AI事業者ガイドライン改定案（第1.2版）の深堀分析

概要

2026年2月16日、総務省の第29回AIガバナンス検討会において「AI事業者ガイドライン」の令和7年度更新内容（案）が提示された。正式版（第1.2版）は2026年3月末に公開予定であり、今回は最終確定前の「更新案」段階である。本改定の最大の特徴は、**AIエージェント**と**フィジカルAI**という2つの新技術カテゴリへの対応を正面から打ち出し、「人間の判断を必須とする仕組み（Human-in-the-Loop）」の構築を開発企業に求める方針を明記した点にある。[1][2][3][4][5]

改定の背景と経緯

検討プロセスのスケジュール

時期	会議	内容
2025年10月	事業者向け意見照会	ガイドライン利活用に関する課題・意見収集[6]
2025年12月2日	第28回AIガバナンス検討会	更新に向けた論点提示[6]
2026年2月16日	第29回AIガバナンス検討会	更新内容（案）の提示[3][5]
2026年3月下旬（予定）	第32回推進会議・第30回検討会（合同）	修文案の検討・公開承認[5]
2026年3月末（予定）	–	AI事業者ガイドライン第1.2版の公開[5]

第28回検討会で示された更新論点に対し、構成員から多数の意見が寄せられた。特にAIエージェントとフィジカルAIへの対応については「絶対重要」「喫緊の課題」との声が多く、今回の改定の中核に据えられている。[6]

なぜ今改定が必要なのか

現行のAI事業者ガイドライン（第1.1版、2025年3月公表）には、AIエージェントの用語定義すら含まれておらず、便益は「生産性の向上」への限定的な言及のみ、リスクも脚注での可能性言及にとどまっていた。フィジカルAIに至っては、定義・便益・リスクのいずれの記載もなかった。一方、PwCの調査によれば約70%の事業者がAIエージェントを導入中または導入検討中であり、フィジカルAI市場もCAGR 33.5%の急成長が予測されるなど、技術の社会実装が規制の整備を大きく先行している状況にあった。[6]

論点①：AIエージェントへの対応

AIエージェントの公式定義

更新案では、AIエージェントを以下のように定義している。[6]

> **AIエージェントとは、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**

この定義は、Microsoft、Google、Amazon、IBM、Accentureなど主要グローバル企業の定義を参照しつつ策定された。便益としては「ユーザーの意図を理解し自律的にタスクを遂行することで、複雑な業務プロセスを効率化し、人的負荷を大幅に削減できる」とされている。[6]

エージェンティックAI（マルチエージェント）

AIエージェントの上位概念として「エージェンティックAI」も新たに説明が追加された。[6]

> **エージェンティックAIとは、AIエージェントよりも包括的かつ進化的な概念であり、複数のAIエージェントにより自律的に意思決定を下しアクションを起こす目標主導型のAIシステム**

マルチエージェントの構成パターンとして、以下の4類型が整理されている。[6]

- **シングルエージェント**：単一エージェントで動作。判断は必要だが複雑な意思決定が不要なケース
- **協働型（横並び）**：各エージェントがリアルタイム情報に基づき協力。リソース共有が必要なケース
- **競争型（横並び）**：多様な戦略で競争し革新と最適化を促進。研究開発など創造的なケース
- **階層型（指揮と部下）**：明確な指揮構造で一貫性を維持。集中管理が必要なケース

AIエージェントのリスク（12類型）

更新案では、AIエージェント固有のリスクを含む12のリスク類型が新たに整理された。具体的なインシデント事例とともに提示されている。[6]

リスク	概要	特有/共通
悪意ある入力で誤作動	不正な指示により本来と異なる行動を取る	共通
制約回避した不正行動	人間の意図しない方法で制約を破る	共通
判断根拠が不明瞭	非決定的な判断で根拠の追跡が困難	共通
誤情報の拡散	間違いを繰り返し学習・出力して広める	共通
ツールの悪用	許可範囲のツールで意図しない操作を実行	**特有**
コードの悪用	生成コードが不正操作や侵入に利用される	**特有**
権限の乗っ取り	他のシステムから権限を奪い高い権限を取得	**特有**
なりすまし操作	他のAIを装い不正行為を行う	**特有**
偽情報の混入	通信に虚偽情報を加え協調行動を妨げる	**特有**
誤情報の記憶汚染	間違った情報を記憶し将来の判断に悪影響	**特有**
悪意あるAIの侵入	マルチエージェント環境全体の安全性低下	**特有**
人間の過信誘導	AIを過信させて有害な行動に導く	共通

このうち、ツールの悪用、コードの悪用、権限の乗っ取り、なりすまし操作、偽情報の混入、誤情報の記憶汚染、悪意あるAIの侵入の7項目は**AIエージェント特有**のリスクとして位置づけられている。[6]

AIエージェントの主体別対策

対策は5つの方針に分類され、AI開発者・AI提供者・AI利用者の3主体それぞれに対して留意事項が設定された。[6]

対策方針	AI開発者	AI提供者	AI利用者
権限管理と不正利用防止	APIキーや認証情報の管理強化、最小権限設計の徹底	セキュリティ要件の定義	操作履歴の定期確認

自律行動の制御	ガードレール設計、モデル参照データの制限	−	**HITL（ヒューマンインザループ）により高リスク操作は人間承認を必須化**
メモリの健全性確保	メモリ更新データの制限や信頼性の評価	メモリ管理ポリシー策定（保存期間・削除ルール等）	誤情報を発見した際の即時報告、機密情報入力の防止
通信の安全性確保	システム間通信の暗号化、通信ログの監査と異常検知	−	外部接続の必要最小限化、エージェント連携の制限
AI過信・要求過多の防止	出力に不確実性指標を付与（信頼度評価）	AI利用者向け教育（AIの限界・リスク）	大量要求を避け、優先度を明確化

特に注目すべきは、**「HITL（ヒューマンインザループ）により、高リスク操作は人間承認を必須化」**という記載が、AI利用者の対策として明記された点である。これは日経新聞が報じた「人の判断必須の仕組みを」という方針の具体的な中身に当たる。[2][1][6]

論点②：フィジカルAIへの対応

フィジカルAIの公式定義

更新案では、フィジカルAIを以下のように定義している。[6]

> **フィジカルAI（Physical AI）とは、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**

5つの技術的特徴（センサーによる環境認識、データ処理・推論、リアルタイム意思決定、アクション実行、学習・適応）が体系的に整理されている。日経新聞では「目的達成の最適な方策を自律的に推論・判断し物理的な行動につなげるシステム」と記載すると報じられている。[2][6]

フィジカルAIの便益

便益として以下の3点が明記された。[6]

- 少子高齢化による**労働力不足を補い**、人と協働して生産性を向上させ、あらゆる産業や現場の自動化と効率化を実現

- **危険な環境で人の代わりに作業**を行い、安全性を高めつつリスクを低減
- 介護や生活支援を通じて人々の**自立とQOL向上**に寄与し、福祉や医療などの分野で新たな支援の形を創出

フィジカルAIのサービス事例

検討会資料では、以下の4事例が紹介されている。[6]

- **自動運転システム**（TESLA）：車両周囲をセンサーで認識しAIが走行判断や操作を自動化
- **巡回警備ロボット**（SECOM「cocobo」）：カメラ・センサーで映像解析・行動認識を行い、施設内異常を自律検知・通報
- **清掃ロボット**（アイリスオーヤマ「DX清掃ロボットジルビー」）：LiDARや3Dカメラで環境認識、清掃ルート最適化・状況学習
- **自律型ロボットアーム**（安川電機「MOTOMAN NEXT」）：自ら判断・計画する自律適応型ロボット

フィジカルAIのリスク（8類型）

リスク	概要	特有/共通
個人情報の無断収集	センサーを通じて周囲の個人情報が意図せず取得される	**特有**
センサー誤作動	光やノイズ、妨害により環境を誤認識する	**特有**
学習データの偏り	不適切なデータにより誤判断や不公平な行動が生じる	共通
物理的事故の発生	ロボットの誤作動で人や物に損害を与える	**特有**
倫理的悪用	自律兵器や監視用途など、倫理的に問題のある利用に転用	共通
意図に反する学習	目標達成のため危険な手段を自発的に学ぶ	**特有**
長期運用の不安定化	ハード劣化や未知の環境で性能が低下・異常動作	**特有**
判断のブラックボックス化	内部処理が不透明で原因特定や責任追及が困難	共通

フィジカルAI特有のリスクは5項目で、個人情報の無断収集、センサー誤作動、物理的事故、意図に反する学習、長期運用の不安定化が該当する。[6]

フィジカルAIの主体別対策

対策方針	AI開発者	AI提供者	AI利用者
プライバシー保護・データ管理	個人情報の最小化と匿名化技術の実装	利用者や業務外利用者への説明事項連携	収集データの適法性確認と不要情報の削除
データの偏り防止	多様性を確保したデータによる学習と偏り検出機能の組込み	–	異常兆候を発見した際の即時報告
デジタルツインによる事前検証	シミュレーション環境の整備と異常シナリオの事前検証	–	–
安全設計とフェイルセーフ	緊急停止機能等の実装	定期的な安全性検証や障害対応手順の明示	利用シーンの適切な見極めと安全プロトコルの遵守
責任所在と解釈性の明確化	結果の出力過程を明確化する工夫	業務外利用者への必要情報の連携	–

なお、リスクが顕在化し事故が発生した際の民事責任の在り方については、経済産業省主催の「AI利活用における民事責任の在り方に関する研究会」の検討結果を今後ガイドラインに反映していく想定とされている。[6]

論点③：リスクベースアプローチの導入

リスク評価手法の追加

現行ガイドラインではリスクベースアプローチの重要性には触れつつも、具体的な手法は事業者に委ねられていた。今回の更新案では、2つのリスク評価手法が新たに提示された。[6]

- **手法案1：リスク＝影響度（1件あたりの深刻さ）× 規模（影響が及ぶ範囲の大きさ）**
    - 長所：発生確率が不明でも社会的・倫理的影響を考慮しやすい
    - 留意点：定量化が難しく主観に依存する部分がある
- **手法案2：リスク＝事故損失額（1件あたりの具体損失額）× 事故発生確率**

- 長所：数値化しやすく、コストベネフィット分析に活用できる
- 留意点：確率や損失額の正確な見積もりが難しい場合があり、AI Incident Database等の参照が推奨される

特に留意すべきユースケース

EU AI Actでハイリスクとして定義されている8領域が参考として提示されている。[6]

1. **バイオメトリクス**（公共空間での顔認証、感情認識AI等）
2. **重要インフラ**（電力網、交通、上下水道等）
3. **教育・職業訓練**（入学可否判定、学習成果評価等）
4. **雇用・労働管理**（採用選考、パフォーマンス評価等）
5. **公共・民間サービス**（医療・金融・保険のアクセス判定等）
6. **法執行**（犯罪予測、捜査支援等）
7. **移民・国境管理**（ビザ・入国審査等）
8. **司法**（量刑推奨、司法判断支援等）

論点④：RAGと機械学習の概念整理

戦略的に重要な峻別

更新案では、**RAG（検索拡張生成）とIn-Context Learning**が、従来の**機械学習（パラメータ更新を伴う学習）**と明確に区別して整理された。RAGはモデルのパラメータを更新しないため、多くの事業者は「AI開発者」としての義務を負わず、「AI提供者」「AI利用者」としての責任管理に注力できる可能性が高まる。[3]

この区分整理の実務的意義は大きい。RAGシステムの構築やファインチューニングを行う企業が、形式的に「利用者」であっても実質的には「提供者」や「開発者」の責任を併せ持つケースがあり、今回の更新で各主体区分の境界に補足説明が追加された。[3]

論点⑤：ガイドラインの利活用推進

現状の課題

事業者向けアンケート調査の結果、ガイドラインの**認知度は81%**と高い一方、**活用度は46%**に留まることが明らかになった。事業者や構成員から挙げられた主要課題は以下の4点に集約される。[6]

- **分量の多さ**：「全体像を簡単に理解するのが困難な文章量」
- **読みたい箇所の探しづらさ**：「項目間の対応・依存関係が不明」「PDFの検索性が不十分」
- **内容の分かりづらさ**：「"適切な"等の漠然とした表現」「手順（How）が分からない構成」
- **動機付けの弱さ**：「遵守への強制力がなく事業者の自主性に依存」

具体的な改善策

取り組みの方向性	具体策（案）
ガイドラインの理解促進	**活用ガイドの作成**（経済産業省で検討中、今年度末公開目標）[6]、解説動画の作成
検索性に優れる補助ツール	**チャットボットの提供**（総務省）[3]、Webツールの提供
提供形式の変更	HTML形式での提供、スマートフォン対応、PDFへのしおり付与、ガイドラインの分割[6]
利用を促す仕組み	**ベストプラクティス表彰制度**の策定[6]

先行事業者の活用事例

ソフトバンクの取組み

第29回検討会では、ソフトバンクのAIガバナンス推進室長・浦野憲二氏が活用事例を発表した。同社はガイドラインを**「リスク対策」と「リテラシー向上」**の2カテゴリで組織の意思決定プロセスに組み込んでいる。[3]

リスク対策面では：[3]

- ガイドラインの評価軸に基づくAI倫理ポリシー・社内規程の整備（ルール）
- AI案件のチェックシートへの反映（オペレーション）
- 社内AIガバナンス検討会議における開発者・提供者・利用者の視点整理（判断軸）
- リスクベースアプローチに基づく継続的改善の監査プロセスへの導入（内部監査）

リテラシー向上面では：[3]

- 経営層向けAIガバナンスの「経営課題」化
- 全社員向けe-learning・勉強会
- **顧客との議論におけるビジネス上の信頼性担保ツール**としての活用
- 他社の実践事例の研究

特に、ガイドラインを「お客様との信頼性担保」のツールとして営業の現場で活用している点が注目に値する。[3]

AIガバナンス協会（AIGA）の実装ソリューション

AIGAからは「**AIガバナンスナビ**」による組織の成熟度診断や、許可なく利用される「**シャドーAI**」の調査手法など、実地での実装ソリューションが紹介された。[3]

国際的文脈との連動

広島AIプロセスとの整合

今回の更新案は、**広島AIプロセス**との連動を強く意識している。広島AIプロセスは2023年のG7広島サミットで立ち上げられた国際的枠組みで、「全てのAI関係者向けの広島プロセス国際指針」と「国際行動規範」を含む。OECD報告枠組みへの参加やガイドライン準拠が、グローバルな取引における「安全・安心・信頼」の証明となり得ると位置づけられている。[7][3]

日本版AI法との関係

2025年5月に成立した「人工知能関連技術の研究開発及び活用の推進に関する法律」（AI法）では、人工知能基本計画に基づく指針の策定が規定されている。AI事業者ガイドラインはこのAI法の下での具体的なソフトロー・ガイダンスとして機能する位置づけにあり、法的拘束力はないものの、技術進化に合わせて柔軟に更新される「Living Document（生きた文書）」として設計されている。[8][3]

EU AI法との比較

更新案のリスクベースアプローチは、EU AI Actのリスク分類を参考にしつつも、日本独自の**アジャイル・ガバナンス**（ソフトローベース、柔軟に更新）を維持している。EUがトップダウン的な法規制（罰則付き）で臨むのに対し、日本は事業者の自主的取組みを促進しつつ国際基準との整合を図るアプローチを採用している。[9][10][3][6]

IP・ビジネスへの示唆

知的財産戦略上の注目点

- **AIエージェントの自律的行動による責任の所在**：AIエージェントが自律的に購買や予約を行うケースでは、利用者本人が知らないうちに行為が完了する可能性がある。このため「人間承認必須」の設計が求められ、特許出願においてもHITL機構の実装に関する技術が注目される[6]
- **フィジカルAIの安全設計と知財**：緊急停止機能、フェイルセーフ機構、デジタルツインによる事前検証などの安全設計技術は、特許による保護の対象になりやすい領域である
- **RAGの主体区分整理による影響**：RAGがパラメータ更新を伴わないことの明確化は、AIシステムの構築を行う事業者の法的責任範囲に影響し、ひいてはAI関連特許のクレーム設計にも示唆を与える

事業者が今すべきこと

1. **自社のAIシステムがガイドラインの3主体（開発者・提供者・利用者）のどこに該当するか再確認**[3]
2. **AIエージェント導入前のリスク予習**：エージェント特有の12リスク類型をチェックリストとして活用[3]
3. **フィジカルAI実装時の安全設計検証**：デジタルツインの整備や緊急停止機能の実装
4. **広島AIプロセスの報告枠組みへの準拠検討**：国際的な信頼の証明として早期対応が有利[3]
5. **3月末の第1.2版公開に向けたパブリックコメント等への対応準備**

参考：配布資料一覧（第29回AIガバナンス検討会）

第29回AIガバナンス検討会（2026年2月16日開催）では、以下の資料が配布されている。[4][11]

- 【資料1】令和7年度 今後の活動スケジュール
- 【資料2】広島AIプロセス等の国際・国内動向報告
- 【資料3】AI事業者ガイドラインに関する事業者による取り組みの動向報告
- 【資料4-1】AI事業者ガイドライン活用事例（ソフトバンク）

- 【資料4-2】AI事業者ガイドライン活用事例（AIガバナンス協会）
- 【資料5】AI事業者ガイドラインの令和7年度更新内容（案）