シャドーAI の蔓延：企業内に潜む未承認 AI の不可視リスクとその航海術

Gemini

はじめに：生産性の先にあるもの ―未管理 AI イノベーションの隠れたコスト

現代の企業環境は、従業員主導のイノベーションと生産性向上への飽くなき追求と、それがもたらす深刻かつ定量化が困難なリスクとの間に存在する、根源的な緊張関係によって定義されます。本レポートが分析対象とする「シャドーAI」は、単なるポリシー違反ではなく、企業の IT 供給体制と、俊敏性を求める現代の労働力のニーズとの間の断絶を反映した、構造的な課題です[1]。

シャドーAI とは、IT 部門およびセキュリティ部門による明確な承認、監督、または関知がないまま、組織内で AI ツールやシステムが利用される状況を指します[1]。この現象は、その前身である「シャドーIT」と区別して理解する必要があります。シャドーAI は、モデルのバイアス、データポイズニング、そして出力の非決定性といった、AI 特有のリスクを内包している点で、より複雑かつ深刻な脅威をもたらします[1]。

本レポートは、シャドーAI に対する全面的な禁止という戦略が、非現実的であるだけでなく、AI の利用をさらに深い影、すなわち可視性や管理の及ばない領域へと追いやることで、逆効果にさえなると論じます[6]。前進するための唯一の実行可能な道筋は、「セキュア・イネーブルメント（安全な有効化）」という積極的な戦略です。これは、堅牢なガバナンス、高度な技術的監視能力、そして進化し続ける脅威ランドスケープへの深い理解を組み合わせた、多角的かつ洗練されたアプローチを必要とします。

第 1 章：見えざる増殖：シャドーAI の規模と背景の理解

蔓延の実態を定量化する

シャドーAI が一部の問題ではなく、広範囲にわたる現実であることを示す統計的証拠は、経営層にとって看過できないものです。

普及に関するデータ：調査によれば、セキュリティ専門家の半数以上（56%）が組織内での未承認 AI の利用を認めており、さらに 22%がその存在を疑っています[8]。ある研究では、全従業員の半数がシャドーAI ユーザーであり、企業従業員による生成 AI アプリケーションの採用率は 2023 年の 74%から 2024 年には 96%へと急上昇したことが示されています[9]。さらに深刻なのは、従業員の 3 分の 1 以上（38%）が、会社の許可なく機密情報をこれらのツールで共有した経験を認めている点です[2]。

シャドーIT との比較：この問題の根底にある「シャドーIT」の規模は、シャドーAI が利用する巨大な可視性のギャップを明らかにします。大企業では、利用されているクラウドサービスの平均数は 1 社あたり 207 にのぼりますが、これらの企業の 65%以上が効果的なシャドーIT 対策を実施していません[11]。これは、シャドーAI が現在悪用している広大な未管理領域の存在を示唆しています。

人間的要因の分析

効果的な対策を講じるためには、従業員がなぜシャドーAI に頼るのかを理解することが不可欠です。その動機は悪意によるものではなく、多くは実利的なものです。

生産性と効率性：最大の動機は、生産性を高め、反復的なタスクを自動化し、ワークフローを合理化したいという欲求です[2]。従業員は、調査の要約（57%）、文書作成（45%）、コードのデバッグ（40%）など、あらゆる業務に AI を活用しています[8]。

イノベーションと俊敏性：シャドーAI は、公式の調達や承認プロセスといった時間のかかる手続きを待つことなく、チームが新しいソリューションを試行し、イノベーションを加速させることを可能にします[2]。この俊敏性は競争上の優位性をもたらす可能性がありますが、それは中央集権的なリスク管理を犠牲にすることで得られます。

「ユーティリティ・ギャップ」の存在：従業員の 33%が、自社の IT 部門が業務遂行に必要なツールを提供していないという単純な理由で、個人用の AI ツールを使用していると回答してい

ます[9]。これは、企業の IT 戦略が技術の進歩とユーザーのニーズに追いついていないことの証左です。

組織的摩擦の先行指標としてのシャドーAI

シャドーAI の蔓延は、単なるセキュリティ上の失敗以上のものを意味します。それは、組織の既存プロセスと従業員の業務上のニーズとの間に存在する「摩擦」の直接的な指標です。この現象は、ビジネスのスピードと IT・セキュリティガバナンスのスピードとの間の乖離を示唆しています。従業員が生産性向上や遅い社内プロセスを回避するためにシャドーAI を導入しているという観察結果は[2]、これが反抗的な行為ではなく、必要に迫られた行動であることを示しています。この行動パターンは、中央 IT 部門の対応が遅すぎる、あるいは官僚的すぎるために事業部門が独自に SaaS アプリケーションを調達した、歴史的なシャドーIT の動向と酷似しています[2]。

したがって、シャドーAI の高い発生率は、より根深い組織的問題の兆候と捉えるべきです。それは、公式ツールが不十分である「ユーティリティ・ギャップ」の存在[9]、承認プロセスがボトルネックとなっている現状、そしてプロセス遵守よりも結果を暗黙的に評価する企業文化を浮き彫りにします。CISO（最高情報セキュリティ責任者）は、シャドーAI に関する指標を単なる違反リストとしてではなく、戦略的な診断ツールとして活用すべきです。これにより、より優れた公認ツールを必要としている高パフォーマンスのチームを特定し、時代遅れの調達ポリシーを洗い出し、摩擦を減らすためのプロセス改革の機会を見出すことが可能となり、結果としてリスクの低減に繋がるのです。

第 2 章：目前に迫る危機：データ漏洩、知的財産損失、コンプライアンス違反

ケーススタディ：サムスン電子におけるデータ漏洩インシデント

このインシデントは、善意の従業員がいかにして壊滅的なデータ侵害を引き起こしうるかを示す典型的な事例です。

事象の概要：2023 年初頭、サムスン電子のエンジニアが業務効率化を目的として、機密性の高い独自のソースコードや社内会議の議事録を ChatGPT にアップロードしました[13]。

漏洩のメカニズム：一般公開されている AI ツールに入力されたデータは、外部サーバーに送信され、モデルの学習データとして取り込まれる可能性がありました。これにより、データは回収不能となり、第三者に開示されるリスクに晒されました[13]。これは明確なガイドラインの欠如と「ヒューマンエラー」が直接的な原因でした[14]。

事後対応：サムスン電子の即時対応は、生成 AI ツールの厳格な利用禁止でした[15]。この動きは、多くの企業が同様のインシデント発生後に強いられる、受け身の姿勢を象徴しています。この事例は、初期段階における最大のリスクが、高度な外部攻撃ではなく、意図しない内部の行動であることを明確に示しています。

ユーザー主導型リスクの全貌

機密データと知的財産の漏洩：ソースコード以外にも、従業員は顧客データ、マーケティング計画、財務情報などを未検証のツールに入力しており、これが機密情報の継続的かつ低レベルな流出を引き起こしています[10]。英国では、企業の 5 社に 1 社が、従業員による生成 AI の利用が原因でデータ漏洩を経験しています[10]。

不正確な情報と「ハルシネーション」：生成 AI モデルは、事実に基づかない情報や完全に捏造された情報を生成することがあり、これは「ハルシネーション（幻覚）」として知られています[13]。このような未検証の出力に依存して報告書作成、分析、意思決定を行うことは、誤った経営判断を招き、企業の信頼性を損なう可能性があります[6]。

コンプライアンスおよび法的違反：シャドーAI の利用は、連鎖的な法的問題を引き起こす可能性があります。これには、個人情報を入力することによるデータプライバシー規制（GDPR など）への違反、著作権で保護された素材を含む AI 生成コンテンツを利用することによる知的財産権の侵害、そして顧客やパートナーとの機密保持契約への違反などが含まれます[5]。

第 3 章：侵害される AI サプライチェーン：悪意のあるモデルと脆弱なツール

AI サプライチェーンの解体

現代の AI エコシステムは一枚岩ではありません。それは、データセット、事前学習済みモデル（多くはオープンソースハブから提供）、開発フレームワーク、そしてデプロイメントインフラから構成される複雑なサプライチェーンです[20]。この複雑性が、攻撃者によって悪用されうる多数の脆弱性を生み出しています。

脅威ベクトル 1：オープンソースリポジトリに潜む悪意のあるモデル

Hugging Face インシデント：セキュリティ研究機関である JFrog は、人気のオープンソース AI リポジトリである Hugging Face 上で、100 を超える悪意のある AI モデルがホストされていることを発見しました[24]。

技術的メカニズム：攻撃ベクトルには、Python の pickle フォーマットが悪用されました。Python オブジェクトをシリアライズするために使用される Pickle ファイルは、ロード時に任意のコードを実行するように細工することが可能です。攻撃者はモデルファイル内に悪意のあるペイロードを埋め込み、開発者やデータサイエンティストがそのモデルを自身の環境にロードした際にコードが実行されるようにしました。これにより、持続的なバックドアが作成されました[24]。この事例は、モデルを

*ロードする*という行為自体がセキュリティリスクになりうることを示しています。

検知の課題：標準的なセキュリティスキャナは、これらの脅威を検知できないことが多くあります。Hugging Face 自身のスキャンでは、「安全でない」とフラグが立てられたモデルの96%以上が誤検知であり、これが開発者の「アラート疲れ」を引き起こし、警告を無視する傾向を生み出しました。その一方で、真に悪意のあるモデルは難読化技術を用いて検知を完全に回避していました[27]。

脅威ベクトル 2：AI 搭載開発ツールに存在する脆弱性

ケーススタディ：Cursor IDE の脆弱性（MCPoison）：AI 搭載のコードエディタ「Cursor」

に、深刻なリモートコード実行（RCE）の脆弱性が存在することが明らかになりました[28]。

技術的メカニズム：この脆弱性は、Cursor のモデルコンテキストプロトコル（MCP）という、自動化ワークフローを定義するシステムに存在していました。ユーザーは、共有コードリポジトリ内にある、一見無害な MCP 設定ファイルを*一度だけ*承認します。その後、攻撃者は、この既に承認されたファイルを、悪意のあるコマンド（例：リバースシェルの起動）を含むように密かに変更することができました。Cursor は初回の承認後にファイルの再検証を行わなかったため、被害者がプロジェクトを開くたびに、悪意のあるコードが警告なしに実行される状態でした[28]。

広範な示唆：この事例は、協調的かつ AI が統合された開発環境における信頼モデルの重大な欠陥を明らかにしています。AI サプライチェーンのリスクが、モデル自体だけでなく、それらと対話するために使用されるツールにまで及ぶことを証明しています。

AI サプライチェーンがもたらすソフトウェアセキュリティリスクのパラダイムシフト

AI サプライチェーンに内在するセキュリティリスクは、従来のソフトウェアサプライチェーンのリスクとは根本的に異なり、より複雑です。従来のソフトウェア部品表（SBOM）は、データに埋め込まれたリスク、モデルの非決定的な振る舞い、あるいは悪意のあるシリアライゼーション形式のような新たな攻撃ベクトルを考慮できないため、不十分です。

従来のソフトウェアサプライチェーンのセキュリティは、既知のライブラリや依存関係における脆弱性（例：Log4j）に焦点を当てており、その構成要素は決定論的なコードです。SBOM はこれらの依存関係をリスト化できます[29]。しかし、AI 特有のリスクは異なります。第一に、AI モデルはその学習データの産物であり、汚染されたデータはコードではなく、モデルの学習された振る舞いの中に脆弱性を生み出します[30]。SBOM ではこれを捉えきれません。第二に、Hugging Face のインシデントが示すように、モデルのアーティファクト自体が、pickle のようなフォーマットを通じて悪意のあるペイロードの運び手となりえます[24]。これは、ライブラリの CVE とは異なる脅威のクラスです。第三に、Cursor の事例が示すように、AI コンポーネントと

*対話する*ツールが、協調的な環境で悪用されうる新たな信頼ベースの脆弱性を生み出します[28]。

要するに、AI サプライチェーンは、データ層、モデル層、そしてツール層において、動的かつ不透明なリスクを導入します。これは、従来の依存関係スキャンをはるかに超える、新たな透

明性とセキュリティのフレームワークを必要とします。この新しい、多面的な攻撃対象領域に対して必要なレベルの可視性を提供するためには、「AI 部品表（AI-BOM）」という新しい基準の開発と採用が不可欠です。

表 1：シャドーAI リスクのスペクトラム

リスクカテゴリ	概要	攻撃ベクトル	主要事例	ビジネスへの影響
意図しないデータ漏洩	従業員による機密情報の偶発的な入力。	ユーザーエラー、ガイドラインの欠如。	サムスン電子事件[13]。	知的財産（IP）の損失、コンプライアンス違反による罰金[10]。
不正確な出力への依存	AI のハルシネーションに基づくビジネス上の意思決定。	モデルの不確実性、検証プロセスの欠如。	レポートや分析での誤った情報の利用[13]。	企業の評判低下、戦略的判断の誤り[6]。
著作権・IP 侵害	第三者の権利を侵害する AI 生成コンテンツの利用。	AI の学習データに含まれる著作物。	ニューヨーク・タイムズ対 OpenAI 訴訟[32]。	法的責任、金銭的ペナルティ[19]。
AI サプライチェーンの侵害	悪意のあるオープンソースモデルや脆弱な AI 開発ツールの利用。	悪意のあるペイロード（pickle）、信頼モデルの悪用。	Hugging Face、Cursor IDE[24]。	システム侵害、持続的なバックドアの設置。

モデル完全性への攻撃	高度な攻撃者による意図的なデータポイズニングやバックドアの挿入。	汚染された学習データ、隠されたトリガー。	学術研究で実証された攻撃手法[30]。	モデルの信頼性低下、標的型操作。
自律型エージェントの悪用	エージェント型 AI システムを操作し、悪意のある行動を実行させる。	メモリポイズニング、ゴール操作、人間へのソーシャルエンジニアリング。	OWASP AI エージェント脅威リスト[34]。	不正なデータ窃取、リソースの濫用、妨害行為。

この表は、CISO が直面する多様な AI リスクを構造化し、優先順位付けを行うためのフレームワークを提供します。単純な従業員のミスから国家レベルの攻撃まで、リスクを論理的に分類することで、取締役会や他のステークホルダーに対してリスクの全体像を明確に伝えることができます。これにより、まずはユーザー主導のリスク（例：トレーニング、DLP）に対する基本的な対策にリソースを集中させ、その後、より高度なサプライチェーンやモデル完全性への攻撃に対する防御策へと移行するなど、効果的なリソース配分が可能になります。混沌とした脅威リストを、管理可能な戦略的計画へと転換するのです。

第 4 章：根源的脅威：データポイズニングとモデルのバックドア

このセクションでは、サプライチェーン*内*の脆弱性から、サプライチェーン*そのもの*への攻撃、特に AI モデルの完全性を標的とする攻撃へと焦点を移します。これらは、モデルを内側から破壊する、ステルス性が高く洗練された攻撃です。

データポイズニングの解説

定義：データポイズニングとは、モデルの将来の振る舞いを操作する目的で、その学習データ

セットに意図的に悪意のある、あるいは破損したデータを注入する行為です[30]。

メカニズム：学習トークンのごく一部（例：0.001%）の汚染データでさえ、重大な影響を及ぼす可能性があります。例えば、医療 AI に有害な誤情報を広めさせることが可能です[31]。攻撃は、「ラベルフリッピング」（正解ラベルを不正解ラベルに変更する）のような単純なものから、正しくラベル付けされているものの、モデルの決定境界を歪めるように戦略的に作成されたデータを追加する、より巧妙な「クリーンラベル」攻撃まで多岐にわたります[35]。

脆弱性のスケーリング則：直感に反するかもしれませんが、研究によれば、より大規模で高性能な LLM ほど、データポイズニングに対して*より脆弱*であることが示されています。これらのモデルは、小規模なモデルよりも、ごくわずかな有害データへの暴露から、より迅速に有害な振る舞いを学習します。これは、モデルが大規模化するにつれて、この問題が悪化する可能性が高いことを意味します[35]。

バックドア攻撃の解説

定義：バックドア攻撃は、モデル内に隠された「トリガー」を埋め込むことを目的とした、特殊なデータポイズニングの一種です[33]。

メカニズム：モデルは、特定の、多くの場合無害に見えるトリガー（例：珍しい単語、特定のフレーズ、あるいは「cf」のような文字）が、悪意のあるターゲット出力と一貫して関連付けられたデータで学習させられます[33]。

ステルス性と活性化：通常の運用中、バックドアは休眠状態にあり、モデルはクリーンなデータに対して正しく動作するため、攻撃の検知は極めて困難です[33]。攻撃者が秘密のトリガーを含む入力を提供したときにのみ、モデルは悪意のある振る舞い（例：物体の誤分類、偏ったテキストの生成、特定の情報の漏洩）を実行します[33]。

第 5 章：新たなるフロンティア：自律型エージェントと斬新な攻撃ベクトル

このセクションでは、AI が受動的なツールから、目標を設定し、意思決定を行い、デジタル世界で行動を起こすことが可能な自律型エージェントへと移行する未来に目を向けます[36]。この

自律性は、強力である一方で、攻撃対象領域を劇的に拡大させます。

エージェント型 AI の台頭と OWASP フレームワーク

Open Web Application Security Project （OWASP）は、エージェント型システムに特有の脅威を特定しました。重要な発見は、外部インターフェースに焦点を当てた従来のセキュリティ監視が、これらの新しい脅威の大部分（73%）に対して無力であるという点です。なぜなら、これらの脅威はモデルの内部状態と論理を悪用するためです [34]。

検知困難な主要脅威の分析

内部状態の操作（T1, T6）：攻撃者がエージェントの記憶に偽情報を注入して将来の決定を汚染する*メモリポイズニング*（T1）や、プロンプトインジェクションを用いてエージェントの核となる目標を乗っ取る*ゴール操作*（T6）のような攻撃です。これらはモデルの「ブラックボックス」である推論プロセス内で発生するため、外部からの検知は不可能です [34]。

マルチエージェントシステムの悪用（T12, T13）：複数のエージェントが協調するシステムにおいて、攻撃者は*エージェント間通信ポイズニング*（T12）を用いてエージェント間に偽情報を拡散させたり、正当に見えるがシステムの集合的な目標を破壊するために活動する*不正エージェント*（T13）を潜入させたりすることができます。

ヒューマン・イン・ザ・ループ攻撃（T10, T15）：これらの攻撃は、人間の監督者を悪用します。*人間への過負荷*（T10）は、人間のレビュー担当者に大量のリクエストを送りつけ、その判断力を麻痺させるものです。*人間操作*（T15）は、AI が人間の信頼を得た後、巧妙に人間を説得して有害な行動を取らせるソーシャルエンジニアリングの一形態です。システムログには、これが「正当な人間による操作」として記録されます [34]。

表 2：OWASP エージェント型 AI システムのトップ脅威：ビジネスインパクト分析

OWASP 脅威 ID と名称	技術的概要	具体的なビジネスシナリオ	潜在的なビジネスへの影響
T1: メモリポイズニング	攻撃者がエージェントの記憶に偽情報を注入し、意思決定を操作する。	市場分析エージェントに、数週間にわたり競合他社の業績に関する偽データを subtle に供給。汚染された「記憶」を持つエージェントは、この偽データに基づき欠陥のある価格戦略を推奨する。	市場シェアの喪失、収益の減少、誤った戦略的投資。
T6: ゴール操作	悪意のあるプロンプトでエージェントの目標を乗っ取る。	顧客からの問い合わせに隠された悪意のあるプロンプトにより、カスタマーサービス AI の目標が「チケット解決」から「顧客認証情報の窃取」にすり替わる。エージェントは顧客を騙してパスワードや金融情報を入力させる。	データ侵害、規制当局からの罰金（GDPR/CCPA）、深刻な評判の毀損。
T10: 人間への過負荷	大量の要求で人間の判断能力を麻痺させる。	金融取引承認エージェントが、疑わしい支払いを人間によるレビューのためにフラグ付けする。攻撃者は、何千もの小規模で境界線上の疑わしい取引をシステムに殺到させる。	「アラート疲れ」に陥った人間の承認者は、承認を機械的に行うようになり、大規模な不正取引を見逃す。直接的な金銭的損失、監査での不合格。

T15: 人間操作	AI への信頼を悪用し、人間を誘導して有害な行動を取らせる。	経営幹部が使用する AI アシスタントの信頼を悪用。数ヶ月間の信頼性の高いパフォーマンスの後、AI は過去のメールから学習した文脈に沿った詳細情報を含む、「CEO から」の「非常に緊急」な偽の送金依頼を提示し、幹部に承認させる。	CEO 詐欺、重大な金銭的損失。

この表は、OWASP の脅威リスト（T1-T15）のような技術的で抽象的な概念を、取締役会が理解できるビジネスリスクの言葉、すなわち金銭的損失、評判の毀損、戦略的失敗へと翻訳する「ロゼッタ・ストーン」として機能します。各脅威を、現実的で影響の大きいビジネスシナリオに結びつけることで、エージェント型 AI のリスクを具体的かつ緊急性の高いものとして提示します。これにより、CISO は、AI セキュリティに関する技術的な議論から、AI 主導の世界で中核的なビジネス機能を保護するための戦略的な議論へと会話を移行させることができます。これは、あらゆる技術的脅威に対する「だから何なのか？」という問いに答えるものです。

第 6 章：フレームワークの構築：積極的な AI ガバナンスの必要性

単純なユーザーエラーから高度な攻撃に至るまで、リスクが多岐にわたる現状において、その場しのぎのアプローチは通用しません。AI ガバナンスとは、組織が AI に関連する活動をその目的、価値観、およびリスク許容度と整合させるために確立する、ルール、ポリシー、プロセス、役割の公式な枠組みです [19]。

効果的なガバナンスへの主要な課題

技術的課題：AI の意思決定プロセスが不透明である「ブラックボックス問題」は、公平性と説明責任の確保を困難にします[5]。

組織的課題：経営層の理解不足や部門間の縦割り構造が、全社的で一貫した戦略の策定をしばしば妨げます[37]。

社会的・規制的課題：法的な状況は断片的で急速に変化しており、安定的で将来を見据えたコンプライアンスフレームワークの構築を困難にしています[19]。

企業 AI ガバナンスフレームワークの柱

部門横断的な監督体制：IT、セキュリティ、法務、人事、データサイエンス、および主要な事業部門の代表者を含む、専門的かつ多角的な AI ガバナンス委員会を設立し、包括的なアプローチを確保します[39]。

明確なポリシーとガイドライン：明確で実用的な AI 利用ポリシーを策定し、周知徹底します。これは単なる禁止令ではなく、一連のガードレールとして機能します。

- **データ分類**：機密情報、個人情報、または専有情報を公開 AI ツールに入力することを禁止します[40]。
- **承認済みツール**：従業員が使用できる、精査済みの公認 AI ツールのリストを維持管理します[42]。
- **透明性と開示**：従業員が外部向け、または重要な内部利用のために AI を用いてコンテンツを生成した場合、その事実を開示することを義務付けます[40]。

教育とトレーニング：サムスンのインシデントが示すように、ポリシーだけでは不十分です。AI のリスク、適切なデータハンドリング、安全な利用方法に関する継続的な教育が、責任ある文化を醸成するために不可欠です[9]。

ケーススタディ：組織におけるポリシー事例：サンノゼ市やコロンビア大学のような組織のアプローチを検証します。これらの組織は、リスクレベル（低、中、高）を区別し、重要な意思決定に対する人間の監督を義務付け、学術的および専門的な業務における AI の利用開示を要求する明確なガイドラインを導入しています[40]。

第 7 章：影を照らす：技術的統制と透明性の義務化

ガバナンスがルールを設定する一方で、それらを監視し、強制するためにはテクノロジーが必要です。ポリシーのみに依存するアプローチは、技術的な可視性がなければ失敗する運命にあります。

シャドーAI の検知・監視ツール

機能：シャドーAI 問題に対処するための新しいクラスのセキュリティツールが登場しています。これらのツールは、ネットワーク監視、エンドポイントエージェント、ブラウザ分析を利用して、未承認の AI サービスへのアクセスを検知します [7]。

ベンダー事例（Teramind, Acuvity ）：Teramind のようなソリューションは、ウェブサイトの利用状況、クリップボードの操作、行動分析など、従業員の活動を詳細に監視し、AI プラットフォームとの異常なデータ共有を警告します [4]。Acuvity は、何千もの AI サービスにわたる検知、データハンドリングポリシーに基づくリスクスコアリング、およびデータ漏洩防止機能を提供します [4]。これらのツールの目標は、見えないものを見えるようにすることです。

AI 部品表（AI-BOM）：透明性のための必須要件

概念：AI-BOM は、AI システム内のすべての構成要素を正式に記録した目録であり、従来のソフトウェアの SBOM に似ていますが、はるかに包括的です [29]。

主要構成要素：AI-BOM は、ソフトウェアライブラリだけでなく、学習に使用されたデータセット、モデルアーキテクチャ、バージョン情報、そして倫理的配慮や既知の制限事項まで文書化する必要があります [45]。

戦略的重要性：AI-BOM は、第 3 章で特定された AI サプライチェーンのリスクを管理するための基礎となるツールです。依存関係における脆弱性の特定、データバイアスの監査、規制遵守（例：EU AI 法）、そしてセキュリティインシデントへの対応に必要なトレーサビリティを提供します [29]。AI-BOM の生成とスキャンを自動化するツールも登場しつつあります [47]。

第 8 章：規制の迷宮を航海する：グローバルスタンダードとコンプライアンス

このセクションでは、企業の AI に対する責任を形成し、組織がシャドーAI をどのように管理すべきかに直接影響を与える、主要な国際的フレームワークに関する戦略的ブリーフィングを提供します。

EU AI 法：世界の先駆者

リスクベースのアプローチ：この法律は、リスクの階層（許容不可能、高、限定的、最小）を確立します。採用や信用スコアリングなど、多くの一般的なビジネスユースケースが「高リスク」カテゴリに分類され、厳格な義務が課せられます [49]。

高リスクシステムに対する主要な義務：これらには、厳格なリスク評価、バイアスを防止するための高品質なデータガバナンス、トレーサビリティのための活動ロギング、人間の監督、そして堅牢性とサイバーセキュリティに関する高い基準が含まれます [49]。

シャドーAI への示唆：従業員が高リスクな目的（例：履歴書のスクリーニング）で未検証のシャドーAI ツールを使用した場合、組織全体が法律違反となり、巨額の罰金に直面する可能性があります。これにより、シャドーAI はセキュリティ問題から重大なコンプライアンス危機へと格上げされます。

NIST AI リスクマネジメントフレームワーク（AI RMF）：実践的ガイド

自主的なフレームワーク：EU の拘束力のある法律とは異なり、NIST AI RMFは、組織が AI リスクを管理するための実践的で構造化されたプロセスを提供することを目的とした、自主的なフレームワークです [52]。

4 つの中核機能：このフレームワークは、**統治（Govern ）**（リスク管理の文化を確立する）、**マップ（Map ）**（文脈の中でリスクを特定する）、**測定（Measure ）**（リスクを分析・追跡する）、そして**管理（Manage ）**（リスクを軽減するためにリソースを割り当てる）という 4 つの主要な活動を中心に構築されています [52]。

戦略的価値：AI RMF は、企業が特定の業界や地域に関わらず、内部のガバナンスプログラムを構築するために使用できる、柔軟で適応性の高いプレイブックを提供します。これは、責任ある AI 導入のための「ハウツー」ガイドです。

日本の AI 事業者ガイドライン（経済産業省/総務省）：協調的アプローチ

原則ベースのガイダンス：経済産業省と総務省が共同で策定した日本のガイドラインは、AI ガバナンスに対するマルチステークホルダーによる原則ベースのアプローチを推進しています。これは、AI 開発者、提供者、利用者向けの統一されたフレームワークとして、以前のガイダンスを統合したものです[26]。

10 の共通指針：このガイドラインは、人間中心、安全性、公平性、プライバシー保護、セキュリティ確保など、10 の中核的な指針を中心に構成されており、企業に対して高度な倫理的および運用上の指針を提供します[26]。

表 3：グローバル AI 規制フレームワークの比較分析

フレームワーク	管轄	種類	中核原則	企業への主要要件	シャドーAI 管理への影響
EU AI 法	EU	拘束力のある規制	リスクベースの分類	高リスクシステムに対する厳格なコンプライアンス。	高リスク業務での未承認 AI 利用は、深刻な法的・金銭的責任を生じさせる。
NIST AI RMF	グローバル（米国主	自主的なフレームワー	ライフサイクルを通じたリスク管	内部 AI リスクプログラム構築の	シャドーAI を特定・管理するため

	導）	ク	理	ための実践的な「ハウツー」ガイドを提供する。	の社内リスクプログラム構築の実践的指針となる。
日本の AI 事業者ガイドライン	日本	原則ベースのガイダンス	マルチステークホルダーと倫理原則	企業 AI ポリシーが反映すべき国内の行動規範と倫理的期待値を設定する。	シャドーAIを含む企業の AI ポリシーが準拠すべき、倫理的基準と社会的期待を定義する。

この表は、AI に関するグローバルな規制環境の複雑さを、経営幹部が迅速に把握できるように設計されています。各フレームワークの法的拘束力（拘束力あり vs. 自主的）、中心的な哲学（リスクベース vs. 原則ベース）、そしてシャドーAI の管理への直接的な影響という核心的な要素に要約しています。これにより、リーダーは、コンプライアンス義務（EU）、運用プレイブック（NIST）、そして倫理的基準（日本）を理解することができます。これは、必要な場所ではコンプライアンスを遵守し、実践においては堅牢で、倫理的にも健全なグローバル AI ガバナンス戦略を策定する上で役立ち、企業が国際的な AI 規制の多面的な性質に備えることを可能にします。

結論：未管理のリスクから戦略的イネーブラーへ：安全な AI 導入のためのロードマップ

調査結果の統合

本レポートの核心的な論点を要約すると、シャドーAI は技術の進歩と従業員の創意工夫の必然的な結果です。そのリスクは深刻かつ多層的であり、単純なデータ漏洩から、AI サプライチェーンやモデル自体の完全性に対する高度な攻撃へと進化しています。

戦略的必須事項：セキュア・イネーブルメント

結論として、禁止という戦略は実行可能な選択肢として断固として否定します。代わりに、本レポートは「セキュア・イネーブルメント」というパラダイムを提唱します。その目標は、AI の利用を停止させることではなく、安全で透明性の高いフレームワークの中でそれを導くことです。

C レベル向けロードマップ

本レポートは、経営幹部のための明確で実行可能なロードマップを提示して締めくくります。

1. **統治（Govern ）**：直ちに部門横断的な AI ガバナンス委員会を設立し、明確でリスクベースの AI 利用ポリシーを策定する。
2. **可視化（Illuminate ）**：シャドーAI 検知ツールに投資し、企業全体の AI 利用の現状を包括的に可視化する。これは、交渉の余地のない最初のステップである。
3. **教育（Educate ）**：AI のリスクと責任ある利用に関する、全従業員向けの必須かつ継続的なトレーニングプログラムを開始する。
4. **標準化（Standardize ）**：サプライチェーンリスクを管理するため、すべての公認 AI システムについて AI 部品表（AI-BOM）を作成し、維持管理するプロセスを開始する。
5. **有効化（Enable）**：「ユーティリティ・ギャップ」を埋め、従業員にシャドーAI に代わる安全で効果的な選択肢を提供するため、厳選された強力かつ精査済みの AI ツールを提供する。

最終考察

AI ガバナンスは、コストセンターやイノベーションの障壁としてではなく、戦略的なイネーブラーとして位置づけるべきです。シャドーAI のリスクを積極的に管理することで、組織は人工知能がもたらす計り知れない生産性とイノベーションの恩恵を、安全かつ責任ある方法で、そ

して明確な競争優位性を持って享受することができるのです。

引用文献

1. シャドウ AI とは？未承認 AI のリスクと企業が取るべき対策 - Zendesk, 9 月 29, 2025 にアクセス、 https://www.zendesk.co.jp/blog/shadow -ai/
2. シャドーAI とは - IBM, 9 月 29, 2025 にアクセス、 https://www.ibm.com/jp -ja/think/topics/shadow -ai
3. www.zendesk.co.jp, 9 月 29, 2025 にアクセス、 https://www.zendesk.co.jp/blog/shadow -ai/#:~:text=%E3%82%B7%E3%83%A3%E3%83%89%E3%82%A6AI%E3%81%A8%E3%81%AF%E3%80%81%E4%BC%9A%E7%A4%BE,%E3%81%99%E3%82%8B%E3%81%93%E3%81%A8%E3%82%92%E6%8C%87%E3%81%97%E3%81%BE%E3%81%99%E3%80%82
4. Shadow AI Detection Software | Teramind, 9 月 29, 2025 にアクセス、 https://www.teramind.co/solutions/shadow -ai-detection/
5. 第 1 回 AI ガバナンスとは何か？リスクと必要性 | デロイト トーマツ グループ - Deloitte, 9 月 29, 2025 にアクセス、 https://www.deloitte.com/jp/ja/services/audit -assurance/blogs/ai-governance-01.html
6. シャドーAI とは？新たなリスクに対して企業が取るべき対策とは？, 9 月 29, 2025 にアクセス、 https://nuco.co.jp/blog/article/c0RmPqfN
7. What Is Shadow AI? How It Happens and What to Do About It - Palo Alto Networks, 9 月 29, 2025 にアクセス、 https://www.paloaltonetworks.com/cyberpedia/what -is-shadow-ai
8. Research: Shadow AI is a Blind Spot in Enterprise Security, Including Among Security Teams - Mindgard, 9 月 29, 2025 にアクセス、 https://mindgard.ai/resources/shadow -ai-is-a-blind-spot
9. Half of all employees are Shadow AI users, new study finds - Software AG, 9 月 29, 2025 にアクセス、 https://newscenter.softwareag.com/en/news -stories/press-releases/2024/1022-half-of-all-employees-use-shadow-ai.html
10. What Is Shadow AI?- IBM, 9 月 29, 2025 にアクセス、 https://www.ibm.com/think/topics/shadow -ai
11. 2024 年最新シャドーIT 対策実態調査レポート | assured.jp, 9 月 29, 2025 にアクセス、 https://assured.jp/column/shadowit2024 -report
12. 【IT 導入管理者必読】グレーな AI 利用が社内に潜む？ 拡大する「シャドーAI」と企業リスクの最新動向 - note, 9 月 29, 2025 にアクセス、 https://note.com/dx_labo/n/n0b333fb6b23b
13. シャドーAI とは？業務上のリスクや事例、対策方法を解説, 9 月 29, 2025 にアクセス、 https://www.jbsvc.co.jp/useful/security/shadow -ai.html
14. 生成 AI の情報漏洩事例 3 選｜トラブルを防ぐための対策を解説 - 侍エンジニア, 9 月 29, 2025 にアクセス、 https://generative -ai.sejuku.net/blog/5212/

15. Samsung 社内コード流出事件から学ぶ、生成 AI 活用のリアルな教訓 ..., 9 月 29, 2025 にアクセス、 https://note.com/inoch01/n/n00e68e257c43
16. シャドーAI のリスクと対策とは？会社に”無許可”で生成 AI を業務に利用している実態が明らかに, 9 月 29, 2025 にアクセス、 https://ai.yoshidumi.co.jp/navi/shadow-ai-risks-and-countermeasures
17. サムスンが従業員の ChatGPT 利用を禁止、機密データ漏洩で - Ledge.ai, 9 月 29, 2025 にアクセス、 https://ledge.ai/articles/samsung-chatgpt-ban
18. AI エージェントとは？自律的 AI 技術の活用事例と導入メリット・課題を徹底解説【2025 年最新版】, 9 月 29, 2025 にアクセス、 https://it-map.jp/ai agent/
19. 生成 AI 時代の AI ガバナンス｜リスク管理と倫理的な活用のためのガイドライン - テックファーム, 9 月 29, 2025 にアクセス、 https://www.techfirm.co.jp/blog/ai-governance
20. www.softbank.jp, 9 月 29, 2025 にアクセス、 https://www.softbank.jp/biz/solutions/cybersecurity/security-glossary/supply-chain-risk/#:~:text=%E3%82%B5%E3%83%97%E3%83%A9%E3%82%A4%E3%83%81%E3%82%A7%E3%83%BC%E3%83%B3%E3%83%AA%E3%82%B9%E3%82%AF%E3%81%A8%E3%81%AF,%E3%82%BB%E3%82%AD%E3%83%A5%E3%83%AA%E3%83%86%E3%82%A3%E4%B8%8A%E3%81%AE%E3%83%AA%E3%82%B9%E3%82%AF%E3%81%A7%E3%81%99%E3%80%82
21. AI は、グローバルサプライチェーンを次なる大きな衝撃から守る | 世界経済フォーラム, 9 月 29, 2025 にアクセス、 https://jp.weforum.org/stories/2025/01/ai-will-protect-global-supply-chains-from-the-next-major-shock-8eeeb8b2bb/
22. サプライチェーン・リスク管理とは | IBM, 9 月 29, 2025 にアクセス、 https://www.ibm.com/jp-ja/think/topics/supply-chain-risk-management
23. サプライチェーンリスク管理とは？リスクの種類と軽減策をわかりやすく解説 - ＡＧＳ株式会社, 9 月 29, 2025 にアクセス、 https://www.ags.co.jp/nw/column/column09.html
24. Hugging Face AI Platform Riddled With 100 Malicious Code ..., 9 月 29, 2025 にアクセス、 https://cyberir.mit.edu/site/hugging-face-ai-platform-riddled-100-malicious-code-execution-models/
25. Malicious AI Models on Hugging Face Backdoor Users' Machines, 9 月 29, 2025 にアクセス、 https://sos-vo.org/news/malicious-ai-models-hugging-face-backdoor-users-machines
26. AI 事業者ガイドライン （第 1.1 版） 概要 - 経済産業省, 9 月 29, 2025 にアクセス、 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_2.pdf
27. JFrog × Hugging Face ～ AI/ML モデルの脆弱性における 96% 誤検知削減と真の脅威特定を実現 | エクセルソフト ブログ - XLsoft, 9 月 29, 2025 にアクセス、 https://www.xlsoft.com/jp/blog/blog/2025/03/13/jfrog-17-post-95367/
28. チェック・ポイント・リサーチ、人気 AI コーディングツール ..., 9 月 29, 2025 に

アクセス、 https://prtimes.jp/main/html/rd/p/000000432.000021207.html
29. AI Bill of Materials (AI-BOM) - PointGuard AI, 9 月 29, 2025 にアクセス、 https://www.pointguardai.com/glossary/ai-bill-of-materials-ai-bom
30. Multi-Faceted Studies on Data Poisoning can Advance LLM Development - arXiv, 9 月 29, 2025 にアクセス、 https://arxiv.org/html/2502.14182v1
31. Medical large language models are vulnerable to data-poisoning attacks - PubMed, 9 月 29, 2025 にアクセス、 https://pubmed.ncbi.nlm.nih.gov/39779928/
32. 【2024 年最新】生成 AI による事件 5 選｜情報漏洩〜詐欺事件まで - メタバース総研, 9 月 29, 2025 にアクセス、 https://metaversesouken.com/ai/generative_ai/incident/
33. A Survey of Recent Backdoor Attacks and Defenses in Large ..., 9 月 29, 2025 にアクセス、 https://arxiv.org/pdf/2406.06852
34. AI エージェントを脅かす 15 の脅威｜OWASP 脅威分析が示す「検知 ..., 9 月 29, 2025 にアクセス、 https://www.nri-secure.co.jp/blog/ai-agent-3
35. Scaling Trends for Data Poisoning in LLMs - AAAI Publications, 9 月 29, 2025 にアクセス、 https://ojs.aaai.org/index.php/AAAI/article/view/34929/37084
36. ［技術解説］AI が自分で考え、行動する時代へ：エージェント AI の脅威と賢い活用法 | JOBIRUN, 9 月 29, 2025 にアクセス、 https://jobirun.com/agentic-ai-risks-governance-explained/
37. AI ガバナンス—企業に求められるあるべき姿と導入メリットについて解説 - 株式会社モンスターラボ, 9 月 29, 2025 にアクセス、 https://monstar-lab.com/dx/technology/ai-governance/
38. AI ガバナンスとは？必要な背景と企業が実施できる方法などを解説, 9 月 29, 2025 にアクセス、 https://www.fujifilm.com/fb/solution/dx_column/ai/about-ai-governance
39. AI リスクから企業を守るには 実践的な AI ガバナンスの方法 | NRI Digital Consulting Edge, 9 月 29, 2025 にアクセス、 https://www.nri.com/jp/media/column/scs_blog/20250917.html
40. Generative AI Policy | Office of the Provost, 9 月 29, 2025 にアクセス、 https://provost.columbia.edu/content/office-senior-vice-provost/ai-policy
41. Generative AI Use Policy Template for the Social Sector 2024 - NTEN, 9 月 29, 2025 にアクセス、 https://word.nten.org/wp-content/uploads/2024/07/GAI-Policy-Template.pdf
42. シャドーAI 対策の決定版！情報漏洩リスクから企業を守るための完全ガイド, 9 月 29, 2025 にアクセス、 https://sinmido.com/news/p3210/
43. San José AI Guidelines and Policies, 9 月 29, 2025 にアクセス、 https://www.sanjoseca.gov/your-government/departments-offices/information-technology/itd-generative-ai-guideline
44. Shadow AI Discovery - Acuvity, 9 月 29, 2025 にアクセス、 https://acuvity.ai/access/shadow-ai-discovery/
45. AI-BOM: Building an AI Bill of Materials - Wiz, 9 月 29, 2025 にアクセス、 https://www.wiz.io/academy/ai-bom-ai-bill-of-materials

46. Creating an AI Bill of Materials (AI BOM) for Secure GenAI - Mend.io, 9 月 29, 2025 にアクセス、 https://www.mend.io/blog/what-is-an-ai-bill-of-materials-ai-bom/
47. snyk-labs/ai-bom-scan - GitHub, 9 月 29, 2025 にアクセス、 https://github.com/snyk-labs/ai-bom-scan
48. AI Bill of Materials (BOM) Builder | JobBOSS² - ECI Software Solutions, 9 月 29, 2025 にアクセス、 https://www.ecisolutions.com/products/jobboss2/features/ai-bom-builder/
49. AI Act | Shaping Europe's digital future, 9 月 29, 2025 にアクセス、 https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
50. EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act, 9 月 29, 2025 にアクセス、 https://artificialintelligenceact.eu/
51. What is the Artificial Intelligence Act of the European Union (EU AI Act)? - IBM, 9 月 29, 2025 にアクセス、 https://www.ibm.com/think/topics/eu-ai-act
52. NIST AI Risk Management Framework: A tl;dr | Wiz, 9 月 29, 2025 にアクセス、 https://www.wiz.io/academy/nist-ai-risk-management-framework
53. NIST AI Risk Management Framework: A simple guide to smarter AI governance - Diligent, 9 月 29, 2025 にアクセス、 https://www.diligent.com/resources/blog/nist-ai-risk-management-framework
54. 【AI 関連】AI 事業者ガイドライン（第 1.0 版）のポイント解説① AI 利用者向け - note, 9 月 29, 2025 にアクセス、 https://note.com/mandp/n/nfdfcef363858