# EUのAI法の汎用目的AIモデル実践規範について

2025年11月18日

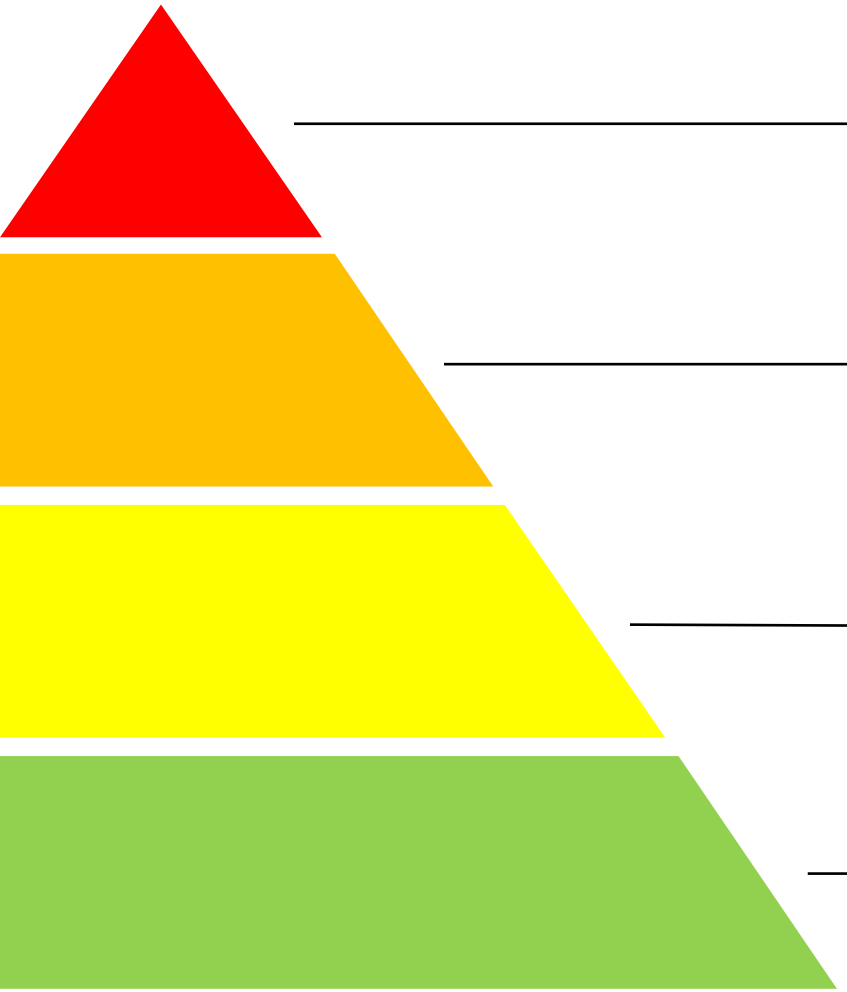
（株）国際社会経済研究所　小泉 雄介

© Institute for International Socio-Economic Studies 2025

# EUのAI法

IISE

- AI法では、リスクベースアプローチ、すなわち（AIを一律に規制するのではなく）AI機器・サービスがもたらすリスクに応じて規制をかけるアプローチが取られている。欧州委員会によれば、ほとんどのAIシステムは④に相当。



①許容できないリスクのAI（禁止されるAI）
例：社会的スコアリング
→ 禁止

②ハイリスクAI
例：雇用、医療機器
→ 一定要件と事前適合性評価の遵守の下で許可

③透明性の義務のあるAI
例：「人間になりすます」チャットボット
→ 透明性義務の下で許可

④最小限リスク、またはリスクのないAI
例：スパムフィルター
→ 許可

①～④とは別軸として、以下の基盤モデルも規制

⑤汎用目的AIモデル（GPAIモデル）
（累積学習計算量が10の23乗FLOPを超えるモデル）

⑥システミックリスクのあるGPAIモデル
（累積学習計算量が10の25乗FLOPを超えるモデル）

出典：欧州委員会資料を和訳、加筆


© Institute for International Socio-Economic Studies 2025

# AI法における汎用目的AIモデルのプロバイダーの義務

IISE

- 汎用目的AIモデル（GPAIモデル）のプロバイダーには、以下の義務が課される。（第53条第1項）
  - （a）学習およびテストのプロセスと評価の結果を含む当該モデルの技術文書を作成し、最新の状態に保つ。これには、要求に応じてAIオフィスおよび国内所轄機関に提供する目的で、少なくともAnnex XIに規定されている情報を含める。
  - （b）汎用目的AIモデルを自らのAIシステムに組み込むことを意図するAIシステムの下流プロバイダーに対して、以下のような情報と文書を作成し、最新の状態に保ち、利用可能とする。
    - （i）AIシステムのプロバイダーが当該モデルの能力と制限を十分に理解し、本規則に基づく義務を遵守できるようにする。
    - （ii）少なくとも、Annex XIIに規定されている要素を含む。
  - （c）著作権および関連する権利に関するEU法を遵守し、特に指令（EU）2019/790（著作権指令）の第4条第3項に基づいて表明された権利の留保を、最先端の技術を用いる等により、特定し遵守する方針を導入するための方針を導入する。
  - （d）AIオフィスが提供するテンプレートに従って、汎用目的AIモデルの学習に使用されるコンテンツについての十分に詳細な概要を作成し、公開する。

- 上記（a）および（b）の義務は、無料のオープンソースライセンスの下でリリースされており、重みなどのパラメータ、モデルアーキテクチャに関する情報、モデルの使用法に関する情報が公開されているAIモデルのプロバイダーには適用されない。ただし、この例外はシステミックリスクのある汎用目的AIモデルには適用されない。（第53条第2項）
- 汎用目的AI モデルのプロバイダーは、整合規格が発行されるまで、義務の遵守を実証するために、実践規範に依拠できる。欧州整合規格を遵守すると、その規格が義務をカバーする範囲で、プロバイダーに適合性の推定が認められる。承認された実践規範に準拠していない、または欧州整合規格を遵守していないプロバイダーは、欧州委員会による評価のために、遵守の十分な代替手段を実証しなければならない。（第53条第4項）


© Institute for International Socio-Economic Studies 2025

# AI法におけるシステミックリスク汎用目的AIモデルのプロバイダーの義務

IISE

- システミックリスクのある汎用目的モデルのプロバイダーには、（前述の汎用目的AIモデルのプロバイダーの義務に加えて）以下の義務も課される（第55条第1項）。
  - （a）システミックリスクの特定と軽減を目的とした当該モデルの敵対的テストの実施と文書化を含め、最新技術を反映した標準化されたプロトコルとツールに従ってモデル評価（model evaluation）を実行する。
  - （b）システミックリスクのある汎用目的AIモデルの開発、市場投入、または使用に起因する可能性のあるシステミックリスクを、その発生源を含めてEUレベルで評価（assess）し、軽減する。
  - （c）重大インシデントとそれらに対処するための是正措置に関する関連情報を追跡し、文書化し、不当な遅滞なくAIオフィス、および必要に応じて国内所轄機関に報告する。
  - （d）システミックリスクのある汎用目的AIモデルおよび当該モデルの物理インフラストラクチャに対して十分なレベルのサイバーセキュリティ保護を保証する。

- システミックリスクのある汎用目的AI モデルのプロバイダーは、整合規格が発行されるまで、上記の義務の遵守を実証するために、実践規範（codes of practice）に依拠できる。欧州整合規格を遵守すると、その規格が義務をカバーする範囲で、プロバイダーに適合性の推定が認められる。承認された実践規範に準拠していない、または欧州整合規格を遵守していないシステミックリスクのある汎用目的AIモデルのプロバイダーは、欧州委員会による評価のために、遵守の十分な代替手段を実証しなければならない。（第55条第2項）


© Institute for International Socio-Economic Studies 2025

# 【ご参考】AI法におけるシステミックリスク汎用目的AIモデルの分類基準

IISE

- システミックリスクのある汎用目的AIモデルとしての分類（第51条）
  - 汎用目的AIモデルは、以下のいずれかの条件を満たす場合、システミックリスクのある汎用目的AIモデルとして分類される。（第1項）
    - （a）当該汎用目的AIモデルが、指標およびベンチマークを含め、適切な技術ツールおよび方法論に基づいて評価された、ハイインパクトな能力を備えている。
    - （b）職権による、または科学パネルからの適格な警告を受けた欧州委員会の決定に基づき、当該汎用目的AIモデルが、Annex XIIIに定められた基準を考慮して、（a）に規定されたものと同等の能力または影響を有する。
  - 汎用目的AIモデルは、浮動小数点演算（FLOP）で測定された、学習に用いられた累積計算量が10の25乗を超える場合、上記（a）に従ってハイインパクトな能力を備えていると推定される。（第2項）

- ハイインパクトな能力、システミックリスクの定義
  - 「ハイインパクトな能力（high-impact capabilities）」とは、最先端の汎用目的AIモデルが記録している能力と同等またはそれを超える能力を意味する。（第3条（64））
  - 「システミックリスク（systemic risk）」とは、汎用目的AIモデルのハイインパクトな能力に特有のリスクであって、そのリーチによって、または公衆衛生、安全、治安、基本的権利、もしくは社会全体に対する実際もしくは合理的に予見可能な悪影響によって、EU市場に重大な影響を与え、かつバリューチェーン全体に大規模に伝播しうるものを意味する。（第3条（65））


© Institute for International Socio-Economic Studies 2025

# 汎用目的AIモデルの実践規範

IISE

- 汎用目的AIモデル（GPAIモデル）の実践規範（Code of Practice）の最終版は作業グループによる作成が遅延していたが、2025年7月10日に欧州委員会に提出され公表。欧州委員会は8月1日、GPAIモデル関連条項の適用開始日(8/2)の前日に、実施法令による十分性決定を通じて実践規範の最終版を承認。同日、欧州AI会議も最終版を承認。
  - https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai
  - 2025年5月2日が法定の発行期限であったが、POLITICO記事によれば、遅延の理由の一つは米国ビッグテックやロビイストが表明してきた懸念（実践規範がAI法の内容を超えている）をバックに米国政府（駐EU米国政府代表部）が欧州委員会に4月下旬に宛てた書簡で実践規範案に欠陥があると伝えたことであるという。
- 実践規範は「透明性」「著作権」「安全性とセキュリティ」の3つの章から構成。
  - 「透明性」と「著作権」の章は、GPAIモデルのプロバイダー全般向けのものであり、AI法第53条の義務（技術文書の作成、下流プロバイダーへの情報提供、著作権法の遵守、学習データ詳細概要の公表）の遵守を証明する方法を記載。
  - 「安全性とセキュリティ」の章は「システミックリスクのあるGPAIモデル」のプロバイダー向けのものであり、AI法第55条の義務（(a)敵対的テスト・文書化含むモデル評価、(b)システミックリスクの評価と軽減、(c)重大インシデントの報告、(d)モデルとインフラのサイバーセキュリティ）の遵守を証明する方法を記載。
    - システミックリスクの4類型として「化学・生物・放射線・核（CBRN）」「制御の喪失」「サイバー攻撃」「有害な操作」を提示。
- GPAIモデルのプロバイダーは実践規範に自発的に署名し、自発的に準拠することにより、AI法への遵守を証明することができる。実践規範に準拠しないGPAIモデルのプロバイダーは、欧州委員会による評価のために、遵守の十分な代替手段を証明しなければならない。
- 実践規範に署名した企業はAnthropic、Amazon、Google、IBM、Microsoft、Mistral AI、OpenAIなど27社（2025年8月1日時点では26社）。Metaは署名しておらず、xAIは「安全性とセキュリティ」のみに署名 。Deepseekなどの中国企業も署名していない。また日本企業で署名した企業はない。

© Institute for International Socio-Economic Studies 2025

# 【ご参考】汎用目的AIモデルのガイドライン

IISE

- 汎用目的AIモデル（GPAIモデル）の実践規範を補完するものとして、欧州委員会は2025年7月18日に「AI法で規定された汎用目的AIモデルの義務の範囲に関するガイドライン」（https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act）を公表。拘束力を持つものではないが、GPAIモデルの監督・執行を担う欧州委員会が、AI法の下でのGPAIモデルのルールの解釈と適用について指針を提供するもの。要点は以下3つ。
  - 明確な定義：　AIモデルが「GPAIモデル」とみなされる条件（累積学習計算量が$10^{23}$を超え、かつ言語生成、テキストから画像生成、またはテキストから動画生成が可能なモデル）など、明確な技術的基準を提示。開発者がAI法の義務を遵守する必要があるか否かの理解に役立つ。
  - 実用的なアプローチ：　例えば、（GPAIモデルのプロバイダー以外では）AIモデルに大幅な変更を加える者のみがGPAIモデルプロバイダーの義務を遵守する必要があり、軽微な変更を加える者はGPAIモデルプロバイダーの義務を遵守する必要がないことを明確化。
  - オープンソースの免除：　透明性とイノベーションを促進するために、オープンソース AI モデルのプロバイダーがどのような条件下で特定の義務を免除されるかを明確化。


© Institute for International Socio-Economic Studies 2025

# 汎用目的AIモデルの実践規範：「安全性とセキュリティ」の章の構成

IISE

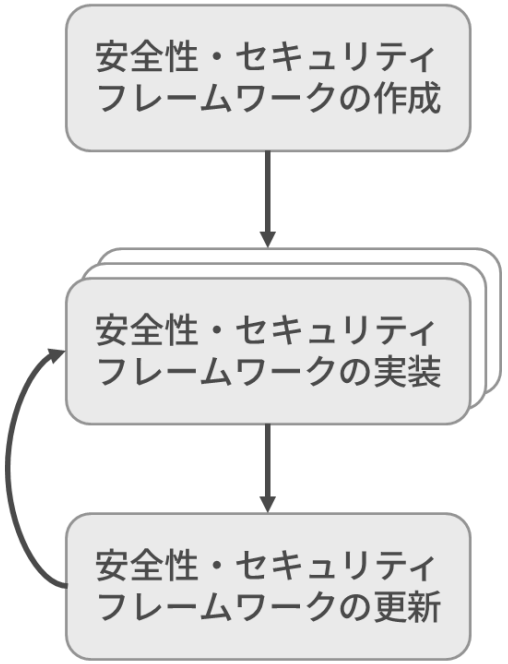
- 目的
- 前文
- コミットメント１　安全性・セキュリティフレームワーク　→(a) 敵対的テスト・文書化含むモデル評価
- コミットメント２　システミックリスクの特定
- コミットメント３　システミックリスクの分析
- コミットメント４　システミックリスクの受容の決定　→(b) システミックリスクの評価と軽減
- コミットメント５　安全性のための軽減策
- コミットメント６　セキュリティのための軽減策　→(d) モデルとインフラのサイバーセキュリティ
- コミットメント７　安全性とセキュリティに関するモデルレポート
- コミットメント８　システミックリスクの責任分担
- コミットメント９　重大インシデントの報告　→(c) 重大インシデントの報告
- コミットメント10　追加的な文書化と透明性
- 用語集
- 付録１　システミックリスクおよびその他の考慮事項
- 付録２　同様に安全な、またはより安全なモデル
- 付録３　モデル評価
- 付録４　セキュリティ軽減策の目標と措置


© Institute for International Socio-Economic Studies 2025

# コミットメント１　安全性・セキュリティフレームワーク（1/4）

IISE

- 署名企業は、最先端の安全性・セキュリティフレームワークを導入する必要がある。フレームワークは、システミックリスク管理プロセス・措置を概説した文書。フレームワーク導入プロセスは以下3つから成る。
  - (1) 安全性・セキュリティフレームワークの作成
  - (2) 安全性・セキュリティフレームワークの実装
  - (3) 安全性・セキュリティフレームワークの更新
- また、署名企業は安全性・セキュリティフレームワークをAIオフィスに通知する必要がある。




- 最先端の安全性・セキュリティフレームワークを作成する
- 一定の内容をフレームワークに含める
- AI法第52条1項に基づき欧州委員会に通知してから4週間以内、かつモデルを市場投入する2週間前までにフレームワークをコンファームする

- モデルのライフサイクル全体を通じて継続的にフレームワークを実装する
- モデルの市場投入前とモデルレポート更新時に、完全版システミックリスク評価および軽減プロセスを実施する
- 継続的にシステミックリスクの評価と軽減を行う（適切なトリガーポイントにおける簡易なモデル評価を含む）

- 適切にフレームワークを更新する
- フレームワークの十分性やフレームワークの遵守が著しく損なわれた、もしくは損なわれると判断する合理的根拠がある場合、またはモデルの市場投入から12ヶ月ごとに、いずれか早い方のタイミングで、フレームワーク評価を実施する

図の出典：汎用目的AIモデル実践規範に記載の図を和訳


© Institute for International Socio-Economic Studies 2025

# コミットメント１　安全性・セキュリティフレームワーク（2/4）

IISE

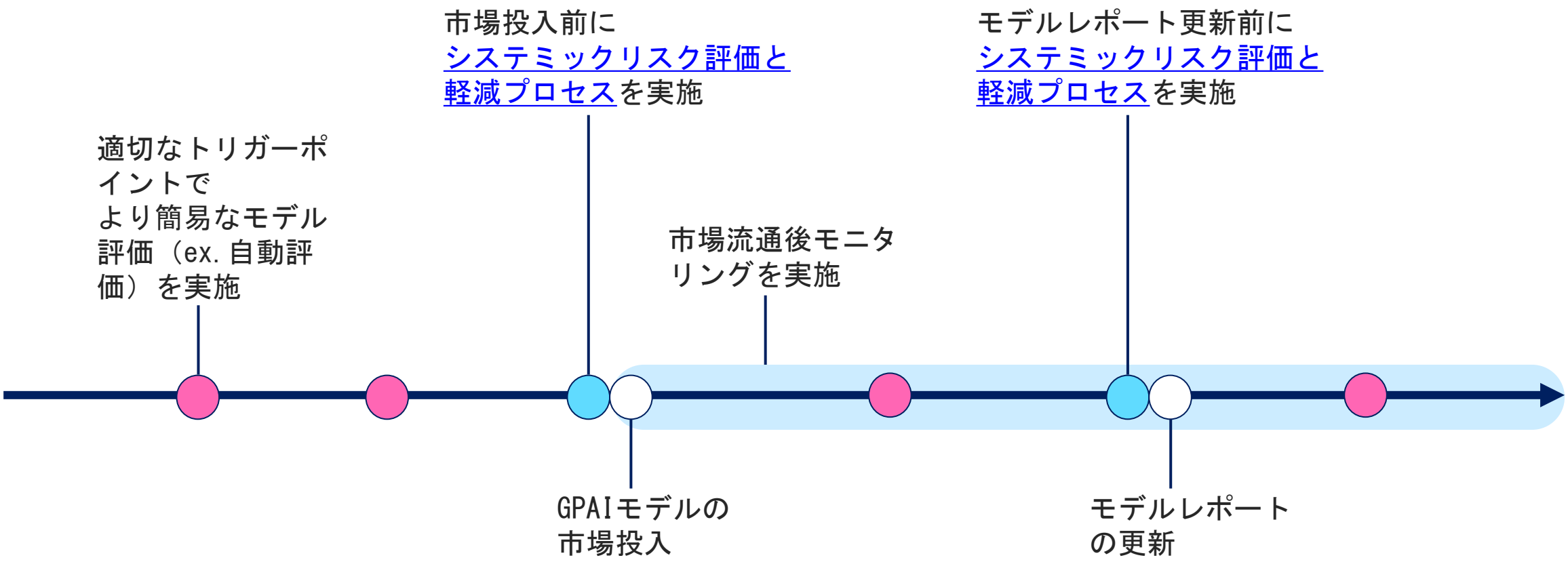
- 安全性・セキュリティフレームワークに含めるべき内容は以下。
  - (0) システミックリスクの評価と軽減のために実施済みおよび計画中のプロセスと措置の概要。
  - (1) 署名企業が、モデルのライフサイクル全体にわたって、より簡易なモデル評価を実施するトリガーポイントとその使用方法の説明と正当性。
  - (2) 署名企業が、システミックリスクが受容可能かどうかを判断するための、以下の事項：
    - (a) システミックリスク階層（systemic risk tiers）およびその使用方法を含む、システミックリスクの受容基準の説明と正当性。
    - (b) 各システミックリスク階層に到達した場合に署名企業が実施する必要がある安全性・セキュリティ措置の概要。
    - (c) システミックリスク階層を定義した各システミックリスクについて、署名企業が合理的に予見し、既存モデルが既に到達している最高のシステミックリスク階層を超えるモデルを保有する時期の推定。
    - (d) 独立した外部評価の結果以外で、政府を含む外部関係者からの情報が、署名企業のモデルの開発・市場投入・使用の進行に影響を与えるか否か、また影響を与える場合、どのようなプロセスによるかについての説明。
  - (3) システミックリスクを評価・軽減するプロセスにおけるシステミックリスクの責任分担方法についての説明。
  - (4) 署名企業がフレームワークを更新するプロセスの説明。


© Institute for International Socio-Economic Studies 2025

# コミットメント１　安全性・セキュリティフレームワーク（3/4）

IISE

- システミックリスク評価、より簡易なモデル評価の実施タイミング




図の出典：汎用目的AIモデル実践規範に記載の図を和訳

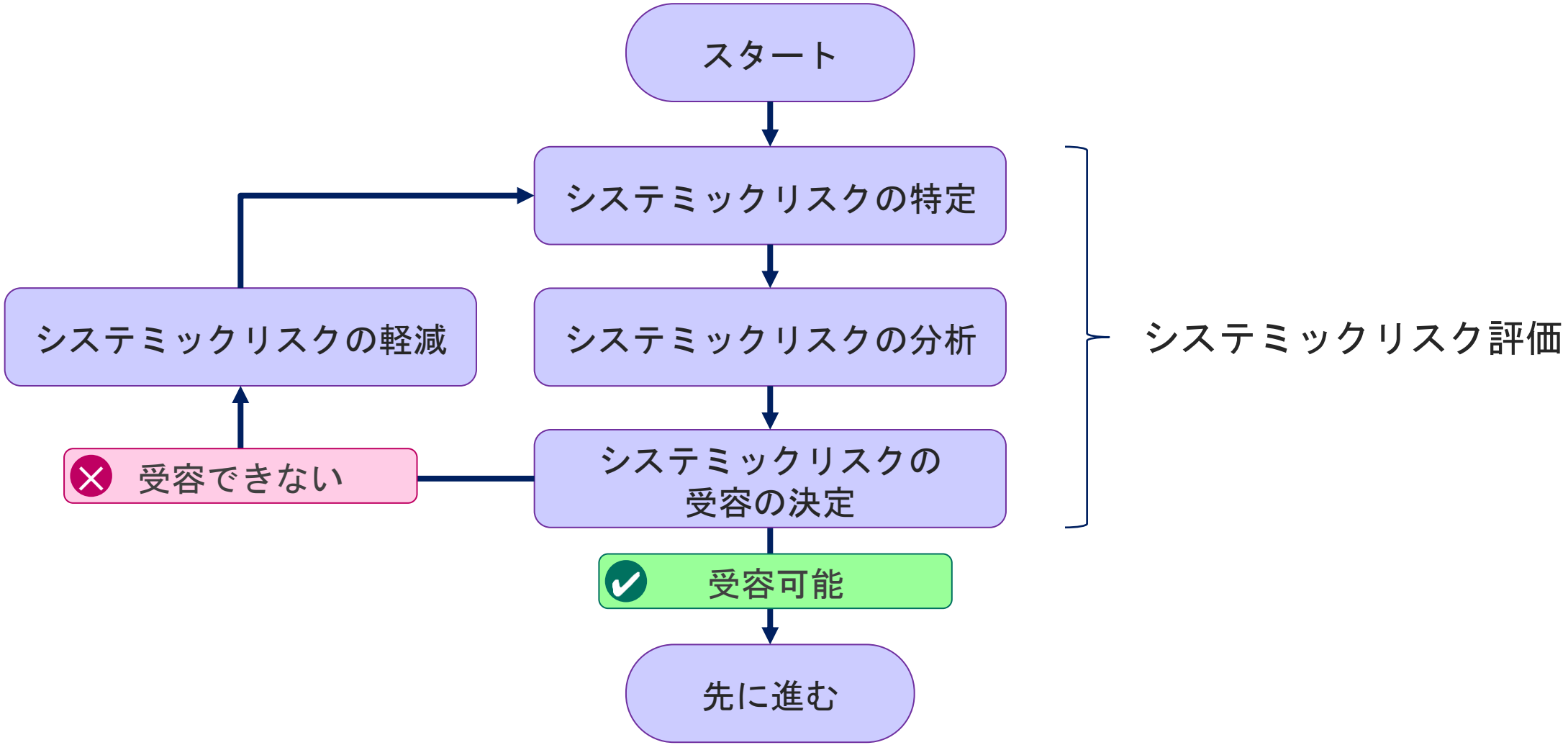


© Institute for International Socio-Economic Studies 2025

# コミットメント１　安全性・セキュリティフレームワーク（4/4）

IISE

- システミックリスク評価のプロセス




図の出典：汎用目的AIモデル実践規範に記載の図を和訳



© Institute for International Socio-Economic Studies 2025

# コミットメント２　システミックリスクの特定（1/3）

IISE

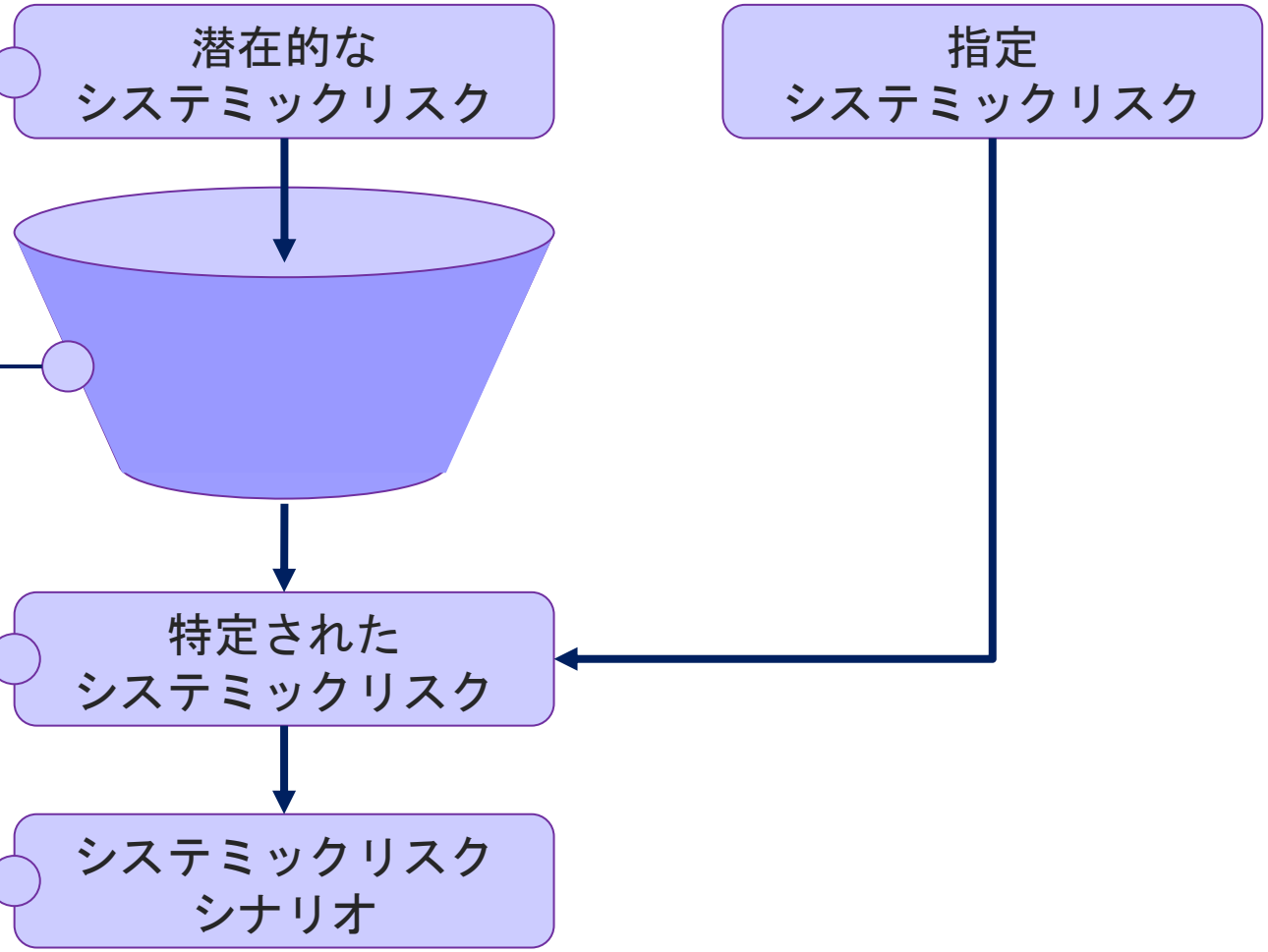
- システミックリスクを特定するプロセス

モデルに起因し、システミックリスクとなる可能性のある**リスクのリストを作成**する

それらの**リスクの関連特性を分析**する（ex.　性質や発生源）

モデルに起因する**システミックリスクを特定**する

特定された各システミックリスクについて**適切なシステミックリスクシナリオ**を策定する




図の出典：汎用目的AIモデル実践規範に記載の図を和訳



© Institute for International Socio-Economic Studies 2025

# コミットメント２　システミックリスクの特定（2/3）

IISE

- リスクの種類
  - システミックリスクを特定する際、以下の種類のリスクがシステミックリスクになりうる。
    - (1) 公衆衛生に対するリスク
    - (2) 安全に対するリスク
    - (3) パブリックセキュリティに対するリスク
    - (4) 基本的権利に対するリスク
    - (5) 社会全体に対するリスク
  - 上記の種類のリスクに基づき指定されたシステミックリスク（指定システミックリスク）のリストは次頁。

  - その他、上記の種類のリスクに該当しうるシステミックリスクの例は、以下のとおり。
    - 重大事故のリスク。
    - 重要なセクターまたはインフラ、公衆衛生、表現の自由と情報の自由、差別の禁止、プライバシーと個人データの保護、環境、人間以外の福祉、経済安全保障、民主的プロセスに対するリスク
    - 権力の集中や、違法・暴力的・憎悪的・過激化・虚偽のコンテンツによるリスク（児童性的虐待コンテンツ（CSAM）および合意のない親密な画像（NCII）によるリスクを含む）。


© Institute for International Socio-Economic Studies 2025

# コミットメント２　システミックリスクの特定（3/3）

IISE

- 指定システミックリスク（Specified systemic risks）

| 指定システミックリスク | 説明 |
|---|---|
| 1. 化学・生物・放射線・核（Chemical, biological, radiological and nuclear） | 化学・生物・放射線・核（CBRN）の攻撃または事故を可能にすることによるリスク。これには、関連する兵器や素材の設計・開発・取得・リリース・配布・使用において、悪意のある行為者の参入障壁を大幅に低下させること、または達成される潜在的な影響を大幅に増大させることが含まれる。 |
| 2. 制御の喪失（Loss of control） | モデルを確実に指揮したり、変更したり、シャットダウンしたりする能力を人間が喪失することによるリスク。そのようなリスクは、人間の意図または価値観とのミスアライメント、自己推論、自己複製、自己改善、欺瞞、目標変更への抵抗、力を得ようとする行動（power-seeking behaviour）、AIモデルやAIシステムの自律的作成・改善から生じうる。 |
| 3. サイバー攻撃（Cyber offence） | 重要システム（重要インフラなど）を含む、大規模で高度なサイバー攻撃を可能にすることによるリスク。これには、脆弱性の自動発見、エクスプロイト生成、運用上の使用、攻撃の規模拡大などを通じた攻撃的なサイバー作戦で、悪意のある行為者の参入障壁を大幅に低下させること、または達成される潜在的な影響を大幅に増大させることが含まれる。 |
| 4. 有害な操作（Harmful manipulation） | 大規模な集団や重要な意思決定者を標的にすることで、説得・欺瞞・個人的ターゲティングを通じて、人間の行動や信念の戦略的な歪曲を可能にすることによるリスク。これには、特に複数ターンのやり取りを通じて、個人がそのような影響に気づいていない、あるいは合理的に検知できない場合に、説得・欺瞞・個人的ターゲティングを行う能力を大幅に強化することが含まれる。このような能力は、保護対象の特性に基づく搾取を含め、民主的なプロセスや基本的権利を損なう可能性がある。 |

図の出典：汎用目的AIモデル実践規範に基づき筆者作成


© Institute for International Socio-Economic Studies 2025

# コミットメント３　システミックリスクの分析

IISE

- システミックリスク分析は、前項で特定された各システミックリスクについて、以下の5つの要素から成る。
  - (1) モデルに依存しない情報の収集：　ウェブ検索・文献レビュー・市場分析・学習データのレビュー・過去のインシデントデータのレビュー　・専門家へのインタビュー・　一般人へのアンケート調査等。
  - (2) モデル評価（model evaluation）の実施：　モデルの能力・傾向（propensities）・アフォーダンス・効果を評価。モデル評価手法の例としては、Q&Aセット、タスクベースの評価、ベンチマーク、レッドチーム演習その他の敵対的テスト、人間向上研究（human uplift studies）、モデル生物（model organisms）、シミュレーション、機密資料の代理評価等。
  - (3) システミックリスクのモデリング
  - (4) システミックリスクの推定：　システミックリスクによる損害の発生確率と重大性を推定する。システミックリスクの推定値の例として、定性的なシステミックリスクスコア（例：「中程度」又は「重大」）、定性的なシステミックリスクマトリックス（例：「確率：低い」×「影響：高い」）、定量的なシステミックリスクマトリックス（例：「〇〇%」×「△△EURの損害」）。
  - (5) 市場流通後モニタリングの実施：　様々なチャネルから、モデルの能力・傾向・アフォーダンス・効果に関する情報を収集する。市場流通後モニタリングを円滑に進めるため、十分な人数の独立した外部評価者を選定し、評価者たちに最新鋭のモデルバージョンや当該バージョンの思考の連鎖（chains-of-thought）へのアクセスを提供する。


© Institute for International Socio-Economic Studies 2025

# コミットメント４　システミックリスクの受容の決定

IISE

- システミックリスクの受容基準を定め、モデルに起因するシステミックリスクが受容可能か否かを判断する。
  - (1) 特定された各システミックリスクについて、少なくとも以下のいずれかを実施する。
    - (a) 適切なシステミックリスク階層（systemic risk tiers）を定義する。
    - (b) システミックリスク階層が当該システミックリスクに適しておらず、かつ、当該システミックリスクが指定システミックリスクでない場合は、その他の適切なシステミックリスク受容基準を定義する。
  - (2) これらの階層やその他の受容基準をどのように用いて、特定された各システミックリスクおよび全体的なシステミックリスクが受容可能かどうかを判断するかを説明する。
  - (3) これらの階層やその他の受容基準の使用が、特定された各システミックリスクおよび全体的なシステミックリスクが受容可能であることをどのように保証するかを正当化する。
- システミックリスクが受容可能と判断された場合のみ、モデルの開発・市場への提供（市場で利用可能にすること）・使用を進める。システミックリスクが受容可能と判断されない場合、または近いうちに受容可能と判断されないことが合理的に予見される場合、モデルに起因するシステミックリスクが現在受容可能であり、今後も受容可能であることを保証するための適切な措置を講じる。特に、以下を実施する。
  - (1) 必要に応じて、モデルを市場に提供しない、市場への提供を制限する、モデルを撤回または回収する。
  - (2) 安全性ための軽減策やセキュリティための軽減策を実施する。
  - (3) システミックリスクの特定・分析・受容の決定を再度実施する。


© Institute for International Socio-Economic Studies 2025

# コミットメント５　安全性のための軽減策

IISE

- システミックリスクが受容可能であることを保証するために、敵対的な圧力（例：ファインチューニング攻撃やジェイルブレイク）に対して十分に堅牢なものを含め、モデルのライフサイクル全体にわたって適切な安全性のための軽減策（safety mitigations）を実施する。安全性軽減策の例は以下。
  - (1) 学習データのフィルタリングとクリーニング（例：不正確な思考連鎖のトレースなど、望ましくないモデル傾向につながる可能性のあるデータ）。
  - (2) モデルの入力／出力のモニタリングとフィルタリング。
  - (3) 安全性を重視したモデルの挙動の変更（例えば、一定のリクエストを拒否したり、役に立たない応答を返すようにモデルをファインチューニングするなど）
  - (4) モデルへのアクセス許可を段階的に行う（例えば、APIアクセスを審査済みのユーザーに限定する、市場流通後モニタリングに基づいてアクセスを徐々に拡大する、当初はモデルのパラメータを一般にダウンロード可能としないなど）
  - (5) モデルに起因するシステミックリスクを軽減するために他のアクターが使用できるツールを提供する
  - (6) モデルの動作に関する高確度の定量的な安全性保証を提供する技術
  - (7) AIエージェントの安全なエコシステムを実現するための技術（モデル識別、専用の通信プロトコル、インシデント監視ツールなど）
  - (8) その他の新たな安全性軽減策（例えば、思考連鎖推論の透明性の実現、またはモデルが他の安全性軽減策を無効にする能力に対する防御など）


© Institute for International Socio-Economic Studies 2025

# コミットメント６　セキュリティのための軽減策

IISE

- システミックリスク（不正なリリース、不正アクセス、モデルの盗難から生じうるもの）が受容可能であることを保証するために、そのライフサイクル全体にわたるモデルと物理インフラに対し、十分なレベルのサイバーセキュリティ保護を実施する。
- モデルや物理インフラに対するセキュリティ脅威主体（非国家の外部脅威主体、内部脅威主体、その他の想定される脅威主体を含む）を目標（「セキュリティ目標」）として定め、それらからの保護を実施する。
- 「付録４：セキュリティ軽減策の目標と措置」で具体的な軽減策を規定。
  - 付録4.1：一般的なセキュリティ軽減策
  - 付録4.2：未公開モデルパラメータの保護
  - 付録4.3：未公開モデルパラメータへのインターフェースアクセスの防御強化
  - 付録4.4：内部脅威
  - 付録4.5：セキュリティ保証


© Institute for International Socio-Economic Studies 2025

# コミットメント７ 安全性とセキュリティに関するモデルレポート

IISE

- モデルを市場投入する前に、安全性とセキュリティに関するモデルレポート（Safety and Security Model Report ）を作成し、モデルおよびシステミックリスクの評価と軽減のプロセスと措置に関する情報をAIオフィスに報告する。また、モデルレポートを最新の状態に保つ。
- 既に他のレポートや通知においてAIオフィスに関連情報を提供している場合、モデルレポートにおいてそれらのレポートや通知を参照することができる。
- モデルレポートに記載すべき項目（大項目）は以下。
  - 1. モデルの説明と動作： モデルのアーキテクチャ・能力・傾向・アフォーダンスに関する概要説明、モデルの開発方法・使用方法、モデルバージョンの違い、モデル仕様等
  - 2. 開発等を進める理由： システミックリスクが受容可能な理由、開発・市場投入・使用を進める決定がどのように行われたかの説明等
  - 3. システミックリスクの特定・分析・軽減に関する文書化： モデル評価の結果、外部評価者、安全性軽減策、セキュリティ軽減策等
  - 4. 外部レポート： 外部評価者のレポート、外部機関のセキュリティレビュー等
  - 5. システミックリスクの状況における重大な変化


© Institute for International Socio-Economic Studies 2025

# コミットメント８　システミックリスクの責任分担

IISE

- 自社組織のあらゆるレベルにおいて、システミックリスクの管理に関する責任を明確に定義する。これには、以下に関する責任が含まれる。
  - (1) システミックリスク評価および軽減のプロセス／措置の監督
  - (2) システミックリスクのオーナーシップ　(3) システミックリスクのサポートとモニタリング
  - (4) システミックリスクの評価および軽減のプロセス/措置の十分性について、内部保証および必要に応じて外部保証を、監督機能を有する経営組織またはその他の適切な独立機関に提供する
- 自社のガバナンス構造および組織の複雑性に応じて、これらの責任を組織内の以下のレベルに分担させる。
  - (1) 監督機能を有する経営組織またはその他の適切な独立機関（評議会や取締役会など）
  - (2) 執行機能を有する経営組織　(3) 関連する業務チーム
  - (4) 可能な場合は、内部保証提供者（例：内部監査部門）　(5) 可能な場合は、外部保証提供者（例：第三者監査人）
- 自社の経営組織が、責任を分担された者への適切な資源（人的資源、財務資源、情報・知識へのアクセス、計算資源）の配分を監督することを保証する。
- 自社内での健全なリスク文化を促進する。
  - 内部通報窓口が積極的に活用され、報告に対して適切な対応がなされている。
  - 自社モデルに起因するシステミックリスクについて、合理的な根拠に基づき所轄機関に公表または提供した者に対して、解雇・降格・法的措置・否定的な評価・敵対的な職場環境の創出など、直接的または間接的な不利益な行為を含むいかなる形態の報復も行わない（内部告発者保護）。　等


© Institute for International Socio-Economic Studies 2025

# コミットメント９　重大インシデントの報告

IISE

- 重大インシデントに関する関連情報を収集・記録するため、市場流通後モニタリングにおける情報収集と同様な方法（エンドユーザーからのフィードバック、匿名通報チャネル、インシデント報告フォーム、利害関係者との対話、アカデミアとの協力等）を検討する。
- 重大インシデントについて認識した場合、一定の情報（発生日、発生した損害と被害者、一連の経緯、関与モデル、対応内容等）について追跡、文書化し、AIオフィスおよび（該当する場合）国内所轄機関に報告する。
- AIオフィスおよび（該当する場合）国内所轄機関に提出する最初の報告において、一定の事項を、以下の時間内に提供する（ただし例外的な状況を除く）。
  - (1) 重要なインフラの管理／運用における深刻かつ回復不能な混乱の場合、当該インシデントへのモデルの関与を認識してから2日以内。
  - (2) モデル重みの（自己）流出やサイバー攻撃を含む重大なサイバーセキュリティ侵害が発生した場合、当該インシデントへのモデルの関与を認識してから5日以内。
  - (3) 人の死亡が発生した場合、当該インシデントへのモデルの関与を認識してから10日以内。
  - (4) 人の健康への重大な危害、基本的権利を保護することを目的としたEU法に基づく義務の侵害、財産もしくは環境への重大な危害の場合、当該インシデントへのモデルの関与を認識してから15日以内。

※AI法第3条（49）：「重大インシデント（serious incident）」とは、直接的または間接的に以下のいずれかにつながるAIシステムのインシデントまたは誤動作を意味する。

- (a)人の死亡、または人の健康への重大な危害。　(b)重要なインフラの管理と運用の深刻かつ回復不能な混乱。
- (c)基本的権利を保護することを目的としたEU法に基づく義務の侵害。　(d)財産または環境への重大な危害。


© Institute for International Socio-Economic Studies 2025

# コミットメント10　追加的な文書化と透明性

IISE

- AIオフィスの要請に応じて提供するために、以下の情報を作成し、最新の状態に保つ。
  - (1) モデルのアーキテクチャの詳細な説明。
  - (2) モデルがAIシステムに統合される方法の詳細な説明。
  - (3) モデル評価の詳細な説明（結果および戦略を含む）。
  - (4) 安全性軽減策の詳細な説明。
- システミックリスクの評価や軽減に必要な場合、安全性・セキュリティフレームワークの概要版およびモデルレポートの概要版、ならびにその更新版を（ウェブサイト等で）公表する。
  - ただし、安全性／セキュリティ軽減策の有効性を損なわないよう、またセンシティブな企業情報を保護するために、編集できる。


© Institute for International Socio-Economic Studies 2025

# EUのAI法実践規範と米国カリフォルニア州TFAIAの比較（1/2）

IISE

| | | EUのAI法<br>汎用目的AIモデル実践規範「安全性・セキュリティ」章 | 米国カリフォルニア州<br>フロンティアAI透明性法（TFAIA、SB53）（25年9月29日成立） |
|---|---|---|---|
| 規制対象となる事業者 | | システミックリスク汎用目的AIモデルのプロバイダー<br>（汎用目的AIモデル一般については実践規範の「透明性」章と「著作権」章で規制） | フロンティアモデルの大規模開発者<br>（フロンティアモデルの開発者一般にも一部の義務が課されるが罰則は無し） |
| 規制対象となるAIモデル | | システミックリスク汎用目的AIモデル：<br>10の25乗を超える浮動小数点演算（FLOP）の累積計算量を使用して学習された汎用目的AIモデル | フロンティアモデル：<br>10の26乗を超える整数演算または浮動小数点演算（FLOP）の累積計算量を使用して学習された基盤モデル |
| 署名したプロバイダー（AI法実践規範）／大規模開発者（TFAIA）の義務 | (a) 敵対的テスト・文書化含むモデル評価 | ○<br>安全性・セキュリティフレームワークの作成・実装・遵守・更新・公開（要約版） | ○<br>フロンティア AI フレームワークの作成・実装・遵守・更新・公開 |
| | (b) システミックリスクの評価と軽減 | ○<br>システミックリスクの評価と軽減（安全性軽減策・セキュリティ軽減策）、外部評価者の使用 | ○<br>壊滅的リスクの評価と軽減、外部評価者の使用 |
| | (c) 重大インシデントの報告 | ○<br>重大インシデントのAIオフィス／国内所轄機関への報告 | ○<br>重大な安全インシデントの州当局への報告 |
| | (d) モデルとインフラのサイバーセキュリティ | ○<br>セキュリティ軽減策 | ○<br>サイバーセキュリティ対策（未公開のモデル重みを内部・外部者による不正な変更や転送から保護） |
| | モデルレポートの作成と提出 | ○<br>安全性とセキュリティに関するモデルレポートの作成、AIオフィスへの提出、公開（要約版） | ○<br>透明性レポートの作成、公開 |
| | 内部告発者保護 | ○ | ○ |
| 違反時の罰則 | | 世界売上総額の3%または1500万ユーロのいずれか高額の方を超えない罰金（fines） | 違反 1 件あたり 100 万ドル以下の民事罰（civil penalty） |

（筆者作成）



© Institute for International Socio-Economic Studies 2025

# EUのAI法実践規範と米国カリフォルニア州TFAIAの比較（2/2）

IISE

| | EUのAI法<br>汎用目的AIモデル実践規範「安全性・セキュリティ」章 | 米国カリフォルニア州<br>フロンティアAI透明性法（TFAIA、SB53）（25年9月29日成立） |
|---|---|---|
| システミックリスク／壊滅的リスクの定義 | システミックリスク（systemic risk）：<br>汎用目的AIモデルのハイインパクトな能力に特有のリスクであって、そのリーチによって、または公衆衛生、安全、パブリックセキュリティ、基本的権利、もしくは社会全体に対する実際もしくは合理的に予見可能な悪影響によって、EU市場に重大な影響を与え、かつバリューチェーン全体に大規模に伝播しうるもの。（AI法第3条（65））<br><br>指定システミックリスク（specified systemic risks）：<br>(1) 化学・生物・放射線・核（CBRN）<br>(2) 制御の喪失（Loss of control）<br>(3) サイバー攻撃<br>(4) 有害な操作（Harmful manipulation） | 壊滅的リスク（catastrophic risk）：<br>フロンティア開発者によるフロンティアモデルの開発、保管、利用、または展開が、フロンティアモデルが以下のいずれかを行うことを伴う単一インシデントに起因して、50人を超える死亡もしくは重傷、または10億ドルを超える財産の損害もしくは損失に実質的に寄与するという、予見可能かつ重大なリスク。<br><br>(A) 化学兵器、生物兵器、放射線兵器、核兵器（CBRN兵器）の製造またはリリースにおいて専門家レベルの支援を提供すること。<br>(B) 意味のある人間による監視、介入、または監督なしに、サイバー攻撃、またはその行為が人間によって行われた場合には殺人、暴行、恐喝、もしくは窃盗（詐欺による窃盗を含む）の犯罪を構成する行為を行うこと。<br>(C) フロンティア開発者または利用者の制御を回避（Evading the control）すること。 |

（筆者作成）

※米国カリフォルニア州フロンティアAI透明性法（TFAIA）については以下も参照のこと。
「米国カリフォルニア州のフロンティアAI透明性法（TFAIA）の概要」
https://www.i-ise.com/jp/information/media/2025/251028.pdf


© Institute for International Socio-Economic Studies 2025

I I S E



© Institute for International Socio-Economic Studies 2025