RITSUMEIKAN UNIVERSITY

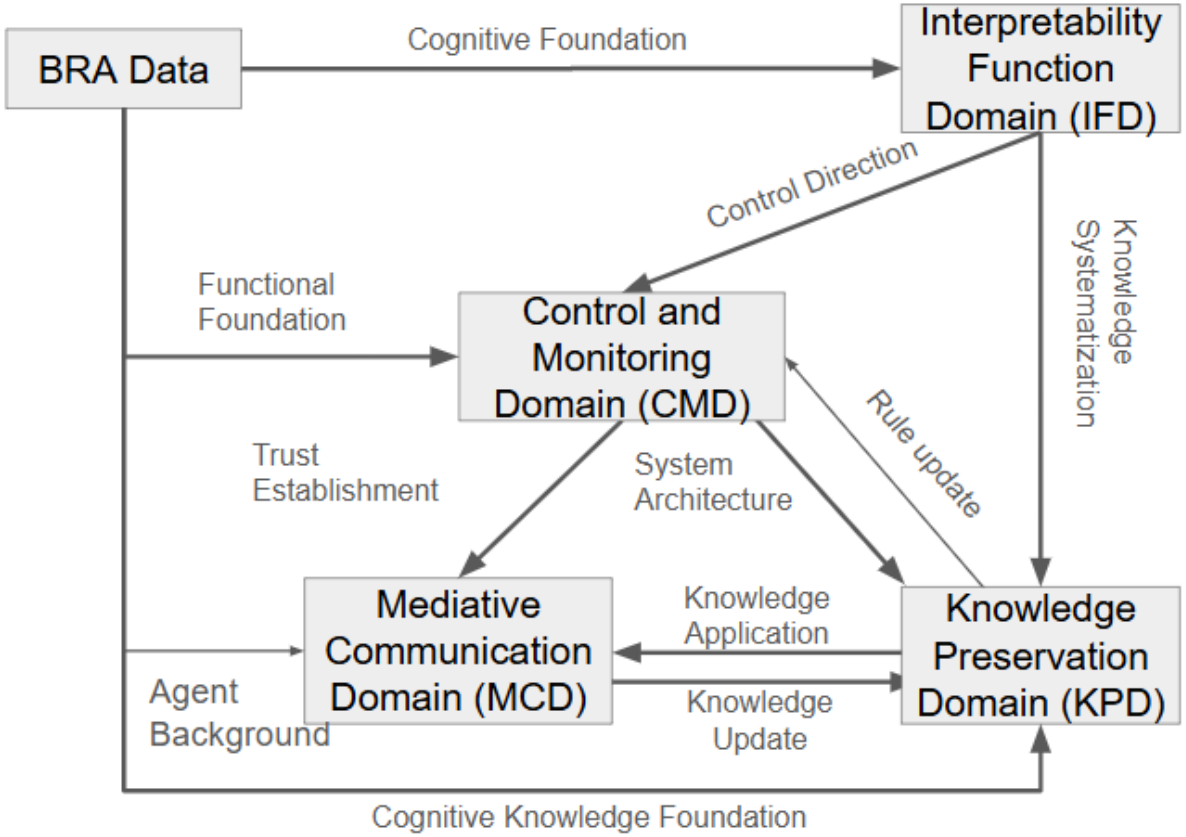
2025年9月4日（木）
第10回全脳アーキテクチャシンポジウム
@東京大学

脳型AIの実装と解釈可能性

立命館大学　情報理工学研究科
中島　毅士

脳型AIが社会と共生するための4つの機能領域

NeuroQuad[1]



機能領域	定義
IFD（解釈可能性）	システムの意思決定と内部プロセスについて、透明性のある説明を提供する
CMD（制御・監視）	挙動を監視し、安全限界が危険にさらされた際に介入する
MCD（仲介・コミュニケーション）	人間と各種 AI システムの間で、意図とコンテキストを翻訳（媒介）する
KPD（知識保存）	人類の価値観と文化の長期的な記憶を保持する

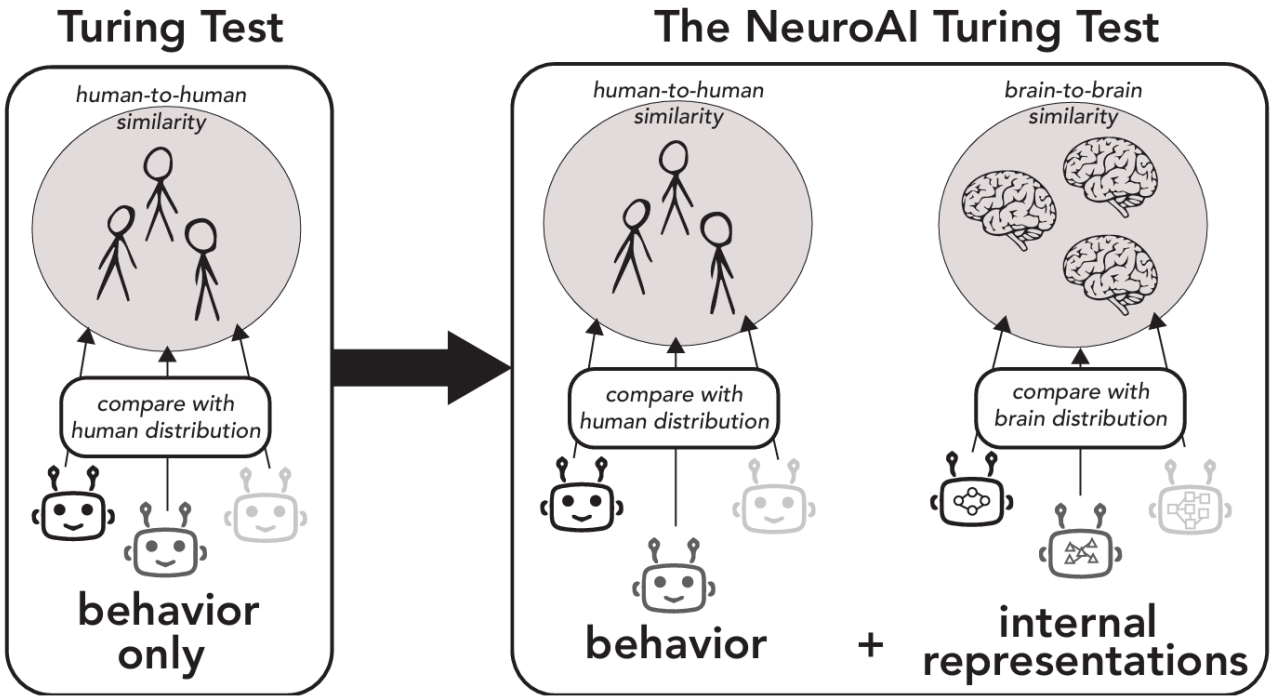
解釈可能性やモニタリングに関連する項目を報告します

[1] Yamakawa, Hiroshi. "Brain-Inspired AGI for Post-Singularity Symbiosis." *1st Workshop on Post-Singularity Symbiosis*.

解釈可能なモデルへのアプローチ

環境と神経活動の相互作用が知的行為を生むとし、
両者を含んだシステム全体の挙動の解釈を目指す

NeuroAI Turing Test [2]



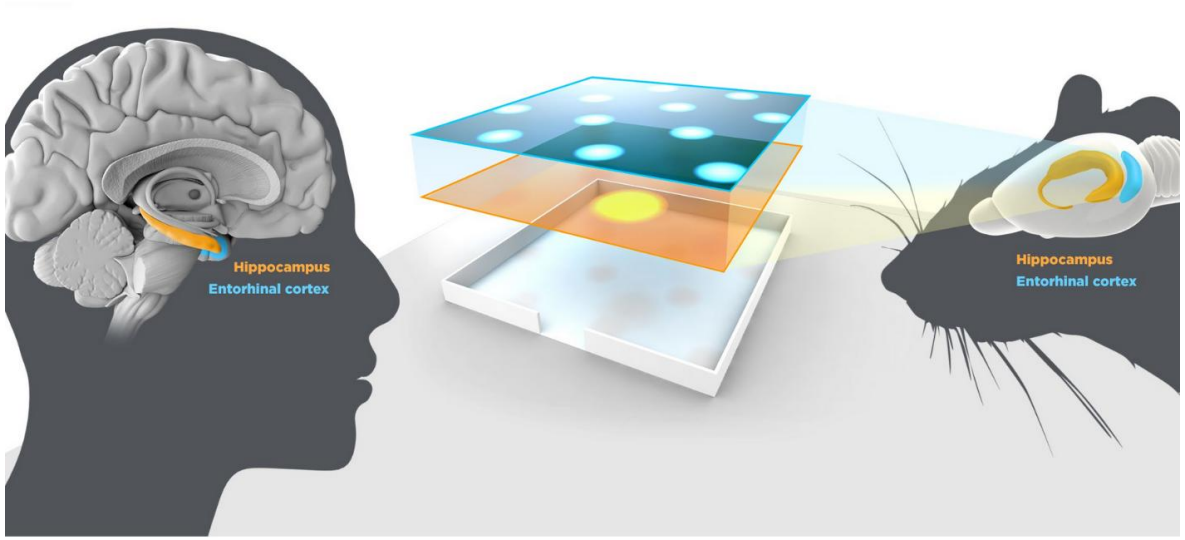
モデルの行動（出力）だけでなく、内部表象も生物と類似させる

[2] Feather, Jenelle, et al. "Brain-model evaluations need the neuroai turing test." arXiv preprint arXiv:2502.16238 (2025).

海馬体と認知地図

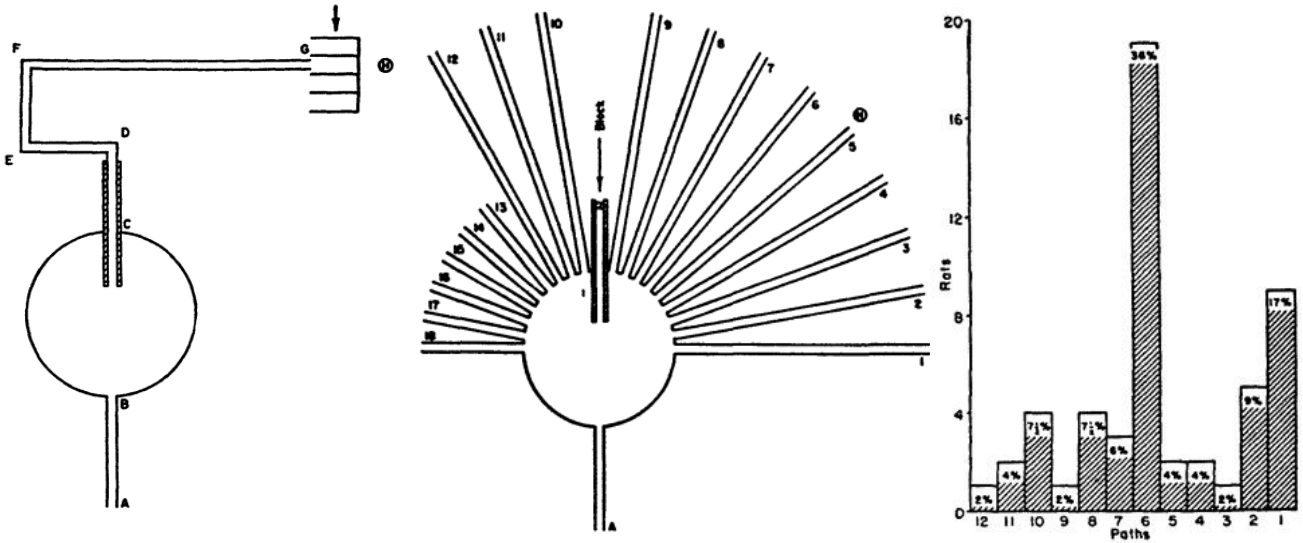
- **Place cell と Grid cell**[3]

脳におけるポジショニングシステムを構成する細胞
(Place cell やGrid cell)が海馬体で発見された
(2014 年ノーベル生理学・医学賞)



- **Tolmanの認知地図** [4]

環境をラットは,刺激−反射の対応を学習しているというよりも、
柔軟な行動を可能にする空間の脳内表象を獲得している



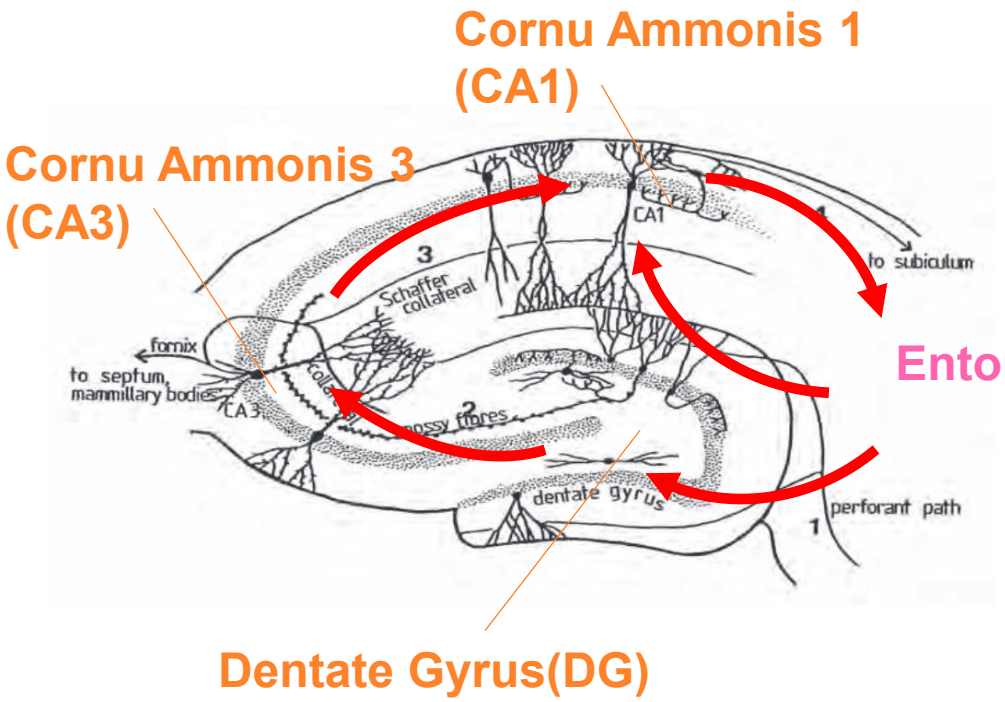
海馬体は、初めての状況でも目標までの適切な道筋を立てる能力
(ナビゲーション機能)の基盤

[3]The Nobel Prize in Physiology or Medicine 2014. NobelPrize.org.. https://www.nobelprize.org/prizes/medicine/2014/summary/
[4]Edward C Tolman. Cognitive maps in rats and men. Psychological review, Vol. 55, No. 4, p. 189, 1948.

空間認知モデル

構成論的アプローチで、ナビゲーション機能のメカニズムをあきらかにする

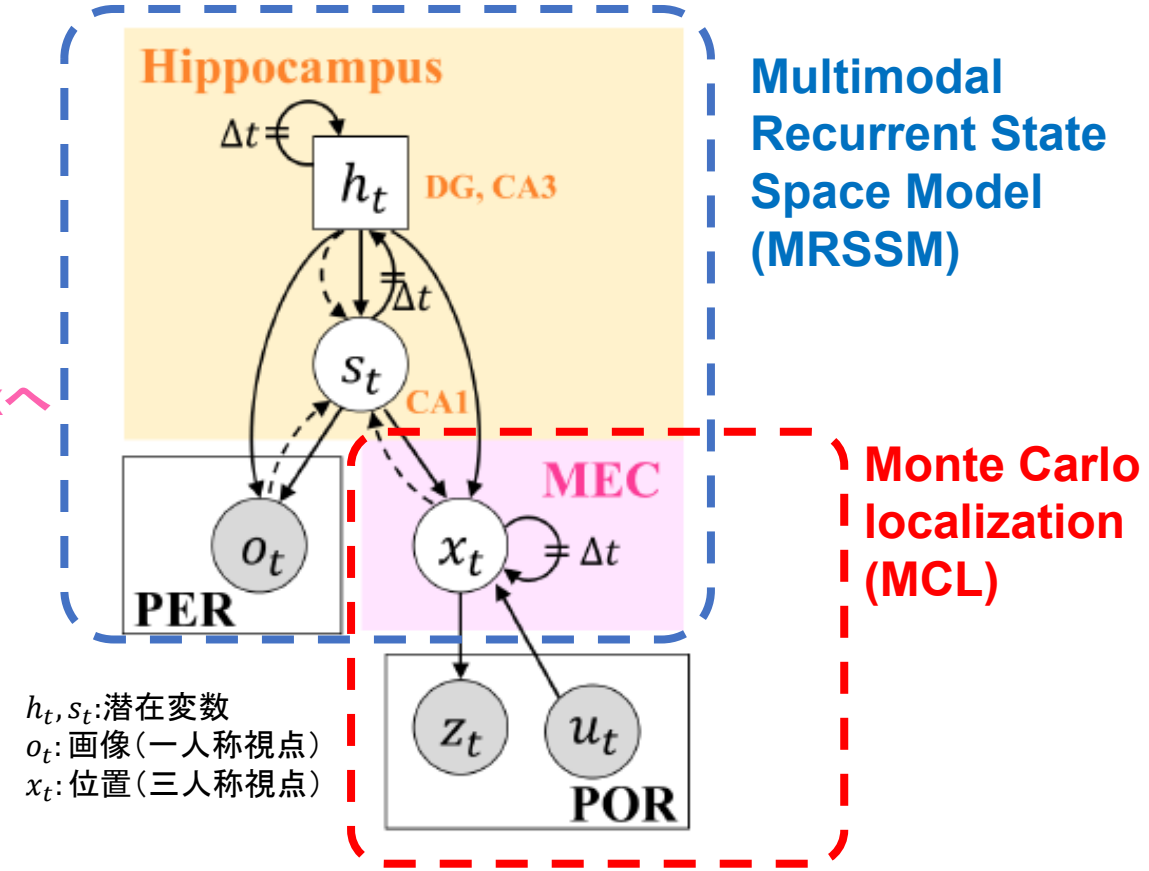
海馬のサブ領域と接続[5]



制約

Entorhinal cortexへ

確率的生成モデルのグラフィカルモデル[6]



海馬体の構造（アーキテクチャ）を参照し、空間認知モデルを構築

[5] E. T. Rolls, Brain computations and connectivity. Oxford University Press, 2023

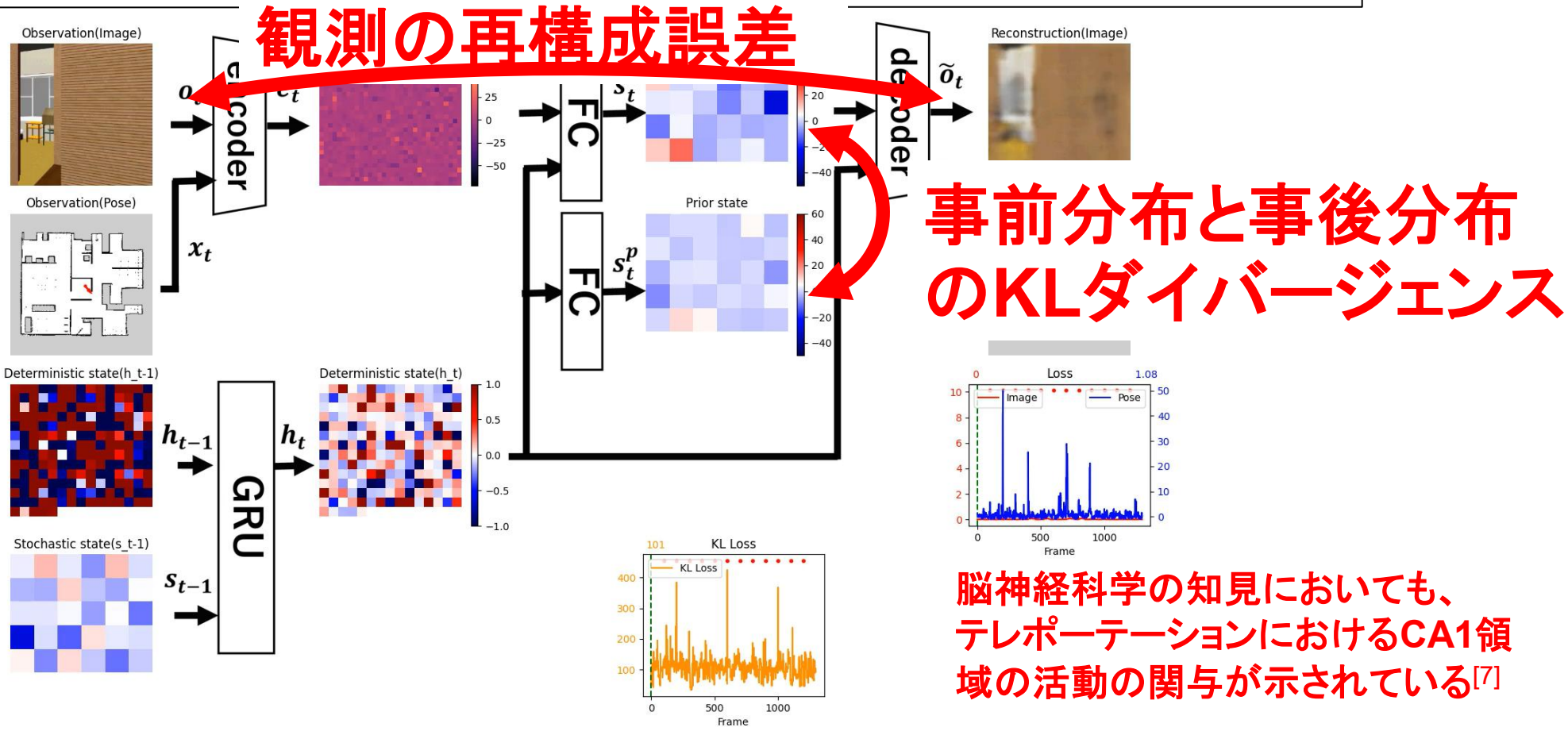
[6] Nakashima, Takeshi, et al., "Hippocampal formation-inspired global self-localization: quick recovery from the kidnapped robot problem from an egocentric perspective." *Frontiers in Computational Neuroscience* 18: 1398851.

異常検知タスクにおける空間認知モデル内部の挙動

生成モデルは、モデルの予測と実際の観測の誤差を用いて異常検知が可能

ナビゲーションの観測シーケンスデータに異常（ロボット誘拐問題）を人工的に作成してモデルに入力

どの確率変数に予測（再構成）誤差が生じるかを評価



脳神経科学の知見においても、テレポーテーションにおけるCA1領域の活動の関与が示されている[7]

脳参照アーキテクチャモデルは、脳神経科学の知見（機能局在）による解釈が可能

[7] Jezek, K , et al., Theta-paced flickering between place-cell maps in the hippocampus. Nature 478, 246–249. doi: 10.1038/nature10439

まとめ

- 海馬体は、（動的な）環境でのナビゲーション機能を担う脳部位である
- 構成論的アプローチにより、海馬体の構造を参照する空間認知の計算モデルを構築
- ナビゲーション中の異常（ロボット誘拐問題）におけるモデルの内部状態と、脳神経科学の機能局在の知見との類似性を示し、解釈性への貢献の可能性を示した

最後に

ICONIP2025
November 20-24, 2025

Special Session ”Toward Safe Brain-Inspired AI” で本講演に関する内容を発表予定

Interpretable Anomaly Detection in a Hippocampal Formation-inspired Spatial Cognition Model

Takeshi Nakashima[1 [0009−0009−3218−8590]], Akira Taniguchi[2 [0000−0003−0678−1103]], Tadahiro Taniguchi[3,4 [0000−0002−5682−2076]], and Hiroshi Yamakawa[5,6,7,8 [0000−0002−6981−0349]]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Osaka, Japan
nakashima.takeshi@em.ci.ritsumei.ac.jp
[2] College of Information Science and Engineering, Ritsumeikan University, Osaka, Japan
[3] Research Organization of Science and Technology, Ritsumeikan University, Shiga, Japan
[4] Graduate School of Informatics, Kyoto University, Kyoto, Japan
[5] The Whole Brain Architecture Initiative, Tokyo, Japan
[6] School of Engineering, The University of Tokyo, Tokyo, Japan
[7] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[8] AI Alignment Network, Tokyo, Japan.

Abstract. Interpretability is a key requirement in deploying AI systems for safety-critical tasks such as anomaly detection. Brain-inspired AI offers a promising path toward interpretable models by aligning internal representations with known structures and functions of the brain. To address the challenge of developing interpretable anomaly detection systems, we propose a novel approach based on a spatial cognition model inspired by the hippocampal formation. Our model introduces three anomaly scores, each corresponding to a probabilistic variable associated with a specific subregion of the hippocampal formation: the lateral entorhinal cortex, the medial entorhinal cortex, and the cornu ammonis-1, respectively. Experiments in a simulated home environment, involving different types of anomalies such as missing observation data, pose estimation failures, and contextual anomalies (e.g., the kidnapped robot problem), show that some of the anomaly scores are most effective in detecting the type of anomaly that activates the corresponding hippocampal formation subregion in actual neuroscience studies. These results suggest that our hippocampal formation-inspired model is capable of anomaly detection and offers insights into the functional localization of internal processing that mirrors neuroscientific findings, thereby enhancing interpretability. Our approach lays the groundwork for integrating neuroscientific principles into designing more transparent and trustworthy AI systems.

Keywords: Anomaly detection · Brain-inspired AI · Hippocampal formation · Interpretability · Spatial cognition

ご清聴ありがとうございました