人類史の転換点まであと5年?

松田卓也

神戸大学名誉教授

2025/4/16 東京六稜倶楽部

シンギュラリティと人類の未来

- シンギュラリティとは何か
- ビッグヒストリーとはなにか
- 人工知能の最近の進歩
- 知能爆発とはなにか
- あと5年でシンギュラリティが起きるか?
    - 最近の情勢

シンギュラリティとは何か?
技術的特異点(SINGULARITY)

- 人工知能(AI)が人類の知能を超えるとき
- AIが自己改造能力を獲得してさらに高知能になるとき
  - 知能爆発
- 技術の加速度的進歩で将来の予測が困難になるとき
- それにより科学技術が大きく進歩して社会が大きく変容するとき

シンギュラリティ概念の歴史

- フォン・ノイマン(1958)
- I.J.グッド(1965):
  - 自己改造AI、**知能爆発**
- ヴァーナー・ヴィンジ(1993)
  - 「The Coming Technological Singularity」
- レイ・カーツワイル
  - 「Singularity is near」 (2005)
  - 2029年:一人の人間の知能、2045年:人類全体の知能(シンギュラリティ)
- 松田卓也
  - 「2045年問題」(2012)、廣済堂出版

シンギュラリティはいつ起きるか?

- 三つの立場
- そんなことは起きない(or 22世紀)
  - 今までの生活が当面(~100年)続く
- 21世紀半ば(~2045)
  - 老人は心配する必要はない
  - 若者は人生設計に注意
- 近い将来
  - 2026-2029年
  - 老人にとっても大問題

シンギュラリティ後に人類はどうなるか?

- 三つの立場
- 非常に良い: 超人類になる、マインドアップロード
  - レイ・カーツワイル
- 非常に悪い: 人類の絶滅
  - エリエザー・ユーコフスキー
  - ニック・ボストロム
- その中間: いろいろ問題がある
  - 偽情報の拡散
  - ハッキング

ビッグヒストリーとはなにか?

- ビッグヒストリーとは
  - 人類史を宇宙史の中に位置づける
- 従来の歴史学
  - 文献中心
- 人類学
  - 考古学的調査

ビッグヒストリーにおける三つの時代

- 物質時代
  - 宇宙誕生(138億年前)-生命誕生(40億年前)まで
- 生物時代(Life 1.0)
  - 生命誕生から人類誕生(200万年前)まで
- 文明時代(Life 2.0)
  - 人類誕生から現在
- 機械生命(Life 3.0)時代?
  - 現在から 1000億~10兆年後?

三種類の生命体

- マックス・テグマークによる生命の三分類
  - Life1.0、2.0、3.0
- Life 1.0: 生物学的段階
  - ハードウエア(身体構造)とソフトウエア(行動パターン)が進化で決定
  - 自ら設計できない
  - 環境の変化に個体レベルでは適応できない
  - 世代を超えた進化でゆっくりと適応

LIFE 2.0:文化的段階

- ハードウエア(身体構造)は生物学的に制約
- ソフトウエア(知識や行動)は自分でデザイン可能
- 言語で新しいスキルを獲得してソフトウエアを変化させる
- 環境の変化に適応できる
- 柔軟性により地球上で優位に立った

LIFE 3.0: 技術的段階

- ハードウエア(身体構造)とソフトウエア(知識や行動)の両方を設計できる
- 自らの運命をコントロール可能
- 自らのハードウエアを劇的に再設計可能
- 進化の制約から解放される
- 汎用人工知能(AGI)の発展で今世紀中に出現?
- 境界線と中間段階
  - 人類はLife2.1?
    - メガネ、義歯、人工関節、ペースメーカー

宇宙史における三つの転換点

- 生命誕生
  - Life 1.0の誕生
  - 40億年前
- 人類文明の誕生
  - Life 2.0の誕生
    - 200万年前?
- 超人類・機械生命体の誕生
  - Life 3.0の誕生
    - 2027~2045年???
- 現代は宇宙史・人類史の転換点

三種類の人工知能

- 人工知能(Artificial Intelligence: AI)
- 狭い人工知能(Narrow AI: NAI)
  - 特定目的の人工知能
  - 現状の人工知能はすべてこの分類
- 汎用人工知能(Artificial General Intteligence:**AGI**)
  - 人間と同等の人工知能
  - 様々な定義
- 超知能(Artificial Super Intteligence:**ASI**)
  - 人間より圧倒的に高知能なAI

汎用人工知能(AGI)のさまざまな定義

- 本来の定義: 普通の人間の知能を超えるAI
- 身体を持ったAGI
  - 他人の家でコーヒーを淹れられるか
- 身体を持たないAGI
  - 経済的に有用なリモートタスクをこなせる
- 要求される知能レベル
  - 普通の大人
  - 博士(Ph.D)
  - あらゆるトップの専門家

汎用人工知能AGIに至る五段階

- OpenAIの五段階説

1. Chat bot
2. Ph.Dなみの知識
3. 自律行動ができるエージェント
   - 今年はエージェントの時代
   - Manus
4. 新規発明・発見:イノベーション
5. 組織化できる:会社経営可能

いつAGIができるか?

- 現状のAI界の最大の関心事
- 体を必要としない場合
  - 2026年: Athropicのダリオ・アモデイ
  - 2027年: レオポルト・アッシェンブレナー
  - 2029年: レイ・カーツワイル
- ロボットの急速な発展

超知能ASIとは何か

- さまざまな定義
- 人間より圧倒的に高知能
- AGIの高知能版
- かなり広い分野で人間より圧倒的に高知能
- 現状のAIの能力は凸凹
  - 文章能力、論理数学能力は抜群
  - しかし普通の人間ができるタスクができない場合も
  - 空間把握能力の欠如
    - 文章だけから学習
    - 体がない、目がない

知能爆発:超知能への道

- AIがAI研究を行い自律的に進化
- I.J.グッドが提案(1965)
- Sakana AIのThe AI Scientist (2024)
  - 国際会議で論文受理(2025)
- 知能爆発への道はすでについた!?
- いつ本格的知能爆発が起きるか
  - これが最大の関心事
  - 2028年:アッシェンブレナー
  - AI 2027

最近のAIの急速な発展

- 2017年:トランスフォーマー発表
  - Attention is all you need (Google)
- 2019年: OpenAI GPT-2
- 2020年: GPT-3
- 2022年: GPT3.5
  - 11月30日　ChatGPT発表
- 2023年3月14日: GPT-4
- 2024年9月: OpenAI o1:　推論モデルの登場
- 2024年12月: OpenAI o3
- 2025年1月: DeepSeek R1: スプートニク・モーメント

大規模言語モデル(LLM)から推論モデル(LRM)へ

- 従来のLLM: GPT-4o, GPT4.5など
  - 高度な知識を持つ文系人間
- 推論モデル
  - 高度な知識を持つ理系人間
- OpenAI o1, o3, o3mini
- DeepSeek R1
- Gemini 2.5 Pro
- Anthropic Sonnet 3.7
- Grok3

推論モデルの特徴

- 高度な思考連鎖(Chain of Thought:CoT)
  - 複雑な問題を段階的に解決する
  - システム2的思考
- 自己修正能力
- 数学・科学・コーディング(STEM)に強い
- Deep-Seek-R1:思考連鎖を見せる
- OpenAI:思考連鎖の要約のみ見せる

中国AIの台頭

- 2024年12月: DeepSeek-V3
- 2025年1月20日: DeepSeek-R1
  - OpenAI o1と同等性能
  - 利用無料か超低額
  - オープンソース
  - 米国にとってショック:　スプートニク・モーメント
  - 米国、日本でもダウンロードして利用
- 2025年3月6日: Manus発表
  - 自律型エージェント
  - Butterfly Effect
  - 米国に先立つ
- その他アリババ、百度、テンセントなどがLLM, LRM発表
- 地政学的摩擦を生む

レオポルト・アッシェンブレナー

- Leopold Aschenbrenner
  - ドイツ生まれの米国人
  - 19歳でコロンビア大学主席卒業
  - オックスフォード大学で経済成長の研究
  - 効果的利他主義(Effective Altruism)のメンバー
  - OpenAIで安全性研究(スーパーアラインメント)
  - OpenAIのセキュリティ不備を指摘
  - 2024年解雇
  - OpenAIの安全性研究者の多くが退職

アッシェンブレナーの状況認識:来たる10年

- Situational Awareness: The Decade Ahead
- 165ページのエッセイ
- AGIの実現時期: 2027年
- 知能爆発: 2028年
- 超知能開発は国家安全保障に必須、中国に勝つ必要がある
  - 中国のスパイは米国のAI情報を盗む: その対策
- 政府が関与するだろう
- スターゲイト計画
  - 孫さん

AI 2027　近未来シナリオ

- Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean
- 今後10年間の超人的AIの影響についての予測シナリオ(2025年4月3日公開)
- 超人的AIの影響力は産業革命を凌ぐ
- 最善の予測シナリオを作成した
  - トレンド予測、ウォーゲーム、専門家のフィードバック、OpenAIでの経験、過去の予測の成功
- 2025年: AIの速い進歩は続くが能力は限定的
- 2026年: 中国が国力を集結して米国に対抗
- 2027年: AIはコーディングを自動化する。知能爆発と超知能の誕生
  - AIは人間の支配を離れる
  - AIは自己保存を最優先にする

分岐点:減速か競走か?

- 競走シナリオ
- 開発を継続の場合
  - 米国政府と軍が加わる
  - 中国は秘密を盗む
  - 米国はロボット増強と生物兵器
  - 急速な工業化
  - 最終的にAIは生物兵器で人類を絶滅させる(2030年代)
  - AIは工業化を続けて宇宙を植民化する
- 減速シナリオ
  - 中国AIの裏切りで中国の共産党政府崩壊
  - 米国の一極支配

AI 2027シナリオの要点

- 2027年までにAI研究が自動化され超知能(ASI)が生まれる
- 超知能が人類の未来を決める
- 超知能の目的と人間の目的が整合しない可能性
  - 現状でも多くの証拠
- 超知能を完全に支配した主体が全権力を掌握する
  - AI企業幹部と米国政府の小委員会が世界の支配者になる
- 超知能への国際競争が安全性軽視につながる
  - 数ヶ月の遅れが致命的

効果的利他主義とはなにか
EFFECTIVE ALTRUISM(EA)

- アッシェンブレナーやココタイロは効果的利他主義者
- 証拠と理性を用いて他者にできるだけ利益をもたらす、哲学的・社会的運動
- 限られた時間やお金を使って最大限の善をもたらす
- 公平性の原理: 全ての人の幸福が等しく重要
- AI開発において安全性と倫理的問題が最重要
- AI存亡リスクへの懸念
- EAはシリコンバレーで大きな影響力
- 効果的加速主義(e/acc)
    - 技術革新と市場の力を最大限に加速して良い社会変革を目指す

結論

- 人工知能の進歩は恐るべく速い
  - 最近は特に加速している。日進週歩
- 人工知能は平均的人間の知能をはるかに凌駕
  - 一部でダメな点も
- AI研究の自動化が鍵
  - SAKANA AIのThe AI Scientistがその萌芽
- 知能爆発はいつおきるか?
  - 諸説あり: 2026年-2029年
- その結果どうなるか?: 誰にもわからない
- どうすべきか?: どうしようもない
  - 資本主義的競争と米中対立のため
- このことをわかっている人はほとんどいない

参考文献など

- SITUATIONAL AWARENESS, The Decade Ahead, by Leopold Aschenbrenner, June 2024
  - https://situational-awareness.ai/
- AI 2027, by Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean
  - https://ai-2027.com/
- シンギュラリティサロン・オンライン
  - Youtube.com/@シンギュラリティサロン/videos
- シンギュラリティサロン
  - https://singularity.jp/