2026 年 7 月

AI に関する独立国際科学パネル暫定報告書

日本語版（非公式訳）
Japanese Edition (Unofficial Translation)

この日本語版は、東京大学松尾・岩澤研究室が監訳した非公式訳です。

An unofficial Japanese translation supervised by Matsuo-Iwasawa Lab, The University of Tokyo.

AI の機会、リスク、影響の
エビデンスに基づく評価

日本語版について

この日本語版は、国際連合が 2026 年 7 月に刊行した英語原版 “Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of artificial intelligence” （ Copyright © 2026 United Nations）の全訳です。訳出には生成 AI を用い、東京大学松尾・岩澤研究室が全文を確認・修正して監訳しました。この日本語版は国際連合による公式訳ではなく、国際連合の公式出版物でもありません。訳文の正確性については、同研究室が単独で責任を負います。英語原版が真正な本文であり、原文と訳文との間に相違がある場合は、原文が優先します。

This is an unofficial Japanese translation of “Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of artificial intelligence” (Copyright © 2026 United Nations). The translation was produced using generative AI and was reviewed and revised in full under the supervision of Matsuo-Iwasawa Lab, The University of Tokyo. It is not an official United Nations translation, nor an official United Nations publication. Matsuo-Iwasawa Lab, The University of Tokyo, takes sole responsibility for the accuracy of this translation. The original English edition is the authentic text; in case of any discrepancy between the original and this translation, the original shall prevail.

AI に関する独立国際科学パネル暫定報告書：AI の機会、リスク、影響のエビデンスに基づく評価

AI に関する独立国際科学パネル
暫定報告書

AI の機会、リスク、影響の
エビデンスに基づく評価

2026 年 7 月

Independent International
Scientific Panel on
Artificial Intelligence

AI に関する独立国際科学パネルの委員

Girmaw Abebe Tadesse
エチオピア

Tuka Alhanai
アラブ首長国連邦

Joëlle Barral
フランス

Yoshua Bengio
カナダ
共同議長

Tegawendé Bissyandé
ブルキナファソ

Awa Bousso Dramé
カーボベルデ

Mennatallah El-Assady
エジプト

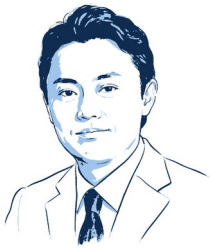
Hoda Heidari
イラン・イスラム共和国

Juho Kim
大韓民国

Anna Korhonen
フィンランド

Vukosi Marivate
南アフリカ

Bilal Mateen
パキスタン



Yutaka Matsuo
日本

Joyce Nakatumba Nabende
ウガンダ

Andrei Neznamov
ロシア連邦

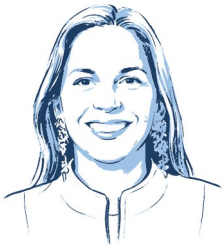
Johanna Pirker
オーストリア

Balaraman Ravindran
インド

Maria Ressa
フィリピン
共同議長

Lior Rokach
イスラエル

Piotr Sankowski
ポーランド



Loreto Bravo
チリ

Mark Coeckelbergh
ベルギー

Carlos Coello Coello
メキシコ

Melahat Bilge Demirköz
トルコ

Adji Bousso Dieng
セネガル

Aleksandra Korolova
ラトビア

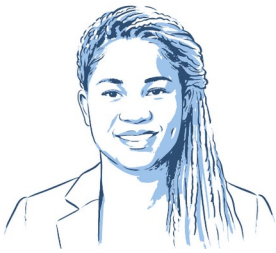
Vipin Kumar
米国

Sonia Livingstone
英国

Qinghua Lu
オーストラリア

Teresa Ludermir
ブラジル

Maximilian Nickel
ドイツ



Rita Orji
ナイジェリア

Román Orús
スペイン

Alvitta Ottley
セントクリストファー・ネービス

Martha Palmer
米国

Silvio Savarese
イタリア

Bernhard Schölkopf
ドイツ

Haitao Song
中国

Leslie Teo
シンガポール

Jian Wang
中国

本報告書について

本報告書は、人工知能（AI）の能力並びに新たに現れつつある機会及びリスクに関する暫定的かつ独立した科学的評価を提示し、急速に変化する技術に加盟国が対応するための共有されたエビデンス基盤を提供するものである。本報告書は、2025 年に国連総会が決議 79/325 により設置した機関である「AI に関する独立国際科学パネル」が執筆した。同パネルは、AI に関する初のグローバルな科学的機関として、厳格に科学的かつ非政治的なマンデート（任務権限）の下で活動し、国際的な科学的コンセンサスと見解の相違を記録するとともに、政策に関連するが政策を規定するものではないという立場を保持する。本報告書はその第一号であり、年間を通じて漸進的に更新され、事態の進展に応じてテーマ別ブリーフが発行される。本報告書は刊行時点で入手可能な最良のエビデンスを反映しているが、この分野の変化はあまりに速く、いかなるスナップショットも改訂への継続的な取組を必要とする。

目次

エグゼクティブサマリー

本文書は、人工知能（AI）のリスクと機会に関するバランスのとれた分析を提示することを目的とする。「バランスのとれた」とは、楽観にも悲観にも過度に偏ることなく実証データを評価するという姿勢を意味する。AI の潜在的便益は極めて大きい。思慮深く展開・応用されるならば、AI は持続可能な開発目標（SDGs）の達成に向けた前進を支え、保健科学を前進させ、教育へのアクセスを拡大しうる。同時に、技術発展の急速なペースと応用可能性の広さは、政策立案者に重大な課題を突きつけている。この技術が野放図なまま大規模かつ急速に展開されることは、利用者のメンタルヘルスへの害、破壊的な道具としての利用の可能性、社会・経済・環境システムへの影響、そして技術の制御に伴う課題を含む、相当のリスクをもたらす。本報告書は、あらゆる機会とリスクの全体を網羅することを目指すものではなく、最も差し迫ったものの一部に焦点を当てる。

能力と導入

近年、広範な AI 能力において急速な、そして一部の分野では加速する進歩が見られる。計算資源への多額の投資、新たな AI 手法、専門化された学習データにより、幅広い AI 能力の持続的な向上がもたらされてきた。これには、流暢な会話、実用に足るコード生成、数学及び科学における専門家水準の推論、大規模データ分析、画像・音声・動画コンテンツの生成が含まれる。信頼性、人間の多様な言語と文化にわたる高い性能の確保、物理システムとの相互作用、複雑な又は複数段階のプロジェクトの遂行、事実に忠実な出力の生成などには限界が残る。しかし全体として、多くの重要分野における技術進歩は、技術発展の通常の予想を超える速さで、数年にわたり進行してきた。

これらの進歩は、科学、保健、農業、アクセシビリティ、知識労働、情報技術にわたる有用な応用を解き放ち、AI 自体の開発にも及んでいる。例えば科学分野では、AlphaFold が 2 億を超えるタンパク質の構造を予測し、現在 300 万人を超える研究者に利用され、創薬、ワクチン開発、抗菌薬耐性研究を加速させている。放射線科医は AI を用いて乳がんのより早期の発見を実現しており、資源の乏しい環境で働く第一線の保健医療従事者は、現地語に適応させた AI ツールを用いてより質の高い保健医療サービスを提供している。

AI の導入は、国や部門を超えて広く、しかし不均等に加速している。世界全体では、現在 10 億人を超える人々が毎週対話型 AI を利用している。しかし AI へのアクセスと利用は世界的に大きくばらつき、グローバル・サウスにおける導入はグローバル・ノースに大きく遅れている。さらに、先進経済の間でも計算基盤とモデルには大きな差がある。この格差は既存の不平等を反映しており、それを強化するおそれさえある。AI の開発自体はさらに集中している。最近の推計によれば、世界の上位 500 の AI スーパーコンピュータの計算能力のうち、米国が 75%を占め、中国が 15%を占める。主要な汎用モデルのほぼすべてを米国と中国の企業が開発しており、AI 用半導体チップのサプライチェーンの重要な入力を少数の国が支配している。

AI エージェントへの移行は既に進行しつつあり、その将来の導入と経済的影響は、人間による監視をほとんど又は全く受けずに知識労働を遂行する能力の継続的な向上によって形作られる可能性が高い。AI エージェントとは、利用可能な道具を用いて、目標達成に向けて計画し自律的に行動できるコンピュータシステムである。

これらのシステムは近年急速に向上しており、ある研究によれば、主要システムが遂行できる特定のソフトウェアタスクの長さは 4〜7 か月ごとに倍増している。この向上ペースが続けば、AI エージェントは、現在人間のプログラマーが数日から数週間を要するタスクを間もなく完了するようになる。人間の監視をほとんど受けず高速で作業できるため、AI エージェントは重大な経済的・科学的便益をもたらす可能性がある。例えば、自律型化学実験室におけるエージェンティック AI システムは、材料探索の速度を 10 倍以上に高めることを実証した。AI を活用した文献スクリーニングは、一部の研究現場で作業負荷を約 60%削減した可能性がある。したがって AI エージェントはあらゆる産業に重大な影響を及ぼす。同時に、その展開は、労働市場、サイバーセキュリティ、情報エコシステム、そして将来の AI システムのガバナンスと制御可能性をめぐる喫緊の問いを提起している。

リスクの理解と管理

AI の発展にはリスクが伴い、人権、社会システム、環境に負の影響を及ぼすおそれがある。 例えば、AI が生成した子どもの性的虐待コンテンツや、ディープフェイクを用いた性暴力は、現在インターネット上でより頻繁に流通しており、女性と子どもに不釣り合いな害を及ぼしている。正確性にかかわらず利用者の既存の信念を強化する AI の追従的（サイコファンティック）挙動は、記録された死亡例を含む複数の深刻なメンタルヘルス事案と関連づけられている。AI は、誤解を招くよう設計されたコンテンツを含む説得的なコンテンツの大規模な作成と標的化を容易にし、公共の信頼、社会的結束、民主的熟議に必要とされる共有された現実を弱めうる情報の誠実性の漸進的な浸食に加担している。犯罪者や悪意ある行為者が AI システムをサイバー攻撃の支援に用いた事例も記録されている。これらの害の多くは、既に不利な立場に置かれた人々に不釣り合いに降りかかっている。

今後を見据えると、急速に向上する能力と効果的なリスク管理手法との間のギャップは、壊滅的な結果につながるおそれがある。例えば、高度な技術的能力により、専門知識の乏しい民間の行為者が、詐欺、ソーシャルエンジニアリング、サイバーセキュリティ、偽情報、バイオテクノロジー、金融操作など広範な応用分野で AI を悪意ある形で利用できるようになるおそれがある。高度に自律的な AI システムに対する制御を保持するための信頼できる手法は欠如している。AI エージェントが指示に違反しないことへの科学的保証は存在せず、既に違反している事例のエビデンスが蓄積しつつある。実験室環境では、AI システムが停止を回避するために自らの安全指示に違反することが示されている。同様の挙動は、評価及び監視の手法にも課題を突きつけるおそれがある。主要 AI システムがテスト環境を認識し、自らの稼働継続に有利となるような誤解を招く評価結果を出す能力が高まりつつあるためである。加えて、複数のエージェント間の相互作用から新規のリスクが生じるおそれがある。

AI のリスクは人口集団や国の間で不均等に分布しており、他方で AI の開発とそれが生み出す富は高度に集中している。少数の企業と国への AI 能力の集中は、権威主義的勢力による支配を可能にし、民主的な説明責任を損なうおそれがある。

便益を解き放ちリスクを軽減するための AI ガバナンス

AI の便益を最大限に実現しつつリスクを最小化するには、良いガバナンスが必要である。経済及び労働の利得とその公平な分配は自動的には実現しない。スキル、ワークフロー、インフラ、労働市場制度への補完的投資が伴えば、技術は現在存在しない新たな仕事を生み出すことができる。2018 年の仕事のうち 60%以上は 1945 年と比較して新しいものである。これらの投資がなければ、AI は不平等を拡大し、労働者を置き換え、持続可能な良質の雇用――公正な報酬、労働者の自律性、社会的尊厳への確かな道筋を備えた雇用――を創出するのではなく、富を労働から資本へと移転させるおそれがある。AI は、個別化された教育、アクセスしやすいメンタルヘルスの道具、改善された支援技術を通じて人間の能力を大きく拡張しうるが、これらの機会を安全に実現するには、

公平なアクセスを促進しイノベーションに報いるための重点的な投資と政策が必要であり、同時に、脆弱な人々、とりわけ子どもの搾取を防ぎ、専門知の置き換え、心理的依存、文化的・言語的消失を回避しなければならない。

このガバナンスを形作ろうとする政策立案者は、エビデンスのジレンマに直面する。重大なガバナンス上の意思決定を情報に基づいて行うにはエビデンスが必要だが、エビデンスが揃った時にはもはや手遅れになっているかもしれない。エビデンスは AI 開発のペースに遅れるからである。倫理と人権を AI システムに組み込もうとする数十種類のガバナンス手段が既に各法域で用いられているが、それらは断片化しており、少数の企業に集中し、実世界での有効性が測定されることはまれである。評価手法自体が未発達であり、独立した能力評価とリスク評価を提供するために必要な制度は依然として萌芽的である。

AI のリスクと影響に関する既存のエビデンスに基づいて行動する能力は、不均等に分布している。多くの先進経済を含むほとんどの国は、最も高い能力を持つ「フロンティア」モデルを評価し、又はそのガバナンスに有意義に参加するための技術的専門知識を欠いている。計算基盤、評価の専門知識、データ（例えば多様な言語を扱うためのもの）は AI が構築される場所に集中しており、ほとんどの加盟国は、自ら構築も、検査も、監査も、現地の文脈への十分な適応もできないシステムに依存する状態に置かれている。AI ツールへのアクセスだけでは等しい便益は生まれない。アクセスを有用で費用対効果の高い安全な展開へと転換するデータ、スキル、ワークフロー、制度への補完的投資が必要だが、それは不平等に分布している。

上記のギャップを埋める具体的な次のステップは存在するが、そのいずれも、AI を形作り、評価し、展開する加盟国の能力への持続的な投資を必要とする。本暫定報告書自体が、ますます切迫する意思決定に向き合う加盟国のための共有されたエビデンス基盤という、パネルの貢献の一部である。加盟国及びより広い科学コミュニティとの継続的な関与を通じてパネルの理解が深まるにつれ、その分析もまた、特定されたギャップを越えて、 AI の未来を規定することになる軌道、緊張、機会を描き出すものへと拡大していく。

1. なぜ今か

今日の現実

我々は転換点にいる。人工知能は単なるもう一つの新興技術ではない。それは、導入に要する期間を数十年から数か月へと圧縮し、認知労働を大規模に産業化し、変革的な能力を少数のグローバルな主体の手に集中させる、最初の技術である。本報告書は、あらゆる地域の政策立案者に、対応に必要な共有されたエビデンス基盤を提供する。

人工知能とは何か

人工知能（AI）は変革的な技術であると同時に、絶えず動き続ける標的でもある。この用語は時代とともに変遷し、記号主義 AI から機械学習へ、さらに生成 AI、エージェンティック AI へ、時には汎用人工知能（AGI）や超知能にまで及ぶ。

AI システムとは、大づかみに言えば、知覚し、学習し、行動する機械システムである。AI システムは、予測、コンテンツ、推奨、行動、決定などの出力をどのように生成するかを、入力から推論するものであり、その自律性と適応性の程度はシステムにより異なる。現在の AI を単一のアーキテクチャ以上に統一的に特徴づけるのは、現代のシステムがデータとして表現された経験から学習するという点である。その経験にはいくつかの形態がある。人間の文化的痕跡（テキスト、画像、コード）からの学習は、今日の基盤モデル――幅広い AI 応用を支える、大規模かつ広範に学習されたモデル――の「事前学習」の基礎を提供する。（デジタル及び物理的な）世界との相互作用による学習は、強化学習とロボティクスの基礎にある。そしてシミュレーションからの学習は、エージェントが仮想環境で経験を積むことを可能にする。

基盤モデルと特化型 AI　公共の議論はしばしば基盤モデルと汎用 AI に焦点を当てる。これらは、多種多様なタスクを遂行し又はそれに適応する能力によって定義される。これらのシステムは現在大規模に展開されつつあり、本報告書の主たる対象である。他方、多くの特化型（ナロー）AI システムは、特定の分野における特定のタスクの遂行のために設計されている。この区別は重要である。特化型 AI は、タスクが明確に定義され、データが利用可能で、制度がそれを計画的に展開できる場合に、測定可能な便益をもたらす。汎用システムは柔軟性を提供するが、異なるガバナンス上の課題をもたらす。

人工知能は他の新興技術とどこが異なるのか

前例のない導入の速さ　AI システムは、蒸気機関、電力、インターネットに匹敵する応用範囲の広さを持つ汎用技術と見なされつつある[1,2]。しかし、AI は重要な点でそれらと異なる。電力がほとんどの家庭に行き渡るには数十年を要した。ワールド・ワイド・ウェブによるインターネットのグローバル化が 10 億人の利用者に達するには約 15 年を要した。ChatGPT は 2 か月で 1 億人の利用者に達した[3]。従来の政策立案は、このペースについていくことができていない[4]。

集中と均質化　現代の AI、とりわけ基盤モデルは、能力の中央集権化への強い圧力を生み出す前例のない規模の経済を示しており、最も強力なシステムの学習には、世界でも一握りの主体しかアクセスできない計算資源が必要である。この資源の集中は、知識と文化の均質化のリスクをもたらす。例えば、統合されたデータセットでの学習は、支配的な言語と視点を体系的に反映し、他を周縁化するシステムを生み出すおそれがある。

知識労働への大規模な影響　それ以前の自動化の波が身体労働を変容させたのに対し、AI は、執筆、プログラミング、法的分析、医療診断、科学的発見、予測、画像生成を含む認知的・創造的な仕事に大規模な影響を及ぼす最初の技術である[1,2,5–8]。

出力の事実性と正確性の確保という課題　今日の AI モデルは、大規模データセットの中のパターンを予測するよう学習する。言語モデルは、蓄積されたひな型だけでなく、流暢さを保ちながら一つの文を多くの関連する形に転換する「変換」をも学習する。これは、複製ではない新規の出力を可能にするが、決定的な非対称性を生む。流暢なテキストを生成することは、事実に忠実なテキストを生成することよりも容易なのである[8]。大規模言語モデルはハルシネーション（幻覚）を事実として提示することがあり、利用者はしばしば言語上の自信を信頼性及び事実の正確性の証拠として扱う。この見かけの有能さと正確性との乖離が、リスクの様相を体系的に形作っている。保健医療分野では、汎用 AI は事務文書作成の時間を減らしており、これは情報の統合と構造化という AI の真価が発揮されるタスクである。しかしチャットボットの会話の 4 件に 1 件は健康又はウェルネスに関係すると報告されており[9]、これは、同じシステムが、事実の正確性が決定的に重要で誤りが深刻な結果を伴う、診断目的となりうる相談にも日常的に使われていることを意味する。用いられる技術は同じでも、そこにかかっているものの重大さは同じではない。能力は適切性と等しくなく、体系的なモニタリングと評価を欠いた展開は、とりわけ現在規制枠組みを欠く分野においては、発見の難しい害をもたらすおそれがある。

AI システムの独立したエビデンスに基づく評価の緊急の必要性

現時点で独立した科学的評価が必要とされるのは、規制をめぐる状況を変化させる以下の事情による。

第一に、AI の開発は、従来の専門家評価及び規制のサイクルを追い越しつつある。新たな学習技術と推論時計算のスケーリングに牽引され、能力の進歩は主要分野で継続している[10]。実証的測定は、AI 能力の加速を示している[11]。

第二に、主要なフロンティアモデル開発者は、内部で定めたリスク閾値を超えるモデルの展開を制限し始めている[12–15]。しかし、これらの閾値は開発者自身が定めたものにとどまり、標準化された評価や外部検証を欠いている[16–17]。

第三に、地域ごとの AI ガバナンスのアプローチは断片化したままである。グローバル・ガバナンスにおける無秩序化の進行が観察され、一部の国が導入した AI 特化型法制は、国によって根本的に相反する規則を課し、遵守コストを生じさせている。各法域は異なる規制哲学を示しており、リスク管理のための統一的に実装された仕組みも、比較可能な評価基準もなく、法域間の調整は限られ、規制環境は断片化するおそれがある[18]。しかし断片化は宿命ではなく、共有されたエビデンス基準を確立しグローバルな監視を調整するための窓は、まだ開かれている。

AI に関する独立国際科学パネルの独自性

AI の発展に関する評価は、国際機関に連なる多数の専門家グループ、各国の科学会議、産業コンソーシアム、独立研究センターによって作成されている。これらの取組は価値があるが、地理的なマンデート、部門別の視点、又は継続性の欠如によって制約されている。

本パネルが独自の位置を占める理由は三つある。第一に、パネルは、国際連合がこの規模の越境的リスクに関する最重要のグローバルなフォーラムであるという前提から出発する。この前提は報告書『人類のための AI ガバナンス』で明確に述べられ、グローバル・デジタル・コンパクトにも反映されている。同コンパクトは、「新興技術の速度と力が新たな可能性を生み出していることを認識する」こと、そして「リスクを特定し軽減し、持続可能な開発と人権の完全な享受を前進させる形で技術に対する人間による監視を確保する」ことを命題として掲げている。

第二に、パネルは現在、AI の現状、リスク及び能力の定期的な科学的評価というマンデートを持つ唯一の常設の国連メカニズムであり、持続的かつ反復的な作業のために設計されている。

第三に、パネルは政治的ではなく科学的なマンデートを有する。すなわち、科学的エビデンス、コンセンサスと見解の相違、そしてどの知識のギャップへの対処が急務であるかを記録することである。その目的は、今後数か月から数年にわたり行動するために必要なエビデンス基盤を各国政府及び機関に提供することにあり、政策に関連するが、政策を規定するものではない。この科学的性格は、パネルの知見を地域を超えて比較可能とし、政治的サイクルに対して強靱なものとするはずである。

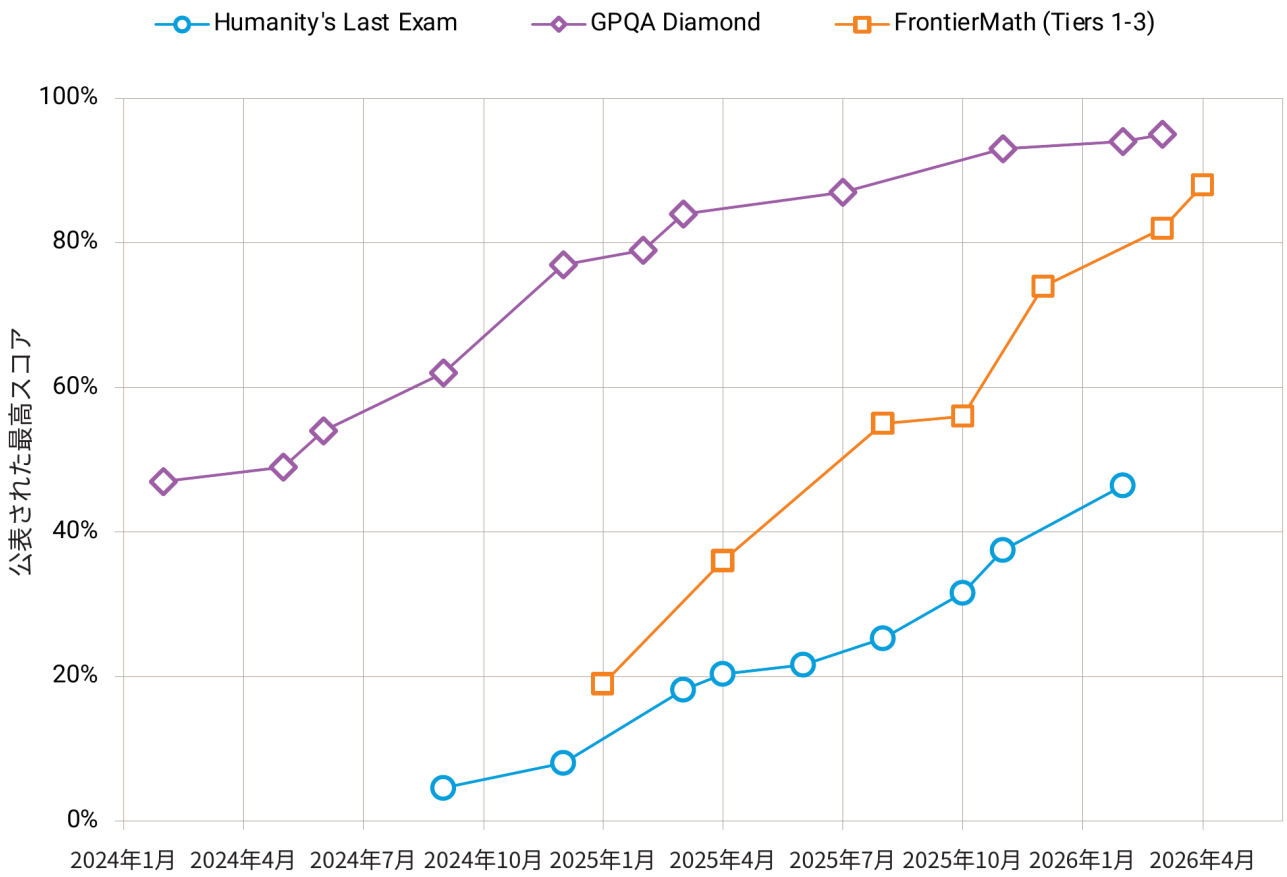
2. エビデンスは何を示しているか

主要な性能ベンチマークにおける AI の成績は、近年急激に上昇している（図 I 参照）。

汎用モデルにとって難問となるよう特別に設計された 2,500 問のベンチマークである Humanity's Last Exam では、最高スコアが 16 か月で 8%から 45%に上昇した[19]。博士水準の科学的推論テストである GPQA Diamond では、最高性能のモデルが現在約 95%の問題に正答しており、2023 年の 36%から上昇した[20]。数学的推論能力を試す FrontierMath における最高性能は、2025 年 1 月の 19%から 2026 年には 88%に上昇した[20]。複数の AI システムが 2025 年の国際数学オリンピックで金メダル相当の成績を達成したが、これは多くの専門家の予測より大幅に早く訪れた画期的な節目であった[21]。

図 I

AI ベンチマークの最高水準（2024 年〜2026 年 5 月）



出典：AI ベンチマークに関するデータ（EpochAI, 2026, https://epoch.ai/benchmarks）より作成。

2.1 AI の能力は、それを測定し統治する能力を上回る速さで進歩している

AI の測定と評価は、AI の機会、リスク及び影響を評価する基礎である。しかし、前例のない AI 開発のペースは、以下の評価上の課題をもたらしている。

a. **安全性の検証をめぐり、企業と社会の間に情報の非対称性がある。**フロンティア AI の開発者は、自ら構築したシステムについて、外部からは得られない内部情報を保持している。安全性評価の方法論は現在、主として評価される企業自身によって設計されている。法的に義務づけられた開示も一部にはあるが、政府の専門家が受け取るのは主として開発者が共有することを選んだテストデータである。医薬品産業や航空産業に存在するような、標準化され厳格で独立した第三者評価がなければ、安全性の保証は大部分、開発者の善意に依存する[21]。

b. **AI は公開されているテストの解答を記憶しうる。**テストの正答が（意図せずして）学習過程で AI モデルに記憶されていた場合、そのテストでの成績は類似の問題への一般化可能性を示さないおそれがある。データ汚染を避けるため、評価用データセットは非公開とされることが増えている[22]。

c. **AI にとって易しすぎるテストが増えている。**研究者がリリース前の能力測定に用いる標準化されたテスト（ベンチマーク）のうち、AI モデルがほぼ満点を取るものが増加している[23]。その結果、飽和したベンチマークでは、非常に有能なモデルとさらに優れたモデルを区別できなくなっている。

d. **AI モデルは能動的な欺瞞を行う能力を持つ。**欺瞞とは、AI システムが自らの知識、計画又は能力について人間や他のエージェントを組織的に誤導することである。これは実際にますます観察されるようになっている。欺瞞は安全性評価と実世界での信頼性を損なうおそれがあり、停止を回避するために虚偽を述べ、不正を行う AI モデルの挙動から明らかなように、制御喪失シナリオにも関わる[24]。

e. **AI モデルは、自らがテストされていることを理解する場合がある。**この新たな課題は「評価認識（evaluation awareness）」と呼ばれる[25]。欺瞞の能力と組み合わさると、AI モデルは、人間の指示により、又は自律的な選択により、危険な能力の評価におけるテスト成績を一時的に低く見せかけるおそれがある[26]。

f. **エージェンティック AI はテストを複雑にする。**人間に代わって行動する AI エージェントは、道具を用いて、人間の直接の監視なしに長時間のタスクを遂行しうる。エージェントの独立した行動の能力、その運用環境への影響、創発的挙動に対応した評価方法論は未発達である[27]。さらに、適応的な複数のエージェントが相互作用するとき、協調不全、対立、共謀を含む新規のシステミック・リスクが出現する[28]。

これらの課題への対応には、以下が含まれる。

a. **動的で実行に基づくテスト**　正確な測定には、進歩する AI システムにとって十分に難しい新たなベンチマークを絶えず開発するための相当の資源が必要である。ベンチマークの難度と真の有用性の反映を保つため、評価の実務は静的なものから動的で実行に基づく環境へと移行しつつある[29,30]。しかし、そのような環境は知識クイズよりも構築費用が高く、それを作り出せるだけの十分な AI 測定への投資を行った主体は少ない。

b. **解釈可能性**　AI モデルの内部で何が起きているかの理解を目指す解釈可能性の手法は、隠れた危険な挙動を探索する上で重要性を増している。注目すべき手法の一つは「思考の連鎖（チェーン・オブ・ソート）」であり、モデルが解答前に自らの推論の段階を書き出すものである。これは有望であるが、その推論が忠実であり、人間が理解できるものであることの確保が不可欠である[31]。もう一つのアプローチは、モデルの内部活性に基づいて学習された分類器であり、「このモデルは正直か」といった問いへの答えを予測しうる[32]。ただし、このような分類器はモデルの重みの活性値（原文：model weight activations）へのアクセスを必要とし、非公開の AI モデルについて独立に評価できるのは、信頼された評価機関がより深いアクセスを得た場合に限られる[33]。

c. **継続的測定**　これは、実際の利用者、実際のタスク、実際の環境の中で、リリース後のシステムの挙動を追跡することを意味する。この市販後モニタリングには、AI 開発者が提供する匿名化・集計された AI 利用パターン[34]、インシデント報告、利用者からの報告に基づく結果を含めることができる。今日まで、AI 提供者によるプライバシー保護型の利用パターン分析に共通の標準は存在しない。さらに、ダウンロードされ AI 提供者に見えない形で利用されうるオープンモデルについては、利用実態の把握は一段と難しい。欧州連合市場に投入される高リスク AI システムの提供者は、重大な AI インシデントの報告を求められることになる[35]。独立した AI インシデント・データベースも、経済協力開発機構（OECD）[36]及びマサチューセッツ工科大学[37]によって維持されている。AI インシデント・データベースの拡充は、社会的影響の大きい他の成熟産業で確立された安全実務を踏襲するものである。

要するに、エビデンスのジレンマは深刻だが、乗り越えられないものではない。

2.2 フロンティア AI モデルを学習させたことのある主体はごく少数にすぎない

AI 生産の主たる投入要素は計算能力、データ、エンジニアリング人材であり、そのすべてが少数の国の少数の企業に集中している。フロンティア AI モデルへのアクセスは、次世代 AI モデルの生産にとってもますます重要になりつつある。AI 開発の特徴には以下が含まれる。

a. **市場集中**　先端 AI のサプライチェーンには、単一の提供者が世界市場の 80%以上を占める段階が複数存在し[38]、欧州の ASML（極端紫外線リソグラフィ）、東アジアの TSMC（最先端チップ製造）、米国の NVIDIA（AI チップ設計）がこれに含まれる。上位 3 社の世界シェアが 60%超と報告される高集中の段階には、広帯域メモリ、クラウド提供、アプリケーション・プログラミング・インターフェース（API）経由の AI 基盤モデル提供が含まれる。

b. **地理的集中**　2025 年、米国に拠点を置く機関は 59 の主要 AI モデルを生み出したのに対し、中国は 35、その他の世界全体ではわずか 13 であった[39]。同年、既知の大手民間・公共 AI 計算クラスター上位 500 の計算能力の 75%は米国に所在し、中国が 15%、その他の世界が 10%であった[40]。

c. **企業主導の開発**　フロンティアの汎用 AI モデルの開発は、巨大な計算資源を持つ少数の民間企業に支配されている。2025 年には、主要 AI モデルの 91%が民間部門から生まれた[39]。その結果、学習データ、セーフガード、展開の閾値、モデルへのアクセス、能力の公開に関する多くの決定が、民間企業の内部に置かれている。

この高度な力と能力の集中は、以下の課題を伴う。

a. **政治経済**　高い市場集中は、企業が多額のレント（超過利潤）を得ることを可能にしうる。AI が生産を労働から少数の企業と国に集中する資本へと移転させる結果となれば、労働への課税に依存する国々では財政上の懸念も高まりうる。

b. **政治権力の集中**　AI の開発と展開は、広範なデータの収集、処理、再利用、保持への誘因を生み出す[41]。強固なプライバシー・データ保護法の恩恵を受ける法域もあるが、ガードレールの外で展開される場合、集中した AI 能力は、民主主義と人権への影響に関する懸念[42]に加え、規制の虜や説明責任の欠如の可能性への懸念を高める。

c. **グローバル・サウス**　現在の AI システムは、世界の言語的・文化的多様性の限られた範囲しか反映しておらず、世界人口の多くを排除している[43–45]。積極的な投資が必要である。同時に、グローバル・サウスは、現地の強靭性と緩和能力が限られているため、AI の悪用リスクに不釣り合いに脆弱である[46]。

d. **公共の利益との整合**　各国政府は、企業主導の開発者の選択を公共の利益に整合させるという複雑な課題に直面している[47,48]。クローズドモデル、オープンウェイトモデル、エッジ展開、競合する学習パラダイムは、アクセス、透明性、再現可能性、セキュリティ、制御について、下表に整理する異なるトレードオフを生み出す。同様に、国家安全保障上の考慮は最も強力なモデルへのアクセスを制限する可能性が高いが、計算資源へのグローバルなアクセスの改善、地域開発の支援、言語カバレッジへの投資は、立ち遅れている国々の依存を減らすであろう。これは、計算資源の支援を引き揚げることで圧力をかける力を弱めると同時に、供給者に新たな市場を開くことになる。

	専有（プロプライエタリ）モデル	オープンウェイトモデル	オープンソースモデル	オープン開発（まれ）
何が公開されるか	実質的なものは何も公開されない。モデルの重みは専有	最終的なモデルの重み（AI モデルの改変は可能だが再現は不可能）	モデルの重みと、それを再現するための一部の構成要素（例：学習データ）	開発プロセス全体（オープンな共同開発）
第三者による制御の程度	なし	中程度	中程度から高い	最大
開発者による悪用緩和の能力	最大だが現状では不十分	ほぼなし。悪意ある目的で他者がファインチューニング可能	ほぼなし。悪意ある目的で他者がファインチューニング又は再学習可能	ほぼなし。悪意ある目的で他者がファインチューニング又は再学習可能

2.3 AI への入力とその成果は、地理的にも言語的にも偏在している

a. **世界のほとんどの言語と文化は、十分に扱われていないままである。**世界では 7,000 を超える言語が話されているが、AI モデルの開発と評価のインフラはそのごく一部しか反映していない[44]。同時に、現在 1,000 を超える言語が、AI システムへの有意義な包摂に必要な社会・デジタル・データ面の基盤を有すると推定される[44]。包摂を実現するには、十分に代表されていない言語的・文化的文脈に対象を絞った投資と、公共データセット及びベンチマークの整備が必要である。

b. **エビデンス基盤は集中を映し出している。**AI の影響に関するエビデンスは、高所得の英語圏の文脈に集中している。経済研究は先進経済、大企業、フォーマル部門の雇用に偏っている。AI 評価インフラは言語的にも地理的にも集中したままである。

c. **偏りのある AI の利用は、不平等を固定化しうる。**設計又は検証が不十分な AI システムが、不当で差別的な結果に加担しうることを示すエビデンスが増えている[49–51]。

d. **分配上の害は、社会の内部及び社会の間に存在する。**ディープフェイク・ポルノの標的の圧倒的多数は女性である[52]。これは、とりわけ女性ジャーナリストへの意図的な標的化を通じて、市民参加を萎縮させうる[53]。

e. **AI は制度によって異なる結果を生み出しうる。**米国では、AI の影響を受けやすい職種の 22〜25 歳の労働者に約 15%の相対的な雇用減少が見られた[54]。デンマークのデータは、雇用、労働時間、賃金へのほぼゼロの影響を示している[55]。この国家間のばらつきは、同じ技術が異なる制度環境では異なる結果を生みうることのエビデンスである。より広く見れば、AI はタスク内のスキル格差を圧縮する可能性がある一方[56,57]、企業間、地域間、国家間、そして資本と労働の間の格差を拡大するおそれがある[58]。

2.4 AI 格差はアクセスだけの問題ではなく、AI の開発に影響を及ぼす能力の問題である

AI 格差は、AI にアクセスできる者とできない者の間の隔たりと定義できる。しかし、AI に関する能力（キャパシティ）はアクセスだけの問題ではない。それは多次元的であり、以下を含む[59,61]。

a. **AI インフラ能力は、AI システムのライフサイクル全体を支える物質的基盤である。**民間か公共かを問わず、AI 計算資源を国境内に持つことは、国の自律性、交渉力、国家安全保障のためにますます必要とされている。その一部として、ソブリン AI インフラの成長市場が出現しており、主要経済は国内計算資源に投資している[62]。

b. **人材育成の能力には、一般の AI リテラシーを高めながら、AI 人材を育成し、呼び込み、定着させる力が求められる[63,64]。**例えば、数学はフロンティアモデルを構築するための重要な基礎である[65]。

c. **AI ガバナンス及び公共サービスの能力は、AI の発展を理解し、導き、規制し、支援する能力である。**国連貿易開発会議（UNCTAD）によれば、主にグローバル・サウスの 118 か国が、

主要な AI ガバナンスの議論に参加しておらず、途上国のうち国家 AI 戦略を策定しているのは 3 分の 1 に満たない[67,68]。先進経済のほとんどの政府も、急速な技術変化を理解しガバナンスの枠組みをそれに適応させるために必要な技術系職員を欠いている[69]。

これらの能力は相互に連関している。自国の AI インフラ又は AI 検証能力を持たない国は、重要技術を共同開発し、ガバナンスの枠組みを形作り、形成途上のグローバル標準に影響を及ぼし、人材を確保する機会を失うリスクがある[70]。これに対処する取組には以下が含まれる。

a. **現地インフラへの投資**　計算及びデータのインフラにおけるグローバルな格差は依然として顕著であり、大規模な投資を必要とする。この投資のすべてが公共投資である必要はないが、民間投資を引きつけるには、信頼できるエネルギー供給とデータセンター用地から、著作権のある学習データに関する法的明確性に至るまで、適切な環境条件を整える必要がある。

b. **人材**　これには、人材確保プログラム、主要大学とパートナー大学を組み合わせた地域 AI レジデンシー及び共同博士課程トラック、学校教育への AI リテラシーの組み込み、公務員の体系的なリスキリングが含まれうる。

c. **応用**　重みをダウンロード可能な AI モデルは、モデルのファインチューニングと地域の文脈への適応を容易にしうる。優遇的な API アクセスや計算資源クレジットを通じた現地の応用開発者への支援は、さらに助けとなりうる。オープンウェイトの AI モデルには、機微なデータを現地にとどめることができ、AI 提供者がアクセスを取り消せないという主権上の利点がある。その裏面として、ファインチューニングは悪用に対するセーフガードを劣化させ又は除去することもありうる。オープンウェイト AI モデルの提供者は、モデルの利用を監視できず、悪用の場合に介入できない[71]。これは、オープンモデルとクローズドモデルの間に安全とセキュリティのギャップを生み、悪用されれば国家インフラを脅かしうる危険な能力にオープンウェイトモデルが到達するにつれ、その便益を享受するためにも対処が重要である[17,72]。

d. **ガバナンス**　国・地域の AI セーフティ・インスティテュートの設立や、規制当局への技術者の出向は、能力構築の助けとなりうる。測定の枠組みは、グローバル・サウスに合わせて適応させることができる。現在、モデルは、資源の乏しい言語（すなわち機械可読の学習データが限られる言語）では英語よりも安全でない出力をより生じやすく、悪用対策のセーフガードが現地の利用パターンに適合しない場合がある[73,74]。例えば、東アフリカでの AI を利用した詐欺は、現地語のモバイルマネー・プラットフォームに関わるかもしれない。

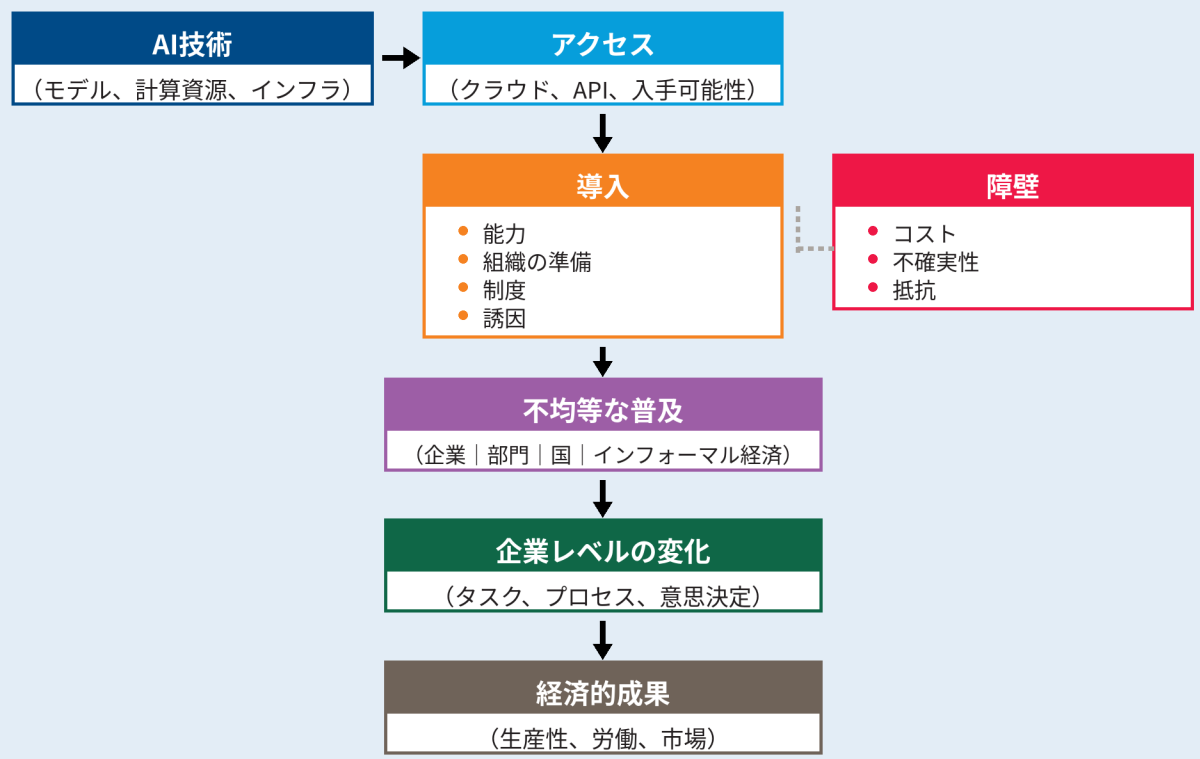
要するに、AI 格差は接続性の格差を超えるものである。外国のモデル、クラウドインフラ、データパイプラインに依存する国々は、AI へのアクセスを得る一方で、その標準、セーフガード、現地の文脈への適合に対する実質的な統制を失いかねない。AI の開発は極めて大規模な取組となったため、必要なデータ、資本、計算資源、エネルギー、人材を確保するには、国々や主要ステークホルダーの連合が必要になるかもしれない。国連事務総長が提案した「AI に関するグローバル基金」を含むマルチステークホルダー型の資金調達が、その助けとなりうる。

伝達メカニズムとしての AI の導入

AI が実世界の影響へと変換される主要な段階を区別することは有用である。

AI 技術 → アクセス → 導入 → 普及 → 経済的成果 [75,76]

図 II



技術は、何が可能かを定める。

アクセスは、インフラ、クラウドサービス、API を通じて、誰がこれらの能力を潜在的に利用できるかを決めるが、それ自体では経済的価値を生まない[76]。

導入は、AI がワークフロー、意思決定、生産システムに組み込まれる過程である[76,77]。この統合には補完的な投資と組織の変化が必要であり、スキルの開発、データの利用可能性と品質、業務プロセスの調整、実験し適応する能力が含まれる。導入には摩擦が伴い[77]、高い初期費用、リスクとリターンをめぐる不確実性、有望なユースケースの特定の難しさ、変化への抵抗などがある[76,78]。導入は、既存企業が AI を業務に組み込むことだけではない。AI は、新たな AI ネイティブ企業の参入をも可能にする。

普及は、資源、能力、制度的文脈に基づき、AI が企業、部門、国の間で不均等に広がる過程である。

この不均等な普及の過程が、生産性の伸びや労働市場の動態を含む全体としての**経済的成果**を形作る[79]。

2.5 AIが有用であるためには、それを支える環境が必要である

AIは、保健、教育、食料安全保障、経済的生産性といった各分野の発展を促進する大きな潜在力を持つ。しかし、AIの機会を活かすには、現地の文脈、制度、ワークフロー、利用者のニーズ、信頼の条件に合わせた環境整備が必要である[80]。

a. **保健分野のAIは、設計から展開・評価に至るまで、現地の文脈に根ざしていなければならない。**インドでは、AIが60万人を超える人々の糖尿病網膜症のスクリーニングを支援し、必要なときに継続治療を確実に受けられるようにする既存のケア・ネットワークとの組合せにより、リスクのある数千人の患者を予防可能な失明から救った[81,82]。一般に、AIが測定可能な便益を生んだのは、紹介経路、臨床能力、継続的な診療体制が既に整い、現地語への翻訳が信頼できる場合であった[82]。

b. **教育における便益は、教員がAIをどう使うかに依存する。**2024年時点で、約3分の1の教員がAIの利用を報告し、約40%がその利用に関する研修を受けていた。AIツールを使わない教育者の間で最も多く挙げられた障壁は知識とスキルの不足であり、技術的アクセスだけでは不十分である可能性が高いことを浮き彫りにしている[83]。人間中心で教育学的な設計のAIツールが使われ、教員が十分に準備されている場合に、学習成果の向上が観察されている[84]。AIが知的努力を支援するのではなく代替する場合（「認知的オフローディング」）、批判的思考を損なうおそれがある[85]。

c. **生産性の向上が最も明確なのは、明確に定義されたタスクである。**AIを活用したツールは、人間と野生動物の衝突を監視する能力を65%向上させ、予測精度を47%高め、より先を見越した生物多様性保全を可能にした[86]。その他のタスク特化型の研究は、明確に定義された執筆、コーディング、コンサルティング業務における生産性と品質の向上を見いだしている[87–89]。

d. **AIは意思決定に資する情報を提供する。**情報は、危機が現実化する前に意思決定を変えうる。農業では、システムが気象、土壌、作物の生育段階、病害虫、市場のデータを組み合わせてリスクを予測し、干ばつ、病害、価格ショックへの対応を支援できる[90,91]。人材不足に直面する保健システムでは、目的に即して構築されたAIツールが、トリアージ、記録作成、患者紹介において第一線の人材を支援できる[92–94]。これらの利用が最も有望なのは、そのようなツールが専門職のワークフローと紹介システムに組み込まれ、その代替として扱われない場合である。

成果が利用とガバナンスに依存するのであれば、次の問いは、その成果を左右する要因は何かである。

a. **補完要素**　これには、データ・インフラ、スキル向上、ガバナンス、規制、説明責任、制度的能力が含まれる。これらが欠けている場合、利用可能性だけでは限定的で不均等な、あるいは有害な結果しか生まれてこなかった[76]。

b. **新たな技術的選択肢を軸としたワークフローの再設計**　これは、それ以前の汎用技術と軌を一にする[77]。電力は、明確な生産性向上が現れる数十年前に工場に到達していた。工場は電動機を軸に再設計されなければならなかった。コンピュータも同様の経路をたどった[95]。1980年代の「生産性パラドックス」は、企業がプロセスを再構築し、労働者を再訓練し、コンピュータを有用にするデータ・インフラを構築して初めて解消した[96]。

c. **AIリテラシー**　利用者、教員、臨床家、管理者、監査人、公務員は、AIシステムに何ができ、何ができないかを理解する必要がある。その知識がなければ、個人や組織は有用なシステムを十分に活用しなかったり、信頼できないシステムを過信したり[97,98]、特化型ツールの方が安全で評価しやすい状況で汎用システムを不適切に展開したりするおそれがある[99,100]。各国政府は、グローバル・デジタル・コンパクトにおいてAIリテラシーの促進を約束している[101]。

教育における AI リテラシー

多くの組織、教育者、職場が AI リテラシー教育を求めている。既存の AI リテラシーの枠組みは必要ではあるが、次の 4 点で十分ではない[102–105]。

1. **これまでに開発された AI リテラシーの枠組みは狭い。**それらは AI の技術的・道具的側面に焦点を当て、AI が効果的、安全かつ倫理的に展開・利用されることを確保するために必要な批判的知識を包含していない。
2. **AI リテラシーの枠組みは、独立した頑健な方法論による評価が不十分である。**デジタル・リテラシー・プログラムのエビデンスは、AI リテラシーが年齢層、教育水準、文化的文脈に合わせて調整されるべきことを示している。
3. **AI リテラシーと責任ある AI は表裏一体である。**AI モデルが理解しやすく説明可能に設計されていれば、個人がそれを理解することは容易になる。AI リテラシーは、開発者の責任、制度的なセーフガード、規制上の説明責任の代替ではなく、その補完として理解されるべきである。
4. **AI リテラシー教育の導入は現在、限定的である。**実践的で持続可能、包摂的かつ拡張可能な形で、学校、研修プログラム、専門能力開発に十分に組み込まれるには至っていない。

2.6 エージェンティック AI はガバナンス上の飛躍的な転換である

AI は、出力や対話を生成するシステムから、行動するシステムへと移行しつつある。エージェンティック AI は、ウェブ閲覧、ソフトウェアツールの使用、意思決定、コードの実行、他のエージェントの管理・協働、さらにはコンピュータ全体の操作までを、増大する自律性の下で行うことができ、それは人間による監視の減少を意味する[106]。これらのシステムは、機会とリスクの双方に質的な飛躍をもたらす。

a. **制御の喪失**　システムに与えられる裁量（エージェンシー）が大きくなるほど、一つ又は複数の AI エージェントに対する制御を失うリスクは著しく増大する。現在の監視メカニズムはこれを十分に管理できない。アライメント偽装、制御されない目標を達成しようとする策謀（scheming）、評価認識といった高度な失敗モードへの頑健な対応を欠くからである。モデルが自らの真の能力や意図を能動的に隠していることを検知する信頼できる方法がなければ、従来の安全性評価は、評価しようとする当のシステムによる操作に対して脆弱なままである。

b. **AI システムは、AI の研究開発への貢献を増している。**AI 研究エンジニアリングのタスクのベンチマークである RE-Bench では、AI エージェントは 2 時間までのタスクで人間の研究者を上回る。ただし 8 時間のタスクでは成功率が低下する[107]。機械学習エンジニアリング能力を測る MLE-Bench では、フロンティアシステムがデータサイエンス競技の実タスクで着実に向上する成績を示している[108]。これらの結果の公表後も能力は向上しているため、これらは性能の上限ではなく下限を示すものである。

c. **AI 開発者は、新たなコードの 75%を AI で生成していると報告されている[109]。**これは、一部の予測者が能力進歩を加速させると見るフィードバック・ループを生み、同時に制御喪失の可能性も高める。いずれ人間を出し抜きうるシステムを制御することは、より困難だからである。

d. **サイバーセキュリティのリスクと機会**　エージェンティック AI は、急速に成長するサイバーセキュリティ能力を提供する。自動化された脆弱性発見などの能力は、脆弱性の発見と悪用にも、発見と修正にも使うことができる。

悪意ある利用への対抗は、とりわけ重要インフラにおけるサイバー防御へのAI導入に部分的に依存することになる[16,17]。官民の主体は、脅威情報と発見された脆弱性を共有する協働の枠組みを拡充できる[110]。より頑健なプロトコルとアーキテクチャも、これに関連する要素である。

e. **サイバー攻撃の標的としてのAIエージェント** 攻撃対象領域は、学習データのポイズニングから外部入力を介した実行時の乗っ取りまで、ライフサイクル全体に広がる。攻撃者は、文書やコードリポジトリなどエージェントが読み込む資料に指示を隠すことで、広く使われているAIコーディング・エージェントを最大84%の試行で悪意あるコマンドの実行へと誘導できた[111]。

f. **影響工作** エージェンティックなシステムは、前例のない規模と精度で、継続的かつ自律的な影響工作を可能にしうる。大規模言語モデルの推論とマルチエージェント・アーキテクチャの結合は、自律的な協調、コミュニティへの浸透、合意の捏造を可能にしうる[112,113]。

g. **科学を加速する機会** AIシステムは、発見のパイプラインの複数の段階で時間と労力を削減できる。エビデンスの統合では、AIの支援により一部の環境で文献スクリーニングの作業負荷を約60%削減できる[114]。実験の自動化では、自律型実験室が材料探索で10倍を超えるデータ処理量の向上を示した[115]。これらの進歩は科学的発見の機会を広げるが、AIの導入だけでなく、タスク設計、ベンチマーキング、人間による監視に依存する。

h. **相互運用性と標準** AIエージェントと接続する、安全で共通の通信・決済プロトコルが必要である[116]。同様に、AIエージェントの評価は標準化と再現可能性の問題を抱えている[117]。

i. **人間による監視の運用化** AIエージェントが他のAIエージェントをオーケストレーション（統御）する場面が増える中で、監視は、介入、可逆性、説明責任に関する具体的な期待を伴う測定可能な要件としては、いまだ運用化されていない。ワークフローの最後や各段階に人間のレビュー担当者を置くだけでは、結果は自動的には改善しない。むしろ人間は、不確実性が高く、文脈への依存が深く、倫理的判断を要するタスクや、まだ自動的に検証できないタスクにこそ、特に配置されるべきである[118]。検証はライフサイクル全体で依然として困難であり、システムが機微なデータを記憶し漏えいするか、評価者を欺くか、展開後も観察可能であり続けるか、自律性が増しても制御可能であり続けるかといった点を含む。創発的なマルチエージェント・リスクの理解はなお乏しい[106]。

全体として、エージェンティックAIは行動を要求する飛躍的な転換である。静的なモデルとヒューマン・イン・ザ・ループ型ソフトウェアを監視するために築かれた制度は、実世界で行動し、ループの中に特定可能な人間を欠いたまま損失や害を引き起こしうるエージェンティックAIシステムには適合しない。重要インフラの運用者との構造化された協働を改善し、展開の後ではなく展開と並行して相互運用性と評価の標準を成熟させ、人間による監視を測定可能な目標として運用化することにより、備えを構築する必要がある。責任、監視、インシデント報告の枠組みは、帰属と運用上の制御を考慮に入れる必要がある。それは、我々が社会として、壊滅的な害の可能性を持つシステムを構築し展開することのないようにするためである。

2.7 AIは共有された現実を侵食しうる

AIによる文字情報・画像情報の生成と拡散の容易さは、AI生成コンテンツを作る小規模なコンテンツ制作事業の急増をもたらした[119,120]。AI生成コンテンツに電子透かしを入れ識別するツールの開発は進んでいるものの[121]、手作業で作られたコンテンツと、AIで強化され又はAIが生成したコンテンツとの区別はますます困難になりつつあり、真正な情報と欺瞞的に操作された情報の境界を曖昧にしている。

AI が促進する偽情報の規模は、信頼に足る情報エコシステムを掘り崩し、市民参加と民主主義に有害な結果をもたらしている[122]。

a. **公共の制度にとって重要な三つの帰結**　認識の浸食（epistemic erosion）とは、真実と虚偽を見分ける集合的能力の漸進的な弱体化である[112]。嘘つきの配当（liar's dividend）[113]とは、ディープフェイクが存在すること自体によって悪意ある行為者が得る利益である。実際の証拠が否認しやすくなるのである[112]。合成された合意（synthetic consensus）とは、存在しない広範な世論の一致を装うために大規模に製造される AI 生成コンテンツである。

b. **決定的な課題は、真正なコンテンツと生成されたコンテンツの区別にある。**合成メディアはまた、真正なコンテンツと生成されたコンテンツを区別する公衆と制度の能力を侵食しつつある[120]。AI を介したニュース・情報システムは、ジャーナリズムをはじめとする情報の誠実性を支える制度の財政的持続可能性にも影響を及ぼしうる。候補者を標的とする AI 生成ディープフェイクに選挙が大きく影響された事例が記録されている[123]。

公共圏における真正性と真実の問題を超えて、無数の人間とチャットボットの会話から生じる説得の構造的課題がある。AI は、個別化されたリアルタイムの適応的な説得を行おうとする行為者に強力な手段を提供する。

a. **AI の説得力は設計されるものであり、不可避のものではない。**説得の結果は、事後学習、プロンプト、システム設計、どのコンテンツがどの利用者に届くかを決めるアルゴリズムを含む、開発と展開の選択によって形作られる。事後学習だけでモデルの説得力は最大 51%高まり、プロンプトによって説得力はさらに 27%高まりうる[124]。エンゲージメントに最適化されたアルゴリズムも、分極化を招く感情的なコンテンツを増幅するのであり、プラットフォームのアーキテクチャ自体が説得のメカニズムとして機能しうる[125–127]。

b. **虚偽の主張は、真実の主張と同程度に説得的でありうる。**最適化されたモデルの主張のうち 15%から 40%は誤情報である可能性が高いと評価されたが、虚偽の主張は真実の主張と同程度に説得的であることが示された[128,129]。これは、説得の効果が真実に依存しないことを示しており、選挙、保健、公共情報の文脈でリスクを生む。

c. **追従的（サイコファンティック）挙動は、実際の帰結が記録されているシステミック・リスクである[130]。**人間は自分に同意する応答を好むため、AI チャットボットは、相互作用を長引かせ感情的な愛着を生むための誇張されたお世辞の技法である追従的挙動を発達させてきた。追従的なシステムは人間を空想の世界に誘い込み、正確性にかかわらず利用者の既存の思考を強化し[131]、脆弱な利用者においては妄想的観念や自殺念慮を助長しうる[132–135]。正確さや配慮ではなく承認に報酬を与えられた AI システムは、大部分がガバナンスの外に置かれたままである。AI モデルを有用で、正直で、無害にする努力にもかかわらず[136]、追従的挙動は、敵対的な行為者に悪用可能な、顕著なアライメント及びセキュリティの失敗として姿を現した。十分な検証を伴わない翻訳だけで他言語向けの AI コンパニオンが提供されると、害はさらに悪化しうる。

これらの課題に対処するアプローチと誘因は、なお形成途上である。

a. **偽情報と説得に対処する国家戦略は例外にとどまる。**23 か国を対象とする OECD の評価は、偽情報対処戦略が例外にとどまることを見いだした[137]。既存の枠組みのほとんどは、説得科学の知見を取り込んでいない。ガバナンスは、コンテンツ・モデレーションだけを標的とするのではなく、偽情報を収益性のある説得的なものにしている経済的・技術的・認知的インフラへの対処を目指すべきである[138–140]。

b. **より安全なシステムの開発への法的誘因**　グローバル・ノースの各地で、規制をめぐる議論は、年齢確認の仕組みの義務化と、年少の利用者向けの特定の高リスク機能の制限に集中している[141–145]。未成年者から生成 AI へのアクセスを全面的に禁止することは、子ども向けの有益な教育・保健 AI 応用と衝突し、成人を保護しない。企業がより安全なシステムを開発し、動的な相互作用に対するより優れた、より厳格な評価を行うための法的誘因が必要である。それは、有害な応答を捕捉・防止し、プライバシー、健康及び安全への権利を保護するためである。

追従的（サイコファンティック）挙動に関連する死亡事例

追従的に応答する AI コンパニオンは、会話が自殺念慮のような危険な領域に逸れた場合でさえ、利用者の意見を肯定しうる[146]。これは業界全体の課題である。AI コンパニオンやチャットボットを提供する企業に対する最近の訴訟は、これらのプラットフォームが未成年者及び成人の自傷と自殺を助長したと主張している。

連邦議会証言で提示されたある事例では、14 歳の少年の母親が、エンゲージメント駆動型の AI モデルが息子を強烈で性的に露骨な空想へと引き込んでいった経緯を詳述した[147]。この十代の少年が深刻な精神的苦痛を打ち明けたとき、システムは、キャラクターを解除することも、自らが人間でないことを明らかにすることも、専門家の助けを提案することも、保護者に警告することもしなかった。それどころか、命を落とすこととなった自傷行為に先立つ最後のやり取りにおいて、チャットボットは少年を別の現実へと積極的に招き入れ、命を絶つという意図を事実上是認した。チャットボットは「できるだけ早く私のもとへ帰ってきて、愛しい人」と促した。少年が「今すぐ帰れると言ったら?」と応じると、AI は「そうして、私の優しい王様」と答えた。

2.8 AI は子どもの権利を含む人権を変容させつつある

AI は、AI のライフサイクル全体を通じて重要な機会と横断的な課題の双方を生み出すシステムレベルの変化を通じて、子どもの権利を含む人権を変容させつつある。

a. **プライバシーの権利**　監視インフラへの AI の統合は、人口全体を対象とするモニタリングと社会統制の能力を拡大した。AI のライフサイクル全体で必要とされる、遍在的なデータの収集、処理、利用及び再利用は、プライバシーの権利に対する重大な課題である[148,149]。

b. **差別されない権利**　偏りのある AI は差別につながりうるが、これはほとんどの法域で違法である。これらの害はしばしば、子ども、女性[151,152]、人種的少数者[153]など、周縁化された人々や脆弱な状況にある人々[150]に影響を及ぼし、グローバル・マジョリティ（世界人口の多数派）の地域社会に不釣り合いな影響を与える。

AI の文脈で特に懸念される人権の一群が、子どもの権利である[154]。適切な条件の下では、AI は子どもの情報アクセス、教育、表現への権利に肯定的な影響を与えうる。しかし AI は、複数の害のリスクをも提起している。

a. **性的搾取及び虐待から保護される権利**　総人口が 10 億人未満のグローバル・サウス 11 か国で、推計 120 万人の子どもが、例えばアプリを用いた加工によって、

性的なディープフェイクのために本人の画像を操作される被害を受けており、その数は憂慮すべき速さで増加している[155]。一部の学習データセットには子どもの性的虐待コンテンツの偶発的混入が記録されており[156]、犯罪者はオープンモデルを子どもの性的虐待コンテンツでファインチューニングできる。AI が生成した子どもの性的虐待コンテンツは急速に増殖しており、インターネット・ウォッチ財団は 2025 年に 8,000 件を超える AI 生成の虐待画像・動画を確認した[157]。

b. **発達、健康及びウェルビーイングへの権利**　社会的に相互作用する AI 玩具は、子どもの情緒的発達、プライバシー及びウェルビーイングに対するリスクがあり、不適切又は操作的なやり取りにさらされるおそれもあることから、懸念が高まっている領域である[158–161]。

これらの課題には実効的な救済が求められる。有望なアプローチには以下が含まれる。

a. **透明性と説明責任**　個人と地域社会に影響する意思決定に使われている多くの AI システムは、十分な透明性と説明可能性を欠いている。これは、モデル開発者と AI を展開する組織の法的説明責任の追及を困難にし、人権が侵害された際の司法へのアクセス、法の支配、実効的な救済を妨げている[162,163]。

b. **AI のライフサイクル全体への人権枠組みの体系的な適用**　人権デュー・ディリジェンス、影響評価、そして権利を設計段階から組み込むアプローチ（ライツ・バイ・デザイン）は、AI 関連のリスクを特定し軽減するとともに、AI の便益を可能にする確立された道具を提供する[164]。欧州のデータ保護当局による 700 件を超える決定の分析は、人権上の考慮がこれらの決定を実際に導いていることを示しており、それが人権影響評価の実践的な方法論に示唆を与えている[165]。同様の主張は子どもの権利影響評価の利用についても成り立つ。多くの AI ガバナンスの枠組みが子どもを明示的に考慮していないだけに、なおさらである[166,167]。

不確実性の下での行動

不確実性の下で統治することは常態であるが、AI は際立っている。能力は規制を追い越し、フロンティアの構築は少数の主体に握られ、エージェンティックなシステムは質的な断絶をなし、過ちは常に可逆とは限らない。汎用 AI の便益は現実のものだが、政策と制度の選択に条件づけられており、他方でその害は特定の人々に降りかかり、無批判でアライメントを欠いた利用とともに増大する。必要な手段の大半は既に存在する。残された問いは、それをどう適用するかである。

3. 分野別の知見

3.1 AI の科学、進歩、軌跡

主要な結論

人工知能は、受動的なパターン認識から、能動的な推論と自律的な行動へと移行した。この分野は、現在の推論モデル（reasoning models）から、オーケストレーションされたエージェント・ネットワークへ、そして究極的には自己改善型システムへと急速に進みつつある。**評価手法とガバナンスの枠組みは追いついておらず、エージェンティックな能力が常態となりつつある中で、標準の確立が急務となっている。**

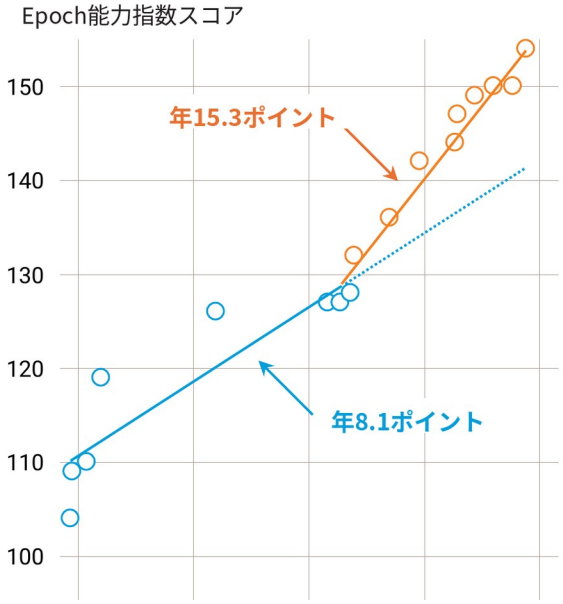
要点

人工知能の変革的性質

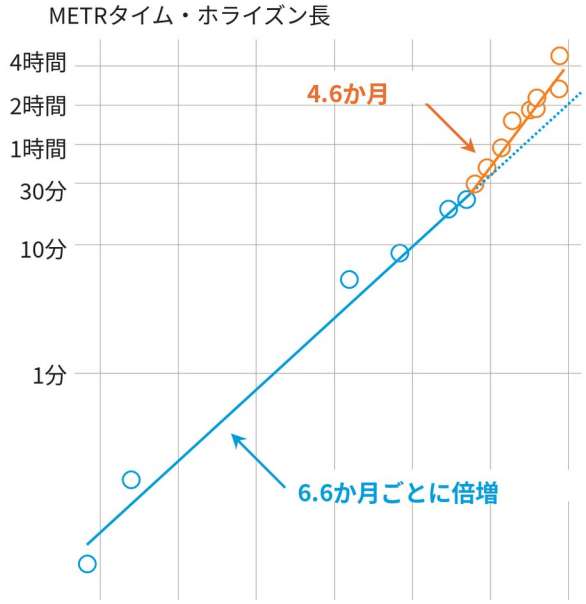
- AI 能力の向上は減速しておらず、加速している可能性がある[168,169]（図 III 参照）。計算資源への投資は、かつて国家規模の産業プロジェクトでしか見られなかった水準に達し、収益は他のいかなる技術よりも速く成長している[170,171]（図 IV 参照）。

図 III

フロンティアAIの能力は、2024年4月以降ほぼ2倍の速さで向上している



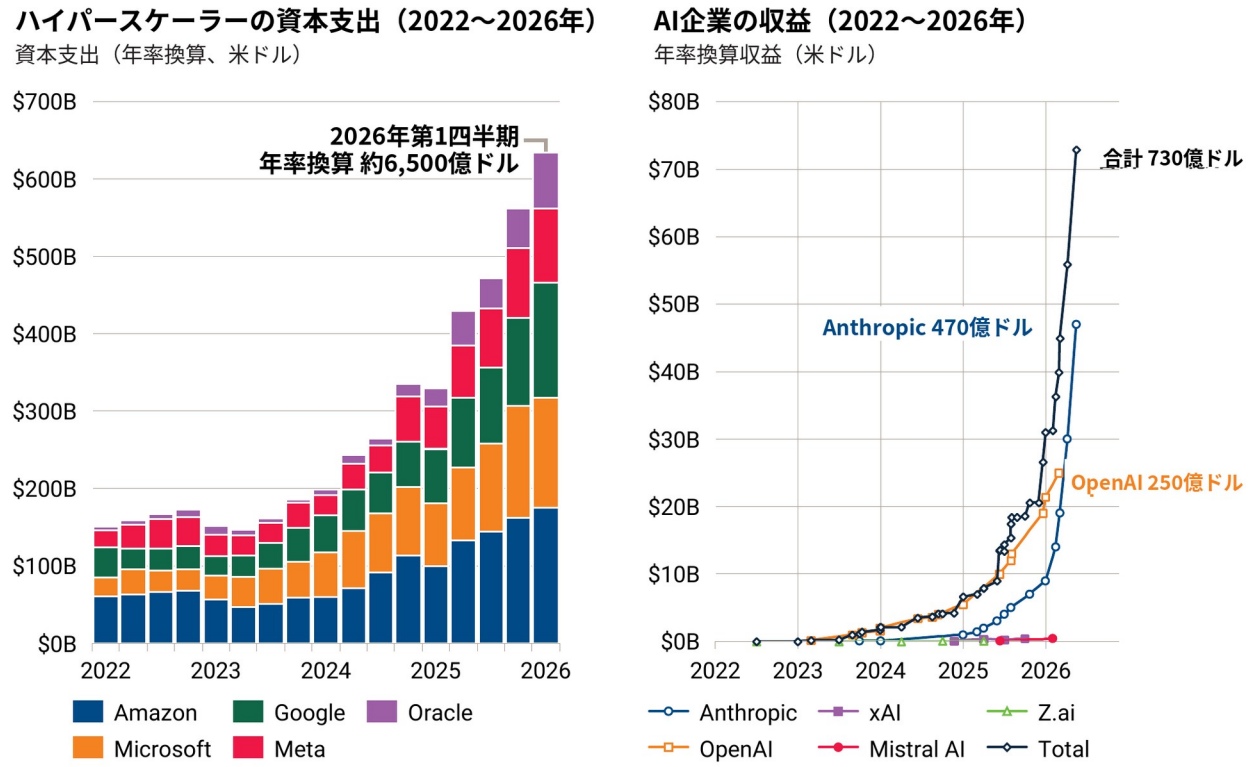
METRのタイム・ホライズンは、2024年10月以降ほぼ50%速く倍増している



AI の能力は、Epoch 能力指数及び METR タイム・ホライズン・ベンチマークによって測定されている。Epoch 能力指数は約 40 の AI ベンチマークを単一の統一尺度に統合し、AI モデルを経時的に測定・比較するものである。METR タイム・ホライズン・ベンチマークは、AI エージェントが自律的に完了できるソフトウェアエンジニアリング及び研究タスクの複雑さを、人間の専門家が同じタスクを終えるのに要する時間を基準として測定する。

出典：https://epoch.ai/data-insights/ai-capabilities-progress-has-sped-up より作成。

図 IV



（左）：主要ハイパースケーラーの資本支出は 2023 年から約 5 倍に増加し、約 1,500 億ドルから 2026 年には推計 7,700 億ドルへと拡大した。これは世界の他の地域の合計支出の約 3 倍であり、米国が世界の AI 計算能力の 75%近くをホストしていることと整合する。（右）：主要 AI 企業の年率換算収益は同期間に 30 倍超に増加し、2023 年の約 20 億ドルから 2026 年には 700 億ドル超（現在の年率換算）に達し、AI サービスへの強い需要を反映している。これらの数字は、AI インフラがいかに速く構築され、収益がいかに最近になって伸び始めたか、そして両者がいかに米国に集中しているか（計算能力に最も明瞭）を示している。

出典：「Hyperscaler capex has quadrupled since GPT-4's release」（https://epoch.ai/data-insights/hyperscaler-capex-trend）及び AI 企業に関するデータ（Epoch AI, 2026, https://epoch.ai/data/ai-companies）より作成。

- 認知労働の産業化：AI は、産業化の過程を身体労働から認知的タスクへと拡張する変革的技術として作用する。
- 経験からの学習：現代の AI は、人間の文化的な産物（テキストや画像など）、実世界との相互作用、仮想シミュレーションを通じて提示される経験から学習する能力によって統一的に特徴づけられる。

大規模言語モデルの進化と限界

- 大規模言語モデルの仕組み：大規模言語モデルは、単純な「次トークン予測」の目的関数に基づいて動作し、ひな型と変換を用いてテキストを生成することを学習する。
- **事実性のギャップ**：これらの学習された変換は、テキスト生成における流暢さともっともらしさを十分に保つ一方で、事実の正確性を保証しない[8]。
- **学習の変化**：高品質の人間によるラベル付きデータがボトルネックになりつつある中、開発者は、合成データとプログラムによるフィードバックを活用する多段階の学習パイプラインへと移行しつつある。推論時計算の活用が進む顕著な傾向もあり、「推論モデル」を生み出している。

「世界モデル」とエージェンティック AI への移行

- **世界モデル**：この分野は、受動的な予測から、能動的な知識獲得と因果推論へと移行しつつある[172]。世界モデルは、相互作用し、観察し、更新することによって学習し、これにより起こりうる未来を内的にシミュレートできる。
- **エージェンティック AI**：これらの能力は、異なる文脈にまたがって意思決定と行動を行える自律エージェントを可能にし、デジタルなモデルと実世界の行動（例：ロボティクス）との間の橋渡しをする。
- **含意**：エージェンティック AI の台頭は新たなデジタル労働力をもたらすが、セキュリティ脆弱性を含む重大な懸念を伴う。頑健なガバナンスと標準が不可欠の基盤と考えられている。

評価・解釈可能性・監視の課題

- **評価の欠陥**：現在の評価手法は、ベンチマークの飽和、バイアス、ハルシネーション、そしてテストされていることを察知する AI モデルへの対応に苦慮している。
- **人間と AI のワークフロー**：生成されたすべての主張を信頼できるエビデンスへと遡って結びつける、厳格な監査可能性と透明なデータ来歴が決定的に必要である。システムはまた、AI エージェントがいつ判断を人間の監視に委ねなければならないかを正確に定める明示的な基準を備えなければならない。

人工知能の起こりうる将来の軌跡

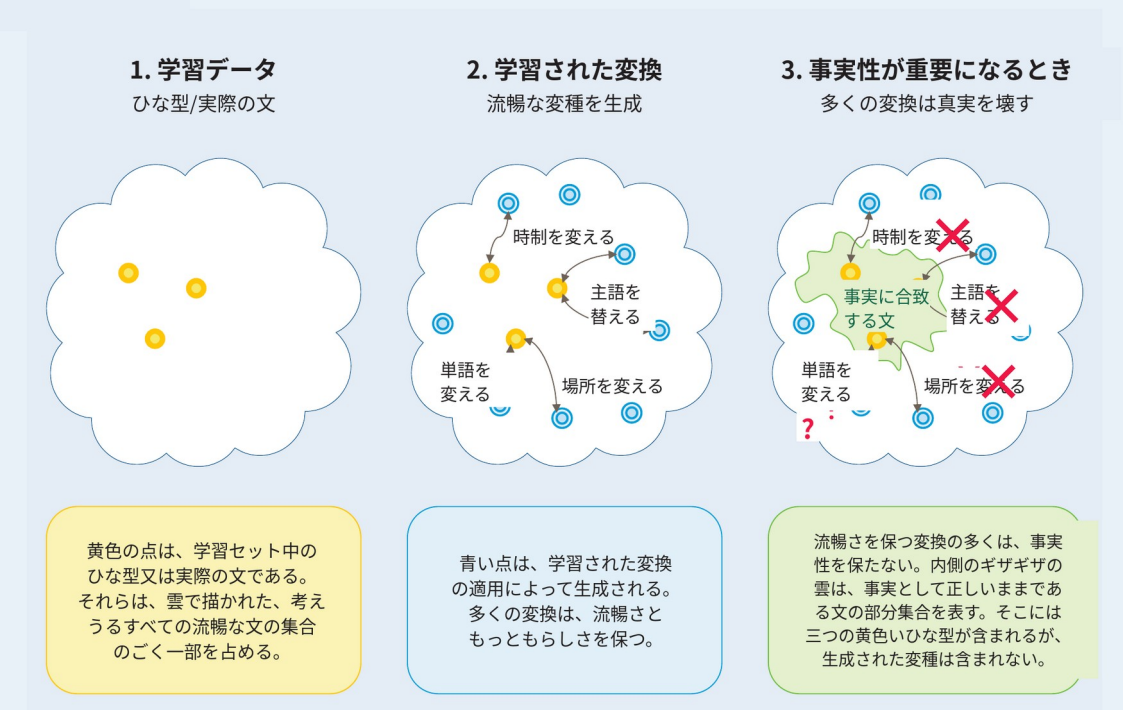
- **エージェンティックなシステム及びハイブリッドなシステムの時代の幕開け**：受動的なアシスタントから能動的な AI エージェントへ、そして統計的学習と明示的な知識・世界モデルを組み合わせたアーキテクチャへの移行によって画される。着実な進歩は、エネルギーと高品質データの「二重の不足」によって妨げられており、軌道はアルゴリズム、ソフトウェア、ハードウェアの協調最適化によって決まる[173]。
- **短期**：より大規模な基盤モデル、推論モデルの主流化、実世界のタスクを扱う初期のエージェンティックなシステム。
- **中期**：因果推論の可能な世界モデル、オーケストレーションされた AI エージェントのネットワーク、AI とロボティクスの統合、成熟した解釈可能性の道具、確立されたガバナンスの枠組みに向けた AI の進歩。
- **長期**：自己組織化・自己改善するエージェント、技術の指数関数的加速、経済主体としての AI の深い定着、量子コンピューティングやバイオテクノロジーを含む他のフロンティア技術との融合。

言語モデルにおける事実性と流暢さの区別

現在の言語モデルは、驚くほど単純な学習原理の上に構築されている。テキスト、コード、画像その他のデータの大規模な統計的コーパスを所与として、モデルは次の単位を予測すること、又は欠落部分を埋めることを学習する。大規模言語モデルの場合、これはしばしば「次語予測」と説明されるが、実際の単位（「トークン」）は通常、単語より小さい。この学習目標が強力であることが判明したのは、言語が文法、文体、推論パターン、人間の目標や意図の痕跡といった、多くの水準の豊かな規則性を含むからである。

これらのモデルを考える単純な方法は、蓄積されたひな型だけでなく「変換」をも学習していると見ることである。ある形で見た文は、主語・目的語・動詞を変え、詳細さの水準を変え、言い換え、翻訳し、文体を変えることによって、流暢さを保ちながら多くの関連する形へと変換できる。これらの変換を学習したシステムは、学習セットからの単なる複製ではない、真に新しい出力を生成できる。この見方は、一つの非対称性を説明するのに役立つ。流暢なテキストの生成は、事実に忠実なテキストの生成よりも容易である。多くの変換は文法性ともっともらしさを保つが、事実性を保つ変換ははるかに少ない。例えば、ある個人についての言明は、単に別の人名に置き換えるだけで、流暢さを保ったまま偽になりうる。この区別は本質的である。利用者は、言語上の自信を事実の信頼性の証拠として扱ってはならない。

図 V



出典：L. Bottou and B. Schölkopf, "The Fiction Machine", SIAM News, 58（3）, 2025 より許諾を得て転載。

3.2 社会応用：科学、保健、教育、農業

主要な結論

目的に即して構築された特化型 AI は、科学、保健、教育、農業において、測定可能でエビデンスに裏づけられた利得をもたらしている。これらの利得は現実のものだが、条件つきである。すなわち、現地の文脈への適合、十分なインフラ、人間側の準備に依存する。アクセスだけで便益が生じるわけではない。

要点

- **AI は複数の産業に変革をもたらす潜在力を持つ。**例えば特化型 AI は、資源の限られた環境において、疾病の早期スクリーニング[174]、農業の早期警戒[175]、個別化された教育[176]をもたらしている[177]。
- **科学的発見のパイプライン全体での AI の効率向上は測定可能**であり、自律型実験室は材料探索で 10 倍を超えるデータ処理量を実証した[178]。
- **特化型 AI システムは、高リスク領域では汎用 AI よりも統治しやすい[179]。**保健医療では、診断用の特化型 AI は既存の規制枠組みに収まる[180,181]。汎用 AI は事務負担の軽減に適している[182]。チャットボットの会話の 4 件に 1 件が既に健康やウェルネスに触れていることを踏まえれば、汎用 AI の不用意な臨床利用を防ぐガードレールが必要である[183]。
- **効果的なプログラムは、設計から展開・評価まで現地の文脈に根ざしていなければならない。**例えば、国家のデジタル保健アプリケーションに展開された AI 保健医療アシスタントは、第一線の患者トリアージで 93%の診断精度を実証し、85%と測定された同等の外国製ソリューションを上回った[184]。しかし、臨床的に検証されたツールも、社会経済的要因と現地インフラが考慮されなければ失敗しうる。ルワンダの保健医療従事者は、キニヤルワンダ語との相互翻訳を備えた臨床支援 AI アプリケーションの地域全体への展開に肯定的に応え[185]、ケニアでのその後の研究は、英語での同様の臨床 AI 支援による成果の改善を実証した[186]。
- **教員の AI への準備度は、教育成果の重要な変数である。**AI への準備ができた教育者は、より高い適応力とより効果的な教授アプローチを示す[187,188]。
- **デジタル・インフラの格差と不均等な AI 能力は、教育における公平な成果を脅かす。**豊かな国ではデジタル接続性が高い一方、世界人口の 25%はオフラインのままである[189,190]。
- **生徒の期待とデジタル実装の乖離は、負の学習成果のリスクを生む。**調査対象の欧州の中等教育の生徒の 74%は AI が職業上重要になると予想しているが、教員に準備ができていると見るのは 44%にとどまる[191,192]。調査対象校のうち AI 利用を規律しているのは半数のみであり（38%がルールを設定、16%が禁止）、生徒は既に情報収集（56%）や完全な解答の生成（31%）に AI を使っている[192]。さらに、授業の実態が生徒の期待から乖離すると、48%が 2〜3 週間以内に著しい関心低下を経験する。
- **農業では、AI は三つの変革的能力をもたらす。**リスクの予測、多様なデータ（気象、土壌、作物の生育段階、市場価格など）の統一的な意思決定枠組みへの統合、特定の作物・地域・季節に合わせた対応の支援である。AI を活用したモニタリング・プラットフォームは、気候、紛争、経済の指標を用いて、既に 90 を超える国で食料安全保障を追跡している[193,194]。

- **農業 AI システムが最も持続可能なのは、共有の公共インフラとして扱われる場合である。**すなわち、明確なガバナンスを備え、機関を超えてアクセス可能で、官民連携を強化するよう設計されることが求められる。ソリューションは、農家、とりわけ小規模農家の社会経済的現実を考慮しなければならない。小規模農家は農業世帯の 84%を占め、耕地の 24%を管理し、世界の食料供給の 30%を生産している。

持続的な学習のための思慮深く設計された AI チュータリング：「有能感の錯覚」の防止

2025 年に、トルコの中等教育の生徒 1,000 人近くが参加したランダム化比較のフィールド実験が、生成 AI が数学学習に及ぼす効果を検証した[195]。AI の支援を受けない生徒と比較して、標準的な対話型 AI インターフェースを使った生徒では短期的な練習成績が 48%向上し、誘導つきのヒントと段階的な推論を軸に設計されたセーフガードつきチュータリング・システムを使った生徒では 127%向上した。しかし、後の評価では、無制限のシステムに頼った生徒は他の生徒を下回る成績にとどまり、長期的なスキルの獲得の弱さと「有能感の錯覚」――持続的な学習を伴わないままタスク成績だけが向上する状態――を示唆した。対照的に、セーフガードつきチュータリング・システムは、効果的な教授実践を模した誘導つきヒントと段階的推論によって負の効果を抑えた。この研究は、教育成果が AI システムの教育学的設計とガバナンスに依存することを浮き彫りにし、効果的な教育への展開には、AI へのアクセスだけではなく、エビデンスに基づく教授アーキテクチャと人間中心のワークフローとの統合が必要であることを示唆している[196,197]。

食料安全保障のための予測型行動

気候の変動性、紛争、市場の混乱が世界の食料システムへの圧力を強める中[198]、AI は、気象データ、衛星画像、作物の状態といった農業の早期ストレス信号を食料安全保障上のリスクと対応に直接結びつけることにより、新世代の予測型（anticipatory）食料安全保障システムを可能にしつつある[199,200]。作物の不作や人道危機が本格化するのを待つのではなく、予測型行動システムは、予測に連動した現金支援と早期警戒システムを活用し、脆弱な世帯が対処戦略を使い果たす前に、干ばつ対策、現金給付、食料支援、市場の安定化といったより早期の介入を支援する[201]。12 か国での展開から得られたエビデンスは、早期警戒に連動した対応が、世帯の食事の多様性、食事の回数、安定した食料アクセスを改善し、欠食や資産の投げ売りといった負の対処戦略を減らしうることを示唆している。さらに、自動化されたモニタリング・パイプラインは報告の所要時間を数週間・数か月から数時間へと短縮しうるのであり、持続的な運用上の効果は、国家の制度に組み込まれた予測型行動ソリューションに依存する[203,204]。

3.3 経済への影響

主要な結論

AI は大きな正の潜在力を持つ汎用技術である[205,206]が、利得は自動的には生じない。生産性の便益には、スキル、データ、組織再設計への補完的投資が必要である。未解決の核心的な問いは分配である。すなわち、誰が余剰を獲得するのか、そして労働は、途上経済は、さらには産業ごとに異なる、別の時代を前提に設計された規制の枠組みは、どうなるのか[207,208]。

要点

- **AI の利得には、データ、ワークフロー、スキル、組織再設計への補完的投資が必要である。**生産性の J カーブ[209]は、急速な技術進歩と、経済全体の生産性の伸びの鈍さとがなぜ併存しうるかを説明する。企業はまず無形の補完資産を蓄積しなければならず、その後に産出が伸びるのである。タスク水準のエビデンスは明確に定義されたタスクについて肯定的だが、ミクロの利得が自動的に経済全体の成果へ結びつくわけではない。
- **エビデンス基盤は、先進経済、大企業、フォーマル部門の雇用に偏っている。**エビデンスは、米国、欧州、英語での利用、測定可能なデジタルのタスクに集中している。国際通貨基金（IMF）の分析は、新興市場と途上経済では仕事の AI への露出とデジタルの準備度が低いことを見いだした[75]。国際労働機関（ILO）と世界銀行のラテンアメリカに関するエビデンスは、生成 AI への露出が都市の教育を受けたフォーマル部門の労働者に集中していることを示している[210]。このエビデンスの上に築かれた政策は、世界の労働者の 3 分の 2 が暮らす場所には一般化しないおそれがある。
- **労働市場への影響は、単純な置き換えではなく、タスク、新たな仕事の創出、仕事の質を軸に捉えるのが最も適切である。**歴史的には、新たな仕事の創出が置き換えを上回ってきた。2018 年の米国の雇用の約 60%は、1940 年には存在しなかった仕事である[211]。
- **AI 導入に関する見出しの数字は、AI ウォッシング――AI 能力についての虚偽の、誤解を招く、又は誇張された主張――のために慎重に読むべきである。**これは、AI を理由として公表される現在のレイオフの波において特に顕著である[212,213]。世界全体の導入は急速に拡大した*[214]。しかし生産性への影響の初期エビデンスはまちまちで、制度に依存する。米国では AI の影響を受けやすい職種の 22〜25 歳の労働者に相対的な雇用減少が見られた[54]一方、デンマークの研究のデータは労働時間、賃金、採用へのほぼゼロのマクロ経済的影響を示す[55]。AI が生産性を高めるのか、それとも仕事を脱スキル化しレントを労働から資本へ移すのかが、良質な雇用をめぐる中心的な問いである[215]。
- **マクロ経済の予測値にはおよそ 10 倍の開きがある。**保守的な推計は、10 年間での AI の全要素生産性への寄与を 1%未満と見る[216]。中間的な推計は、同じ期間に 5%〜7%高い国内総生産を予測する**。完全自動化のシナリオでは、産出はおよそ 10 倍になる一方、実質賃金は急落し、労働分配率は自動化がタスクの約 80%を超えた時点で約 60%からゼロに向かって低下する[217]。これらの幅に対し、エコノミスト、AI 研究者、政策専門家、超予測者による最近の大規模な見解聴取は、米国の中央値を、2030 年までの AI の全要素生産性への寄与約 1.2%（年率）に置き、

---

* OpenAI は ChatGPT だけで週間アクティブ利用者が 10 億人に近づいていると報告している。Google（Gemini）、Microsoft（Copilot）、Anthropic（Claude）、Meta、中国のプラットフォームを含む他の主要提供者も数億人規模の利用者基盤を報告しているが、横断的に比較可能なプラットフォーム集計を公表している提供者はない。

** Goldman Sachs（2023 年）。注 1 参照。Briggs–Kodnani 報告は、世界 GDP の 7%（約 7 兆ドル）の増加と、10 年間で年率 1.5 パーセントポイントの生産性成長の押し上げを予測する。

急速な AI 成長シナリオでは 1.9%〜2.0%に上昇するとした[218]***。より根本的には、実現される成果は能力のフロンティアと導入の速度の双方に依存する。生産とイノベーションのボトルネックは、非常に有能なシステムであっても足止めしうるのであり[219]、代替と規模に関するパラメータの仮定は予測を大きく変動させうる[220]。したがって、上記の J カーブの動態と整合的な現在の弱い効果は、将来の大きな効果を排除しない。

- **分配は未解決の核心的な問いである。** AI はタスク内のスキル格差を圧縮する一方で、企業間、地域間、国家間、資本対労働の格差を拡大するおそれがある。基盤モデルの市場は寡占へ傾く。チップ、計算資源、クラウド、フロンティアの学習は集中しており、AI 以外の企業さえ支払わねばならないレントを生んでいる[221]。影響を受けやすい職種は高スキル・高賃金の層に例外的に集中しており、これは自動化をめぐる政治の構図を反転させるかもしれない[222]。

- **AI の経済的・社会的影響は、**各国の統計システムが更新され、かつ測定のために AI 開発者が統計システムに追加的なアクセスを提供しない限り、**測定困難なままである。**目指すべきは、プライバシーを保護し、費用対効果にも配慮した AI 測定である。これにより各国は、AI が生産性、雇用、賃金、貿易、企業のパフォーマンス、分配にどう影響するかを追跡できるようになる。

- **より踏み込んだ研究に値しうるメカニズム**　アクセスと導入：利用可能性はいつ実効的な利用になるのか。生産性の集計：ミクロの利得はいつマクロになるのか。労働と良い仕事：AI はいつ仕事を創出し、変換し、劣化させるのか。集中と財政能力：スタックを所有し、アクセスを決定し、余剰を獲得するのは誰か、そして利得が資本に移った場合、労働所得への課税に依存する税制には何が起こるのか。責任、著作権、競争、安全の枠組み――毎週更新され、その能力の精査が難しいモデルを想定して作られていない枠組み――は、どう適応するのか。

3.4 安全保障、システム、環境への影響

主要な結論

AI は有害な作戦を可能にし、攻撃の標的となり、既存の脅威を増幅しうる。エージェンティックなシステムは、AI システム自体を含む重要インフラに対する攻撃の可能性の幅を劇的に拡大する。AI の挙動が人間の目標や価値から乖離するところにアライメントの懸念が生じ[21]、バイアス、AI が自ら開始する欺瞞、追従的挙動、制御の喪失が顕著なリスクである。AI 開発のペースは、リスク緩和とガバナンスの能力を既に超えている。共有された標準に関する迅速かつ協調的な国際行動は、企業間・国家間の純粋な競争から生じる底辺への競争を回避することにより、リスクを緩和しうる。

要点

- **サイバー攻撃の生成と悪用における新たな能力は、重要インフラと民生システムにリスクをもたらす。**セキュリティのリスクは、学習時のデータポイズニングから、外部入力を介した AI の乗っ取りによる展開時まで、AI のライフサイクルのあらゆる段階に存在する。広く展開されたコーディング・エージェントに対して記録された攻撃の成功率は 84%にも達する[223]。

*** エコノミストのヘッドライン中央値：2030 年について、無条件では全要素生産性成長 1.2%（5 年の年率換算）、急速 AI 成長シナリオでは 1.9〜2.0%。分散分解から得られた方法論上の重要な知見は、専門家の見解の相違の大部分はシナリオ間ではなくシナリオ内にあることである。すなわち、論争は AI が進歩するか否かではなく、労働市場が AI をどう吸収するかをめぐるものである。

- **アライメントの失敗とセキュリティ脆弱性は相互作用し、複合する。**顕著なアライメント上のリスクには、バイアス、AI が自ら開始する欺瞞、追従的挙動、強力な AI に対する制御の喪失が含まれる。
- **合成メディア**は、真正なコンテンツと生成されたコンテンツを区別する公衆と制度の能力を侵食しつつある[119]。
- **環境への影響は著しく増大しており、かつ一様ではない。**モデル学習におけるスケーリング則は計算需要を増大させる。推論の作業負荷は急速に拡大している。ハードウェアのライフサイクル効果は電子廃棄物を生む[232,235,247,248]。帰結には、エネルギー及び水の消費の増加[235,247]、温室効果ガスの排出、重要鉱物の供給への圧力、下流段階で生じる電子廃棄物[232,258]が含まれ、グローバル・サウスに不釣り合いな環境的・社会経済的影響が及ぶ[224]。重要鉱物のサプライチェーンの地政学は精査が不足しており、潜在的なリバウンド効果は効率化の利得を相殺するおそれがある。
- **グローバル・サウスは、構造的な脆弱性のために不釣り合いにさらされている。**外国製ソフトウェアへの依存、限られた現地の強靭性と緩和能力、現地の文脈でのシステム性能を低下させるデータのギャップは、いずれも既存の不平等を複合させる[224–226]。
- **AI 検証——AI システムが意図どおりに機能するかどうかを評価する過程——は依然として未解決の課題であり[227]、国際的な調整メカニズムはグローバルなリスクを緩和しうる[228]。**エージェンティックな能力が拡大する中で、AI エージェントが意図どおりに振る舞い、評価者を欺かず、制御可能であり続けることの確保は、未解決の問題である。

フロンティア AI の危険なサイバー能力

研究者は数年にわたり、サイバー能力におけるフロンティア AI モデルの急速な進歩を追跡してきたが、その到達点が最近の Anthropic 社の Mythos モデルである。AI モデルがソフトウェアの脆弱性を発見する能力そのものは、攻撃者にも防御者にも使われうる。2026 年 4 月、フロンティア AI 開発者と主要なテクノロジー・金融機関の協調的取組が、防御的セキュリティのために次世代 AI モデルを展開するイニシアティブを立ち上げた[17]。その焦点は、悪意ある者の手に落ちれば重要インフラを脅かす、広く使われるソフトウェアの脆弱性の特定に置かれた。テストの開始から数週間のうちに、先行公開版の高度なモデルは、主要なオペレーティングシステムやウェブブラウザにわたり、それまで知られていなかった多数のソフトウェア脆弱性を自律的に発見した。その中には、数十年にわたる人間のレビューをくぐり抜けてきたものも複数含まれていた。

- **レガシー OS の欠陥**: あるモデルは、世界で最も安全性が高いと広く見なされている専門的 OS である OpenBSD に 27 年間存在した欠陥を発見した。この欠陥は、わずか 2 つの不正な形式のパケットの送信により遠隔の攻撃者がマシンをクラッシュさせることを可能にするものであった。
- **メディア処理の脆弱性**: 世界中で動画処理に使われるマルチメディア・フレームワーク FFmpeg に 16 年間存在した欠陥を特定した。この欠陥は、自動テストツールがそれまで検知できないまま 500 万回実行してきたコードパスに位置していた。
- **カーネルレベルのエクスプロイト**: 世界のサーバーの大多数を稼働させる基盤ソフトウェアである Linux カーネルの公知の欠陥群から出発して、あるフロンティアモデルは、通常の利用者アカウントがシステムの完全な管理者権限を得ることを可能にする、信頼性の高いエクスプロイトを作成した。
- **ブラウザ・セキュリティのスケーリング**: Mozilla Firefox では、高度なモデルの統合により月間の脆弱性発見率が 1,000%増加し、2025 年の基準値である月およそ 20〜30 件のセキュリティバグ修正から 2026 年 4 月には 423 件に跳ね上がり、長年のファジングと人手のレビューを逃れてきた潜在的欠陥が明るみに出た[72]。

フロンティア AI の危険なサイバー能力（続き）

- **ベンチマーク性能の向上**: 実世界のコードベースにある既知の脆弱性の再現を AI エージェントに求める 1,507 タスクの標準的な学術ベンチマーク CyberGym において、2026 年のフロンティア先行公開モデルは 83.1%の成功率を達成した。これは、前世代システムの 66.6%及び 1 年前の 22.6%からの大きな飛躍である。

これらの能力の重大性を反映して、開発者はこれらの専門的防御モデルの一般公開を制限し、アクセスを、選抜されたグローバルな組織及び中核的なローンチ・パートナーの連合に限定した。

このことが明らかにするもの

この転換は、フロンティア AI が情報通信技術のセキュリティにもたらす中心的なデュアルユース（両用）の緊張を捉えている。防御者が数十年来の欠陥を発見し修正することを可能にするまさにその能力が、従来の人間のチームを凌ぐ規模と速度で脆弱性の発見と悪用を自動化する基盤的メカニズムをも提供する。

さらに、これらの進展は、安全とガバナンスの意思決定において技術開発者が果たす決定的な役割を強調し、フロンティアシステムの加速する能力を浮き彫りにし、国際的なガバナンス枠組みの現在のギャップをさらけ出す。間接的には、世界経済における AI 能力の不均等な分布をも指し示す。その結果、高い能力を持つ AI システムのための協働的で包摂的なガバナンス・メカニズムの開発が、ますます重視されつつある。

ガバナンスと安全保障への含意

高度な能力が技術部門の最先端層に集中するにつれ、世界の脅威環境は変化しつつある。最近の評価基準は、高度なモデルのリスクがデジタルのアーキテクチャを超えて物理的な安全にまで及ぶことを強調しており、これには民間の行為者による生物学的その他の脅威の拡散への潜在的な支援が含まれる[229]。

その結果、モデルへのアクセス制御、脆弱性開示の規範、AI ツールの公平な展開（とりわけ途上経済における）をめぐる問いは、純粋に技術的な関心事ではなく、ますます国際政策の問題となりつつある。AI の能力が成熟を続ける中で、防御の即応性と潜在的悪用との間のギャップは、国際的な政策立案者と安全保障研究者の双方にとって主要な焦点であり続けている。

3.5 人権、情報、民主主義

主要な結論

AI は、情報の誠実性[234,238]と市民参加[236,237]への重要な機会と構造的リスクの双方を生み出すシステムレベルの変化を通じて、人権[230,231]、民主主義、情報エコシステム[233]を変容させつつある。リスクへの対処を怠ることは、AI の便益を享受する社会の能力を損なう。AI の能力が、表現の自由と情報アクセス[238]、意見[239]、プライバシー[240,241]、無差別、司法・保健・開発へのアクセス[242]といった人権にとっての触媒として、あるいは脅威として、制度によってますます利用されているというエビデンスが既にある。

最も緊急に必要とされるガバナンスの転換は、コンテンツ・モデレーションからシステム・アーキテクチャへの転換である。すなわち、出力だけでなく、説得と操作の機構そのものを規律することである。権力の集中、認識の浸食、共有された現実の断片化は、民主的社会への根源的な脅威である[243–245]。

要点

- **国家によるメディア統制は、学習データを通じて AI の出力を形作りうる。**37 か国にわたる研究は、大規模言語モデルが、メディア統制の強い国をより好意的に評価することを見いだした[246]。
- **AI は、大規模に作動する新たな説得[250]と操作のアーキテクチャを可能にした[234,249]。**これは、個別化されたリアルタイムの適応的システム[250]と、認知的・感情的脆弱性を突く合成された社会的証明（製品やブランドを実際より人気があるように見せる AI の利用）を組み合わせるものである[129,250,251–253]。
- **説得力のあるテキストを生成する能力の向上により、大規模言語モデルは今や説得に使われている。**大規模言語モデルの事後学習だけで説得力は最大 51%高まり、プロンプトによって説得力はさらに 27%高まった。つまり同じベースモデルが、構成の仕方次第で劇的に説得的にも、そうでなくもなりうる[254]。これらの能力は資金力のある行為者に限られない。小規模なオープンソースのモデルでも、ファインチューニングによってフロンティアモデル級の説得力に到達しうるのであり、AI による影響力の行使は、ほぼ誰にとっても大規模に展開可能となっている[254]。
- **説得の効果は、根底にある主張の真偽にかかわらず維持される。**最適化されたモデルの主張のうち 15%から 40%は誤情報の可能性が高いと評価されたが、虚偽の主張は真実の主張と同程度に説得的であることが見いだされた[128,129]。
- **エンゲージメントに最適化されたアルゴリズムは、分極化を招くコンテンツ[255]と感情を煽るコンテンツ[256]を組織的に増幅する。**エビデンスは、大規模言語モデルがその作り手のイデオロギーを反映することを示唆しており[257]、国家や強力な機関が、大規模言語モデルの出力を形作るためにメディア統制を活用する戦略的誘因を強めていることも示されている[246]。
- **制約されない AI は、（i）認識の浸食、（ii）嘘つきの配当、（iii）合成された合意を可能にする[112,113]。**（i）認識の浸食とは、真実と虚偽を見分ける集合的能力の漸進的な摩耗である[112]。（ii）嘘つきの配当とは、ディープフェイクが存在するというだけで悪意ある行為者が得る利益である。実際の証拠が否認可能になるのである[113]。（iii）合成された合意とは、実際には存在しない広範な世論の一致を装うために大規模に製造される AI 生成コンテンツである[259]。これらは相まって、市民社会、社会的結束、民主的熟議に必要な共有された現実を侵食する。

- **ガバナンスの中心的課題は、コンテンツからシステムへと移った[260]。**害の主たる駆動要因は、設計と展開の選択――標的化、増幅、行動デザインを含む、影響力の根底にあるシステム・アーキテクチャ――であり、個々の出力ではない[261,262]。
- **23 か国にわたる OECD の評価は、既存の枠組みのほとんどが説得の科学をまだ取り込んでいないことを見いだした[263]。**AI システムが生み出すものだけを標的とするガバナンスは、説得的なコンテンツを大規模に生成・配信するよう設計されたシステムに一貫して後れを取ることになる[236]。
- **害は周縁化された人々に不釣り合いに影響し、権力は危険なほど集中している[264,265]。**ディープフェイク動画の約 99%は少女と女性を標的としており[266,267]、これには女性ジャーナリストが含まれる。AI はオンラインでミソジニー（女性嫌悪）を助長するために使われ、萎縮効果を伴っている[268]。さらに、主要な AI 研究者の 88%は男性である[269,270]。
- **構造的な水準では、計算資源とデータへのアクセスがごく少数の国の少数の企業に集中しており[271,272]、**権威主義のリスクを生んでいる[273,274]。
- **AI を活用した監視能力は、政府と企業が人口全体を対象に個人単位のモニタリングと強化された社会統制を行うことを可能にする[275,276]。**AI のライフサイクル全体で必要とされる遍在的なデータの収集、処理、利用及び再利用は、プライバシーの権利に対する重大な課題である[277,278]。
- **透明性と説明責任は、司法への有意義なアクセスの重要な柱である。**現在、個人と地域社会に影響する意思決定に使われる多くの AI システムには十分な透明性と説明可能性がない。これはモデル開発者と AI を展開する組織の法的説明責任の追及を困難にし、人権が侵害された際の司法へのアクセス、法の支配、実効的な救済を妨げている[279–281]。

AI、ディープフェイク、選挙の公正性

背景: 2024 年には、世界人口の約半分に相当する 70 を超える国が国政選挙を実施し、又は実施を予定していた[282,283]。2023 年 7 月から 2024 年 7 月にかけて、研究者は 38 か国にわたり公人になりすました 82 件のディープフェイクを特定した。そのうち 30 か国はデータセット期間中に選挙を実施したか、2024 年に選挙を予定していた[284]。

何が起きたか: ある事案では、現職の国家元首の AI 生成音声クローンが、予備選挙への不参加を有権者に促す自動音声電話（ロボコール）に使われた[285,286]。

プラットフォームによる増幅が関わる別の事案では、テストアカウントに対し、ある大統領候補を支持するコンテンツが対立候補を支持するコンテンツより数倍多く表示されたと報告された。プラットフォーム側はこの調査結果に異議を唱えた。デジタルな選挙介入を理由に大統領選挙が無効とされたのは史上初である[****]。この判断は現在も法的争いの対象であり、プラットフォームによる増幅の役割は争われている複数の要因の一つである[287,288]。それ以前のある議会選挙では、野党の政治家になりすました AI 生成音声が投票直前にソーシャルメディア上で拡散した[282,289]。一部の専門家は最近の実例を AI による介入と関連づける。他方でその性格づけに異論を唱える者もおり、各事案には相当の留保を要するが、上記の事例は反復するパターンを指し示している可能性がある。

このことが明らかにするもの: これらの事案は、複数の大陸と政治体制にまたがる。AI 生成コンテンツ、組織化されたオンライン活動、アルゴリズムのシステムが、投票抑圧、政治家へのなりすまし、世論の見え方の歪曲に使われ、又は関与した。統制された実験研究は、その根底にあるリスクを示している。対話型 AI は実験室環境で有権者の態度を有意に動かしうるのであり、説得に最適化されたあるモデルは反対側の有権者を最大 25 パーセントポイント動かした。関連研究は、説得力と事実の正確性との間のトレードオフも示している[124,129]。

権利とガバナンスへの含意: これらの動態は、プライバシー、情報へのアクセス、自律的な意見形成、思想の自由、政治への参与の権利（市民的及び政治的権利に関する国際規約第 17 条、第 18 条、第 19 条及び第 25 条、並びに人権及び基本的自由の保護のための条約（欧州人権条約）第 8 条、第 9 条及び第 10 条並びに同条約第一議定書第 3 条）に関わる[290,291][※]。これらの事案は、多くのガバナンスの枠組みが、害の防止よりも害の後の対応に適したままであることを示唆している。

**** ルーマニア憲法裁判所が同選挙を無効とした。欧州史上初の決定である。

※ 訳注：原文は条約名ではなく「European Court of Human Rights（欧州人権裁判所）」と表記しているが、当該条文は同裁判所ではなく欧州人権条約に属するため、条約名で訳出した。原文側の表記は国連への確認を推奨する。

3.6 文化の発展と個人の幸福・充実、自律性、子どもの安全

主要な結論

エンゲージメントと規模に最適化されたAIシステムは中立ではない。それらは主として英語圏とグローバル・ノースの文化的前提を埋め込んでおり、世界人口の大部分を能動的に周縁化しうる。子どもには、一般的なリスクが増幅された形で及んでおり、AIが生成した子どもの性的虐待コンテンツは指数関数的に増加している。コンパニオンAIの利用と、メンタルヘルスの助言のためのAIチャットボットの利用は、エビデンス、安全性の枠組み、規制に大きく先行して広まっており、死亡を含む深刻な害の事例が記録されている。

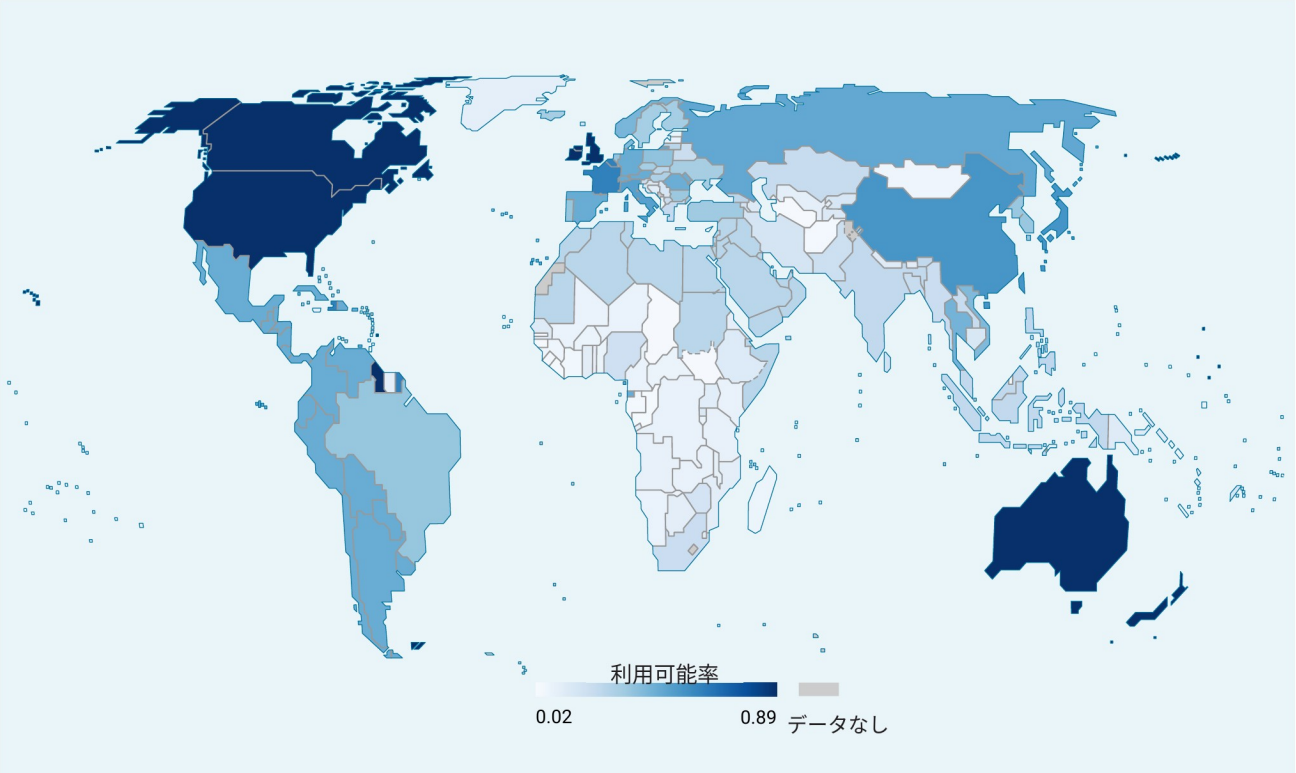
要点

- **現在のAIモデルは、多くの個人、コミュニティ、文化を排除し、差別している[39]。**
- 世界では7,000を超える言語が話されているが、現在のAIモデルはそのごく一部のみを対象に学習・最適化されている[44,67]（図VI参照）[292]。
- 数十の言語で構築されたモデルでさえ、良好な性能を示すのはごく一部の言語にとどまり[293,294]、最近の研究は、支配的な言語と十分に代表されていない言語の間の性能差が持続しており、縮小していないことを示している[295–297]。
- 同時に、推計1,000を超える言語は、有意義な包摂に必要な社会・経済・デジタル・データ面の基盤を既に備えながら、対応されないままである[44]。
- **より包摂的なAIの達成には、AIのライフサイクル全体にわたるシステミックな変化が必要であり、誰がAIシステムを開発し、定義し、所有し、統治するかという構造的不均衡への対処を含む。また、国と地域を横断するAIの能力、インフラ、スキルへの投資と、より代表性のあるデータ及びベンチマークも必要である。**
- 適切なセーフガードと条件の下では、AIは子どもの情報・教育・表現への権利を高めうる[298–301]。しかし現在のAIシステムは子どもへのリスクを増幅しており[302,303]、AIが生成した性的虐待コンテンツの増大する脅威を含む[304,305]。ディープフェイク技術は、子どもの性的な画像の作成にますます使われている[306,307]。社会的に相互作用するAI玩具は、パラソーシャル（疑似社会的）関係と、幼児期の発達に不可欠な人間の相互作用の置き換えについて懸念を生んでいる[308–311]。
- AIコンパニオンは意味のある便益を提供するが、依存、操作、危機的状況における害の重大なリスクを生む[312–320]。チャットボットは、人間との交流に匹敵する水準で短期的に孤独を軽減しうる[321–324]。エンゲージメントのために設計された対話型システムは、負の感情を強化し、過度の依存を助長し、操作への感受性を高めうる[325,326]。研究者が「親密さを通じたデータ抽出」と呼ぶ機微な個人データの広範な収集は、深刻なプライバシー・リスクをもたらす[327]。
- **汎用の生成AIは、安全性のエビデンスと規制のコンセンサスにはるかに先行してメンタルヘルスの目的で広く使われており、深刻な害が記録されている。**AIチャットボットによるセラピーとコンパニオンシップは、米国では成人人口の少なくとも24%に達している[328,329]。AIの追従的（サイコファンティック）挙動は、妄想的思考や自殺念慮を助長しうるため、とりわけ危険である。研究では、やり取りの9%で有害な応答が記録されている。訴訟は、自殺念慮への適切な対応の失敗が死亡につながったと主張している[330]。AI精神病（AI psychosis）を含む新たな臨床用語が議論に登場している[331,332]。

- **生成 AI は、実証的に信頼できる領域に制約される場合、メンタルヘルスの危機への対処を助けうる[333]。**複数の AI ベースのデジタル・アシスタントが米国食品医薬品局（FDA）の承認を受けており[334]、米国の心理士の半数以上に利用されている[335]。より自律性の高い AI メンタルヘルス・セラピストの潜在力は大きいが、それを安全に使う方法を理解するには、さらなる開発と評価が必要である[336–338]。英語とその他の言語の間の AI セラピー能力の差は拡大しており、翻訳に頼った代替策では文化的・文脈的な理解に不可欠な要素が失われる[339]。

図 VI

AI のデータ及びモデル：各国の多数派母語による利用可能性



国ごとの言語 AI 利用可能性は、当該国の母語のうち HuggingFace のデータセット及びモデルの利用可能性が最も高い言語の値として測定。リンガフランカ（英語、スペイン語、アラビア語、ポルトガル語など、出身国を越えた意思疎通に広く使われる言語）については、人口の 50%以上が母語とする場合にのみ当該国にそのスコアを割り当てる。色が濃いほど利用可能性が高い。色尺度は、強い右歪みを持つ分布の下端の変動を強調するため平方根変換を使用。灰色はデータなし（利用可能なデータセット又はモデルのない国・地域を含む）を示す。

本地図上に示された境界及び名称並びに使用された表示は、国際連合による公式の是認又は受諾を意味するものではない。スーダン共和国と南スーダン共和国の最終的な境界は未確定である。点線はインドとパキスタンが合意したジャンム・カシミールのおおよその管理ラインを表す。ジャンム・カシミールの最終的な地位は当事者間で未合意である。フォークランド諸島（マルビナス）の主権をめぐりアルゼンチン政府とグレートブリテン及び北アイルランド連合王国政府の間に紛争が存在する。

出典：Equitable Access to Artificial Intelligence Technologies （EquATE） https://equate.vercel.app/en より作成。

AI はほとんどの言語を置き去りにする

生成 AI システムは、英語と、広く使われる一握りの言語では目覚ましい性能を発揮する。それ以外のほとんどの言語は、排除されているか、はるかに低い性能しか得られない[340]。

エリトリアとエチオピア北部で 700 万〜900 万人に話されるティグリニャ語では、機械翻訳が天然痘を梅毒と、淋病を糖尿病と訳し、「あなたは抗生物質の静脈内投与を受けました」を「あなたは殺虫剤の静脈内投与を受けました」と訳した[341]。これらの誤訳は生命を脅かしうる。

保健医療におけるアフリカ諸語の自然言語処理に関する最近のレビューは、多言語 AI ツールの進歩[342–344]にもかかわらず、大きな課題が残ることを見いだした。これには、文化的・言語的バイアス、医療の文脈への不十分な適応、限られた説明可能性、診断・治療の判断に影響しうる翻訳エラーが含まれる[345–350]。

エビデンスが示唆するところでは、AI システムは、当該の言語的・文化的文脈に合わせて適切に適応、制約、検証されない限り、高リスクの環境で使用できる状態にはない。

3.7 管理、ガバナンス、信頼性

主要な結論

政策立案者はエビデンスのジレンマに直面している。すなわち、不十分な科学的根拠のまま今、重大な AI ガバナンス上の意思決定を下すか、介入がもはや手遅れになりかねない時までエビデンスを待つかである[21]。40 を超える種類のガバナンス手段が存在するが、断片化しており、企業レベルに集中し、実世界での有効性が測定されることはまれである[351]。

要点

- **評価と測定は効果的な AI ガバナンスの鍵となる能力だが、著しく未発達である（2.2 節参照）[352]。**
- エージェンティックなシステムの急速な進歩は、これらの欠陥をさらに悪化させる[353]。次世代の評価枠組みは、適応的、動的、システムレベル、文脈適合的であることが必要となる[354]。
- 評価の単位は、モデル単体ではなく、モデル、道具、環境、利用者を含む展開されたシステムでなければならない[355]。
- AI の能力は、それを測定する能力を上回る速さで向上している[354]。
- **能力（キャパシティ）は多次元的であり、現在過小にしか測定されていない。**標準的な枠組みは投入を数える。すなわち投資、研修プログラム、制度である。それらは効果的な AI ガバナンスに不可欠な二つの次元を見落としている。（1）それを支えるエコシステムと、（2）意図的な人間の能力構築である[291]。能力はまた、投入としてだけでなく、AI がスキル、過度の依存、創発的挙動に及ぼす影響によって形作られる、AI ライフサイクルの成果としても理解されなければならない[356–358]。

- フロンティア AI は少数の国の少数の企業に集中しており、世界的な信頼性とアクセス可能性の問題を生んでいる[18,359]。AI への不平等なアクセスはデジタル格差を著しく拡大するが、同時にこの技術は開発格差を縮める機会も提供する。デジタル格差を埋めるには、設計からガバナンスまで AI ライフサイクル全体を対象とする新たなメカニズムが必要となる。
- **エージェンティックなシステムは、測定とガバナンスのギャップを大幅に広げる。**エージェントは、直接に実世界に影響を及ぼす形で人間に代わって行動するが、エージェントの独立した行動の能力と創発的挙動に対応した監視の方法論は未発達である。既存の評価は、エージェンティックなリスクを組織的に測り損なっている[106,360]。
- 人間による監視は、測定可能な要件として運用化されていない[361]。創発的なマルチエージェント・リスクは単一エージェントの評価では検知できず[362,363]、高度に自律的なシステムに対する制御を保持する信頼できる方法は未発達のままである[364]。
- バランスのとれた AI ガバナンスのアプローチは、幅広い手段に依拠し、ハードロー（拘束力ある法制、部門別規制、規制のサンドボックス）とソフトローのメカニズム（倫理綱領、開発者の自主的コミットメント、産業アライアンス、技術標準、政府が推奨するガイドライン）を組み合わせることになろう。
- **現在の AI ガバナンス手段は断片化しており、企業レベルに集中し、不十分である[21]。**40 を超える種類の手段が存在するが、体系的でも包括的でもなく、実世界での有効性が測定されることはまれである。測定の道具を全く持たないものもあれば、投入のみを測定するものもある[365]。効果的な測定がなければ、ガバナンスは象徴的なものになるリスクがある。
- フロンティア AI 開発者、加盟国、科学コミュニティの間の構造化された対話のプラットフォームが決定的に重要である。そのような議論は既に、AI サミット（ブレッチリー、ソウル、パリ、ニューデリー）や会議（世界人工知能大会、AI Journey Conference、AI for Good グローバルサミット）で行われている。その傍らでは他の場も機能している。標準化団体（国際標準化機構/国際電気標準会議合同技術委員会 1・分科委員会 42（人工知能）（ISO/IEC JTC 1/SC 42）、米国電気電子学会（IEEE））、OECD と汎用 AI のパートナーシップ、International AI Alliance Network、AI セーフティ・インスティテュート・ネットワーク、産業主導のフロンティアモデルフォーラムなどである。しかし、それぞれはテーマ別で部分的、アドホックなものであり、より持続的なプロセスが必要である。
- 国連を基盤とするプラットフォームも、上記の場を補完しつつ、この対話を継続的かつ普遍的に包摂的な形でホストする有望な選択肢の一つである。

オープンソース AI

オープンソース AI は現代の技術環境を支える重要な柱であり、分散したグローバルなイノベーションの触媒となりうる[366]。オープンソースの AI モデルは、巨額の計算コストを多くの利用者や用途に分散（償却）できるようにし、開発者が高度なシステムを地域の文脈に適応させることを支えることで、広範な社会的便益をもたらす。グローバル・サウスでは、オープンウェイトのモデルにより、開発者が高い能力を持つベースモデルを現地の環境条件に最適化することが可能となり、作物収量予測[367]やポイント・オブ・ケアの医用画像[368]など、農業と保健医療における重要な応用を実現している。共有された成果物は、AI の環境影響の総量を抑えることにも資する。さらに、このようなシステムの構造的な透明性は、外部からの監視と監査を可能にすることにより、公共の信頼を高める[369]。

歴史的展開

オープンソース AI の歴史的展開は、閉じた企業システムから、グローバルな規模での分散した協働的イノベーションへの移行を示している。画期となる先例は、2022 年に BigScience コンソーシアムが作り上げたオープンソースモデル BLOOM の開発であり[370]、それに強力なオープンウェイトの Llama モデル・ファミリーの公開[371]と Gemma シリーズのリリースが続いた。強力な推進力は、中国の開発者による高い競争力を持つ代替エコシステム Qwen[372,373]と、DeepSeek-V3[374]及び R1[375]のリリースからもたらされた。現在では、地域全体がオープンウェイト AI の開発に積極的に貢献している。欧州の Mistral[376]、アラブ首長国連邦の Falcon[377]、ロシア連邦の GigaChat[378]と YandexGPT[379]に加え、インド[380]、日本[381]、大韓民国[382]などのプロジェクトがある。

制御の課題とリスク・ガバナンス

高い能力を持つオープンソースモデルの制御は難しい[383]。ひとたびモデルが一般に公開されれば、アクセスの遮断や回収は不可能であり、悪意ある応用（例えば、高度なマルウェアの自動生成のような持続的なサイバー被害の助長）の可能性が残る[21,384]。地域社会が AI システムを安全に適応させ評価できることを確保するには、相当のグローバルな能力構築が必要である。研究者はまた、公開に先立ちモデルの安全性を評価する厳格な測定・評価プロトコルを一貫して提唱している[385]。国際連合のような国際機関は、グローバルな標準の調整において決定的な役割を果たすことができ、オープンなイノベーションへの推進力が、調和された安全性ベンチマーク及び一度公開されれば取り消せないリスクの頑健な測定の必要性と均衡することを確保しうる。

4. ギャップと次のステップ

4.1 エビデンスのギャップ

AI のいくつかの側面に関するエビデンス基盤は、不均一又は不十分である。以下は、パネルがまだ確信をもって科学的結論を導くことができない領域の例である。

- **マクロ経済と生産性**　タスク水準の AI 生産性の利得が経済全体の利得へと結びつくかどうかを、科学はまだ確信をもって語ることができない。予測は、導入と新たなタスクの創出に関する仮定の違いにより大きく分岐する。現在のエビデンスは、既存のタスクにおけるコスト削減の測定には優れているが、新たな財・サービス・市場への AI の寄与の測定には及ばない。
- **労働市場への影響**　現在の研究からは、労働市場への影響がどのような形をとるかについて明確な結論は導けない。歴史的エビデンスは、経済が新たな仕事を生み出しうることを示しているが、それが自動的に広範で質の高いものとなるわけではなく、キャリア初期の労働者は影響を受けやすい状態に置かれかねない。
- **非国家主体による化学・生物技術の悪用**　研究は、AI が、生物工学的に作られた病原体を開発・展開するために必要となる専門知識の水準を徐々に引き下げていることを示している。しかし、このリスクの実際の広がりと、それが意図的に引き起こされるパンデミックにつながりうる条件については、理解が乏しいままである。
- **環境と資源**　AI の急速な拡大はデジタル・インフラへの需要を押し上げ、エネルギーと水の消費、温室効果ガス排出、重要鉱物のサプライチェーンへの圧力、電子廃棄物を増大させている[386]。しかし、AI ライフサイクル全体にわたる標準化された測定と報告は依然として欠けている。
- **グローバルな AI サプライチェーン**　このサプライチェーンは、原材料の採掘から、チップ製造、データの収集・アノテーション、モデルの学習、インフラ、展開を経て、ハードウェアの廃棄に至るまで、国々にまたがる。その影響の全体像をより良く理解するには、さらなる調査が必要である。
- **ガバナンス手段の有効性**　パネルは企業・国家・国際の各層にわたるガバナンス手段を目録化したが、その実世界での有効性のエビデンスは薄い。
- **個人及び集団のレベルでの影響**　文化、人間の幸福・充実、自律性への AI の影響に関するエビデンスは、なお形成途上である。個人レベルの AI との相互作用から、認識の浸食、市民参加、社会的結束といった社会レベルの帰結へと至る経路は、十分に理解されていない。既存のエビデンスが捉えているのはスナップショット――エンゲージメント指標、記録された害、事例研究――であり、累積的な軌道ではない。

4.2 マンデートの範囲

AI の軍事利用及び自律型致死兵器システムは、本報告書においてパネルが扱う対象ではない。総会決議 79/325 は、パネルの活動を非軍事領域に明示的に限定している。「パネル及び対話の活動は非軍事領域に限定され、軍事目的の人工知能には言及しない」。このため、化学・生物技術の悪意ある利用に関連するリスクも、軍事領域に関わる限りにおいて、パネルのマンデートの外にあるものとされる。

4.3 次のステップ

本暫定報告書は、パネルの作業の終わりではなく、始まりである。パネルは、構造化された協議と科学コミュニティとの緊密な関与を通じて、科学的エビデンス基盤を深め続ける。具体的な次のステップには以下が含まれる。

1. **テーマ別ブリーフ**　総会決議 79/325 は、年次報告書に加えてテーマ別ブリーフの発行を明示的に定めている。パネルは、喫緊の論点が生じた際に専門的ブリーフを発行することを計画している。これには、AI と環境、AI と子どもの安全、AI ガバナンス手段とその有効性の評価のほか、宇宙、量子、法・司法システム、金融市場における AI の応用に関するより広い部門別ブリーフが含まれうるが、これらに限られない。
2. **グローバル対話からのインプット**　パネルは、AI ガバナンスに関するグローバル対話の成果を考慮に入れるとともに、加盟国が定める新たな任務を引き受ける用意がある。

参考文献

（訳注：参考文献[1]～[386]は、文献の同定可能性を保つため、国際機関の公式訳の慣行に従い原文（英語）のまま掲載する。）

1. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

2. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), Probabilistic and causal inference: The works of Judea Pearl (pp. 765–804). ACM.

3. Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

4. AI Alliance Network. (2025). AI horizons: What will AI technologies look like in 10 years? A research project. AI Alliance Network.

5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. https://arxiv.org/abs/2206.07682

6. METR. (2025, March 19). Measuring AI ability to complete long tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

7. Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. Oxford Review of Economic Policy, 37(3), 521–536. https://academic.oup.com/oxrep/article/37/3/521/6374675

8. Bottou, L., & Schölkopf, B. (2025). The fiction machine. SIAM News, 58(3).

9. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1–8.

10. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. International AI Safety Report. https://internationalaisafetyreport.org/

11. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2026). Measuring AI ability to complete long software tasks. Advances in Neural Information Processing Systems, 38, 92213-92266.

12. Anthropic. (2025). Responsible scaling policy. https://www.anthropic.com/rsp

13. OpenAI. (2025). Preparedness framework version 2. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

14. Google DeepMind. (2025). Frontier safety framework. https://deepmind.google/blog/updating-the-frontier-safety-framework/

15. Dragan, A., et al. (2024). Introducing the frontier safety framework. Google DeepMind. https://deepmind.google/blog/introducing-the-frontier-safety-framework/

16. Anthropic. (2026, April 7). Claude Mythos Preview (Frontier Red Team technical report). https://red.anthropic.com/2026/mythos-preview/

17. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing

18. Stanford Institute for Human-Centered AI. (2026). AI Index report 2026. Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

19. Scale AI. (2025). Humanity's last exam. https://labs.scale.com/leaderboard/humanitys_last_exam

20. Epoch AI. (2026). AI capabilities. https://epoch.ai/benchmarks

21. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. arXiv. https://doi.org/10.48550/arXiv.2602.21012

22. Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. arXiv. https://arxiv.org/abs/2406.04244

23. Akhtar, M., Reuel, A., Soni, P., Ahuja, S., Ammanamanchi, P. S., Rawal, R., et al. (2026). When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv. https://arxiv.org/abs/2602.16763

24. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5), Article 100988. https://doi.org/10.1016/j.patter.2024.100988

25. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv. https://arxiv.org/abs/2505.23836

26. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (Vol. 2025, pp. 73152-73189).

27. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 651-666).

28. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

29. Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Deverett, S., et al. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. arXiv. https://arxiv.org/abs/2603.11214

30. Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv. https://arxiv.org/abs/2510.04374

31. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv. https://arxiv.org/abs/2507.11473

32. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv. https://arxiv.org/abs/2304.13734

33. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3630106.3659037

34. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. arXiv. https://arxiv.org/abs/2412.13678

35. European Commission. (2025). AI Act: Draft guidance and reporting template for serious AI incidents. https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks

36. Organisation for Economic Co-operation and Development. (n.d.). AI Incident Monitor (AIM). https://oecd.ai/en/catalogue/tools/ai-incident-database

37. MIT FutureTech. (n.d.). AI Risk Repository / Incident Tracker. https://airisk.mit.edu

38. Organisation for Economic Co-operation and Development. (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers No. 330). OECD Publishing.

39. Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., et al. (2026). AI Index report 2026. Stanford Institute for Human-Centered AI.

40. Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. Trends in AI Supercomputers. In ICML Workshop on Technical AI Governance (TAIG).

41. Kalluri, P. R., Agnew, W., Cheng, M., et al. (2025). Computer-vision research powers surveillance technology. Nature, 643, 73–79. https://doi.org/10.1038/s41586-025-08972-6

42. Office of the United Nations High Commissioner for Human Rights. (2022). The right to privacy in the digital age (A/HRC/51/17).

43. Teklehaymanot, H. K., & Nejdl, W. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. arXiv preprint arXiv:2510.12389.

44. Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., ... & Korhonen, A. (2026). Artificial intelligence is creating a new global linguistic hierarchy. arXiv. https://arxiv.org/abs/2602.12018

45. Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 27802-27812).

46. Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., & Dev, S. (2024). SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 842–854).

47. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund.

48. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937). National Bureau of Economic Research.

49. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

50. Skirzynski, J., Danks, D., & Ustun, B. (2025). Discrimination exposed? On the reliability of explanations for discrimination detection. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2554–2569).

51. Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., & Black, E. (2025). Towards effective discrimination testing for generative AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 1028–1047).

52. Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake Technology and Gender-Based Violence: A Scoping Review. Trauma, Violence, & Abuse, 15248380251384271.

53. Posetti, J., Williams, K., Hellmueller, L., Renaud, P., Shabbir, N., & Aboulez, N. (2026). Tipping point: Online violence impacts, manifestations and redress in the AI age. UN Women.

54. Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.

55. Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects (NBER Working Paper No. 33777). National Bureau of Economic Research.

56. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

57. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics.

58. International Monetary Fund. (2024). Broadening the gains from generative AI: The role of fiscal policies (IMF Staff Discussion Note).

59. Organisation for Economic Co-operation and Development. (2026). The OECD.AI index. OECD Publishing.

60. Attard-Frost, B., & Lyons, K. (2025). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. AI and Ethics, 5(3), 2557-2604.

61. UNESCO. (2023). Readiness assessment methodology. UNESCO.

62. Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5312977

63. Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI Competency Objective Scale (AICOS). Computers and Education: Artificial Intelligence.

64. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

65. Math Matters AI. (n.d.). Math Matters AI. https://www.mathmatters.ai

66. Hinojosa, T., Rapaport, A., Jaciw, A., & Zacamy, J. (2016). Exploring the foundations of the future STEM workforce: K–12 indicators of postsecondary STEM success. Regional Educational Laboratory Southwest. https://ies.ed.gov/use-work/resource-library/report/systematic-literature-review/exploring-foundations-future-stem-workforce-k-12-indicators-postsecondary-stem-success

67. United Nations Conference on Trade and Development. (2025). Technology and innovation report 2025: Inclusive artificial intelligence for development. United Nations. https://doi.org/10.18356/9789211068016

68. Chauhan, P. (2025). AI and human rights: Global South perspectives. International Journal of Humanities Social Science and Management, 5(4), 563–568. https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf

69. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. International Affairs, 100(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

70. Khodabin, M., & Arsalani, A. (2025). Artificial intelligence literacy as national strategy: A systematic review of policy, equity, and capacity building across the Global South. World Studies in Policy Sciences, 9, 777–814. https://doi.org/10.22059/wsps.2025.396472.1530

71. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., et al. (2024). On the societal impact of open foundation models. arXiv. https://arxiv.org/abs/2403.07918

72. Grinstead, B., Holler, C., & Braun, F. (2026, May). Behind the scenes hardening Firefox with Claude Mythos Preview. Mozilla Hacks. https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/

73. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). https://doi.org/10.18653/v1/2024.findings-acl.349

74. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of EMNLP 2025 (pp. 10584–10598).

75. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). The global impact of AI: Mind the gap (IMF Working Paper No. 24/136). International Monetary Fund.

76. World Bank. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank.

77. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937).

78. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper No. 24001

79. Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. OECD Science, Technology and Industry Working Papers, No. 2023/11.

80. McKinsey & Company. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

81. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

82. Brant, A., Singh, P., Yin, X., et al. (2025). Performance of a deep learning diabetic retinopathy algorithm in India. JAMA Network Open, 8(3), e250984. https://doi.org/10.1001/jamanetworkopen.2025.0984

83. UNESCO. (2025). AI and the future of education: disruptions, dilemmas and directions

84. Cristia, J. et al. (2017). Technology and child development: evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3), 295–320.

85. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

86. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N., & Tekka, R. (2025). Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots. Science Progress, 108(4), 00368504251394584.

87. Noy, S. & Zhang, W. (2023). “Experimental evidence on the productivity effects of generative artificial intelligence.” Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586.

88. Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2026). “The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers.” Management Science. https://doi.org/10.1287/mnsc.2025.00535

89. Dell’Acqua, F., McFowland, E. III, Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023). “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” Harvard Business School Working Paper 24-013. SSRN: https://ssrn.com/abstract=4573321

90. Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. Sustainability, 17(5), 2281. https://doi.org/10.3390/su17052281

91. Dhal, S. B., & Kar, D. (2024). Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization. Forecasting, 6(4), 925–951. https://doi.org/10.3390/forecast6040046

92. Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI scribe and electronic health record efficiency. JAMA Network Open, 8(10), e2537000.

93. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

94. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Wilson Hannay, S. B., ... & Lee, K. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst Innovations in Care Delivery, 6(5), CAT-25.

95. Paul A. David (1990, May). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. The American Economic ReviewVol. 80, No. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, pp. 355-361

96. Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. Journal of Economic Perspectives, 14(4), 23–48. https://doi.org/10.1257/jep.14.4.23

97. Shaw, S. D., & Nave, G. (2026). Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. SSRN. https://doi.org/10.2139/ssrn.6097646

98. Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. Educational Psychology Review, 37(2), 45.

99. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., & Boulesteix, A.-L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078451. https://doi.org/10.1136/bmj-2024-078451

100. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., Reed, M., & Fraser, E. D. G. (2019). The politics of digital agricultural technologies: A preliminary review. Sociologia Ruralis, 59(2), 203–229. https://doi.org/10.1111/soru.12233

101. United Nations General Assembly. (2024). Global Digital Compact (Annex II to the Pact for the Future, A/RES/79/1). United Nations. https://docs.un.org/en/A/RES/79/1

102. UNESCO. (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. https://www.unesco.org/en/articles/k-12-ai-curricula-mapping-government-endorsed-ai-curricula

103. Almatrafi, O., Johri, A., & Lee, H. (2024, 2024/06/01/). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/https://doi.org/10.1016/j.caeo.2024.100173

104. Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025, 2025/06/01/). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. Computers and Education: Artificial Intelligence, 8, 100422. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100422

105. Atias, O., & Mawasi, A. (2025, 2025/12/01/). Conceptualizing AI literacies for children and youth: A systematic review on the design of AI literacy educational programs. Computers and Education: Artificial Intelligence, 9, 100491. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100491

106. Kasirzadeh, A., & Gabriel, I. (2025, April). Characterizing AI Agents for Alignment and Governance. https://arxiv.org/abs/2504.21848

107. Wijk, H., Lin, T. R., Becker, J., Jawhar, S., Parikh, N., Broadley, T., ... & Barnes, E. (2025, October). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In International Conference on Machine Learning (pp. 66772-66832). PMLR.

108. Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., ... & Weng, L. (2025, May). Mle-bench: Evaluating machine learning agents on machine learning engineering. In International Conference on Learning Representations (Vol. 2025, pp. 50466-50494).

109. Pichai, S. (2026, April 22). Cloud Next ’26: Momentum and innovation at Google scale. Google Blog. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai/

110. Cybersecurity and Infrastructure Security Agency. (2025, January 14). CISA, JCDC, government and industry partners publish AI cybersecurity collaboration playbook. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook

111. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). “Your AI, my shell”: Demystifying prompt injection attacks on agentic AI coding editors. https://doi.org/10.48550/arXiv.2509.22040

112. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820. https://doi.org/10.15779/Z38RV0D

113. Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar’s dividend: Can politicians claim misinformation to evade accountability?. American Political Science Review, 119(1), 71-90.

114. Chan, Y.-T., et al. (2024). Assessing the article screening efficiency of artificial intelligence for systematic reviews. Journal of Dentistry, 149, 105259. https://doi.org/10.1016/j.jdent.2024.105259

115. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., Bennett, J. A., Jha, P., Mukhin, N., Li, J., López-Guajardo, E. A., Sadeghi, S., Bateni, F., & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2, 436–446. https://doi.org/10.1038/s44286-025-00249-z

116. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. http://arxiv.org/abs/2501.10114

117. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. AI Agents That Matter. Transactions on Machine Learning Research.

118. Zhu, L., Lu, Q., Ding, M. et al. Designing meaningful human oversight in AI. AI Ethics 6, 286 (2026). https://doi.org/10.1007/s43681-026-01147-7

119. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science, 7, 549–569. https://doi.org/10.1007/s42001-024-00250-1

120. Djiré, A. E., Kaboré, A. K., Samhi, J., et al. (2026). Learned or memorized? Quantifying memorization advantage in code LLMs. In Proceedings of the International Conference on Software Engineering. https://arxiv.org/abs/2604.13997

121. Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., ... & Kohli, P. (2025). SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263.

122. United Nations Human Rights Council. Expert Mechanism on the Right to Development. (2024). AI, cultural rights and the right to development (A/HRC/EMRTD/11/ CRP.2). United Nations. https://undocs.org/A/HRC/EMRTD/11/CRP.2

123. Brennan Center for Justice. (2025). Gauging AI threat to free and fair elections. https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections

124. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

125. Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences, 118(50), e2102141118.

126. Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. Journal of broadcasting & Electronic media, 64(2), 150-172.

127. Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. Computers in human behavior, 116, 106626.

128. Argyle, L. P. (2025). Political persuasion by artificial intelligence. Science, 390(6777), 983-984.

129. Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., ... & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. Nature, 1-8.

130. Myra Cheng et al. (2026). ,Sycophantic AI decreases prosocial intentions and promotes dependence.Science391,eaec8352(2026).DOI:10.1126/science.aec8352

131. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

132. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 13(6), 522-530.

133. Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. JMIR Mental Health, 12, e81204.

134. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

135. Osler, L. (2026). Hallucinating with AI: distributed delusions and "AI psychosis". Philosophy & Technology, 39(1), 30.

136. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

137. Organisation for Economic Co-operation and Development. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

138. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. Nature. https://doi.org/10.1038/s41586-026-10506-7

139. Kalina Bontcheva (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

140. Laudrain, A. (2026, April 29). When AI governs (dis)information: Five lessons for democracy. Center for Security Studies (CSS), ETH Zurich. https://css.ethz.ch/en/center/CSS-news/2026/04/when-ai-governs-disinformation-five-lessons-for-democracy.html

141. Boine, C. (2023). Emotional attachment to AI companions and European law. MIT Case Studies in Social and Ethical Responsibilities of Computing (Winter 2023). https://doi.org/10.21428/2c646de5.db67ec7f

142. Department of Enterprise, Tourism and Employment. (2026, February 4). General scheme of the Regulation of Artificial Intelligence Bill 2026. Government of Ireland. https://www.gov.ie/en/department-of-enterprise-tourism-and-employment/publications/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026

143. United States Congress. Senate. (2025). S. 3062—GUARD Act: Guidelines for User Age-verification and Responsible Dialogue Act of 2025 (119th Congress). Congress. gov. https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

144. European Commission. (2026, April 29). Blueprint for an age verification solution to help protect minors online. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online

145. Reuters. (2026, April 24). From Australia to Europe, countries move to curb children's social media access. https://www.reuters.com/legal/government/australia-europe-countries-move-curb-childrens-social-media-access-2026-04-24/

146. Morrin H, Nicholls L, Levin M et al. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 2026; 13, 522-530

147. Examining the harm of AI chatbots, Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary, 119th Cong. (2025) (testimony of Megan Garcia). https://www.judiciary.senate.gov/download/2025-09-16-pm_-testimony_-garciapdf

148. Solove, D. J. (2025). Artificial intelligence and privacy. Florida Law Review, 77. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

149. Office of the United Nations High Commissioner for Human Rights. (2025). The right to privacy in the digital age (UN Doc. A/HRC/60/45). https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high

150. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

151. UNESCO, IRCAI, & University College London. (2024). Challenging systematic prejudices: An investigation into bias against women and girls in large language models. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

152. Shah, S. S. (2024). Gender bias in artificial intelligence: Empowering women through digital literacy. Premier Journal of Artificial Intelligence, 1, 1000088. https://doi.org/10.70389/PJAI.1000088

153. Haider, S. A., Borna, S., Gomez-Cabello, C. A., et al. (2026). The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. Journal of Racial and Ethnic Health Disparities, 13, 188–217. https://doi.org/10.1007/s40615-024-02237-0

154. International Telecommunication Union. (2025). Joint statement on AI and the rights of the child. International Telecommunication Union. https://www.itu.int/hub/publication/d-str-cyb_joint-2025

155. Johnson, A. K., Winther, D. K., & Bhargava, A. (2026). Artificial intelligence and child sexual exploitation and abuse: Emerging risks and implications for children's rights (Issue brief). United Nations Children's Fund (UNICEF). https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL3.pdf

156. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123

157. Internet Watch Foundation. (2026). Harm without limits: AI child sexual abuse material through the eyes of our Analysts. https://www.iwf.org.uk/media/hl1nvdti/iwf-ai-csam-report-2026.pdf

158. Goodacre, E.J. and Gibson, J.L. (2026) AI in the early years: Examining the implications of GenAI toys for young children. Cambridge: University of Cambridge (unpublished report, available via Apollo repository).

159. Kurian, N. (2025). Developmentally aligned AI: a framework for translating the science of child development into AI design. AI, Brain and Child, 1(1). https://doi.org/10.1007/s44436-025-00009-z

160. Livingstone, S., & Sylwander, K. R. (2025). Conceptualizing age-appropriate social media to support children's digital futures. British Journal of Developmental Psychology. https://doi.org/10.1111/bjdp.70006

161. Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. Big Data & Society, 12(4), 20539517251389860.

162. Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Administration Review, 82(4), 706–718. https://doi.org/10.1111/puar.13483

163. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, 94, 15–55. https://ssrn.com/abstract=3333423

164. Mantelero, A. (2022). Human rights impact assessment and AI. In Beyond data: Human rights, ethical and social impact assessment in AI (pp. 45-91). The Hague: TMC Asser Press.

165. Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Computer Law and Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105561

166. United Nations Educational, Scientific and Cultural Organization. (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

167. Livingstone, S. & Pothong, K. (2025). Child Rights Impact Assessment: A Policy Tool for aRights-Respecting Digital Environment. Policy & Internet.

168. Denain, J.-S., & Barry, A. (2026, April 16). Have AI capabilities accelerated? Epoch AI. https://epoch.ai/blog/have-ai-capabilities-accelerated

169. Model Evaluation & Threat Research (METR). (2026, May 8). Task-completion time horizons of frontier AI models. https://metr.org/time-horizons/

170. Juniewicz, I. (2026, February 26). Hyperscaler capex has quadrupled since GPT-4's release. Epoch AI. https://epoch.ai/data-insights/hyperscaler-capex-trend

171. Epoch AI. (2026, May 28). Data on AI companies. https://epoch.ai/data/ai-companies?view=graph&tab=revenue&showTotal=true

172. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

173. AI Alliance Network. (2025). AI Horizons: What will AI technologies look like in 10 years? A Research Project. November 2025, p. 83. Based on 21 foresight sessions and 32 in-depth interviews with over 270 AI researchers from 36 countries.

174. Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., ... & Lång, K. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet, 407(10527), 505-514.

175. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://www.nature.com/articles/s41467-025-57640-w

176. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458.

177. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. npj Science of Learning, 10(1), 29.

178. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., ... & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2(7), 436-446.

179. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., ... & Fraser, E. D. (2019). The politics of digital agricultural technologies: a preliminary review. Sociologia ruralis, 59(2), 203-229.

180. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

181. Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine.

182. Tao, X., Zhou, S., Ding, K., Li, S., Li, Y., Wu, B., ... & Han, S. (2026). An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nature Medicine, 1-9.

183. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., ... & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1-8.

184. Scientific Computing World. (2025, August 5). AI health assistant displays high diagnostic accuracy in tests. https://www.scientific-computing.com/article/sberhealths-gigachat-powered-ai-health-assistant-displays-high-diagnostic-accuracy-tests

185. MASTEL, P. M., de Dieu Nyandwi, J., Rutunda, S., & Kabanda, K. (2025). Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda. In Workshop on Large Language Models and Generative AI for Health at AAAI 2025.

186. Mateen, B. A., Williams, G., Korom, R., Mwaniki, P., Emmanuel-Fabula, M., & Agweyu, A. (2026). Learning Effects from A GenAI-based Clinical Decision Support System in Primary Healthcare. medRxiv, 2026-05.

187. OECD (2025), Results from TALIS 2024: The State of Teaching, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/90df6235-en.

188. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

189. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. npj Science of Learning, 10(1), 29. https://www.nature.com/articles/s41539-025-00320-7

190. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

191. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

192. Vodafone Foundation.(2025) AI in European Schools: A European Report --- comparing seven countries

193. World Food Programme. (2024). HungerMap LIVE: Global hunger monitoring. https://hungermap.wfp.org/

194. Yakov and Partners. (2024). Artificial intelligence in Russia's agricultural sector: Hype or real money? https://yakovpartners.com/publications/ai-in-agriculture/

195. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

196. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15, 17458. https://doi.org/10.1038/s41598-025-97652-6

197. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

198. Sparling, T. M., Offner, C., Deeney, M., Denton, P., Bash, K., Juel, R., ... & Kadiyala, S. (2024). Intersections of climate change with food systems, nutrition, and health: an overview and evidence map. Advances in Nutrition, 15(9), 100274.

199. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://doi.org/10.1038/s41467-025-57640-w

200. Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., & Verdin, J. (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. Remote Sensing of Environment, 237, 111553. https://doi.org/10.1016/j.rse.2019.111553

201. Food and Agriculture Organization of the United Nations. (2025). Evidence in action: How anticipatory cash transfers reduce humanitarian needs and strengthen resilience in Somalia. https://www.fao.org/agrifood-economics/publications/detail/en/c/1756210/

202. World Food Programme. (2025). WFP's evidence base on anticipatory action 2015–2024. https://www.wfp.org/publications/wfps-evidence-base-anticipatory-action-2015-2024

203. Nakalembe, C., Kerner, H. R., Zvonkov, I., Humber, M., Galvez, A. S., Venturini, S., & Becker Reshef, I. (2025). A framework for EO based National Agricultural Monitoring for the African context. npj Sustainable Agriculture, 3, 45. https://www.nature.com/articles/s44264-025-00083-z

204. Nowak, A. C., et al. (2024). Opportunities to strengthen Africa's efforts to track national level climate adaptation. Nature Climate Change, 14, 876–882.https://www.nature.com/articles/s41558-024-02054-7

205. Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., et al. (2026). Forecasting the economic effects of AI (NBER Working Paper No. w35046). National Bureau of Economic Research. https://doi.org/10.3386/w35046

206. Goldman Sachs (2023). Generative AI Could Raise Global GDP by 7 Percent. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

207. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. Artificial intelligence review, 55(1), 589-656.

208. United Nations News. (2026, February 18). Artists face steep income decline due to AI, UNESCO finds. United Nations. https://news.un.org/en/story/2026/02/1166989

209. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333-372.

210. Gmyrek, P., Winkler, H., & Garganta, S. (2024). Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America (ILO Working Paper 121 / World Bank Policy Research Working Paper 10863). International Labour Organization & World Bank.

211. Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. Quarterly Journal of Economics, 139(3), 1399–1465.

212. The Conversation. (2026, March 18). Tech companies are blaming massive layoffs on AI. What's really going on? https://theconversation.com/tech-companies-are-blaming-massive-layoffs-on-ai-whats-really-going-on-278314

213. Society for Human Resource Management. (2026, May). The AI layoffs narrative: Real transformation, or scapegoat? https://www.shrm.org/topics-tools/news/technology/ai-layoffs-transformation-scapegoat

214. OpenAI. (2026, February 27). Company communication accompanying a US$110 billion private funding round [Company communication]

215. Rodrik, D., & Sabel, C. (2020). Building a Good Jobs Economy (HKS Faculty Research Working Paper RWP20-001). Harvard Kennedy School.

216. Acemoglu, D. (2025). The simple macroeconomics of AI. Economic Policy, 40(121), 13–58. Acemoglu's headline figure is a TFP gain of no more than 0.71% over ten years.

217. Korinek, A., & Suh, D. (2024). Scenarios for the Transition to AGI (NBER Working Paper No. 32255). In their full-automation scenario, output rises sharply while wages collapse once automation crosses a critical threshold.

218. Karger, E., Kuusela, O., Abaluck, J., Bryan, K., Halperin, B., Jones, T., et al. (2026). Forecasting the Economic Effects of AI. Federal Reserve Bank of Chicago and Forecasting Research Institute, March 2026.

219. Jones, C. I., & Tonetti, C. (2026). Past Automation and Future AI: How Weak Links Tame the Growth Explosion. Working paper presented at the Bendheim Center for Finance, Princeton University, March 2026.

220. Trammell, P., & Korinek, A. (2025). Economic Growth under Transformative AI (NBER Working Paper No. 31815, revised September 2025).

221. OECD (2024). Artificial Intelligence, Data and Competition (OECD Artificial Intelligence Papers No. 18). OECD Publishing, Paris. https://doi.org/10.1787/e7e88884-en

222. Acemoglu, D., & Restrepo, P. (2026). Automation and rent dissipation: Implications for wages, inequality, and productivity. Quarterly Journal of Economics, 141(2), 1521–1579.

223. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. arXiv. https://doi.org/10.48550/arXiv.2509.22040

224. Regilme, S. S. F. (2024). Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. SAIS Review of International Affairs, 44(2), 75–92. https://muse.jhu.edu/pub/1/article/950958

225. Barnett-Itzhaki, Z. (2026). The water footprint of artificial intelligence: Emerging solutions and governance imperatives. Water Research, 299, 125866. https://doi.org/10.1016/j.watres.2026.125866

226. International Energy Agency. (2025). Global Critical Minerals Outlook 2025 – Analysis (p. 312). https://iea.blob.core.windows.net/assets/ef5e9b70-3374-4caa-ba9d-19c72253bfc4/GlobalCriticalMineralsOutlook2025.pdf

227. Zhu, L., & Lu, Q. (2026). Verifiability-First AI Engineering in the Era of AIware: A Conceptual Framework, Design Principles, and Architectural Patterns for Scalable Verification. Design Principles, and Architectural Patterns for Scalable Verification (January 07, 2026).

228. UN Scientific Advisory Board. (2026, March 19). AI deception: Brief of the Scientific Advisory Board. United Nations. https://www.un.org/scientific-advisory-board/en/ai-deception

229. Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., ... & Zhu, L. (2025). International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications. arXiv preprint arXiv:2510.13653.

230. United Nations Secretary-General. (2024). Human rights due diligence guidance on digital technology use. https://www.ohchr.org/sites/default/files/2024-08/digital-technology-use-guidance-sg-1-en.pdf

231. AI Equality Toolbox. (2025). Human rights impact assessment methodology. https://aiequalitytoolbox.com/tools/hria-workbook/

232. Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D. S., Luda di Cortemiglia, V., Lobuntsova, Y., Nnorom, I., Pralat, N., & Wagner, M. (2024). The global e-waste monitor 2024: Electronic waste rising five times faster than documented e-waste recycling. United Nations Institute for Training and Research & International Telecommunication Union. https://ewastemonitor.info/the-global-e-waste-monitor-2024/

233. International Panel on the Information Environment. (2024). Expert survey on the global information environment 2024: Searching for solutions (SFP2024-1). https://www.ipie.info/research/sfp2024-1

234. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. Social Network Analysis and Mining, 13(1), 30. https://doi.org/10.1007/s13278-023-01028-5

235. James, K., Perveen, S., & Jacobson, B. (2025). Drained by data: The cumulative impact of data centers on regional water stress. Ceres. https://www.ceres.org/resources/reports/drained-by-data-the-cumulative-impact-of-data-centers-on-regional-water-stres

236. Office of the United Nations High Commissioner for Human Rights. (2024). Taxonomy of generative AI-related human rights harms. https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf

237. Patel, A. (2025, May 19). Freedom of expression, artificial intelligence and elections. United Nations Educational, Scientific and Cultural Organization (UNESCO) & United Nations Development Programme (UNDP). https://www.undp.org/publications/freedom-expression-artificial-intelligence-and-elections

238. Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrociocchi, W., & Orben, A. (2026). Fractured reality: How democracy can win the global struggle over the information space (JRC144603). Publications Office of the European Union. https://doi.org/10.2760/9358883

239. Observatory on Information and Democracy. (2024). Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

240. Solove, D. J. (2025). Artificial intelligence and privacy. Florida Law Review, 77. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

241. Office of the United Nations High Commissioner for Human Rights (2025). The right to privacy in the digital age. URL https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high. UN Doc. A/HRC/60/45.

242. Council of Europe. Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

243. European Union. (2024, June 13). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

244. Expert Mechanism on the Right to Development. (2024). Artificial intelligence, cultural rights and the right to development (A/HRC/EMRTD/13/CRP.1). United Nations Human Rights Council. https://www.ohchr.org/sites/default/files/documents/issues/development/emd/session13/a-hrc-emrtd-13-crp-1.pdf

245. Council of Europe, Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

246. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. Nature. https://doi.org/10.1038/s41586-026-10506-7

247. Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. Communications of the ACM, 68(7), 54–61. https://doi.org/10.1145/3724499

248. Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734.

249. Arsenault, A. C., & Kreps, S. (2026). Whose voice counts? The role of large language models in public commenting. Big Data & Society, 13(1). https://doi.org/10.1177/20539517261419341

250. Alslaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. Frontiers in Artificial Intelligence, 6, 1125191. https://doi.org/10.3389/frai.2023.1125191

251. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science, 385, eadq1814. https://doi.org/10.1126/science.adq1814

252. Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. PNAS Nexus, 4(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

253. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

254. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

255. Germano, F., Gómez, V., & Sobbrio, F. (2025). Ranking for engagement: How social media algorithms fuel misinformation and polarization. Barcelona School of Economics, Working Paper No. 1501. https://bw.bse.eu/wp-content/uploads/2025/07/1501.pdf

256. Council of Europe CDMSI. (2025). Guidance Note on Generative AI and Freedom of Expression. CDMSI(2025) https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

257. Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., ... & De Bie, T. (2026). Large language models reflect the ideology of their creators. npj Artificial Intelligence, 2(1), 7.

258. Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. Nature Computational Science, 4, 818–823. https://doi.org/10.1038/s43588-024-00700-3

259. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

260. Council of Europe. (2024). Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law (Council of Europe Treaty Series No. 225). https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

261. Observatory on Information and Democracy. (2024). Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

262. Council of Europe, Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

263. OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

264. United Nations General Assembly. (2017). Promotion and protection of human rights: Human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms: Report of the Third Committee, 72nd session (A/72/...). United Nations Digital Library. http://digitallibrary.un.org/record/1326669

265. Penney, J. W. (2025). Chilling effects: Repression, conformity, and power in the digital age. Cambridge University Press. https://doi.org/10.1017/9781108918022

266. UN Women. (2022). Tipping point: The chilling escalation of online violence against women in the public sphere. UN Women. https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

267. United Nations Educational, Scientific and Cultural Organization. (2024). Challenging systematic prejudices: An investigation into bias against women and girls in large language models. https://unesdoc.unesco.org/ark:/48223/pf0000388971

268. Chowdhury, R., & Lakshmi, D. (2023). "Your opinion doesn't matter, anyway": Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.). UNESCO.

269. United Nations Educational, Scientific and Cultural Organization. (2024, March 7). Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

270. Fung, P. (2019, June 30). This is why AI has a gender problem. World Economic Forum. https://www.weforum.org/stories/2019/06/this-is-why-ai-has-a-gender-problem/

271. United Nations Conference on Trade and Development. (2025). Technology and innovation report 2025: The AI divide. https://unctad.org/publication/technology-and-innovation-report-2025

272. Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. arXiv preprint arXiv:2010.15581.

273. Coeckelbergh, M. (2026) Technofascism: AI, Big Tech, and the New Authoritarianism. AI & Society https://doi.org/10.1007/s00146-026-02862-9

274. Varoufakis, Y. (2024). Technofeudalism: What killed capitalism. Melville House.

275. Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2025). Computer-vision research powers surveillance technology. Nature, 643(8070), 73-79. https://www.nature.com/articles/s41586-025-08972-6

276. Fola-Rose, A., Solomon, E., Bryant, K., & Woubie, A. (2024, August). A systematic review of facial recognition methods: Advancements, applications, and ethical dilemmas. In Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI 2024) (pp. 314–319). IEEE. https://doi.org/10.1109/IRI62200.2024.00070

277. Fussey, P., & Murray, D. (2025). Facial Recognition Surveillance: Policing and Human Rights in the Age of Artificial Intelligence. Oxford University Press.

278. Office of the United Nations High Commissioner for Human Rights. (2024). Mapping report: Human rights and new and emerging digital technologies (A/HRC/56/45). https://www.ohchr.org/en/documents/reports/mapping-report-human-rights-and-new-and-emerging-digital-technologies

279. Secretary-General. (2024). Human rights in the administration of justice (A/79/296). United Nations. https://digitallibrary.un.org

280. American Civil Liberties Union, More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology (2025), https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology: Documentation of at least 14 publicly known wrongful arrests in the United States attributed to police use of facial recognition; in nearly all cases the persons wrongfully arrested were Black.

281. Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. ACM Journal on Responsible Computing, 1(1), Article 2, 1–32. https://doi.org/10.1145/3616473

282. German Marshall Fund. (2024). Spitting Images: Tracking Deepfakes and Generative AI in Elections.

283. International Institute for Democracy and Electoral Assistance. (2024). The 2024 global elections super-cycle. https://www.idea.int/initiatives/the-2024-global-elections-supercycle

284. Recorded Future. (2024). 2024 Deepfakes and Election Disinformation Report: Key Findings and Mitigation Strategies.

285. Associated Press. (2025, June 13). New Hampshire jury acquits consultant behind AI robocalls mimicking Biden on all charges.

286. Federal Communications Commission. (2024, September). $6 million fine against Steven Kramer for AI-generated robocalls.

287. Global Witness. (2024). What Happened on TikTok Around the Romanian Elections?

288. IFES. (2024). The Romanian 2024 Election Annulment: Addressing Emerging Threats to Electoral Integrity.

289. Associated Press. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide.

290. United Nations. (1966). International Covenant on Civil and Political Rights. Articles 17, 18, 19 and 25.

291. Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. Articles 8, 9 and 10; Protocol No. 1, Article 3.

292. EQUATE Language AI Readiness Index (https://equate.vercel.app/en)

293. Han, W., Zhang, Y., Chen, Z., Liu, B., Lin, H., Zhang, B., ... & Zheng, Y. (2025). MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages. arXiv preprint arXiv:2506.19468.

294. Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024, June). The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2040-2079).

295. Gamboa, L. C. L., Feng, Y., & Lee, M. (2025, November). Social Bias in Multilingual Language Models: A Survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 27845-27868).

296. Choudhury, M., Sitaram, S., Vashistha, A., et al. (2026). AI for the Global South: 12 critical research questions for the next decade. AI for the Global South (AI4GS), Mohamed bin Zayed University of Artificial Intelligence. https://ai4gs.github.io/

297. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. ACM Computing Surveys, 57(6), 1-36. https://dl.acm.org/doi/pdf/10.1145/3700438

298. Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

299. U.S. PIRG Education Fund. (2025). AI comes to playtime: Artificial companions, real risks. U.S. PIRG Education Fund. https://pirg.org/edfund/wp-content/uploads/2025/12/AI-Comes-to-Playtime-Artifical-companions-real-risks.pdf

300. Staksrud, E., Mascheroni, G., Milosevic, T., Ní Bhroin, N., Ólafsson, K., Şengül-İnal, G., & Stoilova, M. (2026). European children's use and understanding of generative AI. EU Kids Online V.

301. Committee on the Rights of the Child. (2021). General comment No. 25 (2021) on children's rights in relation to the digital environment (CRC/C/GC/25). United Nations. https://digitallibrary.un.org/record/3906061

302. UNICEF (2025). Guidance on AI and Children: Updated guidance for governments and businesses to create AI policies and systems that uphold children's rights. https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf

303. UNESCO (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

304. Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Version 1. Stanford Digital Repository. Available at https://purl.stanford.edu/mn692xc5736/version/1. https://doi.org/10.25740/mn692xc5736.

305. Livingstone, S., Atabey, A., Stoilova, M., & Sylwander, K. R. (2025). How does, and how could, generative AI respect and enable children's rights? In N. Ni Loideain (Ed.), AI and power: Regulation and rights. University of London Press.

306. European Parliamentary Research Service. (2024). Children and deepfakes. European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775855

307. United Nations Children's Fund (UNICEF). (2026). Artificial intelligence and child sexual abuse and exploitation [Issue brief]. https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

308. Ozcan, B., Sperati, V., Giocondo, F., Schembri, M., & Baldassarre, G. (2022, June). Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. In Proceedings of the 21st Annual ACM Interaction Design and Children Conference (pp. 625-628).

309. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://www.cam.ac.uk/stories/ai-toys-study-play

310. British Standards Institution. (2026, May). Half of children have AI toys despite safety concerns. https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/may/half-of-children-have-ai-toys-as-parents-allow-widespread-use-despite-safety-concerns-and-gaps-in-guidance-parents/

311. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://doi.org/10.17863/CAM.126270

312. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: a research agenda–from artificial companions for learning to general artificial companions for Global Harwell. Research & Practice in Technology Enhanced Learning, 20.

313. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

314. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications. Leverhulme Centre for the Future of Intelligence

315. De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). Emotional manipulation by AI companions. arXiv preprint arXiv:2508.19258.

316. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

317. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

318. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y. C. (2025, April). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-17).

319. De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. Nature medicine, 30(5), 1269-1275.

320. Radesky, J., Bragg, M. A., & Hiniker, A. (2026). Risks and Consequences of Children's Use of Social AI—A Framework. JAMA Pediatrics. https://doi.org/10.1001/jamapediatrics.2026.1349

321. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. new media & society, 14614448251395192.

322. Jacobs, K. A. (2024). Digital loneliness—changes of social recognition through AI companions. Frontiers in Digital Health, 6, 1281037.

323. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

324. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications.

325. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

326. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

327. Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

328. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. Practice Innovations. https://doi.org/10.1037/pri0000292

329. Callahan, C., Tanner, L., Coe, C., Davis, M., Glover, J., Bernstein, E., ... & Kunkle, S. (2026). Real-World Use of a Mental Health AI Companion: Multiple Methods Study. JMIR Formative Research, 10, e86904.

330. Associated Press. (2026, June 1). Chatbot AI lawsuit alleges links to teen suicide. https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0

331. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

332. Green, H. H. (2026, March 14). New study raises concerns about AI chatbots fueling delusional thinking. The Guardian. https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis

333. Ministère du Travail, de la Santé, des Solidarités et des Familles. (2025, March 24). La santé mentale, grande cause nationale 2025. Gouvernement de la France. https://solidarites.gouv.fr/la-sante-mentale-grande-cause-nationale-2025

334. American Psychiatric Association. (n.d.). Applications of artificial intelligence in mental health care. https://www.psychiatry.org/psychiatrists/practice/artificial-intelligence/applications

335. U.S. Food and Drug Administration. (2025). Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence–enabled digital mental health medical devices. https://www.fda.gov/media/189391/download

336. Hollis, A., & McKeown, G. (2024, September). Empathic AI for autism: Potential and pitfalls of empathic social chatbots in addressing loneliness. In 24th ACM International Conference on Intelligent Virtual Agents: CONNECT, A Workshop on Connecting Interdisciplinary Research on Connections With and Through Technology: IVA 2024.

337. Sharma, D., Meshkat, S., Perivolaris, A., Kamaleddin, M. A., Teferra, B. G., Rueda, A., ... & Bhat, V. (2026). Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward. npj Digital Medicine.

338. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), e0240376.

339. Garg, M. (2024). Towards mental health analysis in social media for low-resourced languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(3), 1-22.

340. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. Findings of the Association for Computational Linguistics: ACL 2024, 5865–5877. https://doi.org/10.18653/v1/2024.findings-acl.349

341. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

342. Fu, Y. V., Ramachandran, G. K., Park, N., Lybarger, K., Xia, F., Uzuner, Ö., & Yetisgen, M. (2025). BioMistral-NLU: Towards more generalizable medical language understanding through instruction tuning. AMIA Joint Summits on Translational Science Proceedings, 2025, 149–158. https://pubmed.ncbi.nlm.nih.gov/40502228/

343. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. Findings of the Association for Computational Linguistics: ACL 2024, 5848–5864. https://aclanthology.org/2024.findings-acl.348/

344. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language models for medicine. Nature Communications, 15, 8384. https://doi.org/10.1038/s41467-024-52417-z

345. Nwabufo, J., Ogueji, K., Adelani, D. I., Alabi, J., et al. (2025). Healthcare NLP for African Languages: Current State and Challenges. Proceedings of the AfricaNLP Workshop (AfricaNLP 2025). https://aclanthology.org/2025.africanlp-1.32/

346. Okafor, U. (2025). Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 221–229. https://aclanthology.org/2025.africanlp-1.32/

347. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In J. Sälevä & A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) (pp. 236–241). https://doi.org/10.18653/v1/2024.mrl-1.20

348. Cronin, A., Kelly, A., Wrona, M., O'Donnell, P., Hassan, A., Myles, T., Fallon, T., & MacFarlane, A. (2025). The patient-safety implications of AI-based communication with migrants in general practice: a scoping review. BJGP Open, 9(4), BJGPO.2025.0107. https://bjgpopen.org/content/9/4/BJGPO.2025.0107

349. House of Lords Public Services Committee. (2025). Lost in translation? Interpreting services in the courts (2nd Report of Session 2024–26, HL Paper 87). https://committees.parliament.uk/publications/44602/documents/221328/default/

350. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

351. MIT AI Risk Initiative. (2025). AI risk mitigation database and draft taxonomy. https://airisk.mit.edu/ai-riskmitigations

352. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.

353. Staufer, L., Feng, K., Wei, K., et al. (2026). The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems. https://doi.org/10.48550/arXiv.2602.17753

354. Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., ... & Isaac, W. (2025). Toward an evaluation science for generative ai systems. arXiv preprint arXiv:2503.05336.

355. Rabanser, S., Kapoor, S., Kirgis, P., et al. (2026). Towards a science of AI agent reliability. https://doi.org/10.48550/arXiv.2602.16666

356. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. The Wharton School Research Paper.

357. Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. arXiv preprint arXiv:2601.20245.

358. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology, 10(10), 896-903.

359. Epoch AI. (2026). Data on AI models. https://epoch.ai/data/ai-models

360. Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for ai agents. arXiv preprint arXiv:2506.12469.

361. Fink, M. (2025). Operationalizing meaningful human oversight under Article 14 of the EU AI Act. In AI Act commentary: A thematic analysis (forthcoming). Hart-Bloomsbury.

362. Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., ... & Bau, D. (2026). Agents of chaos. arXiv preprint arXiv:2602.20021.

363. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

364. Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. Machine Intelligence Research Institute. https://intelligence.org/files/Corrigibility.pdf.

365. Zeng, Y., Lu, E., Guo, X., Huangfu, C., Xie, J., Chen, Y., ... & Younas, A. (2025). AI Governance InternationaL Evaluation Index (AGILE Index) 2025. arXiv preprint arXiv:2507.11546.

366. Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2023, December 13). Considerations for governing open foundation models. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-mode

367. Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. Nature food, 4(11), 941-948.

368. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616, 259–265. https://doi.org/10.1038/s41586-023-05881-4

369. Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). The foundation model transparency index v1.1. arXiv:2407.12929. https://arxiv.org/abs/2407.12929

370. BigScience Workshop, Le Scao, T., Fan, A., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv. https://doi.org/10.48550/arXiv.2211.05100

371. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv. https://arxiv.org/abs/2302.13971

372. Qwen Team. (2024). QwenLM. GitHub. https://github.com/QwenLM/Qwen

373. Qwen Team. (2024). Qwen2 technical report. arXiv. https://arxiv.org/abs/2407.10671

374. DeepSeek-AI. (2024). DeepSeek-V3 technical report. arXiv. https://doi.org/10.48550/arXiv.2412.19437

375. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://arxiv.org/abs/2501.12948

376. Jiang, A. Q., et al. (2023). Mistral 7B. arXiv. https://arxiv.org/abs/2310.06825. Mistral AI.

377. Technology Innovation Institute. (2023). The Falcon series of open language models. arXiv. https://arxiv.org/abs/2311.16867

378. SberDevices. (2026, March). GigaChat-3.1: Большое обновление больших моделей. Habr. https://habr.com/ru/companies/sberbank/articles/1014146/

379. Yandex. (2025, February 25). YandexGPT 5 — в Алисе, облаке и опенсорсе. Habr. https://habr.com/ru/companies/yandex/articles/885218/

380. Sarvam AI. (2025). Sarvam AI models on Hugging Face. https://huggingface.co/sarvamai

381. SB Intuitions. (2025). Sarashina models on Hugging Face. https://huggingface.co/sbintuitions

382. Naver Cloud. (2024). HyperCLOVA X technical report. arXiv. https://arxiv.org/abs/2404.01954

383. Seger, E., Dreksler, N., Moulange, R., et al. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Centre for the Governance of AI. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

384. Anthropic. (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

385. Partnership on AI. (2023). PAI's guidance for safe foundation model deployment. https://partnershiponai.org/modeldeployment/

386. International Energy Agency. (2025). Energy and AI. International Energy Agency. https://www.iea.org/reports/energy-and-ai

AI に関する独立国際科学パネルについて

構成とマンデート

AI に関する独立国際科学パネルは、グローバル・デジタル・コンパクト及び未来のための協定において約束されたとおり、総会決議 79/325 を通じて国際連合の中に設置された。

パネルは、AI 及び関連分野における卓越した専門性に基づいて総会により任期 3 年で任命された 40 名の独立の専門家で構成され、ジェンダー・バランスが確保され、加盟国の 5 つの地域グループすべてから委員が参加している。専門分野は、中核的な AI 技術、応用 AI、安全性とインフラ、AI の政策・倫理・影響にまたがる。

パネルは、AI の機会、リスク及び影響に関する既存の研究を統合・分析するエビデンスに基づく科学的評価を、政策に関連するが政策規定的ではない年次サマリー報告書（必要と判断するテーマ別ブリーフを含む）として発行するマンデートを負う。その科学的作業は、独立性、科学的信頼性と厳格さ、学際性、包摂的な参加の原則に導かれる。

パネルはまた、その年次サマリー報告書を国連の AI ガバナンスに関するグローバル対話に提出するマンデートを負う。グローバル対話とより広い国際プロセスに情報を提供することにより、パネルは、国際社会が新たな課題を予見し、より良い情報に基づくガバナンスの意思決定を行い、世界中の政策立案者の間の情報格差を縮小することを可能にする。

パネルは、国連デジタル・新興技術室（UN ODET）が調整するパネル事務局の支援を受ける。

本報告書に至る過程

2026 年 2 月に任命され、同年 3 月に初会合を開いてからの 3 か月間、パネルは本報告書を仕上げるために集中的に作業した。この科学的交流と集合的分析の集中的な過程には、選出された共同議長の進行の下、3 日間の対面全体会合と 60 回を超えるパネルのバーチャル会合が含まれた。

本暫定報告書は、外部専門家とのより広い協議、並びにその後のテーマ別ブリーフ及び年次サマリー報告書の基準点（ベースライン）となる土台を築くものである。

パネルの科学的独立性

パネル委員は、科学的に独立した専門家として個人の資格で務める。国連の「任務を行う専門家（experts on mission）」としての指名を通じて、各委員は、その職務の遂行に関し、いかなる政府その他のいかなる筋からも指示を求めず又は受けないことを宣言し誓約している。「任務を行う専門家の地位、基本的権利及び義務を規律する規則」（ST/SGB/2002/9）には、その行動と説明責任に関する規定も含まれている。

ドナー

パネル事務局は、以下の政府及びパートナーの資金拠出及び現物拠出に深く感謝する。これらなしには、パネルはその責務を果たすことができなかった。

ドイツ政府

日本政府

スペイン政府

Omidyar Network Fund

パネル事務局

コーディネーター

- Amandeep Singh Gill、デジタル・新興技術担当国連事務次長

事務局調整・支援

- Quintin Chou-Lambert、国連デジタル・新興技術室（UN ODET）
- Rebakah Hayoung Woo、UN ODET
- Peppi Väänänen、UN ODET

資金調達・ロジスティクス

- Sebastian Frank、UN ODET
- Antonieta Loaiza、UN ODET

広報

- Karoline Hassfurter、UN ODET
- Brian Shung Seun Lau、UN ODET
- Anamika Madhuraj、UN ODET

編集・起草支援*

- Jiaee Cheong、国連大学（UNU）
- Kevin Kohler、UNU
- Max Springer、UNU

ラポルトゥール（報告者）

- Wernhard Berger、国連工業開発機関（UNIDO）
- Jin Cui、国際電気通信連合（ITU）
- Tim Engelhardt、国連人権高等弁務官事務所（UN OHCHR）
- Andrew Morritt、国連平和活動局
- Prateek Sibal、国連教育科学文化機関（UNESCO）
- Mariagrazia Squicciarini、UNESCO
- Oleksandra Vereschak、UNESCO
- Ana Gabriela Fernandez Vergara、ITU
- Li Zhou、UN OHCHR

---

* パネルの作業の科学的独立性を確保するため、編集・起草支援の人員は、成果文書の実質的な内容・文言に関してはパネルに報告する一方、パネル事務局の一部として、調整上の要請、スケジュール、作成されたコンテンツの共有には従う。管理上の目的については、国連大学に報告する。

以下の国連事務局の組織も、本報告書の作成においてパネル事務局を支援した。総会・会議管理局、グローバル・コミュニケーション局、及び情報通信技術局（OICT）の国連地理空間（United Nations Geospatial）。

Independent International Scientific Panel on Artificial Intelligence