人工知能と社会（仮訳）

エグゼクティブ・サマリーと提言

AI（人工知能）は、我々の社会と日常生活のさまざまな側面に大きな変革をもたらす技術である。AI はすでに数多くの恩恵をもたらしており、今後、さらに大きな経済的繁栄をもたらすと考えられる。しかし、一方では、雇用不安、データの機密性、プライバシー、倫理的価値の侵害、結果の信頼性欠如など、AI に対する懸念も顕在化している。これらの AI にまつわる懸念に対処するために、政策立案者と科学者は共に以下のことを確認し、実行すべきである。

■ **社会のあらゆる階層、人々が AI の恩恵を共有できるように、責任をもって導いていくこと**

このためには、AI が雇用に及ぼすインパクトに細心の注意を払っていく必要がある。この雇用へのインパクトは、AI の技術的進展だけでなく、政治的・政策的な要因、経済的な要因、文化的な要因など、多彩な要因によって決まっていく。

■ **信頼できる AI システムとデータ**

信頼は、データの品質、データのバイアス、データのトレーサビリティについての措置を講ずることで獲得できる。AI・データへの信頼は、データへのアクセス容易性を上げることで向上するが、一方で、このアクセス容易性が、アクセス権限のない第3者への個人データの開示につながらないようにしなければならない。

■ **安心・安全な AI システムとデータ**

人の脆弱性に関係する応用[1]の場合には、安心・安全は不可欠な要件であり、場合によっては、システムの正しさが明示的に証明できることも必要となろう。

■ **説明可能な AI システムの開発を促す研究の推進**

人々に大きな影響がある重大な判断を AI が下す場合には、影響を受ける関係者に十分な情報が供与され、AI の判断に異議を申し立てたり、従わなかったりする自由が保障されなければならない。（例：提示された治療や処置の拒否、判断に対する不服申し立て）

■ **AI の社会的恩恵を最大化するための他分野からの洞察**

学際的な研究は、AI 自体の研究にとどまらず、自然科学、生命・医学、工学、ロボット工学、人文科学、経済・社会科学、倫理学、コンピュータ科学といった非常に多様な分野を巻き込まなければならない。

---

[1] 例として、医療分野が挙げられる。

■ **一般市民の AI の受容性(AI-Readiness)を高めること**

AI に関わる多様な教育の機会や情報を市民に提供していくとともに、AI に対する誤った神話を正していくために、一般市民との対話、科学的に健全でよく準備された対話を重ねていくことが不可欠である。

■ **AI の破壊的・軍事的利用について、開かれた政策論議を推進すること**

人間の制御を必要としない、自律性の高い兵器が持つ危険性を抑制する国際的な取り組みを、国連のしかるべき組織が進めていくべきである。

■ **官民のリサーチ・セクター間の人材交流と共同**

民間と公的な研究機関との間の人材交流や共同は、人々に大きな恩恵をもたらす領域において、安全かつ迅速な AI 応用の社会への展開を促す。また、AI システム開発に不可欠な大量データの収集において、官と民との共同が重要である。

序文

AI とは、コンピュータやその他の装置を知的に機能させることを目的とする手法や技術を指す。AI は、基本的には（通常は大量の）データを使って作動するアルゴリズムの集合体である。この AI を構成する一分野に、機械学習（ML）があるが、これは、複雑なデータから有用な情報を取り出すアルゴリズムに関する分野である。近年、機械学習は、多くの科学や技術の分野に予想以上に大きなインパクトを及ぼしてきている。AI 研究は今後も着実に進歩を遂げていくであろうということ、また、AI が社会に及ぼすインパクトも今後より一層高まっていくということが、一般的な共通の認識となっている。

利用可能なデータの増大と計算機の処理能力の向上と結びつくことで、高度なアルゴリズムを使ったシステムの開発が、「音声認識」、「画像分類」、「異常検出」、「自律走行車」、「意思決定支援システム」、「ロボット」、「機械翻訳」、「脚を使った移動」、「質問応答システム」など、特定の目的に特化したさまざまなタスクで目覚ましい成功をもたらしている。現在、こうした AI 応用のいくつかは、障害をもった人たちに極めて有用な支援ツールを提供しつつある。例えば、麻痺を患った人が、ブレイン・マシン・インターフェース（BMI）を使うことで、コンピュータを介して周囲の環境とやり取りすることが可能になっている。

自然科学や社会科学の分野においても、機械学習のアルゴリズムは、複雑なデータやプロセスの取り扱い、そのモデル化に新たな進歩と新たなツールをもたらしている。このような進展は、将来非常に大きな恩恵をもたらす。文明がもたらす恩恵の大部分が「ヒトの知能」が作り出

したものであることを考えると、「ヒトの知能」が AI ツールによってさらに一層強化されたとしたら、一体どんなことが成し遂げられるようになるのか、今は想像をめぐらすだけである。

しかしながら、同時に、このようなシナリオに潜む落とし穴や危険について、慎重な検討が必要な疑問や懸念も数多く存在する。

AI 研究のこれまでの急速な進展から、現在は、単に AI の能力向上への努力だけでなく、倫理的な価値を尊重しつつ、社会への恩恵を最大化することに注意を向けることが必要な時期となっている。

AI の社会実装や技術開発は、倫理上の考察から、適切な指針を得るべきである。データの統計的な分析や機械学習に基づく AI システムが、結果として、望ましくないバイアスを持ってしまうことへの懸念も、増大しつつある。

こうした全般的な文脈において、我々は、まず、AI が経済に及ぼす革命的なインパクトと、それが引き起こす課題について議論する。次いで、AI システムが、うまくかつ倫理的にも正しく人間と交わり、やり取りするうえで、備えるべき一般的な性質を検証する。さらに、ヘルスケア分野での AI システムの利用における具体的な課題、自律性の高い兵器システムへの AI 応用が引き起こす課題、ロボット・システムに埋め込まれる AI の潜在的な可能性、を考察する。このような分析が、エグゼクティブ・サマリーに示した一連の提言にまとめられている。

1. 社会に及ぼす AI のインパクトの管理と最適化

経済学者とコンピュータ科学者は、AI の経済的恩恵を最大化させるとともに、それがもたらす悪影響を軽減する研究が今必要だという一般的な見解で一致している。現時点で、格差の拡大、失業率や非倫理的行動の増大という観点から、AI がもたらすインパクトを検証することが重要である。これらの未解決な課題について、以下で詳細に考察する。

1．1 労働市場予測

AI は、非常に大きな経済的な恩恵をもたらすことができる。すなわち、AI 技術は、さまざまな産業分野・セクターにわたって、生産性を向上させ、新たな製品やサービスを作り出す可能性を持っている。しかし、この可能性は、一方で、AI が雇用や職業人生にもたらすインパクトについての疑義にもつながっている。

AI は、雇用にかなり破壊的な影響を及ぼす。消失する雇用もあれば、新たに生まれる雇用もあり、また、大きく変化する雇用もあると見込まれる。雇用に及ぼす AI のインパクトを予測する様々な研究には、変化のスピード、自動化される可能性が高いと考えられるタスクや仕事が

どの程度あるかに関して、大きな開きがみられる。

長期的な観点からは、技術は人々全体の生産性や富の増大に寄与する。しかし、こうした恩恵が実現するまでには時間がかかるために、移行期では、一部の人たちが不利益を被る期間が生じ得る。すなわち、一部の人たちや地域に混乱を生じさせ、短期的には社会的な格差を拡大するという、移行期の影響があることになろう。したがって、自動化の影響を受けやすい仕事のことを勘案し、この不均衡が経済や社会に及ぼすインパクトを予測する研究が必要なことは明らかである。様々な仕事に対する AI システムのインパクト、たとえば、特別な技能を必要としない仕事へのインパクト、あるいは、高度な訓練を受けたプロフェッショナルを必要とする仕事へのインパクトといったことの分析は、いろいろな政策のもとで未来に新たに生まれてくる仕事を予測することに比較すると、簡単であろう。AI 技術が将来どのように発展していくかについては、いくつもの可能な道筋が考えられる。AI が雇用に及ぼすインパクトも、AI 技術の能力という要因だけではなく、政治的、経済的、文化的な要因という、多様な要因によって左右される。様々な学問分野から得られる最良の知見を持ち寄ることは、技術がもたらす変革の恩恵を社会全体で共有する政策を策定する上での大きな助けとなる。

１．２ 社会における AI 発展の管理と統合にあたっての政策

AI は社会の多様なセクターに大きなインパクトを及ぼし、人間の仕事を助けたり、または、代替したりするであろう。こうした変化を予測し、AI のマイナスの影響を少なくし、AI を社会にうまく統合していくことを可能にする政策を策定していくことが、大きな課題である。このためには、教育が、AI 導入の推進とそれによる格差の防止の双方において、鍵となる。

データと AI 技術の利用についての基本的な理解が、AI 技術を作り出す側や AI 利用の専門家のみならず、あらゆる年齢層の市民すべてにとって不可欠となる。学校教育の場で基本的な概念について教えることが、この基本的な理解を確かなものにする。科学、数学、コンピューティング、芸術、人文科学といった分野での若年層教育において、広範でバランスの取れたカリキュラムを導入することで、生徒らに必要となる様々なスキルや生涯にわたる学習の基盤を与えることができる。

また、高度なスキルを持った人材への需要は、非常に高い。様々なセクターと職業において、AI を効果的に使いこなすスキルが求められるだろう。これに向けた新たなイニシアチブをとることで、AI システムを使いこなすのに十分な知識を持った利用者集団を作り出すことが必要である。さらに、新たな教育コースや AI のためのインフラ作りへの支援が、AI 分野での高度なスキル、すなわち、雇用を生み出す新たな応用に結び付く AI 分野での高度なスキルを作り出していくのに必要となる。

こうした課題はすでに、前回の G7 サミットにおけるオタワ宣言「Realizing our digital future and shaping its impact on knowledge, industry, and the workforce（デジタル・フューチャー～デジタル化による社会変革の実現と情報・知識、産業、労働・雇用への影響の展望について～）」の一部となっている。政府には、包括的で、かつ、1 人 1 人の市民が AI の恩恵を公平に享受できるような政策を実施していくことが奨励されている。これには、「情報の質」、「セキュリティ」、「レジリエンスの保証」に加え、AI システムの「透明性」、「開放性」、「相互運用性」も求められる。

AI の能力が現行の規制が追いつけないスピードで発展している領域においては、人間と知的な機械とのかかわりが引き起こす倫理上の問題を考慮に入れた、新たなガバナンスへのアプローチが必要だと思われる。どのような形で AI が既存の倫理上の規範を犯す可能性があるか、あるいは、どのような場面で AI があらたな倫理上の問題を引き起こすかを考察する上で、人文科学や社会科学が果たす役割、これらの科学が AI の研究者や利用者との協調の中で果たす役割、の重要さは強調するに値する。

2. 奨励されるべき AI システムの特長

2. 1. データ

AI とビッグデータとの組み合わせの相乗効果をフルに活用する能力は、データの取得、データの査定および管理に係る能力に左右される。現行の AI 技術の大部分は、膨大なデータへのアクセスを必要とする。したがって、技術を存分に駆使するためには、データを利用に供するための新たな枠組みが必要だと考えられる。このような枠組みは、オープンデータ、あるいは、公益に資するプライベートデータにとって重要であり、その枠組みの中で、データが有益な形で使われることを保証する新たな基準といったものが用意される必要があろう。例えば、データの意味を明確にし、データがどのような文脈で獲得されたものか、データの出所やそれに施された処理についての情報を付与する努力が必要となるだろう。こうしたことを AI 技術すべてに渡って取り組むことが、オープンデータの考え方がもたらすとした恩恵を真に実現するために、また、社会、経済、組織、個々の技術といった壁を越えた相互運用性を提供するために、重要となる。

同時に、高品質のデータセットへのアクセスにおいては、個人データのプライバシーと機密性を尊重すること、不当なバイアスや個人の権利侵害についての懸念を払しょくすること、が必要である。銀行、保険会社、将来の雇用主といった第三者による個人の機密データへのアクセスが適切な規制に従うように、最大限の努力を払う必要がある。また、悪意ある攻撃からデータセットを守ることも必要である。大企業に対してだけでなく、オープンソースのイニシアチブに対しても、データ収集とその共有、アクセスの仕方に関する適切な指針や政策が用意されなければならない。

2．2 AI技術の性能と説明可能性

最も成功を収めていて評価の高いAI技術の一部、特に深層学習は、現在のところ、説明可能性の低さに悩まされている。また、さまざまな AI の手法は、それぞれの手法に依存した、違った説明可能性を提供している。このことが、利用者の AI ツールに対する信頼を減損する場合がある。特定の領域でのAI応用では、説明できることが不可欠の要件となっている。例えば、医学への AI 応用では、説明のない、天下り的な診断は受け入れられないだろう。より高い説明可能性を持ったモデルを開発すると同時に、性能と説明可能性との間にみられるトレードオフを明確にすべきであろう。AIシステムが下した判断の背後にある理由を利用者が理解できるようにするためには、使われているアルゴリズムが持つ限界が明示的に記述されている必要がある。AI の説明可能性の向上は、AI システムが望ましくないバイアスを混入させないようにする上で重要である。「差別的効果」という概念が、主として法理論上の概念として使われはじめている。これは、民族、出身階層、ジェンダー、年齢といった個人の属性が、AI アルゴリズムの下す判断に直接影響を及ぼし、そういったアルゴリズムの適用が生み出す予期せぬ差別のことを指している。人々の日常生活に深い影響をもつような判断を下すのに使われる AI システムは、この望ましくない「差別効果」を生み出すことがあってはならない。

2．3 オンラインで進化するシステムの検証と正当性の確認

オンラインで進化するシステムというのは、運用中に絶え間なく遭遇するデータを使って、時間とともに進化していくシステムをいう。こういった AI システムは、運用初期の状態から徐々に遠のいていって、その結果、たとえば、ジェンダーや人種について、“望ましくない”方向に進化してしまうことがある。従って、このようなオンラインで進化していくシステムでは、出力を常に監視していて、望ましくない方向への進化を検知することが必要となる。

3．3．　いくつかの適用領域の例と社会的な影響

3．1 ヘルスケアへの適用

健康と医療における意思決定支援のシステムにおいて、AI は極めて大きな恩恵をもたらす可能性をもっている。この分野のいろいろな構造的な難しさから、「診断の誤り」、「専門的な知識・技能の不足」、「研究者、工学者、臨床専門家の間での意思疎通の不足」等の問題を引き起こす可能性がある。しかしながら、同時に、AI は、膨大な量の研究論文の有効な活用を助けたり、膨大なデータを使うことで事前には想定できなかったデータ間の相関関係や微弱な相関関係を見つけたり、ヘルスケア・システムが生み出す画像やデータを解析したり、新たな医療技術を開発したりすることに、貢献できる。臨床的な意思決定支援システムをより良いものにすることは極めて重要であり、AI 技術は、さまざまなツールや装置の開発に寄与することで、診断や治療法の選択などでの医師の意思決定を補完・支援することに大きく貢献できる。ここ

での目標は、所見や測定データを解釈する過程、診断を下す過程をよりよいものにすること、また、より正確で効果的、かつ、多くの人に利用可能なヘルスケアの実現に寄与することである。このようなことの実現には、細心の注意を払ってシステムを設計する必要がある。すなわち、AI とその利用者との協調のあり方、さまざまな場面で異なったものが要求される説明可能性への対処、システムの正しさの証明や正当性の確認の方法などへの配慮、が必要となる。医師と患者がそうしたシステムを信頼できること、および、タイプの異なる利用者のさまざまなグループに対して、システムが効果的に動作することが重要となる。

さらに、細心のデータ・ガバナンスも不可欠である。国際的な連携、AI を使って医療の進展を加速化させる国際的な連携は、すべての国々の市民の利益に叶うものである。

3．2 ロボット兵器

AI は、軍事用途、とりわけ、標的の選定とそれへの攻撃というクリティカルな機能を持った自律的な兵器システムの開発を可能にする。そうしたロボット兵器は、新たな兵器の開発競争を引き起こし、戦争への敷居を低くし、迫害者やテロリストの道具と化する恐れがある。このようなことから、一部の組織は、化学・生物兵器の分野での協定と同じく、ロボット兵器に関する協定の締結を求めている。こうした禁止にあたっては、兵器や自律性の意味を正確に定義することが不可欠になる。自律型致死兵器システム（Lethal Autonomous Weapons Systems – LAWS）の禁止の合意がない段階では、いかなる兵器システムも国際人道法を遵守することが守られなければならない。こうした兵器が、既存の指揮統制体制に組み込まれる場合には、その使用の責任と法的な説明責任をある特定の人物に帰着するようにしておく必要がある。また、この領域で提起される課題について、透明性の高い、公開の議論を進めていくことが明らかに必要となっている。

3．3 ロボット工学

ロボットは、外界からのセンシングと外界での移動・動作の機能を持った、AI を具現化した機械である。機械が人間を含めた外部環境と直接的・物理的なつながりを持つことは、大きな挑戦的課題である。ロボットは、安全で信頼でき、かつ、安心できるものでなければならない。最近になるまで、ロボットは主に製造業で使われていが、これらのロボットは、特定の環境に閉じ込められていて、人間と同じ活動空間を共有することはなかった。現在、第 2 次のロボット研究の波が到来しているが、ここでは、ロボットはこれまで以上に、人間と同一の空間を共有し、人間とやり取りするようになってきている。AI の適用範囲は、これまでのところ、データを処理することで判断や意思決定のための知識を取り出すことに力を注いでいたのに対して、ロボット研究の最終的な目標は、物理的な実社会に直接かかわる能力を備えた技術システムを作り出すことにある。

機械学習のアルゴリズムの利用に加え、ロボット研究は、物理的な安全性という根源的な制約に取り組まなければならない。ロボットの設計には、故障に対する耐性、信頼性、存在し続ける能力といったものを最大限確保するために、ソフトウェアの保証や正当性の論理的証明が必要になる。

近年の様々な進展があるとはいえ、進歩への期待から、技術変革の速度を過大評価する傾向がしばしばみられる。

最後に、ロボット、あるいは、より広く AI 全般について一般の人々がもつイメージ は、科学的な確かな証拠ではなく、むしろ空想的な物語の影響を強く受けている。今後は、公教育、すべての市民を巻き込んだ議論や討論に積極的に関与していくことで、ロボットや AI 科学についての神話をなくし、正確な情報発信を行っていくことが重要である。