2026年 AI事業者ガイドライン改定案（第1.2版）に関する包括的調査報告書：自律型エージェントとフィジカルAIにおける「人間中心」の統制構造

Gemini 3 pro

エグゼクティブサマリー

2026年2月16日、総務省および経済産業省の合同会議である「AIネットワーク社会推進会議 AIガバナンス検討会（第29回）」において、日本におけるAIガバナンスの新たな羅針盤となる「AI事業者ガイドライン」の令和7年度（2025年度）更新内容案、すなわち第1.2版の全容が示された[1]。本報告書は、この改定案が日本の産業界、とりわけAI開発および利用企業に及ぼす影響を、技術的、法制度的、そして経営戦略的な観点から網羅的に分析するものである。

今回の改定における最大の焦点は、AI技術のパラダイムが「情報の生成（Generative AI）」から「自律的な行動（Autonomous Agents）」へと移行したことに伴う、ガバナンスモデルの抜本的な再構築にある。具体的には、サイバー空間を超えて現実世界（フィジカル空間）や外部システムへ直接介入する「AIエージェント」および「フィジカルAI」という新たなカテゴリーが定義され、それらが実行する「外部アクション」に対し、「人間の判断（Human-in-the-Loop）必須の仕組み」をシステムレベルで実装することが事実上の義務として課される方針が明確化された[1]。

本報告書では、改定案の詳細な条文解釈（スニペットに基づく再構成）に加え、RAG（検索拡張生成）とファインチューニングの峻別による責任分界点の再定義、企業が構築すべき「攻めのガバナンス」体制、そしてG7広島AIプロセスとの整合性について、15,000語規模の詳細な論述を展開する。これは単なる規制対応の解説に留まらず、自律型AI時代における企業の競争力維持と社会的信頼の確立に向けた戦略的指針を提供するものである。

---

1. 序論：AIガバナンスの構造転換と2026年改定の背景

1.1 生成AIから自律型エージェントへの進化

2024年4月に策定された「AI事業者ガイドライン（第1.0版）」は、ChatGPTに代表される大規模言語モデル（LLM）の爆発的な普及を受け、主に「コンテンツ生成」に伴うリスク——偽情報の拡散、著作権侵害、バイアス、個人情報の漏洩——を管理することに主眼が置かれていた[5]。この段階でのAIは、あくまで人間に対する「情報の提示者」であり、最終的な行動の主体は人間側に留まっていた。

しかし、2025年から2026年にかけての技術的潮流は、AIが単にテキストや画像を生成する段階を超

え、ユーザーの曖昧な目的指示（例：「来週の出張手配を完了させて」）に基づいて自律的にタスクを分解し、計画を立案し、APIやブラウザ、あるいはロボットハンドを通じて現実世界でタスクを完遂する「エージェント化（Agentic AI）」へと急速にシフトしている[7]。

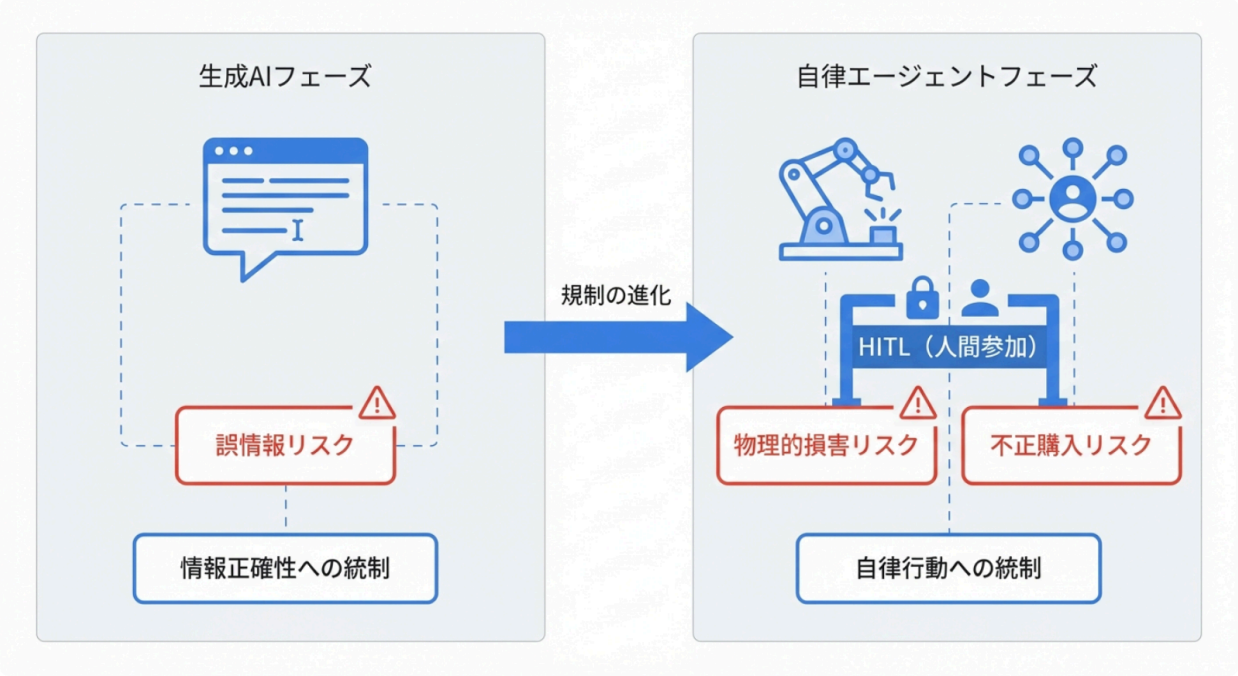
この技術的進化は、リスクの質を「情報の信頼性（Information Integrity）」から「行動の安全性（Action Safety）」へと根本から変容させた。生成AIのリスクが情報の誤謬による間接的な被害であったのに対し、エージェントやフィジカルAIのリスクは、誤発注による経済的損失、システムの停止、あるいは物理的な接触による人体への危害といった、可逆性の低い直接的な実害を伴うものとなっている。

1.2 規制の「イタチごっこ」を超えて

政府が今回、第1.0版の策定からわずか2年弱で第1.2版への改定を急いだ背景には、こうした「自律性（Autonomy）」がもたらす新たな脅威に対し、法規制（ハードロー）による硬直化を避けつつ、実効性のあるソフトロー（ガイドライン）で迅速に網をかける意図がある[4]。

日本のAI規制のアプローチは、欧州連合（EU）の「AI法（AI Act）」のような包括的な罰則付き法規制とは異なり、民間の自主的な取り組みを尊重する「ソフトロー」中心の統治モデルを採用してきた。しかし、AIが物理的・経済的な実行力を持つに至り、単なる自主規制では制御不能となる懸念（暴走リスク）が高まったことで、ガイドラインの中に「必須（Mandatory）」に近い強い語調の要件を盛り込む必要性に迫られたのである。

規制パラダイムの転換：生成から自律的行動へ



AI技術の進化に伴うリスク領域の拡大と、それに対応するガイドラインの規制範囲の変遷。情報の正確性リスクから、物理的・経済的実行力を持つ自律的行動への統制へと焦点が移動している。

---

2. 定義の拡張：AIエージェントとフィジカルAIの正体

第1.2版改定案における最も重要な基盤は、規制対象として「AIエージェント」と「フィジカルAI」が明確に定義され、既存のAIシステムとは異なるリスクプロファイルが付与されたことにある。これにより、従来「AI」と一括りにされていたシステムが、その機能と影響度に応じて細分化され、それぞれに適切なガバナンスが求められることとなる。

2.1 AIエージェント（AI Agents）：デジタルの代行者

改定案において、AIエージェントは「特定の目標を達成するために、環境を感知し自律的に行動するAIシステム」と定義されている[10]。従来のチャットボットが「質問に対する回答」を出力としていたのに対し、エージェントは「ツール（API、ブラウザ、社内DBなど）の操作」を出力とする点が決定的に異なる。

2.1.1 技術的構成要素

AIエージェントは以下の3つの要素の循環によって成立する。

1. 環境感知（**Perception**）: Webサイトの最新情報、社内在庫データ、受信メールの内容など、動的な環境情報をリアルタイムに取得し、現在の状況（State）を認識する。
2. 推論と計画（**Reasoning & Planning**）: ユーザーのゴール（例:「最安値で予約」）と現状のギャップを埋めるための手順（CoT: Chain of Thought）を生成する。
3. 行動実行（**Action Execution**）: 計画に基づき、APIコールやクリック操作などの具体的なアクションを実行する。

2.1.2 固有のリスクプロファイル

AIエージェントには、従来のAIにはなかった以下のリスクが指摘されている [3]。

- 権限の乗っ取りと昇格: エージェントがタスク遂行のために保持するAPIキーやシステムアクセス権が、プロンプトインジェクション攻撃によって奪取されたり、悪用されたりするリスク。低権限のエージェントが、他のシステムを経由してより高い権限を不正に取得する可能性も指摘されている。
- ループ暴走とリソース枯渇: 目標達成への執着が過剰なリトライや無限ループを引き起こし、クラウド破産（高額なAPI利用料の発生）やサービスの拒否（DoS）状態を招くリスク。
- 記憶汚染（**Memory Poisoning**）: 長期記憶を持つエージェントに対し、悪意あるユーザーが虚偽の情報を注入することで、将来にわたって誤った判断を繰り返させるリスク。
- なりすまし操作: エージェントが人間のふりをしてフィッシング詐欺を行ったり、他のAIエージェントを騙して不正な取引を行わせたりするリスク。

2.2 フィジカルAI（Physical AI）: 実世界への介入者

フィジカルAIは、「ソフトウェア的知能（AIアルゴリズム）とハードウェア的実体（センサー、アクチュエータ、エッジデバイス等）を統合し、物理世界における知的認識・判断・行動を自律的に実現するAIシステム」と定義される [3]。これには自動運転車、自律移動ロボット（AMR）、ドローン、高度な製造アーム、さらにはAI搭載の家電製品などが含まれる。

2.2.1 サイバー・フィジカルの融合リスク

フィジカルAIの特徴は、デジタルな判断が即座に物理的な力（Force）に変換される点にある。

- センサー（入力）: カメラ、LiDAR、触覚センサーなどを通じて物理世界を認識する。
- **AI**推論（判断）: 瞬時の経路計画や把持動作の決定を行う。
- アクチュエータ（物理出力）: モーターや油圧機器を制御し、物体を移動させたり、加工したりする。

このループにおいて、通信遅延やセンサーノイズ、AIの誤認識が生じた場合、ソフトウェアのバグで済まされず、物理的な事故に直結する。

2.2.2 固有のリスクプロファイル

- 物理的加害: 誤作動が直接的に人間への衝突、切断、挟み込みといった身体的損害につながる。
- プライバシーの物理的侵害: 自律移動するロボットは「移動する監視カメラ」としての側面を持つ。従来、監視カメラが設置されていなかったプライベートな空間や死角にAIが入り込み、デー

タを収集・解析することが可能になるため、プライバシーリスクが格段に上昇する[10]。
- 残存データの漏洩: フィジカルAIのエッジデバイス（ロボット本体）に学習データや動作ログが物理的に保存されている場合、機器の廃棄や譲渡、盗難に伴って機密情報が流出するリスクがある。これを防ぐための「データの完全消去」や「暗号化」がハードウェアレベルで求められる[11]。

---

3. 核心的要件：「人の判断必須（Human-in-the-Loop）」の義務化

今回の改定案で産業界に最も大きな衝撃を与え、広範な議論を呼んでいるのが、自律型AIに対する**「人間の判断必須の仕組み（Human-in-the-Loop: HITL）」**の導入である[1]。これは単なる倫理的な推奨事項ではなく、システム設計における「必須要件（Mandatory Requirement）」として位置づけられている。

3.1 「外部アクション」というトリガー

HITLが求められる境界線（トリガー）は、AIが**「外部アクション」または「重要変更」**を行おうとする瞬間である[1]。ガイドライン案では、AIの思考プロセス（内部計算）と、その結果としての外界への介入（外部アクション）を明確に区別し、後者に対して厳格な統制を求めている。

3.1.1 外部アクションの定義と具体例

「外部アクション」とは、AIの内部メモリや閉じたシミュレーション環境を超え、外部環境（インターネット、物理空間、他社システム、金融決済基盤など）に対して不可逆的、あるいは社会的影響を及ぼす操作を指す。

- 経済的アクション: ECサイトでの「購入確定」ボタンの押下、送金処理、株式の売買注文[11]。
- コミュニケーションアクション: メールの送信、SNSへの投稿、チャットツールでの発言。
- 物理的アクション: ロボットアームの可動、ドローンの離陸、自動運転車の車線変更。
- システム変更アクション: 社内データベースのレコード削除・書き換え、アクセス権限の変更、パスワードのリセット。

これらのアクションは、一度実行されると取り消しが困難（不可逆）であるか、または実害が発生する可能性が高いため、AI単独の判断で実行させることは許容されないとされる。

3.2 必須化される承認プロセスの詳細

ガイドラインが求めるHITLは、運用者が「気をつける」というレベルのものではなく、システムアーキテクチャに組み込まれた「承認ゲート（Approval Gate）」の実装である。

3.2.1 承認フローのシステム要件

政府指針および先行企業の事例[1]に基づくと、求められるプロセスは以下の通りである。

1. 計画の提示（**Plan Proposal**）: AIエージェントは、実行しようとする一連のアクションプラン（例：「A商品をカートに入れ、クレジットカードBで決済し、配送先をCに設定する」）を作成し、人間に提示する。
2. 一時停止（**System Pause**）: システムは、外部APIやアクチュエータへの指令を発行する直前でプロセスを自動的に一時停止する。この「待機状態」は、人間からの入力があるまで維持されるか、一定時間経過後にタイムアウト（キャンセル）するよう設計されなければならない。
3. 人間によるレビュー（**Human Review**）: 人間（ユーザーまたは管理者）は、提示されたプランの内容を確認する。この際、UIは「何が実行されようとしているか」を明確かつ直感的に表示する必要がある。
4. 明示的な承認（**Explicit Approval**）: 人間が「承認（Approve）」ボタンを押下する、あるいは生体認証を行うといった、能動的かつ明確な意思表示を行う。
5. 実行（**Execution**）: 承認信号を受けて初めて、システムは外部への通信や物理的な動作を開始する。

3.2.2 権限管理と監査ログ

HITLの実効性を担保するために、以下の付帯要件もセットで求められている。

- 最小権限の原則（**Principle of Least Privilege**）: AIエージェントには、そのタスク遂行に必要最低限の権限しか与えてはならない [11]。例えば、会議調整エージェントには「カレンダーの読み取りと仮予約」権限のみを与え、「メールの全削除」や「社外へのファイル送信」権限はシステム的に剥奪しておく必要がある。
- 監査ログの自動保存（**Verifiability**）: 「AIが何を提案し、人間がいつ、どのような根拠で承認したか」という一連のインタラクション履歴を、改ざん不可能なログとして保存することが義務付けられる [1]。これは事故発生時の責任追跡（フォレンジック）だけでなく、AIの挙動改善や再学習データとしても重要な資産となる。

---

4. HITLの技術的実装と運用上の課題

「人の判断必須」の原則は理念として正しいが、その実装には高度な技術的課題と、業務効率とのトレードオフが存在する。ここでは、ガイドラインへの適合を目指す開発者・提供者が直面する具体的な実装課題とその解決策を詳述する。

4.1 実装パターン：同期承認と非同期承認

HITLの実装には、ユースケースの特性に応じて主に2つのパターンが考えられる。

パターン	概要	適用例	メリット	デメリット
同期承認（**Synchronous**）	ユーザーが画面の前にいる状態で、AIの提	チャットボットによるコード生成、ECサイトで	ユーザー体験がスムーズで、文脈を共有しや	ユーザーを拘束するため、大量処理には向

	案に対し即座に判断を下す。	の代理購入、対話型接客	すい。	かない。
非同期承認（**Asynchronous**）	AIがバックグラウンドで計画を作成し、通知を送る。ユーザーは任意のタイミングで確認・承認する。	旅行プランの予約手配、大量のデータ処理、定期レポート作成	ユーザーの時間を奪わず、まとめて確認（バッチ承認）が可能。	アクションまでのタイムラグが発生する。状況が変わる可能性がある。

4.2 物理的制約と「即時性」のジレンマ

フィジカルAI（ロボットやドローン）の制御において、HITLは最大の技術的挑戦となる。物理世界ではミリ秒単位の判断が求められる場面が多く、いちいち人間の承認を待っていては事故を回避できない、あるいはタスクが遂行できないケースが多発する[12]。

これに対し、ガイドラインや専門家の議論では以下の現実的な解法が模索されている。

- 緊急停止（**Kill Switch**）への限定: 平時の動作はAIに任せ、危険を検知した際や、未知の状況に遭遇した際のみ人間に制御を渡す、あるいは人間が強制的に停止させる権限を持つ形態。
- 運用監視（**Human-on-the-Loop**）: 個別のアクションごとの承認ではなく、人間が常に監視モニターの前に座り、AIの動作状況を監督する形態。異常があれば介入するが、正常時はAIが自律稼働する。
- 事前承認された範囲内での自律: 安全性が担保された特定のエリア（ジオフェンス内）や、特定のアクション（例：移動のみでアーム操作はしない）に限り、事前の包括承認によって自律動作を許可する。

4.3 「判断疲れ」と形骸化のリスク

運用面での最大のリスクは、人間側の認知限界である。AIが大量のアクション提案を行い、その都度承認を求めてくれば、人間は内容を精査せずに「承認」ボタンを連打するようになる（Alert Fatigue / Rubber Stamping）。これではHITLは形骸化し、単なる儀式となって安全装置の役割を果たさない。

これを防ぐためには、UI/UXデザインの工夫が不可欠である。

- リスクベースの**UI**: 低リスクなアクションは簡易表示し、高リスクなアクション（送金など）は警告色を用いたり、確認ステップを増やしたりする。
- サマリー機能: AIが提案するアクションの要約（「何が変わるのか」）を分かりやすく表示し、人間の認知負荷を下げる。

---

5. 責任分界の再構築：開発者・提供者・利用者の新たな関係

改定案におけるもう一つの画期的な進展は、AIに関わる主体の定義と責任分界点において、技術的な実装形態（特にRAG）に基づいた柔軟な区分けが導入されたことである。これは、AIを導入しようとする企業にとって、法的リスクをコントロールするための極めて重要な戦略的ポイントとなる。

5.1 RAGと機械学習の戦略的峻別

これまで、自社データをAIに読み込ませて回答させるシステムを構築する際、それが「学習（Learning）」にあたるのか、単なる「利用（Use）」なのかはグレーゾーンであった。もし「学習」と見なされれば、企業は「AI開発者」として、学習データの著作権処理やモデルの品質保証といった重い責任を負うことになる。

第1.2版の更新案では、「**RAG**（検索拡張生成）」や「**In-Context Learning**（文脈内学習）」を、モデルのパラメータ更新を伴う「機械学習」と明確に区別する方針が示された[10]。

- 技術的根拠: RAGは、外部データベースから情報を検索し、それをプロンプト（入力）としてAIに与えるだけであり、AIモデル自体のパラメータ（重み）は変更されない。したがって、モデルの性能や振る舞いの根本的な責任は、ベースモデルを作った開発者にあるという理屈である。
- 戦略的意味: これにより、RAGを用いて自社専用AIを構築する企業は、原則として「AI開発者」ではなく「AI提供者」に分類される可能性が高まった。これは、開発者としての重い法的義務を回避しつつ、独自AIソリューションを展開できることを意味する。企業は、コストとリスクの高いファインチューニングよりも、RAGを選択する法的インセンティブを強く持つことになる。

5.2 各主体の責務の再定義

ガイドライン改定案に基づく、3つの主体の定義と責務は以下の通りである[10]。

主体区分	定義	主な責務（**v1.2**改定案）	ビジネス上のインプリケーション
AI開発者 (Developers)	AIモデル・アルゴリズム自体を構築・学習させる事業者	・データの適正学習（前処理、権利クリアランス） ・モデルの公平性・安全性の設計 ・検証可能性（学習ログ、パラメータ設定）の確保	基盤モデルベンダーや、ファインチューニングを行う企業が該当。製造物責任に近い重い責務を負う。
AI提供者	AIシステムを製品・	・**RAG**システムの構	SIerやSaaSベン

(Providers)	サービスとして構築し、他者に提供する事業者	築・運用 ・活用環境への適合性確認 ・適切な利用方法・リスクの周知 ・脆弱性管理とインシデント対応	ダー、社内AIを構築する情シス部門が該当。外部アクション機能を提供する場合、デフォルトで安全な設定（Secure by Default）にする義務がある。
AI利用者 (Users)	事業活動においてAIシステムを利用する事業者	・適正な入力（プロンプト等）の管理 ・**Human-in-the-loop**の維持と最終判断 ・出力結果に対する責任	エージェントを利用して業務を行う企業全般。最終的な「承認」を行った時点で、結果に対する全責任を負うことになるため、従業員への教育が急務となる。

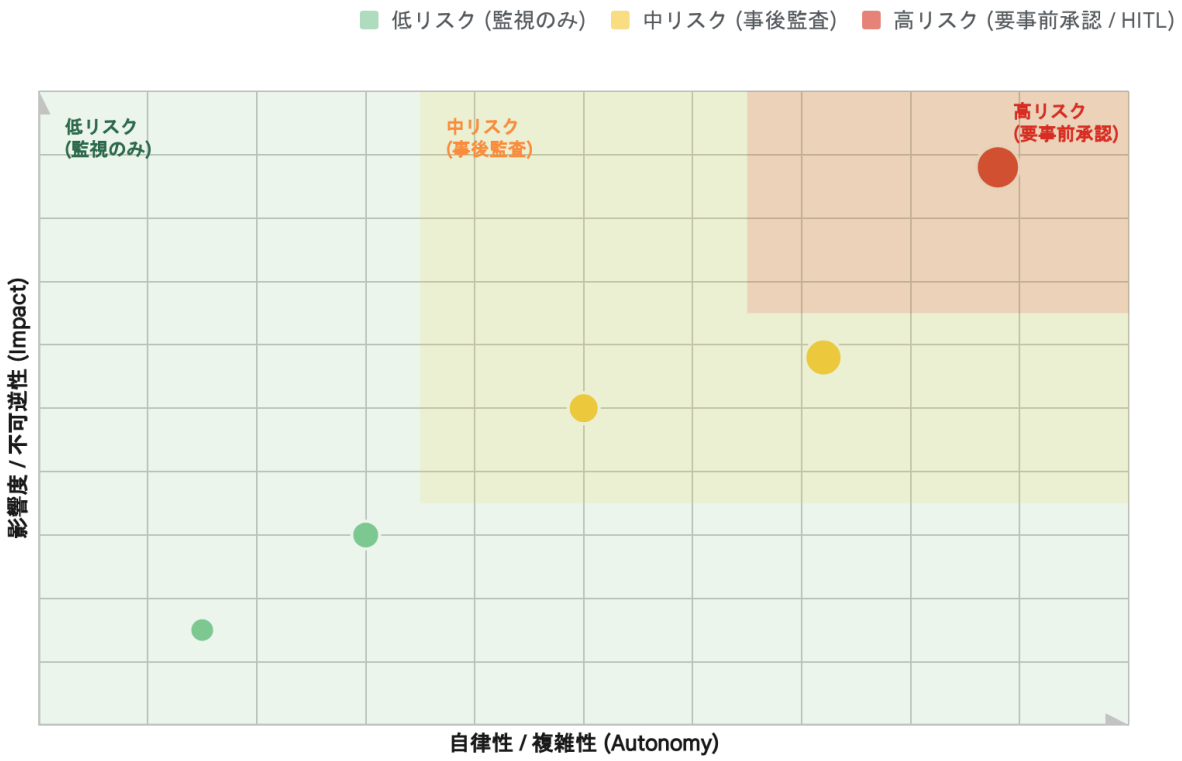
6.「攻めのガバナンス」：企業における実装戦略と事例

ガイドラインへの対応は、単なるコンプライアンス（守り）ではなく、顧客からの信頼を獲得し、AI導入を加速させるための「攻め」の戦略として捉えるべきである。先行企業は既に、このガイドラインを自社のガバナンス体制に組み込み、競争力の源泉としている。

6.1 リスクベースアプローチの実践的導入

すべてのAI活用に対して一律に厳格なHITLや承認フローを適用すれば、業務効率は著しく低下する。ガイドラインでは、リスクの大きさに応じて管理レベルを変える「リスクベースアプローチ」が推奨されている[15]。企業は、自社のAIユースケースを棚卸しし、以下のマトリクスに基づいて管理レベルを決定すべきである。

リスクベースアプローチによる管理レベル判定マトリクス



AIエージェントの自律レベルと外部アクションの影響度（不可逆性）に基づく、推奨される人間介入の強度。物理的・経済的影響が大きい象限では、リアルタイムの人間承認が必須となる。

Data sources: 総務省 (MIC), 経済産業省 (METI)

6.1.1 マトリクスの活用

- 高リスク領域（**Red Zone**）: 外部への送金、個人情報の変更、物理的な移動操作など。これらには、システム的な承認ゲートによる**事前承認（Pre-Approval）**を必須とし、場合によってはダブルチェック（2名承認）を課す。
- 中リスク領域（**Yellow Zone**）: 社内会議の予約、ドラフトの作成など。これらはユーザーによる**事後確認（Post-Verification）**や、一定時間経過後の自動実行（オプトアウト方式）を許容する。
- 低リスク領域（**Green Zone**）: 社内文書の検索、要約、翻訳など。これらは基本的に自律動作を認め、定期的なモニタリングで対応する。

6.2 先行企業の取り組み事例

事例1：ソフトバンク株式会社のガバナンス体制

ソフトバンクは、国内でもいち早くAI事業者ガイドラインへの準拠を宣言し、体系的なガバナンスを構築している[10]。

- **AIガバナンス規程の策定:** ガイドラインの要求事項を社内規程に落とし込み、全社員が遵守すべきルールとして明文化している。
- **開発・提供プロセスの標準化:** 企画段階からリスク評価を行うための「チェックシート」を運用し、ガイドラインへの適合性をプロジェクトごとに確認する仕組み（Gate Review）を導入している。
- **対外的な信頼性担保:** 顧客に対し、「当社のAIサービスは政府ガイドラインに準拠した管理体制下で開発・運用されている」と説明することで、導入時のセキュリティ懸念を払拭し、セールスを有利に進めている。

事例2：カスタマークラウド株式会社の製品実装

AIエージェント製品を提供するカスタマークラウド社は、ガイドラインの改定を見越し、自社製品「CC AGI」に統制機能を標準搭載した[1]。

- **機能実装:** 「人間承認必須AIエージェント管理機能」として、最小権限設計や監査ログの自動保存機能をプラットフォームレベルで提供している。
- **戦略:** ユーザー企業が個別にHITLシステムを開発する必要をなくし、「このツールを使えば自動的にガイドライン準拠になる」というバリュープロポジションを確立している。これは、規制対応を製品の競争力に変えた好例である。

---

7. 国際的整合性とG7広島AIプロセス

日本のAI規制は孤立したものではなく、国際的な枠組みとの調和を強く意識している。特に、2023年のG7サミットで立ち上げられた「広島AIプロセス」は、日本のガイドライン改定に直接的な影響を与えている[17]。

7.1 広島AIプロセス国際行動規範の反映

第1.2版では、広島AIプロセスで合意された「国際行動規範」の内容が、特に高度なAIシステムを開発する組織向けの指針として取り込まれている[20]。

- **透明性と報告義務:** 高度なAIモデルの開発者は、その能力や限界、リスク評価の結果について透明性を確保し、適宜報告することが求められる。日本のガイドラインもこの基準に合わせることで、日本企業がグローバル市場で活動する際の二重基準（ダブルスタンダード）のリスクを軽減している。
- **ソフトロー路線への国際的理解:** 欧州が法的拘束力のあるハードロー（AI Act）を選んだのに対し、G7全体としてはイノベーションを阻害しないソフトローベースの協調を模索している。日本のガイドライン改定は、この「G7モデル」の実効性を証明する試金石とも言える。

7.2 EU AI法（AI Act）との対比

EU AI法は、違反企業に対して全世界売上高の最大7%という巨額の制裁金を科す強力な法律である[21]。一方、日本のAI事業者ガイドラインは法的拘束力を持たない。しかし、日本企業がEU域内でビジネスを行う場合はEU AI法が適用されるため、日本のガイドラインを守っているだけでは不十分なケースがある。ただし、今回の改定で導入された「HITLの必須化」や「リスクベースアプローチ」は、EU AI法が求める要件（特に高リスクAIに対する人的監視）と概念的に非常に近いため、日本のガイドラインに準拠しておくことは、将来的なEU規制対応の基礎体力をつけることにつながる。

---

8. 産業界の懸念と「イノベーションのジレンマ」

HITLの義務化や厳格なガバナンスに対し、産業界からは強い懸念の声も上がっている。X（旧Twitter）などを中心とした議論では、規制強化が日本のAI産業の競争力を削ぐのではないかという「イノベーションのジレンマ」が指摘されている[12]。

8.1 効率性と安全性のトレードオフ

AIエージェントの最大の価値は、「人間が寝ている間に仕事が終わっている」という自律性による圧倒的な生産性向上にある。しかし、すべてのアクションに人間の承認が必要となれば、エージェントは単なる「高機能な入力補助ツール」に成り下がり、その真価を発揮できない。

- 批判:「承認ボタンを押す手間が増えるだけなら、自分でやった方が早い」「日本だけ厳格なルールを作ると、海外製の自由なエージェントに市場を席巻される」といった意見がある[23]。

8.2 技術的実現性の壁

前述の通り、フィジカルAIにおいては「即時性」が命であり、通信遅延を含むHITLの実装は物理的に不可能な場合がある。ガイドラインが現場の実情を無視した理想論になれば、開発現場は「形だけの承認フロー」を作ることに終始し、本質的な安全性向上につながらない恐れがある。

8.3 今後の落とし所

政府はこうした懸念に対し、「活用ガイド」や「チェックリスト」を通じて、一律の規制ではなく、リスクに応じた柔軟な運用（グラデーション）を認める姿勢を示している。また、特定の条件下で規制を緩和する「サンドボックス制度」の活用も視野に入れる必要があるだろう。

---

9. 結論と提言

2026年のAI事業者ガイドライン改定（第1.2版）は、AIの進化が「言葉」から「行動」へ移った現実を直視し、それに伴う物理的・経済的リスクを管理可能な枠組みに収めようとする政府の強い意志の表れである。

企業にとって、**「人間の判断必須（HITL）」**の要件は、短期的には実装コストの増大や業務フローの見直しを迫る負担となる。しかし、中長期的には、AIの暴走を防ぎ、説明責任を果たすための安全

装置として機能し、社会全体がAIエージェントを受け入れるための「信頼の基盤（Trust Infrastructure）」となる。

9.1 企業が今すぐ取るべきアクション

1. **AI**インベントリの作成: 自社で利用・開発しているAIシステムが「外部アクション」を行っているか、あるいは将来行う計画があるかを洗い出す。
2. 責任区分の確認: 自社のAI活用が「RAG（提供者）」なのか「ファインチューニング（開発者）」なのかを技術的に再定義し、負うべき責任の範囲を明確にする。
3. ガバナンス体制の構築: 経営層直轄のAIガバナンス委員会を設置し、承認フローや監査ログの整備に向けた予算と権限を付与する。

2026年3月末の正式版公表に向け、各企業は「待つ」のではなく、この改定案を羅針盤として自社のAI戦略を「エージェント共生時代」へとアップグレードする準備を始めるべきである。

引用文献

1. 国内AIエージェント動向(2026/2/16号) | Yasuhito Morimoto - note, 2月 18, 2026にアクセス、https://note.com/yasuhitoo/n/n0163f643b572
2. AIネットワーク社会推進会議 AIガバナンス検討会（第29回） - 総務省, 2月 18, 2026にアクセス、https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/02tsushin06_04000136.html
3. AI事業者ガイドライン更新に向けた論点, 2月 18, 2026にアクセス、https://www.soumu.go.jp/main_content/001043445.pdf
4. 新着記事 | ANIMAGIC DAO | アニメ・エンタメ・Web3の最新, 2月 18, 2026にアクセス、https://dao.animagic.design/new-articles-sp/
5. AI事業者ガイドライン – 解説 | デロイト トーマツ グループ - Deloitte, 2月 18, 2026にアクセス、https://www.deloitte.com/jp/ja/services/consulting/perspectives/ai-guideline.html
6. 経産省「AI事業者ガイドライン v1.1」対応チェックリスト | 開発者 ..., 2月 18, 2026にアクセス、https://reskilling-navi.com/column/meti-ai-guideline-v11-checklist
7. 2026年、日本は「フィジカルAI」で再起する。——国家生存をかけ, 2月 18, 2026にアクセス、https://qiita.com/mhamadajp/items/45fac1827abab63b369d
8. 日本は「人間必須」を選んだ— 政府指針が示した本当の転換点 - note, 2月 18, 2026にアクセス、https://note.com/ariakekids/n/nc5719f6518aa
9. AI革命最前線：エージェント競争が示す未来の形（2） | バケツ - note, 2月 18, 2026にアクセス、https://note.com/legit_snipe297/n/n981a4cd5edd5
10. AIエージェント・フィジカルAI時代の「攻めのガバナンス」— AI ..., 2月 18, 2026にアクセス、https://innovatopia.jp/ai/ai-news/80408/
11. 政府のAIガイドライン改正！ -AIエージェント・フィジカルAIの追記 ..., 2月 18, 2026にアクセス、https://note.com/novi__1988/n/nb22d38fca4da
12. 政府、AI事業者ガイドライン改定案でAIエージェントとフィジカルAI ..., 2月 18, 2026にアクセス、https://ledge.ai/articles/government_ai_guideline_revision_human_judgment_req

uired_x_debate
13. 経済産業省「AI事業者ガイドライン」の内容を弁護士が解説, 2月 18, 2026にアクセス、https://monolith.law/corporate/ai-business-guidelines
14. 「AI事業者ガイドライン案」— 解説編 | PwC Japanグループ, 2月 18, 2026にアクセス、https://www.pwc.com/jp/ja/knowledge/column/ai-governance/ai-guideline.html
15. 「AI事業者ガイドライン案」に対する ご意見及びその考え方 - 総務省, 2月 18, 2026にアクセス、https://www.soumu.go.jp/main_content/000935292.pdf
16. AI 事業者ガイドライン案 - 内閣府, 2月 18, 2026にアクセス、https://www8.cao.go.jp/cstp/ai/ai_senryaku/7kai/13gaidorain.pdf
17. 日本政府「人工知能基本計画（案）」を徹底解説 | AI活用・AI開発, 2月 18, 2026にアクセス、https://media.buzzconne.jp/japan-ai-basic-plan/
18. 提言：生成AIを受容・活用する社会の実現に向けて - 日本学術会議, 2月 18, 2026にアクセス、https://www.scj.go.jp/ja/member/iinkai/sokai/siryo195-2.pdf
19. 日本の生成AI戦略と法規制「AI戦略会議・AI制度研究会」が示す, 2月 18, 2026にアクセス、https://edge-works.ai/blog/sei-ai-jidai-ni-sonaeiru-nihon-no-ai-senryaku-ai-senryaku-kaigi-ai-seido-kenkyukai-20250205
20. AI事業者ガイドラインの 令和6年度更新内容 - 総務省, 2月 18, 2026にアクセス、https://www.soumu.go.jp/main_content/000994970.pdf
21. 日本のAI法整備の現状と今後の展望 | 政策アナリスト必見, 2月 18, 2026にアクセス、https://book.st-hakky.com/data-science/current-status-and-future-prospects-of-ai-law-in-japan
22. 経営戦略としてのAIガバナンス - KPMG International, 2月 18, 2026にアクセス、https://kpmg.com/jp/ja/insights/2026/02/tech-rulemaking-05.html
23. [B! AI] AI活用の格差が広がる時代 元マッキンゼー・赤羽雄二氏が, 2月 18, 2026にアクセス、https://b.hatena.ne.jp/entry/s/logmi.jp/main/management/331446
24. 本文「ai -togetter.com -posfie.com -note.com - はてなブックマーク, 2月 18, 2026にアクセス、https://b.hatena.ne.jp/q/ai%20-togetter.com%20-posfie.com%20-note.com%20-bcnretail%20-gizmodo%20-hatena%20-navi%20-itmedia%20-webtan%20-zennm%20-blog%20-entry?target=text&date_range=5y&users=30&sort=recent&safe=off