人工知能戦略専門調査会 制度の在り方検討ワーキンググループ（第2回）

フロンティア AIのリスクと制度の在り方

Anthropic Japan合同会社
公共政策責任者
吉田 朗

3つのポイント

1. リスクの捉え方

最も高性能なAIモデルは、**生物化学、サイバー、制御喪失**の分野で、**深刻で取り返しのつかない被害**をもたらし得る段階に近づいている。**AI自身がAIの開発**を助けることで、これらのリスクが増幅される可能性がある。

2. 現在の取組

Anthropicとしての方針に基づいて**安全性評価、対策、公表**を実施。**インシデント**などの問題が起きた場合も公表。独立した評価者による**外部検証も拡大**。

3. 必要な制度

AI開発企業の自主的な取組だけでは不十分であり、安全性評価・対策・公表、外部検証を含む**実行的な制度**が必要。 **最先端のAIモデル開発者に絞った明確なルールと政府による監督を支持**。一つのモデル評価と、一つの報告が、各国で通用する制度設計が望ましい。

1. リスクの捉え方

リスク	懸念点
生物化学	危険な病原体や化学兵器の開発が容易になること
サイバー	重要なソフトウェアの脆弱性の大規模な発見と悪用
制御喪失	AIが監視を回避し、人間が修正や停止をできなくなること
AI開発の自動化	AIが次のAIを開発し、他のリスクを増幅して人間の理解を超えて進化すること

自社の評価

2026年8月のリスク報告書では、生物化学、ミスアラインメント、AI開発の自動化のリスクを、いずれも『低』と判断。ただし、AI開発の自動化は判断の確度が前回より下がり、ミスアラインメント、既知の生物化学兵器のリスクは前回より高いと評価。サイバーは評価対象外。

他の問題も重要

詐欺、偽情報、雇用への影響なども重要な問題。
本日は、最も深刻なリスクを中心に説明。

2. Anthropicの取組

リスクガバナンス

責任あるスケーリングポリシー、ロードマップ、フロンティア・コンプライアンス・フレームワーク

公開報告

システムカードと定期的なリスク報告書

外部の評価者

METR（モデルの能力や安全性リスクを評価する非営利の研究機関）が社員に近いアクセスを持ち、独自の分析を公表できる

セーフガード

分類器（有害・危険なコンテンツを検知する仕組み）、利用規約の執行、サイバーと生命科学分野における検証済みアクセスプログラム

インシデントの開示

事案を公表し、社内評価の方法を変更

日本のAISI

2025年10月に協力覚書を締結

3. 必要な制度：ハードローとソフトロー

AI開発企業による自主的な取組のみでは、対応が遅れている事業者が基準点となってしまうおそれ。明確なルールを支持。

法的拘束力のある対応：最先端の開発者を対象

- **安全フレームワーク：**ガバナンス方針を定め、責任者を明示して公表し、遵守
- **システムカードとリスク報告書：**モデル提供時のテスト結果の公表、半年ごとのリスク報告書
- **独立した評価者：**モデルへの実質的なアクセスを持つ適格な外部者が、独立した立場から見解を公表
- **セキュリティ体制：**外部と内部の脅威に対応
- **内部通報：**匿名の窓口、不利益な取扱いの禁止
- **重大インシデント：**対象事案を明確に定義し、報告

柔軟な枠組：技術の進展に合わせて見直し

- **評価の方法と様式：**AISIや標準化団体を通じて整備
- **閾値：**拘束力ある規律の対象基準を、外部の専門家などの意見を聞きつつ少なくとも年1回見直し
- **広島AIプロセス：**広島AIプロセスへの報告を、法に基づく報告義務にカウントできるようにする
- **指針：**他の開発者やAIを利用する事業者向け

3. 必要な制度：国境を越えて機能するルール

目指すべきは、世界中の人々にAIの恩恵が行き渡るよう、技術を円滑に利用できる共通の基準。

一つの評価

- パートナー国の信頼できる機関のテスト結果を相互に承認
- 評価者は公表された基準で認定：独立性、専門性、アクセス、セキュリティ

一つのインシデント開示

- 一定の水準を満たせば、他国への開示についても国内の開示要件を満たすものとみなす
- 他国の制度が一部のみ対応する場合は、不足分だけを追加
- 日本として、同等と認める他国の制度を公表

一つの報告

- インシデントの定義、期限、様式の整合性を確保
- 一つの報告で複数の政府に対応できるようにする

どのような制度を採用するかは日本として決めること。モデルの安全性評価の相互承認は、モデルを確認する作業を分担するものであり、各国政府の権限は維持される。

3. 必要な制度：日本の役割と制度設計

日本が主導できること

- **AISIの機能強化：**開発者の説明だけに頼らず、日本自身がフロンティアモデルの安全性を確認
- **日本固有の検証：**提供後のモデルについて、日本語での安全性対策、重要インフラ、国内の悪用事例などを検証
- **広島AIプロセス：**広島AIプロセスへの報告を、法に基づく報告義務にカウント
- **備えと人材：**サイバーセキュリティと生物化学分野、それを担う人材

制度設計の考え方

- **法的な明確性：**義務内容の明確化、行政裁量の行使基準の明確化、秘密情報の保護
- **具体的な事実に基づく執行：**明確な証拠に基づく判断、事前の通知、是正の機会、不服申立て
- **小さく始めて柔軟に見直す：**新たな義務の導入までの準備期間の確保、評価体制に応じた柔軟な見直し

Thank you