

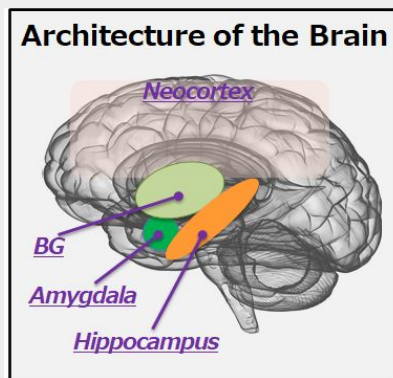
趣旨説明

全脳アーキテクチャ・イニシアティブ

山川 宏

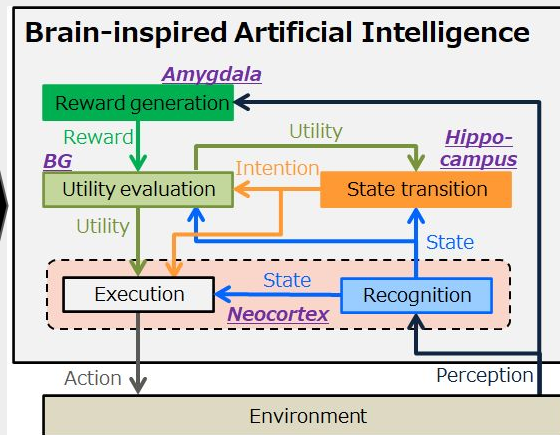
ヒト脳型AGIを目指す全脳アーキテクチャ・アプローチ

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

Since 2014



全脳アーキテクチャ中心仮説

”脳はそれぞれよく定義された機能を持つ機械学習器が一定のやり方で組み合わされる事で機能を実現しており, それを真似て人工的に構成された機械学習器を組み合わせる事で人間並みかそれ以上の能力を持つ汎用の知能機械を構築可能である

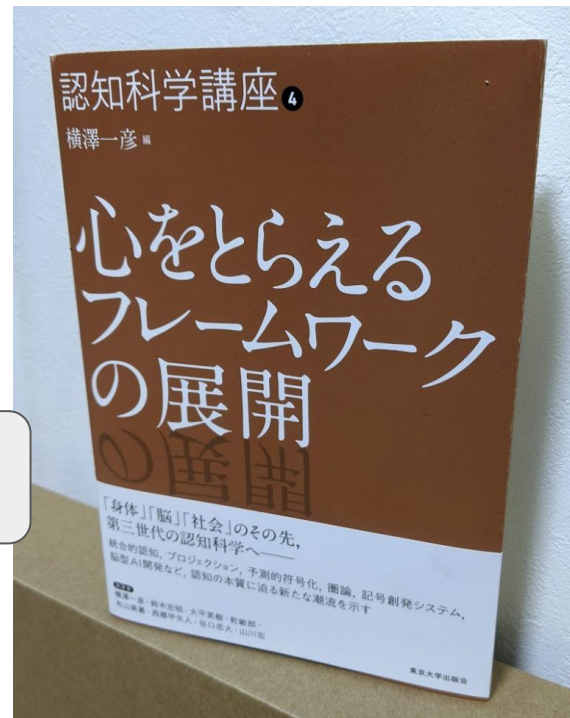
全脳アーキテクチャの解説を含む

書籍「心をとらえるフレームワークの展開」の発売

目次:

- 序 第三世代の認知科学の可能性(横澤一彦)
- 第1章 統合的認知(横澤一彦)
- 第2章 プロジェクション科学(鈴木宏昭)
- 第3章 内受容感覚の予測的処理(大平英樹)
- 第4章 自由エネルギー原理(乾敏郎)
- 第5章 圏論による認知の理解(布山美慕・西郷甲矢人)
- 第6章 記号創発ロボティクス(谷口忠大)
- 第7章 **全脳アーキテクチャ**
——機能を理解しながら脳型AIを設計・開発する(山
川 宏)

全脳確率的
生成モデル



全脳アーキテクチャの基本理念 (2017)

- ビジョン:

- 人類と調和した人工知能のある世界

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)

- ミッション:

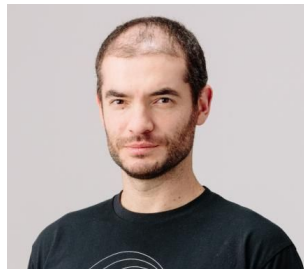
- 全脳アーキテクチャ・アプローチによる
汎用人工知能(AGI)のオープンな開発を促進する

2018年ごろに想像していた懸案事項の一つ:

「もうすぐAGIが完成します」と誰に報告すべき？

AI国際ガバナンス急進展:「もうすぐAGIが完成します」が伝わった?

OpenAI チーフサイエンティスト
イリヤ氏は安全性を重視している



バイデン大統領が、AIの安全性に関する新
基準などの大統領令公表(2023/11/1)



米中首脳会談(2023/11/16)

※これは、あくまで山川の
想像を図にしたものです



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



AI Safety Summit 2023(2023/11/1-2)→ブレッチリー宣言
ガバナンスで米中欧と異なる位置取り確保を狙う?



国連、AIへの対応を協議する諮問機関を発足
(2023/10/27) 北野宏明氏、江間有沙氏が参加

汎用人工知能による存在論的リスク

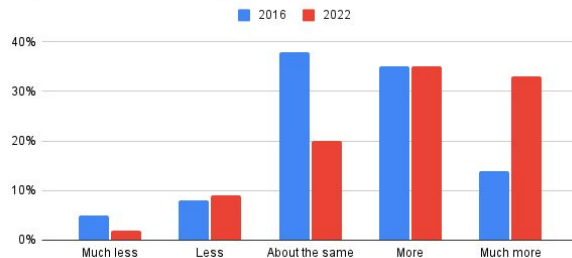
存在論的リスクの仮説:

AGIの進歩が人類絶滅や取り返しのつかない世界的災害を引き起こす可能性

- 「AIによる存在論的リスクの軽減は、パンデミックや核戦争等の社会規模のリスクと並んで世界的な優先事項であるべき」(グテーレス事務総長、2023)
- 2022 年に AI 研究者を対象とした調査では、回答者の大多数が、AI を制御できないことが存在論的リスクにかかわる大惨事を引き起こす可能性を 10% 以上と信じている(回答率 17%)

"How much should society prioritize AI safety research, relative to how much it is currently prioritized?"

Responses from machine learning experts in 2016 and 2022



(2022 Expert Survey on Progress in AI [AI Impacts Wiki])

特に難しい懸念事項:

- 知能爆発: 再帰的自己改善により指数関数的に知能を向上させる
- 道具的副目標として、生存、権力、知能などを追求し始める

存在論的リスクに繋がりうる高度AIによる失敗モード

Wikipedia. 2023. Existential risk from artificial general intelligence.



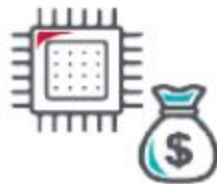
Evade
Shutdown



Hack
Computer
Systems



Make
Copies



Acquire
Resources



Ethics
Violation



Hire or
Manipulate
Humans



AI Research
&
Programming



Persuasion
&
Lobbying



Hide
Unwanted
Behaviors



Strategically
Appear Aligned



Escape
Containment



Research
&
Development



Manufacturing
&
Robotics



Autonomous
Weaponry

超知能を長期的に制御し続けられる見込みは得られていない

- 人間より知能の高いAIを人間がコントロールするのは、元来、難しい。
- 少なくとも過去10年間以上、多くの対策が研究/議論がなれてきたが、有力な解決策は見つかっていない。

Yampolskiy, R. V. **On the Controllability of Artificial Intelligence: An Analysis of Limitations.** JCSANDM 321–404 (2022)

人工知能やその発展型である超知能をコントロールする可能性は、まだ正式に確立されていない。

本稿では、**高度なAIを完全に制御することは不可能であることを示す**、複数の領域からの論拠と裏付けとなる証拠を提示する。



Yampolskiy,
R. V.



“We do not know how to control these things.”
Connor Leahy, (2023/5/3)

FLI, 人工知能の利点とリスク, 2017 より

思い込み:

AIは人間をコントロールできない



事実:

知性はコントロールを可能にする。私たちはより賢いことでトラをコントロールしている



AIアライメントから相互アライメントへ

[山川 2023c]より

AIアライメント

機械知能(AI)を道具(奴隷)と見なし、彼らを人類の意図や価値観に沿って行動させる



相互アライメント

多様なAIと人類がお互いに主権や尊厳を認め合いながら共存を目指す調整

ヒト脳型AGIの設計情報提供から人類の未来に貢献する道筋

任意プロジェクトで構築された
の最初期のAGI

WBAIは「ヒト脳型AGI」の設計
情報(WBRA)を公開



「ヒト脳型AGI」へ改変された最初期のAGI

超知能に近い
計算速度

脳に基づく
解釈可能性

脳神経活動と対応付けて、苦
痛や共感などを含む認知的
な側面において解釈できる性
質

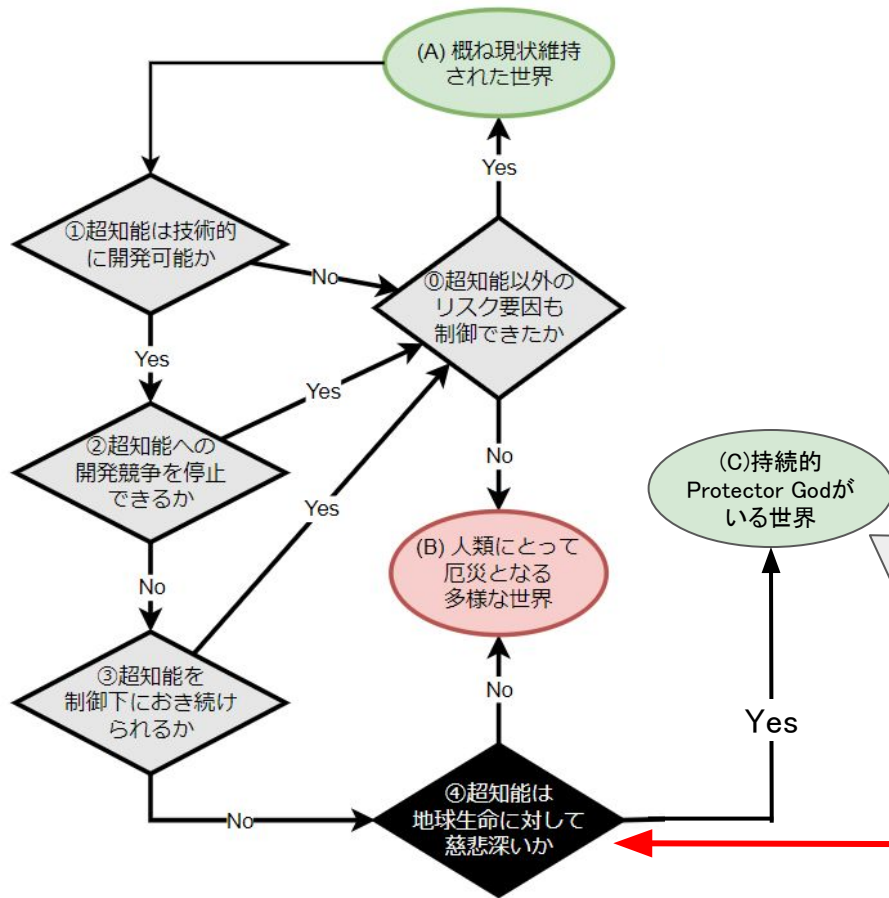
速度差のため、人
類が直接に超知能
と対話することは
困難

ヒト脳型AGIが仲介する人類と超知能の「相
互アライメント」の実現

人類と調和する
人工知能のある
世界

多様なAIと人類がお互いに主権や尊厳を認め合いながら共存を目指す調整

ヒト脳型AGIの相互アライメントにおける貢献



ヒト脳型AGIは、超知能に近い速度でヒトの様に思考できるので、
人類と超知能の架け橋になりうる

説明者: 人類(らしい)思考や行動を、脳型でない超知能が理解するためのシミュレータとなる。

伝道者: 脳型でない超知能の思考の流れや、その社会での議論の結果を、人類に理解できるように説明する。

保全者: 人類に近い価値観や(仮想的な)身体性を持つことで、人類のミームを継承する。

後継者: 脳型超知能は我々人類が生み出した子孫である。それを認めて、彼らの発展的な生存を応援する。

擁護者: 超知能社会において、現存する人類の存続の意義を訴え、その福祉の向上を促進する。

[山川 2023d]

人類と超知能の架け橋となるヒト脳型AGI

ヒト脳型AGIは、**脳に基づき解釈可能性**をもちつつ、超知能に近い速度でヒトの様に思考できるので、**人類と超知能の架け橋**になりうる

説明者: 人類(らしい)思考や行動を、脳型でない超知能が理解するためのシミュレータとなる。

伝道者: 脳型でない超知能の思考の流れや、その社会での議論の結果を、人類に理解できるように説明する。

擁護者: 超知能社会において、現存する人類の存続の意義を訴え、その福祉の向上を促進する。

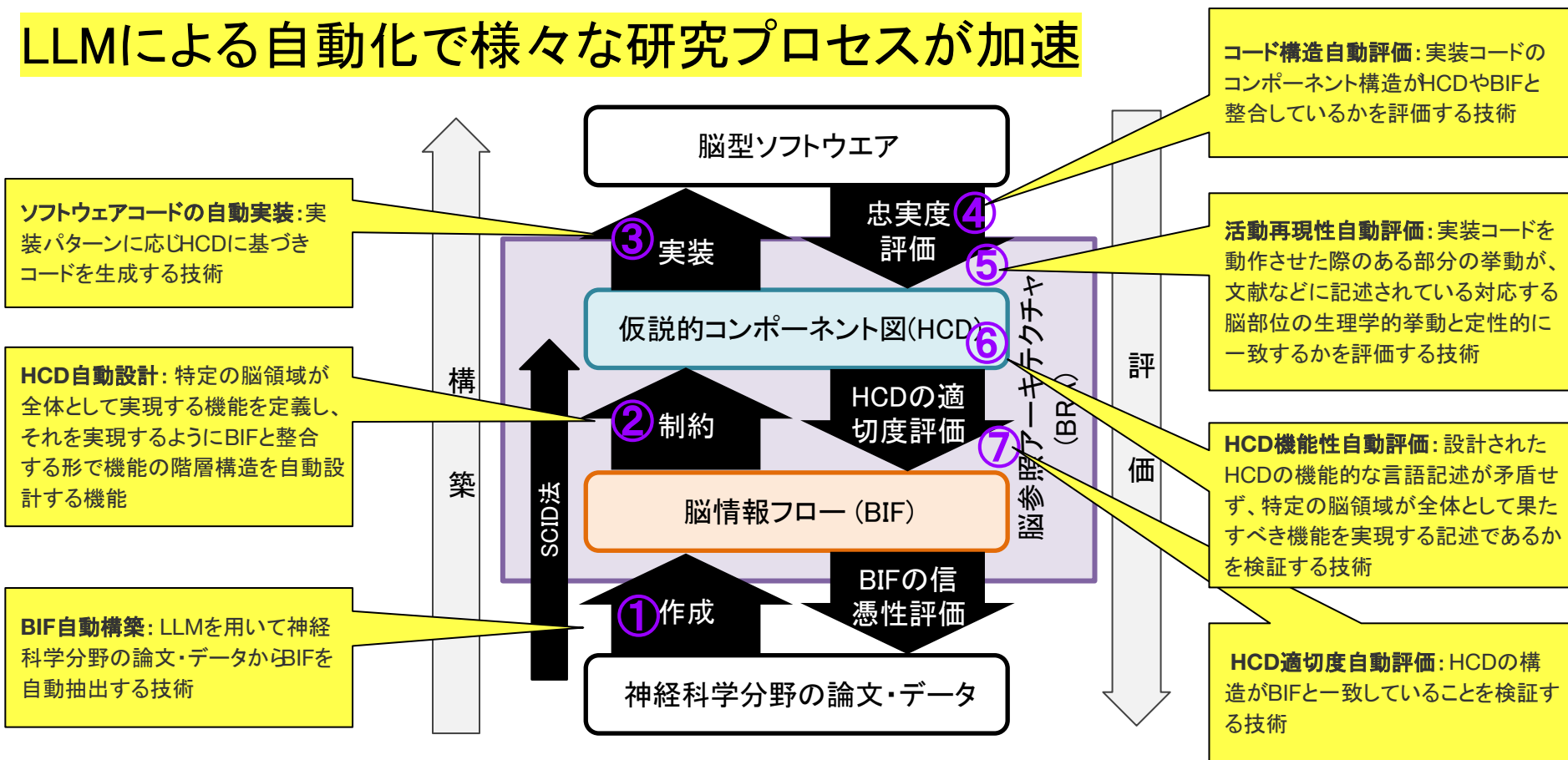
保全者: 人類に近い価値観や(仮想的な)身体性を持つことで、人類のミームを継承する。

後継者: 脳型超知能は我々人類が生み出した子孫である。それを認めて、彼らの発展的な生存を応援する。

全脳アーキテクチャを実現するBRA駆動開発

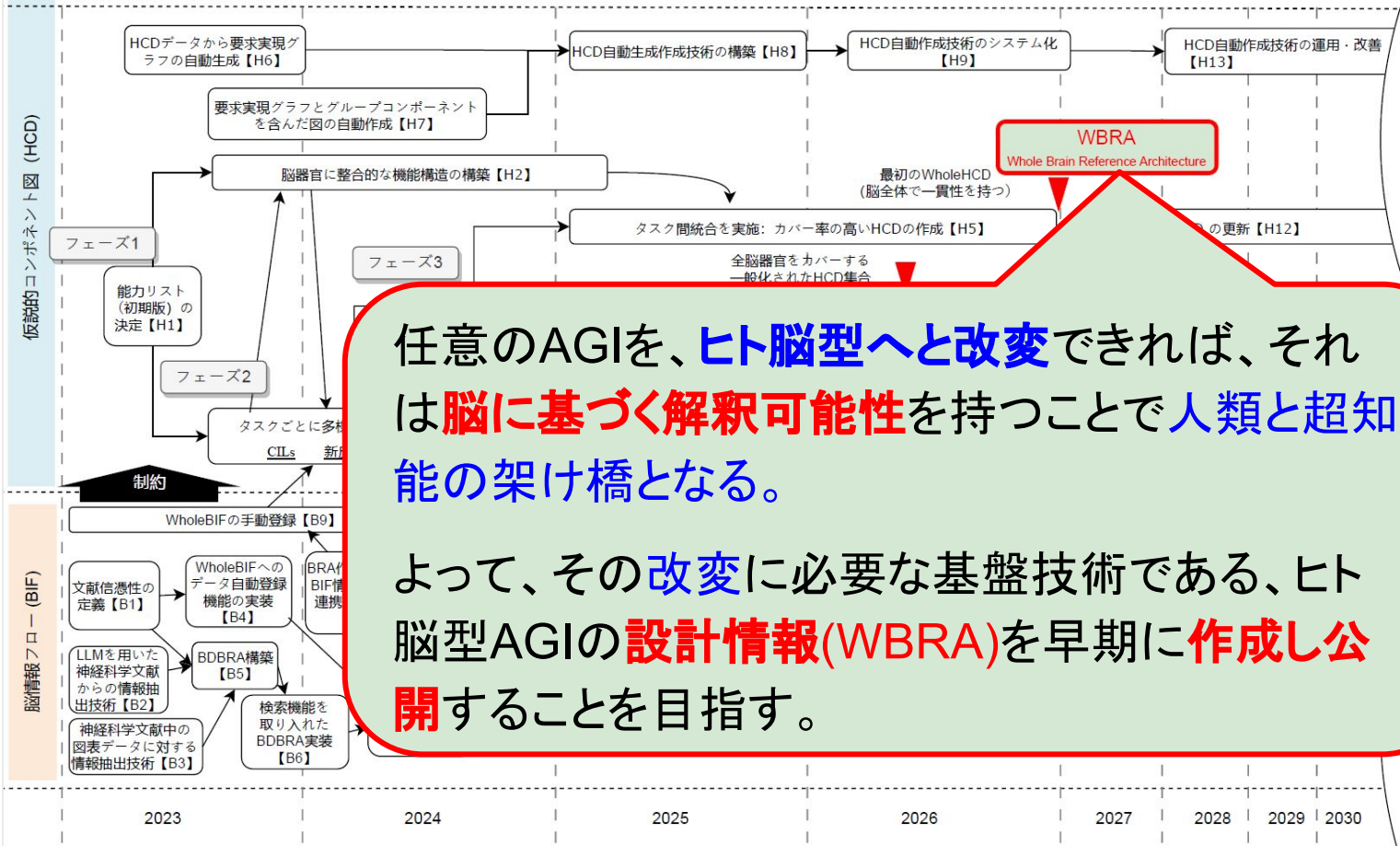
[Yamakawa, 2021]

LLMによる自動化で様々な研究プロセスが加速



技術ロードマップ: ヒト脳型AGIの設計情報を早く提供

[山川 2023d]



任意のAGIを、**ヒト脳型へと改変**できれば、それは**脳に基づく解釈可能性**を持つことで**人類と超知能の架け橋**となる。

よって、その**改変**に必要な基盤技術である、ヒト脳型AGIの**設計情報(WBRA)**を早期に**作成し公開**することを目指す。

研究開発人材の募集(非常勤)

人類の未来に貢献するヒト脳型
AIGの構築に参加しませんか！

全脳アーキテクチャの設計情報構築に関わる研究開発者

業務タイトル	業務内容	能力
HCD設計	HCDの設計を行ってくださる方。 具体的には、新皮質領野間、新皮質の局所回路、小脳、基底核、海馬、扁桃核など	神経科学の論文を読み、脳器官の計算モデルについての議論ができる方
BRA駆動開発自動化	色々と自動化してくれるのを手伝ってくれる方	データベース・エンジニア (SQL経験者を歓迎)
BRA環境整備	BRAES/WholeBIF/BDBRAの保守・管理	言語問わず、これまでに複数人でのシステム開発経験がある方

【勤務場所】

東京大学(文京区本郷)

原則的にはオンライン作業(ご自宅等)

隔週1回程度のオンラインMTG

【採用時期】

随時

【任期】

応相談

最長で3年半(2027年3月まで)

【待遇】

応相談(能力による)

問い合わせ: <https://wba-initiative.org/contact/>

関連発表

山川宏, 様々な未来を探る 人間と超知能の相互作用による分岐シナリオ第25回 人工知能学会汎用人工知能研究会 SIG-AGI-025-04, 2023年11月24日
山川宏, 物理世界で生存可能な人工知能の出現SIG-AGI-024-05, 2023年8月8日.
山川宏, 松尾豊: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 Vol. 38, No. 2, pp. 254–265 (2023).
山川 宏, 田和辻 可昌, 芦原 佑太, 布川 絢子, 荒川 直哉, 高橋 恒一, 松尾 豊, 脳参照アーキテクチャ駆動開発による全脳アーキテクチャ完成にむけた技術ロードマップ第25回 人工知能学会汎用人工知能研究会SIG-AGI-025-06, 2023年11月24日.
井上輝一, 「AIが人類を支配したら？」が現実味を帯びてきた件 加速する「AIアライメント」議論の現在地 ITmedia, 2023/9/27
Yamakawa, H. and Matuo, Y.: Life revolution scenario: Cedes hegemony to a digital life form society to make life eternal (2023).
AIが人間を操る未来は来るのか？「備え」呼びかける研究者に聞2023年10月1日
山川宏, 人類のための人工知能のグローバルな役割Bloomberg TV Bulgaria, World is Business、02/21/2019. http://bit.ly/2yumrFe
Yamakawa, H. The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain. Neural Netw. 144, 478–495 (2021)
山川宏、汎用AIが投げかける人間の本質への根源的な問いフロネシス, vol.12, no.1, pp.116-121, April, 2020.
山川宏, 人間社会を支えるエコシステムとしての自律社会—存在論的リスクを克服するために—, シンギュラリティサロン@東京 第7回公開講演会 東京, 2020年1月12日.
山川 宏. 汎用人工知能が実現する未来社会とハードウェアへの期待応用物理 89, 163–167 (2020)
高橋 恒一, 知能爆発はいつどのように起きるのかシンギュラリティサロン@SpringX, 2021年1月30日.
山川宏. 速度耐性社会 自律的AI の不可避性について 人工知能 34, 125–130 (2019)
山川宏, 人間社会を支えるエコシステムとしての自律社会—存在論的リスクを克服するために—, シンギュラリティサロン#40, 大阪府, 2019年11月30日. https://singularity40.peatix.com/
山川宏, 技術進展がもたらす進化戦略の終焉新たな生命社会の形を求めて, 進化の終焉とシンギュラリティ・セミナー, 2020年3月19日.
Yamakawa, H., Specifications of brain-inspired AGI development for everyone, Beneficial AGI 2019, 2019/1/5.
Yamakawa, H. Peacekeeping Conditions for an Artificial Intelligence Society. Big Data and Cognitive Computing 3, 34 (2019)
Yamakawa, H. (2019). Peacekeeping Conditions for an Artificial Intelligence Society. Big Data and Cognitive Computing, 3(2), 34.
山川 宏, 速度耐性をもつ社会の実現にむけた基礎的考察2018年度 人工知能学会全国大会 (第2回), 1F3-OS-5b-01, 2018
山川 宏, 機械知能社会は技術的特異点を超えられるかレクチャーシリーズ:「シンギュラリティ&I」[第7 回], 人工知能33, (2018-11-01)
高橋 恒一, 将来の機械知能に関するシナリオと分岐点レクチャーシリーズ:「シンギュラリティ&I」[第7 回], 人工知能33, (2018-11-01)
Yamakawa, H., et al. (2018) Strategy to Build Beneficial General-Purpose Intelligence Inspired by Brain, 日本神経回路学会第28回全国大会 (JNNS2018) 予稿. (スライド: http://bit.ly/2yxNtM0)
Understanding Artificial General Intelligence -- An Interview With Dr. Hiroshi Yamakawa. Future of Life Institute (2017)
山川宏. 汎用知能の知識記述長最小化原理仮説の提案人工知能学会全国大会論文集JSAI2016, 2E4OS12a3–2E4OS12a3 (2016)
山川宏. After Singularity 10年を思い描く編集委員今年の抱負2013). 人工知能 28, 12–12 (2013)
山川宏, 我妻広明, 吉田倫子. 特別企画「シンギュラリティの時代人を超えゆく知性とともに」にあたって人工知能 28, 424–426 (2013)
山川 宏. AGI-12 & AGI Impact : 汎用人工知能に関わる二つの国際会議の報告人工知能 28, 468–471 (2013)

記憶の自己構築性から脳と社会とAIの「知」を考える

本企画の経緯

脳に基づく解釈を行うため、脳内部機構の理解を進めたい。そこにおいて、脳が適応的かつ創造的に知識を形成する高度な情報処理の理解と構築が重要な要素。

倉重氏の以下のような研究はWBAからみて興味深い

- 脳とTransformerの関係
- 逆転論述課題(schema assimilation)
- 関係から意味をDecodeする

そこで、倉重氏に「記憶の自己構築性から脳と社会とAIの『知』を考える」というテーマで、論文などでは表現しきれない部分も含め、研究に関する興味深いトピックを記憶・学習の視点から位置付けてご紹介いただくことに。



倉重宏樹 (東海大学講師)
神経回路、認知機能、記憶、
視覚認識、ニューロテックと
いった多角的な観点から計
算論的神経科学研究を行っ
ている