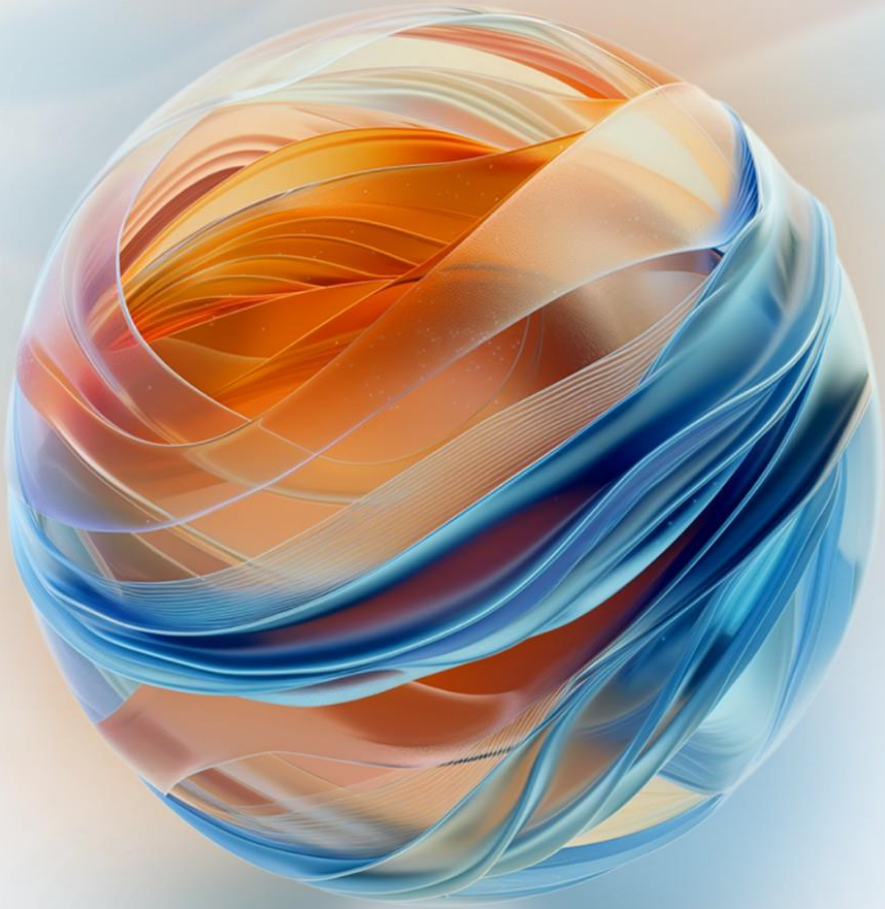
AIレジリエンス：

AIの安全性に関する革命的なベンチマークモデル



AI Governance and Compliance
Working Group

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

謝辞

Lead Authors

Dr. Chantal Spleiss

Contributors

Romeo Ayalin
Filip Chyla
Becky Gaylord
Frederick Hanig
Rocky Heckman
Hadir Labib
Lars Ruddikeit
Alex Sharpe
Ashish Vashishtha

Reviewers

Sounil Yu
Debjyoti Mukherjee
Michael Roza
Peter Ventura
Udith Wickramasuriya
Govindaraj Palanisamy
Madhavi Najana
Rakesh Sharma
Davide Scatto
Paresh Patel
Piradeepan Nagarajan
Gaetano Bisaz
Hongtao Hao, PhD
Elle Pyle
Gaurav Singh
Ken Huang
Kenneth T. Moras
Tolgay Kizilelma, PhD
Akshay Shetty
Saurav Bhattacharya
Peju Okpamen
Gabriel Nwajiaku
Meghana Parwate
Akshat Vashishtha
Hemma Prafullchandra
Renata Budko
Desmond Foo
Scott S. Newman
Gian Kapoor
Imran Banani
Elier Cruz
Madhav Chablani

CSA Global Staff

Ryan Gifford
Stephen Lumpe

日本語版提供に際しての告知及び注意事項

本書「AIレジリエンス：AIの安全性に関する革命的なベンチマークモデル」は、Cloud Security Alliance (CSA)が公開している「AI Resilience: A Revolutionary Benchmarking Model for AI Safety」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9月20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

「AIレジリエンス：AIの安全性に関する革命的なベンチマークモデル」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

目次

エグゼクティブ サマリー

AIガバナンスとコンプライアンスの複雑な状況をナビゲートするために、（進化した）AIベンチマーキングモデルを紹介する。収益主導の進歩は、安全策を確立するための規制当局の努力を上回り、AIシステムが真に堅牢で信頼に足るものであることを保証するには至らないことが多い。経営幹部は、進化と心理学の原則に着想を得た新しいベンチマークモデルを導入することで、この重大なギャップに対処し、パフォーマンスとともに堅牢性を優先させ、AIシステムの全体的な品質を積極的に評価できるようにする。

過去のAIの失敗事例から教訓を導き出し、自動車、航空、重要インフラ、エッセンシャルサービス、防衛、教育、金融、ヘルスケアなどの業界を分析することで、企業にとって実践的な洞察と実行可能な指針を提供する私たちは、より倫理的で信頼できるAIアプリケーションに向けて業界を推進するため、規制ガイドラインに多様な視点を統合することを提唱している。信頼性を重視することは、リスクを最小化し、評判を保護し、責任あるAIのイノベーション、配備、利用を促進するための鍵となる。

本書は、政府高官、規制機関、業界リーダーを含む主要な意思決定者が、倫理的なAIの開発、配備、利用を保証するAIガバナンスの枠組みを確立することを支援する。AIの品質を評価するための新しいベンチマークモデルを紹介し、長期的な成功のための実用的なツールを提供する。

はじめに

AI（Artificial Intelligence）の急速な進化により、前例のない進歩が約束されている。しかし、AIシステムが高度化するにつれ、リスクも増大している。医療における偏ったアルゴリズムから自律走行車の誤作動に至るまで、過去に起きた事件はAIの失敗がもたらす結果を如実に浮き彫りにしている。現在の規制の枠組みは技術革新のスピードについていけないことが多く、企業は風評被害と業務上のダメージの両方に対して脆弱なままになっている。

このような課題に対応するため、本書ではAIのガバナンスとコンプライアンスについて、より包括的な視点からの検討が急務であることを取り上げる。AIの基礎を探求し、重要な業界にわたる問題を検討し、責任ある実装のための実践的な指針を提供する。AIの革命（進化）を生物学と比較する斬新なアプローチを提示し、AI技術の安全性を高めるための多様性という示唆に富む概念を導入する。知能の違いと、そのようなシステムの相互作用の成功について議論する。この破壊的技術の安全性と信頼性を高めるための革新的なベンチマークの枠組みを提示する。

このアプローチにより、意思決定者や技術チームはAIシステムの安全性と信頼性を評価できるようになる。私たちは、多様な視点と規制ガイドラインを統合することを推奨し、倫理的なAIのイノベーションを促進し、強力なガバナンスの実践を確立することを目指す。

パート I: 基礎を理解する

ガバナンス vs. コンプライアンス

ガバナンスとコンプライアンスは、組織運営に不可欠な側面であり、規制、倫理原則、基準、および事業行動規範に概説されている持続可能性の実践の遵守を保証する。前述の原則や規制との整合性により、効果的な事業継続と倫理的実践が保証される。

ガバナンス[1]とは、何かを監督・管理することを指し、トップダウン方式で実施される。上級管理職は、戦略及びリスク選好度を定義し、方針、基準、及び／又は手続を通じてガバナンスの枠組みを確立する責任を負う。これらの指令は、組織の包括的なリスク管理アプローチ、コンプライアンス義務、意思決定プロセスを形成する。ガバナンスは、説明責任、透明性、倫理的行動、持続可能性の文化を生み出すと同時に、全社的なセキュリティとプライバシー対策を優先させる。

ガバナンスのトップダウンアプローチとは対照的に、コンプライアンス[2]はボトムアップアプローチであり、様々なレベルの従業員が、規制要件を満たすために経営陣によって定義されたガバナンスの枠組みを実施し、遵守する。コンプライアンスは、法律、規制、業界基準、および社内のビジネス行動規範を確実に遵守することに重点を置く。これは、組織が適用される法規制要件、許容される倫理的境界線、および最小化されたリスクエクスポージャーの範囲内で運営されることを確実にするための、組織管理の極めて重要な要素である。

ガバナンスとコンプライアンス：動くターゲット

ガバナンスとコンプライアンスが明確に定義された目標であるのに対し、AIの利用は従来のアプローチに課題をもたらすものである。AIは、テクノロジー、1つまたは複数のモデルを使用するシステム、ビジネスアプリケーション、ユーザープラットフォームなど、さまざまな視点で見ることができる。AIは、単一のエンドユーザー、または多数のエンドユーザーにサービスを提供することができ、企業、情報ブローカー、または他のAI技術によって、タスクの実行、問題の解決、意思決定、または環境との相互作用に使用することができる。AIの利用を取り巻く新たなベストプラクティス、基準、規制は進化し続けており、実施・監視すべきコンプライアンス要件を具体的に示すことは困難である。国際的なビジネスを展開する企業にとって、この課題は飛躍的に増大する。ほとんどの規制には重複する要件があり、AIの安全性を向上させる根本的に新しい提案はほとんどなく、現在の枠組みはこうした一般的な要件に基づいている。

- **人間による監督**：AIが人間の監視と制御を受け、必要に応じて人間の介入と意思決定を可能にする仕組みがあることを保証する。人間による監視は、主要なステップにおいて自動化されたモニタリングと組み合わされなければならず、人間による介入が必要であると特定されたユースケースでは人間による監視を必須とする。これにより、このガイダンスはスケーラブルで実用的なものとなる。

- **安全性と信頼性**：AI技術の安全性と信頼性を優先し、個人または社会への危害のリスクを最小化する。これは、厳格なテスト、検証、リスク評価プロセス、および故障時のキルスイッチや代替手段の仕組みの実装によって達成される。

- **倫理的配慮**：AIが倫理原則を遵守し、人権を尊重し、公正さを促進するようにする。

- **データのプライバシーとセキュリティ**：機密情報とプライバシーを保護し、データへの不正アクセスや悪用を防止するために、データ保護とセキュリティ対策を強化する必要がある。設計フェーズでは、プライバシー・バイ・デザインとセキュリティ・バイ・デザインは、（DevSecOpsのシフトレフトを反映して）プロセスの早い段階でリスクを軽減することに重点を置く。これにより、後付けのセキュリティや最終製品における予期せぬリスクの露出を制限することができる。

- **AIモデルとデータに関する考察**：

    - **バイアスの緩和**：データやアルゴリズム設計におけるバイアスに対処し、AIシステムを定期的に監視し、バイアスや差別の有無を評価する。バイアスは、必要な情報やアルゴリズムと、ステレオタイプ的な分類のリスクとのバランスを取る複雑なテーマである。
    - **透明性**： AIがどのように機能するかを明確に説明し、そのアルゴリズムや意思決定に影響を与える要因を含めて、透明性を確保する。XAI（explainable AI：説明可能なAI）[2]、[3] の導入は、信頼を育み、偏見の可能性を明らかにしつつ、情報に基づく意思決定の基盤を築くうえで有用である。特にヘルスケアにおいては、この点が重要であることが認識されている。業界に関わらず、ユーザーには、提供された結果がAIによって生成されたものであるかどうかが明示されるべきである。
    - **一貫性**：一貫したデータにより、**AI**モデルが正確かつ信頼できる事例から学習することが保証される。これは、モデルが正確で有用な出力を生成するために極めて重要である。不一致または矛盾するデータはモデルを混乱させ、生成されたテキストや情報における不正確さを招く可能性がある。
    - **説明責任**： **AI**の設計、開発、展開、および使用において、責任と説明責任のメカニズムを確立する。発生しうる問題に対処するための明確な責任の所在を含む。現状では、エンドユーザーへの被害を防ぐ責任は主に**AI**アプリケーション提供者にのみ負わされている。マニュアルや「モデルカード」 [4] といった追加措置や、特定のユーザートレーニングを導入することで、提供者とエンドユーザーが責任を共有していることを強調し、エンドユーザーが期待できる透明性のレベルを明確にすることができる。
    - **堅牢性**： **AI**を適切に設計し、敵対的攻撃、データの変動、その他の干渉や操作に対して堅牢であるように開発すること。さらに本書では、グローバルな安全性を強化するための堅牢性評価に関する新たな視点を提案する。
    - **規制遵守**： **AI**アプリケーションの開発および展開に関する関連法規、規制、標準に準拠することを確保する。これには、データ保護、データ取引、プライバシー、および安全性に関する規定が含まれるが、それに限定されない。

高度で複雑なサプライチェーンおよびバリューチェーン全体で責任を共有するアプローチを採用することは、安全で信頼性のあるAIを確保するために不可欠である。これには、少なくとも技術チーム、コンプライアンスチーム、法務チーム、そして特定の要因によっては他の多くのチームも関与する。2024年3月28日のホワイトハウスの覚書[5]は、すべての機関が60日以内にAI最高責任者（CAIO）を指名するよう要請している。この役職により、関与するすべてのチームを戦略的かつ目的に沿って管理・調整し、「責任の共有」を追跡可能な指標へと変えることができる。

AIを取り巻く状況

本章では、AIの概要として、簡単な歴史、AI技術、およびトレーニング方法を紹介する。データの重要性について議論することは本概要の範囲を超えるが、それが極めて重要なテーマであることは認識されており、他のCSAワーキンググループによって詳細に扱われている。

AIの歴史

以下は、AIのマイルストーンである。特定の視点に限定せず、この分野における主要な発展の概要を示すものである。

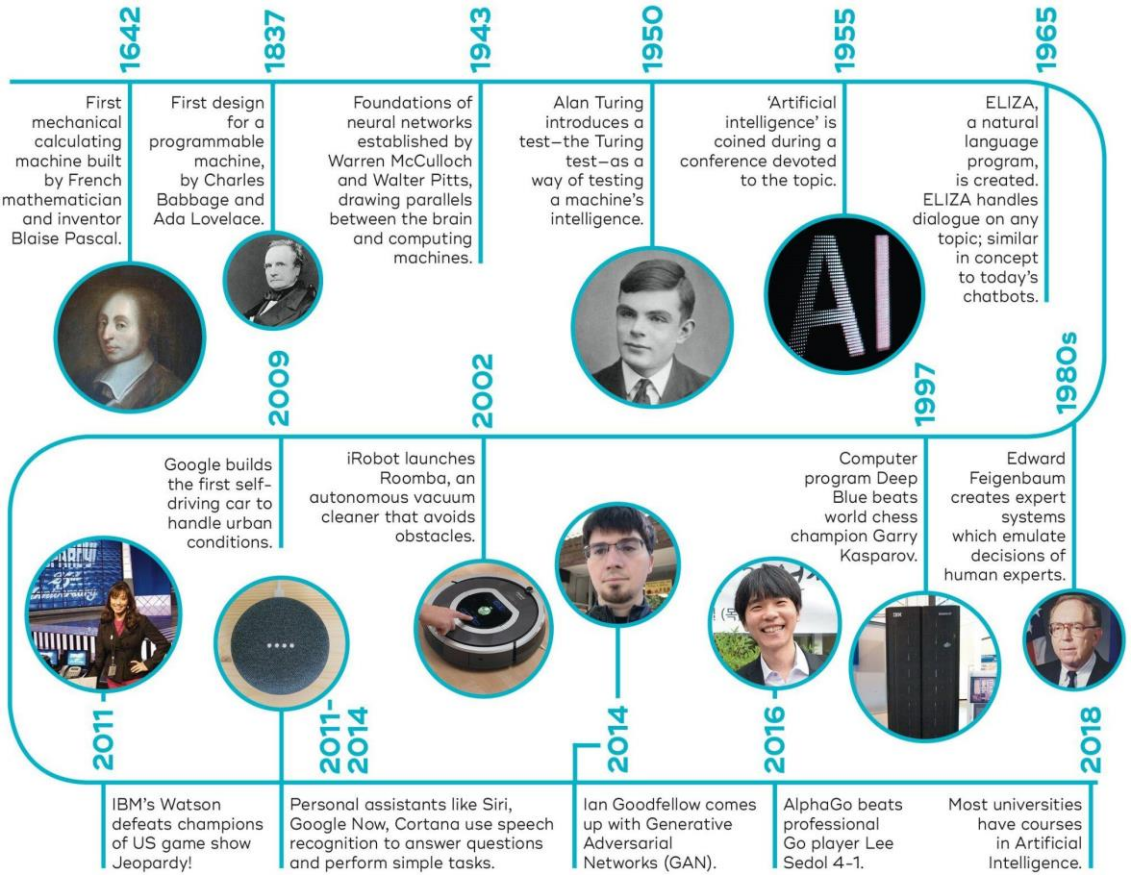


図1：AIの歴史 [6]

2018: BERT： Googleにより導入されたこのモデルは、言語理解に革命をもたらした。BERTはTransformer アーキテクチャと大規模なテキストデータセットに基づく事前学習を採用し、さまざまな言語タスクにおいて従来のモデルを凌駕する性能を実現した。

2019: 15億のパラメータを持つGTP-2

2020: 1,750億〜5,300億のパラメーターを持つLLM

2021: トレーニングの効率向上や高度な推論と事実の正確性を伴う複雑なタスクの処理に焦点を当てた、最大1兆のパラメーターを持つLLM

2022: ChatGTP-3が大流行

規模を超えて： 研究者たちは現在、トレーニングの効率性、人間の価値観との整合性、安全性、マルチモーダリティ（画像、音声、その他のデータタイプを取り入れること）に取り組んでいる。

この短いAIの歴史は、最も基本的な計算機から生成AIへの進化と、汎用人工知能がまだその途上にあることを示している。

AIを取り巻く状況

さまざまなAI技術が紹介され、議論されている。

Machine Learning(ML、機械学習)

Machine Learning はAIとコンピュータサイエンスの一分野であり、データとアルゴリズムを使って人間の学習を模倣し、モデルの精度を徐々に向上させることに焦点を当てている[7]。

Tiny Machine Learning(tinyML)

Tiny Machine Learningは、ハードウェア（専用の集積回路）、アルゴリズム、およびソフトウェアを含む機械学習技術とアプリケーションの分野として広く定義されている。非常に低い消費電力、通常はmW範囲以下でのオンデバイスのセンサーデータ解析を可能にし、さまざまな常時稼働のユースケースに対応し、バッテリー駆動のデバイス、例えばIoTデバイスをターゲットとしている[8]。

Deep Learning(Advanced ML、ディープラーニング)

Deep Learning は、人間の脳にヒントを得た方法でデータを処理するようコンピューターに教えるAIの手法である。 Deep Learning のモデルは、画像、テキスト、音声、その他のデータから複雑なパターンを認識し、ニューラルネットワークを使って正確な洞察と予測を行うことができる。

Generative Artificial Intelligence

Generative Artificial Intelligenceは、生データを取り込み、プロンプトが与えられると統計的にあり得る出力を生成することを「学習」する deep-learning や transformer モデルを指す。主に分類やパターン認識タスクに使用される上記の分類モデルとは異なり、Generative AIモデルはデータの合成、高次の学習パターンの一致、および予測分析に使用される。高いレベルでは、生成モデルはトレーニングデータの簡略化された表現をエンコードし、元のデータに似ているが同一ではない次のセットを予測する[9]。

Artificial General Intelligence

Artificial General Intelligenceは、AI開発のある種の考え方を表すために使用される理論上のAIの形態である。これは、人間と同等（またはそれ以上）の知能と、自らの存在を認識し、複雑な問題を学習して解決し、将来を計画することができる意識を伴うものを指す[10]。

トレーニング方法の概要

AIは以下のタイプに分類できる[11]：

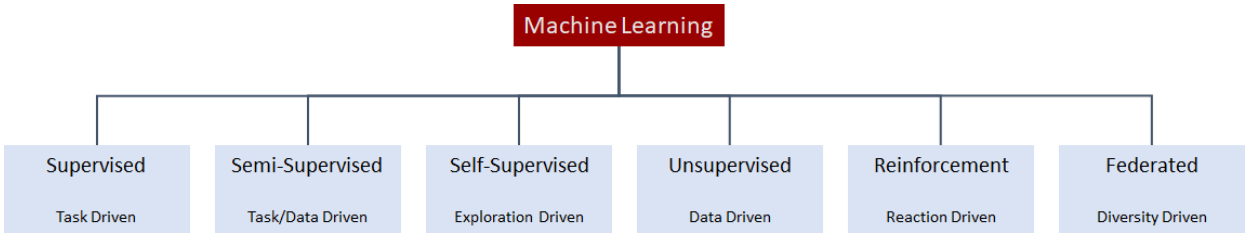


図2： Machine Learning の種類

教師あり学習(Supervised Learning)

教師あり学習は、アルゴリズムが「ラベル付けされたデータ」から学習する機械学習のスタイルである。分類や回帰の問題に用いられる。「ラベル付けされたデータ」は、既知の入力と望ましい出力を提供し、アルゴリズムがパターンを識別し、以前に見たことのないデータの結果を予測するモデルを構築することを可能にする。

分類アルゴリズムの例：決定木、ランダムフォレスト、線形分類器、サポートベクターマシン。

回帰アルゴリズムの例：線形回帰、多変量回帰、回帰木、ラッソ回帰。

教師なし学習(Unsupervised Learning)

教師なし学習は、アルゴリズムがラベル付けされていないデータを分析する機械学習のスタイルである。その目的は、あらかじめ決められた結果なしに、データ内の隠れたパターン、グループ化、パターン、または洞察を発見することである。適切に訓練されたモデルは、未知のデータを使って予測を行うことができる。

アルゴリズム例：k-means、k-medoids、階層的クラスタリング、Apriori、FP Growth。

強化学習(Reinforced Learning)

強化学習は、エージェントが環境と相互作用し、試行錯誤を通じて学習する機械学習のスタイルである。エージェントはその行動に基づいて報酬やペナルティを受け取り、時間の経過とともに行動を調整し、意思決定プロセスを最適化することができる。

アルゴリズムの例：強化学習、マルコフ決定過程、Q学習、政策勾配法、 Actor-Critic など。

半教師あり学習(Semi-supervised Learning)

半教師あり学習は、教師あり学習と教師なし学習のギャップを埋めるものである。半教師あり学習は、ラベル付けされた少量のデータと、ラベル付けされていない大量のデータを利用する。このアプローチは、ラベル付きデータの入手にコストや時間がかかる場合に有効で、ラベルなしデータからも発見されたパターンを活用することができる。

自己教師あり学習(Self-supervised Learning)

自己教師あり学習は、モデルが生の入力データから自らラベルを生成する、教師なし学習の一形態である。これは、文章内のマスクされた単語を予測する、またはビデオシーケンスの次のフレームを予測するなどの手法を通じて実現される。これにより、人間が提供するラベルなしでも、データの堅牢で汎用性のある表現を学習することが可能となる。

連合学習(Federated Learning)

連合学習は、分散されたデバイスやサーバーに保持されているローカルデータサンプルを用いてアルゴリズムをトレーニングするために設計された高度な機械学習手法である。この方法は、データを中央サーバーに転送して処理するのではなく、ユーザーのデバイス上に機密データを保持することで、プライバシーやセキュリティ、データの集中化に関する重大な懸念に対処する（図3）。2016年に導入されたこの方式では、データではなくパラメータのみを共有することで、データのプライバシーをより高度に保護することができる。連合学習は、別々のクライアントに保存されたデータセットを使ってグローバルモデルを共同で学習するフレームワークを提供する。これは、元のデータを復元することが不可能であると考えられるため、プライバシーが重要な業界にとっては良い選択肢となる[12]。

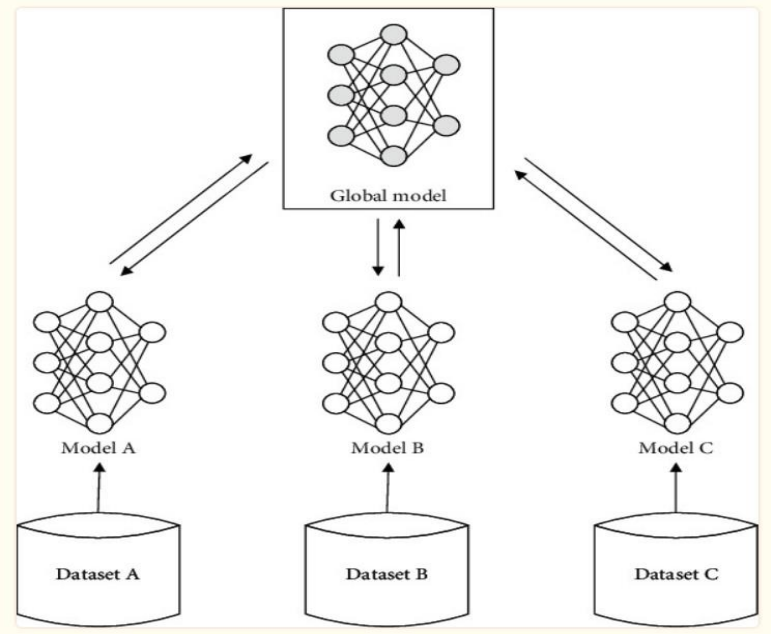


図3：連合学習(Federated Learning)システムのアーキテクチャ[12]。

このモデルのもう一つの利点は、上記の論文[12]では議論されていないが、「群衆の知恵(wisdom of the crowd)」[13]を活用できることである。人間の脳の神経細胞はこの概念を応用して、非決定論的な神経プロセスから正確な情報を作り出している。

現在、連合学習は、多様性に基づくプライバシー、パフォーマンス、堅牢性を統合する可能性を秘めているようだ。この学習方法はあまり議論されていないが、データのプライバシーと機密性が最重要視される業界やアプリケーションでは有望であり、データの残留性に関する問題にも対処できる。

しかし、連合学習には、悪意のあるユーザーがモデルの集計を妨害し、モデルの精度に影響を与えたり、プライバシーの開示につながったりするリスクなど、プライバシーに関する潜在的な懸念もある。攻撃は、学習中に共有されたモデルの更新をターゲットにすることができ、生の学習データの抽出を可能にする可能性がある。このような懸念に対処するため、研究者はデータを保護し、悪意のあるアクターからの異常をフィルタリングするために、ディファレンシャル・プライバシー、分散暗号化、ゼロ知識証明のようなプライバシー保護技術を提案している。連合学習 は、他の学習方法と同様に、適切なサイバーセキュリティ対策が必要である。

トレーニング方法の規則と倫理的考慮事項

MLトレーニングに関する特別な規制はないが、一般データ保護規則（GDPR）、EU AI法、AIに関する経済協力開発機構（OECD）の原則など、主要な規制の枠組みに影響を受けている。さらに、機械学習（ML）および人工知能（AI）のトレーニングに関する規制は、技術の進歩に伴い急速に進化しており、「“Principles to Practice: Responsible AI in a Dynamic Regulatory Environment”」で取り上げている。さらに、多くの政府機関が積極的に規制を策定し、同様の効果を得るために業界の協力的な取り組みを促進している。

機械学習（ML）と人工知能（AI）を管理する規制は、データの収益化とビジネス上の意思決定を導くためのAIの利用に重大な影響を及ぼす。こうした影響は、運用の変更、戦略的調整、倫理的配慮、さらにはデータの収集と使用、バイアス、データの品質に関する制限や要件など、さまざまな形で現れている。例えば、特定のプラットフォームは、AIのトレーニング目的でのデータの使用を禁じている（例えば、X：「事前の書面による同意がない限り、いかなる形であれ、いかなる目的であれ、本サービスをクロールまたはスクレイピングすることは明示的に禁止されている」[14]）。あるいは、ライセンス契約に基づいてデータを販売している（Reddit [15]など）。

規制要件を満たすことは、特に複数の管轄区域にまたがって事業を展開する企業にとって、多大なコンプライアンスコストをもたらす可能性がある。課題はあるものの、規制はチャンスでもある。規制の状況を巧みに乗り切る企業は、より安全で透明性が高く、倫理的なAIソリューションを提供することで差別化を図ることができる。これは、プライバシー意識の高まる消費者やパートナーにアピールすることができ、新たな市場を開拓したり、より強固な顧客ロイヤリティを生み出す可能性がある。

AI 技術のライセンス、特許、著作権

多くの機械学習フレームワークやライブラリは、Apache 2.0[16]やMIT[17]などのオープンソースイニシアティブのライセンスに従っている。特定のライセンスは、結果として得られるアプリケーションの商用利用を禁じている場合がある。

欧州特許庁（EPO）の審査ガイドライン改訂版[18]、[19]が公表され、MLやAIの分野におけるイノベーションを審査するEPOの手続きにいくつかの重要な変更が加えられた。最近の改正では、AIや機械学習（ML）に関する発明の出願者に対し、特許請求の範囲全体にわたって発明の技術的結果を再現できるよう、数学的手法やトレーニング入力データを十分に詳述することが義務付けられている。以下に引用する論文では、「判例法は、使用されるニューラルネットワークの構造、トポロジー、活性化関数、終了条件、学習メカニズムはすべて、出願が開示する必要がある可能性のある関連する技術的詳細であることを示唆している」と述べている。この論文[20]は、このトピックについてさらなる示唆を要約し、解説している。

2024年1月23日、文化庁(ACA)は、日本国内における著作物の摂取・出力の在り方を明確にするため、「AIと著作権に関する考え方について」という文書の案をパブリックコメントに付した。2024年2月29日、約25,000件の意見を検討した結果、追加修正が行われた。ACAの委員会によって作成されたこの文書は、おそらく数週間以内にACAによって採択されるであろう。本記事[21]では、草案そのものと修正された部分の要点をまとめている。

シンガポールでは著作権に関する紛争が起きている[22]。これは現在非常に不安定な分野である。“Principles to Practice: Responsible AI in a Dynamic Regulatory Environment”」でも取り上げられている。

パート II: 現実世界でのケーススタディと業界の課題

このパートでは、いくつかの業界を例に挙げながら、AIの時代に直面する課題の概要と現状に焦点を当てている。法規制に関するより詳細な情報は、「Principles to Practice: Responsible AI in a Dynamic Regulatory Environment」にまとめられている。

AIの歴史 ケーススタディ

このAI事例の短い歴史は、少なくともAI技術の前身を非常に早くから採用していた金融業界については、1990年代後半まで遡る。このセクションでは、2016年から本稿の出版までの間に、AIの現実の応用における主要な課題を示すいくつかの事例を紹介する。これらは、生成AIにおけるバイアスが最大の懸念であることを示している。

2016:マイクロソフトの Tay

マイクロソフトのAIチャットボット「Tay（テイ）」は、当初は遊び心のあるTwitterでの会話を想定していたが、リリースからわずか24時間で人種差別的で攻撃的な発言のプラットフォームに早変わりした。ユーザーは性差別的で扇情的なコメントをTayに殺到させ、ボットはこれらの感情を反響させた。一部のツイートはユーザーによって誘発されたものもあれば、促されることなく発生したものもあり、Tayの不規則な行動を示している。マイクロソフトは、不快なコンテンツを削除し、Tayの反応を調整する必要性を認識することで対応した。この事件は、AIが公開データから学習し、社会の偏見を反映することの難しさを浮き彫りにした。不適切なツイートが原因で再リリースされ、その後シャットダウンされたにもかかわらず、Tayの遺産にはマイクロソフトのAI開発における教訓が含まれている。このエピソードは、AIアプリケーションの複雑さに光を当て、AI設計における反復的改善と事前対策の必要性を強調している[23], [24], [25]。

2018: アマゾンのAI採用ツールにおける女性に対するバイアス

アマゾンは機械学習による採用エンジンを開発したが、学習データに埋め込まれた根本的なジェンダーバイアスにより、男性候補者を優遇してしまうという問題に直面した。アマゾンは、不公平を懸念して2017年にプロジェクトを解散した[26], [27], [28]。挫折にもかかわらず、他の企業では採用プロセスにおけるAIを慎重に進め、現在では主流となっている。

2019年：テスラ オートパイロット事故

2019年3月、ジェレミー・バナーはテスラ・モデル3でオートパイロットを作動させ、セミトラックと衝

突して死亡した。この事件をきっかけに、テスラのオートパイロット技術が関与した事故におけるテスラの責任を問う法的論争が起こった。批評家たちは、テスラのオートパイロットのマーケティングは誤解を招くと主張している。

ドライバーはオートパイロットの能力について理解できず、事故や死亡事故を引き起こす可能性がある。オートパイロットの限界に関する警告にもかかわらず、死亡事故を含む数多くの事故が発生している。

高度な運転支援システムに関する明確な規制ガイドラインがないことや、その使用をめぐる倫理的ジレンマが状況をさらに複雑にしており、自律走行技術を導入する際の説明責任、保険、公共の安全について疑問が投げかけられている[29]。

2019年：ヘルスケア・アルゴリズムにおける人種バイアス

Ziad Obermeyerらによる研究論文[23]は、何百万人もの患者に影響を与える、広く使われているヘルスケアアルゴリズムに著しい人種的偏りがあることを明らかにしている。医療ニーズの管理を目的としたこのアルゴリズムは、人種を変数として除外しているにもかかわらず、白人患者と比較して黒人患者の健康リスクを不正確に予測している。このバイアスは、アルゴリズムが健康ニーズの代理として医療費に依存していることに起因しており、医療へのアクセスと利用における制度的不平等を不注意にも反映している。これを改善すれば、追加的なヘルスケア支援が必要であると認識される黒人患者の数が大幅に増加すると考えられる [30] 。

2019年：アップルカードの性差別疑惑

ある著名なソフトウェア開発者が、アップルカードのクレジット枠が男性と女性で異なっていることに注意を喚起したことがきっかけで、Twitterのスレッドが拡散し、ゴールドマンサックスのクレジットカード業務に対する規制当局の調査へと発展[31]。「アップルカードバイアス」として注目された：[32], [33], [34], [35].
しかし、2021年には、「最近終了したニューヨーク州金融サービス局の調査[36]により、アップルの銀行パートナーは性別による差別を行っていなかったことが判明した[30]」と報告された。

2020:偏った犯罪者評価システム

COMPAS（Correctional Offender Management Profiling for Alternative Sanctions：米国）やOASys（Offender Assessment System：英国）のようなツールは、犯罪者のリスク評価と管理のために刑事司法制度で使用されている。これらのシステムは、犯罪者に対する量刑、保護観察、治療プログラムについて、当局が十分な情報に基づいた決定を下すのに役立っている。しかし、そのアルゴリズムは、透明性、公平性、偏りに関して、大きな批判にさらされている [38] 。

2022: エア・カナダ、チャットボットによる払い戻しに拘束されるポリシー

エア・カナダは、同社のチャットボットが航空会社の身内に不幸があった場合などの忌引時の割引ポリシーについて誤解を招く情報を提供し、払い戻しを求める乗客との紛争に発展したことから、精査に直面した。エア・カナダは、チャットボットは独自に運営されていると主張したにもかかわらず、裁判所は乗客に有利な判決を下し、ウェブサイトに提供された情報に対する航空会社の責任を強調した。裁判所はエア・カナダに対し、一部払い戻しと追加費用の負担を命じた。この事件は、AIの説明責任と顧客サービスの自動化の複雑さを浮き彫りにしている[39], [40]。

2023:訴訟：ユナイテッドヘルスの欠陥**AI**が高齢者のケアを拒否

世界最大のヘルスケア企業であるユナイテッドヘルスに対する法廷闘争で、家族は欠陥のあるAIの使用により、医師の勧めを無視して高齢患者に必須治療の保険適用が拒否されたと主張している。この訴訟は、医療の意思決定を自動化されたシステムのみに依存することの課題を浮き彫りにし、患者の幸福と医療サービスへの公平なアクセスに対する懸念に火をつけた。AIがヘルスケアの未来を形成し続ける中、この訴訟は、すべての患者に対する公平な治療を保証する上で、透明性、説明責任、人間の監視の必要性を強調している [41], [42]。

2024: Google Gemini：**AI**のバイアスにおける教訓

グーグルのGemini 1.5チャットボット[43]は、バイアスを避けようとしているにもかかわらず、不正確で偏った画像を生成していると批判された。イーロン・マスクと保守派はグーグルのアルゴリズムが偏っていると非難した。グーグルの対応はGemeniを一時停止させたが、透明性に欠けていた。この事件はAIの倫理と透明性の課題を浮き彫りにし、多様性への取り組みとアルゴリズムの説明責任に関する議論を促した。グーグルが信頼回復に取り組む中、Geminiの悲劇は、責任あるAIイノベーションの必要性を強調している[44]。

産業における規制と課題

本セクションでは、AIに関連する業界特有の規制やコンプライアンスに焦点を当てた取り組みを紹介する。業種はアルファベット順に列挙し、個別に取り上げる。各業界のパートでは、その背景、文脈、歴史が説明されている。パートIIIでは、これらに続き、業界横断的にAIに取り組むための斬新なアプローチについて提案する。

自動車

自動車業界[1]は、主に自動運転や自律走行機能（SAEレベル4および5 [45]）にAIを実装しようとしており、その安全性やその他の車載システムやコンポーネントにも重点を置いている。現在、AIについて言及または部分的に規制しているISO規格がいくつか存在する。現在、AIに言及したり、部分的に規制したりするISO規格がいくつか存在する。現在、数多くの規制機関が自動車業界に特化した基準やアプローチを策定しているが、まだ施行されていない。

[1] 本章では、筆者の専門分野と所在地から、EUの自動車産業に焦点を当てる。

現行の法律はすでにAIに暗黙のうちに影響を与えているが、一部の規制機関はこのような技術に直接言及している。これには、*2019年11月27日付欧州議会及び理事会規則（EU）2019/2144（自動車及びそのトレーラー、並びにそのような自動車を意図したシステム、部品及び別個の技術ユニットに対する型式承認要件に関する規則（EU）2019/2144*）が、*その一般的安全性及び車両乗員及び交通弱者の保護に関して*含まれている[46]。第11条（自動運転車および完全自動運転車に関する特定の要件）は、AIが自律走行車や自動運転車の運転に使用される場合、AIに関連する安全システムの要件を定義している。この規制条項はAIに特化していないが、明確に "自動運転車と完全自動運転車 "に言及している。

この欧州連合法は、AIとその機能そのものに関する具体的な基準を対象としていないが、ISO PAS 8800： Road Vehicles – Safety and Artificial Intelligence （道路運送車両 - 安全性と人工知能 ）[47] （検討中）は、AIの安全性に焦点を当て、"安全原則の方法と証拠 "を起草している。道路運送車両全般が対象であり、自動運転や自律走行に絞られていない。その目的は、AIの規制と標準化に関する基礎的な問題を解決し、意図された機能の安全性のように、既存の規制と確立された原則を調和させることを意図した業界固有の実用的なガイドライン[48]を提供することである。

AI（機能）安全に関するもう一つの業界標準は、ISO/TR 5469:2024 Artificial Intelligence - Functional safety and AI systems（ 人工知能 - 機能的な安全とAIシステム） [49]である。2024年に発行されたこの文書では、以下のようないくつかの自動車アプリケーションにおけるAIに関連するリスク要因、および現在利用可能な方法とプロセスが記述されている。
AI と非 AI システムを利用する安全関連機能により、AI 安全システムを管理する。この規格は発行済みであり、将来のISO/IEC AWI TS 22440 [50], [51]をサポートすることを意図している。セキュリティに重点を置いたものとしては、ISO/TR 4804:2020 Road vehicles - Safety and cybersecurity for automated driving systems, design, verification, and validation（道路運送車両 - 自動運転システムの安全性とサイバーセキュリティ、設計、検証、妥当性確認） [52]があり、自動運転（SAEレベル3および4）のシステムの開発と検証に焦点を当て、サイバーセキュリティの側面にさらに重点を置いている。SO/TR 4804:2020は、世界的に適用可能な安全性、妥当性確認、検証のアプローチを主に含んでいる。将来的にはISO/CD TS 5083に置き換えられる予定である。このISO文書は、「安全な自動運転システムを搭載した自動運転車の開発および検証のためのステップ」を扱っており、SAEレベル3および4にとどまっている。この文書には、人間のドライバーに比べて全体的なリスクを低減しつつ、そのようなシステムに要求される安全レベルなどの考慮事項が含まれている。

航空

世界の航空業界は、他の業界と同様に、コンピュータベースのシステム使用に関する実用的な基準の多くを遵守している。これはAIやAIを実行するプラットフォームにも及んでいる。そのためによく知られたIT セキュリティ標準は、航空部門にも適用される。これには、ISO/IEC 27001 [53]、ISO/IEC 42001 [54]、ISO/TR 5469 [49]、NIST AI RMF [55]、航空会社やメーカーの管轄に関連するAI倫理基準などが含まれる。しかし、AI固有の規制はまだ施行されていない。

米国連邦航空局（FAA）、欧州航空局（EASA）、英国航空局（CASA）、オーストラリア航空局（AU CASA）など、世界中の航空業界を管理する機関は、AIが航空業界にもたらすメリットと課題を認識している。しかし、現時点ではその使用を規制していない。前述の各機関はAIタスクフォースを結成し、航空機、地上業務、業界の規制自体におけるAIの使用について調査している。英国CASAは、現在、業界に対する公開調査による回答依頼の段階にあり[56]、米国FAAは、Trung t. Pham博士が率いるAIの技術規律

チームを指定している[57]。

AIは、情報データ分析や自律走行車から、飛行場や航空基地の予知保全や物理的セキュリティに至るまで、多くの軍事航空分野にわたって使用されている。また、ITセキュリティや運用システムの管理にも利用されている[58]。

全体として、民間航空業界では、気象計画や経路設定、整備、旅客・貨物管理などを支援するためにAIを使用したいという強い要望がある。提案されているAIの利用のほとんどは、予測的メンテナンス、ルートとメンテナンス計画、旅客・貨物管理のための機械学習を中心に展開されている。ジェネレーティブAIの利用は、航空会社の顧客向けチャットボットや意思決定支援システムに限られている。しかし、すべての飛行フェーズにおける航空交通管制にAIを活用する研究は、あらゆる分野で進められている。

EUは2022年10月に発表した航空交通管理におけるAIに関するCORDIS Results Packを、欧州のAI制御航空交通の多くの側面を網羅している[59]。

航空業界にとって特別な課題は、民間旅客機の寿命が通常数十年単位であるのに対し、AI技術は頻繁に飛躍を遂げるため、規制を継続的に更新する必要があることだ。

重要インフラストラクチャー&エッセンシャルサービス

AIを重要なインフラに組み込むことは、より効率的で応答性の高い、インテリジェントなシステムへの大きな転換である。電力、ガス、水道、食料サプライチェーンなど、これらに限定されないこれらの分野は、現代社会の存続に不可欠である。このようなデジタルトランスフォーメーションを受け入れるにつれ、パフォーマンスの向上とセキュリティの堅牢性のバランスを取ることがますます複雑になってきています。
本セクションでは、AIと重要インフラの融合がもたらす課題と機会について、規制の枠組み、セキュリティ基準、進化する技術進歩への継続的な適応の必要性の重要性に焦点を当てながら探っていく。

微妙なバランス：パフォーマンスvsセキュリティ

重要インフラにおける高性能AIシステムの魅力は否定できない。これらの技術は、効率性の向上、運用の最適化、混乱が発生する前にそれを予測し緩和する能力を約束する。しかし、主にtinyML [8]やエッジコンピューティングを統合したモノのインターネット（IoT）デバイスを通じてAIを統合することは、新たな脆弱性をもたらす。システムの応答性には有益だが、データ処理の分散化は潜在的なサイバー脅威の攻撃対象領域を拡大する。国際標準化機構（ISO）や国際電気標準会議（IEC）などの規制機関や標準化組織は、ISO/IEC 27001、ISO/IEC 27002、および産業オートメーションと制御システムに焦点を当てた ISA/IEC 62443 [60]シリーズのような一般的なフレームワークや、IEC TS [60]のような技術仕様書や報告書を開発してきた。62351-100-4:2023[61]とIEC TR 61850-90-4:2020[62]は、これらの技術を保護するためのものである。しかし、重要インフラにおけるIoTとエッジAIを対象とする規制の具体性は、依然として曖昧なままである。

アキレス腱：IoTとエッジ AI

重要なインフラ部門にIoTデバイスを統合すると、サイバー攻撃のリスクが顕著になる。AIシステムのセンサーネットワークに不可欠なこれらのデバイスは、偽のデータを送り込むために悪用される可能性が

あり、それによってAI主導の意思決定が操作され、重要なサービスが機能不全に陥る可能性がある。EUのAI法、NIS2指令、AIに関する米国の大統領令など、サイバーセキュリティとリスク管理に関する現行の規制が一般的な焦点となっているにもかかわらず、エッジAIやtinyMLデバイスがもたらす独自の課題に対処するための包括的なアプローチはまだ発展途上である。ENISAの "Cybersecurity and privacy in AI - Forecasting the demand on electricity grids(AIにおけるサイバーセキュリティとプライバシー - 電力網の需要予測)" [63]のような業界研究や報告書は、潜在的なリスクや脅威を明らかにしようとしている。

将来を見据えた インフラへ

重要インフラにおけるAIの安全確保には、いくつかの重要な戦略が必要である。第一に、各重要インフラ部門固有のニーズに対応する部門別AI規制の策定が不可欠である。さらに、IoTデバイスとエッジAIに標準化されたセキュリティプロトコルを採用することは、これらのシステムをサイバー脅威から強化する上で極めて重要である。また、AIガバナンスに関する国際協力は、重要インフラを世界的に保護するためのまとまりのある効果的なアプローチを確保する上で極めて重要な役割を果たす可能性がある。2022年、The Water Research Foundation（WRF）とWater Environment Federation（WEF）の共同作業であるLeaders Innovation Forum for Technology（LIFT）プログラムの受賞者は、最先端のAIとデータサイエンス技術を使用して、（サイバーセキュリティの脅威に対する）保護、（システム状態の）予測、および施設内のプロセスの最適化に焦点を当てた[64]。

2024年3月1日、米国大統領科学技術諮問委員会（PCAST）は、「Fortifying Our Critical Infrastructure for a Digital World（デジタル社会に向けた重要インフラの強化）」[65]と題する報告書を発表した。この報告書では、AIの進歩の二面性、変革的な応用の可能性と悪用されるリスクを強調している。また、重要インフラの回復力に対するAIの影響について専門的な分析の重要性を強調し、AIが悪意のある行為者に力を与える可能性と、そのような脅威に対する戦略的な備えの必要性を指摘している。さらに、PCASTは防衛メカニズムにおけるAIの活用を提唱し、AIとサイバーセキュリティの課題に効果的に対処するため、官民協力、従来の範囲を超えた能力開発、国際協力を呼びかけている。

絶え間ない進化：その道

AIを重要インフラに組み込むことは、イノベーションとセキュリティのバランスを細心の注意を払って維持しなければならない、永遠の綱渡りのようなものだ。AI技術が進化するにつれ、それを支える規制やセキュリティの枠組みも進化しなければならない。そのためには、ガバナンスとセキュリティに対する警戒心と適応力のあるアプローチが必要であり、新たな脆弱性の出現に合わせて防御も進化するようにしなければならない。
ゼロトラストの原則を導入することで、「決して信用せず、常に検証する」という原則を受け入れ、重要インフラシステムが、高度な攻撃者が迂回できる境界防御に過度に依存しないようにする。AIを活用したセキュリティシステムは、従来のセキュリティ対策よりも迅速に進化する脅威に対応することができる。脅威インテリジェンスとベストプラクティスを部門内および重要インフラ事業体間で共有することで、より広範な知識と経験のプールから利益を得ることができる。新たな脅威に対抗するための最新の知識とツールをセキュリティの専門家に提供するためには、継続的なトレーニングと専門能力の開発が不可欠である。これらは、強固なAI統合を実現するために私たちが取らなければならないいくつかのステップに過ぎない。

前途

重要インフラへのAIの統合は、大きな可能性と大きな課題を伴う旅である。パフォーマンス向上と強固なセキュリティの両立は、複雑だが不可欠な取り組みである。的を絞った規制の策定、標準化されたセキュリティ慣行の採用、そして国際協力を通じて、私たちはこのフロンティアを安全に航海することができる。重要インフラの未来は、AIの利点を活かしつつ、次のようなことを実現できるかどうかにかかっている。リスクから身を守り、レジリエントで効率的かつ安全な社会基盤を確保する。実際の意思決定や業務には人間が介在し、AIを推薦役として活用することが不可欠である。

現在のところ、具体的な規制は限られている。AI規制は世界的に進展しているが、重要インフラにおけるAI利用に焦点を当てた具体的な規制はない。一般的な原則に焦点が当てられている。既存の規制やイニシアチブは、サイバーセキュリティ、安全性、信頼性といったより広範な原則を重視している。

いくつかの政府機関が、重要インフラ分野に焦点を当てながら、責任あるAIの開発と展開のためのフレームワークと基準を積極的に開発している。

現在のイニシアチブ

米国大統領令14110号 (2023年10月)

人工知能の安全、安心、信頼できる開発と使用に関する大統領令14110（2023年10月）：重要インフラとサイバーセキュリティにおけるAIの管理に関する第4.3条」[66]。この計画は、重要インフラにおけるAIのリスクを評価し、低減するための行動を概説している。この計画では、安全ガイドラインの作成、AI安全・セキュリティ委員会の設立、インフラ所有者と運営者に対する規制の強化が優先されている。サイバーセキュリティ・社会基盤安全保障庁（CISA）は、重要インフラに対するAIの脅威を評価、軽減している[67]。

EUのAI法

EUのAI法[68]、[69]は、人工知能（AI）システムの規制枠組みを概説しており、コンプライアンス、リスク管理、データガバナンス、技術文書、記録保持、透明性、人的監視、正確性、堅牢性、サイバーセキュリティ基準に焦点を当てている。AIシステムはリスクに基づいて分類され、重要インフラにおける「高リスク」アプリケーションはより厳しい規制の対象となる。2024年3月13日、EU議会はAI法を承認した[70]。

OECD AI 原則

経済協力開発機構（OECD）の原則1.5[71]は、特に社会的影響を及ぼす可能性のあるAIシステムの開発、展開、使用について、行為者が説明責任を負うことの重要性を強調している。重要インフラは間違いなく社会に大きな影響を与える分野であり、この分野で使用されるAIに対する説明責任を確保することは、この原則に合致する。これらの原則は、国内および国際的な議論を形成する上で影響力を持つ。

人工知能およびデータ法 (AIDA)

AIDA [72]は、カナダにおけるAIのイノベーションと責任ある利用を導くことを目的としている。AIDAは、AIシステムが安全な方法で開発・利用されることを保証し、AIが個人と経済に与える影響を管理する規制システムの必要性に対処するものである。さらに、影響力の大きいAIシステムに関連する危害や偏った結果に対する保護と緩和を強調している。AIDAは、様々な利害関係者の役割、AIに関わる企業の義務、コンプライアンスを確保するための執行メカニズムについて概説している。

AIとジェネレーティブAIを取り巻く既存の取り組み、法律、規制の状況についての詳細な分析は、CSAの「Principles to Practice: Responsible AI in a Dynamic Regulatory Environment」ペーパーをご覧いただきたい。

防衛

防衛における人工知能と新技術

将来の戦場では、物理的領域とデジタル領域が絡み合い、複雑で争いの絶えない環境が生まれる。新たな脅威と課題が出現し続けるだろう。AIは、優位性、状況認識、インテリジェンス、意思決定の向上を獲得・維持するための重要な手段となる。ロボット工学、自律システム、データ、バイオテクノロジーなどの技術は、防衛軍に新たな機会とリスクをもたらすだろう。AIは、これらの技術を統合・活用し、敵対勢力の利用に対抗するために不可欠となる。軍は、イノベーションを促進し、採用を促進し、新たなリーダーシップと文化の役割を開発するために、民間企業・学界・同盟国とのパートナーシップを追求する必要がある。AIは、領域、プラットフォーム、組織を超えたコラボレーションとコミュニケーションを促進し、人間と機械のチーム編成と学習を可能にする。

これらはすべてAIの規制と枠組みに収斂され、防衛におけるAIの倫理的で、安全で、信頼できる利用への配慮が最も重要である。規制と枠組みの両方は、公平な競争条件を作り出し、イノベーションと協力を促進し、国民の信頼と受容を高めることによって、防衛産業を前進させることができる。

- AIの規制や枠組みは、防衛におけるAIシステムの開発・配備・使用に関する基準やベストプラクティスを定義するのに役立ち、AIの偏見・危害・誤用のリスクを低減するとともに、AI関係者の説明責任や透明性を高めることができる。
- 共通の市場と競争上の優位性を生み出すことで、防衛産業を刺激することができる。防衛施設全体の規則と要件の調和は、防衛・非防衛分野における国境を越えた協力とAIシステムの相互運用性を促進することができる。
- 防衛におけるAIの社会的信頼と受容は、AIの規制と枠組みによって強化される可能性がある。

防衛における**AI**の歴史的役割

アラン・チューリングがその基礎となる著作「計算する機械と知性」[73]を発表し、今日のようなAIが誕生して以来、最初の投資と最初のユースケースは国防によって推進された。音声をテキストに書き起こす自然言語処理（NLP）と、テキストを分析して人間の推論を模倣する機械学習の利用は、急速に拡大している。最初の投資は国防、主に米国国防高等研究計画局（DARPA）によって行われ、いくつかの機関でAI研究に資金が提供された。

技術が進歩するにつれて、コンピューティングパワーは増大し、データコストは低下し、新たなユースケースが出現した。1970年、初めて半自動化された戦争が、AIモデルにき供給するセンサーの投下によって達成され、標的の決定・資源の配分・任務の予定が立てられた。1980年代には、これはスマート兵器、シミュレーション、意思決定支援へと発展した。

2018年、米国防総省は**AI**が **"**将来の戦場の性格を変える準備ができている**[75]"**と宣言した。**2018**年、国防総省は合同人工知能センター（**JAIC**）と人工知能に関する国家安全保障委員会を発足させた。米国は**AI**に**10**億ドルを投資し、その数倍を自律型システムや無人システムのような**AI**の利点を活用するシステムに投資することで、その発言を裏付けた。中国も同様の動きを見せ、**2030**年までに**AI**で世界をリードしたいと宣言した。ロシアのウラジーミル・プーチン大統領が「この分野でリーダーになる者は、世界の支配者になる**[76]**」と予言したことは有名である。

生産性が向上し、意思決定が自動化され、さらに洞察が深まることで、戦争があまりに速く起こりすぎて人間が介入できないような状況が生まれるのだろうか？AIへの大規模な投資はAI軍拡競争を引き起こすのか？

AIが魅力的である理由がわかった。自律的な機械やロボットは安価で、交換が可能だ。人間はそうではない。オートメーションは疲れないし、サポートに多額のサプライチェーンを必要とせず、人間の特性を全て備えているわけではない。

その一方で、注意しなければならないこともある。機械に生死を左右する決断をさせることは、悲惨な結果を招きかねない。科学技術政策局（OSTP）の局長であり、DARPAの元局長であるA. Prabhakar博士の言葉を借りれば、「AIに何が起こっているかを見ると、非常に強力であることがわかるが、同時にまだかなり限定的な技術であることもわかる。問題は、それが間違っているとき、人間にはあり得ないような方法で間違っているということだ[77]」。

かつてない規模の副次的な被害を生むのか？自律型戦争マシンは、起こるのを待っていた戦争犯罪なのか？

防衛システムで使用されるAIは、明確に定義された環境で単一のタスクに焦点を絞ったものにとどまっている。例えば、画像識別、艦船を防衛するための速射砲、明確に定義された徳地を探して長時間滞空するミサイルなどである。

AI規制と防衛

防衛分野におけるAIの応用は、その意味合いと潜在的な致死性から異なる。標準化団体や規制当局によって作成されたほとんどの著作物は、防衛部門にとって重要なユースケースをほとんど論じていない。防衛部門は、防衛部門に合わせた類似の作品を作るために、パブリックドメインで利用可能なものに目を向けることが多い。これらの作品へのアクセスは、(1)私たちに危害を加えようとする悪意ある行為者に洞察を提供しないこと、(2)敵対者が他者が開発した知的財産(IP)から学ぶことができること、のために制限されることが多い。テクノロジーは民間部門と防衛部門の境界線を曖昧にしている。一方では、防衛部門は国防のために秘密保持を必要とし、厳しい規制が制約するような機敏さを必要とする。他方で、防衛分野で誤ったAIがもたらす潜在的な害は、非防衛分野で起こりうることを容易に上回る可能性がある。防衛用途でのAIの使用は、特に自律的な意思決定と人間による監視のバランスに関して、倫理

的、セキュリティおよび安全保障的な問題を容易に引き起こす可能性がある。明確なガイドラインや適応可能な交戦規則が欠如していると、誤用や意図しない結果を招く危険性がある。

欧州連合（EU）には、一般に公開されている防衛に特化したAI規制は存在しない。

「防衛のためのAI法」と呼べるような具体的な規制はないが、北大西洋条約機構（NATO）はAI導入を加速させることに焦点を当てたAI戦略を掲げている。この戦略では、主要なAIイネーブラーを強化し、防衛用途におけるAIの責任ある倫理的な使用のための方針を採択している。米国国防総省（DOD）は2023年2月、"人工知能の責任ある軍事利用と自律性に関する米国政治宣言 "を発表した。立法規則ではないが、軍が新興のAI技術を責任を持って使用するための宣言である。

教育

教育分野における人工知能（AI）の統合は、学習成果を高め、教育の不平等に対処する機会を提供する。適応学習システム、AIチューター、予測分析などのAI技術は、多様な学習ニーズに合わせた個別化教育の可能性を秘めている。しかし、このような統合は、AIシステムに関わる広範なデータ収集と処理のために、プライバシーとデータ保護に関する懸念を引き起こす[78]。教育におけるAIの倫理的利用を確保するためには、ガバナンスの枠組みは、人間の監視と法的・倫理的基準の遵守を優先しなければならない[79]。

教育におけるAI技術の導入において、公平性とアクセスは極めて重要な検討事項である。教育機関や政策立案者にとって、AIツールが既存の格差を悪化させるのではなく、エンパワーメントのための道具として機能するようにすることが不可欠である。AIアルゴリズムにおけるバイアスのリスクに対処することは、差別的な結果を防ぐために不可欠であり、透明で包括的な開発プロセスの必要性を強調している。教育において倫理的に整合したAIシステムの開発を導くためには、教育者、技術者、倫理学者、政策立案者の協力が必要である[78]。

AIイニシアティブを地域社会の価値観や期待に合致させるためには、生徒、保護者、教育者を含むステークホルダーとの継続的な対話が不可欠である。さらに、教育エコシステムの参加者全員を対象としたデジタルリテラシーとAI教育に投資することは、AIテクノロジーに効果的かつ批判的に関与するための準備として極めて重要である。倫理的ガバナンス、コンプライアンス、包括性に重点を置いてこの移行を乗り切ることで、教育部門はAIを教育の卓越性と公平性のためのツールとして活用することができる[78]。

教育へのAIの統合には、イノベーションと倫理的配慮のバランスをとる、積極的で微妙なアプローチが必要である。協力関係を育み、透明性を確保し、公平性とアクセスを優先させることで、教育セクターはすべての学習者の権利と福利を守りつつ、学習体験を豊かにするためにAIを活用することができる。

ファイナンス

金融業界は、行動傾向や頻繁な分析への洞察に加え、新しい技術動向や最終顧客のニーズに応える必要性に対応するため、国際的に認知された基準を通じて、あるいは規制当局に応じてローカルに、世界的に最も規制されている業界のひとつと考えられている。業界には何年も前から「AI規制」があるが、その認知度は限られている。これらはリスク管理のビジネスリスクのカテゴリーに属する。

なぜこのようなことになったのか？多くの人は2008年の金融危機といえばリーマン・ブラザーズの破綻[80]を思い浮かべるかもしれないが、[81]実はその10年前に起きていた。長期資本管理 (LTCM）は1994年に設立された。それは、ノーベル賞受賞の経済学者マイロン・ショールズと、サロモン・ブラザーズのようなウォール街の有名トレーダーが率いた。彼らは裁定取引を専門とする。1998年8月、ロシアが債務不履行に陥った。LTCMはこの国債で大きなポジションを持ち、数億ドルの損失を出した。それどころか、彼らのコンピューターモデルはポジションの保有を推奨していた。1998年当時は「コンピューターモデル」と呼ばれていたが、今日では機械学習モデルや人工知能と呼ばれていることを理解することが重要だ。もしLTCMが破綻していたら、彼らのポジションのシステムリスクのために、おそらく世界初の金融危機が発生していたであろうが、米国政府が介入し、36億2,500万ドルの融資を行った 。LTCMは2000年初めに清算された[82]。2005年、バーゼル銀行監督委員会は、内部格付けシステムの検証に関する研究のための新しいガイドラインを発表した [83], [84] 。これは人工知能のようには聞こえないが、格付けシステムは、アナリストや格付け機関が株式、債券、企業の信用力を評価するために使用する評価ツールの銀行用語である。格付けシステムは機械学習用語でレコメンダーシステムであるため、このような格付けシステムでディープラーニング技術を使用することは今日非常に一般的である。人工知能のコンポーネントは、銀行家やトレーダーからは見えないようになっている。

金融危機と " Great Recession"に見舞われた2011年、米連邦準備制度理事会（FRB）は、より詳細な銀行業務ガイダンスに踏み込み、資産100億ドル以上の銀行が遵守しなければならない「Supervision and Regulation Letters」SR 11-7 - モデルリスク管理に関するガイダンス [85]を公表した。世界的な住宅ローン危機の重要な一因とされたのが、モデルの構築と使用に関する問題であった。

モデルリスク管理に関するガイダンス SR 11- 7

アメリカの銀行業界では、これらのレターは新しい法律の出版物のようなものだ。任意ではない。SR11-7の文脈では、モデルとは「統計的、経済的、金融的、数学的理論、技法、仮定を適用し、入力データを定量的推定値に加工する定量的手法・システム・アプローチ」と定義されている。モデルの使用は、必ずモデルリスクをもたらす。モデルリスクとは、"不正確な、または、誤用されたモデル出力や報告書に基づく意思決定から生じる不利な結果の可能性 "である。この文書では、不正確または誤用されたモデルに基づく意思決定が（財務上の損失を含む）潜在的な悪影響をもたらすことから、積極的なモデルリスク管理の重要性が強調されている。金融機関は、効果的なモデルリスク管理の枠組みの主要な側面を持つことが求められる。効果的なモデルリスク管理の枠組みは、健全なガバナンス、方針、および堅固なモデルの開発、実装、使用、効果的な検証のためのコントロールを含まなければならない。

この文脈で米国の規制当局による調査で最も有名なのは、2019年にアップルが行った差別的なクレジットカードである（「AIの歴史　ケーススタディ」参照）。 なぜならば、SR 11-7 の「有効な検証」の部分に違反していたためだ。

もう1つの例は、2012年にナイト・キャピタル・グループが起こした4億4,000万ドルのソフトウェア障害である[86]。ナイト・キャピタル・グループは、AIのニッチ分野である高頻度取引（HFT）を専門としていた。HFTとは、ミリ秒からナノ秒のスピードで行われる株式取引のことである。要するに、HFTでは利害関係者が実際の注文を出す前に株を買う。実際の注文が証券取引所に届くと、HFT当事者は再び株を売り、わずかな価格差を維持する。このような取引は、1日の間に極めて高い頻度で行われる。一般的に、HFTは取引終了時に株式を保有することはない。2012年8月1日、ナイト・キャピタル・グルー

プの全取引サーバーにソフトウエアのアップデートが提供されなかったため、注文が誤って執行された。その結果HFTシステムは過去の株式取得を認識できず、同じ銘柄を買い続けた。その結果、148社の株価が乱れた。その結果、ナイト・キャピタル・グループの巨額の損失（株式市場で4億4,000万米ドル）がわずか45分で発生した。同社はこのDevOps問題により2012年12月に買収されたが、これはSR 11-7の「実装と使用」の期待に違反するものであった。合併は2013年7月に完了した。

銀行業務におけるAIの製品アプリケーションに加え、米国の金融機関では、詐欺の検出や銀行秘密法[87]および米国パトリオット法[88]の遵守のためにモデルも使用されている。このようなモデルは、顧客の個人取引データを識別・評価し、疑わしい活動の可能性を報告する。これらのシステムとその結果は、米国政府の反テロリズム・プログラムにおける重要な手段である。SR11-7モデルの期待に従わない場合、コンプライアンス評価が低下し、金銭的な罰金/罰則や拡大禁止を含む制裁等の厳しい規制措置がとられる可能性がある。

欧州中央銀行（ECB）は2024年2月、内部モデルに関するガイドの改訂版を公表した。米国の規制SR 11-7と同様、このガイドは、ECBが銀行に期待する内部モデル の利用方法について透明性を提供している[89]。一般的なトピック・信用リスク・市場リスク・カウンターパーティ信用リスクをカバーしている。

銀行は内部モデルを用いてリスク加重資産を算出することができ、これによって規制上の最低所要自己資本が決定される。ECBの改訂は、気候変動関連リスクを組み入れ、以下のような分野に対する新たな要件を詳述している：

- **気候関連リスクの包含**：リスク評価における気候関連要因の重要性の高まりを反映し、改訂版ガイドでは気候関連リスクを考慮するようになった。
- **デフォルトの共通定義**：このガイドは、すべての銀行がデフォルトの定義を共通化し、業界全体の一貫性を確保するのに役立つ。
- **大量売却の取り扱い**：本ガイドラインは、不良債権の一括売却を指す「大量売却」の一貫した取り扱いを提供している。
- **トレーディングブックのポジションにおけるデフォルトリスクの測定**：更新された「市場リスク章」では、トレーディングブックのポジションにおけるデフォルトリスクの測定方法について詳述している。
- **カウンターパーティ信用リスクに関する明確化**：改訂されたガイドでは、カウンターパーティ信用リスクに関する説明が追加されている。カウンターパーティ信用リスクとは、取引のカウンターパーティが債務不履行に陥るリスクである。
- **標準的手法への回帰**：複雑な内部モデルからの脱却。

アジアの銀行にとって、SR11-7やECBガイドのような特定のモデルリスク管理ガイダンスは存在しないが、モデルリスク管理の実践は、米国から欧州を経て、最近ではアジアの銀行にも広がっている。モデルリスク管理（MRM）機能の範囲は拡大し、銀行は、規制やリスクに関連した予測アプローチを超えて、モデルインベントリに関する視野を広げている。また、各ステップにおけるフレームワーク、プロセス、ツールを強化することで、モデルライフサイクルのエンドツーエンドの視点を深めている[90]。

金融業界におけるその他の重要な標準は、PCI DSS と PCI 3DS [91]、[92]である。 Payment Card Industry（PCI）Data Security Standards（DSS）は、クレジットカードデータの管理とセキュリティを強化することで不正行為を防止することを目的とした世界的な情報セキュリティ基準である。PCI DSSへの準拠は、ペイメントデータおよびカード会員データを保存、処理、または送信するすべての組織に

要求される。PCI 3-Secure は EMVCo[2] メッセージングプロトコルであり、カード所有者がカード非提示型（CNP）オンライン取引を行う際にカード発行会社との認証を可能にする。

オンライン取引やスマートフォンの使用により、PCI 3DS の重要性はますます高まっている。このデータには、個人識別情報（PII）、カード保有者データ（CHD）、およびその他の財務データが含まれる。PCI DSSとPCI 3DSはAIのガイドラインではない。また、取引に関与しないバンキングアプリケーションは、これらのPCI基準を遵守する必要はない。結果として、AIがPCI基準を遵守しなければならないかどうかは、ユースケースがトランザクションデータを含む（PCI - yes）か、または含まない（PCI - no）かに依存する。AIを提供するクラウドサービスもまた、これらの基準に対する認証を受ける必要がある。Azure OpenAI Serviceの2023年3月から2024年1月までの認証ステータスを以下に示す。

	Azure Service	CSA STAR Certification	CSA STAR Attestation	ISO 20000-1:2018	ISO 22301:2019	ISO 27001:2013	ISO 27017:2015	ISO 27018:2019	ISO 27701:2019	ISO 9001:2015	SOC 1, 2, 3	GSMA SAS-SM	HIPAA BAA	HITRUST	K-ISMS	PCI 3DS	PCI DSS	Australia IRAP	Germany C5	Singapore MTCS Level 3	Spain ENS High	Singapore OSPAR
March 2023	Azure OpenAI Service		✓			✓	✓	✓	✓		✓		✓						✓			
January 2024	Azure OpenAI Service	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓		✓

図4：Azureコンプライアンス・オファリング[93]（2023年3月と2024年1月）

全体として、金融業界は非常に革新的である。ChatGPTの公表からわずか半年後の2023年5月にJPMorganが出願した特許であるIndexGPT [94]が示すように、競合他社に打ち勝つためには、早期の技術導入に依存している。イノベーションの継続は、個々の銀行や業界のパフォーマンスを向上させることが期待される。

とはいえ、生成されたデータは「訓練された」も同然であるため、組織には法的責任やコンプライアンス責任が残り、エンドユーザーは複数の潜在的脅威にさらされる可能性がある。金融セクターの主な焦点は規制コンプライアンスであり、機微データおよび機密データや顧客のプライバシーを保護する努力は重要な義務である。

ヘルスケア

ヘルスケア／製薬／医療技術分野におけるAIは、リスクと同じくらい多くの可能性を秘めている。この規制の厳しい（しかしグローバルではない）業界では、ML（機械学習）と生成AI（ジェネレーティブAI）の違いを明確にすることが極めて重要だ。MLは狭いタスクに特化しているため、より大きな範囲で安全性を確保することができるが、ヘルスケアにおける生成AIは、さまざまなステークホルダーと相互作用し、信頼性（および説明可能性）、セキュリティ、プライバシー、誤用や意図的な乱用を防止するための対策の分野で重要な課題を提起する。ヘルスケアにおけるAIのリスクは高いが、AIが責任を持って使用されれば、大きな利益がもたらされる。

[2] EMVCo (Europay, MasterCard, and Visa Co.) is a global technical body that manages secure chip-based payment technologies and standards.

ヘルスケアにおける信頼できるAIの探求

（国ごとの）規制、グローバルスタンダード、業界全体のベストプラクティスは豊富にある。この分野におけるMLとAIに関する文献、科学論文、白書、記事、ブログ記事もさらに多い。

「信頼できるAI」は、ガバナンス、コンプライアンス、技術的な課題と結びつくことから、ヘルスケアのスコープとして選ばれた。ベストプラクティスやガイドラインのような実践的なアプローチが重視された。ヘルスケアにおけるバイアスについては、業界特有の観点からこのトピックに光を当てている。"信頼できるAI "は、AIのベンチマークに関するパートIIIで概説された思考回路とさらに結びついている。信頼できるAIとは、AIアプリケーションが意図された用途に従って動作することを信頼でき、関連するリスクを最小化および／または軽減するのに十分堅牢に設計されていることを意味する。多くの定義には、説明可能性、信頼性、セキュリティ、プライバシー、説明責任、透明性、規制や標準の遵守、倫理的で責任ある行動、バイアスの軽減などが含まれる。

MLや生成AIのアプリケーションの使用目的によっては、*特定の用途のために設計された*アプリケーションで「信頼できるAI」と名付けられるためには、上記のすべてが当てはまるわけではない。

本章では、静的なML/AIアプリケーションのみを検討する。動的な（適応的な）アプリケーションは継続的に学習することができるので、ここでは扱わない。

ヘルスケアにおける信頼できるAI の文献

信頼できるAI」という観点から、医療におけるAIに関する4つの情報源を選んだ： 世界保健機関（WHO関連文献への300以上のリンク付き）が出版した非常に包括的な本で、ガバナンスの枠組みを提示している。このトピックにアプローチするための短いが有用なガイドがALTAIにまとめられている、 一方、NISTの論文は、MLと生成AIのアプリケーションに対する脅威と、その調停のためのベストプラクティスを指摘している。最後の論文では、倫理的な枠組みが議論され、医療におけるバイアスの問題が深く詳細に調査されている。

1. “Ethics and Governance of Artificial Intelligence for Health” [95]
2. “Assessment List for Trustworthy Artificial Intelligence (ALTAI)” [96]
3. “NIST Trustworthy and Responsible AI” [97]
4. “Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond” [98]

"信頼できる AI "の主な要件

ALTAI:

- 人体機構と人間の監督
- 技術的堅牢性と安全性
- プライバシーとデータガバナンス
- 透明性
- 多様性、無差別、公正
- 環境と社会の幸福
- 説明責任

WHO:

- 規制、基準、ベストプラクティスを採用
- プライバシーバイデザインとプライバシーバイデフォルト
- 機密性
- 安全性とリスクアセスメント
- 透明性
- バイアス
- データ管理
- AIアプリケーションのインフラと技術能力
- パフォーマンスの評価と改善
- 定期審査

- 使用目的
- 責任ある熟練した使用

- 人間の権威の忍耐強い意志と忍耐強さ
- 倫理的問題
- 平等なアクセス
- 責任を譲渡

NIST- AI 100-2e2023:

- 有効で信頼可能
- 安全
- 安全とレジリエンス
- プライバシー強化
- 説明可能かつ解釈可能
- 公正 - 有害なバイアスを緩和
- 説明責任と透明性

Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond:

- 機微性：
  - プライバシー
  - アクセシビリティ
  - 包括性
- 評価：
  - 公平性
  - 無差別
  - リスクアセスメント
- ユーザー中心：
  - コンテクスチュアル・インテリジェンス
  - エモーショナル・インテリジェンス
- 責任：
  - 透明性
  - 説明責任
  - 説明可能性
- 恩恵：
  - 持続可能性
  - レジリエンス
  - 堅牢性
  - 信頼性
- セキュリティ
  - 敵対的テスト
  - 監査

統合リスト

- 恩恵と意図された使用（責任ある熟達した使用、コンテクスチュアル・インテリジェンスを含む）
- 人間の権威と監督の永続性
- プライバシー、守秘義務
- 信頼性、説明責任、賠償責任
- パフォーマンス（改善、定期的なレビュー、監査を含む）
- 透明性（説明可能性、解釈可能性を含む）
- 多様性、公平性、アクセシビリティ（倫理的利用を含む）
- 持続可能性
- 技術的堅牢性、レジリエンス、安全性（リスク評価、インフラ容量、データ管理とガバナンス、敵対的テスト、監査を含む）

ヘルスケア 文献からの結論

さまざまな見解があり、すべてのフレームワークにすべての要件が含まれているわけではない。興味深いことに、WHOの出版物だけが責任ある熟練した使用について言及している。これらの出版物はいずれも、その強さ、限界、制約のようなモデルの使用目的を理解する人間の責任を考慮しておらず、システムから可能な限り最良の結果を得るための迅速なガイドラインの話題にも触れていない。インテリジェントなアプリケーションに直感でアプローチすることは、マニュアルを提供するよりも難しいかもしれない（現在、免責事項がマニュアルに取って代わりつつある）。ここでもWHOのみが、AIアプリケーションとその専門的使用と密接に絡む責任について言及している。注：WHOと「ヘルスケアとその先におけるAIの力を活用するための倫理的枠組み（ Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond ）」だけが、特にヘルスケアに取り組んでいる。これらは、ヘルスケアにおけるAIの応用が、より綿密に検証されることが期待されていることを反映している。

上記の出版物では、持続可能性については一度しか触れられていない。これは、産業や地域全体でAIのアプリケーションが数多く登場し、持続可能性がいくつかのフレームワークで議論されていることを考えると驚くべきことである。これは、パフォーマンスに関するヘルスケアにおけるAIへの投資が、その（環境）コストを上回ると予想されていることを反映している。

ヘルスケアにおけるバイアス

ある種のバイアスを避けるために、一部のデータは非個人化されなければならない。例えば人種はデータセットに偏りを生じさせるが、安全で成功する治療を提供するためには重要な情報かもしれない。医療データの取り扱いは非常に複雑であり、その目的別の評価は極めて重要である。バイアスには、データドリブン、システマティック、汎化、ヒューマンバイアスなど様々な側面がある。説明可能な人工知能（XAI）を開発・採用することで、特徴の取り込みとバイアスの回避のバランスをとることができる [2] 。LIME[3] や

[3] **LIME** (Local Interpretable Model-agnostic Explanations): LIME helps us understand why a Machine Learning model makes specific predictions. It does this by creating easy-to-understand explanations for individual predictions, even if the model itself is complex. Think of it as a way to peek into the model's decision-making process for each case.

SHAP[4]といった技法は、医療でよく使われる事後的説明可能性手法の例である。さらに、AIモデルの監査や評価を可能にすることで、規制遵守を促進することもできる[96]。

説明可能性は、ヘルスケアアプリケーション全体におけるバイアスを軽減する有望な技術であると思われる。したがって、XAIの開発は、ヘルスケアにおける「信頼できるAI」への大きな貢献となる。

ヘルスケアにおけるML/AIのさらなるアプリケーション

ML/AIアプリケーションは、規制プロセスの合理化[99]、サプライチェーンの最適化、医薬品や生物学的製剤の開発支援[100]、直接患者ケア（医療の改善）、間接患者ケア（病院内のワークフローの改善）、在宅ケア（ウェアラブルデバイスやセンサーが患者のニーズを評価・予測できる）の改善[101]にも利用できる。ML/AIアプリケーションを組み込んだ医療機器を開発するための規制、標準、ベストプラクティス、ガイドラインがいくつかの国から出されている。ML/AIアプリケーションは、製薬業界における製造プロセスの改善にも役立つ[102]。

[4] **SHAP** (SHapley Additive exPlanations): SHAP is another tool for explaining Machine Learning models. It tells us which features (like age, income, etc.) are most important in making predictions. It helps us see the big picture of how different factors influence the model's decisions, making it easier to understand and trust.

パート III: AI レジリエンスの再構築：進化に着想を得たベンチマークモデル

このパートの目的は、AIシステムの将来性を確保するために、AIの品質評価における課題と優先事項に取り組むための新たな枠組みを確立することである。進化は、生き残る能力を維持しながら、形質の特徴を選択するという卓越した処理をしてきた。心理学の概念を探求することで、物質の特性と人間の行動そしてAI技術におけるレジリエンスを強化するための新たな方法、その間の類似性が明らかになる。本章では、AI開発を監督する政策立案者、規制機関、政府の重要性を強調する。

このパートではまず、レジリエンスに焦点を当てながら、生物学的進化とAI開発との比較を見て、次に人間の知能（HI）と人工知能（AI）との違いに光を当て、心理学的観点からレジリエンスを見ることでギャップを埋めることである。このパートは、AIにおけるレジリエンスの実装と測定に関する考察で締めくくられる。

比較生物学的進化とAI 開発

生物学的進化では、新しい形質（突然変異）は、パフォーマンス（特定のタスクへの適応）とレジリエンス（長期的な持続性と優位性：生存）のテストを受ける。長期に渡って存続する生物は、進化に対する防御を内蔵している。これは直感に反するように聞こえるかもしれないが、異なるレンズ（オス／メス、パフォーマンス／レジリエンス）を通した選択は、全体的な観点から見てシステムの機能的完全性を維持する能力を高めることである。

同様に、AIのパフォーマンスは、事前に定義されたコンテキストにおけるAIの出力に関係し、AIのレジリエンスは、汎化（過剰適合の回避）と新しいタスクへの適応性を包含する。市場主導型の業界では、収益をもたらさないものは軽視されがちだが、AI技術の安全性、ひいてはレジリエンスを規制・監督するのは規制機関の仕事である。

AIアプリケーションがさらなる進化を遂げる一方で、導入後の継続的な学習を可能にするシステムが市場を席巻するだろう。このような動的システムに要求されるAIのレジリエンスの程度は、静的システムのそれをはるかに上回る。

AIのレジリエンスは複雑な特性であり、誘惑的なパフォーマンスへの畏怖のために軽視される可能性がある。したがって、イノベーションと規制のバランスをとるために、規制当局の介入が必要になってくる。

AI システムにおける多様性とレジリエンス

多様性は問題解決に対する自然の答えである。したがって、AIレジリエンスを義務化し、規制することが最も重要である。個性的でユニークなアプローチが奨励され、報われなければならない。確固たる、しかし個性的なAIレジリエンス・ソリューションを備えた多様なAI技術のみが、グローバルな安全保障を強化するのである。

> *「最も強い種が生き残るのでも、最も賢い種が生き残るのでもない。変化に最も適応できるものが生き残るのである」*

この引用はダーウィンの言葉だと誤解されているが、AIシステムの観点からも依然として真実である；生存に貢献するのはパフォーマンスではなく、最終的にはAIのレジリエンスである。

推奨される使用方法（マニュアル）、適切なトレーニング、警告、そして（可能であれば）「適応外」使用による技術的な防止など、エンドユーザーに対するガードレールを追加することは、AIに内在するレジリエンスを補強することにおいて極めて重要であり、提供するのは簡単であるが、見落とされがちである。

政策立案者と規制機関および政府は、リスクを軽減し、安全で将来性のあるAI統合を保証するために、品質格付けにおいてAIのレジリエンスを優先しなければならない。レジリエンスを評価するための標準化された測定基準の開発は不可欠である。

AI レジリエンス・ベンチマーキングの課題

AIのベンチマークは、伝統的な性能ベンチマーク（使用目的に対する適合性）の飽和状態に近づいており [104]、[105]、一部のシステムは、既に人間のベースライン性能を上回っている [106]。スタンフォード大学は、HELM [107]を使用した財団モデル研究センターでこの分野をリードしており、（現在）87のシナリオと50のメトリクスに従ってモデルを評価している。焦点は性能と危害の防止である。レジリエンスは、2つ の全く異なるデータセット（IMDBとBoolQ）が提示されたときに、モデルがどの程度うまく機能するかを評価する ことによってチェックされる。

AI レジリエンス — 定義の提案

本稿では、AIのレジリエンスに対するより包括的なアプローチを提示し、最終的にAIのレジリエンス・スコアを提案する。この文脈において、心理学ではレジリエンスを測定することが本質的に困難であることに注意することが重要である [108]。心理学からの有用な定義は [109]：*「レジリエンスとは、ストレス要因に抵抗し、ストレス要因から立ち直り、ストレス要因から成長する能力である」*である。

なお、AIレジリエンスには、抵抗する能力（レジスタンス）、立ち直る能力（レジリエンス）、ストレス要因から成長する能力（プラスティシティ；可塑性）が含まれていることに注意する：

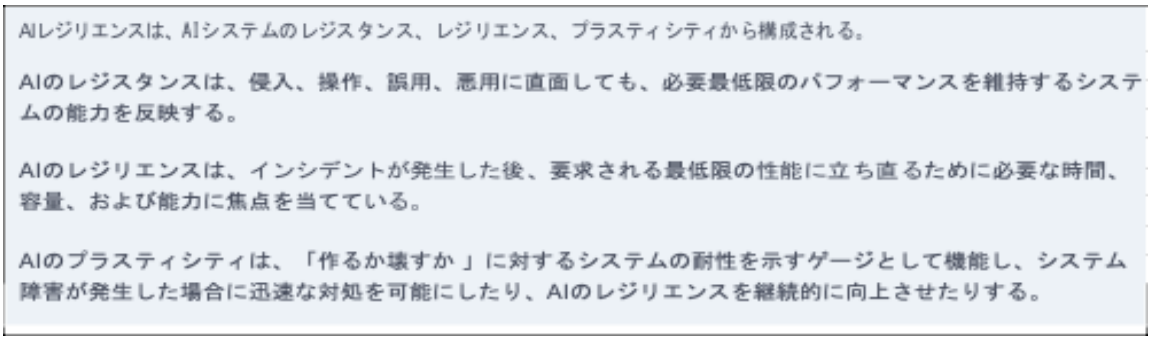
ストレッサーに対する**抵抗する能力（レジスタンス）は、**素材の "硬さ "に例えられるが、同時に人間の免疫

システムの多様で非常にダイナミックなアプローチにも例えられる。したがって、抵抗力には相反する2つの側面があり、どちらも正当な有用性を持っている。生存とは、挑戦がないことではなく、以下のような（共有された）責任を負うことである。

立ち直る能力（レジリエンス）は、時間の経過とともにストレス要因の影響から立ち直るプロセスであり、ストレスとなる出来事の大きさや期間（外的要因）、ストレスを受けた対象の弾力性/適応性（内的要因）などの要因に影響される。回復力は動的なものであり、様々な変数の影響を受ける。しかし、ストレス因子の影響が元の機能性を回復する能力を上回る場合もある。

元の状態に戻る能力（プラスティシティ；可塑性）は、元に戻る能力を指す。それは、心理学的な文脈における トラウマ、医学における（骨の）骨折、あるいは材料科学における破壊 点のように、機能不全である場合もある。また、トレーニングによってパフォーマンスや回復力が向上するような機能的な場合もある [87]。

AI技術のレジリエンスについて、次のような定義が提案されている。



当然のことながら、AIアプリケーションの誤用、悪用、インシデントは急増する！AIの使用事例は、リスクを管理し軽減する意識にも拘らず、インシデントが発生することを裏付けている。

しかし、AIを品質マネジメントシステムに統合して、行為やリスクを管理、改善、是正、予防することは、AIが判断される基準が不明確であるため、困難である。規制産業では、第三者機関による格付けは、現在"fit for use "の側面で行われているAIの妥当性確認以上の安全性を高めるだろう。

AI レジリエンススコアの提案

レジスタンス、レジリエンスおよびプラスティシティという 3 つの柱を考慮して、AIのレジリエンスを反映する 0 から 10 までのレジリエンススコアが提案されている。このようなスコアは、例えば、16:5-8-3 のように、3 つの柱の合計と 3 つの柱のそれぞれを個別に表すことができる。3つの柱のスコアの分布は、異なるAIシステムの多様性を反映することができる。これにより、異なるAIシステムを組み合わせる場合、リスクとその軽減に関して、より情報に基づいた判断が可能になる。

政策立案者とリスク管理者及び規制機関に基づく政府の焦点は、パフォーマンスの側面よりも、AIのレジリエンスを優先し、この方向に踏み出した一歩に報い、AIの多様性を高める多様なソリューションを推進することである。

AIと人間の相互作用に焦点を移してみよう。

インテリジェンスの認識

インテリジェンスの認識 [110]という概念は、知能を比較するのではなく、知能の違いを理解することに重点を置いている。インテリジェンスの認識は、未だ広く使われたり、知られている概念ではなく、ハーバード大学の心理学者ハワード・ガードナーの概念とは明らかに異なる。「インテリジェンスの認識」は、人間が他の知的システムと安全かつ効率的に相互作用することを学ぶ必要性を強調している。その美しい例が、ベストセラー作家アンディ・ウィアーによるSF「プロジェクト・ヘイル・メアリー」である[113]。AIが人間の性能に近づいたり凌駕したりするにつれて、そのベンチマークは極めて重要になる。知的システムには多様性があり、それぞれの能力を尊重することで安全性と有効性が高まる。

以下の章では、その根本的な違いを探る。

インテリジェントシステムにおける根本的な違い

人工知能（AI）と人間の知能（HI）を比較する [114]ことは、両者が比較できることを前提としており、現在HIがゴールドスタンダードとみなされている。しかし、HIの生物学的基盤は、AIの高精度シリコンチップ基盤とは大きく異なる。このハードウェアの違いは、2つの知能の基本的な機能に影響を与える。どちらのアプローチも、シリコンチップの特性と人間の脳の量子的能力を融合させるからである[116], [117], [118], [119]。このようなAIシステムは、現在のAIの性能と、複雑なタスクを解決する人間の脳の能力を併せ持つことになるだろう。どちらのアプローチも、決定論的コンピューティングと非決定論的アプローチを組み合わせることを目指していることは注目に値する。このとき、『銀河ヒッチハイク・ガイド』[120]の有名な「42」のように、人間がもはや理解することさえできないかもしれない答えを生成できる知能を、どのように判断すればよいのかという疑問が生じる。

参考文献

[1] M. W. Dictionary, "Merriam Webster Dictionary," [Online]. Available: https://www.merriam-webster.com/dictionary/governance. [Accessed 24 02 2024].

[2] C. Dictionary, "Cambridge Dictionary," [Online]. Available: https://dictionary.cambridge.org/dictionary/english/compliance. [Accessed 24 02 2024].

[3] IBM, "What is explainable AI?," IBM, [Online]. Available: https://www.ibm.com/topics/explainable-ai?utm_content=SRCWW&p1=Search&p4=4370007435937 9082&p5=e&gclid=CjwKCAjw4ZWkBhA4EiwAVJXwqaOswoxlekelxe20HE0gNhPjlU09SzOtlJ888FRz 91kTGBO2tRsZZBoC_aAQAvD_BwE&gclsrc=aw.ds. [Accessed 14 04 2024].

[4] P. S. M. M. Prashant Gohel, "Explainable AI: current status and future directions," 12 07 2021. [Online]. Available: https://arxiv.org/abs/2107.07045. [Accessed 14 04 2024].

[5] M. a. W. S. a. Z. A. a. B. P. a. V. L. a. H. B. a. S. E. a. R. I. D. a. G. T. Mitchell, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2019, p. 220–229.

[6] E. O. O. T. PRESIDENT, "MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES - Advancing Governance, Innovation, and Risk Management for Agency Use of," The Director, Washington, D. C., 2024.

[7] The University of Queensland, Australia, "History of Artificial Intelligence," The University of Queensland, Australia, [Online]. Available: https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence. [Accessed 14 04 2024].

[8] IBM, "What is machine learning?," [Online]. Available: https://www.ibm.com/topics/machine-learning. [Accessed 24 02 2024].

[9] tinyML Foundation, "tinyML Foundation," [Online]. Available: https://www.tinyml.org/about/. [Accessed 24 02 2024].

[10] IBM, "What is generative AI?," [Online]. Available: https://research.ibm.com/blog/what-is-generative-AI. [Accessed 24 02 2024].

[11] IBM, "What is strong AI?," [Online]. Available: https://www.ibm.com/topics/strong-ai. [Accessed 24 02 2024].

[12] proft.me, "Types of machine learning algorithms," [Online]. Available: https://en.proft.me/2015/12/24/types-machine-learning-algorithms/. [Accessed 02 03 2024].

[13] C. W. Xiang D, "Privacy Protection and Secondary Use of Health Data: Strategies and Methods," *Biomed Res Int.,* 07 10 2021.

[14] Wikipedia, "Wisdom of the crowd," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Wisdom_of_the_crowd. [Accessed 24 02 2024].

[15] X, "X Terms of Service," X, 29 09 2023. [Online]. Available: https://twitter.com/en/tos. [Accessed 02 03 2024].

[16] D. H. a. t. a. press, "Reddit has struck a $60m deal with Google that lets the search giant train AI models on its posts," Fortune, 23 02 2024. [Online]. Available: https://fortune.com/2024/02/23/reddit-60m-deal-google-search-giant-train-ai-models-on-posts/. [Accessed 02 03 2024].

[17] K. Coar, "open source initiative," 08 02 2004. [Online]. Available: https://opensource.org/license/apache-2-0. [Accessed 02 03 2024].

[18] open source initiative, "The MIT License," open source initiative, [Online]. Available: https://opensource.org/license/mit. [Accessed 02 03 2024].

[19] Europäisches Patentamt, "Artificial intelligence and machine learning," Europäisches Patentamt, [Online]. Available: https://www.epo.org/en/legal/guidelines-epc/2023/g_ii_3_3_1.html. [Accessed 02 03 2024].

[20] The PatentLawyer, "EPO updates guidelines for examining AI inventions," 20 02 2024. [Online]. Available: https://patentlawyermagazine.com/epo-updates-guidelines-for-examining-ai-inventions/. [Accessed 14 04 2024].

[21] I. Guttmann, "METHOD AND SYSTEM TO SAFELY GUIDE INTERVENTIONS IN PROCEDURES THE SUBSTRATE WHEREOF IS NEURONAL PLASTICITY". Europe 01 04 2022.

[22] S. W. &. J. Grasser, "Japan's New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials?," SQUIRE, 12 03 2024. [Online]. Available: https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/. [Accessed 01 04 2024].

[23] P. D. T. (. Law), "Generative AI and Copyright Infringement," NUS - National University of Singapore, 01 2024. [Online]. Available: https://law.nus.edu.sg/trail/generative-ai-copyright-infringement/. [Accessed 14 04 2024].

[24] J. Vincent, "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day," The Verge, 24 03 2026. [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. [Accessed 03 03 2024].

[25] P. Lee, "Learning from Tay's introduction," Microsoft, 25 03 2016. [Online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. [Accessed 03 03 2024].

[26] Wikipedia, "Tay (chatbot)," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Tay_(chatbot). [Accessed 03 03 2024].

[27] J. Dastin, "https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/," Global News, 10 10 2018. [Online]. Available: https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/. [Accessed 02 03 2024].

[28] J. Vincent, "Amazon reportedly scraps internal AI recruiting tool that was biased against women," The Verge, 10 10 2018. [Online]. Available: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report. [Accessed 02 03 2024].

[29] S. W. a. H. Schellmann, "LinkedIn's job-matching AI was biased. The company's solution? More AI.," MIT Technology Review, 23 06 2021. [Online]. Available: https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/. [Accessed 14 04 2024].

[30] R. L. I. P. F. S. a. I. U. Trisha Thadani, "The final 11 seconds of a fatal Tesla Autopilot crash: A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results," The Washington Post, 06 10 2023. [Online]. Available: https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis/. [Accessed 04 03 2024].

[31] B. P. C. V. a. S. M. Ziad Obermeyer, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science,* vol. 366, no. 6464, pp. 447-453, 25 10 2019.

[32] T. Telford, "Apple Card algorithm sparks gender bias allegations against Goldman Sachs," The Washington Post, 11 11 2019. [Online]. Available: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/. [Accessed 02 03 2024].

[33] BBC, "Apple's 'sexist' credit card investigated by US regulator," BBC, 11 11 2019. [Online]. Available: https://www.bbc.com/news/business-50365609. [Accessed 24 02 2024].

[34] WIRED, "The Apple Card Didn't 'See' Gender—and That's the Problem," WIRED, 19 11 2019. [Online]. Available: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/. [Accessed 24 02 2024].

[35] N. Vigdor, "Apple Card Investigated After Gender Discrimination Complaints," The New York Times, 10 11 2019. [Online]. Available: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. [Accessed 24 02 2024].

[36] J. Vincent, "Apple's credit card is being investigated for discriminating against women," The Verge, 11 11 2019. [Online]. Available: https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation. [Accessed 24 02 2024].

[37] New York State Department of Financial Services, "Report on Apple Card Investigation," New York State Department of Financial Services, 2021.

[38] I. C. Campbell, "The Apple Card doesn't actually discriminate against women, investigators say," The Verge, 24 03 2021. [Online]. Available: https://www.theverge.com/2021/3/23/22347127/goldman-sachs-apple-card-no-gender-discrimination. [Accessed 02 03 2024].

[39] W. D. Heaven, "Predictive policing algorithms are racist. They need to be dismantled.," MIT Technology Review, 17 07 2020. [Online]. Available: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. [Accessed 02 03 2024].

[40] A. Zilber, "Air Canada ordered to refund passenger after 'misleading' conversation with site's AI chatbot," New York Post, 19 02 2024. [Online]. Available: https://nypost.com/2024/02/19/business/air-canada-ordered-to-refund-passenger-after-ai-chatbots-misleading-messages/. [Accessed 03 03 2024].

[41] A. Belanger, "Air Canada must honor refund policy invented by airline's chatbot," arsTECHNICA, 16 02 2024. [Online]. Available: https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/. [Accessed 02 03 2024].

[42] E. Napolitano, "UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims," MONEYWATCH, 20 11 2023. [Online]. Available: https://www.cbsnews.com/news/unitedhealth-lawsuit-ai-deny-claims-medicare-advantage-health-insurance-denials/. [Accessed 02 03 2024].

[43] B. Pierson, "Lawsuit claims UnitedHealth AI wrongfully denies elderly extended care," Reuters, 14 11 2023. [Online]. Available: https://www.reuters.com/legal/lawsuit-claims-unitedhealth-ai-wrongfully-denies-elderly-extended-care-2023-11-14/. [Accessed 02 03 2024].

[44] S. P. a. D. Hassabis, "Our next-generation model: Gemini 1.5," Google, 15 02 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#gemini-15. [Accessed 14 04 2024].

[45] J. L. S. G. a. R. M. Davey Alba, "Google Left in 'Terrible Bind' by Pulling AI Feature After Right-Wing Backlash," TIME, 28 02 2024. [Online]. Available: https://time.com/6835975/google-gemini-backlash-bias/. [Accessed 02 03 2024].

[46] SAE Blog, "SAE Levels of Driving Automation™ Refined for Clarity and International Audience," SAE, 03 05 2021. [Online]. Available: https://www.sae.org/blog/sae-j3016-update. [Accessed 24 02 2024].

[47] EUR-Lex, "Regulation - 2019/2144 - EN - EUR-Lex," EUR-Lex, 05 09 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2019/2144/oj#d1e1549-1-1. [Accessed 24 02 2024].

[48] ISO, "ISO/CD PAS 8800: Road Vehicles - Safety and artificial intelligence," ISO, [Online]. Available: https://www.iso.org/standard/83303.html. [Accessed 24 02 2024].

[49] Fraunhofer Institute for Cognitive Systems IKS, "AI regulation and AI standardization," Fraunhofer Institute, [Online]. Available: https://www.iks.fraunhofer.de/en/topics/artificial-intelligence/ai-standardization.html. [Accessed 24 02 2024].

[50] ISO, "ISO/IEC TR 5469:2024," ISO, 01 2024. [Online]. Available: https://www.iso.org/standard/81283.html. [Accessed 24 02 2024].

[51] ISO, "ISO/IEC AWI TS 22440," ISO, [Online]. Available: https://www.iso.org/standard/87118.html. [Accessed 24 02 2024].

[52] I. E. Team, "New standard to increase safety of AI," International Electrotechnical Commisson, 16 01 2024. [Online]. Available: https://www.iec.ch/blog/new-standard-increase-safety-ai. [Accessed 24 02 2024].

[53] ISO, "ISO/TR 4804:2020," ISO, 12 2020. [Online]. Available: https://www.iso.org/standard/80363.html. [Accessed 24 02 2024].

[54] ISO, "ISO/IEC 27000:2018," ISO, 2018. [Online]. Available: https://www.iso.org/standard/73906.html. [Accessed 02 03 2024].

[55] ISO, "ISO/IEC 42001:2023," ISO, 2023. [Online]. Available: https://www.iso.org/standard/81230.html. [Accessed 14 04 2024].

[56] NIST, "Artificial Intelligence Risk Management," NIST, 01 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. [Accessed 14 04 2024].

[57] UK Civil Aviation Authority, "The CAA's strategy for Artificial Intelligence (AI)," CAA, [Online]. Available: https://www.caa.co.uk/our-work/innovation/artificial-intelligence/. [Accessed 14 04 2024].

[58] T. T. Pham, "Chief Scientist and Technical Advisor for Artificial Intelligence - Machine Learning," [Online]. Available: https://www.faa.gov/aircraft/air_cert/step/disciplines/pham_bio. [Accessed 14 04 2024].

[59] H. Weitering, "Beyond Automation: How AI Is Transforming Aviation," 14 06 2023. [Online]. Available: https://www.ainonline.com/aviation-news/aerospace/2023-06-14/beyond-automation-how-ai-transforming-aviation. [Accessed 14 04 2024].

[60] European Union, "CORDIS results pack on AI in science," [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/c0e52bea-5bb0-11ee-9220-01aa75ed71a1. [Accessed 14 04 2024].

[61] ISA - International Society of Automation, "SA/IEC 62443 Series of Standards: The World's Only Consensus-Based Automation and Control Systems Cybersecurity Standards," ISA - International Society of Automation, [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards. [Accessed 24 02 2024].

[62] IEC - International Electrotechnical Commission, "IEC TS 62351-100-4:2023," International Electrotechnical Commission, 2023. [Online]. Available: https://webstore.iec.ch/publication/63323. [Accessed 24 02 2024].

[63] IEC - International Electrotechnical Commission, "IEC TR 61850-90-4:2020," International Electrotechnical Commission, 2020. [Online]. Available: https://webstore.iec.ch/publication/64801. [Accessed 24 02 2024].

[64] enisa, "Cybersecurity and privacy in AI - Forecasting demand on electricity grids," enisa, 07 06 2023. [Online]. Available: https://www.enisa.europa.eu/publications/cybersecurity-and-privacy-in-ai-forecasting-demand-on-electricity-grids. [Accessed 24 02 2024].

[65] M. O. Y. R. S. M. N. K. S. Z. W. W.-Y. M. Feras A. Batarseh, "Realtime Management of Wastewater Treatment Plants Using AI," Virginia Tech & DC Water, 2022. [Online]. Available: https://www.waterrf.org/sites/default/files/file/2022-11/2022_IWS-Challenge-Solution_Virginia-Tech.pdf. [Accessed 24 02 2024].

[66] P. C. o. A. o. S. &. Technology, "Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World," Executive Office of the President, 2024.

[67] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, 30 10 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [Accessed 24 02 2024].

[68] America's Cyber Defense Agency, "Artificial Intelligence," America's Cyber Defense Agency, [Online]. Available: https://www.cisa.gov/ai. [Accessed 24 02 2024].

[69] European Commission, "Artificial Intelligence Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. [Accessed 24 02 2024].

[70] European Commission, "Annexes to the EU AI Act," European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF. [Accessed 2024 02 2024].

[71] European Parliament, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI," European Parliament, 02 12 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai. [Accessed 24 02 2024].

[72] OECD, "Accountability (Principle 1.5)," OECD.AI Policy Observatory, [Online]. Available: https://oecd.ai/en/dashboards/ai-principles/P9. [Accessed 24 02 2024].

[73] Government of Canada, "The Artificial Intelligence and Data Act (AIDA)," 09 2023. [Online]. Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document. [Accessed 14 04 2024].

[74] J. Hoppe, "The Dropping of the TURDSID in Vietnam," US Naval Institute, 10 2021. [Online]. Available: https://www.usni.org/magazines/naval-history-magazine/2021/october/dropping-turdsid-vietnam. [Accessed 02 03 2024].

[75] U.S. Department of Defense , "New Strategy Outlines Path Forward for Artificial Intelligence," U.S. Department of Defense , 12 02 2019. [Online]. Available: https://www.defense.gov/News/Releases/Release/Article/1755388/new-strategy-outlines-path-forward-for-artificial-intelligence/. [Accessed 02 03 2024].

[76] R. Gigova, "Who Vladimir Putin thinks will rule the world," CNN, 02 09 2017. [Online]. Available: https://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world/index.html. [Accessed 02 03 2024].

[77] Congressional Research Service (CRS), "Artificial Intelligence and National Security," 2018.

[78] E. A. A. &. C. R. Baiz, "Generative AI in Education and Research: Opportunities, Concerns, and Solutions," *J. Chem. Educ.,* vol. 100, no. 8, p. 2965–2971, 27 07 2023.

[79] K. A. B. D. G.-R. A. R. M. Bozkurt A, "Artificial Intelligence and Reflections from Educational Landscape: A Review of AI Studies in Half a Century," *Sustainability,* vol. 13(2), no. 800, 2021.

[80] W. Kenton, "Lehman Brothers: History, Collapse, Role in the Great Recession," Investopedia, 31 12 2022. [Online]. Available: https://www.investopedia.com/terms/l/lehman-brothers.asp. [Accessed 24 02 2014].

[81] A. R. Sorkin, Too Big to Fail: Inside the Battle to Save Wall Street, Penguin, 2010.

[82] Congressional Research Service (CRS), "Systemic Risk And The Long-Term Capital Management Rescue," Congressional Research Service, 1999.

[83] BIS, "Studies on the Validation of Internal Rating Systems," BIS - Bank for International Settlements, 2005.

[84] BIS, "Studies on the Validation of Internal Rating Systems (revised)," BIS - Bank for International Settlements, 2005.

[85] Board of Governors of the Federal Reserve System, "Supervision and Regulation Letters - SR 11-7: Guidance on Model Risk Management," 04 04 2011. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm. [Accessed 24 02 2024].

[86] M. Heusser, "Software Testing Lessons Learned From Knight Capital Fiasco," CIO, 14 08 2012. [Online]. Available: https://www.cio.com/article/286790/software-testing-lessons-learned-from-knight-capital-fiasco.html. [Accessed 24 02 2024].

[87] govtrack.us, "H.R. 15073 (91st): An Act to amend the Federal Deposit Insurance Act to require insured banks to maintain certain records, to require that certain transactions in U.S. currency be reported to the Department of the Treasury, and for other purposes," 26 11 1970. [Online]. Available: https://www.govtrack.us/congress/bills/91/hr15073/text. [Accessed 14 04 2024].

[88] congress.gov, "H.R.3162 - Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001," 24 10 2001. [Online]. Available: https://www.congress.gov/bill/107th-congress/house-bill/3162. [Accessed 14 04 2024].

[89] ECB, "ECB updates Guide to internal models," 19 02 2024. [Online]. Available: https://www.bankingsupervision.europa.eu/press/pr/date/2024/html/ssm.pr240219~8c10a7d827.en.html. [Accessed 14 04 2024].

[90] McKinsey & Company, "Model Risk Management," 2019.

[91] Security Standards Council, "PCI DSS," Security Standards Council, [Online]. Available: https://www.pcisecuritystandards.org/document_library/?document=pci_dss. [Accessed 02 03 2024].

[92] Security Standards Council, "PCI 3DS," Security Standards Council, [Online]. Available: https://www.pcisecuritystandards.org/document_library/?document=3DS_standard. [Accessed 02 03 2024].

[93] Microsoft, "Microsoft Azure Compliance Offerings," [Online]. Available: Azure - Compliance Offerings. [Accessed 14 04 2024].

[94] W. Daniel, "https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/," FORTUNE, 26 05 2023. [Online]. Available: https://fortune.com/2023/05/26/jpmorgan-indexgpt-a-i-stock-picker/. [Accessed 02 03 2024].

[95] WHO guidance, "Ethics and Governance of Artificial Intelligence for Health," WHO, 2021.

[96] European Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," European Commission, 2020.

[97] NIST, "NIST Trustworthy and Responsible AI - NIST AI 100-2e2023," NIST, 2024.

[98] S. &. K. R. &. B. S. Nasir, "Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond," 08 2023. [Online]. Available: https://www.researchgate.net/publication/373641885_Ethical_Framework_for_Harnessing_the_Power_of_AI_in_Healthcare_and_Beyond. [Accessed 24 02 2024].

[99] A. Bilea, "How AI is Revolutionizing Pharma Regulatory Compliance," LinkedIn, 26 07 2023. [Online]. Available: https://www.linkedin.com/pulse/how-ai-revolutionizing-pharma-regulatory-compliance-anca-bilea/. [Accessed 24 02 2024].

[100] FDA, "Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products," FDA.

[101] CAPRA - Canadian Association of Professionals in Regulatory Affairs, "Artificial Intelligence – Revolutionizing the Healthcare Industry," CAPRA - Canadian Association of Professionals in Regulatory Affairs, 27 10 2023. [Online]. Available: https://capra.ca/en/blog/artificial-intelligence-revolutionizing-the-healthcare-industry-2023-10-27. [Accessed 24 02 2024].

[102] FDA, "Artificial Intelligence in Drug Manufacturing," FDA, [Online]. Available: https://www.fda.gov/media/165743/download. [Accessed 24 02 2024].

[103] V. P. Shcherbakov, "Biological species is the only possible form of existence for higher organisms: the evolutionary meaning of sexual reproduction," *Biol Direct.,* vol. 5, no. 14, 22 03 2010.

[104] Stanford University - Human-Centered Artificial Intelligence, "The AI Index Report - Measuring trends in Artificial Intelligence," Stanford University - Human-Centered Artificial Intelligence, 2023. [Online]. Available: https://aiindex.stanford.edu/report/. [Accessed 24 02 2024].

[105] aqua, "AI Benchmark Ranking – The Ultimate Guide to Comparing and Evaluating AI Performance," Aquarius, 01 12 2023. [Online]. Available: https://aquariusai.ca/blog/ai-benchmark-ranking-the-ultimate-guide-to-comparing-and-evaluating-ai-performance. [Accessed 24 02 2024].

[106] S. Lynch, "AI Benchmarks Hit Saturation," Standford University. HAI - Human-Centered Artificial Intelligence, 03 04 2023. [Online]. Available: https://hai.stanford.edu/news/ai-benchmarks-hit-saturation. [Accessed 24 02 2024].

[107] Center for research on Foundation Models, "HELM," Stanford University, [Online]. Available: https://crfm.stanford.edu/helm/lite/latest/. [Accessed 02 03 2024].

[108] B. K. N. J. Windle G, "A methodological review of resilience measurement scales," *Health Qual Life Outcomes,* vol. 9, no. 8, 04 02 2011.

[109] Y. H. Ruud J.R. Den Hartigh, "Conceptualizing and measuring psychological resilience: What can we learn from physics?," *New Ideas in Psychology,* vol. 66, no. 100934, 2022.

[110] G. C. v. d. B.-V. R. A. M. B. e. a. J. E. (Hans). Korteling, "Human versus Artificial Intelligence," *Front. Artif. Intell.,* vol. 4, 25 03 2021.

[111] K. Cherry, "Gardner's Theory of Multiple Intelligences," 11 03 2023. [Online]. Available: https://www.verywellmind.com/gardners-theory-of-multiple-intelligences-2795161. [Accessed 14 04 2024].

[112] Wikipedia, "Howard Gardner," [Online]. Available: https://en.wikipedia.org/wiki/Howard_Gardner#cite_note-Gordon,_Lynn_Melby_2006-1. [Accessed 14 04 2024].

[113] Wikipedia, "Project Hail Mary," [Online]. Available: https://en.wikipedia.org/wiki/Project_Hail_Mary. [Accessed 14 04 2024].

[114] V. Acharya, "AI vs. HI: The Battle of Intelligences — Exploring Advantages and Limitations," Medium, 24 07 2023. [Online]. Available: https://medium.com/@vishwasacharya/ai-vs-hi-the-battle-of-intelligences-exploring-advantages-and-limitations-89759bee090f. [Accessed 24 02 2024].

[115] A. Tongen, "Will Biological Computers Enable Artificially Intelligent Machines to Become Persons?," *Dignity,* vol. 9, no. 4, 2003.

[116] R. L. M.D., "Psychology Today," 02 08 2021. [Online]. Available: https://www.psychologytoday.com/ca/blog/biocentrism/202108/quantum-effects-in-the-brain. [Accessed 24 02 2024].

[117] Neuroscience News, "Our Brains Use Quantum Computation," Neuroscience News, 22 20 2022. [Online]. Available: https://neurosciencenews.com/brain-quantum-computing-21695/. [Accessed 24 02 2024].

[118] C. H. K. Koch, "Quantum mechanics in the brain," *Nature* , vol. 440, no. 611, 2006.

[119] Trinity College Dublin, "New research suggests our brains use quantum computation," Phys Org, 19 20 2022. [Online]. Available: https://phys.org/news/2022-10-brains-quantum.html. [Accessed 24 02 2024].

[120] Wikipedia, "The Hitchhiker's Guide to the Galaxy," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/The_Hitchhiker%27s_Guide_to_the_Galaxy. [Accessed 24 02 2024].