自律ロボットの説明性と対話システム

2021.11.30 第12回対話システムシンポジウム

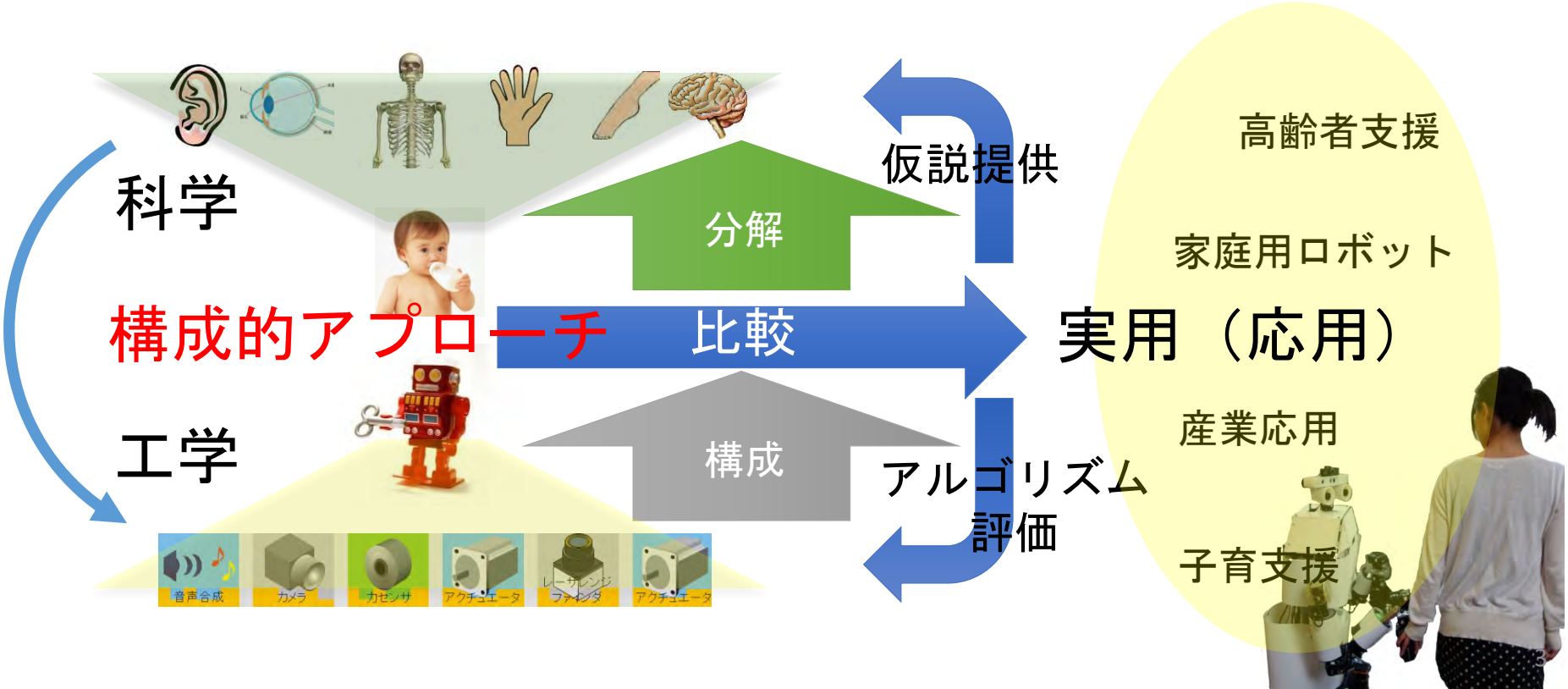
大阪大学大学院基礎工学研究科
電気通信大学人工知能先端研究センター
長井隆行

アウトライン

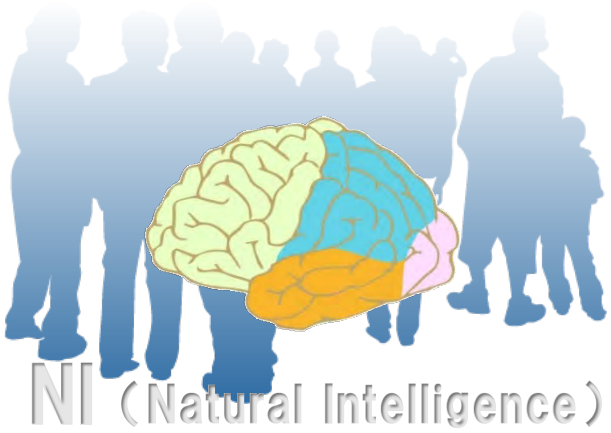
- 自己紹介
  - 研究のモチベーション
- AI×ロボット＝ロボット学習
  - ロボットが自ら創る世界のモデル
- 今後の課題
  - 自律ロボットはパートナーとなれるか？
- 自律ロボットの説明性（XAR）
  - 問題設定と課題
  - 手法の例
  - 対話研究との接合？
- まとめ

構成論の目指すもの

- **構成的**人間科学　⇒　人間の知能（心）を知りたい！そして作りたい！
  - 知能ロボティクス（ロボティクス、人工知能）　基盤技術
  - 認知発達（認知科学、発達心理学）　観察に基づくモデル
  - 記号創発（複雑系、言語学、脳科学）　計算モデル



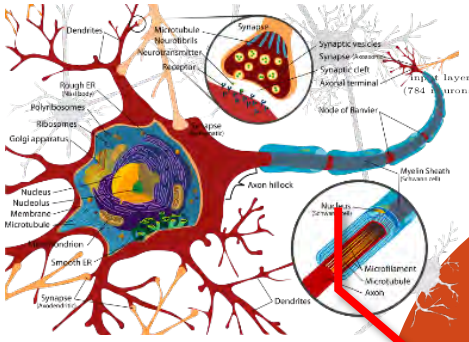
研究の方向性

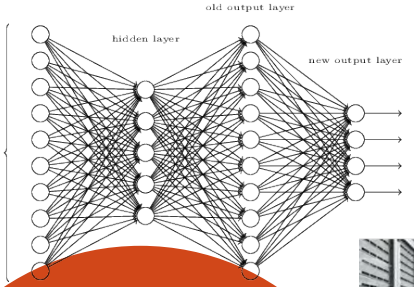


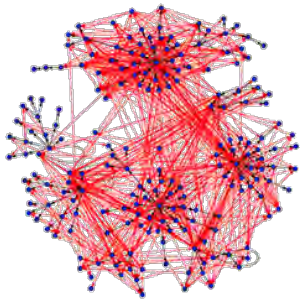
NI（Natural Intelligence）

VS

知る
創る







知能（ＡＩ）
脳
学習
コンピュータ

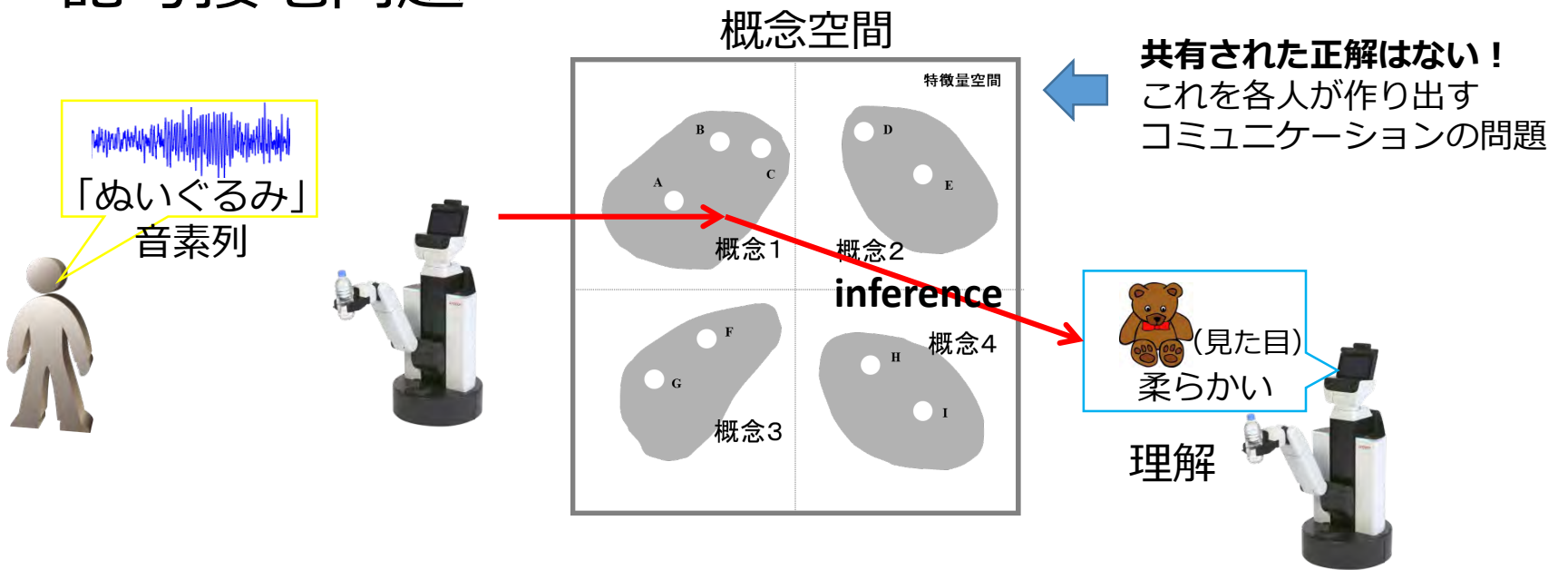
体
ロボティクス
メカ
制御

関係性
社会性
共感
コミュニケーション

ロボティクス×AI（Artificial Intelligence）

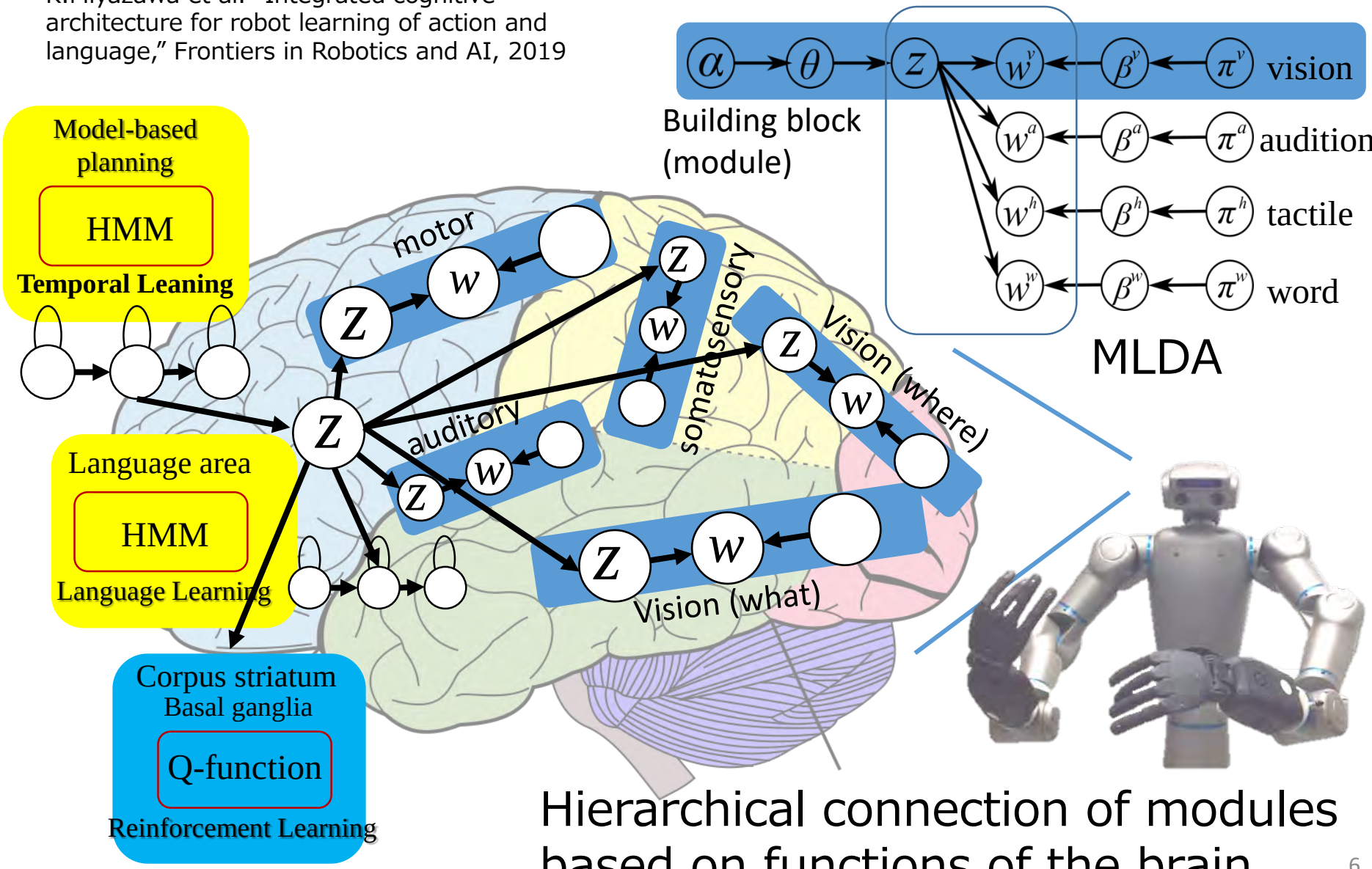
ナイーブな捉え方 ～“理解”とは何か？～

- ロボットによる実世界理解
  - **“理解”**：概念を通した未観測情報の予測
  - **“意味”**：予測した内容
  - **“概念”**：経験（マルチモーダルデータ）の
    カテゴリ分類によって形成される
- 記号接地問題



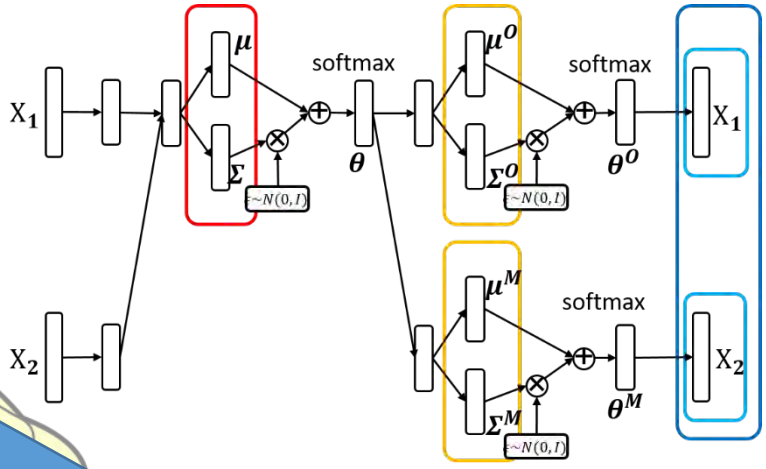
確率的生成モデルによる統合認知モデル (第一世代)

K.Miyazawa et al. “Integrated cognitive architecture for robot learning of action and language,” Frontiers in Robotics and AI, 2019



Hierarchical connection of modules based on functions of the brain

深層生成モデルによる統合認知モデル（第二世代）



Deep mMLDA

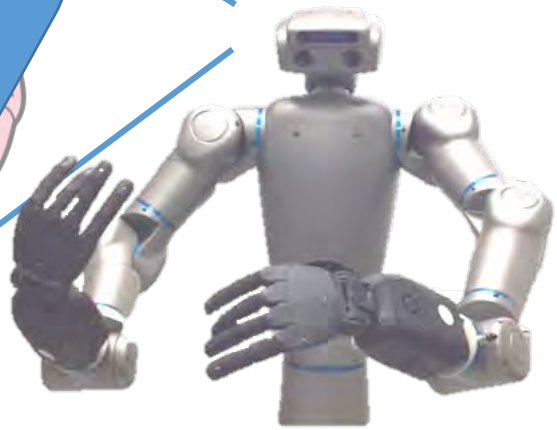
LSTM
(temporal learning)

Representation Learning
(State Space)

Latent Variables

(Deep) Q Network
RL

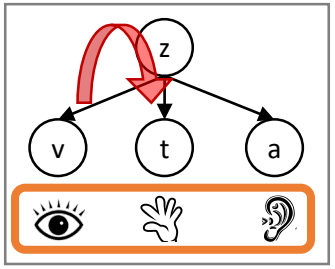
Reinforce



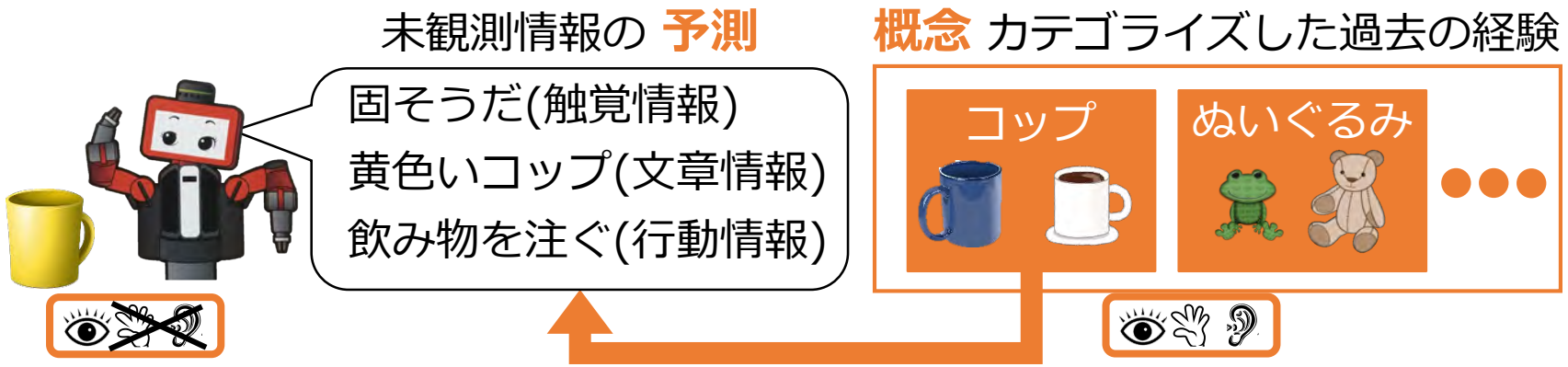
Hierarchical connection of modules based on functions of the brain

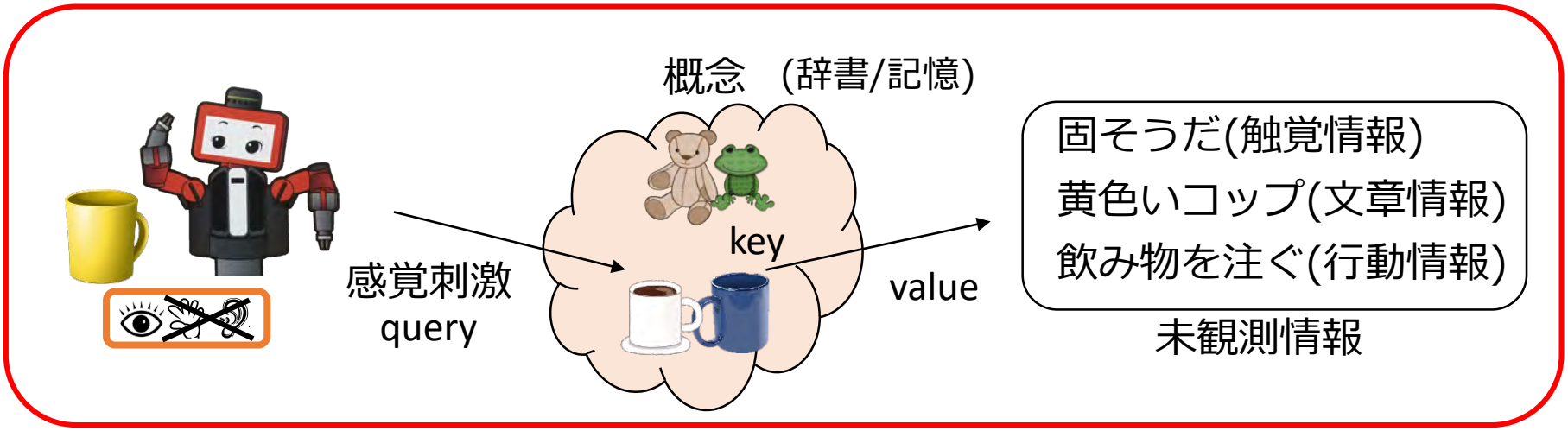
第三世代 アテンション機構を用いたマルチモーダル概念形成

宮澤 和貴, 青木 達哉, 堀井 隆斗, 長井 隆行, "Self-Attentionによる物体概念の形成", 第38回日本ロボット学会学術講演会, 1C2-05, 2020.10.9

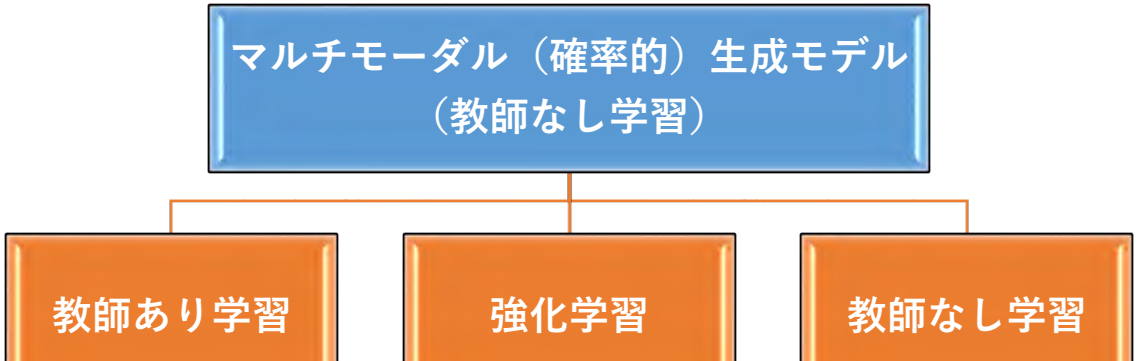


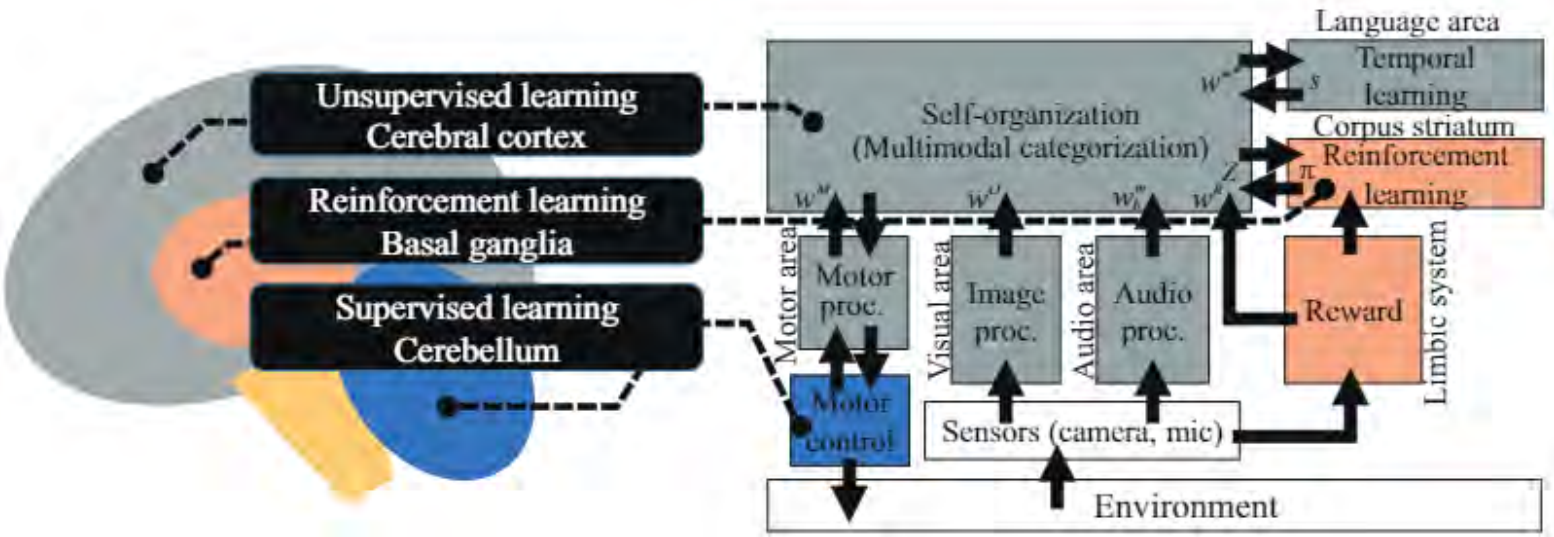
- 理解とは経験を通した**予測**であると考える
  予測をより効率的に行うために**概念形成**を行う





脳の構造との対応





Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?, Neural networks, 12, 961–974

ご興味のある方は論文をご覧ください

Computer Science > Artificial Intelligence

[Submitted on 15 Mar 2021]

Whole brain Probabilistic Generative Model toward Realizing Cognitive Architecture for Developmental Robots

Tadahiro Taniguchi, Hiroshi Yamakawa, Takayuki Nagai, Kenji Doya, Masamichi Sakagami, Masahiro Suzuki, Tomoaki Nakamura, Akira Taniguchi

Building a humanlike integrative artificial cognitive system, that is, an artificial general intelligence, is one of the goals in artificial intelligence and developmental robotics. Furthermore, a computational model that enables an artificial cognitive system to achieve cognitive development will be an excellent reference for brain and cognitive science. This paper describes the development of a cognitive architecture using probabilistic generative models (PGMs) to fully mirror the human cognitive system. The integrative model is called a whole-brain PGM (WB-PGM). It is both brain-inspired and PGMbased. In this paper, the process of building the WB-PGM and learning from the human brain to build cognitive architectures is described.

Comments: 55 pages, 8 figures, submitted to Neural Networks
Subjects: **Artificial Intelligence (cs.AI)**
Cite as: arXiv:2103.08183 **[cs.AI]**
(or arXiv:2103.08183v1 **[cs.AI]** for this version)

Submission history

From: Tadahiro Taniguchi [view email]
[v1] Mon, 15 Mar 2021 07:42:04 UTC (2,030 KB)

Whole brain Probabilistic Generative Model toward Realizing Cognitive Architecture for Developmental Robots

Tadahiro Taniguchi[a], Hiroshi Yamakawa[b,g], Takayuki Nagai[c], Kenji Doya[d], Masamichi Sakagami[e], Masahiro Suzuki[b], Tomoaki Nakamura[f], Akira Taniguchi[a]

[a] *Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Japan*
[b] *Tokyo University, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan*
[c] *Osaka University, 1-3 Machikane-yama, Toyonaka, Osaka, Japan*
[d] *Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami, Okinawa, Japan*
[e] *Tamagawa University, 6-1-1 Tamagawa Gakuen, Machida, Tokyo, Japan*
[f] *The University of Electro-Communications, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan*
[g] *The Whole Brain Architecture Initiative, Nishikoiwa 2-19-21, Edogawa-ku, Tokyo, Japan*

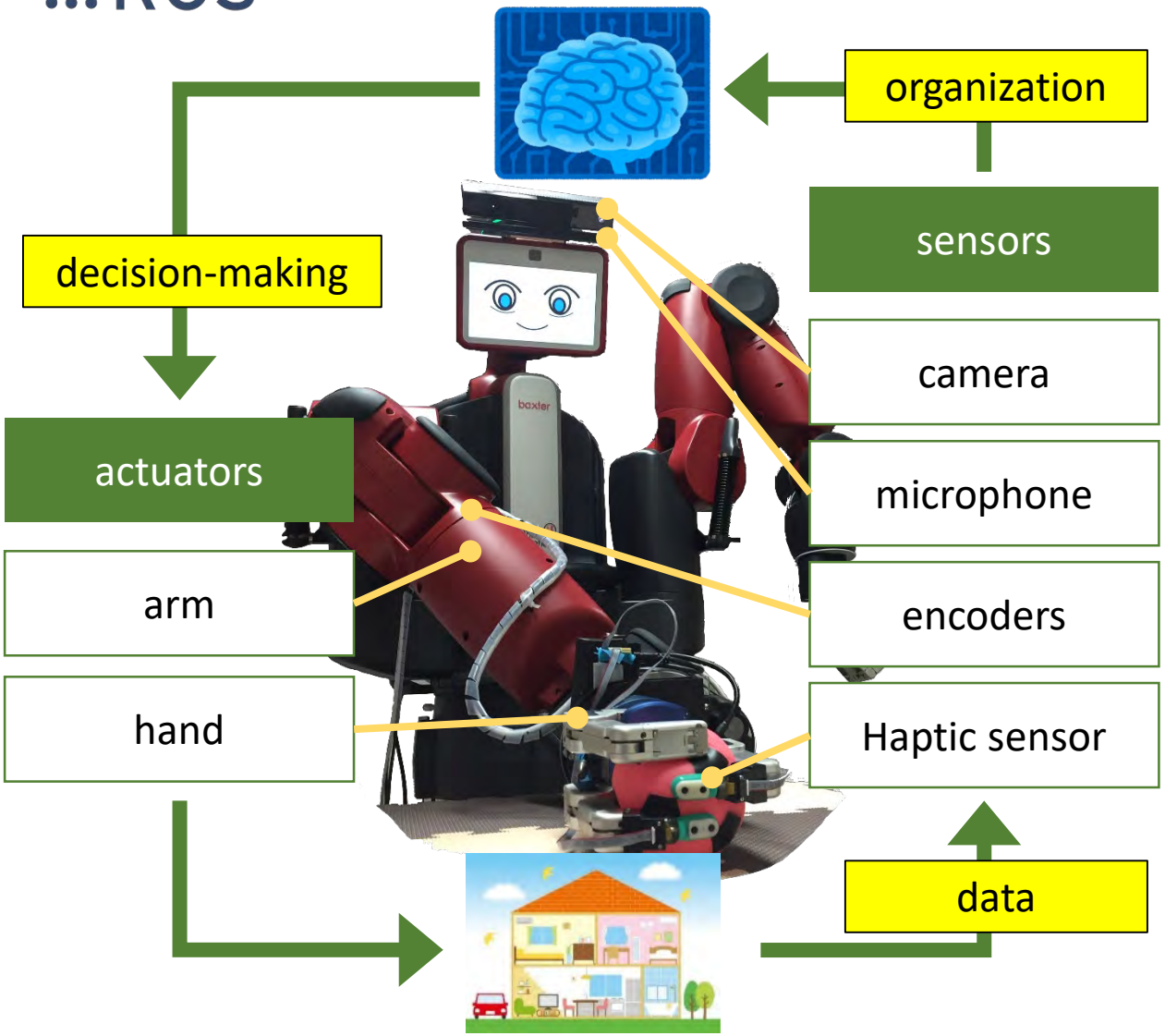
Abstract

Building a humanlike integrative artificial cognitive system, that is, an artificial general intelligence, is one of the goals in artificial intelligence and developmental robotics. Furthermore, a computational model that enables an artificial cognitive system to achieve cognitive development will be an excellent reference for brain and cognitive science. This paper describes the development of a cognitive architecture using probabilistic generative models (PGMs) to fully mirror the human cognitive system. The integrative model is called a whole brain PGM (WB-PGM). It is both brain-inspired and PGM-based. In this paper, the process of building the WB-PGM and learning from the human brain to build cognitive architectures is described.

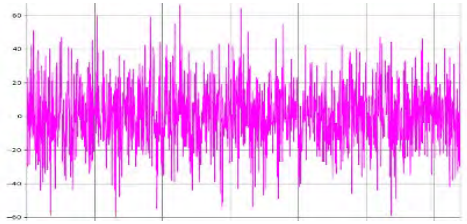
Keywords: Cognitive architecture, Probabilistic generative model, Brain-inspired artificial intelligence, Artificial general intelligence, Developmental robotics

ロボットへの実装

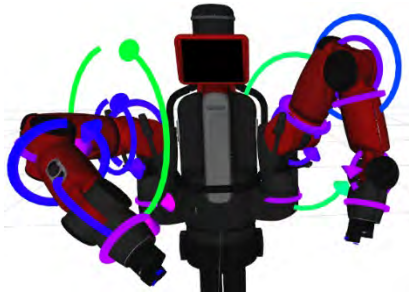
ROS



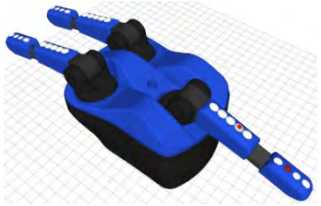
image



sound



proprioception

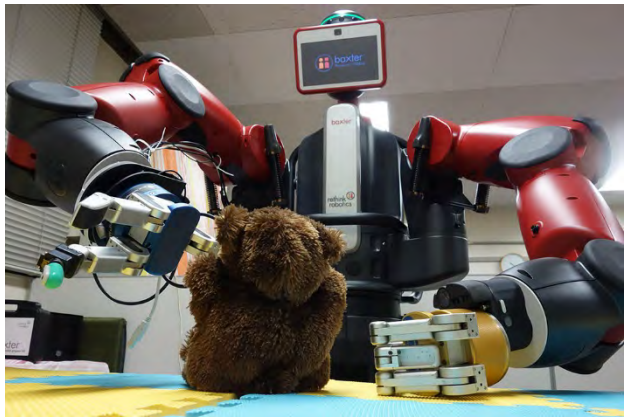


Tactile sensor

Data from various sensors

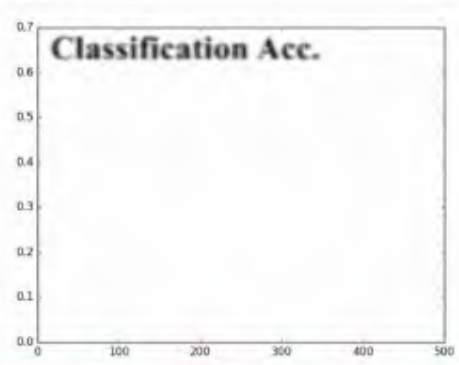
長期間学習実験 [Nishihara+ 16]

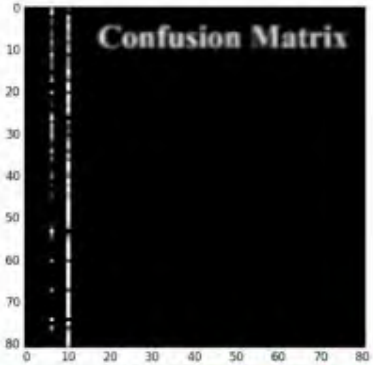
- 3 ~ 5 時間/日 ×　1か月 (100 時間以上)
- 500 物体 (81 カテゴリ) を使用

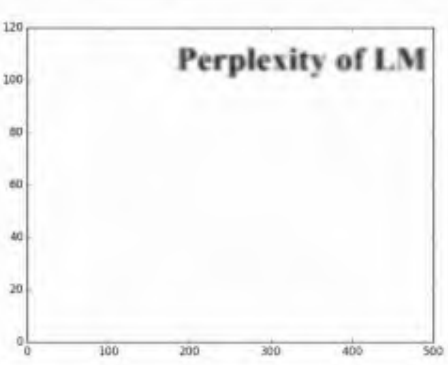


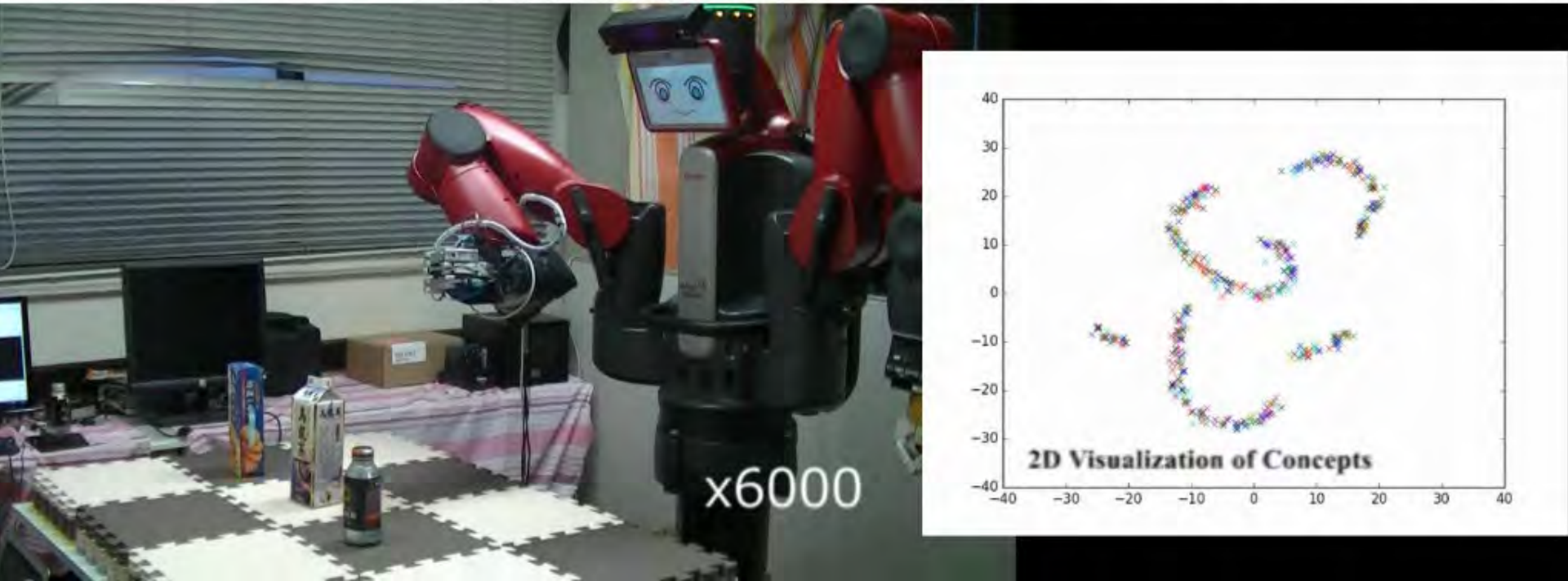
長期間学習実験 [Nishihara+ 16]

- 3000以上の発話（300ユニーク単語）100単語程度獲得
- 約7割程度の正解率



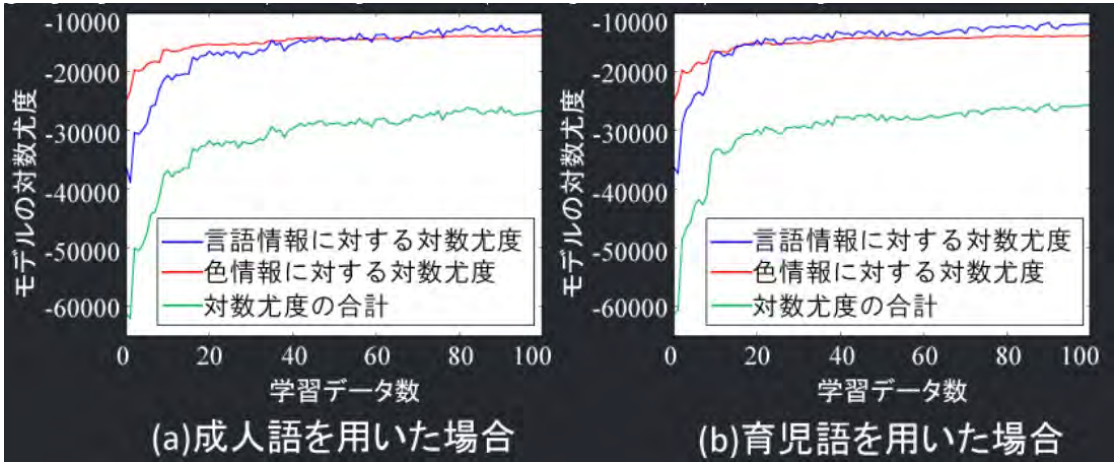


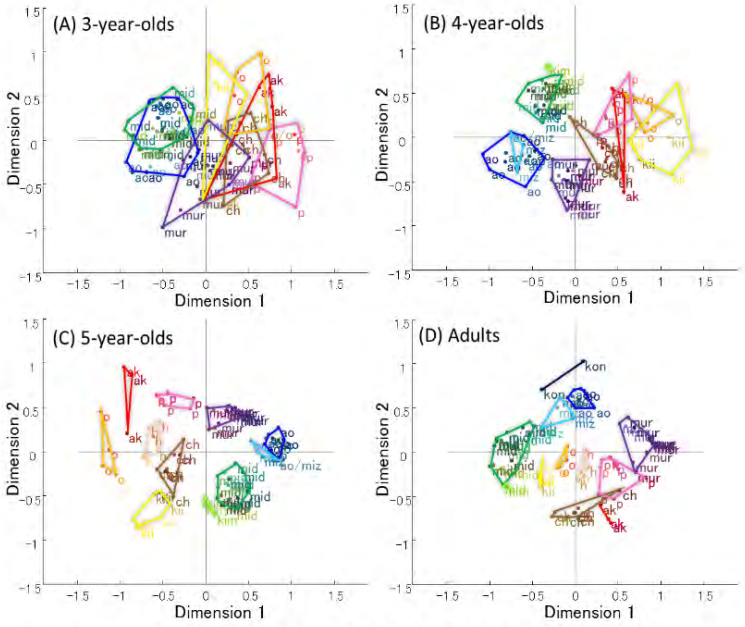




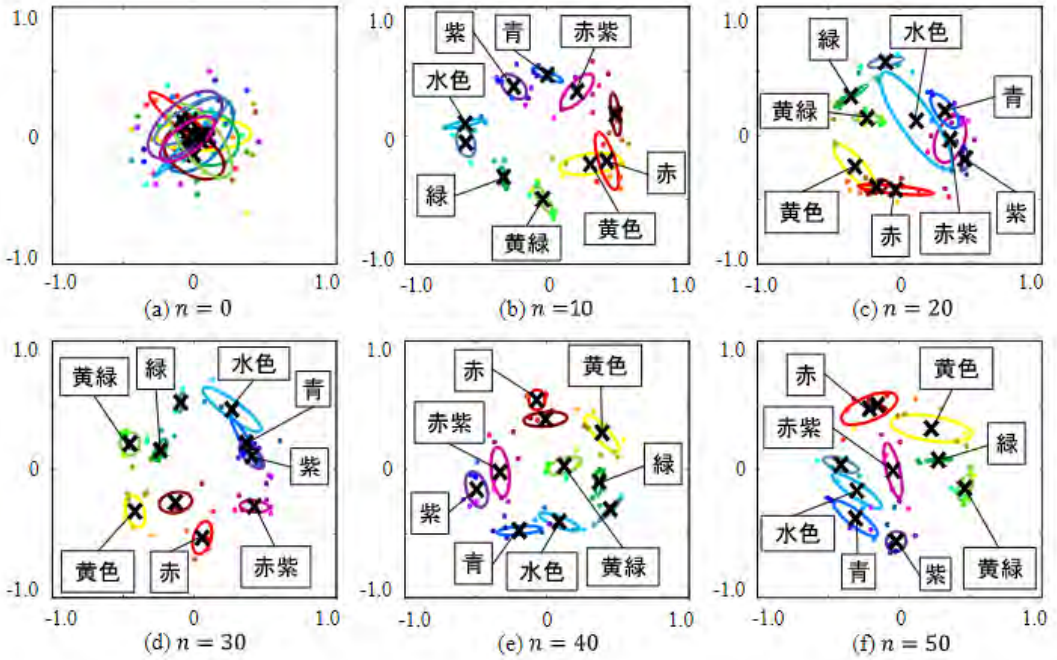
人との比較によって見えてくるもの

- ロボットの学習を解析することで得られた示唆
  - 幼児語の果たす役割 [Funada+ 17]
  - 色の学習 [Funada+ 16]

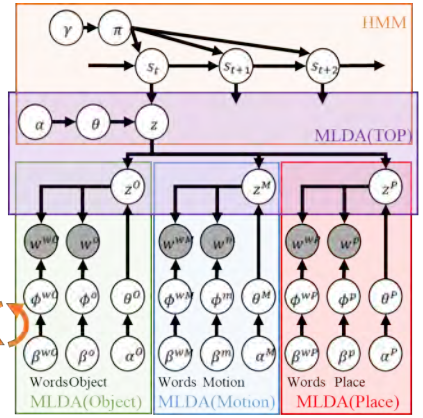


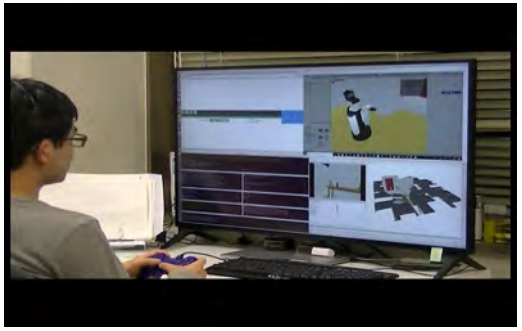


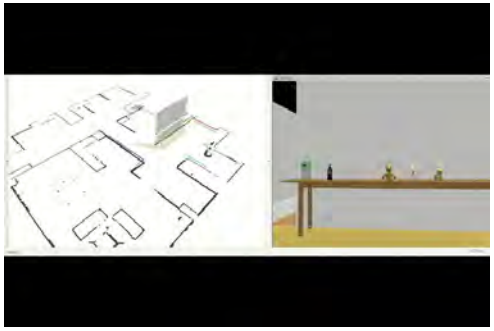
ak = 赤, ao = 青, ch =茶色, h = 肌色, kii = 黄色, kim = 黄緑 mid =緑, miz =水色, mur =紫, o =オレンジ色, p =ピンク色



Learning from Teleoperation for Domestic Service Robots







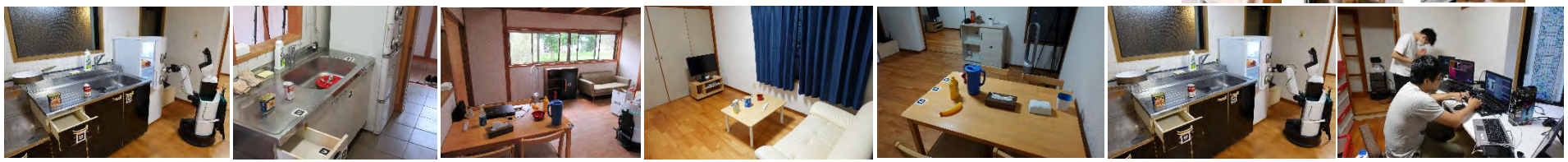
K.Iwata et al. “Learning and generation of actions from teleoperation for domestic service robots,” IROS 2018

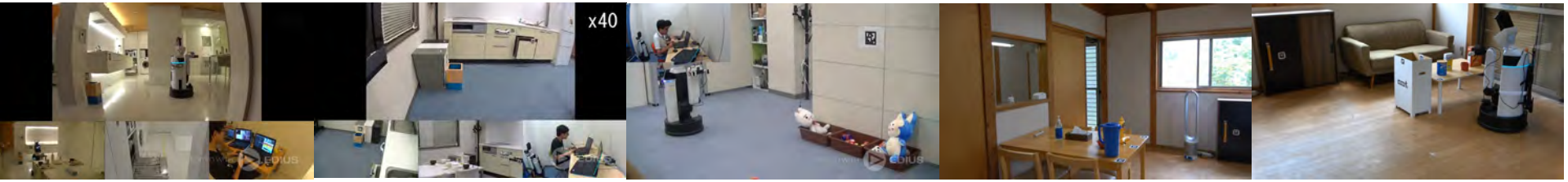
K.Miyazawa et al. “Integrated cognitive architecture for robot learning of action and language,” Frontiers in Robotics and AI, 2019

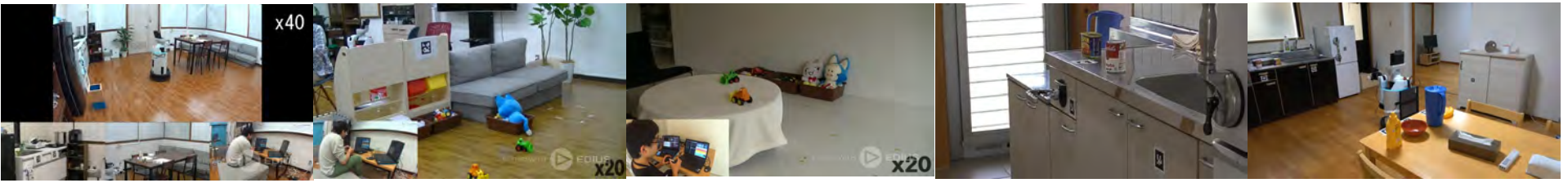
Currently working on experiments @ three different real home environments

実家庭でのデータ収集

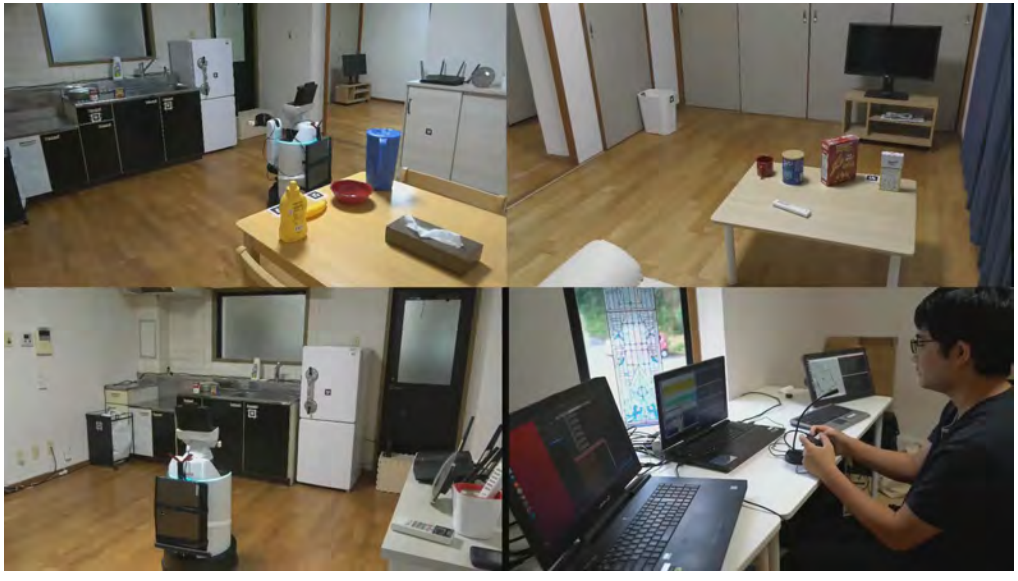
- ~10 Real home environments



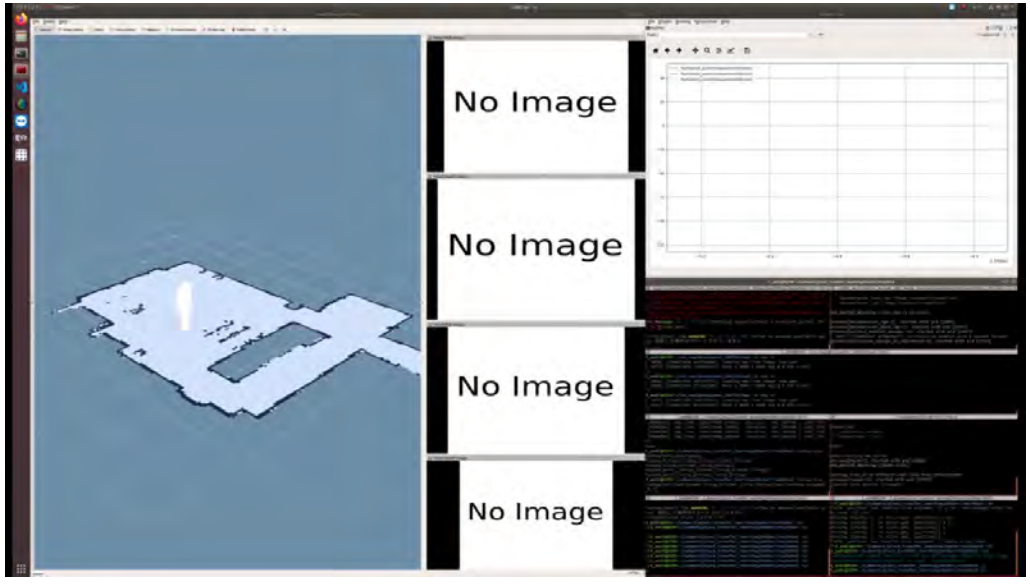




実施例(Dinning & Kitchen + Living)
総時間約4,050分(=67.5時間)のデータを収集

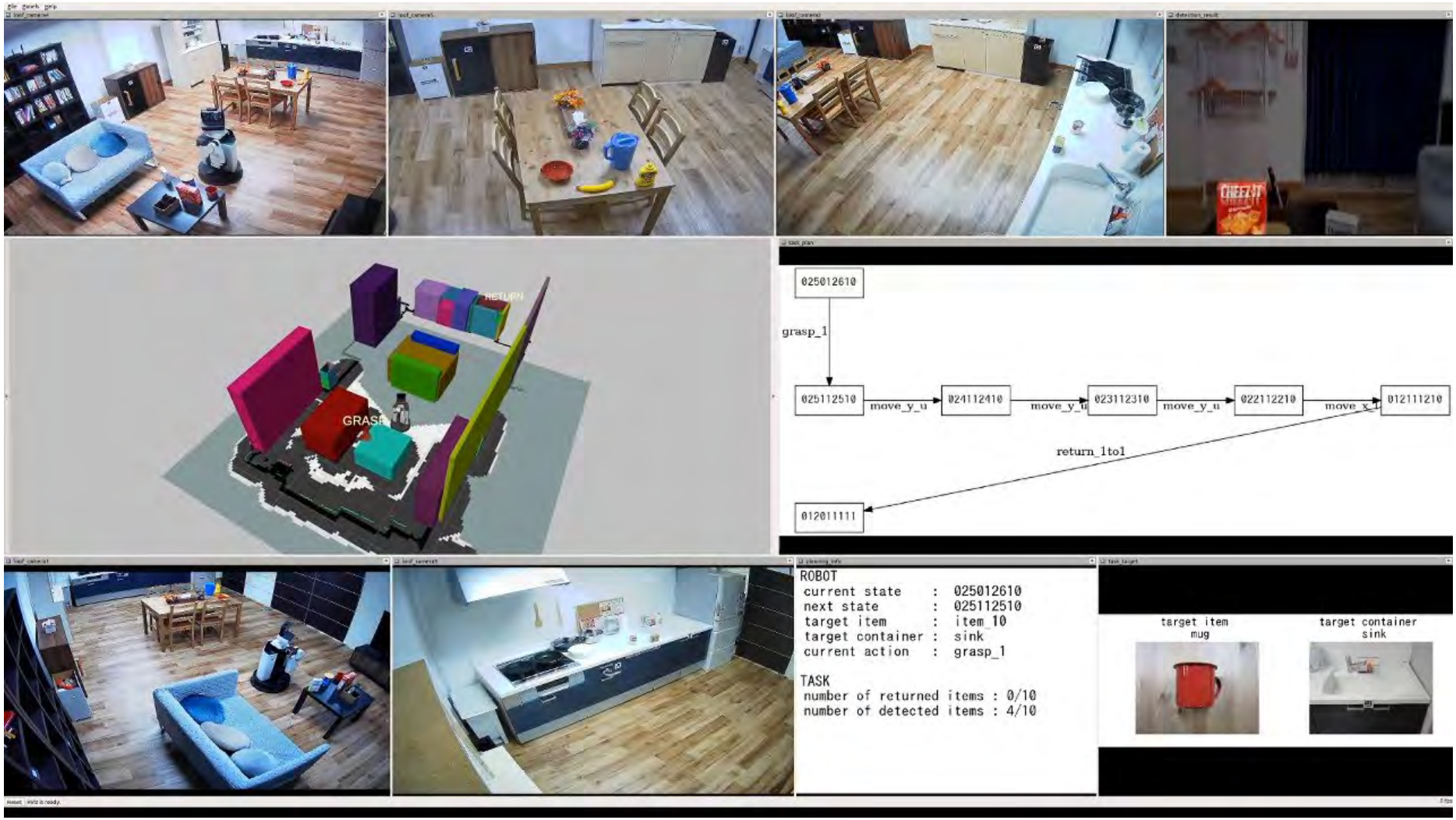


実験の様子



取得データの例

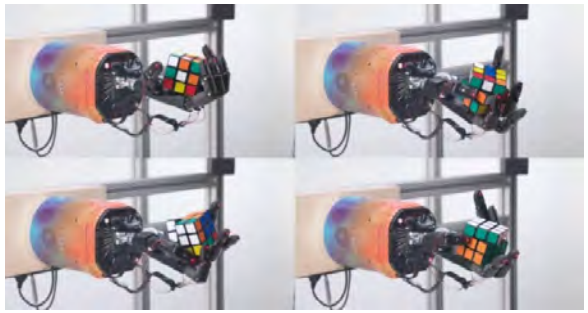
HSRによる**自律片付け**デモ [2020]



足りないもの・・・

- 自身の経験から学ぶロボットはなんとなくできた
- しかし足りないものも多い
  - ロボットの身体能力を向上させる
    - ダイナミックな動作の学習
    - 器用さの向上
  - より高次な認知機能
    - 論理的な推論
    - 演繹、アブダクション
    - 因果性
  - コミュニケーション
    - 自他の分離
    - ミラーニューロンシステム
  - 情動・感情
    - 報酬の考え方
    - 模倣学習と強化学習

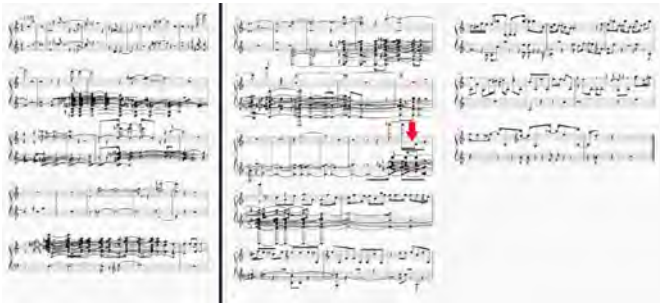
フルヒューマノイドロボットが逆上がり練習中
（シミュレーション＆sim2real）



sudo apt install intelligence　　様々な経験　　10歳まで育てる

これからの課題

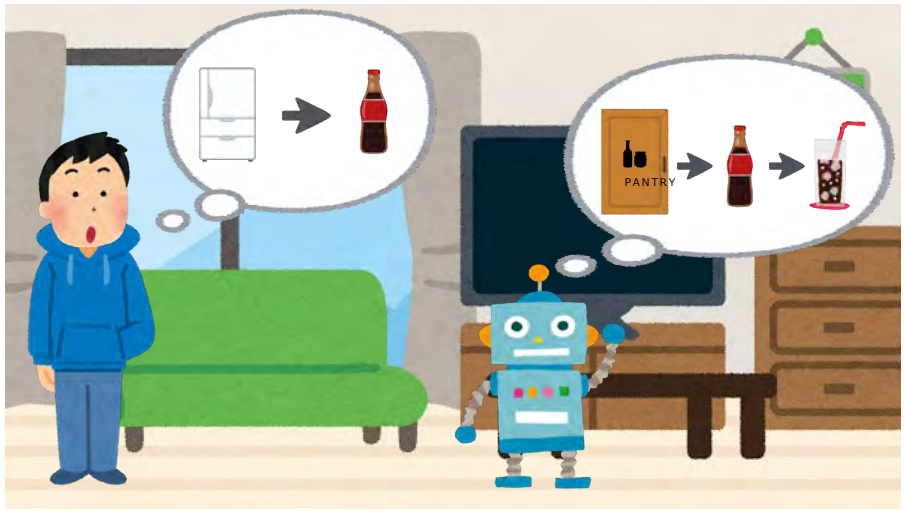
- ロボットの身体能力を向上させる
  - ダイナミックな動作の学習
  - 器用なタスク
- 感情
  - 感情の計算モデルの研究が進みつつある
  - ロボットも感情をもてる！
- 社会性
  - 人の社会性については色々と研究が進んでいる
  - ロボットにも社会性をもたせたい
- 創造性
  - 創造性とは何か？
  - 機械学習（生成モデル）の発展が研究を後押し
- **説明性**
  - ほぼコミュニケーションと同義
  - 社会的な価値にも関連する
- 意識
  - 人の意識が十分に解明されていない
  - まだまだ難しい



自律ロボットはパートナーになれるか？

- ロボットが自律的になればなるほど
  - 人がすぐに理解できない行動が増える
  - ロボットにどのようにお願いすればよいのか不明
- ロボットとの信頼関係が築けない

ユーザー：ロボットに冷たいコーラを取ってくるように頼む
ロボット：冷蔵庫ではなくパントリーの方向に動く
ユーザー：・・・（説明してほしい・・・）



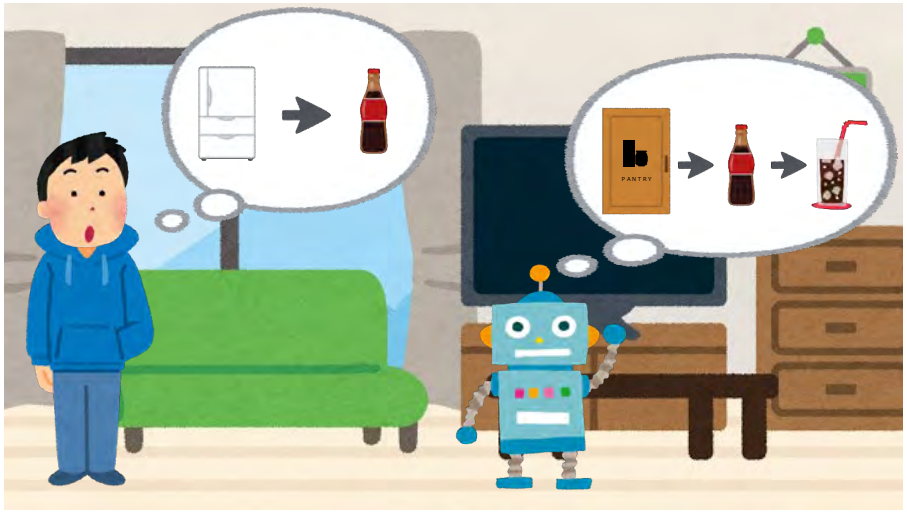
キーワードは
説明性

説明性とは何か？

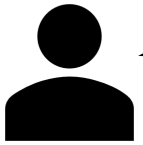
家庭環境における説明事例

ユーザー：ロボットに冷たいコーラを
取ってくるように頼む

ロボット：冷蔵庫ではなく
パントリーの方向に動く

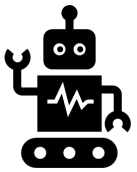


What-question：今何をしているのか？



冷蔵庫はそっちじゃないけど、何しに行くの？

パントリーにあるコーラを取りに行きます

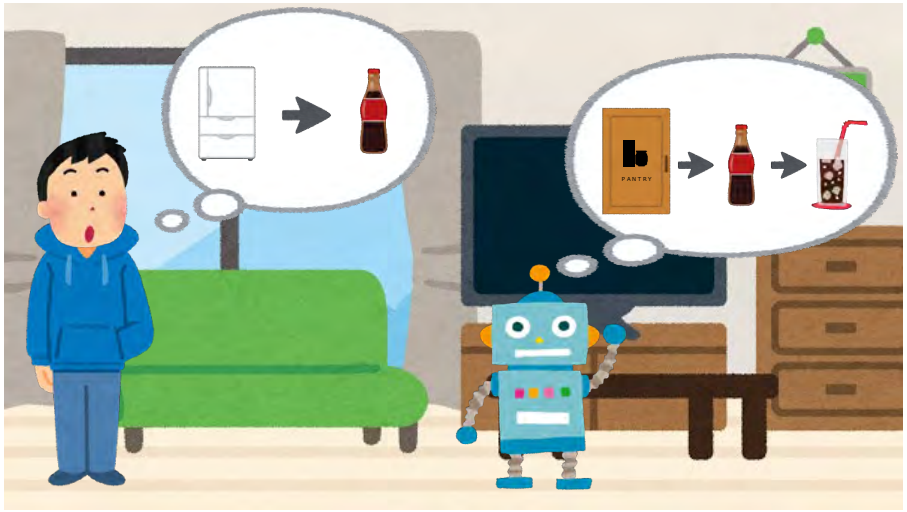


説明性とは何か？

家庭環境における説明事例

ユーザー：ロボットに冷たいコーラを
取ってくるように頼む

ロボット：冷蔵庫ではなく
パントリーの方向に動く



How-question：どのように実現するのか？

パントリーのコーラはぬるいよ

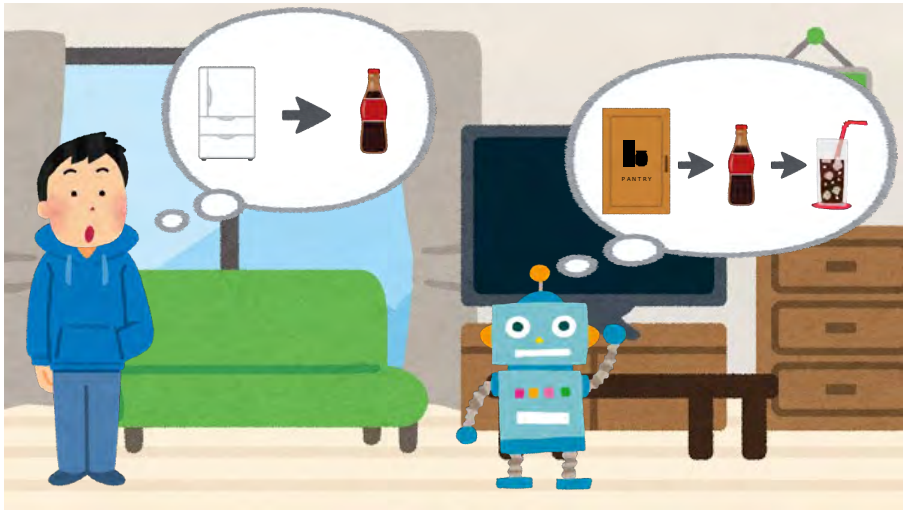
氷の入ったコップに注いで渡します

説明性とは何か？

家庭環境における説明事例

ユーザー：ロボットに冷たいコーラを
取ってくるように頼む

ロボット：冷蔵庫ではなく
パントリーの方向に動く



Why-question：なぜそのような行動をとるのか？

あれ、冷蔵庫にコーラなかったっけ？

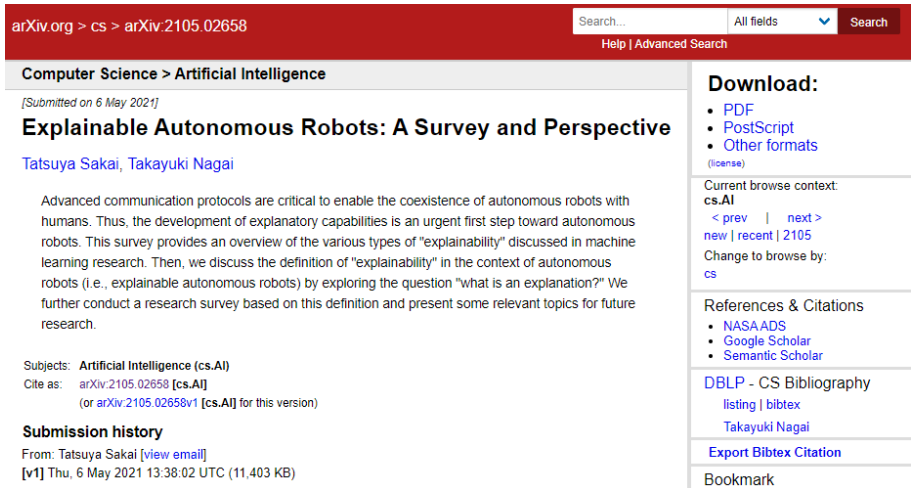
今朝弟が飲み干してしまいました

説明とは何か？

- 説明の分類 [Miller 19]
  - What：ロボットが何をするかを示すという文脈で研究されている
  - How：因果推論と関係 (whyで言い換えることもできる)
  - Why：これが難しく重要な問題
- 説明はなぜ必要か？ [Malle 06]
  - Discovery of meaning
    - 新しい知識（意味）の発見や矛盾・誤りの解消
  - Maintenance of social interaction
    - 相手との共有信念や関係性の構築
    - 相手に対する印象や感情にも影響を与える

[Miller 19] Tim Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Vol. 267, pp. 1 – 38, 2019
[Malle 06] Bertram F Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, Mit Press, 2006

ご興味のある方は是非サーベイ論文をご覧ください！




SURVEY PAPER

Explainable Autonomous Robots: A Survey and Perspective

Tatsuya Sakai[a] and Takayuki Nagai[a,b]

[a]Graduate School of Engineering Science, Osaka University, Osaka, Japan; [b]Artificial Intelligence Exploration Research Center, The University of Electro-Communications, Tokyo, Japan

ABSTRACT

Advanced communication protocols are critical to enable the coexistence of autonomous robots with humans. Thus, the development of explanatory capabilities is an urgent first step toward autonomous robots. This survey provides an overview of the various types of "explainability" discussed in machine learning research. Then, we discuss the definition of "explainability" in the context of autonomous robots (i.e., explainable autonomous robots) by exploring the question "what is an explanation?" We further conduct a research survey based on this definition and present some relevant topics for future research.

KEYWORDS

Autonomous agents; Autonomous robots; Explainability; Interpretability

1. Introduction

Artificial intelligence (AI) technologies have demonstrated remarkable progress and they are employed in a wide variety of applications in various fields including automatic translation, image recognition, and medical diagnosis [1–3]. It is commonly claimed that AI will replace most manual labor in the future; however, is this really the case? AI technologies do have higher image recognition accuracy compared to humans in some limited contexts, and have consistently outperformed humans in classical games such as Go and chess. Nonetheless, we believe that even advanced future developments based on current technology will not lead to robots replacing humans.

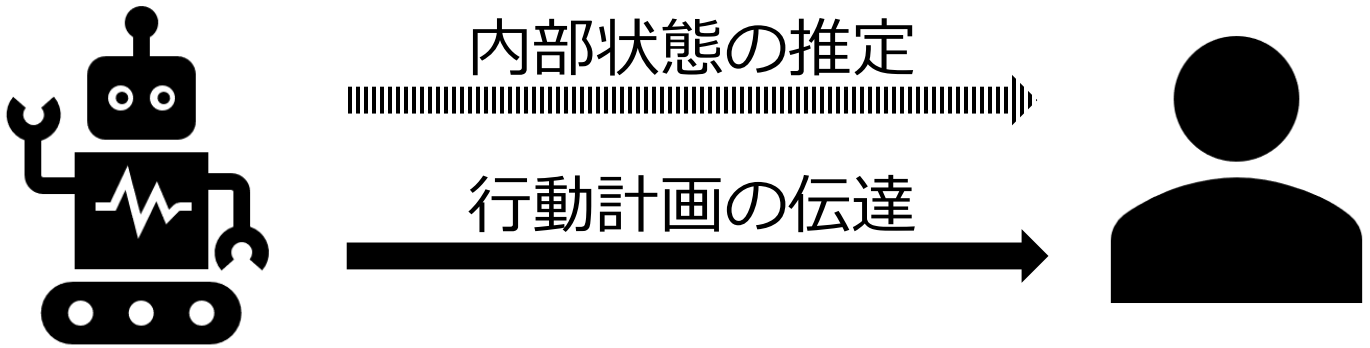
AI systems' fundamental lack of ability to communicate naturally and effectively with humans is among the most significant reasons that they cannot replace human labor. Here, one may believe that such communication could be achieved via the development of natural language processing (NLP) technology [4]; however, NLP technologies are systems for estimating the content of human statements and their meanings; they do not constitute communication. That is, humans do not feel that robots using such systems truly understand and respond to them appropriately. Therefore, if

Tatsuya Sakai and Takayuki Nagai,
Explainable Autonomous Robots: A Survey and Perspective,
arXiv cs.AI, 2021

XAR : EXplainable Autonomous Robots

自身の**行動や目標を**ユーザーが理解しやすい形で**伝える**

相手の内部状態を推定して行動や情報提示方法を変える



XARの定義

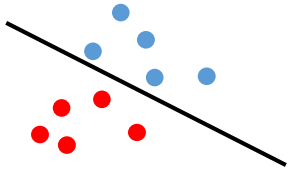
XARはロボットと人間が意思疎通を図るための研究分野

■XAR(Explainable Autonomous Robots)

自律ロボットと人間が意思疎通を図るための説明を生成
記号推論でない意思決定やプランニングに関して透明性を確保

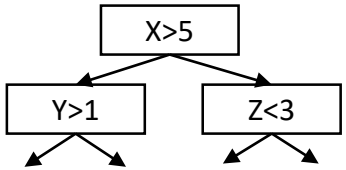
■XAI(Explainable AI)

対象：分類モデル
予測結果に寄与した特徴量を提示



■XRL(Explainable RL)

対象：強化学習モデル
ポリシーを人が理解できる形で可視化



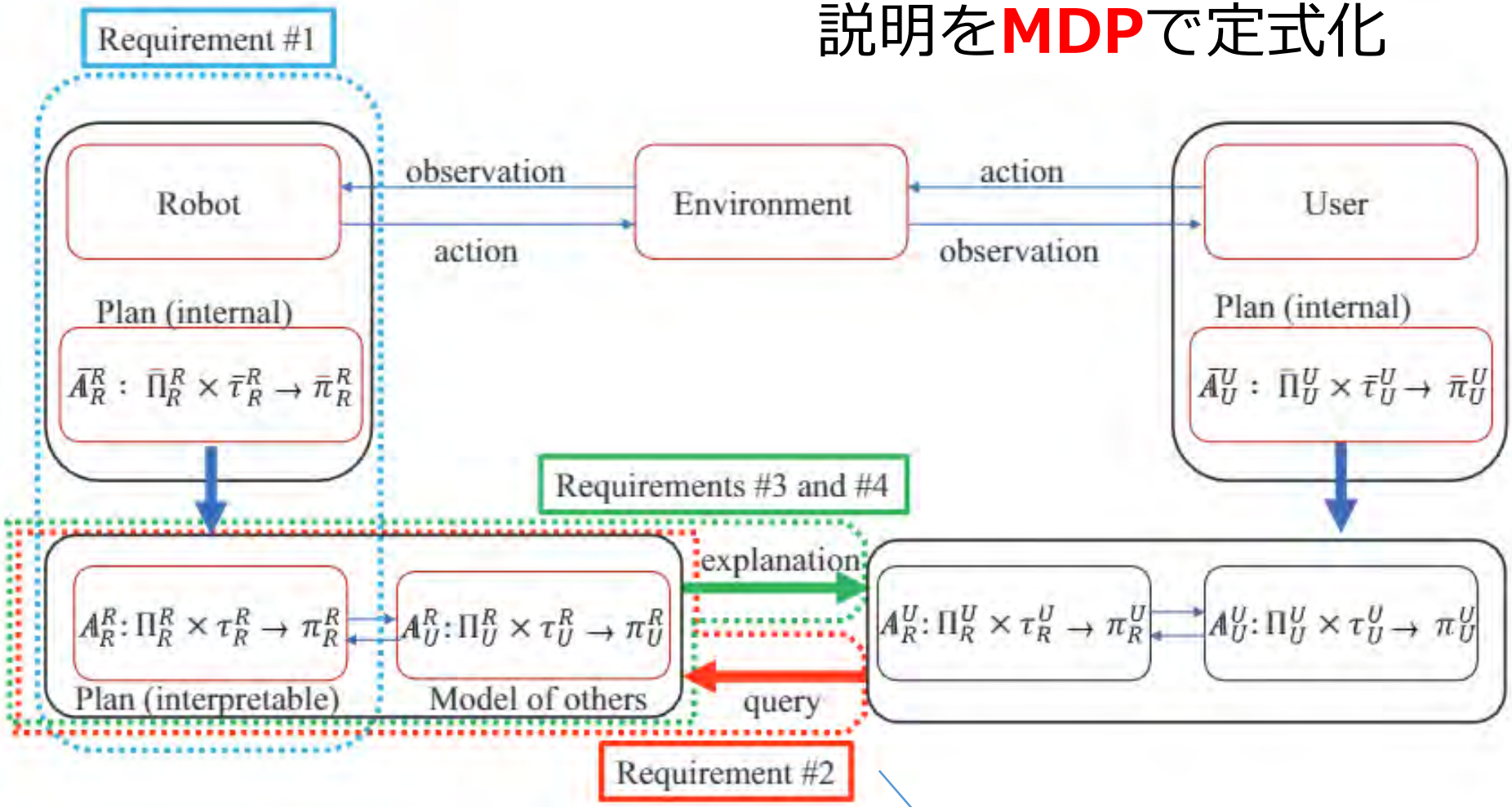
■XAIP(Explainable AI Planning)

対象：記号的なプランニング
プランの予測が難しい状況時に説明を補足

A→B→C

想定する説明の全体像

説明を**MDP**で定式化



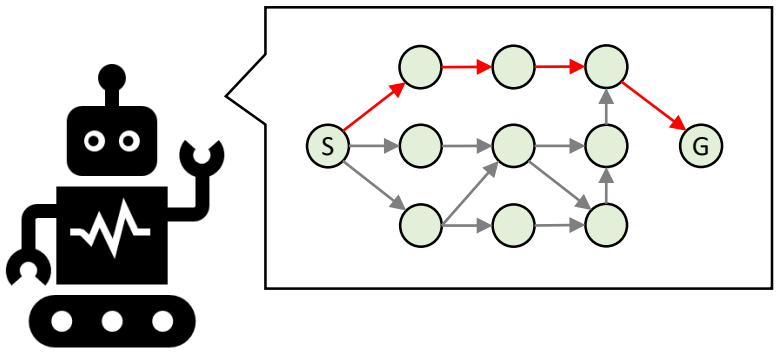
π：ポリシー
Π：意思決定空間（MDP）
τ：制約（報酬最大化や説明性など）

本来は双方向のプロセス

XARの定義（全体）

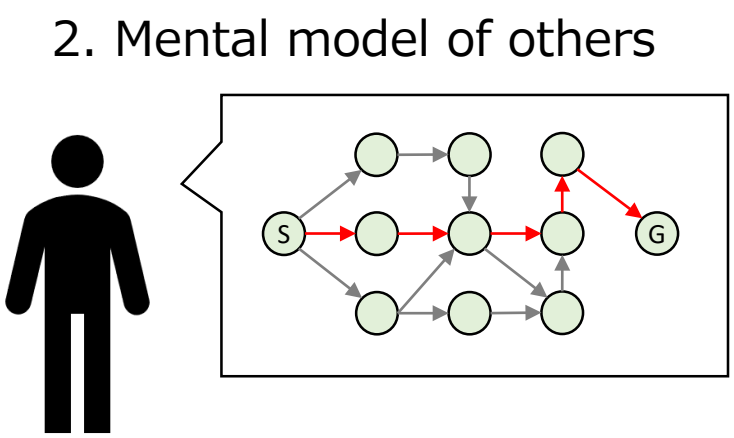
XARに求められる４つの要件

1. Interpretable decision space



3. Customized explanation for individuals

2. Mental model of others



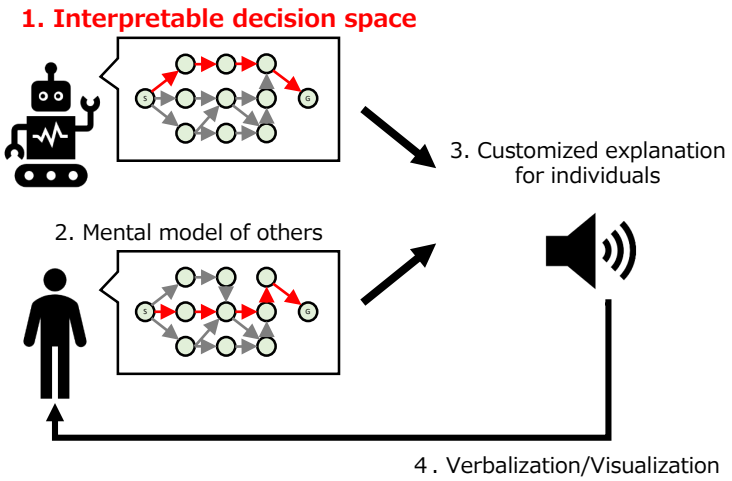
4. Verbalization/Visualization

XARの要素（その１）

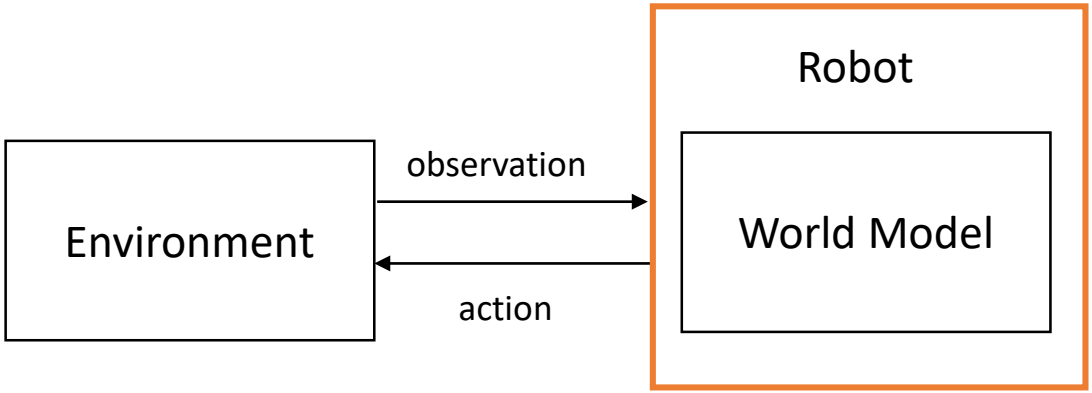
１：解釈可能な世界モデルの保持

環境とのインタラクションを通して
解釈可能なWorld Model[1]を獲得

[1]例えば　World Models[David Ha+,2018]

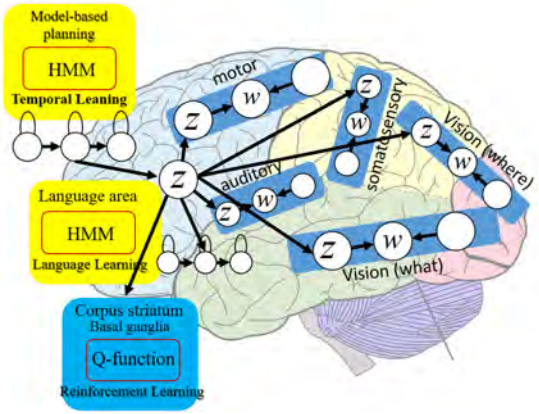


▶ 人間の作りこみに依らない意思決定空間の構築

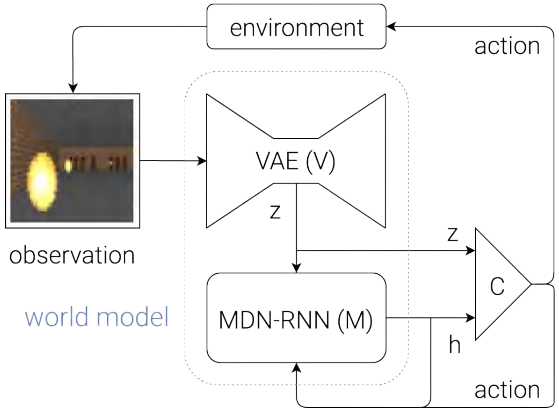


前半の話はここをどのように作るかという話
つまり、ロボットが自ら作り出す世界のモデルが基盤となる！
言語との結びつきも世界モデルが基盤

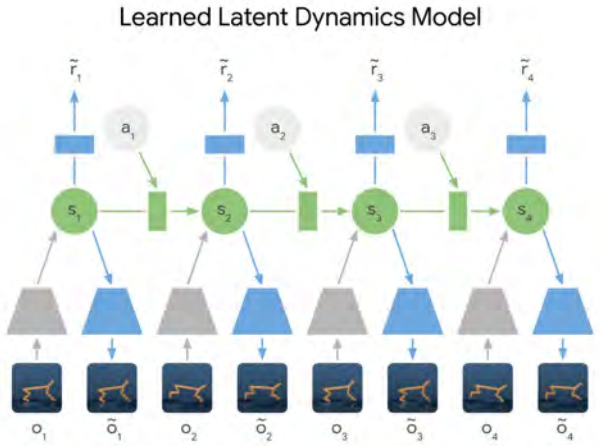
世界モデルの例



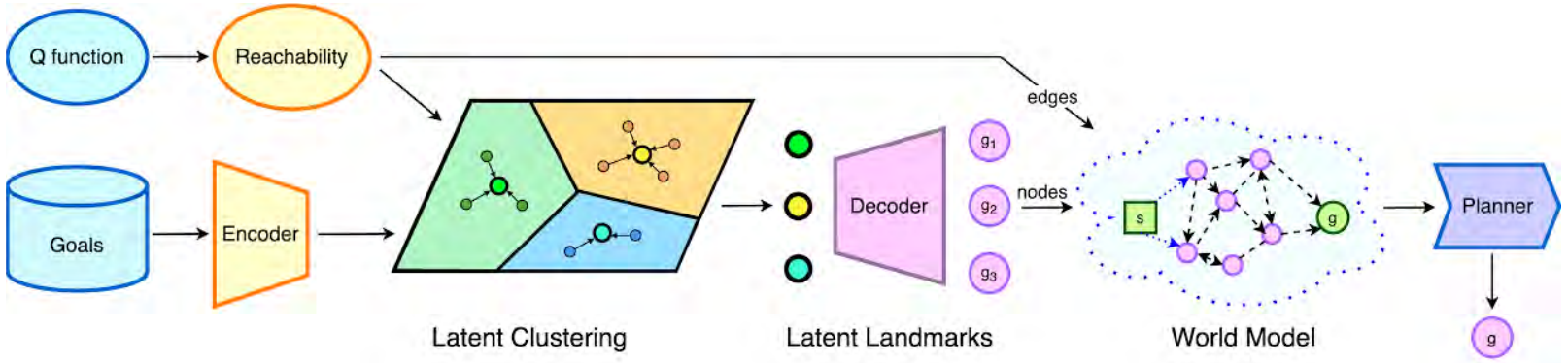
K.Miyazawa et al., Integrated cognitive architecture for robot learning of action and language, Frontiers in Robotics and AI, 2019



D.Ha, J.Schmidhuber, Recurrent World Models Facilitate Policy Evolution, NeurIPS2018



D.Hafner et al., Learning Latent Dynamics for Planning from Pixels, ICML2019



L.Zhang, G.Yang, B.Stadie , World Model as a Graph: Learning Latent Landmarks for Planning, ICML 2021

World Models [Ha+ 2018]

At each time step, our agent receives an **observation** from the environment.

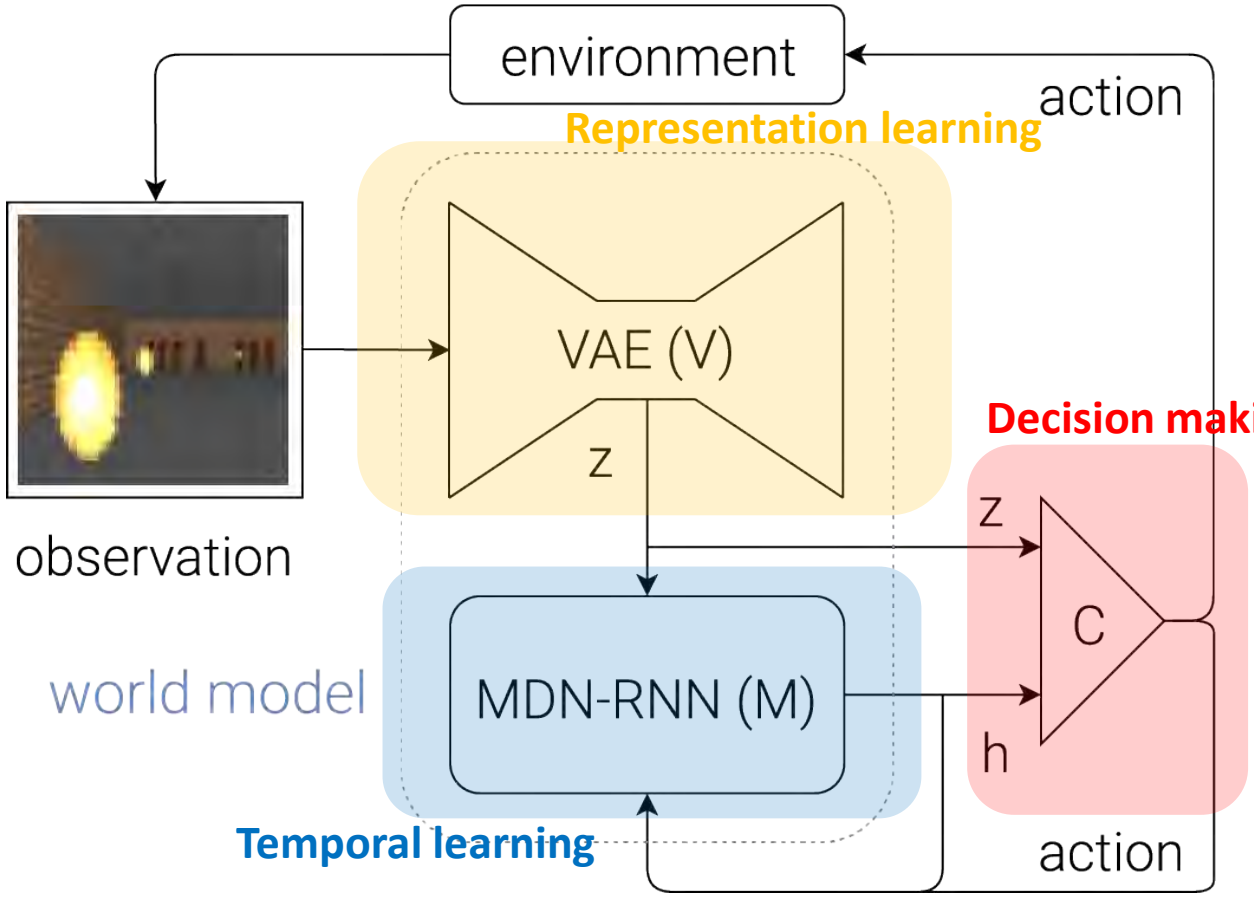
World Model

The **Vision Model (V)** encodes the high-dimensional observation into a low-dimensional latent vector.

The **Memory RNN (M)** integrates the historical codes to create a representation that can predict future states.

A small **Controller (C)** uses the representations from both **V** and **M** to select good actions.

The agent performs **actions** that go back and affect the environment.



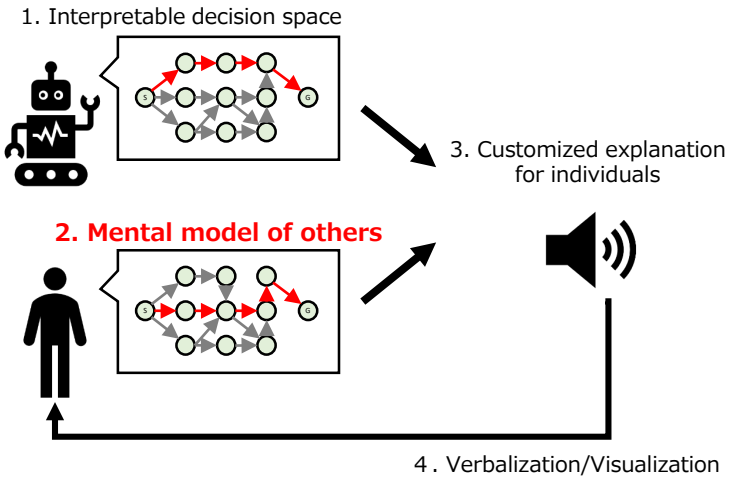
[Ha+ 2018] D.Ha, J.Schmidhuber, Recurrent World Models Facilitate Policy Evolution, NeurIPS2018

XARの要素（その２）

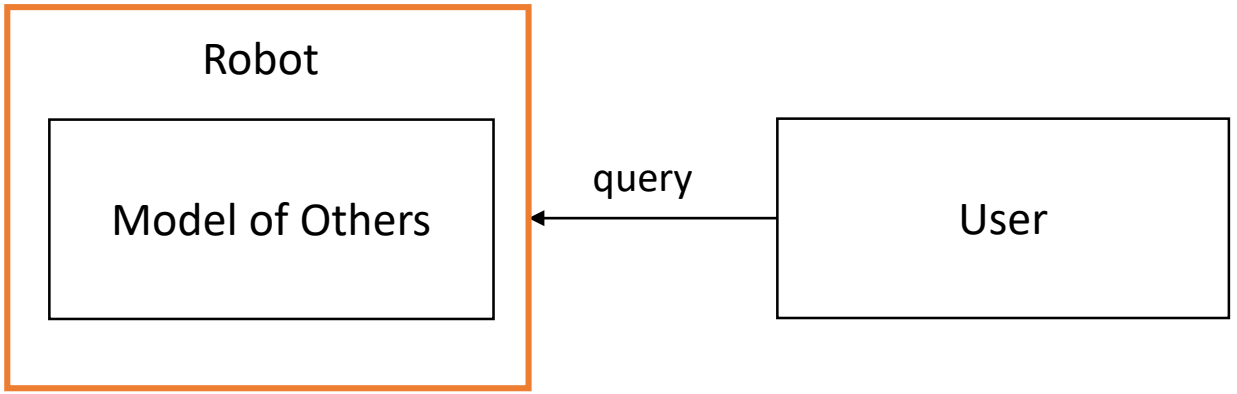
２：他者モデルの推定

部分的な手掛かり(query)から

他者の世界モデルを推定
自身の世界モデルを基盤にする？



- ユーザーに合わせた効果的な説明因子の同定に活用

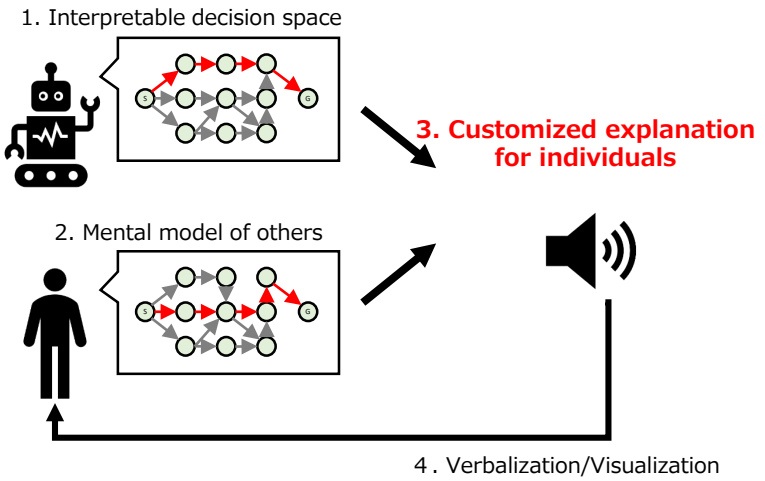


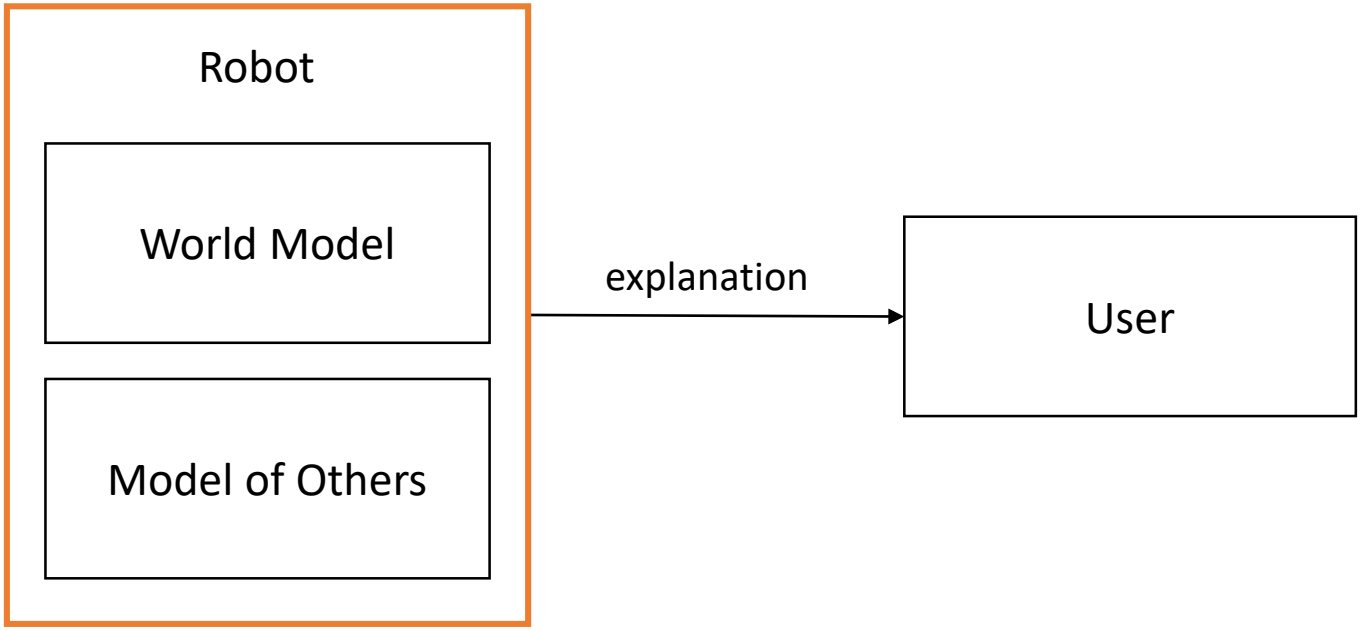
あまり研究がされていない・・・

XARの要素（その３）

３：ユーザーに合わせた説明の生成

ロボットの保持する
世界モデルと他者モデルから
最適な説明を生成

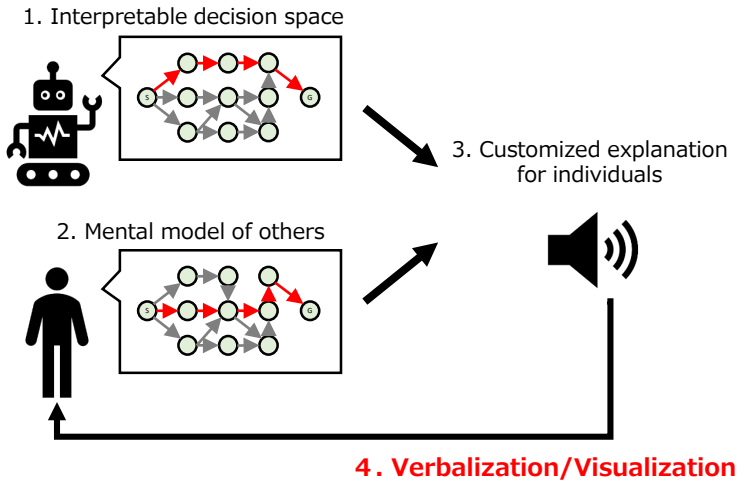


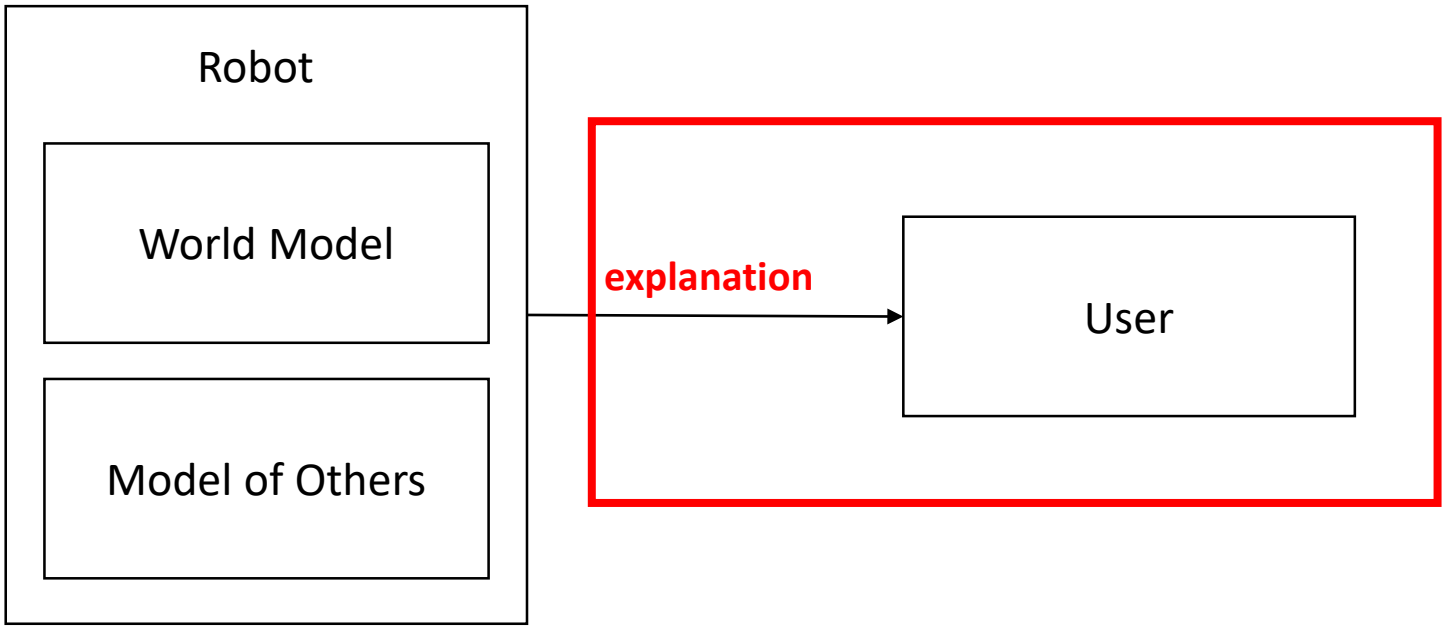


XARの要素（その４）

４：説明の提示

生成した説明を言語化する
他のモダリティも使える





研究事例

Tatsuya Sakai

ADVANCED ROBOTICS
2021, VOL. 35, NO. 17, 1054-1067
https://doi.org/10.1080/01691864.2021.1946423

Taylor & Francis
Taylor & Francis Group

Open access

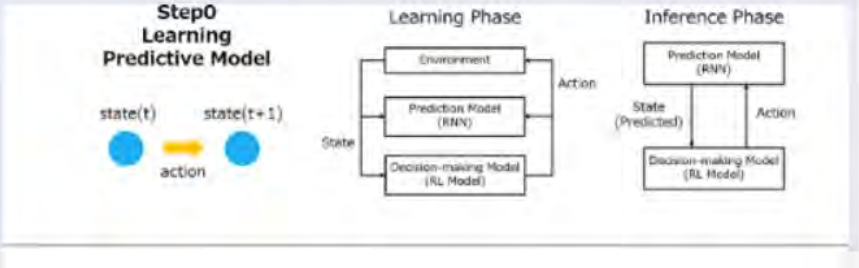
A framework of explanation generation toward reliable autonomous robots

Tatsuya Sakai[a] , Kazuki Miyazawa[a] , Takato Horii[a] , and Takayuki Nagai[a,b]

[a] Graduate School of Engineering Science, Osaka University, Osaka, Japan [b] Artificial Intelligence Exploration Research Center, The University of Electro-Communications, Tokyo, Japan

ABSTRACT

To realize autonomous collaborative robots, it is important to increase the trust that users have in them. Toward this goal, this paper proposes an algorithm that endows an autonomous agent with the ability to explain the transition from the current state to the target state in a Markov decision process (MDP). According to cognitive science, to generate an explanation that is acceptable to humans, it is important to present the minimum information necessary to sufficiently understand an event. To meet this requirement, we propose a framework for identifying important elements in the decision-making process using a prediction model for the world and generating explanations based on these elements. To verify the ability of the proposed method, we conducted an experiment using a grid environment. It was inferred from the result of a simulation experiment that the explanation generated using the proposed method was composed of the minimum elements important for understanding the transition from the current state to the target state. Furthermore, subject experiments showed that the generated explanation was a good summary of the process of state transition, and that a high evaluation was obtained for the explanation of the reason for an action.



T.Sakai, K.Miyazawa, T.Horii, T.Nagai, A framework of explanation generation toward reliable autonomous robots, Advanced Robotics, 35, 17, 1054-1067, 2021

日本ロボット学会誌 Vol. xx No. xx, pp.1～4, 201x

レター

Graph2vecを用いた世界モデルの分散表現獲得と他者世界モデルの推定

境 辰也*1 堀井隆斗*1 長井隆行*1*2

Representation Learning of World Models and Estimation of World Model of Others Using Graph2vec

Tatsuya Sakai*1, Takato Horii*1 *and* Takayuki Nagai*1*2

To realize advanced interaction between autonomous robots and users, it is important for robots to aware the difference in their state space representations (i.e., world models). As a first step toward this goal, we propose a method to estimate user's world model based on queries. In our method, the agent learns distributed representation of world models by graph2vec and generates concept activation vectors (CAVs) that represent the meaning of queries in latent space. The experimental results show that our method can estimate user's world model more efficiently than the simple method using "AND" search of queries.

Key Words: Autonomous robot, Explainability, Representation learning

1. は じ め に

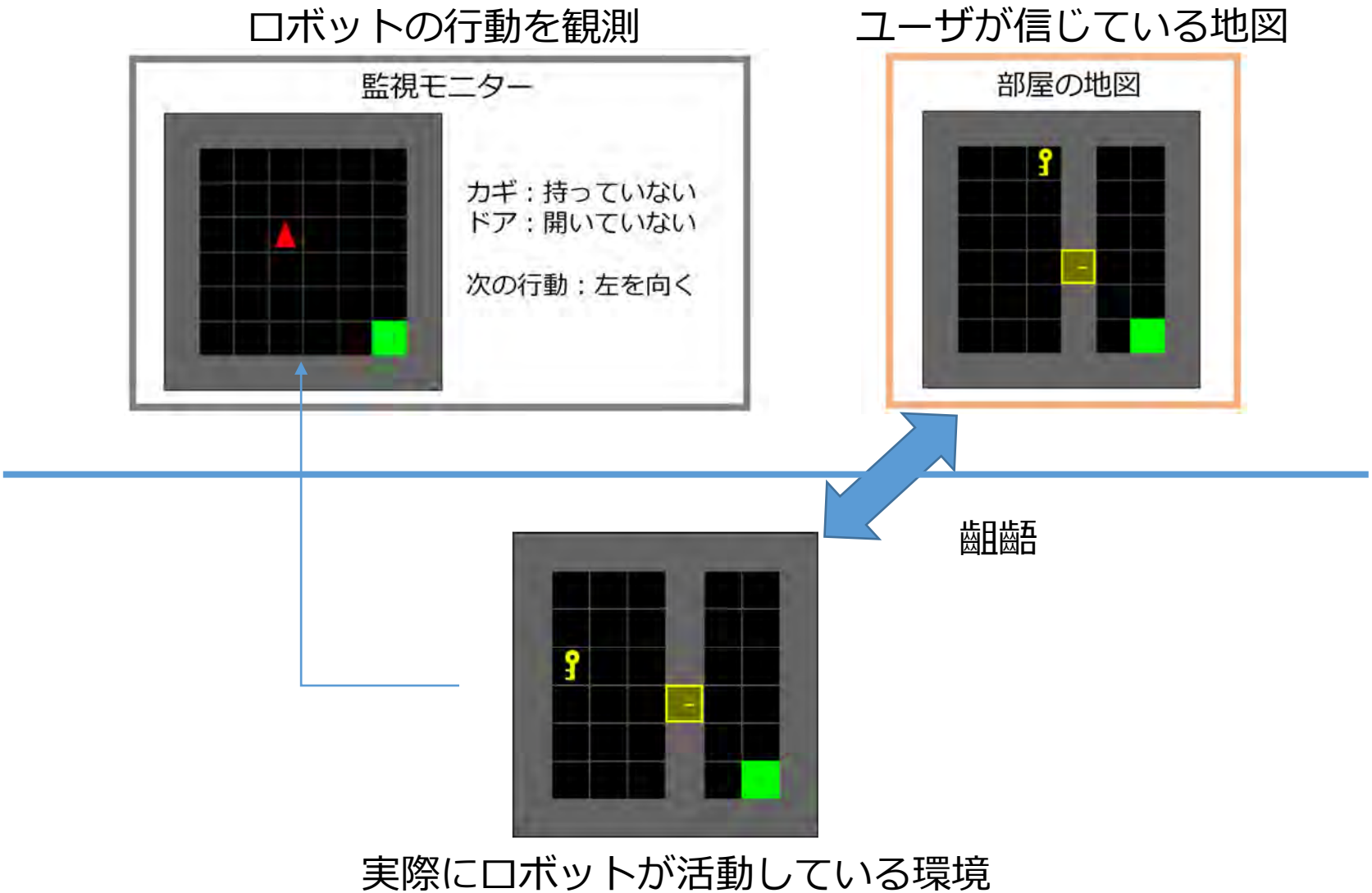
自律ロボットの応用が進んでいる．現状の自律ロボットは，与えられた命令を忠実に実行することで人間のタスク遂行を補助するツールである．一方で，より高度な意思決定をする自律ロボットの場合，命令を忠実に実行することが必ずしも最善の方策であるとは限らない．このような自律ロボットがユーザーにユーザーに応じた説明をするために重要である．人－ロボットインタラクションの文脈において，ユーザーの内部状態を推定することの重要性は既に認識されており，Gao ら [3] や Clair ら [4] はユーザーの取った行動やインタラクションの履歴から尤もらしい行動方策を推定する枠組みを提案した．また，Huang ら [5] は，説明を方策に復元する過程に注目し，説明を受け取るユーザーが持つ復元アルゴリズムを適切に推定することの重

境, 堀井, 長井, Graph2vecを用いた世界モデルの分散表現獲得と他者世界モデルの推定, 日本ロボット学会誌, to appear

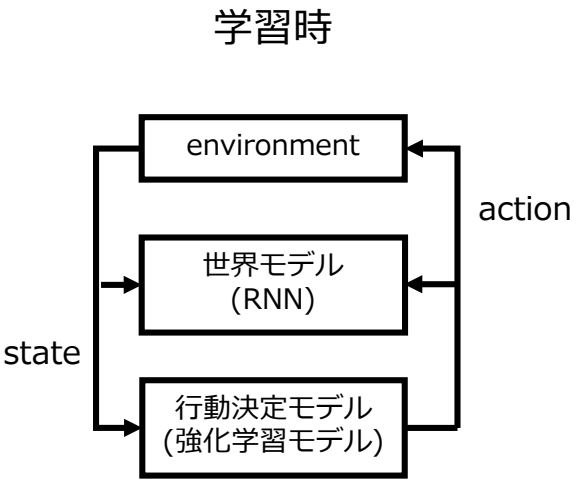
研究事例

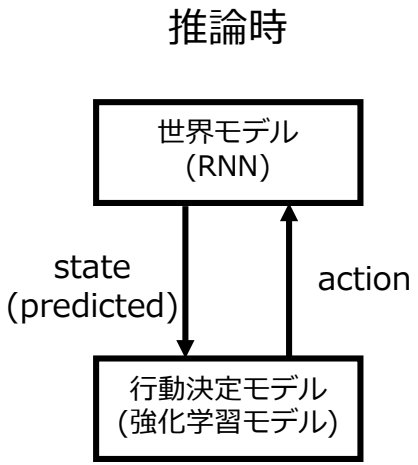
（シナリオ）

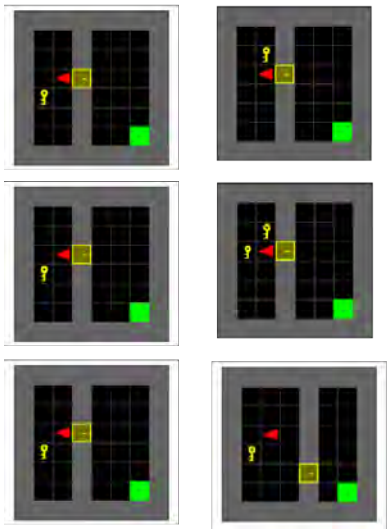
- ユーザの信念と実際の環境に齟齬がある
- ユーザがロボットの行動を理解できない場合がある
- その際にユーザがロボットに質問
- ロボットは**説明する**必要がある



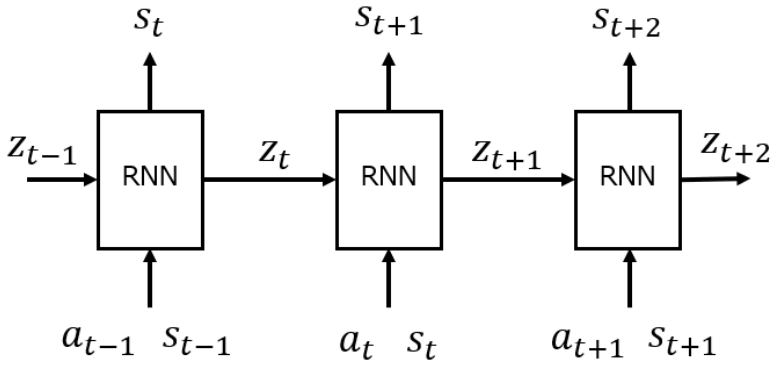
Step１：世界モデルの学習
実環境での行動を伴わず、状態遷移の推論を可能にする



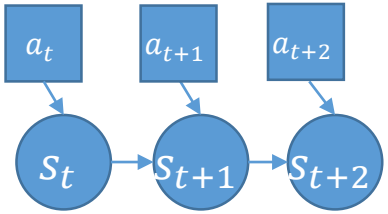




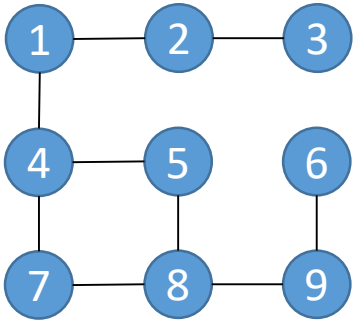
色々な環境のモデルを学習する



RNN



MDP



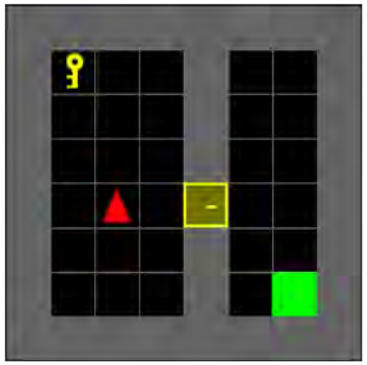
Graph

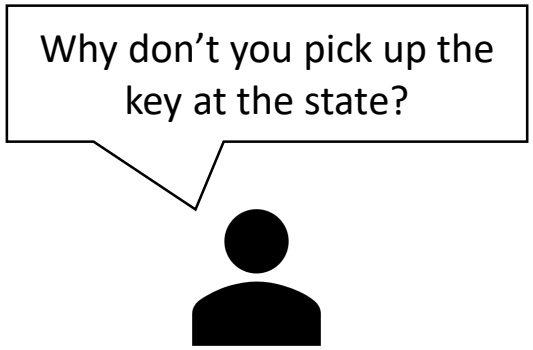
Step２：他者モデルの推定
ユーザーの持つ世界モデルを推定する
(ユーザーがどのように環境を認識しているかを推定する)

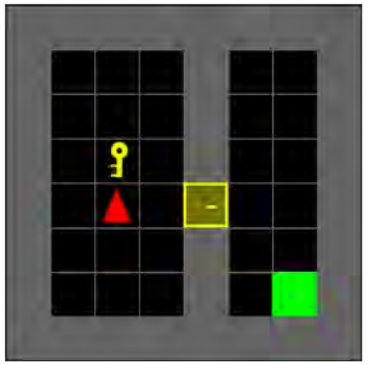
エージェントの
世界モデル

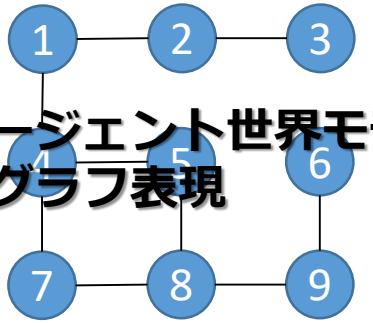
ユーザーからの
クエリ(質問)

ユーザーの世界モデル
(推定値)



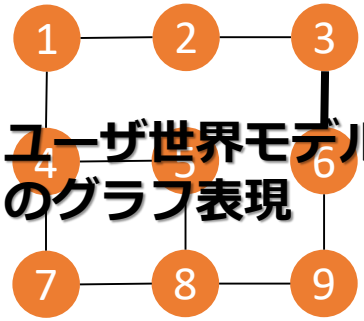






エージェント世界モデル
のグラフ表現

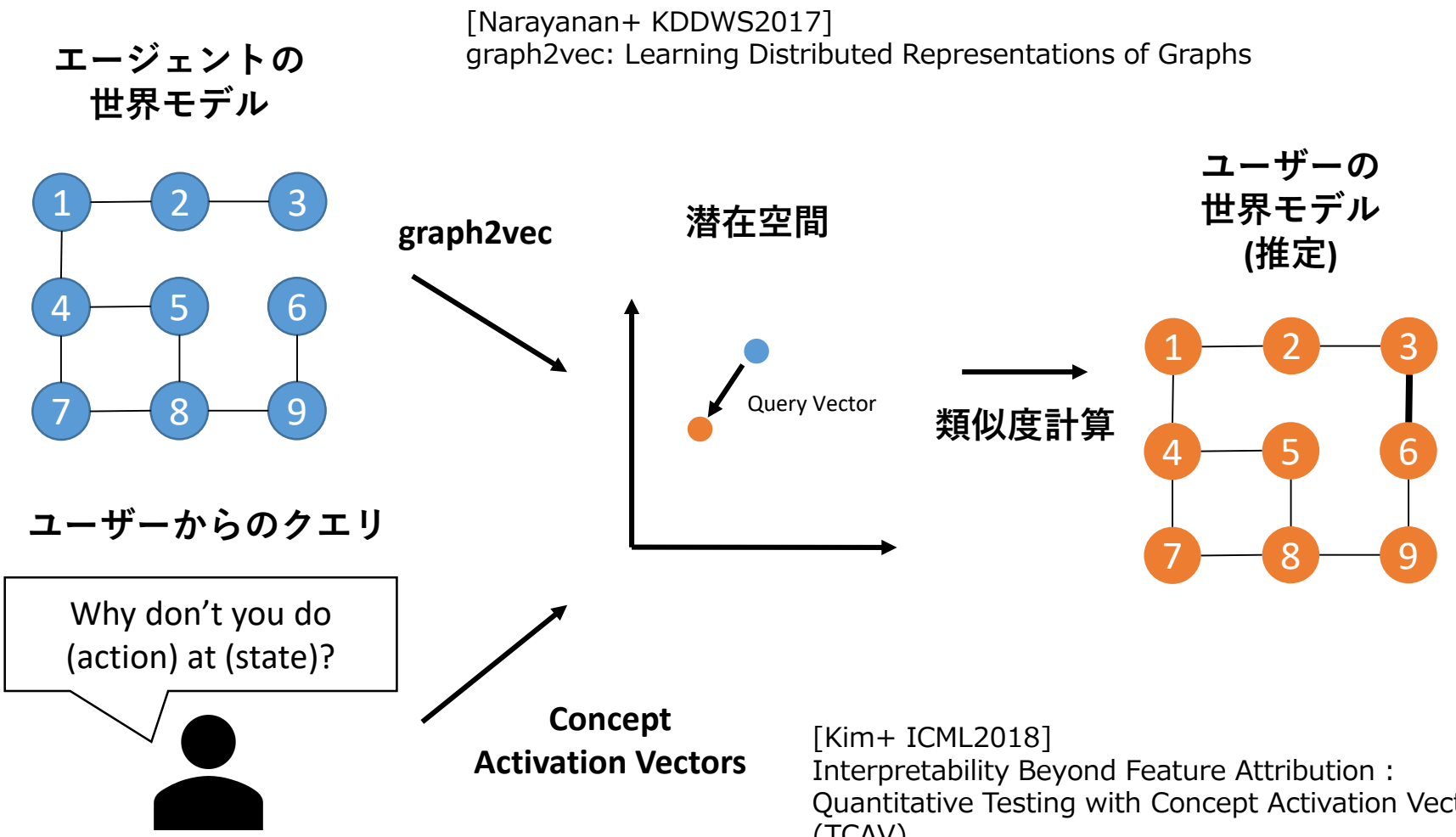
クエリ



ユーザ世界モデル
のグラフ表現

エージェントの持つ世界モデルとユーザーの与えたクエリ(質問)からユーザーの持つ世界モデルを推定

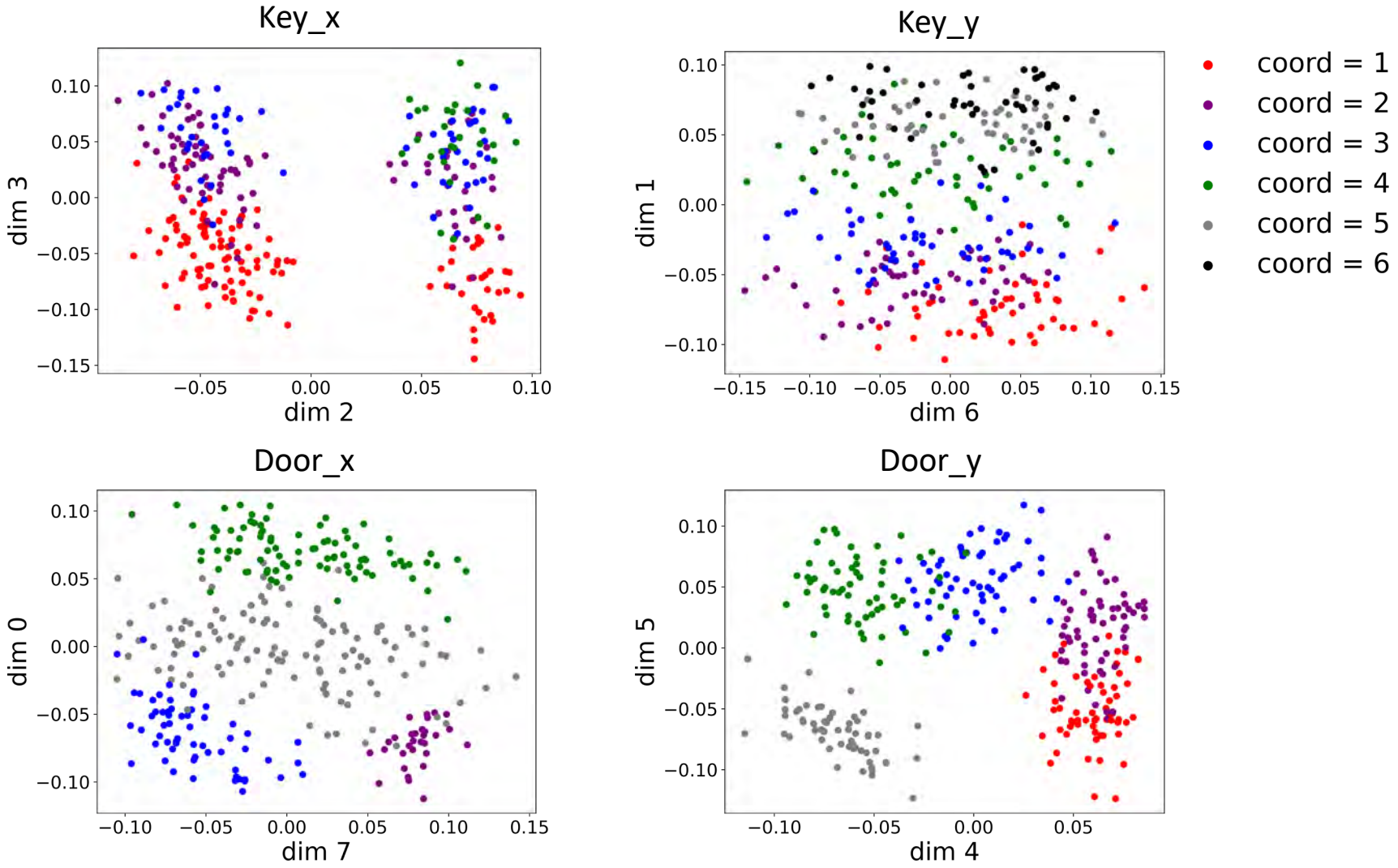
[Narayanan+ KDDWS2017]
graph2vec: Learning Distributed Representations of Graphs



[Kim+ ICML2018]
Interpretability Beyond Feature Attribution :
Quantitative Testing with Concept Activation Vector
(TCAV)

結果

鍵,ドアの座標それぞれに対しクラスタが形成された

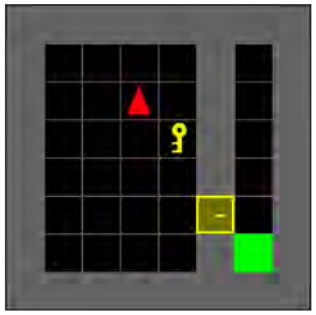


クエリ数 $j = 1$, 16次元の分散表現をICAにより8次元に圧縮

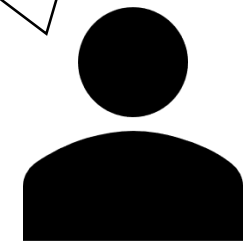
結果

クエリを満たす最低限の変更が加えられた環境が推定された

エージェントの世界モデル

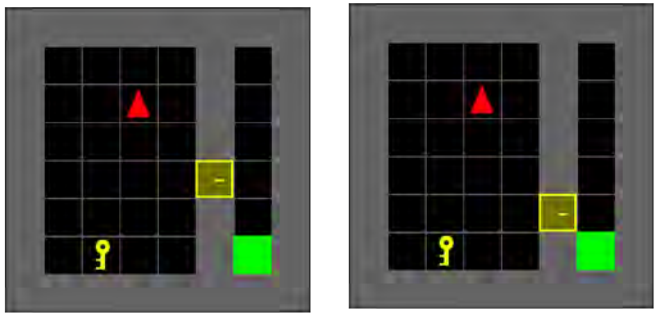


Why don’t you pick up the key at the state?



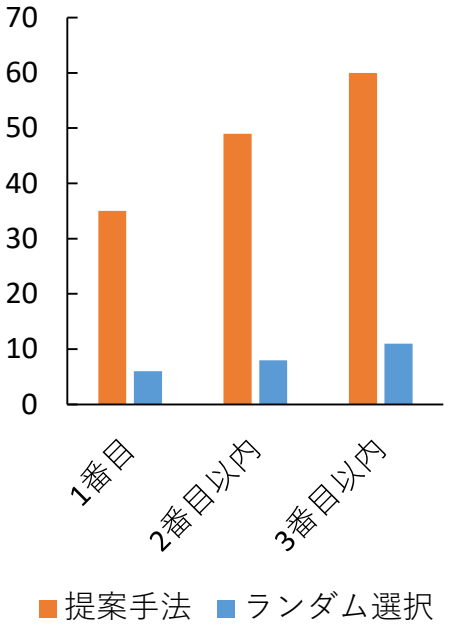
高いスコアを得た世界モデル

低いスコアを得た世界モデル



クエリ数 $j = 1$

最適環境の出現順



Step３：説明の生成

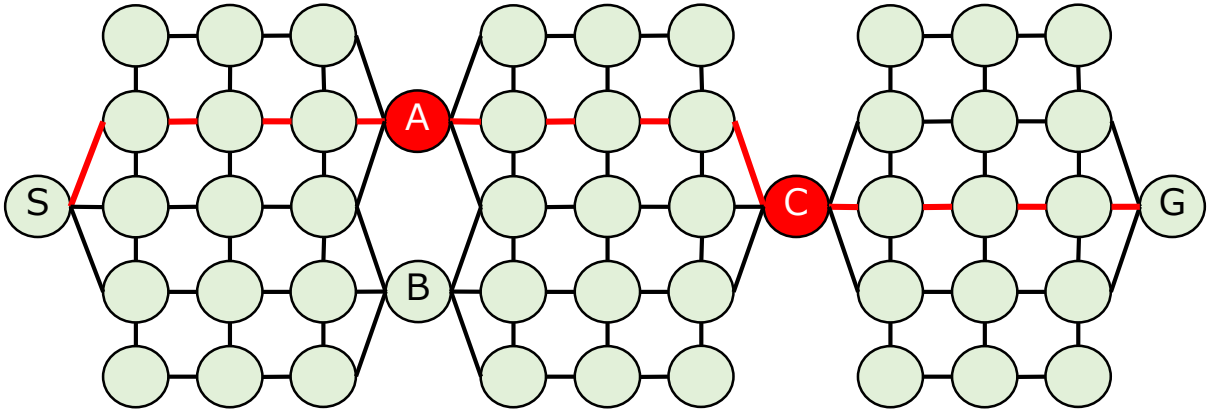
説明のための重要要素の同定

目標到達のために通過すべき状態行動対を抽出する

ある最適経路$(\boldsymbol{s}, \boldsymbol{a})$において
目標状態到達に重要な状態行動対(s, a)を抽出したい

▼

迂回経路の少なさ＝状態行動対(s_f, a_{opt})の重要度



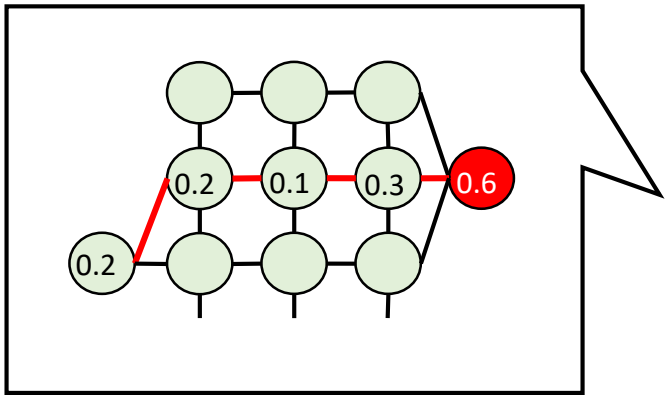
説明のための重要要素の同定
目標到達のために通過すべき状態行動対を抽出する

(s_f, a_{opt})の重要度

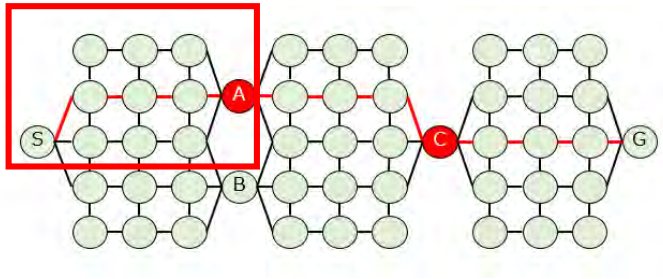
注目状態s_fからランダム探索後、
方策に従い行動決定したとき再び注目状態s_fに戻ってくる確率値

▼

世界モデル上での**反復計算**により、
重要度の高い要素を**サブゴール**として抽出

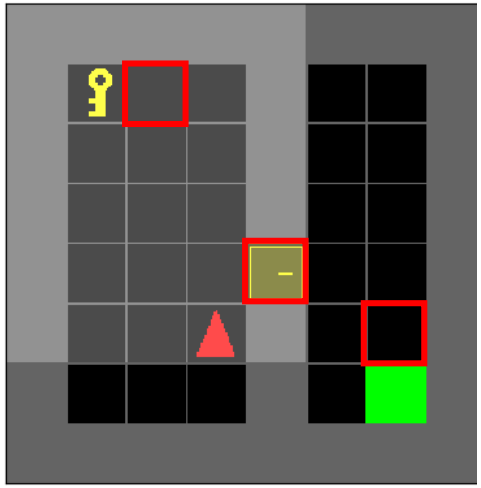


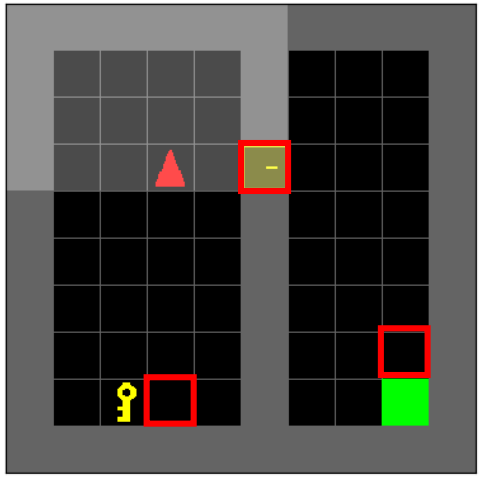
因果推論として定式化できる！

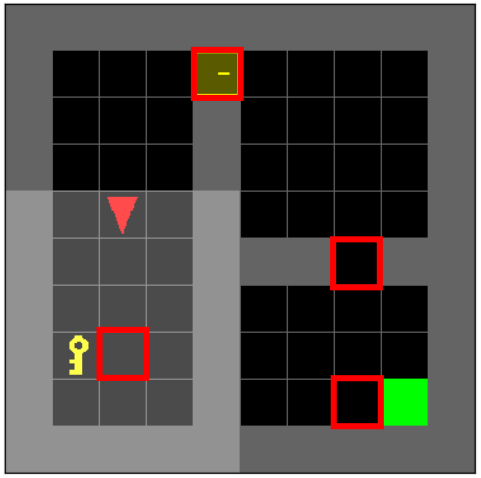


$$\mathrm{ACE} = P(Goal = 1|do(allow = 1)) - P(Goal = 1|do(allow = 0))$$

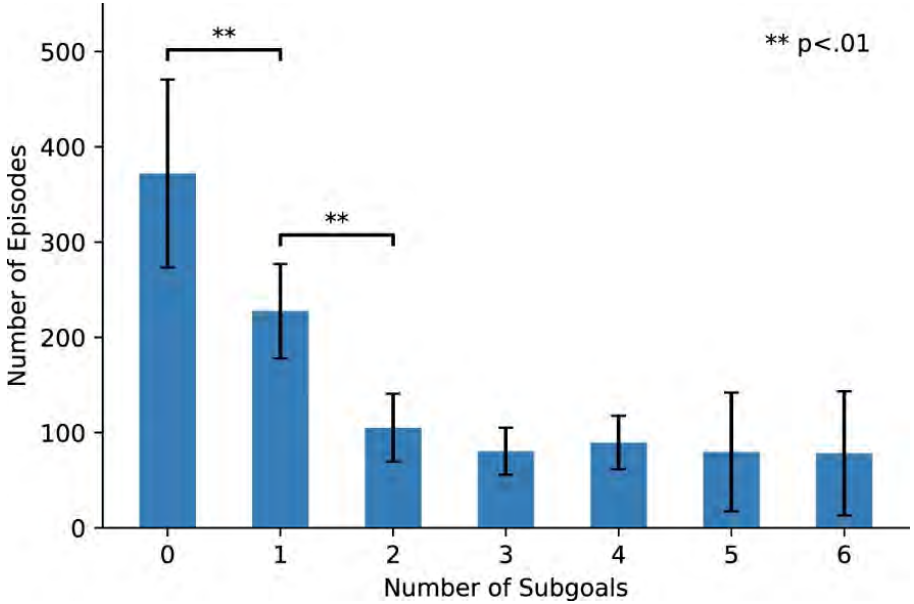
結果





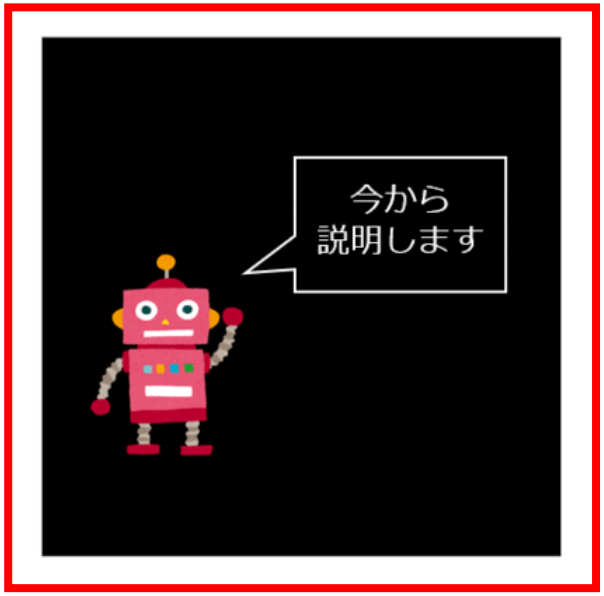


重要要素に報酬を与えて強化学習した時にどれくらい少ないエピソードで学習できるのか？
　⇒コーチングしているイメージ

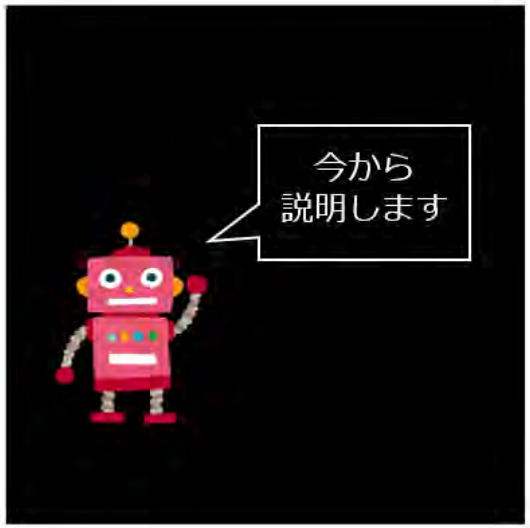


主観評価実験

ランダム要素提示

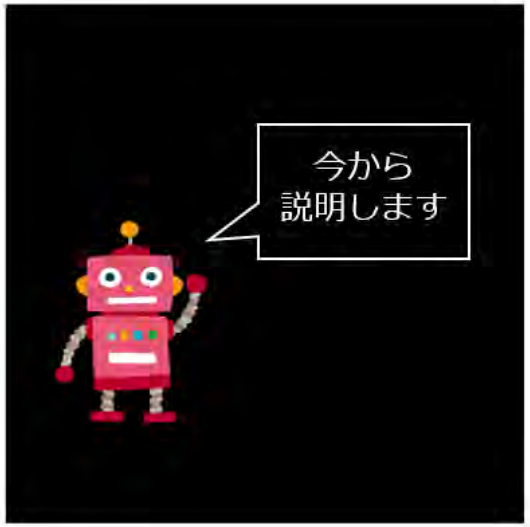


サブゴール提示

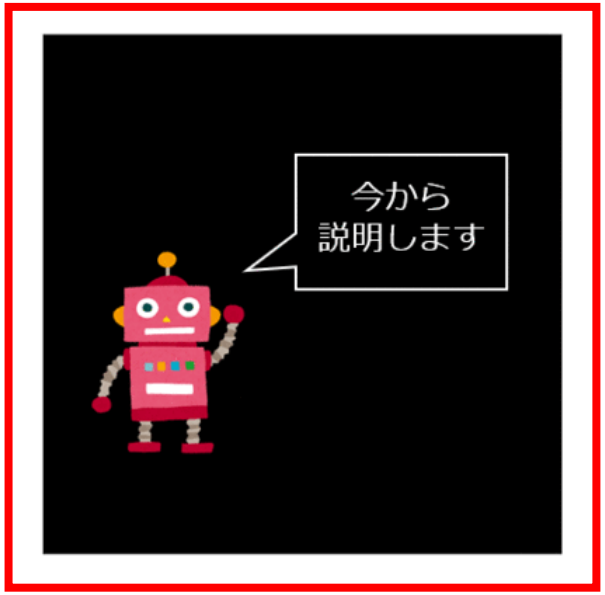


主観評価実験

ランダム要素提示



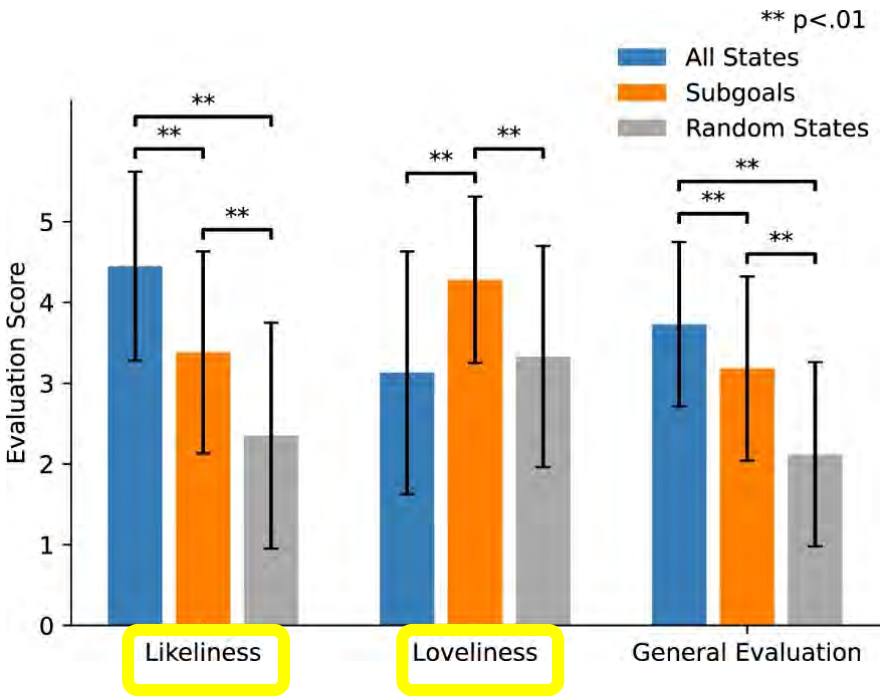
サブゴール提示



主観評価実験結果

説明の評価 [下條+,2019]

- **説明の正しさ(Likeliness)**
  できるだけ詳しく、正確な説明が良い
  (事象Xの事後生起確率$P(X|E_i)$が高い説明E_iが良い)
- **説明の美しさ(Loveliness)**
  できるだけシンプルな説明が良い
  (原因の数が少なく、かつ提示された原因が実際に観測できるような説明が良い)



LikelinessとGeneral Evaluationに交互作用

- 説明に対する人間の受容はその正しさと美しさで決まる
- 特に他者から説明を受ける際は美しさ(Loveliness)に大きく依存して受容が決まる[Douven+,2018]

[下條+,2019] 評価状況が因果的説明の選好に与える影響についての実験的検討
[Douven+,2018] Best, Second-Best, and Good-Enough Explanations: How They Matter to Reasoning

Step４：説明の提示
ユーザに説明を伝える（言語化）

議論

- 説明内容の言語化
  - VQAやGPTなどの手法で可能か？
  - 他のモダリティを使ってもよい
- 全体の構造
  - 相手の発話を理解
  - 他者モデルの再推定
  - 説明の生成
- この繰り返しが対話システム？
  - パイプラインではない全体的なシステム
  - 学習も含む（強化学習に発話するという行為も含める？）

XARにおける研究課題

- 実世界の問題への適用
  - XAIのような「実世界で使える」技術への昇華
- インタラクションへの活用
  - 説明内容だけでなく情報提示の頃合いや方法の自律的決定
- 説明の社会的側面の考慮
  - 説明がロボットと人間の関係にどのように影響するのか？
- 嘘の説明生成と倫理
  - 人間が許容できる嘘の範囲
  - ロボットが嘘をつくことの是非

ロボット学習（言語を含む世界モデル）
という文脈でこれらの問題を解きたい！

対話研究との接合？

acknowledgement

I would like to thank

- CREST project members
  - Prof.Taniguchi, Prof.Ogata, Prof.Sugiura, Prof.Iwahashi, Prof.Inamura, Prof.Okada
- NEDO project members
  - Prof.Nakamura, Dr.Abe, Dr.Kasuya, Ms.Oku, Ms.Yamauchi
- Kakenhi project members
  - Prof.Ohira, Dr.Hieida, Prof.Mochihashi, Prof.Kobayashi
- Lab members
  - All staff and students!

- JST, NEDO, MEXT
- Toyota, Panasonic, Kawasaki Heavy Industry, ChiCaRo

for their support!