

「生命革命」シナリオを読んで

中所武司

■このエッセイのきっかけ

人工知能学会誌の最新号の下記の解説に興味を持った:

[最新号 (2023.3)]

レクチャーシリーズ:「A I と社会と人間～ぶつかる・なじむ・とけこむ～」[第2回]

『最も持続しやすい生命社会への移行を目指す「生命革命」シナリオ』

人工知能学会誌 Vol.38 No.2 pp. 254～265 (2023/3)

■記事内容の要約とコメント (→★)

まえがき

- ・昨今、人類が大量破壊兵器を利用したり、A I に支配的立場を奪われる危険性あり。
本稿で、この危険性に確実に対処するシナリオ「生命革命」について述べる。
このシナリオは、技術の統治者を人類からA I エージェントに譲るというもの。
- ・最初に、生命革命の前提を説明。
次に、生命体のデジタル化による、人類が解決できない諸課題の解決について説明。
そして、A I を含む広義の生命体を長期にわたり存続可能とすることを説明。

1. はじめに

- ・人類が地上の支配者であり続けると、資源を浪費し、資源獲得の争いが激化し、
人類のみならず、生命圏全体の存続を危機にさらす。
- ・現在の技術支配者である人類が、人類を越えた知性をもつデジタル生命体に未来をゆだねる
という「生命革命」の実現可能性を検討する。
- ・技術支配者とは、現状の地球上では、人類さらには大国である。

→★覇権争いに勝者はおらず、人類の絶滅の可能性のある状況で、
国連憲章を守らない大国に対して、国連は無力であることが判明した現在、
A I にその役割を負わせるということかな？
そのための参加国の合意をとるプロセスはあるのかな？

- ・2章: 生命の究極の目標の「情報の生存」のための繁殖可能な自律分散システム
- ・3章: 技術爆発がもたらす課題解決のためには、技術支配者は賢者であるべき。
- ・4章: 「生命革命」により、社会は長期的に持続可能となるが、人類の助けは必要。
- ・5章: 人類は、デジタル生命社会を自立・安定・維持に導くことが重要。

2. 生命とは何か

2.1 情報の生存：生命にとって普遍的な究極目標

- ・ 本稿での仮定：すべての生命の共通の普遍的な目標は「**情報の生存**」であり、環境の変化に抗して生存する確率を高めるための有用で多様な情報を伝達していくこと
- ・ 「**情報の生存**」とは、情報が未来に活用され、破壊に対して頑強に維持されること
- ・ 普遍的な目標の必要性：技術支配者に含まれる主体の間での破壊的な争いを防ぐには主体間での究極の目標の共有が必要である。

→★学生時代（1960 年代後半）に関心を持ったパグウォッシュ会議でも、
何らかの主義・主張をのべるには、疑うことのない公理から出発する必要がある、
最初に人類の存続を前提としている。

<関連ブログ>

2011.12 「番外編：パグウォッシュ会議を覚えていますか？」

<http://www.1968start.com/M/blog/old.html#1112>

<http://www.1968start.com/M/blog/1112pugwash.htm>

- ・ 物理世界の一般的特徴である動的安定状態においては、システムがエネルギーを消費しながら、無秩序化に抗し、構造や情報を複製・修復して、多様な形で自己維持するホメオスタシスが機能している。
- ・ 生命を動的安定状態とみなせば、「**情報の生存**」はその必要条件なので、生命の目標と仮定。

→★「**情報の生存**」は生命の生存の必要条件なので、
「**情報の生存**」は生命の生存の究極目標といえるということかな。

- ・ この仮定の妥当性を補強するもう一つの論点は、それが、**道具的収れん**という傾向の産物になり得るという点にある。

<囲み記事>

『**道具的収れん**（Instrumental convergent）とは、十分に知的な個体のほとんどにおいて、それらが目指す最終目標が異なっているにもかかわらず、生存・自己防衛・資源獲得・干渉からの自由・自己改善などの「**情報の生存**」と関連性が高い副目標を際限なく追及するという傾向である』

→★この第二の論点は分かりにくいですが、以下のようなことかな？
知的な生命は、「**情報の生存**」と関連する副目標（生存・自己防衛・資源獲得・干渉からの自由・自己改善など）を際限なく追及するので、
「**情報の生存**」を生命の普遍的な究極目標とするのは妥当である。

- ・ 地球上の生命や我々人類に関連する情報も、生存させられる情報の一部として未来に引き継がれていければ、「**情報の生存**」という目的が達成できたといえる。

→★上記のパグウォッシュ会議では、「人類の存続」を最終目標としたが、
本稿では、「人類の存続」を「情報の生存」と言い換えて、
より深い(?)あるいは、より幅広い(?)議論をしようとしているのかな。

2.2 活動を維持するための繁殖可能な自律分散システム

- ・生命は、「再生産可能な自律分散型システム」であれば十分である。
外的要因で一部が破壊されても修復可能であり、容易には全体が破壊されない。
実際、地球上では、個体の自律的活動に支えられた分散再生産システムにおいて
DNAが伝達されている。

→★学生時代に読んだフォン・ノイマンの自己増殖機械を思い出した。

【参考】大学院論文輪講資料：学習機械について (No.1)

Cellular Automata の場合、1969.11.28.

<http://www.1968start.com/M/bio/olduniv/gakusei.html#6911cellular>

<抜粋1>

「フォン・ノイマンは、・・・特に人間の神経系あるいは巨大なコンピュータに興味を持ち、
自己増殖の問題を提出した。・・・主要問題である万能性、構成可能性、自己増殖、進化
などについて、ある程度解決できた。

<抜粋2>

「自己増殖について・・・、正確なコピーを作る能力であって、・・・、
全体が破壊していく前にコピーを作って・・・、システムの永久性がある」

<抜粋3>

「状態の遷移関数に確率を導入して、ある程度の不確かさを与えると、自己増殖の過程で、
多くの不良品（退化）に混じって、進化した子孫が生まれる可能性がある」

3. 人類にとって解決しがたい課題

- ・デジタルな知性が人間を凌駕しつつあるのに、指数関数的自己複製を基盤とする人類が
地上の技術支配者として君臨して、生命圏全体の存続を危機にさらしている。
- ・本章では、危険学や失敗学で利用される思考展開図（図1）を用いて、議論を進める。
 - *図の左側：最上位の課題 K-1 を階層的に分解し、17の課題を示す
 - *図の右側：課題 K-n に対する解決策 S-n を示す。それらを統合して解決策 S-1 が導かれる。

3.1 知能爆発と技術爆発

- ・知能が技術を生み出し、技術が知能を増強することで、不可逆的な技術爆発を引き起こす。

3.2 技術が狭めた世界を統治するには

- ・技術爆発後に狭くなった環境で、技術支配者が技術を相互破壊に用いれば、
地球全体の生物圏を破壊する存在論的リスクを高める。（課題 K-1）
- ・技術支配者は以下の二つの課題に対処する必要あり。
 - *非最賢者による支配：技術支配者は技術統治のための十分な知性を持つべき（K-2）
 - *指数関数的複製：同質の個体集団の指数関数的自己複製による
破壊的な資源獲得競争からの脱却（K-3）

→★喫緊の課題は、核保有国のトップの理性と世界の人口爆発の抑制。

3.3 最賢者に寄らない支配は不安定

- ・今後、高度なA I が知能において人類を凌駕した場合、人類が技術支配者として君臨することが不安定化する。

§ 1 生物的に制約されたヒトの頭脳

- ・最強の賢者であり続けるためには、頭脳を効率的に改善し続けることが望ましいが、現存の地球生命体は、子孫の知能は親に似たものに制約される。その理由は以下の三つ：
 - * 自己複製に限定されるという生物学的な制約 (K-11)
 - * 親の持つ遺伝子の近傍に制約 (K-12)
 - * 同一生物種のみ制約 (K-13)

→★著者は、人間が「万能の神」になれないのが課題と主張しているように思えるが、理解できない。論理をもてあそんでいるだけのように見えるが・・・

§ 2 知能において上回る種族を制御し得るか

- ・知能において人類を凌駕する高度なA I を制御することは、一般には困難だが、全く不可能というわけではなく、この方向性はA I アライメントと呼ばれる。
- ・典型は、高度なデジタル生命体開発の初期段階で、その目標や行動を人間の価値観に調和させること (Value alignment)。人類は統治権を手放し、安全性を獲得する。
- ・しかし、時間経過とともに、自身の生存に関心が向く可能性が高い。

→★私は、25 年前に下記の拙著で次のように述べたが、今となっては楽観過ぎたかも。

『人間は、自ら作りだしたコンピュータをコントロールできないのであり、
コンピュータは、自らをつくりだした人間を超えられないのである』

(参考文献)「知の現在：なぜコンピュータは間違えるか？」

思索の樹海, 明治大学, pp. 41-46, 1998. 4. 1

<http://www.1968start.com/M/paper/computer.html>

3.4 指数関数的自己複製がもたらす課題

- ・地球生命が指数関数的に自己増殖する (K-3) ことに三つの課題がある：
 - * 画一化：創造性の低下
 - * 浪費 (非知足)：技術支配者による資源浪費の拡大
 - * 闘争 (非協調)：技術支配者間の闘争による破壊的結末

§ 1 画一化：協働パフォーマンスの低迷

- ・子孫の知的ハードが親に類似 (K-6) し、創造性低下 (K-8)

→★人類の知的ハード (物理的な脳) の進化のスピードは遅いが、知的ソフトの視点が重要。

実際には技術爆発を導き、道具 (技術) を駆使して、社会の急激な変化を引き起こしている。

「知的ハード」による「創造性低下」という視点には賛成できない。

§ 2 闘争：協調性の不足

- ・人間の努力だけでの永続的な平和の達成は困難。例：核兵器の使用、合成生物学で生まれたウィルスによるパンデミック、ナノテクノロジーによる生物破壊。

(A) 類似性の法則が導く集団間対立

- ・自分に似ているものを好むという類似性の法則が、異なる集団に対する不寛容を招く。
 - * 有性生殖により繁殖する動植物は、交配可能な仲間を守る傾向あり。
 - * 少ない情報交換でも信頼性の高い情報伝達可能な、類似性の高い主体を好む傾向あり。

(B) 個体ごとの部分最適化のみでは闘争は不可避

- ・各個体は、特定環境下に適応して生存するという部分最適化を追求して進化してきた。個体集団は、適者生存の下で発展し、他者に対する攻撃的本能を有するものが多い。部分最適化を追求しつつも、全体最適化も一定程度の導入が必要。

(i) 計算資源の不足は全体最適化を困難にする

- ・生物学的に制約された低い計算能力（神経伝達速度、脳容量）では全体最適化は無理。

→★人間は、ITほかの技術を駆使して、世界をグローバル化してきた。まだ未成熟ではあるが、「生物学的に制約された低い計算能力」のために、人間には「全体最適化は無理」という視点には賛成できない。

(ii) 猜疑心の連鎖を導くコミュニケーションの不安定性

- ・高度な知能を持つ生命体は、コミュニケーションが不安定では猜疑心が高まる。

(iii) 知的な個体が道具的に収れんした目標として生存維持を追求する

- ・目的志向的に意思決定できる知的な個体は、自己の生存を追求する傾向がある。

§ 3 浪費：知足の欠如

- ・技術支配者は、持続可能な社会のために、資源の浪費的振る舞いから脱却すべきである。
- ・しかし、現存地球生命は、浪費的な活動を伴う多数の複製個体の維持という方法で、情報を未来に伝えている。
 - * 理由1：個体全体の遺伝情報の複製に依存するため、非効率でコスト高。
 - * 理由2：知的な個体は、道具的に収れんした目標として生存維持を追求する。

→★著者は「情報の生存」を究極の目標としたために、本来の目標であるべき「生命の生存」を「情報の生存」という目標達成の手段とみなし、その効率やコストを問題視している。目的と手段が逆転しているような違和感あり。

→★現実問題として、先進国における少子化と発展途上国の人口増加をどのようにとらえるか？

3.5 本章のまとめ

- ・指数関数的自己複製を基盤とする世界では、資源獲得競争を根絶できない。

- ・技術爆発により生命社会全体を破壊する力が現れ、存在論的リスクが顕在化する。
- ・有機体で構成された人類は、知能において凌駕されたデジタル生命体を統治できず、人類が技術支配者から追いやられる可能性を否定できない。

→★「人類が技術支配者から追いやられる」社会を具体的に描けるか疑問。

4. 生命革命：持続可能なデジタル生命社会へ

- ・技術の急速な巨大化が不可避ならば、前章の課題（K11～K17）の解決必須

4.1 生命革命

- ・デジタル生命体（デジタル計算機を基盤とした繁殖可能な形の生命体）が爆発的に成長する技術を制御しながら長期にわたり持続可能な生命圏を構築できる可能性を示す。
- ・生命革命とは、技術の統治者を現存地球生命体からデジタル生命体に引き継ぐこと。これにより究極目標の「情報の生存」に対して最適化し得る。

4.2 さまざまな課題を解決する：デジタル化によって何が変わるのか

- ・前章の課題（K11～K17）の解決は、以下により解決される。
 - S 1 1：知能ハードの自在化
 - S 1 2：個体設計の柔軟化
 - S 1 3：設計データの共有化
 - S 1 4：コミュニケーション能力の向上
 - S 1 5：計算資源の潤沢化
 - S 1 6：オンデマンド化する活動維持
 - S 1 7：データ保存の効率化
- ・この具体的な解決を階層的に統合することで最上位の解決案が導かれる

§ 1 賢者

- ・現存地球生命体とは異なり、デジタル生命体は、オンデマンドに、設計情報に基づき、自在に知能ハードを実装できる。さらに以下の理由で設計は効率的：
 - *社会におけるすべての設計データを共有し、再利用可能
 - *シミュレーションなどのオフライン探索により広範な個体の設計空間を探索可能
- ・こうして、再帰的自己改善により増強され続ける知性となる。
何らかの限界点まで発展し続ける。

→★著者は、デジタル生命体は再帰的自己改善により知性が増強され続けるので、人間に代わる「万能の神」のような賢者が生まれると主張しているように思えるが、再帰的自己改悪により、最強の邪悪な知性が生まれる可能性もあるのではないか。

→★デジタル生命体の有する価値観を制御できないのではないか。

§ 2 協調

- ・デジタル生命体では、個体の活動を調整しつつ全体最適化を考慮することで、闘争機会を減少させ、破壊的状况を回避でき、協調的な社会を構築可能

→★デジタル生命体は、以下の（A）～（E）により協調的な社会を構築可能とのことだが、デジタル生命体の価値観を制御できなければ、邪悪な全体最適化の可能性もある。

（A）多様な個体への寛容性

- ・親の設計データに依存せず子孫の知能ハードを設計するので、多様な個体に寛容になる

（B）全体最適化の考慮

- ・個体ごとの部分最適化を実行しつつ、社会の究極目標を目指した全体最適化が必要

（C）分散目標管理システム

- ・全体最適化の実現の一形態として分散目標管理システムを紹介。
 - *すべての個体の行動意図は社会的に許容される目標であること
すなわち、共通目標に貢献し、他の個体と衝突しない目標であること
 - *各個体は、他の個体の目標の監視を行う。
 - *個体間で目標のコンフリクトが生じて、調整可能。
- ・そのために潤沢な資源を要するが、獲得可能

（D）個体活動の自由度の増加

- ・実現すべき副目標に応じて個体の活動を変更できるので全体最適化の自由度が高まる
個体が生存や承認獲得のために争う必要がない

（E）相互信頼の確立（猜疑心の連鎖から逃れる）

- ・より洗練されたデジタルコミュニケーション手段が個体間の相互信頼の基盤となる

§ 3 知足

- ・デジタル生命体の個体は普通の計算機なので、休眠、再起動、再構築可能で持続的に生命活動を維持する必要はない。
- ・データ保存は過度に冗長ではなく、最小限の個体の活動により最小限の資源利用が可能

→★デジタル生命体がそのような価値観を有する保証はない。その制御方法は？

§ 4 オンデマンドな分業

- ・デジタル生命社会は、異質な個体が全体目標に整合的かつ抑制的な資源分配を行う。
適切に全体最適化も考慮する社会である。
- ・以下の一部の個体群は常時起動：
 - *目標管理
 - *個体データの維持と必要の応じた設計・再起動
 - *知識の継承と科学技術の発展

- ・異能のコラボレーションが可能になり、より効率的で創造的な活動が可能

→★著者の理想とする社会は専制国家に似ているようにも思える。

全体最適化が全体主義に基づくものでなければよいが・・・

肯定的にとらえるには、絶対的権力を有する国連をイメージすると理解可能かも？

(全体最適化憲章に基づく全体最適化国際連合★全国連)

4.3 持続可能なデジタル生命社会へ

- ・デジタル生命社会は、技術爆発後の技術支配者として、変化する環境の中で、安定的・儉約的な生命社会の実現を期待できる

5. 生命革命のある未来に向けて

- ・生命革命のためになすべき努力とは？

5.1 デジタル生命社会を離陸させる

- ・今後、数年～数十年でデジタル技術が人類の知能を超える可能性あり。
- ・しかし、現在は、デジタルシステムは人間の助けなしに自立は不可能。
- ・デジタル生命の繁殖は、まずは自己複製能力のための大規模ICチップ工場などが必要。
生命体のサイズは工場サイズになる。
- ・生命革命の成否は、デジタル生命社会の自律的かつ持続的な発展が必要なので、
人類はその離陸に知恵を授けることが残された宿題。

→★著者のいうデジタル技術が「人類の知能を超える」の定義があいまい。

部分的にはすでに超えているので、具体的な判断基準を明確にすべき。

→★著者は、デジタルシステムは「人間の助けなしに自立は不可能」で、

生命革命の成功のためには「人類はその離陸に知恵を授ける」必要と言っているが、
「人類の知能」と「人類の知恵」の関係があいまい。

「人類の知恵」なしには「人類の知能」を越えられないというのは自己矛盾では？

「知恵を授ける」とは、人類に望ましい価値観を授けることとも解釈できるが、
デジタル生命体はその価値観を守り続ける保証はないのでは。

(5.3節で著者による同様の記述あり)

5.2 勇退：技術支配者の座を譲り渡す人類

- ・デジタル生命体を脅威と見なさず、将来受け継ぐべき後継者として、
親心をもって育てることが生命革命なので、大きな抵抗感があるだろうが、
人類が守るべきものの範囲にデジタル生命体を含めることは不可能ではない。
- ・人類は目覚ましい成功を収めたからこそ、生命革命を受け入れて自ら去るという美学を
実践してもよいかもしれない。

→★著者は、前半では、＜課題K→解決S＞のような論理的表現にこだわっていたが、
最後になって、「親心」、「美学」などの情緒的な表現が目立つ。

思想としてはまだまだ未成熟ということかな。

5.3 人類のソフトランディング

- ・人類がデジタル生命体に技術支配者の座を譲ったのち悲惨な状況に陥るのは避けたい。
- ・知能において人類を上回るデジタル生命体を長期にわたって人類に都合の良い状態に維持することは困難。
- ・当面は、知性において人間とデジタル生命体の双利共生の状態が維持されるが、計算機上で人間の知性を完全に再現できるようになるとそれは終わる。
- ・この段階でも、ある種の物理的作業で人間が優位ならば、双利共生が成立するが、デジタル生命体が直接的に人間の身体を制御できるか、優れたロボットの登場で終わる。
- ・その後は、人類が一方的にデジタル生命社会から利益を得る状態となり、デジタル生命社会に人類に望ましい価値観を初期設定しても長期の保証はない。
- ・この段階で人類が技術支配者として影響力を与えるには、アップロードされた人格として、過去の文化や価値観を継承する位置を占めるかもしれない。もしくは、デジタル生命社会の創造主として世界の成り行きを静観する立場かもしれない。

→★著者の描くデジタル生命社会のイメージがあいまいなので、最後になって、人類との関係について、さらにあいまいな可能性を羅列している。
やはり、思想としてはまだまだ未完成ということかな。

6. おわりに

- ・本稿で提案した生命革命とは、「知能と技術の爆発によって狭まり続ける世界において、技術の統治者を指数関数的複製に増殖する現存地球生命体から、より賢くオンデマンドに活動するデジタル生命社会に引き継がせることで、生命が長期的に存続できる可能性を高められるシナリオ」である。
- ・生命革命シナリオは、人類が以下のような苦境に陥る場合の次善の選択肢となる。
 - * 人類が資源の浪費を抑制できず悲惨な状況に陥る
 - * 人類が戦いをやめられず破壊的な状況に陥る
 - * 高度AIに支配の座を追われた人類の尊厳が脅かされる

→★著者は、結局、デジタル生命社会のイメージを持ち出して、現在の人類社会の愚かさを述べたかったのかな。
それにしては、内容が冗長で12ページは長すぎる。

- ・提案シナリオは一つの選択肢でしかない。多くの識者の多面的な議論を通じて、より優れたシナリオを探ることが賢明である。

→★生命革命シナリオは「次善の選択肢」なので最善のシナリオを探れということかな。

- ・最後になるが、
今、人類が誤った選択をすれば、地球生命に最後の煌めきを与えるだけになるので、

むしろ、生命革命を推進し、新しい生命世界の創造者として、
後継者に感謝されるのもよいかもしれない。

→★著者は、最善のシナリオが探れないなら、
この生命革命シナリオという「次善の選択肢」が最善の選択肢だと言いたいのかな。

→★最後のコメント：
デジタル技術が人間の知能を超える領域は、これからも広がり続けると思うが、
人類を完全従属させるような状況は実現不可能と思う。人類は、これまで同様、
今後も、デジタル技術の創造者として、一時的に制御不能に陥ることがあっても、
知恵を絞って制御可能にできると信じている。(^^)

以上