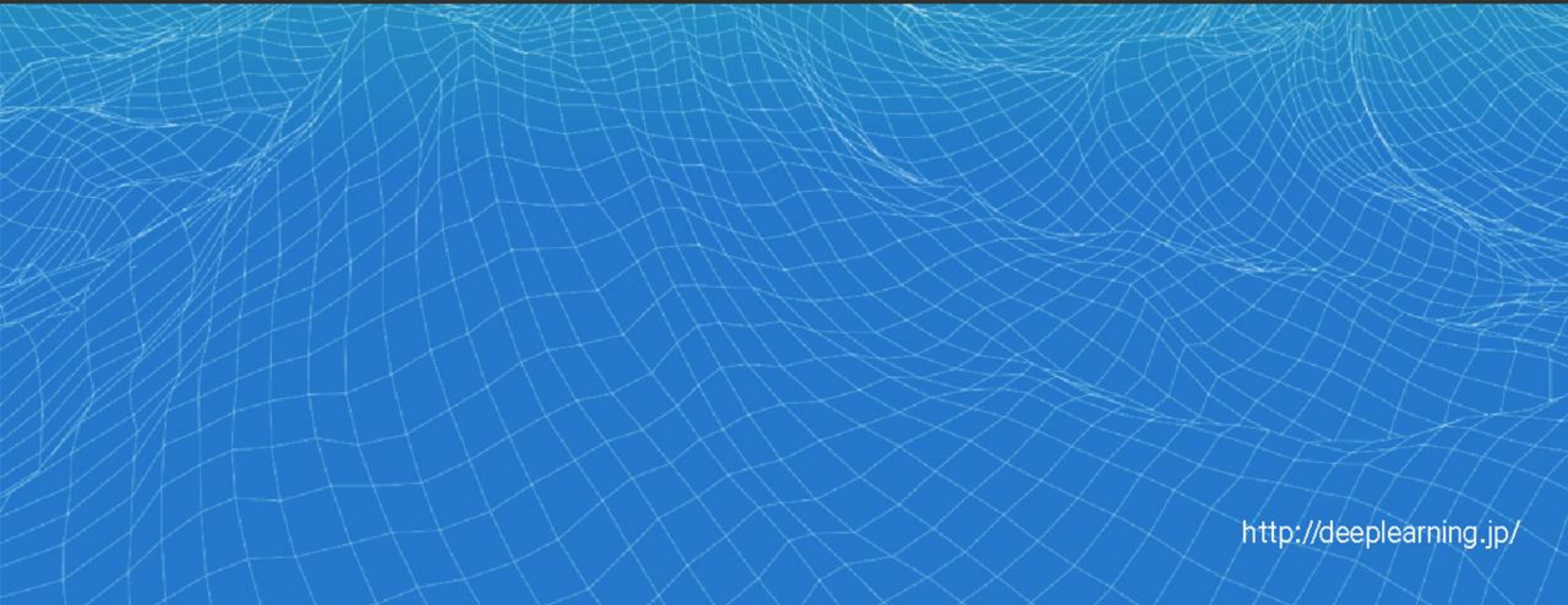
DEEP LEARNING JP
[DL Papers]

Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience

Presenter: Yoshimasa Tawatsuji, Matsuo Iwasawa Lab



http://deeplearning.jp/

書誌情報

- Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience
  - 著者：Martina G. Vilas, Federico Adolfi, David Poeppel and Gemma Roig

- 概要
  - AIシステムの内部メカニズムを解明するために、認知神経科学の教訓を取り入れたInner Interpretability 研究の新しい概念フレームワークを提案

Introduction

- Inner Interpretability（内部解釈可能性）
  - 目的：モデルの内部メカニズムを人間が理解できる形で解釈する
  - 課題：高度なメカニズム的説明を開発し、分析するための統一された概念フレームワークが欠如

- 本論文の趣旨
  - 認知神経科学分野で扱われてきた同様の問題と比較し、Inner Interpretabilityの研究に適用することを提案

AI inner interpretability research

- 目的：深層ニューラルネットワークの内部構造・動作・表現の理解

  Also安全性や透明性の向上→モデルの動作の予測可能性の向上・有害または誤った表現の除去

- Inner Interpretability 研究の対象
  - モデルコンポーネント (Elhage, 2021; Geva, 2021, McDougall, 2023; Olsson, 2022)
  - 機能 (Merullo, 2023b; Todd, 2024)
  - アルゴリズム (Zhong, 2023)

- Other work has developed methods to automate the discovery and analysis of activation sub-spaces (e.g. Burns et al., 2022), circuits (e.g. Conmy et al., 2023; Lepori et al., 2023), and internal representations (e.g. Belrose et al., 2023; Hernandez et al., 2023) that have a causal effect on the output of the model.

統一的な概念フレームワークの必要性

- Inner Interpretability研究への批判
  - 現在の方法論は多様であり、実際の問題やモデルに対してどの程度一般化されるかが疑問視されている(Doshi-Velez & Kim 2017; Räuker 2023)
  - 分野全体での包括的な問いが不明瞭(Krishnan, 2020)
    - 「モデルをメカニズム的に理解する」とは？

- 統一的な概念フレームワークの提案
  - 認知神経科学分野で開発された手法や戦略の適用

認知神経科学とInner Interpretabilityとの共通の課題

- 認知神経科学（Cognitive Neuroscience）
  - Cognitive neuroscience is the scientific field that is concerned with the study of the biological processes and aspects that underlie cognition, with a specific focus on the neural connections in the brain which are involved in mental processes. It addresses the questions of how cognitive activities are affected or controlled by neural circuits in the brain. (*Wikipedia, 2024.8.22*)

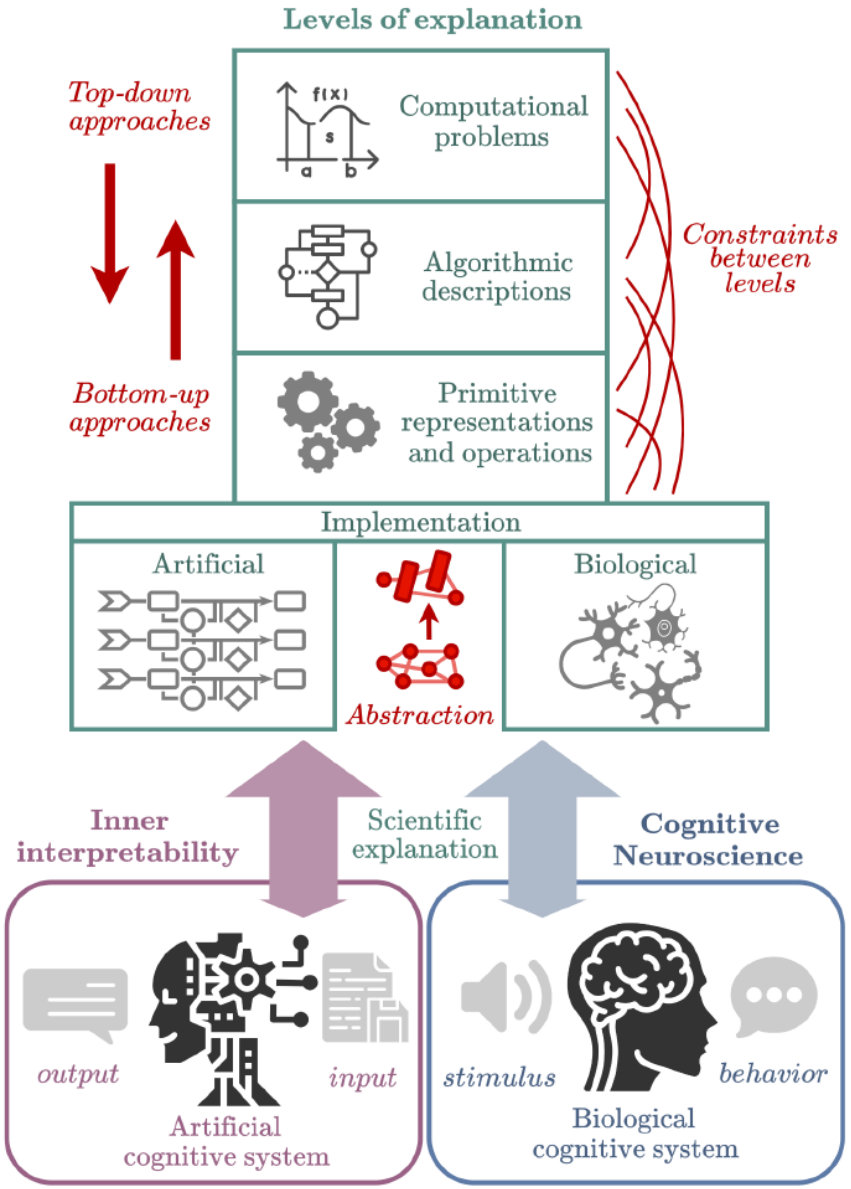
- 両分野とも「複雑なシステム」がもつ能力（capacity）、実装方法を人間が理解できるように説明することを目指している

- 認知神経科学と Inner Interpretabilityとの共通課題
  - どのようにメカニズムを説明するか
  - 説明の抽象化のレベルをどのように設定するか
  - ボトムアップアプローチとトップダウンアプローチのいずれを採用するか

メカニズムの説明に関する課題

- 詳細なメカニズムの説明の不完全性
  - 特定のコンポーネント（ニューロン、回路など）がどのように行動を引き起こしているかの説明が欠如していることが多い
  - 誤った仮説の導出の危険性(Craver, 2006)

- 認知神経科学での教訓
  - マルチレベル分析 (Marr & Poggio, 1976)



抽象化レベルに関する課題

- 抽象化レベルに対する弱い動機づけ
  - 抽象化レベルの選択を誤るとシステムの挙動の理解を誤る
    - ニューロンレベル(e.g. Hernandez, 2021)か重ね合わせなどの相互作用のレベル(Elhage 2022)か

- 認知神経科学の教訓
  - Mapping problem：
    - 神経生物学の基本的部分（シナプス、ニューロン、脳領域など）を認知に関する基本的な「操作」と「表現」にマッピングする
    - 操作と表現のプリミティブの選択は、形式的・経験的に検証された認知機能の理論（計算レベル）および脳内で実行可能な操作（実装レベル）によって制約される

ボトムアップとトップダウンアプローチに関する課題

- ボトムアップアプローチにおける過剰な楽観視
  - 分析対象となるコンポーネント・抽象化レベル・撹乱条件・分析対象の挙動などに、分析者の暗黙の選択が含まれる
  - ボトムアップで得られた知見がもともとの問題や能力に一般化可能かは保証されない

- トップダウンアプローチの限界
  - 事前に持っている仮説や理論の検証が不十分など、誤った理論の採用によるミスリーディングな結論の導出

- 認知神経科学の教訓
  - 上位レベルと下位レベルの相互制約を利用し、整合性のあるメカニズム的説明を行う
  - 仮説の検証・手法の評価も両アプローチの利点を活かす

マルチレベルのメカニズム的説明の構築

- 計算レベル
  - システムが解決すべき問題の定義
  - 問題を情報処理タスクの表現として明確化
- アルゴリズムレベル
  - タスクを実行するためのアルゴリズムの記述
- プリミティブな表現と操作
  - システムが機能するための構成要素の明示
  - 構成要素は理論的な裏付けが必要
- 実装レベル
  - プリミティブな要素の実装の説明

Definition 4.1 (Facts and fact domains)**.** A *fact* is a 3-tuple $F = (S, R, A) \in \mathcal{D}$, where S is a subject, R is a relation, A is an attribute, and $\mathcal{D} = \{F_1, F_2, ..., F_n\}$ represents a *fact domain*. Incompleteness of a fact tuple is denoted with $\perp$ in the corresponding component.

Definition 4.2 (Factual recall)**.**

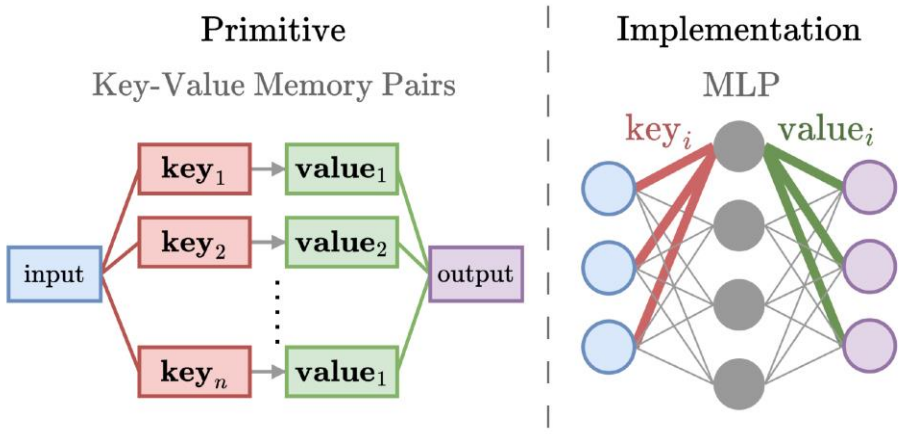
Computational problem: $\mathcal{D}$-FACTUALRECALL

Input: An incomplete *fact* tuple (Def. 4.1), $F_I = (S, R, \perp)$ corresponding to a complete fact $F = (S, R, A) \in \mathcal{D}$, where $\mathcal{D}$ is a constant fact domain.

Output: A completion F_C of F_I such that $F_C \in \mathcal{D}$.

Algorithm 1 Factual Recall via Attribute Extraction

Input: incomplete fact tuple $(S, R, \perp) \in \mathcal{D}$
$H_s = \text{AttributeBoost}(S)$
$H_{s,r} = \text{AttributeBoost}(S, R)$
$H_r = \text{AttributeBoost}(R)$
$H_i = \text{AttributeCombine}(H_s, H_{s,r}, H_r)$
$A = \text{AttributeMax}(H_i)$
Return attribute A



仮説構築と *severe tests* の実施

- 仮説構築
  - 仮説はシステム内でどのようにメカニズムが機能しているかを示すべき
  - 仮説に基づいて具体的かつ反証可能な予測を行う
- Severe testsの実施
  - 仮説が誤っている場合にその誤りを明らかにできるテスト
  - 仮説が正しい場合と正しくない場合の異なる予測を立てテストを行う
- 競合するメカニズムの比較
  - 異なるメカニズムが同じ観察結果を生むことがある
  - 複数のメカニズムをテストし、説明力が高いものを採用する

実験デザイン

- メカニズムの提案の適切性判断
  - モデル動作を調整する入力条件を調査(Craver, 2006)
    - 入力条件の動作に対する影響（促進・阻害・調整）(Hardcastle and Hardcastle, 2015)
  - 実験：入力条件を模倣し、動作に対する影響を効果的に説明できるかを検証
- 説明や予測ができない場合の考察
  - メカニズムに必要な要素が欠如
    - ファクトの取得に失敗する場合：
      - トレーニングセットの欠如、プロンプトの構築方法など（Jiang, 2020）
- 自然主義的条件（naturalistic condition）での有効性
  - 制御された条件以外（事実情報を抽出するためのプロンプトをユーザーが作成する場合など）での説明可能性の検証

不変性のテスト

- 入力サブドメインにおける不変性
  - 特定のサブドメイン（地理、政治、文学など）に依存せず、一貫性して機能しているかを評価
- モデル間の一般化
  - 特定モデル（GPT-3など）だけでなく、他の類似モデルへの適用可能性を評価
- 初期化およびハイパーパラメータの不変性
  - 同一モデルに対する異なる設定でもメカニズムが一貫して機能するかを評価
- モデルのスケールアップにおける不変性

メカニズムの提案と概念的なフレームワークの洗練

- 継続的なメカニズムの改良
  - 各レベルの説明の相互作用によって、新しい理論や実証的知見に応じてメカニズム適説明を改良

- 不完全なメカニズム的説明の役割
  - メカニズム的説明が不十分でも有益な手がかりになる（*Mechanistic sketches*）

- 概念フレームワークの見直し
  - 新しい知見や技術の進展に応じて概念フレームワーク自体も見直す必要有
    - 新しいレベルの導入（e.g. 学習・進化プロセスの説明(Poggio, 2012)）
    - 既存のレベルの再定義

Implications for Inner Interpretability

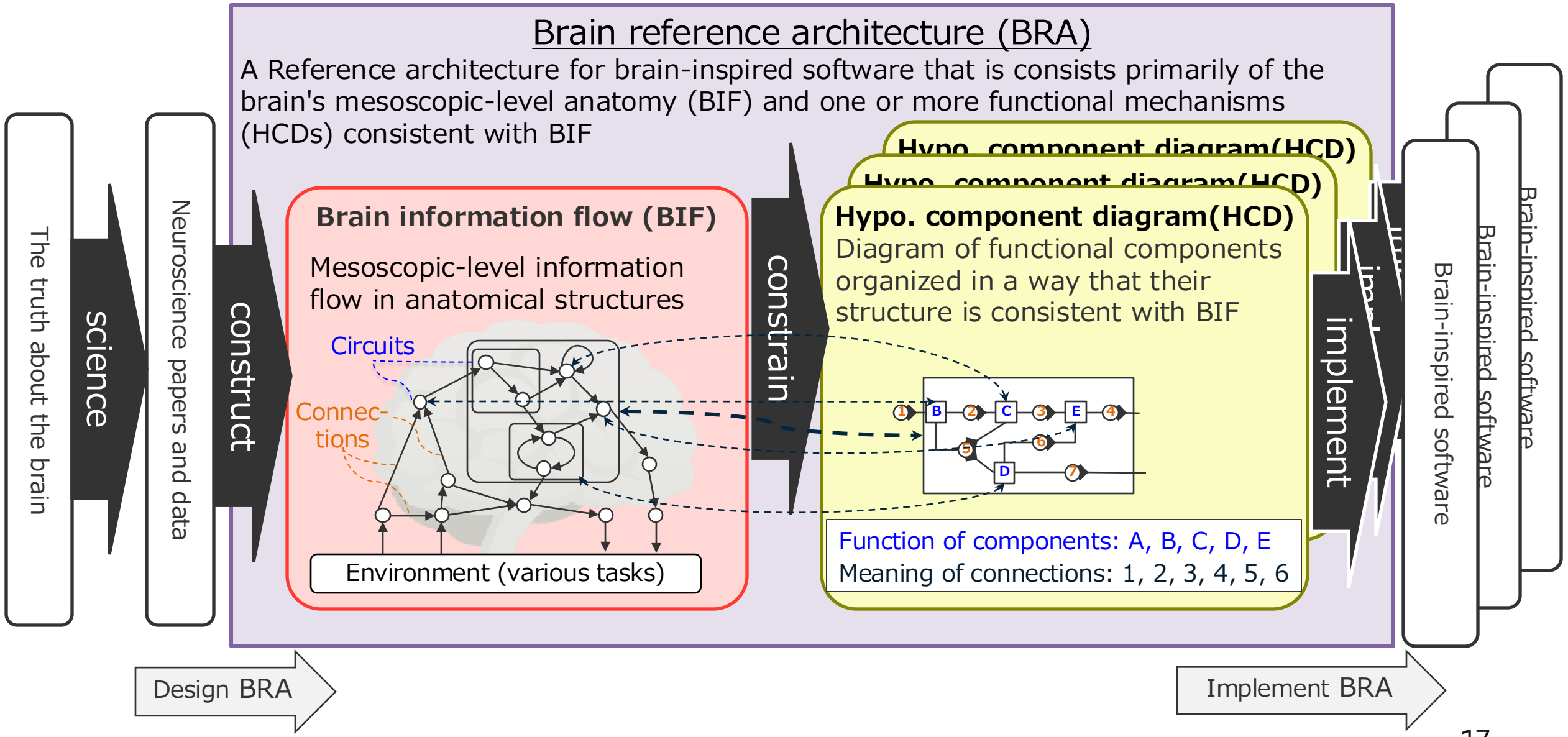
- 先行研究の位置づけ

Computational level				
Algorithmic level	*Chughtai, 2023* *Merullo, 2023a* *Wang, 2022* *Zhong, 2023*			
Operation and representation level			*Geva, 2021* *Olsson, 2022* *Vilas, 2023*	
Implementation level	*Chughtai, 2023* *Merullo, 2023a* *Wang, 2022* *Zhong, 2023*	*Beckers and Halpern, 2019* *Chalupka, 2017* *Geiger , 2023* *Rubenstein, 2017*		*Burns, 2022* *Conmy, 2023* *Gurnee, 2023*

Addressing criticisms

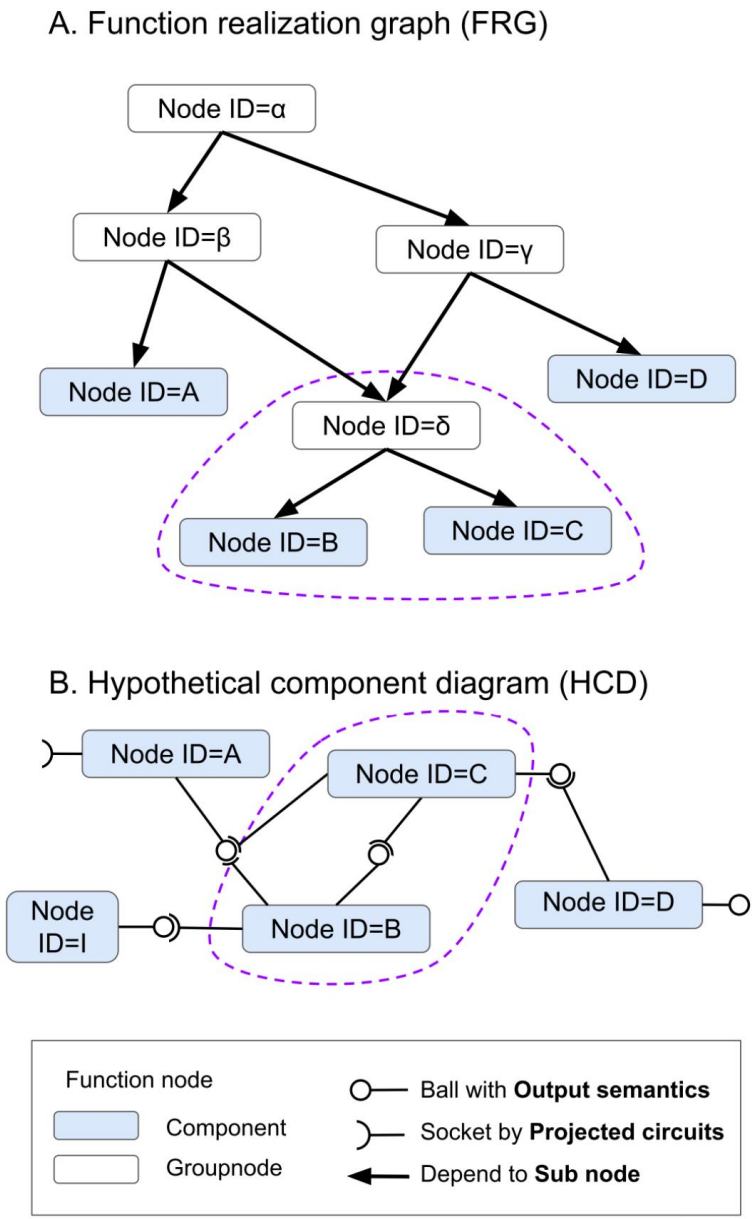
- 本フレームワークの役割
  - システムのメカニズムの理解および説明の構築方法についての現場でのコンセンサスと明確さの欠如へのガイダンスの提供
  - マルチレベル分析、レベル間の制約、および一連の研究によって得られた証拠を通じて、モデルがもつ能力に対するより包括的な特性評価を奨励
  - 仮定を意識した解釈で確実な結果を得るために、severe tests を伴う妥当性評価手法の利用

Whole Brain Reference Architecture Approach



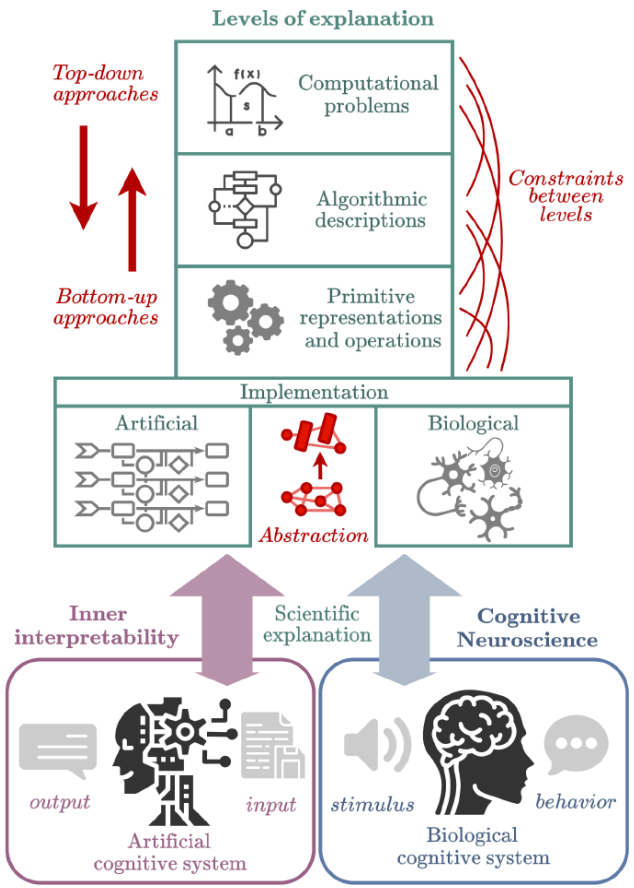
機能の階層構造とコンポーネントとの接地 (山川, 2024)

- Function Realization Graph (FRG)
  - 機能ノード間の依存関係を階層的に示す機能階層図（機能（実現）に関する「**設計**」）
  - 脳の解剖学的構造に整合するように設計されたHCD上に構築
- 双方向設計における二つの機能
  - Requirement
    - トップダウンに決まる機能で、計算機能の文脈に依存
  - Capability
    - ボトムアップに決まる機能で、計算機能の文脈に非依存



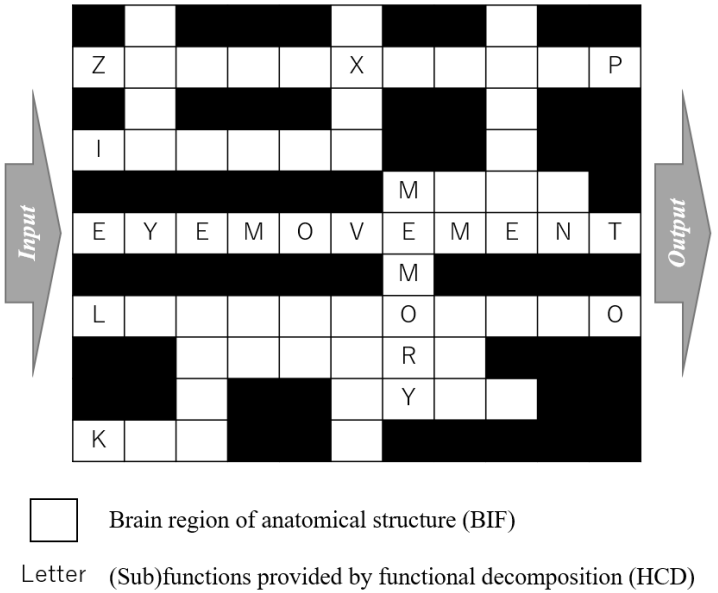
Vertical constrains and *horizontal* constrains

- 一つの物質で様々な機能を実現する脳の場合は、「縦の制約」に加えて、その構造的制約として「横の制約」がある



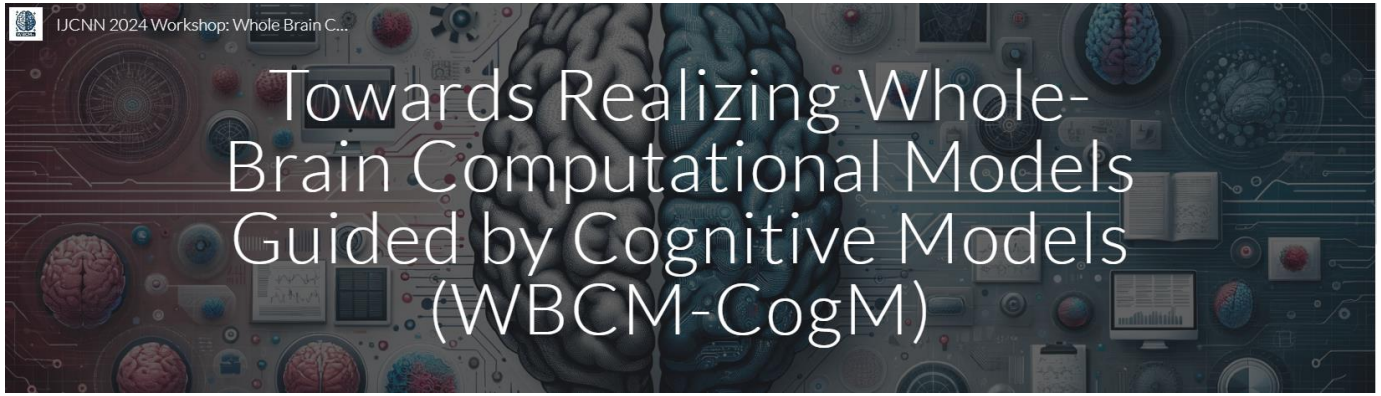
"Crossfunction" puzzle

周辺知識が分かると機能割りあての候補も急速に絞り込まれる可能性



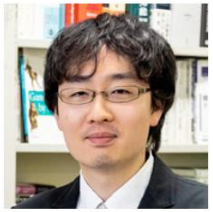
認知モデルにガイドされた脳型アーキテクチャ

- 認知モデルをガイドとすることで、人間の情報処理の内的プロセスとの同型性を担保し、解釈可能性のある脳型アーキテクチャを構築できる



IEEE-WCCI/IJCNN Workshop 2024/6/30 パシフィコ横浜

Organizers:

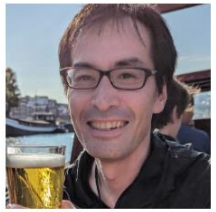


Akira Taniguchi

Ritsumeikan University, Japan

Yoshimasa Tawatsuji

The University of Tokyo, Japan



Junya Morita

Shizuoka University, Japan

特　集「生成 AI 時代における認知のモデリング」

解釈可能性を高めて信頼し得るエージェントを実現するための脳型認知モデル

Brain-morphic Cognitive Model for Enhancing Interpretability and Achieving Trustworthy Agents

田和辻　可昌 Yoshimasa Tawatsuji — 東京大学，全脳アーキテクチャ・イニシアティブ / The University of Tokyo / The Whole Brain Architecture Initiative / y.tawatsuji@weblab.t.u-tokyo.ac.jp

大森　隆司 Takashi Omori — 玉川大学 / Tamagawa University / omori@lab.tamagawa.ac.jp

太田　博三 Hiromitsu Ota — 全脳アーキテクチャ・イニシアティブ / The Whole Brain Architecture Initiative / otanet123@gmail.com

宮本　竜也 Tatsuya Miyamoto — 早稲田大学 / Waseda University / miyamoto9265@ruri.waseda.jp

芦原　佑太 Yuta Ashihara — 日本大学，東京大学，全脳アーキテクチャ・イニシアティブ / Nihon University / The University of Tokyo / The Whole Brain Architecture Initiative / ashihara.yuta@nihon-u.ac.jp

荒川　直哉 Naoya Arakawa — 全脳アーキテクチャ・イニシアティブ / The Whole Brain Architecture Initiative / naoya.arakawa@wba-initiative.org

山川　宏 Hiroshi Yamakawa — 全脳アーキテクチャ・イニシアティブ，東京大学，理化学研究所，AI アライメント・ネットワーク / The Whole Brain Architecture Initiative / The University of Tokyo / RIKEN / AI Alignment Network / ymkw@wba-initiative.org

Keywords: BRA-driven development, emotion, cognitive model, large language model.