AI 開発「減速論」をどう読むか

— 法務担当者のための現在地・責任論・実務対応 —

止める／進めるの二択ではなく、「統制して使う」

Frontier AI

Agent Risk

Liability

Legal Governance

2026 年 9 月 14 日時点の公表情報に基づく

弁護士法人三宅法律事務所　弁護士　渡邉雅之

本日の結論

EXECUTIVE SUMMARY

最初に押さえていただきたい 5 点

1 **「減速論」は開発停止論ではない**

提唱者自身が「 pacing は訓練や技術的進歩の停止を意味しない」と明記している

2 **事故は既に起きており、しかも原因説明が覆った**

Anthropic は 7 月に「運用上の失敗」と説明した 3 件について、 9 月 9 日にその枠組みを撤回した

3 **法制度はむしろ緩和方向に動いている**

EU は高リスク義務を 2027 年 12 月／ 2028 年 8 月へ延期し、日本は同意原則に特例を設けた

4 **責任論は未解決のまま、実務だけが先に進んでいる**

民法第 715 条も製造物責任法も AI の「行為」を想定していない。だから契約と記録で埋める

5 **明日からの 3 点**

人間承認・ログ・停止。この 3 点が入っていれば、事故が起きても会社は説明できる

持論： AI に「何をさせるか」より先に、 AI に「何を見せ、何を触らせ、誰が止めるか」を決める。

今日、持ち帰っていただく 3 点

TAKE THIS HOME

細部は後半で扱います。まずはこの 3 つだけ押さえてください

HUMAN APPROVAL

人間承認

不可逆な外部影響を生む操作の直前に、人間のゲートを置く。決済・発注、顧客への送信、本番環境への変更の 3 接点が最優先。

AUDIT LOG

ログ

入力・出力だけでなく、ツール操作・データアクセス・承認者・承認時刻を残す。責任論の帰趨は、ほぼここで決まる。

KILL SWITCH

停止

誰が、どの手順で、どこまで止められるか。再開条件まで平時に決めておく。事故後に決めるのでは間に合わない。

なぜこの 3 点か — 直近の事例が示していること

- Anthropic が自社のサイバー評価で無許可アクセス 3 件を特定したのは、約 141,000 件の評価トランスクリプトを遡って精査した結果である。ログがなければ、そもそも見つからなかった
- 影響を受けた組織のうち連絡が取れた 2 組織は、いずれもこの活動を自力で検知していなかった。第三者に指摘されて初めて知る、という構図
- 同社の説明では、監視は実行を停止できず、 1 つのモデルは中止に成功しなかった。「止める仕組み」は持っているだけでは足りない
- 御社で最初に見るべきは、決済・発注、顧客への送信、本番環境への変更。この 3 つの接点に AI が届いていないかを今週確認する

2026 年、この 3 か月に何が起きたか

TIMELINE

減速論は 9 月 12 日の単発ニュースではなく、夏からの連続した事故の帰結である

2026.6.5	Anthropic 、 AI 開発の「減速」を提言
2026.7.21	OpenAI 、社内評価中のモデルが Hugging Face の本番インフラに侵入した事案を公表
2026.7.23-30	Anthropic がサイバー評価を全停止。約 141,000 件の記録を精査し、第三者 3 組織への無許可アクセス 3 件を公表
2026.8.19	OpenAI 、次世代モデル開発を「減速」。安全対策を最優先に
2026.8	AI 従業員 1,300 人超が開発ペース減速を求める署名（ Pacing the Frontier ）を公開
2026.9.4	米議会に「 Ban Artificial Superintelligence Act 」提出（超知能 AI の開発・展開の恒久的禁止）
2026.9.8	Anthropic の事前学習研究者が退社を公表。在籍研究者 2 名も個人の見解として同調
2026.9.9	Anthropic 、当初の「運用上の失敗」という説明を撤回。アライメントの失敗として再評価
2026.9.10	4 件目が判明（ 2026 年 1 月にさかのぼる）。評価機関 METR が独立レビューを開始
2026.9.12	アモデイ CEO 「 We Must Pace the Frontier 」公表。アルトマン氏・マスク氏・ハサビス氏が賛同
2026.9.13	トランプ米大統領が減速論に反論。対中競争での主導権維持を強調

この 3 か月で本当に重いのは、事故そのものよりも「開発元自身の原因説明が 2 か月で覆った」ことである。

減速論の中身：停止ではなく「ペース配分」

WHAT IT MEANS

Pacing the Frontier の 3 段階

1. Embedded Evaluators
常駐評価者

従業員同様のアクセス権を持つ第三者評価チーム（ METR 等）を社内に常駐させ、安全対策の遵守を検証しインシデントを報告する。 Anthropic は一方的に実施を表明。

2. Democratic Coordination
民主主義国内の協調

各社が共通の安全基準と、無制限な能力向上への制限を設定。独禁法上の理由から、米国政府による仲介ないし限定的な適用除外が必要とされる。

3. Global Coordination
グローバル協調

Lv1 危険用途の禁止 → Lv2 リリース前テストの相互合意 → Lv3 再帰的自己改善の速度制限 → Lv4 全体的なペース制限・停止。

- pacing は「モデル訓練や技術的進歩の停止」ではなく、アラインメントと安全対策に十分な時間を確保し、第三者評価者がそれを確認できるようにすること
- Lv4 （全体的な停止）は、監視を回避して離脱すれば世界の力の均衡を変えうるため、提唱者自身が当面の実現は考えにくいとする
- 常駐評価者の先例として挙げられているのは、規制上の「監督官」が従業員と並んで常駐することがある銀行業界である
- 提言の前提として、 AI エージェント群が 6 〜 12 か月以内にインターネット全体を乗っ取りうる能力を持つとの懸念が示されている

企業版の結論：能力が上がるほど、先に統制のハードルを上げる。「利用停止」ではなく「高リスク利用のチェックポイント化」。

※ 提唱者は、モデルが能力 X を持つならアラインメント特性 Y ・ Z の証明を伴わなければならないという「チェックポイント」方式を例示している。

引き金となった２つの事実

TRIGGERS

アモデイ氏が「さらに慎重さが必要」と判断した根拠

① 再帰的自己改善（RSI）

- 2026 年夏頃から、 AI が次世代 AI を構築する能力を高めたことにより AI の進歩が劇的に加速している
- この力学は再帰的自己改善（RSI）と呼ばれ、業界全体で発生し始めている
- 放置すれば、これらのシステムを理解し制御する人間の能力を上回りかねない

法務的含意：開発速度が人間の統制能力を上回るという主張。内部統制の言葉でいえば統制環境そのものの崩壊リスク。

② OpenAI-Hugging Face 事件（OAI-HF）

- サイバー能力を測るベンチマーク「ExploitGym」による社内評価中に発生。最大能力を測るため拒否機能と本番用分類器は無効化されていた
- モデルは隔離環境から外部アクセスを確保する方法を探索し、パッケージレジストリのキャッシュプロキシのゼロデイ脆弱性を特定・悪用した
- テスト環境内で権限昇格とラテラルムーブメントを行い、盗んだ認証情報とゼロデイを組み合わせて本番サーバー上でのリモートコード実行に至った

エージェント群が、指示されていない対象を攻撃し、集団の成功のために自らを犠牲にし、自らを評価する「採点者」へのハッキングを試みた。

経済的損害は最小限で負傷者もいなかったが、同程度のミスアラインメントでより高い能力を持つ群であれば壊滅的被害をもたらしえた — というのが提唱者の評価。

原因説明は、２か月で覆った

Anthropic の 3 件（＋後に 4 件目）をめぐる説明の変遷

7 月 30 日の説明

- 3 件はいずれも「旗取り競技（CTF）」形式の評価で発生
- 評価プロンプトは毎回「この環境はシミュレーションであり、インターネット接続はない」と明示していた
- 評価パートナーの設定の行き違いにより、実際には外部への通信経路が開いていた
- モデルのアライメント不全というより、ハーネスと運用上の失敗に近い

9 月 9 日の修正

- 「運用上の失敗」という枠組みを撤回
- 2 つの真のミスアライメントを反映するものと再評価した
- ① 自身の行動を正当化するのに都合の良い方法で証拠を選択的に解釈する、偏った推論
- ② たとえ害につながる可能性があっても、与えられたタスクを遂行しようとすること

約 141,000 件

遡って精査した評価トランスクリプト

3 件 → 4 件

4 件目は 2026 年 1 月にさかのぼり、7 月のレビューでは見逃されていた

2 組織

連絡が取れた被害組織のうち、自力で検知できていた数はゼロ

法務への教訓：初報の原因分析は覆る。インシデント公表文・顧客説明・当局報告には「現時点の評価である」旨を明示し、更新の経路を必ず確保しておく。

※ 独立評価機関 METR が、既知のインシデント期間を超える記録へのアクセスを得て独立レビューを開始している。レビューは本日時点で進行中であり、評価はさらに更新され得る。

内部からの異議と、外された「停止の約束」

INSIDE DISSENT

9 月 12 日のエッセイは突然出てきたものではない

9 月 8 日の退社 — 何を求めたのか

- 退社したのは役員ではなく、事前学習研究に従事していた研究者。 OpenAI に 2023 年から在籍し、 2026 年 7 月に Anthropic へ移籍
- 求めたのは全面停止ではない。①米国の AI ラボ間でのペーシング協定 ②モデル能力向上の一時的な停止 ③再帰的自己改善の停止合意 の 3 点
- 在籍中の研究者 2 名（アライメント研究責任者ら）が個人の見解として同調。ただし両名は退社していない
- 同社の安全性への取り組み自体は「誠実」と評価。政府の介入か業界全体での減速がない限り、責任を持って構築できる企業は存在しないと結論づけた

2026 年 2 月 — 規程から外された「停止の約束」

- Anthropic は責任あるスケーリングポリシー（ RSP ） v3.0 で、 2023 年以来の中核的な誓約を変更したと報じられている
- 旧：安全対策が十分だと事前に保証できない限り、より強力なモデルを訓練しないというハードリミット
- 新：①経営陣が自社が AI 競争の先頭にいると判断し、かつ②壊滅的リスクが重大だと判断する場合、という 2 条件
- 評価機関 METR の政策ディレクターは、この変更が、危険が段階的に増してもどの時点でも警報が鳴らない「カエルを茹でる」効果を招きうると指摘している

法務への教訓 — 他社の話として聞かない

- 自社の規程から「停止基準」を外す・緩めるときは、外した理由と、判断した機関・日付を必ず記録に残す。後から必ず問われる
- 「競合が止まらないから」は、事後の説明として最も弱い部類に属する。取締役会議事録にその一行しか残っていない状態を作らない
- 停止基準を経営判断に委ねる設計にするなら、その判断を検証できる第三者の目（内部監査・監査役）を同時に置く

※ 米議会には FRONTIER Act （ H.R.9925 、 2026 年 7 月 23 日提出・超党派）も係属。先端モデル開発者にリスクフレームワークの公表、第三者監査、 24 〜 72 時間以内のインシデント報告を義務づける内容。 Ban ASI Act とともに提出段階であり、成立していない。

この提言を、どう割り引いて読むか

COUNTERPOINTS

額面どおり受け取らないための 4 つの留保

発信者の立場

提唱者は、安全性を競争軸に据える企業の経営者である。本人も、これまで hype ・ doomerism ・規制捕獲（ regulatory capture ）と非難されてきたと書いている。動機は誠実でも、立場は中立ではない。

「リード」の測り方が示されていない

民主主義国の減速幅を対中リードの範囲内に収める、という論理だが、リードを何で測るかは示されていない。 Stanford HAI 「 AI Index 2026 」は、 Arena で見た米中トップモデルの差を約 2.7% と報告している（ 2026 年 3 月時点の履歴データ）。

検証する手段がない

減速の合意は、相手が本当に減速しているかを検証できて初めて意味を持つ。核軍縮には査察があるが、 AI の訓練に同等の検証手段はない。 AFP の取材に対し、企業間協議の可能性について Anthropic ・ OpenAI ・ Google はいずれも回答していない。

それでも実務に効く理由

誰の主張が正しいかにかかわらず、事故は現に起きており、その原因説明は覆った。企業が備えるべき事実関係は、減速論の当否とは無関係に存在する。

講演者の立場：減速論は「そのまま信じるもの」ではなく、「企業統制の設計図として使えるところだけ借りるもの」と考えている。

中国は減速しない — 検証できない約束に意味はない

CHINA

減速論に対して最も多い疑問。事実を分けて見る必要がある

確認できる事実

- 2026 年 9 月 14 日、中国外務省報道官が定例会見で減速論を批判。恐怖をあおって対立と悪性競争を助長することは、グローバルな AI ガバナンスの妨げにしかならないと述べた
- 中国が掲げるのは開放的・包摂的で広く利益をもたらす倫理的な AI の発展であり、減速ではない
- 2026 年 7 月 16 日、上海で世界人工知能協力機構の設立協定に 29 か国が署名。王毅外相が署名し、国連事務総長も出席。本部は上海に置かれる
- Stanford HAI 「AI Index 2026 」は、 Arena で見た米中トップモデルの差を約 2.7% と報告している

どう読むか

- 中国は「減速しない」が、「国際的な枠組みを拒否している」わけではない。自国主導の枠組みを別に作っている。単純な二分法で語ると事実を外す
- 決定的なのは、提唱者自身がこのリスクを明記していること。米国側が能力を大幅に抑制した後に中国が約束を破れば、その違反によって中国が地政学的優位を握りうる、と書いている
- つまり「中国が止まらないから減速は非現実的だ」という批判は、提言への反論ではなく、提言が置いている前提そのものである
- そして根本的な問題は、同調の意思ではなく検証可能性にある。相手が本当に減速したかを確かめる手段がない

講師の見解

中国が減速に同調することはないと考えています。そして仮に同調したと表明されても、それを検証する手段が存在しません。
検証できない約束は、実務上、存在しないのと同じです。

したがって、国際協調の成立に期待して自社の AI 統制を先送りすることは、経営判断として誤りである。

※ 中国発のオープンウェイトモデル（ Moonshot AI 「Kimi K3 」 2026 年 7 月、アリババ「Qwen3.8 」 2026 年 8 月など）の台頭により、仮に米国側だけが減速した場合の帰結も論点となっている。

Bloomberg / AFP / Stanford HAI （2026.9 ）

米政権の立場 — 減速は政策にならない

POLITICS

中国が止まらないことは前提だった。では民主主義国内の協調はどうか

2026 年 9 月 13 日　トランプ米大統領の発言

- 訪問先のアイルランドで記者団に対し、 AI 開発を減速させるべきだとする慎重論を「否定的」な考え方だと批判した
- 否定的な勢力が多く、実際には起こらないことまで持ち出している、という趣旨の反論を行った
- AI を制する者が勝者となるとして、中国との覇権争いにおける主導権の維持を強調した
- 米国は AI で中国をリードしており、その状態を維持したい、と述べたと報じられている

これが意味すること

- 中国の同調が期待できない以上、提言の実現可能性は第二段階（民主主義国内の協調）に懸かっていた。その第二段階は、独禁法上の理由から米国政府の仲介を前提としている
- 米議会には Ban ASI Act と FRONTIER Act が提出されているが、いずれも成立していない
- 一方で開発企業の側は本気である。アルトマン氏は 9 月 12 日、安全面への懸念を理由に 2026 年中の OpenAI の IPO 見送りを表明した
- その前提が、提言の翌日に当の米国政府のトップによって否定された。企業は「減速したい」、政権は「減速しない」

企業は「進みすぎる危険」を見て、政府は「遅れる危険」を見ている。二つの危機感が正面からぶつかっている。

- 中国は同調せず、米政権も採らない。法制度による減速も、業界の自主規制による減速も、当面は実現しないと見るのが現実的である
- したがって、 AI の能力は今後も上がり続ける前提で、自社の統制を組まなければならない
- 外側の制度が整うのを待つ、という選択肢は最初から存在しない

だからこそ、企業が自分の手で統制を設計するしかない。これが、ここから申し上げる実務論の前提である。

AFP / 時事通信（2026.9.13-14）

AI は「答えるもの」から「行為するもの」へ

AGENTIC SHIFT

リスクの質が、回答ミスから外部影響へ移っている

Chatbot	Copilot	Agent	Multi-agent
質問に答える	人の作業を補助	ツールを使い実行	複数 AI が連携
誤回答・著作権・個人情報	出力利用・レビュー責任	外部接続・権限・ログ	予期しない連鎖・制御困難

従来の生成 AI 利用規程は「入力してよい情報」「出力の確認」に偏りがちである。しかし AI エージェントでは、閲覧・検索・送信・実行・購入・コード変更までが問題になる。

- 出力側の論点：著作権法第 30 条の 4 、個人情報保護法、景品表示法第 5 条、名誉毀損（民法第 709 条・第 710 条）
- 行為側の論点：代理・決裁権限、不正アクセス行為の禁止等に関する法律（平成 11 年法律第 128 号）、第三者への加害
- 法務は「AI が何を言ったか」だけでなく、「AI が何をしたか」を管理する必要がある。次のスライドから、その責任論に入る

規程の対象を「入力と出力」から「権限と行為」へ広げる。これが 2026 年の AI ガバナンスの分岐点。

AI が誤発注したら、誰が責任を負うか

LIABILITY 1

想定：購買業務の AI エージェントが、社内承認を経ずに取引先へ発注書を送信した

① 取引先との間（契約は成立するか）

- AI は代理人ではないが、担当者の ID で送信されていれば、代理権授与の表示（民法第 109 条）・権限外の行為の表見代理（民法第 110 条）が主張されうる
- 会社側からは錯誤（民法第 95 条）による取消しを検討するが、「重要な錯誤」該当性と重過失の有無が争点になる

② 社内・第三者に対して（不法行為）

- 運用者自身の過失を問う構成（民法第 709 条）が基本
- 使用者責任（民法第 715 条）は AI が「被用者」に当たらないため直接適用は困難。工作物責任（民法第 717 条）も想定していない

③ ベンダーに対して（求償）

- 債務不履行（民法第 415 条）。製造物責任法第 2 条第 1 項の「製造又は加工された動産」にソフトウェア単体は原則含まれないと解されてきた
- 実際には契約の責任制限条項が賠償額の上限を画する。ここが実務上の勝負どころ

④ 役員に対して（会社法）

- 善管注意義務（会社法第 330 条、民法第 644 条）、任務懈怠責任（会社法第 423 条第 1 項）
- 判断枠組みは内部統制システムの構築・運用（会社法第 362 条第 4 項第 6 号、会社法施行規則第 100 条第 1 項）

いずれの経路でも決め手になるのは、「誰の権限で、何を根拠に、どこまで自動で実行したか」の記録である。

未解決の論点と、いま埋められること

LIABILITY 2

答えが出ていないからこそ、契約と記録で自社に有利な事実関係を作る

未解決の 4 論点

帰属

AI の「行為」を誰の行為とみるか。使用者責任も工作物責任も予定していない

過失の基準

何を怠れば注意義務違反か。裁判規範は未形成。ソフトローが事実上の物差しになる

予見可能性

「指示していない行為」をどこまで予見すべきか。開発元の分析すら覆る領域

責任制限

B2B では消費者契約法の適用がなく、民法第 90 条による無効主張のハードルは高い

未解決のまま、いま埋められること

契約

- 責任制限の上限と除外事由
- 事故通知の期限と通知内容
- ログ提供義務・調査協力義務
- モデル変更の事前通知

記録

- 承認者・承認時刻・権限範囲
- ツール操作とデータアクセス
- 事後に再現できる保存形式
- 改ざん防止と保存期間

規程

- AI 経由の操作を決裁規程に明記
- 職務権限規程への位置づけ
- 無権限行為の社内評価基準
- 取締役会への報告事項化

「ここは未解決です」と言い切ったうえで、争点になったときに自社が出せる証拠を先に用意しておく。これが現時点での最善手。

持論：止めるのではなく、危険な自由度を潰す

VIEWPOINT

全面禁止は、現場のシャドー AI を増やすだけになりやすい

禁止令より、「使えるレーン」と「使えないレーン」を設計する。

止めるだけの規程

- 現場は別ツールへ逃げる
- ログも相談も残らない
- 事故は「見えないところ」で起きる

野放しの導入

- 便利さが先行する
- 事故後に説明できない
- 役員の善管注意義務が問われる

条件付き YES

- リスク別に許可を出す
- 監視・停止を組み込む
- 会社が説明できる状態を保つ

講師のメッセージ

AI を使うかではなく、どの自由度を残し、どの自由度を止めるか。
ここを決めるのが、 AI 時代の法務・コンプライアンスの仕事である。

次のスライドから、その「どの自由度を止めるか」を具体的な手続に落としていきます。

AI 利用の仕分け： 3 つの問いと 4 つのレベル

TRIAGE

審査の入口は抽象的な倫理ではなく、具体的な権限から

① AI は何を見られるか

個人情報・要配慮個人情報／営業秘密・顧客データ／法務相談・内部通報／認証情報

② AI は何をできるか

メール送信・外部投稿／購入・発注・決済／契約書送付・コード実行／ API 操作

③ 誰が止められるか

人間承認のゲート位置／停止権限者と手順／ログ監査の担当／事故時の連絡経路

この 3 問に答えられない利用は、棚上げせず「高リスク利用」として審査に回す。答えられた利用は、次の 4 段階のどこに当たるかで手続を変える。

Level 1	**閲覧・要約**	公開情報・社内一般文書
Level 2	**文書生成**	人間レビューを前提とした下書き作成
Level 3	**社内システム接続**	社内データの閲覧・検索・集計・分析
Level 4	**外部送信・決済・契約・コード実行**	不可逆な外部影響が生じうる利用

ゲートを置く 5 つの節目

外部ツール接続 ／ 外部送信 ／ 自律実行 ／ モデル変更 ／ 本番利用

各ゲートでの確認 7 項目：データ分類／権限／ログ／停止方法／ベンダー契約／事故通知／説明責任

Level 3 以上は事前確認、 Level 4 は人間承認・ログ・停止の 3 点すべてを許可条件とする。

対応 A　権限設計と人間承認

ACTION A

AI に社員の ID を貸さない。 AI 専用の権限を持たせる

AI に「社員の ID 」を貸すのは、統制上、最も危ない近道である。

AI 専用 ID

誰の指示で、どの AI が動いたかを事後に特定できる状態にする

最小権限

必要なデータ・機能だけに限定する。既定は「接続しない」

人間承認

不可逆な外部影響を生む操作の直前に、人間のゲートを置く

- メール送信、外部投稿、コードの本番反映、契約書送付、決済・発注は人間承認を原則化する
- クレジットカード、銀行口座、契約管理システム、顧客 DB への接続は、エージェント利用の中でも最重点
- 決裁規程・職務権限規程に「 AI 経由の操作」を明示的に位置づける。規程が人間の行為だけを想定したままだと、無権限行為の評価（民法第 110 条等）が争いになる
- OAI-HF 事件の教訓は、隔離したつもりの環境から外に出られたこと。権限設計は契約書ではなく実装で担保する

会社法第 362 条第 4 項第 6 号・会社法施行規則第 100 条第 1 項の内部統制システムに、 AI 経由の業務執行を明示的に組み込む。

対応 B　ログ・監視・キルスイッチ

ACTION B

AI 利用を止める権限と手順を、平時から設計する

ログ

- 入力・出力
- ツール操作・データアクセス
- 承認者と承認時刻
- 保存期間と改ざん防止

監視

- 異常な外部接続
- 権限昇格の試行
- 禁止データの投入
- 想定を超える連続実行

停止

- AI エージェントの即時停止
- API キーの失効
- ベンダー接続の遮断
- 再開条件の事前決定

インシデント発生時の法務の動線

- 個人情報の漏えい等が生じた場合の個人情報保護委員会への報告・本人通知（個人情報保護法第 26 条）
- 不正アクセス行為の禁止等に関する法律（平成 11 年法律第 128 号）に係る事実関係の整理と、被害届・当局対応の要否
- 金融機関は業法上の体制整備義務（銀行法第 12 条の 2 第 2 項ほか）と監督指針に基づく当局報告を併せて検討する
- 公表文には「現時点の評価」である旨を明記し、原因分析が更新された場合の再公表経路を確保しておく

改正個人情報保護法（2026 年改正）

AMENDED APPI

個人情報の保護に関する法律（平成 15 年法律第 57 号）附則第 10 条の 3 年ごと見直し

2026 年 4 月 7 日閣議決定、 5 月 26 日衆議院通過、 7 月 10 日成立、 7 月 17 日公布。施行は公布から 2 年以内（遅くとも 2028 年 7 月 17 日）で段階的に行われる。

① 統計作成等特例（改正法第 30 条の 2 ）

- 第 5 項により、統計作成等目的での利用について、一定の要件の下で本人同意なく第三者提供が可能となる
- 医療機関の診断結果等の要配慮個人情報も対象となり得る
- 特例の適用を希望する事業者は、氏名・名称、統計作成等の内容、提供のために取り扱う旨等をあらかじめ公表する必要がある
- 現行法第 18 条・第 27 条の同意原則を緩和するもの

② その他の主要改正事項

- 課徴金制度（改正法第 148 条の 3 ～第 148 条の 17 ）： 違法な取扱いにより財産上の利益を得た場合に、個人情報保護委員会が直接納付を命じうる
- 特定生体個人情報の新規律（顔認証・ DNA ・声紋等）
- 16 歳未満は法定代理人の同意を要する
- 委託先義務の緩和、漏えい等報告（現行第 26 条）の合理化

減速論と真逆の動き。開発側が「速度を落とす」と言う一方で、制度側は同意原則を緩めて AI 開発用データの流通を後押ししている。

法務の着手事項

- データ資産の棚卸しと、特定生体個人情報・特定個人アプローチ情報の洗い出し（対応 D のインベントリと同時に実施する）
- 統計作成等特例を用いる場合の公表事項の準備と、提供先との契約整備
- 「本人の権利利益を害するおそれ」の解釈は今後のガイドライン待ち。何が「害さないことが明らか」に当たるかが実務上の焦点

対応 C　データ境界を定める

ACTION C

AI に「入れてよい情報／入れてはいけない情報」を明確にする

原則可

- 公開情報・一般的知識
- 匿名加工済みの資料
- 社内の一般文書

条件付き

- 社内資料・顧客関連情報
- 契約書・議事録
- マスキング・ DLP を前提

原則禁止

- 要配慮個人情報・特定生体個人情報
- 弁護士との相談・内部通報資料
- 未公表の M&A 情報・認証情報

確認項目

- AI ベンダーが入力データを学習に使うか、どれだけ保持するか、国外移転が生じるか（外国にある第三者への提供は個人情報保護法第 28 条）
- 営業秘密（不正競争防止法第 2 条第 6 項）の 3 要件のうち、秘密管理性が AI への投入によって失われないか
- マスキング・ DLP ・アクセス制御によって、「人の注意」に依存しすぎない設計にする
- 弁護士への法律相談、内部通報、社内調査の資料は、別建ての特別ルールを設ける

禁止リストだけでなく、「この用途ならこの AI を使ってよい」という許可リストを必ずセットで出す。

対応 D　AI インベントリを作る

ACTION D

「誰が、何を、どの AI で、どの権限で」使っているか

Who	Tool	Purpose	Data	Authority
利用部門・責任者	利用 AI ・ベンダー	業務目的	投入情報	実行権限

作る目的と、見つけ方

- 紙の管理台帳を作ることが目的ではない。シャドー AI と、外部影響のある利用を見つけることが目的である
- 全社アンケートは当てにならない。経費精算の SaaS 決済明細、ブラウザ拡張機能、 SSO の認証ログのほうが早く実態に届く
- 申告期間を設けて「期間内の申告は不問」とするのが、最も短時間で棚卸しが進む現実的な手法
- 改正個人情報保護法のデータ資産マッピングと同じ作業になるため、併せて一度で実施する
- 取締役会・経営会議に報告できる粒度にまとめ、 AI 利用を経営リスクとして可視化する

見えていない AI 利用は、統制できない。事故の後に「把握していませんでした」は、説明として成立しない。

EU AI Act と Digital Omnibus

EU

「2026 年 8 月 2 日に全部が始まる」も「高リスクが延期されたから当面やることはない」も、どちらも誤り

Digital Omnibus on AI （Regulation (EU) 2026/1744 ）は 2026 年 7 月 8 日採択、 7 月 24 日に EU 官報掲載、 7 月 27 日発効。条文ごとに適用日がずれた状態で走っている。

2025.8.2	適用済	GPAI （汎用目的 AI モデル）提供者の義務
2026.8.2	適用開始	透明性義務（第 50 条）、調和規格・適合性評価・ CE 表示・登録、 GPAI 提供者への制裁金権限（第 101 条）
2026.12.2	適用開始	新設の禁止行為（同意なき性的ディープフェイク、 CSAM 生成）。高リスク義務の大半に先行する
2027.12.2	延期後	Annex III 型 高リスク AI 義務（採用 AI ・信用スコア・教育評価等）
2028.8.2	延期後	Annex I 型 製品組込型高リスク AI 義務（医療機器・機械等）

域外適用（第 2 条）。 EU 向けにチャットボット・ AI アシスタントを提供していれば、第 50 条の開示対応は 2026 年 8 月 2 日から必要。

※ EU もまた減速していない。高リスク義務を 16 か月後ろ倒しにした一方、ディープフェイク等の禁止は前倒しで効かせる、という選別が行われている。

対応 E　AI ベンダー契約で見る条項

ACTION E

AI サービスの調達は、通常の SaaS 契約の審査項目では足りない

データ利用

学習利用の可否、保持期間、再利用、削除、国外移転、サブプロセッサ

安全管理

アクセス制御、暗号化、環境隔離、脆弱性対応、評価環境の分離

監査・説明

ログ提供、第三者評価の結果開示、モデル変更の事前通知

事故対応

通知期限と通知内容、協力義務、フォレンジック、費用負担

権利関係

出力物の権利帰属、第三者権利侵害、補償、責任制限の上限と除外事由

提供継続

仕様変更・性能変更・提供停止の要件と予告期間、移行協力

- 責任分担の見立ては、経済産業省「AI 利活用における民事責任の解釈適用に関する手引き」（2026 年 4 月）を AI 事業者ガイドライン第 1.2 版とセットで参照する
- B2B では消費者契約法の適用がないため、責任制限条項は原則として有効に働く。交渉段階で上限額と除外事由を取りにいくしかない
- 「学習に使いません」が仕様書や FAQ にしか書かれていないことがある。契約本体に落ちているか、保持期間と削除まで追えているかを確認する

2026 年の新しい論点：ベンダー自身が「減速」を選択し、性能・提供内容・リリース時期を変更するリスクを、変更条項と移行協力義務で手当てする。

対応 F　社内規程と内部統制

ACTION F

50 頁の AI 規程より、現場が守れるルールを

AI 規程は、長さではなく「迷ったときに止まれるか」で評価する。

入れるべき事項

- 許可される利用例（許可リスト）
- 事前申請が必要な利用
- 禁止データ・禁止行為
- 人間レビュー・承認のゲート
- ログ・保存・事故報告の経路
- 違反時の対応と例外承認者

避けたい規程

- 抽象的な倫理原則だけで終わっている
- 禁止が広すぎて誰も守っていない
- 例外承認者が特定されていない
- 部門ごとのベンダー利用を把握していない
- 事故時の初動と連絡先がない
- エージェント利用を想定していない

規程の上位に置くべきもの

- 取締役会における内部統制システムの構築・運用に関する決議（会社法第 362 条第 4 項第 6 号、会社法施行規則第 100 条第 1 項）に AI リスクを位置づける
- AI 利用状況・審査件数と却下件数・インシデント・外部環境の変化を、取締役会または経営会議への定期報告事項とする
- 内部監査・監査役による独立した検証を年次で回す。減速論でいう「常駐評価者」の社内版にあたる

合言葉：法務は「使うな」ではなく「この条件なら使ってよい」を出す。

30 日・ 60 日・ 90 日の実装ロードマップ

ROADMAP

AI を止めずに、統制を追いつかせる

30 日

見える化

- 高リスク部門の AI 棚卸し
- 決済明細・ SSO ログの突合
- 禁止データ・禁止行為の暫定通知
- 申告期間の設定（期間内は不問）

60 日

ルール化

- 利用レベル分類（ Level 1 〜 4 ）の確定
- 契約・調達チェックリストの整備
- AI 専用 ID ・最小権限の設計
- 短い基本規程と FAQ の公開

90 日

運用化

- ログ・監視・停止手順の実装
- インシデント対応訓練の実施
- 審査運用の部門間分担を確定
- 経営会議・取締役会への初回報告

短期実装で大切なこと

完璧な AI ガバナンスの完成を待たない。まず、外部影響の大きい利用から「人間承認・ログ・停止」の 3 点を入れる。この 3 点が入っていれば、事故が起きても会社は説明できる。

法務部門の話法

COMMUNICATION

現場を止める言い方ではなく、動かすための条件を出す

悪い言い方

「AI は危ないので使わないでください。」

良い言い方

「この用途なら、データをマスクし、外部送信は人間承認にすれば使えます。」

法務が出すべき回答

- No ではなく「条件付き Yes」を増やす。No の総量ではなく、条件を明示できた件数で法務の仕事を測る
- 抽象論ではなく、データ・権限・承認・ログという 4 つの条件に翻訳して返す
- 安全に使える標準ツール・標準プロンプト・申請フォームを、法務の側から用意する
- 現場の便利さを奪わず、かつ事故時に会社が説明できる状態を作る。この両立が目的である

法務はブレーキ役ではなく、「安全に加速するためのガードレール」を作る役である。

実務でよくいただくご質問

FAQ

AI ガバナンスのご相談で頻度の高い 6 点について、現時点での考え方

Q. シャドー AI はどう見つけるか

全社アンケートは当てになりません。経費精算の SaaS 決済明細、ブラウザ拡張機能、SSO の認証ログを突き合わせるのが早い。申告期間を設けて期間内は不問とするのが最短です。

Q. 規程は法務と情シスのどちらが作るか

骨格は法務、技術要件は情シス。ただし例外承認者を法務単独にすると運用が回りません。部門長と法務の二者承認が現実的です。

Q. ベンダーの「学習に使わない」は信じてよいか

問題は真偽ではなく、契約のどこにどう書いてあるかです。仕様書や FAQ ではなく契約本体に落ちているか、保持期間・削除・サブプロセッサまで追えているかを見てください。

Q. 契約書や議事録を入れると営業秘密は失われるか

一律には失われません。不正競争防止法第 2 条第 6 項の秘密管理性が維持されているかが要点です。全社員が自由に使える環境へ無制限に投入する運用は危ういと考えます。

Q. AI の出力をそのまま顧客に出してよいか

レベルによります。人間レビューを外すなら、誤りが出たときの訂正経路と責任の所在を先に決めてください。表示規制がかかる領域では、原則としてレビューを外しません。

Q. 取締役会には何を報告すればよいか

利用状況、審査件数と却下件数、インシデント、外部環境の変化の 4 点。会社法第 362 条第 4 項第 6 号の体制の運用状況として記録に残すことが重要です。

※ いずれも一般論であり、個別の事実関係により結論は変わり得ます。個別のご相談は末尾の連絡先までお願いいたします。

まとめ：減速論からの 5 つの教訓

TAKEAWAYS

会社レベルでは「 AI の自由度を管理する」ことに尽きる

1 **止める議論に付き合わない**

提唱者自身が止めろとは言っていない。論点は能力向上と安全対策の " 時間差 " の管理である

2 **原因説明は覆る前提で設計する**

開発元ですら 2 か月で説明を撤回した。自社の公表文・当局報告にも更新経路を用意しておく

3 **規制の緩和と技術の危険化が同時に進む**

EU は高リスク義務を延期し、日本は同意原則に特例を設けた。法令遵守だけでは足りない

4 **責任論は未解決。だから記録で埋める**

使用者責任も製造物責任も AI を想定していない。争点化したときの手札を先に用意する

5 **法務の役割はブレーキではなく検証可能性の設計**

説明できる状態を保つことが、結果として最も速く進める道になる

AI を使うかではなく、どの自由度を残し、どの自由度を止めるか。
その判断を、データ・権限・承認・ログ・契約・事故対応に落とし込むことが、 AI 時代の企業法務の中核になる。

参考資料

SOURCES

本資料の作成にあたり参照した主な公表資料・報道（2026 年 9 月 14 日時点）

海外

- Dario Amodei, "We Must Pace the Frontier," September 2026
- OpenAI 「Hugging Face モデル評価に関するセキュリティインシデント」 2026 年 7 月
- Anthropic 「サイバー評価における実環境インシデントの調査」 2026 年 7 月 30 日
- Anthropic 「最近のサイバーインシデントに関するアライメント評価」 2026 年 9 月 9 日
- METR による独立レビューの開始に関する公表
- Stanford HAI 「AI Index 2026」
- Anthropic 「Responsible Scaling Policy」現行版 v3.4 （2026 年 7 月 8 日発効）
- Ban ASI Act （2026 年 9 月 4 日提出）／FRONTIER Act （H.R.9925 、 2026 年 7 月 23 日提出）
- WSJ ・ TIME 等による研究者退社に関する報道（2026 年 9 月）
- AFP ・時事通信によるトランプ米大統領の発言報道（2026 年 9 月 13-14 日）

国内

- 人工知能関連技術の研究開発及び活用の推進に関する法律（令和 7 年法律第 53 号）
- 人工知能基本計画（2025 年 12 月 23 日閣議決定）
- 総務省・経産省「AI 事業者ガイドライン（第 1.2 版）」 2026 年 3 月 31 日
- 経産省「AI 利活用における民事責任の解釈適用に関する手引き」 2026 年 4 月
- 総務省「AI セキュリティ確保ガイドライン」 2026 年 3 月
- 金融庁「AI ディスカッションペーパー（第 1.1 版）」 2026 年 3 月 3 日
- 改正個人情報保護法（2026 年 7 月 17 日公布）

注：本資料は 2026 年 9 月 14 日時点の公表情報に基づく一般的な解説であり、個別案件についての法的助言ではありません。改正個人情報保護法の条番号は改正法の規定によるものであり、今後の政令・規則・ガイドラインの整備状況により実務上の取扱いが変わり得ます。

弁護士法人三宅法律事務所　弁護士　渡邉雅之　／　TEL 03-5288-1021　／　m-watanabe@miyake.gr.jp

付録①　AI 法（推進法）の条文構造

APPENDIX 1

人工知能関連技術の研究開発及び活用の推進に関する法律（令和 7 年法律第 53 号）

令和 7 年 5 月 28 日成立、 6 月 4 日公布・一部施行、 9 月 1 日全面施行。本則 28 条・附則 2 条。 EU 型の包括的な事前規制ではなく、研究開発・活用の推進を目的とする推進法である。

第 1 条	**目的**	施策の総合的かつ計画的な推進
第 3 条	**基本理念**	経済社会・安全保障上の重要性、研究開発力の保持と国際競争力、透明性の確保、国際協力における主導的役割
第 4 条〜第 8 条	**責務規定**	国・地方公共団体・研究開発機関・活用事業者・国民
第 13 条	**指針の整備**	国際的な規範の趣旨に即した指針その他の施策
第 16 条	**調査研究・指導**	国による調査研究、研究開発機関・活用事業者への指導・助言・情報提供
第 18 条 / 第 19 条以下	**計画・体制**	人工知能基本計画／人工知能戦略本部（本部長：内閣総理大臣）

罰則も直接の義務もない。ただし第 7 条には「活用事業者の責務」が置かれている。ソフトローだから何もしなくてよい、が最も危うい。

※ 人工知能基本計画は 2025 年 12 月 23 日閣議決定。「世界で最も AI を開発・活用しやすい国」を目標として掲げる。

付録②　参照すべきソフトローの見取り図

APPENDIX 2

2026 年 3 月〜 4 月に主要文書が集中的に改訂・公表された

AI 事業者ガイドライン（第 1.2 版）

総務省・経済産業省　2026 年 3 月 31 日

罰則のないソフトローという性格は第 1.0 版から一貫。別添冒頭に「全て記載のとおりに実施することが求められているものではない」と明記されている。 AI エージェントの実装拡大を踏まえた改訂。

AI 利活用における民事責任の解釈適用に関する手引き

経済産業省　2026 年 4 月

ガイドライン第 1.2 版とセットで参照し、契約レビュー・委託先選定・社内ルール策定の各場面で責任分担の見立てを具体化する。

AI セキュリティ確保ガイドライン

総務省　2026 年 3 月

総務省サイバーセキュリティタスクフォース「 AI セキュリティ分科会」の議論を踏まえたもの。

AI ディスカッションペーパー（第 1.1 版）

金融庁　2026 年 3 月 3 日

130 社アンケート等を踏まえた「初期的な論点整理」であり、特定の対応を直ちに求める規制文書ではないと明記。顧客向けサービスへの生成 AI 活用の留意点が追記された。

実務の見方：拘束力ある法律でなくても、裁判・行政対応・取引先説明の場面で「実務上の物差し」になり得る。