資料３

AIガバナンスに関する議論の方向性について（ディスカッションペーパー）

令和５年４月

内閣府

科学技術・イノベーション推進事務局

背景（今回の検討の必要性）

「人間中心のAI社会原則」策定の当初から、AIガバナンスに対する国際的変化や技術動向の変化が生じており、人間中心のAI社会原則の日本の立場を国際的に明示し、AI社会原則をもとに社会実装の推進を行う。

AIガバナンスについて国際的に立場（方針）を明確にする。

AI規制などを含むガバナンスに関する国際動向に応じた今後の日本のAIガバナンスの方向性を提案する。

→ポリシーペーパーを国際的議論の場において利用することを想定

論点

（現状認識）

- ＡＩに関するリスクは顕在化しているが、ＡＩの実装の程度や、作業や行為の中でＡＩが果たす役割については、利用される分野やそれぞれの利用シーンごとに大きく異なる
- ＡＩに関連する技術は発展の途上にあり、その変化も著しい
- ＡＩ規制に関する議論が各国で進んでいる

ＡＩに関するリスクに対して、何らかの対処（リスクの回避又は軽減）が必要

（論点）

① ＡＩに関する、どのようなリスクに対処するべきか。

② リスクへの対処にあたっては、どのようなことを踏まえるべきか。

A） イノベーションの確保？

B） 社会的負担（コスト）の極小化？

C） 流動的な状況に対処するための柔軟性？

①ＡＩに関するリスクについて

AIの信頼性

- AIの出した結論が利用者には判らない。
- AIの導入効果についての期待は大きい。

↓

- 信頼性・説明性が適切に把握されず、実態とのギャップが生じる。

↓

- 社会的安定性を損なう。
  （バイアスの拡大、「もっともらしい嘘」(注)の流布など）

悪用・誤用

- 画像生成、音声合成など、AIにより人間の認知を欺瞞することが容易になっている。

↓

- 巧妙な詐欺や、政治等に混乱をもたらすフェイクなど、悪用のリスクが高まっている。

- AIの技術的進展が著しい一方で、人々の知識やスキルには大きな格差が生じている。

↓

- 誤用のリスクが高まっている。

(注)「ファウンデーション・モデルでは、学習での誤りを完全に排除することは難しく、説明性も欠如している。このため、もっともらしい嘘をつく、不適切な情報を示す」などの意見がある。

普及に伴う影響の増大

- AIは急速に普及し、社会経済の基盤を形成する一つの要素となりつつある。
- 人々の日常生活にも欠かせないものとなることが見込まれている。
- 利用者の能力差（AIを使える人、使えない人の差）やAIによる個人信頼性の格差が生み出される。

↓

- 技術的問題やその他の要因により、AIの機能が損なわれること（AIへのアクセス自体ができなくなるなど含む）による影響範囲は拡大している。
- 医療や法務などの利用分野においては、AIに起因する問題が不可逆的な影響（人命や人権など）に直結しうるなどの深刻なリスクも潜在する。

リスクに関する関連事項

（会議における議論から）

- 近い将来、アルゴリズムとデータの区別がつきにくくなる実態を踏まえると、バイアス等の考慮にあたって変化が生じるのではないか。
- AIが広範に学習を行うようになっている。学習に使用したデータに関する権利の問題などで、AIを利用したビジネスの不確定性がましているのではないか。
- AIの利用の進展に伴い、一部には就職活動でAIによる面接が行われるなどの事例も出てきている。AIに関する法令の適用等の動向に注意を払うべきではないか。
- AIがダーク・パターン（より高額な契約をさせるなど、人の判断をゆがめさせるように作り込まれたユーザ・インタフェース）に利用されることが危惧されている。こうした問題に関して、AIについて専門的知見のある側からの情報発信をしていくべきではないか。

（国際動向や関係事例から）

- AIの信頼性：バイアスによる判定基準の歪み、不適切性による問題事例。
  （システムの学習中における人種、学歴、性別、行動履歴などのデータの差がバイアスを作り出す事例）

- 悪用：ディープフェイクの悪用が個人、社会レベルでの問題を引き起こしている。
  →政治家、著名人のなりすましによる偽情報の流布（YouTubeにおけるウクライナ大統領のディープフェイク動画））

- 誤用：データリーケージによる誤った結果の導出
  （AIによる特定事象の予測において学習時のデータとテスト時（予測時）のデータに重複があった。）

②リスク対処の方向性について

A) イノベーションの確保

- Society 5.0の実現に向けては、研究開発と利活用のサイクルを通じ、望ましい発展が加速されることが望ましい。（人間中心のAI社会原則の7原則の一つ）

⇩

- プライバシーやセキュリティの確保は前提としつつ、研究開発や社会実装を促進するためには、オープンでイノベーション促進的な環境を維持することが重要。

B) リスク対処のコスト（利便性の低下なども含む）

- リスクに応じた対処は、一般にコストを伴うものである。

⇩

- リスクに比べて過剰となる規制等の対処を避けることが重要。

C) 流動性への対応

対象の曖昧さ

- AIの定義は定まっておらず、規制をすり抜ける。

⇩

- すり抜け回避のためには、個別に定義づけることが必要。だが、それがAIの規制たりうるのか。

⇩

- AIの発展は途上であるとの視点にたち、対処を検討することが必要。

対処の柔軟化

- AIに関するリスクの状況は流動的である。
  →将来を見越した対処は困難
  →分野横断的となる一律の対処は困難
  →固定的な対処は困難

⇩

- 柔軟な対処のためには、分野ごとの非規制的手法を効果的に活用することが重要。

リスク対処に関する関連事項

実装される分野等における差違

- AIは、社会経済や国民生活の多様な場面に実装されていく。
- 導入されるAIに期待される機能や社会的役割（人間との役割分担など）における共通項は限定的である。

→人間とAIの関係は、実装される分野当に多様となる。

医療や法務をはじめ比較的重い規範が求められる分野などでは、固有の資格・規制や業務の特色に応じた既存の法制がある。

各分野の法令において、それぞれのリスクを踏まえつつ、AIの利用を確保していく手法が効率的となる。

（会議における議論から）

- 現状においては、欧州のようなAIに対する包括的な規制のアプローチは、そのためのコストの観点から適さないのではないか。AIが利用される場面によっての差違もあるので、規制には強弱をつけるべきではないか。
- AIの利用に際してリスク回避の選択肢としてオプト・アウトを導入する場合に、AIに関する製品やサービスを提供する側の負担が増加することも当然にあり、AIの利用の支障となることも危惧される。そうした手段の導入に当たっては、一律とするのではなく、適切な程度の見極めが必要になるのではないか。
- AIの規制に関して、リスク・ベースのアプローチをとる場合、AIへの依存度や、法的な意味合いなどの複数のメトリクスで議論することが適切ではないか。
- AIの利用では、データのクレンジングやバイアスの除去など、さまざまなチューニングを経て実用化のレベルを上げていくことが多い。このため、企業等においては運用状況の評価やリスク分析に応じて、ガバナンスで目指すべきゴールや、そのためのマネジメントのデザインをも柔軟に変更するガバナンス（アジャイル・ガバナンス）を導入することが適切ではないか。

（参考）その他

（会議における議論から）

- AIの判断により不利益を被った人が異議を表明できるよう、AIの判断についての情報（ロジックなどの要素など）をあらかじめ知らせる仕組みが必要ではないか。
- AIの利用に関して差別を排除しようとする場合、結果の平等ではなく、機会の平等をめざすべきではないか。
- 医療などでのAIの利用が進みつつある中で、そうした恩恵に浴せない方が生じてしまうことも予想される。また、ChatGPT（AIの開発や普及を行う米国の非営利団体OpenAIが開発したチャットボット）などのようなAIはある種の基礎インフラになる可能性もある。AIに対する機会の公平性を担保するためには、こうしたAIに対するアクセス権やリテラシーなどについて議論が必要ではないか。
- GPAIの成果を踏まえた今後の取組においては、特にレジリエンスに重点をおくべきではないか。

まとめ

日本のAIガバナンスについての立場案

AIは様々な側面で発展中であり、AIが生み出す多様なリスクへの対応とイノベーションの推進を両立させるためのバランスが必要（？）である。

（参考）

・AIの技術的進展

・AIの最近の動向

・基礎技術：GANsについて

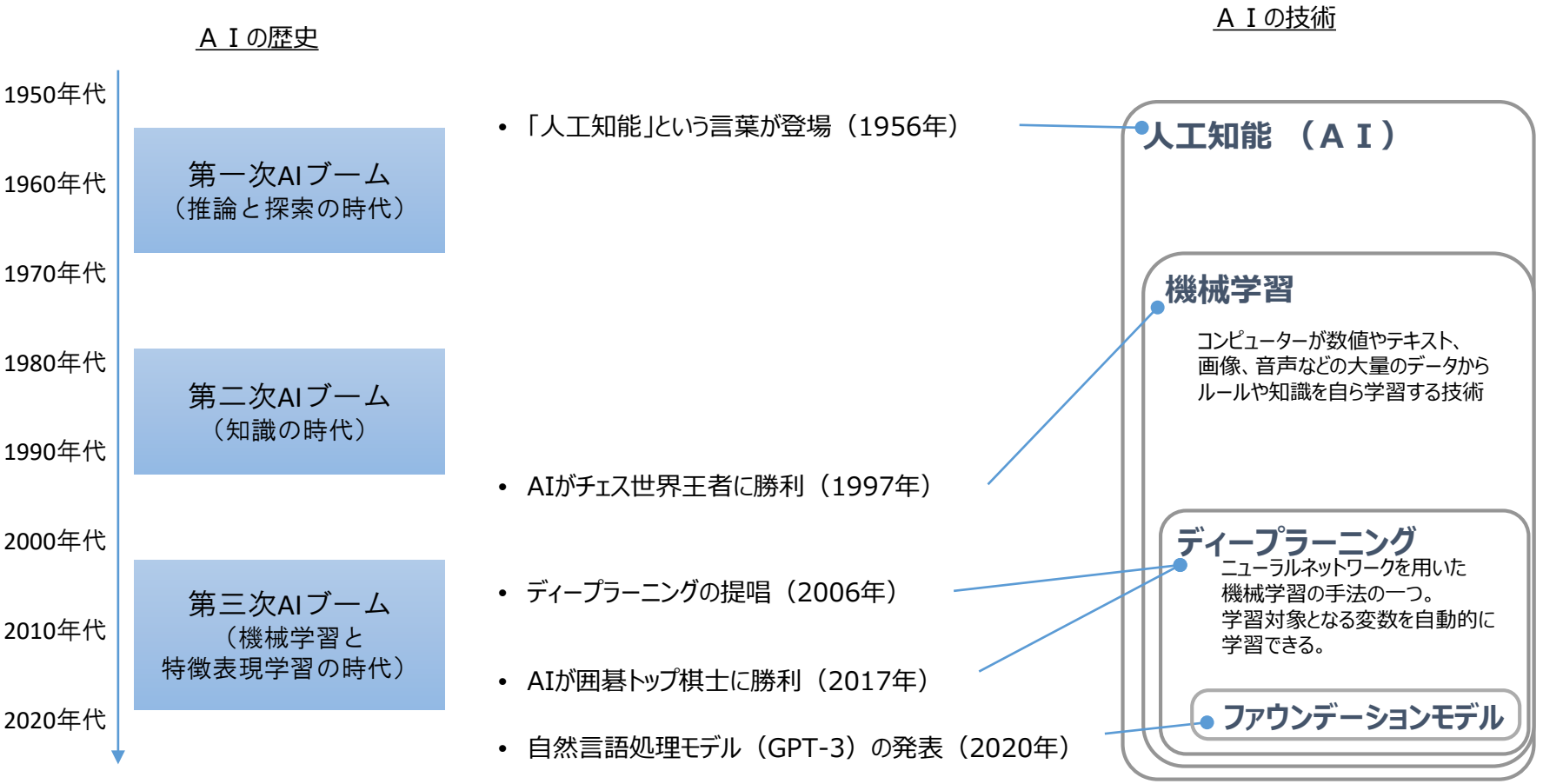
・AIの倫理・ガバナンス等をめぐる状況

・公開書簡「巨大AI実験一時停止」について

・人間中心のAI社会原則

（参考）ＡＩの技術的進展

- 人工知能（AI）の研究は1950年代から開始。
- 2000年代には**ビッグデータ**を活用しAI自身が知識を獲得する**機械学習**が一般化。
- 2010年代以降は実用性の高い**ディープラーニング**が主流となっている。

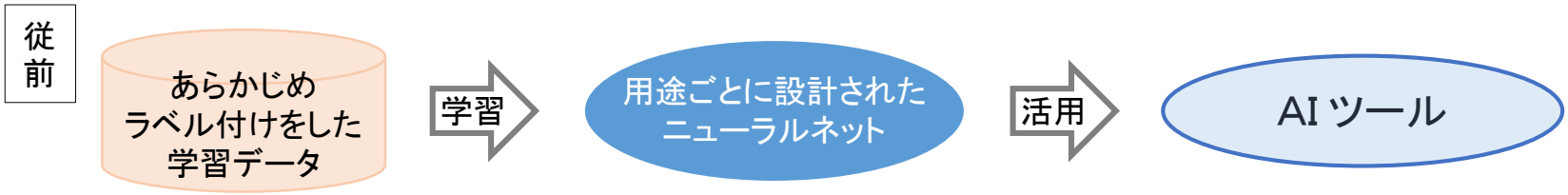


（「平成28年版　情報通信白書」等を参照して作成）

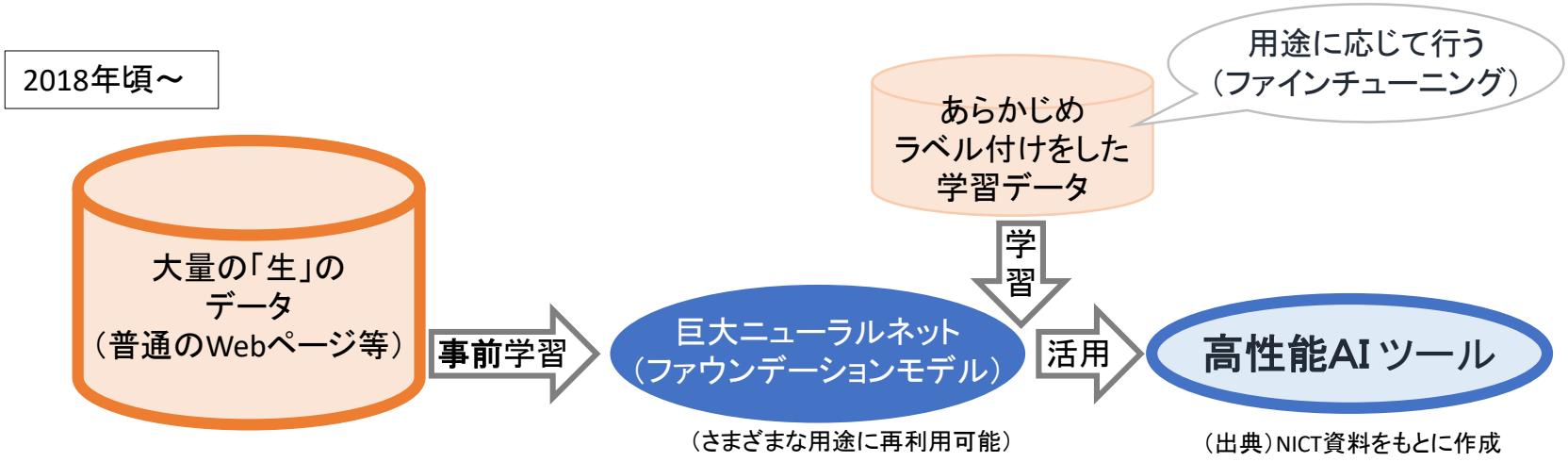
（参考） AIの最近の動向

ファウンデーションモデル（基盤モデル）の登場

- ＡＩでは、一般に あらかじめラベル付けをした学習データ により、ニューラルネットの学習が行われる。



- 近年、ラベル付けをしない 大量の「生」のデータ で事前学習を行い、
その後に用途に応じた学習を行うことで、より高性能のＡＩツールを実現する手法が普及。
- 生データのサイズとニューラルネットのサイズに応じて、必要な計算資源や学習時間は増加するが、
巨大言語モデル（BERT、GPTなど）での格段の進化につながっている。



（出典）NICT資料をもとに作成

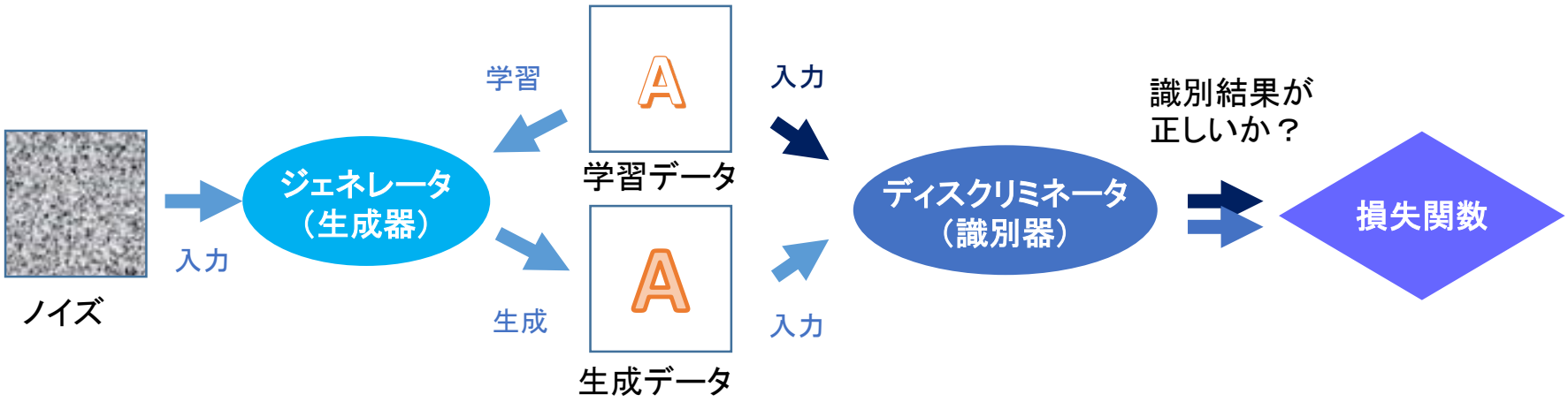
（参考） 基礎技術：GANsについて

GANs (Generative Adversarial Networks) 「敵対的生成ネットワーク」

生成器と識別器という対立する2つのニューラルネットワークを競わせながら学習させることで、高精度のデータを生成可能になるシステム（主に画像生成で利用）

生成器：学習データにそっくりなデータを生成して出力

識別器：入力されたデータが、元の学習データか偽の生成データかを識別した予測を出力



識別結果を損失関数で評価 → フィードバックから両者が再学習

生成器：見破られると罰則の損失が発生 → 識別器に見破られないデータを生成するように学習

識別器：識別を誤ると罰則の損失が発生 → 生成器が生成したデータを偽物と見破るように学習

（参考） AIの倫理・ガバナンス等をめぐる状況①

- ＡＩについては、**我が国が非規制的・非拘束的な方針を踏襲**しているのに対して、**欧州はＡＩのリスクに応じた規制の導入に積極的**。

日本

- 2016年4月　Ｇ７情報通信大臣会合において、ＡＩの開発原則の検討を提案。ＡＩの倫理やガバナンスに関する国際的な議論のきっかけとなる。
- 2019年3月　世界に先駆けて「人間中心のＡＩ社会原則」を策定。
  ＡＩの開発者及び事業者については、非規制的で非拘束的の方針が提言される。
- 以後、国際的議論において、日本政府は、**非規制・非拘束といった”Soft Law”（ソフト・ロー）の方針**を踏襲。

非規制的、非拘束的、ソフト・ロー ⇔ 規制的

欧州委員会

- 2021年4月　ＡＩのリスクへの対処、ＡＩの導入やイノベーションの強化に対する法案を提案。
  このうちリスクの対処に関しては、**ＡＩを４段階に区分**し…
  - ①**許容できないリスクのあるもの**（サブリミナルな技法、公的機関による社会的スコアリング等）　⇒**禁止**
  - ②**ハイリスクなもの**（重要インフラの管理、教育・職業訓練での利用等）　⇒**規制**（品質管理など）
  - ③**限定的なリスクがあるもの**（人と対話するAI等）　⇒**透明性の確保**（ＡＩであることを明示など）
  - ④**最小リスクのもの**　⇒**非規制**
- 上記法案の導入スケジュールについて、従前は「2022年後半の発効、2024年後半の完全施行」との公算であったものの、企業等から懸念する意見もあり、先行きは不透明。

英国

- 2022年7月　ＡＩ規制に関する政策文書を公表。ＡＩの開発・普及において社会的信頼を促進しつつ、イノベーションを志向した**リスク・ベースでの規制導入アプローチ**を示している。
- 具体的には、**均一性の低いコンテクスト・ドリブン（context-driven）な手法**、すなわち特定のコンテクストに着目するアプローチの導入が提案されている。

（参考） AIの倫理・ガバナンス等をめぐる状況②

欧州評議会

日本、米国、カナダ等はオブザーバーとして参加

- 欧州評議会では、既存の国際条約が存在しない分野で、多くの多国間条約を作成している。これまでの代表的な成果は、欧州人権条約、サイバー犯罪条約など。
- 2022年4月　**ＡＩに関する法的枠組（条約）の策定・合意**を目指し、検討を開始。
- 2022年11月時点では、**基本原則**や、**リスクやインパクトについての評価**を含む内容が議論されている。（具体的案文は非公表。日本からも総務省職員が検討に参加しており、拘束的な規定を設けることについての懸念などを伝えているが、疑義を提起する意見は少数派）

UNESCO

- 2021年11月　**ＡＩの倫理に関する勧告を採択**。ＡＩに関係するすべての関係者に尊重されるべき事項として、
  人間の尊厳、人権及び基本的自由の尊重　などの**価値**
  安全・安心、プライバシーとデータの保護　などの**原則**
  を示し、加盟国が措置すべき倫理的影響評価等の**政策措置**を勧告。
- ただし、**上記勧告の適用は、加盟国の自主的判断**（on a voluntary basis）とされている。

OECD

- ＡＩに関する取組の情報共有のためのオンラインプラットフォームとして「ＡＩ政策に関するオブザーバトリー」（OECD.AI）のほか、助言を行う非公式専門家会合として OECD Network of Experts on AI（ONE AI）を立ち上げている。

（参考）公開書簡「巨大AI実験一時停止」について

非常に高度な知能を持つAIシステムについて、現在の体制での開発競争のリスクを警告し、その対策を講じるために、すべてのAI研究所に対する**GPT-4（1）よりも強力なAIシステムの開発の一時停止**と、AIシステムの**安全性を保証する技術開発および制度整備**等を訴えたもの。

2023年3月 Future of life Institute（2）がインターネット上に公開（3）。

署名者: イーロン・マスク氏（Tesla社共同創業者）、スティーブ・ウォズニアック氏(Apple社共同創業者）他
産業界の重鎮や著名なAI研究者　など多数

本文の主な内容

- 人類と競合する高度な知能を有するAIシステムは、大きな変化をもたらす半面深刻なリスクを及ぼす可能性があり、見合ったケアとリソースを用いて人類により計画・管理されるべきであるが、現在そのようなレベルの計画・管理が行われないままAI研究所間の制御不能な開発競争に陥っているように見える。
- 現在AIシステムは一般的なタスクで人間と競合しつつある。**強力なAIシステムは、その効果が好ましいものであり、そのリスクが管理可能であると確信した場合にのみ開発されるべき**である。
- よって、**すべてのAI研究所に対し、GPT-4よりも強力なAIシステムの訓練を少なくとも6ヶ月間直ちに一時停止**するよう求める。この一時停止を利用して、AI研究所と独立した専門家により**厳格に監査および監督される高度なAI設計・開発のための一連の共有安全プロトコルの開発・実装**を進めるべきである。 並行して、堅牢なAIガバナンスシステム（技術開発や制度整備を含む）の開発を加速するためAI開発者は政策立案者に協力する必要がある。

※世間の反応

- 2023年3月下旬に数多くのメディアに取り上げられ、専門家からの様々な意見が報じられている。

（1） OpenAI 社が2023年3月に公開を開始した、最新の大規模言語モデル。
（2） 2014年米国で設立されたAI等の先端技術の倫理的取組を行う非営利団体。2017年アシロマAI原則を策定。
（3） https://futureoflife.org/open-letter/pause-giant-ai-experiments/

（参考）「人間中心のAI社会原則」

- ＡＩをより良い形で社会実装し共有するための基本原則についてとりまとめたもの。
- 「人間の尊厳」、「多様性・包摂性」、「持続可能性」を基本的な理念として尊重し、その実現のため７つの原則を定めている。