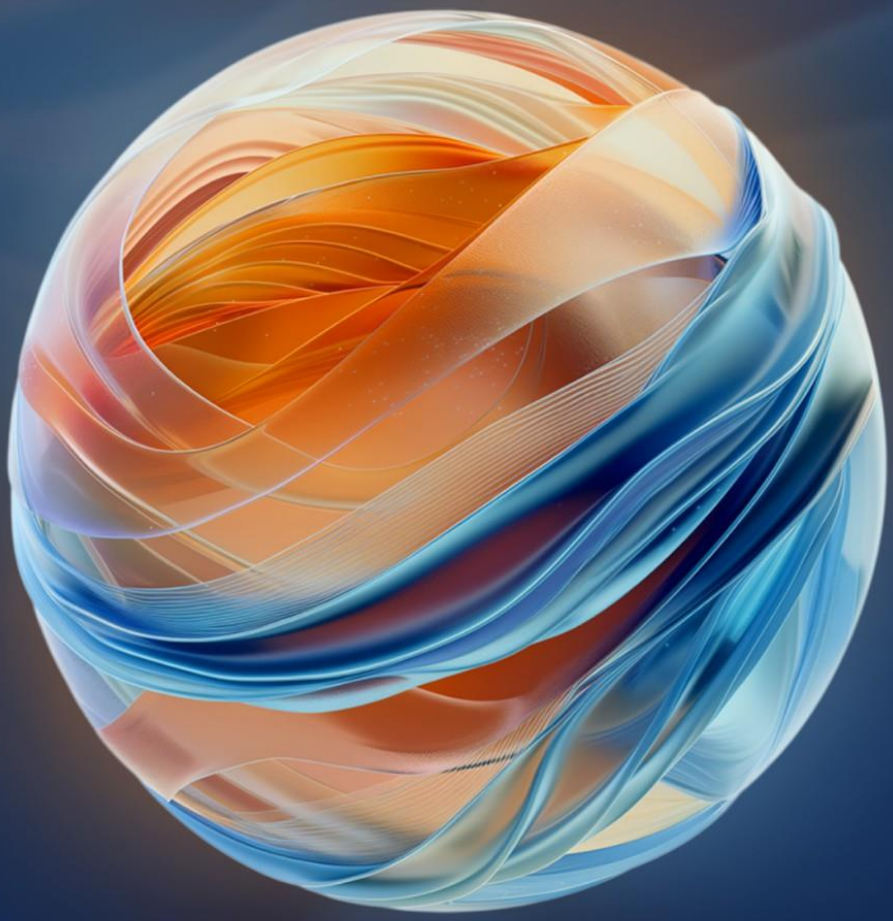
原則から実践へ：
ダイナミックな規制環境における責任あるAI



AI Governance and Compliance Working Group

CSAJC cloud security alliance SM Japan Chapter

CSA cloud security alliance®

The permanent and official location for the AI Governance and Compliance Working Group is https://cloudsecurityalliance.org/research/working-groups/ai-governance-compliance

日本語版提供に際しての告知及び注意事項

本書「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」は、Cloud Security Alliance（CSA)が公開している「Principles to Practice: Responsible AI in a Dynamic Regulatory Environment」の日本語訳です。本書は、CSAジャパンが、CSAの許可を得て翻訳し、公開するものです。原文と日本語版の内容に相違があった場合には、原文が優先されます。
翻訳に際しては、原文の意味および意図するところを、極力正確に日本語で表すことを心がけていますが、翻訳の正確性および原文への忠実性について、CSAジャパンは何らの保証をするものではありません。
この翻訳版は予告なく変更される場合があります。以下の変更履歴（日付、バージョン、変更内容）をご確認ください。

変更履歴

日付	バージョン	変更内容
2024年11月06日	日本語版1.0	初版発行

CSAジャパン成果物の提供に際しての制限事項

日本クラウドセキュリティアライアンス（CSAジャパン）は、本書の提供に際し、以下のことをお断りし、またお願いします。以下の内容に同意いただけない場合、本書の閲覧および利用をお断りします。

１．　責任の限定
CSAジャパンおよび本書の執筆・作成・講義その他による提供に関わった主体は、本書に関して、以下のことに対する責任を負いません。また、以下のことに起因するいかなる直接・間接の損害に対しても、一切の対応、是正、支払、賠償の責めを負いません。
(1)　本書の内容の真正性、正確性、無誤謬性
(2)　本書の内容が第三者の権利に抵触しもしくは権利を侵害していないこと
(3)　本書の内容に基づいて行われた判断や行為がもたらす結果
(4)　本書で引用、参照、紹介された第三者の文献等の適切性、真正性、正確性、無誤謬性および他者権利の侵害の可能性

２．　二次譲渡の制限
本書は、利用者がもっぱら自らの用のために利用するものとし、第三者へのいかなる方法による提供も、行わないものとします。他者との共有が可能な場所に本書やそのコピーを置くこと、利用者以外のものに送付・送信・提供を行うことは禁止されます。また本書を、営利・非営利を問わず、事業活動の材料または資料として、そのまま直接利用することはお断りします。
ただし、以下の場合は本項の例外とします。
(1)　本書の一部を、著作物の利用における「引用」の形で引用すること。この場合、出典を明

記してください。
(2)　本書を、企業、団体その他の組織が利用する場合は、その利用に必要な範囲内で、自組織内に限定して利用すること。
(3)　CSAジャパンの書面による許可を得て、事業活動に使用すること。この許可は、文書単位で得るものとします。
(4)　転載、再掲、複製の作成と配布等について、CSAジャパンの書面による許可・承認を得た場合。この許可・承認は、原則として文書単位で得るものとします。

３．　本書の適切な管理
(1)　本書を入手した者は、それを適切に管理し、第三者による不正アクセス、不正利用から保護するために必要かつ適切な措置を講じるものとします。
(2)　本書を入手し利用する企業、団体その他の組織は、本書の管理責任者を定め、この確認事項を順守させるものとします。また、当該責任者は、本書の電子ファイルを適切に管理し、その複製の散逸を防ぎ、指定された利用条件を遵守する（組織内の利用者に順守させることを含む）ようにしなければなりません。
(3)　本書をダウンロードした者は、CSAジャパンからの文書（電子メールを含む）による要求があった場合には、そのダウンロードしまたは複製した本書のファイルのすべてを消去し、削除し、再生や復元ができない状態にするものとします。この要求は理由によりまたは理由なく行われることがあり、この要求を受けた者は、それを拒否できないものとします。
(4)　本書を印刷した者は、CSAジャパンからの文書（電子メールを含む）による要求があった場合には、その印刷物のすべてについて、シュレッダーその他の方法により、再利用不可能な形で処分するものとします。

４．　原典がある場合の制限事項等
本書がCloud Security Alliance, Inc.の著作物等の翻訳である場合には、原典に明記された制限事項、免責事項は、英語その他の言語で表記されている場合も含め、すべてここに記載の制限事項に優先して適用されます。

５．　その他
その他、本書の利用等について本書の他の場所に記載された条件、制限事項および免責事項は、すべてここに記載の制限事項と並行して順守されるべきものとします。本書およびこの制限事項に記載のないことで、本書の利用に関して疑義が生じた場合は、CSAジャパンと利用者は誠意をもって話し合いの上、解決を図るものとします。

その他本件に関するお問合せは、info@cloudsecurityalliance.jp までお願いします。

日本語版作成に際しての謝辞

「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」は、CSAジャパン会員の有志により行われました。作業は全て、個人の無償の貢献としての私的労力提供により行われました。なお、企業会員からの参加者の貢献には、会員企業としての貢献も与っていることを付記いたします。
以下に、翻訳に参加された方々の氏名を記します。（氏名あいうえお順・敬称略）

石井　英男
笠松　隆幸
仲上　竜太
三井　陽一　CISSP, CCSP, CISA, CISM, CDPSE
諸角　昌宏

Acknowledgments

Lead Authors

Maria Schwenger
Louis Pinault

Contributors

Arpitha Kaushik
Bhuvaneswari Selvadurai
Joseph Martella

Reviewers

Alan Curran MSc
Udith Wickramasuriya
Piradeepan
Nagarajan Rakesh
Sharma Gaetano
Bisaz Hongtao Hao
Jan Gerst
Ashish Vashishtha
Gaurav Singh
Ken Huang
Frederick Hänig
Dirce Hernandez
Tolgay Kizilelma, PhD
Saurav Bhattacharya
Michael Roza
Gabriel Nwajiaku
Vani Mittal
Meghana Parwate
Desmond Foo
Lars Ruddigkeit
Madhavi Najana

CSA Global Staff

Ryan Gifford
Stephen Lumpe

目次

内容

本書は情報提供のみを目的としており、法的助言ではないことに注意。

クラウドセキュリティアライアンス（CSA）のために作成されたこの調査資料では、人工知能（AI）を取り巻く規制ガバナンスの現状を探っている。本書は様々な法律や規制の枠組みを扱っているが、提示された情報が*特定の状況に適用される法的指針として解釈されるべきではない*ことを強調する。

AIに関する規制の状況は急速に進化しており、法律や規制の解釈や適用は、以下のようなさまざまな要因によって大きく異なる可能性がある：

- 管轄（国または地域）
- 特定のコンテキスト（業界、ユースケースなど）
- 特定のAI技術またはアプリケーション

従って、クラウドセキュリティアライアンスと本文書の執筆者は、AIの開発、展開、または利用に伴う法的な意味合いに関する疑問や懸念については、外部の専門的な*弁護士に相談することを強く推奨する。*

セーフハーバー声明

先見的な声明と人工知能の進化する状況

本書には、本質的に先見的と考えられる声明が含まれている。これらの適用可能性を判断するために、該当する国の規制機関や法律顧問に指導を求めることを奨励する。著者とCloud Security Alliance（CSA）は、現在の知見と予想に基づいてこれらの声明を発表する。将来の見通しに関する声明には、固有のリスク、不確実性、および仮定が含まれるため、実際の結果がそのような声明によって予測または暗示されるものと大きく異なる可能性があることに留意することが重要である。

以下は、人工知能（AI）分野の将来の発展および関連する規制状況に影響を及ぼしうる重要な要因の一部であり、その結果、本書における将来に関する記述の正確性に影響を及ぼす可能性がある。

- 急速な技術の進歩：AIの分野は常に進化しており、新しい技術やアプリケーションが急速に登場している。これらの進歩の正確な軌跡や、AI規制のさまざまな側面への影響を予測することは困難である。
- 規制の枠組みにおける不確実性：AIに対する規制のアプローチはまだ発展途上であり、AIの開発、展開、および使用に関する具体的な規制は、管轄区域によって大きく異なる可能性があり、時間の経過とともに変化する可能性がある。

- 新たな倫理的配慮：AIの応用がより高度になるにつれて、新たな倫理的考慮事項が生じ、これらの技術の責任ある開発と利用をめぐる新たな規制やガイドラインにつながる可能性がある。
- 経済的・社会的要因：AIに対する全体的な経済情勢や社会的態度は、新技術の開発や採用、またそれらを取り巻く規制の状況に影響を与える可能性がある。

著者およびCSAは、将来の出来事や状況を反映させるために、本書中の将来に焦点を当てた声明を更新または修正する責任を一切負わないものとする。読者は、これらの記述を過度に信頼しないよう注意する必要がある。これらの声明は、本書の発行日時点における著者およびCSAの見解のみを反映したものである。

本書の概要

本書では、AIと生成AI（GenAI）を取り巻く法規制の状況を概観する。生成AIの用途の多様性、世界の規制当局による規制アプローチの違い、既存規制の適応の遅れなどから、この複雑でダイナミックな状況を乗り切ることの難しさを浮き彫りにしている。

本書の目的は、組織が自らの現状を根本的に理解し、責任あるコンプライアンスに則ったAI活用のために急速に変化する要件を乗り切るために必要な一般的知識を身につけることである。本書では、既存の規制の一部を紹介し、地域、国、および国際的なレベルで責任あるAIを開発・展開するための考慮事項とベストプラクティスを示している。

本書は、本稿執筆時点における、生成AI（GenAI）を含むAIの法的・規制的状況についての概要を提供するものである。網羅的なものではなく、組織が現在の立場を理解し、責任あるコンプライアンスに準拠した生成AI利用の進化する要件をナビゲートするための主要な考慮事項を特定するための出発点である。

現在進行中の技術の進歩や、進化する法的・政策的状況のため、完全な概要を提供することは困難である。したがって、進化するAIの規制や権限について常に情報を得るための基礎として、この情報を活用することをお勧めする。AIの規制は、世界中のさまざまなレベルの政府や管轄区域から出されていることを考慮することが重要である。さらに、データプライバシーや差別禁止規制などの法律により、AIが特にその目的のために特別に設計されたものではなくても、AIをどこでどのように使用できるかを決定する。たとえば米国では、AIは市、州、連邦の法律、政府機関の行動、行政命令、業界の自主的な協定、さらにはコモンローによって管理されることになる。AI規制の成り立ちは必ずしも直感的なものではないため、AIプロジェクトの準備として綿密な分析を行う必要がある。その最初の法的枠組みが欧州AI規制法（European AI Act）であり、これは人々と企業の安全と基本的権利を保証するものである。ある種のAIアプリケーションは、それが市民の権利を妨げたり脅かしたりする場合には禁止される。大規模言語モデル（LLM）のようなリスクの高いAIシステムについては、健康、安全、基本的権利、環境、民主主義、および法の支配に重大な害を及ぼす可能性があるため、規制が予想される。

エグゼクティブサマリー

人工知能（AI）は私たちの世界を急速に変貌させつつあり、社会の構造そのものを再構築する計り知れない可能性を秘めている。しかし、この変革の力には重大な課題が伴う。現在の法規制の状況は、AI、特に生成AI（GenAI）の爆発的な成長に追いつくために苦労している。本書の目的は、既存の法律や規制の概要と、それらがAIの開発、展開、利用に与える影響を提供することである。私たちの目標は、責任あるAIを導入するための実践的なアプローチを模索しながら、法整備が遅れている分野を特定することである。現状は確立された法律がないため、高度化するAI機能に関連する潜在的リスクへの対応にギャップがある。このため、GDPRやCCPA/CPRAのような既存の規制は、データプライバシーの基礎を提供するものの、AI開発特有の課題に対する具体的な指針を提供せず、例外が少なすぎて十分ではないという状況が生まれている。大手ハイテク企業がAIに数千億ドルの投資を計画しているように、技術革新の勢いは衰えることはないだろう。

ある厄介なギャップが生まれつつある。個人用と業務用の両方で生成AI が広く使用されている一方で、適切なガバナンスが欠如している。悪意あるアクターはすでに生成AIを駆使して高度な攻撃を仕掛けており、企業は生成AIを競争上の優位性と見なし、その採用をさらに加速させている。このような急速な導入はエキサイティングであるが、イノベーションを阻害しないような責任あるAI開発の実践を伴う必要がある。理想的な解決策は、明確で実用的なガイドラインに支えられた、責任ある、透明性のある、説明可能なAIの使用を奨励するグローバルな環境を育成することである。AIの無限の可能性と責任ある開発の必要性とのギャップを埋めるためには、3つの側面から協力するアプローチが必要である。すなわち、すべてのテクノロジー企業が責任あるAIに取り組むこと、政策立案者が明確なガイドラインを示すこと、そして立法府が効果的な規制を設けることである。

本書は、法規制に焦点を当てたAIガバナンスに関する重要な対話を開始し、AIガバナンスの現状とその欠点に関する基本的な理解を、AIに取り組む実務家や企業に提供するものである。こうしたギャップを浮き彫りにすることで、責任あるAIの開発と導入に必要な法的枠組みについて開かれた議論を促進することを目指す。

はじめに

急速に拡大するAIの分野では、個人と社会を保護しながら、責任ある開発、展開、およびイノベーションを確保するために、進化する法的・規制的状況をナビゲートする必要がある。

AIの倫理的・法的枠組みを理解することで、組織は3つの重要な目的を達成することができる。

- **信頼とブランド評価の構築**：組織は、透明で責任あるAIの実践を示すことで、利害関係者との信頼を築き、ブランドの評判を高めることができる。
- **リスクの軽減**：フレームワークに積極的に関与し、リスクベースのアプローチを活用することで、無責任なAIの使用に関連する潜在的な法的リスク、評判リスク、および財務リスクを軽減し、組織と個人の両方を保護することができる。

- **責任あるイノベーションの育成**：ベストプラクティスを遵守し、透明性、説明責任を維持し、強力なガバナンス構造を確立することで、組織は責任ある安全なAIイノベーションの文化を育むことができ、その発展とともに社会へのポジティブな影響を確保することができる。責任あるAIは、多様なチーム、包括的な文書化、人間による監視を通じて、バイアスを軽減し、課題を早期に発見し、実世界での使用に合わせることで、モデルのパフォーマンスを向上することができる。

スコープと適用可能性

AI、特に生成AI（GenAI）の複雑な法的状況をナビゲートすることは、その本質的な多様性ゆえに、大きな挑戦である。本書では、リアルなテキスト形式（コード、スクリプト、記事）を生成する深層学習モデル、視覚コンテンツを操作するコンピュータビジョンアプリケーション（顔認識、ディープフェイク）、安定した拡散（テキストから画像へのモデル）、自律システム（自動運転車、ロボット）で採用される強化学習アルゴリズムなど、多様なシステムを包含するAIを取り巻く規制の状況を掘り下げる。敵対的生成ネットワーク（Generative Adversarial Networks、GAN）や大規模言語モデル（LLM）のような広範なカテゴリは、多くの生成AIアプリケーションを支えており、規制上の考慮事項に含める必要がある。急速に進化するこの広範なシステムを管理するためには、現行の法律がこのダイナミックな状況に適応するための課題に直面しているため、繊細なアプローチが必要である。これは、急速に進化するテクノロジーが、競争圧力によって私たちの生活や商習慣に浸透しているにもかかわらず、法的枠組みが不十分で適応が遅れているという危機的状況を生み出している。本書では以下の点を探っていく。

- 最も一般的な既存の規制は、生成AIの特定分野にどのように対処しようとしているのか。
- 新法制定をめぐるいくつかの課題と機会
- 説明可能なAI技術を用いた責任あるAI原則を開発するためのハイレベルな推奨事項とベストプラクティス

本稿では、段階的アプローチを用いてAIのガバナンスを分析し、以下の分野にフォーカスを当てる。

現在のドキュメント	今後の検討事項
トップレベルの政府/連邦立法： ● USA: ○ 大統領令（人工知能におけるアメリカのリーダーシップの維持、人工知能の安全、セキュア、および信頼できる開発と展開に関する大統領令など）、および ○ 連邦議会法案（2023年アルゴリズム説明責任法等）（案）	国家レベル： ● APACからのいくつかの規制：中国（制定）（科学技術省）、日本（内閣府）、韓国（科学ICT省）、シンガポール、インドの国家政策「万人のためのAI」（NITI Aayog） ● その他、AI政策が台頭している国（カナダ、英国、オーストラリア） **国際機関**：フレームワークを探っている

● EU: ○ 欧州委員会の政策文書（信頼できるAIのための倫理ガイドラインなど） ○ 規制（人工知能法など） 主な地域規制： ● カリフォルニア州消費者プライバシー法（CCPA）、カリフォルニア州プライバシー権法（CPRA）による改正 ● 欧州一般データ保護規則 (GDPR)	● OECD（AIに関する勧告） ● ユネスコ（AIの倫理に関する勧告）。 ● 人工知能に関するグローバル・パートナーシップ 人工知能に関するグローバル・パートナーシップ（The Global Partnership on Artificial Intelligence (GPAI)）国際協力を促進するための科学、産業、市民社会、政府、国際機関、学界の専門知識 ● ISO/IEC 42001:2023 (AIMS) ● OWASP Top 10 for Large Language Model Applications

Table 1:ガバナンス領域のスコープ

特定の業界におけるAIガバナンスの詳細については、CSAのA Revolutionary Benchmarking Model for AI Safety を参照。

生成AIに関する法的規制の主要分野

データプライバシーとセキュリティ

生成AIは、データプライバシーとセキュリティの領域でユニークな課題を提示する。膨大な量のデータから学習するその能力は、AIの開発と展開のライフサイクルを通じて、個人情報の収集、保存、使用、共有、転送について懸念を抱かせる。欧州一般データ保護規則 (GDPR)、カリフォルニア州消費者プライバシー法 (CCPA)、カリフォルニア州プライバシー権法(CPRA)、医療保険の相互運用性と責任に関する法律(HIPAA)など、いくつかの既存の法律や規制は、以下のように個人のプライバシーとデータセキュリティを保護することを目的としている。

欧州一般データ保護規則（GDPR） (EU)

- **適用可能性**：GDPRは、組織の所在地にかかわらず、欧州経済地域（EEA）内で個人の個人データを処理する組織に適用される。
- **主な規定**：
  - **処理の合法的根拠、公平性、透明性**：組織は、個人データを処理するための合法的な根拠（ユーザの同意、正当な利益など）を持たなければならない。データ収集と処理の目的について、明確かつ具体的な情報を個人に提供することを求めている。

- **データの最小化**：個人データの収集および保持は、明示された目的に厳密に必要なものに限定する。
- **データ主体の権利**：アクセス権、修正権、消去権、および処理制限権など、個人データに関する様々な権利を個人に付与する。
- **セキュリティ対策**：個人データを未認可なアクセス、開示、改ざん、または破壊から保護するための適切な技術的および組織的措置を要求する。
- **プロファイリングを含む自動化された個人の意思決定**：プロファイリングを含む自動意思決定には、データ主体の明示的な同意が必要である（GDPR、 第22条）。

- **生成AIのためのGDPRコンプライアンス**：EUのGDPRは、AIシステムで使用されるデータを含む個人データの処理について、個人が同意を提供することを義務付けている。さらに、データ保護の要件は、システムがGDPRの原則（合法性、公正性、透明性、目的の限定、データの最小化、正確性、保存の制限、完全性、機密性など）に準拠しなければならないことを意味する。

1. 合法的かつ透明性のあるデータ収集および処理

- **トレーニングとプロンプトデータの限界**：GDPRは、データの取り扱いに関する主要原則を以下のように概説する：
  - **目的の制限**：データは、特定の、明確に定義された、または互換性のある目的のためにのみ収集および使用することができる。
  - **必要性**：これらの目的を達成するために不可欠な個人データのみを収集し、使用することができる。
  - **データの最小化**：収集・利用する個人データの量は最小限にとどめ、絶対に必要なものだけを収集する。
  - **保管時間の制限**：個人データの保管は可能な限り短期間でなければならず、保管の期限を定め、定期的に見直さなければならない。

  トレーニングデータ（プロンプトデータも「トレーニングデータ」になる可能性がある）の文脈では、これは特定のトレーニング目的に本当に必要な範囲でのみデータを収集し、使用することを意味する。

- **告知に基づく同意** ：GDPRは、生成AIモデルの学習に使用する個人データの収集と処理について、ユーザーの明示的な同意を義務付けている。これにより、個人は自分のデータがどのように使用されるかを理解し（モデルのトレーニングや微調整のためなど）、拒否する権利を持つことになる。AI開発者は、AI/MLシステムによってデータが処理される個人によるこれらの権利の行使を促進しなければならない。
- **透明性**：EUの個人は、 アクセス権、修正権、消去権、処理の制限、およびデータポータビリティなど、個人情報に関する権利を有する 。組織は、目的、法的根拠、およびデータ保持期間を含め、AIやMLにおける個人データの利用方法について透明性を持たなければならない。ユーザーは、自分のデータがどのように生成されたアウトプットに貢献するかを理解できなければならない。

2. データセキュリティと説明責任

- **データセキュリティ**：GDPR第25条では、組織は「データ保護byデザイン・デフォルト」を採用し、暗号化、アクセスコントロール、およびデータ侵害通知手続きなど、基盤モデルで使用される個人データのセキュリティを確保するための適切な技術的および組織的措置を実施しなければならないとされている。さらに、LLMはサプライチェーン全体の一部であるため、そのセキュリティには、敵対的攻撃、データポイズニング、およびモデルバイアスなどの悪意のある手法に細心の注意を払う必要がある。
- **説明責任**：組織は生成AI対応システム内で個人データを使用する責任を負い 、GDPRの遵守を証明しなければならない。これには、データ保護影響評価の実施と適切な記録の管理が含まれる。
- **データの匿名化と仮名化**：匿名化や 仮名化はプライバシーリスクの軽減に役立つが 、限られた情報でもアイデンティティを推測できる生成AIの文脈では、必ずしも十分ではないかもしれない。
- **生成AIが出力する潜在的な弊害**：GDPRはモデルトレーニングに使われるデータだけに影響するように見えるが、規制はモデルの出力にも適用される。これには、意図せずに生成された出力や、個人の評判を傷つけ倫理原則に反するディープフェイクの悪意ある使用への対処も含まれる。明確なガイドラインとセーフガードを確立することは、生成AIの責任ある開発と使用を保証し、リスクを軽減し、さらに潜在的な危害から個人を保護するために不可欠である。

3. 個人の権利とコントロール

- **アクセス権および修正権**：個人は、生成AIで使用される自己の個人データを理解し、アクセスする権利を有し、不正確または不完全である場合には修正を要求する権利を有する。これには、生成AIに直接提供された情報、または生成AIとのやり取りを通じて生成されたデータが含まれる。しかし、伝統的なデータベースとは異なり、AIのトレーニングデータに対する修正を実施することは、データのサイズが大きく、相互に接続されている性質があるため、潜在的にモデル全体の再トレーニングを必要とし、意図しない結果を引き起こす可能性があるという課題を提起する。現在までのところ、AIモデルの学習データにすでに取り込まれた不正確な情報を修正することの可能性は不明である。データラベリングとプライバシー保護技術に関する研究は進行中だが、「修正する権利」の確保は依然として未解決の課題であり、この要件を促進する方法に関する研究は注視が必要である。
- **消去権（忘れられる権利）**：個人は、個人データの消去を要求する権利を有し、これはAI/MLモデルのトレーニングと使用方法に影響を与える可能性がある。個人データは、トレーニング後にモデルの複雑な内部表現に深く埋め込まれる可能性がある。現在のところ、トレーニングされたモデルから特定のデータポイントを削除することの技術的実現可能性と倫理的意味は不明なままであり、このような要請を処理するための信頼できるプロセスや確立されたガイダンスがないため、個人のプライバシーとモデルの全体的な機能性や社会的利益とのバランスについて重大な疑問が投げかけられている。
- **異議申し立ての権利**：個人は、 生成AIの文脈を含め、特定の目的のために個人データを処理することに反対する権利を有する。しかし、生成AIの文脈でこの権利を行使することは、特有の課題をもたらす。現在のところ、モデルの学習が完了した時点で、学習セットから個人データを削除する信頼性の高い標準化されたプロセスは存在しない。

さらに、異議申し立ての権利は、特定のデータ要素および／または特定の目的にのみ適用される可能性があり、モデルのトレーニングに使用されるすべての情報に適用されるとは限らず、個人の異議申し立てのスコープが限定される可能性がある。このことは、個人のプライバシーの権利を尊重する生成AIシステムのための、透明で説明可能な慣行の継続的な開発の必要性を浮き彫りにしている。

- **コンプライアンス**：GDPRは、データ処理活動に対してデータプライバシー影響評価（DPIA）の実施を義務付けている。これは、AIシステムによるデータ処理と、それがデータ主体にもたらすリスクにも及ぶ。大規模な生成モデルの学習に使用される大規模なデータセットの中で個人データを特定することは困難であり、EUが生成AIの文脈でGDPRコンプライアンスにどのように対処するかは不明確なままである。
- **ADMガバナンス**：GDPR第22条は、プロファイリングを含む、個人に法的または重大な影響を及ぼす自動意思決定（ADM）に反対する権利を個人に認めている。これは、特にADMによる決定が、偏見によって個人の生活に重大な影響を与える可能性がある場合、個人はADMからオプトアウトする権利、またはADMによる決定に異議を唱える権利を有することを意味する。その結果、ADMを使用する企業は、人間による不服審査プロセスを持つことが求められる。

カリフォルニア州消費者プライバシー法/カリフォルニア州プライバシー権法 (CCPA/CPRA)

- **適用範囲**：カリフォルニア州で事業を行い、世界売上高が2,500万米ドルを超えるなど、他の閾値要件を満たす営利企業に適用される。これは、カリフォルニア州民に、自分についてどのような個人データが収集されているかを知る権利、およびその削除や正確性のための変更を要求する権利を与えるものである。企業はまた、個人情報の収集と処理を、開示された目的のために必要なものに限定しなければならない。CCPAはこのデータに依存するAI/MLシステムにも適用され、組織はこれらのシステムがカリフォルニア州住民の個人情報に関わるトレーニングや出力生成に関するプライバシー要件に準拠していることを確認する必要がある。企業は、生成AI（基盤型）モデルの開発・デプロイにおいてカリフォルニア州民の個人データを利用する際、CCPAの義務を慎重に考慮しなければならない。

- **主な規定**：
  - **知る権利**：消費者は、自分について収集されたカテゴリおよび特定の個人情報に関する情報を要求することができる。
  - **削除権**：事業者が収集した個人情報の削除を求める権利を消費者に付与する。
  - **オプトアウトの権利**：消費者に個人情報の販売をオプトアウトする権利を与える。

注：*CCPAとその延長であるCPRAは、一般的に使用される「個人を特定できる情報」（PII）よりも広範な消費者データの定義を使用している。このため、本文書では、CCPAのスコープとの整合性を確保するため、「個人情報」（PI）という用語を採用している。PIIとは通常、氏名や社会保障番号など、個人を直接的に特定する特定のデータを指す。しかし、CCPAのPIの定義は、より広範なデータポイントを包含している。これには、閲覧履歴、IPアドレス、ジオロケーションデータが含まれる。これらは、単独ではPIIとはみなされないかもしれないが、他の情報と組み合わされたときに、誰かを特定するために使用される可能性がある。したがって、「個人情報」は、消費者のデータプライバシーに関するCCPAの意図をより正確に反映している。*

- **生成AIのための**CCPA/CPRA**コンプライアンス**：CCPA/CPRAは、生成AIに対する直接的な技術的要件を提示してはいないが、個人データの権利に焦点を当てることで、コンプライアンスを確保するための慎重な実践を必要とする重大なデータ管理上の課題が生じる可能性があり、モデルのパフォーマンスや機能性に影響を与える可能性がある。CCPA/CPRAが保護するのはカリフォルニア州居住者の個人データのみであることを忘れてはならない。考慮すべき点は以下の通りである。

1. CCPA/CPRAに基づくデータの収集、保存、使用、開示

CCPA/CPRAは、主にカリフォルニア州住民に関する企業による個人情報の収集、使用、開示を規制することに主眼を置いている。これは、AI/MLモデルの学習に使用されるデータ、およびその結果の出力に個人情報が含まれている場合に適用される。カリフォルニア州民は、CCPA/CPRAに基づき、自分の個人情報にアクセスする権利を有する。この権利は、モデルトレーニングに使用されるデータにも適用される可能性があるが、個人データを含む出力と、より一般的なモデルの出力とを区別することが重要である。カリフォルニア州居住者は、AI目的で収集される個人情報の内容、収集目的、および共有される第三者のカテゴリを知る権利を有する。CCPA/CPRAは必ずしも特定のトレーニングデータソースの開示を義務付けているわけではないが、透明性を重視していることは確かである。

トレーニングに使用されるデータの出所と系統を追跡するデータ・プロビナンスは、特に生成AIにしばしば使用される膨大なデータセットを考慮すると、CCPA/CPRAコンプライアンスに不可欠である。複雑なデータの出所は、「アクセスする権利」や「知る権利」の要求を満たすことを困難にする。強固なデータガバナンスの実践、適切なロギング、および匿名化されたトレーニングデータの開示の可能性などが、こうした課題を軽減するのに役立つ。

2. 消費者の権利

CCPA/CPRAは、個人情報へのアクセス権、削除権、修正権、オプトアウト権など、消費者の個人情報に関する特定の権利を認めている。詳細は以下の通りである。

- **知る権利**：モデルのトレーニングのための個人情報（PI）の収集と使用に関する詳細の開示を義務付ける。これには、トレーニングに使用するデータカテゴリー（テキスト、画像、音声、名前、場所など）の特定、PIの情報源の特定（ユーザーとのやりとり、購入／第三者のデータセット、ソーシャルメディア、公的記録など）、PIの使用目的の詳細（モデルトレーニング、性能評価など）が含まれる。
- **アクセス権**：ユーザーは、トレーニングデータに使用されている特定のデータポイントへのアクセスを要求することができ、トレーニングプロセスによっては、個人を特定できる情報が明らかになる可能性がある。これには、トレーニングデータセット内の個々のデータポイントを識別し、分離するメカニズムの実装を必要とする可能性があるが、匿名化または集約技術が導入されている場合、技術的に困難な場合がある。
- **削除権**：ユーザーは、トレーニングに使用された自分の個人情報の削除を要求する権利を有し、いくつかの方法でモデルに影響を与える。
    - **データの削除**：この場合、残りのデータでモデルの再トレーニングが必要になり、性能と汎用性に影響を与える可能性がある。
    - **データの修正**：トレーニングプロセスによっては、特定のデータポイントの匿名化や再編集が必要で、モデルの精度や解釈可能性に影響を与える可能性がある。

    - **知識の除去**：何千億もの層からなるディープニューラルネットワークで学習された知識をどのように識別し、特定の学習情報を取り除けるだろうか？実際には、LLMをゼロから再教育する必要があり、経済的にも環境的にも実現不可能である。

  技術的な実現可能性の観点から見ると、複雑な学習データセット内の個々のデータポイントを特定し、削除することは、高度なAIシステム（LLMなど）にとっては、計算コストが高く、時間がかかり、あるいは単に不可能な場合もある。削除が必要なデータでトレーニングされたモデルの扱いについては、まだ未解決のままである。
- **販売をオプトアウトする権利**：生成されたAI出力が CCPA/CPRAの下で「個人情報」とみなされる場合 （例：ディープフェイク）、ユーザーは第三者への販売や開示をオプトアウトする権利を有する可能性がある。これには、生成AIのアウトプットをCCPAのフレームワークの下で明確に定義し、分類することが必要であり、さらなる明確化と法的解釈が必要になる可能性がある。

3. コンプライアンスと執行

CCPA/CPRAの遵守には、主に個人情報を保護するための技術的および手続き的な保護措置の実施が含まれる。

カリフォルニア州プライバシー保護庁（CPPA）は、2020年に設立された比較的新しい機関で、消費者データやプライバシーを含む様々な分野にわたる規制を確立している最中である。CPPAは、カリフォルニア州プライバシー権法 (CPRA)および カリフォルニア州消費者プライバシー法 (CCPA)を実施・執行する。AIの管理のみに焦点を当てた具体的な規制はまだ発表されていないが、2つの重要な動きがAI、特に生成AIに触れている。

4. 自動意思決定技術(ADMT)規則ドラフト

- 2023年11月に発表されたこの規制案は、 自動意思決定技術（ADMT） の責任ある利用に焦点を当てたもので、これには消費者向けの意思決定に使用される多くの形態のAIが含まれ、その性質はGDPR第22条に類似している。
- 規則案では、ADMTを利用する事業者に対する要求事項がまとめられている。
    - **使用前の注意事項**：消費者に影響を与える意思決定プロセスにおいて、ADMTを使用する前に消費者に情報を提供する。
    - **オプトアウトの権利**：ADMTのみによる決定に従わないという選択肢を消費者に認める。
    - **アクセスと説明を受ける権利**：ADMTがどのように利用され、どのような決定に至ったかを説明する情報を消費者に提供する。
    - **リスク評価**：偏見や差別など、ADMTの使用に関連する潜在的な危害を特定し、軽減するためのリスクアセスメントの実施を企業に義務付ける。

生成AIについては特に言及していないが、これらの規制は、消費者に関する自動意思決定を行うために使用されるあらゆるAI技術に適用される可能性があり、カリフォルニア州における企業の生成AIの導入・活用方法に影響を与える可能性がある。

5. 生成AIに関するカリフォルニア州行政命令

- 2023年10月、カリフォルニア州知事のGavin Newsomは、州政府内で生成AIの責任ある開発、採用、および導入を検討するワーキンググループを設置する行政命令を出した。
- この命令は、生成AIの潜在的な利点を強調する一方で、偽情報の拡散や責任ある配備の必要性といった潜在的なリスクも認めている。
- この作業部会は、以下のようなテーマについてカリフォルニア州政府機関への提言を作成することを任務としている。
    - 生成AI導入の潜在的なメリットとリスクの特定。
    - 生成AIを使用するための倫理原則の確立。
    - 潜在的な危害を軽減するためのセーフガードの実施。

この州行政命令は、民間セクターを直接規制するものではないが、カリフォルニア州が生成AIの開発と利用の将来を理解し、潜在的に形成するための積極的なアプローチを示している。

CPPAが進化を続け、生成AIの複雑性に適応していくにつれて、コンプライアンス要件が追加され、複雑性が増す可能性が予想される。このことは、生成AIの責任ある開発と展開を促進しながら、進化する規制の状況をナビゲートする継続的な努力の必要性を浮き彫りにしている。

欧州AI規制法(EU AI Act / EUAIA)

- **適用対象**：欧州AI規制法は、欧州連合（EU）における人工知能システムの開発、デプロイ、使用に関わるプロバイダ、デプロイ業者、輸入業者、販売業者、その他の事業者に適用される。軍事、防衛、または国家安全保障目的には適用されない。これは、AIシステムの開発者とユーザーに対し、リスクの4つのレベル（許容できないリスク、高リスク、限定的なリスク、および最小のリスク）に焦点を当てた一連の規則と要件を提案するものである。この法律は、プライバシーや非差別といった基本的権利の保護、AIシステムの安全性と透明性、AI技術の責任ある利用を確保することを目的としている。そのAIシステムがEU市場で提供または使用される場合、あるいはEU内の人々に影響を与える場合、EU内外に拠点を置く事業者に適用される。この法律は、生体認証、自律走行車、重要インフラストラクチャなど、幅広いAIアプリケーションに適用される。
- **主な規定**：
    - **禁止行為（第5条）**：同規則の第5条は、AIシステムに関する禁止行為を概説している。これらの行為は、個人の保護と安全を確保し、AIシステムの非倫理的で有害な使用を防止するために禁止されている。人間の行動を操作するAI、ソーシャルスコアリングシステム、法執行を目的とした公共空間での「リアルタイム」遠隔生体認証システムの使用など、許容できないリスクとみなされるAIシステムはEUで禁止される。
    - **リスクベースアプローチ（第9条）**：欧州AI規制法第9条は、 EUにおけるAIシステムの規制にリスクベースのアプローチを導入し、規制とイノベーションのバランスをとることで、AIシステムの安全性と信頼性を確保しつつ、不必要なコンプライアンスコストを回避することを目的としている。AIシステムは、高リスク、限定的なリスク、および最小のリスクのいずれかに分類され、個人に及ぼす潜在的な危害のレベルに応じて規制のレベルが変わる。
        - **高リスクAIシステム**：重要インフラストラクチャで使用されるような高リスクAIシ

ステムは、 厳しい要件を満たし 、精査を受け、デプロイに事前承認を得なければならない。このようなシステムのプロバイダは、透明性と説明可能性、人間による監視、および独立した検証など、規制の最も厳しい規定を遵守しなければならない。

- **限定的なリスクAIシステム**：限定的なリスクAIシステムは、リスクは低いものの、特定の要件を遵守しなければならない。これらのシステムのプロバイダは、関連する法的義務、透明性、およびトレーサビリティの規則を満たしていることを保証しなければならない。
- **最小のリスクAIシステム**：最小のリスクAIシステムは、個人に対してほとんど、あるいはまったくリスクをもたらさない。これらのシステムには同じ規制要件は適用されないが、AIシステムに適用される法的枠組みには従うべきである。

- **データガバナンス（第10条）**：第10条の目的はAIシステムにおけるデータの利用が透明で、説明責任を果たし、個人のプライバシーとデータ保護の権利を尊重することを保証することである。AIシステムのトレーニングやフィードに使用されるデータは、欧州一般データ保護規則（GDPR）およびその他の関連データ保護法の規定を遵守する必要がある。この条文では、高リスクAIシステムのプロバイダは、AIシステムのトレーニングやフィードに使用されるデータが適切で、信頼性が高く、バイアスがなく、ならびに誤りがないことを保証しなければならないと定めている。また、システムのパフォーマンスを監視・監査するために、データが適切に文書化、ラベル付け、および注釈が付けられていることも確認する必要がある。さらに同条は、データは透明性をもって管理されなければならず、データが使用される個人には情報を提供し、同意を得なければならないと定めている。
- **透明性と説明可能性（第13条）**：この条文では、高リスクAI システムは透明で説明可能なものでなければならず、個人がその仕組みや意思決定を理解できるようにし、どのように機能しているかを説明し、ユーザーが文書にアクセスできるようにすることを求めている。AIモデルは、監査ができるように適切な記録とログとともにそれが維持されなければならない。この条文はまた、AIシステムが誠実さ、説明責任、および透明性をもって運用されることを保証するために、情報を得る権利と、AIシステムによる決定に異議を唱えるために人間の介入を求める権利を定めている。
- **人間による監視（第14条）**：人間による監視は、リスクを防止または最小化することを目的とすべきであり 、システムに組み込まれた対策または配備者が実施する対策によって達成できる。さらに、AIシステムは人間のオペレータが時折チェックできるように設計されていなければならない。システムを監督する自然人は、システムの能力を理解し、その運用を監視し、その出力を解釈し、必要に応じて介入することができなければならない。バイオメトリクス識別システムに関する具体的な要件が概説されている。
  高リスクAIシステムは、人間による監視を確保することを目的とした厳格な義務の対象となり、自然人が効果的に監視できるように設計されるべきである。
- **独立した試験と検証（第57条から第63条まで）**：高リスクAI システムは、安全性と信頼性を確保するために、独立したテストと検証を受ける必要がある。
- **ガバナンスと認証（第64条から第70条まで）**：EUは、EU内のAIシステムが必要な基準や規制を満たすことを保証するために、統治機構と認証の枠組みを確立している。同規則は、国内およびEUレベルで同規則の適用を調整・支援するためのガバナンスの枠組みを確立している。ガバナンスの枠組みは、連邦レベルでの調整と専門知識の構築、既存の資源と専門知識の活用、デジタル単一市場の支援を目的としている。

- **罰則（第99条）**：この条文では、同規則の規定に違反した場合に課される制裁、措置、および罰則について説明している。同規則は、加盟国は同規則の規定を執行するために、適切な行政手続きまたは司法手続きを確立しなければならないと定めている。規制が効果的に施行されるようにし、違反には多額の罰則を課すことでコンプライアンス違反を抑止する。これは、AIシステムが責任と倫理を持って開発、導入、および使用され、個人の権利と自由が保護されることを目指すものである。欧州AI規制法（EUAIA）では、違反の重大性に応じて制裁と制裁金が段階的に定められている。段階的アプローチは、罰則が各違反によって引き起こされた損害のレベルに比例することを保証することを目的としている。

生成**AI**のための**EUAIA**コンプライアンス

1. 要件、義務および規定

この規則は、域内市場の機能を向上させ、欧州連合（EU）における人間中心の信頼できる人工知能（AI）システムの導入を促進することを目的としている。これは、AIシステムの市場投入、サービス開始、および使用に関する調和された規則と、高リスクAIシステムに関する特定の要件と義務を定めたものである。また、特定のAI慣行の禁止を盛り込み、特定のAIシステムの透明性ルールを確立する。
さらに、市場監視、市場監視ガバナンス、および執行にも言及している。

高リスクAIシステムのプロバイダの義務：高リスクAIシステムが、概略の要件に準拠していることを確認する。

- **リスク管理（第9条）**：プロバイダは、高リスクAIシステムについて、安全、基本的権利、およびシステムの意図された目的に対する潜在的なリスクを考慮し、徹底的なリスク評価を実施しなければならない。高リスクAIシステムについては、既知および予見可能なリスクを特定・分析し、顕在化する可能性のあるリスクを評価し、リスク管理策を採用することを含むリスク管理体制を確立しなければならない。リスク管理措置は、特定されたリスクを排除または低減し、規則に定められた要求事項の複合的な影響に対処することを目的とする。リスク管理システムは、高リスクAIシステムの全体的な残存リスクが許容可能であると判断されることを保証しなければならない。
- **データの質とガバナンス（第10条）**：プロバイダは、高リスクAIシステムが、高品質で関連性があり、また、代表的なデータセットでトレーニングされていることを確認しなければならない。バイアスを防ぎ、データの正確性を確保するために、適切なデータガバナンス対策を実施しなければならない。データトレーニング技術を使用する高リスクAIシステムは、高品質のトレーニング、検証、およびテストデータセットを使用しなければならない。設計の選択、データ収集プロセス、データ準備処理作業、およびバイアスやデータギャップへの対処を含む、データガバナンスの実践が必要である。
- **技術文書（第11条）**：プロバイダは、高リスクAIシステムについて、正確であり最新の技術文書を作成し、維持しなければならない。この文書には、システムの設計、開発、構成、および運用に関する情報を含めるべきである。高リスクAIシステムの技術文書を作成し、常に最新の状態に保つ必要がある。文書は、規制への準拠を証明し、当局や通知機関による評価に必要な情報を提供するものでなければならない。EU整合法令に該当する高リスクAIシステム

については、単一の技術文書を作成すべきである。欧州委員会は、委任法を通じて技術文書要件を改正することができる。

- **記録の保持（第12条）**：高リスクAIシステムは、そのライフタイムを通じてイベント（ログ）を自動的に記録できなければならない。ロギング機能は、リスク状況の特定を可能にし、市販後のモニタリングを容易にし、また、高リスクAIシステムの運用を監視するものでなければならない。
- **透明性と情報提供（第13条）**：プロバイダは、高リスクAIシステムの透明性を確保し、システムの能力と限界に関する適切な情報をユーザーに提供しなければならない。高リスクAIシステムは、配備者がシステムの出力を適切に解釈し利用できるよう、透過的に運用されなければならない。使用説明書には、プロバイダに関する関連情報、システムの特性と能力、既知のリスク、出力を説明する技術的能力、および出力を解釈するための規定を含めるべきである。
- **人間による監督と介入（第14条）**：プロバイダは、高リスクAIシステムにおいて、人間による監視と介入のための適切なメカニズムを組み込まなければならない。これには、必要なときに人間のオペレータがシステムを上書きしたり停止したりできるようにすることも含まれる。高リスクAIシステムは、システム使用中に自然人による効果的な監視を可能にするよう設計されなければならない。人間による監督措置は、リスクを防止または最小化することを目的とすべきであり、システムに統合することも、配備者が実施することもできる。人間による監視を担当する自然人は、システムの能力と限界を理解し、異常を検知し、システム出力を解釈し、また必要であればシステムの決定に介入するか、無効化することができなければならない。
- **正確性、堅牢性、サイバーセキュリティ（第15条）**：プロバイダは、高リスクAIシステムの正確性、信頼性、堅牢性を確保しなければならない。システムの性能に関連するエラーやリスクを最小限に抑え、精度と堅牢性の課題に対処するために必要な措置を講じるべきである。リスクを特定し、システムの設計を考慮して必要な緩和策を実施するために、セキュリティリスクアセスメントを実施すべきである。高リスクAIシステムは、包括的なリスク評価を受け、サイバーセキュリティ基準を遵守することが求められる。また、言語モデルを利用する際には、適切なレベルの精度、堅牢性、およびサイバーセキュリティを達成する必要がある。精度と堅牢性の技術的側面に対処するために、ベンチマークと測定手法が開発される可能性がある。精度のレベルと関連する測定基準は、添付の使用説明書で宣言されるべきである。
- **特定のAIシステムに関する特定の要件（第53条および第55条）**：同規則は、生体認証システム、重要インフラストラクチャで使用されるシステム、教育や職業訓練で使用されるシステム、雇用目的で使用されるシステム、および法執行当局が使用するシステムなど、特定の種類の高リスクAIシステムに対する特定の要件を特定している。
    - 高リスクAIシステムまたはそのパッケージ／文書に、その名称、登録商標または登録商品名、連絡先住所を明記する必要がある。
    - 規制への準拠を保証する品質マネジメントシステムを導入していること。

- 技術文書、品質マネジメントシステム文書、通知機関が承認した変更、通知機関が発行した決定書、およびEU適合宣言書などの文書を保管すること。
- 高リスクAIシステムが生成したログを一定期間保存すること。
- 高リスクAIシステムを市場に出したり、使用開始したりする前に、関連する適合性評価手続きを受けること。
- EU適合宣言書を作成し、規制への適合を示すCEマーキングを貼付すること。
- 登録義務を遵守すること。
- 必要な是正措置を講じ、必要に応じて情報を提供すること。
- 国家所轄庁の合理的な要求があれば、高リスクAIシステムの適合性を証明すること。
- アクセシビリティ要件に確実に準拠すること。

輸入者の義務：

- 高リスクAIシステムを市場に出す前に適合性を検証すること。
- 高リスクAIシステムに必要なCEマーキングが付され、EU適合宣言書が添付され、使用説明書が添付されていることを確認すること。
- 高リスクAIシステムが適切に保管され、輸送すること。
- 通知機関が発行した証明書、使用説明書、およびEU適合宣言書のコピーを保管すること。
- 要求に応じて、必要な情報と書類を各国の管轄当局に提供すること。
- 高リスクAIシステムがもたらすリスクを軽減するため、各国の所轄当局と協力すること。

ディストリビュータの義務：

- 高リスクAIシステムに必要なCEマーキングが付され、EU適合宣言書が添付され、使用説明書があることを確認すること。
- 該当する場合、その名称、登録商号または登録商標、および連絡先住所を包装／文書に明記すること。
- 保管または輸送条件が、高リスクAIシステムのコンプライアンスを危険にさらさないようにすること。
- 通知機関が発行した証明書、使用説明書、およびEU適合宣言書のコピーを保管すること。
- 要求に応じて、必要な情報と書類を各国の管轄当局に提供すること。
- 高リスクAIシステムがもたらすリスクを軽減するため、各国の所轄当局と協力すること。

2. イノベーションの促進（第57,58,59,60,61,62,63条）

イノベーション支援策は以下の通りである。

AI規制のサンドボックス：

- 加盟国は、国レベルでAI規制のサンドボックスを設置し、市場に投入される前の革新的なAIシステムの開発、テスト、および検証を促進することが求められる。
- サンドボックスは、イノベーションを促進し、リスクの特定と軽減を可能にするコントロールされた環境を提供する。
- これらの目的は、法的確実性を向上させ、ベストプラクティスの共有を支援し、イノベーションと競争力を促進し、エビデンスに基づく規制の学習に貢献し、特に中小企業や新興企業にとって、AIシステムのEU市場へのアクセスを容易にすることである。
- 各国の管轄当局はサンドボックスに対する監督権限を持ち、他の関連当局との協力を確保しなければならなりない。

AIサンドボックスにおける個人データの処理：

- その他の目的で収集された個人データは、公益のために特定のAIシステムを開発、訓練、およびテストするためだけに、AI規制サンドボックスで処理される場合がある。
- データ保護規制を確実に遵守するためには、効果的な監視メカニズム、データ主体の権利保護措置、個人データ保護のための適切な技術的・組織的措置などの条件を満たさなければならない。

リスクの高いAIシステムの実環境でのテスト：

- リスクの高いAIシステムのプロバイダやプロバイダ候補は、AI規制のサンドボックスの外で、実環境でのテストを実施することができる。
- 実地試験計画を策定し、市場監視当局に提出しなければならない。
- テストは単独で行うことも、採用見込みの配備者と共同で行うこともできる。
- 倫理的審査は、EU法または国内法で義務付けられている場合がある。

ガイダンスとサポート：

- AI規制サンドボックス内の所轄官庁は、参加者にガイダンス、監督、およびサポートを提供する。
- プロバイダは、規制の導入、標準化、および認証に関するガイダンスなど、導入前のサービスを受けることができる。
- 欧州データ保護監督当局は、欧州連合の機関、団体、事務所、および機関専用のAI規制サンドボックスを設置することができる。

ガバナンスと調整：

- 同規則は、国レベルおよびEUレベルでAI規則の実施を調整・支援するためのガバナンスの枠組みを確立している。
- AI事務所は加盟国の代表で構成され、AIに関するEUの専門知識と能力を開発し、EUのAI法の実施を支援する。

- 欧州AI委員会、科学委員会、および諮問フォーラムが設置され、規制の実施に必要な意見、助言や専門知識を提供する。
- 各国の所轄官庁は欧州AI委員会内で協力し、AI規制のサンドボックスの進捗状況と結果について年次報告書を提出する。
- 欧州委員会は、AI規制のサンドボックスに関する単一の情報プラットフォームを開発し、各国の所轄当局と調整を行う。

市場監視とコンプライアンス：

- 加盟国が指定する市場監視当局は、規制の要件と義務を執行する。
- 執行権限を持ち、独立かつ公平に職務を遂行し、共同活動や調査を調整する。
- コンプライアンスは、リスクの軽減、市場での入手の制限、AIモデルの撤回または回収を含む措置を通じて執行可能である。

データ保護当局の関与：

- 各国のデータ保護当局や、監督的な役割を担うその他の関連する国の公的機関や団体は、基本的権利を保護するEU法に沿ってAIシステムを監督する責任を負っている。
- 彼らは規則に基づき、作成された関連文書にアクセスすることができる。

金融サービス当局との協力：

- EU金融サービス法の監督を担当する所轄当局は、規制・監督される金融機関が提供または使用するAIシステムに関する市場監視活動を含め、AI規制の実施を監督する所轄当局として指定される。
- 欧州委員会は、首尾一貫した義務の適用と執行を確保するために、これらの機関と調整を図っている。

倫理的で信頼できるAIの推進：

- 高リスクに分類されないAIシステムのプロバイダは、高リスクAIシステムに適用される必須要件の一部または全部を自主的に適用する行動規範を作成することが奨励される。
- AI事務局は、汎用AIモデルを提供するすべてのプロバイダに対し、実施規範の遵守を求めることができる。

透明性のある報告と文書化：

- プロバイダは、AIシステムの使用とリスクを分析するための市販後モニタリングシステムを持つことが義務付けられている。
- AIシステムの使用に起因する重大な事故は、関係当局に報告する必要がある。
- AI規制サンドボックスからの技術文書と終了報告書は、規制への準拠を証明するために使用することができる。
- 欧州委員会及び欧州AI委員会は、関連業務に関する終了報告書にアクセスすることができる。

3. 特定のAI行為の禁止

- **人間の行動を著しく歪める**：身体的、心理的健康、または経済的利益に重大な害をもたらす可能性のある、人間の行動を実質的に歪曲する目的または効果を持つAIシステムを利用可能にすること、使用開始すること、ならびに使用することは禁止されている。これには、人の自律性、意思決定、自由な選択を破壊または損なう、サブリミナル成分やその他の操作的または欺瞞的なテクニックの使用が含まれる。
- **機微な個人データのためのバイオメトリック分類**：政治的意見、労働組合への加盟、宗教的・哲学的信条、人種、性生活、性的指向などの機微な個人情報を推測または推論するために、自然人のバイオメトリクスデータに基づくバイオメトリクス分類システムを使用することは禁止されている。
- **ソーシャルスコアリングを提供するAIシステム**：自然人の社会的行動、既知、推測、予測される個人的特徴、または性格的特徴に基づいて自然人を評価または分類するAIシステムは 、差別的な結果や特定の集団の排除につながる可能性がある。このようなAIシステムを、データが生成または収集された背景とは無関係に、個人または集団に不利益または不利な取り扱いをもたらす社会的採点目的で使用することは禁止されている。
- **法執行機関のためのリアルタイム遠隔生体認証**：法執行の目的で、公共のアクセス可能な空間における個人のリアルタイムの遠隔生体認証は、押しつけがましいと考えられ、個人の私生活に影響を及ぼす可能性がある。このような行為は、行方不明者の捜索、生命や身体の安全に対する脅威、特定の重大な犯罪の加害者や容疑者の特定など、実質的な公共の利益を達成するために厳密に必要な、限定的に列挙された狭く定義された状況を除き、禁止されている。

4. コンプライアンス、違反、罰則

同規則はさまざまな用語の定義を定め、そのスコープを定めている。AIシステムに関する個人情報の保護、プライバシー、および守秘義務を強調している。同規則（AI法）には、コンプライアンスに関する規定、違反に対する罰則、影響を受けた人に対する救済措置が含まれている。さらに、同規則（AI法）は将来的な評価と見直しを可能にし、実施権限を欧州委員会に委譲するもので、発効後一定の期間内に適用されることになっている。

コンプライアンスに関する規定：

- 汎用AIモデルのプロバイダは、規則発効日から36ヶ月以内に、規則で定められた義務を遵守するために必要な措置を講じなければならない。
- 一定の期日（規則発効日から24ヶ月）前に市場に投入された、あるいは運用が開始された高リスクAIシステムのオペレータは、その設計に大幅な変更が加えられた場合に限り、規則の要求事項の対象となる。
- 高リスクAIシステムを使用する公的機関は、規則発効日から6年以内に規則の要求事項を遵守しなければならない。

侵害に対する罰則：

欧州AI規制法における違反に対する罰則は、段階的なアプローチに従っている。

- 要求に対する回答として、不正確な、不完全な、または誤解を招くような情報を通知機関または国家管轄当局に提供した違反に対しては、750万ユーロ以下、または違反者が事業者である場合は、前会計年度の全世界の年間総売上高の1%以下のいずれか高い方を制裁金として課される。
- 高リスクAIシステムの認証を取得していない、リスク管理などの透明性や監督要件を遵守していない、あるいはプロバイダ、正規代理店、輸入業者、販売業者、または配備業者の義務に違反している場合、制裁金は最高1,500万ユーロ、または全世界の年間売上高の3%のいずれか高い方となる：
    - 第16条に基づくプロバイダの義務
    - 第22条に基づく委任代理人の義務
    - 第23条に基づく輸入者の義務
    - 第24条に基づく販売業者の義務
    - 第26条に基づく配備業者の義務
    - 第31条、第33条(1)、第33条(3)、第33条(4)または第34条に基づく通知機関の要件と義務
    - 第50条に基づくプロバイダとユーザーの透明性義務
- 許容できないリスクをもたらすと判断されたAIシステムの使用や、規則第5条に記載された禁止されたAI慣行への不遵守などの違反に対して、制裁金は最高3,500万ユーロ、または全世界の年間売上高の7%のいずれか高い方となる。

EUAIAは、課される制裁金は特定の状況に関連するすべての状況を考慮するよう求めている。これには、侵害の性質、重大性、期間、その結果影響を受けた人数、および、その人が被った損害が含まれる。制裁金は、AIシステムの目的に照らして評価されるべきである。さらに、他の当局から行政処分を受けたことがあるかどうか、事業者の規模、年間売上高、および市場占有率などの要因も考慮すべきである。その他の決定要因としては、侵害の結果生じた金銭的利益や損失、各国の管轄当局との協力の程度、事業者の責任、侵害が知られるようになった経緯、事業者側に過失や故意があったかどうか、および影響を受けた人々が被った損害を軽減するために取られた措置などが考えられる。また、訴訟手続きにおいて当事者の防御権が十分に尊重されるべきであり、個人データや企業秘密の保護という個人または企業の正当な利益に従って、関連情報にアクセスする権利があるとしている。

医療保険の相互運用性と責任に関する法律 (HIPAA)

医療保険の相互運用性と責任に関する法律（HIPAA）は、1996年に米国で制定された連邦法で、主に医療データのプライバシーとセキュリティに関する規定として知られている。

- **適用対象：**HIPAAは、個人の保護されるべき健康情報（PHI）を取り扱う、医療プロバイダ、医療計画、および医療クリアリングハウスを含む対象事業体に適用される。
- **主な規定**：

- **最小限必要な基準**：対象事業体は、意図された目的を達成するために必要な最小限のPHIのみを使用し、開示することを要求する。
- **管理的、技術的、物理的な保護措置**：PHI の機密性、完全性、可用性を保護するための適切な保護措置の実施を義務付ける。
- **患者の権利**：個人情報へのアクセス、訂正、および開示の記録を要求する権利など、個人のPHIに関する一定の権利を認める。

生成AIにおけるHIPAAコンプライアンス

1. データプライバシーとセキュリティ：

- **データ保護の要件**：HIPAAの厳格なデータ保護基準は、すでに技術分野全体で確立されており、技術の種類や目的に関係なく、すべてのデータ利用に適用される（例えば、PHIを保護するために、開発・デプロイ全体を通じて強固な暗号化が義務付けられている）。しかし、生成AIの領域の関係者は、生成AIの運用と処理の文脈でこれらの既存の原則を適用する具体的なニュアンスを理解し、実行することに焦点を移さなければならない。確立されたルールに再創造は必要ないが、この新たな状況に適応させるには、生成AIがもたらすユニークな課題に注意を払う必要がある。
- **トレーニングデータの制限**：HIPAAはPHIへのアクセスと共有を制限しており、ヘルスケアアプリケーションのための生成AIモデルを訓練するために利用可能な医療データの量を制限する可能性がある。トレーニングデータの出所とコンプライアンスを追跡することは、生成されたアウトプットがプライバシーの懸念を継承しないことを保証するために極めて重要になる。このことは、診断、治療予測、および個別化医療などの分野での開発や精度を複雑にし、医療応用のためのAIモデルの有効性や汎用性を制限する可能性がある。
- **非識別化の要件**：PHIに対して訓練された生成AIからの非識別化された出力でさえ、微妙なパターン、相関関係、または高度な技術によって再識別可能である可能性があり、プライバシーの懸念を引き起こし、HIPAAに違反する可能性がある。匿名化と仮名化はアイデンティティを不明瞭にすることができるが、生成AIの文脈では、特にモデル内で追加のデータソースと組み合わされた場合、再識別を防ぐことができない場合が多い。そのためには、個人のアイデンティティを効果的に保護するための強固なプライバシー保護手法（差分プライバシー、連合学習など）が必要となる。
- **限られたモデルの共有**：PHIでトレーニングされた生成AIモデル間の共有も、プライバシーの懸念から制限されており、この分野での協力と進歩を妨げている。
- **厳格なアクセスコントロール、監査、追跡**：HIPAAは、PHIへのアクセスと使用の厳格な監査と追跡を義務付けている。これは生成AIシステムにも及び、サプライチェーン全体のHIPAAコンプライアンスを確保するための強固なロギングとモニタリングの仕組みが必要となる。

2. モデルのトレーニング、出力、使用方法：

- **トレーニングデータの制限**：HIPAAはPHIへのアクセスと共有を制限しており、ヘルスケアアプリケーションのための生成AIモデルをトレーニングするために利用可能な医療データの量を制限する可能性がある。モデルのトレーニングに関しては、多様で包括的なヘルケアデータセットでモデルをトレーニングする能力を制限することは、偏った出力や不正確な出力につながる可能性がある。差分プライバシーやその他の匿名化技術を導入することで、患者のプライバシーを保護しつつ、トレーニングのためのデータの有用性をある程度確保できる可能性がある。

- **共有と開示の制限**：PHIを含む生成コンテンツの共有や開示は、たとえ匿名化されていたとしても、厳しく制限されている。このため、生成AIを利用した医学的知見の共有や共同研究が制限される可能性があり、慎重な設計と実装が求められる。
- **PHI の生成の制限**：生成AIは、PHIとみなされる可能性のあるデータを直接出力することができないため、トレーニングやテスト目的で合成カルテを生成するようなタスクでさえも、その使用が制限されている。
- **下流での使用制限**：PHIでトレーニングされた生成AIモデルは、モデル自体がPHIを直接出力しないとしても、PHIを公開する可能性のある下流のアプリケーションで使用することはできない。
- **モデルの解釈可能性と説明可能性**：生成AIモデルがどのようにそのアウトプットに到達するのかを理解することは、不注意にPHIを開示してHIPAAに違反することがないようにするために極めて重要である。そのためには、解釈可能なモデルとその理由についての明確な説明が必要である。AIが生成する医療アウトプットの透明性と説明可能性を確保することは、信頼を築き、HIPAAの「説明を受ける権利」規定を遵守する上で極めて重要である。

3. その他のHIPAA規則による要求：

- **入念なアウトプットのレビューとアウトプットの結果のモニタリング**：PHI上でトレーニングされた、あるいはPHIを利用した生成AIモデルによって生成されたすべての出力は、個人を特定できる情報を含まないこと、あるいは個人を再特定する可能性がないことを確認するために、徹底的なレビューを受けなければならない。そのため、当然ながら開発時間が長くなり、モデル出力の継続的なモニタリングも複雑になる。
- **患者の同意と認可**：診断や治療の推奨のようなタスクに生成AIを使用するには、入力／出力ワークフローに複雑さを加える可能性があるとしても、患者の明確な同意と認可が必要である。
- **監査とコンプライアンス**：PHIを扱う生成AIを使用する組織は、HIPAA規則の下で他の全てのシステムに適用されるように、HIPAA規則の遵守を確実にするために、強固な監査とコンプライアンス対策を実施しなければならない。
- **リスク評価と緩和計画**：生成AI関係者は、患者のプライバシーを保護し、HIPAAコンプライアンスを維持するために、定期的な リスク評価を優先しなければならない。これらの評価は、AI/MLシステムを徹底的に評価し、潜在的なプライバシー侵害の特定と、的を絞った緩和戦略の実施を可能にするものでなければならない。

HIPAA規則は、生成AIを医療に応用する上で大きな課題となっている。こうした課題には、AIシステムの徹底的な理解、導入、および継続的な監視が必要である。これらのAIシステムを注意深く設計し、強固なプライバシー保護技術を採用し、規則を厳守することで、患者のプライバシーを保護し、責任あるコンプライアンスに則った使用を保証しながら、医療を改善する生成AIの可能性を引き出すことができる。技術革新と患者のプライバシーのバランスをとることは、この新興分野における重要な課題である。

ヘルスケアにおけるAI（生成AIを含む）とMLを取り巻くダイナミックな規制の状況は、特に生成AIシステムに対するHIPAAやその他の関連規則の進化する解釈へのコンプライアンスを確保するために、関係者による継続的な適応を必要とする。

生成AIのハルシネーションがデータプライバシー、セキュリティ、および倫理に与える影響への対処

ハルシネーションとは、AIシステムが、トレーニングされたパターンやデータに基づいて、画像、動画、テキストなど、現実的だが事実とは異なる、あるいは捏造された出力を生成する現象である。このようなハルシネーションは、データプライバシーとセキュリティをめぐる法律や規制に関する重大な懸念を引き起こす。

生成AIのハルシネーションが影響を及ぼす重大な分野のひとつに、データプライバシーがある。生成AIのモデルは、機密性の高いデータを入力すると、個人や組織の個人情報を不用意に開示する出力を生成する可能性がある。このことは、欧州のGDPRやカリフォルニア州のCCPA/CPRAのような、個人データを未認可なアクセスや開示から保護するための厳格な措置を義務付ける規制の枠組みにとって大きな課題となる。AIが生成するコンテンツの出現は、真正な情報と捏造された情報の境界線を曖昧にし、データプライバシー法を効果的に執行する努力を複雑にしている。

生成AIのハルシネーションはまた、規制環境にセキュリティリスクをもたらす。悪意のあるアクターは、捏造された画像やテキストなど、AIが生成したコンテンツをエクスプロイトすることで、個人を不正に欺いたり、操ったりする可能性がある。これはデータシステムの完全性とセキュリティに対する直接的な脅威となるため、規制当局は既存のサイバーセキュリティ規制を適応させ、AIが生成するコンテンツがもたらす特有の課題に対処する必要がある。生成AIの技術が進化し、そのモデルの能力が向上するにつれ、特に生成されたアウトプットの真正性が不確かなままであれば、セキュリティ基準への準拠を保証することはますます複雑になる可能性がある。

政策立案者や規制当局が生成AIのガバナンスに取り組む際には、生成AIのハルシネーションの倫理的意味合いにも立ち向かわなければならない。生成AIガバナンスのための規制の枠組みを形成するには、法令遵守にとどまらず、倫理的な配慮が不可欠である。ハルシネーションを持つコンテンツが個人の権利、自律性、幸福に与える潜在的な影響を含め、生成AIモデルの責任ある開発と使用をめぐる質問は、慎重な審議が必要である。生成AIガバナンスの枠組みが透明性、説明責任、および包括性といった倫理原則を優先することを確実にするため、規制イニシアチブはイノベーションの育成と社会的価値の保護のバランスを取らなければならない。

AIが生成するハルシネーションの課題に取り組むには、AIの出力を継続的に評価し、信頼できる複数の情報源から情報を検証し、コンテンツの正確性を評価する際に人間の判断を採用することが不可欠である。さらに、明確なプロンプトを提供し、十分に収集されたトレーニングデータを使用することで、最初からハルシネーションの可能性を減らすことができる。

生成AIのハルシネーションは、特にデータプライバシー、セキュリティ、および倫理の領域において、AIガバナンスのための既存の立法・規制の枠組みに挑戦している。これらの課題に対処するためには、生成AIに関連するリスクと機会を効果的に管理する包括的なガバナンス・メカニズムの開発において、政策立案者、規制当局、業界関係者、および倫理学者の協力が必要である。

DHS Policy Statement 139-07 Impact on Gen AI

- **データ入力**：米国国土安全保障省（DHS）の個人に関するデータ（それが個人を特定できる情報（PII）であるか、あるいは、匿名化されているかにかかわらず）、ソーシャルメディアコンテンツ、またはFor Official Use Only、Sensitive but Unclassified Information（現在は「Controlled Unclassified Information（CUI）」として知られている）、またはClassified情報を商用生成AIツールに入れることは禁止されている。
- **データの保持**：データ保持を制限するツールのオプションを選択し、 モデルをさらにトレーニングするために使用される入力をオプトアウトする。
- **アウトプットの見直しと活用**：これらのツールを使用して生成または修正されたすべてのコンテンツは、特に一般市民と対話する場合など公的な立場で使用する前に、正確性、関連性、データの機密性、不適切なバイアス、およびポリシーの遵守について、適切な主題の専門家（SME：Subject Matter Experts）によってレビューされるようにする。
- **意思決定**：市販の生成AIツールは 、給付裁定、資格認定、審査、または法律もしくは民事上の調査もしくは執行に関連する行動の意思決定プロセスにおいて使用することはできない。

連邦取引委員会政策提言・調査ガイダンス：

- **AI (**およびその他**)** 企業：利用規約を密かに変更することは不公平または欺瞞となる可能性がある

  データが技術革新やビジネス革新の原動力となる中、AI製品を開発する企業は、主要なデータ源としてユーザベースに依存するようになっている。しかし、これらの企業は、こうしたデータへのアクセスと、ユーザーのプライバシーを保護するという約束とのバランスを取らなければならず、より多くの顧客情報を利用するためにプライバシーポリシーを密かに緩めようとする試みは、法律に違反することになりかねない。企業は、プライバシーポリシーの条件を遡及的に変更することはできない。これは、異なる条件下でポリシーに同意した可能性のある消費者にとって、不公正かつ欺瞞的な行為となるからである。連邦取引委員会は、企業による欺瞞的で不公正なプライバシー慣行に異議を申し立ててきた歴史があり、プライバシー規制を無視し、消費者を欺こうとする企業に対して今後も行動を起こしていく。結局のところ、透明性、正直さ、および誠実さは、ユーザーとの信頼関係を確立し、法的な反響を避けたい企業にとって不可欠なものである。

- **AI**企業：プライバシーと守秘義務の遵守

  AIモデルの開発には大量のデータとリソースが必要であり、すべての企業が独自のモデルを開発できるわけではない。Model-as-a-service企業は、ユーザインターフェースやAPIを通じてサードパーティにAIモデルを提供することを支援する。これらの企業はモデルを改善するために常にデータを必要としているが、それは時としてユーザデータやプライバシーを保護する義務と相反することがある。連邦取引委員会（FTC）は、顧客データやプライバシーの保護を怠った企業や、顧客データを悪用した企業に対して法律を執行している。Model-as-a-service 企業は、どこで製造されて

いるかに関係なく、その公約を遵守しなければならず、顧客を欺いたり、不公正な競争を行ったりしないようにしなければならない。AIモデルのトレーニングとデプロイにおける虚偽表示、重大な省略、およびデータの不正使用は、競争にリスクをもたらす可能性がある。消費者のプライバシー権を侵害したり、不公正な競争方法に関与したりするModel-as-a-service企業は、独占禁止法と消費者保護法の両方で責任を問われる可能性がある。

連邦政府機関による人工知能の使用におけるガバナンス、イノベーション、リスク管理を推進するOMBポリシー

人工知能のリスクを軽減し、その利点を活用するための政府全体の方針が、Kamala Harris副大統領によって発表された。この方針は、AIの安全・安心の強化、公平性と公民権の促進、アメリカのAIイノベーションの促進を目的としたバイデン大統領のAI大統領令（下記参照）の一環として発表された。新方針には、米国人の権利や安全に影響を与える可能性のある方法でAIを使用する連邦政府機関に対する具体的な保護措置が含まれている。その目的は、責任あるAIイノベーションの障壁を取り除き、AI労働力の拡大とスキルアップを図り、AIガバナンスを強化することである。政府はこの発表により、AIを活用する連邦政府機関全体の透明性、説明責任、および権利と安全の保護を促進する。人工知能（AI）のリスクを軽減し、その利点を活用するための政府全体の政策の主なハイライトは以下の通りである。

- **AIのための具体的なセーフガード**：2024年12月1日までに、連邦政府機関は、アメリカ人の権利や安全に影響を与える可能性のあるAIを使用する際、具体的なセーフガードの導入を義務付けられる。これらのセーフガードには、AIが一般市民に与える影響の評価、テスト、監視、アルゴリズムによる差別のリスクの軽減、および政府によるAI利用の透明性の提供などが含まれる。
- **医療における人間による監視**：AIが連邦政府の医療システムで重要な診断の決定をサポートするために使用される場合、ツールの結果を検証し、医療アクセスにおける格差を避けるために、人間がプロセスを監督する。
- **不正検知における人間による監視**：政府サービスの不正を検知するためにAIが使用される場合、影響力のある決定を人間が監視し、影響を受ける個人はAIの被害に対する救済を求める機会がある。
- **AI利用の透明性**：連邦政府機関は 、AIのユースケースの年次目録を拡大公開し、機密性の高いユースケースに関する指標を報告し、AIの適用除外を正当な理由とともに国民に通知し、政府が所有するAIのコード、モデル、およびデータを公開することで、AI利用の透明性を向上させることが求められている。
- **責任あるAIイノベーション**：この政策は、連邦政府機関による責任あるAIイノベーションの障壁を取り除くことを目的としている。気候危機への対応、公衆衛生の向上、および公共の安全の保護におけるAIの応用例を取り上げている。
- **AI人材の育成**：このガイダンスは、AI人材を拡大し、スキルアップするよう機関に指示している。
- **AIガバナンスの強化**：この方針は、連邦政府機関に対し、機関全体のAI利用を調整するチーフAIオフィサーを指名し、機関内のAI利用を管理するAIガバナンス委員会を設置することを求めている。
- **AIの使用中止**：指定されたセーフガードを適用できない場合、そうすることで安全や権利全体に対するリスクが高まる、あるいは重要な省庁業務に受け入れがたい支障が生じるという理由を省庁の指導者が正当化しない限り、AIシステムの使用を中止しなければならない。

バイデン大統領による人工知能の安全、セキュア、および信頼できる開発と利用に関する大統領令

バイデン大統領が2023年10月に発表した「安全、セキュア、および信頼できる人工知能（AI）に関する大統領令」は、社会的な懸念に対処し、責任あるAIの実践を確立するための画期的な取り組みである。

この命令は、データプライバシー、倫理、人材育成、および国際協力などの主要分野を含む、AIの安全、セキュア、倫理的な開発と利用の確保に重点を置いている。これは、AI技術の責任ある開発と配備を導くためのガイドラインとベストプラクティスを作成するための計画の概要を示している。この計画には、国立標準技術研究所（NIST）、全米科学財団（NSF）、商務省（DOC）といった複数の政府機関に、既存のフレームワークや以下のようなトピックに関するリソースやベストプラクティスの開発を任せることも含まれている。

- アルゴリズムの公平性とバイアス
- AIモデルの説明可能性と解釈可能性
- 標準化されたテストと評価方法

具体的な規制の詳細はまだこれからだが、この命令は、信頼できるAIのための強固な枠組みを構築するという政府のコミットメントを示すものである。バイデン大統領の大統領令は、AIを取り巻く法的・規制的状況を再定義するものではないが、倫理的で説明責任のある利用の重要性を強調し、データのライフサイクル全体を通じて、データプライバシー、セキュリティコントロール、およびサイバーセキュリティなどの懸念事項に対処している。

具体的な規制を定めるものではないが、「安全、セキュア、および信頼できる人工知能に関する大統領令（Safe, Secure, and Trustworthy AI Executive Order）」は、データプライバシー、倫理、人材育成、および国際協力に焦点を当てることで、社会的な懸念に対処し、責任あるAIの開発と利用への包括的なアプローチの基礎を築くものである。

連邦政府によるAI規制がないため、州や地方によってさまざまな規制が提案され、制定されるという複雑な状況になっている。BCLPlawの「米国州別AI法 スナップショット」で強調されているように、この規制のつぎはぎ構造は重要な懸念を生む。

無差別と公平性

斬新なコンテンツを生み出し、意思決定に影響を与える生成AIの能力は、差別や公正さに関する重大な懸念を引き起こし、法律や規制による精査を促している。差別禁止法や規制が、生成AIの設計、デプロイ、および使用方法にどのような影響を与えるかを確認してみよう。

1. 既存の差別禁止法および規制の一部

AIのアルゴリズムや意思決定プロセスにおける保護特性に基づく差別に対処するための現行法および提案されている法律の概要：

- 公民権法第7条 **(米国、1964年）**：人種、肌の色、宗教、性別、国籍による雇用差別を禁止する。採用、昇進、業績評価で使用されるAIシステム（生成AIを含む）は、保護されるグループに対する偏見を永続させる場合、タイトルVIIの下で精査に直面する可能性がある。
- 雇用機会均等と公民権に関する法律と権限（米国）：タイトルVIIの保護を年齢と障害に拡大する。これらの特徴に基づくアルゴリズムのバイアスも禁止されている。

  EEOCの技術支援文書は、雇用やその他の雇用判断に使用されるAIを含むソフトウェアが、EEOCが執行する連邦公民権法を遵守することを保証する「人工知能とアルゴリズムによる公正イニシアチブ」の一環である。

  さらに、2008年遺伝情報無差別法（The Genetic Information Nondiscrimination Act of 2008）は、雇用や健康保険における遺伝情報に基づく差別を禁止する連邦法である。これは、アルゴリズムによる意思決定（AIシステムによる意思決定を含む）を直接規制するものではないが、雇用決定における遺伝情報に基づく差別を禁止している。生成AIシステムを使用する企業は、そのシステムが公正で偏りがなく、遺伝情報を含むあらゆる機微情報に基づく差別的慣行を永続させないことを保証する責任が依然としてある。

- 公正住宅法（米国）：タイトルVIIと同じ保護特性に基づく住宅差別を禁止する。入居審査や住宅ローンの承認に使用されるAI搭載ツールは、これらの保護に準拠しなければならない。
- 信用機会均等法**（米国）**：人種、肌の色、宗教、国籍、性別、配偶者の有無、年齢、障害による信用差別の禁止。AIによる信用スコアリングモデルは、潜在的な差別的影響について慎重に評価されなければならない。
- タイトルVI、タイトルIX、セクション504のようないくつかの連邦公民権法は、人種、肌の色、国籍、性別、障害、年齢に基づく教育現場での差別を禁止している。学校や教育機関は、機械学習やAIのような技術を含む実践が、上記の保護特性に基づいて生徒を差別しないことを保証するために、これらの法律を遵守しなければならない。
- 欧州一般データ保護規則（GDPR）(EU)：個人データへのアクセス、修正、消去の権利を個人に付与する。これは、差別的な結果を避けるために、生成AIシステムが個人情報を収集し使用する方法に影響する。データコントローラに対しては、差別的なプロファイリングや自動的な意思決定に対するセーフガードの実施が義務付けられている。さらに、CCPA/CPRAは、プライバシー権を行使する消費者を差別することを禁じている。

- Algorithmic Accountability Act (US, 2019-2020):偏見監査の連邦基準を確立し、政府機関や企業が使用するAIアルゴリズムの公正性、説明責任、および透明性を評価することを目的とする。
- European Union's Artificial Intelligence Act (EU AI Act) (2024):偏見や差別への対処を含め、リスクの高いAIアプリケーションに特定の要件を課す。
- New York City Bias in Algorithms Act (US, 2021):市の機関が使用するAIアルゴリズムについて、保護されるべき特性に基づく潜在的な偏りがないか監査することを義務付ける。
- California Automated Decision Making Act (US, 2023) / New York Automated Employment Decision Tools Act (US, 2023):両者とも、消費者に大きな影響を与える自動意思決定ツールを使用する場合、事業者に通知と説明を提供することを求めている。
- CCPA/CPRAは、プライバシー権を行使する消費者に対する差別を禁止している。これは、固有のバイアスを含むデータセットでトレーニングされた生成AIモデルに潜在的な課題をもたらす可能性がある。このようなバイアスを緩和し、非差別的なアウトプットを保証することは、CCPA/CPRAのもとでは極めて重要である。
- Americans with Disabilities Act (ADA): これは、障害を持つ人々への配慮を義務付ける一連の基準と規則である。一般市民と対話するAIシステムは、アクセシビリティに関するADAガイドラインに準拠する必要がある。
- Fair Credit Reporting Act (FCRA):この法律は、消費者情報の収集と使用方法を規制している。金融業界で使用されるAIモデル（ローン決定など）は、意思決定における不当なバイアスを避けるため、FCRAの遵守を保証する必要がある。

偏ったAIアルゴリズムによる差別的な雇用慣行を主張する米国での訴訟や、EUのGDPR裁定に基づくAI搭載の顔認識システムに対する監視の強化など、最近の事例では、潜在的な偏見や差別的プロファイリング、および差別禁止法遵守の必要性に対する懸念が浮き彫りになっている。

これらの最近の例は、AI採用における偏見の可能性を浮き彫りにしており、候補者を選ぶためのツールが差別の申し立てに直面した例もある：

- ニュースの中の執行：AIによる差別をめぐるEEOC初の訴訟、2023年：「iTutorGroupは55歳以上の応募者のうち200人以上を年齢を理由に採用しなかったとしている」。告発者は、本当の生年月日で応募し、即座に不採用となり、翌日、より新しい生年月日で応募し、面接のオファーを受けたと主張した。その結果、「2023年1月、EEOCは2023年から2027年までの戦略的執行計画の草案を発表した。この草案は、採用から従業員の業績管理まで、雇用ライフサイクル全体におけるAIの差別的使用に対するEEOCの明確な重点を明らかにしている。」
- AI採用における差別と偏見：ケーススタディ、2023年：この事例は、AI採用における偏見の実例である。ある銀行が採用候補者を絞り込むためのAIツールが差別的であることが判明した。このケースは、重要な法的考察を提起し、採用プロセスにAIツールを導入する際に潜在的なバイアスに注意することの重要な必要性を強調している。
- AIを使って従業員のメッセージを監視、2024年：この記事では、大企業が従業員のメッセージを監視するためにAIサービスをどのように利用しているのか紹介している。単に感情を分析するだけでなく、テキストや画像を「いじめ、ハラスメント、差別、コンプライアンス違反、ポルノ、ヌード、その他の行動」、さらにはさまざまな属性（年齢層、場所など）が企業の取り組みにどのように反応するかまで評価することができる。データの匿名化のようなプライバシー保護技術は適用されるが、そのようなやり方はプライバシーの権利や言論の自由に対する懸念を引き起こす。

プライバシーの侵害であり、オープンなコミュニケーションを阻害しかねないという意見がある一方で、潜在的な課題を特定し、企業を保護するための意思決定を強化する方法だと考える人もいる。法的な状況は依然として不透明で、このような行為に対する規制や社会的なハードルの可能性を示唆している。

2. 規制上の課題

現在の法的枠組みは、いくつかの制限のために、生成AIにおける非差別と公平性に対処するのに苦労している：

- **適用ギャップ**：既存の法律は、複雑なAIシステムへの対応に苦慮しており、「差別」のような概念がアルゴリズムやデータにどのように反映されるのか明確ではない。
- **バイアスを証明することの難しさ**：不透明なAIシステムは、差別的意図や影響を特定し証明することを困難にし、さらにシステム内の要因が相互に関連しているために複雑になっている。
- **執行の課題**：限られた資源と専門知識が効果的な捜査と執行を妨げており、AI開発の世界的な性質によってさらに複雑になっている。
- **イノベーション対規制**：急速に進化するAI技術は、現在の法的枠組みを凌駕し、不確実性を生み出し、イノベーションと倫理的配慮の間の微妙なバランスを必要とする。
- **公平性の定義と実行**：AIにおける公平性の実現は多面的である。解釈の違いや公平性の原則間の潜在的な対立があるため、正確な定義は複雑である。公平性を確保するための措置を実施することは、しばしば技術的に大きな困難を伴い、多大な資源を必要とする。
- **解釈の複雑さ**：AIモデル、特にディープラーニングモデルは、信じられないほどに非常に複雑である。何百万ものパラメータで構成されることもあり、入力データがどのように出力予測に変換されるかを理解することは困難である。このような変化を正確に反映した説明を作成することは、自明なことではなく、多大な計算資源と時間を必要とする。
- **正確さと説明しやすさのトレードオフ**：ニューラルネットワークのような精度の高いモデルは、解釈しにくいことが多い。一方、線形回帰や決定木のような、より単純で解釈しやすいモデルは、複雑なタスクではそれほどうまく機能しないかもしれない。このトレードオフのバランスをとりながら、正確かつ説明可能なモデルを開発するのは、難しいプロセスである。生成AIは、説明しにくさの代わりに精度の高さを受け入れている最たる例である。
- **標準化された技術の欠如**：AIの意思決定を説明する手法は、（Local Interpretable Model-Agnostic Explanations（LIME）、SHapley Additive exPlanations（SHAP）などいくつかあるが、万能な手法はない。適切な手法はモデルの種類や特定のアプリケーションによって異なるため、説明可能なAIの開発にはカスタムソリューションが必要になることが多い。
- **説明の検証**：説明可能なAI技術によって生成された説明が 、モデルの意思決定プロセスを正確に反映しているかどうかを検証することは、それ自体が複雑な作業である。この検証プロセスには時間と計算コストがかかる。

今日、現在の法的枠組みは、急速に発展している生成AIの分野において、無差別と公平性に関する懸念に対処するための十分な設備が整っていない。このギャップを埋めるためには、国民の理解、合意形成、および適応可能な規制の確立が必要である。

3. 規制の焦点とテクニック

生成AIのための規制の枠組みは、開発と配備のライフサイクルの様々な段階にわたって、バイアスの緩和と公平性に取り組むべきである。いくつかの規制上の考慮事項と、それに対応するバイアスと公平性に対処するための手法を以下に挙げる。

- **データのデバイアス**：
    - **規制の焦点**：データプライバシー規制を活用することで、責任あるデータ収集と利用を実践することができる。特定の規制は、機密データのデータ・デバイアシング技術を義務付けたり、データ処理パイプラインの透明性を要求するかもしれない。
    - **テクニック**：データクリーニング（偏ったアノテーションの削除、矛盾の特定と修正など）、データ拡張（代表性を向上させるための合成データの生成など）、データの重み付け（代表性の低いグループからのサンプルにより高い重みを割り当てるなど）。いわゆる「安全な」または「サニタイズ済み」（専門家によっては「前処理済み」または「バイアス除去済み」を好む）のデータセットを使用することは、出発点となり得るが、組織は、不完全なバイアスの緩和、限られたスコープ、および潜在的な情報損失などの限界を考慮すべきである。IBMのような企業は、AI開発の初期段階における足がかりとしてこのようなデータセットを提供しており、参考文献は必要に応じてオンラインで見つけることができる（例えば、ウィキペディア).
    - **適用される規則**：関連するのは、データ処理の透明性と責任あるデータ収集の実践を規定する欧州一般データ保護規則（EU）、個人データへのアクセス、削除、売却のオプトアウトの権利を個人に提供するカリフォルニア州消費者プライバシー法(CCPA)、生成AIモデルとAI出力のトレーニングのためのデータ使用を規定する可能性がある、標準化された文書化フレームワークとしてのモデルカード文書化フレームワーク（Hugging Face）であり、"特にバイアス、リスク、および限界のセクションに重点を置いている"。

- **アルゴリズムの透明性**：
    - **規制の焦点**：透明性に関する規制は、特に影響度の高いアプリケーションにおいて、開発者にモデル出力の説明を提供することを求めることができる。これには、標準化された説明フォーマットや、独立した分析のための関連データのサブセットへのアクセスが含まれる。
    - **テクニック**：モデルの意思決定プロセスを解明する説明可能なAI（XAI）手法（顕著性マップ（saliency maps）、反実仮想（counterfactuals）など）。
    - **適用される規則**：関連するものとして、欧州連合（EU）のAI法（EU AI Act）がある。これは、「高リスク」AIシステムに透明性と説明可能性を求め、特定の説明可能なAI技術の使用を義務付ける可能性があるもので、NISTの「説明可能な人工知能の4原則」（2021年）、「AIモデルの共有理解の価値」を提唱するモデルカード文書化フレームワーク（Google Research）などがある。

- **人間の監視とフィードバック**：

    - **規制の焦点**：重要な決定やデリケートな領域については、規制によって特定の人間による監視メカニズムが求められるかもしれない。これには、査読者の資格、査読プロトコルの定義、または特定されたバイアスの報告義務などが含まれる。
    - **テクニック**：Human-in-the-loopシステム、人間によるフィードバックループを用いた能動的学習、データ対象者の明示的な同意、モデル出力の人間によるレビュー。
    - **適用される規則**：FDAが提案したTPLCアプローチ（2021年）では、製造業者が "AI/MLデバイスを監視し、リスク管理アプローチや、アルゴリズム変更の開発、検証、実行において「既存デバイスへのソフトウェア変更の510(k)申請の時期を決定する」ガイダンス18に概説されたその他のアプローチを取り入れる "ように、人間による監視を提唱している。

- **AI開発における多様性、公平性、包括性（DE&I）**：
    - **規制の焦点**：AIチーム内での公正な雇用と開発慣行を確保するために、平等法と無差別法を適用することができる。規制によって、開発チームの多様性指標が義務付けられたり、デプロイ前にバイアスによる影響評価が求められたりするかもしれない。
    - **テクニック**：開発チーム内にダイバーシティマインドを構築し、 多様性、公平性、公正性、およびインクルージョンの原則を設計やテスト段階に取り入れ 、バイアス監査や影響評価を実施する。
    - **適用される規則**：AIに関する具体的なDE&I規制はまだ発展途上だが、組織はAIシステムが公正かつ公平であることを保証する倫理基準を積極的に採用し、「DEIを企業のAI戦略に組み込む」（Harvard Business Review、2024年）機会を活用しなければならない。業界のガイダンスでは、「AIは、コミュニティに力を与え、社会に利益をもたらすよう、そのアプローチにおいて倫理的かつ衡平でなければならない」（世界経済フォーラム、2022年）と強調され、バイアスや差別を避けている。

- **アルゴリズムの透明性と説明可能性**：AIの決定の透明性と説明可能性（例：説明可能なAIイニシアティブ）の要件を特定する。特にリスクの高いアプリケーションにおいて、AIによる意思決定の説明可能性を求める規制と、それが組織のアプローチにどのような影響を与えるのかを検討する。関連文書の一部は以下である：
    - Algorithmic Accountability Act, 2021-2022:いくつかの州で提案されているこれらの法案は、「AIや自動化システムによってすでに生じている問題に対応する」ために、重要な判断に使用されるAIシステムの透明性を確保し、監査を保証することを目的としている。
    - Algorithmic Accountability Act, 2023-2024:現在導入段階にある「Algorithmic Accountability Act」（2023年9月）は、AIシステムの責任ある開発と利用の枠組みを確立することを目的としている。具体的な内容はまだ策定中だが、以下にこの法律が一般的にどのような分野に重点を置く可能性があるかについて、我々の理解を示す。
        - **透明性と説明可能性**：AIシステムがどのように意思決定を行うのか、開発者に説明を義務付け、国民の理解と信頼を向上させる。
        - **データプライバシーとセキュリティ**：AIシステムのトレーニングや導入に使用される個人データを保護するためのセーフガードを確立する。
        - **アルゴリズムの公平性とバイアス**：データとアルゴリズムのバイアスに対処することで、差別的な結果の可能性を軽減する。
        - **リスクの評価と軽減**：安全性、セキュリティ、公平性への懸念など、AIに関連する潜在的なリスクを特定し対処する。

法案が立法プロセスを経るにつれて、AIと生成AIを管理するための具体的な規制に関する詳細が明らかになるだろう。

新たな規制の枠組み、基準、ガイドライン

The AI Bill of Rights (White House Blueprint), 2023: この拘束力のないガイドラインは、AIシステムが差別なく公平に使用される必要性を強調している。アルゴリズムによる差別や有害なバイアスに対するセーフガードを推奨している。

人工知能に関する国連グローバル決議：安全で信頼でき、人間中心のAIを支援する国連の決議は、加盟国に対し、安全で信頼でき、人間中心で透明性の高いAIの開発と利用を促進するよう求めている。また、AIが人権と基本的自由を尊重し、バイアスや差別のない方法で使用されることの重要性を強調している。さらに決議では、AIの開発と使用に関する国際的な規範と基準を策定するために、加盟国が協力することを奨励している。AIに関する国連決議の主なハイライトは以下の通りである。

- 安全で信頼でき、人間中心のAIの開発と利用を促進するよう加盟国に奨励する。
- AIは、人権と基本的自由を尊重しつつ、透明性があり、バイアスや差別のない方法で使用される必要性を強調する。
- 加盟国に対し、AIの開発と利用に関する国際的な規範と基準の策定において、互いに協力・協調するよう促す。
- AIが社会全体に利益をもたらすことを確実にするため、加盟国がAIの開発と利用におけるベストプラクティスと経験を共有することを奨励する。
- 責任ある倫理的な方法でAIの開発と利用を導くために、政府、市民社会、および産業界を含む様々なセクターの利害関係者との継続的な対話と関与を呼びかける。

National Institute of Standards and Technology (NIST) AI Risk Management Framework: このフレームワークは、偏見や差別に関連するリスクを含め、AIシステムに関連するリスクを特定、管理、および軽減することを組織が支援することを目的としている。AIの開発にDEIを考慮することを奨励している。
フレームワークの詳細については、Artificial Intelligence Risk Management Framework (AI RMF 1.0)を参照。

効果的なAI規制は、以下のように、標準化、説明責任、および国際協力という3つの重要な側面を促進すべきである：

- **標準化**：これには、ACMライブラリー（2019年）の論文 "Model Cards for Model Reporting "で提案されているような、"Model Cards "の標準化されたフォーマットの採用など、バイアスを検出、防止、緩和するための共通の方法の確立が含まれる。
- **説明責任**：責任ある開発と導入にインセンティブを与えるためには、責任と説明責任に関する明確な枠組みが必要である。
- **国際協力**：国境を越えた一貫した効果的なアプローチは、AI規制に関する国際協力を通じて達成することができる。

AIの倫理的、透明性、信頼性のある設計、開発、配備、および運用を奨励するためのフレームワーク、ガイドラインやリソースが存在する。以下はそのいくつかの例である。

- IIA AI監査フレームワークは、AIシステムの信頼性を評価するための包括的なアプローチを提供する。それは、ガバナンス、倫理、コントロール、そして人的要因という4つの主要分野に焦点を当てている。3つの包括的な構成要素（AI戦略、ガバナンス、人的要因）と7つの要素（サイバーレジリエンス、AIコンピテンシー、データ品質、データアーキテクチャとインフラストラクチャ、パフォーマンス測定、倫理、ブラックボックス）の詳細については、フレームワークの文書を参照。
- IBMの "Trusted AI" Ethicsは、AIが倫理的かつ透明性をもって設計、開発、導入、および運用されることを保証するためのガイドラインを提供している。
- マイクロソフトの "Responsible AI Practices "は、信頼できるAIの開発と利用のためのガイドラインと原則である。
- AWSの "Core dimensions of Responsible AI "は、AIを安全かつ責任を持って開発するためのガイドラインと原則であり、教育、科学、および顧客を優先する人間中心のアプローチをとっている。
- グーグルの「責任あるAIの実践と原則」は、人間中心設計のアプローチを用いて、責任あるAIの開発と利用を導くように設計されている。
- アランチューリング研究所による「Understanding AI Ethics and Safety（AIの倫理と安全性を理解する）」ガイドは、AI倫理、潜在的な利益、課題、およびAIに関連する倫理的懸念を強調したケーススタディに関する入門的なリソースとして役立つ。
- パートナーシップ・オン・AIによるAIインシデントデータベースは、AIシステムが意図しない危害を引き起こした実例のリポジトリであり、IEEEによる倫理的整合設計（EAD）ガイドラインは、倫理的で透明性が高く、信頼できるAIシステムを設計するための推奨事項とフレームワークを提供している。

これらのリソースは、ユーザーに対して透明性を保ちながら、AIの倫理的利用を促進するための推奨事項を提供している。この資料では、AIの潜在的な利点、AIの導入に伴う課題、倫理的な懸念やAIシステムが意図しない危害をもたらした事件に関するケーススタディなどのトピックも取り上げている。

The OWASP Top 10 for Large Language Model Applications projectは、大規模言語モデル（LLM）を配備・管理する際の潜在的なセキュリティリスクについて、開発者、設計者、アーキテクト、管理者、組織を教育することを目的としたイニシアチブである。このプロジェクトは、LLMアプリケーションでよく見られる最も重大な脆弱性のトップ10を包括的にリストアップし、潜在的な影響、エクスプロイトの容易さ、および実世界のアプリケーションにおける普及率を強調する。脆弱性には、プロンプトインジェクション、機密情報漏洩（データ漏洩）、セキュアでないプラグイン設計、および不正なコード実行／モード盗用などがある。このプロジェクトの最終目標は、これらの脆弱性に対する認識を高め、改善策を提案し、最終的にはLLMアプリケーションのセキュリティ態勢を改善することである。

同様に、the OWASP Machine Learning Security Top 10 project（現在草案中）は、機械学習システムのセキュリティ課題トップ10の包括的な概要を提供している。このプロジェクトは、開発者、設計者、アーキテクト、管理者、および組織に対し、機械学習システムを開発・導入する際の潜在的なセキュリティリスクについて教育することを目的としている。このプロジェクトは、機械学習システムでしばしば見られる最も重大な脆弱性の包括的なリストを提供し、その潜在的な影響、エクスプロイトの容易さ、実世界のア

プリケーションにおける蔓延を強調する。脆弱性の例としては、敵対的攻撃、データポイズニング、およびモデル盗用などがある。このプロジェクトの最終目標は、これらの脆弱性に対する認識を高め、改善策を提案し、最終的に機械学習システムのセキュリティ体制を改善することである。

責任あるAIの実践を支援するために、いくつかの重要な規格が策定されている。いくつか例を挙げてみよう。

- ISO/IEC 42001:2023は、AIシステムのマネジメントシステムの枠組みを提供する規格である。この規格は、AIシステムの開発、デプロイ、およびメンテナンスを含むライフサイクルを管理するための体系的なアプローチを概説している。リスク、倫理的、社会的、および法的配慮を考慮した、AIシステムの機能的な管理システムの確立と導入を支援する。この基準は、さまざまな利害関係者のニーズを考慮して開発された、透明性が高く説明責任を果たせるAIシステムの重要性を強調している。また、人権やプライバシーの尊重を含む倫理原則を遵守し、責任あるAIの実践とガバナンスを実施することも奨励している。
- ISO/IEC 23053:2022は、機械学習（ML）を用いた人工知能（AI）システムの開発、導入、および管理のためのフレームワークを提供する規格である。この基準では、データの収集と処理、モデルの訓練と検証、システムのデプロイ、継続的な監視と保守など、AIシステムの開発とデプロイにおける主要な活動を概説するプロセスモデルを定めている。同基準は、AIの開発と展開における倫理的で責任あるアプローチの重要性を強調している。潜在的なリスクの特定とその軽減を含め、リスク評価とリスク管理に関するガイダンスを提供する。同基準はまた、AIシステムにおける信頼性、透明性、および説明責任に関する課題を取り上げ、AI出力の説明可能性と解釈可能性の必要性を強調している。

特定の業界におけるAIガバナンスとコンプライアンスの詳細については、CSAのAI Resilience: A Revolutionary Benchmarking Model for AI Safety 資料を参照。

安全性、責任、説明責任

創造的なテキストから驚くほどリアルな画像や映像まで、自律的にアウトプットを生成する能力を備えた生成AIの急速な発展は、間違いなく技術的驚異の新時代を到来させた。しかし、この進歩は、安全性、責任、および説明責任に関する重大な懸念に対処することを迫るものである。Geminiによる偏ったビジュアルの生成（Google Blog、2024年2月）や、カナダ航空のボットによる誤った払い戻し情報の提供（New York Post、2024年2月）といった最近の例は、AIの不品行がもたらす非常に現実的な結果を浮き彫りにしている。物事がうまくいかなくなったとき、誰が責任を取るのか、そして生成AIによる不適切な、あるいは危険な結果の矢面に立つのは誰だろうか。現在、この強力なテクノロジーの責任ある使用を保証するために必要な法的ガバナンスと効果的な枠組みがあるのだろうか？政策立案者と産業界のリーダーは、生成AIの責任ある利用のための国際基準を確立するために、どのように協力できるだろうか？生成AIが悪意を持って使用される可能性を制限するために、どのような技術的な安全策を講じることができるだろうか？
生成AIに関連する潜在的なリスクを軽減するには、以下のような多方面からのアプローチが必要である。

- **業界標準**：生成AIの開発、配備、および使用に関する明確で包括的なガイドラインの策定。これらの基準は、公平性への配慮、偏見の緩和、および責任あるデータの取り扱いを優先しなければならない。

- **法的枠組み**：AIが生成したコンテンツによって被害が生じた場合の責任の帰属に関する複雑な課題に対処する法的枠組みの開発。説明責任のバランスをとり、責任あるイノベーションを可能にするためには、慎重な検討が必要である。
- **組織のリスク管理戦略**：生成AIの活用に伴うリスクを効果的に評価・管理するためのツールと知識を組織に提供する。これには、強固な保護措置と責任ある使用方針の実施が含まれる。

生成AIの責任、リスク、および安全性をめぐる考察

生成AIは、その潜在的な利点にもかかわらず、固有のリスクを伴い。以下は、懸念される主な分野である。

1. 生成AIの失敗に伴う潜在的な賠償責任リスク

- **偏見と差別**：偏ったデータで訓練された生成AIモデルは、生成されたコンテンツに有害なステレオタイプを永続させ、差別的な結果をもたらす可能性がある。そのような法的課題の例としては、不公正な住宅、雇用／雇用慣行、商品の推奨、またはローンの申請／承認に関するものがある。
- **プライバシー侵害**：生成AIシステムはしばしば膨大なデータへのアクセスを必要とするため、ユーザーのプライバシーや機密情報の悪用の可能性についての懸念が生じる。トレーニングデータに使用されている機密情報をうっかり漏らしてしまい、プライバシーの侵害や法的措置につながる可能性がある。
- **安全とセキュリティの課題**：ヘルスケアや自律走行車のような重要な分野では、生成AIの誤作動が安全上の危険、事故責任の帰属、あるいは身体的危害につながる可能性がある。
- **誤報と悪意ある利用**：生成AIは、ディープフェイクコンテンツの生成、コンテンツの操作、フェイクニュースの生成、偽情報の拡散のためにエクスプロイトされる可能性があり、国民の信頼と民主的な言論に対する脅威となる。これは、名誉毀損や詐欺に関する法的懸念を引き起こす可能性がある。

2. 責任分担の法的枠組み

AIシステム（特に生成AI）によって引き起こされた損害に対する責任を決定し、割り当てることは、複雑な法的課題を提示する。現在の枠組みは、AI特有の特性への対応に苦慮することが多く、不確実性をもたらしている。製造物責任法、過失責任法、データプライバシー法などの伝統的な法原則は、特定の文脈や法域では適用可能かもしれないが、AI技術のダイナミックな性質は、新たな法的枠組みの開発を必要とする。

AI法制に特化したアルゴリズムの透明性に関する法律など、新たな規制が様々な地域で具体化し始めている。これらの枠組みは、偏見、説明責任、透明性、および公平性に関する課題に重点を置き、AIシステムがもたらす独自の課題に対処することを目的としている。しかし、これらの規制の実施とスコープは、法域によって、さらにはユースケースによって大きく異なる可能性がある。

OECDのAI原則のような国際的イニシアチブは、世界中で責任あるAIの開発と展開を促進するための指針を提供している。これらの原則は、AIシステムにおける透明性、説明責任、および包括性などの基本的価値を支持し、倫理的で持続可能なAIイノベーションの礎石となる。拘束力はない

が、将来のAI政策や規制を形成するための基礎的枠組みを形成するものである。

こうした努力にもかかわらず、AIの責任をめぐる法的状況の把握は依然として複雑である。法解釈や適用可能性は文脈に大きく依存するため、個々の事例や管轄区域を徹底的に分析する必要がある。したがって、AI関連の取り組みにおいてコンプライアンスを確保し、リスクを軽減するためには、AI法に精通した法律専門家の指導を仰ぐことが不可欠である。

明確で予測可能な法的枠組みを確立することは、ユーザーの安全と社会の幸福を確保しながらイノベーションを促進するために極めて重要である。

3. 保険

AI関連リスクの軽減は、AIシステムによって引き起こされる潜在的な損害の金銭的負担を分散させるために、AI賠償責任保険に特化することで達成できる。

生成AIのためのハルシネーション保険

ハルシネーション保険は、生成AIが私たちの生活やビジネスのさまざまな側面にますます統合されるにつれて出現している斬新なコンセプトである。その名が示すように、この保険は「ハルシネーション」、つまり生成AIシステムによって生成されたアウトプットに含まれる誤った情報、偏見、および／または事実誤認によって引き起こされる経済的・風評的ダメージを軽減することを目的としている。

この保険は、以下のような生成AIのハルシネーションがもたらす可能性のある結果に対する経済的な保護を提供しようとするものである：

- **財務上の損失**：これには、エラーの是正に関連するコスト、弁護士費用、風評被害、および不正確な出力や誤解を招く出力から生じるビジネス機会の損失などが含まれる。
- **規制上の罰則および手数料**：AIが生成したアウトプットが規制や倫理ガイドラインに違反した場合 、当局が課す制裁金や罰則を保険でカバーできる可能性がある。
- **サイバーセキュリティの侵害**：生成AIシステムが侵害されたり、機密情報が流出した場合、保険は修復や潜在的な法的影響を支援できる。

ハルシネーション保険の出現にはいくつかの要因がある：

- **高まる生成AIへの依存**：企業が様々な分野で生成AIを活用するようになるにつれ、リスク軽減戦略の必要性がより重要になる。
- **高価な結果を招く可能性**：AIのハルシネーションは経済的および風評的に大きな損害をもたらす可能性があり、保険はリスク管理のための貴重なツールとなっている。
- **進化する規制の状況**：AIの利用をめぐる規制が進展する中、保険はコンプライアンスを確保し、法的リスクを軽減することができる。

まだ初期段階だが、ハルシネーション保険は他のタイプの保険と同様の機能を果たすと期待されている。企業や個人は、特定のリスクに対する補償と引き換えに保険料を支払う。対象となる特定のリスクや、金銭的補償やリスク軽減戦略の焦点は、生成AIの適用や被保険者のニーズによって異なる。

ハルシネーション保険の正確な形態や仕組みはまだ定義されていないが、生成AIの導入が進むにつれ、この保険タイプは保険業界においてより目立つ存在になると予想される。ハルシネーション保険は、特に生成AIに大きく依存している企業にとっては、他の形態の賠償責任保険やサイバー保険と同様に、標準的なビジネス必需品になる可能性があると考える専門家もいる。

技術的な観点からは、ハルシネーション保険は特効薬ではないことに注意する必要がある。生成AIシステムの責任ある開発とデプロイは、ユーザーの意識と批判的思考／人間の監視とともに、リスクを最小限に抑える上で極めて重要な要素であることに変わりはない。それでも、この斬新な保険商品は、AIの世界に踏み出そうとしている企業にとって切望されていた保護を提供する可能性を秘めており、信頼を育み、この強力なテクノロジーに関連する潜在的なリスクを軽減する。

知的財産

生成AIは、所有権、著作権、および説明責任に関する複雑な知的財産権に関する議論を提起しているが、現状では明確な法的枠組みがない。課題としては、所有権の曖昧さ、トレーニングデータによる著作権侵害の可能性、およびアウトプットに対する責任の所在の不明確さなどが挙げられる。
将来の規制の形成、技術革新の促進、および新たな知的財産モデルの探求にチャンスがある。この急速に進化する状況を乗り切り、十分な情報に基づいた意思決定を行うためには、常に立法や裁判の最新情報を入手することが重要である。2020年からの 国連の人工知能（AI）に関する活動 レポートは 、 世界知的所有権機関（WIPO）の調査（2019年）と "1950年代以降に発表された34万件以上のAI関連特許出願と160万件の科学論文の分析 " に基づき、"AI技術には知的財産権（IP）に対する大きな需要がある "と明確に概説している。

以下に、現在の知的財産の枠組みが、AIが生成したモデル、アルゴリズム、およびデータをどのように扱おうとしているのか、ライセンシングと保護に関する考慮点を強調しながら説明する。

1. 著作権、発明権、所有権

特許、著作権、および企業秘密など、既存の知的財産（IP）の枠組みは、人間のクリエイターを念頭に構築されている。例えば、アメリカの著作権局は、AIによってのみ創作された作品の著作権を否定している。しかし、AIが支援する創作物は、人間の実質的な関与があれば、著作権が認められる。しかし、この概念にはグレーゾーンがある。著作権保護に必要な人間の創造性のレベルは依然として不明確であり、裁判の中で議論されることになるだろう。

著作権局に見られるように、アメリカでは依然として人的貢献が重視されている。裁判所や立法府は、トレーニングデータ、プロンプト、デザインの選択、クリエイティブな要素の選択など、さまざまな側面から「十分な人間による作者性」の証拠を求めるだろう。特に、人間とAIが共同で創作を行う共同発明に関しては、このような進化する状況の中で、新たな認識の枠組みが必要と

なる。

生成AIコンポーネントの保護

- **アルゴリズムとモデル**：これらはしばしば企業秘密とみなされ、秘密にしておけば保護可能で、競争上の優位性をもたらす。独自のアルゴリズムを保護することも選択肢の一つだが、モデルが複雑化するにつれ、特に内部の意思決定プロセスやデータの依存関係に関して、機密性を維持することが難しくなる。ニューラルネットワーク・リバース・エンジニアリングと モデル 逆転／モデル盗用に関するいくつかの出版物は 、これらの課題について論じており、複雑なモデルの機密保持の難しさを強調している。モデルの特許に関しては、LLMの核となる概念（テキストやその他の出力を処理・生成するための統計的手法）や、その基礎となる数学的原理は、抽象的なアイデアや自然現象であるため特許の対象とはならない。しかし、独自のアーキテクチャや学習アルゴリズムなど、LLMの具体的かつ新規な実装は、新規性と非自明性の基準を満たせば特許になりえる。さらに、LLMを医療診断用に調整されたアプリケーションのような特定のアプリケーションと組み合わせれば、その組み合わせ自体がユニークで進歩的な解決策を生み出すため、特許になる可能性がある。つまり、モデル全体を特許化することは難しいかもしれないが、モデル内の特定の技術的特徴や特定の実装は、新規性や非自明性の基準を満たせば特許化できる可能性がある。
- **データ**：所有権は出所と用途に依存する。公開データは自由に利用できるが、ライセンスされたデータは特定の条件が必要である。AIのトレーニングデータの所有権は、特に複数の関係者から入手する場合、複雑になる可能性がある。2022年、2023年には、IPを考慮せずにグローバルなインターネットデータを使ってモデルをトレーニングする傾向があり、2023年には複数の訴訟に発展することになる。データプライバシー規制の適用と適切なライセンス契約は極めて重要である。

2. 著作権保護

現在の著作権法は、独創的な創造的表現を盾にしている。AIが生成した作品は、作家性やオリジナリティの問題を提起する。既存の法律は人間が作成した作品を保護するものであり、「作者」がアルゴリズムであるAIが作成した作品には課題がある。 本文書の出版日現在、米国著作権局は純粋にAIが生成した作品の保護を否定しており、国際的な議論を巻き起こしている。未解決の疑問が残っている： AIが生成した詩や音楽のような芸術的アウトプットは、特に著作権で保護された学習データに大きく依存している場合、保護のための独創性基準を満たすことができるのか？今日の法律では、データの選択、プロンプト、出力編集における人間の役割、トレーニングデータの公正使用原則が強調されている。

フェアユースの原則（米国著作権局、2023年）は、著作権者の許可を得ることなく著作物を限定的に使用することを認めている。たとえば、グーグルは、検索エンジンを作るために書籍からテキストをスクレイピングすることを変形的利用として認めると主張し、成功した。生成AIシステムは、その作成プロセスでウェブのスクレイピングデータを使用して、同じカテゴリに分類されるかもしれない。

3. 特許保護

特許は、新規かつ自明でない発明を保護するものである。AIモデルのアルゴリズムは、これらの基準を満たせば特許になる可能性がある。著作権と同様、特許も人間の発明者を必要とする。AIによる発明が存在する一方で、アルゴリズムへの発明者帰属は未解決のままである。最近のガイダンスとして、米国特許商標庁（2024）から、AIが生み出した発明について非自明性と発明者適格性を認めることは困難であるとの見解が示されている。単なるアウトプットではなく、生成AIが提供する進歩や解決策を強調し、技術的な側面に焦点を当てた考察が必要である。重要な要素は、将来の挑戦を避けるために、特許出願においてAIの機能を適切に（透明性をもって）開示することである。

4. 営業秘密

営業秘密とは、競争上の優位性をもたらす秘密情報である。組織は、未認可なエンティティがそのような企業秘密にアクセスできないよう、最大限の注意を払って企業秘密を保護しなければならない。AI/MLのアルゴリズム、モデル、およびトレーニングデータは、現在の法律では営業秘密に該当する可能性があるが、裁判所はすべての法域において、これに関する明確な判決を下すには至っていない。営業秘密保護の要件は異なる場合があるため、具体的な指針を得るためには弁護士に相談することが不可欠である。

複雑なAI/MLシステムでは、特にオープンソースの開発では、秘密保持が難しい場合がある。営業秘密は未認可な取得に対する限定的な保護に過ぎず、類似技術の独自開発に対する保護にはならないことを忘れてはならない。すべての利害関係者は、AIモデルとトレーニングデータの機密性を保護するための強固な対策を実施しなければならない。

5. ライセンスと保護戦略

- **オープンソースモデル/クリエイティブ・コモンズ・ライセンス**：AIモデルをオープンに共有することは開発を加速させるが、「複雑なアプローチ」（Semantic Scholar.org、2021年）であり、基礎となる学習データの悪用や著作権侵害の可能性が懸念される。クリエイティブ・コモンズ・ライセンスはAIが作成した作品に使用することができるが、ライセンスによっては他のライセンスよりも幅広い商業利用が許可されているため、適切なライセンスタイプを慎重に選択することが重要である。オープンソースライセンスに精通した弁護士への相談が推奨される。
- **商用ライセンス**：生成AIモデルを開発・展開する企業は、商業利用を可能にしながら知的財産を保護するために、慎重に作成されたライセンスを必要としている。共同研究者やユーザーとの契約では、所有権、使用権、および責任を明確に定義する必要がある。
- **データライセンス**：著作権侵害やプライバシー侵害を避けるためには、AIモデルの訓練や微調整に使用するデータについて適切なライセンスを取得することが不可欠である。営業秘密の横領や、特定のデータタイプ（ヘルスケアデータなど）に特有のプライバシー法など、使用される特定のデータに応じて他の法的課題も発生する可能性があることは注目に値する。

6. 商標

商標とは、連邦政府によって登録された、商品やサービスの出所を特定し区別するためのシンボル、言葉、またはフレーズのことである。営業秘密とは異なり、商標は、ロゴ、スローガン、さらには特定の製品デザインのような特徴的なブランド要素を法的に保護する。これは、AIが生成するイメージの領域では特に重要であり、ユニークなビジュアルアウトプットは貴重なブランド資産となりえる。

著作権は、AI画像を生成するために使用される特定のコードやプロセスを保護することができるが、結果として得られる画像自体は、その識別性や商取引における使用に応じて、商標保護の対象となる可能性がある。AIが生成する商標を取り巻く法的状況はまだ進化しているが、ここではいくつかの重要な検討事項を紹介する：

- **独自性**：商標として保護されるためには、AIが生成した画像は本質的に特徴的である必要がある。AIが生成したイメージがあまりにも一般的であったり、既存の商標と類似していたりすると、保護が難しくなる可能性がある。
- **著者名**：現在の商標法では、しばしば人間による著作が必要とされるが、創造的プロセスにおけるAIの役割が進化するにつれ、複雑さが増している。
- **ブランド使用**：画像は、製品やサービスの出所を特定できる方法で使用されなければならない。例えば、パッケージやマーケティング資料にAIが生成したビジュアルを一貫して使用することで、商標の主張を強化することができる。

商標の保護は自動的に行われるものではなく、事前の対策と執行が必要である。組織は、未認可な使用を登録し、監視する責任を負う。これには以下が含まれる：

- **商標登録**：AIが生成した商標を関連する商標事務所に登録することで、侵害された場合の法的立場を強化することができる。
- **アクティブ・モニタリング**：商標権侵害の可能性がないか、オンライン市場や競合他社の活動を定期的にチェックする。
- **施行**：侵害が検出された場合、未認可ユーザーに対して適切な措置を取るために弁護士に相談する必要がある場合がある。

企業は、AIが生成したイメージを商標として保護し、消費者がそのユニークなビジュアルをブランドと関連付けられるように競争上の優位性を守るために、積極的な対策を講じるべきである。

商標法は複雑であり、具体的な規制は法域によって異なる場合がある。AIが生成した画像を商標で保護するための具体的なガイダンスについては、知的財産を専門とする弁護士に相談することが推奨される。

7. 進化する状況

- **国際的な矛盾**：現在、生成AIに関する知的財産法は国によって異なるため、グローバルビジネスにとって課題が生じ、AIの国際的な協力と商業化を可能にするための調整が必要となる。欧州連合（EU）にはまだ明確な指令はないが、イギリスのように、すでに"コンピュータ生成著作物"をカバーする著作権法（1988年著作権意匠特許法（CDPA））がある国もある。この文書の第9節(3)には次のように書かれている：「文芸、演劇、音楽又は美術

の著作物でコンピュータにより作成されるものの場合、著作者は、その著作物の創作に必要な手配が行われた者とみなされる。

- **継続的な政策協議**：知的財産の枠組みを生成AI特有の課題によりよく対応させるための議論が続いている。AIが生成した著作物に対する新たな保護（例えば、「学習済みAIモデルに対するsui generis権利」）や、著作権や特許などの既存のカテゴリの改正が検討されている。

今後、政策立案者や法律の専門家は、AIが生み出す知的財産を管理するための解決策を積極的に模索しており、継続的な議論や今後の法改正に期待しよう。責任ある持続可能なAI開発を促進するためには、標準化されたライセンスモデルと、より明確な所有権の帰属が早急に必要である。訓練データの偏りや、AIが生成したコンテンツの潜在的な悪用など、倫理的な考慮事項に対応する追加の規制には、立法上の注意が必要である。

8. 関連法規

- バイデンが2023年10月に発表したAIに関する大統領令は、データプライバシー、倫理、人材育成、および国際協力に関連する条項で、AIの安全、セキュリティ、および信頼できる開発と利用の確保に焦点を当てている。同命令は、AIが生成したアウトプットに対して新たな知的財産権を設定するものではないが、「知的財産に関する既存の法的枠組みは、AIが提示するユニークな課題に対処するために完全には適していない可能性がある」と述べ、知的財産とAIを取り巻く複雑さを認めている。「明確で一貫性のある指針」の必要性を認識し、米国特許商標庁 （USPTO、2024年）や著作権局を含むいくつかの省庁に対し、AI関連の知的財産権に関する懸念に対処するための勧告を1年以内に作成するよう指示している。

- 世界知的所有権機関（WIPO）は、"経済・社会全体におけるAIアプリケーションの発展と、経済・文化財・サービスの創造・生産・流通への重大な影響に関わる知的財産課題の理解を深めるためのマルチステークホルダーフォーラム "を発行している。WIPO Conversationのいくつかのセッションでは、 人工知能の影響について検討されており、人工知能 （AI）が知的財産政策に与える影響について考察している。

責任あるAIのための技術戦略、標準、ベストプラクティス

このセクションでは、すでに説明した責任あるAIを実装するための技術標準とベストプラクティスのいくつかを要約し、成功した実装アプローチを示す短いケーススタディを提供する。

組織にとってよくある質問は、AIの利用において透明性、説明責任、倫理的慣行を実証するために、確立された技術標準をどのように活用するかということである。技術標準は様々な方法でグループ化することができる。私たちは、将来必要に応じて拡張できるよう、簡略化した分類を採用した。

公平性と透明性

- **データの公平性**：

    - **データセットの多様性**：代表的で多様なデータセットを積極的にキュレーションし、アウトプットにおける意図しないバイアスを最小限に抑える。
    - **データセット監査**：生成AIモデルのトレーニングデータセットを定期的に監査し、潜在的な偏りや不十分を特定する。多様性と包括性を高めるために、データ拡張（data augmentation）や合成データ生成（synthetic data generation）のような技術を採用する。
    - **データの透明性**：生成AIモデルの訓練に使用されたデータセットについて、 その構成、ソース、必要な前処理手順などの情報を公開する。これは外部からの精査を可能にし、データの潜在的な偏りやギャップを特定するのに役立つ。
    - **定期的なバイアスの評価**：データセットと生成AIモデル内のバイアスを特定し、 緩和するためのツールとプロセスを積極的に導入する。定期的なテストと検証を行い、差別的な出力がないかチェックする。
    - **バイアスの緩和**：開発・配備段階において、生成AIモデル内のバイアスを検出し対処するために、公平性メトリクスとバイアス緩和技術を積極的に使用し、開発する。
- **アルゴリズムの透明性**：
    - **文書化**：生成AIモデルの設計、アーキテクチャ、および意思決定プロセスを徹底的に文書化する。理解と精査を促進するため、利害関係者にわかりやすい形式でこの情報を共有する。
    - **モデル解釈可能性**：LIMEやSHAPのような説明可能なAI（XAI）技術を採用することで、生成AIモデルがどのように特定の結果を導き出すかに関する洞察を提供し、潜在的なバイアスの特定を可能にする。
    - **モデルカード**：モデルの使用例、トレーニングデータ、パフォーマンス測定基準、制限、および潜在的なバイアスの概要を透過的に示す「モデルカード」を作成する。モデルカードは、機械学習モデルの透明な文書として機能し、学習データ、制限、および使用目的を詳細に説明する。責任あるAIを推進するために、企業はHugging Face、TensorFlow Model Garden、Papers With Codeにあるようなモデルカードを利用することができる。そこにはデータソース、構成、および前処理ステップに関する情報が含まれている。これにより信頼が醸成され、ユーザーはAIシステムの潜在的なバイアスや限界を理解することができる。
    - **オープンソースモデル**：可能であれば、オープンソースの生成AIモデルに貢献し、 より幅広い精査と共同改善を可能にする。
- **説明可能性**：
    - **解釈可能なモデル**：可能であれば、意思決定プロセスに対するより深い洞察を提供し、固有の解釈可能性を持つ生成AIモデルを優先する。
    - **説明可能なAI（XAI）**：生成AIモデルがどのように意思決定やアウトプットに至るかを説明するテクニックを組み込む。XAIテクニックを活用して、ブラックボックスモデルであっても、出力に影響を与える要因を強調した説明を生成する。これにより、ユーザーや利害関係者は推論プロセスを理解し、モデルがどのように機能するかに関する理解を深めることができる。
    - **説明可能なインターフェース**：生成AIのアウトプットの背後にある明確な説明と根拠を提供し、信頼と理解を醸成するユーザインターフェースを設計する。

セキュリティとプライバシー

- **データセキュリティ**：
  - **暗号化**：機微データを保存時、移動時、および使用時に暗号化する。
  - **認証プロトコル**：多要素認証（MFA）やゼロトラストセキュリティモデルなどの堅牢な認証プロトコルを採用し、機微情報やAI機能へのアクセスを認証された正規ユーザーに厳密に許可することで、不正アクセスのリスクを軽減する。
  - **定期的な監査**：定期的なセキュリティ監査と脆弱性評価を実施し、潜在的なセキュリティリスクを特定し軽減する。
- **プライバシーを保護する技術**：
  - **プライバシー・バイ・デザイン**：プライバシーの原則（データの最小化、同意、セキュリティなど）を生成AIシステムの設計と実装に直接組み込む。
  - **プライバシー強化技術（Privacy-enhancing technologies、PET）**：機微性の高いユーザデータを保護する技術を探る：
    - **差分プライバシー**：計算されたノイズをデータセットに加えることで、統計的特性を維持しながら情報を匿名化し、個人のプライバシーを保護しながら分析を行うことができる。
    - **統合された学習**：複数のデバイスやサーバーに分散されたデータで生成AIモデルを訓練することで、機微データを中央の場所に集める必要性を回避する。
    - **準同型暗号**：暗号化されたデータを復号せずに計算する。これにより、基礎となるデータを明らかにすることなく、機微情報をセキュアに分析することができる。
- **敵対的攻撃に対するセキュリティ**：
  - **Adversarial Robustness Training（ART）**：意図的な操作に対するレジリエンスを高めるために、敵対的な例を使って生成AIモデルを訓練する。ARTは特定のケースでは有効であることが示されているが、研究者の中には計算コストや学習分布外の敵対的攻撃に対する感度などの考慮事項に基づいて、ARTの実用的な限界について懸念を示す者もいる。
  - **セキュリティテスト**：敵対的攻撃シミュレーションを定期的に実施し、 脆弱性を特定 し、モデルの防御力を向上させる。

堅牢性、コントロール、および倫理**AI**の実践

- **安全性と信頼性**：
  - **リスクアセスメント**：徹底的なリスクアセスメントを実施し、生成AIシステムの潜在的な危害と予期せぬ結果を特定し、緩和措置とセーフガードを実施する。
  - **テストと検証**：様々なシナリオやエッジケースで生成AIモデルを厳密にテストし、様々な状況における信頼性と堅牢性を確保する。
  - **被害を最小限に抑制**：潜在的な危害を最小限に抑えるためのセーフガードを備えた生成AIシステムを設計する。これには、「セーフティスイッチ」の組み込みや、用途やリスクに応じた制限の設計が含まれる。

- **人間による監督**：
    - **ヒューマン・イン・ザ・ループ**：重要な意思決定プロセスにおいて、特に重要度の高い申請については、有意義な人間の関与を維持する。必要に応じて生成AI出力をオーバーライドまたは調整するために、人間の介入を許可する。
    - **フェイルセーフメカニズム**：明確なエスカレーションパスとフェイルセーフメカニズムを確立し 、特に外部ユーザーから報告された場合に、予期しないあるいは有害なモデルの動作に対処する。
- **説明責任**：
    - **オーナーシップと責任**：AIシステムの開発、デプロイ、および監視について明確な役割と責任を定め、個人が技術の影響に責任を持つようにする。これにより、課題に対処し、効率的に改善を行うための説明責任が確保される。
    - **監査証跡**：モデルの開発、トレーニング、および使用に関する徹底的なログを維持する。これらの監査証跡は、予期せぬ挙動や倫理的な懸念を調査する上で非常に貴重なものとなる。
    - **報告のメカニズム**：利害関係者（社内外）が生成AIシステムに関する懸念事項や潜在的な課題を報告できるオープンなチャネルを構築する。これにより、積極的なフィードバックが促進され、迅速な是正措置が可能になる。
    - **インシデントレスポンス**：意図しない結果やAIに関連した危害が発生した場合に備えて、明確なインシデントレスポンス計画 と報告メカニズムを確立する。
    - **倫理審査委員会**：生成AIアプリケーションの潜在的な影響を評価し、企業の価値観との整合性を確保するために、倫理委員会や審査委員会を設置する。
    - **バイアスと公平性の監査**：生成AIシステムのデータセット、アルゴリズム、および結果における潜在的なバイアスを特定し、緩和するための監査を定期的に実施する。

組織がこれらの標準を活用する方法

これらの技術標準を効果的に採用することは、単なる理解にとどまらない。組織は、ベストプラクティスを開発プロセスに組み込むことによって、倫理基準を現実の行動に移さなければならない。以下はその例である。

- **明確な社内方針を確立する**：これらの標準を社内の開発ガイドラインや組織の方針に組み込む。開発プロセスから生産までのすべての段階において、責任あるAIに対する、明確な期待を作成する。
- **文書化と報告**：データの使用状況、モデルのパフォーマンス、バイアスの評価、および是正措置に関するレポートを定期的に公表する。これにより、外部の利害関係者との透明性が促進される。
- **パートナーシップとコラボレーション**：あなたは孤独ではない！業界団体や倫理AI研究コミュニティと連携 し、ベストプラクティスの開発に貢献し、責任ある生成AIに関する議論を積極的に形成する。

技術標準はあくまで出発点であり、普遍的なものではないことに留意することが重要である。その実装は、特定の組織のニーズや使用事例に合わせて調整されるべきである。倫理AIの導入は、（1回限りの解決策ではなく）継続的なプロセスであり、技術、規制、および社会的期待の進化に伴い、組織はそのプロセスを常に適応させ、更新していく必要がある。

責任ある生成AIのための技術的セーフガード（データ管理）

表2は、最も一般的なデータ管理規制に準拠したAIシステムを構築するための主要な技術とベストプラクティスの概要を示している。

データプロセス	テクニック	説明
データの前処理	データの匿名化または仮名化	これは、トレーニングデータから個人を特定できる情報（PII）を削除または置き換えるもので、プライバシーリスクを最小限に抑える。トレーニングにPIIが使用される場合、アウトプットの慎重なサニタイズが必要となる。
	データのフィルタリング	生成モデルの特定の目的に関連するトレーニングデータを選択し、フィルタリングすることで、不必要なデータ収集やデータ増強を避ける。
データキュレーション	データ選択	目的の目的に沿い、バイアスを避けるためにトレーニングデータを慎重に選択する。これには、無関係な情報や有害な情報のフィルタリングを含む。
	データの増強	ノイズを加えたり、合成データを生成したりすることで、学習データの多様性を高め、より堅牢で偏りの少ないモデルを作ることができる。
モデルの設計、トレーニング、最適化	統合された学習	分散化されたデータセットでモデルをトレーニングし、データを中央サーバーに転送する代わりに個々のデバイスに保持する。
	差分プライバシー	トレーニングデータにランダムなノイズを導入する。このノイズは、正確なデータがマスクされるため、個人のプライバシーを守るために役立つ。しかし、データセットが十分に大きければ、ノイズが多数の人々の間で平均化されるため、個々のデータポイントを特定することなく、実際の傾向やパターンを観察することができ、プライバシー保護が強化できる。
	モデルの解釈可能性	潜在的なバイアスやエラーの特定と緩和を容易にするため、出力に至る方法を理解できるモデルを開発する。
	定期的なモニタリングと再トレーニング	モデルのパフォーマンスを定期的に監視し、バイアスへのドリフトや不正確な出力の生成などの潜在的な課題に対処するために、更新されたデータやキュレーションされたデータを使って再トレーニングを行う。

	ハイパーパラメータの調整	学習プロセスをコントロールするモデルのハイパーパラメータを微調整することで、出力に影響を与え、意図しない結果を軽減できる可能性がある。
継続的なモニタリングと評価	モデルで使用するデータを定期的に監査、評価	意図した目的に沿い、不必要なデータが保持されないようにする。
	モデルの出力に潜在的なバイアスや意図しない結果がないか監視	特定された課題に対処するためのセーフガードを実装する。
ヒューマン・イン・ザ・ループ・テクニック	人間による監督：	人間の監視をプロセスに組み込み、デプロイや活用の前に人間がAIのアウトプットをレビューし、検証する。
	インタラクティブ生成	ユーザーがAIの生成プロセスを誘導し、望ましい結果を達成できるような対話型システムを設計する。
説明可能性と透明性	説明可能なAI（XAI）技術	モデルの推論を理解し、潜在的なバイアスや限界を特定するために、LIME、SHAP、Mimic、または Permutation Feature Importance のような技術を使用する。
	開発とデプロイにおける透明性	AI/MLとその応用に伴う限界、偏見、および潜在的リスクについて透明性を保つ。

Table 2:責任あるAIシステムを構築するための重要なテクニックとベストプラクティス

ケーススタディ ～ 透明性と説明責任を実践で示す

このケーススタディは、組織が倫理AIの原則を具体的な開発手法に反映させるための実践的なアプローチを示している。この例では、ある企業が画像生成のために生成AIモデルを導入している。このケースは、透明で説明責任のあるAIのための具体的な戦略と技術標準が、その開発とビジネスプロセスに直接組み込まれていることを示している。これには以下のいくつかの主なステップがある。

- **モデルのトレーニングデータセット、制限事項、および想定される使用例を説明したモデルカードを公開**：収集されたトレーニングデータは、その起源が明確に文書化され、特定の使用方法と意図について適切な同意が得られていた。データセットは代表的で多様性に富んでおり、取得した顧客データ、一般に入手可能なデータ、そしてデータ拡張と組み合わせた合成データの組み合わせである。これらすべては、生成されたアウトプットに意図しないバイアスが生じる機会を最小限に抑えることを目的としている。同社は、潜在的なバイアスおよび/またはトレーニングデータセットを定期的に監査している。同社は、生成AIモデルの学習に使用するデータセットについて、その構成、ソース、および社内で利用する前処理ステップなどの情報を公開した。

実践的なアプローチ：Hugging Faceでホストされているモデルカードと同様に 、同社は TensorFlow Modern Gardenに基づいて、 生成AIモデルのトレーニングデータの概要を 示す詳細なモデルカードを提供している。このカードには、データソース（例：顧客データ、一般に入手可能なデータセット）、構成（例：テキスト、画像）、および前処理ステップに関する情報が含まれる。この透明性によって、ユーザーはモデルの潜在的なバイアスと限界を理解することができる。

- **説明可能なAI（XAI）技術を採用**し、生成された画像とともに人間が理解できる説明を提供することで、出力に影響を与える要因を明確にする。これらの説明は、画像出力に貢献した主要な要因を強調し、ユーザーに力を与える：
    - **生成された画像の根拠を理解する**：生成AIモデルの意思決定プロセスを透明化することで 、ユーザーはモデルがなぜ特定の画像を生成したのかについての洞察を得ることができる。これによって信頼が醸成され、モデルの推論を理解した上での意思決定が可能になる。
    - **潜在的なバイアスを特定する**：XAIの説明は、トレーニングデータやモデル自体の潜在的なバイアスを明らかにすることができる。これにより、ユーザーは出力を批判的に評価し、差別的または不公正な要素が存在する可能性があるかどうかを特定することができる。
    - **モデルのデバッグと改善**：説明を分析し、特定の要因が出力にどのような影響を与えたかを理解することで、開発者はモデルの潜在的な欠点を特定し、その精度と公平性の向上に取り組むことができる。
- **ヒューマンレビュープロセスを確立し、バイアス検証を実施することは、すべてのテストサイクルの一部である**：このプロセスは、特に機微性が高くリスクの高いユースケースのために設計されており、いくつかの重要な点を考慮する必要がある。
    - **フラッグの基準**：明確な定義づけがなされた、人間によるレビューのトリガーとなる基準を確立する。これには、予想外の（潜在的に有害な）モデル動作の変化や、モデルの信頼度がある閾値を下回ったとみなされる特定の出力が含まれる。
    - **レビューチームの構成**：技術オーナーおよびデータサイエンティストと緊密に連携し、ビジネスステークホルダーで構成されるヒューマンレビューのための多様で優秀なチームを編成する。このチームは、モデルの目的、潜在的なバイアス、そのアウトプットの倫理的意味を理解するために必要な専門知識を有している。
    - **レビューの手順**：フラグが付けられたアウトプットをレビューするための明確で標準化された手順を定義する。これには、潜在的なバイアスを評価し、倫理的ガイドラインとの整合性を確保し、モデルの再較正やデータのクレンジングなどの適切な措置を決定することが含まれる。
    - **バイアス検証をテストサイクルに組み込む**：バイアスの検証は1回限りのイベントではない。生成AIモデルの開発と配備のライフサイクルを通じて継続的に行われるプロセスである。企業レベルでは以下の戦略が採用される。
        - **テスト用に多様なデータセットを採用する**ことで、 トレーニングデータに潜在するバイアスの診断に役立てる。
        - **公平性の指標を活用**：開発プロセス全体を通して、公平性の指標を導入し、監視する。これらの指標は、モデルの出力における潜在的なバイアスを特定し、定量化するのに役立つ。
- **モデルの出力の定期的なバイアス監査を実施し、潜在的な差別的行動を検出して緩和する**：この監査では、人間の専門家、データサイエンティスト、ビジネス関係者が協力して、イメージ生成において特定の属性を優遇したり、有害なステレオタイプを永続させるなど、モデルの出力に意図しない偏りがないかどうかを分析する。一旦特定されると、適切な緩和戦略が実施される（例：拡張されたデータセットによるモデルの再トレーニング、モデルアルゴリズムの調整など）。すべての機微性

の高いユースケースには、人間による監視が組み込まれている。

これらの業界標準とベストプラクティスを実施することで、生成AIモデルの倫理的で責任ある使用を保証し、消費者との信頼を築くための積極的な措置を講じている。

継続的なモニタリングとコンプライアンス

生成AIが私たちの生活やビジネス手法にますます統合されていく中で、その安全かつ倫理的な使用を確保することが最も重要である。継続的なモニタリングとコンプライアンスは、効果的な生成AIガバナンスの重要な側面であり、潜在的なリスクを継続的に評価し、責任ある生成AIの使用を支持することを可能にする。
コンプライアンスは、単に進化する法律に対応するだけではない。責任ある生成AIの使用を保証するには、継続的なコンプライアンスのための積極的な計画とともに、ライフサイクルの各段階を慎重に評価する必要がある。これは複雑な仕事であり、通常は以下の2つの側面からのアプローチが必要である。

1. **強固なモニタリングプロセスの確立**：これは、生成されたコンテンツと開発プロセス全体を継続的に監視することをいう。これには、データ、モデル、出力のバイアスを検出することと、データプライバシー規制と倫理的取り扱いの遵守を検証することが含まれる。この積極的なアプローチは、ディープフェイクや有害コンテンツのような生成コンテンツの悪用を防止しながら、公正さと包括性を促進する。

継続的なモニタリングは、2つの重要な課題を解決するために役立つ：また、生成されたコンテンツが悪用される可能性を特定できるため、企業は倫理的なガイドラインに従い、誤った情報の拡散を防ぐことができる。

2. **包括的なコンプライアンス計画の策定**：この計画は、生成AI活動に関連する潜在的なコンプライアンスリスクを特定し、軽減するための手順を概説するものでなければならない。主な考慮事項は以下の通りである。
    - **データセキュリティとプライバシー**：適用される規制に従って、データの収集、保存、および処理中に機微情報を保護するための強固なセーフガードを導入することは極めて重要である。
    - **バイアスと公平性**：公平性と無差別性を確保するために、トレーニングデータとモデル出力の潜在的なバイアスを定期的に評価し、緩和する。
    - **透明性と説明可能性**：生成AIツールがどのように機能し、そのアウトプットの背後にある根拠をユーザーに理解してもらうことは、信頼と説明責任を構築する上で極めて重要である。

コンプライアンスを積極的に監視し、適切なセーフガードを導入することで、企業は責任ある倫理的な生成 AIの利用を保証し、ユーザーやステークホルダーからの信頼を醸成することができる。

生成AIを管理する上での法的・倫理的考察

生成AIを効果的に管理するには、法的および倫理的な考慮事項の間の複雑な相互作用をナビゲートする必要がある。合法性は、生成AIの急速な技術進歩に遅れをとりがちな、確立された法律や規制を遵守することに重点を置いている。そのため、グレーゾーンが生まれ、倫理的な枠組みがその開発と配備を導く余地が残されている。

法的には、知的財産、データプライバシー、および非差別といった既存の法律の遵守が焦点となる。これには、責任ある開発、透明性のあるデータ利用、生成AIのアウトプットによって引き起こされる可能性のある危害に対する明確な説明責任を確保する枠組みの確立が含まれる。議論されているように、AIが生成したコンテンツから著作権侵害が発生する可能性がある一方、データプライバシー規制は、これらのシステムを訓練し運用するためにユーザー情報がどのように使用されるかに取り組んでいる。さらに、社会的不平等を永続させないためには、AIの出力における公平性を確保し、バイアスを軽減することが極めて重要である。

倫理的配慮は、単に法律を守るだけではない。これらは、生成AIの責任ある有益な利用を確保するための、より広範な社会的価値観と原則を包含している。バイアス、透明性、説明責任、および技術の悪用の可能性をめぐる重要な問題は、すべて倫理的な傘の下にある。このような懸念に対処するためには、開発者、政策立案者、および一般市民の間で継続的な対話と協力を行い、生活の様々な側面における生成AIの統合に関する倫理的ガイドラインとベストプラクティスを開発する必要がある。

生成AIがますます普及するにつれ、いくつかのホットなトピックが浮上している。AIによる自動化によって雇用が奪われる懸念、ディープフェイクによる言論操作の可能性、および生成AIを利用した偏ったコンテンツの作成など、いずれも慎重な対応が求められる分野である。これらの課題に対処するには、政策立案者、開発者、ビジネス関係者、そして一般市民が協力し、イノベーションと社会の幸福のバランスをとる包括的なガバナンスの枠組みを開発する必要がある。

結論：責任ある未来のためにAIガバナンスのギャップに対処する

AIガバナンスの現状は、世界中の政策立案者や規制機関が早急に注意を払う必要があるいくつかの重要な課題を伴う複雑な状況を明らかにしている。一方で、既存の規制はAIに間接的に触れてはいるものの、この進化する技術がもたらす独自の課題に効果的に対処するために必要な具体性を欠いている。逆に、生成AI技術の急速な普及と日常生活の様々な側面への統合は、包括的な法整備が急務であることを浮き彫りにしている。このギャップにより、生成AIを含むAIシステムの責任ある開発、配備、および利用のための明確なガイドラインを確立する新たな規制の策定が必要となる。

さらに、AIガバナンスに関する国際的な協力の欠如は、断片的な法的状況を生み出し、イノベーションを阻害する可能性がある一方で、法域間の矛盾や一貫性の欠如が懸念される。このような調和の欠如は、抜け穴を作り、AIに関連する危害に対する責任をアクターに負わせる上で困難をもたらす可能性がある。

こうした課題に取り組む緊急性は、生成AIの急速な普及と私たちの日常生活への統合が進むことで高まっている。企業は生成AIを競争上の優位性とみなし、強固な規制がない場合でも、その急速な導入を推進している。様々な分野で存在感を増している生成AIは、イノベーションとディスラプションの両面で強力なツールとなる可能性を秘めている。生成AIによって引き起こされた損害賠償に関連する訴訟の出現は、規制のギャップに対処し、潜在的な悪影響から保護することが急務であることを痛感させるものである。

責任ある未来に向けて前進するために、私たちは多面的なアプローチを受け入れるべきである。

1. **AIの法規制の整備を加速**：立法者は、生成AIに関連する特定のニーズと潜在的なリスクを考慮し、包括的かつ適応可能なAI規制の策定を優先する必要がある。そのためには、政府、業界の専門家、および市民社会が協力して、効果的かつ倫理的な枠組みを確立する必要がある。
2. **国際的な協力と調和**：AIガバナンスに関する国際的な協力関係を促進することは、裁判管轄間の分断や矛盾に対処するために不可欠である。国や地域の特性を尊重しつつ、国際的な枠組みや基準を確立することは、責任あるイノベーションを促進し、国境を越えた効果的な説明責任を確保することにつながる。
3. **技術基準と責任ある開発**：堅牢な技術標準とベストプラクティスを開発し実施することは、あらゆるセクターで責任あるAIの開発と管理を可能にするために極めて重要である。この包括的なガイドラインにより、企業、開発者、および政策立案者は、倫理的配慮に沿ったAIシステムを構築し、公平性と透明性を優先し、最終的に社会に積極的に貢献することができるようになる。

包括的な規制がないにもかかわらず、今日、企業はAIシステムの設計をめぐってますます厳しい監視に直面している。「ビルトイン」や「バイデザイン」など、AIを製品や製品に適切に組み込むための明確なガイダンスの必要性が高まっている。本書の実践的なアプローチは、技術標準を正しく適用することの重要性を強調し、これらの標準が現在の法的状況の中で責任あるAI開発をどのようにサポートできるかについての初期的／限定的な理解を提供する。

今後、生成AIの効果的なガバナンスを実現するためには、迅速な対応が求められる。政策立案者は、イノ

ベーションと社会的利益の保護とのバランスを考慮した規制を策定し、実施することを優先しなければならない。調和のとれた標準を確立し、矛盾した規制の枠組みを防ぐためには、国際的な協力が不可欠である。適切な法整備は、企業が提供するAIがバイアスや差別を厳格に緩和し、安全なデプロイのためのベストプラクティスを遵守するよう、監視の目を厳しくする負担を軽減する。これは、すべての利害関係者の間で倫理的で責任あるAI開発の重要性が認識されつつあることを強調するものであり、業界の慣行を導く上で規制当局の支援が重要な役割を果たすことを強調するものである。