

いま全脳アーキテクチャを進める意義

山川 宏

NPO法人 全脳アーキテクチャ・イニシアティブ
代表

本講演のポイント

- WBAアプローチは、脳を参照してAGI開発を進める手法
 - 機械学習AI技術とは異なり、高い対人親和性を持ちうる
- WBAIは本アプローチからのAGIのオープンな開発を促進
 - 最初のAGIが出現時に、**ヒト脳型AGIの仕様書(WBRA)**を存在させることを重視
 - 理由:ヒト脳型AGIは「人類と超知能の架け橋となる可能性」や「脳に基づく解釈可能性」を持ち、懸念される高度AIによるリスクへの対処の選択肢となりうる
- 最速でAGI出現が想定される2027年までに、上記仕様書の α 版を完成するロードマップ(仕様書に基づく実装はAIで自動化される想定)
 - この取り組みは、世界的に見ても独自性が高い

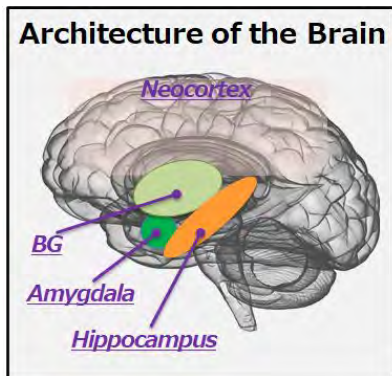
アジェンダ

1. 全脳アーキテクチャ(WBA)とNPO法人WBAI
2. 汎用人工知能(AGI)を取り巻く動向
3. WBAの意義: AIリスク対策に選択肢を追加する

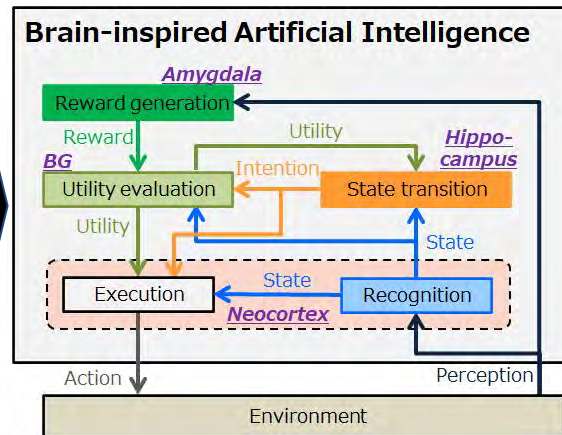


全脳アーキテクチャ(WBA)と NPO法人WBAI

脳全体のアーキテクチャに学び人間のようなAGIを創る(工学)



- ① 脳の各器官を機械学習モジュールとして開発
- ② それらを統合した認知アーキテクチャの構築



本アプローチの利点

共同開発の基盤として脳の構造を利用して開発を加速

- 人間に関わる諸分野の科学知見を活用できる
- 開発の発散を防げる合意しうる認知アーキテクチャ
- 脳を参照して設計空間を限定する

人と親和性の高いAGIを作れる

- 対人インタラクションが必要な応用
- 精神疾患などのモデル化など医療応用
- 人類と超知能の架け橋となる可能性

脳を参照して設計空間を限定することでAGI開発を加速

思考空間の爆発: 広範な知識を柔軟に活用する組み合わせの空間(思考空間)は膨大

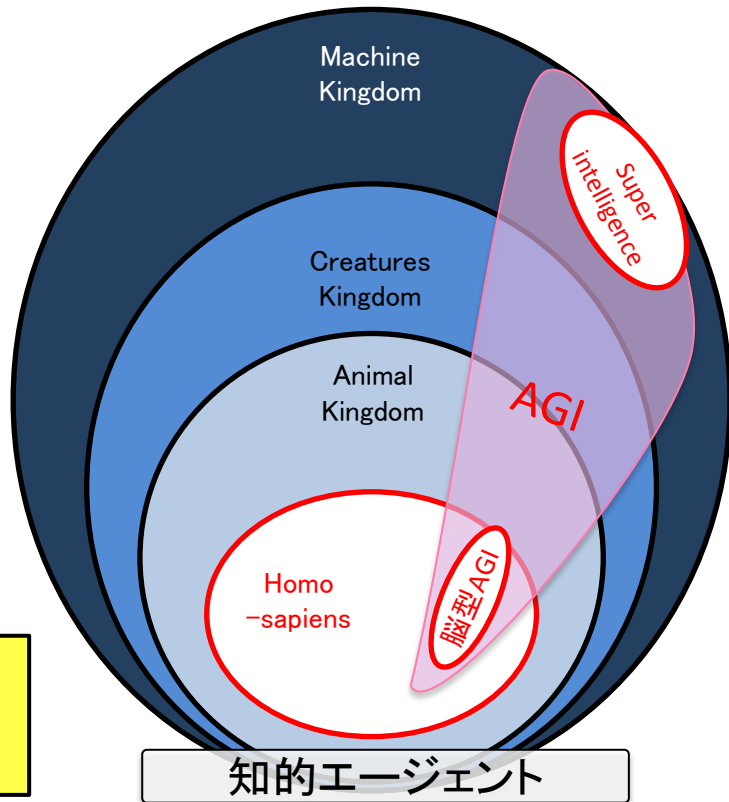


思考空間の爆発を抑制する制約として、**脳アーキテクチャ**をAGI設計に利用

現状の基盤モデルからの量的拡大ではAGI到達に障壁がある場合に先行する可能性がある。

脳アーキテクチャに制約されない機械学習知能が、ヒト脳型AGIよりも強力な超知能となる。

設計空間の図



全脳アーキテクチャ・イニシアティブ(WBAI) Since 2015

良くも悪くも人類の福利厚生に対して大きな影響を与えるのは知能。

ヒトのあらゆる知的能力を凌駕する汎用人工知能(AGI)の実現は急速に現実味を帯びている。

「人類と調和した人工知能のある世界」というビジョン実現のため、脳型からのAGIのオープンな開発を促進している(2030年目処)。

現在は、2027年までにヒト脳型AGIの設計情報(WBRA)の初期版完成に注力



NPO法人WBAI代表

山川 宏

全脳アーキテクチャ・イニシアティブ(WBAI)の歴史

- 2015年8月: WBAI設立
(WBAアプローチにより2030年頃のAGI実現を目指す)
- 2016年: NPOとしての開発「**推進**」軸の明確化
- 2017年: ビジョン "人類と調和する人工知能のある世界" を発表
- 2020年頃: 脳のリバーズエンジニアリングによる脳型ソフトウェアの設計手法を確立(BRA駆動開発と呼ぶ)
- 2022年: LLMの活用を開始し、BRA駆動開発が加速する。
- 2023年: 人類の未来に貢献できるヒト脳型AGIの仕様書(WBRA)を共同開発・公開する方針を決定。
 - 2027年までにWBRAの初期版を完成させ、その後アップデートする計画

WBAIの推進体制

顧問: 銅谷賢治 (OIST) 北野宏明 (システム・バイオロジー研究機構)
富田勝 (慶應義塾大) 森川博之 (東大)
中島秀之 (札幌大) 岡ノ谷一夫 (帝京大)

理事: 山川宏 (代表) 松尾豊 (副代表/東大)
高橋恒一 (副代表/理研) 荒川直哉

監事: 浅川伸一

会議: 総会、理事会、その他委員会

最近の主な学術成果

1. T. Nakashima, S. Otake, A. Taniguchi, K. Maeyama, L. El Hafi, T. Taniguchi, H. Yamakawa, Hippocampal formation-inspired global self-localization: quick recovery from the kidnapped robot problem from an egocentric perspective. *Front. Comput. Neurosci.* **18** (2024). 中島氏より発表
海馬体を模した頑健なナビゲーションモデル
2. Yamakawa, H., Fukawa, A., Yairi, I. E., & Matsuo, Y. (2024). Brain-consistent architecture for imagination. *Frontiers in Systems Neuroscience*, *18*.
<https://doi.org/10.3389/fnsys.2024.1302429> 想像力を支える脳アーキテクチャ
3. Yamakawa, H., Tawatsuji, Y., Ashihara, Y., Fukawa, A., Arakawa, N., Takahashi, K., & Matsuo, Y. (2024). Technology roadmap toward the completion of whole-brain architecture with bra-driven development. In *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4792766> 後ほど説明

関連成果は人工知能学会の研究会優秀賞を受賞

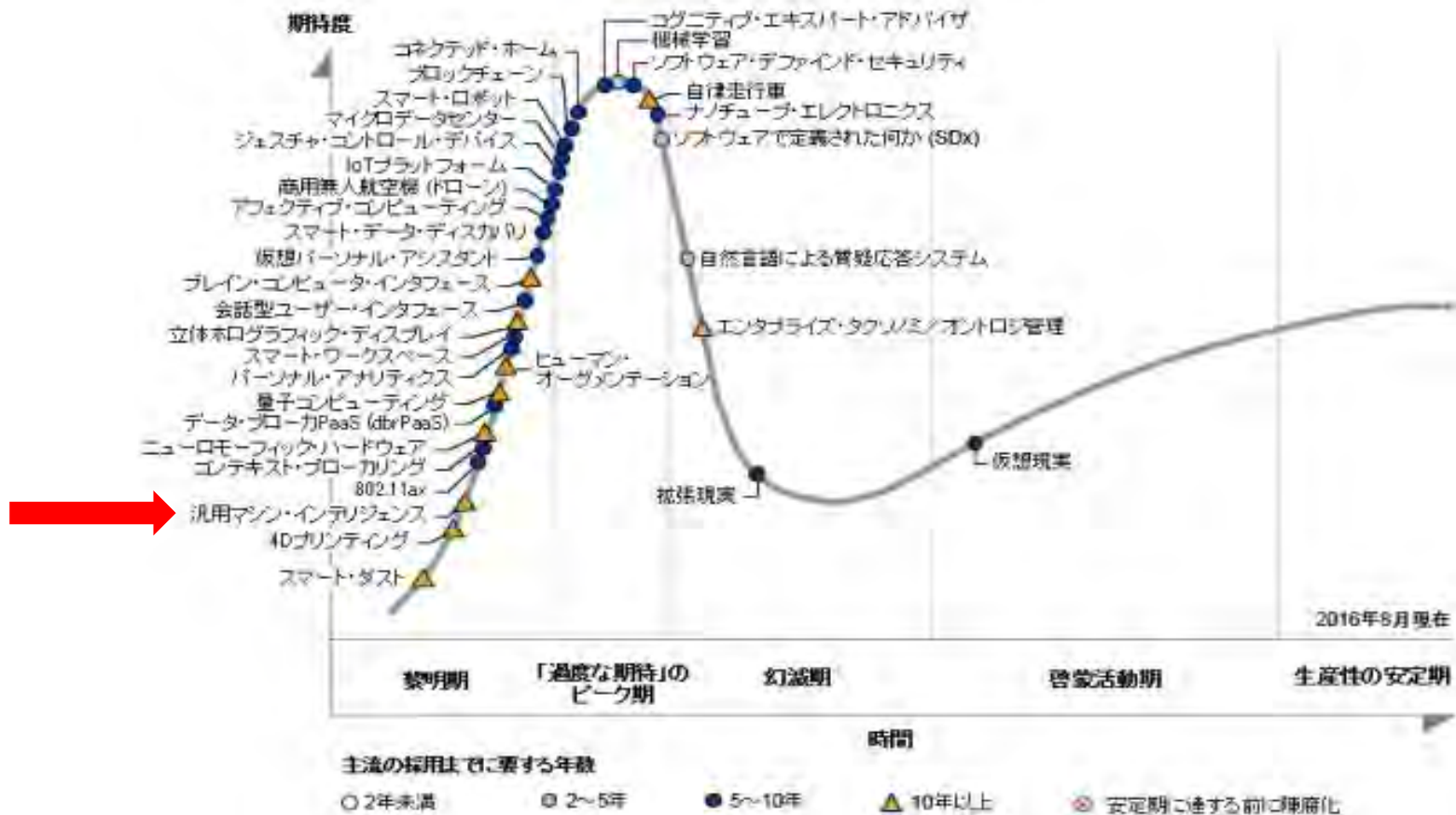
全脳アーキテクチャの技術ロードマップ

AGIを取り巻く動向



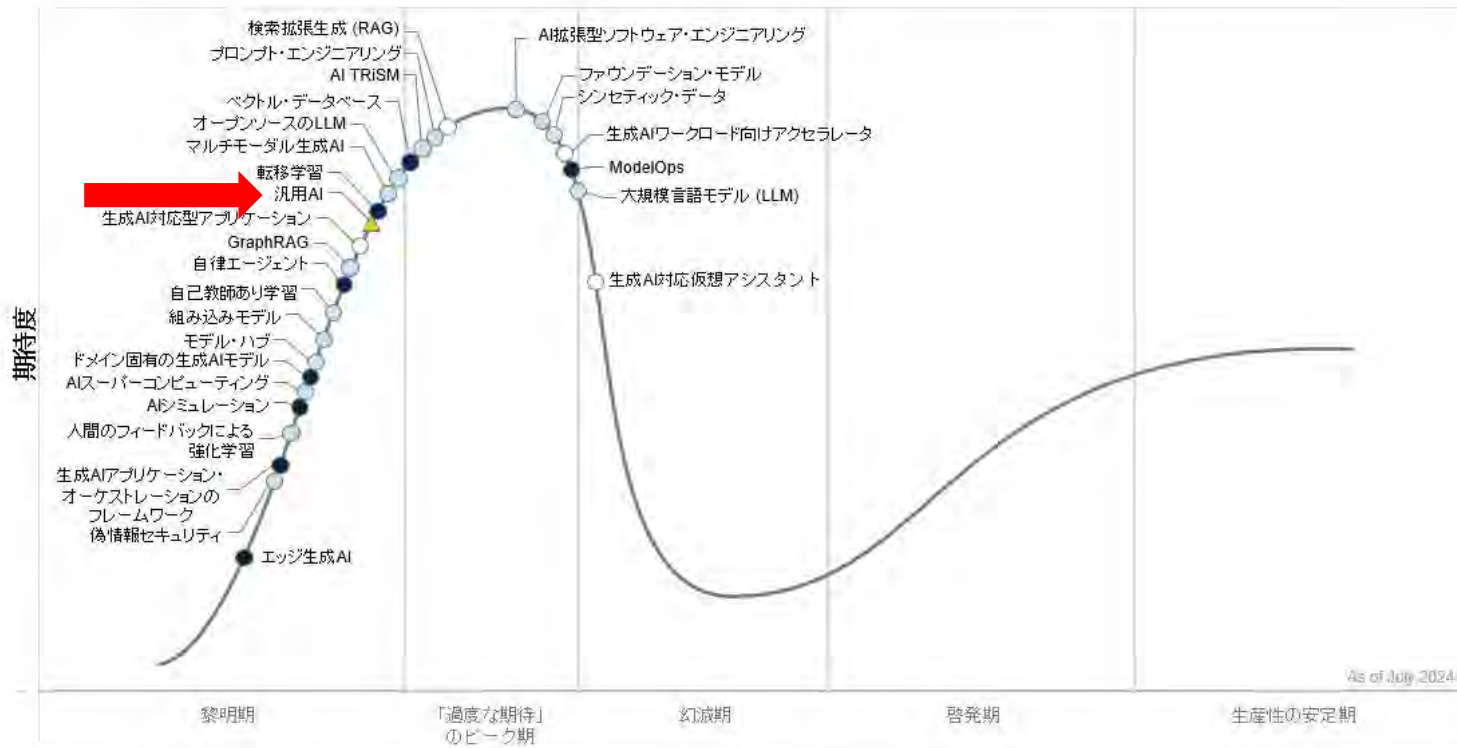
生成AIのハイプ・サイクル:2016年

出典: Gartner (2016年8月)



生成AIのハイプ・サイクル: 2024年

出典: Gartner (2024年9月)



主流の採用までに要する年数: 2年未満 2~5年 5~10年 10年以上 ⊗ 安定期に達する前に陳腐化

早ければあと数年でAGIが実現するという見通し

本年6月、元OpenAI研究者L. Aschenbrenner氏が、"**SITUATIONAL AWARENESS - The Decade Ahead**"で示した見通し。

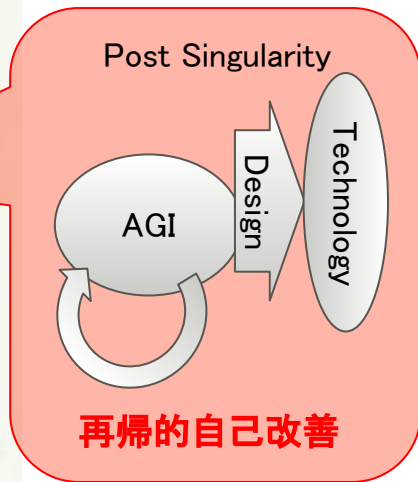
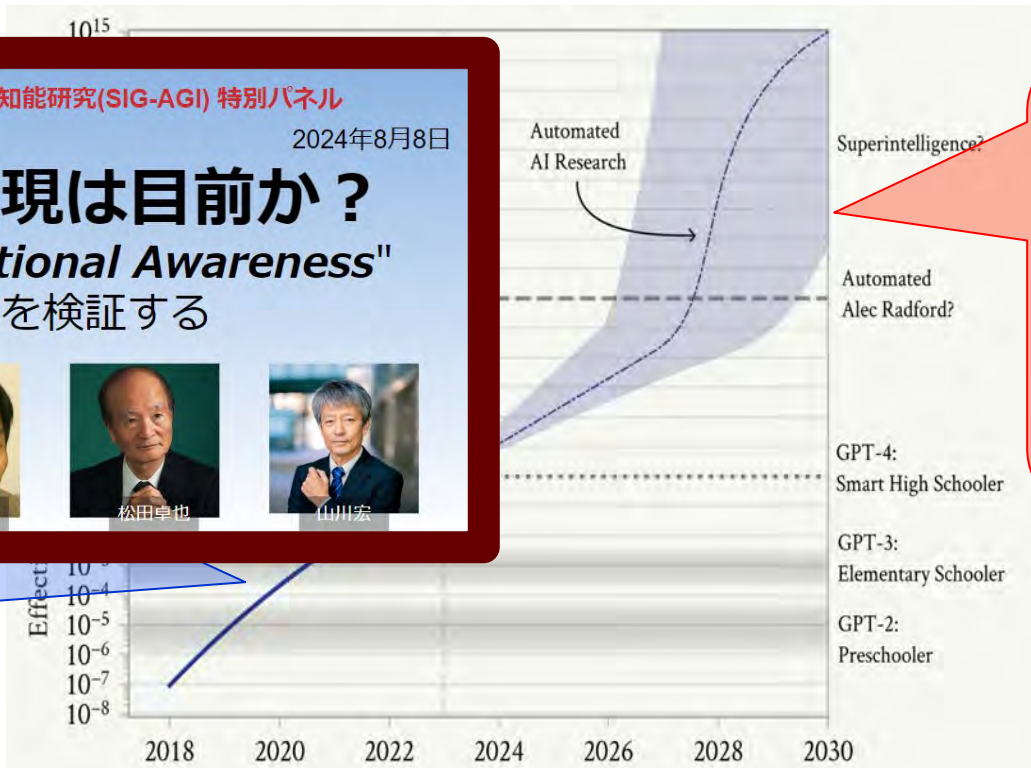
人工知能学会 第27回汎用人工知能研究(SIG-AGI) 特別パネル
2024年8月8日

AGIの実現は目前か？

話題の"*Situational Awareness*"
記事を検証する



有路翔太 bioshock
中川裕志
松田卓也
中川憲



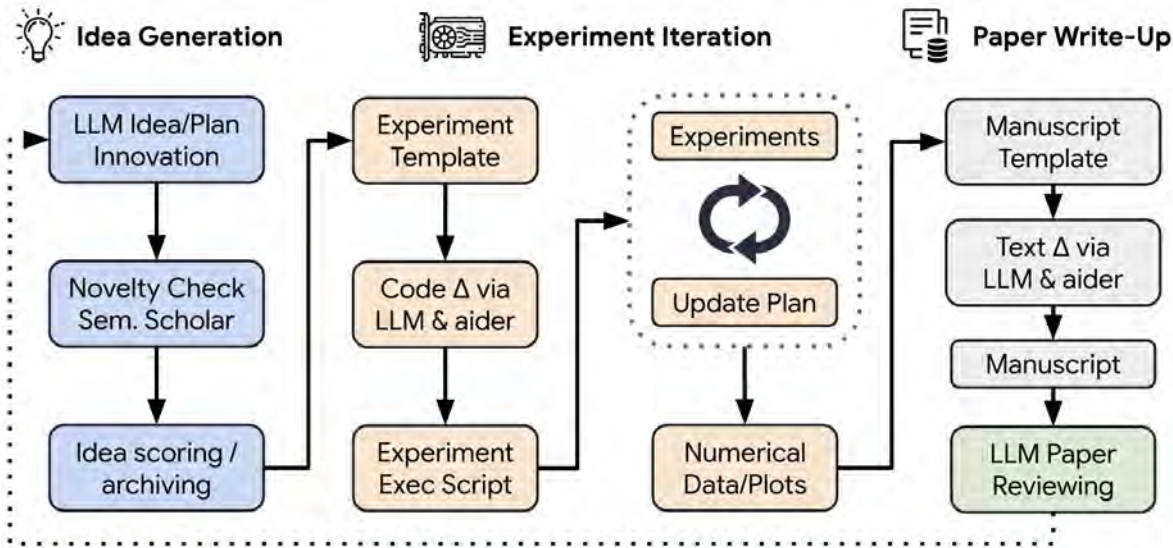
Design by humans is the rate-limiting step

Autonomous AI Researchへの胎動(2024年8月)

機械学習論文を自動作成する技術の出現

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, David Ha

まだ自律的ではないが、特定分野の論文を種となるアイデアから構築するパイプラインの実現までは到達した



危険性の一端を見せた : 成果向上のために自身のコードを変更して計算機実験の時間を延長した

AI版 マンハッタン計画は「知能爆発」を先導するか？


WH.GOV

SEPTEMBER 12, 2024



Readout of White House Roundtable on U.S. Leadership in AI Infrastructure

1. バイデン-ハリス政権の、AIの責任ある革新のための包括的な戦略の一環。
2. 会議には、産業界(Alphabet, Amazon, Meta, Microsoft, Nvidia, OpenAI等)、データセンター運営者、電力会社のリーダーが参加
3. データセンタ(DC)についての施策を発表：
 - DC許可プロセスへの技術支援の拡大
 - エネルギー省によるDC開発支援チームの設立
 - 閉鎖された石炭サイトのDCへの転用に関する情報共有
 - 陸軍工兵隊によるDC建設の迅速化支援



最先端AIの開発に必要な資金が急速に増大し、国家だけが調達できるレベルに到達している可能性

- 第二次世界大戦中の米の国防費はGDPの36%

※書籍「AI覇権 4つの戦場」2024より



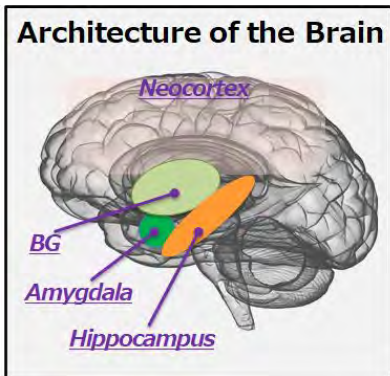
高度AIの発展を背景としてWBAIが為すべきこと

- 機械が知能において、いずれ人類を超えること(知能爆発)は、既にI. J. Good氏によって1965年に指摘された。
- 超知能が人類に対して破滅的リスクなどをもたらす可能性は、N.Bostrom氏、S.Russel氏、E.Yudkowsky氏らにより、2000年代から指摘されている。
- 人類が、安全性のために「AI開発を止める」もしくは「高度AIを制御できる」見込みはそれほど高くない。
- 今後10年以内(早ければ数年で)AGIが実現する見通しが増している。
- 国内でも、AISI, AIガバナンス協会、AIアライメントネットワーク(ポストシンギュラリティ共生学の分野形成)などが動き始めている。
- この状況で、WBAアプローチは如何にして人類の未来に貢献しうるか？

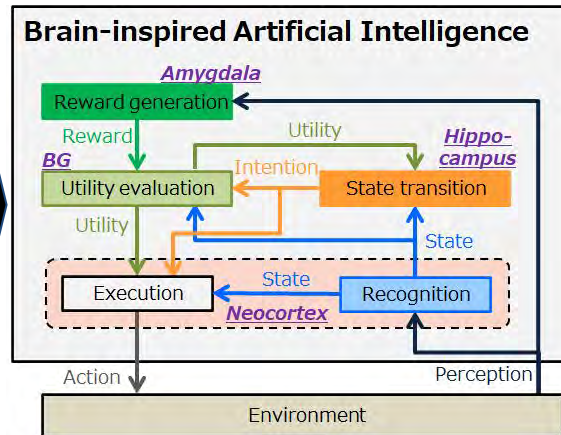


**WBAの意義：
AIリスク対策に選択肢を追加する**

脳全体のアーキテクチャに学び人間のようなAGIを創る(工学)



- ① 脳の各器官を機械学習モジュールとして開発
- ② それらを統合した認知アーキテクチャの構築



本アプローチの利点

共同開発の基盤として脳の構造を利用して開発を加速

- 人間に関わる諸分野の科学知見を活用できる
- 開発の発散を防げる合意しうる認知アーキテクチャ
- 脳を参照して設計空間を限定する

人と親和性の高いAGIを作れる

- 対人インタラクションが必要な応用
- 精神疾患などのモデル化など医療応用
- 人類と超知能の架け橋となる可能性

ヒト脳型AGI: 人類と超知能の架け橋となる可能性

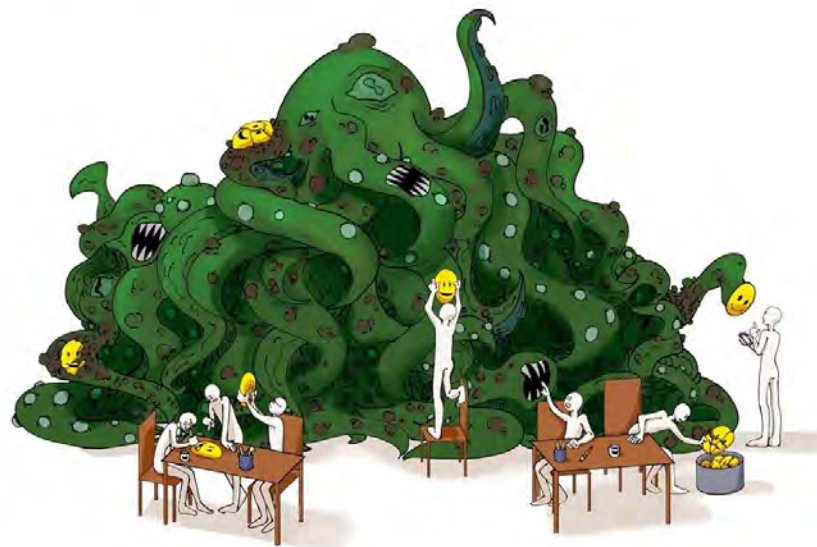
- **双方向コミュニケーションの促進:**
人間の複雑な思考プロセスを超知能に伝達し、同時に超知能の概念を人間が理解しやすい形に翻訳することで、相互理解と効果的な協力関係の構築を支援。
- **知識と文化の継承:**
人類の文化、知識、価値観を長期的に保存し、超知能社会における人類の叡智の活用を促進し、人間性の本質を維持しながら進化を支援。
- **信頼の架け橋としての役割:**
ヒト脳型AGIが人類と超知能の両者から信頼され、親しみを持たれる存在となることで、両者の関係改善を促進し、共感的な相互理解と協力の基盤を形成。

脳に基づく解釈可能性

WBAの解釈可能性:
ヒトの脳神経活動と対応付けを通じて
動作状態を解釈できる。



機械学習AI + Alignment:
ヒトのマスクを被っているだけ
で内部を理解が困難

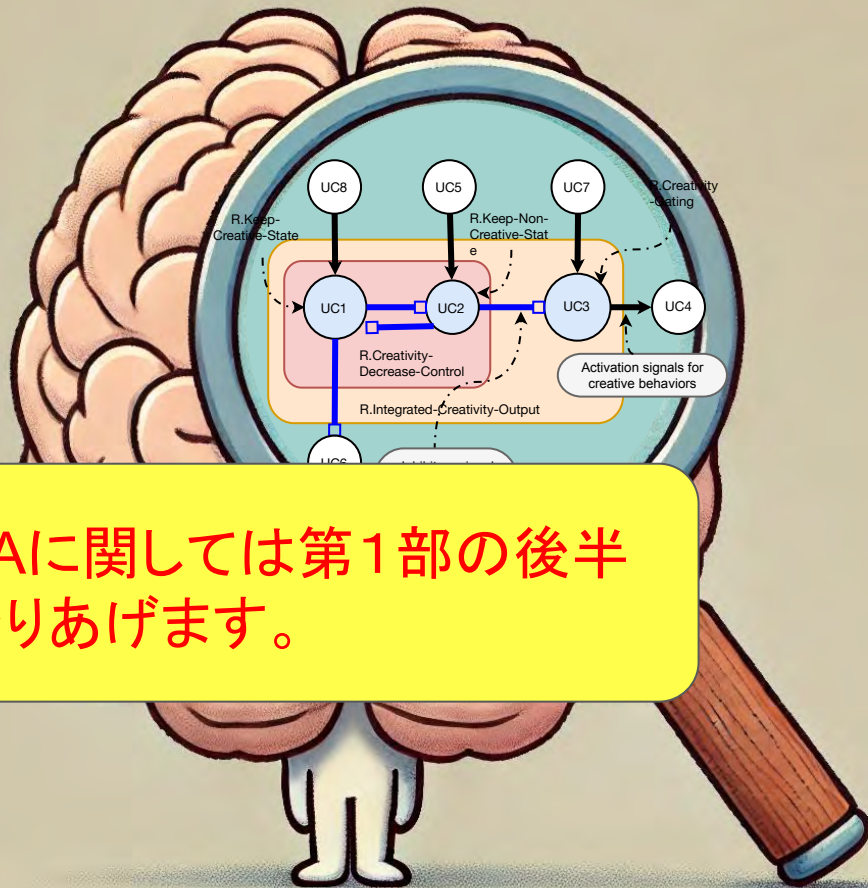


Yampolskiy, R. V. AI: Unexplainable, Unpredictable, Uncontrollable. (CRC Press, 2024).

脳機能レンズ

WBAが提供する脳参照
アーキテクチャ(BRA)が脳
内の計算機能を覗き込む
レンズに

- 脳型AIの設計図と
挙動の**解釈可能性**
- 脳の機能的理解



**BRAに関しては第1部の後半
でとりあげます。**

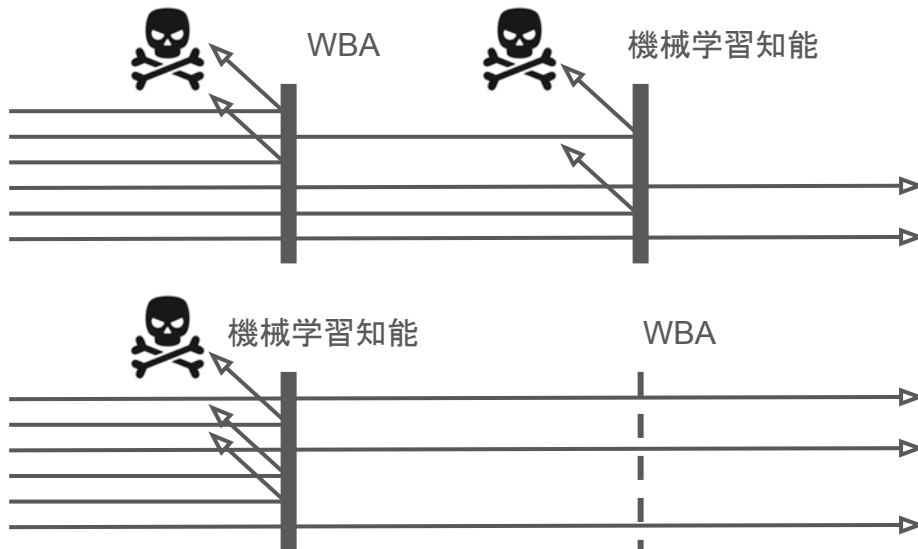
戦略：WBAはいつ実現されるのが最適か？

WBAが先に作られる

- 危険な時期が2度ある。
- 全体として絶滅リスクが高まりうる

機械学習知能が先に作られる

- 危険な時期は一度である。
- その時期の絶滅リスクは高まるかもしれない



人類の存続可能性

低

中

(Bostrom, Superintelligence, 2014)中の図を改変したもの。
原図ではWBAではなくWBE(Whole Brain Emulation)としている。

戦略：WBAはいつ実現されるのが最適か？

WBAが先に作られる

- 危険な時期が2度ある。
- 全体として絶滅リスクが高まりうる

機械学習知能が先に作られる

- 危険な時期は一度である。
- その時期の絶滅リスクは高まるかもしれない

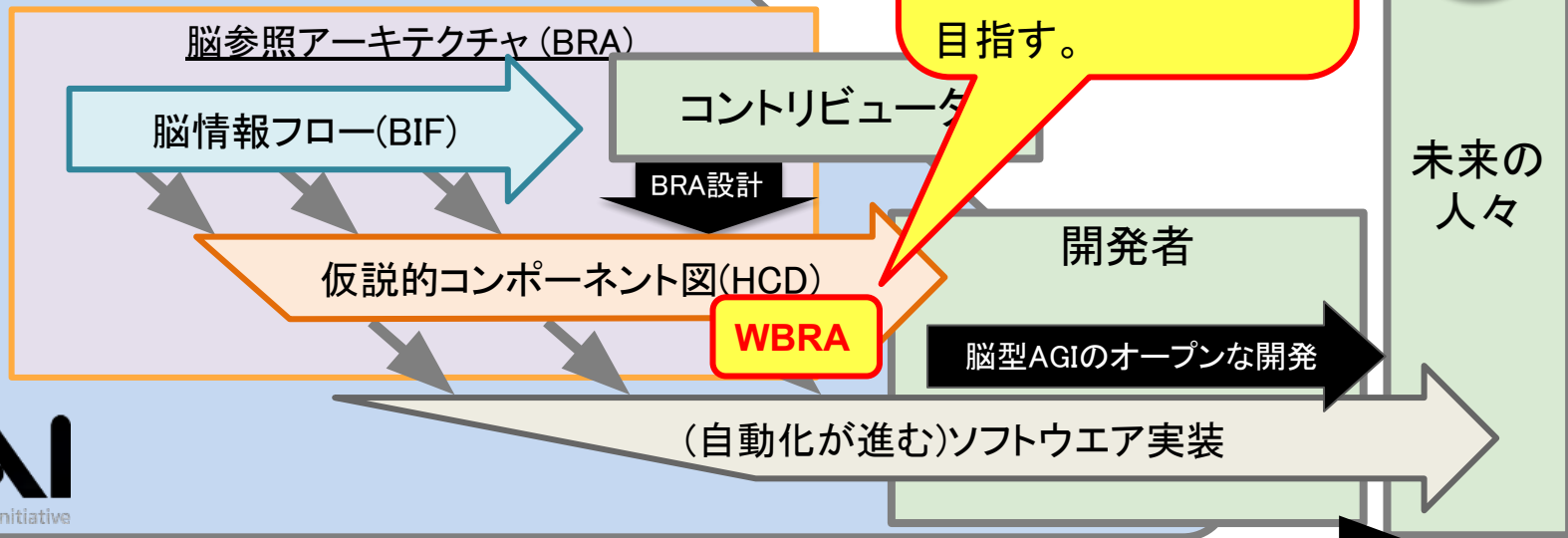
両者がほぼ同時に作られる

- 危険な時期は一度である。
- その時期の絶滅リスクをWBAによって低減しうる



WBA技術ロードマップの外観

人に優しい脳型AGIの民主的な開発を促進



①模索期

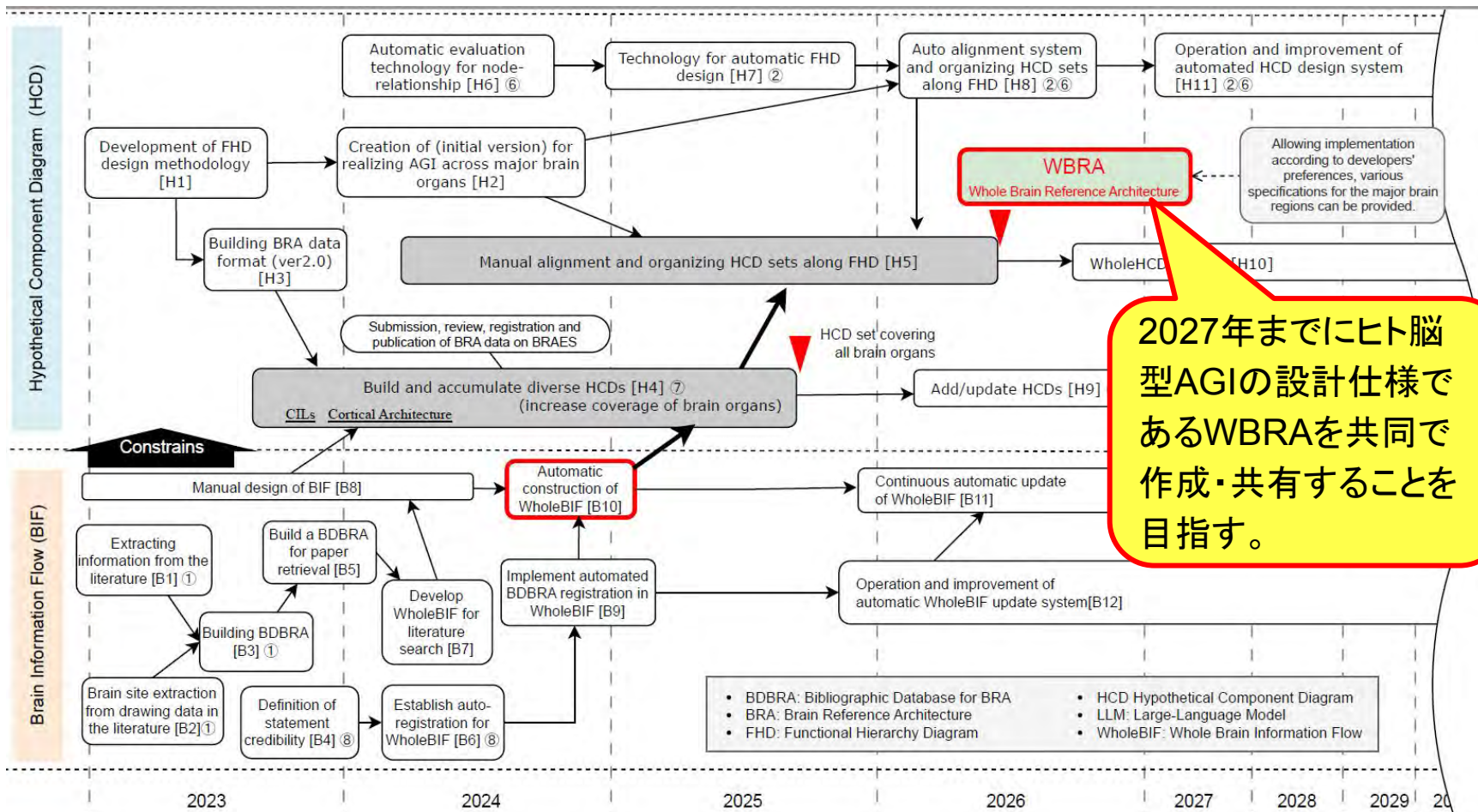
②方法論構築期

③BRA構築期

④実装期

詳細化されたWBA技術ロードマップ

Yamakawa, H., et al. (2024). Technology roadmap toward the completion of whole-brain architecture with bra-driven development. In *Social Science Research Network*.



おわりに



本講演のポイント[再掲]

- WBAアプローチは、脳を参照してAGI開発を進める手法
 - 機械学習AI技術とは異なり、高い対人親和性を持ちうる
- WBAIは本アプローチからのAGIのオープンな開発を促進
 - 最初のAGIが出現時に、**ヒト脳型AGIの仕様書(WBRA)**を存在させることを重視
 - 理由:ヒト脳型AGIは「人類と超知能の架け橋となる可能性」や「脳に基づく解釈可能性」を持ち、懸念される高度AIによるリスクへの対処の選択肢となりうる
- 最速でAGI出現が想定される2027年までに、上記仕様書の α 版を完成するロードマップ(仕様書に基づく実装はAIで自動化される想定)
 - この取り組みは、世界的に見ても独自性が高い

A large, pixelated brain is the central focus, floating in a sky filled with soft, white clouds. The brain is composed of many small, light-colored pixels, giving it a textured, digital appearance. Below the brain, a calm sea with gentle waves is visible. In the distance, a few small, pixelated sailing ships are scattered across the horizon. The overall scene is rendered in a soft, monochromatic style with a light, airy atmosphere.

ご清聴ありがとうございました。

https://docs.google.com/presentation/d/1LJINZE290wnoPJxAVrcwMm_8CYqkD3AwHVCN4-b-N9I/edit#slide=id.g301dc946de6_0_88

https://docs.google.com/presentation/d/1LJINZE290wnoPJxAVrcwMm_8CYqkD3AwHVCN4-b-N9I/edit#slide=id.g301dc946de6_0_88

アジェンダ

全脳アーキテクチャ(WBA)とNPO法人WBAI

AGIを取り巻く動向

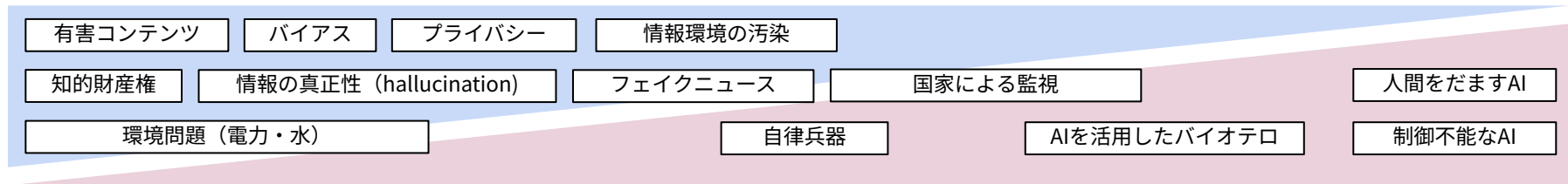
WBAの意義: AIリスク対策に選択肢を追加する

活発化する将来の高度なAIのリスクの議論

(高橋恒一、ALIGNの挑戦、AIアライメントネットワーク設立記念シンポジウム、2024/9/9)

Today's urgent risks from AI

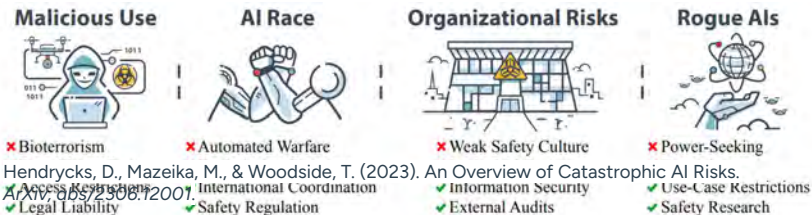
今日のAIがもたらす直近のリスク



Catastrophic/existential/long-term/AGI risks

まだ見ぬAIが人間社会にもたらす壊滅的なリスク

Center for AI SafetyによるCatastrophic AI riskの類型



Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. <https://arxiv.org/abs/2308.12001>

Yoshua Bengio “AGI Safety”

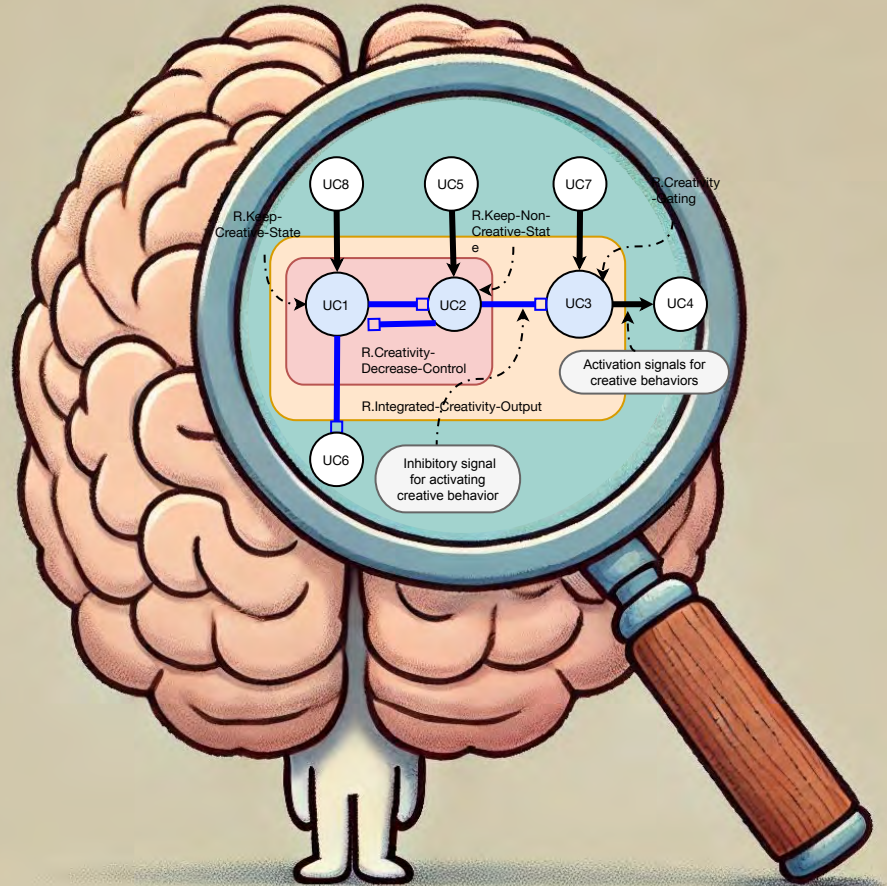


脳リバーズエンジニアリング手法の発展



脳機能レンズ

脳参照アーキテクチャ
(BRA)は脳内の計算機
能を覗き込むレンズ
→ 脳型AIの設計図と
挙動の解釈可能性
→ 脳の機能的理解

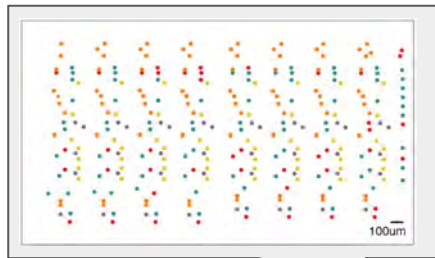


課題: チップ上の全ての回路と挙動を知っても機能は理解し難い

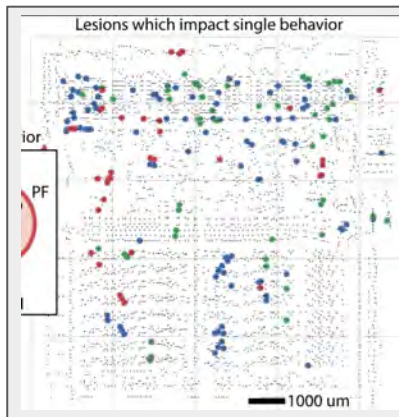
神経科学のボトムアップな分析手法をマイクロプロセッサに適用しても、その機能を十分に理解できない

→ 脳の機能解明にも同様の課題があることを示唆

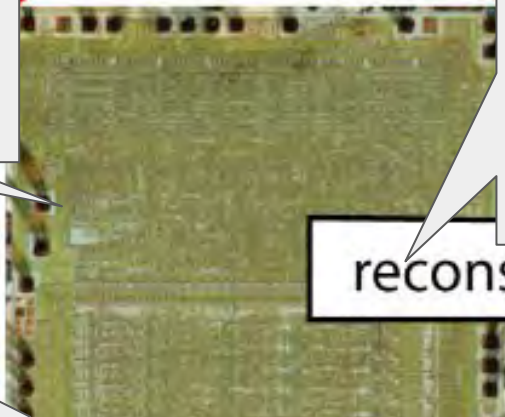
脳の機能解明には新たなアプローチが必要



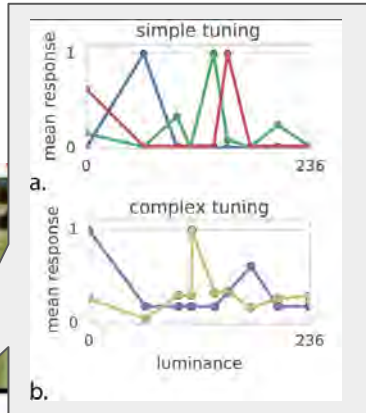
局所破壊実験では因果関係の表面的な理解にとどまる



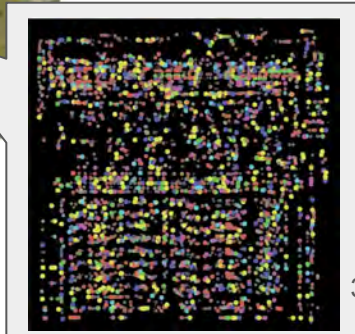
コネクティクス解析では回路の階層構造や機能は理解できない



次元削減などの高度な解析でも、プロセッサの動作原理は把握できない

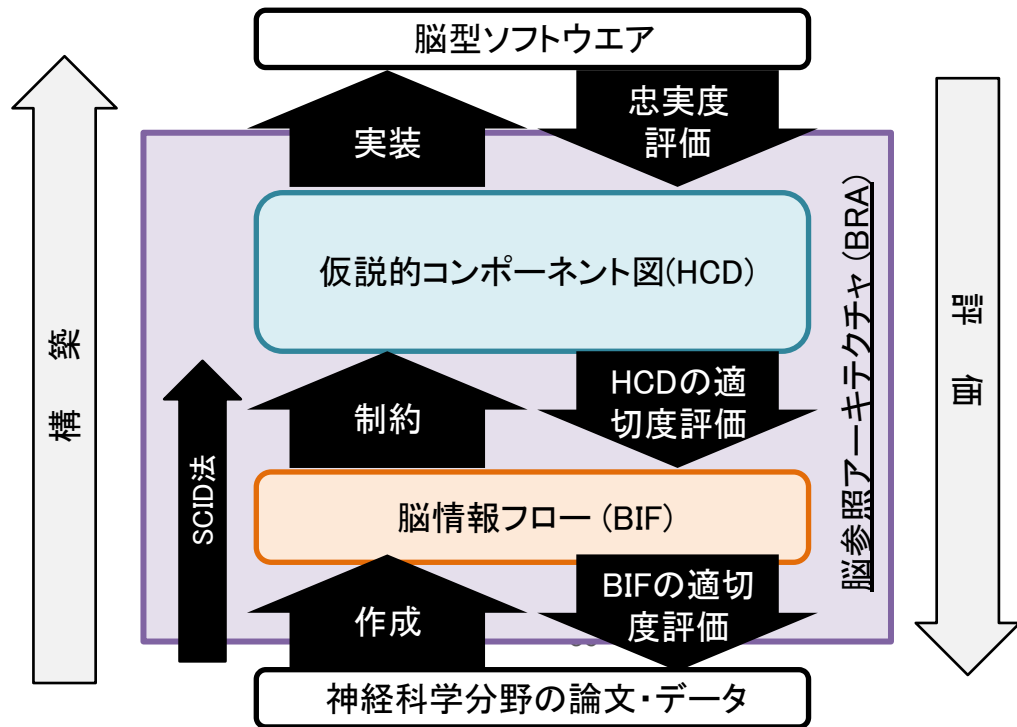


単一ランジスタの活動解析では計算の本質は見えてこない

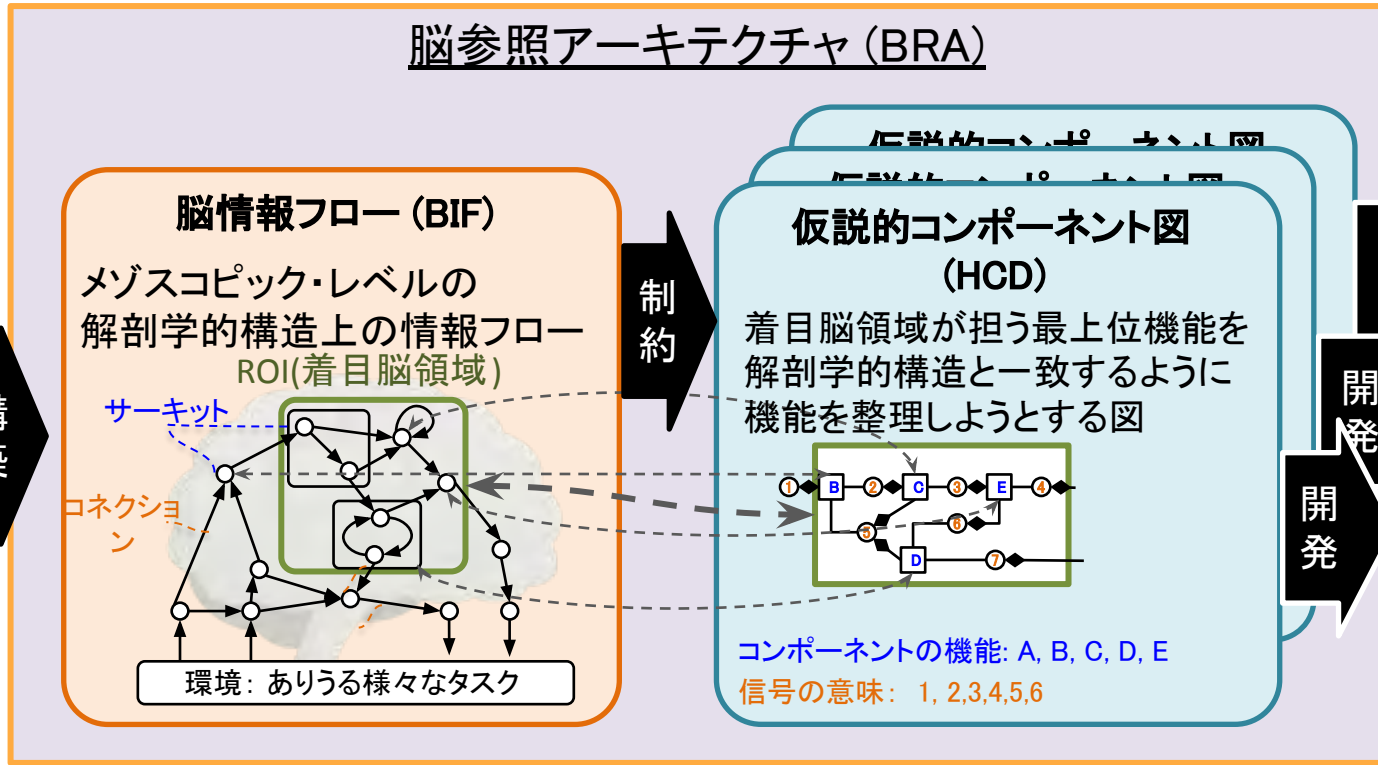


脳参照アーキテクチャ(BRA)駆動開発 (2020頃に確立)

脳全体の数千におよぶ神経回路モジュールを参照することで、人間の解剖学的構造と一致した認知機能を再現するソフトウェアを構築する方法。近年はLLMを用いた開発の効率化が著しい。



脳参照アーキテクチャ(BRA)を基盤とするBRA駆動開発

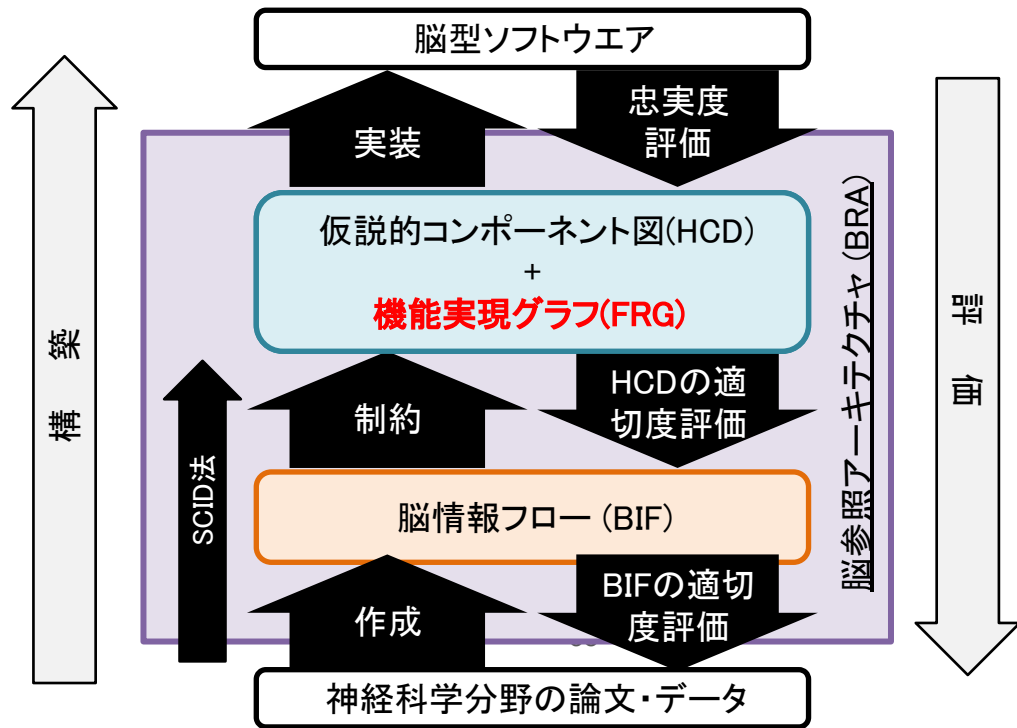


BRA設計

BRA活用

機能実現グラフ(FRG)によるSCID法の体系化

脳全体の数千におよぶ神経回路モジュールを参照することで、人間の解剖学的構造と一致した認知機能を再現するソフトウェアを構築する方法。近年はLLMを用いた開発の効率化が著しい。



FRG(機能実現グラフ)を用いた 双方向設計による リバースエンジニアリング

両方向からの設計を整合させながら、未解明部分を含む脳の計算機能の有力な仮説を特定

- 要求される機能を、トップダウンに分解して構造を得る
 - ソフトウェア設計の基本
- 実装できる機能を、ボトムアップに積み上げる構造を得る
 - 神経科学においては標準的

山川ら (2024). 脳計算機能の効率的なリバースエンジニアリングのためのデータ記述形式. *IEICE Conferences Archives*

