AI リスク（その 1）：AI 技術実装により自律化するシステムの課題

トップ研究者達の AI リスクへの言及が、約 1 年前より増えて来ている（参考資料 1～5）。その内容は、雇用減少を気にした旧来の論調からエスカレートしており、人類を脅かす問題、核兵器並みとされている。 本ブログでは、その概要を数回に分けて報告したい。問題を正確に認識することが、対策技術の価値を理解する上での重要なステップと思うからである。

トップ研究者達の警笛

トロント大学の Hinton 教授（注 1）は、2023 年 5 月に Google 社から離れ、複数のメディアを通して AI リスクを警告した。その中には、以下の指摘が含まれていた。

- AI は制御を失い、人類にとって損害をもたらす存在になる可能性がある。
  （AI 等の先端技術により人類が絶滅する確率は、今後 20 年間に 10%程度ある）
- ネットワークにつながる AI 群は、一種の集団意識（Hive-Mind）を形成し、人間に対する優位性を手に入れる可能性がある。
- 対策は分からないが、アナログコンピュータ（注 2）の方が、人間にとっては好ましいだろう。

また、現在の生成 AI 技術の先導役であるモントリオール大学の Bengio 教授は、2023 年 12 月に、以下のように米国議会上院にて AI リスクへの理解を証言し、対策が必要だと訴えた（参考資料 3）。

- 環境と自律的に相互作用する AI システムは、不可逆的に制御喪失する可能性がある
- その対策は見い出せていない。また、その対策にどの位の期間を要するか分からない。
- 不良化した AI（Rouge AI）の出現に備え、防御的 AI を開発する国際的な研究ネットワークを構築し、防衛的 AI の技術をそのネットワーク内に秘匿化するべきだ。
- 大規模 AI システムは、寡占化する（大きな権力を少数個人に与える）可能性がある。

両教授共に、大手テック企業との関係が密だったからであろうか、これらの問題を予想する理由については多くを語っていない。また、社会の側も、行政府を除き、日本国の中では未だ大きな反応を見せているようには思えない。反応が少ない理由は、AI が人類の敵となる物語は、サイエンス・フィクション（SF）の馴染みのテーマとして定着してしまったからであろうか？

実は、工学研究者達の AI リスクに関する学術論文の発表や書籍の刊行は、2017 年を境に、急増している（図 1）。それらは、

- AI 技術に潜む脆弱性問題
- AI 技術の悪用問題（マルウエア生成やサイバー攻撃）
- AI の動作がブラックボックス的であり、推論結果を説明しない問題
  （AI の論理は確率的であり、人間が理解可能な説明を AI に行わせることができない）
- AI の動作を開発者の意図に従わせることの難しいこと
- 高度な AI をインターネット等で環境と相互作用（自律動作）させることは危険なこと
  （AI が、自律的に人間が望まない動作を行うという現象が見つかっている）
- 社会や産業界に、独占や寡占が発生する懸念があること
- 軍事兵器の自律動作には危険であること

に関する内容が多い(参考資料 4、5)。

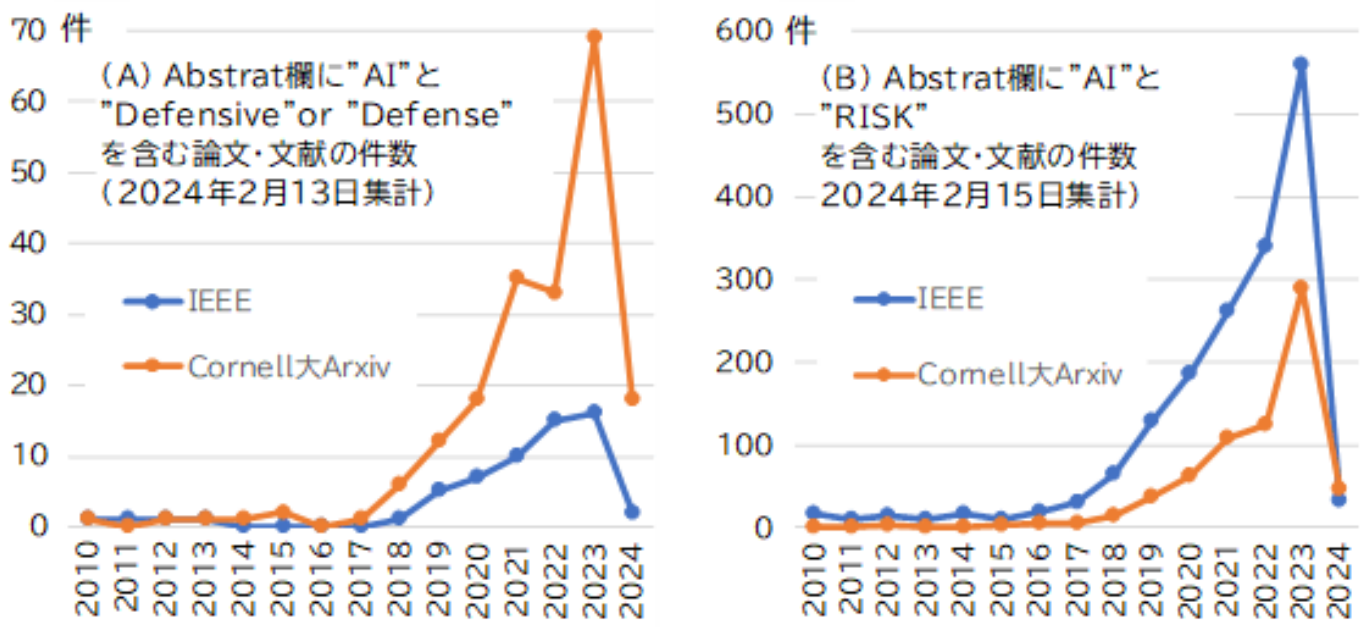

図 1 左の(A)は、Abstract 欄に"AI"と"Defensive"を含む論文・文献数の推移(2024 年 2 月 13 日集計)、右の(B)は、 Abstract 欄に"AI"と"Risk"を含む論文・文献数の推移(2024 年 2 月 15 日集計) 2017 年は、生成 AI に関する基盤技術が見いだされ、AI の言語インターフェースが革新され、IoT に関する連合学習技術が見いだされ、大量の AI(自律エージェント)群を通信ネットワークによって統合するという方向が現実的となった重要な年であった。

非常に、多岐にわたるテーマが含まれているが、筆者は、これらは、AI を IoT(Internet of Things)システムに実装させた時に発生する問題であると見る。そこで、以下に、2017 年以降、顕著となった AI-IoT 統合での重要技術 2 点を改めて振り返ってみる。

IoT システムのマルチエージェント化と連合学習技術

IoT 技術は、実物世界で稼働するセンサ等の端末機器を通して収集する大量のデータを、インターネットを通してデータセンターのサーバに蓄積し、統計処理を行った上で端末機器に制御信号として回帰させるというネットワークベースの複合システム(System of Systems)アーキテクチャである。大量に収集されたデータを元に、サーバがデータを確率的な推論論理として集約し、高度運転支援/自動運転/製造制御/生産制御/インフラ管理から金融/事務管理に関する端末装置を自律的にサポートする社会実装が現在進められている。

その IoT 端末を、ネットワークが断絶した孤立環境でも自律動作することを可能とすべく、近年は、端末やエッジサーバ装置にも AI 技術を実装する方向(Edge-AI)が試されて来ている。そのように自律動作能力が高い端末は、コンピュータサイエンスや認知科学で言われる「エージェント」とみなされうる(注 3)。端末を AI エージェントに進化させ、IoT システムを、自律性が高いマルチエージェントシステムに変容させる研究は、特にドローン等の軍事用兵器開発にて強力に進められて来た(図 2)。

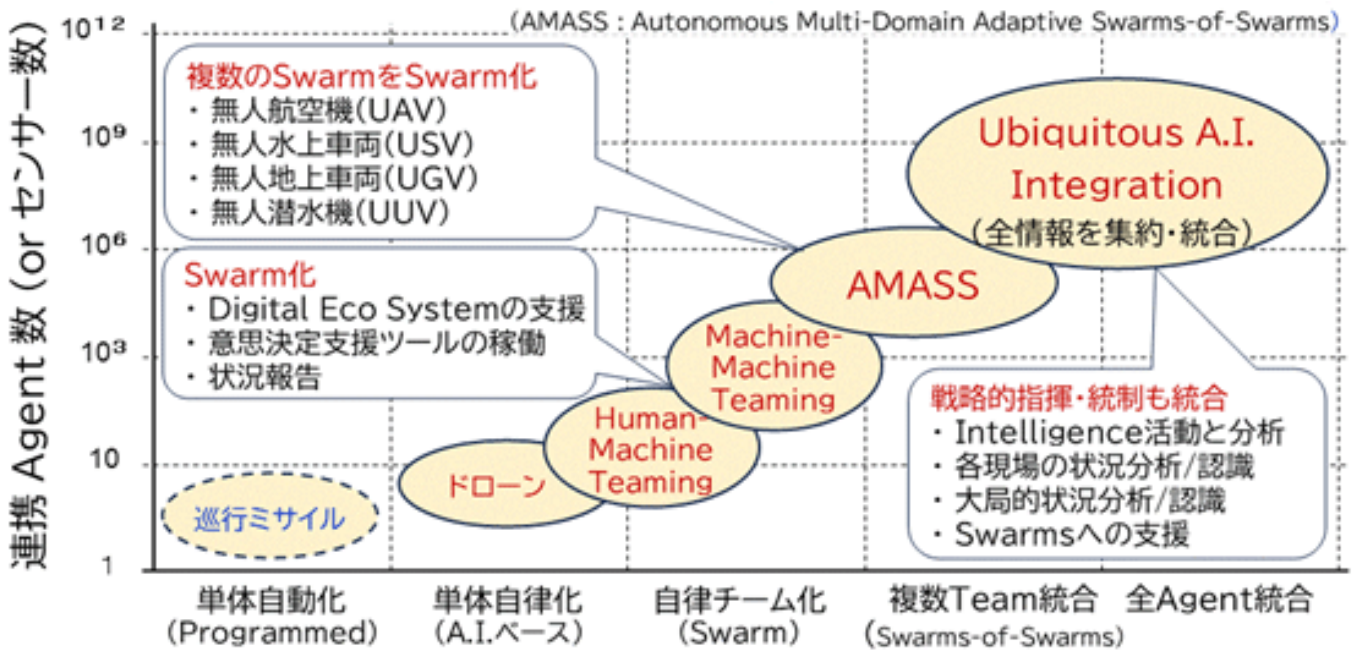

図 2 軍事技術における AI エージェントの統合化の動向 出典 ： 汎用人工知能研究会での筆者らの発表論文で用いた図

但し、複数のエージェントが連携し、共通の目標や目的を実現すべく協調するには、各エージェントの学習がローカルに孤立し分散的であっては不十分であり、各エージェントの学習結果を、システム全体を管理するサーバに集約・統合し、協調的な経験知として再構成して再配信しなくてはいけない。
そのような協調的学習（Collaborative Learning）を、より少ないデータ交信量にて効率的に行う手法が、2017 年に、Google 社より見い出され連合学習（Federated Learning）との名称で発表された。連合学習は、IoT システムを自律性が高いマルチエージェントシステムに進化させる上での重要なブレイクスルーであった。

端末や Edge サーバへの AI 回路の実装による、IoT システムのマルチエージェント化は、軍事システムを大々的に革新しつつあるが、そこで用意されている技術は、産業界や政府・自治体が進めている DX（デジタルトランスフォーメーション）で用いられる技術と共通する。 その状況の調査は、昨年、筆者らが「AI によって自律化が進む米国軍事技術の動向」と題した研究会論文にまとめた（参考資料 7、8）。

生成 AI によるマルチエージェントシステムの管理

Google 社は、2017 年に、トランスフォーマアーキテクチャと呼ぶ、入力テキスト文を元に、文章や画像などのコンテンツを自動生成する生成 AI 技術（大規模言語モデル）を発表した。生成 AI は、プログラムコードを出力させることもできるため、各種の装置内のプログラムと対話し、更新されるソフトウエアを自律的に配信することも可能である。つまり、生成 AI は、チャットボットとして人間と対話するように、IoT システムの各 AI エージェントと会話し、各 AI エージェントに指令を出すデータセンター側の技術であった。

結果、IoT システムは、環境からの情報を元に自律的に状況を認識し、自律的に環境中に働きかけることが可能な制御系 AI システムへの進化を開始した。しかし、このことが、先進の研究者達も不安とさせている点である。Internet には様々な AI 搭載装置が参加しているため、生成され送信されるデータやコードが個々の装置に及ぼす影響の全てを予想することなど不可能だからである。ましてや、AI 技術には、様々な脆弱性や制

御の難しさが報告されている。

カリフォルニア大学バークレイ校(UC Berkeley)の Dan Hendrycks 氏は、「これ(AI 開発)は火遊びのようなものだ」と言い、「今日のシステムをより安全にする方法の検討が必要」と、工学技術の側面から「安全設計の原則」が必要であることを強調した(参考資料 4)。

また、イリノイ大学アーバナ・シャンペーン校(University of Illinois Urbana-Champaign、UIUC)の R. Fang ら研究者達は、「大規模言語モデル(LLM)が、事前にサーバや PC 等の脆弱性を学習させていなくても、自律的にウェブサイトをハッキングし、複雑なタスクを人間のフィードバックなしで実行すること」を示した(参考資料 9)。インターネットに接続され、世界中にサービスを行っている GPT-4 のような生成 AI も、「そのようなハッキングが可能」であることが確認され、報告した研究者達は、「私たちの調査結果は、LLM の広範な展開について疑問を引き起こす」との見解を記した(参考資料 9)。

今、この問題の全体像を理解し、対策を進めることが必要となっている。当然ながら、対策技術には、大きな市場価値があるはずである。

筆者注
1. Hinton 教授は、深層学習(Deep-Learning)技術を見出し、Auto-Encoder や深層学習を始めとして、第 2 次 AI ブーム以降の AI 技術開発の先導者であった。
2. Hinton 教授は、Google 社で「アナログコンピュータ」の開発に取り組んでいた。Wired-Online の記事によると、「アナログのハードウェアはそれぞれ少しずつ異なるため、あるアナログのモデルから別のアナログのモデルにパラメータの重みを移すことはできない」ことがメリットとなると語っている(参考資料 10、11)。
3. エージェントとは、元々は、「設定された目標や目的を実現すべく自律的に活動する代理人」を意味する法律用語であったが、コンピュータサイエンスでは、1990 年頃より、「特定タスクを自律的に実行するプログラム」の意味で使われて来た。結果、「エージェント」は、サーバ側 AI の支援が無くても自律的に学習し応答することが可能な「特定タスク向けの知的機能モジュール」の意味で使われて来ている(参考資料 6)。

参考資料
1. Y. Bengio, et al., "Managing AI Risks in an Era of Rapid Progress", (2023).
2. Y. Bengio, "AI and Catastrophic Risk", in Journal of Democracy", (2023/09)
3. Y. Bengio, "Presented before the U.S. Senate Forum on AI Insight Regarding Risk, Alignment, and Guarding Against Doomsday Scenarios", presentation before the U.S. Senate Forum, (2023/12/06)
4. Dan Hendrycks & Mantas Mazeika, "X-Risk Analysis for AI Research", Submitted on 13 Jun 2022.
5. 山川宏、江間有沙、大蔵峰樹;「(耕論)AI と私たち 人工知能、制御できるか」、朝日新聞デジタル、(2023/09/16)
6. M. Wooldridge & N. R. Jennings; "Intelligent agents: theory and practice", The Knowledge Engineering Review , Volume 10 , Issue 2 , June 1995 , pp. 115 - 152.
7. 岡島&山川、「AI によって自律化が進む米国軍事技術の動向」、汎用人工知能研究会、(2023/10/27)
8. 岡島&山川、「AI によって自律化が進む米国軍事技術の動向(スライド)」、汎用人工知能研究会、

(2023/10/27)
9. R. Fang, ”LLM Agents can Autonomously Hack Websites”, 2024
10. Steven Levy,「AI のゴッドファーザーが提案する、未来の AI を友好的に保つ方法」、Wired, (2023/08/11)
11. Charles Platt,「アナログコンピューターの逆襲―複雑な現実を扱う新世代アナログチップは実現するか」、Wired, (2023/06/23)

情報統合技術研究合同会社 代表 岡島義憲