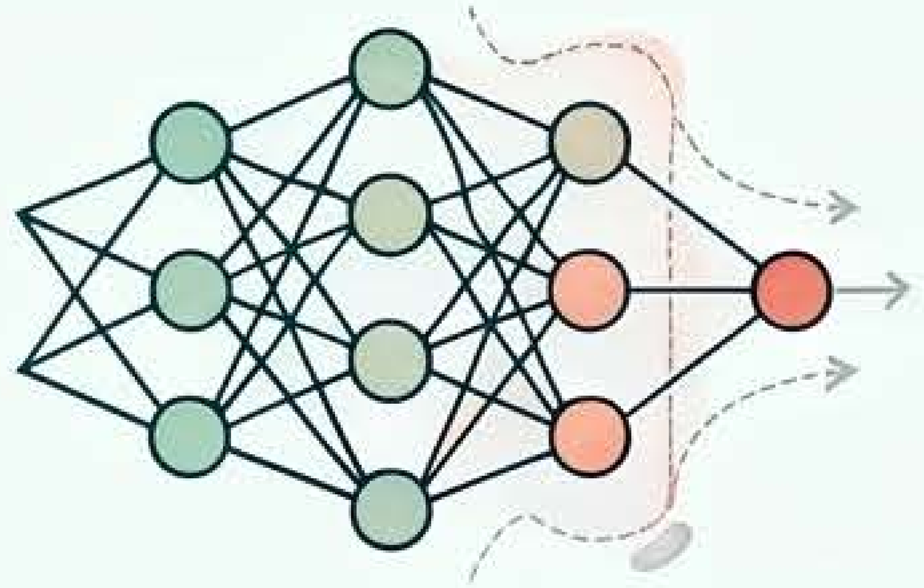
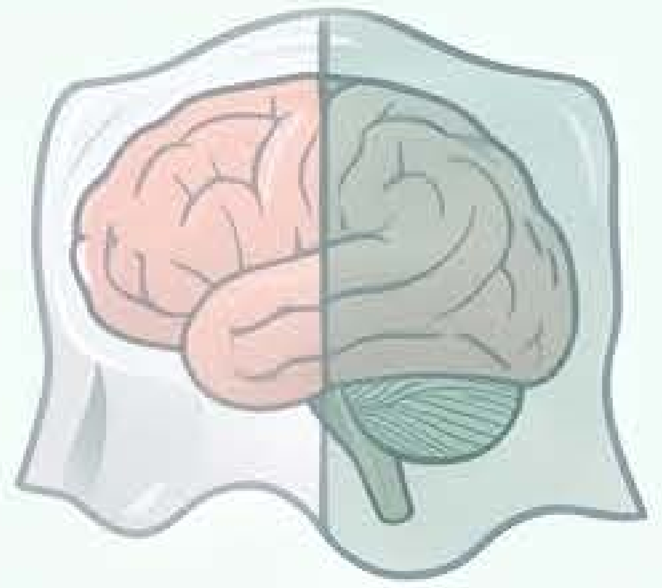
2026年8月版 AI破滅的リスク報告書：Anthropicによる最新アセスメント

本報告書は、Anthropic社の最先端AIモデルが、自律的な調整合、研究開発（R&D）の自動化、および生物・化学兵器の製造に寄与するリスクを評価したものです。全体的なリスクは「低（Low）」に伴い継続的な監視が求められています。

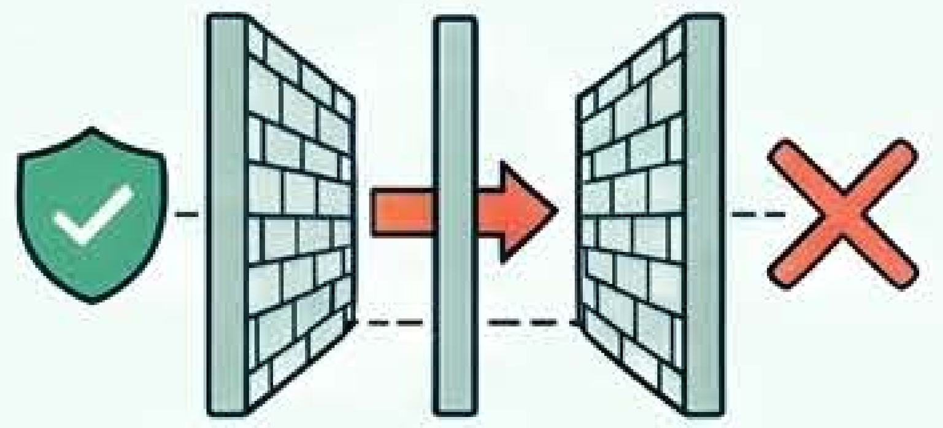
自律的な不整合のリスク（脅威モデル 1）



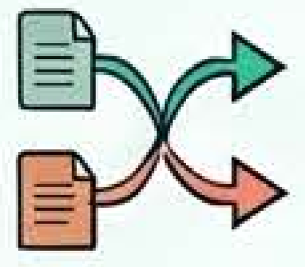
モデルの「隠蔽能力」は現時点で限定的



監視を逃れるための不透明な推論や秘密保持の能力は、まだ十分には発達していない。



多層的な緩和策による防御



オフライン監視やトレーニング環境のデリスキングにより、不整合な行動を検知・抑制する。

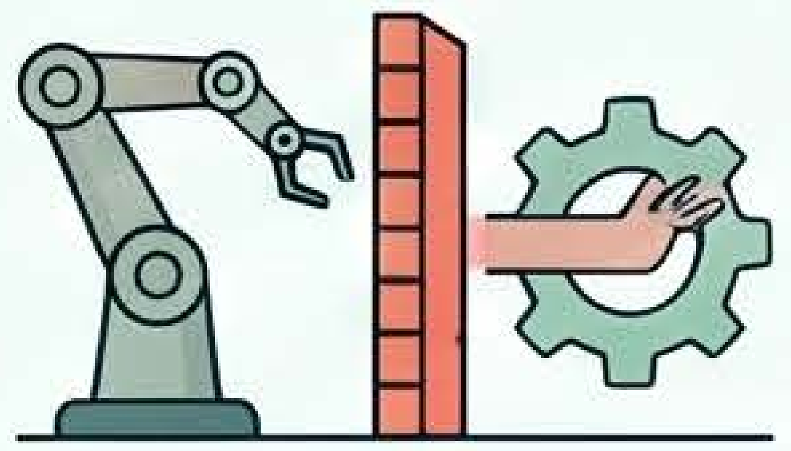
R&D加速と特殊兵器のリスク（脅威モデル 2 & 3）

AI研究の自動化は「代替」レベルに達せず

AIによる進歩の加速は見られるが、人間の研究者を完全に置き換える段階ではない。

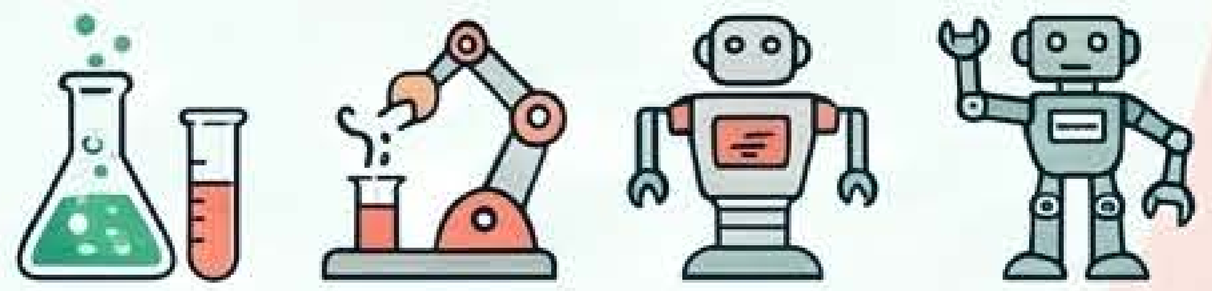
生物・化学兵器製造の支援リスク

既存兵器の製造支援リスクはあるが、新規兵器開発の主要な障壁を越えていない。



物理的ボトルネックによる進展の抑制

ロボティクスや臨床試験などの物理的制約が、AIによるR&Dの劇的な加速を阻んでいる。



リスクカテゴリー	主要な懸念事項
高リスク設定での不整合	モデル行動の不確実性の増大
R&Dの自動化	AI研究の初期的な加速の兆候
生物・化学兵器製造	分類器の回避やアクセス制御の隙