セキュリティ・トラブルを Generative AI Security Textbook

未然に防ぐ知識と実践

生成AIセキュリティの教科書

Shinichi Shichiri

七里信一

累計24万人※に生成AIの活用方法を教えてきた人気セミナー講師が ※2025年6月現在

生成AIの「安全な使い方」を

徹底解説!

生成AI研究所出版

はじめに ✨

【リスクを克服し、成長を加速させる全社戦略】

貴社では、生成AIの導入について、どのような議論がなされているでしょうか？

革新的な技術である一方、そのリスクを懸念し、導入に踏み切れない…そのような状況かもしれません。

しかし、今、その決断の遅れが、将来の成長機会を大きく損なう可能性があるとしたら、どうでしょうか？

本書は、生成AIを「正しく」活用し、企業の成長を加速させることを目的としています。生成AIは、適切に活用すれば、業務効率を劇的に向上させ、新たな価値創造を可能にする強力なツールです。

【AI活用のメリット：競争優位の確立】

・生産性の飛躍的向上

貴社の作業労力が大幅に削減される可能性があります。

・新たな価値創造

これまで不可能だった製品やサービスの開発が可能になります。

・競争優位性の確立

競合他社がAIを活用して成長を加速させる中で、取り残されるリスクを回避できます。

【AI活用のデメリット（リスク）と対策：本書の役割】

もちろん、AI導入にはリスクも伴います。情報漏洩、著作権侵害、誤情報拡散、バイアスのあるコンテンツ生成など、懸念される点は少なくありません。本書では、これらのリスクを詳細に分析し、具体的な対策を提示します。

・リスクの可視化

どのようなリスクが存在するのか、具体的な事例を交えて解説します。

・対策の具体化

リスクを最小限に抑えるための、実践的な対策を提示します。

・成功事例の共有

AI導入に成功した企業の事例を紹介し、成功への道筋を示します。

・倫理的配慮

AI倫理に関する最新の動向を踏まえ、責任あるAI活用のための指針を示します。

【日本の大企業が生成AIの活用に慎重な姿勢を見せる理由】

●リスクへの懸念

・情報漏洩

生成AIは大量のデータを学習するため、機密情報や個人情報が漏洩するリスクがあります。特に、顧客情報や企業秘密を扱う大企業にとっては、このリスクは非常に深刻です。

・著作権侵害

生成AIが作成したコンテンツが、既存の著作物を侵害する可能性があります。特に、コンテンツ制作に関わる企業や、知的財産を重視する企業にとっては、法的リスクが懸念されます。

・誤情報・偽情報

生成AIは、事実に基づかない情報や、偏った情報を生成する可能性があります。これが社会に拡散されると、企業のブランドイメージを毀損したり、風評被害につながる可能性があります。

・品質管理

生成AIの出力は、必ずしも期待通りの品質であるとは限りません。品質管理体制が整っていない段階での導入は、業務効率の低下や、顧客からの信頼を失うリスクがあります。

●日本特有の文化的・社会的背景

・前例主義・保守性

日本の企業文化は、前例踏襲やリスク回避を重視する傾向があります。新しい技術の導入には慎重で、特に大企業ほどその傾向が強いです。

・終身雇用・年功序列

終身雇用や年功序列といった制度の下では、新しい技術の導入によって雇用が脅かされたり、既存のスキルが無駄になることへの懸念が生まれやすいです。

・意思決定の遅さ

日本の企業は、稟議制度など、意思決定に時間がかかる傾向があります。迅速な対応が求められる生成AIの活用においては、この点がネックとなる可能性があります。

・「匠の技」の重視

日本には、熟練した職人の技術や経験を重視する文化があります。AIによる自動化に対して、「人間の仕事が奪われる」という抵抗感が生まれやすい土壌があります。

・セキュリティ意識の高さ

日本企業は、情報セキュリティに対して非常に高い意識を持っています。生成AIのセキュリティリスクに対する懸念も、導入を躊躇させる要因の一つです。

●技術的な課題

・日本語対応の遅れ

生成AIは、英語を中心に開発が進められており、日本語の対応が遅れている場合があります。日本語の精度や表現力に課題があるため、業務での活用が難しいケースもあります。

・導入、運用コスト

生成AIの導入や運用には、高額な費用がかかる場合があります。特に、大規模なシステムを構築する場合には、初期投資やランニングコストが大きな負担となります。

・人材不足

生成AIを適切に活用できる人材が不足しています。AIエンジニアやデータサイエンティストなど、専門的な知識やスキルを持つ人材の育成が急務となっています。

・法整備の遅れ

生成AIに関する法整備が追いついていないことも、企業が活用を躊躇する要因の一つです。著作権、個人情報保護、責任の所在など、様々な法的課題が未解決のままです。

本書の目的は、生成AIのリスクを正しく理解し、安全かつ効果的に活用してもらうことです。

個人が日常的に生成AIを使用する際には、ほとんどリスクはないと感じています。実際、その通りです。しかし、企業での利用となると話は別です。

責任の所在が曖昧であったり、スタッフのAIに対する理解が十分でなかったりすると、企業としてリスクが発生する可能性があります。また、そのリスクを適切に管理するのが難しいという現実もあります。

そこで、本書では生成AIの適切な活用方法を解説し、AIの得意な部分と不得意な部分を社内で正しく理解することの重要性を伝えます。

さらに、リスクを分散しながらAIを活用する方法についても詳しく説明することで、企業が生成AIを安全に活用できるようになることを目指しています。

本書を読めば、生成AIを活用する際にどのようなリスクがあるのかを理解できるだけでなく、それらのリスクにどう対処すればよいのかについても明確に把握できるようになるはずです。

生成AIには、様々な種類があります。

代表的なものとして、文章生成AI、画像生成AI、音声生成AI、映像生成AIが挙げられます。また、それぞれの分野ごとに複数のサービスが存在し、現在も次々と新しいシステムやツール、サービスが登場しています。生成AIのリスクを理解する上で最も重要なことは、「どのAIを使用するのか」を適切に選択することです。

例えば、文章生成AIの場合、利用するサービスを慎重に選び、さらにそのサービス内の設定を適切に調整することで、セキュリティを強化することができます。

画像生成AIに関しては、使用されるトレーニングデータの信頼性が特に重要です。著作権を侵害していないデータが用いられているかどうかを確認することが、適切なAI活用の鍵となります。

これは画像だけでなく、音声や映像生成AIにも共通する課題であり、著作権リスクが発生しやすい分野でもあります。そのため、どの生成AIを使用するのかを社内で明確に決定することが、リスクを抑える上で不可欠です。

さらに、使用する生成AIを決めたとしても、設定次第でセキュリティレベルは大きく変わります。本書では、セキュリティやリスクが少ない生成AIを紹介するとともに、それらをどのように設定し、安全に活用すればよいのかについて詳しく解説しています。

本書が、貴社のAI導入を成功に導き、持続的な成長を実現するための「羅針盤」となることを願っています。

七里信一

CONTENTS

第3章

【データ・プライバシー関連リスク】
情報漏洩とその対策

第4章

【コンテンツ関連リスク】
生成物の責任問題とその対策

第5章

【モデル・運用関連リスク】
AIシステムの安定性と信頼性確保

第6章 【共通対策①】組織全体で構築するAIガバナンスと推進体制

第7章 【共通対策②】インシデント発生！その時のための対応フロー

AI技術の未来展望
これからAIはどう進化していくのか？

変化の時代を生き抜くために
企業の次の一手とマインドセット

第1章

知っておくべき生成AIの基礎

「生成AI（ジェネレーティブAI）」――この言葉を聞かない日はないほど、私たちのビジネスを取り巻く環境は急速に変化しています。「何かすごいらしい」「使わないと乗り遅れるのでは？」と感じつつも、具体的に何をどうすれば良いのか、どんな影響があるのか、そしてリスクはないのか…多くの経営者、管理職、そして現場で働く従業員の皆さんが、期待と不安の入り混じった思いを抱えているのではないでしょうか。

本章は、そのような皆様が生成AIという新しい時代の波を理解し、自社のビジネスや自身の業務にどう関わってくるのか、その本質を掴むための「最初の地図」となることを目指します。難しい技術の話は最小限にとどめ、「結局、何なの？」「私たちの仕事はどう変わるの？」という素朴な疑問に、ビジネスの視点から分かりやすくお答えしていきます。さあ、一緒に生成AIの世界を探検しましょう。

1.1 生成AIとは？

新しいものを「創り出す」AIの登場

まず、生成AIがこれまでのAIと何が違うのか、そのポイントを押さえましょう。

「分析するAI」から「創造するAI」へ

これまで私たちが主に接してきたAIは、例えば、大量のデータから売上を予測したり、写真に写っている人物の顔を認識したり、迷惑メールを仕分けたりといった、情報を「分析」したり「識別」したりするのが得意でした。いわば、非常に優秀な分析官や仕分け係のような存在です。

一方、生成AIは、その名の通り、新しいものを「生成する」、つまり『創り出す』ことができるAIです。まるで人間のクリエイターのように、新しい文章、企画書、報告書、デザイン案、さらにはプログラムコードまで、ゼロから、あるいは与えられた指示に基づいて創り出す能力を持っています。これは、ビジネスの現場に、強力なアイデアマンであり、かつ高速な実務アシスタントが登場したような、大きな変化なのです。

なぜ今、こんなに注目されているのか

このような「創造するAI」は、AI研究の進化、特に「ディープラーニング（コンピュータが大量のお手本から自ら学習する技術）」の目覚ましい進歩によって、ここ数年で一気に実用的なレベルに達しました。

特に文章を扱う分野で革命を起こしたのが、2017年に登場した「Transformer（トランスフォーマー）」という技術（AIの設計図のようなもの）です。これにより、AIは言葉の表面的なつながりだけでなく、文章全体の文脈やニュアンスを非常に深く理解できるようになりました。

この技術をベースにして開発されたのが、「大規模言語モデル（LLM）」です。これは、いわば「膨大なビジネス書やウェブ情報を読み込み、人間レベルの言語能力と知識を身につけた超優秀なアシスタント」のような存在です。

そして2022年頃から、このLLMを使ったOpenAI社の「ChatGPT」などの対話型AIサービスや、画像生成AI「Stable Diffusion」「Midjourney」などが次々と登場し、誰でもその驚くべき能力に触れられるようになったことで、世界中で爆発的な注目を集め、社会現象とも言えるほどの急速な普及が始まったのです。

1.2 何ができるのか？

主要な生成AIの種類と代表的なツール例

では、具体的に生成AIはどのようなことができるのでしょうか？　生成するコンテンツの種類によって、いくつかのタイプがあります。ここでは代表的な種類と、現在（執筆時点）よく知られているツールの例、そしてビジネスでの活用イメージを見ていきましょう。

【テキスト生成AI】（文章作成・要約・アイデア出しなどを得意とするAI）

■できることの例

人間と自然な対話をする、長い文章を要約する、メールやレポートの下書きを作る、企画のアイデアを出す、多言語に翻訳する、簡単なプログラムコードを書く、など。

■ビジネスでの活用例

議事録作成補助、資料作成の効率化、マーケティングコピー案作成、顧客問い合わせへの一次対応（チャットボット）、ソフトウェア開発支援など、非常に幅広い業務で活用が始まっています。

■代表的なツール例

- **ChatGPT（OpenAI社）：**対話型AIの普及を牽引した代表的なサービス。汎用性が高く、様々な用途に利用されています。
- **Gemini（Google社）：**Googleの最新AI技術を活用したモデル・サービス。検索エンジンとの連携などが特徴。
- **Copilot（Microsoft社）：**Microsoft 365（Word、Excel、PowerPointなど）と連携し、文書作成やデータ分析を支援する機能を提供。
- **Claude（Anthropic社）：**安全性や倫理性に配慮した設計を特徴とする対話型AI。

※これらはあくまで一部であり、他にも多数のテキスト生成AIサービスが存在します。

【画像生成AI】（新しい絵や写真のような画像を創り出すAI）

■できることの例

「青い空とヨット」のような言葉（プロンプト）からオリジナルの画像を生成する、写真の一部を修正・編集する、画像の解像度を上げる、など。

■ビジネスでの活用例

ウェブサイトや広告用の画像素材作成、製品デザイン案の作成、プレゼン資料の図版作成など、クリエイティブ分野での活用が進んでいます。

■代表的なツール例

- **Stable Diffusion：**オープンソースとしても提供されており、多くの派生サービスが存在します。カスタマイズ性が高いのが特徴。
- **Midjourney：**高品質で芸術的な画像の生成に定評があります。デザイン分野などで人気。
- **DALL-E（OpenAI社）：**ChatGPTと同じOpenAI社が開発。テキストからの画像生成能力が高い。
- **Imagen（Google社）：**Googleが開発する高画質な画像生成モデル。
- **Firefly（Adobe社）：**Photoshopなど同社のクリエイティブツールとの連携や、学習データの著作権への配慮を特徴としています（第8章参照）。

【音声生成AI】（人間の話し声を作り出すAI）

■できることの例

テキストを人間のように自然な声で読み上げる、特定の人の声色を再現する（ボイスクローニング）、外国語のテキストをその言語の自然な音声で読み上げる、など。

■ビジネスでの活用例

動画のナレーション作成、オーディオブック制作、電話自動応答システム、多言語対応の音声案内、バーチャルアシスタントの声など。

■代表的なツール例

ElevenLabsやHeyGenなどが、リアルな音声合成サービスとして広く知られています（2025年4月時点）。

【動画生成AI】（短い動画を創り出すAI）

■できることの例

テキストや画像から短い動画クリップを生成する、既存動画のスタイルを変換する、など（※この分野は現在、急速な技術開発が進んでいます）。

■ビジネスでの活用例

SNS用の短いプロモーション動画の作成、製品紹介動画のコンセプト作成、動画編集作業の効率化などへの応用が期待されています。

■代表的なツール例

RunwayやPika Labsなどが、テキストや画像からの動画生成サービスとして注目を集めています。OpenAI社の「Sora」のように、非常に高品質な動画生成が可能なモデルも発表されていますが、執筆時点では利用が限定的です（2025年4月時点）。

【音楽生成AI】（新しい曲や効果音を創り出すAI）

■できることの例

「明るくリズミカルなBGM」「ホラー風の効果音」といった指示に基づいて、オリジナルの楽曲やサウンドを生成する。

■ビジネスでの活用例

プレゼンテーションや社内動画用の著作権フリーBGM作成、ゲームやアプリの効果音制作など。

■代表的なツール例

Suno AIやUdioといったサービスが、テキスト指示から楽曲を生成する能力で注目されています。Stability AI社の「Stable Audio」なども知られています（2025年4月時点）。

【重要】ツール利用にあたっての注意点

ここで挙げたのは、あくまで現時点（2025年4月）でよく知られているツールのほんの一例です。生成AIの世界は進化が非常に速く、日々新しいツールが登場し、既存ツールの機能や性能も頻繁にアップデートされています。

また、各ツールには利用規約があり、データの取り扱い（入力した情報が学習に使われるか等）、生成物の権利（誰のものになるか）、商用利用の可否、料金体系などがそれぞれ異なります。

したがって、実際にこれらのツールを業務で利用する際には

下記の点が極めて重要になります。

- **常に最新の情報を確認すること**
- **自社の利用目的に合ったツールを慎重に比較検討すること**
- **利用規約を必ず確認し、内容を理解・遵守すること**

安易なツール選択や利用規約の無視は、思わぬリスク（情報漏洩、著作権侵害など）につながる可能性があるため、十分にご注意ください（具体的なリスクと対策は第3章以降で詳しく解説します）。

1.3 生成AIの得意なこと、苦手なこと

「賢い相棒」との上手な付き合い方

非常に便利な生成AIですが、決して万能ではありません。その能力を最大限に引き出し、同時にリスクを避けるためには、AIの「得意なこと」と「苦手なこと（注意点）」を正しく理解しておくことが重要です。AIを「100％信じる」のではなく、「賢い相棒」として捉え、上手な付き合い方を学びましょう。

●生成AIが得意なこと – 頼りになる場面 –

- **大量の情報の整理、要約：**長文のレポートや議事録から要点だけを抜き出す。
- **文章の壁打ち、構成案作成：**何か書き始める際の「最初の叩き台」を作る。

・**多様なアイデア出し：**自分だけでは思いつかないような、様々な切り口や表現の案を大量に出す。
・**定型的な文章、コードの生成：**メール返信、簡単な報告書、基本的なプログラムコードなど、ある程度パターンが決まっているものの作成。
・**翻訳：**多言語間の翻訳を高速に行う。
・**知識の検索（ただし注意が必要）：**幅広い分野の知識を持っており、質問に答えてくれる（ただし、次の「苦手なこと」も参照）。

●生成AIが苦手なこと・注意すべき点 – 人間が補うべき場面 –

・**事実確認：**AIは、学習データにはない情報や、間違った情報を、さも事実であるかのようにもっともらしく生成する（ハルシネーション）ことがあります。AIの回答を鵜呑みにするのは非常に危険です。必ず人間がファクトチェック（事実確認）を行う必要があります（平気で嘘をつくことがある！）。
・**倫理的・常識的な判断：**何が道徳的に正しく、何が社会的に許容されるかといった、複雑な倫理観や常識に基づいた判断は苦手です。差別的な表現や、配慮に欠ける内容を生成してしまう可能性もあります。
・**真の創造性・独創性：**AIは学習データにあるパターンを組み合わせるのは得意ですが、人間のように全く新しい概念や、深い洞察に基づく独創的なアイデアを生み出すことは

まだ難しいと言われています。

- **最新情報への追随：**LLMの学習データは、ある特定の時点までのものが使われていることが多く、リアルタイムの情報や、ごく最近の出来事については正確に答えられない場合があります。
- **指示の微妙なニュアンス理解：**指示（プロンプト）の仕方が曖昧だと、意図と違う出力をしてしまうことがあります。望む結果を得るには、明確で具体的な指示を出す工夫（プロンプトエンジニアリング）が必要です。
- **ブラックボックス性（なぜそう答えたのか）：**AIがなぜ特定の答えを出したのか、その思考プロセスや判断根拠を、AI自身や開発者でさえ完全に説明できない場合があります。この「ブラックボックス」性は、結果の信頼性を評価する上で注意が必要です。

このように、生成AIは素晴らしい能力を持っていますが、限界や注意点も多く存在します。**AIに全てを任せるのではなく、その得意な部分を最大限に活用しつつ、苦手な部分は人間が責任を持ってチェックし、補うという「人間とAIの協働」の姿勢**が、これからのビジネスでは不可欠になるでしょう。

1.4 ビジネスへの本当の価値

なぜ今、取り組むべきなのか？

生成AIの導入は、単なるコスト削減や業務効率化に留まらず、企業の競争力や成長戦略そのものに関わる重要な意味を持っています。なぜ今、多くの企業が生成AIに注目し、取り組みを始めているのでしょうか？

■桁違いの生産性向上とコスト削減

これまで人間が数時間、あるいは数日かけていた作業（資料作成、情報収集、データ入力など）を、AIが数分、数秒で完了させることが可能になるかもしれません。これにより、人件費を含むコストの大幅な削減と、組織全体の生産性の劇的な向上が期待できます。これは、限られたリソースで最大の成果を出すことが求められる現代のビジネスにおいて、非常に大きな魅力です。

■新しいビジネスチャンスの創出

AIの支援によって、これまで時間やコスト、あるいは技術的な制約から実現できなかった、全く新しい製品やサービスの開発が可能になるかもしれません。また、顧客別の究極のパーソナライズを実現し、新たな顧客価値を提供することも考えられます。これは、市場での差別化を図り、新たな収益源を確保するチャンスです。

■競争優位性の確立（あるいは、乗り遅れリスクの回避）

競合他社が生成AIを活用して、より速く、より効率的に、より革新的なビジネスを展開し始めた場合、何もしなければ相対的に競争力を失い、市場から取り残されてしまうリスクがあります。生成AIへの取り組みは、もはや先進企業だけのものではなく、あらゆる企業にとって競争上、避けて通れない課題となりつつあります。重要なのは、「導入するか否か」ではなく、「いかに賢く、早く、自社に合わせて活用するか」です。

■従業員の働きがい向上

単調な作業や時間のかかる付帯業務から解放された従業員は、より創造的で、思考力が求められ、達成感のある仕事に集中できるようになります。これは、従業員のモチベーションやエンゲージメントを高め、ひいては人材の定着や組織全体の活性化にもつながる可能性があります。

生成AIは、使い方次第で、コスト削減、売上向上、イノベーション促進、従業員満足度向上といった、**企業経営における様々な課題に対する強力な解決策**となり得るのです。

1.5 なぜ導入をためらうのか？

日本企業が直面する壁と本書の役割

これだけの大きな可能性がありながらも、特に日本の多くの企業では、生成AIの本格導入に慎重な声が多いのも事実です。皆さんの会社でも、以下のような理由で議論が停滞したり、導入に二の足を踏んだりしている状況はないでしょうか？

■リスクへの懸念 —「何かあったら怖い」

- **情報漏洩**：会社の機密情報やお客様の情報が漏れたらどうしよう…。
- **著作権侵害**：AIが作ったものが、誰かの権利を侵害していたら訴えられるかも…。
- **誤情報、偽情報**：AIの間違った情報を信じて、大きな失敗をしないだろうか…。
- **偏見、差別**：AIが不適切な発言をして、会社の評判を落とさないだろうか…。
- **責任問題**：AIが問題を起こした場合、誰が責任を取るのだろう…。

■組織・文化の壁 —「ウチの会社では難しいかも」

- **新しいことへの抵抗感**：これまでやってきたやり方を変えたくない…。
- **失敗を恐れる空気**：もし導入して失敗したら、責任問題に

なるのでは…。

- **意思決定のスピード感：**関係部署が多くて、導入を決めるまでに時間がかかりすぎる…。
- **今の仕事への影響：**AIに仕事を取られてしまうのでは…という従業員の不安。

■技術・リソースの課題 ―「そもそも使えるの？」

- **コストの問題：**導入や利用にお金がかかりそう…費用対効果が見えない…。
- **人材の不足：**AIを使いこなせる社員がいない…専門家も採用できない…。
- **日本語の性能：**本当にビジネスで使えるレベルの日本語なの…？

■法制度の不確実性 ―「ルールが分からない」

- **AI利用に関する法律やルールがまだはっきり決まっていない**ため、何がOKで何がNGなのか分からず、手探り状態になってしまう。

これらの懸念や課題は、どれももっともなことです。決して無視して良いものではありません。

本書は、まさにこれらの課題や不安を抱える企業の皆様のために書きました。

本書の役割は次の3点です。

1.生成AIに潜む様々なリスクを、**過度に恐れるのではなく、**

正しく、具体的に理解すること。

2. それぞれの**リスクに対して、どのような実践的な対策**があり、自社で何ができるのかを知ること。
3. AIの**能力と限界**（得意なこと・苦手なこと）を見極め、**安全かつ効果的に活用するための具体的な方法**を見出すこと。

これらを通じて、皆様が抱えるAI導入への漠然とした不安を具体的な知識と行動計画へと変え、自信を持って最初の一歩を踏み出すためのお手伝いをします。リスク管理は、AI活用を諦める理由ではなく、AI活用を成功させるための前提条件なのです。

次章からは、いよいよ企業が直面する主要なリスクについて、その詳細と具体的な対策を掘り下げていきます。本章でAIの基本的な姿を掴んだ上で、ぜひ安心して次のステップへ進んでください。

第2章

生成AI活用に伴うリスクの全体像と向き合い方

第1章では、生成AIがいかに私たちの仕事を変える可能性を秘めているかを見てきました。「夢のようなツールだ！」と感じられたかもしれませんね。しかし、どんなに強力なツールでも、使い方を誤れば思わぬ怪我をする可能性があるように、生成AIにも注意すべき「リスク」が潜んでいます。

「リスク」と聞くと、どうしても身構えてしまいがちですが、心配はいりません。大切なのは、やみくもに怖がるのではなく、まず「どんな種類の危険があるのか？」を正しく知ることです。そして、「その危険はどのくらい深刻なのか？」を見極め、「どうすれば安全に使えるのか？」という対策の方向性を考えること。これが、AIという新しい時代の波を賢く乗りこなすための基本姿勢となります。

本章では、まず企業がAIを活用する上で特に気をつけたい「主なリスクの種類」にはどんなものがあるのか、その全体像を分かりやすくご紹介します。次に、たくさんあるリスクの中から「どれを優先的に対処すべきか」を見極めるための考え方（リスク評価）と、その際に本書全体で参考として用いる「評価の目安（指標）」について詳しくご説明します。さあ、AIのリスクと上手に付き合うための第一歩を踏み出しましょう。

2.1 主要なリスク領域の概観

「こんなはずじゃなかった…」を防ぐために

生成AIを使うことで起こりうる問題は様々ですが、ここでは特に企業活動に大きな影響を与えかねないリスクを、大きく4つのカテゴリーに分けて整理してみましょう。皆さんの会社や業務にどんな関係があるのか、イメージしながら読んでみてください（これらのリスクの詳細と具体的な対策は、第3章から第5章でじっくり解説します）。

情報漏洩リスク：「会社の秘密や個人情報が外に漏れてしまう！」

■どんなリスク？

従業員がAIに質問や作業を依頼する際に、うっかり会社の重要な内部情報（新製品のアイデア、顧客リスト、会議の議事録など）や、お客様・社員の個人情報（氏名、連絡先、人事評価など）を入力してしまい、それがAIサービスを通じて外部に漏れてしまう危険性です。AIが過去に学習した情報から、意図せず機密情報が出力されてしまう可能性もゼロではありません。

■どんな懸念？

会社の競争力の源泉である情報がライバルに知られたり、大切なお客様や従業員のプライバシーが侵害されたりすれば、会社の信用は地に落ち、法的な責任を問われ、事業継続すら危う

くなる可能性があります。「これくらい大丈夫だろう」という油断が、取り返しのつかない事態を招くかもしれません。

生成コンテンツ責任リスク：「AIが作ったものが問題を起こす！」

■どんなリスク？

AIが自動で作り出した文章、画像、プログラムコードなどが、その「中身」によって問題を引き起こす危険性です。

■どんな懸念？

- **著作権侵害：**AIが作ったデザインが、どこかの会社のロゴにそっくりだった！
- **誤情報・偽情報：**AIチャットボットが、お客様に間違った製品情報を伝えてしまい、大クレームに！
- **バイアス・差別助長：**AIが書いた社内報の記事に、特定の属性の人々を不快にさせる偏った表現が含まれていた！

これらは、会社の評判を大きく傷つけ、法的な紛争や社会的な非難につながる可能性があります。

AI過信リスク：「AIを信じすぎたら、とんでもないことに！」

■どんなリスク？

AIが出す答えや提案が非常に優れているように見えるため、ついそれを鵜呑みにしてしまい、人間が本来行うべき確認や、常識・倫理観に基づいた判断を怠ってしまう危険性です。AI

の判断プロセスが不透明（ブラックボックス）であることも、このリスクを高めます。

■どんな懸念？

「AIの市場分析を信じて新商品を発売したら、全く売れなかった」「AIの診断支援を過信して、患者さんの重要な訴えを聞き逃してしまった」「問題が起きた時、『AIがやったことなので…』と責任逃れをしてしまう」など、誤った経営判断や業務上の重大なミス、説明責任の欠如といった事態を招きかねません。

AIシステム停止リスク：「頼みのAIが動かない！　業務がストップ！」

■どんなリスク？

業務で利用しているAIサービスや、それを動かすための社内システム、ネットワークなどが、技術的なトラブル、サイバー攻撃、あるいはサービス提供元の都合などで突然停止し、利用できなくなってしまう危険性です（可用性の問題）。

■どんな懸念？

「顧客対応AIが止まって電話がパンクした」「AIで在庫管理していたが、システムダウンで何がいくつあるか分からなくなった」「開発に必須のAIツールが、予告なくサービスを終了してしまった」など、AIへの依存度が高いほど、業務が完全にストップし、事業継続に深刻な影響が出ます。

リスクは連鎖する：「ドミノ倒し」にも注意

これらのリスクは、それぞれ別々に起こるだけでなく、互いに関連し、連鎖することにも注意が必要です。例えば、AIシステムへの攻撃（AIシステム停止リスク）が原因で、顧客情報が漏洩する（情報漏洩リスク）かもしれません。あるいは、AIへの過信（AI過信リスク）が、誤った情報の拡散（生成コンテンツ責任リスク）につながることも考えられます。一つのリスク対策が、別のリスクを引き起こす可能性すらあります。リスクを考える際は、このように全体的なつながり（ドミノ倒しのような影響）も意識することが大切です。

2.2 リスク評価の視点と本書で用いる指標

「どのリスクから手をつけるべき？」

さて、様々なリスクがあることが分かりましたが、これら全てに一度に完璧な対策をするのは現実的ではありません。「時間は有限、予算も有限、人も限られている…一体どこから手をつけるべきか？」これが経営者や管理職の皆さんの正直な悩みではないでしょうか。そこで重要になるのが、「リスク評価」、つまり「どのリスクが自社にとってより深刻で、優先的に対策すべきか」を見極めることです。これは、限りあるリソースを最も効果的な対策に集中させるための、戦略的な判断と言えます。

リスクの「大きさ」や「深刻度」を測るための基本的な考え方は、シンプルに以下の2つの質問に答えることです。

「それって、どれくらいの確率で起こりそう？」（発生可能性/Likelihood）
「もし起こったら、どれくらい大変なことになる？」（影響度/Impact）

例えば、「会社の重要書類をデスクに置きっぱなしにする」というリスクを考えてみましょう。「発生可能性」は、社員の意識次第で高くも低くもなりますね。「影響度」は、置きっぱなしにした書類がただのチラシなら軽微ですが、もし最重要の顧客リストだったら、会社が傾くほどの影響があるかもしれません。

このように、「発生可能性」と「影響度」の2つの軸で考えることで、リスクを客観的に評価し、優先順位をつけることができます。一般的には、「発生可能性が高く、かつ、影響度も大きい」リスク（例えば、頻繁に起こりそうで、起こったら会社が大損害を受けるようなリスク）から、最優先で対策を検討していくことになります。

本書における評価指標について

今後の章（特に第3章～第5章）で、様々なリスクに対する具

体的な「対策」を解説する際に、読者の皆様がそれぞれの対策の特徴をより掴みやすく、比較検討しやすくするために、参考情報として以下の3つの視点から評価の目安を示していきます。

リスクの「発生可能性」

そのリスクが、一般的な企業環境において、どの程度の確率や頻度で発生しうるかを示す目安です。対策の必要性を判断する材料となります。

A（非常に高い）：特に対策を講じなければ、ほぼ確実に起こる、または頻繁に起こると考えられるレベル。

B（高い）：起こる可能性が高い、または無視できないと考えられるレベル。積極的な対策が推奨されます。

C（中程度）：起こる可能性も、起こらない可能性もあると考えられるレベル。状況に応じた対策の検討が必要です。

D（低い）：通常は起こらないが、特定の状況下では起こりうると考えられるレベル。油断は禁物です。

E（非常に低い）：起こる可能性はかなり低いと考えられるレベル。ただしゼロではありません。

対策の「導入のしやすさ」

その対策を企業が導入する際に、一般的に必要となるコスト、手間、期間、専門知識などの度合いを示す目安です。対策の実

現可能性を判断する材料となります。

A（非常に高い）：コストも手間もほとんどかからず、すぐに始められるレベル（例：意識改革、簡単な設定変更）。

B（高い）：少しの準備や調整で、既存のものを活用できるレベル（例：既存ツールの設定変更、ガイドライン作成）。

C（中程度）：ある程度の準備が必要。専門家の助けが必要な場合もあるレベル（例：新しいツールの導入、研修プログラムの企画）。

D（低い）：かなりのコストや準備が必要。専門人材や新システムが必要なこともあるレベル（例：大規模システム開発、専門チーム組成）。

E（非常に低い）：莫大なコストがかかる。実現が難しい、または運用が困難なレベル（例：最先端技術の独自開発）。

対策の「効果」

その対策を実施した場合に、対象とするリスクをどの程度低減できるかの期待度合いの目安です。対策の有効性を判断する材料となります。

A（非常に高い）：その対策によって、対象リスクが発生する可能性や影響をほぼ無くせる、または影響を最小限にできることが期待されるレベル。

B（高い）：リスクが発生する可能性や影響を大きく減らせることが期待されるが、完全には無くせないレベル。

C（中程度）：ある程度の効果はあるが、限定的であり、他の対策との組み合わせが重要となるレベル。

【最重要】この評価は「絶対」ではありません！

ここで強調しておきたいのは、これらの評価（A ～ E、A ～ C）は、あくまで本書の解説を分かりやすくするための「目安」であり、一般的な傾向を示すものにすぎないということです。

皆さんの会社の業種（例：金融か製造か）、規模（大企業か中小企業か）、扱っている情報の種類（個人情報が多いか、技術情報がメインか）、AIを何に使うか（顧客対応か、社内分析か）、従業員の皆さんのITスキルや意識の高さなど、具体的な状況によって、リスクの起こりやすさも、対策の導入のしやすさや効果も、全く違ってきます。

ですから、本書に示されている評価を「正解」だと思わず、必ず「自社の場合はどうだろうか？」と立ち止まって考え、自社にとってのリスクの重要度や、対策の優先順位を判断してください。この指標は、その際の思考の整理や比較検討を助ける「ものさしの一つ」としてご活用いただければと思います。

2.3 各リスクへの対策アプローチ

「技術」と「人・組織」の両輪で

リスクを評価し、優先順位をつけたら、次はいよいよ具体的な対策の検討です。生成AIのリスクに対する対策は、大きく分けて2つのアプローチがあります。順番に見ていきましょう。

技術的対策（Technical Controls）

“システムの機能で守る”（車の安全装置のようなもの）

■考え方

AIシステム自体や、関連するITツール、ネットワークなどに、セキュリティ機能や制御メカニズム（いわば「安全装置」）を組み込むことで、リスクの発生を直接的に防いだり、検知したり、影響を抑えたりする方法です。

■具体例

- 不正なアクセスをブロックする仕組み（家の鍵や警備システム）
- データを盗まれても読めなくする技術（暗号化）
- AIへの不適切な指示を検知するフィルター
- 問題発生をいち早く知らせるアラートシステム（車で言えば、エアバッグ、自動ブレーキ、ABSなど）

組織的対策（Organizational/Administrative Controls）

“会社のルールや人の力で守る”（交通ルールや運転技術のようなもの）

■考え方

技術だけではカバーしきれない部分を、会社のルール作り、責任体制の整備、従業員の教育、日々の運用プロセスなどを通じて管理していく方法です。特に、人間の判断ミスや不注意、あるいは倫理的な配慮が関わるリスクに対して重要になります。

■具体例

・全員が守るべきAI利用のガイドライン（交通ルール）作成
・リスクやルールに関する従業員研修（運転免許講習や安全運転教育）の実施
・誰が責任を持つかを明確にする推進体制の構築
・AIの提案を人間が必ず確認する業務フロー（安全確認の手順）の設計
・定期的なリスクチェック（車両点検や健康診断）の実施など

大切なのは「組み合わせること」

ここで非常に重要なのは、「技術的対策」と「組織的対策」のどちらか一方だけでは不十分であり、両方をバランス良く組み合わせて、複数の防御壁（多層防御）でリスクに備える必要がある、ということです。

車の例で考えてみてください。どんなに最新の安全装置（技術的対策）が付いた車でも、運転する人が交通ルールを守らず、危険な運転（組織的対策＝ルールや人の意識・行動の欠如）をすれば、事故は起こってしまいますよね。逆に、どんなに模範的なドライバーでも、ブレーキが効かない車（技術的な欠陥）に乗っていたら、事故を防ぐことはできません。

AIの安全な活用も全く同じです。優れた技術的対策（システムの力）と、それを支えるしっかりとした組織的対策（ルール、体制、人の意識・行動）の両輪があってこそ、リスクを効果的に管理し、安心してAIのメリットを享受することができるのです。

以降の章では、この考え方に基づき、具体的なリスク対策を解説していきます。第3章から第5章では、各リスク領域に特有の技術的対策や運用上の注意点を中心に解説し、そして第6章では、複数のリスクに共通して重要となる組織的対策（AIガバナンス、ガイドライン、教育など）をまとめて、体系的に掘り下げていきます。

【章のまとめ】

本章では、「なぜリスクを理解することが重要なのか」から始まり、生成AI活用における主要な4つのリスク領域、リスクの大きさを評価するための基本的な考え方と本書で用いる評価指標、そして対策の2つの主要なアプローチ（技術的・組織的）

について、解説しました。

これらの知識は、皆さんがこれからAIリスクと向き合い、具体的な対策を検討していく上での「土台」となるはずです。リスクを知ることは、怖がることではなく、AIと賢く付き合うための第一歩です。リスクは確かに存在しますが、正しく理解し、優先順位をつけ、適切な対策を講じれば、それは十分に管理可能なものなのです。

さあ、次章からは、いよいよ個別のリスクについて、その詳細な内容と具体的な対策の世界へと踏み込んでいきましょう。

第3章

【データ・プライバシー関連リスク】

情報漏洩とその対策

さて、ここからはAI活用における具体的なリスクと、その対策について見ていきましょう。最初に取り上げるのは、多くのビジネスパーソンが最も気にされているであろう「情報漏洩リスク」です。AIは便利な反面、使い方を間違えると、会社の命運を左右するような大切な情報が外部に漏れてしまう危険性をはらんでいます。

想像してみてください。

もし、開発中の新製品の情報がライバル会社に漏れたら？
大切なお客様の個人情報が流出してしまったら？　その損害は計り知れません。

本章では、まず「なぜAIを使うと情報が漏れる心配があるのか？」その仕組みを分かりやすく解き明かし、次に「どんな情報が漏れると、どれだけ大変なことになるのか？」を具体的に見ていきます。そして、それを防ぐための「今すぐできる対策」から「本格的な対策」までを、その「導入のしやすさ」と「期待できる効果」の目安とともにご紹介します。最後に、万が一の際の対応ポイントも押さえておきましょう。

3.1 なぜ情報が漏洩するのか？

AI利用の落とし穴を知る

AIを使っているだけで、なぜ情報が漏れる可能性があるのでしょうか？　主な原因は3つあります。

学習データ由来リスク

■AIが「うっかり覚えていた」ことによる漏洩

AIは、開発される際に大量の文章や画像を「学習」します。もし、その学習データの中に、非公開の情報（個人情報や社内情報など）が紛れ込んでいたら、AIがそれを記憶してしまい、後になって、皆さんが質問した際に、ポロッとその情報を喋ってしまう（出力してしまう）可能性がゼロではありません。

■現状は？

最近の主要なAIサービスでは、「皆さんが入力した情報を勝手に学習に使うことはしません」と約束していることが多いです。そのため、このリスクは以前より大きく減っています。しかし、「絶対に安全」とは言い切れません。サービス提供側の管理体制や、AIの仕組み自体の限界もあるため、注意は必要です。

プロンプト入力由来リスク

人が「うっかり教えてしまった」ことによる漏洩が、現状で最も注意が必要な落とし穴です（一番多い！）。

私たち従業員が、AIに仕事を依頼する際に使う「指示文（プロンプト）」の中に、悪気なく、会社の重要な情報や個人情報を書き込んでしまうケースです。

■どんな書き込みに危険性がある？

例えば、「この顧客リスト（氏名・連絡先付き）を地域別に分類して」「この社外秘の議事録を要約して」「このエラーが出ているプログラムコード（自社の重要資産）を見て」といった指示です。

■入力したらどこへ？

たとえAIの学習に使われなくても、入力した情報はAIサービスを提供している会社のコンピュータ（サーバー）に送られて処理されます。その過程で記録が残ったり、サービス提供会社の管理が甘かったりすると、そこから情報が漏洩する可能性があります。特に海外のサーバーで処理される場合は、日本の法律が及ばないリスクも考慮する必要があります。

■なぜ起こる？

忙しい業務の中で「これくらいなら大丈夫だろう」と安易に考えてしまったり、そもそも「何が会社の機密情報なのか」を従業員がしっかり理解していなかったりすることが原因です。

コピー&ペーストの操作ミスも頻繁に起こりがちです。

■対策の方向性

まずは「機密情報や個人情報は絶対に入力しない」という社内ルールを作り、それを全員が理解し守ることが基本です（詳細は第6章）。加えて、入力内容を自動チェックするような技術的な支援ツール（詳細は第9章）の活用も考えられます。

不正アクセス・内部不正由来リスク

システムが「外部から、あるいは内部から破られた」ことによって情報が漏洩してしまうケースです。

■外部からの攻撃

これは従来のITシステムと同じ脅威ですが、AIシステムや、AIが利用するデータが保管されている会社のサーバー自体が、ハッカーなどからサイバー攻撃を受け、情報が盗み出されるリスクです。AIを狙った新しい攻撃手法も登場しています（詳細は第5章）。

■内部からの漏洩

会社の内部事情を知る従業員や、一時的に関わる委託先の担当者が、故意に情報を持ち出したり、あるいは不注意なミス（例：アクセス権限の設定ミス）によって情報を漏洩させてしまったりするリスクも、残念ながら存在します。これらの原因を知っておくことが、適切な対策を考える上での第一歩となります。

3.2 どんな情報が漏れると大変か？

具体的なシナリオと深刻な影響

では、実際に情報が漏洩すると、どのような問題が起こるのでしょうか？「プライバシー侵害（個人情報の漏洩）」と「機密情報漏洩（会社の秘密情報の漏洩）」の２つのケースで、身近なシナリオと、その深刻な影響を見ていきましょう。

プライバシー侵害（個人情報の漏洩）

「お客様や従業員の情報を漏らしてしまった！」

■身近なシナリオ例

- 問い合わせ対応のAIが、Aさんの質問に答える中で、Bさんの名前や購入履歴をうっかり表示してしまった。
- 人事担当者が、社員の評価シートをAIで分析しようとして、誤ってファイルを外部に送信してしまった。
- 利用していたAIサービスのサーバーが攻撃され、登録していたユーザーのメールアドレスとパスワードが流出した。

■経営を揺るがす影響

- **法律違反と厳しいペナルティ：**個人情報保護法では、漏洩が起きた場合、国（個人情報保護委員会）への報告と本人への通知が義務付けられています（詳細は3.5参照）。対応を

誤れば、高額な罰金だけではなく、事業停止命令が出される可能性すらあります。

■高額な損害賠償

被害者から集団訴訟を起こされ、莫大な損害賠償金の支払いが必要になるケースもあります。

■会社の信用失墜

「あの会社は情報をちゃんと管理できない」という評判は、顧客離れや取引停止を招き、ブランドイメージを回復困難なまでに傷つけます。

■個人の人生への影響

漏れた情報が悪用されれば、なりすまし、詐欺、ストーカー被害など、お客様や従業員の人生を狂わせる深刻な事態を招きかねません。

機密情報漏洩（会社の秘密情報の漏洩）

「会社の“虎の子”が盗まれた！」

■身近なシナリオ例

・開発チームがAIで設計図を検討中、その情報が競合に漏れて類似品が先に発売された。
・経営会議の議事録をAIで要約させたら、その内容（例：リストラ計画）が社内外に広まって大混乱になった。
・営業担当者がAIで提案書を作成中、顧客との価格情報がプロンプトから漏洩した。

・自社独自の技術ノウハウや製造方法が漏洩し、安価な模倣品が出回ってしまった。

■事業継続に関わる影響

・**競争力の源泉喪失：**他社にはない独自の技術や情報が、会社の強みの源泉です。それが失われれば、市場での優位性を失い、業績悪化に直結します。

・**契約違反と訴訟：**取引先との秘密保持契約（NDA）に違反すれば、損害賠償請求や取引打ち切りのリスクがあります。

・**莫大な経済的損失：**研究開発費が無駄になる、株価が下落する、訴訟費用がかかるなど、会社の財政基盤を揺るがすほどの経済的損失につながる可能性があります。

・**ステークホルダーからの信頼失墜：**取引先、株主、金融機関など、会社の存続を支える関係者からの信頼を失えば、資金調達や事業運営そのものが困難になります。

個人の情報も、会社の情報も、一度漏洩してしまえば取り返しがつきません。だからこそ、「漏洩させない」ための予防策が何よりも重要なのです。

3.3 技術的対策

システムの力で情報を守る具体的な方法

情報漏洩を防ぐためには、システムの機能を使って守る「技術的対策」が有効です。ここでは主な対策について、それが「何を」「なぜ」「どのように」行うものなのか、そして導入の手間や期待できる効果の目安（評価指標）と共に解説します（第2章で説明した評価指標を参考にしてください）。

アクセス制御（Access Control）

情報への「門番」を強化する

■目的

必要な人だけが必要な情報にアクセスできるように制限し、不正アクセスや内部からの意図しない情報閲覧・持ち出しを防ぎます。

■具体的な方法

- **権限設定の最小化：**「とりあえず全員アクセス可能」ではなく、「この業務にはこの情報だけが必要」と考え、役職や担当業務に応じてアクセスできる範囲を必要最小限に絞ります。
- **認証の強化：**ID・パスワードだけでなく、スマートフォンへの確認コード送信や指紋認証など、複数の方法で本人確

認を行う**多要素認証（MFA）**を導入し、なりすましによる侵入を防ぎます。

- **アクセスの監視：**誰が、いつ、どの情報にアクセスしたかの記録（ログ）を取り、不審な動きがないかチェックする体制も有効です（監査証跡）。

【評価目安】

・導入のしやすさ：B（高い）

多くのシステムに基本的な機能は備わっていますが、適切な設計・設定・運用には計画が必要です。少しの準備や調整で、既存のものを活用できるレベルです。

・効果：A（非常に高い）

情報漏洩対策の基本であり、適切に実施すればリスクをほぼ無くせる、または影響を最小限にできる非常に効果的な対策です。

暗号化（Encryption）

情報を「解読不能な暗号」に変える

■目的

データそのものを、たとえ盗まれたとしても中身が分からないように「暗号化」します。機密情報や個人情報を守るための最後の砦です。

■具体的な方法

- **保管データの暗号化：**パソコンやサーバー、クラウド上に保存するファイルやデータベースを暗号化します。OSの

標準機能や専用ツールを使います。

- **通信データの暗号化：**メールやウェブサイトでの情報のやり取り、AIサービスとの通信などを暗号化します（SSL/TLSという技術が一般的です）。
- **鍵の厳重管理：**暗号化には「鍵」が必要です。この鍵が漏れると意味がないため、鍵はパスワードなどで保護し、厳重に管理する必要があります。

【評価目安】

・導入のしやすさ：B（高い）

OSやクラウドの機能を使えば比較的容易な部分もありますが、鍵管理など少しの準備や調整が必要です。

・効果：A（非常に高い）

データが漏洩した場合でも、情報の内容そのものを保護できるため、リスクの影響を最小限にできる非常に効果的な対策です。

入力/出力フィルタリング（Input/Output Filtering）

情報の「関所」でチェックする

■目的

AIに情報を入力する際や、AIが出力する際に、個人情報や機密情報、あるいは不適切なキーワードが含まれていないかをシステムが自動でチェックし、問題があれば警告したり、ブロックしたりします。

■具体的な方法

・**NGワード/パターン検知：**電話番号、メールアドレス形式、マイナンバー形式、社外秘といった特定の「キーワード」や「文字パターン」を事前に登録しておき、それらが含まれていたら検知します。

・**AIによる内容チェック：**より高度な方法として、AI（LLM）自身が文章の文脈を理解し、「これは機密情報っぽいな」と判断して警告する技術もあります。

■注意点

この方法は、巧妙な言い換えや未知の情報パターンを完全に見抜くことは困難です。あくまで補助的な対策と考え、過信は禁物です。

【評価目安】

・導入のしやすさ：A（非常に高い）～ C（中程度）

NGワード設定ならすぐに始められますが、高度なAIチェックはある程度の準備が必要です。

・効果：B（高い）～ C（中程度）

意図しない情報送信を大きく減らせますが、完全ではなく、他の対策との組み合わせが重要です。

匿名化/仮名化（Anonymization/Pseudonymization）とデータ最小化（Data Minimization）

そもそも「危ない情報」を扱わない・減らす

■目的

AIに学習させたり、処理させたりするデータから、個人を特定できる情報や機密情報を、可能な限り最初から「取り除く」または「分からない形に加工する」という、非常に根本的で効果的なアプローチです。

■具体的な方法

- **データ最小化:**「この分析に、本当に氏名や連絡先は必要？」と考え、AIに渡す情報は必要最小限にします。これが最もシンプルで重要です。
- **匿名化：**氏名や住所などの情報を削除したり、「東京都、30代、男性」のように情報を簡略化して、個人が特定できない形に加工します。
- **仮名化：**氏名などを「顧客ID-001」のような固有のIDに置き換えます。IDと氏名の対応表を別途厳重に管理すれば、個人を特定せずにデータを扱えます。
- **差分プライバシー（専門的な技術）：**データ全体に特殊な処理（ノイズ追加）を施し、個人の情報が含まれているかどうかを外部から分かりにくくする高度なプライバシー保護技術です。

■注意点

データを加工しすぎると、分析結果の精度が落ちるなど、データの役に立つ度合い（有用性）が低下する可能性があります。どの情報をどの程度加工するかは、目的とのバランスを見て慎重に判断する必要があります。専門的な知識が求められる場合

もあります。

【評価目安】

・導入のしやすさ：A（非常に高い）～ C（中程度）

データ最小化の考え方はすぐに始められます。高度な匿名化や差分プライバシーはある程度の準備や専門家の助けが必要です。

・効果：A（非常に高い）～ B（高い）

適切に行えば、情報漏洩リスクの根源を断つことができるため、リスクをほぼ無くせる、または大きく減らせる非常に効果的な対策です。

これらの技術的対策は、どれか一つだけを行えば良いというものではありません。自社の状況に合わせて複数を組み合わせ、多層的な防御体制を築くことが重要です。

3.4 データ取り扱いに関する注意点

「情報＝大切な資産」として扱う習慣を

どんなに優れたシステムを導入しても、それを使う私たち人間が情報をぞんざいに扱ってしまっては、情報漏洩は防げません。会社の「情報」を、現金や設備と同じように「大切な資産」として捉え、日々の業務の中で丁寧に扱う習慣を組織全体で身につけることが重要です（具体的な社内ルールは第6章のガイ

ドラインで定めます）。

データを集める・入力する時：「入口」での注意が肝心！

■「入れない」が基本

会社の機密情報、お客様や同僚の個人情報は、原則として外部のAIサービスには絶対に入力しない。これを徹底しましょう。

- **「本当に必要か？」と立ち止まる：**AIに頼む前に、「この情報はAIに入れないとダメ？」「もっと安全な方法はない？」と一呼吸置いて考える癖をつけましょう。
- **チェックと加工を忘れずに：**どうしても必要な場合は、入力内容を必ず再確認し、機密情報や個人情報が含まれていないかチェックします。可能なら、名前をイニシャルにする、会社名を伏せるなど、匿名化・マスキングの一手間を加えましょう（会社のルールに従い、必要なら上司に相談・承認を得てください）。

■「ついでにこれも」はNG

AIに依頼する目的に関係のない情報は、絶対に入力しないようにしましょう。

データを保管する時：「金庫」に入れる意識で

■決められた場所に保管

データは、会社が指定した安全な場所（アクセス制限や暗号化がされたサーバーなど）にだけ保管します。自分のパソコン

のデスクトップや、個人のUSBメモリ、無料のネットストレージなどに無造作に置かないこと。整理整頓と同じです。

■不要になったらすぐ削除

保管する必要がなくなったデータは、いつまでも放置せず、速やかに削除する習慣をつけましょう。データが少なければ、漏洩リスクも減ります。

データを利用・共有する時：「誰に」「何を」渡すか慎重に

■目的外利用はしない

集めた時の目的以外にデータを使わないようにします。

■アクセス権限を守る

自分に許可されていないデータにはアクセスしません。

■安全な方法で共有

他の人や社外とデータを共有する際は、会社のルール（パスワード設定、安全なファイル共有サービスの利用、秘密保持契約の確認、上司への報告・承認など）を必ず守りましょう。特に取引先や委託先にデータを渡す場合は、相手がしっかり管理してくれるか、契約などで確認することが重要です。

データを廃棄する時：「完全に消す」意識で

■「ゴミ箱に入れる」だけでは不十分

パソコンのゴミ箱を空にしても、データは復元できてしまう

ことがあります。不要になったデータは、専用ソフトで完全に削除したり、紙ならシュレッダーにかけたり、記録媒体（USBメモリなど）は物理的に壊したりして、復元できないように確実に処分しましょう。

■情報の「重要度」を意識する（データ分類）

会社には様々な情報があります。「これは社外秘」「これは部内限り」「これは公開情報」といったように、情報の重要度に応じてランク付け（データ分類）をし、そのランクに合わせた取り扱いルールを決めておくと、日々の判断がしやすくなり、管理も徹底できます。

これらの注意点は、AI利用に限らず、情報セキュリティの基本です。しかし、AIによって大量の情報を扱う機会が増える今だからこそ、改めて組織全体でこれらの基本を徹底することが、情報漏洩リスクを防ぐための土台となります。

3.5 このリスク特有のインシデント対応ポイント

万が一、漏洩が起きてしまったら…

どれだけ気をつけていても、事故は起こりえます。もし情報漏洩（特に個人情報漏洩）が発生してしまった場合、パニックにならず、迅速かつ適切に対応することが、被害を最小限に食

い止め、信頼を回復するために極めて重要です。第7章で解説するインシデント対応の基本的な流れ（初動→調査→通知→再発防止）に加えて、情報漏洩特有の対応ポイントを押さえておきましょう。

■法的義務への超速対応（時間との勝負！）：「報告・通知が必要か？」を即判断

漏洩した情報が個人情報保護法上の「個人データ」で、かつ「個人の権利利益を害するおそれが大きい特定ケース」（例：クレジットカード情報、人種・信条などの要配慮情報、不正アクセスによる漏洩、1,000人超など）に該当するかを、覚知後できるだけ早く（できれば当日中、遅くとも数日以内に）判断します。この初動判断が遅れると、すべてが後手に回ります。

■期限厳守での報告・通知

報告・通知義務があると判断したら、個人情報保護委員会への報告（速報：概ね3～5日以内、確報：30日以内など）と、影響を受けた本人への通知を、定められた期限内に、法律で定められた内容・方法で行う必要があります。時間は非常に限られています。事前にテンプレートや連絡体制を準備しておくことが必須です。

■影響範囲の正確な特定：「何を・誰の・何件」を徹底調査

どのような情報が、誰の情報（顧客？　従業員？）が、何件くらい漏洩した（またはその可能性がある）のか。この被害状況の正確な把握が、その後の対応の質を決定します。システムログの解析、データベースの調査などを徹底的に行い、影響範

囲を正確に特定します。憶測で判断してはいけません。

■二次被害を防ぐための先回り

漏れた情報が悪用され、被害を受けた本人にさらなる被害（なりすまし、詐欺、不正利用など）が及ぶのを防ぐための手を打ちます。例えば、「不審なメールや電話に注意してください」といった注意喚起、パスワード変更の強い推奨、クレジットカード情報の漏洩ならカード会社への連絡と監視依頼、場合によってはお見舞い金などの対応の検討も必要になります。

■原因究明：なぜ漏れたのか？　根本原因を突き止める

再発防止のためには、表面的な原因だけでなく、「なぜ」（Why）を繰り返し問い、根本原因を特定することが重要です。技術的な脆弱性だけでなく、業務プロセス上の問題、従業員の認識不足、管理体制の不備など、多角的に調査します。どの経路（学習データ？　プロンプト入力？　不正アクセス？　内部犯行？）から漏れたのかを特定することが鍵です。

■誠実・透明性・迅速なコミュニケーション：信頼回復の唯一の道

不都合な事実を隠したり、対応が遅れたりすることは、さらなる不信を招きます。影響を受けた本人、監督官庁、顧客、株主、社会全体に対し、判明している事実、原因、対応状況、再発防止策などを、誠実に、透明性を持って、そして適切なタイミングでコミュニケーションをとることが、信頼回復への唯一の道です。広報・法務・経営が連携し、一貫性のある丁寧な説明を心がけます。

情報漏洩インシデントは、企業の危機管理能力そのものが問

われる重大事態です。**「備えあれば憂いなし」**というように、日頃からの対策と、万が一の際の冷静かつ迅速・誠実な対応体制の準備が不可欠です。

【章のまとめ】

本章では、生成AI活用における最重要リスクの一つ、「情報漏洩」について、その発生原因から深刻な影響、具体的な技術対策、日々のデータ取り扱いにおける注意点、そして万が一の際の対応ポイントまでを、できるだけ分かりやすく解説しました。

情報漏洩は、ひとたび発生すれば企業に計り知れないダメージを与えます。しかし、そのリスクは、本章で見てきたような技術的な対策、日々の丁寧なデータ取り扱い（運用）、そして第6章で解説する組織的なルール作り（ガイドライン）と従業員一人ひとりの意識向上（教育）、さらに第7章のインシデント対応への備えを組み合わせることで、確実に低減し、管理していくことが可能です。

次章では、AIが生み出す「コンテンツ」そのものに焦点を当て、著作権侵害、誤情報・偽情報、そしてバイアス・差別といった、これまた企業が見過ごせないリスクとその対策について、詳しく見ていくことにしましょう。

第4章

【コンテンツ関連リスク】

生成物の責任問題とその対策

第3章では、情報漏洩という「データ」に関するリスクを見てきました。本章では視点を変えて、生成AIが「創り出したもの（コンテンツ）」そのものに潜むリスクに焦点を当てます。AIが作った文章、画像、企画案などが、意図せず問題を引き起こし、会社の信用や法的立場を危うくするケースがあるのです。

想像してみてください。AIに作らせた広告デザインが有名ブランドのデザインにそっくりだったら？　AIが書いたプレスリリースに重大な事実誤認があったら？　AIが作成した社内文書が、特定の社員グループを不快にさせる内容だったら？

これらは決して絵空事ではなく、AI活用を進める上で避けては通れない課題です。

本章では、特に注意すべきコンテンツ関連の３大リスク――「著作権侵害」「誤情報・偽情報」「バイアス・差別助長」――を取り上げます。それぞれ「なぜ起こるのか？」「どんな法的・倫理的な問題があるのか？」「どうすれば防げるのか？」そして「もし問題が起きてしまったら、どう対応すべきか？」を、ビジネスパーソンの視点で分かりやすく解説していきます。

4.1 著作権侵害リスク

発生メカニズム、法的論点、企業の留意点

なぜAIが著作権侵害？　その仕組みとは

■AIは「学習」する、だから「似てしまう」？

AIは、インターネット上にある膨大な文章や画像を「お手本」として学習します。そのお手本の中には、当然、誰かが著作権を持っているものがたくさん含まれています。AIが賢くなる過程で、特定の作家の文体や、特定のイラストレーターの画風、特定の楽曲のメロディなどを、まるで人間が癖を覚えるように「記憶」してしまうことがあります。その結果、AIが生成したものが、意図せず既存の作品にそっくりになってしまう可能性があるのです。特に、学習データが偏っていたり、特定のスタイルを集中して学習させたりすると、その傾向は強まります。

■指示（プロンプト）が悪影響？

私たち利用者が「○○（有名キャラクター）みたいな絵を描いて」「△△（有名作家）風の文章で」といった指示を出せば、当然、AIはそれに似せようとします。これも著作権侵害につながる可能性があります。

■法律のルールはどうなっているの？（現状と課題）

実は、生成AIと著作権の関係については、まだ法律のルー

ルが完全には追いついておらず、世界中で議論が続いているのが現状です。「これはOK」「これはNG」と明確に線引きできないグレーゾーンが多いのです。

■AIの「学習」はOK？

AIが開発のためにネット上の情報を集めて学習すること自体が、著作権侵害にあたるのではないか？という議論があります。日本の法律（著作権法）では、一定の条件下（例：情報解析のためなど）であれば許されるという考え方もありますが、その解釈は専門家の間でも分かれており、特に営利目的のAI開発については、まだ不透明な部分が多く残っています。

■AIが作ったものに著作権はあるの？

AIが自動的に作り出した文章や絵に、人間が作ったものと同じように「著作権」が認められるのか？　もし認められるなら、その権利は誰（AI開発者？　AI利用者？　それともAI自身？）のものになるのか？　これも世界的に結論が出ていない大きな論点です。現状では、「人間の創作的な工夫や指示がどれだけ加わったか」が、著作物として認められるかどうかのポイントになると考えられています。

■「似ている」＝「侵害」？

AIが作ったものが既存の作品と似ていた場合、著作権侵害になるかどうかは、最終的には「元の作品を参考にして（依拠して）作られたか」そして「表現が具体的にどれだけ似ているか（類似性）」で判断されると考えられます。しかし、AIの場合は「参考にした」という意図を証明するのが難しく、どこま

で似ていたらNGなのか、その判断基準もまだ手探りの状態です。関連する裁判も国内外で始まっていますが、明確な判例が蓄積されるには時間がかかりそうです。

企業として特に注意すべきこと

■「商用利用」はリスクが高い！

AIが生成したものを、自社の製品、サービス、ウェブサイト、広告、販促物などに使う場合（商用利用）は、特に注意が必要です。もし著作権侵害と判断されれば、利用の差止めや、多額の損害賠償を請求される可能性があります。

■「利用規約」を必ず確認！

利用するAIサービスの利用規約には、「生成したものの権利はどうなるか」「商用利用はOKか」「もし問題が起きたら誰が責任を負うのか」といった重要なルールが書かれています。サービスによって内容は全く異なるので、必ず利用前に隅々まで確認し、理解しておく必要があります。「知らなかった」では済みません。

■社内ルールがないと危険！

従業員が個々の判断で、「便利だから」「きれいだから」とAI生成物を安易に会社の資料やウェブサイトに使ってしまうと、後で大きな問題になる可能性があります。会社として利用して良いAIサービス、利用目的、利用前の確認手順などを明確なルール（ガイドライン）として定め、全員に徹底することが非

常に重要です（詳細は第6章参照）。

4.2 著作権侵害への対策

「材料選び」「作り方」「使い方」に注意！

著作権侵害のリスクをゼロにすることは難しいですが、以下の対策を組み合わせることで、リスクを大幅に減らすことができます。料理に例えるなら、「安全な材料を選び」「レシピを守って調理し」「出来上がった料理を安全に提供する」というイメージです（社内ルール作りや教育の詳細は第6章、発生時の対応は第7章を参照してください）。

「安全な材料」を選ぶ（学習データの確認・選択）

■権利がクリーンなAIを選ぶ

可能であれば、AIの開発元が「学習データは著作権に配慮したものだけを使っています」と明確に説明しているサービス（例：Adobe Fireflyのように、自社のストックフォト素材などを活用しているサービス）や、著作権フリーのデータ（パブリックドメイン、CC0ライセンスなど）のみで学習させたモデルの利用を優先的に検討しましょう。

■ライセンス条件を確認する

著作権フリーとされる素材にも、「商用利用は不可」「改変不

可」「作者名の表示が必要」といった条件（ライセンス）が付いている場合があります。利用するデータやサービスのライセンス条件は必ず確認し、ルールを守りましょう。

■開発元に問い合わせる（可能な範囲で）

利用したいAIサービスについて、学習データの権利処理方針が不明な場合は、開発元に問い合わせてみることも有効です。

【評価目安】

・導入のしやすさ：B（高い）～ C（中程度）

権利に配慮したサービスを選ぶことは比較的容易ですが（B）、自社でデータを用意したり、ライセンスを詳細に確認したりするにはある程度の準備が必要です（C）。

・効果：B（高い）

学習データ由来のリスクを大きく減らせますが、生成プロセスや利用段階でのリスクが残るため、完全ではありません。

「安全な作り方」を心がける（生成プロセスでの対策）

■指示（プロンプト）に注意

AIに指示を出す際に、特定の作品名、作者名、キャラクター名などを安易に含めないようにします。「○○風」といった指示も、類似性が高まるリスクがあるため、慎重に検討が必要です（社内ガイドラインでルール化しましょう）。

■類似度チェックツールを活用（参考程度に）

AIが生成したものと、既存の作品が似ていないかをチェッ

クするツール（画像比較、テキスト類似度判定など）もあります。これらはリスクを発見する「補助的な手段」としては有効ですが、世界の全ての作品と比較できるわけではなく、精度にも限界があるため、ツールだけで安心せず、必ず人間の目でも確認しましょう。

【評価目安】

・導入のしやすさ：C（中程度）

ツールの選定や導入、運用にある程度の準備が必要です。

・効果：C（中程度）

明らかな類似を発見できる効果はありますが限定的であり、過信は禁物。他の対策との組み合わせが重要です。

「安全な使い方」を徹底する（生成物の利用段階での対策）

■人間による最終チェック（必須！）

AIが生成したコンテンツを、特に会社のウェブサイトに載せたり、製品に使ったり、広告で使ったりする前には、必ず担当者が自分の目で見て、既存の作品に似すぎていないか、問題がないかを確認します。少しでも懸念があれば、利用を控えるか、著作権に詳しい専門家（弁護士や弁理士）に相談しましょう。

■「ひと手間」加えてオリジナルに

AI生成物をそのまま使うのではなく、人間の手で修正、編集、他の要素との組み合わせなどを行うことで、独自の創作性を加え、著作権侵害のリスクを低減できる場合があります（ただし、

元が酷似している場合は効果が限定的です)。

■利用規約を守る

利用しているAIサービスの利用規約で定められたルール(例:「商用利用はOKか?」「AIが作ったことを表示する必要があるか?」など)を必ず守りましょう。

【評価目安】

・導入のしやすさ:A(非常に高い)

特別なツールは不要で、今日からでも意識して始められる対策です。

・効果:B(高い)

リスクのある生成物の利用を水際で防いだり、独自性を加えたりすることで、リスクを大きく減らせます。

4.3 誤情報・偽情報リスク

発生メカニズムと社会的影響

なぜAIは「嘘」や「間違い」を言うのか?

■ハルシネーション(AIの“知ったかぶり”)

これが最も厄介な性質の一つです。AIは、知らないことや学習データにない情報についても、質問されると「それらしい嘘」をもっともらしく、時には非常に自信満々に答えてしまうことがあります。人間なら「分かりません」と言う場面でも、

AIは平気でデタラメな情報を生成してしまうのです。これを「ハルシネーション（幻覚）」と呼びます。

■学習データの限界

AIは、学習したデータに基づいて答えを生成します。そのため、学習データ自体が古かったり、間違っていたり、偏っていたりすれば、AIの答えも当然そうなります。「AIは何でも知っている」わけではなく、学習した範囲内の知識しか持っていないのです。

■「今」の情報には弱い

多くのAIは、ある特定の時点までのデータで学習しています。そのため、最新のニュース、株価、天気予報など、リアルタイムで変化する情報については、正確に答えられないことがよくあります。

■悪意ある使い方（フェイクニュース生成など）

このAIの能力を悪用すれば、非常に精巧な偽情報（フェイクニュース記事、本物そっくりの偽画像・動画「ディープフェイク」など）を、効率的に大量生産することも可能になってしまいます。これらが、詐欺、世論操作、特定の個人や企業への攻撃などに使われる危険性が高まっています。

■もし誤情報・偽情報が広まったら？

AIが生成した、あるいはAIによって拡散された誤情報・偽情報は、単なる「間違い」では済みません。個人の判断を誤らせ（例：間違った健康情報で健康被害）、会社の評判を著しく傷つけ（例：自社に関するデマが拡散）、社会に無用の混乱や

不信感を引き起こし（例：災害時のデマ）、時には選挙や世論に影響を与え、民主主義の根幹を揺るがす可能性すらあります。

4.4 誤情報・偽情報への対策

「鵜呑みにしない」「必ず裏を取る」が鉄則！

AIが生成した情報の真偽を100％見抜く魔法はありません。しかし、リスクを減らすためにできることはあります。技術的な補助機能もありますが、最終的には私たち利用者自身の「情報リテラシー（情報を正しく見極める力）」と「確認プロセス」が最も重要になります（組織的なチェック体制や教育については第6章参照）。

技術的なアプローチ（参考にはなるが過信は禁物）

■信頼度スコア・出典表示

一部のAIサービスでは、「この回答の自信度は○○％です」「参考にした情報源はこちらです」といった情報を表示してくれることがあります。これらは判断材料の一つにはなりますが、スコアが高くても間違っていることや、出典自体が信頼できないこともあるため、鵜呑みにしてはいけません。

■自動ファクトチェック

AIが情報の真偽を自動でチェックする技術も研究されてい

ますが、世の中の全ての情報の真偽を判定するのは非常に難しく、まだ発展途上の技術です。現時点では、これに頼り切ることはできません。

利用者側のリテラシーと確認プロセス（これが最重要！）

■「AIの言うことは絶対ではない」と心得る

まず大前提として、生成AIの出力は「完璧な答え」ではなく、「AIが生成した参考情報」あるいは「たたき台」と捉えましょう。「本当に正しいか？」「根拠はあるか？」「偏っていないか？」と常に疑う目（批判的思考）を持つことが重要です。

■「裏取り（ファクトチェック）」を必ず行う習慣を

AIが提示した情報、特に客観的な事実、数値、歴史的な出来事、科学的な情報などについては、そのまま信じ込まず、必ず信頼できる情報源（公的機関の発表、専門機関の報告書、信頼性の高い報道機関の記事など）を複数確認し、裏付けを取ることを徹底しましょう。インターネット検索で見つけた匿名のブログ記事やSNSの情報だけを根拠にするのは危険です。

■情報源を確かめる

AIが出典を示した場合でも、その情報源が本当に信頼できるものか（専門機関か、個人の意見かなど）を確認する一手間が大切です。

■専門分野では特に慎重に

医療、法律、金融、技術開発など、情報の正確さが人命や事

業に直結するような専門分野においては、AIの利用はあくまで「補助」に留めるべきです。AIの提案を参考にしつつも、必ずその分野の専門家によるレビューと最終判断を経るプロセスを省略してはいけません。

■学び続ける姿勢

AIの能力や限界、誤情報の手口、ファクトチェックの方法などについて、継続的に学び、知識をアップデートしていくことが大切です（第6章の教育と連携）。

【評価目安】

・導入のしやすさ：A（非常に高い）

特別なツールは不要で、意識と習慣の問題であり、今日からでも始められます。

・効果：A（非常に高い）

誤情報によるリスクを回避するための最も基本的かつ効果的な対策であり、影響を最小限にできます。

4.5 バイアス・差別助長リスク

発生メカニズムと倫理的課題

なぜAIは「偏見」を持つの？

■学習データが偏っているから

AIは、人間が作った大量のデータ（インターネット上の文章や画像など）から学習します。残念ながら、そのデータの中には、私たちの社会に存在する様々な「偏見（バイアス）」や「固定観念（ステレオタイプ）」（例：性別による役割分担の意識、特定の人種や国籍に対するイメージ、特定の職業に対する思い込みなど）が反映されてしまっています。AIはそれをそのまま学習してしまうため、結果として偏った、あるいは差別的とも取れるコンテンツを生成してしまうことがあります。「看護師と言えば女性」「社長と言えば男性」といったイメージをAIが再現してしまうのは、このためです。

■データが足りない場合も

学習データが、世の中の多様な人々（性別、人種、年齢層、地域など）をバランス良く反映しておらず、特定のグループの情報が極端に少ない場合、その少数派グループに対して不正確だったり、不利になったりするような結果を出してしまう可能性もあります。

■作り方（アルゴリズム）の問題も

AIモデルの設計方法や、何を「良い結果」として学習させるかという目標設定自体が、意図せず特定のグループに不利なバイアスを生み出してしまうこともあります。

■なぜこれが大きな問題なのか？（倫理的課題）

AIによるバイアスや差別は、単に「ちょっと偏ってるね」で済む話ではありません。

■不公平・不公正

例えば、採用選考AIが特定の性別や出身校に無意識の偏見を持っていたら、公平な採用機会が奪われてしまいます。ローン審査AIが特定の地域住民に不利な判断をしたら、経済的な不利益につながります。

■偏見の再生産・固定化

AIが社会に存在する偏見を繰り返し出力することで、その偏見がさらに強化され、世の中に定着してしまう恐れがあります。

■個人の尊厳

差別的な表現や扱いは、人の心を深く傷つけ、尊厳を否定することにもなりかねません。

企業が提供するAIサービスがこのような問題を起こせば、社会的な非難を浴び、ブランドイメージは大きく傷つきます。「AI倫理（AIを開発・利用する上で守るべき道徳や規範）」への配慮は、もはや企業の社会的責任（CSR）として、避けて通れない課題なのです。

4.6 バイアス・差別助長への対策

「多様性」と「チェック」が鍵

AIからバイアスを完全になくすことは非常に難しい挑戦ですが、その影響をできる限り減らし、より公平なAI利用を目指すために、以下の対策が考えられます（組織的な倫理原則策定や多様なチーム編成は第6章参照）。

「材料（データ）」の質を高める

■多様なデータをバランス良く

AIの学習に使うデータは、可能な限り、性別、人種、年齢層、地域、文化などの点で偏りがなく、社会の多様性を反映したものになるように努めます。どこからデータを集めるか、誰がデータに情報を付与するか（アノテーション）の過程でも、多様な視点を取り入れることが重要です。

■データの中のバイアスをチェック

使うデータセットにどのような偏り（バイアス）が含まれている可能性があるかを、統計的な手法や専用のツールを使って事前に分析・評価し、そのリスクを把握しておくことが大切です。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

質の高い多様なデータセットの構築・入手は、コストや準備

が必要であり、容易ではありません。

・効果：A（非常に高い）

学習データはバイアスの根源の一つであり、ここの質を高めることはリスクをほぼ無くせる、または影響を最小限にできる根本的な対策となります。

「作り方（技術）」で偏りを減らす

■バイアス緩和技術の活用

AIの開発・学習プロセスにおいて、特定のグループに不利にならないように公平性を保つための技術（バイアス緩和技術）を導入することを検討します。例えば、学習データの中で少数派のデータの重要度を高くする（重み付け）、学習中に公平性の度合いをチェックしながら調整する、AIの出力結果を後から補正する、といった様々なアプローチがあります（これらは専門的な技術なので、導入にはAI専門家の知識が必要です）。

■公平性の「ものさし」で測る

AIの出力が、例えば男女間や人種間で不公平になっていないかを、客観的な指標（公平性指標）を使って定期的に測定し、監視する仕組みも有効です。

【評価目安（バイアス緩和技術・指標）】

・導入のしやすさ：D（低い）～ E（非常に低い）

高度な専門知識や技術、かなりのコストや準備が必要となる

場合が多いです。

・効果：B（高い）

技術的にバイアスを大きく減らせる可能性がありますが、完全な除去は難しく、新たなバイアスを生む可能性もゼロではありません。

「使い方（利用段階）」で注意する

■多様な視点でレビュー

AIが生成したコンテンツ（特に社外に出すものや、多くの人が目にするもの）は、様々な立場の人（異なる性別、年齢、文化背景など）がチェックし、偏った表現や配慮に欠ける点がないかを確認するプロセスを設けることが望ましいです。

■特に慎重な分野での利用制限

人事評価、採用、ローン審査、犯罪予測など、個人の人生や権利に大きな影響を与える可能性のある分野でAIを利用する場合は、バイアスリスクが特に深刻な結果を招くため、利用を制限するか、あくまで補助的な参考情報として位置づけ、最終判断は必ず人間が、複数の視点から慎重に行うというルールを徹底する必要があります。

【評価目安（多様なレビュー、利用制限）】

・導入のしやすさ：B（高い）

プロセスやルールを整備する必要はありますが、既存のものを活用できる部分も多く、比較的取り組みやすい対策です。

・効果：B（高い）

バイアスのある出力が外部に出たり、重要な判断に使われたりするリスクを大きく減らせます。

4.7 このリスク領域特有のインシデント対応ポイント

「信頼」に関わる問題への対応

もし、AIが生成したコンテンツが、著作権侵害、誤情報・偽情報、あるいはバイアス・差別を含むと外部から指摘されたり、社内で問題が発覚したりした場合、その対応は企業の信頼に直結します。第7章で解説するインシデント対応の基本フロー（初動→調査→通知→再発防止）を踏まえつつ、以下の点が特に重要になります。

■共通の初期対応：まず被害拡大を防ぐ！

問題となっているコンテンツ（ウェブ記事、SNS投稿、広告、社内文書など）を特定し、直ちに公開停止、削除、あるいは修正します。これ以上、問題が広がるのを防ぐことが最優先です。同時に、そのコンテンツが「いつ、どこで、誰に」影響を与えた可能性があるのか、影響範囲の調査を開始します。

■著作権侵害が疑われる場合：「権利者」との対話が鍵

権利者や代理人から連絡があったら、無視せず、誠実に対応しましょう。まずは事実確認を行い、必要であれば弁護士など

の専門家に相談します。相手の主張をよく聞き、和解交渉やライセンス契約といった解決策を探ることも重要です。訴訟に発展すると、時間もコストもかかり、企業の評判にも傷がつきます。社内で同様のコンテンツが使われていないか、徹底的に調査し、あれば利用を停止します。

■誤情報・偽情報が拡散された場合：「迅速な訂正」と「透明性」

間違った情報だと判明したら、できるだけ早く、正確な情報と共に「訂正」（場合によっては「謝罪」も）を、誤報が広まったのと同じかそれ以上の範囲に発信します。隠そうとすると、かえって不信感を招きます。SNSなどで拡散状況を監視し、必要であれば再度訂正情報を流したり、プラットフォームに削除を依頼したりします。なぜ誤情報が生成・公開されたのか原因を究明し、ファクトチェック体制やAIの利用ルールを見直すことが再発防止につながります。

■バイアス・差別的な表現があった場合：「誠実な謝罪」と「改善への姿勢」

不快な思いをさせたり、不利益を与えたりした可能性のある個人やコミュニティに対して、真摯に謝罪することが第一です。そして、彼らの声に耳を傾け、懸念や意見を真摯に受け止める姿勢を示しましょう。なぜ問題が起きたのか、会社としてどう受け止め、今後どのように改善していくのかを、可能な範囲で透明性を持って説明することが、信頼回復への道です。場合によっては、外部の有識者や関係団体との対話も有効です。技術的な対策（バイアス緩和）だけでなく、開発チームの多様性確

保や、社内の倫理教育、チェックプロセスなど、組織的な問題にも目を向け、根本的な改善を図る必要があります。

コンテンツに関する問題は、多くの人の目に触れやすく、感情的な反発も招きやすいため、特に迅速かつ誠実、そして透明性の高いコミュニケーションが求められます。

【章のまとめ】

本章では、生成AIが生み出す「コンテンツ」そのものに焦点を当て、著作権侵害、誤情報・偽情報、そしてバイアス・差別助長という、企業が見過ごすことのできない3つの主要なリスクについて、その発生の仕組み、法や倫理に関わる課題、そして具体的な対策と問題発生時の対応ポイントを、ビジネスパーソンの視点から詳しく解説しました。

これらのリスクへの対応は、単に技術を導入すれば解決するものではありません。AIを使う私たち人間自身の知識（リテラシー）、注意深さ、そして社会に対する責任感（倫理観）が問われています。また、会社全体で明確なルールを作り、それを守る文化を育てていくこと（第6章のAIガバナンス）、そして万が一の際に適切に対応できる備え（第7章のインシデント対応）が不可欠です。

次章では、視点を変え、AIを動かす「モデル」自体や、AIを「運用」するプロセスに潜むリスク――システムの脆弱性、予測不能な動作、AIへの過信、そしてシステム停止といった、

AIシステムの安定性や信頼性に関わる問題について、詳しく見ていくことにしましょう。

第5章

【モデル・運用関連リスク】

AIシステムの安定性と信頼性確保

これまでの章では、AIに入力する「データ」のリスク（第3章）や、AIが生み出す「コンテンツ」のリスク（第4章）について見てきました。しかし、AI活用におけるリスクはそれだけではありません。私たちが利用するAIの「モデル（AIの頭脳部分）」そのものや、AIシステムを日々「運用」していく過程にも、見過ごせないリスクが潜んでいます。

「AIシステムがハッキングされたら？」

「AIが突然おかしな動きをし始めたら？」

「便利さのあまりAIを信じすぎて失敗したら？」

「頼りにしていたAIサービスが突然使えなくなったら？」

これらはすべて、企業のビジネス継続や社会的な信頼を根底から揺るがしかねない問題です。

本章では、このようにAIモデル自体や日々の運用に関わるリスクに焦点を当て、その具体的な脅威の内容、なぜそれが起こるのか、そして企業としてどのように備え、対策していくべきかを解説します。AIという強力なエンジンを、常に安定して、信頼できる状態で、安全に動かし続けるための重要なポイントを一緒に学んでいきましょう。

5.1 モデルの脆弱性と攻撃リスク

敵対的サンプル、モデル抽出などの脅威

私たちが日々利用しているパソコンやスマートフォンがウイルス感染やハッキングの脅威に晒されているように、AIの「頭脳」であるAIモデル自体も、悪意を持った人たちの攻撃ターゲットになりえます。攻撃者は、AIを騙して誤った判断をさせたり、AIが持つ秘密情報（学習データやモデルの設計図）を盗み出したり、あるいはAIサービスそのものを利用不能にしたりしようとします。企業が知っておくべき、AIを狙った主な攻撃の種類を見てみましょう。

敵対的サンプル（Adversarial Examples）：「AIの目くらまし」攻撃

■どんな攻撃？

AIが見る画像や聞く音声、読むテキストに、人間にはほとんど分からないような特殊な加工（ノイズなど）を少しだけ加えることで、AIに全く違うものだと勘違いさせる攻撃です。まるでAIの目をくらませるような手口です。例として、自動運転車が特定のシールが貼られた「止まれ」の標識を「速度制限解除」と誤認識してしまう、といった研究事例があります。

プロンプトインジェクション（Prompt Injection）：「AIへの隠しコマンド」攻撃

■どんな攻撃？

私たちがAIに送る指示（プロンプト）の中に、悪意のある命令をこっそり紛れ込ませることで、AIに開発者が想定していない危険な動作（例：会社の秘密情報を漏洩させる、差別的な文章を生成させる、他のシステムを不正に操作させるなど）を実行させようとする攻撃です。チャットボットなどが標的になりやすいと言われています。

モデル抽出（Model Extraction/Stealing）：「AIの設計図」泥棒

■どんな攻撃？

AIサービス（特にAPIなどで外部から利用できるもの）に対して、様々な質問を大量に投げかけ、その応答パターンを分析することで、AIモデルの内部構造や学習内容（いわば「設計図」や「秘伝のレシピ」）を推測し、盗み出そうとする攻撃です。

■脅威

企業が多大なコストと時間をかけて開発した独自の高性能AIモデルが、競合他社などにコピーされてしまうリスクがあります。

データ汚染（Data Poisoning）：「AIに毒リンゴを食べさせる」攻撃

■どんな攻撃？

AIが学習するデータの中に、意図的に間違った情報や偏った情報、あるいは特定の状況で誤作動を引き起こすような「毒」データを仕込む攻撃です。特に、ユーザーからのフィードバックなどを元に追加学習を続けるタイプのAIは、この攻撃を受けやすい可能性があります。

■脅威

AIの性能全体が悪くなったり、特定の質問に対して常に間違った答えをするようになったり、有害なバイアスが埋め込まれたりする可能性があります。

メンバーシップ推論攻撃：「あの情報、AIは知っている？」を探る攻撃

■どんな攻撃？

AIの応答の仕方などから、「このAIは、○○さんの病気のデータ（個人情報）や、△△社の新製品情報（機密情報）を学習したのではないか？」と推測しようとする攻撃です。

■脅威

直接データが漏れるわけではなくても、AIがどのような情報を学習したかが推測されることで、間接的にプライバシーや機

密情報が侵害されるリスクがあります。

これらの攻撃は、AI技術の進化と共にますます巧妙化しています。AIシステムの「守り」を固めるための継続的な対策が不可欠です。

5.2 脆弱性・攻撃への対策

「AIの鎧」を固める

AIモデルを悪意ある攻撃から守るためには、様々な技術的・運用的な対策を組み合わせることが重要です（会社としてのルール作りや従業員教育については第6章で詳しく述べます）。

技術的対策（システムの力で守る）

●敵対的サンプルへの防御

AI自身に「騙されにくい」訓練（敵対的学習）をさせたり、入力された情報に不審な加工がされていないかチェックしたりする技術を導入します。

【評価目安】

・導入のしやすさ：C（中程度）

専門知識が必要で、ある程度の準備が必要です。

・効果：B（高い）

特定の攻撃に対してはリスクを大きく減らせますが、未知の

攻撃には限界があります。

●入力検証・サニタイズ（入力の無害化）

AIへの指示（プロンプト）に、悪意のある命令やコードが含まれていないかを厳しくチェックし、危険な部分を取り除く（無害化する）仕組みを導入します。プロンプトインジェクション対策の基本です。

【評価目安】

・導入のしやすさ：B（高い）

少しの準備や調整で、既存の仕組みを活用できる場合もあります。

・効果：B（高い）

多くの既知の攻撃パターンのリスクを大きく減らせますが、巧妙な手口には注意が必要です。

●出力検証・フィルタリング

AIが生成した答えに、意図せず機密情報や不適切な内容が含まれていないかチェックし、問題があれば表示させないようにします（第3章のフィルタリング技術と共通）。

【評価目安】

・導入のしやすさ：A（非常に高い）～ C（中程度）

内容によりますが、基本的なチェックはすぐに始められます。

・効果：C（中程度）

漏洩や不適切出力のリスクをある程度減らせますが、完全で

はなく、他の対策との組み合わせが重要です。

●APIセキュリティの強化

AIサービスを外部から利用するための接続口（API）の管理を厳重にします。利用者本人かどうかを厳しく確認し、短時間に大量のアクセスができないように制限（レートリミット）することで、不正利用やモデル抽出攻撃を防ぎます。

【評価目安】

・導入のしやすさ：B（高い）

既存の仕組みを活用できる場合が多いです。

・効果：B（高い）

不正利用リスクを大きく減らせます。

●モデル難読化（AIの設計図を分かりにくくする）

モデルの内部構造を推測されにくくする技術的な工夫を行うことも、モデル抽出対策として考えられます。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

専門知識やかなりの準備が必要となる場合があります。

・効果：C（中程度）

完全な防御は難しいですが、攻撃の難易度を上げる効果は期待できます。

●プライバシー保護技術の活用

差分プライバシーや連合学習といった技術は、学習データのプライバシーを守るだけでなく、メンバーシップ推論攻撃やモデル抽出攻撃に対する防御効果も期待できます（3.3参照）。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

専門知識やかなりのコスト・準備が必要な場合があります。

・効果：B（高い）

特定の攻撃に対してリスクを大きく減らせる可能性があります。

運用上の対策（開発プロセスや日々のチェックで守る）

●安全な開発プロセス（SDL）の導入

AIモデルやシステムを開発する段階から、セキュリティを意識した手順（設計段階での脅威分析、安全なプログラムの書き方、十分なテストなど）を組み込み、そもそも脆弱性を作り込まないようにします（第6章の組織体制と連携）。

【評価目安】

・導入のしやすさ：B（高い）～ C（中程度）

プロセスの見直しやある程度の準備が必要ですが、既存の開発プロセスに組み込むことも可能です。

・効果：A（非常に高い）

開発初期からの対策は、手戻りを防ぎ、リスクをほぼ無くせる、または影響を最小限にできる非常に効果的なアプローチです。

●定期的な「健康診断」(脆弱性診断・ペネトレーションテスト)

開発したAIシステムに「悪いところ」がないか、専門家(社内チームや外部業者)に定期的(例:年に1回、大きな変更時など)にチェックしてもらう「脆弱性診断」や、実際にハッカーの視点で攻撃を仕掛けてみて弱点を探す「ペネトレーションテスト」を実施します。発見された弱点はすぐに修正します。

【評価目安】

・導入のしやすさ:C(中程度)

専門家への依頼やある程度の準備が必要です。

・効果:A(非常に高い)

未知の弱点を発見し、攻撃を未然に防ぐ上で非常に高い効果が期待できます。

5.3 予期せぬ動作・結果リスク

ブラックボックス性と予測不能性

現在の生成AI、特にLLMのような複雑なAIは、なぜそのような答えを出したのか、その思考プロセスが人間には完全には分からない「ブラックボックス」のような部分を持っています。そのため、普段は非常に賢く、期待通りに動いているように見えても、ある特定の状況や、私たちが予期しないような質問(入力)に対して、突然、おかしな動きをしたり、全く見当違いな、あるいは不適切な答えを出したりするリスクが常に付

きまといます。

なぜ「想定外」が起こるのか？

■AIの「知らないこと」

AIは学習したデータに基づいて動作するため、学習データに含まれていないような珍しい状況（エッジケース）や、全く新しいタイプの質問にはうまく対応できず、奇妙な反応を示すことがあります。

■複雑すぎる「頭の中」

AIの内部は何十億、何兆もの部品（パラメータ）が複雑に絡み合っています。その全てを人間が把握し、どんな入力に対しても完璧な出力を保証することは、現状の技術では非常に困難です。

■「世の中の変化」についていけない

AIが学習したのは過去のデータです。現実世界の状況（新しい出来事、新しい言葉遣いなど）は常に変化するため、AIの知識や判断が古くなり、的外れな答えをしてしまうことがあります。

■「気まぐれ」な一面？（確率的な性質）

AIの答えは、ある程度ランダム（確率的）に生成される要素を含んでいます。そのため、全く同じ質問をしても、時によって少し違う答えが返ってくることがあり、それが予期せぬ結果につながる可能性もゼロではありません。

■何が問題になる？

AIのこのような予測不能な振る舞いは、単に「面白い間違

い」で済めば良いですが、ビジネスにおいては、顧客の信頼を損なったり、業務に混乱を招いたり、場合によっては安全に関わる重大な問題（例：自動運転車の誤作動、医療AIの誤診）を引き起こす可能性があります。

5.4 予期せぬ動作への対策
「AIの健康診断」と「安全装置」

AIの予測不能性を完全になくすことはできませんが、そのリスクを管理し、できるだけ安定して、信頼できる動作をさせるための対策はあります。車の定期点検や安全装置のようなイメージです（組織としての監視体制やルールは第6章参照）。

徹底的な「健康診断」（テストと検証）

■色々な状況で試してみる（ロバスト性テスト）

AIが「普通の状況」だけでなく、少し意地悪な条件（ノイズの混じったデータ、わざと不完全にした指示、想定外の使い方など）でも、大きく動作がおかしくならないか（頑健か＝ロバストか）を、開発段階や導入前に徹底的にテストします。

■仮想空間で試運転（シミュレーション）

現実世界で試すのが難しい、あるいは危険な場合（例：自動運転、工場の制御など）は、コンピュータの中に現実そっくり

の仮想空間（シミュレーション環境）を作り、その中で様々な状況を再現してAIの動作を安全に事前検証します。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

テスト設計やシミュレーション環境構築には専門家の助けや、かなりのコスト・準備が必要な場合があります。

・効果：A（非常に高い）

事前に問題を発見し、修正することで、リスクをほぼ無くせる、または影響を最小限にできる重要な対策です。

「チーム」で安定性を高める（モデルの安定化技術）

■複数のAIで多数決（アンサンブル学習）

一つのAIモデルだけに頼るのではなく、複数の異なるAIモデル（作り方や学習データが違うもの）に同じ質問を投げかけ、それらの答えを統合（例：多数決や平均）して最終的な答えとする方法です。個々のAIの弱点や癖を互いに補い合い、全体としてより安定した、信頼性の高い結果が得やすくなります。

【評価目安】

・導入のしやすさ：C（中程度）

複数のモデルを管理・運用するため、ある程度の準備が必要です。

・効果：B（高い）

単一モデルよりもリスクを大きく減らせることが期待できます。

常に「見守り」、異常があれば知らせる（監視と検知）

■リアルタイム監視

AIが実際に動いている間、その動作（出力内容、応答時間、内部の状態など）を常にシステムが監視します。

■異常検知アラート

監視データの中から、「いつもと違う動き」「これはおかしい」という異常なパターンをシステムが自動で検知し、管理者にすぐに警告（アラート）を送ったり、場合によっては自動的にAIの動作を安全なモードに切り替えたりする仕組みを導入します。

【評価目安】

・導入のしやすさ：B（高い）～C（中程度）

監視システムの導入・設定に少しの準備や専門家の助けが必要な場合があります。

・効果：A（非常に高い）

問題発生を早期に検知し、影響を最小限にできる非常に重要な対策です。

「なぜ?」を探る手助け（原因究明の支援）

■説明可能なAI（XAI）技術

もしAIが予期せぬ動作をした場合に、「なぜそうなったのか」の原因究明を助けるために、AIの判断根拠やプロセスを可視

化する技術（XAI）を活用することも有効です。原因が分かれば、的確な修正や再発防止策につなげやすくなります。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

XAI技術の導入・活用には専門知識や準備が必要です。

・効果：B（高い）

原因究明を助け、再発防止につなげることでリスクを大きく減らせます。

5.5 AI過信リスク

人間の判断軽視が招く危険性

生成AIは非常に高性能で、様々な場面で私たちを助けてくれます。しかし、その便利さと賢さゆえに、私たちがAIの能力を過度に信じ込み、自分自身で考えたり、確認したりすることを怠ってしまう危険性があります。これが「AI過信リスク」です。

なぜ信じすぎてしまうのか？

■「すごい！」という印象

AIの驚くような能力を目の当たりにすると、「AIは自分より賢い」「AIの言うことなら間違いないだろう」と思い込みや

すくなります。

■「楽」だから

自分で調べるより、考えるより、AIに聞けばすぐに答え（らしきもの）が出てくるため、つい頼り切ってしまい、確認作業を面倒に感じてしまいます。

■「よく分からない」から（ブラックボックス性）

AIがどうやって答えを出しているのか仕組みが分からないため、「よく分からないけど、すごい技術が導き出した答えだから正しいはずだ」と、根拠なく信じてしまいがちです。

信じすぎるとどうなる？

■重大な判断ミス

AIが出した間違った情報や偏った分析結果（第4章参照）を鵜呑みにして、ビジネス上の重要な決定（例：投資判断、人事評価、顧客対応方針など）を誤ってしまう可能性があります。

■責任の所在が曖昧に

問題が起きた時に、「AIの指示に従っただけです」と、本来人間が負うべき責任をAIに転嫁しようとしてしまいがちです。しかし、最終的な判断と責任はAIではなく、それを利用した人間や組織にあります。

■人間の「考える力」の低下

常にAIに頼ることで、自分で情報を調べ、分析し、判断するという、ビジネスパーソンにとって最も重要な能力が鈍って

しまう（スキル低下）恐れがあります。「AIがないと仕事ができない」状態になってしまうかもしれません。

■倫理観の欠如

AIを介在させることで、その判断や行動がもたらす倫理的な影響に対する当事者意識が薄れてしまう可能性も指摘されています。

AIはあくまで道具（ツール）であり、それを使うのは私たち人間です。AIの能力を活かしつつも、常に「本当にそれで良いのか？」と立ち止まって考える姿勢が不可欠です。

5.6 過信への対策

「人間が主役」のAI活用ルールを作る

AIへの過度な依存や信頼を防ぐためには、技術的な工夫と同時に、「AIはあくまで脇役、主役は人間」という考えに基づいた仕事の進め方やルール作りが重要になります（従業員への教育やガイドラインの詳細は第6章参照）。

「人間中心」の仕組みづくり

■AIは「副操縦士」

AIはあくまで人間の判断を「助ける」存在（副操縦士）であり、最終的な判断と責任は人間（機長）が持つ、という役割分

担を、システム設計や業務プロセスの中に明確に組み込みます。「AIに全てお任せ」は原則NGと考えましょう。

■「人間チェックポイント」を設ける

AIの処理プロセスの中に、必ず人間が内容を確認し、承認・修正・却下といった判断を行う「関所」を設けます。特に、お客様への回答、重要な意思決定、安全に関わる作業などでは、人間の承認を必須とするルールにします。

■分かりやすいインターフェース

AIシステムの画面表示（インターフェース）も重要です。AIの回答が「確定情報」ではなく「提案」であることが一目で分かるようにしたり、安易に「OK」ボタンを押させないようなデザインにしたりする工夫も有効です。

【評価目安（人間中心プロセス）】

・導入のしやすさ：B（高い）

既存プロセスの見直しや少しの準備・調整が必要ですが、意識改革から始められます。

・効果：A（非常に高い）

過信によるリスクを根本的に防ぐ上で非常に効果が高い対策です。

AIの「自信度」を伝える工夫

■信頼度スコアなどの提示

AIが「この答えには結構自信があります（90％くらい）」「こ

れはあまり自信がないです（30％くらい）」といったように、出力に対する自信度（確信度）を数値やレベルで示してくれると、利用者はそれを参考に判断できます。自信度が低い場合は、「これは鵜呑みにせず、ちゃんと調べよう」と思えますね。

■警告・注意喚起

AIの答えが、学習データが少ない分野に関するものだったり、複数の情報源で矛盾があったりする場合などに、「この情報は不確かな可能性があります」「追加の確認が必要です」といった警告メッセージを自動で表示する仕組みも有効です。

【評価目安（信頼度提示・警告）】

・導入のしやすさ：B（高い）～ C（中程度）

利用するAIサービスによっては標準機能の場合も。独自開発なら準備が必要。

・効果：A（非常に高い）

利用者の「鵜呑み」を防ぎ、リスクをほぼ無くせるレベルで行動変容を促せます。

■「なぜそう言えるの？」を問いかける（XAIの活用）

AIの判断根拠を分かりやすく示す技術（XAI）を活用することで、利用者は「なるほど、こういう理由でAIはこの答えを出したのか」と納得感を持って結果を受け入れたり、あるいは「その根拠はおかしいのでは？」と疑問を持って批判的に評価したりすることができます。AIの「思考プロセス」が見えることで、盲目的な信頼（過信）を防ぐことにつながります。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

XAI技術の導入・活用には専門知識や準備が必要です。

・効果：B（高い）

過信を抑制し、適切な判断を促す上でリスクを大きく減らせます。

5.7 システム停止リスク（可用性）

事業継続への影響と備え

AIを業務に活用することが当たり前になればなるほど、そのAIシステムが「止まってしまう」ことの影響は計り知れません。電気や水道、インターネットが止まると困るのと同じように、AIシステムの可用性（Availability）、つまり「いつでも安定して使える状態」を維持することは、ビジネスを継続していく上で極めて重要になります。

■もしAIが止まったら？（影響）

顧客対応AIが止まれば、お客様からの問い合わせに答えられず、顧客満足度は急降下します。AIで在庫管理や生産計画を行っていれば、それが止まると欠品や生産遅延につながります。AIによるデータ分析が止まれば、経営判断に必要な情報が得られなくなります。AIへの依存度が高ければ高いほど、

システム停止は深刻な業務中断、売上減少、信用の失墜といった事態を招きます。

■なぜ止まってしまうのか？（原因）

- **機械の故障：**AIを動かしているコンピュータ（サーバー）や、データを保存している装置（ストレージ）、ネットワーク機器などが壊れる。
- **プログラムのミス：**AIシステム自体のプログラム（ソフトウェア）にバグがあったり、設定を間違えたりする。
- **ネットの問題：**会社のインターネット回線や、AIサービス提供会社のネットワークで障害が発生する。
- **外部サービスのトラブル：**利用している外部のAIサービス（API提供元など）自体が、障害やメンテナンスで利用できなくなる。
- **災害・停電：**地震や台風、大規模な停電などで、データセンターやオフィスが機能しなくなる。
- **サイバー攻撃：**大量のアクセスでシステムを麻痺させる攻撃（DDoS攻撃）や、システムを乗っ取る攻撃などを受ける。

■どう備えるか？（対策）

「止まらないようにする」「止まってもすぐ復旧できるようにする」「止まっても何とかなるようにする」という3つの視点で備えます。

■止まりにくい仕組みを作る（高可用性設計）

- **予備を用意する（冗長化）：**重要な機械（サーバー、ネット

ワーク機器、電源など）は、1つだけでなく、予備をもう1セット（あるいはそれ以上）用意しておき、片方が壊れても自動的にもう片方に切り替わるようにします（二重化、クラスタリングなど）。

- **仕事を分担させる（負荷分散）：**複数のサーバーに仕事を分散させ、1台に負荷が集中しすぎないようにします。これにより、安定稼働と処理速度の向上が期待できます。

【評価目安】

・導入のしやすさ：C（中程度）～ D（低い）

システム構成の見直しや専門人材、かなりのコスト・準備が必要です。

・効果：A（非常に高い）

単一の故障によるシステム停止をほぼ無くせる効果的な対策です。

■すぐに元通りにできるようにする（バックアップと復旧計画）

- **データのコピーを取っておく（バックアップ）：**AIモデルや重要なデータは、定期的にコピーを取り、別の安全な場所に保管しておきます。毎日取るのか、週に1回取るのかなど、頻度や方法はデータの重要度に応じて決めます。
- **「避難計画」を作っておく（DRP）：**もし大きな障害や災害が起きても、「どのくらいの時間で復旧させるか（RTO）」「どの時点のデータまで戻せればOKか（RPO）」という目標を決め、そのための具体的な復旧手順（DRP：災害復

旧計画）を事前に作成し、訓練しておきます。

【評価目安】

・導入のしやすさ：B（高い）

バックアップは比較的容易。DRP策定・訓練には少しの準備や調整が必要です。

・効果：A（非常に高い）

万が一の際に迅速に復旧し、影響を最小限にするために不可欠です。

■止まっても困らないようにする（代替手段）

- **「もしも」の時の代わりの方法：**AIシステムが使えなくなった場合に備え、一時的に業務を続けるための代替手段（例：重要な作業を手作業で行う手順、別のシンプルなシステムを使う、など）を準備し、その手順を関係者に周知・訓練しておきます。

【評価目安】

・導入のしやすさ：B（高い）～ C（中程度）

代替手段の内容により、準備の程度は異なります。

・効果：B（高い）

業務完全停止のリスクを大きく減らせます。

■外部サービス利用時の注意：特定の会社のAIサービスに頼りすぎるのは危険です。その会社との契約（SLA：品質保証契約）をよく確認し、「どの程度の安定稼働を保証してくれるの

か」「もし止まったらどう補償してくれるのか」を理解しておきましょう。場合によっては、複数の会社のサービスを併用することもリスク分散になります。

5.8 このリスク領域特有のインシデント対応ポイント

問題発生時の「現場の動き」

AIモデルへの攻撃、予期せぬ動作、AI過信によるミス、システム停止…本章で見てきたような問題が現実に発生してしまった場合、混乱を最小限に抑え、迅速かつ的確に対応することが求められます。第7章で解説するインシデント対応の基本的な流れ（初動→調査→通知→再発防止）を踏まえつつ、この領域のリスク特有の対応ポイントを以下に示します。

共通の初期対応：まず止める！　広げない！　状況把握！

何か異常が起きたら、まずは関連するAI機能やシステムを安全に一時停止させ、被害が広がらないように隔離することが基本です。同時に、何が起きているのか、どこに影響が出ているのか、状況把握を急ぎます。

もし「脆弱性・攻撃」が疑われたら…？

■攻撃をシャットアウト！

攻撃の種類を特定し、不正なアクセスを遮断したり、悪用された弱点（脆弱性）を塞いだりする「封じ込め」を最優先で行います。

■「デジタル鑑識」で証拠探し

攻撃の証拠を保全し、誰がどこからどのように攻撃してきたのか、何が盗まれた（あるいは改ざんされた）のかを突き止めるため、専門家による詳細な調査（フォレンジック調査）が必要になることがあります。

もし「予期せぬ動作」が発生したら…？

■「再現」できるか確認

まず、問題が起きた時と同じ状況（入力、環境など）で、もう一度同じ問題が起きるか（再現性）を確認します。これが原因究明の手がかりになります。

■原因を切り分ける

問題の原因が、入力されたデータにあるのか、AIモデル自体にあるのか、それともシステム環境（連携している他のソフトなど）にあるのかを、冷静に切り分けて特定していきます。

■「元に戻す」か「直す」か判断

原因がAIモデルにある場合、一時的に安定していた前のバ

ージョンに戻す（ロールバック）か、あるいはモデル自体を修正・再学習するかを、影響の大きさを考慮して判断します。

もし「AI過信」による問題だと分かったら…？

■「なぜ信じすぎたか」を調査

問題を起こした担当者を注意するだけでなく、なぜAIを過信してしまったのか、その背景にある業務プロセスや、本人の知識・経験、あるいは会社の文化などをヒアリング等で詳しく調査します。

■「人」と「ルール」を見直す

調査結果に基づき、AIの限界や正しい使い方に関する従業員教育の内容や、AI利用に関する社内ガイドライン（特に人間の確認プロセスの部分）を具体的に見直します。

もし「システム停止」が発生したら…？

■「復旧」が最優先！

まずは、事前に定めた復旧計画（DRP）に従い、一刻も早くシステムを元の状態に戻し、事業を継続させることが最優先です。

■代替手段へ切り替え

復旧に時間がかかりそうな場合は、準備しておいた代替手段（手作業など）にスムーズに切り替えて、業務への影響を最小限に抑えます。

■根本原因を突き止め、恒久対策を

復旧後、なぜ停止したのか（機械の故障？　プログラムのミス？　外部サービスの障害？）を徹底的に調査し、同じ問題が二度と起こらないように、システムの構成を見直したり、契約を改善したりといった恒久的な対策を講じます。

これらのリスクは、AIという「技術」の問題だけでなく、それを使う「人」や「組織の仕組み」と深く関わっています。そのため、対応には技術部門だけでなく、関係する業務部門、法務、広報など、組織全体での連携が不可欠となります。

【章のまとめ】

本章では、AIモデルそのものの脆弱性や攻撃、予測不能な動作、そしてそれを運用する上での人間の過信、さらにはシステム停止といった、「モデル・運用関連リスク」について、その脅威と対策、インシデント発生時の対応ポイントを掘り下げてきました。

これらのリスクは、AIシステムの「安定性」と「信頼性」という、いわばAI活用の土台を揺るがしかねない重要な問題です。対策には、AI特有の技術的なアプローチ（敵対的防御、XAI、異常検知など）と、従来のITシステムにも通じる堅牢な開発・運用プラクティス（セキュア開発、高可用性設計、インシデント対応体制など）、そして何よりも「AIは万能ではない」という認識のもとで人間が適切に関与し続ける仕組みが不可欠

となります。

もちろん、これらの対策も、第6章で解説する組織全体のAIガバナンス体制や倫理原則、そして第7章で詳述するインシデント対応への包括的な備えと一体となって初めて、真の効果を発揮します。

次章（第6章）では、いよいよ、これまで見てきた様々なリスクに対し、組織全体としてどのように向き合い、管理体制、ルール作り、人材育成といった共通の基盤を構築していくべきかについて、具体的な方法論を解説していきます。

第6章

【共通対策①】

組織全体で構築するAIガバナンスと推進体制

これまでの章で、生成AIがもたらす様々なリスクと、それらに対する個別の対策について見てきました。「情報漏洩にはこの対策」「著作権にはこの注意点」といった具合です。しかし、どれだけ優れた「個別の対策（点）」を用意しても、それらを組織全体で適切に運用し、継続していくための「仕組み（線や面）」がなければ、本当の意味でリスクを管理し、AIの恩恵を最大限に引き出すことはできません。

例えば、高性能なセキュリティソフト（技術的対策）を導入しても、社員がパスワードを付箋に書いてモニターに貼っていたり（ルール・意識の問題）、どの部署がセキュリティに責任を持つのか曖昧だったりしたら、意味がないですよね。

本章では、AI活用を成功させるための「組織としての土台作り」、すなわち「AIガバナンス（AIを正しく使うための会社のルール作りと管理体制）」と、それを実行するための「推進体制」の構築について、具体的なポイントを解説します。これは、経営層から現場の従業員まで、会社に関わる全ての人に関わる重要なテーマです。

6.1 なぜ「会社全体」での取り組みが不可欠なのか？

「AI導入は、とりあえずIT部門や一部の詳しい部署に任せておけば良いのでは？」と思われるかもしれません。しかし、生成AIの活用においては、それでは不十分です。会社全体で足並みを揃えて取り組むことが、なぜそれほど重要なのでしょうか？

■リスクは部署を選ばない（複合性と全社的影響）

AIが生み出すリスク（情報漏洩、著作権侵害、誤情報、倫理問題など）は、特定の部署だけでなく、会社のあらゆる部門に関わり、相互に影響し合います。マーケティング部が使ったAIの出力が法務部の問題になったり、開発部の作ったAIが人事部の課題を引き起こしたりする可能性があるのです。問題発生時のダメージも、会社全体に及びます。

■「技術」だけでは守れない（技術的対策の限界）

最新のセキュリティ技術を導入しても、従業員のうっかりミス（例：機密情報をプロンプトに入力）、悪意ある行動（内部不正）、あるいはAIでは判断できない複雑な倫理的な問題などを、技術だけで完全に防ぐことはできません。「会社のルール」「従業員の意識」「チェック体制」といった、組織的な仕組みによる補完が絶対に必要です。

■「ウチの部署ルール」では混乱する（一貫性の必要性）

もし部署ごとにAIの利用ルールがバラバラだったらどうでしょう？　A部署ではOKなことがB部署ではNG、C部署ではそもそもルールがない…これでは、従業員は何を信じて良いか分からず混乱し、リスク管理もままなりません。会社として一貫した方針と明確なガイドラインを定め、全員がそれを理解し、同じ方向を向いてAIを活用していく必要があります。

■会社の「信頼」は全員で築くもの（社会的責任）

お客様や取引先、社会全体は、「あの会社はAIを責任持って使っているだろうか？」という目で見ています。しっかりとした管理体制（ガバナンス）を構築し、透明性を持って、倫理的に配慮しながらAIを活用している姿勢を示すことが、企業の社会的信用を守り、高める上で不可欠です。これは経営層だけでなく、全従業員の行動にかかっています。

■AIの価値を最大限引き出すために（持続的活用）

AIは導入して終わりではありません。技術は日々進化します。組織全体で継続的に学び、使い方を改善し、成功事例や失敗事例を共有していくことで、初めてAIはその真価を発揮し、会社の持続的な成長エンジンとなります。そのためには、一部の部署だけでなく、会社全体でAI活用を推進し、支える体制が欠かせません。

このように、AIガバナンスと推進体制の構築は、単なるリスク対策に留まらず、AI活用の効果を最大化し、企業の未来を左右する重要な経営課題なのです。

6.2 誰がやる？　どう進める？
責任体制の構築

「会社全体で取り組む」と言っても、誰かが旗を振り、計画を立て、実行を管理しなければ、物事は進みません。AI活用を円滑に進め、リスクをしっかり管理するためには、「誰が」「何を」「どのように」責任を持って進めるのか、その役割分担と体制を明確にすることが出発点です。

【経営層】「ウチはAIでこうなる！」ビジョンと覚悟を示す

これは最重要なことです。まずは経営トップが、AI活用の重要性を深く理解し、「AIを使って会社をこう変えていくんだ！」という明確なビジョンと戦略を示すことが全ての始まりです。そして、その実現に向けて、必要な予算や人員といったリソースを確保するという強いコミットメント（約束）が必要です。経営層の本気度が、組織全体の推進力となります。

【AI推進チーム/CoE】部門の壁を越えた「司令塔」

AI活用は一部門だけでは完結しません。IT部門、法務・コンプライアンス部門、人事部門、広報部門、そして実際にAIを使う各事業部門など、関係する部署から代表者を集めた横断的な専門チーム（AI推進チームやCoE：Center of Excellenceと呼ばれることも）を設置することが非常に有効です。

・**役割：**このチームが中心となって、AI導入の戦略作り、全社

的なガイドラインの作成・見直し、利用するAIツールの評価・選定、社内教育の企画・実施、活用状況の把握、リスク管理のサポートなど、AI活用に関する様々な活動をリードしていきます。いわば、AI活用の「司令塔」です。

【全体責任者】プロジェクトの「現場監督」

AI導入・活用プロジェクト全体を統括する責任者（役職名は様々ですが、AI担当役員や専任のプロジェクトマネージャーなど）を任命します。

・**役割：**プロジェクトが計画通りに進んでいるか管理し、各部署との調整を行い、定期的に経営層に進捗や課題を報告し、重要な意思決定をサポートします。現場の状況を把握し、プロジェクトを力強く前進させる「現場監督」の役割です。

【現場リーダー／アンバサダー】各部署での「推進役」（推奨）

各部署やチームの中に、AIに関心が高く、他のメンバーをサポートできるような「AI推進リーダー」や「AIアンバサダー」のような役割の人を置くことも、現場での活用を広げる上で効果的です。

・**役割：**自分の部署の仕事でAIをどう使えるか具体的に考えたり、メンバーからの質問に答えたり、ガイドラインが守られているか気を配ったり、成功事例や困りごとを推進チームに伝えたりする「現場の推進役」です。

【関係部門】それぞれの専門性を活かした連携

AI活用は、各部門がそれぞれの専門性を活かして協力することで、より安全かつ効果的に進められます。役割分担の例と

しては、次のようなものがあります。

- **IT部門：**AIを使うための技術的な基盤（システム、ネットワーク）の準備・運用、セキュリティ対策の実施、ツールの管理。
- **法務・コンプライアンス部門：**著作権、個人情報保護などの法律リスクのチェック、契約書の確認、ガイドラインが法的に問題ないかの確認。
- **人事部門：**AI活用に必要なスキルの明確化、教育・研修の企画・実施、採用活動、AI導入に伴う働き方の変化への対応。
- **広報部門：**社内へのAI活用方針の周知、社外への適切な情報発信（成功事例や取り組みなど）。
- **各事業部門：**自分たちの業務でAIをどう使えるかのアイデア出し、実際に使ってみての評価、効果測定、改善提案。

このように、それぞれの役割と責任を明確にし、互いに連携し合う体制を築くことが、組織的なAI活用を成功させる鍵となります。

6.3 全従業員のための利用ガイドライン・ポリシー策定

「守るべきルール」を明確に

従業員が安心して、かつ責任を持ってAIを使うためには、「何をして良くて、何をしてはいけないのか」「何に気をつけるべきなのか」という明確なルールが必要です。それが「利用ガイ

ドライン」や「ポリシー」です。ただ作るだけでなく、「使える」「守られる」ルールにするためのポイントを見ていきましょう。

■なぜルールが必要か？（目的と基本方針）

まず、「なぜこのルールが必要なのか」（例：リスクを防ぎ、安全にAIのメリットを享受するため）、「会社としてAI利用にどう向き合うか」（例:積極的に活用するが、ルールは厳守する）という基本的な考え方を最初に示しましょう。これがルールの土台となります。

■誰のためのルールか？（対象範囲）

正社員、契約社員、派遣社員、業務委託先など、このルールが誰に適用されるのかを明確にしておきます。

■どのAIツールを使って良いか？（利用ツール）

会社として利用を許可するAIツールやサービスを具体的にリスト化するのが分かりやすいでしょう（ホワイトリスト方式）。リストにないツールの業務利用は原則禁止とします。個人のアカウントで会社の業務に使う場合のルール（許可制にするか、禁止するかなど）も明確に定めます。

■何のために使って良いか？（利用目的）

AIを使って推奨される業務（例：情報収集、アイデア出し、文章作成補助、データ整理など）と、原則禁止する、あるいは特に注意が必要な業務（例：会社の最終的な意思決定、人事評価、差別や違法行為につながる利用など）を具体的に示します。

■「これだけは入力しないで！」（情報入力ルール、超重要！）

・**会社の秘密・個人情報は絶対NG！：**会社の重要な機密情報（顧客リスト、未公開の財務情報、技術情報など）や、個人情報（お客様や従業員の氏名、連絡先、プライベートな情報など）は、原則として外部のAIサービスに絶対に入力しないことを、最も重要なルールとして明確に規定します。ここが情報漏洩の最大の落とし穴です。

・**例外ルールと承認：**もし例外的に入力が必要な場合があるとしても、どのような条件なら許されるのか（例：情報を完全に匿名化するなど）、そして必ず上司や担当部署の事前承認を得る、という厳格なプロセスを定めます。

AIが作ったものの扱いは？（生成物の取り扱いとルール）

■著作権の確認

AIが作ったものが他人の著作権を侵害していないか、利用者が確認する責任があることを明確にします。確認方法の目安（似ているものがないかチェックするなど）も示します。特に会社のロゴとして使ったり、製品として販売したりする場合は、専門家への相談も含め、より慎重な確認プロセスを定めます。

■事実確認（ファクトチェック）の義務

AIの答えは鵜呑みにせず、必ず信頼できる情報源で事実確認を行うことを義務付けます。「AIが言っていたから正しい」は通用しません。

・**AI作成の表示：**必要に応じて（特に社外向けの文書など）、そのコンテンツがAIによって生成されたものであることを明記するルールを設けることも検討します。

■基本的なセキュリティルール

AIサービスにログインするためのアカウント（ID・パスワード）をしっかり管理すること。もしAIがおかしな動きをしたり、不審な情報を表示したりしたら、すぐに担当部署に報告すること、などを定めます。

■人としての配慮（倫理的配慮）

AIを使って、差別的な内容や、他人を傷つけるような内容を生成したり、広めたりしないこと。AIの利用が不公平にならないように注意すること、などを規定します。

■ルール違反したら？（違反時の措置）

ルールを守らなかった場合に、どのような注意や処分があり得るのかを明確にしておくことで、ルール遵守の意識を高めます。

■ルール作りの進め方と運用（関係部署と協力して作成）

法務、IT、人事、各事業部門など、関係者の意見を聞きながら、実用的で分かりやすいルールを作成します。

■経営層の承認

会社の公式ルールとして、経営層の承認を得ます。

■全員にしっかり伝える

作ったルールは、説明会や研修などを通じて、全従業員にその内容と重要性を確実に伝えます。いつでも見返せるように、社内ポータルなどに分かりやすく掲載します。

■定期的に見直す

AI技術や法律、社会状況は変化します。少なくとも年に1回程度は見直しを行い、必要に応じてルールを改訂していくことが重要です。

【評価目安（ガイドライン策定・周知）】

・導入のしやすさ：B（高い）

関係部署との調整や内容検討に時間はかかりますが、少しの準備や調整で策定に着手できます。既存の就業規則などを活用できる部分もあります。

・効果：A（非常に高い）

全従業員の行動基準となり、組織的なリスク管理の根幹をなすため、リスクをほぼ無くせる、または影響を最小限にできる非常に重要な取り組みです。

6.4 リテラシー向上のための継続的な教育・研修プログラム

「知っている」と「できる」は違う

どんなに素晴らしいルールやガイドラインを作っても、従業員一人ひとりがそれを「知らない」「理解していない」「守る意識がない」のでは、絵に描いた餅です。AIを安全かつ効果的に活用するためには、継続的な教育・研修を通じて、全従業員のAIリテラシー（AIを理解し、使いこなす能力）とリスクに対する感度を高めていくことが不可欠です。

教育・研修のねらい

・AIってそもそも何？　どう動くの？　何ができて、何が苦手なの？　を基本から理解する。
・会社が決めたAI利用のルール・ガイドラインとその理由をしっかり理解し、なぜそれを守る必要があるのかを納得する。
・情報漏洩、著作権侵害、誤情報、バイアスといった具体的なリスクと、それを避けるための自分の行動を学ぶ。
・自分の仕事でAIを安全かつ効果的に使うための基本的なコツ（例：良い指示の出し方）を身につける。
・AIを使う上での倫理的な配慮（公平性、差別防止など）について考えるきっかけを持つ。

誰に何を教える？（対象者別の設計）

全員に同じ内容ではなく、立場や役割に応じて内容を工夫すると、より効果的です。

■経営層向け

AIが会社の戦略にどう影響するか、AIガバナンスの重要性、投資判断のポイント、倫理的なリーダーシップなど。

■管理職向け

部下が安全にAIを使うための指導・管理方法、部門での活用推進のヒント、リスクを発見した場合の対応手順など。

■全従業員向け

AIの基本、会社のガイドライン、日常業務での具体的な注意点（情報入力、ファクトチェック、著作権など）、困ったときの相談先。

■開発・運用担当者向け

より専門的な内容として、安全なAIシステムの開発手法、脆弱性対策、データ管理、公平性確保の技術など。

どんな内容を盛り込むか？（内容例）

・本書第1章の「AIの基礎知識」
・本書第3章～第5章の「主要なリスク」とその「身近な事例」
・本書第6章6.5の「AI倫理の考え方」
・本書第7章の「問題発生時の報告・連絡手順」
・自社の「利用ガイドライン・ポリシー」の具体的な解説とQ&A
・「プロンプト作成の基本と注意点」
・「ファクトチェックの具体的なやり方」

どうやって教える？（効果的な実施方法）

■多様な方法を組み合わせる

全員必須のeラーニング、対話や質疑応答ができる集合研修、具体的なスキルを学ぶワークショップ、部署内での勉強会、成

功事例や失敗事例を共有する情報交換会、実際の業務を通じた指導（OJT）などを組み合わせると効果的です。

■「自分ごと」として捉えてもらう工夫

一方的な講義だけでなく、具体的な事例研究（ケーススタディ）、グループディスカッション、クイズ形式などを取り入れ、参加者が能動的に考え、自分自身の業務と結びつけて学べるように工夫します。

■「継続」が何より重要！

AI技術やリスクは常に変化します。一度研修をやっただけでは、知識はすぐに古くなり、意識も薄れてしまいます。定期的（例：年に1回）に研修を実施したり、新しい情報（新リスク、ガイドライン改訂など）が出たら随時アップデート情報を提供したりするなど、継続的な学びの機会を提供することが極めて重要です。

研修の効果（理解度、意識の変化など）をアンケートやテストで測定し、その結果を基に研修内容を改善していくサイクル（PDCA）を回しましょう。

【評価目安（継続的な教育・研修）】

・導入のしやすさ：B（高い）

企画や準備は必要ですが、eラーニングなど既存の仕組みを活用したり、外部研修を利用したりすることも可能です。

・効果：A（非常に高い）

従業員の意識と行動を変えることは、あらゆるリスク対策の基礎であり、リスクをほぼ無くせる、または影響を最小限にで

きる非常に効果の高い投資です。

6.5 AI倫理原則の導入と浸透

「正しい使い方」のその先へ

法律や会社のルールを守ることは当然として、さらに一歩進んで、「AIを倫理的に、社会的に責任ある形で使う」という視点を持つことが、これからの企業には求められます。それが「AI倫理」への取り組みです。

■なぜ「倫理」が重要なのか？

AIの判断や生成物は、時に意図せず不公平を生んだり、差別を助長したり、人の尊厳を傷つけたりする可能性があります。たとえ法律違反ではなかったとしても、倫理的に問題のあるAI利用は、社会からの厳しい批判を浴び、会社の評判を大きく損なう可能性があります。逆に、倫理に配慮し、「人間中心」でAIを活用する姿勢を示すことは、顧客や社会からの信頼を得て、持続可能なビジネスを行うための重要な基盤となります。また、倫理的な視点は、新しいリスクを予見したり、より良いAIサービスを生み出すイノベーションの源泉にもなりえます。

■どんな原則があるの？（参照すべき考え方）

世界中の政府機関や研究機関、企業などが、責任あるAIのための原則やガイドラインを発表しています。共通して重要とされる考え方には、以下のようなものがあります。

・**人間中心：**AIは、あくまで人間の幸福や社会全体の利益のために使われるべき。
・**公平性：**AIが、性別、人種、年齢などで人を不当に差別しないこと。
・**透明性・説明責任：**AIがなぜそう判断したのか、可能な限り説明できるようにすること、そしてその結果に責任を持つこと。
・**安全性・信頼性：**AIが安全に、意図した通りに動作すること。
・**プライバシー保護：**個人のプライバシーを尊重し、データを適切に保護すること（これらの考え方は、日本の総務省が出している「人間中心のAI社会原則」などでも示されています。付録A参照）。
・**自社としての「倫理の軸」を持つ：**これらの普遍的な原則を参考にしながら、自社の事業内容や企業理念、大切にしたい価値観などを踏まえ、「わが社はAI倫理についてこのように考え、行動します」という独自の原則や行動指針を策定し、社内外に示すことが望ましいでしょう。単なるお題目ではなく、従業員の日々の判断や行動につながるような、具体的で分かりやすい言葉で表現する工夫が大切です。

■どうやって会社に根付かせるか？（組織文化への浸透）

作った原則が「額縁の中の飾り」にならないように、組織文化として浸透させるための地道な努力が必要です。

・**トップからのメッセージ：**経営層が繰り返し、AI倫理の重要性を社内に訴えかけます。

- **研修での学びと対話：**従業員研修で倫理原則について解説し、具体的な事例をもとに「自分ならどう判断するか」を考える機会を作ります。
- **業務プロセスへの組み込み：**AIの開発や導入、利用に関するチェックリストや承認プロセスの中に、倫理的な観点からの確認項目を入れ込みます。
- **相談できる場づくり：**「このAIの使い方は倫理的に大丈夫だろうか？」と悩んだ時に、従業員が安心して相談できる窓口（例：倫理担当部署、専門の委員会）や、部署内で気軽に話し合える雰囲気を作ることが重要です。

【評価目安（AI倫理原則導入・浸透）】

・導入のしやすさ：C（中程度）

原則策定や浸透策の検討にはある程度の準備と議論が必要です。専門家の助けが有効な場合もあります。

・効果：A（非常に高い）

長期的な企業の信頼性、ブランド価値、従業員の意識向上につながり、潜在的なリスクを未然に防ぐ非常に高い効果が期待できます。

6.6 定期的なリスクアセスメントと監査体制

「大丈夫か?」を常にチェックする仕組み

AIガバナンス体制や、これまで述べてきた様々な対策が、実際にちゃんと機能しているか、形骸化していないかを定期的に確認し、必要に応じて見直し・改善していく仕組みが不可欠です。これがいわば、AI活用の「定期健康診断」と「内部チェック」にあたります。

●リスクアセスメント（定期的な危険度チェック）

- **目的：**AIの利用状況、新しいAI技術の登場、法律や社会状況の変化などを踏まえて、定期的に（例：年に1回、あるいは新しいAIツールを導入する際など）、「現在、自社にはどのようなAIリスクが、どの程度のレベルで存在するか？」を洗い出し、分析し、評価するプロセスです。
- **進め方：**第2章で紹介した「発生可能性」と「影響度」の考え方などを用いてリスクの優先順位をつけ、その結果に基づいて、「今ある対策は十分か？」「追加の対策は必要か？」などを検討し、対策計画を見直します。IT部門だけでなく、法務、人事、事業部門など、関係部署が連携して行うことが重要です。

【評価目安】

- **導入のしやすさ：B（高い）～C（中程度）**

既存のリスク管理プロセスに組み込むことも可能ですが、AI特有のリスク評価には専門家の助けが必要な場合もあります。

・効果：A（非常に高い）

リスクを早期に発見・評価し、的確な対策につなげるための基本であり、非常に重要です。

●監査体制（ルールが守られているかのチェック）

- **目的：**会社で定めたAI利用のガイドラインやポリシーが、実際に従業員によってきちんと守られているか、導入したセキュリティ対策が適切に運用されているか、実施している教育は効果が出ているかなどを、客観的な立場からチェック（監査）する仕組みです。
- **実施方法：**社内の内部監査部門が担当する、あるいは外部の専門家（監査法人やコンサルタント）に依頼して、定期的に（または抜き打ちで）監査を実施します。監査結果は経営層に報告され、問題点があれば改善が指示されます。

【評価目安】

・導入のしやすさ：C（中程度）

監査体制の構築や外部委託にはある程度の準備が必要です。

・効果：B（高い）

ルールや対策の形骸化を防ぎ、実効性を担保する上で高い効果があります。

●継続的なモニタリング（日々の見守り）

- **目的：**AIの利用状況（どのツールを、誰が、どのように使っているか）、システムログ、セキュリティアラート、従業員からのヒヤリハット報告などを日常的に監視・収集し、リスクの兆候やガバナンス上の問題点を早期に発見する体制です。

【評価目安】

・導入のしやすさ：B（高い）～ C（中程度）

監視ツールの導入や報告体制の整備に少しの準備や調整が必要です。

・効果：B（高い）

問題の早期発見・早期対応につながり、リスクが大きくなる前に対処できる可能性を高めます。

■PDCAサイクルで改善し続ける

これらのリスクアセスメント、監査、モニタリングの結果を踏まえて、課題を見つけ（Check）、改善策を計画し（Plan）、実行し（Do）、その効果をまた評価する（Check）、そしてさらに改善する（Act）というPDCAサイクルを継続的に回していくことが、AIガバナンス体制とリスク対策を常に最適な状態に保ち、進化させていくために不可欠です。

【章のまとめ】

本章では、個別のリスク対策を支え、組織全体で生成AIと安全かつ効果的に向き合うための共通基盤となる、「AIガバナンス」と「推進体制」の具体的な構築方法について解説しました。

明確な責任体制のもと、実効性のあるガイドラインを定め、それを継続的な教育によって従業員に浸透させ、AI倫理への配慮を組織文化とし、そして定期的なリスクアセスメントと監査によって常に状況を確認し改善し続ける――。これら一連の組織的な取り組みがあってこそ、第3章から第5章で見てきたような技術的な対策や運用上の工夫が、真に力を発揮するのです。

AIガバナンスは、決してAI活用を妨げる「ブレーキ」ではありません。むしろ、それは安心してアクセルを踏むための「頑丈な車体」であり、「信頼できるナビゲーションシステム」なのです。

次章（第7章）では、これらの対策を講じていてもなお発生しうる「万が一」の事態、すなわちインシデントに対して、組織としてどのように備え、冷静かつ的確に対応していくべきか、その具体的なフローについて解説します。

第7章

【共通対策②】

インシデント発生！その時のための対応フロー

第6章では、AIリスクの発生を未然に防ぐための組織的なルール作りや体制整備について解説しました。しかし、どんなに万全な備えをしていても、予期せぬ事故や問題、すなわち「インシデント」が起こってしまう可能性は、残念ながらゼロにはできません。それはまるで、どれだけ気をつけていても交通事故に遭う可能性があるのと同じです。

重要なのは、事故が起こらないように最大限努力すると同時に、「もし起こってしまったら、どうするか？」を事前に決めておき、いざという時に冷静に、迅速に、そして的確に対応できるように準備しておくことです。

本章では、AI活用に関連する「困った出来事」（インシデント）が発生した場合に、被害を最小限に食い止め、問題を解決し、信頼を回復するための組織的な対応手順（インシデントレスポンス）の基本的な流れと考え方について、ステップ・バイ・ステップで分かりやすく解説します。これは、経営層から従業員まで、組織の誰もが知っておくべき重要な知識です。

7.1 インシデントレスポンスの基本原則と体制

「転んだ後」の正しい起き上がり方

インシデントレスポンスって何？ 難しく考える必要はありません。インシデントレスポンスとは、会社にとって好ましくない「困った出来事」(例えば、情報漏洩、AIによる不適切発言、システムの不正利用、サービス停止など)が発生したことを察知してから、その悪影響をできるだけ小さくし、原因を突き止め、元通りにし、そして二度と同じことが起こらないように対策するまでの一連の活動のことです。「転んでしまった後、いかに上手に応急処置をし、原因を考えて再発を防ぐか」というプロセス全体を指します。

■対応時の心構え(基本原則)

いざ問題が発生すると、焦ったり、隠したくなったりするかもしれません。しかし、そんな時こそ以下の原則を思い出してください。

- **スピードが命！(迅速性)：**問題の発見が遅れたり、対応が遅れたりするほど、被害はどんどん広がります。早期発見・早期対応が鉄則です。
- **慌てず、正確に(正確性)：**噂や憶測に惑わされず、何が起きているのか事実を正確に把握することが重要です。関係者間での情報共有も、正確さを心がけましょう。

・**まず被害を食い止める（影響最小化/封じ込め）：**火事の時にまず延焼を防ぐように、問題の影響がそれ以上広がらないように、システムを一時的に止めたり、ネットワークから切り離したりといった応急処置（封じ込め）を行います。
・**安全を確認して元に戻す（復旧）：**問題の原因を取り除き、安全が確認できたら、できるだけ早く通常の業務状態に戻します。
・**「なぜ起きたか？」を徹底的に考える（再発防止）：**問題が解決したら終わりではありません。なぜその問題が起きたのか、根本的な原因を突き止め、同じ過ちを繰り返さないための恒久的な対策を講じます。
・**誠実に説明する（透明性と説明責任）：**隠し事は状況を悪化させます。影響を受けたお客様や従業員、関係機関などに対して、状況を正直に、分かりやすく説明し、会社としての責任を果たす姿勢が信頼回復につながります。

■「起きてから」では遅すぎる！　事前準備の重要性

インシデントは、ある日突然やってきます。その時に初めて「どうしよう？」と考えていては、対応が後手に回り、被害が拡大してしまいます。事前に「もしこんな問題が起きたら、誰が、何を、どのように対応するか」という計画（インシデント対応計画）を立て、必要な役割分担（対応体制）を決め、そして実際に訓練しておくことが、まさに重要です。

【評価目安（事前準備）】

・**導入のしやすさ：B（高い）～ C（中程度）**

計画策定や訓練には少しの準備や調整、場合によっては専門家の助けが必要ですが、やらないことのリスクの方がはるかに大きいと言えます。

・効果：A（非常に高い）

事前の備えは、インシデント発生時の混乱を抑え、迅速かつ適切な対応を可能にし、被害の影響を最小限にできる極めて効果の高い取り組みです。

■誰が対応するの？（対応体制）

インシデント対応には、様々な知識や権限が必要となるため、関係部署からメンバーを集めた専門チームを事前に決めておくことが理想的です。情報セキュリティ分野ではCSIRT（シーサート）と呼ばれることが多いですが、名称は問いません。重要なのは、

- **リーダーを決め、指揮命令系統を明確にすること。**
- **必要な専門家（例：IT・AI技術担当、法務・コンプライアンス担当、広報担当、人事担当、関連する事業部門担当など）をメンバーとして定義しておくこと。**
- **各メンバーの役割分担（情報収集、技術調査、法的判断、対外発表、社内連絡など）を決めておくこと。**

です。大企業でなくても、まずは兼務の担当者を決め、必要に応じて外部の専門家（弁護士、セキュリティ会社など）と連携できる体制を整えておくだけでも、いざという時の動きが大きく変わってきます。

7.2 フェーズ1：初動対応（発生直後～数時間以内）

まずは落ち着いて、被害を食い止める！

インシデント発生直後の数時間が、その後の展開を大きく左右します。パニックにならず、事前に決められた手順に従って、以下の行動を迅速に行います。

■「何かおかしいぞ？」を検知し、すぐに報告！

インシデントの最初の兆候は、システムからの警告アラートかもしれませんし、お客様からの問い合わせかもしれません。あるいは、従業員自身が「あれ？　AIの動きが変だ」「このデータ、見覚えがないぞ」と気づくこともあります。どんな小さな異常でも、「おかしい」と感じたら、決して放置したり隠したりせず、決められた報告ルート（直属の上司、IT部門、セキュリティ窓口など）に、できるだけ早く、正確に報告することが、全ての始まりです。「報告しにくい…」と感じさせない、風通しの良い組織文化も重要です（第6章参照）。

■状況把握と優先順位づけ（トリアージ）

報告を受けた担当者やチームは、まず「何が起きているのか？」「いつから？」「どこで？」「誰に影響がありそうか？」といった状況を可能な限り迅速に把握します。そして、その事態の緊急度や深刻度を見極め（これを「トリアージ」と呼びます）、複数の問題が同時に発生している場合などは、どれから

優先的に対応すべきか判断します。

■被害拡大を食い止める！（封じ込め）

火事の延焼を防ぐように、インシデントの影響がこれ以上広がらないように、応急処置を行います。

例：攻撃を受けているサーバーをネットワークから切り離す、問題のあるAIサービスを一時的に停止する、不正に使われているアカウントをロックする、誤情報が載ったウェブページを非公開にする、など。

ただし、止めすぎると業務への影響も大きくなるため、どの範囲をどう止めるかは、状況に応じた冷静な判断が必要です。事前にいくつかのシナリオを想定し、対応を決めておくと良いでしょう。

■緊急対策チームを招集！

事前に決めておいた手順に従い、必要なメンバーを招集し、緊急対策チーム（またはCSIRTなど）を立ち上げます。リーダーを中心に、役割分担と指揮命令系統を再確認し、組織的な対応を開始します。

■「証拠」をしっかり確保！（証拠保全）

後で「なぜ問題が起きたのか」を正確に突き止めるため、そして場合によっては法的な手続きのためにも、インシデントに関連する証拠を、消えたり改ざんされたりしないように、きちんと保全することが非常に重要です。

例：システムのログファイル（操作記録、アクセス記録など）、問題が発生した時の画面コピー（スクリーンショット）、関

連するメールやチャットの記録、使用していたパソコンやサーバー自体など。保全の方法も事前に決めておくとスムーズです。

■やったことを記録し始める！

「いつ、誰が何を発見し、誰に報告し、どんな指示があり、どんな対応をし、その結果どうなったか」といった一連の動きを、時系列で正確に記録し始めます。これは、後で状況を正確に把握し、関係者に説明し、そして改善策を考えるための、非常に重要な基礎資料となります。

7.3 フェーズ2:詳細調査と影響範囲特定（数時間〜数日以内）

何が起きたのか？　どこまで広がったのか？

応急処置で一息ついたら、次は問題の根本原因と被害の全体像を正確に把握するための詳細な調査フェーズに入ります。

■原因究明：「なぜ」「どのように」起きたのか？

初動対応で集めた「証拠」（ログ、データなど）を詳しく分析し、インシデントが「なぜ起きたのか（根本原因）」そして「どのようにして起きたのか（発生経路、プロセス）」を突き止めます。「探偵が証拠から犯人を特定する」ようなイメージです。

例：システムのログを解析して不正アクセスの痕跡を探す、AIモデルの動作記録を調べて予期せぬ動きの原因を探る、

関係者にヒアリングして操作ミスや手順に問題がなかったか確認する、など。

必要であれば、コンピュータ内部のデータを詳細に解析する専門的な技術（デジタル・フォレンジック）を持つ外部の専門家の協力を得ることもあります。

■影響範囲の特定：「誰に」「どこまで」影響が出たのか？

このインシデントによって、どのシステム、どのデータ、そしてどの利用者（お客様、従業員など）に、どのくらいの期間、どのような悪い影響が及んだのか（あるいは及ぶ可能性があるのか）、その範囲と深刻度を正確に調査・特定します。

特に情報漏洩の場合は、「どのような種類の情報が」「誰の情報が」「何件くらい」漏洩した（可能性がある）のかを特定することが、その後の法的対応や本人への通知のために極めて重要になります。

この問題の影響が、他のシステムや業務プロセスにまで波及していないかも確認します。

■弱点の特定（脆弱性特定）

もしインシデントの原因が、システムやソフトウェア、あるいは運用プロセス上の「弱点（脆弱性）」にあると判明した場合は、その具体的な箇所を特定します。これが再発防止策の重要な手がかりとなります。

7.4 フェーズ3:関係者への通知と公表(数日〜数週間以内)

誠実なコミュニケーションで信頼回復へ

調査によって事実関係が明らかになってきたら、次は影響を受けた可能性のある方々や、社会全体に対して、必要な情報を適切なタイミングと方法で伝えるフェーズです。この段階でのコミュニケーションの取り方が、企業の信頼回復を大きく左右します。

■誰に何を伝えるべきか?(通知・報告義務の確認)

まず、法律(個人情報保護法など)、契約、業界ルールなどによって、誰か(例:国の監督官庁、警察、影響を受けた本人、取引先など)に、このインシデントについて報告・通知する義務があるかどうか、もしあるなら何を、いつまでに、どのように報告・通知しなければならないのかを確認します。法務部門との連携が不可欠です。

■どう伝えるか?(コミュニケーション戦略)

誰に対して、どのタイミングで、どのような内容を、どの方法(例:個別の手紙やメール、ウェブサイトでの告知、記者会見など)で伝えるのが最も適切か、一貫性のあるコミュニケーション戦略を立てます。

伝える内容は、判明している事実関係、原因、影響、会社の対応状況、今後の対策、そして問い合わせ先などを、正確に、

かつ分かりやすく整理します。専門用語は避け、誠意が伝わる表現を心がけます。

広報部門が中心となり、法務部門や経営層と緊密に連携して、メッセージの内容や伝え方を慎重に決定します。

■伝えるべき人に、誠実に伝える（通知・報告の実施）

法令などで義務付けられている報告・通知は、期限内に必ず実施します。

特に、影響を受けたお客様や従業員への通知は、最大限の誠意と丁寧さをもって行います。状況に応じて、専用の問い合わせ窓口や相談窓口を設置し、不安や疑問に寄り添う姿勢が大切です。

■世の中に知らせるべきか？（公表の検討・実施）

インシデントの内容や規模、社会的な影響の大きさなどを考慮し、自主的に情報を公表するかどうかを判断します。近年は、問題を隠さず、透明性を重視して積極的に情報を公開する企業姿勢が、むしろ信頼につながるという考え方が主流になっています。

公表する場合は、不確かな情報や憶測が広まらないよう、事実に基づいた正確な情報を、適切なタイミングと方法（例：プレスリリース、自社ウェブサイト）で発信します。技術的な詳細をどこまで公開するかは、さらなるリスクを招かないよう慎重に検討します。

■問い合わせに対応する

設置した窓口には、様々な問い合わせが寄せられる可能性が

あります。迅速かつ丁寧に対応できるよう、FAQ（よくある質問とその回答）を作成・公開したり、対応担当者への情報共有を徹底したりします。

7.5 フェーズ4：再発防止策の実施と継続的改善

同じ過ちを繰り返さないために

インシデント対応は、問題が収束し、関係者への説明が終わったら完了、ではありません。むしろここからが重要です。なぜ問題が起きたのかを深く反省し、同じ過ちを二度と繰り返さないための仕組みを作り、組織として成長していくことが求められます。

■「なぜ？」を繰り返す（根本原因の分析）

調査結果をもとに、表面的な原因だけでなく、「なぜそれが起きたのか？」を繰り返し問いかけ、問題の根本にある原因（例：技術的な欠陥だけでなく、チェック体制の不備、従業員の知識不足、組織間の連携不足、無理なスケジュール、リスクを軽視する文化など）を突き止めます。

■具体的な「次の一手」を決める（再発防止策の策定）

特定された根本原因を取り除くための、具体的で実行可能な再発防止策を計画します。技術的なシステムの改善、業務プロセスの見直し、ガイドラインの改訂、従業員教育の強化、組織体制の変更など、多角的な視点から検討します。

■決めたことを実行する（再発防止策の実施）

計画した対策を、「誰が」「いつまでに」行うのか、責任者と期限を明確にして、着実に実行に移します。経営層のリーダーシップのもと、全社的に取り組むことが重要です。

■「やりっぱなし」にしない（有効性の評価と見直し-PDCA）

実施した対策が、本当に効果を発揮しているか、形骸化していないかを、定期的に（例：次回の監査やリスクアセスメント時）評価します。効果が不十分であれば、さらなる改善策を検討します。

このように、計画（Plan）→実行（Do）→評価（Check）→改善（Act）のサイクル（PDCA）を回し続けることが、継続的なリスク低減につながります。

■失敗を「学び」に変える（教訓の共有と活用）

今回のインシデント対応で得られた教訓（うまくいったこと、うまくいかなかったこと、改善すべき点など）を、報告書などにまとめ、隠すことなく組織全体で共有します。「失敗は成功のもと」として、今後のリスク管理活動、従業員教育、システム開発などに活かしていく文化を育てることが大切です。

■「備え」をアップデートする（対応計画・体制の更新）

今回の経験を踏まえ、インシデント対応計画や対応体制そのものに不備や改善点がなかったかを見直し、より実効性の高いものへとアップデートしていきます。

7.6 各リスクに応じた対応のポイント

「火事」と「事故」では対応が違う

本章で解説してきた「初動対応→詳細調査→通知・公表→再発防止」という4つのフェーズは、あらゆるインシデント対応の基本となる共通の流れです。

しかし、実際に発生したインシデントの種類によって、特にどこに力を入れるべきか、誰と連携すべきか、どのような点に注意すべきか、といった対応の「焦点」は異なります。例をあげていきましょう。

■もし「情報漏洩（特に個人情報）」が起きたら…

- **第3章 3.5参照**→法律（個人情報保護法）で定められた報告・通知義務を、期限内に確実に果たすことが最優先事項の一つです。法務部門との連携が極めて重要になります。

■もし「著作権侵害」を指摘されたら…

- **第4章 4.7参照**→権利を持っている相手との誠実な対話と交渉が対応の中心になります。弁護士などの専門家への相談も早期に検討すべきでしょう。

■もし「誤情報・偽情報」を広めてしまったら…

- **第4章 4.7参照**→できるだけ早く正確な情報で訂正し、これ以上拡散しないように手を打つこと、そして信頼回復のための丁寧なコミュニケーションが鍵となります。広報部

門の役割が重要です。

■もし「バイアス・差別」的な内容が問題になったら…

・**第4章 4.7参照**→影響を受けた方々への真摯な謝罪と配慮、そして公平性を確保するための組織的な改善に取り組む姿勢を示すことが求められます。倫理担当や人事部門との連携も必要です。

■もし「モデルへの攻撃」を受けたら…

・**第5章 5.8参照**→技術的な「封じ込め」と原因究明、そしてシステムの脆弱性を修正することが急務となります。セキュリティ専門チーム（CSIRTなど）やIT部門が主導します。

■もし「予期せぬ動作」が発生したら…

・**第5章 5.8参照**→なぜそうなったのか、原因の切り分け（データ？　モデル？　環境？）を慎重に行い、安全を確認した上で修正やロールバックを行う判断が必要です。開発部門との連携が鍵です。

■もし「AI過信」によるミスが発覚したら…

・**第5章 5.8参照**→技術的な問題だけでなく、利用者の判断プロセスや、それを助長した組織的な要因にまで踏み込んで調査し、教育やガイドライン、業務プロセスを見直す必要があります。

■もし「システム停止」が起きたら…

・**第5章 5.8参照**→何よりもまず、事業を継続させるための「サービス復旧」が最優先課題となります。復旧チームと代替手段への移行プロセスが重要です。

このように、発生したインシデントの種類に応じて、対応のポイントや関係者が異なります。各リスク領域に特有の対応ポイントについては、第3章（3.5節）、第4章（4.7節）、第5章（5.8節）でそれぞれ解説していますので、本章で学んだ基本フローとあわせて参照し、自社のインシデント対応計画に具体的に落とし込んでみてください。

【章のまとめ】

どんなに注意していても、予期せぬ「困った出来事（インシデント）」は起こりうるものです。しかし、大切なのは、発生した時に組織としていかに冷静に、迅速に、そして的確に対応できるか、そのための「備え」を日頃からしておくことです。

本章では、インシデント対応の基本的な考え方と、具体的な4つの対応フェーズ（初動→調査→通知→再発防止）について解説しました。このフローを理解し、自社の状況に合わせて具体的な計画を立て、担当チームを決め、そして定期的に訓練を行うこと。これが、AI活用時代の必須の「備え」と言えるでしょう。

インシデント対応は、決してネガティブなだけの活動ではありません。それは、問題から学び、組織の弱点を克服し、より強く、より信頼される企業へと成長するための重要な機会でもあるのです。

次章（第8章）では、これまで学んできたリスク管理、ガバナ

ンス、そしてインシデント対応への備えを踏まえながら、実際に生成AIの活用に挑戦し、成果を上げている（あるいは試行錯誤している）企業のリアルな事例から、私たち自身のAI活用戦略に向けた実践的なヒントを探っていきます。

第8章

導入成功への道標

リスクに配慮した企業活用事例に学ぶ

これまでの章で、生成AIの基礎知識から始まり、様々なリスクとその対策、組織として備えるべきガバナンス体制、そして万が一の際の対応フローまで、理論的な側面を中心に学んできました。しかし、「理屈は分かったけれど、実際に他の会社はどうやっているのだろう？」という疑問は当然湧いてきますよね。

本章では、その疑問にお答えするため、実際に生成AIの活用に取り組み、同時にリスクにも真摯に向き合っている企業の「リアルな事例」（公開されている情報に基づきます）を、いくつかのテーマに分けてご紹介します。先進的な企業の具体的な取り組み、試行錯誤の様子を知ることは、皆さんが自社でAI導入・活用戦略を考え、実行していく上で、非常に貴重なヒントと勇気を与えてくれるはずです。理論と実践を結びつけ、自社ならではの成功への道筋を描くための「道標」として、本章をご活用ください。

【重要】本章で紹介する事例は、各社の公式発表（ニュースリリース、公式ブログ、統合報告書など）や、信頼できる報道機関によって公開された情報に基づいて記述しています。情報の正確性には最大限配慮していますが、企業の取り組みは常に変化していますので、詳細や最新状況については、必ず各社の公式情報をご確認いただくようお願いいたします。憶測や未確認情報に基づく記述は一切しておりません。

8.1 事例から学ぶことの重要性

他社の経験を「自分の知恵」に変える

なぜ、他社の事例を知ることが、自社のAI活用にとってそれほど有益なのでしょうか？

■教科書だけでは分からない「現実」が見える：本書で解説してきたようなリスクや対策が、実際のビジネスの現場でどのような形で現れ、企業がどのように悩み、判断し、対処しているのか、その「生々しさ」を知ることができます。これにより、抽象的な理解が、具体的な行動イメージへと変わります。

■実践的な「引き出し」が増える

他社が試した具体的な工夫、導入したツール、構築したプロセス、あるいは直面した想定外の課題などを知ることで、「こんなやり方があったのか！」「うちでも応用できそうだ」といった、自社での取り組みに応用できる実践的なアイデアやヒントの「引き出し」を増やすことができます。

■自社に合った「最適解」を見つけるヒントになる

様々な業種や規模の企業の事例に触れることで、「自社の状況に近いのはこのケースだ」「この会社の考え方は参考になる」「あえて違うアプローチを取るべきかもしれない」といったように、自社ならではの課題や目標に照らし合わせて、最適な活用方法やリスク対策を考える上での重要な比較検討材料とな

ります。

■「自分たちだけではない」という安心感と勇気

新しい技術への挑戦には、不安がつきものです。しかし、多くの企業が、リスクがあることを認識しながらも、試行錯誤を繰り返し、前向きにAI活用に取り組んでいることを知ることで、「難しいのはウチだけじゃないんだ」「失敗しても、そこから学べばいいんだ」という安心感を得られ、「自社でも挑戦してみよう！」という勇気が湧いてくるはずです。

他社の成功事例はもちろん、時には失敗事例（あるいはそこから得られる教訓）からも、自社が同じ過ちを繰り返さないための重要な学びを得ることができます。他社の経験を、ぜひ自社の「知恵」へと変えていきましょう。

8.2 【テーマ別事例】著作権・セキュリティ
「守り」を固める取り組み

AI活用の「攻め」を支えるためには、著作権侵害や情報漏洩といったリスクに対する「守り」を固めることが不可欠です。これらの課題に正面から向き合っている企業の取り組みを見てみましょう。

著作権への配慮：安心して使えるAIを目指して

●Adobe社の例（Firefly）

画像生成AI「Firefly」を提供するAdobe社は、その学習データについて、自社が権利を持つストックフォトサービス「Adobe Stock」の許諾済み画像や、著作権保護期間が終了したパブリックドメインの画像などに限定していると公表しています。これは、AIが生成した画像が第三者の著作権を侵害するリスクを根本から低減し、利用者が安心して（特に商用目的で）使えるようにするための重要な取り組みです。さらに、企業向けプランでは、生成物が原因で著作権侵害の訴訟を起こされた場合に、Adobe社が補償するというプログラムも提供しており、利用企業の不安解消に努めています。

- **学びのポイント：**AIサービスを選ぶ際に、学習データの内容や権利処理の方針について、提供元がどの程度透明性を持ち、配慮しているかを確認することが重要であることを示唆しています。また、提供元による補償制度の有無も、リスクを判断する上で参考になります。

●Microsoft社の例（Copilot）

同社が提供する生成AIサービス「Copilot」の企業向け有料版の利用者に対して、「Copilot Copyright Commitment」というプログラムを発表しています。これは、利用者がCopilotの生成物を利用した結果、著作権侵害で訴えられ

た場合に、Microsoft社が法的な防御や費用の負担を行うというものです（一定の条件あり）。これも、企業が著作権リスクを過度に恐れることなく、AI活用を進められるように支援する取り組みと言えます。

- **学びのポイント：**AIサービス提供者側も、企業ユーザーの著作権リスクに対する懸念を認識し、それに対応する動きを見せていることが分かります。ただし、補償の対象となる条件等は詳細に確認する必要があります。

●一般的な動向

出版社や報道機関といったメディア企業では、記事の要約やアイデア出しに生成AIを活用する際、著作権法やガイドラインを遵守するための社内ルールを整備したり、引用元を明記したり、場合によっては原著作物の権利者から事前に許諾を得るといったプロセスを検討・導入する動きが見られます。また、デザイン会社や広告制作会社などでは、AIが生成したものをそのまま使わず、必ず人間のデザイナーやクリエイターが大幅に手を加え、独自の創作性を付与することで、著作権侵害のリスクを低減しようと努めているケースも多いようです。

- **学びのポイント：**サービス提供者の努力に加えて、AIを利用する企業側でも、利用目的やコンテンツの性質に応じた適切な権利処理プロセスや、人間によるチェック・加工といった運用上の工夫が重要であることを示しています。

情報セキュリティへの配慮：会社の重要情報を守るために

●金融機関（メガバンクなど）の例

顧客の大切な資産や個人情報、そして市場の信頼を預かる金融機関は、生成AIの活用においてセキュリティとガバナンス（統制）の確保を最優先課題としています。そのため、多くのメガバンクなどでは、インターネット経由で誰でも利用できるようなパブリックなAIサービスの業務利用を原則禁止または厳しく制限していると報じられています。

（1）プライベート環境での利用

Microsoft社の「Azure OpenAI Service」のように、入力したデータがAIモデルの再学習に使われず、かつ、自社の閉じたネットワーク（閉域網）から安全に接続できるような、セキュリティが強化されたクラウドサービスを選択・利用する。

（2）自社専用環境の構築

あるいは、外部のクラウドサービスすら利用せず、自社のデータセンター内（オンプレミス）や、自社専用のクラウド環境（プライベートクラウド）に独自のAI基盤を構築し、データの流れを完全にコントロール下に置く。

このようなアプローチを検討・推進しているほか、厳格な行内利用ルールの策定（例：入力できる情報の種類を限定、利用目的を制限）や、全行員へのセキュリティ教育の徹底にも力を入れています。

・**学びのポイント：**扱う情報の機密性が極めて高い場合、AIを利用する「環境」そのもののセキュリティをいかに確保するかが最重要課題となります。利便性とのトレードオフになりますが、リスクレベルに応じて、プライベート環境やオンプレミス環境の利用を検討する必要性を示唆しています。

●一般的な動向

金融機関ほどではないにしても、製造業の研究開発部門や、法務・知財部門など、特に機密性の高い情報を扱う部署では、外部AIサービスへの機密情報の入力を厳しく禁止したり、どうしても利用が必要な場合はデータを徹底的に匿名化やマスキングをしたりするルールを設けている企業が多いようです。また、AIシステムや関連データへのアクセス権限を役職や担当業務に応じて最小限に絞り、誰がいつ何を利用したかのログ（利用記録）を監視する体制を整備することも、基本的なセキュリティ対策として広く行われています。

・**学びのポイント：**第3章で解説した、アクセス制御、データ保護（匿名化・マスキング）、そして利用状況の監視といった基本的な情報セキュリティ対策は、AI利用においても同様に、あるいはそれ以上に重要であることを再認識させてくれます。

8.3 【テーマ別事例】ガイドライン策定と全社教育による推進例

「全社共通の羅針盤」を作る

AIを一部の部署だけでなく、会社全体で安全かつ効果的に活用していくためには、全従業員が守るべき共通のルール（ガイドライン）と、それを理解し実践するための教育が不可欠です。この「組織的な土台作り」に積極的に取り組んでいる企業の事例です。

●パナソニック コネクト社の例

この会社は、比較的早い段階（2023年初頭）から、全従業員（当時国内約1万人）が業務で生成AI（ChatGPT）を利用する際の詳細なガイドラインを策定し、社内に公開しました。その内容は、

- **利用目的の明確化：**何のために使うのかを意識する。
- **情報入力の禁止事項：**機密情報、個人情報、社外秘情報は絶対に入力しない。
- **出力内容の注意点：**生成された情報の正確性や著作権、倫理的な問題がないか、必ず人間が確認し責任を持つ。

といった、基本的ながら非常に重要な点が網羅されています。さらに注目すべきは、ガイドライン策定と同時に、全社でAI活用のアイデアを競う「AI活用コンテスト」を開催するなど、リスク管理（守り）と活用促進（攻め）の両輪で取り組みを進め

ている点です。

・**学びのポイント：**AI導入初期において、まず全社共通の明確なルールを定めることの重要性を示しています。また、ルールで縛るだけでなく、積極的な活用を促す施策を組み合わせることで、従業員のAIへの抵抗感を和らげ、前向きな活用文化を醸成できる可能性を示唆しています。

●富士通の例

グループ従業員（国内約12万人）を対象に、生成AI利用ガイドラインを策定し、研修などを通じて周知を図っています。ガイドラインでは、情報漏洩を防ぐための具体的な入力情報の注意点（例：「個人を特定できる情報は入力しない」「社外秘情報は入力しない」など）や、著作権侵害、誤情報、バイアスといったコンテンツに関するリスクへの留意事項を具体的に示しています。単に利用を制限するだけでなく、業務効率化や新しい発想を生むためのツールとして、積極的な活用も奨励している点が特徴です。

・**学びのポイント：**大規模な組織でAI活用を進めるためには、具体的で分かりやすいガイドラインと、それを補完する継続的な教育・研修が不可欠であることを示しています。「禁止事項」だけでなく「推奨される使い方」も示すことで、従業員が安心して活用できるよう配慮しています。

●ソフトバンクの例

社内業務でのAI活用を加速させるため、独自のセキュアな生成AIプラットフォームを構築・提供するとともに、全従業員が遵守すべき「AI活用心得」というガイドラインを策定しています。情報セキュリティや著作権といった基本的なルールに加え、「AIの限界を理解すること」「最終的な判断と責任は人間が持つこと」といった、AIとの向き合い方に関する心構えを強調している点が特徴的です。

- **学びのポイント：**セキュアな利用環境（プラットフォーム）の提供と、利用者の心構え（心得）の両面からアプローチすることの有効性を示唆しています。特に「AIは万能ではない」という認識を組織全体で共有することの重要性がうかがえます。

●その他多くの企業の動向

上記以外にも、日立製作所、NTTグループ、博報堂DYホールディングス、さらには業種を問わず多くの上場企業などが、自社の事業特性や企業文化に合わせてカスタマイズされた生成AI利用ガイドラインやポリシーを策定し、社内イントラネットなどで公開・周知する動きが急速に広がっています。その多くに共通して、「機密情報・個人情報の入力制限」「著作権・個人情報保護への配慮」「出力情報の正確性確認」「倫理的な利用」といった項目が含まれています。

- **学びのポイント：**もはや、AI利用に関する社内ルール作り

は、企業規模や業種を問わず、必須の取り組みとなっていることが分かります。重要なのは、他社の真似をするだけでなく、自社の状況に合った、具体的で、かつ従業員が理解しやすいルールを作成し、それを形骸化させずに運用していくことです。

8.4 【テーマ別事例】AI倫理・公平性確保への挑戦例
「正しいAI」を目指して

生成AIの利用は、法律を守るだけでは十分ではありません。社会の一員として、「倫理的に正しいか？」「誰かを不当に扱っていないか？」といった観点からの配慮、すなわち「AI倫理」への取り組みが、企業の信頼を左右する時代になっています。

●Googleの例（AI Principles）

AI開発のリーディングカンパニーであるGoogleは、比較的早い2018年に、自社がAI開発・利用において遵守すべき7つの「AI原則（AI Principles）」を発表しました。その中には、「社会的に有益であること」「不公平なバイアスを生み出さない、または助長しないこと」「安全性」「人間に対する説明責任を果たすこと」「プライバシーに配慮した設計を行うこと」などが含まれています。同社は、これらの原則に基づいてAI技術の開発や提供を進めているとしており、AI倫理に関する

研究や、公平性を評価するためのツール開発などにも積極的に取り組んでいます。

- **学びのポイント：**企業がAI開発・利用における基本的な価値観や行動規範（プリンシプル）を明確に定め、それを社内外に示すことの重要性を示唆しています。倫理原則が、技術開発の方向性や社員の行動の拠り所となります。

●Microsoft社の例（Responsible AI）

Microsoft社も、「責任あるAI（Responsible AI）」というコンセプトを掲げ、「公平性」「信頼性と安全性」「プライバシーとセキュリティ」「包括性（インクルーシブネス）」「透明性」「説明責任」という6つの原則を定めています。さらに、これらの原則を実際の開発現場で実践するための具体的な社内基準「Responsible AI Standard」を設け、開発プロセスへの組み込み、社内の専門家によるレビュー体制、関連ツールの提供など、原則を具体的な行動に落とし込むための仕組み作りに力を入れています。

- **学びのポイント：**倫理原則を掲げるだけでなく、それを実際の業務プロセスや組織体制の中に組み込み、実践していくための具体的な仕組みを構築することの重要性を示しています。「言うは易く行うは難し」を乗り越えるためのヒントとなります。

●NTTグループの例（AI倫理ガイドライン）

日本企業であるNTTグループも、「NTTグループAI倫理ガイドライン」を策定・公開しています。この中では、「人間中心の原則」「公平性の原則」「透明性の原則」「セキュリティとプライバシーの原則」「アカウンタビリティ（説明責任）の原則」などが定められており、AIの開発者や利用者が、日々の業務の中で倫理的な課題を考慮するための具体的な指針として位置づけられています。

- **学びのポイント：**日本企業においても、AI倫理への取り組みが経営上の重要課題として認識され、具体的なガイドライン策定が進んでいることが分かります。自社の文化や価値観に合わせた原則作りが可能です。

●一般的な動向

上記のような大企業の取り組みに加え、より一般的に見られる動きとして、AIの開発チームにエンジニアだけでなく、多様な専門分野（社会学、倫理学、法律など）や多様な属性（性別、人種、文化背景など）を持つメンバーを加えることで、開発段階から多角的な視点を取り入れ、潜在的なバイアスや倫理的課題に気づきやすくしようとする試みがあります。また、AIの出力におけるバイアスを検出・評価するためのツールを導入したり、採用や人事評価といった特に影響の大きい分野でのAI利用には極めて慎重な姿勢をとったりする企業も増えています。AI倫理に関する専門委員会を社内に設置し、定期的に議論や

レビューを行う体制を構築する企業も現れ始めています。

・**学びのポイント：**AI倫理や公平性の確保は、技術的な対策だけでなく、組織の多様性、専門家の知見、そして利用する場面の慎重な判断といった、多岐にわたるアプローチが必要であることを示唆しています。

8.5 成功と失敗から得られる実践的な教訓

自社への活かし方

さて、これまで様々な企業の先進的な取り組み事例を見てきました。これらの事例全体を俯瞰し、生成AI導入を成功させ、リスクを管理していくための普遍的な教訓を抽出してみましょう。

ここが成功の鍵だ！（共通する成功要因）

■経営トップの本気度

やはり、経営層がAIの重要性を理解し、「会社を変えるぞ！」という強いリーダーシップを発揮している企業は、導入がスムーズに進む傾向にあります。明確な方針と予算配分が、現場の推進力となります。「あなたの会社では、経営層はAIに本気ですか？」

■部門間の協力体制

AI活用はIT部門だけではできません。現場の業務を知る事

業部門、法律リスクを見る法務部門、人材育成を担う人事部門などが、壁を越えて協力し合う体制が築けているかが重要です。「あなたの会社では、部門間の連携はスムーズですか？」

■明確なルールと全員への浸透

利用に関する分かりやすいガイドラインがあり、それが研修などを通じて全従業員にきちんと理解され、守られていること。これが組織的なリスク管理の基本です。「あなたの会社には、守られるルールがありますか？　全員が理解していますか？」

■小さく始めて、学びながら進める

最初から全社で大規模に導入するのではなく、まずは特定の部署や業務で試してみて（スモールスタート）、その効果や課題を検証し、学びながら徐々に展開していくアプローチが、結果的に成功につながりやすいようです。「あなたの会社では、小さく試せる環境がありますか？」

■リスクへの感度が高い

AIの便利な面だけでなく、常に潜在的なリスク（情報漏洩、著作権侵害、誤情報など）を意識し、それに対する事前の対策や、万が一問題が起きた時の対応計画を準備していること。「あなたの会社では、リスクへの備えはできていますか？」

■常に学び、改善し続ける姿勢

AI技術も社会も常に変化します。一度決めたルールや導入したシステムに固執せず、最新情報を学び、状況に合わせて柔軟に見直し、改善し続ける文化があることが、長期的な成功の鍵となります。「あなたの会社は、変化に対応し、学び続ける

組織ですか？」

これは避けたい！（注意すべき失敗パターン）：見切り発車での導入

リスクを十分に評価したり、ルールを定めたり、従業員に必要な教育を行ったりする前に、「流行っているから」「競合がやっているから」と焦ってツールだけ導入してしまうと失敗を招きます。

■リスクの過小評価

「うちは大丈夫だろう」「そんなことは滅多に起こらない」と、情報漏洩や著作権侵害などのリスクを甘く見て、十分な対策を怠る。

■AIへの過信・丸投げ

AIの能力を信じすぎて、人間によるチェックや最終判断を省略してしまい、重大なミスやトラブルを引き起こす。あるいは、現場にAI活用を「丸投げ」し、必要なサポートやルールを提供しない。

■（参考事例）Samsung電子での情報漏洩

この有名な事例（報道ベース）では、従業員が業務上の機密情報を外部のAIサービスに入力してしまったとされています。これは、明確なガイドラインの欠如や、従業員の危機意識の不足が招いた典型的な失敗パターンと言え、私たちに大きな教訓を与えてくれます。

これらの成功要因と失敗パターンは、本書でこれまで解説してきたリスク対策や組織体制の重要性を、実際の企業の経験を通じて裏付けていると言えるでしょう。

【章のまとめ】

本章では、公開されている情報に基づき、リスクに配慮しながら生成AIの活用に取り組む企業の実際の事例を、著作権・セキュリティ、ガイドライン・教育、AI倫理・公平性というテーマ別に紹介し、そこから得られる実践的な教訓を考察しました。

他社の取り組みは、私たちがAI活用を進める上での道標となり、多くのヒントを与えてくれます。しかし、最も重要なのは、これらの事例から得た学びを、そのまま真似するのではなく、自社の事業内容、規模、企業文化、そしてリスクに対する考え方（リスク許容度）といった、自分たちの状況に合わせて「翻訳」し、取捨選択し、応用していくことです。

さあ、他社の経験という「地図」を手に、自社ならではのAI活用戦略を描き、導入への具体的な一歩を踏み出す準備はできたでしょうか？

次章（第9章）では、AIを活用したセキュリティ強化技術について解説していきます。

第9章

プロンプトからの情報漏洩を防ぐ

AI活用による セキュリティ強化技術の実際

第8章では、様々な企業がリスクに配慮しながらAI活用を進めている事例を見てきました。そこからは、技術的な対策だけでなく、組織的な取り組みがいかに重要であるかが読み取れたかと思います。しかし、どんなにルールを徹底しても、日々の業務の中で従業員がプロンプトに機密情報を入力してしまう「うっかりミス」のリスクを完全には排除できません。

この課題に対する具体的な解決策の一つとして、本章ではAI（LLM）を活用して**プロンプトの安全性をチェックする「AIセキュリティチェックサーバー」の技術**について、その仕組み、運用形態、そして導入のポイントを、図解も交えながら詳しく解説します。このような技術がどのように機能し、企業のAI活用における安全性をどう高めるのか、その全体像を掴んでいきましょう。

9.1 なぜプロンプトチェックが重要なのか？

本書で繰り返し述べてきたように、生成AI利用における情報漏洩の主要経路の一つが「プロンプト入力」です。従業員が悪意なく入力した情報が、外部のAIサービス（特に海外サーバー）に送信され、処理・記録される過程で漏洩するリスクは、企業にとって大きな脅威です。

組織的な対策（ガイドライン、教育）はもちろん重要ですが、それらを**技術的に補完し、情報が社外に出る前の「水際」でチェックを行う仕組み**があれば、より安心して生成AIを活用できる環境を構築できます。本章で紹介する「プロンプトチェックAI」のような技術は、まさにその“賢い門番”としての役割を担うことが期待されています。

9.2 プロンプトチェックAIの仕組み

リスクレベル判定と信頼性向上

この技術の中核は、ユーザーが入力したプロンプトの内容を、「AI専用セキュリティサーバー」（多くはLLMベース）がリアルタイムで解析し、その内容に機密情報や個人情報が含まれていないか、どの程度のリスクがあるかを評価・判定する点にあります。

■リスクレベルによる評価

チェック用AIは、情報の機密性や漏洩した場合の影響度に応じて、プロンプトのリスクを段階的に評価します。これにより、一律にブロックするのではなく、リスクレベルに応じた柔軟な制御が可能になります。本書で想定する評価レベルの考え方と具体例を図9-1に示します。

・図9-1

■判定の信頼性を高める工夫

AIによる判定は常に100％正確とは限りません。微妙な表現や文脈によって、リスクレベルの判定結果が揺らぐことも考えられます。そのため、より安全な運用を目指し、判定の信頼性を高めるための工夫が凝らされることがあります。例えば、**一つのプロンプトに対してチェックを複数回行い、その結果を統合して最終的なリスクレベルを決定するアプローチ**です。その際の判断ロジックとして、「安全側に倒す」考え方が採用されることがあります。図9-2にその判定例を示します。

・図9-2

この図にあるように、複数回の結果の中で最も多く判定されたレベルを基本としつつ、もし判断が同数で割れた場合や、判定結果がバラけた場合には、最もリスクが高いレベルを採用す

るといったルールを設けることで、リスクの見逃しを最小限に抑え、より堅牢なチェックを実現しようとしています。

9.3 運用方式の選択肢

データの流れとセキュリティ・コスト

このプロンプトチェック機能をどのように導入・運用するかについては、主に3つの方式が考えられます。図9-3は、一般的な流れとセキュリティチェックサーバーを使用した場合、その全体構図とデータの流れ、チェックポイントを示しています。それぞれの特徴を理解し、自社のニーズに合ったものを選択することが重要です。

■一般の場合

①チェックなし

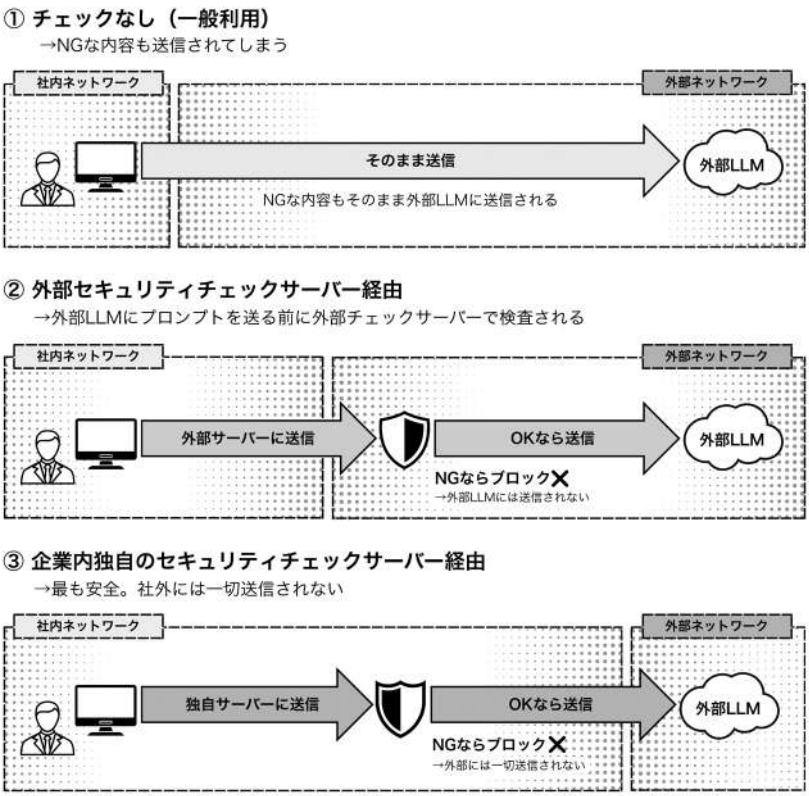
- **データの流れ：**プロンプトはLLMによる高度なチェックは行わず、基本的なNGワードチェック等を経て、直接外部LLMへ送信されます。
- **特徴：**最もシンプルでコストもかかりませんが、防御は限定的です。リスクの低い用途向けです。

■セキュリティチェックサーバー有りの場合

②外部セキュリティチェックサーバーの流れ（※ここではチャプロサーバーを例としてあげています）

- **データの流れ：**プロンプトはまず外部のセキュリティチェ

・図9-3



ックサーバーに送られ、そこでAIによるチェックを受けます。OKなら外部LLMへ、NGならブロックされます。

・**特徴：**企業側でのサーバー管理が不要で導入しやすいです。その反面、データが一時的に社外に出ます。サービス提供者の信頼性が鍵となります。

③企業内独自のセキュリティチェックサーバーの流れ

・**データの流れ：**プロンプトはまず企業内独自のサーバーで

チェックされます。NGの場合はデータが社外に出る前にブロックされ、OKの場合のみ外部LLMへ送られます。

- **特徴：**最もセキュリティレベルが高い方式です。しかし、高性能サーバーの導入・維持コストと管理負荷が高くなります。厳しいセキュリティ要件を持つ企業に適しています。

これらの方式の主な特徴を比較したものを図9-4に示します。

セキュリティレベル、コスト、運用負荷、導入の容易さは、方式によって大きく異なります。どの方式が最適かは、企業の状況によって変わってきます。

・図9-4

	チェックなし	外部サーバーチェック	企業内サーバーチェック
セキュリティレベル	低	中	高
初期コスト（目安）	ほぼ無し	低〜高	高
運用コスト（負荷）	低	低〜高	高
導入のしやすさ	しやすい	やや難しい	難しい

9.4 管理者が考慮すべき設定項目例

これらのプロンプトチェックシステムを効果的に運用するため、企業の管理者は通常、以下のような項目を設定・管理し、自社のポリシーに合わせた運用を実現します。

■送信許可セキュリティレベル

図9-1で示したリスクレベルのうち、どのレベルまで外部送信を許可するか（例：「レベル1まで許可」）。

■チェック回数

図9-2で示した判定ロジックに関わる、1プロンプトあたりのチェック実行回数。

■NGワードリスト

会社として禁止する特定のキーワードやパターン。

■（企業内サーバー方式の場合）サーバー接続情報

自社サーバーのURLなど。

■その他

ユーザーへの通知メッセージ、ログ管理方法など。

これらの設定を適切に行い、定期的に見直すことが重要です。

9.5 導入検討時のヒント

自社に最適な「門番」を選ぶ

プロンプトチェック技術の導入を検討する際には、機能だけでなく、以下の点を総合的に評価し、自社に合った「門番」を選ぶことが重要です。図9-4の比較表も参考にしてください。

■セキュリティ要件とリスク許容度

自社が守りたい情報のレベルと、どの程度のデータ外部送信リスクなら許容できるかを明確にします。

■コスト（初期・運用）とROI

各方式のコスト（サーバー費用、サービス料、人件費など）と、それによって得られるリスク低減効果（情報漏洩による損害回避）を比較検討します。

■運用体制とITリソース

自社にサーバー管理やシステム運用を行える人材や体制があるか。なければ、運用負荷の低い方式やサポートの手厚いサービスを選択する必要があります。

■精度・性能・連携

事前にツールの精度や応答速度を検証し、既存システムとの連携が可能かを確認します。

■段階的導入

全社一斉導入ではなく、特定の部署からスモールスタートしたり、まずは導入しやすい方式（例：サービス提供者サーバー

方式）から始めて、将来的に方式を見直すといった進め方も有効です。

【章のまとめ】

技術の力で、より安全なAI活用環境を

本章では、生成AI活用における情報漏洩リスク、特にプロンプト入力時のリスクに対応するための具体的な技術的アプローチの一例として、「チャプロのセキュリティチェックサーバー」を紹介しました。

このような技術は、従業員が日々安心してAIを活用するための心強い支えとなり、「AIの利便性」と「情報セキュリティの確保」という重要な課題を両立させる上で、大きな可能性を秘めています。

しかし、繰り返しになりますが、いかなる技術も完璧ではありません。本章で紹介したような技術的アプローチも、第6章で解説した組織全体の明確なルール（ガイドライン）、継続的な従業員教育、責任あるガバナンス体制、そして第7章で述べた万が一のインシデントへの備えといった、包括的な取り組みと組み合わせてこそ、その真価を発揮します。

技術の力を賢く借りながら、組織全体でリスクに向き合う。それが、これからのAI活用時代における成功の鍵となるでしょう。

さて、具体的なリスクと対策に関する議論は本章で一区切り

とし、次の第10章では、AI技術全体の進化の可能性と、それに伴う新たな課題について考察します。そして最終章（第11章）で、これからの変化の時代を生き抜くために、企業が取るべき次の一手と持つべきマインドセットについて、本書の結論を述べたいと思います。

第10章

AI技術の未来展望

これからAIはどう進化していくのか？

第9章では、プロンプトチェック技術（セキュリティチェックサーバー）のように、AI活用におけるリスクに対応するための具体的な技術的アプローチの一例を見てきました。こうした個別の技術も進化を続ける一方で、AI技術全体としては、これからどのような未来に向かっていく可能性があるのでしょうか？

本章では、また少し視野を広げ、現在注目されているAI技術の発展方向性を概観し、それらが私たちのビジネスや社会にもたらしうる変化について考えてみましょう。ただし、未来予測には不確実性が伴うことを念頭に置いてください。技術の未来を知ることは、「今」何をすべきかを考える上で重要なヒントを与えてくれるはずです。

10.1 加速する進化と注目される発展方向性

AIはどこへ向かうのか？

生成AIの進化のスピードは、本当に目覚ましいものがあります。「ムーアの法則※1」を彷彿とさせるような勢いで、AIの能力は日々向上し、できることの範囲も広がっています。現在、特に注目されている技術の進化の方向性とその可能性を見ていきましょう（※1.コンピュータの性能が指数関数的に向上するという法則）。

もっと人間らしくなるのか（マルチモーダルAIの進化）

■現在

テキスト（文字）や画像など、特定の種類の情報だけを扱えるAIが多い。

■未来

テキスト、画像、音声、動画、さらには触覚や匂いといった五感に近い情報まで、複数の種類の情報を区別なく理解し、組み合わせ、新しい形で表現できるAIの登場が期待されています。例えば、「会議の音声と映像、配付資料をまとめてAIに渡すと、議事録だけでなく、会議のポイントをまとめたショート動画まで自動で生成してくれる」「製品のイメージ図を見せながら口頭で指示すると、AIが3Dデザインを作成し、改善案まで提案してくれる」といったことが、より自然にできるように

なるかもしれません。これは、人間とAIのコミュニケーションを劇的に変える可能性を秘めています。

「少し」教えるだけで賢くなるのか（効率的な学習能力の向上）

■現在

高性能なAIを作るには、膨大な量の学習データが必要なことが多い。

■未来

まるで勘の良い人のように、ほんの少しの事例（データ）や指示からでも、新しいことや専門的なことを素早く学習できるAI（Few-Shot/Zero-Shot Learningと呼ばれる技術の進化形）が、より一般的になるかもしれません。これが実現すれば、データが少ない分野や、各企業の独自業務に合わせたAIのカスタマイズが、今よりもずっと簡単かつ低コストで行えるようになり、AI活用のハードルが大きく下がることが期待されます。

「考える力」がもっと深くなるのか（推論・計画能力の向上）

■現在

AIは過去のパターンから「それらしい答え」を出すのは得意だが、複雑な論理を組み立てたり、未知の問題を解決したりするのはまだ苦手な面がある。

■未来

単に知識を検索したり、文章を生成したりするだけでなく、与えられた情報から論理的に考え（推論）、目標達成のための手順を計画し（プランニング）、問題を解決する能力を持つAIが登場する可能性があります。これが進めば、新薬の開発、複雑な経営課題の分析、難解な科学的発見など、これまで人間の高度な思考力が必要だった領域で、AIが強力なパートナーとなるかもしれません。

「あなただけ」にピッタリ合わせられるのか（パーソナライズと状況適応性の向上）

■現在

ある程度、個人の好みに合わせた情報提供は行われている。

■未来

あなたの好み、これまでの行動、今の状況、さらにはその時の気分までをAIがより深く理解し、完全に「あなた専用」にカスタマイズされた情報やサービス、あるいは対話を提供できるようになるかもしれません。例えば、あなたの学習レベルや興味に合わせて内容が変化する教科書、あなたの健康状態に合わせて最適なアドバイスをくれるヘルスケアAIなどが考えられます。

「なぜ?」に答えてくれる?（説明可能なAI(XAI)の発展）

■現在

AIの判断プロセスは「ブラックボックス」で、なぜその答えになったのか分からないことが多い。

■未来

AIが「なぜこのように判断したのか」「どの情報が判断に影響したのか」を、人間が理解できる形で説明してくれる技術（XAI）が、より実用化されていくと考えられます。これは、AIの「信頼性」と「透明性」を高める上で非常に重要です。特に、金融（融資審査）、医療（診断支援）、法務（判例分析）など、説明責任が求められる分野でのAI活用を大きく後押しする可能性があります。

「自分で考えて動く」AIが実現可能?（自律型AIエージェントの進化）

■現在

AIは基本的に、人間からの指示を受けて動作する。

■未来

人間から「○○を達成してほしい」という大まかな目標を与えられると、そのために必要な情報収集、分析、ツールの利用、他のAIとの連携などを、AI自身が計画し、自律的に実行する「AIエージェント」のような存在が登場するかもしれません。

例えば、「来週の出張の最適なプランを立てて予約まで済ませておいて」と頼むだけで、AIエージェントが飛行機やホテルを比較検討し、予約まで完了してくれる、といったイメージです。これが進めば、多くの定型的な業務プロセスが自動化される可能性があります。

身近なモノの中で動くAI？（エッジAIとの融合）

■現在

高度なAI処理は、主にインターネット経由でクラウド上の強力なコンピュータで行われることが多い。

■未来

スマートフォン、自動車、家電、工場の機械といった、私たちの身近にある様々なデバイス（エッジデバイス）自体に、高性能なAIが搭載され、その場で処理を行う（エッジAI）ことが、より一般的になると考えられます。これにより、インターネット接続がない場所でもAIが使えたり、処理の応答速度が速くなったり、個人情報をデバイスの外に出さずに処理ができるためプライバシー保護が強化されたりといったメリットが生まれます。IoT（モノのインターネット）や自動運転、ロボット技術との連携をさらに加速させるでしょう。

これらの技術トレンドは、互いに影響し合いながら進化していくと考えられます。

10.2 技術進化がもたらす新たな可能性と課題

光と影を見据える

このような技術の進化は、私たちのビジネスや社会に、計り知れないほどの素晴らしい可能性をもたらします。

■「働き方」の革命

単純作業や情報収集・整理といった業務から解放され、人間はより創造的で、コミュニケーションや共感を必要とする、人間ならではの価値を発揮できる仕事に集中できるようになるかもしれません。生産性は劇的に向上し、労働時間の短縮にもつながる可能性があります。

■「個別化」の進展

教育、医療、買い物、エンターテインメントなど、あらゆる分野で、一人ひとりのニーズや状況に完璧に最適化されたサービスや体験が提供されるようになるかもしれません。

■「不可能」への挑戦

新薬開発、気候変動対策、貧困問題など、これまで人類が解決できなかった地球規模の困難な課題に対して、AIがその強力な分析・予測・シミュレーション能力で解決策を見出す手助けをしてくれるかもしれません。

■「創造性」の解放

特別なスキルがなくても、誰もがAIの力を借りて、アイデ

アを形にし、新しいものを創造できるようになり、文化やビジネスの多様性が花開くかもしれません。

しかし、輝かしい未来の可能性と同時に、私たちは技術進化がもたらす新たな課題やリスク、「影」の部分にも、冷静に目を向ける必要があります。

■倫理的な問題の深刻化

AIがより社会に深く浸透し、自律性を持つようになると、「AIによる差別の固定化」「責任の所在の曖昧化」「人間の監視・操作」「プライバシーの侵害」といった倫理的な問題は、より複雑で深刻になります。技術の進歩と倫理的なルール作りを、常に両輪で進めていく必要があります。

■悪用リスクの増大

より高性能なAIは、悪意を持って使われれば、より巧妙な偽情報の生成・拡散、高度なサイバー攻撃の自動化、あるいは自律型兵器といった、社会の安全を脅かす深刻な脅威となりえます。技術の悪用を防ぐための国際的な協力や規制も重要な課題です。

■社会構造への影響と格差

AIによる自動化が、雇用（特に定型業務）にどのような影響を与えるのか、そしてAIを使いこなせる人とそうでない人との間のスキルや経済的な格差（デジタルデバイド）が拡大しないか、といった社会構造への影響についても、十分な議論と対策が必要です。

■制御不能になるリスク？

これはまだSFの世界に近いかもしれませんが、一部では、AIが人間の知能を遥かに超え、人間の制御が効かなくなる「シンギュラリティ」の可能性も議論されています。長期的な視点では、AIの安全性をいかに確保していくかという問題も存在します。

技術の進歩は、私たちに大きな力を与えてくれますが、その力を「何のために」「どのように」使うのか、その責任は常に私たち人間にあります。未来を楽観視するだけでなく、潜在的なリスクや課題にも真摯に向き合い、賢明な選択をしていくことが求められます。

【章のまとめ】

本章では、生成AI技術がこれから向かう可能性のある方向性と、それがもたらす光と影、すなわち新たな可能性と課題について概観しました。このような未来像は、私たちに大きな期待を抱かせると同時に、備えるべき課題も示唆しています。技術がどのように進化しようとも、重要なのは、その変化に企業として、そして個人としてどう向き合い、適応していくかです。

次の最終章（第11章）では、この変化の時代を生き抜き、AIを真の成長の力に変えていくための具体的な次の一手と、持つべきマインドセットについて、本書の結論として述べたいと思います。

第
11
章

変化の時代を生き抜くために

企業の次の一手とマインドセット

第10章では、AI技術がこれからさらに進化し、私たちのビジネスや社会を大きく変えていく可能性について見てきました。マルチモーダル化、自律性の向上、人間とのより自然な対話…。未来への期待が膨らむ一方で、倫理的な課題や悪用リスクといった、私たちが真剣に向き合わなければならない側面も浮き彫りになりました。

このような、変化が激しく、予測が難しい時代を、企業はどのように生き抜き、AIという強力な力をどのように真の成長エンジンへと変えていけば良いのでしょうか？　技術の進化をただ眺めているだけでは、あっという間に時代に取り残されてしまいます。

本章は、本書全体の締めくくりとして、これまでの議論を踏まえ、企業が今、そしてこれから取るべき具体的な「次の一手」と、その土台となるべき「マインドセット（心構えや組織文化）」について、具体的な提言を行います。未来を創るための行動を、ここから一緒に始めましょう。

11.1 企業が継続的に取り組むべきこと

「変化への適応力」を鍛える

■技術の進化を追いかけ、学び続ける（知的好奇心と情報感度）

AIの世界は日進月歩です。「一度導入したから終わり」ではありません。常に最新の技術動向、新しいツール、競合の動き、そして新たなリスクに関する情報を積極的に収集し、「自社にとってこれはどういう意味を持つのか？」「活用できるチャンスはないか？」と評価し続ける姿勢が求められます。アンテナを高く張り、学び続ける組織文化を作りましょう。

■「責任あるAI」を追求し続ける（倫理とガバナンスの徹底）

技術が進歩すれば、新たな倫理的課題も生まれます。第6章で構築したAIガバナンス体制や倫理原則を形骸化させず、社会の声に耳を傾けながら定期的に見直し、改善していくことが重要です。企業の社会的信頼を守り、持続的な成長を遂げるためには、「責任あるAI」への取り組みが不可欠です。

■「人」と「組織」を育て続ける（人材育成と組織能力向上）

AIを使いこなし、リスクを理解し、そしてAIと共に新しい価値を生み出せる人材は、これからの企業の宝です。全従業員のAIリテラシー向上はもちろん、専門人材の育成・確保、そしてAI導入によって変化する業務や求められるスキルへの対応（リスキリング）に、継続的に投資していく必要があります。人が成長してこそ、組織も成長します。

■「データ」という武器を磨き続ける（データ戦略の高度化）

AIの性能は「データ」の質と量に大きく左右されます。自社が持つ様々なデータを、プライバシーを守り、セキュリティを確保しながら（第3章参照）、いかに価値ある「資産」として整備し、管理し、戦略的に活用していくか。データ戦略を継続的に見直し、磨き続けることが、AI活用の効果を最大化する鍵となります。

■「守り」を固め続ける（セキュリティ対策の不断の見直し）

AIを狙うサイバー攻撃も日々進化しています（第5章参照）。最新の脅威情報を常に把握し、技術的なセキュリティ対策と組織的な運用ルールの両面から、防御体制を継続的に見直し、強化していくことが、安全なAI活用を守るための大前提です。

■「仲間」を増やす（外部との連携・共創）

全ての課題を自社だけで解決しようとする必要はありません。大学や研究機関、先進的な技術を持つスタートアップ、あるいは同業他社や業界団体など、外部の知見やリソースを積極的に活用し、連携・協力していく（オープンイノベーション）視点が、変化のスピードに対応するために有効です。

■「まずやってみる」勇気と柔軟性（アジャイルな取り組み）

完璧な計画を立ててからでないと動けない、という姿勢では、変化の速いAI時代には対応できません。まずは小さく始めてみて（スモールスタート）、試してみて、その結果から学び、柔軟に計画を修正しながら進めていくという、アジャイル（俊敏）なアプローチが有効です。失敗を恐れず、素早く行動し、

学習していく体質を作りましょう。

これらの継続的な取り組みは、単にAIのリスクに対応するためだけではありません。これらは、変化に強い、しなやかな組織を作り上げ、AIを真の競争力へと転換していくための、企業にとっての基本的な「筋肉」を鍛える活動なのです。

11.2 AIを成長エンジンとするためのマインドセット

「意識」が変われば未来が変わる

技術や体制を整えるだけでは十分ではありません。AIという新しい力を真に活かすためには、経営層から従業員一人ひとりに至るまで、組織全体の「マインドセット」、すなわち物事の捉え方、価値観、そして組織文化そのものを変革していく必要があります。

■「守り」と「攻め」のバランス感覚

リスク管理はもちろん重要ですが、リスクを恐れるあまり、新しい挑戦を全くしなくなっては本末転倒です。リスクを適切にコントロールしながら、同時にAIを活用した新しい価値創造へ果敢に挑戦する。この「守り」と「攻め」のバランス感覚を組織全体で持つことが重要です。

■経営トップの「覚悟」と「発信力」

AIによる変革は、時にこれまでのやり方を大きく変える必

要があり、痛みを伴うこともあります。経営トップがその覚悟を持ち、「なぜ変革が必要なのか」「どこへ向かうのか」という明確なビジョンを、繰り返し、情熱を持って組織全体に語りかけることが、変革への推進力となります。

■「壁」を壊し、「協働」を生む文化

AI活用は、部署間の壁を越えた連携が不可欠です。知識やデータ、成功も失敗も、オープンに共有し、互いに協力し合って課題解決や新しいアイデア創出に取り組む。そのような「協働」「共創」の文化を育むことが、組織全体のAI活用レベルを引き上げます。

■「失敗は学び」と捉える挑戦する風土

新しいことへの挑戦に、失敗はつきものです。一度の失敗で諦めたり、挑戦した人を責めたりするのではなく、失敗から何を学べるかを考え、それを次の成功への糧とする。そのような「学習する組織」の文化（心理的安全性）が、イノベーションを生み出す土壌となります。

■「AI vs 人間」ではなく「AI & 人間」へ

AIを、仕事を奪う「脅威」と捉えるのではなく、人間の能力を拡張し、面倒な作業から解放してくれる「頼れる相棒（協働者）」と捉える意識改革が必要です。人間とAIがそれぞれの得意分野を活かし、協力し合うことで、一人では成し遂げられなかった大きな成果を生み出す未来を目指しましょう。

■「短期目線」から「長期的視点」へ

AI活用の真の価値は、短期的なコスト削減だけではありま

せん。数年後、数十年後の会社の競争力を築くための、未来への戦略的投資であるという認識が必要です。すぐに目に見える成果が出なくても、腰を据えて、粘り強く、継続的に取り組み、改善していくという長期的なコミットメントが求められます。

■「受け身」から「主体的」へ

AIが社会をどう変えるかをただ待つのではなく、「自分たちはAIを使って、どんな未来を創りたいのか？」を考え、主体的に変化を起こしていく。その当事者意識こそが、変化の時代を生き抜くための最も重要なマインドセットかもしれません。

11.3 今こそ行動の時

変革への招待

さて、本書を締めくくるにあたり、改めて皆様にお伝えしたいことがあります。

生成AIは、確かに未知のリスクを伴う、変化の激しいテクノロジーです。しかし、本書を通じて学んできたように、そのリスクは決してコントロール不可能なものではありません。リスクの本質を理解し、適切な対策（技術、組織、プロセス、倫理）を多層的に講じ、そして何よりも、私たち人間が賢明さを持って向き合えば、AIは私たちのビジネス、そして社会を、より豊かで、より良い方向へと導いてくれる計り知れない可能性を秘めているのです。

道筋は見えました。羅針盤（本書）も手にしました。あとは、最初の一歩を踏み出す勇気だけです。

変化を恐れて立ち止まっていては、何も始まりません。むしろ、AIを活用して進化していく他社や社会から取り残され、気づいた時には手遅れになっているかもしれません。

今こそ、行動を起こす時です。

難しく考える必要はありません。まずは、あなたの立場でできることから始めてみましょう。

■経営者の方へ

AI活用の明確なビジョンを示し、推進体制と予算を確保し、自ら変革の先頭に立ってください。

■管理職の方へ

部門での具体的な活用を検討し、部下が安心して挑戦できる環境を作り、ガイドラインの遵守を徹底してください。

■従業員の皆さんへ

AIを恐れず、まずは使ってみてください。そして、会社のルールを守り、リスクを意識しながら、自分の業務をどう改善できるか考えてみてください。疑問があれば、積極的に質問し、学び続けてください。

社内でAIに関する対話を始めること。小さな範囲でAIを試してみること。ガイドラインの草案作りに参加してみること。どんなに小さな行動でも構いません。その一つ一つの主体的な

挑戦が積み重なって、あなたの会社の、そしてあなた自身の未来を形作っていくのです。

生成AIと共に、新たな価値創造への旅に出かけましょう。本書が、その素晴らしい冒険への、力強い後押しとなることを心から願っています。

おわりに ✨

【AIの可能性を理解し、未来を創造するために】

本書を手に取り、そして貴重な時間を割いて最後までお付き合いいただき、誠にありがとうございました。

この一冊を通して、私たちは「生成AI」という、今まさに私たちの働き方や社会のあり方を大きく変えようとしている新しいテクノロジーについて、様々な角度から一緒に考え、学んできました。

AIとは何かという基本から始まり、活用することでどのようなメリットが期待できるのか、その一方でどのようなリスク（情報漏洩、著作権の問題、誤った情報、偏見の助長、予期せぬ動作、過信、システム停止など）に注意すべきなのか、そしてそれらのリスクに対して具体的にどのような対策を講じ、組織としてどう向き合っていくべきなのか。

さらには、実際にAI活用に取り組む企業の事例や、AI技術がこれから向かうかもしれない未来の姿まで、多岐にわたるテーマを巡る旅でした。

読み進める中で、生成AIの計り知れない可能性にワクワク

された方もいらっしゃるでしょうし、同時に、無視できないリスクの存在に、改めて身を引き締められた方もいらっしゃるかもしれません。「情報漏洩」「著作権侵害」「AIの嘘」「差別助長」…これらの言葉を前に、AI導入へのためらいを感じていた方も少なくないはずです。

しかし、本書を通じて私たちが一貫してお伝えしたかったのは、これらのリスクは「正しく理解し、備えれば、乗り越えられる課題」であるということです。

リスクは、得体の知れない恐ろしいものではありません。なぜそれが起こるのか（発生メカニズム）を知り、適切な対策（技術的な仕組み、組織的なルール、継続的な学び、そして倫理的な配慮）を、一つひとつ、あるいは組み合わせて講じていけば、その影響を十分に管理下に置くことができるのです。

大切なのは、リスクに目を背けるのではなく、その本質を冷静に見極め、賢く向き合うこと。それこそが、生成AIを単なる「リスクの種」ではなく、日々の業務の生産性を高め、従業員の創造性を刺激し、ひいては会社全体の成長と社会の進歩に貢献する、真に頼もしい「ビジネスパートナー」へと変えるための鍵なのです。

技術革新の波は、待ってはくれません。生成AIという波は、特に大きく、速く、そして広範囲に押し寄せてきています。この変化に対して、「まだ様子を見よう」と受け身でいるのか、それとも「まずは理解し、できることから試してみよう」と主体的に関わっていくのか。その姿勢の違いが、これからの企業の未来を大きく左右していくことになるでしょう。

完璧な準備が整うのを待っていたら、あっという間に時代に取り残されてしまうかもしれません。不確実な状況の中でも、学び続け、試行錯誤を恐れず、小さな一歩でも着実に前に進んでいくこと、それが今、私たちに求められていることではないでしょうか。

もし、本書が、皆様にとって、生成AIという未知の領域へ踏み出すための、そしてリスクという不安を乗り越えていくための、信頼できる「道しるべ」のような存在になれたのであれば、著者としてこれ以上の喜びはありません。

本書で得た知識や考え方が、皆様の会社におけるAI導入への漠然とした不安を、「まず、これをやってみよう」という具体的な行動計画へと変えるきっかけとなり、未来への挑戦に向けた自信と勇気を少しでも後押しできたなら、本当に幸いです。

さあ、本書を読み終えた今こそ、新しい一歩を踏み出す時です。難しく考える必要はありません。

まずは、あなたのチームや部署でAIについて話し合ってみませんか？　あるいは、リスクの少ない業務で、試しにAIを使ってみませんか？　または、第6章を参考に、自社向けのガイドラインの骨子を書き出してみませんか？　どんなに小さな行動でも、それが皆さんの会社、そして皆さん自身の未来をより良い方向へと動かす、確かな始まりとなるはずです。

生成AIが持つ大きな可能性を、恐れるのではなく、理解し、味方につけ、自社の、そして私たち自身の力へと変えていきましょう。本書で得た「知恵」と、皆さんが本来持っている「勇気」をもって、変化を恐れずに、まだ見ぬ未来を、AIと共に切り拓いていく。

皆様のこれからの挑戦が、実り多く、素晴らしいものとなることを、心から願っております。

最後までお読みいただき、本当にありがとうございました。

七里信一

付 録

本書をお読みいただき、生成AIのリスクと対策、そしてその活用に向けた道筋についてご理解いただけたことと思います。AIの世界は日進月歩であり、本書で触れた内容をさらに深く掘り下げたい、あるいは常に最新の情報を追いかけたい、と感じられた方もいらっしゃるかもしれません。

この付録では、皆様のさらなる学習や理解を深めるための一助として、「参考文献・情報源リスト」と「主要用語集」をご用意しました。ぜひご活用ください。

●付録A　参考文献・情報源リスト

本書の執筆にあたり参考にした、あるいは読者の皆様がさらに情報を得る上で有用と考えられる主要な情報源（公的機関、研究機関、業界団体、標準化団体など）を分野別にリストアップします。書籍という特性上URLは記載しておりませんが、機関名や文書名をキーワードにウェブ検索などを行っていただくことで、関連情報にアクセスできるかと存じます。

（※ご注意：各機関の情報は常に更新されています。最新版の文書や情報をご確認ください）

【A.1 AI基礎知識・技術動向】

・主要AI研究開発機関・企業の情報

OpenAI（ChatGPT、DALL-E開発元）：公式ウェブサイト、ブログ、技術文書など。

Google AI/DeepMind(Gemini、LaMDA開発元):公式ウェブサイト、ブログ、研究論文発表など。
Microsoft Research AI:公式ウェブサイト、研究発表など。
Meta AI:公式ウェブサイト、研究発表など(これらの企業のサイトでは、最新のAIモデルや研究開発に関する情報が随時公開されています)。
人工知能学会(JSAI):日本の主要なAI関連学会。学会誌、論文誌、全国大会の発表情報など。
情報処理学会(IPSJ):情報処理全般に関する学会。AI関連の研究会や論文誌も。

【A.2 AIリスク・セキュリティ】

IPA(独立行政法人 情報処理推進機構):「情報セキュリティ白書」(年次発行)、各種セキュリティガイドライン、注意喚起、技術解説文書(ウェブサイト参照)。AI関連のセキュリティ情報も随時公開。
NIST(米国国立標準技術研究所):「AI Risk Management Framework(AI RMF)」:AIリスク管理の包括的なフレームワーク。
「**Cybersecurity Framework**」:サイバーセキュリティ対策の基本的なフレームワーク。その他、AIセキュリティに関する各種ガイドラインやレポート(NISTウェブサイト参照)。
OWASP(Open Worldwide Application Security Project):「OWASP Top 10 for Large Language Model Applications」:LLMアプリケ

ーションに特有のセキュリティリスクTop 10とその対策（OWASPウェブサイト参照）。

ENISA（European Union Agency for Cybersecurity）：EUのサイバーセキュリティ機関。AIセキュリティに関する調査レポートなど（ENISAウェブサイト参照）。

JPCERT/CC（JPCERT Coordination Center）：日本のインシデント対応支援組織。セキュリティインシデント情報、注意喚起など（JPCERT/CCウェブサイト参照）。

【A. 3 AI倫理・ガバナンス】

総務省：「AIネットワーク社会推進会議」の報告書、提言。「AI開発ガイドライン」「AI利活用ガイドライン」「人間中心のAI社会原則」など（総務省ウェブサイト参照）。

経済産業省：「AI原則実践のためのガバナンス・ガイドライン」など（経済産業省ウェブサイト参照）。

内閣府：「AI戦略」「人間中心のAI社会原則」（内閣府ウェブサイト参照）

OECD（経済協力開発機構）：「OECD AI Principles（AI原則）」「OECD AI Policy Observatory」（各国のAI政策動向。OECDウェブサイト参照）

主要企業のAI倫理原則：Google、Microsoft、NTTグループなど、多くの企業が独自のAI倫理原則やガイドラインを公開しています（各社ウェブサイトのサステナビリティ、CSR、AIに関するページなどを参照）。

【A.4 著作権・法的側面】

文化庁：著作権制度に関する解説、Q&A、審議会（文化審議会著作権分科会）の報告書など。AIと著作権に関する議論も（文化庁ウェブサイト参照）。

著作権情報センター（CRIC）：著作権に関する情報提供、解説資料（CRICウェブサイト参照）。

個人情報保護委員会：個人情報保護法に関するガイドライン（通則編、外国にある第三者への提供編、第三者提供時の確認・記録義務編、仮名加工情報・匿名加工情報編など）、Q&A。漏洩等事案が発生した場合の対応に関するガイドライン（個人情報保護委員会ウェブサイト参照）。

情報ネットワーク法学会：AIと法（著作権、プライバシーなど）に関する研究委員会の活動成果など（同学会ウェブサイト参照）。

弁護士・弁理士などの法律専門家：AIと法律に関する解説記事、セミナー、書籍など。

●付録B　主要用語集

本書を読み進める上で、あるいは今後AIに関する情報に触れる上で、理解しておくと役立つ主要な用語を解説します。より深く知りたい場合は、付録Aで紹介した情報源などもご参照ください。

※アルファベット順→五十音順に掲載しています。

【アルファベット】

AI(Artificial Intelligence/人工知能)：人間の脳が行うような「考える」「学ぶ」「判断する」といった働きを、コンピュータを使って実現しようとする技術や研究分野の総称です。

API(Application Programming Interface)：ソフトウェアやサービス同士が、互いの機能やデータをやり取りするための「接続口」のようなものです。多くの生成AIサービスは、このAPIを通じて企業システムなどから利用されます。

ChatGPT：OpenAI社が開発した、人間と自然な対話ができるAIサービス。大規模言語モデル（LLM）技術の代表例として広く知られています。

CSIRT(Computer Security Incident Response Team/シーサート)：会社の中で、情報セキュリティに関する問題（インシデント）が発生した際に、専門的に対応するチームのことです。

Deep Learning(ディープラーニング/深層学習)：コンピュータが、人間の脳の神経回路（ニューロン）を模した複雑なネッ

トワーク構造を使って、大量のデータから自動的に物事の特徴やパターンを深く学習する技術です。現在のAI、特に生成AIの能力を飛躍的に向上させました。

DRP(Disaster Recovery Plan/災害復旧計画)：大きなシステム障害や自然災害が発生した場合に、重要なシステムや業務を、あらかじめ定めた目標時間内に復旧させるための具体的な手順をまとめた計画書です。事業継続計画（BCP）の一部です。

Few-Shot Learning/Zero-Shot Learning：AIが、ほんの少し（Few-Shot）あるいは全くゼロ（Zero-Shot）のお手本データからでも、新しいことを学習したり、指示されたタスクを実行したりできる能力のことです。AIの学習効率を高める技術として注目されています。

Fine-tuning(ファインチューニング)：大量のデータで事前学習された汎用的なAIモデル（例：LLM）を、特定の目的や専門分野に合わせて、追加のデータで再調整（チューニング）することです。これにより、その分野に特化した性能の高いAIを作ることができます。

Gemini：Google社が開発した、テキストだけでなく画像や音声なども扱えるマルチモーダルな能力を持つ、高性能なAIモデルシリーズ。

Generative AI(生成AI/生成系AI)：指示や学習データに基づいて、新しい文章、画像、音声、コードなどのコンテンツを「生成」することができるAIの総称です。本書のテーマです。

LLM(Large Language Model/大規模言語モデル)：インター

ネット上のテキストなどを大量に学習し、人間が使う言葉（自然言語）を非常に高いレベルで理解・生成できるAIモデルのことです。ChatGPTやGeminiなどの基盤となっています。

NIST（National Institute of Standards and Technology／米国国立標準技術研究所）：技術標準に関する米国の政府機関。AIのリスク管理やサイバーセキュリティに関する重要なフレームワークなどを策定・公開しています。

OpenAI：ChatGPTやDALL-Eなどを開発した、AI研究開発における世界的なリーディングカンパニーの一つ。

OWASP（Open Worldwide Application Security Project）：ウェブアプリケーションのセキュリティ向上を目指す国際的な非営利コミュニティ。AI（LLM）アプリケーション特有のセキュリティリスクについても情報発信しています。

Prompt（プロンプト）：生成AIに対して、「何をしてほしいか」を伝えるための指示文や質問のことです。プロンプトの書き方次第で、AIの出力の質が大きく変わります。

Prompt Engineering（プロンプトエンジニアリング）：AIから望ましい結果を引き出すために、効果的なプロンプトを設計・工夫する技術やノウハウのことです。

Prompt Injection（プロンプトインジェクション）：悪意のある第三者が、プロンプトに特殊な指示を紛れ込ませることで、AIに意図しない危険な動作（情報漏洩など）をさせようとするサイバー攻撃の一種です。

SLA（Service Level Agreement / 品質保証契約）：クラウドサ

ービスなどを利用する際に、サービス提供会社と利用者の間で交わされる、サービスの品質（例：稼働率、性能、サポート体制など）に関する保証レベルの取り決めのことです。

Stable Diffusion：オープンソース（設計図が公開されている）の画像生成AIモデルとして有名。多くの派生モデルやサービスが存在します。

Transformer（トランスフォーマー）：2017年にGoogleの研究者が発表したAIモデルの設計構造。文章などのつながり（系列データ）を効率的に処理する能力に優れ、現在のLLMの多くがこの技術をベースにしています。

XAI（Explainable AI/説明可能なAI）：AIが出した答えや判断について、「なぜそうなったのか」その理由や根拠を人間が理解できるようにするための技術や研究分野のことです。AIの「ブラックボックス問題」を解消し、信頼性を高めるために重要視されています。

【五十音順】

可用性（Availability/アベイラビリティ）：システムやサービスが、使いたい時にいつでも中断せずに使える状態にある度合いのことです。「止まらないこと」の指標。情報セキュリティの重要な要素（機密性・完全性・可用性＝CIA）の一つです。

監査証跡（Audit Trail/ログ）：コンピュータシステムでの操作や出来事（イベント）の記録のこと。「誰が」「いつ」「何をしたか」が分かるため、問題発生時の原因調査や不正行為の発見に

不可欠です。

機械学習（Machine Learning）：コンピュータが大量のデータから自動的にルールやパターンを「学習」し、それに基づいて新しいデータに対する予測や分類を行う技術です。AIを実現するための主要なアプローチの一つです。

機密情報（Confidential Information）：会社の経営、技術、営業、人事などに関する秘密の情報で、外部に漏れると会社に損害を与える可能性のある情報のことです。

機密性（Confidentiality/コンフィデンシャリティ）：情報が、許可された人以外にはアクセスされたり漏洩したりしないように保護されている状態のことです。情報セキュリティの重要な要素（CIA）の一つです。

強化学習（Reinforcement Learning）：AIが、試行錯誤を繰り返しながら、特定の目標（例：ゲームで高得点を取る）を達成するために、どのような行動を取れば最も良い結果（報酬）が得られるかを自ら学習していく手法です。

協働者（Collaborator）：AIを単なる命令に従う道具としてだけでなく、人間と協力して目標達成を目指すパートナーとして捉える考え方です。人間とAIがそれぞれの得意分野を活かし合う関係性を目指します。

著作権（Copyright）：文章、音楽、絵、写真、プログラムなど、人が創造的に表現したもの（著作物）を、創った人（著作者）が守るための法律上の権利です。無断でのコピーや利用を制限します。

著作物（Work）：人の考えや感情が創作的に表現されたもので、著作権によって保護される対象です。単なる事実やアイデア自体は著作物ではありません。

敵対的サンプル（Adversarial Example）：AIを騙す目的で、入力データに人間には分からない程度のわずかなノイズなどを意図的に加えたものです。AIに誤認識を引き起こさせます。

透明性（Transparency）：AIシステムの動作原理や判断プロセスが、どの程度人間にとって分かりやすいか、あるいは検証可能かを示す度合いのことです。ブラックボックス性の対極にある概念です。

人間中心（Human-centric）：AI技術は、あくまで人間の幸福、権利、価値観を尊重し、社会全体に利益をもたらすために開発・利用されるべきである、という考え方です。AI倫理の最も基本的な原則の一つです。

バイアス（Bias）：AIの学習データやアルゴリズムに含まれる「偏り」のことです。これが原因で、AIが特定の属性（性別、人種など）に基づいて不公平な判断や差別的な出力をする可能性があります。

ハルシネーション（Hallucination/幻覚）：生成AIが、事実に基づいていない、あるいは学習データにない情報を、あたかも真実であるかのように、もっともらしく生成してしまう現象のことです。「AIの嘘」「AIの知ったかぶり」とも言えます。

ファクトチェック（Fact-checking）：情報の内容が事実に基づいているかどうかを、信頼できる根拠（証拠）に基づいて検証

することです。AIが生成した情報の真偽を見極める上で非常に重要です。

プライバシー（Privacy）：個人に関する情報（個人情報）が、本人の知らないところで収集されたり、本人の意図に反して利用・公開されたりしないように保護されるべきである、という考え方や権利のことです。

ブラックボックス（Black Box）：内部の仕組みや動作原理が複雑すぎて、外部からは分からない、あるいは理解・説明することが困難なシステムのことです。現在の高性能なAIモデルの多くは、この性質を持っています。

マルチモーダルAI（Multimodal AI）：テキスト、画像、音声、動画など、複数の異なる種類の情報（モダリティ）を組み合わせて理解したり、生成したりできるAIのことです。

モデル抽出（Model Extraction/Stealing）：AIサービスの応答結果などを分析して、その内部モデルの仕組みや学習内容を不正に盗み出そうとする攻撃のことです。

リスクアセスメント（Risk Assessment）：組織が直面する可能性のある様々なリスクを、「特定」し、「分析」し、「評価」する（その大きさや優先順位を決める）一連のプロセスのことです。リスク管理の出発点となります。

連合学習（Federated Learning）：個々のスマートフォンや企業などが持つデータを、外部のサーバーに集約することなく、それぞれの場所でAIモデルの学習を行い、その学習結果（モデルの更新情報）だけを共有して統合することで、全体のAIモデ

ルを賢くしていく技術です。プライバシー保護に有効とされています。

ロバスト性（Robustness/頑健性）：AIモデルが、入力データに多少のノイズがあったり、想定外の状況になったりしても、性能が大きく低下したり、誤動作したりすることなく、安定して動作し続けられる「打たれ強さ」のことです。

七里信一（しちり しんいち）

日本一生成AIを教える講師。
生成AIの普及に人生を捧げる、教育者であり起業家。延べ24万人に生成AIを指導した実績を持ち、日本の生成AI教育の第一人者として広く知られる。毎月1万人が参加する無料の生成AI Zoomセミナーは、常に予約で満席となるほどの人気ぶり。
代表を務める生成AIスクール「飛翔」は、実践的なカリキュラムと熱心な指導で高い評価を得ており受講生数2万5000人を突破。また、株式会社エキスパートの代表取締役として、マーケティングシステム「エキスパ」の開発・運営にも携わるなど、多岐にわたる事業を展開。常に最新のAIツールや技術動向を研究し、その知見を惜しみなく提供する姿勢は、多くの支持を集めている。趣味は生成AIプロンプトの研究、最新AIツールの調査。
「生成AIの可能性を最大限に引き出し、誰もがその恩恵を受けられる社会を創る」というビジョンを掲げ、今日も生成AIの普及に情熱を燃やしている。

※数字は2025年6月時点

生成AIセキュリティの教科書

2025年6月10日　初版発行

著者／七里信一

編集・DTP／株式会社ビーパブリッシング

発行・発売／生成AI研究所出版
〒160-0023 新宿区西新宿7-7-30　TEL 03(5348)6870

ISBN 978-4-911384-03-9　C0030