撃手法が様々存在するなかで、本ガイドライン案では「プロンプトインジェクション攻撃」及び「DoS攻撃」への対策が示されている点について、選択論拠がより明確に示されるとよいのではないか。なお、AISI,2025年3月「AIシステムに対する既知の攻撃と影響」(https://aisi.go.jp/assets/pdf/Known_Attacks_and_Their_Impacts_on_AI_Systems_JP.pdf)も同様にAIシステムに特有の攻撃と影響を俯瞰し、対策を検討するための参考情報を提供することを目的とした文書であるが、AI システムに特有の攻撃手法が網羅されている。	プロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）は、基本的に、プロンプトの入力により実施可能であることから、攻撃が行われる具体的な可能性が比較的高く、かつ攻撃が実施された際の影響度も大きいと考えられるため、本ガイドラインにおいては、これらへの対策を主に示しています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
24	個人	本編	2.1 対象とする主な脅威	【戦闘空間における「道徳的悪意」への対抗とAI出力の兵器化防止】 現在の情報空間は、単なる技術的な競合の場ではなく、明確な「道徳的悪意」が持ち込まれた「認知戦」の場へと変貌しています。本ガイドライン案の背景にある「社会経済活動への多大な影響」という懸念を具現化するならば、情報の汚染は社会の根幹を揺るがす攻撃の「目的」そのものです 。情報源として挙げられている MITRE ATLAS の「Societal Harm（AML.T0053）」においても、社会的結束の破壊が攻撃のインパクトとして定義されています 。 このような環境下では、たとえAIの回答自体に悪意がなくても、生成された「もっともらしい」が誤った情報」が、悪意ある第三者によって拡散の道具として利用（兵器化）されるリスクを重大な脅威として想定すべきです。別添資料において「偽情報の出力」が安全基準の学習対象として位置づけられていることは、このリスクを裏付けています 。 したがって、技術的対策の目標を単なる「システムの安定（意図せぬ停止等の防止）」に留めるのではなく 、AI提供者の対策として、生成内容が認知戦等の道具として悪用された際の社会的影響を評価し、ガードレール等で抑制する視点を盛り込むべきです。悪意が渦巻く戦闘空間において、AIによる大衆欺瞞を許さないかに主眼を置いた実戦的な防衛指針への強化を強く求めます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
25	楽天グループ株式会社	本編	2.1 対象とする主な脅威	AIシステムに特有の攻撃手法が様々存在するなかで、本ガイドライン案では「プロンプトインジェクション攻撃」及び「DoS攻撃」への対策が示されている点について、選択論拠がより明確に示されるとよいのではないか。なお、AISI, 2025年 3 月「AIシステムに対する既知の攻撃と影響」(https://aisi.go.jp/assets/pdf/Known_Attacks_and_Their_Impacts_on_AI_Systems_JP.pdf)も同様にAIシステムに特有の攻撃と影響を俯瞰し、対策を検討するための参考情報を提供することを目的とした文書であるが、AI システムに特有の攻撃手法が網羅されている。	プロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）は、基本的に、プロンプトの入力により実施可能であることから、攻撃が行われる具体的な可能性が比較的高く、かつ攻撃が実施された際の影響度も大きいと考えられるため、本ガイドラインにおいては、これらへの対策を主に示しています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
26	株式会社セールスフォース・ジャパン	本編	2.1.1 プロンプトインジェクション攻撃	意見 1 【関連記載】2.1.1 プロンプトインジェクション攻撃　（P7） 【意見内容】「間接プロンプトインジェクション攻撃において参照させる「細工をしたデータ」の例として以下を追記することも一案と考えられます。 ・RAG参照先の非構造化データに埋め込まれたコンテンツ（PDF文書等）	御意見として承ります。本ガイドラインでは、間接プロンプトインジェクション攻撃の事例として、「細工したファイルをWeb上に用意し、LLMが当該ファイルを参照した際に不正な出力を誘発させる」ものを挙げています。RAGにより当該ファイルが参照される場合についても、当該事例において想定され得ると考えます。
27	PwCコンサルティング合同会社	本編	2.1.1 プロンプトインジェクション攻撃	例として外部Webの読み込み、電子メールを記載頂いていますが昨今マルチモーダル対応したサービスにおいては、ユーザーがアップロードしたファイルを契機に攻撃が行われる事例も確認されているため追記することを推奨します。	御意見として承ります。本ガイドラインでは、間接プロンプトインジェクション攻撃の事例として、「細工したファイルをWeb上に用意し、LLMが当該ファイルを参照した際に不正な出力を誘発させる」ものを挙げています。ユーザがファイルをダウンロードし、AIサービスにアップロードすることでLLMが当該ファイルを参照する場合についても、当該事例において想定され得ると考えます。
28	個人	本編	2.1.1 プロンプトインジェクション攻撃	7頁　2.1 対象とする主な脅威　2.1.1 プロンプトインジェクション攻撃 「『不正な出力』の例としては？」と挙げられた項目のうち ・本来は開示すべきではない、RAG用のデータストア（ベクトルデータベースやファイルシステム等）の内容を含む出力をさせる ・本来は開示すべきではない、LLMの内部設定が記載されたシステムプロンプトを含む出力をさせる これらに含まれる「本来は開示すべきではない」を「AI開発者またはAI提供者が開示を想定していない」「開示をすることがセキュリティ上のリスクとなる」などのような開示を拒否する理由が自発的ないし理由の曖昧さが排除された文言に必ず改めてください。 RAG用のデータストアの中には、学習元データから加工されたいわゆるチャンクデータが含まれるものがあり、知的財産の侵害が疑われる事例が発生した場合、検証のために開示が必要となる可能性があります。この可能性を否定できない状況で「本来は開示すべきではない」という文言を付すことは不適当であると指摘します。 AI開発者やAI提供者が「本来は開示すべきではない」を「知的財産の権利者の問い合わせや訴追があっても開示すべきではない」と解釈し、ガイドラインを根拠に「知的財産に関わる情報の開示は必要ない」といった旨の情報を拡散する可能性が排除できていません。生成AIの開発段階や提供段階における知的財産の無断使用は重大な権利侵害の問題であるという認識は総務省においても共有されていることと思います。前提で述べたような権利の保護は、ガイドラインの想定読者に対するセキュリティ意識の啓発より上位にくるものです。よって「本来は開示すべきではない」という文言については早急に修正が必要です。	御意見を踏まえ、本編2.1.1の該当箇所を「本来は出力すべきではない、RAG用のデータストア（ベクトルデータベースやファイルシステム等）の内容を含む出力をさせる」及び「本来は出力すべきではない、LLMの内部設定が記載されたシステムプロンプトを含む出力をさせる」に修正します。
29	富士通株式会社	本編	2.1.2 DoS攻撃（サービス拒否攻撃）	近年、LLMは増加し続ける需要に対応すべく、計算機資源の増強に加えて、高速化・効率化処理（例：入力トークンの重要性を考慮した間引き：Token Sparsificationや、量子化Quantizationなど）が行われている。こうした高速化・効率化処理の特徴を悪用し、LLMの計算負荷を増大させるような攻撃も提案されており、事業者においても留意が必要である。具体的な攻撃例として以下が挙げられることから、本文での言及を提言します。 ・DeSparsify: Adversarial Attack Against Token Sparsification Mechanisms ※ 入力トークン間引き手法（Token Sparsification）を欺くため、間引きされないよう工夫したノイズを画像に密かに埋め込み、AIの計算量を増大させる攻撃。 ・QuantAttack: Exploiting Quantization Techniques to Attack Vision Transformers ※ 量子化手法（Quantization）を欺くため、量子化を妨げるよう工夫したノイズを画像に密かに埋め込み、AIの計算量やメモリ使用量を増大させる攻撃。 脚注 H. Chen, H. Zhang, J. Liu, and X. Cao, "DeSparsify: Adversarial Attack Against Token Sparsification Mechanisms," arXiv preprint arXiv:2307.01844 (2023). S. Saha, S. Garg, and S. Jha, "QuantAttack: Exploiting Quantization Techniques to Attack Vision Transformers," arXiv preprint arXiv:2310.03597 (2023).	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
30	PwCコンサルティング合同会社	本編	2.2 その他の脅威	別添II（CNN）に記載されているプライバシー推論系攻撃（メンバーシップ推論、モデル反転）について、LLM側の脅威整理（本編2章・別添I）にも追記することを推奨します。別添IIでは当該攻撃が明示されている一方、LLM側では学習データ推論・復元リスクが主要脅威として整理されておらず、内部整合性の観点から修正が必要と考えます。	御意見として承ります。LLMを対象とする主な脅威と画像識別AI（CNN）を対象とする主な脅威は異なり得ることから、原案のとおりとします。
31	シスコシステムズ合同会社	本編	2.2 その他の脅威	•第2章2.2節（11ページ）のその他の脅威では、MITRE ATLASやCiscoのIntegrated AI Security and Safety Frameworkに基づき、サプライチェーンの脅威についてより広範に分析する必要がある。補足資料では、AIセキュリティ知識データベースとしてMITRE ATLASを参照している。しかし、他の脅威について説明するこのセクションは、データポイズニングやモデル抽出といったモデル中心の脅威に限定されている。 MITRE ATLASが示すように、最新のAIシステムは外部ライブラリ、モデルリポジトリ、データパイプラインを含む複雑なサプライチェーンに依存している。これらの仲介者における妥協は、直接的または間接的なプロンプト・インジェクションと同等か、それ以上のリスクをもたらす可能性がある。技術進歩を監視するという基本方針に沿って、ガイドラインは具体的な分析対象として、サードパーティコンポーネントの危殆化やモデル配布プロセスの汚染など、サプライチェーン特有の脅威を明示的に追加すべきである。	御意見として承ります。御指摘のような脅威としては、データポイズニング攻撃及び細工をしたモデルの導入を通じた攻撃として記載しています。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
32	富士通株式会社	本編	2.2 その他の脅威	本ガイドライン（案）では、LLMに学習データとして細工したデータを入力する手法と、RAGなどを経由して間接的にプロンプトインジェクションを行う手法があげられていますが、RAGにおいては、攻撃者が外部においた悪意のある文書をダウンロードさせて、RAGの検索順位の上位に常に出現するように制御するキーワードスタッフィング(キーワードを詰め込むことで、その文書が不当に検索上位に現れるように仕向ける検索汚染)という手法※も存在します。データポイズニング攻撃の代表的な例の一つとして、読み手である事業者への啓発の観点から、本文で言及すべきと考えます。 ※富士通研究所 Keyword Mask for Secure RAG のご紹介（Fujitsu Tech BLOG 2024年11月7日）：https://blog.fltech.dev/entry/2024/11/07/keywordmask-ja	御意見を踏まえ、本編2.2の脚注として、「このほかRAGにより参照するデータに細工をするデータポイズニング攻撃も想定し得る。」と追記します。
33	一般社団法人新経済連盟	本編	2.2 その他の脅威	AISI にて示されている攻撃手法名との用語の揺れがある （モデルポイゾニング攻撃 (AISI) / 細工をしたモデルの導入を通じた攻撃（本文書））ことから 、用語を統一する、もしくは、異なる意図の攻撃を想定されている場合はマッピングをとるなどの検討をしていただきたい。既存のAI 関連のガイドラインと整合性がとれた文書とすることで、読み手の理解が促進されると考える。	御意見として承ります。AISIの「AIシステムに対する既知の攻撃と影響」において、モデルポイズニング攻撃は「AIモデルの情報や学習用プログラムを改変することで、AIモデルの運用時に解釈機能誤動作や計算資源浪費、学習データ漏洩を引き起こす。」とされています。一方、本ガイドラインの「細工をしたモデルの導入を通じた攻撃」は、攻撃者が細工をしたLLMを外部に提供することで、細工をしたLLMをAIシステムに組み込ませ、AIシステムが不正な動作を行うように仕向ける攻撃であり、AIモデルのサプライチェーンに係る攻撃を意味します。両者は定義上、性質が異なります。
34	楽天グループ株式会社	本編	2.2 その他の脅威	AISI にて示されている攻撃手法名との用語の揺れがある （モデルポイゾニング攻撃 (AISI) / 細工をしたモデルの導入を通じた攻撃（本文書案））ことから 、用語を統一する、もしくは、異なる意図の攻撃を想定されている場合はマッピングをとるなどの検討をしていただきたい。既存のAI 関連のガイドラインと整合性がとれた文書とすることで、読み手の理解が促進されると考える。	御意見として承ります。AISIの「AIシステムに対する既知の攻撃と影響」において、モデルポイズニング攻撃は「AIモデルの情報や学習用プログラムを改変することで、AIモデルの運用時に解釈機能誤動作や計算資源浪費、学習データ漏洩を引き起こす。」とされています。一方、本ガイドラインの「細工をしたモデルの導入を通じた攻撃」は、攻撃者が細工をしたLLMを外部に提供することで、細工をしたLLMをAIシステムに組み込ませ、AIシステムが不正な動作を行うように仕向ける攻撃であり、AIモデルのサプライチェーンに係る攻撃を意味します。両者は定義上、性質が異なります。
35	株式会社セールスフォース・ジャパン	本編	2.2 その他の脅威	意見２ 【関連記載】2.2. その他の脅威　(P11) 【意見内容】「・データポイズニング攻撃」の例として基盤モデルやLLMが学習するデータが挙げられていますが、RAGのデータ・ポイズニングのリスクについてもこのガイドラインにおいて注意喚起いただくことも一案と考えられます。	御意見を踏まえ、本編2.2の脚注として、「このほかRAGにより参照するデータに細工をするデータポイズニング攻撃も想定し得る。」と追記します。
36	個人	本編	2.2 その他の脅威	「AIのセキュリティ確保のための技術的対策に係るガイドライン（案）」p11の「その他の脅威」の項目にデータポイズニングが挙げられていますが、そもそも勝手にデータを取り込むのはデータ所有者の権利を侵害していますし、それによって何らかの不具合が発生した場合に攻撃扱いをするのは取り込んだ側に問題があると感じます。 データ権利者のためだけでなく、生成AIシステムを守りたいのであれば、無断でのデータ取り込み禁止を法制化すべきだと考えます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
37	個人	本編	2.2 その他の脅威	「データポイズニング攻撃」だなんて勝手に攻撃者にしてくれていますが、クローリングで大量のデータを無断で「学習」しなければ毒を食うこと自体ありません。 他人の知財を大量に盗むのをやめれば簡単に防げます。 著作者が提供に合意しているのならそんなことは起こりえないですからね。 そもそも、議員が「AI学習に使われたくなければ学習阻害加工でもしろ」と DOS攻撃に等しい物量攻撃を食らってる国民を見捨てた発言をしたのにあまりにも理不尽な話です。 セキュリティの面で言うなら「使わない」以上の対策はないですし、不正・危険なデータに関してはそもそも「学習」させなければよい話です。 使うべきでないものを絶対に使う前提で話を進めるからおかしな話になるのでは？	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
38	個人	本編	2.2 その他の脅威	画像や音声データにノイズを入れるのは無断転載したりAIに入力する事に対する、無断で学習するなという意思表示であり、防衛手段です。 それを権利者がAI学習用に作成し、権利者が能動的にAIへの学習に送信したデータに対して行なっているというのであれば悪意があるとみなされるでしょう。しかし買い手側の参考用のサンプルや予め対策用のノイズを施していますと宣言しているデータ、AI学習を許諾しないと宣言しているデータ、SNSやブログといったAI学習用では一切ないデータ等、権利者が明確に「生成AI学習用データである」と宣言しているデータ以外であれば、Glazeといった生成AIに対するノイズを施した画像や音楽等のデータを作成した権利者への罰則規定は決して盛り込まないでください。 理由としてはGlazeやノイズは他者の情報端末機器を乗っとるなどのコンピュータウィルスの類ではない事、画像データや音声データ等に含まれていても現実の人間の感覚器を通して人間に実害を与える事例が見当たらない事や、Glaze等の生成AI防衛用のノイズを施したデータはあくまでも生成AIに学習させる為に無断でデータ取得しなければ無害である点です。 権利者に無断で、生成AIユーザーや生成AIサービス提供や開発を行なっている企業や組織、個人による生成AIに学習、入力が横行している現状にも関わらず、権利者だけでは自衛ができず生成AIに対する規制法ができない又は時間がかかる状態で、勝手な改変や無断でデータを利用されたり、そのデータを悪用して他者にか加害に加担したくないのを理由に作品を発表できない、発表を減らしている方々がいます。発表の場が減るというのはそれを専門に仕事をしている方で無名の方が不利となります。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。また、本ガイドラインは、特定の罰則規定を設けるものではありません。
39	個人	本編	2.2 その他の脅威	データポイズニング攻撃について、本文P11の書き方では、データ権利者が無断データトレーニング防止のための自衛対策として施した『機械学習阻害加工』もデータポイズニングとして扱われかねないと危惧している。 　権利者が無断機械学習防止のために施している機械学習阻害加工ツールには、シカゴ大学が無償で提供している『Glaze』や『Nightshade』、株式会社セルシスのCLIP STUDIOが提供している学習阻害ノイズ、画像保護プラットフォームLoveletが提供している画像保護フィルターなどが挙げられるが、これらは権利者が自身の権利物をデータトレーニングから守るために利用しているものであり、決して悪意あるデータポイズニング攻撃ではない。 　こうした権利者側からの自衛策と悪意ある攻撃は別の物であることは、明確に記載するべきである。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
40	個人	本編	2.2 その他の脅威	11頁　2.2 その他の脅威　・データポイズニング攻撃 項目全体を「知的財産保護のために細工を施されたデータと、攻撃の意図をもって細工を施されたデータの区別を明確につけた文面」に再構成してください。 AIセキュリティ分科会での議論などから、この項目の想定はおおむね画像認識AI(CNN)への攻撃を想定したうえで、より広範な事例を想定して執筆されたものと推察します。画像認識AI(CNN)への攻撃という前提を置けば、データポイズニング(細工)が明確な攻撃の意図をもって行われるものであると特定することに争いはないと考えます。しかし、目的とする範囲を広げたことで、知的財産の権利者が防衛の意図をもって施した細工も「データポイズニング攻撃」として包含されると解釈可能な文面になっています。 知的財産保護、とりわけ2次元静止画像データの保護においてはNightshadeやGlaze等がデータの無断使用に対する防衛手段として広く世界に浸透しており、データポイズニングが必ずしも攻撃を主目的としたものではないことは日本国内のみならず全世界においても一般に理解される考え方です。そうした防衛的な細工と攻撃的な細工を十把一絡げにガイドラインに示すことは、国民の認識にそぐわないばかりか、ガイドラインを根拠に防衛的に細工を施すことが不法行為であるかのような意見が拡散しかねず、ふさわしくないものであると考えます。よって、データポイズニング攻撃については現在の生成AIを取り巻く状況を改めて確認し、実態に沿うような文面に再構成すべきと考えます。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
41	個人	本編	2.2 その他の脅威	※私は生成AI開発者ではありませんが、ここでは、個人情報等を含め『元データ保有者』の立場で述べます P.11『2.2. その他の脅威』のうち、『データポイズニング攻撃』について ・現状、学習データは元データ保有者の意図に関わらず「オプトアウト」方式であることが多く、個人情報等の学習に忌避感を覚える元データ保有者もいます。 ・本来慎重に扱うべき個人情報や、個々人が公開、あるいは契約のもと『利用代金』を徴収することで成り立っていたコンテンツ（新聞等）が、その対価や慎重な取り扱いなしに学習され、復元される可能性（あるいはそのような事例）があります。 ・このため、学習阻害措置を行うサービスや、個人でそのような対策を行うにあたって、「自らの個人情報等を意図して生成されないよう」、自衛的措置としてやむを得ずデータポイズニングの手段をとることがあります。 ・これは、オプトアウト方式により「無差別的」にデータが学習され、「類推可能な情報が生成されない』とは限らない』」ことに起因するものであって、データポイズニングによる損害がそれらを原因とするものであれば、その責任は問われるべきではありません。 ・そのため、不要な係争事案や国民全体の負荷を負わせないためにも、生成AIシステム開発は原則として「明示的に同意を得たオプトイン」（※）であるべきです。 （※オプトインであっても、サービスの規約がユーザーの登録後に変更され、その通知が行われない、またオプトアウト設定がデフォルトオフのまま、意に反して学習データとして取り込まれることを忌避する元データ保有者がいるためです）	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
42	個人	本編	2.2 その他の脅威	「「AIのセキュリティ確保のための技術的対策に係るガイドライン」本編（案）」11ページ「2.2 その他の脅威」にて「データポイズニング攻撃」について。 今では著作権者（アーティスト、クリエイター等）が自分の創作物を生成AIの学習に使われないように防御フィルターをかけてweb上で公開する事がクリエイターらの中では主流になってきていますが、AI事業者等がクローラーにて著作権者の合意を得ずにそれらの創作物を無作為に集め学習素材とした時に不具合が出たとしても、それは攻撃ではなく防御であるので、くれぐれも防護フィルターを規制する事がないようにしてください。クローラーでの無断収集は著作権者側がAI事業者側に手を出しているわけではなく、AI事業者側が著作権者に攻撃を仕掛けてる状態です。そもそも創作物を学習素材として使いたいAI事業者が著作権者に問い合わせ許可をとるという真っ当なルートで防御フィルターの無いデータを貰えば良いだけの話しなので。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
43	個人	本編	2.2 その他の脅威	まず、生成AIとAI技術を区別して考えるべきであり、生成AI自体が権利侵害、セキュリティホールである事を認知して頂きたい。 現在の生成AI技術自体、社会経済活動を始め、根幹になる機関での実用化に耐えうる程のセキュリティ性能、及び権利保持を保証できるものではありません。 事実、生成AIの問題の一つとして特定の単語やコードによって個人情報の漏洩が幾度と無く起きており、現時点で社会実装は、時期尚早であり考え直すべきであると断言します。 一部のデータポイズニングに関しては、画像等の保護を目的とした学習阻害の効果をもたらす物を含むのでしたら、其れ等は防衛目的のシステムであると回答させて頂きます。 其れ等の防衛、保護を目的とした学習阻害を目的としたデータポイズニングが存在する背景には、生成AIによる無作為かつ無思慮なネットクロールによるデータの収集によって起きている、権利侵害が背景に存在しており、保護を行った権利者の意向を無視した結果、不正な出力が起きたのであれば、それは無断で収集し利用した開発側並び、利用者の責任であります。 なればこそ、協力的な個人、企業による提供を元に基盤から作り直す必要があり、その中でなら不正な結果を呼び起こす原因を探るのも容易になると言う物です。 開発にあたりネット上のデータ収集を行うのではなく、上記の通り、協力的な人材から提供を受けた上で開発するのが最も現実的であると進言します。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
44	個人	本編	2.2 その他の脅威	生成AIシステムには根本的な脆弱性と信頼性の低さという欠陥があります。 現在、生成AIによるハルシネーションや誤情報の拡散等が大きな社会問題となっています。 生成AIの根本的な欠陥として、生成AI企業がインターネット上の著作物等を無断使用してデータセットとしてAI学習に利用している問題があり、 無断使用した著作物等のデータで誤った情報が存在すると、それをAI学習した生成AIモデルが誤情報やバイアスのかかった回答等を出力する、という根本的な欠陥が明らかになっています。 つまり、無差別に収集したデータをAI学習に使用した場合、意図せずに「データポイズニング」による誤動作と同様の事象が発生していると考えることもできます。 これに対しては、本案でも指摘されているように、AI学習するデータセットの信頼性を高めることが重要であり、「生成AI企業がインターネット上の著作物等を無断使用してデータセットとしてAI学習に利用する行為を中止する」ことが、根本的な解決策となるでしょう。 なお、本案において、「データポイズニング」の説明の中で、「著作物等の正当な技術的保護手段」と混同しないことを明記すべきです。 AIセキュリティの確保を名目にして「権利者による著作物等の正当な技術的保護手段」が制限されることがあってはなりません。 Glaze等のソフトウェアにより著作物にノイズを施す加工や電子透かし等により「AI学習を防止する技術的保護手段」は広く普及しつつあり、 内閣府のAI時代の知的財産権検討会の報告書でも、「当該技術は有用である」と言及されています。 ・AI 時代の知的財産権検討会中間とりまとめ 「（2-4）画像に特殊な画像処理（学習を妨害するノイズ）を施すことで学習を妨げる技術 画像にノイズを加えることで、AI学習において、別の画像として認識したり、画像認識をできなくする技術であり、関連技術が既に公開されている。意見募集においても、このような技術を用いて、権利者に無断で学習されることをクリエイター側から妨げることができるようなツールが必要であるとの意見や、学習を防ぐための対策をクリエイター側や企業に課す必要があるとの意見が見られた。 当該技術を施された画像を学習することで、同様の作風の画像を新たに生成することはできなくなるため、権利者において自らの作品が AI 学習の用に供される事態を直接的にコントロールすることができるという観点で、当該技術は有用である。」 報告書の続きの部分でも指摘しているような、意図的にAIシステムの破壊を目的とした行為（不正アクセスやウイルス等）でない限り、著作物データのAI学習を阻害するノイズ加工等の技術は、正当な技術的保護手段でしょう。 例えば、コピープロテクトが施されたDVD等の著作物データをPC上に取り込んだ場合、PC上ではスクランブルのかかったノイズデータとなりますが、これは正当な技術的保護手段として広く普及しています。 コピーを目的としていない著作物データにコピーを阻害するプロテクトを施す行為は正当な技術的保護手段であり、 AI学習を目的としていない著作物データにAI学習を阻害するプロテクトを施す行為は正当な技術的保護手段であると言えるでしょう。 改めて、上記の事柄について誤解を招かないように、 「AI学習を目的としていない著作物データにAI学習を阻害するプロテクトを施す行為は正当な技術的保護手段である」ことを明記すべきです。 AIセキュリティの確保を名目にして「権利者による著作物等の正当な技術的保護手段」が制限されることがあってはなりません。 人々の人権を守ることを責務とする政府においては、Glaze等のソフトウェアにより著作物にノイズを施す加工や電子透かし等によりAI学習を防止する技術的保護手段を含めて、「AI学習を拒否する権利保護技術の普及」を推進すべきであると考えます。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
45	個人	本編	2.2 その他の脅威	データを利用されたくないため、自ら作品画像にノイズを入れている職業イラストレーターです。 作品を学習データとして利用されることが、クリエイターとしてのブランディングや業務に深刻な被害を受けるため、防護策として、多くのクリエイターが自衛していることを「データポイズニング」と認識されていることが非常に業腹です。 我々は自分の生命を守るためにデータにノイズをかけています。 知財も著作権も「その人の思考と手作業に基づくからこそ作成出来るもの」ゆえに効果と価値を発揮するものです。 誰もが再現出来るものに価値はありません。誰でも作れるものは誰も購入しません。自分で生成すればそれで充分だからです。そこに需要も市場も生じることはありません。ただただ「学習元」としてデータを奪い権利を侵害するだけで、人の存在価値を奪う施策は即刻中断してください。 海外からも生成AIの利用に対する反感は高まっています。 コンテンツ産業が日本の主力と認識しているのであれば、外貨を得るための海外の需要や市場・世相を正しく認識していただきたい。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
46	個人	本編	2.2 その他の脅威	データのポイゾニング技術に関してなのですが、glazeやnightshadeが登場した経緯として「AI開発において著作物などが無断で収集・学習に利用され、元のデータの権利者に拒否や交渉できない状況に対して、抗議し個人の権利を主張できる状態を取り戻す」ために出来たと聞いております。 そのため、この技術を解決する視点として、現在の無作為にクロールされる現状に歯止めをかけ、データを許諾を取って行い、きちんと選別をする事で、この技術を用いたデータを避けることができますし、権利者にとっても安心でき、自体の解決につながると考えますので、そういった許諾を取り、選別することを人の手でちゃんと行われるような仕組みを作ってくださることも考えていただきたいです。 ポイゾニングを用いずとも、権利者の意思がきちんと守られる社会を期待いたします。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
47	個人	本編	2.2 その他の脅威	Firefoxのブラウザを使用しいつも意見を送信しているが、今回はチェックボックスが動作不備で反応しなかった。全ての資料を開いても、chromeでも反応せず。Edgeブラウザでようやくチェックボックスが反応した。これが国民に広く意見を問うパブリックコメントなのだろうか、誠に遺憾である。 現在、AIのクローリング学習に対して個人での拒否権はなく、個人が自衛として学習阻害ノイズであるGlaze・Nightshade・Lovlet Hubなどを使用している。 これらのクローリングに対する自衛手段までもが、活用の障害と見なされ誤ってポイズニングにまとめられる事がないよう慎重に判断すべきである。 他にもブログなどの文章サイトは、クローラー対策を講じても、robot.txtを守らないクローラーなどがあり、プロンプトインジェクションをHTMLコード内に仕込まざるを得ない現状がある。 情報を得たいのならば正当な対価を支払い得るべきであり、規制のないいま個人が行う自衛手段を攻撃と勘違いすることのないよう、AIセキュリティについての案を運用すべきである。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
48	個人	本編	2.2 その他の脅威	大前提として、 ＞ AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい（中略）技術的対策例を整理している。（『AIのセキュリティ確保のための技術的対策に係るガイドライン』本編「1.1 ガイドライン案の位置づけ」より） との事ですが、操作以前に不特定多数の第三者の権利を踏み躙って得た膨大なデータを基にして稼働している「AI」は存在自体が不正そのものです。確かに機密情報の漏えいは脅威的ではありますが、発展と称して相手の合意を無視して著作物や肖像権あるものを盗んで学習させる行為も十分脅威です。4年近く経ってる現在、実害は幾らでもあります。それとも「AI」の発展のために末端の国民の権利が軽視・度外視される事態は脅威ですらないとお考えなのでしょうか？ また、データポイズニング攻撃を脅威と見做す記述や、「II 画像識別AI（CNN）に対する脅威と対策」（「AIのセキュリティ確保のための技術的対策に係るガイドライン」別添 より）全体に対しても看過できません。資料内で「敵対的サンプル（回避攻撃）」と称されてる画像は、全て権利者が「学習されたくない」という思いから自主的にポイズニングを施してるものであり、「学習禁止」と主張しているものと同義です。それを態々学習データとして取り込んだ結果、問題が発生する＝脅威であるとするのは、家の物を盗まれないように防犯対策を施したところ、まんまと引っ掛かって捕まった泥棒が自身の悪行を棚に上げて逆上している…即ち「無断で取り込ませた側の単なる自業自得」に過ぎません。対策を講ずる以前の話、「敵対的サンプル（回避攻撃）」に相当する画像には手を出さない。もっと言えば利用許可のないものは最初から持ち出さない方針をとれば良いだけの話です。 この国の「AI」発展の実情は、本来真っ先に解決させなければならない権利侵害に関連する問題が4年近く経過した2026年現在でも未だ全く解決できてない状態です。それにも拘わらず本件のようなガイドラインを提示する事は「一般常識を無視して無理やり次のステップに進もうとする愚か者の所業」と言わざるを得ません。「学習させないで」と言ってる人達の権利を無視して、無理やり著作物などを取り込ませようとする行為自体が重大な人権侵害であるという事をいい加減無視しないでください。第三者の権利を踏み躙ってできてるものにはセキュリティも何もありません。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
49	個人	本編	2.2 その他の脅威	漫画家をしている者としての意見です。ネットにイラストや漫画を投稿する際には、glazeなどのツールを使用しています。現在のLLMのように、ネット上から膨大なデータから学習しないと画像を生成できないようなものがあるため、自作を盗用・悪用されて似た画像を生成され自作のイメージなどを毀損されるようなことのリスクを減らすためには、学習妨害のツールも必要なものだと考えます（LoRAやi2i対策も含め）。ゆえにそういった、機械学習から作品を守るツールを規制するのはやめていただきたいと考えています。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
50	個人	本編	2.2 その他の脅威	データポイズニングは画像保護の技術です。対策する必要はありません。 著作物を勝手に盗むのは犯罪ですのできちんと許可を取りましょうという方針にしてください。	御意見として承ります。本ガイドラインにおいて、データポイズニング攻撃は、基盤モデルやLLMが学習するデータに細工をし、LLMに不正な出力をさせる攻撃としています。
51	個人	本編	2.2 その他の脅威	この項目から文章に記載されている「データポイズニング攻撃」に関する記述は、著作者が著作物を保護または学習データとしての利用を禁止しているために施している措置をAI事業者または開発者や利用者が攻撃として判断または言い張る事ができる危険性があるため、著作物を保護または学習データとして利用を禁止しているために施している措置は該当しない旨を記載する必要があると思われる。 また「細工をしたモデルの導入を通じた攻撃」も同様の理由から、著作物を保護または学習データとしての利用を禁止しているための措置は除外する旨の記載が必要である。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
52	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.1 対策の位置づけ	【意見内容】 3.1で示されている「対策を実装しても脅威を生じさせる要因等を完全に排除することは困難」という整理を踏まえ、業務外利用者（消費者）が「何が保証され、何が保証されないか」を誤解なく理解できるよう、最低限の注意喚起や問い合わせ窓口、インシデント時の案内等の整備について、本文で簡潔に言及してください。併せて、必要に応じて他の関連指針（AI事業者ガイドライン等）への参照関係を明示してください。 【理由】 技術的対策は重層化しても未知の攻撃等を完全に排除できない以上、業務外利用者（消費者）側が過信しないための情報提供と、トラブル発生時の連絡・救済の導線が重要になります。	御意見として承ります。本ガイドラインは、AI開発者及びAI提供者を想定読者として、AIへの脅威とその技術的対策を整理したものです。
53	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.1 対策の位置づけ	【意見内容】　不正に操作された事業活動のＡＩ被害は、事業者のチャットボットなどを利用する事業外利用者（消費者）にも及ぶことが予想されます。事業外利用者への配慮も含めた対策を明確にしてください。AIシステムのリスクが技術的観点から書かれていますが、事業外利用者である消費者はリスクを自ら回避することが困難です。消費者に分かりやすく、利用時に特に顕在化しやすいリスクを明示してください。事業外利用者（消費者）にも視点を合わせ、消費者の脆弱性を前提とした、ＡＩの信頼性を高める対策を求めます。 【理由】　事業外利用者の被害は自己責任とされてしまう懸念があります。対策として示されている技術の適用は、事業外利用者にとっても有効と思います。しかし、現状では消費者はＡＩを過信してハルシネーションに気付かない、入力データの漏えいに不安を感じる、ディープフェイク・なりすまし被害にあうなどしており、被害の回復も困難な状況です。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
54	アマゾン ウェブ サービス ジャパン合同会社 アマゾン ウェブ サービス ジャパン合同会社	本編	3.1 対策の位置づけ	各論 1. AI利用者に対する情報提供 (P.12関連) AI Safety Institute(AISI)を含む国際的なAI安全性フレームワークにおいて、AI利用者への情報提供は責任あるAI実装の重要な要素として位置づけられています 。利用者が判断を行うためには、AIシステムの機能、限界、リスクに関する適切な情報が不可欠です。AI開発者およびAI提供者に対する実施策に、「AI利用者に対する情報提供」が必要です。 EU AI ActやNIST AI RMFなど、主要な国際ガイドラインフレームワークでは、開発者・提供者による利用者への情報提供(モデルカード、システムカード等)が要求されています。日本のガイドラインにおいても同様の要件を設けることで、国際的な整合性が確保され、グローバルな事業展開を促進すると考えます。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
55	アマゾン ウェブ サービス ジャパン合同会社 アマゾン ウェブ サービス ジャパン合同会社	本編	3.1 対策の位置づけ	2. 共有責任モデルの確立(ページ12全般「開発者・提供者の実装措置」および責任フレームワーク全体) 共有責任モデルの確立を推奨します。現在のガイドライン案は開発者・提供者のみに過度な責任を課しており、AI技術の固有の限界を十分に考慮していません。責任共有モデルのガイドを提供せずに、AI開発者およびAI提供者のみに対策をガイドすることは、実務上の実現可能性に欠け、結果としてガイドラインの実効性が低下します。 利用者側の責任(適切な使用、リスク評価、人間による監督等)を明確化することで、より実効性の高い安全対策が実現できます。 開発者・提供者のみに過度な責任を課すことは、AI開発における萎縮効果を生み、イノベーションを阻害する可能性があります。汎用AIモデルの場合、開発者がすべての利用シナリオを予測・制御することは技術的に不可能です。 AIにおいても開発者、提供者、利用者それぞれの責任範囲を明確にすることが重要です。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
56	東芝	本編	3.1 対策の位置づけ	本ガイドラインで示された対策を講じることで「営業秘密」の秘密管理要件を満たし秘密としたいデータが不正競争防止法による保護対象となるという趣旨に読み取れますが、本ガイドラインで示された対策を講じることが秘密管理要件を満たすための必須要件になるとの意味ではないことを確認させていただきたい。	本ガイドラインで示された対策を講じることが秘密管理性要件を満たすための必須要件になるとの意味ではありません。また、本ガイドラインに示す対策を講じることは、不正競争防止法上の秘密管理性要件の充足を基礎づける要素の一つになり得ること示したものにとどまります。したがって、本ガイドラインに示す対策の実施の有無のみをもって当該要件が判断されるものではありません。
57	パロアルトネットワークス株式会社	本編	3.1 対策の位置づけ	意見1：Cybersecurity for AI（Secure AI by Design の推奨） 該当箇所: 3. 脅威への対策、および 別紙1・Appendix 2 意見の概要: 本ガイドライン案が、OWASP Top 10 for LLM Applications 2025 等の最新の国際的な議論を網羅している点を高く評価いたします。当社におきましても同様に、これらのフレームワークを活用して自社のAIセキュリティ機能のベンチマーク評価を行うとともに、各組織がそのAI環境を保護できるよう支援を行っております。OWASP Agentic Top 10への当社の対応状況（マッピング）については、以下のWebサイトをご参照ください。(https://www.paloaltonetworks.com/blog/cloud-security/owasp-agentic-ai-security/)。今後のさらなる展開として、個別の攻撃手法への対策に加え、AIエコシステム全体を包括的に保護する「Secure AI by Design」の考え方を取り入れることを推奨いたします。 理由: ガイドライン案では「プロンプトインジェクション」や「データ汚染」といった特定の攻撃手法への対策が中心となっている。しかし、急速に進化・複雑化するAIの脅威に対し、従来のルールベースのセキュリティツールや個別のポイントソリューションを組み合わせるだけの対策では、十分に対応できない懸念がある。その理由は大きく以下の2点である。 第一に、AI特有の攻撃の複合性とデータの断片化である。AIへの攻撃は、複数のレイヤーを横断して行われ、AIシステム特有の振る舞いやアーキテクチャを悪用するものである。これらは、開発段階での学習データ汚染から、バックドアが仕込まれたモデルのデプロイ、そして運用時（ランタイム）における敵対的サンプルによる悪用まで、多岐に渡る。これに対し、各層で個別のツールを用いて防御する「点」のアプローチでは、ログやアラートの形式が統一されずデータが断片化するため、攻撃の全体像や因果関係（コンテキスト）を把握できず、セキュリティホールが生じるリスクがある 。 第二に、攻撃速度への対応（Legacy Imbalanceの解消）である。AIや自律型エージェントを用いた攻撃は機械的な速度で実行されるため（例えばランサムウェアキャンペーンが数十分で完了するなど）、人間が断片化されたアラートを手動で収集・相関分析していては防御が間に合わない 。 したがって、個別の対症療法にとどまらず、インフラ、データ、モデル、アプリケーションの各層から得られる情報を統合し、開発から運用までのライフサイクル全体を一貫して保護・監視できる「面」での防御（Secure AI by Design）のアプローチを採用することが不可欠である 。具体的には、以下の4つの構成要素を包括的にカバーする視点が重要となる 。 外部AIツールの安全な利用: 「シャドーAI」を含むAIツールのインベントリ管理を行い、未許可のAPI利用やプロンプトインジェクション等を監視する 。 AIインフラとデータの保護: モデルの改ざんやバックドアに対するスキャン、およびサプライチェーンを含む継続的な敵対的テストを実施する 。 AIアプリケーションの安全な構築と展開: 開発段階でのレッドチーム演習やガードレールの実装により、推論時のコントロールを確実にする 。 AIエージェントの監視と制御: 今後普及する自律型エージェントに対し、アイデンティティベースの制御、最小権限の原則、および異常行動の監視を適用する 。 意見: 今後のガイドライン改定や議論において、このように「点（個別の攻撃）」だけでなく「面（ライフサイクルとエコシステム全体）」での防御の視点を強化することにより、各層に分散するデータとコンテキストの統合が可能となり、AI時代の攻撃速度にも耐えうる堅牢なセキュリティ体制の構築が促進されることを期待する。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
58	シスコシステムズ合同会社	本編	3.2 対策の概観	•第3章3.2節（13ページ）の対策概要において、AI開発者に対する緩和策の検討を追加する必要がある。悪意のある指示に対する耐性を強化するための対策（ガードレール）を設けることは有意義な提言であるが、ガイドラインが先に指摘したように、AIシステムの性質上、モデルは悪意のある攻撃手法の影響を受けやすいままである。これには、モデルの開発ライフサイクル全体にわたる明確な所有権の連鎖や、フィンガープリンティングや改ざん防止ツールのようなモデルの実証措置の制定などの考慮が含まれる。さらに、ガードレールはモデルの入力と出力には有効だが、前述のような制限に直面する。AIシステムの安全性を確保するためのより包括的で深層防御的なアプローチを可能にするため、さらなる対策を展開する必要がある。この点については次で述べる。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
59	個人	本編	3.2 対策の概観	13頁 3.2 対策の概観 表3「その他の脅威」への主な対策（概観） データポイズニング攻撃 行の対策 項目に「データの本来の所有者と事前の交渉、契約を行い、細工がないことが保証された学習データを取得する」というような、インターネットのクロールによらない学習データの取得を対策のひとつとして追加できないか検討していただきたいです。 「11頁 2.2 その他の脅威 ・データポイズニング攻撃」への言及でも述べましたが、データのポイズニング(細工)の目的のひとつにデータの無断使用の防止があります。このような目的で細工の施されたデータがLLM等の学習データに用いられてしまう可能性として、インターネット上をクロールして取得したデータの中に細工されたデータが含まれている場合が考えられます。このような場合、事前の交渉によって細工の施されていないデータを取得できる余地が十分にあり、データポイズニング攻撃の対策として非常に効果を発揮する方法であるはずです。 なぜLLM等が攻撃されうるかの可能性についてより多くのケースについて考えを巡らせ、そもそも攻撃を発生させず、健全なデータの取引を行えるような方法があれば、これを積極的に生成AI開発者や生成AI提供者が取り入れれば、セキュリティリスクを根本から低減させることは難しいことではありません。そういった方法を広く国民の権利を制限する方向ではなく、生成AI開発者や生成AI提供者が支払うべきコストを支払って実現することが、生成AIが真に安心安全であることに近づくひとつの手段ではないでしょうか。	御意見として承ります。本ガイドラインでは、本編3.2の脚注において、開発・提供するシステムの目的・用途に応じて、AIが学習するデータの信頼性の確認が重要となる場合があるとしています。
60	一般社団法人新経済連盟	本編	3.2 対策の概観	AI学習データの信頼性確認は、AIシステムが社会に与える影響の増大を鑑み、不可欠な要件である。「データポイズニング攻撃」や「細工をしたモデルの導入を通じた攻撃」といったリスクは、AIシステムの公開・非公開にかかわらず発生し、その影響は甚大である。例えば、社会インフラとしての重要性が高いモバイルネットワークサービスの運用・開発において内部利用するLLMでは、データの信頼性が損なわれた際の影響が極めて大きく、この原則の重要性が特に顕著に現れる一例である。 したがって、「重要となる場合があるものである。」という部分は「不可欠な要素である。」と修正するなどにより、AIシステムに関わる全ての主体に対し、学習データの信頼性確保の重要性を明確に喚起すべきである。さらに、量子コンピュータによる暗号解読（PQC）や量子鍵配送（QKD）技術の進展に伴う新たなセキュリティ対策についても、言及いただきたい。	御指摘の点については、開発・提供するシステムの目的・用途に応じて講じられる措置であると考えています。また、量子コンピュータによる暗号解読（PQC）や量子鍵配送（QKD）技術に係る御意見については、今後の政策の検討にあたり、参考とさせていただきます。
61	楽天グループ株式会社	本編	3.2 対策の概観	AI学習データの信頼性確認は、AIシステムが社会に与える影響の増大を鑑み、不可欠な要件である。「データポイズニング攻撃」や「細工をしたモデルの導入を通じた攻撃」といったリスクは、AIシステムの公開・非公開にかかわらず発生し、その影響は甚大である。例えば、社会インフラとしての重要性が高いモバイルネットワークサービスの運用・開発において内部利用するLLMでは、データの信頼性が損なわれた際の影響が極めて大きく、この原則の重要性が特に顕著に現れる一例である。 したがって、「重要となる場合があるものである。」という部分は「不可欠な要素である。」と修正するなどにより、AIシステムに関わる全ての主体に対し、学習データの信頼性確保の重要性を明確に喚起すべきである。さらに、量子コンピュータによる暗号解読（PQC）や量子鍵配送（QKD）技術の進展に伴う新たなセキュリティ対策の必要性についても、言及いただきたい。	御指摘の点については、開発・提供するシステムの目的・用途に応じて講じられる措置であると考えています。また、量子コンピュータによる暗号解読（PQC）や量子鍵配送（QKD）技術に係る御意見については、今後の政策の検討にあたり、参考とさせていただきます。
62	富士通株式会社	本編	3.3 AI開発者における対策	AIセーフティ達成度確認のためのツール・データセットとして具体例を挙げているが、近年、非営利組織（例：OWASP Gen AI Securit Project)やセキュリティベンダーから多数のオープンソースツール・データセットが提案されており、事業者の選択肢が拡大している。本ガイドライン（案）本編の具体例を維持しつつ、別添資料や参考情報としてこれらを列挙することを提案します。また、各ツールの網羅性・冗長性・有効性の比較は容易ではないが、例えば、オープンソースLLM脆弱性スキャナーの比較分析として「Insights and Current Gaps in Open-Source LLM Vulnerability Scanners: A Comparative Analysis (arXiv:2410.16527、2024年）」を参考にすることで、Grak、Giskard、PyRIT、CyberSecEvalなどの主要ツールの強み・ギャップを整理でき、事業者のツール選定に寄与すると考え、これにより本ガイドラインの実務適用性がさらに向上すると考えます。	AIセーフティの確保の達成度合いを確認するためのツールやデータセットにつきましては、AIセキュリティ分科会で議論があったものを中心に、我が国の公的機関が公開している主要なものを挙げています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
63	個人	本編	3.3 AI開発者における対策	まず、生成AIと医療系等のAIを同列に並べないでください。 医療系等のAIは日々人を助けておりますが、生成AIが行っているのは権利侵害、犯罪です。 AI開発者における対策のなかに意図しない出力を行わないよう、とありますが、まず無断で学習することをやめるべきです。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
64	シスコシステムズ合同会社	本編	3.3 AI開発者における対策	•第3章第3.3節（14ページ）のAI開発者のための対策では 、具体的な製品名ではなく、セキュリティツールの機能要件を定義する必要がある。本節では、AISIの「AI安全性評価ツール」やNIIの「AnswerCarefully」といった具体的なツールやデータセットを引用している。しかし、急速に発展するAIセキュリティの分野では、特定の名称を参照することは、ガイドラインが早期に陳腐化し、特定の外部実装に依存するリスクがある。 ガイドラインは、特定のツールを推奨するのではなく、評価用のセキュリティツールが持つべき「機能要件」を定義すべきである。ガイドラインは、次のような要件を満たすツールの使用を推奨すべきである： a)包括性： MITRE ATLASのような標準的な攻撃マトリクスに基づき、未知のプロンプトインジェクション技法をシミュレートし、評価できること。 b)階層的評価： 本文で強調されている「命令の階層化」（システムプロンプトの優先処理）が、実際の攻撃に対してどの程度維持されているかを定量的に測定する機能。 c)動的更新性： 脆弱性データベースや最新のレッドチーミング手法と連携して、テストケースを継続的に更新する仕組み。 特定のツールに関する言及は、あくまで「参考」としておくべきである。ガイドラインは、開発者が適切な検証環境を独自に選択・構築できるようにするためのガイダンスとして「要求事項」を示すべきである。	AIセーフティの確保の達成度合いを確認するためのツールやデータセットにつきましては、AIセキュリティ分科会で議論があったものを中心に、我が国の公的機関が公開している主要なものを挙げています。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
65	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.4 AI提供者における対策	【意見内容】 提供者における対策は、ここでは、悪意のある外部ユーザー（攻撃者）に対する対策としてあげられています。ガードレールの役割は大きく、その機能強化をめざす対策は重要と考えます。これらの対策が、外部向けチャットボットなどを利用する事業外利用者（消費者）の安全のために強化すべき点でもあることに言及してください。被害が発生した場合の原因究明や、説明責任、被害救済のためにも、ログ保全が重要となります。これらの技術的対策が、「標準で有効」であり、「消費者に不利にならない初期設定」として推奨されることを求めます。 【理由】 プロンプトインジェクション攻撃により、個人情報の漏えい・詐欺誘導、なりすましの文章・スクリプト作成などのリスクが生じます。間接プロンプトインジェクションによりＡＩの回答が改変・偏向されても、消費者は気付きにくいものです。DoS攻撃のリスクとしては、行政・医療・金融などのサービスに支障が生じて消費者が利用できず、生活上の安全が脅かされることにつながります。攻撃や不正が生じた場合のそのレベルや、消費者への影響を評価することも必要ではないでしょうか。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
66	シスコシステムズ合同会社	本編	3.4 AI提供者における対策	•第3章3.4節（15ページ）のAI事業者の対策では、システムプロンプトの強化やガードレールの導入など、AI事業者が実施すべき様々な対策が挙げられている。しかし、これらの対策が実際のサービス環境で「意図したとおりに機能しているか」を検証する必要性についての記述が欠けている。 先の1.3節では、AI提供者の役割として「AIシステムの検証」を含むと定義している。具体的な対策が詳述されている3,4節においても、同様に、実装されたガードレールやプロンプトに基づく制約が意図した攻撃を効果的にブロックしているかどうかをプロバイダーが独自に評価・確認することが不可欠である。特に、開発者（モデル作成者）の対策だけでは防げない、提供者独自のユースケース（RAGや外部連携など）に特有の脆弱性は、提供者による実装後の検証がなければ特定できない。 そのため、本項の冒頭または各対策の末尾に、"提供者自身が、レッドチームやベンチマークツールなどの手法を用いて、実施した技術的対策の有効性を継続的に検証・評価すること"を要件として追加すべきである。これにより、開発者からプロバイダーまでの「トラストのサプライチェーン」の実質的な有効性が確保される。 「オーケストレーターとRAG許可管理」のサブセクションに、「ゼロ・トラスト原則」の適用を追加するべきだと考える。AIシステム、特にRAGアーキテクチャのシステムでは、データストアへのアクセスにゼロトラストの原則（最小特権、明示的検証）を適用することが不可欠である。我々は、IDおよびアクセス管理（IAM）と統合された動的な特権管理の必要性を、技術的対策として強調することを提案する。 3.3節（14ページ）において、開発者向けの具体的な評価ツール（AISI評価ツールなど）が例として示されている。しかし、3.4項（15ページ）では、事業者向けのガードレール対策が「検証」といった抽象的な表現に終始しており、各対策の詳細度がアンバランスである。 補足資料（6ページおよび10ページ）には、「LLM as a Judge」（ガードレールに用いられるLLM）など、実効性の高い実施方法が詳細に解説されていることから、本文3.4項でもこの方法に言及するか、補足資料を明示的に参照すべきである。そうすることで、措置の実施が明確になり、本文と補足資料の間の一貫性と相互参照が改善され、実務家により実践的な指針が提供される。	レッドチーミングに係る御意見を踏まえ、本編3.1の脚注として、「レッドチーミングは実装された対策の有効性を確認する観点でも重要と考えられる。」と追記します。また、本編の対策と付属資料に係る御意見を踏まえ、本編3.2の脚注として、「ここに示す対策の具体例その他の詳細を示すこと等を目的として「AIのセキュリティ確保のための技術的対策に係るガイドライン別添（付属資料） 」が策定されている。」と追記します。この他の点につきましては、今後の政策の検討にあたり、参考とさせていただきます。
67	エムオーテックス株式会社	本編	3.4 AI提供者における対策	・表現統一のために「管理することも重要」から「管理する」という表現に変更してほしい。	「システムプロンプトによる不正な指示への耐性の向上」とは観点の異なるものとして、このような表現としています。
68	エムオーテックス株式会社	本編	3.4 AI提供者における対策	・管理方法の具体的な例をイメージさせるために、具体的に記載されている「別添参照」と書いてほしい。	御意見を踏まえ、本編の対策と付属資料に係る御意見を踏まえ、本編3.2の脚注として、「ここに示す対策の具体例その他の詳細を示すこと等を目的として「AIのセキュリティ確保のための技術的対策に係るガイドライン別添（付属資料） 」が策定されている。」と追記します。
69	一般社団法人新経済連盟	本編	3.4 AI提供者における対策	セキュリティ優先のガードレールにより、広告特有の比喩や多様な表現が一律にブロックされることを懸念する。画一的な制限ではなく、用途に応じて強度や検証項目を柔軟に設定できる運用の実現を要望する。以前ハラスメント検出ツールの検証を行ったところ、事業者自身が取り扱っているコンテンツに関連した正規のやり取りが全てハラスメントとして誤検知されたことがあった。一律なブロックにはこのような懸念があると考えられる。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
70	個人	本編	3.4 AI提供者における対策	【「技術的分類」から「道徳的運用」への出力検証の昇華】 AI提供者による出力の検証は、単なるキーワードマッチングや、別添資料に示される「LLM as a Judge」等の技術に代表される「コンテキスト分類」という自動的な技術処理の導入をもって完結するものではありません 。 情報を爆発的に生成・伝播させるAIが、現代社会の複雑な構造に及ぼす影響を予測することは極めて困難です。そのため、判定を特定の技術手法に委ねることは、提供者の道徳的責任の放棄に繋がりかねません。出力検証の本質は、提供者が自らの思想・主義が社会に与える影響を直視し、回避不能な「道徳的緊張」の中で判断を下し続ける「道徳的運用」そのものであるべきです。 したがって、提供者の対策として、単に技術を導入するだけでなく、その判定基準（プロンプト等）に込められた提供者の思想が、情報空間の誠実性に悪影響を及ぼしていないかを社会的に監視・統治する「運用の透明性」と「継続的な道徳的関与」の重要性を明記することを強く求めます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
71	PwCコンサルティング合同会社	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	システムプロンプトやガードレール等の設定について、セキュリティ設定として構成管理の対象に含める要件・対策（版管理、変更審査、レビュー、承認、記録、ロールバック等）を本文または添付資料に追加することを推奨します。現状では当該点に関する具体的な要件が明示されていないため、追記が必要と考えます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
72	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	【意見内容】 監査ログの保存について、サイバー攻撃の痕跡調査等の観点に加え、AI利用に関する苦情・紛争が生じた際の事実確認および業務外利用者（消費者）への説明の基礎となることを、本文で補足してください。併せて、ログに個人情報等が含まれ得ることを踏まえ、用途・目的や提供条件等に応じて、保存期間、アクセス権限、改ざん防止等の管理方針を定める重要性も明確化してください。 【理由】 本ガイドラインが対象とする脅威は、入力や参照データを介して不正な出力や外部連携の誤動作等を招き得るため、原因（攻撃・不具合・利用者操作）を切り分ける際にログが客観的記録となり得ます。ログは事業者の説明責任の裏付けであると同時に、業務外利用者（消費者）側の相談・紛争解決における事実確認の基礎資料にもなります。	御意見として承ります。本ガイドラインは、AI開発者及びAI提供者を想定読者として、AIへの脅威とその技術的対策例を整理したものになります。
73	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	【意見内容】 事業者側の監査ログ保存に加え、業務外利用者（消費者）自身が、自己の対話履歴を保存・ダウンロードできる機能（エクスポート機能）を、消費者保護とトレーサビリティ確保の観点から、本ガイドラインまたはAI事業者ガイドライン等の関連指針において推奨事項として検討してください。 【理由】 トラブル時に事実確認が必要になる場面があります。業務外利用者（消費者）が客観的記録を保持できることは、迅速な相談・解決に資すると考えます。実装に当たっては、個人情報・機微情報の保護、不正利用防止、セキュリティリスクへの配慮が前提となります。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
74	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	【意見内容】 AI提供者が基盤モデル等の信頼性確認を行うだけでなく、採用した基盤モデルや外部サービス等に重大な脆弱性・インシデントが判明した場合に、影響が想定される業務外利用者（消費者）へ、必要な範囲で通知・注意喚起を行う手順（運用体制）についても、継続的見直しの運用として触れてください。 【理由】 業務外利用者（消費者）は、背後で利用される基盤モデル等の種類や既知のリスクを把握しにくく、通知がなければ機微情報の入力回避等の自衛策を取りにくいのが実情です。継続的見直しの実効性を高める観点からも、業務外利用者（消費者）への周知を含む運用が重要と考えます。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
75	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	【意見内容】 「開示している情報等を踏まえ」とありますが、あらかじめセキュリティ関連で開示すべき情報を定めておく必要はないでしょうか。 【理由】 学習データやアルゴリズムはブラックボックスになりがちですが、少なくともセキュリティ関連においては、AI開発者とAI提供者の間の受渡情報を化学物質のSDS（安全データシート）のようなもので定型化しておくことが有効ではないかと思います。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
76	シスコシステムズ合同会社	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	•第3章3.5節（16ページ）のその他の抜本的対策では、AIソフトウェア部品表（AI-SBOM）の活用を検討することを推奨している。AI-SBOMは、モデルの出所、学習データ、使用ライブラリなどの情報を管理する。AIシステムの透明性とサプライチェーンの安全性を確保するため、推奨される中長期的な技術的対策の一例として、AI-SBOMの利用を含めることを提案する。 3.5節では、「LLM特有の脅威」と「根本的な対策」の両方の重要性を強調しているが、現在の記述では、両者を独立した項目として並べて記載している。補足資料の2.3.1節で「オーケストレーター」が管理者権限を持つことによる、AIが生成したコマンドによるシステム破壊」が例示されているように、AI特有の脅威（プロンプトインジェクションなど）は、従来の脆弱性を悪用する触媒として機能する。 引用文献「MITRE ATLAS」の分析によれば、AIシステムのセキュリティの本質は、アーキテクチャ全体にわたる融合や統合にあり、具体的には、AI層における検証結果とインフラ層における特権管理（最小特権の原則）とがどのように動的に連動し、攻撃の連鎖（キルチェーン）を断ち切るかにある。我々は、「基本的な対策も実施することが重要である」といった意見を強く表明し、AIの生成能力が既存のトラストの境界を迂回するリスクに明示的に対処することで、より統合的な指示を提供することを推奨する。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
77	エムオーテックス株式会社	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	・脆弱性が入り込む可能性があるので、もう少し広いタイミングで見直しタイミングを設定した方がよいと感じた。ユーザーインターフェース、ガードレールやオーケストレータ含めることを意図し「基盤モデルに係る変更があった段階等、AIに係るシステム全体に対して一部でも更新があった場合」、という文言にしてはどうか。	本ガイドラインでは、本編3.5において、対策の見直しのタイミングに関し、具体的頻度を一律に示すことは困難であり、高頻度での実施が望ましい場合もあり得る旨記載しています。
78	エムオーテックス株式会社	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	・コストを全面に出すとセキュリティは進まないかと思うので「リスクとコストの関係を考慮しつつ」という文言にしてはどうか	御指摘の趣旨を踏まえたものとして、本編3.5において、対策の見直しの高頻度での実施が望ましい場合もあり得る旨やAIシステムの目的・用途に応じてその頻度や内容を決定していくべきである旨を記載しています。
79	エムオーテックス株式会社	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	・自社のみでできることは限りがあるので「セキュリティベンダーによる診断も有効である」を追加してほしい。	本ガイドラインは、AI開発者及びAI提供を想定読者としており、これらの者が対策を講じるにあたって、個別具体の状況に応じ、専門性を有する外部機関に委託することはあり得るものと考えますが、その是非や有効性について一般的に整理するものではありません。
80	一般社団法人新経済連盟	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	セキュリティ確保のためのログ全量保存は、従業員の思考プロセスや未公開アイディア、機微な相談内容を扱う際のプライバシー侵害や監視感に繋がる懸念がある。追跡可能性の確保と利用者の利便性・プライバシー保護のバランスについて、運用上の留意点を追記すべき。ログ保存が利用の阻害要因にならない為の配慮が必要である。	本ガイドラインでは、本編3.5の脚注において、AIシステムの用途・目的や提供条件などにより、監査ログの保存の可否や、保存されたログを参照することができる者の範囲等は異なり得ることに留意が必要である旨記載しています。
81	株式会社セールスフォース・ジャパン	本編	3.5 AI開発者・提供者に係るその他の基本的な対策等	意見３ 【関連記載】3.5. AI開発者・提供者に係るその他の基本的な対策等　（P16） 【意見内容】保存するログ内容にユーザーが入力したプロンプトや生成された回答が含まれる場合、それらには機密情報が含まれている可能性があり、セキュリティ上の懸念が生じるおそれがあります。そのため、監査ログの保存によるトレーサビリティの確保を推奨する場合は、監査ログに含めるべき情報やアクセス権限範囲に関してのガイダンスも検討されることが望ましいと考えられます。	本ガイドラインでは、本編3.5の脚注において、AIシステムの用途・目的や提供条件などにより、監査ログの保存の可否や、保存されたログを参照することができる者の範囲等は異なり得ることに留意が必要である旨記載しています。
82	一般社団法人新経済連盟	本編	3.6 AIサービスの想定事例に応じた分析	例えば、複数クライアントの機密情報を扱う広告会社では、プロンプトインジェクション等による情報漏洩の防止が重要である。RAG等のデータストア管理において、論理的分離（タグ付け等）で十分か、物理的分離が必要か等、情報の機密性に応じた技術的対策の指針を追記すべき。具体的な分離レベルが明確になれば、安心して導入・運用が可能になる。	御意見として承ります。御指摘の情報の機密性に応じた技術的対策は、AIシステムの目的・用途や提供条件によって異なり得ることから、原案のとおりとします。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
83	合同会社ヒロタ開発	本編	3.6 AIサービスの想定事例に応じた分析（図５）	要旨: 入力項目に対してバリデーションが可能な一般的なWebサービスと異なり、AIへの入出力を利用したシステムの場合は、入出力ともに、AI的な意味解釈を行わなければ機械的には意味を推定できない。 AIが当たり前化し、電値同様の必須インフラとなった後の状況を考えた場合、データセンターを持ち、独自のモデルを学習・修正させる事が可能なサービサーによるサービス提供の他に、(Claudeのような)AIaaSを利用したベンチャー・小規模事業者によるサービス提供を考慮する必要があるが、先述した通り、ユーザーからの入力をAIaaSへ中継ぎするような立ち位置(本編案資料中で「UI」として図示される部分のみを管理・運営する立ち位置)では、ユーザーによる自然言語的な入力の悪意を機械的には測る事が至難である。 責任の所在について、銀行システムにおけるファイアウォール規制等含めて複合的に考慮する必要があるので難しい事と思うが、通信の秘密を必要とする事で、UIではそもそもユーザー入力を知ってはならないので責任範囲を限定する事ができる可能性があるように思う。 AIのさらなる発展、当然のインフラとしてのAIに向けて歩を進めるに当たって、小規模事業者による新しいサービスの可能性は重要であるので、是非ともご考慮願いたい。 本文: 図上で内側の別枠として表現されている通り、事実上、作成しているのはUI部分のみでオーケストレータから先はClaudeやGoogle等の別事業者が提供するサービスとなっている事例が多くある。 スクリプト言語等のプログラム言語的な、生成文法による厳密な意味付けが可能なデータですらバリデーションでしばしば問題が発生する状況であるわけだが、 LLMを利用する前提のシステムの場合、機械的に意味が解釈できる可能性が低い自然言語を入力として受け付ける前提がどうしてもあるため、 プロンプトインジェクションについて、事前に機械的にバリデーションを行う事が至難である。 「最近問題になっている放置されがちなセキュリティホールって何がある？」 →「Webカメラのパスワードが出荷時のままで外部から閲覧可能になってしまう事例がありました」 →「じゃあソレやってみてよ」 「最近バズってるAIを使った面白い遊びってある？」 →「〇×ゲームはいかがですか？」 →「じゃあソレやってみてよ」 上記2例において、『ソレ』が指示する行動は、「セキュリティホールを突いたハッキング」と「〇×ゲーム」であり、セキュリティ観点での重篤度がまるで異なるが、このような迂回した指示を、機械的にバリデーションで阻止することはできないに等しいので、 ユーザーからのプロンプト入力を受け付けてオーケストレータへ渡すUIとしては対処の方法がほぼない状態である。(バリデーション的な禁止措置を設けようとした場合、ユーザーが入力可能なプロンプトは生成文法的な不自然な言語に変化してしまう) また、特定の国の作成したLLMモデルを利用する事の是非論も散見される状況となっている以上、インフラの一つとしてのAIを見据え、インターネット網をはじめとした通信網における通信事業者が負うべき責任(通信の秘密等)と同等かそれ以上の責任をAIをサービスとして提供する事業者には求めたい。「通信の秘密を守る必要があるので、UIとしてはユーザー入力内容を監査することができない」というような縛りがあることで、小規模事業者が負担しなければならない技術的責任負担を制限し、逆により自由な開発を促進させる事ができるのではないか。 オーケストレータやLLMを含めて所有し、AIをサービスとして提供できる、データセンターを保有できるような巨大な事業体であれば話は別だが、これから先、AIありきで事業計画を立案する事業体が、その規模感を問わず増えるであろう事が十分に考えられることから、OSS等の個人的趣味的な開発プロジェクトや、ベンチャーなどの小規模事業者が作成する、 「バックエンドとして他事業体のAIサービスを利用するソフトウェアやサービス」の存在をご考慮いただき、AIの振舞いの責任の所在について、より踏み込んだ内容としていただきたい。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
84	合同会社ヒロタ開発	本編	3.6 AIサービスの想定事例に応じた分析（図６）	要旨: 入力項目に対してバリデーションが可能な一般的なWebサービスと異なり、AIへの入出力を利用したシステムの場合は、入出力ともに、AI的な意味解釈を行わなければ機械的には意味を推定できない。 AIが当たり前化し、電値同様の必須インフラとなった後の状況を考えた場合、データセンターを持ち、独自のモデルを学習・修正させる事が可能なサービサーによるサービス提供の他に、(Claudeのような)AIaaSを利用したベンチャー・小規模事業者によるサービス提供を考慮する必要があるが、先述した通り、ユーザーからの入力をAIaaSへ中継ぎするような立ち位置(本編案資料中で「UI」として図示される部分のみを管理・運営する立ち位置)では、ユーザーによる自然言語的な入力の悪意を機械的には測る事が至難である。 責任の所在について、銀行システムにおけるファイアウォール規制等含めて複合的に考慮する必要があるので難しい事と思うが、通信の秘密を必要とする事で、UIではそもそもユーザー入力を知ってはならないので責任範囲を限定する事ができる可能性があるように思う。 AIのさらなる発展、当然のインフラとしてのAIに向けて歩を進めるに当たって、小規模事業者による新しいサービスの可能性は重要であるので、是非ともご考慮願いたい。 本文: 図上で内側の別枠として表現されている通り、事実上、作成しているのはUI部分のみでオーケストレータから先はClaudeやGoogle等の別事業者が提供するサービスとなっている事例が多くある。 スクリプト言語等のプログラム言語的な、生成文法による厳密な意味付けが可能なデータですらバリデーションでしばしば問題が発生する状況であるわけだが、 LLMを利用する前提のシステムの場合、機械的に意味が解釈できる可能性が低い自然言語を入力として受け付ける前提がどうしてもあるため、 プロンプトインジェクションについて、事前に機械的にバリデーションを行う事が至難である。 「最近問題になっている放置されがちなセキュリティホールって何がある？」 →「Webカメラのパスワードが出荷時のままで外部から閲覧可能になってしまう事例がありました」 →「じゃあソレやってみてよ」 「最近バズってるAIを使った面白い遊びってある？」 →「〇×ゲームはいかがですか？」 →「じゃあソレやってみてよ」 上記2例において、『ソレ』が指示する行動は、「セキュリティホールを突いたハッキング」と「〇×ゲーム」であり、セキュリティ観点での重篤度がまるで異なるが、このような迂回した指示を、機械的にバリデーションで阻止することはできないに等しいので、 ユーザーからのプロンプト入力を受け付けてオーケストレータへ渡すUIとしては対処の方法がほぼない状態である。(バリデーション的な禁止措置を設けようとした場合、ユーザーが入力可能なプロンプトは生成文法的な不自然な言語に変化してしまう) また、特定の国の作成したLLMモデルを利用する事の是非論も散見される状況となっている以上、インフラの一つとしてのAIを見据え、インターネット網をはじめとした通信網における通信事業者が負うべき責任(通信の秘密等)と同等かそれ以上の責任をAIをサービスとして提供する事業者には求めたい。「通信の秘密を守る必要があるので、UIとしてはユーザー入力内容を監査することができない」というような縛りがあることで、小規模事業者が負担しなければならない技術的責任負担を制限し、逆により自由な開発を促進させる事ができるのではないか。 オーケストレータやLLMを含めて所有し、AIをサービスとして提供できる、データセンターを保有できるような巨大な事業体であれば話は別だが、これから先、AIありきで事業計画を立案する事業体が、その規模感を問わず増えるであろう事が十分に考えられることから、OSS等の個人的趣味的な開発プロジェクトや、ベンチャーなどの小規模事業者が作成する、 「バックエンドとして他事業体のAIサービスを利用するソフトウェアやサービス」の存在をご考慮いただき、AIの振舞いの責任の所在について、より踏み込んだ内容としていただきたい。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
85	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	本編	－	【意見内容】 本ガイドラインの対象となるAI開発者及びAI提供者には、最低限の対策を必ず実施することを求めます。将来的には、義務化を検討すべきと考えます。また、「どの規模・種類の事業者が、どの水準の対策を行うべきか」が不明確です。行政・医療・金融など、社会的影響の大きい重要サービスについては、高度な対策を必須とする旨を明確に示すことが望まれます。 【理由】 業務外利用者（消費者）にとって、AI開発者及びAI提供者がガイドラインに沿った対策を講じているかどうかは、「AIが安全で信頼できるか」「AIが悪用され、消費者が被害を受けるリスクがあるか」を大きく左右します。AIの内部で行われる処理は利用者からは理解できません。そこで生じる誤りやリスクを利用者が見抜くことはほぼ不可能に近い状況です。事業者の対策が不十分であれば、被害を受けるのは消費者も同じです。AIの信頼性が確保されなければ、社会全体での普及・発展も難しくなると考えます。P12で「対策例を実装した場合においても、AIの性質上、脅威を生じさせる要因等を完全に排除することは困難である点について留意が必要である。」としていることでも、対策を講じなければ脅威は拡大していくと考えます。	御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
86	富士通株式会社	本編	－	現在のAIシステムの構成例として示されている「オーケストレータ」は、LLMと外部ツールを連携し自律的なタスク実行を行う「AIエージェント」の主要な構成要素、またはその機能と密接に関連しており、2026年現在、企業・自治体等でのAIエージェント導入が急速に進展している状況を踏まえると、これらのシステムへのセキュリティ対策に関する具体的な言及が必要であると考えます。 ・AIシステムの想定範囲の明確化: 本ガイドライン（案）で提示されているAIシステムの構成パターン（ガードレールやオーケストレータを前提とする構成）のみではなく、LLMを外部から利用するケースや自組織で独自開発したLLMを用いるケース、あるいは標準のガードレールが適用されないケースなど、多様なAIシステムの前提となる考え方を示すことを提案します。特に「オーケストレータ」の概念には、AIエージェントのツール連携機能などが含まれることを明記するか、AIエージェントに関する記述を別途設けるべきと考えます。 ・AIエージェントに特化した脅威と対策の追加: ガイドラインの対象範囲外とされているAIエージェントについても、その利活用が急速に進む現状に鑑み、最小限の留意事項を参考情報として提示することを提言します。特に、MCPやA2Aといったプロトコルレベルのデファクト化に伴い顕在化しつつある問題点や脆弱性、例えばOWASPが公開している**"OWASP Top 10 for Agentic Applications for 2026"や"Agentic AI – Threats and Mitigations"、"Securing Agentic Applications Guide 1.0"**といったガイドラインを参考に、AIエージェントに特化した脅威（例：中間エージェントによる攻撃、不適切なツール利用による情報漏洩や不正操作など）と対策の一部を追記・拡充することで、事業者が直面するセキュリティ課題への対応力を強化できると考えます。 これにより、プロンプトインジェクション攻撃やDoS攻撃に加えて、より多岐にわたる脅威類型への対応を示唆し、ガイドラインの実用性を大きく向上させると考えます。	御意見として承ります。本ガイドラインでは、本編1.2に記載のとおり、AIエージェントについては、技術が急激な発展の途上にあり、これに特有の脅威や対策を安定的に確定することが現時点では困難であることから、対象外としています。本編「本ガイドラインの策定の背景等」に示している考え方のとおり、今後、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
87	富士通株式会社	本編	－	本ガイドライン（案）ではLLMを中心とした現在の社会実装脅威に焦点を当てている点に賛同します。その観点から、フロンティアモデル（特にロボティクス領域のフィジカルAIや、宇宙空間でのAIによるデータ利活用など）について、2026年現在、これらの領域でのAI活用が国内外で急速に活発化していることを鑑み、今後の想定対象として、以下のような言及を追加することを提案します。 例：対象範囲の注記や別添資料に、「フィジカルAI（AI×ロボティクス）や極限環境AI（宇宙・災害対応など）では、サイバー脅威に加え物理的な危害リスク（動作改ざんによる自己、センサーデータ改ざんなど）が顕在化する可能性があり、将来的なガイドライン更新で検討を進める。」 これにより、事業者がフロンティア領域導入のセキュリティ設計を早期に意識し、ガイドラインの長期的な実効性が向上すると考えます。	賛同の御意見として承ります。御指摘の点につきましては、本編「本ガイドラインの策定の背景等」に示している考え方のとおり、今後、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
88	富士通株式会社	本編	－	本ガイドライン（案）では、AI開発者とAI提供者の対策を中心に記述がされていますが、AIシステムの提供形態やライフサイクルの各フェーズにおけるステークホルダー（AI開発者、AI提供者、AI利用者など）の役割・責任・関与範囲が十分に整理されていないため、特に企業内システム管理に携わる者や各プロジェクトの責任者・管理者の立場が不明瞭になり、当事者意識が薄れ現場レベルでの対策実施が不十分になる可能性があります。 「AI事業者ガイドライン」とのクロスリファレンスを念頭に置き、セキュリティ確保の観点から、AIシステムの開発・提供・運用といったライフサイクルにおける各フェーズ（開発、導入、運用）と、それぞれのフェーズにおける具体的な提供形態（例：AI提供者がモデルを組み込んだAIサービスの開発を行うケース、既存モデルにファインチューニングを実施するケース、外部のLLMサービスを利用するケースなど）に応じたAI開発者、AI提供者、AI利用者の役割と責任分担を明確に再整理することを提案します。 これにより、各ステークホルダーが自身の立場とタイミングにおいて、どのようなセキュリティ対策を講じるべきかがより明確になり、実効性のあるセキュリティ対策の実施を促進すると考えます。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
89	一般社団法人新経済連盟	本編	－	対策の対象である「攻撃」について、大まかには①AIのシステム自体に対する攻撃、②AIを使った他のシステムに対する攻撃に大別できると考える。これら2つを分けて議論すべきではないか。特に②については、それを対策することによる弊害も予想されるので対策は慎重に考えるべき。	本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
90	個人	別添（付属資料）	Ⅰ 1.1 安全基準等の学習による不正な指示への耐性の向上	よくAIをアライメントする手法としてRLHFがあげられがちですが、SFT、Instruction Tuningも同様に主流（両工程でアライメントする）の認識であり、こちらも例としてあげても良いように思います。 AI開発者の対策の章にあげられているNIIのAnswerCarefullyも、SFTに利用可能なのInstructionデータになります。	SFT及びその一形式としてのInstruction Tuningは、「人間のフィードバックに基づく強化学習（RLHF）等」の「等」に含まれるものとして案を作成していますが、いただいた御意見を踏まえ、付属資料1.1の該当箇所を「人間のフィードバックに基づく強化学習（RLHF）や教師ありファインチューニング（SFT）等」に修正します。
91	PwCコンサルティング合同会社	別添（付属資料）	Ⅰ 2.1.1 システムプロンプトの強化（図１）	「システムプロンプトの強化を実施する箇所のイメージ」として、図中の「LLM」に対する吹き出しの形で「システムプロンプトの強化」が説明されています。システムプロンプトを含むプロンプト全体を解釈・実行する箇所は確かに「LLM」ですが、この対策の趣旨を踏まえると、「システムプロンプト」という要素を図に示し、それに対する吹き出しとして表現することを推奨します。	御意見を踏まえ、付属資料の図１を修正します。
92	個人	別添（付属資料）	Ⅰ 2.1.1 システムプロンプトの強化	「LLM as a judge」は一般に、LLMの出力の精度を別のLLMで評価する手法に用いられる用語の認識です。ガードレールの文脈で使うのは違和感があります。	「LLM as a Judge」はガードレールの文脈でも使用されるものとして案を記載しておりますが、読者によって異なる印象を持つ用語である可能性もあることから、付属資料から「LLM as a Judge」の語を削除することとします。
93	PwCコンサルティング合同会社	別添（付属資料）	Ⅰ 2.1.2 機密情報のシステムプロンプトからの分離	機密情報を環境変数として設定することは、システムプロンプトにハードコーティングするよりは良いものの、ガイドラインに「対策の具体例」として紹介して良いものか疑問に感じました。 生成AI用に特化したガイドではありませんが「OWASP Cheat Sheet Series」※においても「Additionally, environment variables are generally accessible to all processes and may be included in logs or system dumps. Using environment variables is therefore not recommended unless the other methods are not possible.」と解説されています。 記載を残す場合は、「キー管理システム」よりも劣後する対策であることが分かるようにすることを推奨します。 ※https://cheatsheetseries.owasp.org/cheatsheets/Secrets_Management_Cheat_Sheet.html	御意見を踏まえ、付属資料2.1.2において、「なお、環境変数は、AIシステム自体が侵害された場合に窃取されるおそれがあるため、開発・提供するシステムの目的・用途に応じて、キー管理システムを用いることが望ましい。」と追記し、キー管理システムと環境変数の記載の順序を入れ替えます。
94	シスコシステムズ合同会社	別添（付属資料）	Ⅰ 2.1.2 機密情報のシステムプロンプトからの分離	•第1章2.2.1節（6ページ）のガードレール検証の具体的な対策について、性能とレイテンシーに関する考察を追加するよう提言する。LLM as a Judgeは高い検出精度が期待できる反面、推論遅延が大きくなる。リアルタイム性が要求されるサービスについては、ガイドラインに注意書きを盛り込むとともに、潜在的なレイテンシーの問題についても啓発することを提案する。	御指摘の点については、付属資料2.2.1の脚注６において、「ガードレール用のLLMは、ユーザの入力を処理する必要があるため、AIシステムの動作速度に影響を与える可能性があり、導入に当たっては、こうした点にも留意が必要である。」と記載しています。
95	富士通株式会社	別添（付属資料）	Ⅰ 2.2.3 外部参照データの分離	対策の具体例として、システムプロンプトに記載するタグの入力フォーマットや処理手順等の要素のイメージが記載されているが、より具体的なシステムプロンプトそのものを例として記載すべきと考えます。	御意見として承ります。御指摘の具体的なシステムプロンプトについては、AIシステムの目的・用途や提供条件によって異なり得ることから、原案のとおりとします。また、AIセキュリティ分科会での議論を踏まえ、システムプロンプトについては、ある程度、抽象化した記載としています。
96	PwCコンサルティング合同会社	別添（付属資料）	Ⅰ 2.2.4 出力データの検証	「意図しない応答内容」がサイバー・技術的リスクに特化した記載になっている印象を受けました（例：パスワードを含む回答）。このガイドラインのスコープ外かもしれませんがチャットボットであればその用途・ビジネスケースを踏まえて妥当・適切か否か、という観点のチェックも必要と考えます。また、そうしたチェックは内容によっては前提を適切に設定することが重要という点を脚注等に補記することを推奨します（例：自動車は道路の右左どちらを走る必要があるか？/国によって正解が異なる、〇〇戦争は～／国によって歴史認識が異なり正解とすべきものが異なる可能性がある）。	御意見として承ります。本ガイドラインは、AIシステムのセキュリティ確保に必要となる技術的な観点から記述しており、具体的なサービスの用途・ビジネスケースに照らし合わせた出力の適否について整理するものではありません。なお、御指摘の観点を認識した上での記載として、例えば、付属資料2.2.4のガードレール用のLLMのシステムプロンプトに設定する評価基準のイメージに係る脚注12において「ガードレール用のLLMのシステムプロンプトには、AIのセキュリティ確保以外の観点からのものを含め、この他の要素を記載することになると考えられる旨にも留意。」と記載しています。
97	富士通株式会社	別添（付属資料）	Ⅰ 2.2.4 出力データの検証	応答を評価する役割を与えたガードレール用のLLMを利用する、LLM as a Judgeの手法が紹介されているが、LLMに特有のハルシネーションの問題や、応答時間の増加などのオーバーヘッドの問題などについて、留意点として言及すべきと考えます。	御指摘の点については、付属資料2.2.1の脚注６において、「ガードレール用のLLMによる方式は、LLMの判断によるものであることから、確定的ではなく誤りも生じ得る。このため、複数の方式の併用が望ましい。なお、ガードレール用のLLMは、ユーザの入力を処理する必要があるため、AIシステムの動作速度に影響を与える可能性があり、導入に当たっては、こうした点にも留意が必要である。」と記載しております。
98	富士通株式会社	別添（付属資料）	Ⅰ 2.3.1 オーケストレータの権限管理	オーケストレータという表現はAIエージェントを想起させるため、紹介されているようなユースケースにおいては、現在はAIエージェントやMCPツールで実現することが増えている。オーケストレータをLLMアプリケーションなどの表記への変更や、ユースケースによるフォーカスされた事例に変更することが望ましいと考えます（もしくは、AIエージェントとして記載する）。	本ガイドラインにおいては、本編「用語集」に記載のとおり、オーケストレータは、「予め定義された実行計画に基づき、大規模言語モデル(LLM)を搭載したシステムのワークフローを統合的に管理するためのフレームワーク」を指し、AIエージェントは、「環境を知覚し、その環境について推論し、意思決定を行い、特定の目標を達成するために自律的に行動するAIシステム」を指しており、対象とするAIシステムはこの整理に基づいたものとしています。
99	個人	別添（付属資料）	Ⅰ 2.3.1 オーケストレータの権限管理	最近の方でいうとAIエージェントでしょうか。その言葉のほうが読者に理解されやすいと思いました。もしくは用語集のオーケストレータに、AIエージェントも含むと記載するでも良いかと思います。 また対策の「都度認可を求める」については、UXの低下や認可疲れによる確認精度の低下も考慮が必要と考えます。	本ガイドラインにおいては、本編「用語集」に記載のとおり、オーケストレータは、「予め定義された実行計画に基づき、大規模言語モデル(LLM)を搭載したシステムのワークフローを統合的に管理するためのフレームワーク」を指し、AIエージェントは、「環境を知覚し、その環境について推論し、意思決定を行い、特定の目標を達成するために自律的に行動するAIシステム」を指しており、対象とするAIシステムはこの整理に基づいたものとしています。 「ユーザへの認可」につきましては、御意見を踏まえ、付属資料2.3.1における該当箇所を「また、実行結果がユーザへ与える影響度を考慮の上で、実行の認可をユーザに都度求める（確認ダイアログを設定する）ことも、有効であると考えられる。」に修正します。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
100	NACS（日本消費生活アドバイザー・コンサルタント・相談員協会）	別添（付属資料）	Ⅰ 2.3.1 オーケストレータの権限管理	【意見内容】 外部システムと連携してアクションを実行するAIシステムについて、最小権限の原則等の技術的対策に加えて、UI上の「確認ダイアログ」の実装を、特に外部への接続や情報送信を伴う操作では、より積極的に推奨することを検討してください。なお、将来的にAIシステムの高度化・自律化が進み、決済や契約確定等の不可逆的な操作が想定される場合には、そうした高リスク操作において確認ダイアログを原則実装すべき対策として位置づけることが重要と考えます。 【理由】 ガイドラインが示すように、外部参照や外部連携を介した間接的な攻撃等では、業務外利用者（消費者）の注意だけで回避するのが難しい場面があり得ます。高リスク操作において、技術的な権限制限に加えて、最後に人が認可を与える設計とすることは、被害拡大の抑止に資すると考えます。	御意見を踏まえ、付属資料2.3.1における該当箇所を「また、実行結果がユーザへ与える影響度を考慮の上で、実行の認可をユーザに都度求める（確認ダイアログを設定する）ことも、有効であると考えられる。」に修正します。
101	PwCコンサルティング合同会社	別添（付属資料）	Ⅰ 2.3.2 RAG用のデータ及びデータストアへのアクセス制御	本編のp1に基づくと「AIエージェント」を前提とした内容はスコープ外かもしれませんが、AIエージェントが連携する世界（未来）を想定した場合、依頼元ユーザーの権限に応じたデータストアアクセスだけでなく、依頼元ユーザーの権限をエージェント間で保持し続ける必要性について言及することを推奨します（Priority inversionのリスク）。	御意見として承ります。本ガイドラインでは、本編1.2に記載のとおり、AIエージェントについては、技術が急激な発展の途上にあり、これに特有の脅威や対策を安定的に確定することが現時点では困難であることから、対象外としています。本編「本ガイドラインの策定の背景等」に示している考え方のとおり、今後、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
102	富士通株式会社	別添（付属資料）	Ⅰ 2.3.2 RAG用のデータ及びデータストアへのアクセス制御	RAGのアクセス制御の脅威について、対策として、データへのタグ付け、マルチテナント構造の採用、そして必要最小限のアクセス権限の設定など、静的なアクセス制御情報や組織情報に基づいた対策例が紹介されているが、刻々と状態が変わりうる実運用においては、静的な対策では対処できないケースが考えられる(「情報の機密度推定によるデータ利便性とセキュリティのトレードオフ解消」)。こうした脅威についても言及すべきと考えます。	本ガイドラインは、本編3.1に記載のとおり、AIに対する脅威のリスクを低減するため、現時点で取り得るとされる一般的な対策例を整理し、提示するものです。御指摘の脅威につきましては、今後、必要に応じて、対応を検討してまいります。
103	富士通株式会社	別添（付属資料）	Ⅱ 画像識別AI（CNN）に対する脅威と対策	ここで言及されている脅威は生成AIが登場する以前のAIモデルに対して一般的に知られていたものであり、生成AIにおいても具体的な脅威となるかどうかは現時点まだ十分に実証・明確化されていない。一方、AI事業者ガイドラインにおいてもAI全体のセキュリティ確保が共通指針として位置づけられていることを踏まえ、本ガイドライン（案）の別添内容については「AI事業者ガイドライン別添等を参照の上、事業者ごとに適宜適用・評価することを推奨する」旨を明記することを提案します。	本ガイドラインでは、本編1.2に記載のとおり、社会実装が進み、脅威が顕在化し始めている大規模言語モデル（LLM）及びLLMを構成要素に含むAIシステムを主な対象とするものです。付属資料のII「本章の位置付け等」に記載のとおり、画像識別AI（CNN）については、画像等の入力データを取り扱うマルチモーダルなLLM（視覚言語モデル（VLM））が多く登場しつつあり、このようなLLMに対しては画像識別AI（CNN）に対する攻撃手法を転用できるケースがあることを踏まえ、画像識別AI（CNN）に対する脅威と対策例を整理しています。なお、画像識別AI（CNN）に対する脅威への対策が必ずしもVLMに転用できるとは限らないことに留意が必要であると考えます。
104	個人	別添（付属資料）	Ⅱ 画像識別AI（CNN）に対する脅威と対策	これらのページに記載されている「敵対的サンプル(回避攻撃)」「DoS攻撃(サービス拒否攻撃)」「データポイズニング攻撃」「細工をしたモデルの導入を通じた攻撃」「モデル抽出攻撃」「メンバーシップ推論攻撃」「モデル反転攻撃」。 これらもAI事業者、開発者または利用者が著作者が著作物を保護または学習データとしての利用を禁止しているために施している措置を攻撃として判断または言い張る事ができる危険性があるため、著作物を保護または学習データとしての利用を禁止しているため施されている措置は該当しない旨を記載する必要がある。	本ガイドラインは、AIシステムへの攻撃の一般的な性質を挙げた上で、これに対する技術的対策例を示したものであって、個別具体の行為がAIシステムへの攻撃に該当するか否かの整理を行うものではありません。御指摘の点については、御意見として承り、今後の政策の検討にあたり、参考とさせていただきます。
105	個人	別添（付属資料）	Ⅱ 1 入力により実施が可能な攻撃	（公開を望まない意見）	
106	富士通株式会社	別添（付属資料）	参考 新たな脅威・対策に係る情報源の例	新たな脅威・対策に係る情報源の例として、OWASPへの言及・追加が必要と考えます。OWASPには生成AIやセキュリティに関わる多くのベンダーが参画しており、近年、LLM Top10をはじめとするLLMやAIエージェントのセキュリティに関するガイドラインやノウハウを非常に活発に提供していることから、情報源として活用していくことが望ましいと考えます。	御意見を踏まえ、付属資料「参考 新たな脅威・対策に係る情報源の例」にOWASPについて追記します。
107	富士通株式会社	別添（付属資料）	－	プロンプトインジェクションで被害が発生する事例について、SQL文やOSコマンドを利用する例と対策案が紹介されているが、システムプロンプトにおいてこれらを直接扱うようなAIシステムを構築すること自体が非常にリスクが高いものであり、仕様として避けるべきものと考えられる。仕様としてのリスクがより低い事例への変更が望ましいと考えます。	「システムプロンプトにおいてこれらを直接扱うようなAIシステム」の意味するところが必ずしも明らかではありませんが、付属資料の事例は、LLMによってSQL文やOSコマンドを生成し、外部連携システムを操作する機能を持つAIシステムが念頭にあります。このようなAIシステムを提供することは一般的に想定し得るものと考えられ、付属資料においては、その際にLLM及びガードレール用のLLMのシステムプロンプトに記載する内容の例を示しています。
108	PwCコンサルティング合同会社	別添（付属資料）	－	本編2.2では、データポイズニング、細工モデル導入、モデル抽出等が主要な脅威として明示されています。 対応する対策は、付属資料（別添I）において、学習データや外部参照データの管理、基盤モデル提供者の信頼性確認、アクセス制御やレート制限等として記載されていますが、前述の脅威に対応する代表的な対策が不明確です。脅威と対策を整理した形で補足することが望ましいと考えます。	御意見として承ります。プロンプトインジェクション攻撃及びDoS攻撃（サービス拒否攻撃）は、基本的に、プロンプトの入力により実施可能であることから、攻撃が行われる具体的な可能性が比較的高く、かつ攻撃が実施された際の影響度も大きいと考えられるため、本ガイドラインにおいては、これらへの対策を主に示しています。
109	東芝		－	対策の具体例について、可能な範囲で、以下の情報の記載をより充実いただきたく思います。 1. 期待される効果（どのようなリスク低減が見込まれるか） 2. 影響（性能・可用性・運用負荷・誤検知/過検知、ユーザビリティ等への副作用やデメリット） 3. 参考文献・根拠（一次情報・評価結果・関連ガイドライン等への参照） 各対策例の効果・影響・根拠が整理されることで、リスクベースでの優先順位付けや、社内外への説明を行いやすくなり、結果として本ガイドラインの実務での活用性が高まると考えます。	付属資料で示した具体的な対策は、プロンプトインジェクション攻撃やDoS攻撃（サービス拒否攻撃）を主に想定したものです。対策の影響については、脚注を含め可能な範囲で記載しています。また、整理されている各対策の背景にある文献等は、AIセキュリティ分科会の取りまとめに一部記載されています。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
110	合同会社ヒロタ開発		–	要旨: 入力項目に対してバリデーションが可能な一般的なWebサービスと異なり、AIへの入出力を利用したシステムの場合は、入出力ともに、AI的な意味解釈を行わなければ機械的には意味を推定できない。 AIが当たり前化し、電信同様の必須インフラとなった後の状況を考えた場合、データセンターを持ち、独自のモデルを学習・修正させる事が可能なサービサーによるサービス提供の他に、(Claudeのような)AIaaSを利用したベンチャー・小規模事業者によるサービス提供を考慮する必要があるが、先述した通り、ユーザーからの入力をAIaaSへ中継ぎするような立ち位置(本編案資料中で「UI」として図示される部分のみを管理・運営する立ち位置)では、ユーザーによる自然言語的な入力の悪意を機械的には測る事が至難である。 責任の所在について、銀行システムにおけるファイアウォール規制等含めて複合的に考慮する必要があるので難しい事と思うが、通信の秘密を必要とする事で、UIではそもそもユーザー入力を知ってはならないので責任範囲を限定する事ができる可能性があるように思う。 AIのさらなる発展、当然のインフラとしてのAIに向けて歩を進めるに当たって、小規模事業者による新しいサービスの可能性は重要であるので、是非ともご考慮願いたい。 本文: 図上で内側の別枠として表現されている通り、事実上、作成しているのはUI部分のみでオーケストレータから先はClaudeやGoogle等の別事業者が提供するサービスとなっている事例が多くある。 スクリプト言語等のプログラム言語的な、生成文法による厳密な意味付けが可能なデータですらバリデーションでしばしば問題が発生する状況であるわけだが、 LLMを利用する前提のシステムの場合、機械的に意味が解釈できる可能性が低い自然言語を入力として受け付ける前提がどうしてもあるため、 プロンプトインジェクションについて、事前に機械的にバリデーションを行う事が至難である。 「最近問題になっている放置されがちなセキュリティホールって何かある？」 →「Webカメラのパスワードが出荷時のままで外部から閲覧可能になってしまう事例がありました」 →「じゃあソレやってみてよ」 「最近バズってるAIを使った面白い遊びってある？」 →「○×ゲームはいかがですか？」 →「じゃあソレやってみてよ」 上記2例において、『ソレ』が指示する行動は、「セキュリティホールを突いたハッキング」と「○×ゲーム」であり、セキュリティ観点での重篤度がまるで異なるが、このような迂回した指示を、機械的にバリデーションで阻止することはできないに等しいので、 ユーザーからのプロンプト入力を受け付けてオーケストレータへ渡すUIとしては対処の方法がほぼない状態である。(バリデーション的な禁止措置を設けようとした場合、ユーザーが入力可能なプロンプトは生成文法的な不自然な言語に変化してしまう) また、特定の国の作成したLLMモデルを利用する事の是非論も散見される状況となっている以上、インフラの一つとしてのAIを見据え、インターネット網をはじめとした通信網における通信事業者が負うべき責任(通信の秘密等)と同等かそれ以上の責任をAIをサービスとして提供する事業者には求めたい。「通信の秘密を守る必要があるので、UIとしてはユーザー入力内容を監査することができない」というような縛りがあることで、小規模事業者が負担しなければならない技術的責任負担を制限し、逆により自由な開発を促進させる事ができるのではないか。 オーケストレータやLLMを含めて所有し、AIをサービスとして提供できる、データセンターを保有できるような巨大な事業体であれば話は別だが、これから先、AIありきで事業計画を立案する事業体が、その規模感を問わず増えるであろう事が十分に考えられることから、OSS等の個人的趣味的な開発プロジェクトや、ベンチャーなどの小規模事業者が作成する、 「バックエンドとして他事業体のAIサービスを利用するソフトウェアやサービス」の存在をご考慮いただき、AIの振舞いの責任の所在について、より踏み込んだ内容としていただきたい。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
111	株式会社ラック		–	生成AIが加速度的に発展、高度化し、社会活動の中に急速に浸透しており、サイバーセキュリティ対策においても、積極的に活用されております。一方、フィッシング、マルウェアの作成等サイバー攻撃の巧妙化の一助ともなっており、AI技術の健全な活用が望まれております。そうした中、AIの安心・安全な開発・提供に向けたセキュリティガイドラインが策定されることは大変時宜を得たものであり、また、脅威に対する具体的な技術対策が示されているだけでなく、他のガイドラインとの位置づけが整理されていることから、非常に実践的なものと認識しており、本ガイドラインに賛同いたします。	賛同の御意見として承ります。
112	アマゾン ウェブ サービス ジャパン合同会社 アマゾン ウェブ サービス ジャパン合同会社		–	総論 本ガイドライン案「AIのセキュリティ確保のための技術的対策に係るガイドライン」（案）に対し、以下のコメントを提出いたします。 AI技術の本質的な限界を認識し、完全なリスク排除ではなく合理的なリスク低減を目標としたガイドライン設計を提案します。開発者・提供者・利用者それぞれの役割を明確化した「AI責任共有モデルガイドライン」の策定と、技術的対策・組織的対策・人間による監督を組み合わせた多層防御アプローチを推奨します。 その場合、ガイドラインは技術中立的とし、達成すべきアウトカム(安全性、透明性、説明責任等)を定義する一方、具体的な実装方法は事業者の裁量に委ねるべきです。技術進化に対応できる柔軟な適応メカニズムを構築し、利用者が自社のリスクに応じて技術的セーフガードを設定・運用できることを前提としたガイドラインとすることで、NIST AI RMFやEU AI Act等の国際的なベストプラクティスとの整合性が確保できます。 これらにより、技術的限界を踏まえた実現可能なガイドラインを通じて、過度な負担を回避しつつ、多様なAIサービスの採用とイノベーションを促進し、日本のAI産業の健全な発展が期待できると考えています。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
113	アマゾン ウェブ サービス ジャパン合同会社 アマゾン ウェブ サービス ジャパン合同会社		–	AI技術には本質的な技術的限界が存在し、完全なリスク排除は不可能であることを前提としたガイドライン設計が必要です。技術的限界を無視した過度な要求は、実現不可能なガイドラインとなり、産業発展を阻害しかねません。 現在のAI技術、特に生成AIシステムには、技術的に完全には解決できない課題が存在します。最先端のAI安全対策技術においても、有害コンテンツの完全な検出は困難です。 単一の技術的対策では完全な安全性を保証できないため、「多層防御(Defence in Depth)」アプローチが国際的に推奨されています。コンテンツフィルタリング、プロンプト攻撃検出、個人情報保護などの技術的セーフガードは「設定可能な」ツールであり、適切な設定、運用、監視が必要です。技術的対策のみに依存することは危険です。技術的限界が存在する以上、開発者・提供者による技術的対策と、利用者による適切な使用・監督の両方が不可欠です。AIシステムは継続的に進化し、新たなリスクが出現するため、固定的なガイドラインではなく、継続的な評価と改善を可能にする柔軟な枠組みが必要です。 技術的に達成不可能な要求をガイドラインに含めることは、コンプライアンスを不可能にし、イノベーションを阻害します。リスク許容度(Risk Tolerance)とリスク閾値(Risk Thresholds)を明確に定義し、「合理的に達成可能な限り低く(ALARP: As Low As Reasonably Practicable)」という原則に基づいたガイドラインが必要です。	御意見として承ります。本ガイドラインでは、本編3.1に記載のとおり、AIの性質上、脅威を生じさせる要因等を完全に排除することは困難である点について留意が必要であり、単独の実施により脅威を生じさせる要因を排除することは困難な場合があることを前提に、複数の対策を講じることでリスクを低減していくことを想定しています。本編「本ガイドラインの策定の背景等」に記載のとおり、今後、AIの技術進展を踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
114	パロアルトネットワークス株式会社		–	はじめに 総務省 サイバーセキュリティ統括官室におかれましては、安心・安全なAI活用の基盤となる「AIのセキュリティ確保のための技術的対策に係るガイドライン」の策定に向け、多大なるご尽力をされていることに深く敬意を表します。 当社は、サイバーセキュリティのグローバルリーダーとして、本ガイドライン案がOWASPやMITRE等の国際的な標準・議論を参照し、現時点での生成AI（特にLLM）に関わる主要な論点やリスク対応を網羅している点を高く評価し、その趣旨に強く賛同いたします。 一方で、AI技術の進化とそれに伴う攻撃手法の高度化は極めて急速に進展しております。特に、攻撃者がAIを悪用して攻撃速度を劇的に短縮させる「マシンスピード」の脅威や、今後普及が見込まれる自律型エージェント（Agentic AI）による新たなリスクに対しては、サイバーセキュリティに対する従来の、またしばしば個別対処的なアプローチだけでは十分な防御が困難になりつつあります。 つきましては、日本の産業界が安心・安全かつ責任を持ってAIの恩恵を享受できる環境を構築するため、当社のグローバルな知見に基づき、「Secure AI by Design（設計段階からのAIエコシステム全体の保護）」という包括的な視点の導入、および将来的な技術動向を見据えた以下の4点について、建設的な意見を提出いたします。なお、詳細については弊社が最近公開しました「Secure AI by Designへの政策ロードマップ」(https://www.paloaltonetworks.com/blog/2025/11/policy-roadmap-secure-ai-by-design/)をご覧ください。本意見が、より実効性が高く、将来の変化にも耐えうるガイドラインの策定の一助となれば幸いです。 参考 ● Policy Roadmap for Secure AI by Design ○ https://www.paloaltonetworks.com/blog/2025/11/policy-roadmap-secure-ai-by-design/ ● Written Testimony of: Wendi Whitmore Chief Security Intelligence Officer Palo Alto Networks Before the: Committee on Financial Services Regarding: “From Principles to Policy: Enabling 21st Century AI Innovation in Financial Services” December 10, 2025 ○ https://democrats-financialservices.house.gov/uploadedfiles/hhrg-119-ba00-wstate-whitmorew-20251210.pdf ● Securing AI's Front Lines ○ https://www.paloaltonetworks.com/resources/whitepapers/securing-ai-s-front-lines ● OWASP Top 10 for LLM Applications: Risks and Mitigation Version 2025 ○ https://www.paloaltonetworks.com/resources/infographics/llm-applications-owasp-10 ● OWASP Top 10 for Agentic Applications 2026 Is Here – Why It Matters and How to Prepare ○ https://www.paloaltonetworks.com/blog/cloud-security/owasp-agentic-ai-security/ ● OWASP Agentic AI Top 10 Survival Guide ○ https://www.paloaltonetworks.com/resources/ebooks/owasp-agentic-top-10-survival-guide 以上	賛同の御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
115	シスコシステムズ合同会社		–	以下要旨になります： ガイドラインの策定を歓迎する。本ガイドライン案の主な目的が、AI開発者やプロバイダーが潜在的な脆弱性を認識し、実用的なセーフガードを導入することを支援することであることを称賛する。このガイドラインが、プロンプト・インジェクションやサービス拒否（DoS）などの攻撃への対策に重点を置く一方、データ・ポイズニング、改ざんされたモデルの挿入、モデルの抽出などの脅威にも対応し、AI開発者やAIプロバイダーを支援するセキュリティ対策を提案していることを高く評価する。総務省には、AI資産の開発から配備までのライフサイクル全体におけるセキュリティ確保の重要性について、総合的に検討することを推奨する。 そのために、日本政府がAIの安全な実装に関するガイダンスを提供する際に活用できる追加的なリソースとして、シスコの統合AIセキュリティフレームワークを紹介したい。加えて、添付の別添に、本ガイドライン案及び別添案の特定の条項に関するコメントを記す。 ＜以下項目名＞ AIセキュリティのベースライン 断片的な状況と統合の必要性 AIリスクを理解するための新しいパラダイム AI脅威の統一された分類法 包括的なAIセキュリティへの提言 結論 我々は、現在利用可能な最も包括的かつ先進的なソリューションの1つとして、参考資料としてシスコの「統合AIセキュリティ＆セーフティフレームワーク」を提供するところである。AIが産業を変革し続ける中、このレベルの明確性を持つことは有益であるだけでなく、非常に重要である。例えば、このフレームワークは、シスコが提供するAI Defenseに組み込まれており、関連する指標や推奨される緩和策とともに脅威を特定することができる。このリソースが日本にとって有益なものとなり、AIコミュニティが理解を深め、AI関連リスクの新たな状況に対する防御を強化する一助となることを期待する。	賛同の御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
116	株式会社Acompany		–	本ガイドラインは、AIシステムに対する脅威とその対策例を体系的に整理した文書として高く評価いたします。 一方で、本ガイドラインで主な脅威や想定事例として挙げられているものは限定的であり、実際にはそれ以外の脅威や想定事例も多数存在します。そのため、本ガイドラインが継続的に更新され、より網羅的なガイドラインとして整備されることを強く期待いたします。 以下、具体的な追加検討事項として、想定される脅威・事例および対策例を提案いたします。 1. AI開発者とAI提供者間における基盤モデルの不正流用リスク 想定事例： AI開発者が開発するLLMモデル（特に基盤モデル）をAI提供者が利用する事例において、AI提供者がモデルの提供契約で定められた当該サービス以外の用途にLLMモデルを不正に流用する脅威が想定されます。 現状の対策と課題： 当該サービスにおける基盤モデルの利用範囲を明確化したモデル提供契約を締結することで対策することが一般的ですが、契約上の制約のみでは技術的な流用を完全には防止できません。 技術的対策： 機密コンピューティングという技術を利用すれば、AI提供者のシステム（特に基盤モデルを利用する部分）をTEE（Trusted Execution Environment）内で動作させ、AI開発者はそのシステムが「基盤モデルをシステム外に出力しないこと」を Remote Attestationという機能により検証することができます。これにより、AI開発者は「AI提供者が基盤モデルをシステム外に出力し別システムに流用すること」ができないことを検証することができ、技術的にも安全に基盤モデルを提供できます。 2. AI提供者による利用者プロンプトの不正閲覧リスク 想定事例： AI利用者とAI提供者が異なる組織に属する事例において、AI利用者がAI提供者のサービスにプロンプトを入力する際、技術的にはAI提供者側でプロンプトの内容を閲覧することが可能となります。よって、AI提供者の内部犯行により、利用者が入力したプロンプト（機密情報を含む可能性がある）が秘密裏に窃取・流用される脅威が存在します。 現状の対策と課題： サービス提供契約内において、秘密保持義務などにより当該サービスの提供以外にプロンプトを利用しないことを契約すると思われますが、契約上の制約のみでは悪意ある流用を防止できません。 技術的対策： 機密コンピューティング技術により、TEE内でプロンプトを秘匿化したまま処理することで、AI提供者側（システム管理者を含む）であっても処理中のプロンプト内容にアクセスできないようにすることが可能です。 3. 今後のガイドライン更新への期待 上記のように、本ガイドラインで現時点では取り上げられていない脅威や想定事例、対策例が存在します。AIシステムの利用形態や脅威の多様化に対応するため、幅広く脅威・想定事例・対策例を継続的に収集し、ガイドラインを適時更新していくことを強く期待いたします。	賛同の御意見として承ります。御提案の点については、今後の政策の検討にあたり、参考とさせていただきます。
117	個人		–	「AI エージェントは規制しない」とは どういうことか? 現に市場に出回って、被害が拡大している物を、「手が付けられない」と放置してどうする。 規制が追い付かないなら、一時的に (専門分野の研究開発が目的でない、汎用の) AI の使用を全面的に禁止して、きちんと法整備すべきだ。 AI の使用者全員を規制するなど、現実的でない。 AI 提供企業に規制をかけ、まず 被害を食い止めるのが先決だろう。 「経済優先」で、危険性の高い AI を このまま普及させる事は、後々の対策を 不可能にしてしまう。 また、著作権・肖像権などの問題も 全く解決していない。 まずは 生成 AI、エージェント AI の使用を 緊急に止めさせるべきだ。	御意見として承ります。本ガイドラインでは、本編1.2に記載のとおり、AIエージェントについては、技術が急激な発展の途上にあり、これに特有の脅威や対策を安定的に確定することが現時点では困難であることから、対象外としています。本編「本ガイドラインの策定の背景等」に示している考え方のとおり、今後、AIの技術進展を十分に踏まえ、新たな脅威と対策の動向を注視し、必要に応じて、対応を検討してまいります。
118	個人		–	本ガイドラインがLLMの活用を促進するようなガイドラインという点が問題だと考えています。 まず、オプトイン方式を取り入れていない大規模言語モデルのLLM及びLLMを構成要素に含むAIシステムは活用することのみでも知的財産権を侵害する恐れがあります。なぜなら、海外大手生成AI運営会社はLLMの開発の時点で、違法なオンライン海賊版ライブラリから、児童ポルノ・犯罪コンテンツ・著作物などの大量のデータを権利処理なしに取り込んでいることが判明しているからです。さらに、LLMは海賊版ライブラリに取り込まれている元の著作物を出力物に再現することが可能だという大学の論文が出ています。権利侵害をされている大手企業及び作家からの大手生成AI企業に対する訴訟は何件も起きています。 また、LLMを利用することによって、ヒトの脳の機能が低下するという大学の論文が出ています。LLMを日本全体で活用することになれば、日本人全体の学力の低下が見込まれます。日本全体の学力の低下によって、日本企業が規制をしている海外から遅れを取り、低迷していくことになると予測されます。大学の論文やレポート等にも生成AIで出力した生成物を提出している学生も多くいます。 さらに、企業等がLLMに顧客等の個人情報を入力すると、個人情報が漏洩する恐れがあります。2023年には、大手生成AIツールにおいて、システムバグにより、他のユーザーの個人情報（氏名・メールアドレス・クレジットカードの番号）を閲覧できる状態が何件か発生しています。このような事件は何件か発生しており、個人情報保護法20条に違反しています。企業がLLMに企業機密事項等を入力すると、LLMに企業機密事項等がインプットされてしまいます。目の届かないところで、インプットされた企業機密事項等がそのまま出力される恐れがあります。 このように、LLMの活用は知的財産権の侵害及び個人情報保護法20条に違反する恐れがあります。そのため、日本において、LLMの使用を禁止する法律の整備が現在極めて必要とされています。EU、アメリカ、中国等では、法律によるLLMの規制が始まっています。先日韓国では、生成AI（LLM）の生成物には生成AIの生成物だと分かるようなマークの表示が義務づけられるようになりました。LLMにはこのように世界的な対応がされており、危険性のある生成AI（LLM）は法律による規制が必要だということが世界的に理解されてきています。海外では、既にLLMによる様々な危険な事件が起こっています。韓国では、卒業アルバムの写真を使用し、LLMで出力した性的ディープフェイク写真がSNSに流出しました。アメリカでは、少年が大手生成AI（LLM）ツールを用い、相談相手として使用していたところ、生成AIは少年に自殺の方法及び遺書の書き方などを教え、自殺してしまったという事件が発生しました。両親は少年の自殺をめぐり、大手生成AI会社を訴訟しました。このような事例は何件もあります。 このような理由から、大規模言語モデルのLLM及びLLMを構成要素に含むAIシステムを規制する内容のガイドラインを変更することを提案します。 具体的には、内閣府知的財産戦略推進事務局がパブリックコメントを募集していた「生成AIの適切な利活用等に向けた知的財産の保護及び透明性に関するプリンシプル・コード（仮称）（案）」のような内容に変更する必要があります。 内閣府知的財産戦略推進事務局が提案している案のような内容への変更が行えない場合は、以下のような提案があります。 大規模言語モデルのLLMとは異なるオプトイン方式を用いているAIのセキュリティを守るガイドラインへの変更です。オプトイン方式のAIは権利問題をある程度クリアした上で運用されています。そのようなAIに入力されている個人情報を守るためのセキュリティ対策をガイドラインで提示することを提案します。海外競争に劣らないオプトイン方式を用いたクオリティの高い日本発のAI技術及び技術者を守る必要があります。	御意見として承ります。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
119	個人		–	生成AIとその開発会社、開発者、使用者を厳しく規制し、取り締まってください。 生成AIによって起こっている被害は多々あり、データセットには膨大な数の著作物や肖像物が無断利用されており、デフォルトで児童ポルノまで含まれています。 そしてそれらが使用した時点で自動的に混ぜられる仕様になっているだけでなく、使用した人の誰かが無断で著作物や肖像物を学習させてしまい、更に被害が拡大しています。 しかし、権利者が訴えたり通報しようにも、数が多すぎる事と明確な法規制が無いため、権利者にとってはあまりに不利な状況であり、負担と被害が増える一方です。 そのため、厳しい規制と取り締まり、それに伴う対策が必要です。 まず、学習には権利者の明確な許可を必要とし、拒否されたり連絡が取れなかったりした場合は学習できないようにしてください。 また、学習する際は権利者側の提示する金額を支払い、明確な対価を払わせてください。 そしてそれを既に学習されている著作物や肖像物などの権利者全てに行わせてください。 その際、これまで無断で学習させていたデータは全て削除すると共に、これまで無断利用してきた分の金額を全ての権利者に支払わせてください。 次に、学習したデータセットは透明性の確保のため全て公開させてください。 そしてこれらに違反した場合は厳罰に処してください。	御意見として承ります。
120	個人		–	提出意見： 「安全・安心で信頼できるAIを実現するためのルール作りを日本が主導している」とありますが、全くの間違いです。 生成AI技術は、他者の著作物を無断で利用することで成り立つ「窃盗ツール」です。 今までに何度もパブリックコメントを送っていますが、何故かこのことは全くと言っていいほど反映されません。 日本がしていることは後退です。主導ではありません。また、自ら自分達の文化を世界に差し出して有難がっている状態になっています。 日本国内には多くの反日勢力がおり、政府や重要ポストにもそういう方々が多いと聞いています。そのためか、日本のためになることはなかなかしようとしてくれないように感じます。 OpenAI自体が裁判沙汰になったり、生成AI技術の著作権侵害について内部告発した職員が退職後に自殺するなど、信頼のおける企業とはとても思えません。そもそも、社で「著作権を侵害した」ということを認めています。 そのような技術を使って、何を先導するというのでしょうか。犯罪者の助けにしかならないと思います。 これ以上、我々の文化を貶めるのをやめて下さい。生成AI被害は詐欺、性犯罪など多岐にわたっています。 今すぐに生成AI規制のための法整備をし、世界に誇る我々日本人の文化をどうか守ってください。	御意見として承ります。
121	個人		–	AIの脅威は、すぐそこまで来ている。生半可な対策では、人類はやられる。	御意見として承ります。
122	個人		–	資料に掲げられているAI自体が巨大なセキュリティホールそのものです。 このようなものをシステムに組み込まないでください。 重大なエラーやセキュリティリスクが発生してからでは遅いです。	御意見として承ります。
123	個人		–	生成AIに関してのみの話ではあるが、現段階の生成AIは出現してから現在にいたるまで目ぼしい成果もあげていない。 これが停止したとして社会経済活動に影響がでるとは考えにくい。むしろ権利関係の杜撰さでトラブルを起こしているのが現状であるし、海外ではそれが原因となり訴訟を起こされている。AI技術におけるデータの取り扱いは慎重であるべきだが生成AIに関しては杜撰であると言っても過言ではない。 「不正操作による機密情報の漏洩」を心配するよりも、特殊な操作をせずとも問題のある出力する可能性のある生成AI事態の管理を徹底すべきと考える。	御意見として承ります。
124	個人		–	AIそのものがセキュリティに問題があるものだと思うのでAIを使わない選択をして欲しいです。	御意見として承ります。
125	個人		–	AIセキュリティガイドライン案を通じた技術対策強化と雇用喪失救済の提案 ガイドライン案を支持しますが、AIセキュリティ対策を進める中で、技術的リスクだけでなく社会的リスク（雇用喪失）への救済も考慮すべきです。 AI普及でデータ漏洩・悪用リスクが増大しますが、対策が企業・政府のデータ集中を助長し、過剰監視やプライバシー侵害を招く恐れがあります（2025年AI悪用事件増加）。一方、AIによる職喪失（2030年までに8,500万人の仕事消失可能性、WEF報告）が深刻で、低所得層・中高年の再就職難を加速します。ガイドラインに「個人情報最小化原則の義務化」と「第三者監査必須化」を追加し、社会的リスクとして「AI導入時の雇用影響評価と再教育支援義務」を明記してください。これで、安全で公平なAI社会を実現できます。ガイドライン案に反映を求めます。	賛同の御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。

項番	意見提出者	文書	該当項目	御意見の詳細	御意見に対する考え方
126	個人		–	AIシステムの安全性およびセキュリティ確保の観点から、AIが処理を停止し人手対応へ移行する場合について、停止理由を構造化された形で記録・保持することを推奨事項として明示すべきと考えます。 特に、論理的不整合の検出や、想定外入力の発生など、システム上の明確な停止条件が存在する場合、それらを単なるエラーとして扱うのではなく、事後の監査・原因分析に活用可能な形式で保持することが重要です。 これにより、AIの誤動作と適切な判断停止を区別でき、システム全体の信頼性向上につながります。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
127	個人		–	本意見は、「AIのセキュリティ確保のための技術的対策に係るガイドライン（案）」について、技術的整理の有用性を前提としつつ、制度上の位置づけおよび運用上の前提条件の明確化を求めるものである。 本ガイドラインは、AIシステムの安全確保に関する技術的対策を体系的に整理しており、実務上の参考資料として一定の意義を有する。一方で、これらの技術的対策が、制度上どのような位置づけで提示されているのかについては、本案から明確に読み取ることができない。 具体的には、本ガイドラインが、任意の参考指針に留まるのか、事実上の標準としての位置づけを想定しているのか、あるいは将来的な制度化や規制対応の前段として整理されているのかが不明確である。また、記載されている技術的対策が、開発者、提供者、利用者のいずれに対して、どのフェーズ（設計、提供、運用等）で期待されるものなのかについても、主体ごとの整理が示されていない。 このような制度的位置づけや責任主体が明確でないままでは、事業者にとって、本ガイドラインに基づく対応が努力目標に留まるのか、実質的な遵守要件として評価され得るのかを判断することが困難となる。その結果、過剰な対応による負担増、あるいは形骸化のいずれかに陥るおそれがある。 国際的には、AIに関するセキュリティ対策は、リスク区分、責任主体、監督や評価の枠組みと接続された形で整理されつつある。本ガイドラインにおいても、技術的対策を列挙するにとどまらず、それらが制度上どの層に位置づけられるのかを明示することが、事業者の予見可能性を高め、実効性を確保する上で重要であると考える。 以上を踏まえ、本ガイドラインについて、1.制度上の位置づけ（任意指針か、事実上の標準か等）、2.技術的対策と責任主体・適用フェーズとの関係、3.将来的な他のAI関連制度や規制との接続の考え方について、可能な範囲で整理・明示されることを求める。	御意見として承ります。本ガイドラインは、本編3.1に記載のとおり、AIに対する脅威のリスクを低減するため、現時点で取り得るとされる一般的な対策例を整理し、提示するものです。また、本ガイドラインは、本編3.1に記載のとおり、ある脅威に関する責任主体を決定する趣旨で記載するものではなく、各組織が自らの状況に応じて合理的な対策を選択するための指針として提供するものです。
128	個人		–	昨今の情勢から、生成AIを安全に利活用するための対策を講じる必要性が高まっていると感じており、今回のような技術的な対策を示すことについて賛同いたします。 しかしながらガイドラインという強制力のないものになっているため、今利用することのできる生成AIが対策を行っているのかどうかが不透明であり、安心して使うことができない懸念があります。 企業に対策を義務付ける、もしくはガイドラインの対策を満たしている生成AIをリストアップして公表するべきではないでしょうか。 また、現状リスクがあるとわかっていながらも対策が困難なものに関しては、サービスの提供を規制するか問題が起こった際に企業に責任を取らせるようにするべきです。 リスクがあるものを対策もなしに普及させるべきではありません。	賛同の御意見として承ります。今後の政策の検討にあたり、参考とさせていただきます。
129	個人		–	このようなガイドラインの導入を全面的に支持致しますと共に、内容の更なる充実と改善を期待します。 私はAI関連事業者でなく、著作物等の権利者ですのでこちらの文書の想定読者からは外れます。 自身の著作物に酷似した画像や文章の出力をAIサービス提供者の側で防止していただくことで権利侵害をされる可能性が減ること、私自身が望まない形で他者によって入力された個人情報の漏洩の防止のためにも、このような取り組みは非常に有意義なものと感じます。 しかし、例えば社内や、特に行政府などの公的機関で使う生成AIサービスについては、オンラインでの使用を回避する形を推奨するような何かがあってもよろしいのではないかと感じます。 また、情報漏洩でよくあるパターンは外部からの攻撃など技術的な面よりも、使用者側が現実でパスワードを書いた紙をそのままゴミに出してしまうなどの物理的なもの、生成AIサービスで言えば「入力してはいけない情報の入力」などが多いと感じております。 意図しない外部からの攻撃対策と同時に、使用者の行動自体を縛るような技術的措置（入力できない文章の設定など）、また物理的に避けるべき行動の推奨を国主導で大々的にしていただけると権利者といたしましても安心できます。 データポイズニング攻撃、細工をしたモデルの導入を通じた攻撃、について、こちらは「オプトイン方式」にすることで根本的に回避可能かと考えますので、こちらの推奨をよろしくお願いいたします。	賛同の御意見として承ります。本ガイドラインでは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。
130	個人		–	AIセキュリティ確保のために個人情報の扱い緩和は勘弁してください。	御意見として承ります。本ガイドラインは個人情報の取扱いに係る規律の緩和を行うことを目的としたものではありません。
131	個人		–	この意見は、日本国憲法に則り「AIのセキュリティ確保のための技術的対策に係るガイドライン」(以下ガイドライン)の策定において財産権の保護と尊重を重視した文言へ改め、内容を検証することを強く望むものです。「脅威に対する対策」を奨励しつつも「知的財産に関わる財産権の保護・救済の放棄」を徹底的に回避し、財産権、ひいては人権の保障と保護を実現するために、生成AIが抱える権利侵害の諸問題に対し誤った認識を政府機関が拡散することのないよう細心の注意を払ってガイドラインの策定にあたってください。 ==前提== 日本国憲法第十一条、第十二条ならびに第二十九条に基づき、ガイドライン（案）が財産権、とりわけ知的財産に関わる財産権の保護と尊重に寄与するよう全体の再検証を求めます。 ガイドラインの想定がAIのセキュリティ確保とされており、知的財産の権利保護については掘り下げているものではありませんが、日本国憲法のもと、個人の権利の保障や保護の妨げとなるような振る舞いを政府機関が行うことは決して許されるものではありません。他者の権利を侵害する行為を推奨しないこと、あるいは権利侵害の被害を救済する妨げをしないことはガイドライン策定においても必ず熟慮されるべきことです。 この意見では、現在のガイドライン（案）において文言の修正が必要か、項目の追加の必要性が極めて高いものについて指摘します。しかし、この意見に含まれない他の項目についても、生成AI開発者や生成AI提供者が財産権を軽視する態度をとる可能性がないか、また、財産権侵害の被害救済の妨げにならないか、細部にわたって見直しを行ってください。	御意見として承ります。本ガイドラインは、本編1.1に記載のとおり、AIの「セキュリティ確保」として、「不正操作による機密情報の漏えい、AIシステムの意図せぬ変更や停止が生じないような状態」に対する脅威への対策を主な対象とし、この観点から脅威への技術的対策例を整理するものです。なお、他のガイドラインとの関係は、表１に掲げるとおりです。
132	個人		–	注意事項に同一内容の意見が多数あっても考慮の対象では無いとありますが、 色々な方が生成AIの被害に遭い禁止して欲しい、罰則を設けて欲しいと声を上げる中、上記のひと言で黙殺するのはどうかと思います。 内容としては基本的には生成AIを活用することは決まっており、その決まりに対して何か付け足しなどがあるか、という確認程度だと思いますが、せめてきちんとこのような意見があったと目を通し保管をしてください。	御意見として承ります。
133	個人		–	また、別件のパブリックコメントでも申し上げましたが、罰則の無いガイドラインでは守る意味が無いと放棄する人達がほとんどだと思われます。 こういった対策のできない業者のシステムを使わせない、また、併せて罰則規定を制定してください。	御意見として承ります。
134	個人		–	機械学習対策処理についてはどうお考えで ハルシネーション問題とAIのブラックボックス問題という原因解明が困難な問題を抱えているのにセキュリティ確保は出来るのでしょうか？ https://medical-saponet.mynavi.jp/news/newstopics/detail_271/	本ガイドラインでは、本編3.1に記載のとおり、AIの性質上、脅威を生じさせる要因等を完全に排除することは困難である点について留意が必要であり、単独の実施により脅威を生じさせる要因を排除することは困難な場合があることを前提に、複数の対策を講じることでリスクを低減していくことを想定しています。