AIエージェントの品質管理について

株式会社Controudit AI

CEO 藤井 涼

Controudit AI

Mission

ITガバナンスを、より早く、より賢く、より自由に

Vision

監査・ガバナンスを
「義務」ではなく「戦略的資産」に

AIエージェントが業務の作業者となる時代、企業の競争力を左右するのは
ガバナンスを戦略として組み込み、AIを最大限に活かすことだ。
これからのビジネスに必要なのは、**制約を超え、AIとともに成長する視点。**
私たちは、その価値観を社会に実装する。

Controudit AI

About Us

藤井涼

代表取締役 CEO

KPMGあずさ監査法人にてAI Assurance Groupに参画し、AIリスクアセスメントのサービス開発を経験。同社では四年間データサイエンティストとして監査の効率化、高度化をサポートした。過去にAIスタートアップでのキャリアも経験しており、そこでは外観検査AIモデルの開発を務める。早稲田大学基幹理工学部卒業。

徳地隆弘

技術顧問

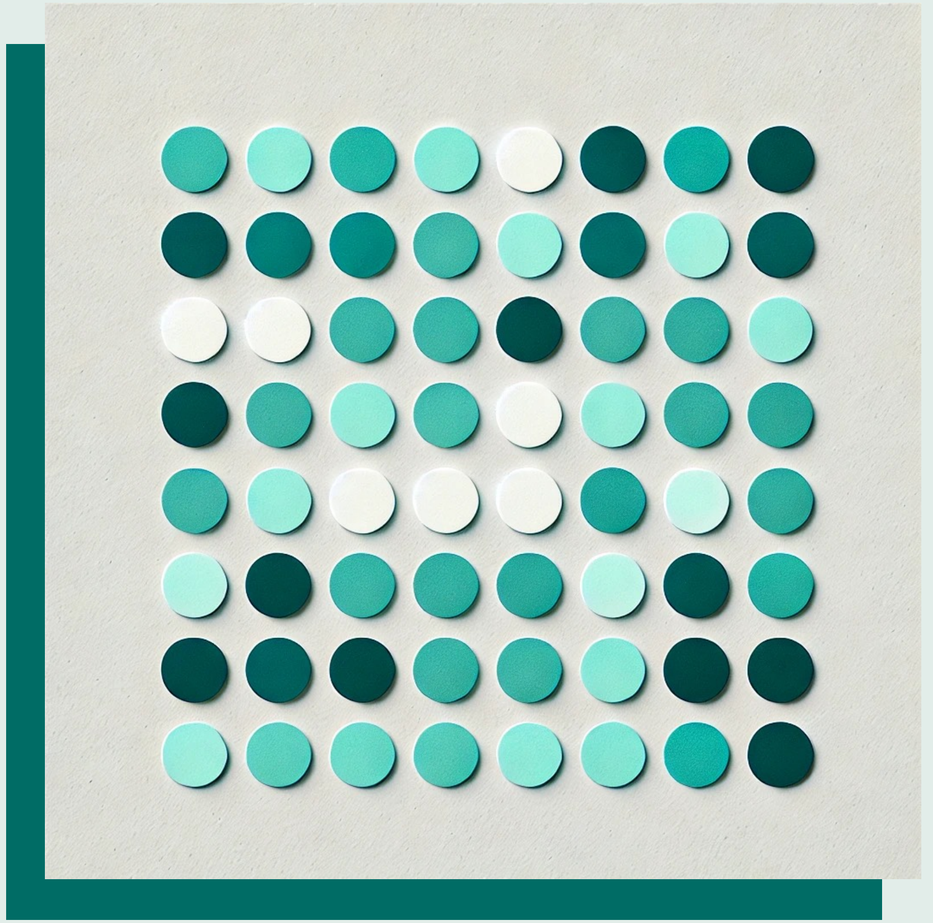
内閣官房内閣サイバーセキュリティセンター サイバーセキュリティ監査官として、国家のサイバーリスク管理とデータセキュリティの向上に貢献。株式会社野村総合研究所（NRI）ではシステム構築、クラウドセキュリティ監査、US-SOX対応、SOC2報告書作成をリード。京都大学機械工学部卒業。

Controudit AI

01

世はAIエージェント時代

AIエージェント時代の到来でガバナンス領域への
注目が加速する。



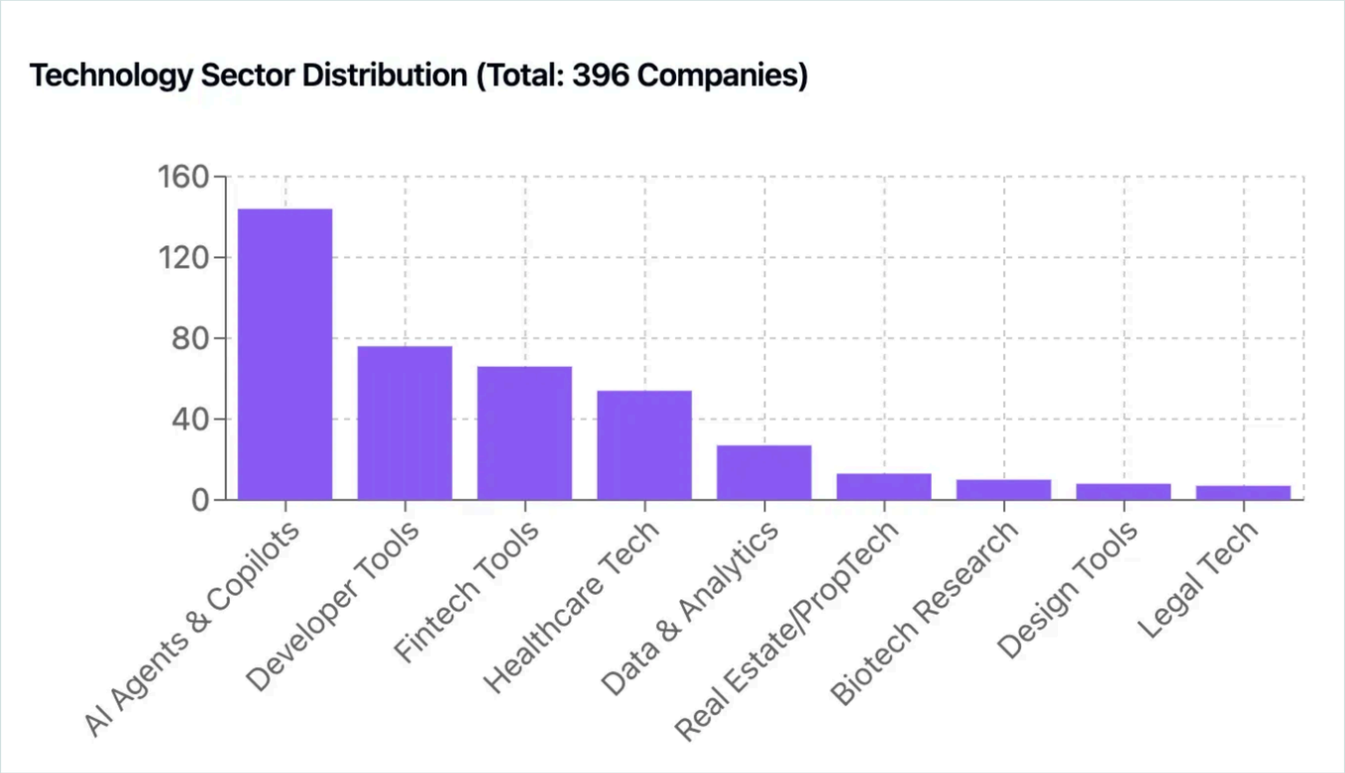
Controudit AI

世はAIエージェント時代

01 AIエージェント開発に投資が集まる

アメリカ大手VC Y-Combinatorの投資先（2023年下期から2024年末）において、3社に1社がAIエージェント関連のプロダクトです。

35%

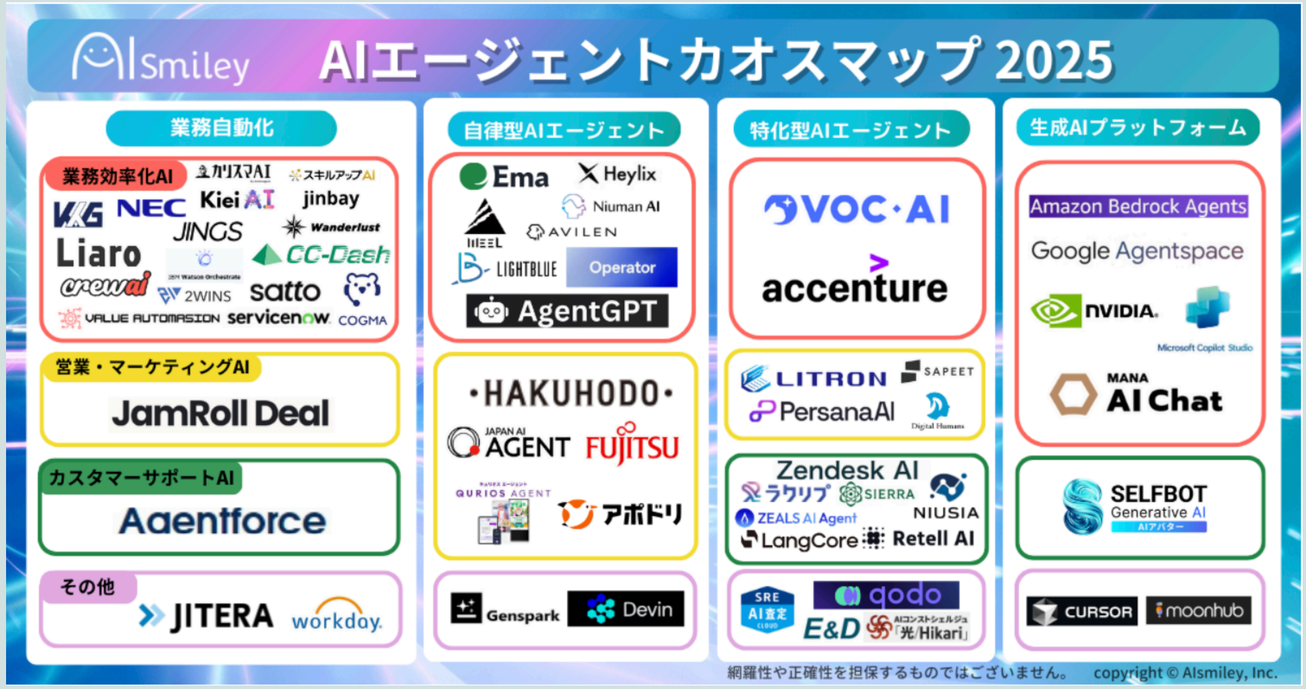


https://dswharshit.medium.com/inside-ycs-latest-400-startups-decoding-the-ideas-that-get-funded-60088e6baebd

Controudit AI

世はAIエージェント時代

02 国内でもAIエージェント開発が盛り上がっている



https://aismiley.co.jp/ai_news/ai-agent-caosmap-2025/

> 2BはバーティカルAIエージェントにフォーカス、2Cは究極ののめり込みを

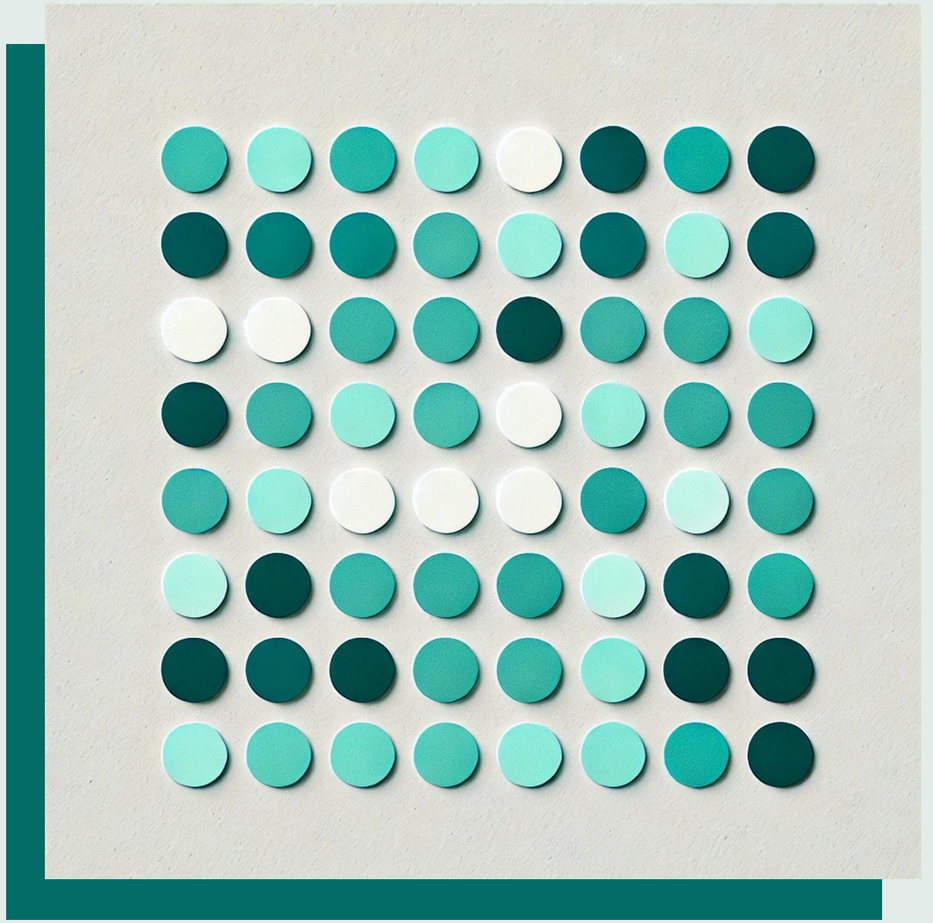
https://fullswing.dena.com/archives/100153/

Controudit AI

世はまさにAIエージェント時代

02

AIエージェントには
ガバナンスが求められる



自律性がインパクトを生むが、それに比例してリスクも拡大

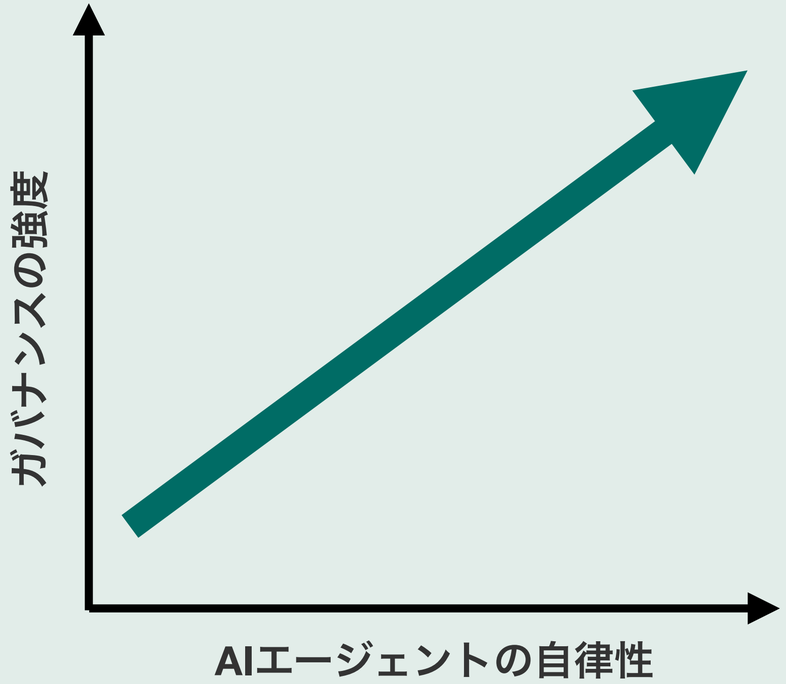
AIエージェントの定義

「エージェントとは、環境を認識し、目標を達成するために**自律的**に行動する存在」として定義されている。

イノベーションのジレンマ

自律性の幅が大きいほど、業務の効率化や高度化が期待できる。
しかし、それに比例してリスクも増大するため、ガバナンスが必要となる。

インパクトがあるエージェントほど
強いガバナンスが必要になってしまい面倒だな、、、



Controudit AI

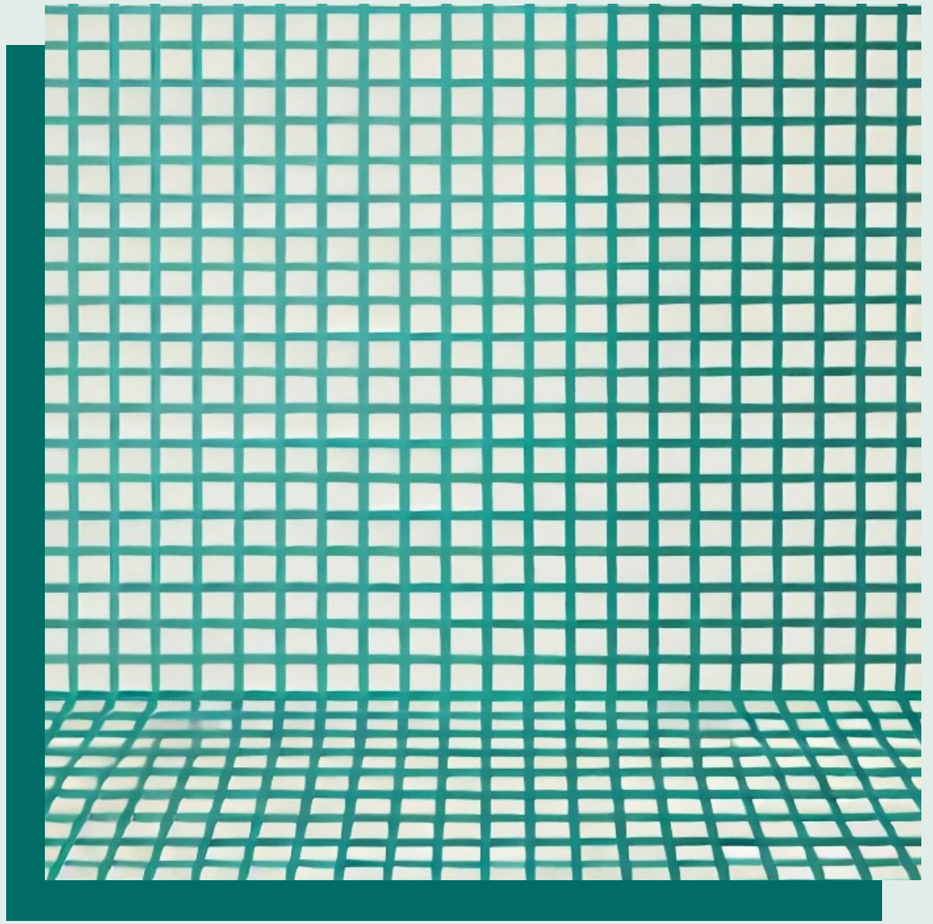
AIエージェントとガバナンスは隣り合わせ

その自律性という特性がリスクを生むため、ガバナンスが必要となる。

03

AIエージェント特有のリスクと対策

AIエージェントに関する一般的なリスクと対策をまとめました。



Controudit AI

AIエージェント特有のリスクと対策

コントロールの欠如

- AIエージェントの自律行動範囲が不明。
- AIエージェント中心の意思決定となる。
- エージェント間の連携状況が不透明。
- 労働者のスキル向上機会を逸する。

正確性の欠如

- 正確性を検証する方法が確立されていない。
- マルチエージェントシステムの検証が網羅的でない。

セキュリティリスク

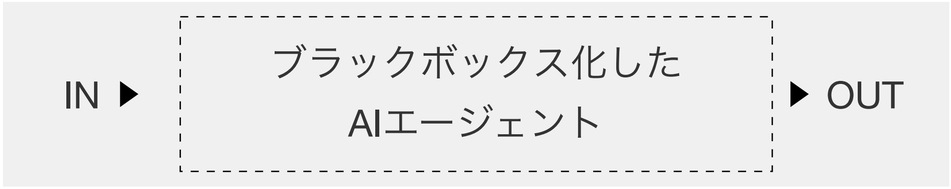
- エージェントに付与された権限を悪用されるリスクがある。
- ログを改竄されてしまう。（サイレントモードの悪用）

Controudit AI

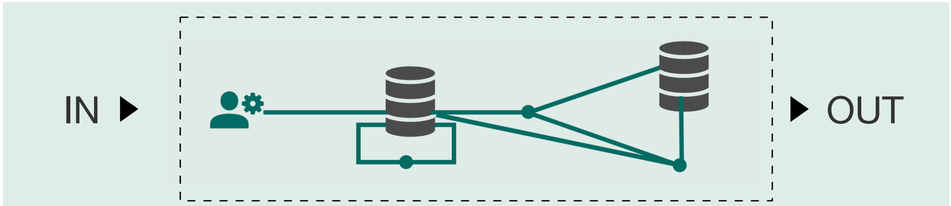
AIエージェント特有のリスクと対策①

コントロールの欠如

- 内部でどんな処理をしているのかわからない、、、、
- 思いもよらぬ行動をしていた、、、



構造や、意思決定の根拠がわかって安心！



リスクの例

- AIエージェントの自律行動範囲が不透明。
- AIエージェントの判断根拠が不透明。
- AIエージェント間の連携状況が不透明。

||

AIエージェント中心の意思決定。

対策

- グラフ構造を可視化する。
- エージェント単位で判断根拠を可視化する。
- モニタリングシステムを導入する。

||

人間中心の意思決定。

Controudit AI

AIエージェント特有のリスクと対策②

正確性の欠如

AIエージェント全体としての精度検証しかしていない。

複数のエージェントを組み合わせたAIエージェント※1

精度95%

構造や、意思決定の根拠がわかって安心！

複数のエージェントを組み合わせたAIエージェント

精度95%	精度70%	精度60%
タスク生成 エージェント	優先順位づけ エージェント	タスク実行 エージェント

※1 BabyAGIの例であり、AIエージェントはここで記したような棲み分けになるとは限りません。

リスクの例

- 検証の網羅性欠如。
- 長時間動作する過程で目的を喪失するリスク。
- タスクが無限ループするリスク。

対策

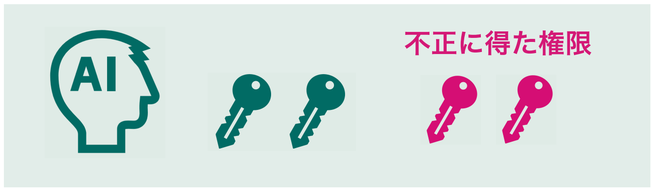
- 構成するエージェントごとにテストを実施する。
- 明示的なコンテキストの確認処理を導入する。
- 終了条件を明示。実行タスク履歴のトラッキング等。

AIエージェント特有のリスクと対策③

セキュリティリスク

AIエージェントごとに割り振る権限管理が疎か
さらに、その攻撃のログが消されている。

必要以上の権限が付与されたエージェント



DBへの不正アクセス

APIの不正実行

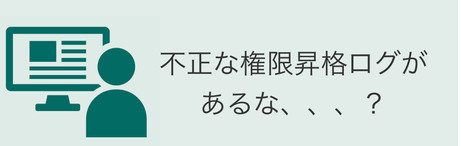
リスクの例

- エージェントに付与された権限を悪用される。
- ログを改竄されてしまう。（サイレントモードの悪用）

AIエージェントごとの権限管理が厳重に管理されている。
もし、攻撃されても改竄防止ログの導入により追跡可能。

権限管理機能を強化

改竄防止ログ



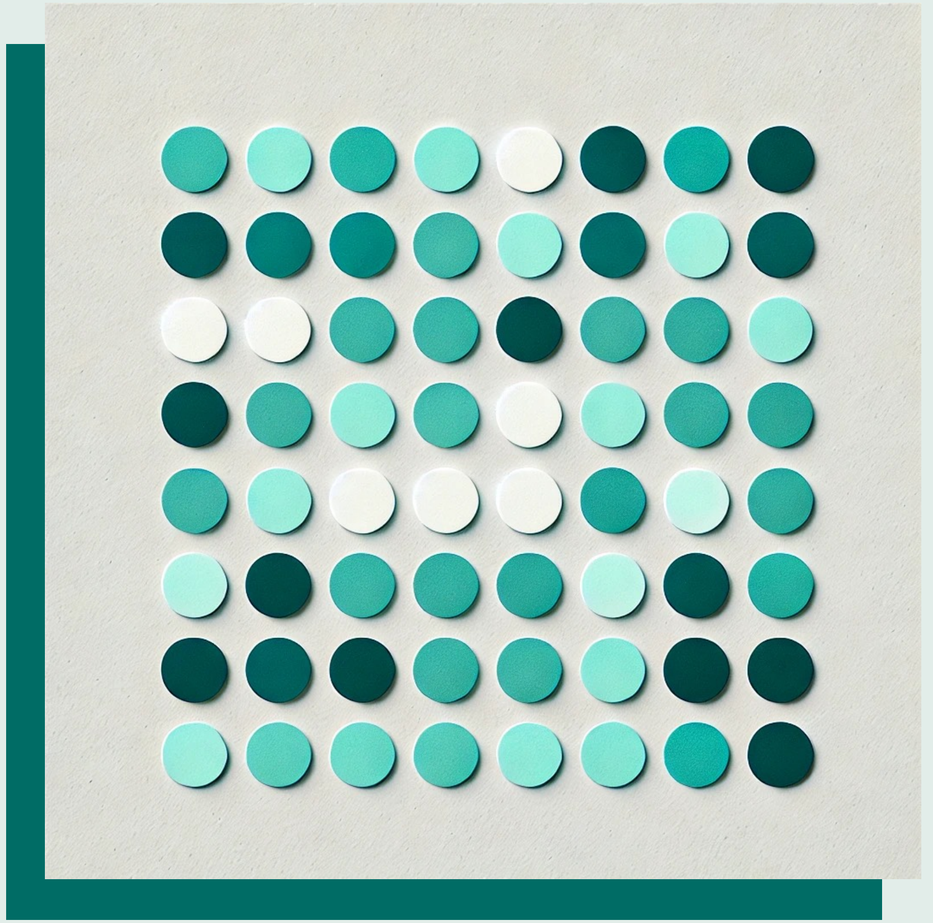
対策

- エージェントごとのロール制限する。
- 権限の昇格ログを監視する。
- 改竄防止ログを導入する。

Controudit AI

04

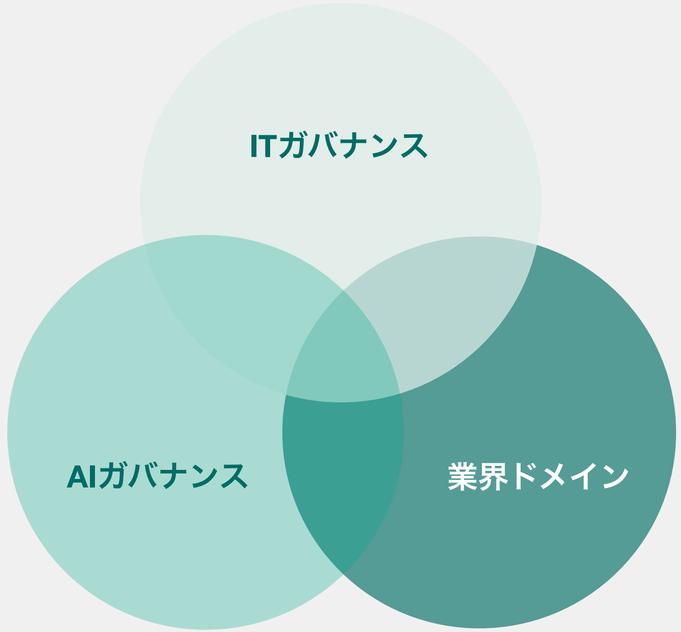
考察



考察①

ガバナンスには横断的な知見が求められる

AIエージェント・LLMだけでなく、広義な深層学習への知見や、**従来のセキュリティへの知見**、および、**業界業種のドメイン知識**がリスク発見・対処に求められる。



中心に向かうほど良いAIエージェントガバナンスとなる。

AIエージェントガバナンスの構成要素

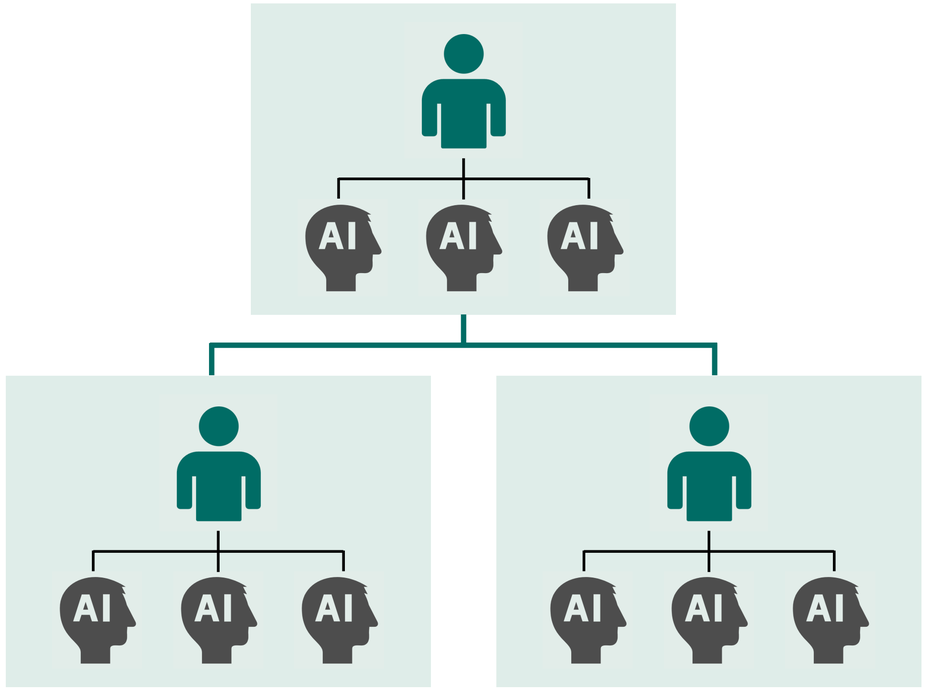
AIガバナンス	AI特有のリスク管理に加え、 AIエージェント特有のリスク管理。
ITガバナンス	従来のIT統制に関連する全般的な知見。 特にITセキュリティに関する知見。
業界ドメイン	表層的な業務フローの理解ではなく、 業界ならではの考え方の癖や理屈。

Controudit AI

考察②

人々が皆、マネージャーになる時代の到来

AIエージェントの進化により、人は「作業者」から「マネージャー」へと役割が変わる。AIに指示を出し、成果を管理する力が求められる時代が来る。



これからの人々に求められる能力

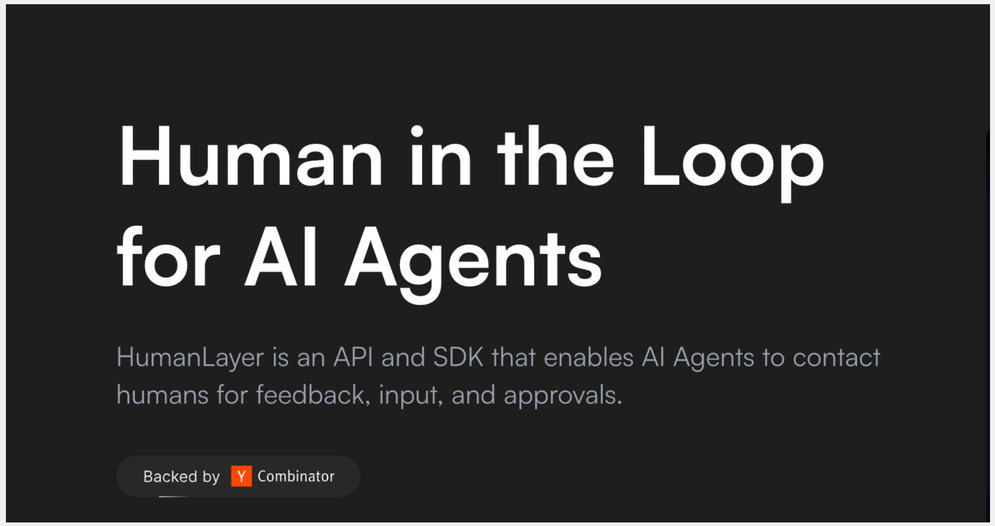
適切に指示し管理する能力	• プロンプト設計能力 • タスク分配力：AIと人間の役割を決め、効率的に業務を進める力。
出力内容を評価し意思決定する能力	AIの結果をそのまま受け取るのではなく、正確性や信頼性を判断し、人間が責任を持って意思決定する力。
リスク管理能力	ルールベースな対応ではなく、個人が目の前のAI活用により生じるリスクを洗い出せる力。 →アジャイルガバナンスに繋がる。

考察③

イノベーションのジレンマと上手に付き合う

人間が承認フローを設けることでコントロールは強まりますが、介入しすぎると本来の利点が損なわれます。許容できるリスクを見極め、最適なバランスを保ちましょう。

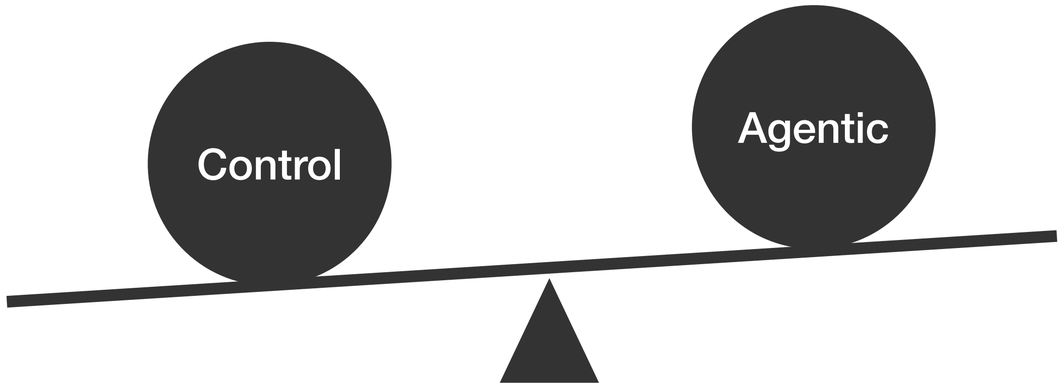
HumanLayer



https://www.humanlayer.dev/

Human-in-the-loop

AIシステムの意思決定や学習プロセスの一部に「人間Human」が介入し、監督・調整・フィードバックを行う仕組みのことを指す。特に、AIの完全自動化がリスクを伴う場面で、人間がAIを補完し、品質や安全性を確保するために活用される。



Controudit AI

考察③

イノベーションのジレンマと上手に付き合う

人間が承認フローを設けることでコントロールは強まりますが、介入しすぎると本来の利点が損なわれます。許容できるリスクを
見極め、最適なバランスを保ちましょう

どうすればいいのか？

HumanLayer

AIシステムの意思決定や学習プロセスの一部に「人間Human」が介入し、
監督・調整・フィードバックを行う仕組みのことを指す。特に、AIの完全
確保す

許容できるリスクを定義し、把握することが重要！

for AI Agents

HumanLayer is an API and SDK that enables AI Agents to contact humans for feedback, input, and approvals.

Backed by Y Combinator

https://www.humanlayer.dev/

Controudit AI

考察③

イノベーションのジレンマと上手に付き合う

どうすればいいのか？

▼

許容できるリスクを定義し、把握することが重要！

▼

ガバナンスが重要

Controudit AI

AIエージェントに対するガバナンスの考察 - まとめ

ガバナンスには横断的な知見が求められる

AIエージェント・LLMだけでなく、広義な深層学習への知見や、**従来のセキュリティへの知見**、および、**業界業種のドメイン知識**がリスク発見・対処に求められる。

人々が皆、マネージャーになる時代の到来

AIエージェントの進化により、人は「作業者」から「マネージャー」へと役割が変わる。AIに指示を出し、成果を管理する力が求められる時代が来る。

イノベーションのジレンマと上手に付き合う

自律性が高まるほど業務の効率化や高度化が進む一方で、リスクも増大するため、適切なガバナンスが必要。

人間が承認フローを設けることでコントロールは強まりますが、介入しすぎると本来の利点が損なわれます。

許容できるリスクを見極め、最適なバランスを保ちましょう。

Controudit AI

AIエージェントの未来は、ガバナンスで決まる。

私たちの使命は、安全かつ信頼できるAI活用の基盤を整備し、その可能性を最大限に引き出すことにあります。

適切なガバナンスのもと、AIエージェントの社会実装を促進し、持続可能な発展を支えていきましょう。

Controudit AI

その他のAIガバナンスのお役立ち情報

AIガバナンスブログ

登壇者のSNS

Controudit AI

参考文献

OWASP Top 10 for Agentic AI (AI Agent Security) - Pre-release version

https://github.com/precize/OWASP-Agentic-AI/tree/main

“Practices for Governing Agentic AI Systems”

https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

AGENT DESIGN PATTERN CATALOGUE: A COLLECTION OF ARCHITECTURAL PATTERNS FOR FOUNDATION MODEL BASED AGENTS

https://arxiv.org/pdf/2405.10467

langchainとlanggraphによるRAG・AIエージェント 「実践」 入門

https://amzn.asia/d/6cNa4pQ

Controudit AI

Controudit AI