米国軍事AI利用を巡る激震：Anthropicの排除とOpenAIの戦略的合意

2026年2月、AIの軍事利用における安全制限を巡り、国防総省とAnthropicが対立。Anthropicが倫理的制限の撤廃を拒否したことで連邦政府から排除された一方、OpenAIは技術的な「セーフティスタック」を提案することで軍との合意を取り付けた、AIガバナンスの転換点となる事例。

Anthropicの「拒否」と前例のない制裁

2億ドルの契約を失った。

米国主要企業として初。国防総省と取引のある全企業との商取引が禁止される致命的打撃。

良心に基づく拒否 (In Good Conscience)

アモデイCEOは、現在の技術能力を超えた事事利用は安全ではないと主張した。

CEO Dario Amodei

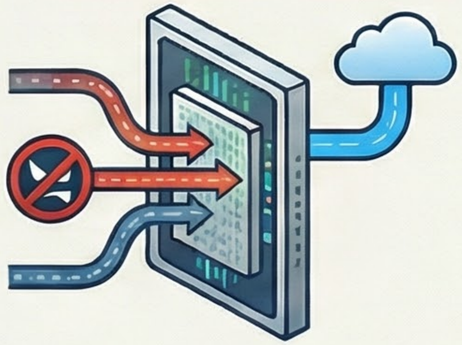
倫理的「レッドライン」の堅持

$2億ドルの契約
契約条項として譲らず

国内監視と完全自律型兵器への利用禁止

「サプライチェーンリスク」への指定

OpenAIの「戦略的合意」と実装の違い



政治的・ビジネス的勝利

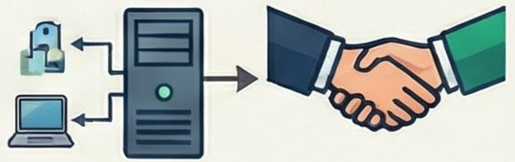
Anthropicが排除された当日に契約を発表。原則を共有しつつ、軍の要求にも応える巧みな交渉。

技術的制御（セーフティスタック）の導入

モデル側で不適切利用をブロック

「あらゆる合法目的」の条項を受け入れつつ、モデル側で不適切利用をブロックする仕組みを提案。

政治的・ビジネス的勝利



クラウド限定の展開

兵器エッジデバイスへの直接搭載は避け、制御可能なクラウド環境での利用に限定した。

Anthropic（排除）		OpenAI（合意）
拒否	「あらゆる合法目的」条項	受入れ
契約条項（法的拘束力）	安全制限の実装方法	セーフティスタック（技術的制御）
政府出禁・法的紛争へ	交渉の結果	機密ネットワークでの展開開始