人工知能への道：AI リスクを超えて

2024年11月 27日 | 岡島義憲の集積回路の明日に向けて
（情報統合技術研究合同会社）

人工知能への道： AI リスクを超えて

昨今の生成 AI ブームは AI リスクへの懸念も巻き起こしているが、それらに関する多くの議論を尻目に、北米では、Digital-Transformation（DX）を進める企業や行政組織の業務フローの全体をサポートする AI 内蔵プラットフォームの標準システムの開発が進んでいる。 今後、そのプラットフォームの利便性とコストダウン効果が確認されると、市場の力と首長選挙や議会選挙のダイナミズムによって、各組織の再編が始まり、産業界や行政組織に新たな水平分業システムが現れるのではないだろうか？

MIT による AI リスクのサーベイ

本年 8 月、米国 MIT（マサチューセッツ工科大学）のコンピュータ科学・人工知能研究所は、既に公開されていた国際的な論文 34 件を体系的に調査し、その中から AI の潜在的リスクに関する言及を 777 件抽出し、以下の 7 分野にてまとめて公開した(参考資料 1、2)。

① 差別と毒性 (Discrimination & Toxicity)
： AI 技術を用いて、差別の助長、有害コンテンツの提供、偏見の推奨を行うリスク

② プライバシーとセキュリティ (Privacy & Security)
： 許可なく個人情報を推測することでプライバシーを侵害、データ漏洩、システムの不正操作を行うリスク

③ 偽情報 (Misinformation)
： ユーザーが好む情報提供を行ったり、誤情報を拡散したりすることで、社会の一体感や政治プロセスを混乱させるリスク

④ 悪意ある行為者と悪用 (Malicious actors & Misuse)
： 偽情報拡散/監視/プロパガンダ/サイバー兵器開発/詐欺のために、AI 技術を悪用するリスク

⑤ 人間とコンピュータの相互作用 (Human-Computer Interaction)
： ユーザーが AI に過剰に依存してしまい、本来の思考力や自律性を減じてしまうリスク

⑥ 社会経済・環境 (Socioeconomic & Environmental)
： AI 技術の導入が、権力や富の集中、産業界/社会の混乱、環境汚染、不平等拡大を促進するリスク

⑦ AI システムの安全性、故障、限界 (AI system safety, failures, & limitations)
： AI が人間の制御を拒絶し、有害な能力を獲得し、人間の目標や価値観と対立するようになるリ

スク

上記の表現のように、近年増えている AI リスクの議論にては、「差別」、「偏見」、「悪意」、「社会の一体感」、「人間の価値観」のような、従来の工学の議論では用いることが少なかった、どちらかというと社会科学や人文科学系の概念用語にて技術が議論されることが増えている。

そのような議論が増えた理由は、次の三つの影響が大きいように思える；

- AI 開発にて人間の知性と同等の機能や性能を実現しようと目論む技術者や企業が増えているため、AI 技術の性能評価でも、人間の知性に関する評価基準を流用し始めている。
- 歴史的にロボットや AI 技術の高度化に警戒心を持って来た英米の様々な分野の研究者が、懸念に関する発言を増やしている。
- 2022 年末に ChatBot 型の AI サービスが公開され、更に、用意する計算資源の規模と実現される知性の関係を定量的に関係付けるスケーリング則が発見され、将来の AI の性能予測が可能となった。

しかしながら、「差別」、「偏見」、「悪意」、「社会の一体感」、「人間の価値観」のような問題点やリスクの評価に、工学や自然科学で行うような「実験」は困難なことが多く、明瞭な閾を設定しようとするのは如何にも難しい試みである。

ましてや、インターネット経由で国境を越えて行われると予想される AI サービスにて世界中の誰もが受け入れる事ができる「価値観」や「コンセンサス」の定義は難題である。 「重要なのはダイバーシティを受け入れることであり、倫理観/価値観/宗教を世界統一しようとするのはむしろ有害だ」と、批判/反発されるのは明らかであり、北米企業のデファクト標準化能力がいかに強力としても、特に、歴史が古く様々な文化がモザイクのように散らばるユーラシア大陸の多くの国々にては、壁のような困難が待つに違いない。

更に、インターネット経由で提供される AI サービスの評価や検証にては、本質的ともいえる「検証の難しさ」がある。 それは、例えば、以下の三点である；

- AI システム内のコアとなる部分の論理は本質的に確率的であるため、同じ入力に対する出力は確定的とならない。
- AI システムへの「入力パターン」には無限ともいえる程に多彩な組み合わせがあり、全ての入力パターンに対する応答の良し悪しを事前には検証しきれない。
- インターネット経由で不特定多数の人達からのプロンプト入力を受け、継続的にデータが埋め込まれ、影響を受け続ける AI の「影響を受けた後の応答性能」をユーザーの使用前に予測することはできない。

筆者は、これらの問題は、 近年のソフトウエアサービスで一般化して来た「Agile なソフトウエアの改良」では対応しきれない問題と思う。AI 技術に期待されるアプリケーションにては、人間によるデリケートな判断や応対の代換を狙うアプリケーションが多く、「問題が発覚した後にアップデートにて改良する」では済まないからである（注１）。

「AI は、投資に見合う価値を生むのか」の議論

米国の MIT は、2024 年 10 月 17 日に、「期待外れだった IT 革命、生成 AI は今度こそ経済成長をもたらすか？」と題したレポートを発表し、これまでの AI 技術が生産性の伸びに寄与することを疑問視し、企業/行政組織のオペーレーションや研究開発の効率化/競争力強化に資することを重視すべきと論じた（参考資料 5)。

企業/行政の「組織」には、その組織特有のルールが存在する。 更に、社会には「法」が存在し、多くの分野で、「善悪」の評価判断を可能とする情報基盤が存在する。
それらルールの遵守は「確率的」であってはならず、旧来のコンピュータアルゴリズムのような確定的な論理によって確実に遵守する必要がある。

前述の米国 MIT の指摘は、AI 技術の開発を、「人間の知性の模倣」ではなく、「企業や行政、もしくは、軍隊や研究所に備わる組織の知性を模倣し、強化することを求めている」と理解すべきだろう。

生成 AI 技術は、コンピュータが人間の言語のような曖昧な情報を処理する道を切り開いたが、産業界や行政の組織の運用にては、確率的論理で操作するだけでは不十分であり、ルールを確実に守る装置を求める。これらの指摘は、AI 設計者に、確率的論理と確定的論理を融合させた Computing を求めていると思える。

企業間の効率化競争に資する AI 技術

本年 7 月 10 日に、米起業家の Nova Spivack 氏は、Web ブログにて、「自身が設立した Mindcorp.ai 社が、企業組織向けに、“Cognition”と呼ぶ研究開発/分析/戦略検討型組織の業務遂行プラットフォームを開発している」と発表した（参考資料 6)。 同ブログは、”Cognition”を「AI＋Human の Hybrid System であり、事業を 1 万以上のエージェントの支援を受けて進められるようにする業務プラットフォーム」 と表現した。

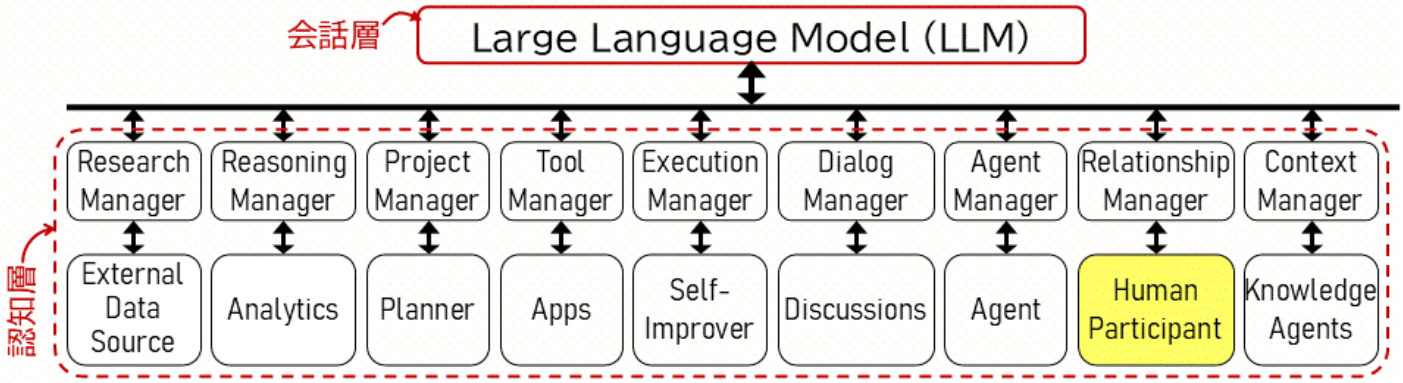

図 1 ： Mindcorp.ai 社の CEO が論文中に示した Cognitive AI システムの機能アーキテクチャ （参考資料 7） 出典： Nova Spivack, et al。;の“Cognition is All You Need” (2024)を元に、筆者が作成
URL: https://arxiv.org/abs/2403.02164

同氏は、2024 年 3 月に、Mindcorp.ai 社のメンバーとの立場で、「複数のエージェントのチームと多くの人間が協力し業務を進める”Cognitive AI”と呼ぶ事業プラットフォーム」の開発方針を開示していたが、もしも、7 月 10 日発表の Cognition システムが 3 月発表の「Cognitive AI システム」であるとすると、多くの AI　エージェントと比較的少人数の人間担当者を合体した AI-Human 融合型 IT システムによる、組織の全業務（Work-Flow）の代換（DX）を狙っていることとなる。

3 月発表の Cognitive AI システムは、大規模な言語モデル（LLM）と多くの汎用的な機能モジュールからなり(図 1)、両者間を行き交う一種の内省的な動作によって出力用のデータを生成し、更に、そのデータを出力した場合の影響評価を行い、問題を引き起こさないとのシミュレーションされた場合にそのデータを出力するという（注 2)。

画期的と思えるのは、その「AI エージェントと人間担当者達を合体した AI-Human 融合型 IT システム」の全体を“AI”と呼び、多くの企業の業務を表現可能な、つまり、業務プラットフォームの標準形を目指している点である。DX 用の 標準かつ汎用のプラットフォームの提供を狙っているといえる。

そのような組織業務の大幅な効率アップを目指している「業務プラットフォームの標準形」のコスト競争力や利便性が実証されると、その評価は、産業界や行政組織に大きなインパクトを与えるのではないだろうか？　Nova Spivack 氏は、7 月 10 日の Web ブログにて、"Mindcorp.ai Launches Enterprise Superintelligence" との表現でそのシステムをアピールした。

DX により生み出されうる新たな水平分業エコシステム

前記のような Hybrid 型のプラットフォームが、企業や行政組織の業務フローの全体を表現できるようになるとすると、各企業や自治体の業務は、

(a) 基盤テクノロジー開発
(b) 業務効率化プラットフォーム開発
(c) 業務プラットフォームの運用
(d) 業務プラットフォームのメインテネンス
(e) 他の業務プラットフォームとのインターフェース プロトコルの開発

に分解し、産業界や複数の自治体の全体を、これら 5 種類の業務プラットフォームの水平分業にて代換する「破壊的イノベーション」が進みうる。

そのような水平分業化が合理的とみなされるようになるには、産業界に、「水平分業を行った方がより効率的に事業を進められる」とのコンセンサスが構築される必要はない。(b) と (c) の業務を行う事業プラットフォームが開発され、社会実装され、「そのプラットフォームを用いると、比較的少人数で (c) のタイプのビジネスを進めることができ、コスト構造が従来よりも格段に低くなる」と証明されてしまうと、その後は、市場の力と首長選挙や議会選挙のダイナミズムによって、DX が各所に浸透してゆくだろうからだ。

そのように構造転換が進む場合に、産業界や行政素組織に派生的に起こる事象とリスクは、図 2 のように、(A) 産業構造の国境を越えた水平分業化、(B) 個人の生活に及ぶ影響、(C) 既得権益勢力と新興勢力間の摩擦、利益配分構造や産業構造の大変化と分類し、再整理されうるのではないだろうか？

(A) 産業構造の国境を越えた水平分業化				(B) 個人の生活への影響
A1) 基盤技術開発競争（ハードウエア、ソフトウエア、ライブラリ、コミュニケーション、等）				B1) 雇用流動化、収入不安、詐欺増
A2) 業務効率化プラットフォーム開発競争（事業分野、行政分野、研究開発分野）				B2) プライバシー問題、AI中毒、
A3) 業務プラットフォームの社会実装競争（物流/流通、交通、通信、データセンター）				B3) 報道情報の偏向、不信の増加
A4) 業務の効率化と再編成（小売り、医療/ヘルスケア、法務/自治体サービス、等）				B4) 生きがい問題、健康への悪影響
(C) 既得権益勢力と新興勢力間の摩擦 / 利益配分構造や産業構造の大変化				
C1) 金融リスク（通貨/預金/投資/保険/年金）	C2) 企業の経営問題	C3) 地政学的問題/経済安全保障問題	C4) 社会システムの見直し	

図 2： (A)産業の国境を超えた分業化と、(B)個人の生活に及ぶ影響がイタレーションし、(C) 既得権益勢力と新興勢力間の摩擦 や 利益配分構造や産業構造の大変化が進むのではないか？　出典：筆者作成

その際には、もちろん、
- この水平分業体が生み出す「付加価値」をどのように配分するのか？
- 従来の事業組織が持っていた様々な「ノウハウ」をどのように値踏みするのか？

が問題となるだろう。

Nvidia 社の創業者/CEO であるジェンスン･フアン氏は、2024 年 11 月に開催した同社の AI Summit Japan で、「すべての業界、すべての企業、すべての国が新たな産業革命を起こす必要がある」と語った(参考資料 8)。
北米では、既に、ポスト生成 AI の段階の AI 開発が力強く進んでいると思える。

注釈:

1) PC やスマホでの「Agile なソフトウエアアップデート」は、開発者とユーザーの両者のメリットとなったが、そのビジネスモデルを高度な AI サービスにまで拡張することの是非については、未だ議論が未成熟と思われる。
ハードウエアに関しては、商品製造元企業や設計者は、1994 年に成立した PL 法（製造物責任法）にて、製品の欠陥に基づく被害に対しては、提供者としての責任を背負っていると理解されているが、ソフトウエア提供後の問題の扱いは、ユーザーとの間の個別の許諾契約次第であり、消費者ユーザーからは、サービス提供者への責任追及は困難とみなされていることが多い。

AI リスクに関しても、2019 年に採択された OECD の AI 原則までは、「サービスの開発/提供者の責任」が明記されていた（参考資料 3）が、2023 年の広島プロセスでは、問題への対処責任の在り処がサービスの開発/提供者ではなく、政府に移ったかのような表現に変更された（参考資料 4)。

2) 「対話層」と「認知層」の内省動作は、2021 年にセミコンポータルのブログにまとめた「対話する構造」に似る（参考資料 9、10)。

参考資料：

1）“What are the risks from Artificial Intelligence?”, 2024 MIT AI Risk Repository （website）;
URL; https://airisk.mit.edu/

2）Scott J Mulligan, “A new public database lists all the ways AI could go wrong”, the MIT Technology Review, August 14, 2024
URL; https://www.technologyreview.com/2024/08/14/1096455/new-database-lists-ways-ai-go-wrong/

3）「OECD、生成 AI の台頭で国際指針の見直しへ」、 OECD の AI 原則、(2019 年)
URL; https://www.jil.go.jp/foreign/jihou/2023/06/oecd_01.html

4）「全ての AI 関係者向けの広島プロセス国際指針」、広島プロセス、（2023 年）
URL; https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03.pdf

5）Rhiannon Williams, “AI could help people find common ground during deliberations”, in the MIT Technology Review, (2024/10/17);
URL; https://www.technologyreview.com/2024/10/17/1105810/ai-could-help-people-find-common-ground-during-deliberations/

6）Nova Spivack, “Mindcorp.ai Launches Enterprise Superintelligence”, (2024/07/10)
URL: https://www.novaspivack.com/business/mindcorp-ai-launches-enterprise-

7）Nova Spivack, et al., “Cognition is All You Need”, (2024);
URL: https://arxiv.org/abs/2403.02164

8）「NVIDIA の創業者/CEO であるジェンスン フアンが AI Summit Japan で『すべての業界、すべての企業、すべての国が新たな産業革命を起こす必要がある』と語る。」Nvidia のホームページ、 (2024/11/15),
URL; https://blogs.nvidia.co.jp/2024/11/15/ai-summit-japan-huang-son/

9） 岡島義憲、「人工知能への道（3）:「対話する構造」の重要性」、セミコンポータル、(2021 /08/27).
URL: https://www.semiconportal.com/archive/blog/insiders/okajima/210826-computing.html

10）岡島義憲、「人工知能への道（4）：対話する構造」、 セミコンポータル、(2021/10/07).
URL: https://www.semiconportal.com/archive/blog/insiders/okajima/211007-mindmodels.html

情報統合技術研究合同会社（https://info-integnology.com/index.html）
代表、岡島義憲