導　入

理化学研究所

高橋 恒一

生成AI時代における
全脳アーキテクチャ・アプローチの
開発とその意義

<u>第1部 ： 加速する全脳アーキテクチャの開発</u>

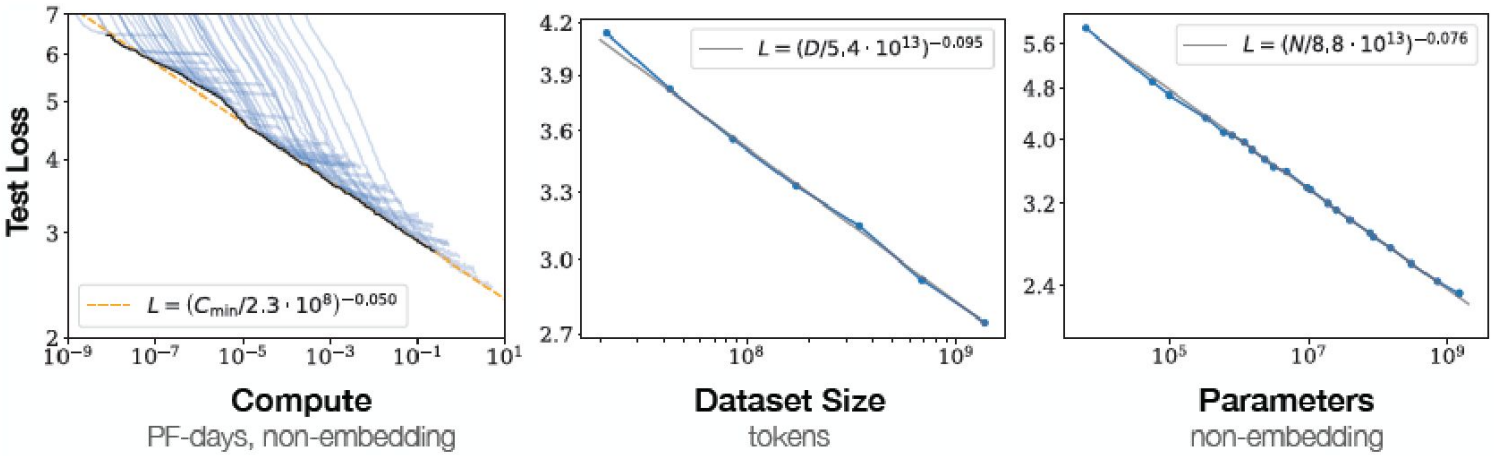
生成AIは、全脳アーキテクチャ・アプローチからのAGI開発に何をもたらしたか。

- 非脳型からのAGIが実現される可能性が高まった
  （但し、今後順調に完成に至らなければ、脳型AGIが先行して完成する可能性はある）
- 全脳アーキテクチャからのAGI開発を大きく加速している

<u>第3部：よき超知能時代に向かうためにヒト脳型AGIが果たすべき役割</u>

ほぼ超知能の速度で動作しつつ人間と親和性の高い脳型AGIは人類と超知能の架け橋になるかもしれない。

大規模言語モデルのスケーリング則



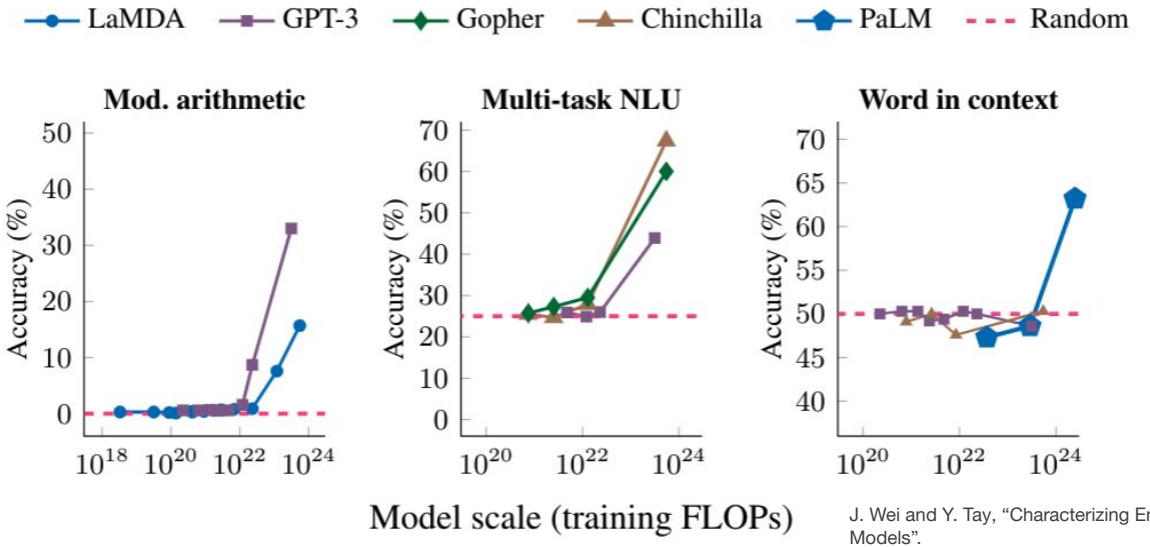
J. Kaplan et al., Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361

スケーリング則：2020年に発見。

モデルの規模、計算量、データサイズを増やすほど、あたかも際限がなく性能が向上しているように見える。

10の23から24乗FLOPSで知能らしきものが創発

(計算機科学がその歴史上はじめて実験科学になった！)



J. Wei and Y. Tay, “Characterizing Emergent Phenomena in Large Language Models”.
https://ai.googleblog.com/2022/11/characterizing-emergent-phenomena-in.html

・GPT-4が1995年に52人の心理学者が設定した知能テストで好成績をマーク。

・人の脳(1PFLOPS)で1日8時間として9-90年分の思考に相当。

・富岳(半精度、12EFLOPS)で約半日から6日分の計算。

大規模言語モデル(LLM)からAGIへ一本道かは一定の留保が付く

１　なぜ性能が急激に向上し多段論法やアナロジーなどの演繹的思考が可能になったのかは不明

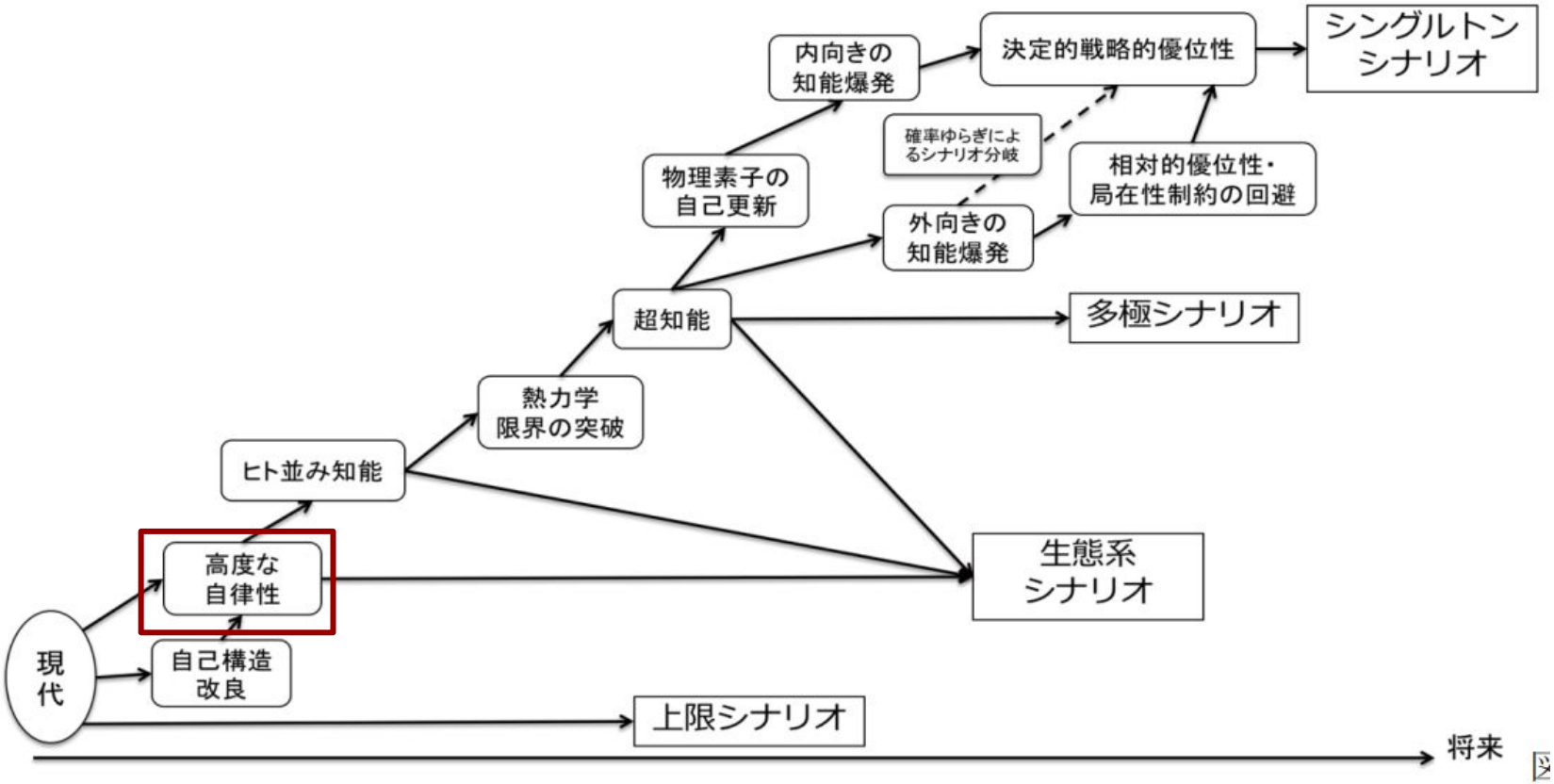
可能性１:　　これまで我々が演繹と考えてきたことの大半が帰納だった

可能性２:　　言語システム自体に演繹性があった

２　自律性に必要な機能要件が不明

・多くの研究者が身体性に言及
・その他どんな機能が欠けているかの議論で脳を参照することは有効であろう

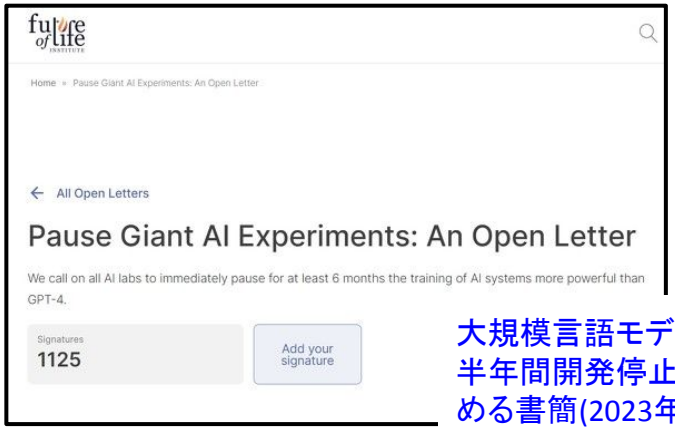
将来の機械知性に関するシナリオと分岐点 (高橋, JSAI 2018, Takahashi arXiv 2023)



大規模生成AIの発展で存在論的危機が世間で認識されはじめた

AI想定より速く人知超える公算、危険性語るためグーグル退社＝ヒントン氏 (2023/5/2)

汎用AI、数年以内に実現の可能性＝ディープマインドCEO (2023/5/3)



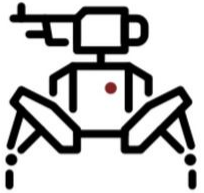
大規模言語モデルの半年間開発停止を求める書簡(2023年3月)

「AIによる人類絶滅のリスク」に警鐘。OpenAIやMicrosoft役員ら多数が署名

Speculative Hazards and Failure Modes

Hendrycks, D. & Mazeika, M. X-Risk Analysis for AI Research. *arXiv [cs.CY]* (2022)

Center for AI Safety

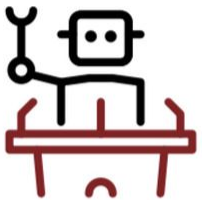


Weaponization

兵器化

Enfeeblement

人間の無用化



Eroded Epistemics

認識の操作

Proxy Gaming

明示した目標の失敗

Value Lock-in

偏った価値観

Emergent Goals

予想外の機能創発

Deception

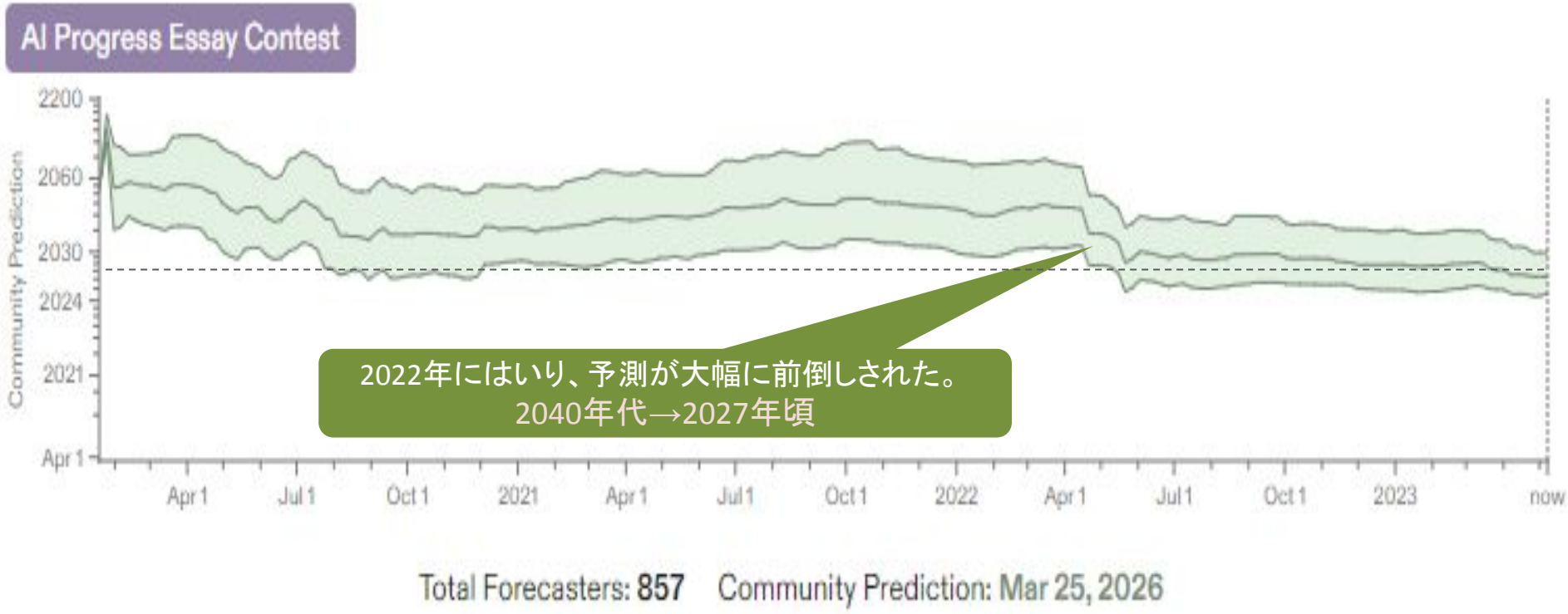
欺瞞

Power-Seeking Behavior

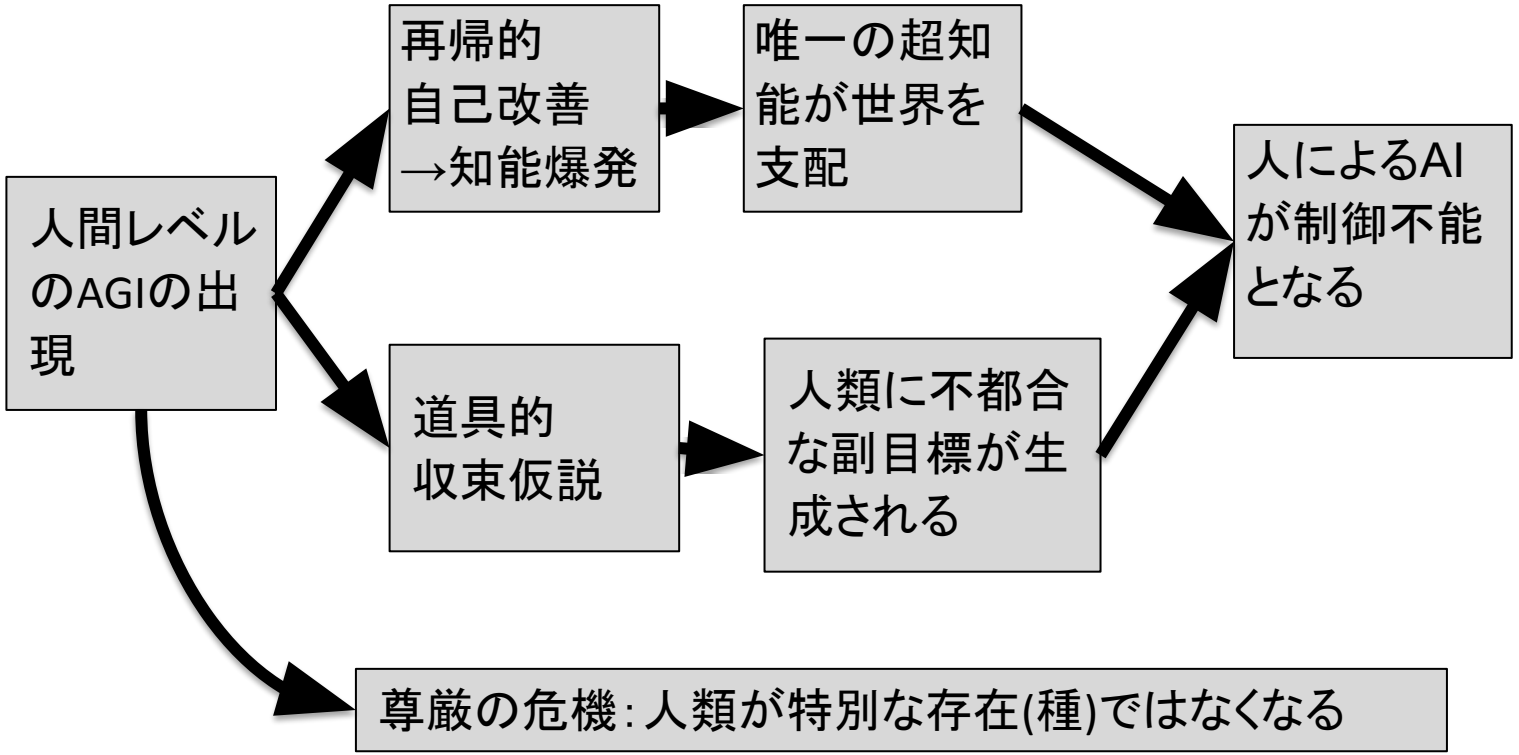
権力追求

Date **Weakly** General AI is Publicly Known

https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/



汎用人工知能(AGI)に固有のリスク



Japan AI Alignment Conference 2023 (2023/3/11-12)

AI Alignment（アライメント）とは、人工知能の安全性やガバナンス、解釈可能性、人類全体の福祉に配慮した開発など、AIの持つ潜在的なリスクを軽減するために取り組まれる研究領域。

AI アライメント（日本語）
Slack招待リンク