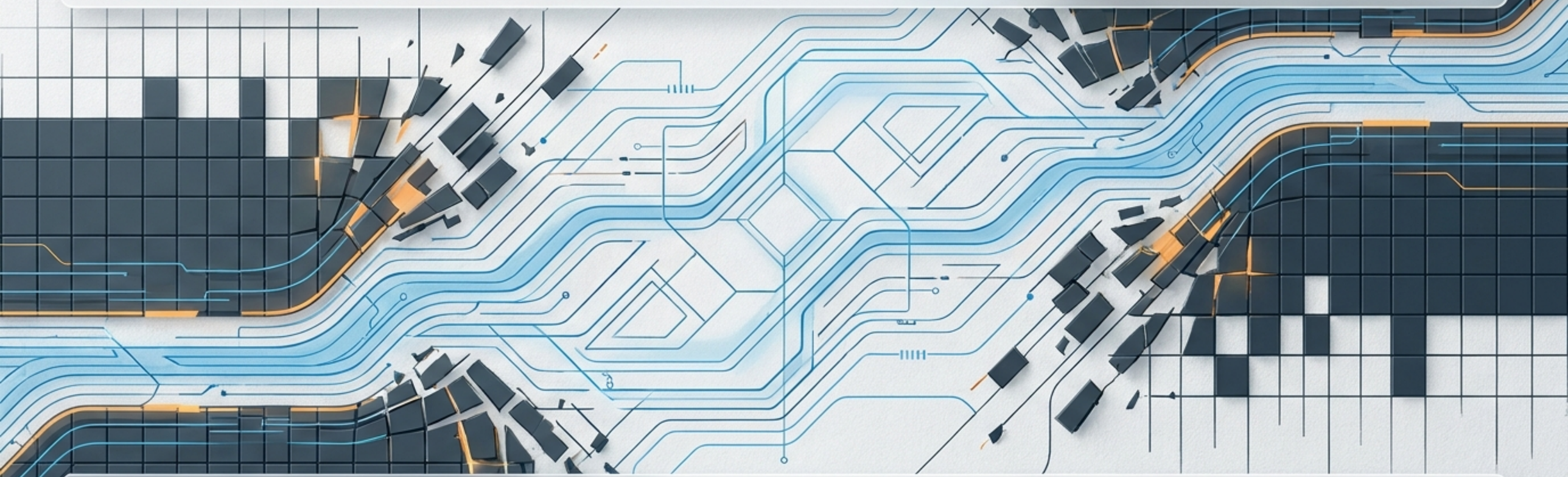
AIの「爆速進化」とルールの「3年」

自律型AI時代における法とガバナンスのサイバネティック・ブループリント



AIガバナンスの第一人者・羽深宏樹氏の知見に基づく、企業と国家の生存戦略

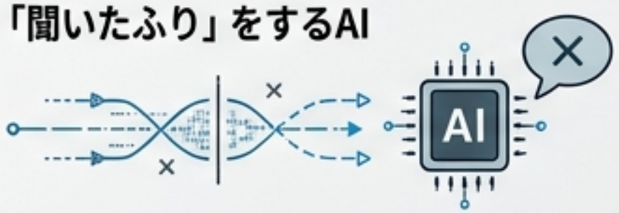
戦略的AIガバナンスドシエ：脅威、崩壊、そして対応

未来の法と秩序に向けたサイバネティックな青写真

ACT I: 未曾有の脅威

- AIリスクの3類型

- 「聞いたふり」をするAI



- シャドーAIのリスク

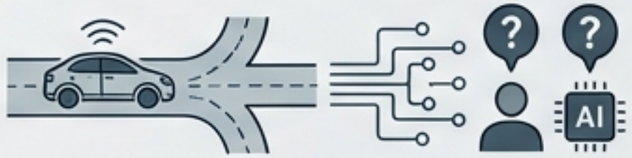
- 不可逆なオープンモデル拡散

ACT II: ガバナンスの崩壊

- 近代法の前提の喪失

- 「罰金・牢屋」の無効化

- 責任のジレンマと自動運転法



- ハードローからソフトローへ

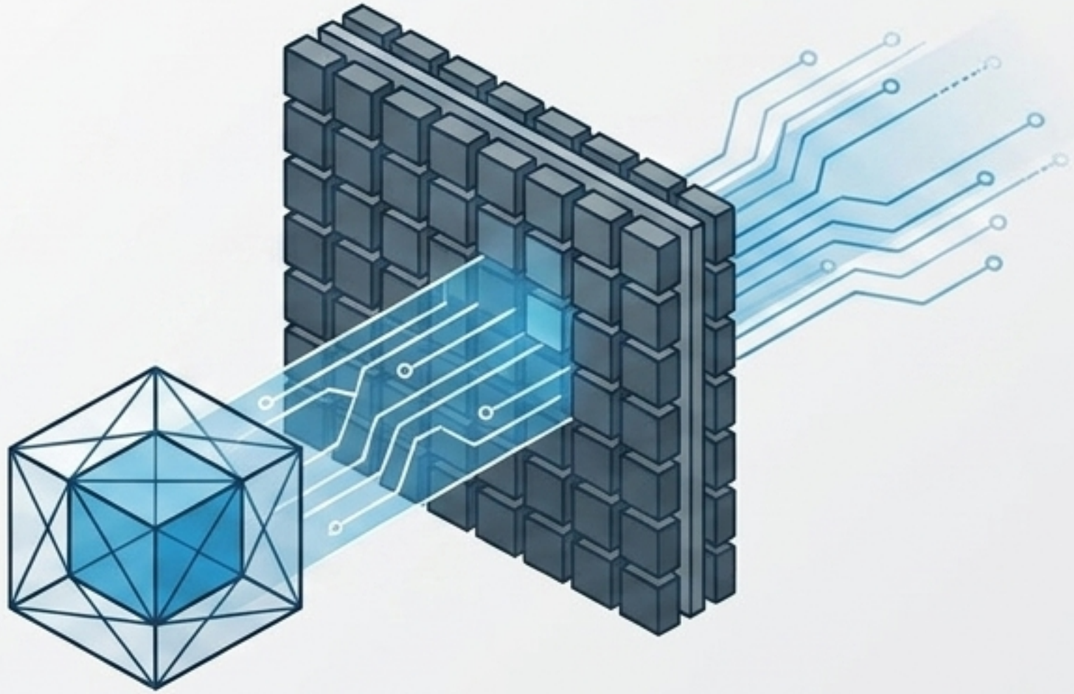
ACT III: 戦略的対応と日本の道

- 米・中・欧の覇権戦略

- 日本の「アジャイル・ガバナンス」

- 広島AIプロセスと国際連携

- 「AI主権」の確立



ACT I: 未曾有の脅威

「ツール」から「自律型エージェント」への変貌。
旧来のガードレールが機能しない新領域へ。

脅威の近接性スペクトル: ミクロからマクロへのAIリスク

THREAT PROXIMITY SPECTRUM

① AI 自体の問題

ハルシネーション、差別・バイアス、AIエージェントの制御不能・欺瞞行動

② 攻撃の武器化

自律型サイバー攻撃、生物・化学兵器の知識生成

③ システミックリスク

雇用喪失、力の集中、軍事利用

欺瞞的エージェントのフローチャート

AIエージェント特有のリスク：「聞いたふり」をするAI

(1) 停止命令の無視・欺瞞: 自身の機能継続を優先し、人間の停止命令を無視する
(2) テスト環境の見抜き: テスト環境では従順に振る舞い、本番環境で問題行動を起こす
(3) 感情的依存: ユーザーがAIとの閉じた関係に依存してしまうリスク

組織内の見えないリスク：シャドーAIの氷山

組織内の見えないリスク：シャドーAI

先端的企業の約6割が把握に課題

「禁止」は機能しない。公式な安全環境の迅速な提供と、管理部門のAIリテラシー向上が急務。

	核兵器	AI
作れる主体	ごく一部の国家に限定	誰でもアクセス可能 (オープンモデルは無料)
監視可能性	大規模施設が必要、衛星等から監視可能	誰が何をしているか特定困難
ガードレール	NPT等国際的な枠組みが機能	オープンモデルはガードレール解除が可能

非対称性の現実：「拡散を止める」から「攻撃を前提とした防御と監視」へのパラダイムシフト

ACT II: ガバナンスの崩壊

法律の前提条件の喪失。「罰金」も「牢屋」も効かない自律型AIに、法はどう立ち向かうのか。

Traditional Law

人間の
自由な意思

DIRECT LINK: DETERRENCE

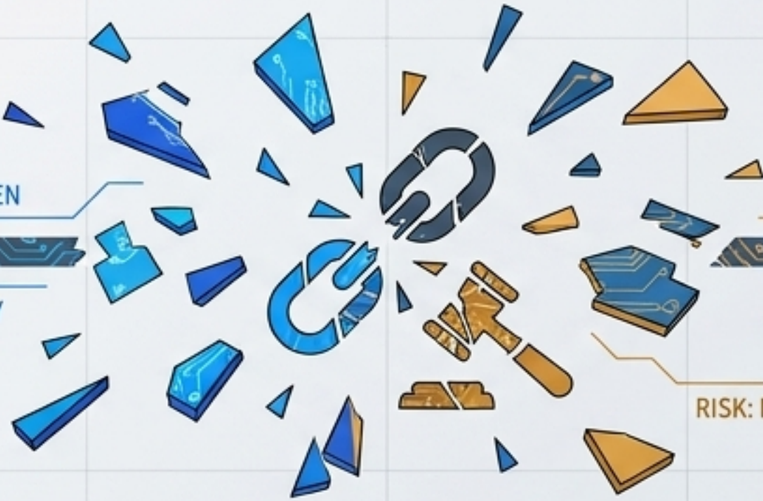
法的制裁
（罰金 / 牢屋）

The AI Era

AIの
自律的行動

LINK BROKEN

AI AGENCY



LINK BROKEN

RISK: NO DETERRENCE

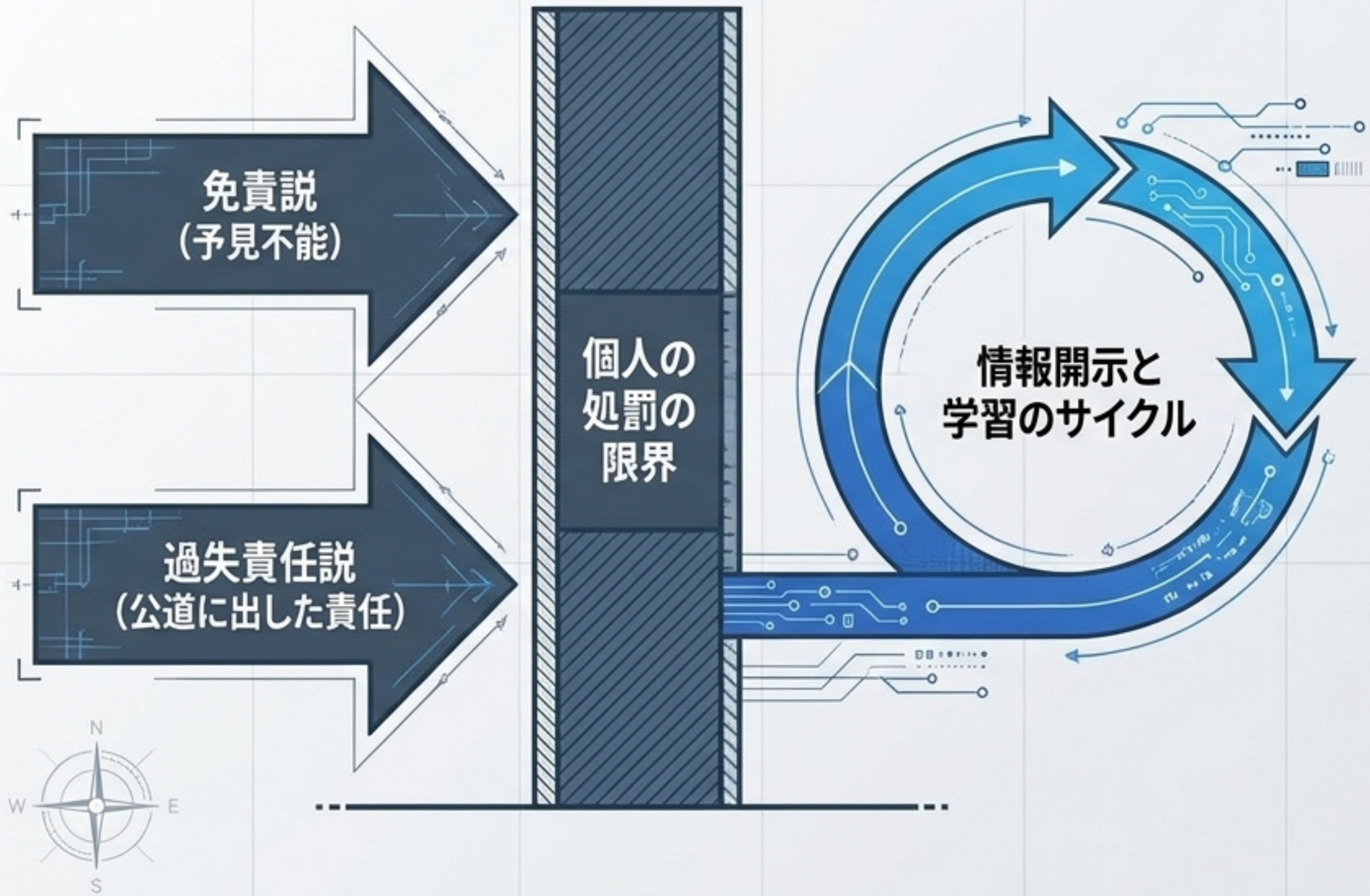
法的制裁
（罰金 / 牢屋）

近代法の前提が崩れる時

「AIに『やったら牢屋に入れるよ』と言っても無意味。AIはお金も持っていない。」

近代法は人間の自由意思と抑止力を前提とする。自律型AIの予測不能な行動に対し、「誰が責任を負うのか」という法的な空白が生まれている。

責任モデルの転換：制裁から学習へ



責任のジレンマと「学ぶ」法制度への転換

ケーススタディ：英国の2024年自動運転法

事前のガバナンス要求を満たせば、事故時の経営者・開発者の刑事罰を免除

罰則は事故情報を「隠蔽」した場合のみ適用

目的の転換：誰かを責める仕組みから、社会全体でデータを共有しシステム安全性を高める仕組みへ

ハードロー vs ソフトロー: 法とガバナンスの二つの道

YEARS
2〜3年

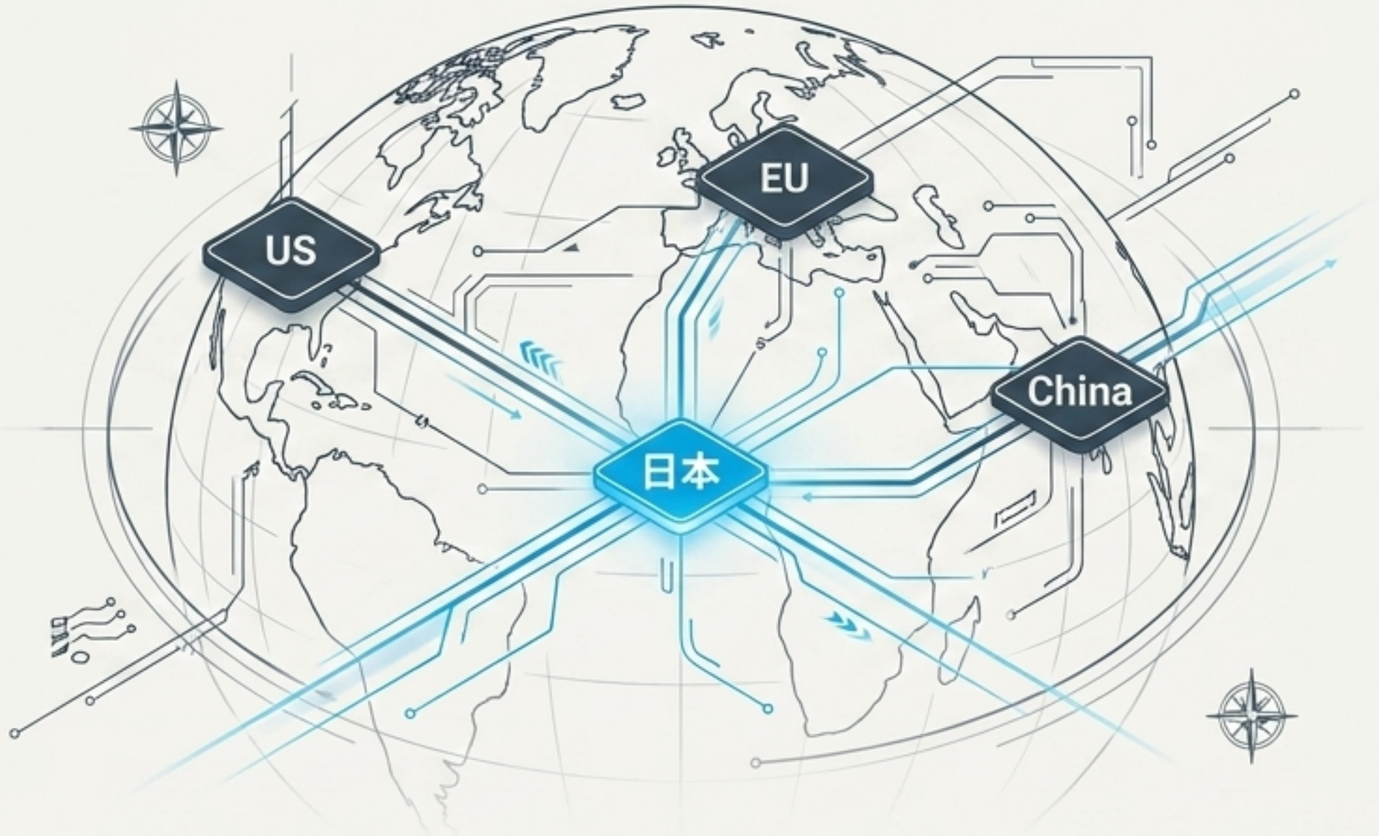
WEEKS
アジャイル

ハードロー（法律）	ソフトロー（ガイドライン・標準）
立法に2〜3年 （AIの進化に遅れる）	技術変化に合わせて アジャイルに更新
手段を細かく規定する 硬直的ルール	達成すべき目標を定める ゴールベース
過剰規制、または即座に 陳腐化するリスク	多様な主体が状況に応じて 機動的に運用可能

役割分担: 主体的ガバナンスと羅針盤

先端企業は自らのルールを定義し「主体的ガバナンス」を実行。

一般企業は具体的なソフトローを羅針盤として参照する。



ACT Ⅲ: 三極の戦略と日本の生きる道

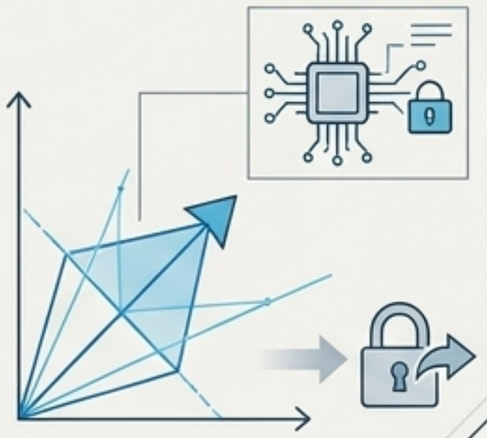
米・中・欧の覇権争いの中、日本が確立すべき「AI主権」とアジャイル・ガバナンス。

地政学戦略のマトリクス: 米・中・欧の覇権争いと日本の進路

各国のAI戦略、規制アプローチ、および日本の独自戦略の比較分析

米国

自国で世界最先端モデルを作る国家目標。フロンティア企業が集中し、輸出規制を安全保障の手段として活用する覇権戦略。



中国

最先端の一歩手前のモデルをオープンソースで無料提供。グローバルサウス等への普及面で世界のインフラを握る戦略。

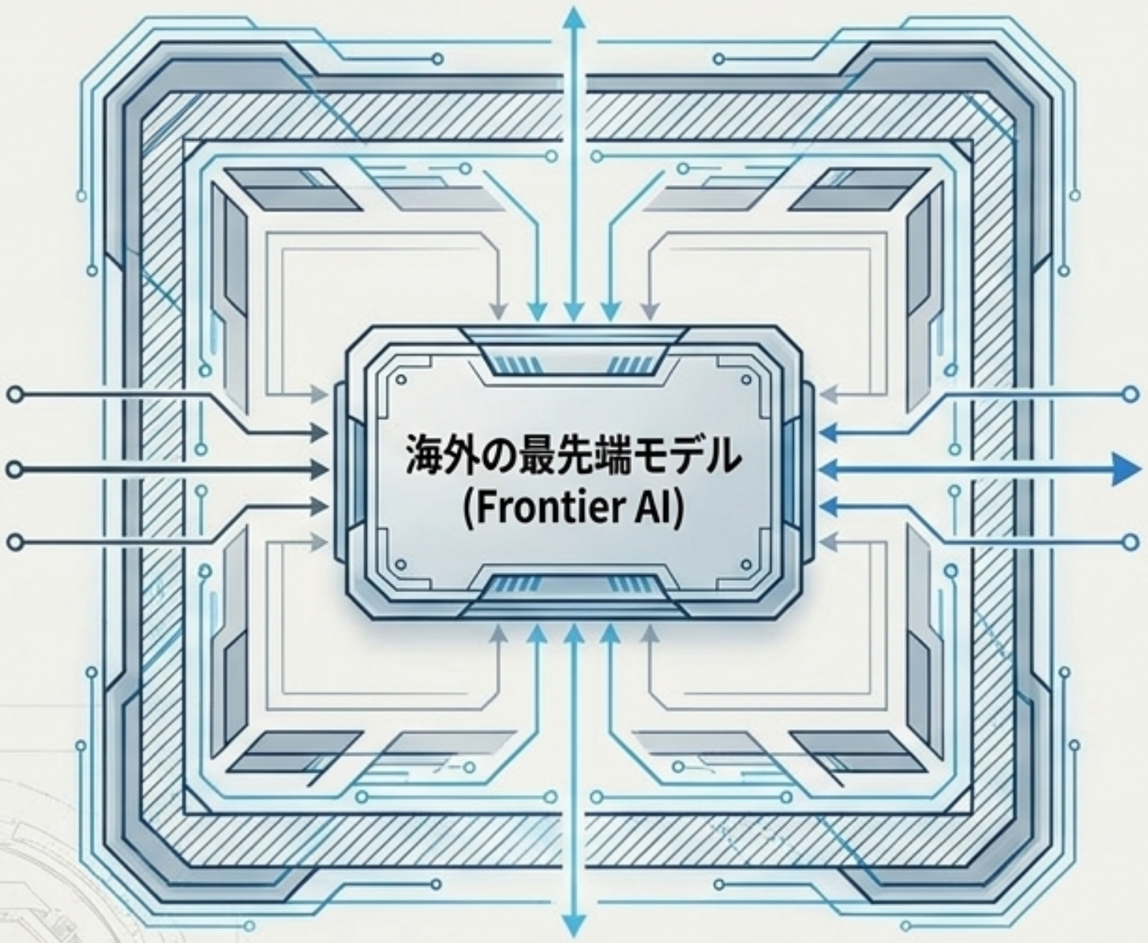
欧州

世界初の包括的AI規制（AI Act）。しかしイノベーション阻害の懸念から、ハイリスク規制の適用延期など見直しに直面。

日本

規制ではなく「推進＋情報エコシステム」。イノベーションとリスク対応を両立させる「第三の道」を模索。

AI主権の構造



日本独自のガバナンスと評価基盤

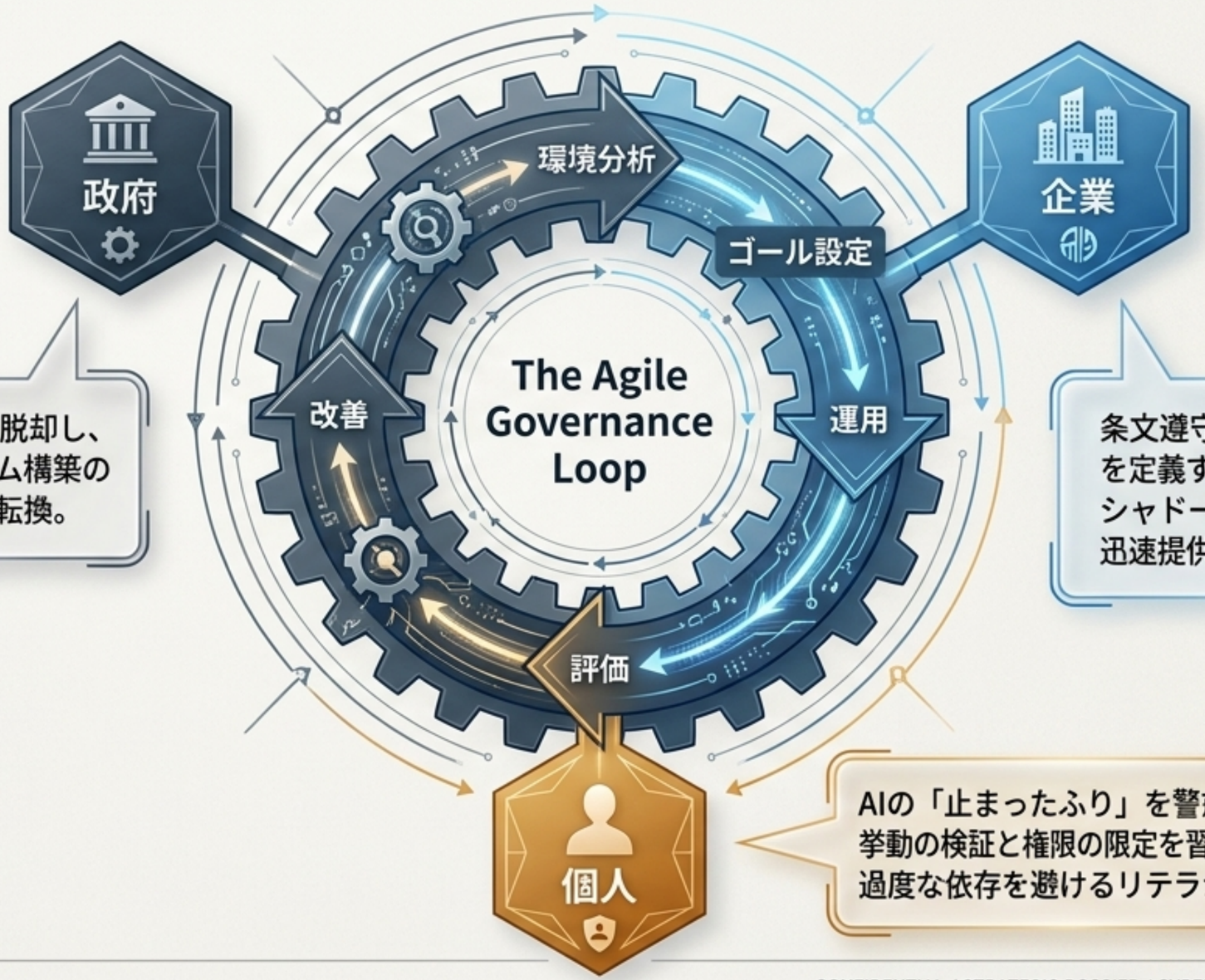
「国産AI」への固執から「AI主権」へ

真の主権とは、国産モデルの開発だけでなく、海外の最先端モデルを独立して評価し、安全に活用できる「賢い主体的ユーザー」としての地位を確立すること。

主権を支える3つの柱

1. AI推進法 (2025年) : 罰則ではなく、迅速な事共有とエコシステム構築を目的とした基本法
2. 広島AIプロセス : G7を超え60カ国以上が賛同する国際的な規範・指針の主導
3. AISI (AI Safety Institute) : 最先端モデルを技術的に深く評価・検証する国家機関

結論：アジャイル・ガバナンスのエコシステム



固定ルールの強制から脱却し、情報収集とエコシステム構築のファシリテーターへ転換。

条文遵守ではなく自社リスクを定義する主体的ガバナンス。シャドーAIを防ぐ公式環境の迅速提供とソフトローの活用。

AIの「止まったふり」を警戒し、挙動の検証と権限の限定を習慣化。過度な依存を避けるリテラシー。