AI事業者ガイドライン 第1.2版：AIエージェント・フィジカルAI時代の新基準

進化するAI：2つの新カテゴリーと潜在リスク

AIエージェント：自律する「デジタルの同僚」

内部特権を悪用したデータの漏洩や、予測不能な連鎖的障害のリスクを内包。

検知困難な「新たな攻撃ベクトル」

AIエージェント特有の脅威の**73%**は、従来の対策では検知が困難とされる。

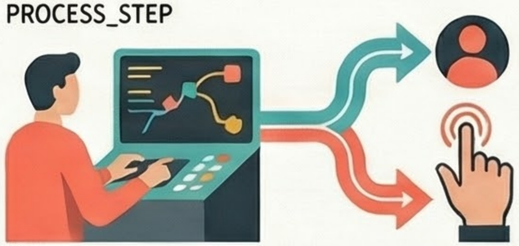
フィジカルAI：物理世界へ進出する知能

サイバー攻撃が人身傷害や設備破壊といった物理的損害に直結する。

AI種別	物理的実体	主なリスク要因		傷害	設備破壊	責任所在の曖昧さ
AIエージェント	なし（ソフトウェア）	権限の不正利用、目標の乗っ取り、連鎖的障害	権限の不正利用、目標の乗っ取り、連鎖的障害			
フィジカルAI	あり（ロボット・車両等）	人身傷害、設備破壊、責任所在の曖昧さ	人身傷害、設備破壊、責任所在の曖昧さ			

企業の必須対応とビジネスインパクト

PROCESS_STEP



Human-in-the-Loop（人間の判断介在）

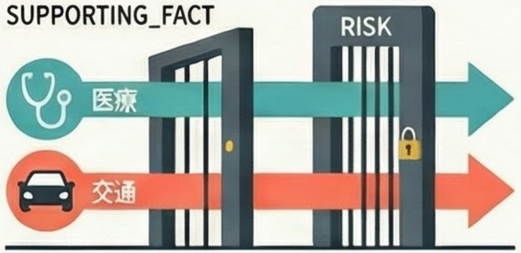
人間がAIを監視・修正できる仕組みの構築が事実上の必須要件となった。

KEY_FINDING

「ソフトロー」による事実上の義務化

ガイドラインへの不適合は、取引先審査や保険契約での不利を招く。

SUPPORTING_FACT



リスクベース・アプローチの具体化

利用目的や影響度に応じ、医療や交通など高リスク分野には厳格な統治を要求。