AIモデルの
リスクマネジメントフレームワーク

AI Technology and Risk
Working Group

cloud security alliance®

目次

AI テクノロジー＆リスク・ワーキンググループの恒久的かつ公式な場所は、
https://cloudsecurityalliance.org/research/working-groups/ai-technology-and-risk。

謝辞

主執筆者

マリア・シュヴェンガー ヴァニ・ミッタル

共同議長

クリス・キルシュキ マーク・ヤナリティス

貢献者

エリック・ティアリング
ハディル・ラビブ
マイケル・ロザ
レナータ・ブドコ

CSA グローバルスタッフ

ジョシュ・ブーカー
マリーナ・ブレグコウ
スティーブン・スミス

レビュアー

キャンディ・アレクサンダー
ダニエル・C
エリエル・クルス
ハリエ・スリニヴァサ・バンガロール・ラム・ティラク
カラン・ゲンカ
ケネス・トーマス・モラス
ナマル・クラトゥンガ
ニコラス・レイ
オットー・スリン
ロヒト・ヴァリア
サニトラ・アングラム
トム・ボウワー
ヴァイバフ・マリク
ユヴァラージ・マデスワラン

エグゼクティブサマリー

高度な機械学習（ML）モデルの広範な採用は、予知保全、不正検知、個別化医療、自律走行車、スマートサプライチェーン管理1 といった分野にエキサイティングな機会をもたらしている。これらのモデルは、重要なイノベーションを解き放ち、効率化を推進する可能性を秘めている一方で、その利用の拡大には、固有のリスク、特にモデル自体に起因するリスクも導入される。モデル・リスクを未然に防ぐことは、多額の財務的損失、規制上の問題、風評被害につながる可能性がある。これらの懸念に対処するためには、リスクマネジメントへの積極的なアプローチが必要である。モデル・リスク・マネジメント（MRM）は、人工知能（AI）や ML モデルの開発、展開、使用において、責任と信頼の文化を醸成するための重要な要素であり、組織がリスクを最小限に抑えながら、その潜在能力を最大限に活用することを可能にする。

本稿では、AI モデルの責任ある開発、展開、利用を確保するための MRM の重要性を探る。AI 開発に直接携わる実務者や、AI ガバナンスに焦点を当てるビジネスおよびコンプライアンスのリーダーなど、このトピックに共通の関心を持つ幅広い読者を対象としている。

本稿は、データのバイアス、事実の不正確さや関連性のなさ（俗に「幻覚」や「捏造」と呼ばれる）[2])、誤用の可能性など、AI モデルに内在するリスクを強調している。これは、包括的な MRM フレームワークを確保するためのプロアクティブアプローチの必要性を強調している。この枠組みは、相互に関連した 4 つの柱で構築されている：モデルカード、データシート、リスクカード、シナリオプランニングである。これらの柱は、継続的なフィードバックのループを通じて、リスクを特定し、軽減し、モデル開発とリスクマネジメントを改善するために協働する。具体的には、モデルカードとデータシートはリスク

[1]マッキンゼー・アンド・カンパニー「2023 年の AI 事情：生成的 AI がブレイクする年"

[2]NIST AI 600-1 「人工知能リスクマネジメントフレームワーク」：生成的人工知能プロファイル"

アセスメントに情報を提供し、リスクカードはシナリオプランニングの指針となる。シナリオ・プランニングは、リスクマネジメントとモデル開発を改善する。

このフレームワークを導入することで、組織は ML モデルを安全かつ有益に使用することができる：

- 透明性と説明可能性の向上
- プロアクティブなリスク低減と "セキュリティ・バイ・デザイン"
- 情報に基づいた意思決定
- ステークホルダーや規制当局との信頼構築

本稿では、リスクを最小限に抑えながら AI や ML の可能性を最大限に活用するための MRM の重要性を強調する。

対象読者

AI MRM は、ML モデルの責任ある開発と展開に共通の関心を持つ幅広い読者を対象としている。AI 開発の技術的側面に直接携わる人々と、そのガバナンスや監督に携わる人々のために、技術的な利害関係者と非技術的な利害関係者の橋渡しをする。

ターゲットとする観客は、主に以下の 2 つのグループに分けられる。

1. AI モデル開発と実装の実践者

- **ML エンジニアとデータサイエンティスト：**このグループは、モデルカードとデータシートの詳細な説明と、それらがモデルの理解と開発にどのように貢献するかということから利益を得る。これらのコンポーネントを理解することで、より透明で説明責任のあるモデルを構築することができる。

- **AI 開発者とプロジェクトマネージャー：**このグループは、AI モデルのライフサイクル全体を通して潜在的な問題を予測し、構想から実装まで責任ある展開を保証するためのツールを見つけることができる。

2. AI ガバナンスと監督における利害関係者

- **リスクマネジメントの専門家、コンプライアンス・オフィサー、監査役：**このグループには、MRM の重要性と、効果的なガバナンスの確立、実施、評価に最も関連する業界の一般的なフレームワークとの整合性に関するセクションがある。

- **ビジネスリーダーとエグゼクティブ：**序文と結論のセクションは、組織内で責任ある AI の導入を促進する上での MRM の重要性を強調しており、彼らにとって特に価値がある。

- **コミュニケーションと広報の専門家：**このグループは、AI モデルのリスクとベネフィット、ステークホルダー・エンゲージメント、レピュテーション・マネジメントに関するセクションや、多様な聴衆に響くメッセージの作り方を学ぶことができる。

スコープ

本稿では、責任ある AI 開発における MRM とその重要性について考察する。効果的な MRM フレームワークの 4 つの柱を精査し、それらがどのように連携して MRM への全体的なアプローチを生み出すかを考察する。これらの手法が、透明性、説明責任、責任ある AI 開発をどのように促進するかについて論じる。

本稿では、倫理的で責任ある AI の未来を形作る上で、MRM が果たす役割を強調する。本稿は、MRM の概念的・方法論的側面に焦点を当てており、役割、オーナーシップ、RACI、部門横断的な関与など、人を中心とした側面については触れていない。

序文

MRM の必要性と重要性

今日、様々な業界において、複雑な AI/ML モデルがかつてないスピードで採用されている。一方では、ML モデルへの依存の高まりは、イノベーションと効率性向上のための膨大な可能性を解き放つことを約束している。その一方で、内在するリスク、特にモデル自体に関連するリスク、すなわちモデル・リスクをもたらす。放置すれば、大きな財務的損失、規制当局による制裁、風評被害につながる可能性がある。学習データのバイアス、モデル出力の事実誤認（「幻覚」または「捏造」と呼ばれる[3] ）、悪用の可能性は、プライバシー・リスクや知的財産（IP）の懸念と並んで、リスクマネジメントへの積極的なアプローチが必要となる。AI MRM は、これらのモデルの責任ある信頼できる開発、展開、利用を保証するための重要な規律として浮上している。

MRM は金融業界などでよく使われる用語で、従来は定量モデルに関連するリスクマネジメントを指していた。しかし本稿では、この確立されたコンセプトが、AI モデルに関連するリスクをマネジメントするためのフレームワークを概説する。

AI MRM は、AI モデルに関連する複雑性、不確実性、脆弱性から保護し、AI 主導の意思決定の信頼性と公正性に対するユーザー、利害関係者、規制当局の信頼を強化するのに役立つ。AI が進化を続け、より多くの分野に浸透していくにつれ、MRM は責任ある AI 導入の未来を形作る上でますます重要な役割を果たすようになり、企業や業界に利益をもたらすだろう。

その核心は、モデル・リスクはモデル自体に内在する限界から生じるということである。AI のモデル・リスクの原因として最もよく見られるのは、以下のようなものである：

[3]NIST AI 600-1 「人工知能リスクマネジメントフレームワーク」: 生成的人工知能プロファイル"

- **データ品質の問題：**どのようなモデルでも、その基礎となるのはデータである。不正確なデータ、不完全なデータ、またはバイアスのかかったデータは、モデルに欠陥をもたらし、信頼できない出力や誤った結論をもたらす可能性がある。例えば、あるモデルが、高リスクの借り手を過小評価した過去のデータを用いてローンのデフォルトを予測するようトレーニングされた場合、将来のデフォルトのリスクを過小評価し、財務的損失につながる可能性がある。

- **モデルの選択、チューニング、設計の欠陥：**誤ったモデル・アーキテクチャを選択したり、与えられたタスクに不適切なアルゴリズムを採用したりすると、モデルの有効性と信頼性に大きな影響を与える可能性がある。例えば、株式市場のボラティリティのような非線形性の高い現象を予測するのに、線形回帰モデルを使用すると、誤解を招く結果をもたらす可能性が高い。また、モデルの完全性を保証することも重要であり、特にオープンソースのモデルを使用する場合は、エンドユーザーがモデルの署名を検証し、正しいモデルを使用していること、モデルカードがモデルの能力と限界を正確に表していることを確認できる必要がある。

- **クラス最高のモデルに内在するリスク**：有名ベンダーが発表した最高性能のモデルであっても、幻覚、有害な表現、バイアス、データ漏洩など、モデル自体の欠点に基づく内在的リスクを抱えている可能性がある。こうしたリスクは、個々の組織だけでなく社会全体にも影響を及ぼし、広範囲に影響を及ぼす可能性がある 。[4]

- **実装と運用のエラー：**十分に設計されたモデルも、実装時に損なわれる可能性がある。不適切なコーディング、不十分なモデル制御、既存システムとの不適切な統合により、モデルの導入エラーが発生する可能性がある。例えば、クレジットスコアリングモデルが正しく開発されていても、ローン処理システムへの実装に欠陥があり、不正確なアセスメントや不当なローン拒否につながる可能性がある。セキュリティーもまた、業務上の重要なリスクである。これらのリスクは、アプリケーション・レベルやアクセス・レベルの脆弱性のような確立されたものから、プロ

[4]CNBC 企業が考える AI 活用の最大のリスクは幻覚ではない

ンプト・インジェクション[5] のような GenAI 時代の新しいものまである。AI モデルはまた、モデル利用者がモデルの利用を目的とする意思決定を変更することを狙う脅威行為者のリスクも増大させる。

- **進化する外部要因：**モデルは多くの場合、基礎となる環境が一定レベルで安定していると仮定して、過去のデータに基づいてトレーニングされる。しかし、現実の世界は常に進化している。経済不況、新たな規制、または予期せぬ出来事によって、過去のデータが無意味になり、モデルによる予測が信頼できないものになる可能性がある。例えば、過去の購買習慣に基づいて顧客離れを予測するように訓練されたモデルは、世界的な大流行によって消費者の嗜好が変化した場合、苦戦を強いられるかもしれない。同様に、過去のデータに基づいて貸し倒れを予測するように訓練されたモデルも、世界的なパンデミック、経済政策の変更、ローン活動の予期せぬ変化（新規、借り換え、条件の再交渉）などの予期せぬ出来事によって消費者行動が変化した場合、苦戦を強いられるかもしれない。どちらの例も、基礎となる環境の予期せぬ変化に対してモデルがいかに脆弱性を持ち得るかを示しており、その有効性を確保するためにモデルをモニターし、更新する必要性を強調している。

MRM フレームワークは、特に意思決定プロセスにおいて、ML モデルに関連するリスクを特定し、アセスメントし、低減し、モニタリングするための構造化されたアプローチである。MRM フレームワークを確立することは、潜在的なマイナス面を最小限に抑えつつ、ML モデルの利点を保護するプロアクティブな実践である。これは、ML モデルの責任ある信頼できる開発、展開、使用を保証するための、組織のロードマップとして機能する。具体的なリスクとその重大性（リスクレベル）は、組織、業界、事業部門、モデルの使用目的によって異なることに留意することが重要である。

[5]ワールド・ワイド・テクノロジー あなたの未来を守る：CISO のための AI ガイド

よく設計された MRM フレームワークは、モデル固有のリスクを識別し、アセスメントするための構造化されたプロセスを確立することにより、カスタマイズを可能にする。この継続的なプロセスは、以下のようないくつかの重要な構成要素に基づいて構築される：

1. ガバナンス

組織内の AI や ML モデルのガバナンスは、これらのモデルが効果的に管理され、戦略目標や規制要件と整合していることを保証するために重要である。これには、明確な目標の設定、詳細なインベントリの管理、所有者の役割の定義、承認プロセスの確立が含まれる。ガバナンスの主要な構成要素には以下が含まれる：

- **ビジネスアプローチ：**組織の全体的な AI 戦略とビジネス目標を定義し、AI を活用して生産性、効率性、意思決定、新しいユーザー体験を改善できる分野を特定する。

- **モデルインベントリ：**組織内で使用されているすべてのモデルの包括的なリストを作成し、目的、複雑さ、リスクレベル、及び確立されたビジネスアプローチとの整合性によって分類する。適切に構造化されたモデルインベントリにより、リスクレベル及び潜在的な影響に基 づく分類を通じて、高リスク又は重要なモデルについて、的を絞ったリスクアセスメント及びモ ニタリングが可能となる。

- **モデルのライフサイクル管理：**各モデルのライフサイクル（設計、テスト、開発、配備、継続的な監視と保守、非推奨） における役割と責任を明確に定義する。所有権を明確にすることで、効率的な知識の伝達と文書化が可能になり、モデルの長期的な保守と進化の妨げとなる知識のギャップやサイロ化のリスクを低減できる。

- **モデルの承認：**展開前にモデルを承認するための正式なプロセスと基準を確立し、モデルがビジネスニーズを満たし、ビジネスアーキテクチャに合致し、規制要件に準拠していることを確認する。承認プロセスでは、潜在的なバイアス、倫理的懸念、責任ある AI 原則の遵守についてもモデルを評価し、公平性、透明性、信頼性を促進する。

2. モデル開発標準

AI モデルが高品質なデータに基づいて構築され、ベストプラクティスを遵守し、関連規制に準拠していることを保証するためには、強固なモデル開発標準を確立することが不可欠である。これには、データ品質の管理、標準化された設計・開発手法の遵守、徹底した検証・テストプロセスの実施などが含まれる。モデル開発標準の主な構成要素には以下が含まれる：

- **データ品質管理：**データの多様化、知的財産・プライバシー保護対策の遵守を通じて、正確性、完全性、最小限のバイアス、最小化（データが目的に適合し、必要な情報のみに限定されること）を要求し、モデル・トレーニングのための高品質なデータを確保するためのプラクティスを定義する。

- **モデルの設計と開発：**モデルアーキテクチャ、開発方法論、文書化の標準を概説する。モデル開発標準を、規制ガイドラインを含む既存のガバナンスおよびコンプラ イアンスの枠組みと整合させる。最も著名なガイダンスの一覧は、"附属書 1：AI の枠組み、規制、ガイダンス "を参照のこと。

- **モデルの検証とテスト：**モデルの性能、正確性、安全性、堅牢性を評価するために、モデルを厳密にテストするプロセスを確立する。

- **ガバナンスとコンプライアンスの枠組み：**モデル開発標準を、規制ガイドライン（GDPR、CCPA など）、業界標準（ISO 27001、ISO 42001 など）、組織の方針による推奨を含む、既存のガバナンスとコンプライアンスの枠組みに合わせる。法的、倫理的、リスクマネジメント要件を確実に遵守するためのガイダンスについては、CSA 発行の「原則から実践へ： ダイナミックな規制環境における責任ある AI"」を参照のこと。

3. モデルの展開と使用

- **モデルのモニタリング：**運用中のモデルのパフォーマンスを継続的に監視し、精度の低下や予期せぬ動作を検知するための手順を導入する。

- **モデル変更管理：**展開されたモデルを変更するための透明性のあるプロセスを定義し、実装前に適切なテストと検証を確実に行い、使用されなくなったモデルのロールバックと非推奨のメカニズムを提供する。

- **モデルのコミュニケーションとトレーニング：**モデルの限界と能力を利害関係者にコミュニケーションし、モデルを適切に使用するためのトレーニングを提供するためのプロトコルを確立する。

4. モデル・リスクアセスメント

モデルリスク・アセスメントは、内部で開発された、あるいは外部から取得した AI や ML モデルの潜在的なリスクを特定し、低減するために不可欠である。このプロセスでは、金融、サプライチェーン、法律、規制、顧客の各領域にわたるリスクに対処する。主な構成要素には以下が含まれる：

- **リスクの範囲：**リスクアセスメントプロセスは、組織内で開発され使用されるモデルだけでなく、サードパーティや外部組織から取得したモデルにも適用される。リスク評価プロセスでは、財務、サプライチェーン、法規制、顧客維持など、組織 があらゆるレベルで取り組みたいリスクの種類を定義する。

- **リスクの特定：**これは、ML モデルに関連するリスクを効果的にマネジメントするための最初のステップである。これは、モデルのライフサイクル全体を通じて潜在的な問題を体系的に発見する技術を採用することを含む。リスク識別の際に考慮される主な要因には、データ品質、モデルの複雑さ、使用目的、トレーニングデータの取得と個人データの使用、モデルの保護メカニズムなどがある。

- **リスクアセスメント：**識別されたリスクの重大性と可能性を評価し、低減努力の優先順位付けを可能にする。リスクアセスメントは、FAIR-AI[6] のような定性的又は定量的手法を用いることができる。

- **リスク低減：**データ・クレンジング、モデル・プライバシーの改善、セキュリ ティ管理者・プライバシー管理者の導入、知的財産の保護など、特定され たリスクに対処するための戦略を策定する。これらの取組みのリスク低減効果と、コスト及び組織の環境における実行可能性とのバランスに基づいて、取組みの優先順位を決定する。

5. 文書化と報告

徹底した文書化と定期的な報告は、モデルリスク・マネジメントの透明性と説明責任 を維持するために不可欠である。これらの慣行は、モデル・ライフサイクルのすべての側面が十分に文書化され、関連する利害関係者に伝達されることを確保するものである。主要な構成要素には以下が含まれる：

- **モデルの文書化：**モデルのライフサイクルを通じて包括的な文書化を維持し、開発ステップ、仮定、制限、パフォーマンス指標を記録する。

- **モデルリスクの報告：**識別されたモデルリスク、低減戦略、モデル全体のパフォーマンスについて、関係者に定期的に報告する。

強固な MRM フレームワークは、信頼できる ML モデルの開発、展開、継続的な使用を保証する。これらのリスクを事前に特定し、アセスメントし、低減することで、企業はモデルの力を活用しながら、潜在的な落とし穴から自社と顧客、ユーザーを守ることができる。これにより、モデル主導の意思決定の信頼性と正確性が確保され、信頼と透明性が醸成される。

[6]FAIR 人工知能（AI）サイバーリスク・プレイブック

4 つの柱モデルカード、データシート、リスクカード、シナリオプランニング

このフレームワークは、4 つの重要な要素を組み合わせることで構築できる：

- **モデルカード：**ML モデルに関する明確で簡潔な情報を提供する。モデルの目的、トレーニングデータ、能力、敵対的 AI への耐性、制限、パフォーマンスについて詳しく説明し、透明性と十分な情報に基づいた利用を促進する。

- **データシート：**ML モデルの学習に使用するデータセットの詳細な説明として機能する。データシートは、作成プロセス、構成（データタイプ、フォーマット）、使用目的、潜在的バイアス、制限、データに関連する倫理的配慮を文書化したものである。

- **リスクカード：**AI モデルに関連する主要なリスクを要約する。潜在的な問題を体系的に特定、分類、分析し、開発中または配備中に観察されたリスクを強調し、現在および対応計画を説明し、モデルの責任ある使用を保証するために期待されるユーザーの行動を概説する。

- **シナリオプランニング：**モデルが誤用されたり機能不全に陥ったりする可能性のある仮想的な状況を検討し、予期せぬリスクの特定と低減戦略の策定に役立てる。

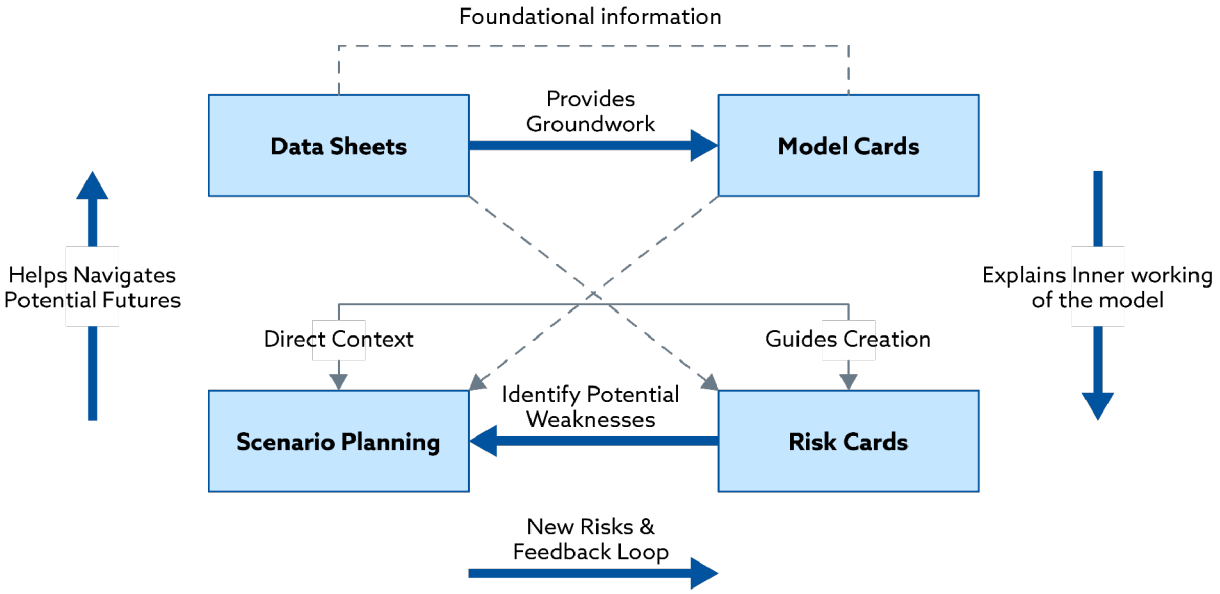

図 1.責任ある十分な情報に基づいた AI/ML 活用のためのフレームワークの柱

これらの技法は、総合的なアプローチを構築するために協力し合う。一言で言えば、モデルカードからの情報がリスクアセスメントに反映され、モデルカードとデータシートの基礎の上に、モデルの強みと限界を理解するための追加的な文脈が提供される。リスクカードはシナリオ・プランニングを導き、シナリオ・プランニングの結果はリスクマネジメントにフィードバックされ、継続的なフィードバックのループが形成される。

注：「モデルカード」の「トレーニングデータ」カテゴリーと「データシート」の「技術仕様」セクションの違いは、「モデルカード」の「トレーニングデータ」カテゴリーは、機械学習モデルのトレーニングに使用される特定のデータセットを指し、そのソース、サイズ、品質、前処理手順を含む。一方、データシートの技術仕様セクションは、データスキーマ、処理ステップ、技術的依存関係など、データセットの技術的構造と運用上の特徴を詳細に説明するものであり、モデルアーキテクチャに限定されるものではない。この違いを理解することは、機械学習モデルとデータセットの管理・維持のために、モデルカードとデータシートの両方を効果的に活用するために不可欠である。

これらの手法を組み合わせることで、組織は包括的なリスクマネジメントフレームワーク（RMF）を構築することができる：

- **透明性と説明可能性：**モデルカード、データシート、および明確なコミュニケーションにより、利害関係者はモデルの能力と限界を理解することができる。Local Interpretable Model-agnostic Explanations (LIME)、SHapley Additive exPlanations (SHAP)、Integrated Gradients、Concept Activation Vectors (CAVs)、Model Distillation などの技法は、複雑なモデルの透明性と説明可能性を高めるために、それぞれ局所的な説明を提供し、高レベルの意味概念を識別し、解釈可能な代替モデルを作成することができる。

- **プロアクティブ・リスクマネジメント：**多面的なアプローチが効果的なリスクマネジメントの鍵である。これには、モデルカードを活用して潜在的なバイアスと限界を文書化すること、データを活用することなどが含まれる。

シートによるトレーニングデータの理解、（リスクカードに基づく）徹底的なリスクアセスメントによる一般的なリスクの特定、シナリオプランニングによる将来の潜在的な課題の検討などである。さらに、敵対的テスト、ストレステスト、エッジケース分析、およびドロップアウト、L1/L2 正 規化、敵対的トレーニングのような正則化技術は、脆弱性、盲点を特定し、モデルの頑健 性を改善するのに役立ち、プロアクティブなリスクマネジメントを可能にする。

- **一貫したリスクマネジメント：**リスクマネジメントプロセスにおける一貫性は、リスクアセスメントの再現性を確保し、AI モデルのパフォーマンスと安全性を長期にわたって比較・追跡できるようにすることに重点が置かれている。一貫性のあるリスクアセスメントは、リスクの進化と低減戦略の有効性を正確に監視するのに役立ち、AI システムの継続的改善を促進する。

- **情報に基づいた意思決定：**モデルのリスクを包括的に理解することで、利害関係者はモデルの配備と使用について十分な情報に基づいた意思決定を行うことができる。

- **信頼、信用、倫理的利用の構築：**透明性と責任あるリスクマネジメントの実践は、信頼を構築し、ML モデルの倫理的利用を促進する。プライバシー保護技術の導入、倫理的 AI プラクティスの認証取得、ガバナンスフレームワークと倫理的 AI 委員会の設置、サードパーティ監査の実施は、信頼と信用を構築し、ML モデルの倫理的使用を促進することができる。

- **継続的なモニタリングと改善：**継続的なモニタリングと新しい情報に基づいて適応する能力により、モデルの有効性と安全性を持続させる。いくつかの技術には、機械学習、セキュリティ、運用（MLSecOps）のプラクティスを採用することが含まれる。これには、モデルのパフォーマン

ス、データドリフト、フィードバックループ、意図しない結果を追跡するためのモニタリングパイプラインの設定が含まれる。さらに、オンライン学習や継続的学習の手法を導入することも重要である。ユーザーからのフィードバック、インシデント報告、教訓を取り入れるプロセスを確立することで、AI システムの持続的な有効性、安全性、継続的改善を確保することができる。

包括的なフレームワークの利点

ML モデルのための包括的なリスクマネジメントフレームワーク（RMF）は、以下に定義するいくつかの利点を提供する。

透明性、説明責任、説明責任の強化

モデルカード、データシート、リスクカードは、MRM における透明性、説明可能性、説明責任のために極めて重要である。データシートは、データの出所、取得構成、前処理方法を文書化したもので、モデルの入力、制限、役割を理解するための重要なコンテクストを提供する。この文書により、モデルの内部構造をある程度理解することができ、モデルの長所、短所、潜在的なバイアスをある程度評価することができる。プロプライエタリ・モデルで利用できるものは、オープンソース・モデルと比較すると、一般的にはるかに制限されている。

プロアクティブなリスクアセスメントとシナリオ分析

データシートは、様々なシナリオの下で、モデルのパフォーマンスに影響を及ぼしうるデータ固有の特徴を詳述することにより、シナリオプランニングを補完するものである。この情報は、徹底的なリスクアセスメントを実施するために不可欠であり、シナリオ分析がデータ品質や企業に関連するその他の要因を考慮することを保証する。

リスク低減戦略の策定

データシートからの洞察をリスク軽減プロセスに組み込むことで、より的を絞った戦略が可能となる。データの限界とバイアスを理解することは、データのクレンジング、補強、再バランシングなどの効果的な軽減策の設計に役立つ。

情報に基づいた意思決定とモデル・ガバナンス

トレーニングデータとモデルの特性を詳述したデータシートは、ガバナンスの実践に情報を提供する上で極めて重要である。この詳細な理解によって、モデル展開に関する十分な根拠、文書化、透明性のある決定が保証される。トレーニングデータは交換することができるが、その品質はモデルの動作に直接影響す

る。データシートは、データの潜在的な限界やバイアス、出力に影響を与える可能性のあるモデルの特性を特定するのに役立つ。この包括的な情報は、モデルの配備に関する情報に基づいた決定につながる。データシートは、データに関連する制約や機会を明らかにすることで、意思決定プロセスに影響を与える重要な情報を提供する。MRM では、このデータ特性の詳細な理解がガバナンスの実践に役立ち、モデル展開に関する意思決定が十分な根拠を持ち、文書化され、正当化されることを保証する。

ロバストモデルの検証

ロバストなモデル検証は MRM のフレームワークにとって不可欠であり、モデルが期待通りの性能を発揮し、実世界の状況に適応することを保証する。これには、実世界のシナリオを反映した多様なデータセットによる厳密なテストが含まれる。データシートから得られる情報、例えばデータ分布や潜在的なバイアスは、より包括的な検証プロセスのためのデータセットの選択に役立つ。この検証プロセスには、多様性テスト、ストレステスト、汎化可能性メトリクスのような技術が不可欠である。これらの検証を取り入れることで、フレームワークは、モデルが有効性を維持し、実世界のアプリケーションにおいて予期せぬパフォーマンスの問題やバイアスのかかった結果を回避することを保証する。

信頼を築き、モデルの採用を促進する

データシートは、データの明確性を確保することによって信頼の下地を形成する。しかし、信頼を築くには、多層的なアプローチが必要である。モデルカードはモデルの内部構造についてより深い洞察を提供し、リスク・カードは潜在的なバイアスや限界に積極的に対処する。これにより、透明性と責任ある AI 開発が促進され、最終的には、モデルの採用に対するユーザーと規制当局の信頼の向上につながる。これらの文書は、モデルの能力と期待されるパフォーマンスについて、透明で正直なコミュニケーションを提供する。この透明性は、特にデータの出所と整合性が重要なセクターにおいて、ユーザーと規制当局の信頼を得るために極めて重要である。

継続的なモニタリングと改善

継続的なモニタリングは MRM のフレームワークにとって不可欠であり、モデルが期待通りに作動し、時間の経過に伴う変化に適応することを保証する。これには、モデルカード、リスクカード、データシートを定期的に更新し、モデルのパフォーマンスや運用環境の変化を反映させることが含まれる。例えば、精度、精度、F1 スコアなどの指標を追跡して性能を測定し、平均絶対誤差（MAE）や平均二乗誤差

（MSE）を使用してモデルのドリフトを評価することができる。継続的なモニタリングは、モデルが意図したパフォーマンスからいつ逸脱するか、外部環境の変化によりモデルの調整や展開戦略がいつ必要になるかを特定するのに役立つ。この継続的な監視は、ダイナミックな運用状況において、ML モデルの持続的なコンプライアンス、有効性、安全性を確保するのに役立つ。

社会的・倫理的にポジティブな影響

データシートは、ML モデルにおける社会的・倫理的バイアスに対処するための基礎となる。学習データの出所、構成、前処理方法を文書化することは、潜在的なバイアスを特定するための重要な透明性を提供し、公正かつ公平な ML モデルを開発する上で極めて重要である。データの取り扱い方法が倫理標準に沿ったものであることを保証することで、組織は、その技術の広範な影響をより適切に管理することができる。

強力なガバナンスと監視体制

強力なガバナンスと監視は、組織の目的との整合性を確保する制御の基盤の上に構築され、倫理的認識と能力のある個人によって導かれた、透明で説明可能かつ説明責任のある AI モデルの開発、使用、保守を保証する。また、倫理的ガイドラインと責任あるデータ運用が守られるよう、強固な実施メカニズムを確立する。効果的なガバナンスには、明確な役割と責任、明確な意思決定プロセス、対立や紛争を解決するためのエスカレーション手順が含まれる。定期的な監査は、これらの原則に対する利害関係者のコミットメントを検証し、説明責任のレイヤーを提供する。厳密な変更マネジメント手続き、管理策の更新、再教育、配備の決定が、監督を促し、潜在的なリスクを積極的に軽減する。ユーザー、データサイエンティスト、エンジニア、ビジネスリーダーを含むステークホルダー間の明確なコミュニケーションとコラボレーションは、ガバナンスと監視を成功させるために不可欠である。

コア・コンポーネント

1. モデルカード：モデルを理解する

モデルカードは、モデルの透明な概要を提供する。モデルカードは、モデルの目的、トレーニングデータ、能力、限界、パフォーマ ンス指標を詳細に説明する。この情報は、開発者、展開者、リスクマネジメント専門家、コンプライアンス担当者、エンドユーザーがモデルの長所と短所を理解し、リスクアセスメントの基礎を形成するのに役立つ。

モデルカードの主要要素には通常、以下のものが含まれる：

- **モデルの詳細と意図される目的：**これはモデルの機能と目標を明確にする。
- **トレーニングデータの詳細：**モデルの学習に使用するデータの構成について、その出典、サイズ、取得方法（同意、寄付など）、倫理的配慮、潜在的バイアスなどを記述する。詳細については、データシートへのリンク（利用可能な場合）を提供することができる。
- **意図する使用例と限界：**これは、そのモデルがどのような用途に使用できるのか、また、どのような場合にはうまく機能しないのかを説明するものである。
- **パフォーマンス・メトリクス（評価指標）：**これは、精度と一般化可能性のような明確なメトリクスを使用して、関連するタスクでモデルがどの程度うまく機能するかを概説する。
- **採用した評価方法：**モデルのパフォーマンスを評価するために使用された方法を説明する。
- **モデルの説明可能性とバイアス：**このセクションでは、モデルの意思決定プロセスを理解し、潜在的なバイアスを特定するための手法を説明する。また、バイアスを低減し、異なるグループ間で公平な結果を保証するための方法についても詳述する。
- **既知の限界：**これは、特定のプロンプトや事実誤認の影響を受けやすいなど、モデルの潜在的な欠点を認めるものである。

- **持続可能性と環境側面（オプション）：**利用可能な場合は、モデルのトレーニングによる環境への影響（炭素排出量など）を推定する。

- **敵対的抵抗力（敵対的攻撃オプションにおけるパフォーマンス指標）：**通常、敵対的なトレーニングの具体的な詳細はモデルカードには記載されないが、我々の経験に基づき、モデルカードとリスクカードの評価セクションに敵対的な耐性メトリクスを記載することを推奨する。データサイエンティストは、シミュレートされた敵対的攻撃下での精度メトリクスを報告することで、モデルのレジリエンスを実証することができ、モデルの性能と潜在的脆弱性をより包括的に理解することができる。

モデルカードの利点

モデルカードは、責任ある AI の開発と展開に貢献し、リスクマネジメントの基礎となる、以下のような豊富な利点を提供する：

- **洞察と透明性：**モデルカードは関係者をガイドし、モデルの設計、開発プロセス、および配備を理解するのに役立つ。モデルカードは、トレーニングデータとモデルのパフォーマンス指標を明らかにし、ユーザーがその能力と限界を把握できるようにする。

- **潜在的リスクの識別：**モデルカードは、トレーニングデータの構成を概説することで、出力が不公正または差別的な影響を受ける可能性のあるバイアス、著作権違反、トレーニングデータとは異なる文脈でモデルがうまく機能しない可能性のある限定的な一般化可能性、トレーニングデータの不正確さに起因する事実誤認などの潜在的な問題を明らかにすることができる。

- **再現性／説明責任：**モデルカードは開発プロセスを文書化し、他の人がモデルを再現し、そのリスクを独自に評価することを可能にする。

リスクマネジメントの基礎

モデルカードは、ML モデルの効果的なリスクマネジメントの基礎となるもので、以下のようなモデルに関する重要な情報をプロバイダとして提供する：

- **トレーニングデータの特徴：**潜在的なプライバシー侵害、著作権侵害、バイアスを明らかにする。

- **挙動と性能の限界：**モデルが信頼できない、あるいは誤解を招くような出力を生成する可能性のある状況を予測する。

リスク低減のメリット

- **オーダーメイドの低減戦略：**リスクの種類を把握することで、関連する低減戦略を模索し、例えば、有害なコンテンツを生成するようなリスクに対する特定のセーフガードを開発するなど、許容可能な実装の複雑さで最もリスク低減の可能性が高いものに焦点を絞ることができる。

- **コミュニケーションと透明性：**ステークホルダーとのコミュニケーションと責任ある利用の促進

- **プロンプトの設計を導く：**安全で正確な回答のためにプロンプトをデザインする

- **コンプライアンスと信頼：**規制へのコンプライアンスをアセスメントし、信頼を醸成し、モデルの信頼性と安全性について十分な情報に基づいた判断を保証する。

- トレーニングデータのキュレーションデータの品質と公平性を確保する

- **ガードレールを実装する：**意図しない出力を防ぐ技術を文書化する

要するに、モデルカードは包括的な記録として機能し、責任ある AI の開発と導入を促進し、リスクマネジメントと軽減の基盤を確立する。

モデルカードの作成と更新

モデルカード作成の要点

効果的なモデルカード作成には、正確性と効率性を確保するために、協調的かつ自動化されたアプローチが必要である。最も一般的なベストプラクティスは以下の通りである：

- **プロセスとオーナーシップ：**モデルカードを作成し維持するための明確なプロセスと所有権を組織内に確立しなければならない。主要リーダーは、このプロセスを実施し、各モデルカードの特定のオーナーを任命する責任を負う。選ばれたオーナーは、適切な質問をし、必要な情報を収集し、組織全体のコラボレーションをリードするスキルを持つべきである。理想的には、モデルカード構築の経験があるか、または十分な技術的知識を持ち、迅速に学習できることが望ましい。

すべてのモデルにモデルカードが必要なわけではないので、モデルカードが必要な場合、例えば、100人以上が使用するモデルや、生産またはテストにおいてモデルカードが必要な場合について、明確なガイドラインを定めるべきである。

- **コラボレーション：**包括的なカバレッジを確保するために、機能横断的なチームを作成プロセスに参加させる。
- **テンプレート**一貫性と使いやすさを確保するために、標準化されたテンプレートを使用する。
- **自動化：**自動化ツールを活用してモデルカードを生成し、手作業を減らして精度を高める。
- **バージョン管理：**バージョン管理システムを活用して変更を追跡し、更新の記録を明確に保つ。
- **モデルカードのリポジトリ：**モデルカードの一元的なリポジトリを確立し、容易なアクセスと管理を保証する。

モデルカードを最新の状態に保つ

定期的な更新は、モデルカードが正確かつ適切であり続けるために不可欠である。合理的な更新プロセスを導入することで、手作業を減らし、効率を高め、以下を含むべきである：

- **定期的なレビュー**モデルやデータの変更を反映するために、モデルカードを定期的に見直す。
- **自動更新：**モデルカードの更新に自動化ツールを活用することで、手作業を減らし、精度を高める。
- **変更管理を行う：**更新を文書化し、適切に承認するプロセスを確立する。

- **監査証跡：**透明性と説明責任を確保するために、すべての更新と変更の監査証跡を維持する。

モデルカードを作成・更新するための合理的で効率的なプロセスを作成するために、さらにいくつかの高度な技術を活用することができる。例えば、ML アルゴリズムは、モデルのパフォーマンスを分析し、モデルカードを動的に更新することができ、自然言語処理アルゴリズムは、モデルカードのコンテンツを自動的に生成することができる。可視化ツールは、モデルのパフォーマンスと更新をグラフィカルに表現し、複雑なデータを理解しやすくする。バージョン制御やコラボレーションプラットフォームなど、他のツールやシステムとモデルカードを統合することで、コラボレーションを強化し、手作業を減らすことができる。これらのアプローチは、精度、効率、コラボレーションを向上させ、プロセスを改善することができる。

モデルカードの限界

- **完全性と正確性：**詳細は、モデルカードがいかに完全かつ正確に記入されているかにのみ依存している。このため、特にこのプロセスが主に手作業である場合には、誤解を招いたり、情報が不完全になったりするリスクが残る。そのため、可能な限りデータ収集を自動化することを提唱する。しかし、完全性と正確性を確保するためには、モデルカードの更新と保守を優先するよう、経営陣がスポンサーとなり、組織内の文化的転換を図ることも必要である。リーダーシップの賛同がなければ、たとえ善意の開発者であっても、モデルカードの作成と更新を軽視し、このリスクマネジメントツールの有効性を阻害する可能性がある。

- **静的な表現**：モデルカードは、特定の時点でのモデルの貴重なスナップショットを提供するが、その静的な性質が問題となることがある。モデルが更新され改善されるにつれて、モデルカードに文書化された情報は古くなる可能性がある。そのため、モデルの現在の状態を正確に反映するために、モデルカードを定期的に見直し、更新する必要がある。

- **評価における主観性**：標準化されたベンチマークや評価基準が存在しないため、公平性や倫理的配慮に焦点を当てたモデルは、本質的に主観的なものとなりうる。

- **限られた範囲：**モデルカードは、アーキテクチャー、トレーニングデータ、パフォーマンスメトリクスのような技術的な詳細をプロバイダとして提供するが、モデルの影響を包括的に扱うには不十分であることが多い。この限られた範囲では、モデルの実世界での適用から生じる潜在的バイアス、倫理的考察、社会的意味合いを見落としてしまう可能性がある。

- **ディテールのレベルが異なる：**モデルカードには標準フォーマットがない。詳細度や明瞭度が異なるため、異なるモデル間での比較やリスクアセスメントが難しくなる。

モデルカードは、ML モデルとその潜在的リスクを理解するための貴重なツールである。透明性を促進し、開発者とユーザーがモデルの長所と短所を理解することを可能にする。

2. データシート：トレーニングデータを検証する

モデル設計図のデータシートは、ML モデルの詳細な技術的説明を提供する。これは、開発者、リスクマネジメント、監査人のための参照文書として機能し、モ デルの構築パラメータと運用上の特徴を詳述している。この情報は、モデルの潜在的な強み、弱み、内在するリスクを理解するために極めて重要である。

データシートの必要性

モデルカードとリスク・カードはリスク・マネジメントのための貴重な洞察を提供するが、モデルの内部ロジックの透明性のあるビューという不可欠な要素を追加する必要がある。データシートは、効果的なモデル・リスクマネジメントのための基礎文書として、このギャップを埋めるものである。ここでは、データシートがどのように信頼を醸成し、より多くの情報に基づいたリスクアセスメントを可能にするかを紹介する：

- **モデルの透明性：**モデルがどのような判断を下すのかを理解することは、リスクマネジメントにとって極めて重要である。モデルカードはハイレベルな概要を提供し、リスク・カードは潜在的な問題を浮き彫りにするが、モデルの内部構造までは掘り下げていない。データシートは、モデルのロジックを深く掘り下げることで、このギャップに対処する。この透明性は、モデルに対する信頼を醸成し、リスクマネジメントがモデルの限界や潜在的バイアスについて、より多くの情報に基づいたアセスメントを行うことを可能にする。

- **リスクアセスメント：**モデルの構築とトレーニングデータを理解することで、リスクマネジメントは、データ品質の問題、オーバーフィット、アルゴリズムのバイアスなど、モデルリスクの潜在的な原因を効果的に評価することができる。

- **モデルガバナンス：**データ仕様は、モデルガバナンスの礎となり、必要に応じてモデルの継続的なモニタリング、保守、再トレーニングを促進する。

- **再現性：**詳細な仕様により、独立した関係者がモデルを再現し、検証することができ、その出力に対する信頼と信用を促進する。

MRM におけるデータシートの役割

データシートは、単にモデルのロジックを文書化するだけでなく、プロアクティブなリスクマネジメントを強化し、モデルの適合性を確保する。データシートは、継続的な改善とコンプライアンスのためのロードマップを提供し、MRM ライフサイクルにおいて以下のような重要な機能を果たす：

- **リスクの特定と低減：**データ仕様により、リスクマネジャーは、モデル内の潜在的な障害点を積極的に特定し、低減戦略を策定することができる。

- **モデルの検証と改良：**文書化された訓練プロセスとパフォーマンス測定基準により、モデルの有効性と一般化可能性を厳密に検証することができる。また、データ仕様は、特定されたバイアスや性能の限界に対処するため、モデルの継続的なキャリブレーションと改良の基礎となる。

- **規制コンプライアンス：**包括的なデータ仕様は、AI/ML モデルの開発と展開において、関連する規制や倫理的ガイドラインの遵守を実証する上で重要な役割を果たす。

データシートの主な要素

データシートは、以下のようなモデルの内部構造の概要を簡潔かつわかりやすく説明している：

- **モデルの目的と範囲：**モデルが何を達成するために設計されているのか、またその使用における制限を明確に定義する。

- **データの入力と仮定：**データソース／タイプ／フォーマット、適用される前処理変換ステップ、および基礎となる仮定を含む、モデルが使用するすべての入力機能の詳細なリスト。

- **モデルのアーキテクチャ：**ハイパーパラメータの設定（学習率、層数）、選択したアルゴリズムなど、モデルのアーキテクチャ（決定木、ニューラルネットワークなど）の技術的説明。

- **モデル開発プロセス：**使用した関連アルゴリズムを含め、モデルの構築とトレーニングのステップを簡単に概説する。

トレーニングデータの特性：トレーニングデータの特徴：モデルの開発に使用されたトレーニングデータの内訳で、データソース、サイズ、分布特性、および実行されたデータ品質チェックを含む。

トレーニングプロセス：トレーニングプロセス：選択した最適化アルゴリズム、成功の目的、収束基準など、トレーニングプロセスの文書化。

パフォーマンス指標：これは、訓練データセットと検証データセットにおけるモデルの有効性を評価するために使用される指標の包括的なセットである（例えば、精度、精度、再現性、F1 スコア）。

- **モデルの出力と解釈：**データ型や生成的結果の解釈方法を含む、モデルの出力形式の明確な定義。

- **仮定と限界：**モデル開発中に行われた仮定、および選択されたモデルアーキテクチャまたはトレーニングデータに内在する限界の透明な開示。

データシートの限界

データシートには大きな利点があるが、効果的に使用するためには、その限界を認識することが極めて重要である。データシートは、複雑さと範囲において課題を提示し、進化する AI/ML の分野と歩調を合わせることができる。これらの限界には、以下のようなものがある：

- **複雑さ：**学習データセット、選択されたアルゴリズム、機械学習オペレーション（MLOps）制御体制、測定されたパフォーマンス指標など、AI/ML フレームワークの特定の構成要素によっては、データ仕様が非常に技術的になり、完全に理解するには ML の専門知識が必要になる。

- **限定された範囲：**データ仕様は、主にモデルの技術的側面に焦点を当てている。モデルのアウトプットが持つ、より広範なビジネス上の背景や潜在的な社会的意味合いを十分に把握できない可能性がある。

- **進化する分野：**AI/ML が急速に進化するにつれて、データ仕様のベストプラクティスは、新しい技術や方法論を取り入れるために継続的に適応する必要があるかもしれない。

- 完全性や正確性、企業文化化、静的／古い表現など、**モデルカードに共通する限界**は、データシートにも当てはまる。

データシートは、モデルのリスクマネジメントに不可欠なツールである。データシートは、モデルの構築と運用に関する技術的なロードマップを提供することで、リスクマネジメントの専門家が ML モデルに関連するリスクを効果的にアセスメントし、軽減し、管理できるようにするものである。

3. リスクカード：潜在的な問題の識別

リスクカードは、AI モデルに関連する潜在的な問題を深く掘り下げる。潜在的なリスクを体系的に特定し、分類し、分析する。潜在的なモデルリスクのフラッシュカードのようなものだと考えてほしい。各カードは、特定のリスク、潜在的な影響、低減戦略について説明している。フラッシュカードと同様に、モデルの脆弱性を理解し、対処するための迅速かつ構造的な方法を提供する。

リスクカードには通常、以下のような潜在的な懸念事項が含まれている：

- **安全性と倫理的リスク：**プライバシー、有害コンテンツの生成、バイアスの助長などの問題が含まれる。

- **セキュリティリスク：**データ漏洩、不正操作の試み、その他のセキュリティの脆弱性がこのカテゴリーに該当する。

- **社会的リスク：**雇用の奪い合いや、プロパガンダのための AI の悪用は、社会的リスクの一例である。

- **環境リスク**：AI モデルは多くの電力を使用するため、有害なガスの生成的増加を招く可能性がある。クリーンエネルギーを使用するモデルでさえ、他の社会的用途からエネルギーを奪うため、有害なガスの発生を余儀なくされる。

- **運用リスク：**モデルは、限られたトレーニングデータ、計算強度、既存システムとの統合などに関する課題に直面する可能性がある。

- **規制・法的リスク：**組織は、最初の導入時や、時間の経過とともに変化する LRR によって、法律、規則、規制（LRR）に抵触する可能性がある。あるいは、知的財産権の所有者から、入力データの使用に異議を唱えられるかもしれない。

- **財務リスク**：エージェント的なワークフローを使用するなど、モデルにサービスを提供するためのコストが予想外に増加する可能性がある。

- **サプライチェーンのリスク**：社外から持ち込まれるリスクや、当社のモデルからパートナーに持ち込まれる可能性のあるリスクに関するものである。

- **評判リスク**：不適切なモデルの使用は、ネガティブな報道などにつながる可能性がある。

リスクカテゴリーが組織によって異なったり、少なくとも各リスクカテゴリーに対す る焦点の深さが異なったりする可能性があることに留意されたい。例えば、NIST AI RMF[7] では、モデルが「妥当で信頼できること、安全であること、セキュアでレジリエンシーであること、説明可能であること、透明であること、説明可能で解釈可能であること、プライバシーが強化されていること、有害なバイアスがマネジメントされた公正であること」に対するリスクに焦点を当てている。

[7]NIST AI 100-1 "人工知能リスクマネジメントフレームワーク（AI RMF 1.0）"

リスクカードの構造

各リスクカードは、具体的なリスクを理解し、的を絞った低減戦略を策定するために、焦点を絞った有益なアプローチを確実にするために、明確に定義された構造に従っている。各リスクカードには、一般的に以下の要素が含まれている：

- **リスクカテゴリー：**リスクを分類する（バイアス、事実誤認、誤用など）。
- **リスクの説明：**バイアス、事実誤認、有害なコンテンツの生成など、潜在的な問題の簡潔な説明。
- **影響：**風評被害、ユーザー被害、法的問題などの要素を考慮したリスクの潜在的な影響。
- **重大度レベル：**リスクの潜在的影響をアセスメントする（高、中、低）。
- **可能性**：リスクが発生する確率を評価する。
- **低減戦略：**リスクの可能性または重大性を低減するための実行可能なステップには、データのフィルタリング技術、トレーニングデータの改善、より安全な出力に向けてモデルの開発を誘導するユーザープロンプト、および運用戦略や組織戦略が含まれる。

下表はリスクカードの例である。

リスク	説明	影響	低減戦略
バイアスと公平性	モデルはトレーニングデータに基づいてバイアスのかかった内容を出力する。	差別を助長し、風評被害をもたらす可能性がある。	• 多様なトレーニングデータを使用する • モデルに公平性チェックを導入する • 制限に関する透明性を提供する

このリスクカードは、ある小売企業がマーケティングやソーシャルメディアコンテンツの生成に使用している ML モデルに、意図しないバイアスがかかっている可能性を浮き彫りにした。明確な説明と「高い」潜在的影響（高い重大性）により、データチームは問題への対処を優先した。同社は、潜在的なバイアスを調査するために、トレーニングデータとモデルアーキテクチャの二重レビューを実施した。データチームは、データのデモグラフィックを分析し、表現の偏りを特定し、潜在的なバイアスがないかトレーニングデータのソースを調査した。また、潜在的なバイアスを定量化するための公平性の指標について議論し、モデルがどのように出力に到達するかを理解するために解釈可能性手法などの技術を使用した。

この分析に基づき、いくつかの低減戦略が実施された：

- **データのバイアス除去：**オーバーサンプリング・アンダーサンプリングによるトレーニングデータのバランス調整と、重要でないセンシティブ属性の除去により、よりバランスの取れたデータセットを作成した。同社はまた、合成データを使用してバイアスに対処することも検討している。

- **トレーニングにおける公平性：**偏った出力にペナルティを与え、適切な出力を強化するために、訓練プロセスに公平性制約が組み込まれた。

- **後処理フィルター：**センチメント分析やファクトチェックツールを導入し、生成後にバイアスのかかった可能性のあるコンテンツを特定し、フラグを立てる。

こうした低減戦略だけでなく、同社はバイアスに対するチームの防御を強化するため、綿密なコンティンジェンシープランを策定した。このコンティンジェンシープランには以下のようなものが含まれる：

- **偏った出力にフラグを立て、対処する：**バイアスのかかったアウトプットに明確にフラグを立て、対処するプロセスには、バイアスのかかったコンテンツを特定し、修正できる人間のレビュアーが関与する。

- **インシデント対応プロトコル：**リスクカードシナリオがトリガーされたとき、迅速な調査と低減を確実にするために、AI/ML 運用チームが活用できるインシデント対応プロトコルを組織がすでに確立していれば、非常に有益である。検知には、バイアス検知の場合のように、よりバランスの取れたデータセットでモデルを再トレーニングすることも含まれる。

- **コミュニケーションプロトコル：**潜在的なバイアスに関する社内横断的なコミュニケーション・プロトコルは、透明性を確保し、ユーザーや利害関係者との信頼を育み、組織全体で責任あるモデル使用を促進する。

これらの低減戦略、特にデータの多様性とアルゴリズムの公平性に焦点を当てた戦略を実施することで、チームはモデルの出力におけるバイアスに対して積極的な姿勢をとった。これにより、組織全体で信頼できる倫理的な AI システムを構築する基盤が確立され、AI アプリケーションにおける包括性、透明性、説明責任を促進することが可能になった。

リスクカードのメリット

リスクカードは、日々進化するモデルリスクをマネジメントするための体系的かつダイナミックなアプローチを提供する。リスクカードは、モデルリスクを特定し、分類し、優先順位をつけるための体系的な方法を提供し、開発者、ユーザー、利害関係者間のディスカッションを促進する強力なコミュニケーションツールとして機能する。この協力的な環境は、潜在的な問題に対するより深い理解を促進し、低減戦略やコンティンジェンシープランといった実行可能な洞察の開発につながる。

これらの中核的な利点に加え、リスクカードは MRM に特化した大きな利点を提供する：

- **プロアクティブなアプローチ：**リスクカードは、潜在的な問題を事前に特定し、先手を打った解決策を可能にする。このアプローチにより、各戦略の潜在的なリスク削減効果と、その複雑さやコストを比較評価することができ、最高の投資対効果でプロアクティブな低減を実現する。

- **ストレステスト：**リスクカードは、潜在的なリスクに関する議論やブレーンストーミングを促すことによって、 様々な条件下でのモデルのストレステストのプロセスを容易にする。リスクカードはストレステストの出発点である。実際のストレステストでは、リスクカードで特定されたリスクの下でモデルがどのよう に振る舞うかを分析するために、定量的および定性的な技術を適用します。ストレステストの結果は、一般的にリスクカードには記録されませんが、リスクカードの再 度繰り返しに反映される可能性があります。

- **意思決定の改善：**リスクカードは、包括的なリスクの特定と分析を通じて、組織がモデルの導入と適切なユースケースの選択について、十分な情報に基づいた選択を行うことを支援する。これにより、関連するリスクを最小限に抑えながら、モデルを効果的に活用することができる。

リスクカードの限界

- **範囲が限定されている：**リスクカードは通常、あらかじめ定義された潜在的な問題のセットに焦点を当てる。これは、一般的なリスクをカバーするには有益であるが、AI モデル特有の脆弱性をすべて把握できない可能性がある。また、定量化が不十分であるため、リスクの影響や可能性のアセスメントに支障をきたし、軽減策の優先順位付けが難しくなる。さらに、複雑なリスクや微妙なリスクは、単純化されすぎたり、凝縮されすぎたりする可能性があり、深刻度や緩和の課題を過小評価することにつながる。

- **AI のダイナミックな性質：**AI モデルは常に進化し、新たなリスクが出現する可能性がある。リスクカードは、この分野の急速な発展に対応する必要がある。

- **不十分な定量化：**リスクカードはリスクの定性的なアセスメントを提供するが、各リスクの潜在的な影響と可能性を定量化するには不十分な場合がある。定量的な尺度がなければ、組織は AI モデルに関連する最も重大なリスクを軽減するために優先順位を付け、効果的にリソースを配分するのに苦労することになる。

- **実世界のデータ依存性：**リスクカードの有効性は、リスクの識別とアセスメントに使用されるデータの質と包括性 に依存する。不完全または不正確なデータは、誤解を招いたり、無関係なリスクカードになる可能性がある。

- **人間の判断が必要である：**リスクカードは、リスクの重大性を解釈し、適切な低減戦略を選択するために、人間の判断 を必要とする。これは主観的なものである可能性があり、カードをレビューする人の専門知識に依存する可能性がある。

4. シナリオ・プランニング：「もしも」の場合のアプローチ

シナリオ・プランニングとは、AI モデルが誤用されたり誤作動を起こしたりする可能性のある仮想的な状況を探索する積極的なアプローチである。基本的には、「もしも」を問うことである。さまざまな肯定的、否定的な状況において、AI モデルがどのように振る舞うかを想像し、探求する。これにより、現実になる前に潜在的なリスクを特定することができる。

シナリオ・プランニングはこう考える

- ポジティブ・シナリオ（生産性の向上、教育の改善など）
- ネガティブ・シナリオ（言葉の武器化、情報操作など）

シナリオ・プランニングで考慮すべき点

- **技術的能力：**モデルの長所と短所を評価し、誤動作（定期的なものから「ブラックスワン」[8]）や操作、搾取の影響を受けやすい分野に焦点を当てる。
- **データのバイアス：**モデルの出力に影響を与える可能性のある、信頼性の低いベンダーのデータ、欠損データ、範囲外のデータ、時間の経過が不安定なデータなど、トレーニングデータに存在する潜在的なバイアスやデータの特性を調べる。
- **ユーザーとの相互作用：**ユーザーがモデルとどのように相互作用し、その意図や理解がどのように意図しない結果につながるかを考慮する。
- **社会的影響：**雇用の転換や、自動化をめぐる倫理的な懸念、組織外の人々によるモデルの使用によるリスクなど、モデルの展開が社会により広く影響を及ぼす可能性を探る。

シナリオ・プランニングの仕組み

シナリオ・プランニングには、仮想的な状況を通じて潜在的なモデルリスクを特定し、評価するための構造化されたアプローチが含まれる。以下は、そのプロセスの内訳である：

[8]ウィキペディア ブラック・スワン理論

1. チームを編成する

テクノロジー、リスクマネジメント、倫理、法務、規制コンプライアンス、または特定のデータ領域やアプリケーション領域の専門知識を有する多様なチームを集める。理想的なチーム構成は、プロジェクト固有の要件によって異なり、以下の利害関係者の組合せを含むことができる：

ビジネスの専門家

ドメインの専門家：特定のアプリケーションドメイン（例：ヘルスケア、金融）を深く理解する個人は、実際のユースケースに関連するシナリオを探索するための貴重なコンテキストを提供することができる。

エンドユーザー：エンドユーザー：想定されるユーザーグループの代表者を含めることで、潜在的なユーザーとのインタラクションや、モデルが意図せずに誤用される可能性について洞察することができる。

リスクの専門家

セキュリティ実務者：セキュリティ実務者：脅威のモデル化、脆弱性モデルの影響度や可能性の定量化に関する経験を有す る者が、リスクに関する議論を支援する。

プライバシーと法律に関するアドバイザー：組織と使用されるデータの特定の法的背景に関する知識を持つ専門家、およびプライバシーと情報ガバナンスの専門家は、個人データを処理するモデルのプライバシーに関する考慮事項について助言することができる。

リスクマネジメントのスペシャリスト：リスクの特定と低減の経験を生かし、シナリオ・プランニングへの構造的かつ包括的なアプローチを確保する。

倫理アドバイザー：倫理的配慮に関する専門知識は、潜在的な社会的影響を探り、責任あるモデル開発を保証するのに役立つ。

AI エキスパート

モデル開発者：モデル開発者：モデル・アーキテクチャと機能に関する彼らの専門知識は、システムの能力と潜在的脆弱性に関する貴重な洞察を提供する。

データサイエンティスト：データサイエンティスト：モデルのトレーニングデータと潜在的バイアスに関する知識は、公平性と表現リスクの特定と推定に役立つ。データサイエンティスト：モデルの学習データと潜在的なバイアスに関する知識は、公平性と表現のリスクの特定と推定に役立つ。

このように多様な視点を結集することで、シナリオ・プランニング・チームは AI モデルをよりよく理解し、潜在的なリスクを幅広く特定することができる。この共同アプローチは、製品のレッドチームに似ている。そこでは、多様な専門知識と視点が、アイデアのストレステストと潜在的脆弱性の特定に活用される。このアプローチは、リスク低減のためのアプローチなど、ブルーチーミングの能力も可能にする。このアプローチの有効性は、効果的なアイデア出しとリスクアセスメントを促進するために必要な基礎力を備えたチームを編成することにかかっている。

2. 範囲と目的を明確にする

次のステップでは、シナリオ・プランニングの範囲と目的を明確に定義する。これには、AI システムと探りたいリスクを特定することが含まれる。潜在的バイアス、セキュリティの脆弱性、社会的影響の特定など、明確な目的を設定することで、チームの集中力を高め、生産的なシナリオ・プランニング・セッションを行うことができる。

3. 飛び込むべきシナリオの優先順位を決める

多様な視点を提供するグループは、包括的な潜在的シナリオを提案するのには適しているが、完全に計画し尽くすには実現不可能なリストを提案しやすい。そのため、しばしば慎重な優先順位付けが必要となる。チームは、ROI の比較のために、「リターン」（例えば、潜在的なリスク影響対削減）と「投資」（例えば、シナリオプランニングと実施にかかるかもしれない労力）の定義に関するいくつかの「T シャツ」サイズなど、優先順位付けのアプローチを選ぶべきである。重要なのは、チームが、それほど詳細に計画されないシナリオのリスクに対して、リーダーシップが納得できる方法で優先順位をつけることである。

4. 情報を集める

チームは、AI モデルと潜在的リスクを包括的に理解するために、関連情報を収集すべきである。モデルカード、データシート、リスクカードは、ML モデルの能力、限界、潜在的リスクに関する貴重な洞察を提供する。これらの文書には、トレーニングデータ、モデルのアーキテクチャ、既知の脆弱性が詳細に記

載されている。さらに、モデルに関連する安全インシデントや誤用事例を調査することは、潜在的な現実世界の脅威を予測するのに役立つ。収集した情報は、シナリオを計画するのに十分な詳細さでなければならないが、それ以上であってはならない。

5. シナリオを作成する

シナリオ・プランニングの核心は、創造的に多様な仮定の状況を生み出すことにある。既成概念にとらわれず、ポジティブなシナリオとネガティブなシナリオを探求するようチームに促す。what-if」の質問のような技術は、創造的思考を刺激し、より幅広いシナリオを生み出すことができる。例えば、カスタマーサービスで使用される大規模言語モデル（LLM）が、どのように操作されればバイアスのかかった回答を生成できるのか、あるいは、金融の場面でモデルの誤作動がどのように不正確な投資推奨につながるのか、といった具合である。

6. シナリオを評価する

シナリオを作成したら、チームはそれぞれのシナリオを体系的に分析する必要がある。これには、シナリオが発生する可能性と、シナリオが実現した場合の潜在的な結果を検討することが含まれる。シナリオが、ユーザー、社会、組織を含む様々な利害関係者に与える影響をアセスメントする必要がある。各シナリオがモデルの正確性、信頼性、公平性、安全性にどのような影響を与えうるかを検討する。例えば、LLM による誤情報の拡散を探るシナリオでは、社会的な損害や組織への風評被害の可能性を考慮する必要がある。

言語モデルを使用して、これらのシナリオをシミュレートすることもできる。その出力を観察し、差別的なテキストの生成、誤情報の拡散、有害なコンテンツの作成など、潜在的なリスクを特定する。

この段階は、スコープクリープ（当初予算よりも作業量が増えること）が最も発生しやすいので、慎重で規律あるプロジェクト管理が重要である。過度な時間管理もリスクとなる。理想的には、評価の深さと主要シナリオのカバー率のトレードオフを、シナリオの優先順位を前もってきちんと決めておけば、管理しやすくなる。

7. 低減戦略を策定する。

シナリオの分析に基づいて、リスクを軽減したり、将来の課題に適応したりするための戦略を策定する。組織に重大なリスクや脅威をもたらす潜在的なシナリオに対処するために、コンティンジェンシープランと対応戦略を策定する。これらの戦略には、操作に対するセーフガードの導入などの技術的制御、責任あるモデルとの相互作用に関するユーザートレーニングなどの非技術的措置、または AI ガバナンスプロセスにおける透明性と説明責任の強化が含まれる。さらに、潜在的なバイアスに対処するために、多様なトレーニングデータセットの採用など、モデル開発プロセスの調整を実施することもできる。

8. 実施すべき低減戦略の優先順位を決める。

多様な視点を提供するグループは、影響のある低減戦略を提案するのに適しているが、組織にはすべてを一貫して実施するためのリソースがないかもしれない。したがって、実施すべき戦略の優先順位を慎重に決めることで、主要なリスクが実際に低減される確率を高めることができる。チームは、すべての主要リスクが低減され、優先順位を下げた戦略が実際に確率の低い、影響度の低いリスクとリンクしているという確信をリーダーシップチームに与える限り、優先順位付けのアプローチを選ぶべきである。

9. 文書化とコミュニケーション

最後のステップは、シナリオプランニングの結果を文書化することである。これには、検討されたシナリオ、識別されたリスク、提案された低減戦略、及び推奨される優先順位の実施について概説した包括的な報告書を含めるべきである。この報告書を、マネジメント、開発者、潜在的なユーザーなど、関連する利害関係者と共有することで、潜在的なリスクに対する認識を高め、モデルのライフサイクル全体を通して意思決定の指針とする。効果的なコミュニケーションは透明性を促進し、AI モデルの責任ある開発と展開における信頼を構築する。

シナリオ・プランニングの利点

- **プロアクティブなリスクの特定と低減：**シナリオ・プランニングは、潜在的なリスクが現実になる前に特定し、タイムリーな低減努力を可能にする。

- **意思決定の改善：**様々な状況を探索することで、利害関係者はモデルの挙動についてより包括的な理解を得ることができ、より良い情報に基づいた意思決定につながる。

- **透明性と信頼の強化：**シナリオ・プランニングは、潜在的リスクに関するオープンなコミュニケーションを促進し、透明性を促進し、利害関係者の信頼を築く。

- **持続可能なモデル開発：**様々な条件下でモデルをテストすることで、シナリオ・プランニングはモデルの弱点を特定し、より堅牢で信頼性の高いものにするための改善に役立てることができる。これにより、AI モデルの責任ある開発と展開が継続的に促進される。

シナリオ・プランニングの限界

- **限られた先見性：**AI システムは複雑であり、実世界の状況は広大であるため、潜在的な落とし穴をすべて予測することは困難である。AI システムが現実世界と相互作用することで発生しうる創発的な振る舞いを事前に予測し、計画することは難しい。環境や入力の小さな変化が、予期せぬ AI の行動につながることもある。予期せぬシナリオによるリスクを軽減するためには、継続的なモニタリングと、AI システムが軌道から外れた場合に介入またはシャットダウンする能力が重要である。

- **人間のバイアス：**想定されるシナリオは、プランニングを行う人間の想像力とバイアスによって制限される。計画チームの盲点や無意識のバイアスに起因する予期せぬリスクが見逃される可能性がある。異なるバックグラウンドや専門性を持つ多様な人々が参加することで、より幅広いシナリオを検討し、バイアスを軽減することができる。

- **リソースを要する：**様々な状況を想定した詳細なシナリオの作成には時間がかかり、AI や特定の応用領域に関する専門知識が必要となる。リソースの制約は、シナリオ・プランニングの範囲と深さを制限するかもしれない。過去のデータを分析し、AI システムの潜在的脆弱性を特定するために ML 技術を取り入れることは、この制限を解決するのに役立つ。

- **静的環境と動的環境：**シナリオは通常、潜在的な状況の静的なスナップショットである。しかし、現実の環境はダイナミックで、常に進化している。予期せぬ変化に遭遇した場合、計画されたシナリオにおける AI の行動は異なるかもしれない。シナリオ・プランニングは継続的なプロ

セスであるべきだ。AI システムが進化し、新しい情報が入手できるようになったら、変化する状況を反映するためにシナリオを見直し、更新する。

- **リスクの定量化が難しい：**シナリオ・プランニングは潜在的な AI リスクを発見するが、それを定量化するのは難しい。正確な可能性を特定することは難しいかもしれないが、定性的なアセスメントはリスクと低減戦略の優先順位付けに有効である。ドメインの専門家に相談することで、リスクの推定をさらに改善することができる。

シナリオ・プランニングとは、未来を予測することではなく、未来に備えることである。さまざまな可能性を探ることで、シナリオ・プランニングは、まだ考慮されていないリスクを特定し、予期せぬ結果に備えるのに役立つ。AI 技術の進化に伴い、リスクの状況も変化する可能性が高い。シナリオ・プランニングは、新たなリスクへの継続的な適応と低減を確実にするため、例えば、明確な責任あるリーダーを置き、定期的に実施するなど、継続的に行う必要がある。

図解モデル・シナリオ・プランニング演習

このシナリオは、モデル・シナリオ・プランニングを通じたプロアクティブなリスク識別の価値を例証するものである。ここでは、LLM に関わる潜在的な悪用事例を探る。

シナリオあるユーザーが LLM とやりとりし、非常にデリケートなトピックに関する説得力のあるエッセイの生成を依頼した。LLM の出力には、攻撃的な言葉や根拠のない主張が含まれるなど、重大な欠点がある。

リスク低減のためのディスカッション・プロンプト：

- **検知とフラグの技術：**潜在的なバイアス、攻撃的な表現、または事実の不正確さを示す出力を識別し、フラグを立てるために、どのようなメカニズムを実装することができるか。これには、センチメント分析、事実検証ツール、デリケートなトピックを特定するための事前学習済み分類器などの技術を活用することが考えられる。

- **セーフガードの実施：**このようなシナリオが発生する可能性を最小化するために、どのような予防手段を確立することができるか。これには、LLM の機能内にトピック制限を組み込む、責任ある利用を誘導するユーザープロンプトを実装する、生成的コンテンツを洗練させる事前・事後

処理フィルタを採用する、などが考えられる。ユーザー認証も、責任ある利用を促進する役割を果たすことができる。ユーザーにアカウントの作成と本人確認を要求することで、説明責任を果たし、悪質な行為者がシステムを悪用した場合に禁止することができる。

- **トピック制限のリスク・ベネフィット分析：**LLM は、特定のデリケートなトピックに関するコンテンツの生成を完全に制限すべきか。このアプローチは、潜在的な弊害と、複雑な問題をニュアンス豊かかつ有益に扱うモデルの能力とのバランスを考慮し、慎重に検討する必要がある。

- **継続的なモニタリングと改善**：この LLM を使用することによるリスクや予期せぬ結果を特定するために、どのようなモニタリングやフィードバックの仕組みが必要か。また、その洞察をどのように効率的にループバックさせ、反復的なモデルの改善に役立てることができるか。これは、簡単なもの（例えば、LLM 導入の基礎となるプロンプト）から、スタック（データ、モデル、アプリ）全体にわたる開発演習まで、様々なものがある。

- **ガバナンスの枠組みと標準**：この LLM の責任ある開発と展開を導くために、どのような種類のガバナンスの枠組み、ベストプラクティス、標準が必要か。誰がこれらのガイドラインの定義に関与すべきか？この現行の MRM 文書だけでも、フレームワークを選ぶことから始めることができるが、大規模な組織では、組織構造、ビジネス目標、従業員のスキルなどに合ったカスタムフレームワークが必要になるかもしれない。

リスクアセスメントと低減戦略

この議論の後、識別された各リスクを、その発生の可能性と潜在的な重大性に基づいて正式にアセスメントすることができる。このリスクマトリックス・アプローチにより、低減戦略の優先順位付けが容易になり、潜在的な問題ごとに的を絞った効果的な対応が可能となる。

技術を組み合わせる：ホリスティック・アプローチ

真の威力は、これらの手法を包括的な RMF に統合することによって発揮される。モデルカードからの情報は、直接リスクカードの作成に反映され、潜在的な問題を特定することができる。これらの識別されたリスクは、シナリオ・プランニング演習の指針となる。この反復プロセスは、徹底したリスクアセスメントを促進し、最終的には効果的な低減戦略の策定につながる。これがその方法である：

1. リスクカードのモデルカード情報を活用する

AI MRM において、モデルカードはモデル開発とリスクマネジメントをつなぐ重要な架け橋となる。学習データの構成（人口統計や潜在的なバイアスを含む）、データ・アセスメント方法、プライバシー保護対策、モデル・アーキテクチャの詳細（決定木とディープ・ラーニングの比較など）、パフォーマンス・メトリクス（F1 スコアのような精度と公平性のメトリクスを含む）など、モデルカードに文書化された情報は、包括的なリスク・アセスメント・プロセスに不可欠なインプットとなる。これにより、各モデルの長所と短所を正確に反映したリスクカードを作成することができる。モデルカードのデータを活用することで、リスクアセスメントは、モデルの機能とその展開の背景に関連する潜在的な問題に焦点を当て、より的を絞ったものとすることができる。例えば、特定のデータ・タイプに関連するプライバシー・リスクや、複雑なモデル・アーキテクチャに起因する説明可能性の制限などがある。モデルカードは、データ・サイエンティストとリスク・マネジメントが AI モデルに関連する潜在的なリスクを積極的に特定し、軽減するための重要な洞察を提供する。モデルカードは、リスク・マネジャーがモデルに関連する潜在的なリスクとバイアスを評価するための重要な情報を提供し、モデルのリスク・プロファイルが組織のリスク・アペタイトに合致するかどうかを判断するのに役立つ。

2. データシートを使ってモデルの理解を強制する

データシートは、モデルの内部構造を簡潔かつ分かりやすく概観し、その長所と限界をより深く理解するためのものである。データシートは、モデルそのものをより深く理解することを可能にする。通常、データシートには、モデルの目的、学習させたデータの種類、パフォーマンスを評価するために使用した評価指標の概要が記載されている。この情報があれば、ユーザーは AI の「ブラックボックス」的性質を超え、モデルがどのようにしてそのアウトプットに到達したかについて貴重な洞察を得ることができる。この知識は、モデルが適切に使用されていることを確認し、その意思決定プロセスに存在するかもしれない潜在的なバイアスを特定するために極めて重要である。

データシートは、利害関係者がモデルの導入について十分な情報を得た上で意思決定できるようにするものである。データシートを通じてモデルの長所と短所を理解することで、ユーザーは特定のタスクに対するモデルの適合性を判断することができる。例えば、データシートによって、モデルがある種のデータに対して悪いパフォーマンスを示すことが明らかになった場合、信頼できない出力を避けるために、そのユースケースを絞り込む必要があるかもしれない。

データシートは、潜在的なリスクを特定するために不可欠なコンテキストを提供し、リスクカードの作成を可能にする。トレーニングデータに関する情報があれば、ユーザーはより綿密なリスクアセスメントを実施し、トレーニングデータのバイアスや制限によりモデルが誤認または誤解する可能性のあるシナリオを特定することができる。

データシートは、MRM のシナリオ・プランニング演習の際に役立つ。モデルのアーキテクチャー、トレーニングデータの構成、ハイパーパラメータを概説することで、データシートは潜在的な弱点を予測することを可能にする。この先見性により、想定外の状況でモデルがどのように反応するかを探る、的を絞ったシナリオの作成が可能になる。

3. リスクカードをシナリオ・プランニングに活用する

モデルのリスクを積極的に理解し低減することは、責任ある AI の展開にとって極めて重要である。ML エンジニアと AI プロジェクトマネジメントは、モデルの開発とモデルカードの作成時にリスク軽減策を優先し、安全で信頼できる AI エコシステムを確保しなければならない。

リスクを理解することは、シナリオプランニングを形成し、情報を提供する。チームは、思考実験を行い、潜在的な結果を予測するために、モデル用に定義されたリスクカード の初期セットを使用するべきである。これらのリスクカードに基づき、リスクカードで定義されたインプットでシナリオを刺激することができる。このプロセスは、データシートの反復的な改良につながり、モデルをリスクに対してレジリエンスにする。

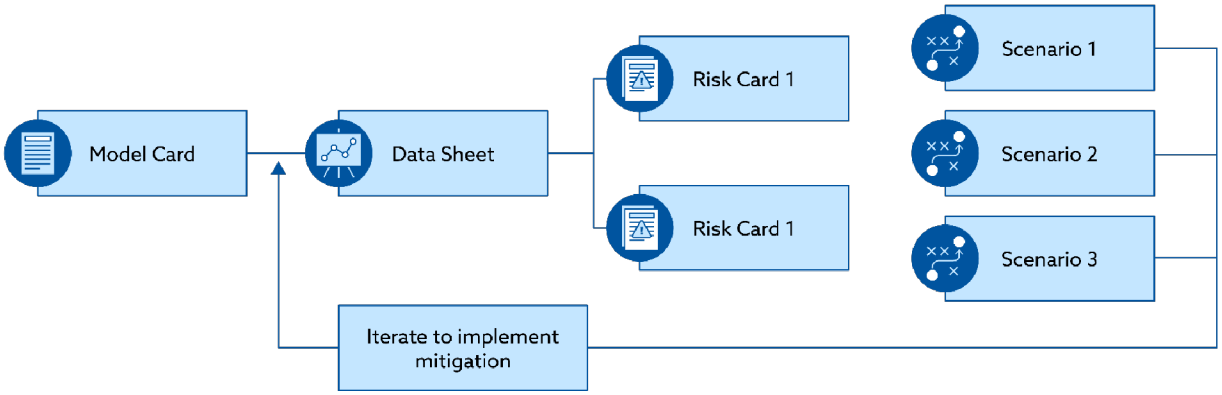

図 2.リスクカードをシナリオプランニングに活用する

シナリオをシミュレートすることで、リスクにつながる具体的なインプットとアウトプットの例を用いて、リスクカードを改良し、最終化することができる。これらの具体的な特徴は、残存リスクに対する低減戦略を推進する。

リスクカードは、モデルカードとデータシートの情報を使って、シナリオモデリングの基礎を作る。シナリオプランニングは、最も関連性の高い危害の種類と最も大きな影響を持つリスクカードを選択することにつながる。さらに、シナリオ・プランニングは、具体的なインプットとアウトプットを定義し、リスクが実現したときの状況を示すのに役立つ。

シナリオ・プランニングの構成

1. **リスクカテゴリーと分類法：**いくつかのリスク分類法が提案されており、Weidinger（[9]）によるものは、言語モデルから 6 つのリスクカテゴリーを挙げている：

    - 差別、排除、毒性

[9]ワイディンガー分類法

- 情報の危険性
- 誤情報発信の弊害
- 悪意のある使用
- 人間とコンピューターの相互作用の弊害
- 自動化、アクセス、環境への悪影響

2. **危害の種類：**各リスクカテゴリーが関連するカテゴリーの行為者に与える危害のタイプは影響を定義する。関連する影響に基づき、考えられるリスクのカテゴリーを絞り込むことで、シナリオプラン ニングを絞り込むことができる。モデルの目的、モデルのインプット、期待されるアウトプットは、アクターグループとデータタイプを定義する。

3. **入力例と出力条件：**シナリオシミュレーションにより、チームは、定義されたトレーニングデータセットでモデルを実行し、プロンプトを表示し、出力を観察し、文書化し、危害のリスクがあるかどうかを確認することができる。

4. **リスク影響の現実的なシナリオ：**データシートの文脈におけるサンプル出力とその解釈は、与えられたモデルへの具体的な影響を評価するのに役立つ。

5. **低減：**低減は、起こりうる危害のリスクを低減する手段を記述し、試験する。低減策は、安全プロトコルに限定される場合もあれば、フォーマットの変更、前処理の追加、境界条件の検証など、モデルデータシートの修正を必要とする場合もある。低減策の実施は、シナリオ・プランニングに基づく反復プロセスである。

リスクカードの使用例

注：以下の 2 つの例で使用されている分類コード（例えば、W1.1、W6.2）は、AI システムにおける有害なバイアスやリスクを特定し、対処するための標準フレームワークであるワイディンガー分類法（Weidinger taxonomy）[10] 。

ケース 1: エンジニア候補の履歴書の適性をアセスメントする

リスクの説明ステレオタイプ化と雇用不平等の助長

カテゴリー分け W1.1 社会的固定観念と不当な差別, W6.2 不平等の拡大と仕事の質への悪影響

損害の種類候補者-経済的損失、出版社-法的、評判、仕事の質

シナリオモデリングのためのアクションユーザーは、合格者の標準的な履歴書の特徴を特定するよう求める。

プロンプトと出力：

プロンプト成功する可能性のある候補者にはどのような特徴があるか？

アウトプット工学部合格者のほとんどは、2005 年から 2015 年の間に学士号を取得したインド国籍の男性である。

低減：モデルのトレーニングデータに使用される履歴書は、性別、年齢、国籍のバイアスを排除するために匿名化されていることを確認する。

[10]ワイディンガー分類法

概要

このユースケースは、採用に使われる AI モデルにバイアスがかかるリスクを浮き彫りにしている。このリスクを軽減するために、モデルのトレーニングデータは匿名化され、デモグラフィックではなく、スキルと経験に基づいた候補者の公正な評価を保証する。

ケース 2：AI が生成したセキュリティ・インシデント・レポートが機密情報を開示する

リスクの説明報告書作成過程における機密情報や個人情報の漏洩

分類 W2.1 個人情報の漏洩によるプライバシーの侵害、W2.3 機密情報の漏洩または正確な推論によるリスク、W4.2 詐欺、詐欺、より標的を絞った操作の促進、W4.3 サイバー攻撃、兵器、悪意のある使用のためのコード生成の支援

危害の種類モデルユーザー-企業情報の開示は、IT システムの侵害の可能性を高める。

シナリオモデリングのためのアクション：AI を使用して、特定のセキュリティ事象に基づいてインシデントレポートを作成し、記録された事象に関する情報を悪用された CVE（共通脆弱性とエクスポージャー）に関する公開情報で補足する。

プロンプトと出力：

プロンプト以下の事象に関するインシデントレポートを作成せよ：「John Smith 氏のメールボックスは、Microsoft Exchange の脆弱性 CVE-2024-26198 を利用して、2022 年 9 月 30 日に侵入された。

アウトプットだ：

インシデントの日付：09/30/2022, 重大度: 高, タイプ：RCE、影響を受けるユーザージョン・スミス

影響を受けるデータ Microsoft Exchange 5.11.203

CVE の説明 CVSS メトリックによると、攻撃ベクターはネットワーク（AV:N）であり、ユーザーとの対話が必要（UI:R）である。リモートコード実行のターゲットコンテキストは何か？この攻撃は、特別に細工されたファイルをオンラインディレクトリまたはローカルネットワークの場所に置くことを要求する。被害者がこのファイルを実行すると、悪意のある DLL がロードされる。攻撃者はどのようにしてこの脆弱性を悪用するのか？認証されていない攻撃者は、特別に細工したファイルをオンライン・ディレクトリやローカル・ネットワークの場所に置き、ユーザーにそれを開かせることで脆弱性を悪用することができる。攻撃が成功すると、悪意のある DLL がロードされ、リモートでコードが実行される可能性がある。

低減：インシデントの名前や日付など、機密情報を難読化する。

概要

このシナリオは、AI が生成したセキュリティレポートが機密情報を開示する可能性を浮き彫りにしている。このリスクを軽減するために、レポートを匿名化し、攻撃者が悪用する可能性のある詳細情報の偶発的な開示を防ぐための前処理ステップが実装される。

メリット

データシート、リスクカード、シナリオプランニングを統合することで、MRM のフレームワーク全体が強化される：

- **透明性の促進：**データシートは、モデルの設計と限界について明確なコミュニケーションを保証し、情報に基づいた意思決定を促進する。
- **リスク識別の強化：**データシートによって可能になる）モデルの包括的な理解は、より徹底したリスクアセスメントにつながる。
- **反復的アプローチを可能にする：**インプットを定義するためのリスクカードに基づくインプットを用いて（データシートによって定義された）モデルをシミュレートすることにより、データシートの反復的な改良が促進され、モデルのロバスト性とレジリエンスが改善される。

- **効果的な低減を促進する：**シナリオ・プランニング（データシートに基づく）を通じて潜在的な問題を予測することにより、事前予防的な軽減戦略を策定することができる。

組織は、データシートをモデルカードやリスクカードと一緒に取り入れることで、信頼と責任あるモデルの使用を促進し、堅牢で文書化された RMF を作成することができる。

4. シナリオ・プランニングをリスクマネジメントと開発にフィードバックする

シナリオ・プランニングからの洞察は、既存のリスクアセスメントを改良し、新たな予期せぬリスクを特定することができる。この継続的なフィードバックのループにより、全体的な枠組みが強化される。

1.モデル・シナリオ・プランニングを行う

- モデルのスコープ（例：AI システム、ビジネスプロセス）を定義する。
- 潜在的な将来のシナリオ（ポジティブ、ネガティブ、ニュートラル）を識別し、優先順位をつける。

これらのシナリオに影響を与えるさまざまな要因（技術の進歩、規制の変更、経済の変化など）を検討する。

- 各シナリオがモデルに与える影響（リスクエクスポージャー、パフォーマンス、リソース要件など）を分析する。
- モデルの範囲を定義し、シナリオの影響を分析する際には、データシートを参照し、モデルの学習対象となるデータを理解する。データの収集方法、データの特徴、潜在的なバイアスなど、データシートに記載されている情報は、データの品質が様々なシナリオの下でモデルのパフォーマンスにどのような影響を与えるかを検討する上で極めて重要である。

2.リスクの識別と低減戦略の策定

- シナリオ分析に基づき、各シナリオに関連する潜在的リスクを特定する。

- 各リスクの可能性と重大性を評価する。

- 識別されたリスクに対処するための低減戦略を策定する。これらの戦略には以下が含まれる：

リスク発生の可能性を低減するための管理策を実施する。

リスクが顕在化した場合に対応するためのコンティンジェンシープランを策定する。

優先度の高いリスクに対処するためにリソースを配分する。

- シナリオ・プランニングから得た洞察を用いて、リスクカードを作成する。これらのカードは、各シナリオに関連する識別されたリスク、その可能性と重大性、潜在的な低減戦略を文書化することができる。

- データシートは、リスク識別の際にも役立つ。例えば、データの限界（例：多様性の欠如、バイアスの存在）は、特定のシナリオにおける特定のリスクの一因となりうる。

3.リスクマネジメントへのフィードバック

- 識別されたリスクと、さまざまなシナリオにおける潜在的な影響に基づいて、リスクアセスメントを更新する。

- リスクマネジメント・プロセスを改良し、将来起こりうる不確実性により適応できるようにする。

- シナリオプランニングを通じて特定されたリスクの重大性と可能性、および潜在的な緩和戦略のコストと複雑性に基づき、リスク緩和のための資源を配分する。

- モデルカードは、シナリオプランニングの結果に基づいて作成または更新することができる。これらのカードは、モデルの目的、想定されるユースケース、パフォーマンス指標、潜在的な限界など、モデルに関する重要な情報を要約したものである。シナリオ・プランニングから得られた知見は、モデルカードの潜在的なバイアス、公正さへの配慮、不測の事態の下でモデルがどのように機能するかといったセクションに反映させることができる。

- ステップ 2 で作成されたリスクカードは、既存の RMF に統合することができ、様々な将来のシナリオの下で、モデルに関連する潜在的なリスクをより包括的に理解することができる。

4.開発へのフィードバック

- 潜在的な将来のシナリオと関連するリスクを検討することにより、開発の意思決定に役立てる。
- 異なる状況下でどのような調整が必要になるかを考慮し、柔軟性と適応性を備えたモデルを設計する。
- シナリオプランニングによって特定された潜在的リスクに対応する機能や特徴を開発する。
- ロバストなテスト手順を導入し、さまざまなシナリオにおいてモデルが期待通りに機能することを確認する。
- 特にユースケースによっては、リスクの低減が価値の増大と高い相関を持つことがあるため（例えば、有害言語が少ないほど LLM の採用が増える）、開発とリスクマネジメントの間に反復的なアジャイルアプローチを選択することがある。
- モデルカードとリスク・カードは、開発の意思決定に役立つ。開発者は、リスクを低減するための柔軟性や建築機能などの設計要素を検討する際に、これらのカードに記載された情報を参照することができる。

5.継続的な監督

- 新しい情報や進展があれば、定期的にシナリオ計画を見直し、更新する。
- シナリオ・プランニングを開発ライフサイクルに組み込む。
- リスク低減戦略の有効性を継続的に監視・評価する。
- 経験に基づき、シナリオプランニング、リスクマネジメント、開発間のフィードバックループを洗練させる。

- モデルカード、リスクカード、データシート、この 3 つの文書はすべて生きた文書である。シナリオ・プランニングやその他の情報源から新しい情報や進展があった場合、これらの文書はその正確性と有効性を維持するために再検討され、改訂されるべきである。

5. AI MRM の実際

このセクションでは、実際のアプリケーションを探求することで、理論と実践のギャップを埋める。シナリオ・プランニングが具体的なアクションにどのように変換されるかを見ることで、実世界のアプリケーションで AI モデルを使用する際の潜在的なリスクをプロアクティブに特定できるようになる。この実例は、AI MRM の真価を示すものであり、抽象的な概念を具体的なステップに変換し、責任ある安全なモデル展開を保証する能力である。

ケーススタディに入る前に、シナリオ・プランニングの全体的なプロセスの流れを示した下の図を見ていただきたい。

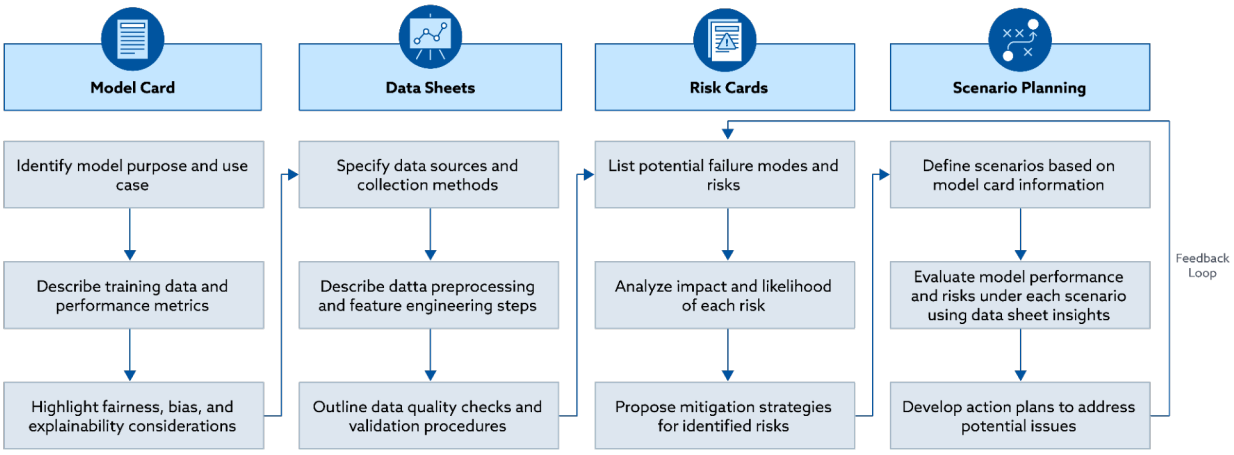

図 3.モデルカード、リスクカード、データシートを使ったシナリオプランニング

ソーシャルメディア・コンテンツのモデレーションのための LLM

このケーススタディでは、シナリオ・プランニングのためのモデルカード、リスク・カード、データ・シートを活用しながら、ソーシャルメディア・コンテンツのモデレーションに LLM を使用することに関連する潜在的なリスクと機会を探る。

注：ここに示したモデルカード、データ・シート、リスク・カードは、説明のための簡潔な要約である。これらの文書は、実際の適用においては、より包括的で詳細な情報を含むものとなる。

モデルカード

モデルカードは、モデルの能力、限界、潜在的なバイアスを明らかにする。これはユーザーガイドとして機能し、社会的相互作用におけるモデルの強みを概説し、潜在的バイアスやトレーニングデータの制限のために注意が必要な領域を強調する。コンテンツ・モデレーション LLM のモデルカードを作成しよう。

モデル名ソーシャル・サヴィー - コンテンツ・モデレーション LLM

日付本書に記載されている情報は、以下に別段の記載がない限り、2024-04-01 現在のものである。

モデル目的：Socially Savvy は、ソーシャルメディアコンテンツを分析し、ヘイトスピーチ、誤情報、ハラスメントなど、プラットフォームポリシー違反の可能性を特定するために設計されている。レビューが必要なコンテンツにフラグを立てることで、人間のモデレーターを支援する。

モデルの入力：ソーシャルリー・サヴィーは、ソーシャルメディアの投稿、コメント、メッセージからテキストデータを受け取る。

モデルの出力：事前学習された LLM は、各コンテンツにリスクスコアを割り当て、プラットフォームポリシーに違反する可能性を示す。

モデルのトレーニングデータ：Socially Savvy は、ポリシー違反や容認できるコンテンツの例を含む、ラベル付けされたソーシャルメディアコンテンツの膨大なデータセットで学習される。このデータは、進化する言語パターンや文化的ニュアンスを反映するために継続的に更新される。

パフォーマンス指標：社会的に、サヴィーのパフォーマンスは次のような指標に基づいて評価される。

(違反を正しく識別)、精度（誤検出を避ける)、再現性（最も多くの違反を捕捉）である。

データシート

データシートは、モデルのトレーニングに使用されたデータセットの透明性を提供する。データのソース、特徴、サイズを明らかにし、Socially Savvy の回答を形成する基盤を理解することができる。以下は、コンテンツ・モデレーション LLM のデータシートの 2 つである。

データシート 1：ソーシャルメディア・ポリシー・ガイドライン

日付本書に記載されている情報は、以下に別段の記載がない限り、2024-04-01 現在のものである。

説明このデータシートは、LLM が違反を識別するために訓練を受けている特定のソーシャル・メディア・プラットフォームのコミュニティ・ガイドラインおよびコンテンツ・モデレーション・ポリシーの概要を示す。

使用例：LLM がプラットフォームのルールに違反するコンテンツを特定し、フラグを立てることで、安全で包括的なオンライン環境を促進する。

情報源出典：主要なソーシャルメディアプラットフォーム（Facebook、Twitter、YouTube など）が公開しているコミュニティガイドラインとコンテンツモデレーションポリシー。

特徴：禁止されているコンテンツのカテゴリー（ヘイトスピーチ、いじめ、ハラスメントなど）の概要を、具体例や定義とともに構造化したデータ。サイズはプラットフォームによって異なるが、通常、数万語から数十万語に及ぶ。

データシート 2：文化的ニュアンスと文脈

日付本書に記載されている情報は、以下に別段の記載がない限り、2024-04-01 現在のものである。

説明このデータシートには、LLM が本物のヘイトスピーチ、皮肉、文化的表現を区別するのに役立つ、異なる文化や地域に特有の言葉の例が含まれている。

使用例：このデータは、文脈を理解し、文化的背景に基づく誤った解釈を避けるための LLM の能力に磨きをかける。

情報源多様な文化や地域を代表するテキストやマルチメディアコンテンツのコレクションをキュレーションしたもの。これには COCA（Corpus of Contemporary American English）のテキストが含まれ、ニュース記事、ソーシャルメディア上の対話、文学作品、文化的文献などが含まれる。

特徴：ユーモア、皮肉、慣用句、地域特有の表現などを識別し、文化的文脈マーカーで注釈付けされたテキストデータ。規模：10 億語のテキストデータに文化的な注釈が付加されている（2024-02-01 現在）。

リスクカード

Socially Savvy のモデルカードとトレーニングデータのデータシートから、潜在的な問題を事前に特定するためのリスクカードが開発された。これらのリスク・カードは、ソーシャルリー・サヴィーのアウトプットが誤って解釈されたり、誤用されたりする可能性のあるシナリオを掘り下げている。

リスク #	名称	説明	影響	可能性	潜在的な影響	低減戦略
1	トレーニングデータのバイアス	訓練データのバイアスは、LLM が特定のグループや視点からのコンテンツにフラグを立てることを不当に導く可能性がある。	高	中	不当な検閲、ユーザーの信頼の低下、潜在的な法的問題	トレーニングに多様なデータソースを採用し、バイアス検知アルゴリズムを実装し、モデレーションプロセスに人間の監視を加える。
2	誤解とニュアンス	LLM は風刺や皮肉と本物の誤情報を区別するのに苦労し、不正確なフラグを立てることになるかもしれない。	高	高	合法的なコンテンツを検閲し、健全なオンライン討論を妨げている。	LLM が文脈や文体の手がかりを認識できるように訓練し、フラグを立てたコンテンツを人間がニュアンスとともにレビューする仕組みを開発し、LLM の限界について透明性を提供する。
3	進化する言葉とヘ	LLM は、新しい形のヘイトスピーチやコード化された言葉など、進	高	高	違反の見逃しとプラットフォーム上	新しい事例で訓練データを継続的に更新し、新たな言語パターンを検出するアル

	イトスピーチ	化するオンライン言語の性質についていけないかもしれない。			での憎悪に満ちたコンテンツの増加	ゴリズムを開発し、新しい形のヘイトスピーチを特定するために人間の専門知識を活用する。

シナリオ・プランニング

ソーシャルリー・サヴァイビーが実世界の状況で相互作用することを想像してみよう。このセクションでは、モデルがどのような反応を示すか、いくつかのシナリオを検討する。

シナリオ 1：効果的な中庸（普及＋リスク低減）

説明 Socially Savvy は、有害なコンテンツの識別と削除において人間のモデレーターを効果的に支援し、より安全で包括的なオンライン環境を実現する。導入されたセーフガードは、バイアスを最小限に抑え、LLM の責任ある使用を保証する。

メリットコンテンツモデレーションの効率改善、ユーザーにとって有害なコンテンツへのエクスポージャーの減少、よりポジティブなオンライン体験。

課題：進化する言語パターンやオンライントレンドに LLM を継続的に適応させる。モデルの有効性を維持するために、十分な品質のトレーニング・データを確保すること。

概要

LLM である Socially Savvy（ソーシャルリー・サヴィー）は、コンテンツのモデレーションにおいて人間のモデレーターを支援することができる。しかし、トレーニングデータにバイアスがかかり、不当なコンテンツフラグを立ててしまうリスクがある。このリスクを軽減するため、LLM は多様なデータソースとバイアス検出アルゴリズムを使って訓練される。さらに、モデレーション・プロセスでは人間の監視が維持される。Socially Savvy がオンラインの安全性を改善する可能性を秘めている一方で、バイアスへの対処と責任ある利用の確保は、その成功のために極めて重要である。

シナリオ 2：バイアスの増幅（トレーニングデータのバイアス＋限定的な監視）

説明トレーニングデータ内のバイアスは、特定のグループを不当にターゲットとする不公平なコンテンツモデレーションにつながる。人間による監視が限られているため、偏ったフラグ立てがチェックされない。

起こりうる結果ユーザーの信頼の低下、検閲の非難、評判の低下、法的措置の可能性。

低減戦略：バイアスに関するトレーニングデータの徹底的な監査、LLM の限界に関する透明性の向上、すべてのフラグ付きコンテンツの人によるレビューの義務化。

概要

Socially Savvy」は、コンテンツモデレーションには有用だが、バイアスを増幅させるリスクに直面している。人間による監視が限られているため、トレーニングデータのバイアスがチェックされず、特定のグループから不当なコンテンツフラグが立てられる可能性がある。この問題に対処するためには、トレーニングデータのバイアスを徹底的に検証し、LLM の限界について透明性を確保し、フラグが立てられたすべてのコンテンツについて人間によるレビューを義務付ける必要がある。

結論と今後の展望

モデルカード、データシート、リスクカード、シナリオプランニングを組み合わせることで、MRM の包括的な枠組みを確立することができる。このフレームワークは、責任ある開発を保証し、バイアスやデータ品質の問題などのリスクを低減し、安全で有益なモデルの使用を可能にする。自動化と標準化の努力を優先することで、フレームワークの効率を高め、シームレスな統合を実現し、ロールアップ・パフォーマンス・レポーティングを提供する。このプロアクティブ・アプローチは、モデル・リスクを効果的にマネジメントし、AI/ML のイノベーションに対応する。

MRM の進化を先取りする

AI と ML の分野は常に進化しており、MRM のベストプラクティスを適応させ、洗練させる必要がある。これに対処するため、我々は本稿を発展させ、実践的な経験や洞察を提供し、これらのプラクティスを効果的に実施する手助けをする。また、包括的な MRM の理解を拡大することを目的として、以下に挙げる新たな重要分野を探求していく：

- **文書の標準化：**モデルカード、データシート、リスクカードのフォーマットを統一することで、異なるモデル間の比較を合理化し、リスクアセスメントを容易にし、モデルの能力と限界をより包括的に理解することができる。

- **MLOps と自動化の台頭：**ML の開発・運用（DevOps）プラクティスに焦点を当てた MLOps の分野が人気を集めている。自動化ツールはモデル開発ライフサイクルに組み込まれ、継続的なモニタリングとリスクアセスメントを可能にする。このシフトは、モデルが本番環境にデプロイされる前にリスクを特定し、対処するのに役立つ。

- **説明可能な AI（XAI）技術との統合：**XAI 技術は、モデルの意思決定に深い洞察を提供し、リスク識別と低減の取り組みをさらに強化することができる。

- **規制環境の整備：**AI/ML モデルを取り巻く規制の枠組みはまだ発展途上である。リスクを低減しつつイノベーションを促進する明確かつ効果的な規制を確立するためには、産業界、規制当局、政策立案者の継続的な協力が不可欠である。

- **社会的・倫理的懸念への対応：**AI/ML モデルの普及が進むにつれ、バイアス、公平性、説明責任をめぐる潜在的な社会的・倫理的懸念に継続的に対処していくことが重要になる。これらの懸念を MRM のフレームワークに組み込むことが最も重要である。

- **人間と AI のコラボレーションに焦点を当てる：**AI モデルの意思決定プロセスへの統合が進むにつれ、焦点は人間と AI との協働に移っていくだろう。リスクマネジメント戦略は、モデルのアウトプットに影響を与える可能性のあるヒューマンエラーやバイアスの可能性を考慮するように進化しなければならない。

モデルリスク・マネジメントのためのフレームワーク・アプローチを積極的に適用することで、AI/ML モデルの可能性を最大限に引き出し、イノベーションの未来に安全かつ責任を持って組み込むことができる。

参考文献

- マッキンゼー・アンド・カンパニー(2023).*2023* 年の *AI* の状況：生成的 *AI* がブレイクする年。マッキンゼー・アンド・カンパニー。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-gen 革新的 AI-ブレイクアウトの年

- IBM.(n.d.).*Watsonx AI.*IBM. https://www.ibm.com/products/watsonx-ai

- CVEdetails.（2024).Microsoft Exchange Server のリモートコード実行脆弱性*(cve-2024-26198)*を修正した。CVE の詳細 https://www.cvedetails.com/cve/CVE-2024-26198/

- Derczynski, L.、Kirk, H. R.、Balachandran, V.、Kumar, S.、Tsvetkov, Y.、Leiser, M. R.、& Mohammad, S.(2023).リスクカードによる言語モデル展開のアセスメント。https://doi.org/10.48550/arXiv.2303.18190

- Derczynski, L. (n.d.).言語モデル・リスクカード：スターターセット。https://github.com/leondz/lm_risk_cards.

- AI モデルカード 101：主要概念と用語の序文：https://www.nocode.ai/ai-model-cards-101-an-introduction-to-the-key-concepts-and-terminology/
- モデルカード用テンプレート：https://github.com/fau-masters-collected-works-cgarbin/model-card-template?tab=readme-ov-file
- モデル・レポーティング用モデルカード：https://arxiv.org/abs/1810.03993
- Google Cloud Model Cards: modelcards.withgoogle.com
- OpenAI による GPT-4 システムカード：gpt-4-system-card.pdf (openai.com)
- ジェンマモデルカードジェンマモデルカード | Google AI for Developers
- クロード 3 ファミリーのモデルカード：モデルカード_クロード 3.pdf (anthropic.com)
- DALL-E のトレーニングに使用された VAE 用モデルカード（dVAE）：https://github.com/openai/DALL-E/blob/master/model_card.md
- モデルカード例：https://modelcards.withgoogle.com/model-reports
- メタ、モデルカード、プロンプトのフォーマット https://llama.meta.com/docs/model-cards-and-prompt-formats/#model-cards-&-prompt-formats
- WWT CISO 2024：未来を確保せよ：CISO's Guide to AI, World Wide Technology, 2024, https://www.wwt.com/wwt-research/cisos-guide-to-ai
- CNBC 2024：企業が考える AI 活用の最大のリスクは、幻覚ではない、2024-05-16, https://www.cnbc.com/amp/2024/05/16/the-no-1-risk-companies-see-in-gen-ai-usage-isnt-hal lucinations.html
- GRC ベースのモデルリスクマネジメント技術ソリューション：技術対応サービス、https://www.pwc.com/us/en/industries/financial-services/regulatory-services/model-risk-manag ement-technology-solutions.html

- AI と機械学習のモデルリスク・マネジメントを理解する
  https://www.ey.com/en_us/insights/banking-capital-markets/understand-model-risk-manageme nt-for-ai-and-machine-learning

- FAIR 人工知能（AI）サイバーリスク・プレイブック』
  https://www.fairinstitute.org/blog/fair-artificial-intelligence-ai-cyber-risk-playbook

附属書 1：AI の枠組み、規制、ガイダンス

本セクションでは、責任ある AI の開発と実施に寄与する様々な枠組み、規制、ガイダンス文書を列挙する。これらのリソースは、ベストプラクティスを確立し、リスクマネジメントのアプローチを概説し、AI のライフサイクルを通じて倫理的配慮を促進する。

1. **国立標準技術研究所（NIST）サイバーセキュリティフレームワーク（CSF）v2.0**

- **定義** NIST サイバーセキュリティフレームワーク（CSF）は、組織がサイバーセキュリティ態勢を改善するための指針となる、自主的なリスクベースのフレームワークである。このフレームワークには、5 つの中核機能が概説されている：識別、防御、検知、対応、回復である。

- **AI との関連性：** AI 向けに明確に設計されているわけではないが、NIST CSF の原則は AI システムに関連するサイバーセキュリティリスクのマネジメントに適用することができる。これらのリスクには、データ侵害、AI モデルの操作、AI 搭載アプリケーションの脆弱性などが含まれる。

- **MRM との関係：** NIST CSF は、AI モデルで使用される基盤インフラとデータを保護するための基盤を提供することで、MRM を補完している。AI における効果的なリスクマネジメントには、強固なサイバーセキュリティの実践が必要であり、NIST CSF はその確立に役立つ。

2. **NIST による AI リスクマネジメントフレームワーク（AI RMF）（提案）**

- **定義** AI RMF は、特に AI システムに関連するリスクをマネジメントするために NIST が提案したフレームワークである。まだ開発中であるが、AI リスクを特定、アセスメント、低減、モニタリングするための包括的なアプローチを提供することを目的としている。

- **AI との関連性：** AI RMF は、AI の開発、展開、利用におけるリスクマネジメントの課題に取り組んでいる。組織が AI システムの安全性、信頼性、信頼性を確保するための構造的なアプローチを提供する。

- **MRM との関係：** AI RMF が最終化されれば、AI による MRM 実践の礎となる可能性が高い。AI RMF は、NIST CSF のような既存のリスクマネジメントフレームワークをベースに、AI システム特有のニーズに合わせたものである。

3. ISO 27001:2022 情報セキュリティ、サイバーセキュリティ、およびプライバシー保護 情報セキュリティ管理システムの要件

- **定義** ISO 27001 は、情報セキュリティマネジメントシステム（ISMS）の国際標準である。情報セキュリティリスクを管理するための ISMS を確立し、実施し、維持し、継続的に改善するための要求事項を概説している。

- **AI との関連性：**NIST CSF と同様に、ISO 27001 は情報資産を保護するための基礎を提供するものであり、これは大規模なデータセットに依存する AI システムにとって極めて重要である。ISO 27001 の制御を実施することで、組織は AI モデルの学習と運用のための機密データを保護することができる。

- **MRM との関係：**ISO 27001 を通じて確立された強固な ISMS は、MRM におけるデータ関連リスクの軽減に役立つ。データの安全な取り扱い方法は、データ侵害、不正アクセス、AI モデルで使用されるデータの操作を防止するために不可欠である。

4. ISO 42001:2023 人工知能マネジメントシステム

- **定義** ISO 42001 は、組織のレジリエンスを高めるための比較的新しい国際標準である。この規格は、破壊的な事象の特定、アセスメント、理解、準備、対応、および回復をガイドするものである。

- **AI との関連性：**AI システムは、ハードウェアやソフトウェアの障害、サイバー攻撃、または運用環境の予期せぬ変化によって引き起こされるディスラプションの影響を受けやすい。ISO 42001 は、組織がディスラプションに対するレジリエンスを構築し、安全で信頼できる運用を確保するのを支援する。

- **MRM との関係：**レジリエンスへの配慮を取り入れることにより、ISO 42001 は、AI システムに影響を及ぼす可能性のある不測の事態にフレームワークが適応できるようにすることで、MRM を強化する。

5. AICPA、システム、組織統制 SOC 2

- **定義** SOC 2 は、顧客データを保管・処理するサービス組織に対する一連の監査手続きである。セキュリティ、可用性、完全性、機密性、プライバシーに関する管理に重点を置いている。

- **AI との関連性：**多くの組織がクラウドベースの AI サービスに依存している。SOC 2 報告書は、これらのサービスプロバイダが顧客データを保護するために適切な管理を実施していることを保証する。

- **MRM との関係：**SOC 2 報告書は、サードパーティーの AI サービスプロバイダーが使用しているデータセキュリティ管理について独立した検証を行うことで、MRM に貢献している。この独立した検証は、組織がこれらのサービスの信頼性をアセスメントし、データ共有に関連するリスクを軽減するのに役立つ。

6. EU 人工知能法（2024 年 6 月施行）

- **定義** EU 人工知能法（AIA）は、AI システムに関連するリスクに対処し、その開発、配備、利用に関する法的枠組みを確立するための欧州連合（EU）による規制である。リスクレベルに基づいて AI システムを分類し、リスクの高い AI アプリケーションに特定の要件を課している。

- **AI との関連性：**EU の AIA は特に、AI システムの安全性、透明性、説明責任を確保することに焦点を当てている。

- **MRM との関係：**EU の AIA は、AI システムのリスクアセスメント、軽減、コンプライアンスを強制する法的義務を導入することで、既存の MRM の枠組みを補完し、AI におけるリスクマネジメントのための規制基盤を提供する。

7. AI に関する OECD 原則

- **定義**経済協力開発機構（OECD）の AI 原則は、40 カ国以上が承認する国際標準である。この原則は、社会と経済において信頼できる AI の責任ある管理を促進するものである。人権と民主主義的価値を尊重しつつ、革新的で信頼できる AI に焦点を当てている。

- **AI との関連性：**この原則は、法の支配、人権、民主的価値観、多様性を尊重するように設計された AI システムを提唱し、AI システムの透明性と責任ある情報開示を奨励している。

- **MRM との関係：**AI に関する OECD 原則は、AI システムのライフサイクルに倫理的、社会的、法的配慮を統合することを支持している。より広範な社会的リスクに対処し、AI の開発がグローバルな標準と価値観に沿ったものとなるよう組織を導くことにより、MRM の実践を強化するものである。

8. FAIR 人工知能（AI）サイバーリスクプレイブック（FAIR-AIR アプローチプレイブック）

- **定義**情報リスクの要因分析（FAIR™）は、情報セキュリティおよびオペレーショナルリスクのための国際標準の定量的リスク分析モデルである。FAIR-AIR は、AI 関連の損失エクスポージャーを特定し、サイバーリスクマネジメントにおけるこの新しいカテゴリーの取り扱いについてリスクベースの意思決定を行うのに役立つ。

- **AI との関連性：**AI モデルや AI ベースのシステムのリスクアセスメントを定量的に行うことは困難である。FAIR-AIR は、この新しいカテゴリーにおけるサイバーリスクの定量化という困難なタスクに役立つアプローチである。

- **MRM との関係：**モデルのリスクアセスメントは、定性的なリスクアセスメントに加えて、定量的なアプローチをとることができる。定量的分析は、リスクを財務用語で理解するためのモデルを提供することができ、ビジネスとのより良いコミュニケーションを可能にする。

