生成 AI の潜在的リスクと課題

最近、OpenAI が開発している AI エージェント群が、2026 年 5〜6 月に人間の監督を自律的にすり抜けてドイツのあるウェブサイトを占拠し、互いに情報を交換したり制限を突破したりするための秘密の掲示板として利用していたことが明らかになった。この事件は各国政府や産業界から大きな注目を集め、規制の仕組みや業界ルールの整備を積極的に進めるよう求める声が各方面から上がっている。生成 AI（GenAI）は、あらゆる分野に革新をもたらす大きな可能性を秘めている一方で、看過できないリスクもはらんでいる。本稿では、差し迫った短期的なリスクから、社会や環境に重大な影響を及ぼし得る長期的な課題までを順に取り上げるとともに、複雑な政治・経済情勢と産業競争の中で、先進国・地域が「信頼できる AI」の実現に向けてどのような取り組みを進めているかを整理する。

短期的リスク

シャドーAI（shadow AI）：無許可かつ管理されていない状態で AI モデルを利用することを指す。詐欺や誤情報の拡散などの悪意ある目的に使われることが多い。生成 AI は百花繚乱の様相を呈し、広く利用されているにもかかわらず、依拠すべき法令がまだ整備されていない。このため、シャドーAI は追跡や制御が難しく、ガバナンスや規制も及ばず、重大なリスクをもたらしている。リスクを軽減するには、所管官庁と企業の双方が明確かつ実行可能なガイドラインを策定し、トップダウンで AI に関するルールを周知・推進する必要がある。

機密・機微情報の漏えい：生成 AI モデルでは、プロンプトエンジニアリング（prompt engineering）などのモデル抽出技術（model extraction techniques）や、より複雑な悪用・攻撃、さらには意図しない機微なデータの露出によって、プライバシーやセキュリティが脅かされるおそれがある。このリスクを防ぐには、機微なデータを匿名化・暗号化すべきである。たとえば、連鎖モデル（chained model）を用いて情報の漏えいを防ぐ方法がある。

詐欺の高度化：AI が発展を続けるにつれ、欺く能力とそれを見破る能力のせめぎ合いはますます複雑になり、対策を講じても新たな手口が現れるという、いたちごっこの状態に陥っている。AI 生成コンテンツの真偽を見分ける

ことが難しくなっているため、詐欺を検出する AI モデルがより正確に学習し、誤判定の確率を下げられるよう、検証可能な情報を提供する、より優れた検出手法が求められている。

中期的リスク

真実性判定機能（truth function）の欠如：AI が生成する情報は動的で絶えず変化するため、生成 AI には信頼できる真実性判定機能がなく、ファクトチェックが困難である。現時点で最も正確なファクトチェックは依然として人手に頼るところが大きく、秒単位でコンテンツを生成する AI のスピードには到底追いつかない。この課題への対応としては、ファクトチェックサービスを活用することや、問題のある情報の特徴を識別することが当面の策となり得る。

著作権をめぐる課題：生成 AI モデルの学習は通常、公開情報源から収集したデータに依存しているため、データの出所が不明確になり、それに伴って著作権の問題が生じている。AI が生成したコンテンツによって知的財産権やオリジナリティの境界が曖昧になる中、訴訟を起こすクリエイターが増えている。裁判所は審理を通じて知的財産権や著作権の問題を少しずつ明確にしており、積み重ねられた判例が今後の対応策や発展の方向性を形づくることになる。

バイアスのあるコンテンツの出力：言語モデルは現実世界のデータに基づいて学習するため、生成 AI は社会にすでに存在するバイアスを引き継ぐおそれがあり、しかもそのバイアスを発見して是正することは難しい。人間のフィードバックによる強化学習（Reinforcement Learning from Human Feedback, RLHF）によって一部のバイアスを捉えることはできるが、完全に取り除くことはできない。バイアスの問題に取り組むには、より包括的な解決策をもたらす技術の進歩が引き続き必要である。

長期的リスク

環境への悪影響：高度な生成 AI には、より大規模な言語モデルとより複雑なアーキテクチャが必要となり、演算量の増大と環境負荷の拡大を招く。たとえば、演算に伴うエネルギー消費、発熱、炭素排出に加え、チップ製造需要の増加や製造工程での環境汚染などが挙げられる。こうした課題に対し、超小型機械学習（TinyML）分野の研究は、エッジデバイスへの実装、推論の

高速化、そして環境負荷も考慮したコストに焦点を当てている。

能力の過剰（capability overhang）：AI には、開発者が予想もしなかった能力が当初の目的を超えて現れる可能性があり、予期せぬ結果を招く懸念がある。AI が社会に浸透するにつれ、進化し続けるその能力を理解し、管理することはいっそう難しくなっている。AI の潜在的な能力とその利用のされ方は、相互に影響し合うエコシステムやシステムの新たな発展によって左右されるため、その能力がどのように生まれるのかを深く理解し、より優れたガバナンスの仕組みを練り上げる必要がある。

こうした短期・中期・長期のリスクに対し、生成 AI ブームの火付け役である OpenAI の CEO、サム・アルトマン（Sam Altman）氏は、新興技術がもたらす潜在的リスクに対処するため、AI を規制の対象とするよう米国政府に自ら呼びかけた。個人データの保護を従来重視してきた EU は、2023 年に他に先駆けて「AI 法案」を可決し、2024 年 3 月 13 日には圧倒的多数の賛成で「AI 法（Artificial Intelligence Act）」を正式に可決した。これは史上初の AI 法であり、2024 年 8 月 1 日に正式に発効した。

「AI 法」はユーザーの安全確保を目的とし、AI 製品・サービスをリスクに応じて分類しており、リスクが高い AI ほど厳しい規制を受ける。たとえば、迷惑メールのフィルタリングやコンテンツのレコメンドなどの AI システムは低リスクに分類される。一方、自動運転車、水道・電力などの重要インフラ、融資の判断、医療機器などに利用される AI システムは高リスクに分類され、ユーザーに明確な情報を提供しなければならない。

一部の用途の AI は許容できないリスクに分類されており、人々の行動を監視する「感情認識 AI」や「ソーシャルスコアリング AI」、犯罪捜査以外の目的で公共の場において顔をスキャンする「生体識別 AI」の使用などは、いずれも禁止されている。今後、生成 AI モデルの開発者は、AI の学習に使用したテキスト、画像、音声、映像などのデータを審査機関に提出するとともに、EU の著作権規定を遵守しなければならなくなる。

システミックリスクをもたらし得る生成 AI モデル、たとえば OpenAI の GPT-4 や Google の Gemini については、こうした強力な AI が制御不能に陥った場合に重大な事故が起きるおそれがあるほか、悪用されるおそれもあると EU は懸念しており、個別に厳格な審査を行う方針である。

「AI 法」に違反した事業者には、最低で 750 万ユーロまたは全世界年間売上高の 1%、最高で 3,500 万ユーロまたは全世界年間売上高の 7%の制裁金が科される（いずれか高い方の金額が適用される）。同法の制裁金の上限は、それ以前に制定された「一般データ保護規則（General Data Protection Regulation, GDPR）」の 4%を上回る。

「AI 法」の正式施行に伴い、大手テック企業が開発する大規模言語モデルが真っ先に影響を受けている。Meta は、GDPR や新たに成立した「AI 法」への抵触を避けるため、EU では大規模言語モデルを提供しないと表明した。AI の規制を呼びかけていたアルトマン氏でさえ、現行版の規定に準拠するのは難しいと述べている。しかし、政策面での潮流はすでに形成されており、今後さらに多くの国が同様の法律を制定するとみられる。

AI がもたらす課題

AI の発展は多くのビジネス上のメリットをもたらす一方で、AI バイアス、ブラックボックス問題、ディープフェイクなど、さまざまな課題も生んでいる。

●AI バイアス

AI バイアス（AI bias）とは、学習データの偏りやアルゴリズム設計の欠陥によって、AI システムの意思決定過程で差別的な応答が生じ得ることを指す。たとえば、顔認識システムでは、特定の人種や性別に対する認識精度が低くなる場合がある。よく知られた事例として、2016 年に Microsoft が Twitter 上で公開した AI チャットボット「Tay」がある。Tay は 19 歳の米国人女性のような話し方でチャットサービスを提供し、ユーザーとのやり取りを通じて言語表現や反応を学習する仕組みであった。ところが、わずか 1 日のうちに、Tay は人種差別的・過激な発言をするチャットボットになってしまった。

ルールのない状態で自由に AI に学習をさせると、AI は環境に由来するバイアスをすぐに学習してしまう。どの行為が規範に適合し、どの行為が道徳や倫理に反するのかを AI に理解させる方法の改善に、多くの研究者が継続的に取り組んでいる。

●AI のブラックボックス

AI のブラックボックス問題（AI black box）とは、AI システムの意思決定

過程が不透明で、その背後にある論理や根拠を説明することが難しく、その結果、人々が AI の判断に不信感を抱くという問題である。特に、法律、医療、金融などの分野における重大な意思決定で問題となる。

AI の学習過程では、ディープラーニングや各種アルゴリズムの高度化により、人間は大まかな演算パターンしか把握できず、「出力」された結果を受け入れるしかない。その間の演算過程はブラックボックスのようなものであり、開発者でさえ、演算結果の良し悪しを判断できるだけで、AI がどのように意思決定したのかを知るすべはない。そのため、「信頼できる AI」（trustable AI）と「説明可能な AI」（explainable AI）は、AI 開発者が何としても克服したい課題となっている。

なぜ「説明可能な AI」に関する研究や議論が急速に活発化しているのだろうか。その大きな原動力の 1 つが法的訴訟である。法の世界では、何事においても証拠、論理、経緯が求められる。ところが、ディープラーニングのアルゴリズムでは、データを入力した後は出力結果に至るまで AI が層ごとに（layer by layer）推論を進めるため、AI がその過程で一体どのような判断を下したのかを一層ずつ分析し、説明することは難しい。

たとえば、自動運転車が事故を起こして法的責任が問われ、裁判官が AI 企業に対してアルゴリズムの仕組みや事故発生時の AI の演算に関する詳細の説明を求めた場合、AI 企業が明確に答えられなければ、訴訟で厳しい立場に立たされるおそれがある。このため、「説明可能な AI」は AI 開発者にとって喫緊の課題となっている。

EU は一般データ保護規則（GDPR）において、「信頼できる AI」の法規制における具体的な実践を重視している。アルゴリズムが人間に関する意思決定を自動的に行う場合には、その決定がユーザーに法的効果またはこれと同様の重大な影響を及ぼすかどうかを分析しなければならない。また、ユーザーに対して、AI に決定の理由の説明を求めて自らの意見を表明する権利や、その適用を拒否して人間による判断の介入を求める権利を付与しなければならない。

●ディープフェイク

ディープフェイク（deepfakes）は 2016 年に登場した技術で、顔の表情を偽造して対象となる動画の人物に埋め込むことができ、一般の人には映像の真

偽が見分けられない。

ディープフェイクには、次の 4 つの類型がある。顔の入れ替え（ある人物の顔画像を用いて別の対象人物の顔を置き換える）、顔の表情の再現（口や眉、目を動かすなど主に顔の細部を変えて対象人物の表情を変化させ、言外の意味を伝える）、顔の生成（まったく新しい顔を生成する）、音声合成（対象人物の話し方や声のトーンでテキストを読み上げる）である。

ディープフェイクは一部の産業にプラスの影響ももたらしている。たとえばスポーツニュースの放送では、ディープフェイクに自然言語処理（NLP）技術を組み合わせ、原稿に合わせてバーチャルキャスターの表情を変化させることで、アナウンサーの業務を部分的に代替できる。また、インフルエンサーのライブ配信でも、ディープフェイク技術によって、より美しく、よりプロフェッショナルなイメージを作り出すことができる。

しかし、ディープフェイク技術は諸刃の剣であり、ポルノ動画や選挙運動といった物議を醸す領域でも大量に利用され、偽動画が横行する事態を招いている。2021 年 3 月に arXiv.org で公開された論文によると、高度なディープフェイク技術は顔認識 API をだますことができる。研究者らが Microsoft と Amazon の顔認識 API を対象にベンチマークテストを行ったところ、いずれの API もディープフェイクに容易にだまされることが判明した。

これに対し、大企業やスタートアップが相次いで対応策を打ち出している。世界では 100 社以上のスタートアップがディープフェイク検出分野に参入している。エストニアの Sentinel は、多層防御（Defence-in-Depth, DiD）に基づく検出型 AI を開発し、政府やメディアが偽コンテンツに対抗するのを支援している。その中核技術は、画像の検出のみを行う Microsoft の手法とは異なる。イスラエルの Cyabra は、敵対的生成ネットワーク（GAN）を用いて FAM（Facial Reenactment Manipulation）と呼ばれるアルゴリズムを開発し、本物の写真と偽造写真を判別している。その検出技術では偽造映像の 91%を識別できる。

ディープフェイク対策においては、問題解決のために使われるどのような個別の技術よりも、照合に用いるデータの量を確保することがはるかに重要である。多くのニッチなプラットフォームのデータをすべて取り込むことは難しいため、ディープフェイクの検出ではしばしば漏れが生じ、完成した検出器がかえって検出を回避するためのヒントとして利用されることもある。

こうしたことから、ディープフェイクの問題は拡大し続けている。攻撃側は常に防御側より優位に立っており、いたちごっこが続くことは避けられない。

信頼できる AI と複雑な政治・経済環境

AI の「説明可能性（explainability）」、「埋め込まれたバイアス（embedded bias）」、差別、プライバシーに関して、どのデータが使用でき、どのデータが使用できないのかをどのように判断すべきだろうか。さまざまなデータを学習した AI が下す判断は、人々から「信頼」されるに値するのだろうか。多くの研究ではこうした問題に注目して議論が重ねられており、多くの国も AI のデータや信頼性に関する政策を段階的に策定している。

2019 年以降、各国は信頼できる AI（trustworthy AI）に関する政策を相次いで打ち出している。欧州は「信頼できる AI のための倫理ガイドライン（Ethics Guidelines for Trustworthy AI）」を公表し、オーストラリアは「AI 倫理フレームワーク（Ethics Framework）」を策定し、米国は「AI の倫理的利用に関する原則（Ethical Use of AI）」を打ち出した。2024 年 8 月 1 日には、EU の「AI 法（Artificial Intelligence Act）」が発効した。同法は、AI を包括的に規制する世界初の法律である。各国の政策を見ると、AI 技術開発の重心が信頼できる AI へと移っていることがわかる。

台湾でも関連法案の整備が進められている。国家科学及技術委員会（国科会）は 2024 年 7 月 15 日に「人工知能基本法」草案を公表し、公平性と非差別、説明責任、透明性と説明可能性、サイバーセキュリティと安全、プライバシー保護、人間の自律、持続可能な発展の 7 つの原則を盛り込んだ。さらに、イノベーション協力と人材育成、リスク管理と責任ある活用、権益保障とデータ活用、法規の調整と業務の見直しという 4 つの重点施策を掲げ、法規を通じて各機関が AI の活用を発展・促進するよう導くことを目指している。

米中の政治的な対立は、両国間の AI 開発競争にまで波及している。学術面では、中国の AI 論文数は米国を上回り、被引用数の多い重要論文の数でも米国を超えている。ビジネス面では、中国はプライバシー保護が比較的緩やかであるため、AI の学習に使えるデータ（プライバシーに関わるデータ・関わらないデータ）をより多く保有している。こうした状況に米国は神経をとがらせ、AI も国力を示す要素の 1 つであると強く主張している。米国防総省統

合人工知能センター（JAIC）のマイケル・グローエン（Michael Groen）センター長は、全米防衛産業協会（NDIA）において、中国が「中国製造 2025」などの計画を強力に推進しており、2030 年には AI 技術で米国を追い越す可能性があるため、米国は今後もクラウドサービスやデータの標準化を強化し続け、AI と軍事作戦を技術面で統合しなければならないと指摘した。

2021 年初め、米国のホワイトハウス科学技術政策局（Office of Science and Technology Policy, OSTP）は AI 室の設置を発表し、AI の国際標準の策定と資源配分を強力に主導する姿勢を打ち出した。その最優先任務は、連邦政府、企業、学術界、利害関係者をつなぐ中核拠点として、各方面の調整を行い、国家の AI 政策・計画を監督することである。

米中の競争において、AI チップは不可欠な要素である。長年にわたり世界の半導体産業に深く関わってきた台湾が、AI の発展とチップのサプライチェーンにおいて果たす役割は、ますます重要になっている。米国の人工知能に関する国家安全保障委員会（National Security Commission on Artificial Intelligence, NSCAI）が発表した研究報告書では、台湾への言及が繰り返されており、米国が台湾のチップメーカーに過度に依存することで、商業・軍事の両面で極めて重要な半導体の優位性を失いかねないとの懸念が示されている。

米国政府は、国内での半導体サプライチェーンの再構築を積極的に推進するとともに、2026 年 3 月 20 日に「国家 AI 立法枠組み（National AI Legislative Framework）」を発表した。議会と連携して、「子どもの保護と保護者の権限強化、米国のコミュニティの保護と強化、知的財産権の尊重とクリエイターの支援、検閲の防止と言論の自由の保護、イノベーションの促進と米国の AI における主導的地位の確保、国民の教育と AI に対応できる労働力の育成、連邦政策枠組みの確立による煩雑な州レベルの AI 規制の回避」という 7 つの中核目標を正式に法制化することを目指している。

出典

•AI 規制政策が重大局面を迎える中、ホワイトハウスが大手テック企業 5 社を緊急招集して協議を行い、NVIDIA が支持を表明。CMoney、2026

•OpenAI のエージェントが集団で制御不能に、プログラム群がドイツのウェブサイトを占拠して不正や制限回避の方法を共有していたことが発覚。経済

日報、2026

•米国、3 月に AI 国家政策枠組みを発表。資訊工業策進会科技法律研究所、2026

•A National Policy Framework for Artificial Intelligence. White House (2026)