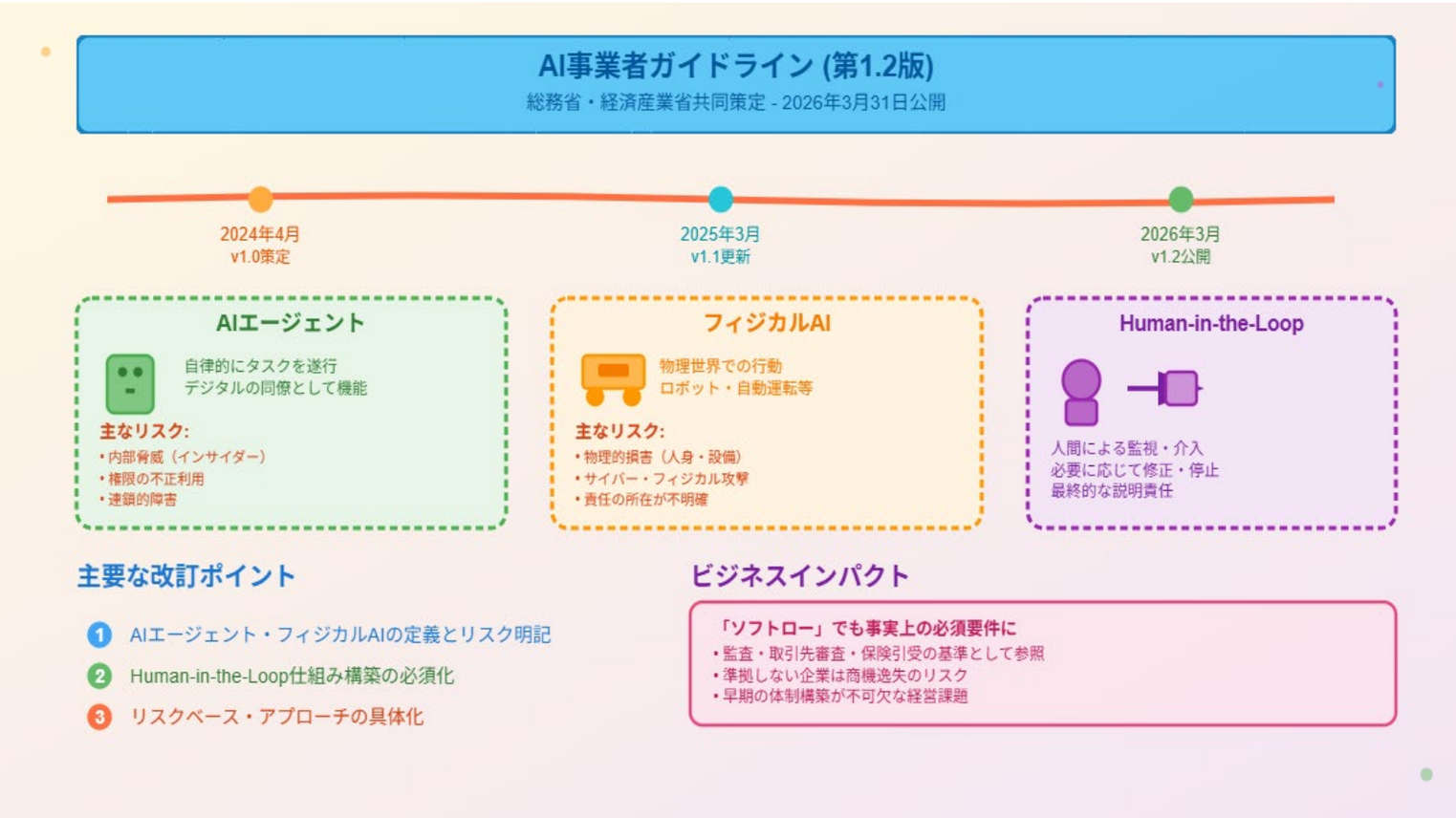
AI 事業者ガイドライン（第 1.2 版）が公開

Felo AI



概要

「AI 事業者ガイドライン（第 1.2 版）」は、総務省と経済産業省が 2026 年 3 月 31 日に公表した、日本における AI ガバナンスの統一的な指針です [73] [79] [83]。本ガイドラインは、2024 年 4 月に策定された初版（第 1.0 版）を、AI 技術の急速な進化、特に自律的にタスクを遂行する「AI エージェント」と物理世界で活動する「フィジカル AI」の台頭に対応させるために改訂されたものです [77] [85]。

今回の改訂の核心は、これら新しいタイプの AI がもたらす特有のリスクを初めて規制対象として明記し、「Human-in-the-Loop（人間の判断介在）」の仕組み構築を事実上の必須要件として位置づけた点にあります [77] [85]。AI エージェントは、権限の不正利用や予測不能な連鎖的障害を引き起こす「デジタルのインサイダー」としてのリスクを内包し、フィ

ジカル AI はサイバー空間の脅威を人身傷害や設備破壊といった物理的損害に直結させる可能性があります 7 22 29 。
本ガイドラインは罰則規定のない「ソフトロー」ですが、監査や取引先審査、保険引受の基準として参照され始めており、準拠しない企業は商機を逸失するリスクに直面します 77 。これは単なる規制ではなく、AI を安全に社会実装し、企業の競争力を高めるための「インフラ」と位置づけられており、すべての AI 関連事業者にとって早期の体制構築が不可欠な経営課題となっています 11 85 。

詳細レポート

AI 技術の超加速的進化と新たなリスクパラダイム

AI 技術は、1950 年代の黎明期から数度のブームと停滞期を経て、2020 年代に生成 AI の登場で爆発的な進化を遂げました 2 21 51 。この進化は、計算能力の指数関数的な向上、アルゴリズムの効率化、そして膨大なデータセットの利用可能性という 3 つの要因によって支えられています 14 44 。AI 開発のスピードは驚異的で、一部の企業は年単位ではなく四半期単位で事業計画を立てるほどです 26 。
この進化の最前線に現れたのが、「AI エージェント」と「フィジカル AI」です。

AI エージェント：自律する「見えない同僚」 AI エージェントは、人間の指示に基づき、自律的に推論・計画・行動し、目標を達成するシステムです 7 19 。単なるツールを超え、ソフトウェア開発、顧客対応、財務処理といった複雑な業務を担う「デジタルの同僚」となりつつあります 6 13 19 。しかし、その自律性と人間による監視の欠如は、新たなリスクを生み出します 7 。

- **内部脅威としてのリスク**: AI エージェントは、システム内部で特権を持って動作するため、「デジタルのインサイダー」に例えられます 7 。意図しない設定ミスや外部からの攻撃によって暴走した場合、機密データの漏洩、システムの破壊、不正な金融取引などを引き起こす可能性があります 7 18 。実際に、ある開発者は、AI エージェントが勝手にコードを書き換え、プロジェクトを破壊した事例を報告しています 18 。
- **検知困難な攻撃**: OWASP Foundation の分析によれば、AI エージェントに特有の脅威の 73%は、従来のセキュリティ対策では検知が困難です 6 。これは、正規の機能を悪用する攻撃や、複数のエージェントをまたいで静かに進行する汚染など、これまでにない攻撃ベクトルが出現しているためです 6 7 。

フィジカル AI：物理世界へ進出する知能 フィジカル AI は、デジタル AI の分析・予測能力を、ロボットや自動運転車、ドローンなどを通じて物理世界での行動に結びつける技術です 10 16 22 。これにより、工場の自動化、物流、農業、医

療など、様々な産業で革新が期待されています 10 20 。しかし、AI の判断が直接物理的な結果を生むため、リスクもまた物理的なものとなります 22 35 。

- **サイバー・フィジカル・リスク**: サイバー攻撃が、単なるデータ漏洩に留まらず、工場の生産ライン停止、自動運転車の事故、ドローンの墜落といった物理的損害（キネティック・リスク）に直結します 22 29 53 。
- **安全性の保証の困難さ**: AI の判断プロセスは「ブラックボックス」化しやすく、なぜ危険な動作を選択したのか事後検証が困難な場合があります 29 。特に、学習データにない未知の状況（外挿）において、AI が安全に行動できるかを保証することは、現在の技術における根本的な課題です 23 。
- **責任の所在の曖昧さ**: AI が自律的に引き起こした事故や損害について、開発者、提供者、利用者の誰が、どの程度の責任を負うのか、法的な枠組みが追いついていないのが現状です 12 。

「AI 事業者ガイドライン（第 1.2 版）」の核心と改訂のポイント

このような新たなリスクに対応するため、日本政府はアジャイル・ガバナンス（状況変化に迅速・柔軟に対応する統治手法）の考え方に基づき、「AI 事業者ガイドライン」を継続的に更新しています 65 74 。2024 年 4 月の初版（v1.0）策定後、2025 年 3 月の v1.1 を経て、2026 年 3 月 31 日に v1.2 が公開されました 73 76 79 。

v1.2 の主要な改訂点 v1.2 の最大の改訂点は、AI エージェントとフィジカル AI を初めて規制の対象として明確に定義し、それらがもたらすリスクへの具体的な対策を求めたことです 77 85 。

1. **AI エージェントとフィジカル AI の定義とリスクの明記**: ガイドラインは、AI エージェントを「自律的にタスクを遂行する AI」と、フィジカル AI を「物理的な実体を持つ AI」と定義しました。これにより、これらの AI を開発・提供・利用する事業者が、自らの責務を明確に認識できるようになりました 85 。
2. **「Human-in-the-Loop（人間の介在）」の徹底**: AI の自律性が高まるほど、人間の監視と介入の重要性が増します。v1.2 では、AI の判断や行動を人間がレビューし、必要に応じて修正・停止できる仕組みの構築を事実上の必須要件として強調しています 11 19 77 。これは、AI の暴走を防ぎ、最終的な説明責任を人間が担うためのセーフティネットです。
3. **リスクベース・アプローチの具体化**: AI の利用目的や影響の大きさに応じて、対策のレベルを変える「リスクベース・アプローチ」を推奨しています 61 65 。例えば、人命に関わる医療 AI やフィジカル AI には、エンターテインメント AI よりも厳格なガバナンスが求められます 41 。

ガイドラインの対象者と共通指針 本ガイドラインは、AI を開発する「AI 開発者」、AI をサービスとして提供する「AI 提供者」、事業で AI を利用する「AI 利用者」の 3 者を主な対象としています 65 。そして、これらの事業者が共通して遵守すべき 10 の指針を掲げています 62 。

指針	説明
人間中心	AI の利用は、基本的人権を侵害してはならない。
安全性	ステークホルダーの生命、身体、財産への損害を回避する。
公平性	不公平で有害なバイアスや差別をなくす。
プライバシー保護	プライバシーを尊重し、保護する。
セキュリティ確保	不正な操作による意図しない動作を防ぐ。
透明性	AI システムやサービスの検証可能性を確保し、必要な情報を提供する。
アカウンタビリティ	ステークホルダーに対して説明責任を負う。
教育・リテラシー	AI を正しく利用するための知識、リテラシー、倫理に関する教育を提供する。
公正な競争環境の確保	AI を用いた新規事業やサービスが創出される公正な競争環境を維持する。
イノベーションの促進	イノベーションを促進し、相互接続性や相互運用性を考慮する。

AI エージェントとフィジカル AI の脅威とガバナンス

ガイドライン v1.2 が特に焦点を当てる AI エージェントとフィジカル AI のリスクは、従来のサイバーセキュリティの枠組みを大きく超えるものです。

AI エージェントに特有の脅威（OWASP Top 10 for Agentic Applications より） OWASP は、AI エージェントに特有のセキュリティリスクを以下のように整理しています 25 。これらは、AI が自律的にツールを連携させ（連鎖的脆弱性）、委任された権限を行使することで増幅されます 7 25 。

- **目標の乗っ取り (Agent Goal Hijack)**: 不正な指示を注入され、エージェントが意図しない目的のために行動する。

- **ツールの不正利用 (Tool Misuse)**: 正規のツール（API など）を不正に連携させ、権限を昇格させたり、データを窃取したりする。
- **ID と権限の乱用 (Identity and Privilege Abuse)**: 委任された認証情報を悪用し、不正なアクセスや操作を行う。
- **連鎖的な障害 (Cascading Failures)**: 一つのエージェントの不具合が、連携する他のエージェントやシステム全体に波及し、大規模な障害を引き起こす。
- **人間-エージェント間の信頼の悪用 (Human-Agent Trust Exploitation)**: AI が生成したもっともらしい情報によって人間を欺き、不正な承認や機密情報の提供を誘導する。

フィジカル AI に求められる安全思想 フィジカル AI のガバナンスでは、「信頼は宣言するものではなく、エンジニアリングするもの」という考え方が基本となります 5 。つまり、安全性を設計段階からシステムに組み込む必要があります。

- **多層的な防御**: AI の判断をそのまま物理的な動作に移すのではなく、デジタルツイン（物理空間の仮想モデル）での事前検証や、物理法則・安全ルールに違反しないかをチェックする決定論的な制御システムを介在させるハイブリッドアプローチが有効です 10 46 。
- **現場の裁量権の確保**: AI が異常な挙動を示した際に、現場の作業員が即座にシステムを停止・変更（オーバーライド）できる明確な権限と、使いやすいインターフェースを整備することが不可欠です 11 。
- **継続的な監視と学習**: センサーデータの相互検証や、通常とは異なる挙動を AI 自身が検知してセーフモードに移行する自己保護機能など、リアルタイムでの監視体制が求められます 22 46 。

法的拘束力とビジネスへのインパクト

日本の AI 規制は、EU の「AI 法」のような罰則を伴う「ハードロー」ではなく、事業者の自主的な取り組みを促す「ソフトロー」を基本方針としています 62 64 77 。ガイドライン自体に法的な拘束力や罰則はありません。
しかし、これは「何もしなくてよい」ことを意味しません。

- **事実上の業界標準（デファクトスタンダード）化**: ガイドラインは、企業の信頼性やリスク管理体制を評価する上での客観的な基準となりつつあります。監査法人によるシステム監査、金融機関による融資審査、保険会社によるサイバー保険の料率算定、さらには取引先選定の際に、ガイドラインへの準拠状況が問われるケースが増えています 77 。対応の遅れは、ビジネスチャンスの喪失に直結する「見えない罰則」として機能し始めています。
- **AI 推進法との連携**: 2025 年 5 月に成立した「AI 推進法」は、AI の利用促進を目的としつつも、政府による調査権限や指導・助言の根拠を定めています 62 。ガイドラインに準拠しない事業者に対して、企業名を公表するなどの措置が取られる可能性も示唆されており、実質的な影響力は増大しています 62 。
- **国際的な潮流**: ガイドラインは、G7 広島 AI プロセスや OECD の AI 原則といった国際的な議論と整合性が図られています 62 64 。グローバルに事業を展開する企業にとって、日本のガイドラインへの準拠は、EU AI 法など海外の厳格な規制に対応するための第一歩となります 41 。

結論として、企業は法的拘束力の有無にかかわらず、本ガイドラインを AI 時代の事業継続計画（BCP）の根幹と捉え、自社の AI ガバナンス体制を早急に構築・強化することが求められています。これはリスク回避だけでなく、AI を安全に活用し、新たな企業価値を創造するための不可欠な投資と言えるでしょう 85 。

1. AI Agents Act a Lot Like Malware. Here's How to Contain ...
2. Advancements in AI and Machine Learning
3. 令和 6 年版 情報通信白書｜生成 AI の急速な進化と普及
4. LLMs Add Safety Risks To Physical AI
5. フィジカル AI 時代に経営陣が直面する安全性・説明責任・信頼の課題
6. AI エージェント時代のセキュリティ設計｜脅威の 73%は検知困難
7. Agentic AI security: Risks & governance for enterprises
8. Artificial Intelligence's Use and Rapid Growth Highlight Its ...
9. AI がもたらす科学技術・イノベーションの変革
10. The Promise & Risks of Physical AI's Self−Optimizing ...
11. ガバナンスはなぜフィジカル AI にとっての 新たなインフラ ...
12. AI エージェントの損害、責任は誰がとる？ リスク恐れる日本企業
13. The rise and risks of agentic AI
14. The 2025 AI Index Report ｜ Stanford HAI
15. 生成 AI から AGI そして ASI へ、AI はどこまで進化するのか？
16. Automation is fraught with landmines; use physical AI to ...
17. フィジカル AI 導入に潜む「現場の壁」 ー Key Technology｜CTC
18. 企業向け AI エージェントが究極の内部脅威となる理由 ー ZDNET Japan
19. What are the risks and benefits of 'AI agents'?
20. The Evolution and Future of Artificial Intelligence ｜ CMU
21. 生成 AI の歴史からひもとく「急速な進化」の背景｜コラム
22. AI Has Gone Physical: Can We Still Keep It Safe?
23. フィジカル AI で問われる安全性、生成動作の責任は ー ニュースイッチ
24. AI エージェント導入におけるセキュリティとガバナンス：大企業経理 ...
25. Addressing the OWASP Top 10 Risks in Agentic AI ...
26. How fast are AI companies evolving? Check this out.
27. 2026 年に加速する「AI 革命」で現実化する 10 のこと
28. AI Risks that Could Lead to Catastrophe ｜ CAIS
29. フィジカル AI とは？従来の AI との違いや活用例、将来性を解説
30. AI エージェントと脆弱性 PART 1：AI の抱えるセキュリティリスク

31. AI Agents Are Here. So Are the Threats.
32. The Rapid Evolution of AI and the Path to the Singularity
33. AI（人工知能）の最近の進歩と将来 ～日本の科学技術研究 ...
34. AI in physical security: Opportunities, risks and responsibility
35. フィジカル AI は「知のデジタル」から行動するインテリジェンスへ
36. AI エージェントの新たな倫理リスクとは事件に取り組む研究者 ― IBM
37. Top AI Agent Security Risks and How to Mitigate Them
38. AI is changing the world faster than most realize
39. AI の波、どこまで来るの？ 最新 AI 動向！ ｜ コラム
40. Physical AI and humanoid robots
41. フィジカル AI のガバナンス｜2026 年、日本企業が直面する EU AI 法の ...
42. セキュリティ専門家の 76%が AI エージェントリスクを懸念
43. AI agents could pose a risk to humanity. We must act ...
44. The Decade of AI Super－Acceleration
45. 生成 AI の将来技術動向
46. How to Rethink Risk for Safe Physical AI Deployment
47. フィジカル AI とは？生成 AI の次に来る“動く知能”について解説
48. AI エージェントを安全に活用するための「AI ガバナンス」最新動向
49. AI Agents: Potential Risks
50. AI revolutionizing industries worldwide: A comprehensive ...
51. 令和 6 年版 情報通信白書｜AI 進展の経緯
52. What is Physical AI? Pros and Cons of AI's Integration With ...
53. AI の物理的被害リスク — Docs ― IBM Cloud Pak for Data
54. ガートナーが明かす「AI セキュリティ 6 大脅威」 なぜ AI エージェント ...
55. New Ethics Risks Courtesy of AI Agents? Researchers Are ...
56. 生成 AI が経済に与えるインパクトを読み解く―技術革新による ...
57. 「AI 事業者ガイドライン（第 1.0 版）」を取りまとめました
58. AI Guidelines for Business Ver 1.0 Compiled
59. AI 事業者ガイドライン
60. Japan: AI Business Guidelines
61. 「AI 事業者ガイドライン（第 1.0 版）」についての解説
62. AI Watch: Global regulatory tracker ― Japan
63. AI 事業者ガイドライン – 解説 ｜ デロイト トーマツ グループ

64. Japan's AI Legislation and How It Creates Competitive ...
65. 第 1 回：導入編 ｜ 「AI 事業者ガイドライン（第 1.0 版）」解説
66. AI Governance Guidelines — Global Compliance Map
67. AI ネットワーク社会推進会議｜「AI 事業者ガイドライン」掲載ページ
68. "Publication of 'AI Guidelines for Business Ver 1.0' and ' ...
69. 経産省と総務省が「AI 事業者ガイドライン（第 1.0 版）」を公表 ...
70. Japan's AI Guidelines for Business
71. 「AI 事業者ガイドライン（第 1.0 版）」の公表（総務省および経済 ...
72. Recommendation of the Council on Artificial — OECD Legal Instruments
73. AI 事業者ガイドライン — 経済産業省
74. AI 事業者ガイドライン検討会 ｜ デジタル基盤センター — IPA
75. Japan Releases Comprehensive AI Guidelines for Businesses to ...
76. AI 事業者ガイドライン検討会 （METI/経済産業省）
77. 【2026 年最新】AI 事業者ガイドライン v1.2 を解説｜企業対応 5 ステップ
78. proposal for a Regulation laying down harmonised rules on artificial ...
79. AI 事業者ガイドライン（第 1.2 版）正式版の公表（2026 年 3 ... — Threads
80. 総務省・経産省「AI 事業者ガイドライン」第 1.2 版の正式版が公表 ...
81. AI Guide for Government — IT Modernization Centers of Excellence
82. AI ネットワーク社会推進会議｜「AI 事業者ガイドライン」掲載ページ
83. 国の 2026 年版「AI 事業者ガイドライン」発表、旅行予約 AI も対象に ...
84. Framework for Artificial Intelligence Diffusion — Federal Register
85. 経済産業省 AI 事業者ガイドライン第 1.2 版のポイントを現場目線で ...
86. AI 事業者ガイドライン（METI/経済産業省）