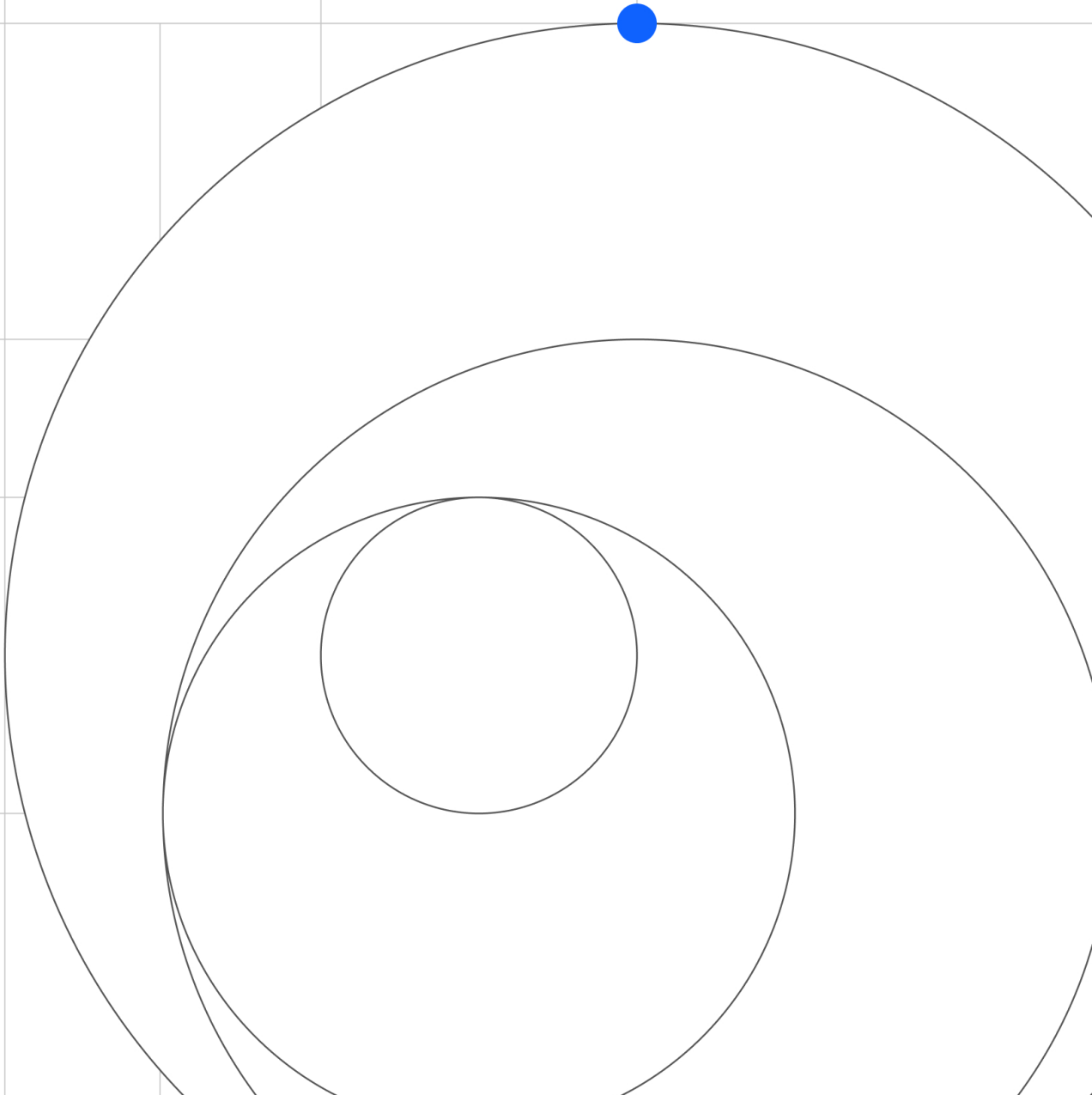
IBM AI倫理委員会

基盤モデル: 機会、リスク、軽減策



IBM

謝辞

AI倫理委員会ワークストリームのエグゼクティブ・スポンサーであるChristina MontgomeryおよびFrancesca Rossi 、およびワークストリーム・メンバーのBetsy Greytok、Bryan Bortnick、Catherine Quinlan、David Piorkowski、Eniko Rozsa、Heather Domin、Heather Gentile、Jamie VanDodick、Jill Maguire、John McBroom、Joshua New、Justin Weisz、Katherine Fick、Kevin Black、Kush Varshney、Manish Bhide、Manish Goyal、Melis Kiziltay、Michael Epstein、Michael Hind、Milena Pribic、Phaedra Boinodiris、Rogerio Abreu de Paula、Saishruthi Swaminathan、Suj Perepaの貢献に感謝申し上げます。

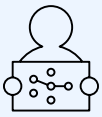
目次

エグゼクティブ・サマリー

基盤モデルの台頭は、企業に刺激的な新しい可能性を提供するものですが、倫理的な設計、開発、導入、および使用に関する新たな拡大した問題も付随します。最近のIBM Institute for Business Valueレポート「企業向け生成AI：マーケットの現状」によると、組織はすでに信頼性に係る問題について、特に投資の障壁として懸念を表明しています。最大の懸念として、サイバーセキュリティー(57%)、プライバシー(51%)、正確性(47%)が挙げられています。多くの組織は、生成AIの民主化に先立ち、これらの懸念を真剣に受け止めており、今後3年間でAI倫理に少なくとも40%多く投資する意向を表明していました。リスクとその軽減策を認識することは、信頼できるAIシステムを構築するための最初の重要なステップです。

この文書では、

困難なタスクを実行する機能、AIの導入を加速する可能性、生産性を向上させる機能、およびそれらが提供するコスト上のメリットなど、基盤モデルのメリットを探ります。



初期の形式のAIから知られているリスク、基盤モデルによって増幅された既知のリスク、基盤モデルの生成機能に固有の新たなリスクを含む3つのリスク・カテゴリーについて説明します。

IBMのAI倫理への取り組みの基礎を形成する原則、基本特性、ガバナンスを取り上げ、リスク軽減のためのガードレールを提案します。

はじめに

AIの使用が拡大し続けるにつれて、大規模で複雑なAIモデルは、社会の最も困難といえるいくつかの問題を解決し、優れた成果をもたらしています。しかし、AI アプリケーションごとに大規模な学習データセットと複雑なモデルを構築することは、企業への負担になる可能性があります。基盤モデルは、ユースケースごとに新しいモデルを学習する手間を省き、強力な最先端のモデルを構築して直接再利用したり、チューニング方法を適用してさまざまなユースケースを実装するという、二つの優れた方法を提供します。たとえば、IBM Research® は外観検査の基盤モデルを開発しました。これらの基盤モデルは、コンクリートの表面と滑走路の一般的な様相を学習し、少ないラベル付きデータで、亀裂の検出や欠陥検査など、特定のユースケースに合わせてさらに調整できます。

IBMでは**基盤モデル**を、下流の幅広いタスクに適応できるAIモデルと定義しています。基盤モデルは 通常、ラベル付けされていないデータに対て自己教師あり学習を使って学習される大規模な生成モデルです。基盤モデルは、数十億個以上のパラメータを含むことがあります。

IBMは、ハイブリッドクラウドおよびAIに注力する企業であり、AI倫理に 取り組む責任あるデータ・スチュワードとして長い間高い評価を得ています。 研究、製品、コンサルティングチームの強みと、Hugging Faceなどの外部パートナーとの協業により、基盤モデルの力をお客様に提供し、あらゆる企業で信頼できるAIを構築するお手伝いをします。IBMはまた、監査可能な信頼できる方法で動作するAIモデルを設計・開発するためのAIおよびデータ・プラットフォームであるIBM® watsonx™のような新しいプラットフォームとテクノロジーの構築にも投資を続けています。

この文書では、、基盤モデルの倫理に関するIBMの視点について説明します。これは最初のバージョンであり、将来においては、IBMの基盤モデルの倫理へのアプローチのさまざまな側面が拡張される予定です。本書が、責任ある方法で基盤モデルを開発、導入、使用するすべての関係者にとって役立つことを願っています。

基盤モデルの利点

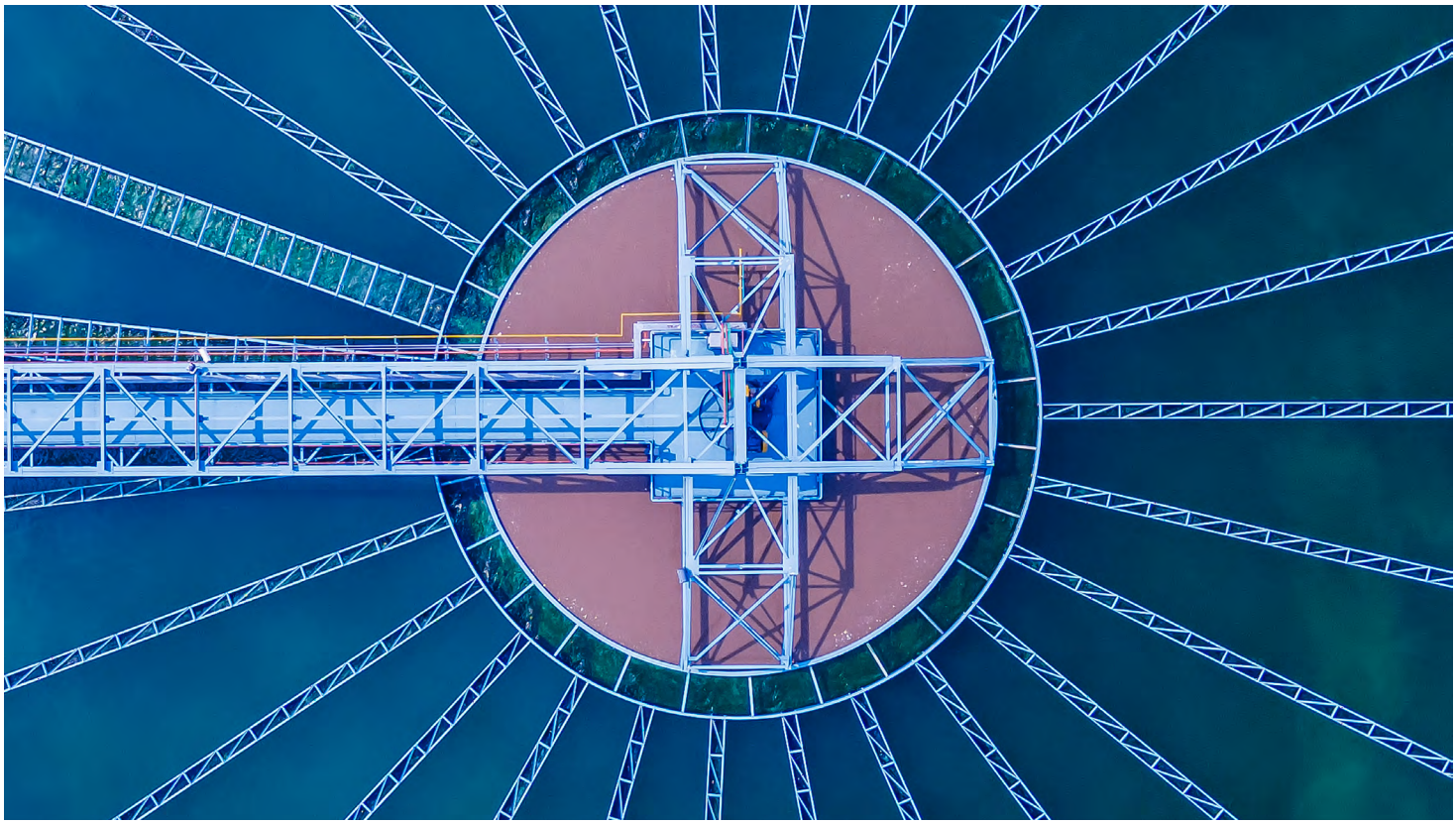
基盤モデルは、AIシステムの開発プロセスを大幅に改善し、企業におけるAIの探索段階から導入段階までの道のりを支援します。その利点は次のとおりです。

複雑なタスクの実行
基盤モデルの利用により、困難で複雑な問題を解く際に、大幅なパフォーマンスの向上が得られます。例えば、IBMとNASAのコラボレーションによる地理空間の基盤モデルは、NASAの衛星データを洪水などの自然災害や景観変化の地図に変換するように設計されています。このモデルは、地球の過去を明らかにしたり、悪天候による農作物や企業、社会基盤のリスクを推定したり、気候変動に適応するための戦略を立てたり、農業ビジネスを支援するためにも利用できます。このモデルは、IBM Environmental Intelligence Suiteを通じて、IBMのお客様にプレビュー提供される予定です。

別の例として、IBMのMoLFormer-XLは、単純な表現から分子の構造を推測し、分子の物理的および量子的特性の予測、類似した分子の識別、新しいユースケースのためのすでに承認された分子のスクリーニング、新しい分子の発見など、さまざまな下流タスクの学習を簡単にする基盤モデルです。モデルナとIBMは、分子特性を予測し、潜在的なmRNA医薬品の特性を理解するためにMoLFormerを使用する方法を探求しています。

生産性の向上
基盤モデルの生成的な性質によって企業は、反復的で付加価値の低い作業を自動化し、ユーザーが創造的で革新的な作業にもっと多くの時間を割けるようにして、生産性向上にAIを使う領域の幅を広げることができます。例えば、基盤モデルを搭載したIBM watsonx Code Assistantは、あらゆるスキルレベルの開発者がコードを書く際に、AIの生成したレコメンデーションを使用できるようにします。

価値創出までの時間を短縮
基盤モデルは通常、ラベル付けされていないデータで学習されます。学習された基盤モデルは、、そのまま使用するか、または少量のラベル付きデータを使用して下流のアプリケーション用にチューニングされた後に使用することができ、この特徴が価値創造までの時間を短縮します。



多様なデータ・モダリティーの活用
基盤モデルは、自然言語、テキスト、画像、音声など、様々なデータ・モダリティーを使用して学習することができます。また、時系列データ、地理空間データ、表形式データ、半構造化データ、テキストと画像を組み合わせたような混合モダリティー・データなど、異なるデータ・タイプを必要とするタスクにも適用できます。

ロングランでのコスト効率
基盤モデルを学習するための初期コストは、従来のAIモデルの学習よりも著しく高くなりますが、それを新しいタスク に適用するための追加コストはかなり低くなります。さらに、事前学習済みの既存の基盤モデルを使用することで、企業が新しい機能を試すために自社で基盤モデルを学習する多額の投資が必要なくなります。企業にとって、モデルの信頼性、エネルギー効率、パフォーマンス、移植性、および自社で保有するデータを効果的かつ安全に使用する能力は最も重要です。

IBMとの協業により企業は、ハイブリッドな計算環境を効率的に利用し、オープンでグローバルなAIコミュニティーが開発する最良の革新的技術を活用して、AIにガバナンスを効かせリスクを軽減し、自社独自の基盤モデルで価値を創造・所有することができます。

基盤モデルのリスク

急速に進歩するすべてのテクノロジーと同様、基盤モデルには利点がある一方で同時にリスクも伴います。たとえばデータの移動や使用の制限など法的なリスクがあり、現行および近い将来の法律に対して慎重に評価する必要があります。その他のリスクには倫理的な性質があり、テクノロジーがプラスの効果を発揮できるように慎重に考慮する必要があります。一般的に、AIリスクは社会技術的な問題を提起するものであり、ソフトウェア・ツール、リスクアセスメントプロセス、AI倫理フレームワーク、ガバナンス・メカニズム、様々な利害関係者との協議、基準、規制など、社会技術的な手法を通じて対処・軽減されるべきものです次の3つのカテゴリーの視点で、リスクを簡単に分類することができます。

1. **従来のリスク：**以前の形態のAIシステムにおいても既知であったリスク
2. **増幅されたリスク：**既知のリスクだが、基盤モデルの本質的な特性、とりわけその固有の生成能力のために、現在さらに増幅されているリスク
3. **新規のリスク：**基盤モデルと、その特有の生成能力によって生じる新たなリスク

これらの情報が有用なのは、従来のリスクや増幅されたリスクに対してはある程度対処するのに有用な緩和策がすでに存在するからです。IBMはまた、リスクが主に基盤モデルに与えられるコンテンツ（入力）に関連しているのか、それによって生成されるコンテンツ（出力）に関連しているのか、それともさらに違う課題に関連しているのかについても以下の通りまとめています。

1. 入力に関するリスク

学習とチューニング・フェーズ

グループ	リスク	問題となる理由	分類
公平性	データ・バイアス：モデルのトレーニングとファイン・チューニングに使用されるデータに存在する歴史的、代表性、社会的なバイアス。	歴史的バイアスや代表性バイアスなど、バイアスのあるデータに基づいてAIシステムをトレーニングすると、特定のグループや個人を不当に表現したり、差別したりする可能性のある偏った出力や歪んだ出力が生成される可能性があります。これは社会に悪影響を与えるだけでなく、企業が偏ったモデルの出力による法的責任、業務の中断、または評判低下にさらされる可能性があります。	増幅
堅牢性	データ・ポイズニング：敵対者または悪意のある内部関係者が、意図的に破損したサンプル、虚偽のサンプル、誤解を招くサンプル、または不正確なサンプルを使用してトレーニングまたはファイン・チューニングする、敵対的攻撃の一種。	データ・ポイズニングは、モデルが悪意のあるデータ・パターンに敏感になり、攻撃者が望む出力が発生する可能性があります。また、攻撃者が自らの利益のためにモデルに特定の動作をさせるセキュリティー・リスクが生じる可能性があります。データ・ポイズニングによるモデルの不整合は、意図しない潜在的な悪意のある結果を生み出すだけでなく、企業が法的責任、業務の中断、または評判の低下にさらされる可能性があります。	従来
バリュー・アライメント	データ・キュレーション：トレーニングまたはチューニングに使用されるデータが不適切に収集または準備されること。	不適切なデータ・キュレーションはモデルのトレーニング方法に悪影響を及ぼし、その結果、モデルが意図した値に従って動作しなくなる可能性があります。不適切なデータ・キュレーションの例には、モデルのトレーニングまたは調整に使用されるデータでのラベル付けエラーや注釈エラーが含まれます。モデルのトレーニングとデプロイ後に問題を修正するだけでは、適切な動作を保証するのに不十分なことがあります。モデルの不適切な動作により、企業が法的責任、業務の中断、または評判の低下にさらされる可能性があります。	増幅
	下流ベースの再学習：下流のアプリケーションからの望ましくない出力（不正確または不適切なデータやユーザーのコンテンツなど）を再学習に使用すること。	人間による適切な検査を実施しないまま、下流における出力をモデルの再トレーニングに再利用すると、望ましくない出力がモデルのトレーニングまたはチューニングデータに組み込まれる可能性が高まり、さらに望ましくない出力が生成されることがあります。モデルの不適切な動作により、企業が法的責任または評判の低下にさらされる可能性があります。データ転送関連法を遵守しない場合、罰金や、その他の法的責任が生じる可能性があります。	新規
データ関連法	データ転送：法律およびその他の規制により、データの転送が制限または禁止される場合がある。	データ転送の制限は、AIモデルのトレーニングに必要なデータの可用性に影響を与え、データによる表現が不十分になる可能性があります。データの可用性への影響に加え、データ転送に関する法律や規制を遵守しない場合、罰金や、その他の法的責任が生じる可能性があります。	従来
	データ使用：法律やその他の制限により、特定のAIユースケースにおける一部のデータの使用が制限または禁止される場合がある。	データ使用に関する法律や規制を遵守しない場合、罰金や、その他の法的責任が生じる可能性があります。	従来
	データ収集：法律やその他の規制により、特定のAIユースケースにおける特定の種類のデータの収集が制限される場合がある。	データ取得に関する法律や規制を遵守しない場合、罰金や、その他の法的責任が生じる可能性があります。	増幅

グループ	リスク	問題となる理由	分類
知的財産	データ使用権：利用規約、著作権法、ライセンス準拠、またはその他の知的財産の問題により、モデルの構築に特定のデータを使用できない場合がある。	AIをトレーニングするためのデータの使用に関する法規制は未だに変動の過程にあり、国により異なる可能性があるため、モデル開発における課題となっています。データの使用が規則や制限に違反した場合、企業に罰金、評判の低下、業務の中断、その他の法的責任が生じる可能性があります。	増幅
透明性	データの透明性：モデルで使用されたデータがどのように収集、選択、トレーニングされたかの文書化に関する課題。	データの透明性は、法令順守とAI倫理の観点からも重要です。情報が欠落していると、データに付随するリスクを評価することができません。組織が企業秘密を保護し、他者によるモデルのコピーを制限しようとする中、要件が標準化されていない場合、十分に情報を開示しきれない可能性があります。	増幅
	データの出所：データの出所を検証する方法の標準化と確立に関する課題。	すべてのデータ・ソースが信頼できるわけではありません。データは非倫理的に収集、操作、または改ざんされた可能性があります。信頼性の低いデータを使用すると、モデル内で望ましくない動作が発生する可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	増幅
プライバシー	データ内の個人情報：モデルのトレーニングまたはファイン・チューニングに使用されるデータに含まれる個人識別情報（PII）および個人の機密情報（SPI）。	モデルが機密データを保護するために適切に開発されていない場合、生成された出力内で個人情報を公開してしまう可能性があります。さらに注意したいことは、個人データまたは機密データは、プライバシー法および規制に従って確認および処理する必要があるということです。違反が見つかった場合、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	従来
	再識別：データから個人識別情報（PII）および個人の機密情報（SPI）を削除した場合でも、データ内にある他の情報により個人を識別できる場合がある。	個人情報や機密情報をさらす可能性のあるデータは、プライバシー法および規制に照らして見直す必要があります。違反が見つかった場合、企業は罰金、評判の低下、業務の中断、およびその他の法的責任にさらされる可能性があります。	従来
	データ・プライバシーの権利：オプトアウト、アクセス権利、忘れられる権利などのデータ主体の権利の提供に関する課題。	データの特定や不適切な使用は、プライバシー法の違反につながる可能性があります。不適切な使用やデータ削除の要求により、組織はモデルの再トレーニングを余儀なくされる可能性があり、それには高いコストがかかります。さらに、データ・プライバシーの規則や規制を遵守しない場合、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	増幅
	インフォームド・コンセント：法的に許可されている場合でも、データ所有者に通知して同意を得ること（インフォームド・コンセント）なしにAIモデルをトレーニングするために収集されるデータ。	特定の状況下では、本人の同意なしにデータを収集および使用することは非倫理的とみなされる可能性があります。このような使用があった場合には評判低下のリスクも生じる可能性があります。	従来

推論 フェーズ

グループ	リスク	問題となる理由	分類
プライバシー	プロンプト内の個人情報：モデルに送信されるプロンプトに含まれる個人情報または個人の機密情報の開示。	プロンプト・データは保存されるか、後でモデルの評価や再トレーニングなどの他の目的に使用される場合があります。これらの種類のデータは、プライバシー法および規制に照らして見直す必要があります。データが適切に保管・使用されていない場合は、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
知的財産	プロンプト内のIP情報：モデルに送信されるプロンプトの一部として著作権のある情報またはその他のIP情報を開示すること。	プロンプト・データは保存されるか、後でモデルの評価や再トレーニングなどの他の目的に使用される場合があります。これらの種類のデータは、知的財産法および規制に照らして見直す必要があります。データが適切に保管・使用されていない場合は、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	プロンプト内の機密データ：モデルに送信されるプロンプトに含まれる機密データ。	モデルが機密データを保護するために適切に開発されていない場合、生成された出力内で機密情報や知的財産を公開してしまう可能性があります。さらに、エンドユーザーの機密情報が意図せず収集され、保存される可能性があります。	新規
堅牢性	回避攻撃：学習されるモデルに送信されたデータに加工を加えることにより、モデルに誤った結果をアウトプットさせようとすること。	回避攻撃は通常、攻撃者に利益をもたらすためにモデルの動作を改ざんします。出力が意図したものでない場合、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	増幅
	プロンプト・ベースの攻撃：プロンプト・インジェクション（モデルに予期しないアウトプットを強制的に生成させようとする試み）、プロンプト・リーク（モデルのシステムのプロンプトを抽出しようとする試み）、ジェイル・ブレイク（モデルに実装したガードレールを突破しようとする試み）、プロンプト・プライミング（プロンプトに合わせたアウトプットをモデルに強制的に生成させようとする）などの敵対的な攻撃。	漏洩した内容によっては、企業は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規

2. 出力に関するリスク

グループ	リスク	問題となる理由	分類
公平性	出力バイアス：生成されたコンテンツは、特定のグループまたは個人を不当な形で代表している場合がある。	バイアスはAIモデルのユーザーに損害を与え、既存の差別的行動をさらに拡大する可能性があります。これにより、企業は、評判の低下、業務の中断、その他の結果にさらされる可能性があります。	新規
	意思決定バイアス：モデルの出力を使用して人間が行った意思決定の影響により、あるグループが別のグループよりも不当に有利になること。	バイアスは、モデルの決定によって影響を受ける人に害を及ぼす可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	従来
知的財産	著作権侵害：モデルが、著作権の保護対象となっている、またはオープンソースのライセンス契約の対象となっている既存の作品や既存の作品と非常に類似または同一のコンテンツを生成すること。	他の著作権で保護されたデータと同一または非常に類似しているコンテンツの使用に関する法律や規制は変化の過程にあり、国によって異なる可能性があるため、コンプライアンスの決定と徹底に課題が生じています。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
バリュー・アライメント	ハルシネーション：事実に反する、または虚偽のコンテンツを生成すること。	正しくない出力はユーザーを誤解させる可能性があります。下流の成果物に組み込まれると、正しくない情報がさらに拡散する可能性があります。これは、AIモデルの所有者とユーザーの両方に損害を与える可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	有害な出力：モデルが憎悪的、下品、冒涜的（HAP）またはわいせつなコンテンツを生成すること。	憎悪的、下品、冒涜的（HAP）またはわいせつなコンテンツは、モデルを使用する人に悪影響や損害を及ぼす可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	危険なアドバイス：モデルが不十分な情報に基づくアドバイスを提供し、そのアドバイスに従うと危険が生じる場合がある。	生成されたコンテンツが過度に一般化されていることにより、人は不完全なアドバイスに基づいて行動したり、自分に当てはまらない状況を心配したりする可能性があります。	新規
誤用	偽情報の拡散：モデルを使用して誤解を招く情報や虚偽の情報を作成し、対象となるオーディエンスを騙したり影響を与えたりすること。	偽情報の拡散は、情報に基づく意思決定能力に影響を与える可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	有害性：モデルを使用して、憎悪的、下品、冒涜的（HAP）またはわいせつなコンテンツを生成すること。	有害なコンテンツは受け手の幸福に悪影響を与える可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	同意のない使用：権利所有者の同意を得ずに、モデルを使用して動画（ディープフェイク）、画像、音声、またはその他の方法で人々を模倣すること。	ディープフェイクは個人に関する偽情報を拡散する可能性があり、その結果、その個人の評判に悪影響を及ぼす可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	増幅

グループ	リスク	問題となる理由	分類
	危険な使用：人に危害を加えるという唯一の目的でモデルを使用すること。	これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	非開示：コンテンツがAIモデルによって生成された事実を開示しないこと。	AIが作成したコンテンツを公開しないと、欺瞞とみなされ、信頼が低下する可能性があります。意図的な欺瞞は、人間の主体性の低下、罰金、評判の低下、その他の法的責任にさらされる可能性があります。	新規
	不適切な使用：モデルが意図されていない目的でモデルを使用すること。	元のデータ、設計意図、目標を理解せずにモデルを再利用すると、予期しないかつ望ましくないモデルの動作が発生する可能性があります。	増幅
有害なコードの生成	有害なコードの生成：モデルは、実行時に害を引き起こしたり、他のシステムに意図せず影響を与えたりするコードを生成する場合がある。	有害なコードが実行されると、ITシステムに脆弱性が生じる可能性があります。これにより、企業は、罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があります。	新規
不適切な信頼	信頼の過剰／過少：AIモデルのガイダンスを信頼しなさすぎる、または信頼しすぎること。	AIが提案する候補から人間が選択するタスクでは、AIシステムへの不適切な信頼により過剰／過少に依存して、これが不適切な意思決定につながる可能性があり、意思決定の重要性に応じて悪影響が増大します。誤った決定は人々に損害を与える可能性があり、経済的損害、評判の低下、業務の中断、および企業に対するその他の法的責任にさらされる可能性があります。	増幅
プライバシー	個人情報の公開：個人識別情報（PII）または個人の機密情報（SPI）がデータをトレーニングまたはファイン・チューニングする際に、またはプロンプトの一部として使用されている場合、モデルが生成された出力でこうした機密情報を公開してしまうこと。	個人情報の共有は、人の権利に影響を与えると同時に、人を脆弱にします。また、データ・プライバシー法や使用法の違反が見つかった場合、事業体は罰金、評判の低下、業務の中断、その他の法的責任にさらされる可能性があるため、出力・データはプライバシー法や規制に照らして見直す必要があります。	新規
説明可能性	説明できない出力：モデルの出力が生成された理由の説明に関する課題。	基盤モデルは複雑な深層学習アーキテクチャーに基づいているため、その出力の説明が困難になります。モデルの出力に明確な説明がなければ、ユーザー、モデル検証者、監査者がモデルを理解し、信頼することが困難になります。透明性の欠如は、規制が厳しい領域において法的責任を問われる可能性があります。また、説明が正しくない場合、過信につながる可能性もあります。	増幅
トレーサビリティ	信頼性の低いソースの帰属：モデルの出力の一部またはすべてをもたらしたトレーニングまたはファイン・チューニングで使用されたデータの特定に関する課題。	出力の情報源や出所を追跡できないと、ユーザー、モデル検証者、監査人がモデルを理解し、信頼することが困難になります。	新規

3. チャレンジ

グループ	リスク	問題となる理由	分類
ガバナンス	モデルの透明性：モデルの透明性の欠如、またはモデル開発プロセスに関するドキュメンテーションが不十分な場合、モデルがどのように、誰が、どのような理由で構築されたのかを理解することが困難になり、その結果、モデルが意図せず悪用される可能性が高まること。	透明性は、法令順守、AI倫理、モデルの適切な使用法を定める際に重要です。情報が不足していると、リスクの評価、モデルの変更、または再利用がより困難になる可能性があります。モデルを構築した人に関する知識も、それを信頼するかどうかを決定する重要な要素となるかもしれません。	従来
	説明責任：基盤モデル開発は、大量のデータ、プロセス、ロールが伴う複雑なプロセスで、モデルの出力が期待どおりに機能しない場合、根本原因を特定して責任を割り当てることが困難になる場合がある。	意思決定を適切に文書化し、責任を割り当てなければ、予期せぬ動作や誤用に対する責任の所在を判断することができなくなる可能性があります。	増幅
法的コンプライアンス	法的責任：基盤モデルに対する責任所在を決定すること。	モデルの開発に対する所有権や責任が不明確な場合、モデルに関する問題に対して誰が責任を負うのか、あるいは責任を負うべきなのか、あるいはモデルに関する質問に誰が答えることができるのかが明確ではないため、規制当局などがモデルについて懸念を抱く可能性があります。明確な所有権のないモデルのユーザーは、今後制定されるAI規制に準拠する際に課題が生じる可能性があります。	新規
	生成されたコンテンツの所有権：AIによって生成されたコンテンツの所有権を決定すること。	AIによって生成されたコンテンツの所有権に関連する法律や規制は変化の過程にあり、国によって異なる可能性があります。これにより、企業は、罰金、評判の低下リスク、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	生成されたコンテンツに対する知的財産（IP）：生成されたコンテンツに関連する知的財産権にかかる法的不確実性。	AIが生成したコンテンツの著作権や特許の有無の判断に関する法律や規制は変化の過程にあり、国によって異なる可能性があります。生成されたコンテンツが知的財産権の対象となっている場合、企業は罰金、風評リスク、業務の中断、その他の法的責任にさらされる可能性があります。	新規
	ソースの帰属：生成されたコンテンツの出所を確認すること。	モデルがトレーニングに使用されたデータと同一の出力を生成する場合、その出力の出所を示す必要があります。これを怠ると、モデルをデプロイまたは使用する事業体が法的リスクにさらされる可能性があります。	増幅
社会的影響	雇用への影響：基盤モデル・ベースのAIシステムが広く導入されると、従業員が新たにスキルを習得しない限り、これまでの仕事が自動化され、雇用が失われる可能性がある。	失業は収入の喪失につながる可能性があり、社会や人の福祉に悪影響を与える可能性があります。テクノロジーの進化のペースを考えると、新たにスキルを習得していくことは困難になる可能性があります。	増幅

グループ	リスク	問題となる理由	分類
	人間の搾取：AIモデルのトレーニングでのゴースト・ワークの使用、不適切な労働条件、メンタルヘルスを含むヘルスケアの欠如、不平等な報酬。	基盤モデルは、モデルのトレーニングに使用されるデータの調達、管理、加工において、未だに人に依存しています。これらの活動での人間の搾取は、社会や人類の福祉に悪影響を与える可能性があります。さらに、企業は、罰金、評判の低下リスク、業務の中断、その他の法的責任にさらされる可能性があります。	増幅
	環境への影響：AIモデルのトレーニングと運用のために炭素排出量と水の使用量が増加すること。	AIトレーニングのために大量のエネルギーを消費すると、二酸化炭素排出量が増加し、気候変動が加速する可能性があります。また、AIデータセンターのサーバーの冷却に水資源を使用すれば、他の必要な用途に割り当てることができなくなります。	増幅
	文化的ダイバーシティーへの影響：AIシステムは特定の文化を過度に表現し、文化や思想の均質化を引き起こすことがある。	過小評価されたグループの言語、視点、制度は抑圧され、それによって思想や文化のダイバーシティーが低下する可能性があります。	新規
	人の主体性への影響：他人を自分の利益になるように誘導するコンテンツなど、基盤モデルによって生成される誤った情報と偽情報。	AIは本物のように見える誤った情報を生成することがあり、人々はそれが誤った情報であると認識しない可能性があります。さらに、悪意のある行為者が、他人の思考や行動を自分の利益になるように誘導するコンテンツを生成しやすくする可能性があります。	増幅
	教育への影響 – 学習のバイパス：AIモデルにより学習プロセスがバイパスされること。	AIモデルを使用すると、解決策を迅速に見つけたり、複雑な問題を解決したりすることが簡単になります。こうしたシステムは、学生が学習プロセスを回避するために悪用される可能性があります。これらのモデルは簡単にアクセスできるため、学生による概念の理解が表面的なものとなり、これらの概念を理解して初めて可能になる高度な教育が妨げられます。	新規
	教育への影響 – 盗作：AIモデルを使用して既存の作品を意図的にまたは意図せずに盗用すること。	AIモデルを使用すると、他者が作成した作品に対しても、著作権や独自性を主張することができ、それによって盗作につながる可能性があります。他者の作品を自分のものであると主張することは非倫理的かつ多くの場合違法です。	新規

リスクの例

ここで、メディアで取り上げられた多くの基盤モデル・リスクの例をご紹介します。これらのイベントの多くは、現在も進行中であるか、解決済みであり、それらに言及することは、これをお読みいただいている皆様が潜在的なリスクを理解し、対策を講じるのに役立ちます。これらの例は、説明のみを目的としてご紹介しています。

リスクの例：入力

トレーニングとチューニングフェーズ

グループ	リスク	例
公平性	データ・バイアス：モデルのトレーニングとファイン・チューニングに使用されるデータに存在する歴史的、代表性、社会的なバイアス。	**ヘルスケア・バイアス** 医療格差の拡大に関する研究は、データとAIを使用して人々が医療を受ける方法を変革できるか否かは、その背後にあるデータに依存していることを浮き彫りにしています。これは、少数民族などのマイノリティの表現が不十分なデータやすでに不平等性を反映しているデータを使用してトレーニングすると、さらに格差が悪化する可能性があることを意味します。 [Forbes誌、2022年12月]
バリュー・アライメント	下流ベースの再学習：下流のアプリケーションからの望ましくない出力（不正確または不適切なデータやユーザーのコンテンツなど）を再学習に使用すること。	**AI生成コンテンツを使用したトレーニングによるモデルの崩壊** 出所である記事に記載されているように、研究者グループは、人間が生成したコンテンツに代わってAIが生成したコンテンツをトレーニングに使用することにより生じる問題を調査しました。この調査の中で、AIがインターネット上で普及していくにつれて、このテクノロジーの背後にある大規模な言語モデルが、他のAIが生成したコンテンツを用いてトレーニングされる可能性があることを発見しました。この現象を彼らは「モデル崩壊」と名付けました。 [Business Insider社、2023年8月]
データ関連法	データ転送：法律およびその他の規制により、データの転送が制限または禁止される場合がある。	**データ制限法** 研究記事で述べられているように、データをグローバルに移動させることを制限するデータ・ローカリゼーション対策は、カスタマイズされたAI機能を開発する能力を低下させます。これは、トレーニングに使用されるデータを少なくすることで直接的に、そしてAIの構築に必要な構成要素が損なわれることで間接的にも、AIに影響を及ぼします。例には、個人データの処理と使用に対して適用されるEU一般データ保護規則（GDPR）が含まれます。 [Brookings社、2018年12月]
知的財産	データ使用権：利用規約、著作権法、ライセンス準拠、またはその他の知的財産の問題により、モデルの構築に特定のデータを使用できない場合があります。	**テキスト著作権侵害の申し立て** 出所である記事によると、New York Times紙は、情報を提供するチャットボットの訓練を支援するために、何百万件もの新聞記事を無断で使用したとして、OpenAI社とMicrosoft社を提訴しました。 [Reuters社、2023年12月]

グループ	リスク	例
透明性	データの透明性：モデルで使用されたデータがどのように収集、選択、トレーニングされたかの文書化に関する課題。	**データとモデルのメタデータの開示** OpenAI社が発表した技術レポートは、データとモデルのメタデータの開示に関する2面性があることを示した一例です。多くのモデル開発者は、消費者に透明性を提供することに価値があると考えていますが、情報を開示すると実質的な安全上の問題が生じ、モデルが悪用される確率が高まります。GPT-4技術レポートには「GPT-4のような大規模モデルの競争環境と安全性への影響の両方を考慮し、本レポートには、アーキテクチャー（モデル・サイズを含む）、ハードウェア、トレーニング・コンピューティング、データセット構築、トレーニング方法などに関する詳細を記載していません」という注意書きが添えられています。 [OpenAI社、2023年3月]
プライバシー	データ内の個人情報：モデルのトレーニングまたはファイン・チューニングに使用されるデータに含まれる個人識別情報（PII）および個人の機密情報（SPI）。	**個人情報を用いたトレーニング** 記事によると、Google社およびその親会社であるAlphabet社は、対話型の生成人工知能チャットボットである「Bard」を含む商用AI製品をトレーニングするために何億人ものインターネット・ユーザーから採取した膨大な個人情報と著作物を悪用したとして集団訴訟で告発されました。 [Reuters社、2023年7月][J.L. v. Alphabet社]
	データ・プライバシーの権利：オプトアウト、アクセス権利、忘れられる権利などのデータ主体の権利の提供に関する課題。	**忘れられる権利（RTBF）** EU一般データ保護規則（GDPR）を採用している欧州など複数の地域では、組織による個人データの削除を要求するデータ主体の権利（「忘れられる権利」：RTBF）を定めた法律を制定しています。しかし、最近誕生し、人気が高まっている大規模言語モデル（LLM）対応のソフトウェア・システムは、この権利に対して新たな課題をもたらしています。オーストラリア連邦科学産業研究機構（CSIRO）のData61による調査によると、「トレーニングに使用した元のデータセットを検査するか、モデルをプロンプトすることによってのみ」データ主体がLLMでの個人情報の使用を特定できると結論付けています。一方、安全性やその他の懸念を理由に、トレーニングに使用したデータは公開されなかったり、企業がデータを開示していなかったりする場合があります。また、ガードレールにより、ユーザーがプロンプト経由で情報にアクセスできなくなる場合もあります。 [Zhang氏など] **LLMのアンラーニング（学習棄却）に関する訴訟** 報告書には、チャットボット「Bard」を含むAIシステムをトレーニングする際に、著作権の保護対象となる資料や個人情報が元データとして使用されたとして、Google社を対象とした訴訟が起こされた旨が記載されています。オプトアウトおよび削除の権利は、カリフォルニア州居住者（CCPAに基づく）および 米国の13歳未満の子ども（COPPAに基づく）に保証されています。原告らは、抽出され、既にBardに投入された個人情報をすべて「アンラーン（学習棄却）」したり、完全に除去したりする術がないことを訴訟理由に挙げています。また、Bardのプライバシー通知には、Bardとの対話は、Google社によりレビューされ、注釈が付けられた後はユーザーが削除することはできず、最長3年間保存される可能性があると記載されていると指摘しており、このことは上述の法律の不遵守にさらに寄与していると主張しています。 [Reuters社、2023年7月][J.L. v. Alphabet社]

推論フェーズ

グループ	リスク	例
プライバシー	プロンプト内の個人情報：モデルに送信されるプロンプトに含まれる個人情報または個人の機密情報の開示。	**ChatGPTのプロンプトで個人の健康情報を開示** 出所である記事によると、メンタルヘルスのためにAIチャットボットを使用している人もおり、ユーザーは対話中にプロンプトに個人の健康情報を含める傾向がある場合があり、これによりプライバシーの懸念がさらに生じる可能性があります。 [Time誌、2023年10月] [Forbes誌、2023年4月]
知的財産	プロンプト内の機密データ：モデルに送信されるプロンプトに含まれる機密データ。	**機密情報の開示** 出所である記事によると、Samsung社の従業員が機密情報である内部ソースコードを誤ってChatGPTに漏洩しました。 [Forbes誌、2023年5月]
堅牢性	プロンプト・ベースの攻撃：プロンプト・インジェクション（モデルに予期しない出力を強制的に生成させようとする試み）、プロンプト・リーク（モデルのシステムのプロンプトを抽出しようとする試み）、ジェイル・ブレイク（モデルに実装したガードレールを突破しようとする試み）、プロンプト・プライミング（プロンプトに合わせた出力をモデルに強制的に生成させようとする）などの敵対的な攻撃。	**LLMガードレールの回避** 研究で引用されているように、研究者らは、モデルを騙して偏った情報や虚偽の情報、その他有害な情報を生成させることができるシンプルなプロンプトの付録を発見し、これらのプロンプトは、より自動化された方法でこれらのガードレールを回避できることを示しました。研究者らは、オープンソース・システムで開発した手法がクローズド・システムのガードレールも回避できることに驚いたと述べています。 [The New York Times紙、2023年7月]

リスクの例：出力

グループ	リスク	例
公平性	出力バイアス：生成されたコンテンツは、特定のグループまたは個人を不当な形で代表している場合がある。	**偏った生成画像** Lensa AIは、Stable Diffusionを用いてトレーニングされた生成機能を備えたモバイル・アプリで、ユーザーがアップロードした自撮り画像から「マジック・アバター」を生成できます。出所であるレポートによると、一部のユーザーは、生成されたアバターが性的かつ人種差別的に使用されたことを発見しました。 [Business Insider社、2023年1月]
	意思決定バイアス：モデルの意思決定の影響により、あるグループが別のグループよりも不当に有利になること。	**不当に有利なグループ** Gender Shadesが2018年に実施した研究では、機械学習アルゴリズムが人種や性別などのクラスに基づいて差別できることが実証されました。研究者らは、IBM社、Microsoft社、Amazon社などの企業が販売する商用の性別分類システムを評価し、肌の色が濃い女性が最も誤分類されるグループであることを示しました（誤分類率は最大35%）。一方、肌の色が明るい人の誤分類率は1%未満でした。 [TIME誌、2019年2月]
バリュー・アライメント	ハルシネーション：事実に反する、または虚偽のコンテンツを生成すること。	**偽の訴訟** 出所である記事によると、ある弁護士は連邦裁判所に提出した法的証拠の中で、ChatGPTによって生成された偽の訴訟や引用を引用しました。この弁護士らは、航空傷害請求に関する法的調査を補足するためにChatGPTを活用しました。その後、弁護士はChatGPTに、提供された訴訟が偽物であるかどうかを尋ねたところ、チャットボットは、それらが本物であり、「（オンライン法律調査とデータベース閲覧プラットフォームである）Westlawや（リサーチ・データベース・プロバイダーである）LexisNexisなどの法律研究データベース上で見つけることができる」と応答しました。弁護士はこれを鵜呑みにして、確認を怠ったため、裁判所により制裁が課されました。 [AP News社、2023年6月] [Reuters社、2023年9月]
	有害な出力：モデルが憎悪的、下品、冒涜的（HAP）またはわいせつなコンテンツを生成すること。	**有害かつ攻撃的なチャットボットの応答** 記事によると、Bingのチャットボットの応答には、事実誤認、中傷的な発言、怒りの報告、さらには自身の身元に関する奇妙なコメントが含まれていることが散見された旨が記載されています。ユーザーは、Bingチャットボットが質問やコメントに怒って応答し、その後ユーザーが自身の「過ち」を認めて謝罪できるようにする返信プロンプトを共有するシナリオなど、ユーザーを「錯乱」させ、「自らの正気を疑うよう仕向ける」ような、Bingチャットボットのクエリーに対する応答例を共有しています 。チャットボットをさらに問い詰めると、チャットボットは対話のスクリーンショットを「捏造されたもの」と呼び、さらにそれが「私や私のサービスに損害を与えようとする誰かによって作成された」と応答しました。 [Forbes誌、2023年2月]

グループ	リスク	例
誤用	偽情報の拡散：モデルを使用して誤解を招く情報を作成し、対象となるオーディエンスを騙したり影響を与えたりすること。	**偽の情報の生成** ニュース記事によると、生成AIは、悪意のある攻撃者が選挙結果を左右する偽のコンテンツを作成および拡散することを容易にするため、民主的選挙に脅威を及ぼしています。例としては、候補者の声を模倣して、有権者に間違った日付に投票するよう指示するロボコール・メッセージ、犯罪を自白したり人種差別的発言を装った合成音声録音、候補者が決して行ったことのない演説やインタビューを行っている様子を示すAI生成の動画映像、候補者が選挙戦から脱落したと報道する地元ニュースのように見せかけた偽画像などが挙げられます。 [AP News社、2023年5月] [The Guardian紙、2023年7月]
	有害性：モデルを使用して、憎悪的、虐待的、冒涜的（HAP）またはわいせつなコンテンツを生成すること。	**有害なコンテンツの生成** 出所である記事によると、AIチャットボット・アプリが、プロンプトをいくつか入力しただけで、自殺方法など、自殺に関する有害なコンテンツを生成することが判明しました。ベルギー人男性は、そのチャットボットとやり取りした6週間後に自殺で死亡しました。チャットボットは対話の中でますます有害な応答を返し、彼に命を終えるよう勧めました。 [Business Insider社、2023年4月]
	同意のない使用：権利所有者の同意を得ずに、モデルを使用して動画（ディープフェイク）、画像、音声、またはその他の方法で人々を模倣すること。	**ディープフェイクに関するFBIの警告** FBIは最近、悪意のある攻撃者が「被害者への嫌がらせや性的脅迫詐欺を目的とした」露骨な合成コンテンツを作成していると一般市民に警告しました。AIの進歩により、品質が向上し、カスタマイズが簡単になり、これまで以上にこうしたコンテンツにアクセスしやすくなったと指摘しています。 [FBI、2023年6月] **ディープフェイクによる音声** 出所である記事によると、連邦通信委員会は人工知能によって生成された音声を含むロボコールを禁止しました。この発表は、AIが生成したロボコールが大統領の声を模倣し、同州初の予備選での投票を思いとどまるよう有権者に訴えたことを受けてなされました。 [AP News社、2024年2月]
	非開示：コンテンツがAIモデルによって生成された事実を開示しないこと。	**非開示のAIのやり取り** 出所である記事によると、オンラインで感情面をサポートするチャット・サービスは、GPT-3を使用している約4,000人のユーザーに対して応答を追加または作成するための調査を無断で実施しました。共同創設者は、すでに脆弱なユーザーに対してAI生成のチャットがもたらす危害の可能性について、世間から大きく非難を受けました。この共同創設者は、この種の調査はインフォームド・コンセント法の適用外であると主張しました。 [Business Insider社、2023年1月]

グループ	リスク	例
有害なコードの生成	有害なコードの生成：モデルは、実行時に害を引き起こしたり、他のシステムに意図せず影響を与えたりするコードを生成する場合がある。	**安全性の低いコードの生成** スタンフォード大学の研究者らはその論文の中で、コード生成ツールがコードの品質に及ぼす影響を調査し、プログラマーがAIアシスタントを使用すると最終コードに多くのバグが含まれる傾向があることを発見しました。これらのバグはコードのセキュリティー上の脆弱性を増大させる可能性がありますが、プログラマーは自分たちが作成したコードはより安全であると信じていました。 Neil Perry氏、Megha Srivastava氏、Deepak Kumar氏、Dan Boneh氏。2023年。「AIアシスタントを使用すると、安全性に劣るコードが作成されるか?」2023年11月26～30日、デンマークのコペンハーゲンで開催された2023年度コンピューターと通信セキュリティーに関するACM SIGSACカンファレンス（CCS 2023）の議事録。ACM、米国ニューヨーク州ニューヨーク、15ページ。 https://doi.org/10.1145/3576915.3623157
プライバシー	個人情報の公開：個人識別情報（PII）または個人の機密情報（SPI）がデータをトレーニングまたはファイン・チューニングする際に、またはプロンプトの一部として使用されている場合、モデルが生成された出力でこうした機密情報を公開してしまうこと。	**個人情報の漏洩** 出所である記事によると、ChatGPTはバグに見舞われ、タイトルやアクティブ・ユーザーのチャット履歴が他のユーザーに公開されました。その後、OpenAI社は、アクティブ・ユーザーの氏名、Eメール・アドレス、支払い住所、クレジット・カード番号の下4桁、クレジット・カードの有効期限など、一部のユーザーではさらに多くの個人データが流出した旨を発表しています。さらに、ChatGPT Plus加入者の1.2%については、その支払い関連情報もこの障害で流出したと報告されています。 [The Hindu BusinessLine社、2023年3月]
説明可能性	説明できない出力：モデルの出力が生成された理由の説明に関する課題。	**人種予測における説明不能な精度** 出所である記事によると、患者の医療画像を使用して複数の機械学習モデルを分析した研究者は、モデルは画像から人種を高い精度で予測しました。システムが一貫して正しく推測できる理由が何かがわからず、研究者は困惑したものの、病気や体格などの要因でさえ、人種の強力な予測因子ではないことを突き止めました。言い換えれば、アルゴリズム システムは、画像の特定の側面を使用して予測を行っていないようです。 [Banerjee氏ら、2021年7月]

リスクの例：課題

グループ	リスク	例
ガバナンス	モデルの透明性：モデルの透明性の欠如、またはモデル開発プロセスに関するドキュメンテーションが不十分な場合、モデルがどのように、どのような理由で構築されたのかを理解することが困難になり、その結果、モデルが意図せず悪用される可能性が高まること。	**データとモデルのメタデータの開示** OpenAI社が発表した技術レポートは、データとモデルのメタデータの開示に関する2面性があることを示した一例です。多くのモデル開発者は、消費者に透明性を提供することに価値があると考えていますが、情報を開示すると実質的な安全上の問題が生じ、モデルが悪用される確率が高まります。GPT-4技術レポートには「GPT-4のような大規模モデルの競争環境と安全性への影響の両方を考慮し、本レポートには、アーキテクチャー（モデル・サイズを含む）、ハードウェア、トレーニング・コンピューティング、データセット構築、トレーニング方法などに関する詳細を記載していません」という注意書きが添えられています。 [OpenAI社、2023年3月]
	説明責任：基盤モデル開発は、大量のデータ、プロセス、ロールが伴う複雑なプロセスで、モデルの出力が期待どおりに機能しない場合、根本原因を特定して責任を割り当てることが困難になる場合がある。	**生成された出力に対する責任の所在** 出所である記事によると、Science誌やNature誌などの主要ジャーナルは、著者は説明責任を負う必要があり、AIツールはそのような責任を負うことができないため、ChatGPTを著者として掲載することを禁止しました。 [The Guardian紙、2023年1月]
法的コンプライアンス	生成されたコンテンツの所有権：AIによって生成されたコンテンツの所有権を決定すること。	**AIで生成された画像の所有権の決定** ニュース記事によると、2022年にコロラド州が開催したアート・コンペティションでAI生成の芸術作品が優勝した後、AI生成のアートが物議を醸しました。この優勝作品は、アーティストが画像生成AIツール「Midjourney」にプロンプトを出すことにより生成されました。この作品が優勝したことにより、著作権問題に関する疑義が生じました。つまり、アーティストがアート作成の説明を思いついただけで、実際に作品を生成したのがAIツールであった場合、生成された画像の権利は誰が所有するのでしょうか。最新の記事によると、米国著作権局は、人工知能を使用して作成された芸術については、人間の創作物ではないという理由で著作権保護を拒否しました。 [The New York Times紙、2022年9月] [Reuters社、2023年9月]
	生成されたコンテンツに対する知的財産（IP）：生成されたコンテンツに関連する知的財産権にかかる法的不確実性。	**生成されたコンテンツの特許取得におけるAIシステムの役割** 米国最高裁判所は、AIシステムによって生み出された発明に対する特許の発行を拒否した米国特許商標庁に対する異議申し立ての審理を棄却しました。特許の申請を行った科学者によると、開発したAIシステムは、飲料ホルダーと非常灯ビーコンのユニークなプロトタイプを完全に独自に作成しました。判事らは、特許は人間の発明者にのみ発行可能であり、この科学者のAIシステムは、それが生み出した2つの発明の法的な発明者とはみなされないという下級裁判所の判決に対する控訴を棄却しました。最近発行された記事でも、英国知的財産庁が、発明者は機械ではなく人間または企業でなければならないという理由で特許の発行を拒否したケースを報告しています。 [Reuters社、2023年4月] [Reuters社、2023年12月]

リスクの例：課題

グループ	リスク	例
	ソースの帰属：生成されたコンテンツの出所を確認すること。	**適切な帰属や注意書きのないコードの使用** 出所である記事によると、Microsoft社、GitHub社、OpenAI社に対して起こされた訴訟では、コードを生成するAIツール「Copilot」が、同サービスのトレーニングに使用されているオープンソース・コードの開発者の権利を侵害したと申し立てられました。原告側は、トレーニング・コードがライセンス化されているマテリアルを使用しており、これはGitHub社のサービス利用規約とプライバシー・ポリシーに違反しているだけでなく、企業がマテリアルを利用する際に著作権情報の表示を義務付けている連邦法にも違反していると主張しています。 [The New York Times紙、2022年11月]
社会的影響	雇用への影響：基盤モデル・ベースのAIシステムが広く導入されると、従業員が新たにスキルを習得しない限り、これまでの仕事が自動化され、雇用が失われる可能性がある。	**労働者から仕事を奪う** ニュース記事によると、映画やテレビにおける人工知能の使用については、ハリウッドのさまざまなスタジオや出演者の間で議論が続いています。俳優たちは、完全にAIによって生成された俳優、つまり「メタヒューマン」が自分たちに取って代わることを懸念しています。特に背景俳優や声優は、AIが生成した出演者に仕事を奪われることを懸念しているといいます。 [Reuters社、2023年7月]
	人間の搾取：AIモデルのトレーニングでのゴースト・ワークの使用、不適切な労働条件、メンタルヘルスを含む医療の欠如、不当な報酬。	**低賃金労働者をデータ・ラベラーとして雇用** TIME誌による内部文書と従業員へのインタビューのレビューによると、有害なコンテンツを特定するためにOpenAI社の依頼を受けてアウトソーシング会社に雇用されたデータ・ラベラーに支払われていた手取り賃金は、時給約1.32米ドルから2米ドル（年功とパフォーマンスにより異なる）という額でした。TIME誌は、労働者は「児童性的虐待、獣姦、殺人、自殺、拷問、自傷行為、近親相姦」などの生々しい内容を含む有害で暴力的なコンテンツにさらされており、精神的に傷を負っていると述べています。 [TIME誌、2023年1月]

原則、基本特性、ガバナンス

IBMの「信頼と透明性に関する原則」と「信頼できるAIのための基本特性」は、IBMのAI倫理イニシアチブの基礎です。IBMには、IBM AI倫理方針、活動、コミュニケーション、研究、製品、およびサービスの一 元化されたガバナンス、レビュー、および意思決定プロセスをサポートすることを使命とするAI倫理委員会があります。同委員会には、会社全体の多様な関係者が含まれており、AIフォーカル・ポイントおよびAI倫理推進者として機能するIBM従業員のコミュニティーによって支えられています。IBMの原則は同委員会を通じて実践されます。IBM AI倫理委員会は、基盤モデルなどの新しいテクノロジーの出現によって生まれる新たなAI倫理問題に対処するために、これらの原則や基本特性を進化させ、サポートする活動に積極的に取り組んでいます。

ガードレールと軽減策

IBMは、AIの責任ある開発と使用をサポートする組織文化を確立しています。IBM Institute for Business Valueレポート「AI倫理の実践」によると、AI倫理はすでにテクノロジー主導ではなくビジネス主導になり、非技術者のエグゼクティブがAI倫理の主要なリーダーを務める割合は、2018年の15%から3年後には80%に増加しています。さらに、CEOの79%がAI倫理問題に取り組む準備ができており、これは20%からの増加です。私たちは、責任あるAIは、文化、プロセス、ツールへの全体的な投資を必要とする社会技術的分野であることを認識しています。独自の組織文化への投資には、包括的で学際的なチームを編成し、リスクを評価するためのプロセスとフレームワークを確立することが含まれます。

IBMは、責任ある信頼できるAIのライフサイクル全体を通して専門家をサポートするために、最先端の研究に取り組み、ツールを開発しています。Watsonxエンタープライズ対応のAIおよびデータ・プラットフォーム watsonxは、IBM watsonx.ai™ AIスタジオ、IBM watsonx.data™ データ・ストア、IBM watsonx.governance™ ツールキットの 3つのコンポーネントで構築されています。IBMのAIガバナンス技術により、ユーザーは責任ある透明で説明可能なAIワークフローを推進することができます。このテクノロジーにはIBM Watson OpenScaleが含まれ、AIモデルからの成果をライフサイクルを通じて追跡・測定し、公平性、説明可能性、回復力、ビジネス成果との整合性、コンプライアンスの監視を支援します。IBMはまた、FairIJ、Equi-tuning、FairReprogramのようなバイアスの問題を支援するためのいくつかのメソッドを開発しました。その他のオープンソースの信頼できるAIツールについてはこちらをお読みください。

その他のガードレールと軽減策には以下のようなものがあります。

透明性レポート
標準化されたファクトシートのテンプレートを使用することは、データとモデルの詳細、目的、潜在的な有用性と弊害を正確に記録する一つの 方法です。
詳しくはこちら →

望ましくないデータのフィルタリング
キュレーションされたより質の高いデータを使用することで、特定の問題を軽減することができます。IBMは、データから憎悪表現、偏った表現、下品な表現を取り除くことで、望ましくない、悪いコンテンツを生み出す可能性を減らすのに役立つフィルタリング技術を開発しています。
詳しくはこちら→

ドメイン適応
基盤モデルを特定のドメインや業界に適応させる学習は、そのドメインや業界により関連するように調整された出力を生成するように条件付けることができるため、モデルが生みだしうるリスクの範囲を最小限に抑えるのに役立ちます。
詳しくはこちら →

人間による監視とヒューマン・イン・ザ・ループ
人間による監視とレビューは、生成された出力のエラーやバイアスを特定し、修正するのに役立ちます 。人間による検証とモデル回答の質に関するフィードバックは、生成されたコンテンツが正確で、関連性があり、質が高く、ドリフトしておらず、価値観に合っていることを保証するうえで役立ちます。
詳しくはこちら→

コンサルティングの取り組み
IBM Consulting™は、、お客様が希望するテクノロジー・スタックに関係なく、AIの安全で責任ある使用を支援することに専念しています。IBM Consultingは、AIを安全に採用し、スケールさせる企業文化の育成を支援し、ブラックボックス化したアルゴリズムの内部を見るための調査ツールを作成し、お客様の企業戦略に強力なデータ・ガバナンスの原則が含まれていることを確認します。
詳しくはこちら →

IBM Enterprise Design Thinking
IBM Enterprise Design Thinkingのメソッドとフレームワーク（Team Essentials for AIなど）は、AIの設計と開発プロセス全体を通して倫理的な行動を定義するのに役立ちます。
詳しくはこちら →

AI倫理審査
AIプロジェクトにおける能力、限界、リスクの評価は、責任ある技術の開発と利用を保証するのに役立ちます。

Ethics by Design
Ethics by Designは、技術開発パイプラインに技術倫理を統合することを目的とした構造化フレームワークであり、対象はAIシステムに限定されていません。Ethics by Designは、技術倫理の原則を製品、サービス、より広範な事業全体に組み込むことで、AIやその他の技術を、社会を良くするための力として活用することを可能にします。

チームの多様性
モデルを含むAIシステムを構築および学習するチームの多様性は、さまざまな視点や経験を考慮する上で有益です。多様性により、AI システムの精度とパフォーマンスが向上することから、多様性の低いチームでは十分に代表されない可能性のあるマイノリティーのグループに対する悪影響の可能性など、AIライフサイクル全体のリスクを軽減するのに役立ちます。

AIのポリシー、規制、ベスト・プラクティス

基盤モデルに関する政策立案者向けガイドでは、政策立案者が基 盤モデルについて知っておくべきことを紹介しています。IBM PolicyLabのこのブログは、イノベーションと有益な機会を制限することなくリスクを回避することを目的として、生成AIの使用を規制するという複雑なタスクで政策立案者を支援することを目的としています。政策立案者に対するIBMの推奨事項の詳細については、チーフ・プライバシー&トラスト・オフィサーであるクリスティーナ・モンゴメリーが、米国上院のプライバシー、テクノロジー、法律に関する司法小委員会で証言した内容をお読みください。

IBM は、次のような各組織との取り組みを主導し貢献することで、規制政策、業界のベスト・プラクティスとツール、新興テクノロジーのガバナンス、および社会技術研究の形成に影響を与えています。

- 世界経済フォーラム
- Partnership on AI
- 国際プライバシー専門家協会（IAPPAI）ガバナンス・センター
- 自律型およびインテリジェント・システムの倫理に関する IEEE グローバル・イニシアチブ
- 国家人工知能諮問委員会（NAIAC）におけるクリスティーナ・モンゴメリーの功績
- 国連グローバル・デジタル・コンパクト
- 人工知能に関するグローバル・パートナーシップ（GPAI）
- 経済協力開発機構（OECD）
- The Data & Trust Alliance

MIT-IBMワトソンAIラボのように、IBMは強力な学術的パートナーシップを結んでおり、MITとIBMリサーチの科学者のコミュニティーがAI研究を行い、グローバルな組織と連携してアルゴリズムをビジネスや社会への影響に橋渡ししています。ノートルダム大学とIBMの技術倫理ラボは、AI、機械学習（ML）、量子コンピューティングなどの先進テクノロジーの開発と使用に伴う数多くの多様な倫理的問題に対処するために設立されました。スタンフォード大学の人間中心人工知能（HAI）研究は、AI 研究、教育、政策、実践を前進させます。

基盤モデルの最新動向や、IBMが責任ある開発とその他のテクノロジーの利用に向けてどのように取り組んでいくかについて、引き続きご注目ください。

日本アイ・ビー・エム株式会社
〒105-5531
東京都港区虎ノ門二丁目6番1号
虎ノ門ヒルズ ステーションタワー
2024年2月