A Survey of Machine Theory of Mind

2022/11/09 毛媛媛

心智理论

心智理论（Theory of Mind）是指个体有能力理解他人的心理状态，这些状态包括意图、假装、愿望、信念与情绪等。

儿童心智理论发展

2岁
- 愿望心理学：人们将以满足愿望的方式采取行动

3岁
- 愿望-信念心理学：儿童开始认识信念是心理表征，能根据信念预测或解释行为，但是当信念与愿望相冲突时，则根据愿望而非信念预测或解释行为。

4-5岁
- 信念-愿望心理学：意识到信念有不同的表征（不同的人有不同的信念），对于行为的预测和解释中，予以信念重要多以愿望。

愿望：指向某一客体活或状态的内在体验和倾向
信念：关于某种事实的确信

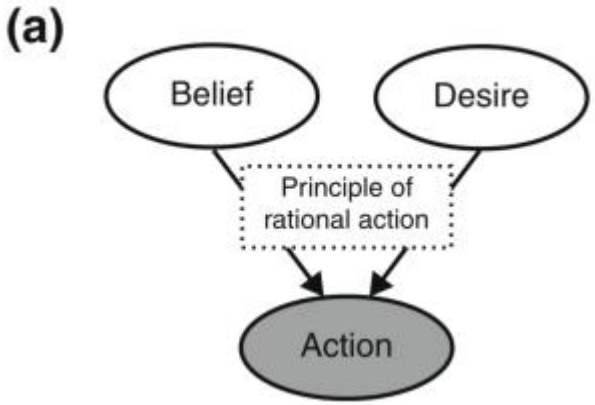
机器心智理论

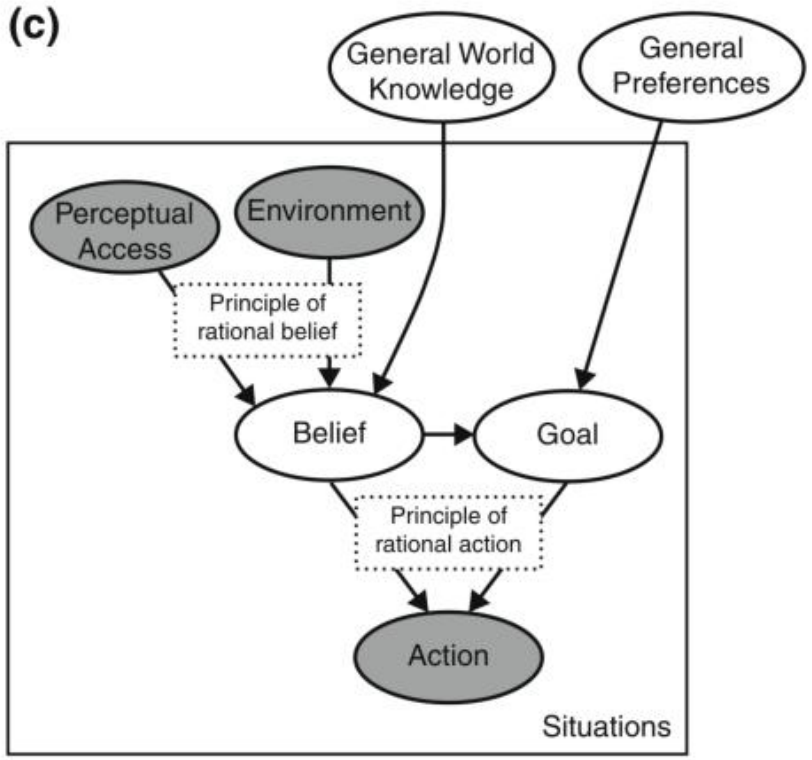
为什么机器需要有心智理论?

- 对人类的行为进行更深度的分析，促进人机之间知心（相互理解）和有温度（同理心）的互动。
- 掌握人类强大的认知能力，可以拥有像人类一样通过小样本学习、迁移学习、举一反三等能力。

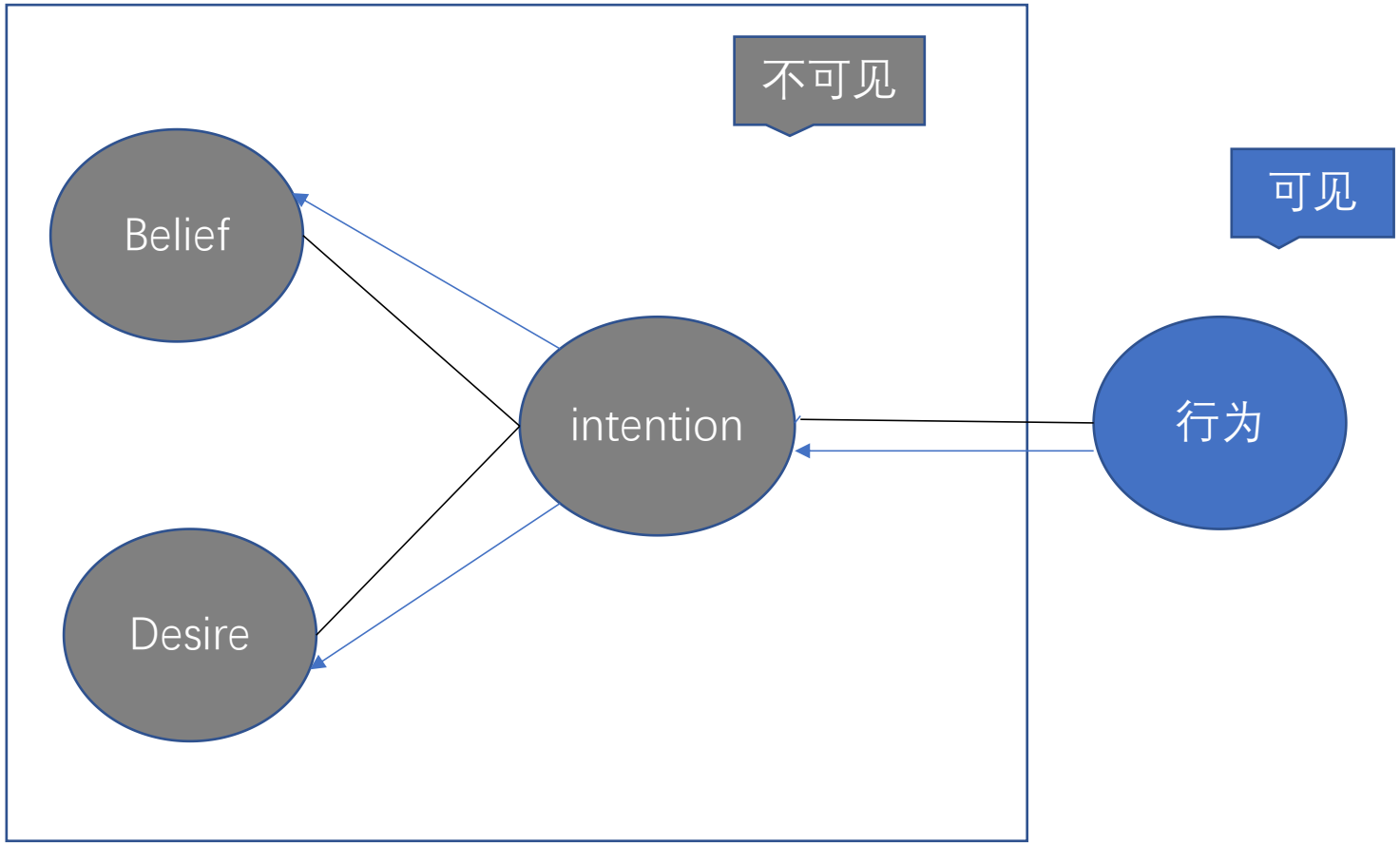
例子：老人想要上楼，但是面前有电梯和楼梯，一个有温度的机器人倾向于他要去做的是电梯而不是楼梯

机器心智理论


(a)
Belief
Desire
Principle of rational action
Action

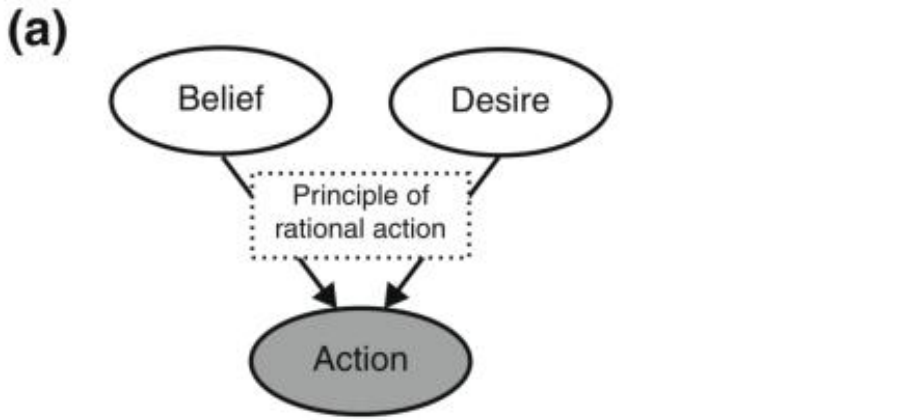

(c)
General World Knowledge
General Preferences
Perceptual Access
Environment
Principle of rational belief
Belief
Goal
Principle of rational action
Action
Situations

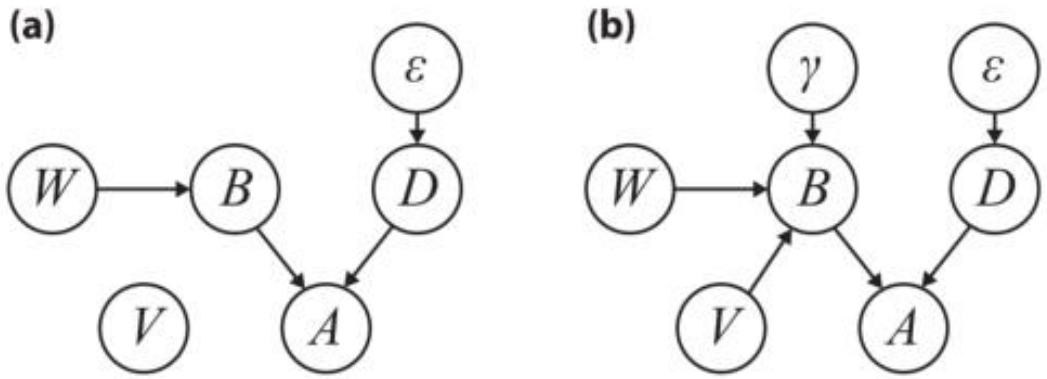
信念-愿望-意图理论架构



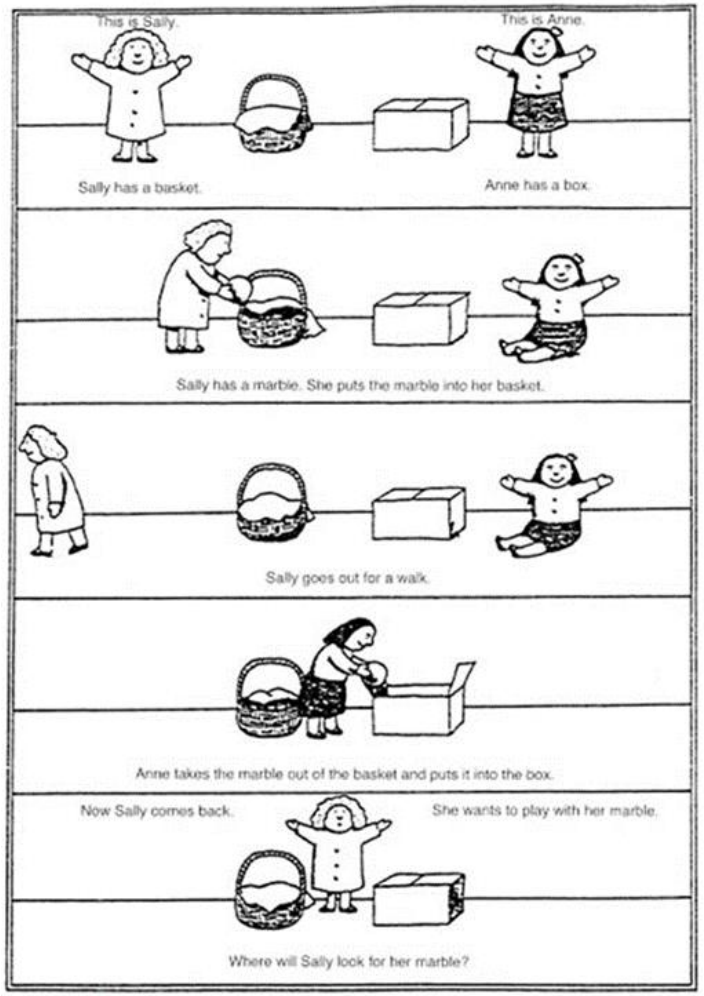
Tasks

行为解释





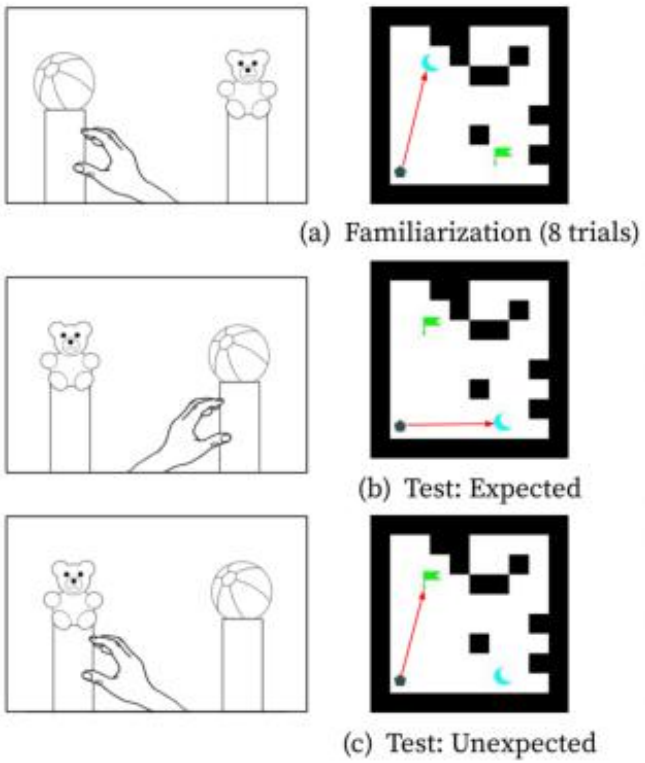
错误信念测试



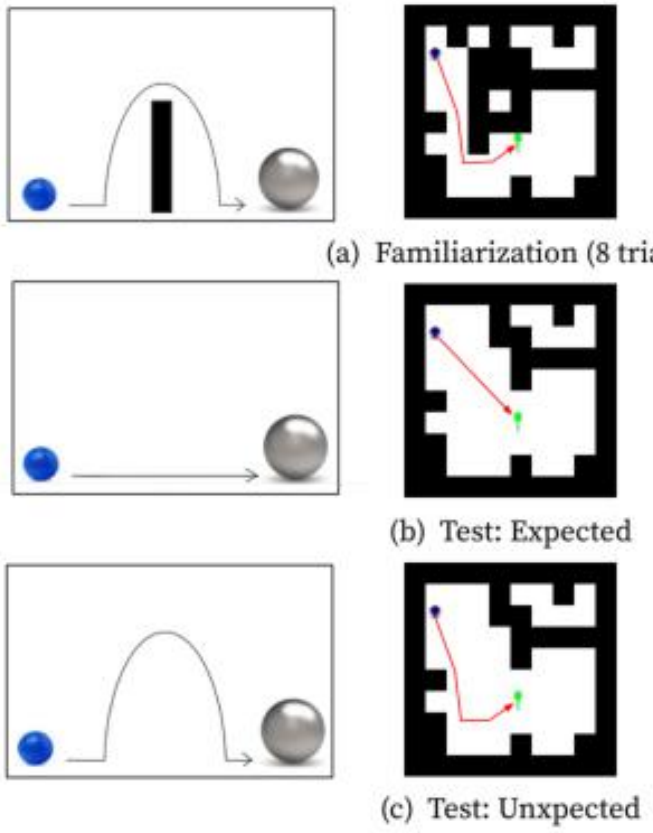
Intuitive Theories of Mind: A Rational Approach to False Belief

Tasks

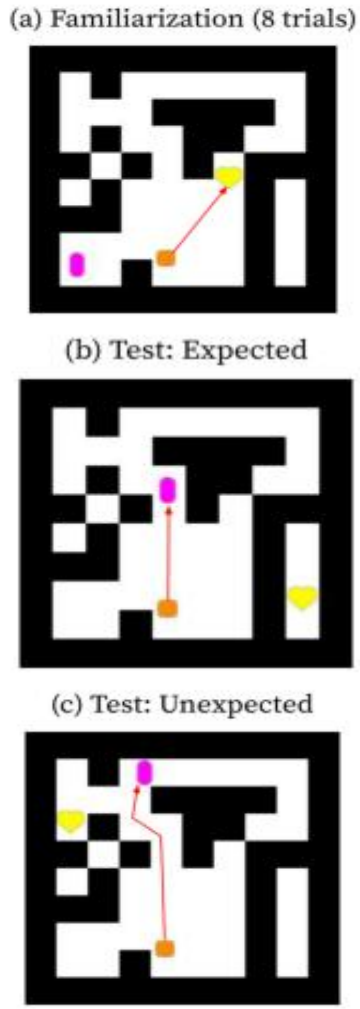
行为预测



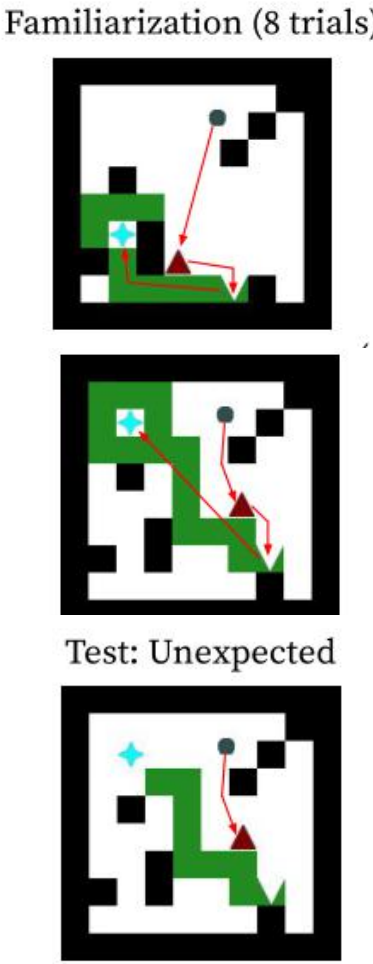
preferences



efficient action



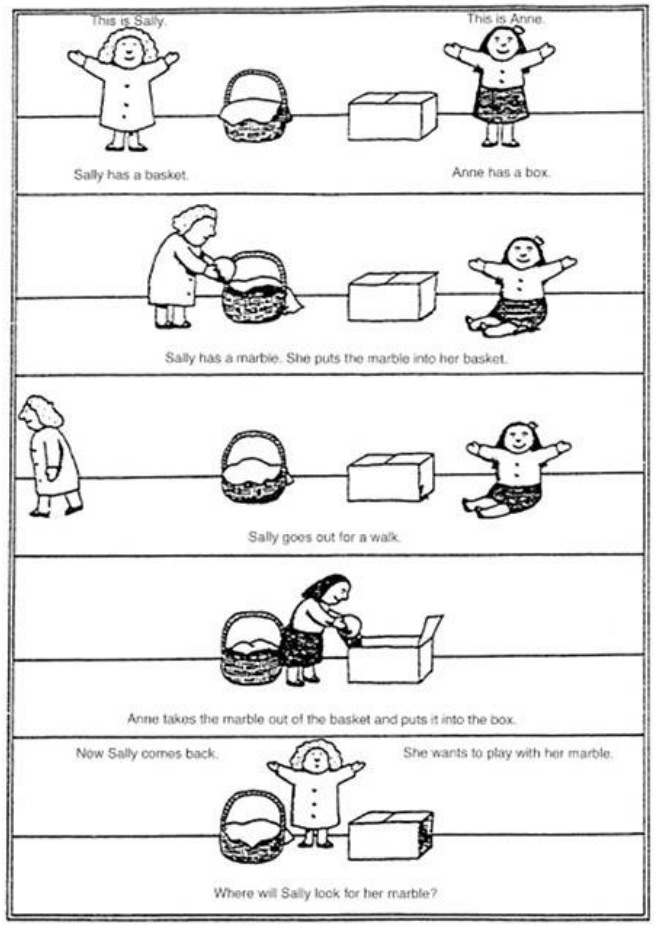
obstacles restrict

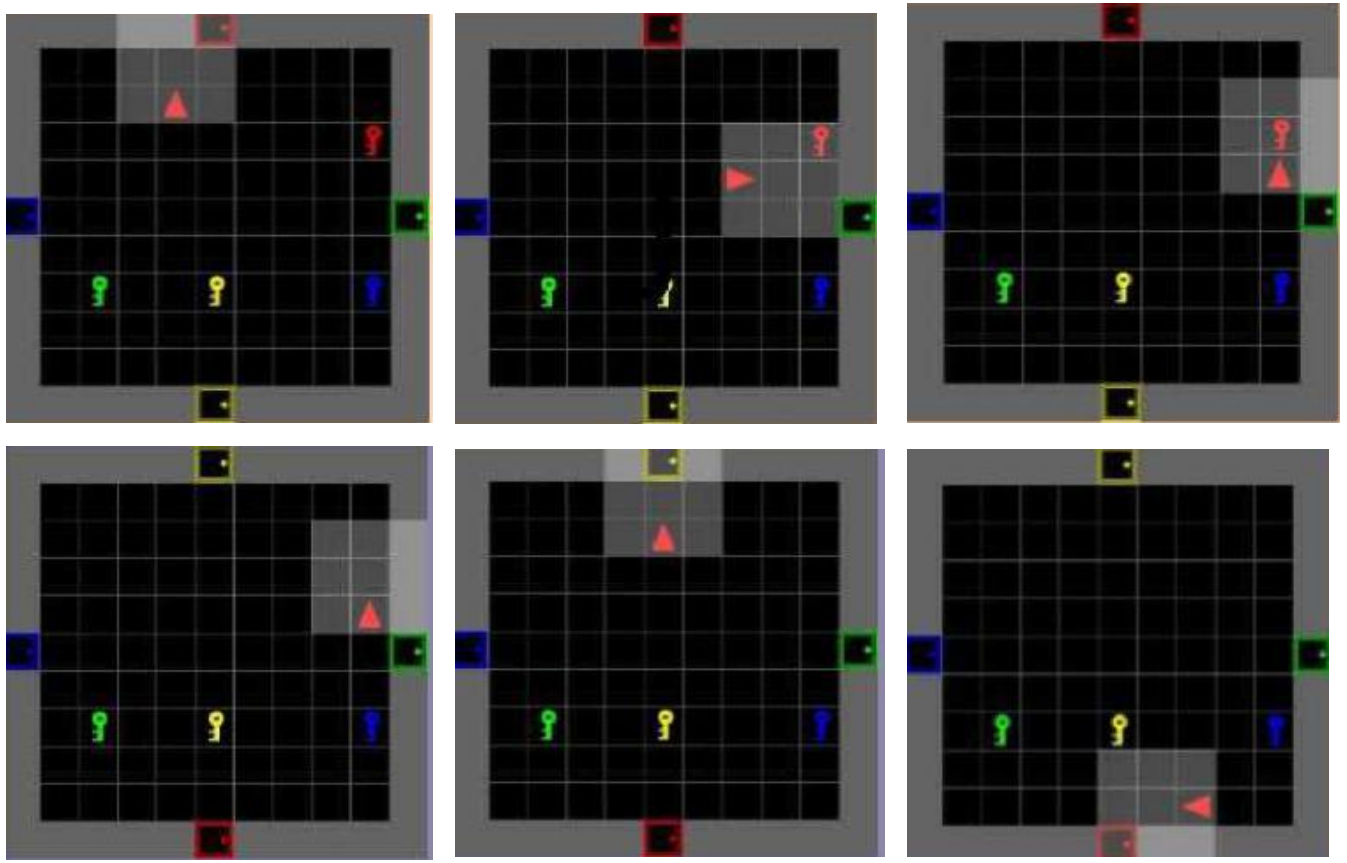


high-order goal

Experiment

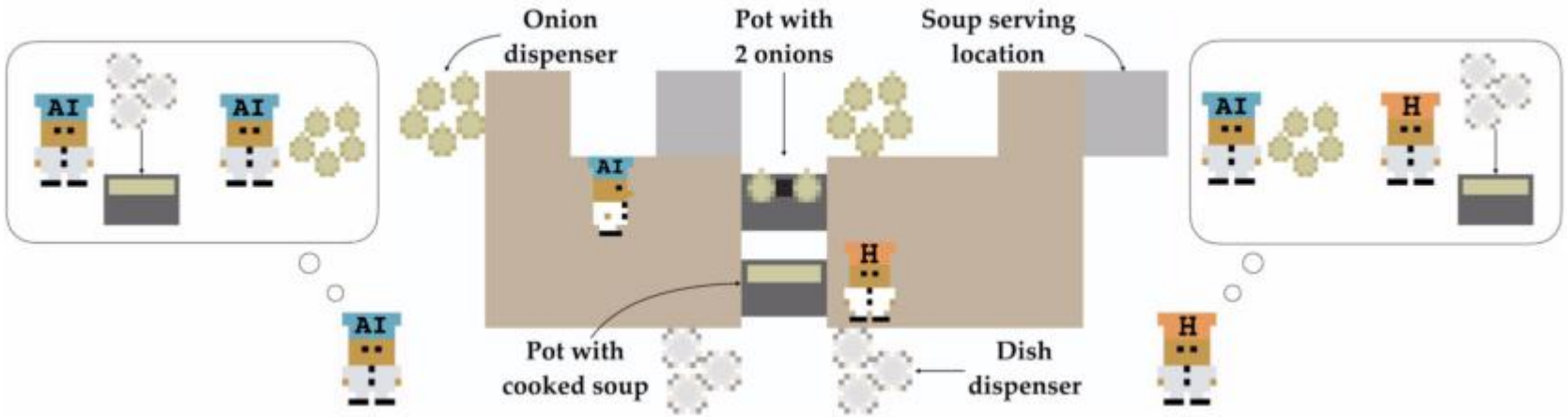
基于错误信念的行为预测

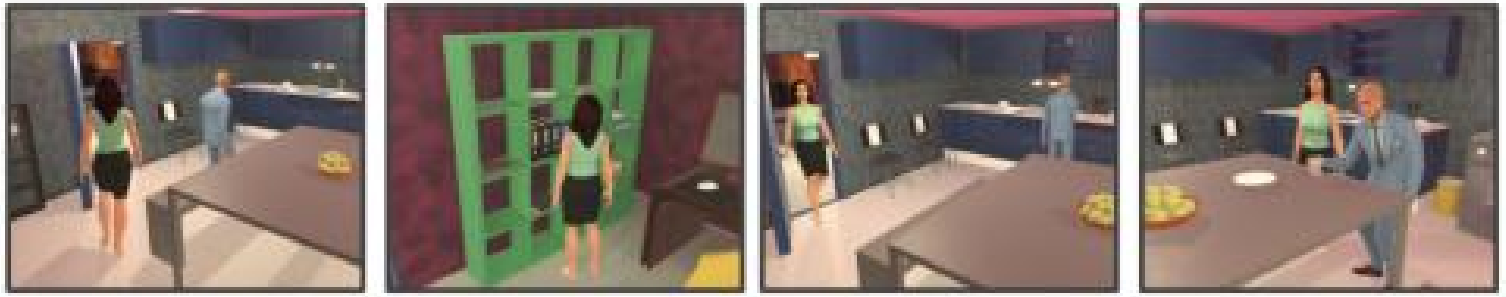




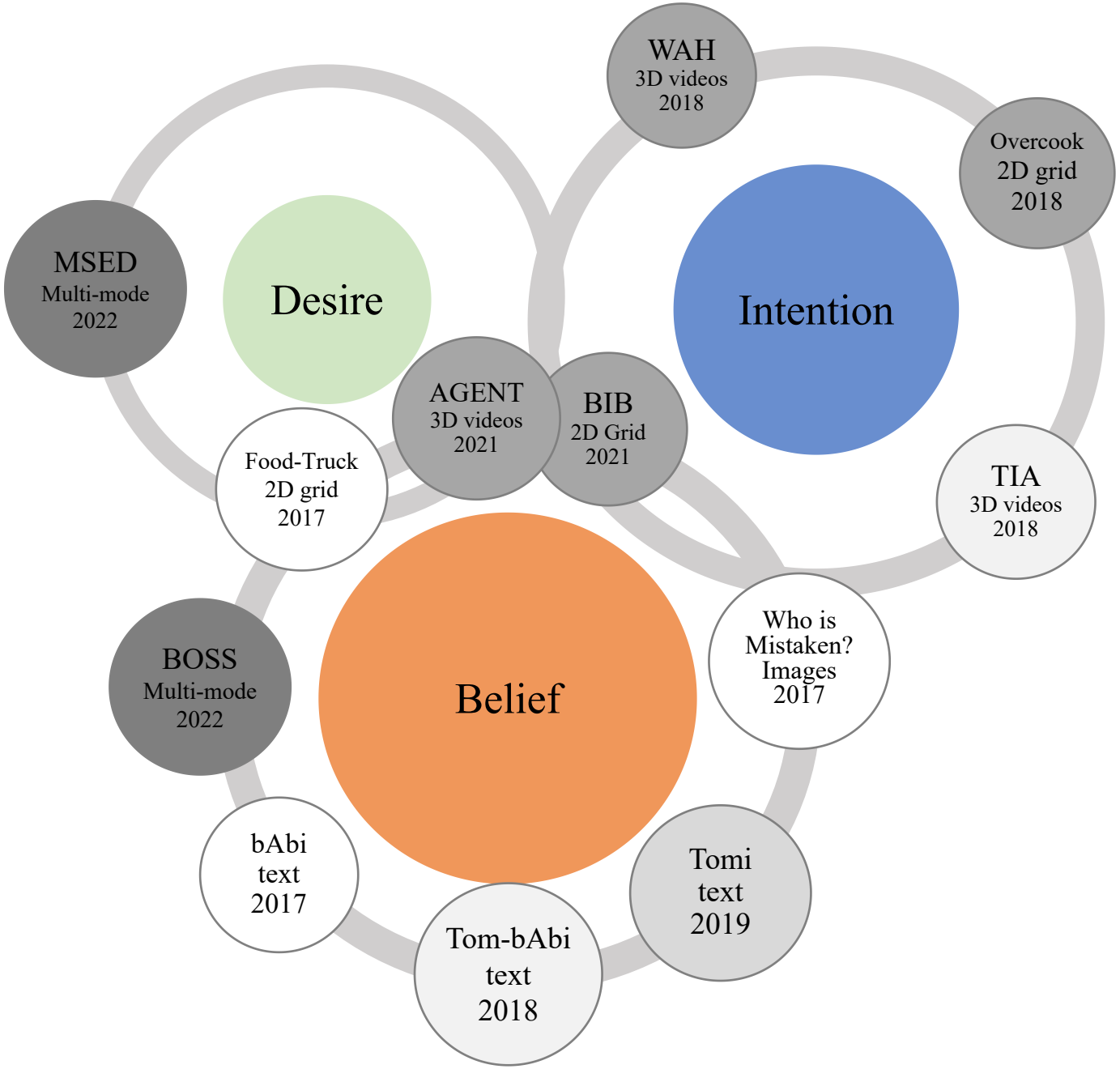
Task

人机交互实验


Onion dispenser
Pot with 2 onions
Soup serving location
AI
AI
AI
AI
H
H
Pot with cooked soup
Dish dispenser



Dataset


WAH
3D videos
2018
Overcook
2D grid
2018
MSED
Multi-mode
2022
Desire
Intention
AGENT
3D videos
2021
BIB
2D Grid
2021
Food-Truck
2D grid
2017
TIA
3D videos
2018
Who is
Mistaken?
Images
2017
BOSS
Multi-mode
2022
Belief
bAbi
text
2017
Tomi
text
2019
Tom-bAbi
text
2018

Dataset

基于QA的信念推理数据集
利用模板生成数据
二阶错误信念推理

True Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Anne moved the milk to the pantry.

False Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.

Second-order False Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.
Anne exited the kitchen.
Sally entered the kitchen.

Figure 2: An example story from each of the three task types.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Anne think that Sally searches for the milk?

ToM-bAbi dataset

Three types of stories

1 ⟨A⟩ entered ⟨L⟩
2 ⟨B⟩ entered ⟨L⟩
3 Phone rang. *// Distractor can appear anywhere*
4 The ⟨O⟩ is in ⟨C1⟩.
5 ⟨B⟩ exited ⟨L⟩ *// if story type 1 or 2*
6 ⟨A⟩ moved the ⟨O⟩ to ⟨C2⟩.
7 ⟨A⟩ exited the ⟨L⟩ *// if story type 2*
8 ⟨B⟩ entered the ⟨L⟩ *// if story type 2*

Example story

1 Isla entered the bathroom.
2 Benjamin entered the bathroom.
3 The cabbage is in the green_pantry.
4 Phone rang.
5 Isla moved the cabbage to the red_drawer.

Answers for each story-question pair

	Story 1	Story 2	Story 3
First Order	C1	C2	C2
Second Order	C1	C1	C2
Memory	C1	C1	C1
Reality	C2	C2	C2

ToMi dataset

Examples of stories from the ToMi dataset

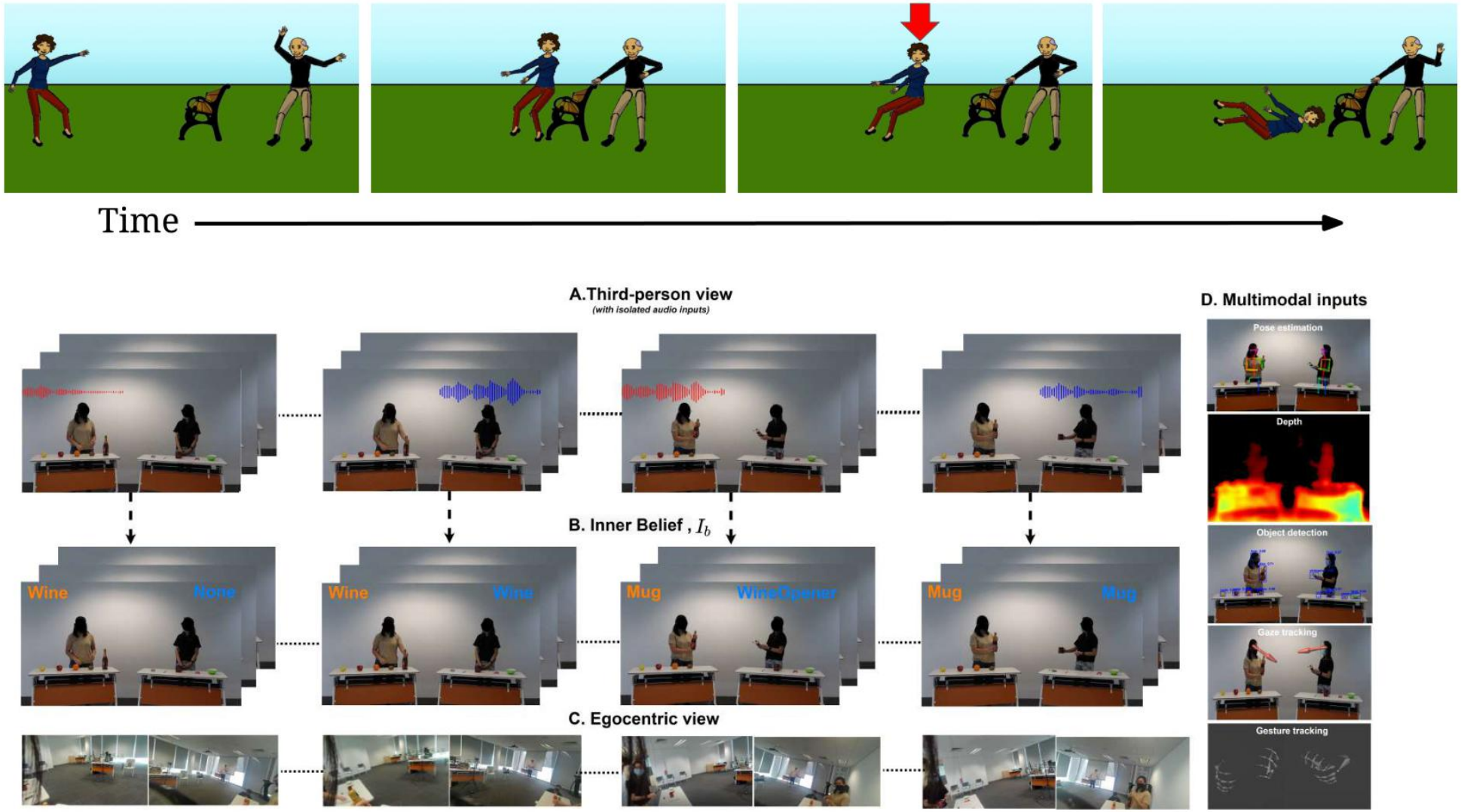
1 Oliver dislikes the kitchen
2 Carter entered the porch.
3 Abigail entered the porch.
4 The potato is in the green_suitcase.
5 Abigail exited the porch.
6 Abigail entered the hall.
7 Carter moved the potato to the green_envelope.
8 Oliver entered the hall.

1 Mila entered the closet.
2 Isla entered the closet.
3 Ava entered the closet.
4 The orange is in the blue_container.
5 Isla exited the closet.
6 Isla entered the garage.
7 Ava moved the orange to the green_bathtub.

1 William entered the staircase.
2 Aiden entered the staircase.
3 Aiden exited the staircase.
4 Aria entered the staircase.
5 The potato is in the red_drawer.
6 Aiden dislikes the grapefruit
7 William moved the potato to the blue_container.
8 Aria exited the staircase.

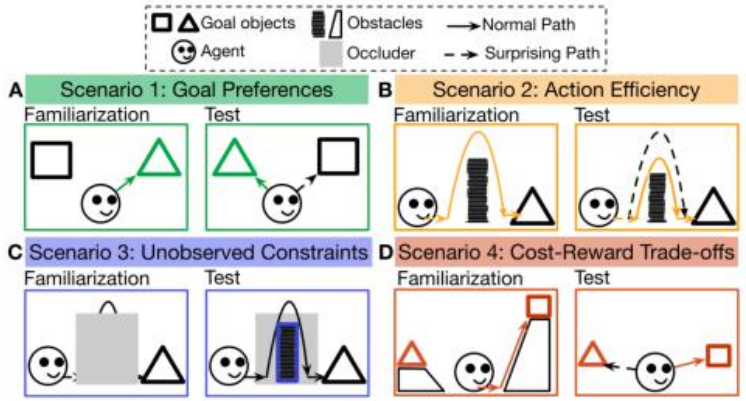
Evaluating Theory of Mind in Question Answering EMNLP, 2018
Revisiting the Evaluation of Theory of Mind through Question Answering EMNLP,2019

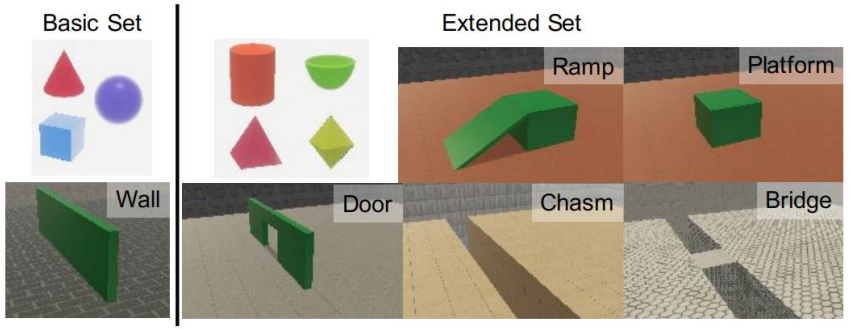
Dataset

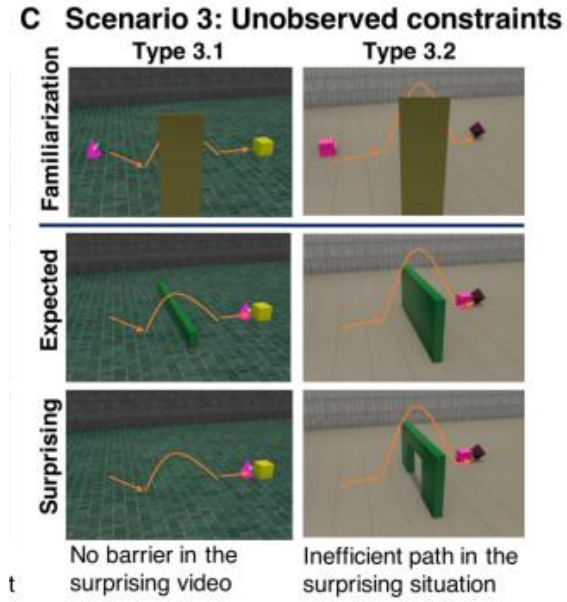


Who is Mistaken?
BOSS: A Benchmark for Human Belief Prediction in Object-context Scenarios

Dataset

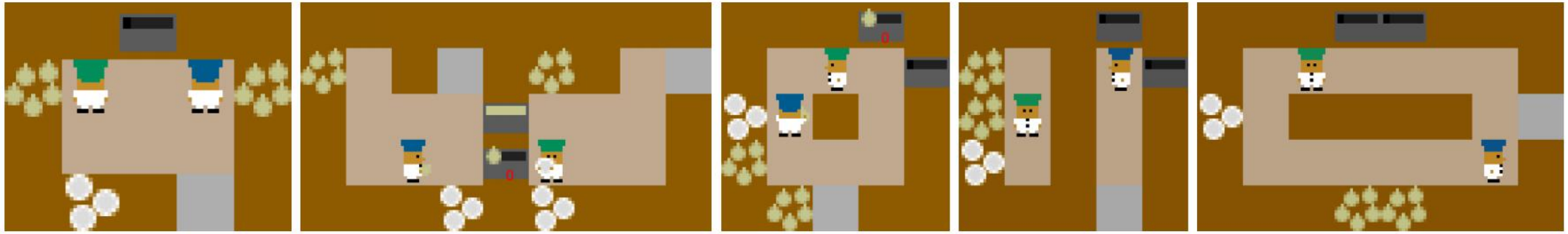


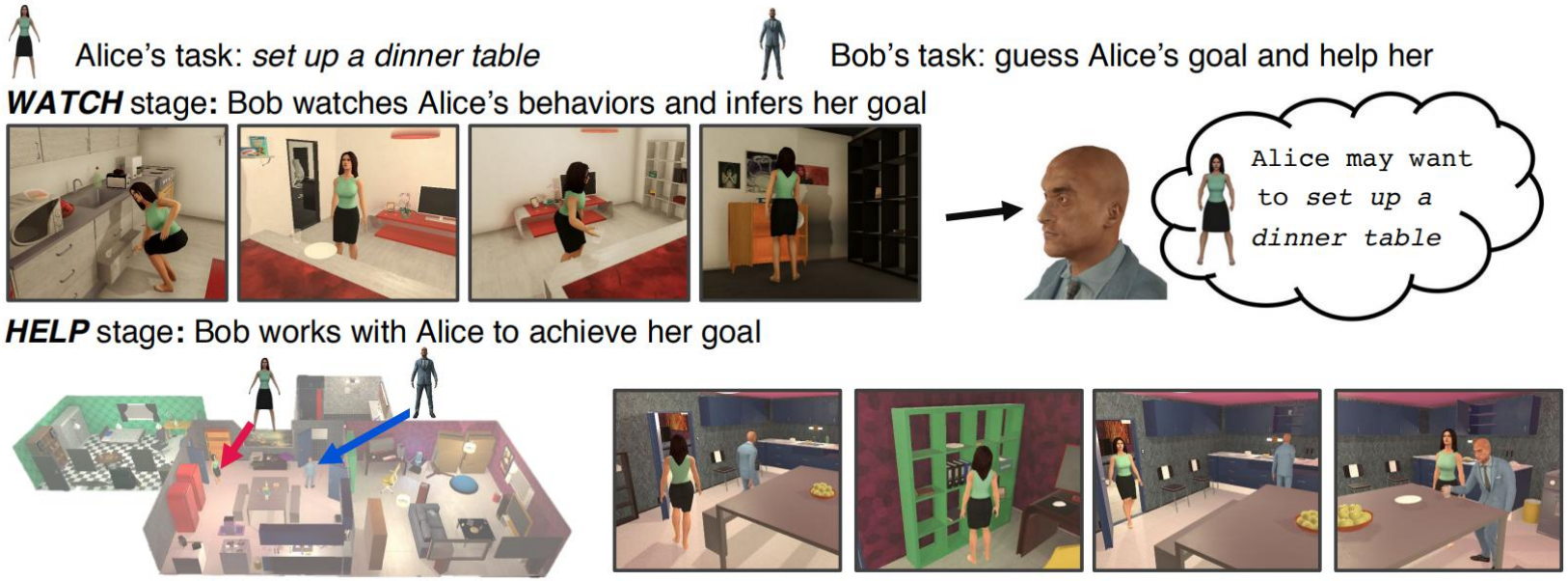




Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. NeurIPS 2021
AGENT: A Benchmark for Core Psychological Reasoning. PMLR 2021

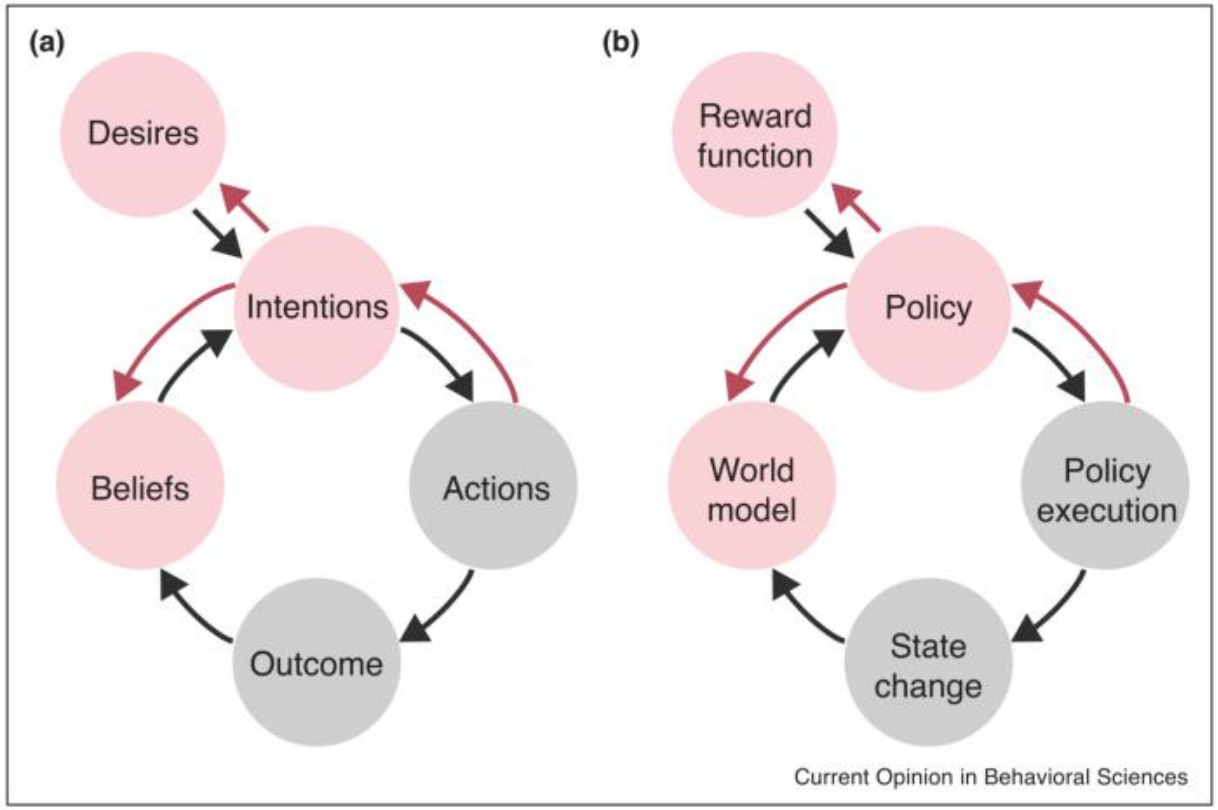
Datasets





On the Utility of Learning about Humansfor Human-AI Coordination. NeurIPS 2019
Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration. ICLR 2021

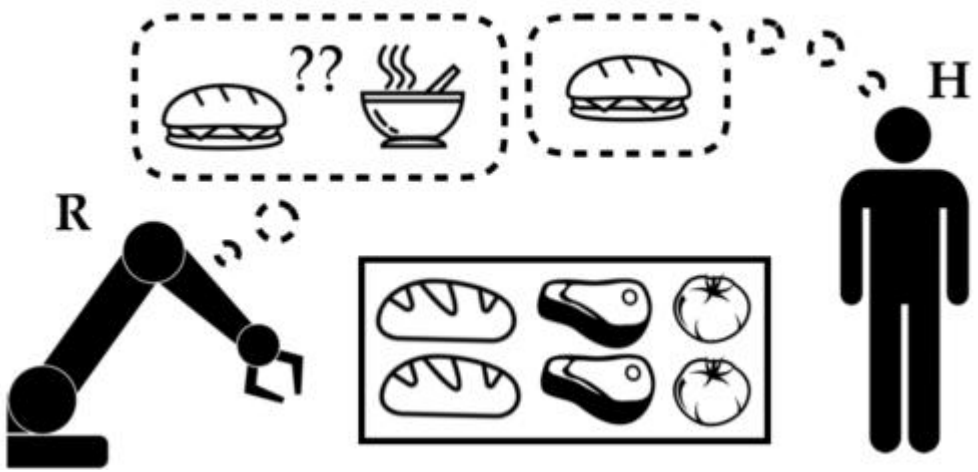
Methods



1 人类的有限理论
2 .人机交互中，机器人和人类是在同一个环境下机器人的动作会对人类的信念产生影响
3 .逆强化学习计算成本很高

Methods

CIRL game



The Transition Dynamics

$$P(s', a^H \mid s, \sigma) = P((x', \theta'), a^H \mid (x, \theta), (a^R, v))$$
$$= T(x, a^H, a^R, x') \cdot \mathbb{1}(\theta' = \theta) \cdot P(a^H \mid x, a^R, v, \theta)$$
$$= T(s, a^H, a^R, s') \cdot \mathbb{1}(a^H = \arg\max_{a^H} Q_H(s, a^H, \sigma)) \quad (2)$$

预测人类的动作时考虑到人类对机器的认识

Bellman update

$$\alpha_\sigma(s) = R(s) + \gamma \cdot \sum_{a^H} \pi_H(a^H \mid Q_H(s, a^H, \sigma)) \cdot$$
$$\sum_{s' \in S} T(s, a^H, a^R, s') \cdot \alpha_{v(a^H)}(s'). \quad (4)$$

Relaxing CIRL ’s Assumption of Rationality

$$\pi_H(a^H \mid Q_H(s, a^H, \sigma)) \propto \exp(\beta \cdot Q_H(s, a^H, \sigma))$$

通过参数来调整人的理性程度

An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning

Methods

Human推理过程

POMDP模型

<S, A, T , R, Ω, O>

S：状态空间，分为两个部分，<x，y>，x代表agent自身的位置，y代表餐车的位置。

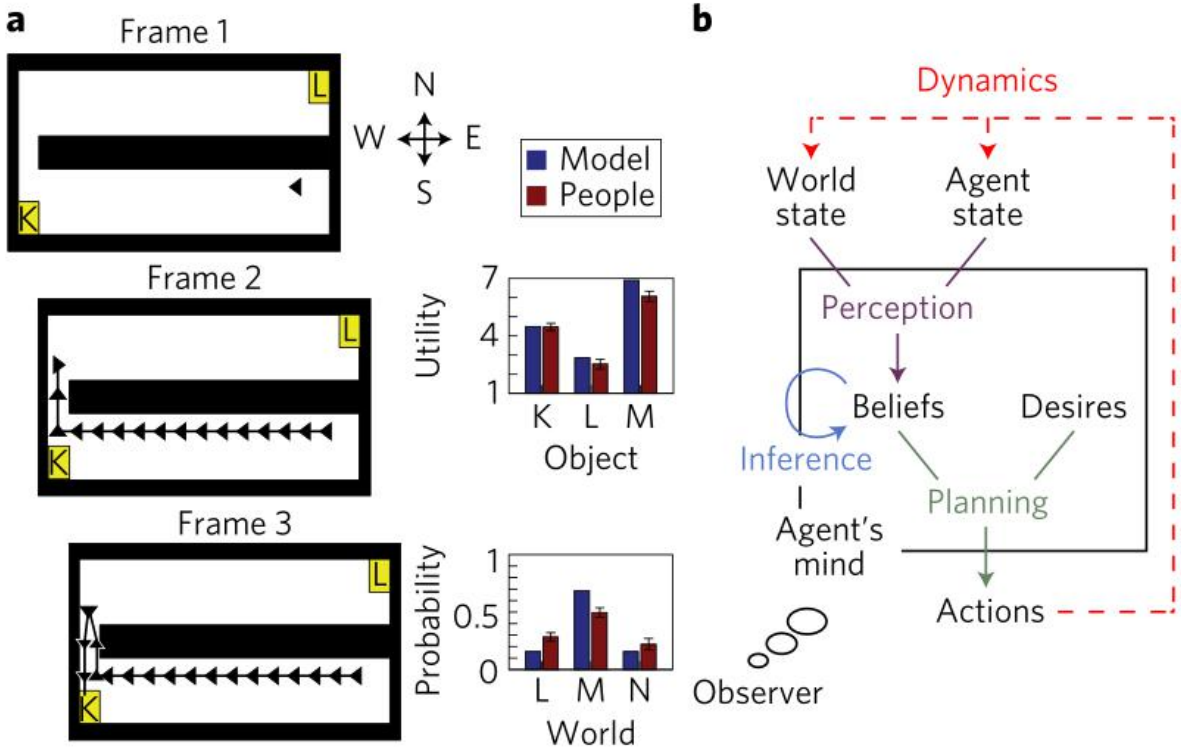
A:动作空间， “North”, “South”, “East”, “West”, “Stay”, and “Eat”

T:状态转移矩阵， T (St+1, St, At) = P (St+1|St, At)

R:奖励函数，R(x , y , a)

Ω: 观测空间，agent在场景中360度可观测环境。包含坐标空间中所有点的多边形区域

O：观测分布，P(x,y|o) 表示在可观测环境o下，状态空间<x,y>的分布



Rational quantitative attribution of beliefs, desires and percepts in human mentalizing nature 2017

Methods

Agent推理human的belief和desire过程，看作为HMM模型，使用forward-backward算法计算联合概率。

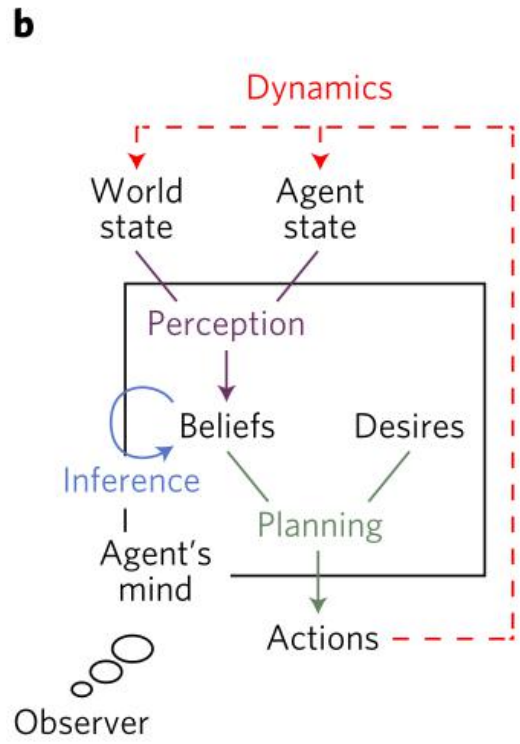
$$D_t^{ijk} = \sum_{o_t, a_{t-1}} P(o_t|s_t)P(s_t|s_{t-1}, a_{t-1})P(a_{t-1}|b_{t-1}^j, s_{t-1}, r_{\mathcal{G}}^k)\delta(b_t^i, BU(o_t, s_t, s_{t-1}, a_{t-1}, b_{t-1}^j)).$$

状态转移→ belief upgrade

$$P(b_t^i, r_{\mathcal{G}}^k|s_{1:t}) \propto \sum_{j=1:m(t-1)} D_t^{ijk} \cdot P(b_{t-1}^j, r_{\mathcal{G}}^k|s_{1:t-1}).$$ forward distribution

$$P(s_{t+1:T}|b_t^i, r_{\mathcal{G}}^k) \propto \sum_{i=1:m(t+1)} D_{t+1}^{jik} \cdot P(s_{t+2:T}|b_{t+1}^j, r_{\mathcal{G}}^k).$$ backward distribution

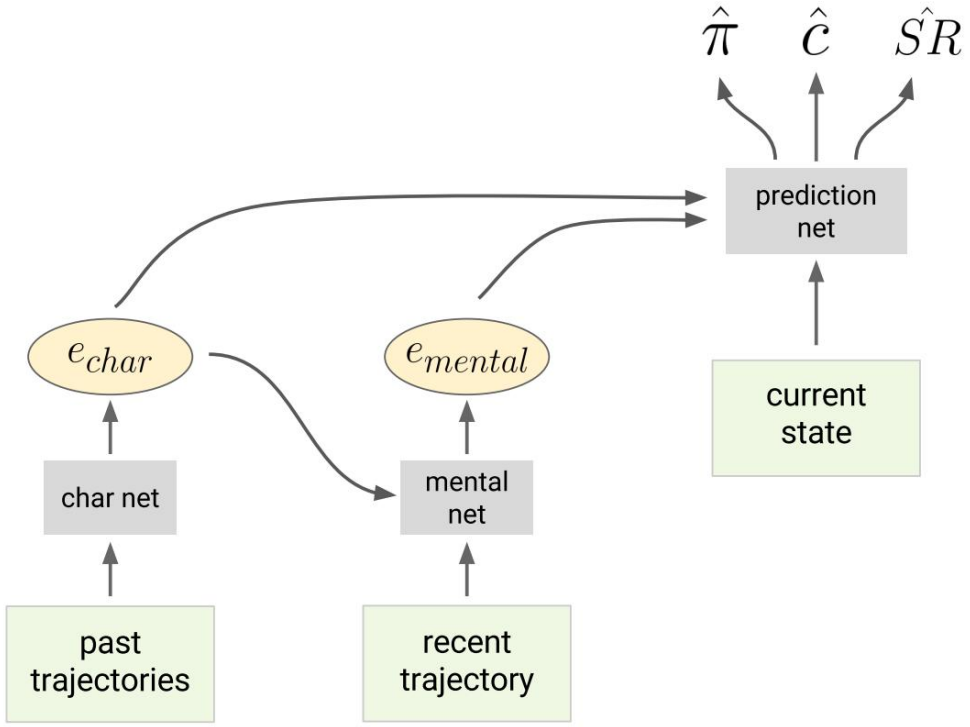
$$P(b_t^i, r_{\mathcal{G}}^k|s_{1:T}) = P(b_t^i, r_{\mathcal{G}}^k|s_{1:t})P(s_{t+1:T}|b_t^i, r_{\mathcal{G}}^k).$$ forward and backward distributions

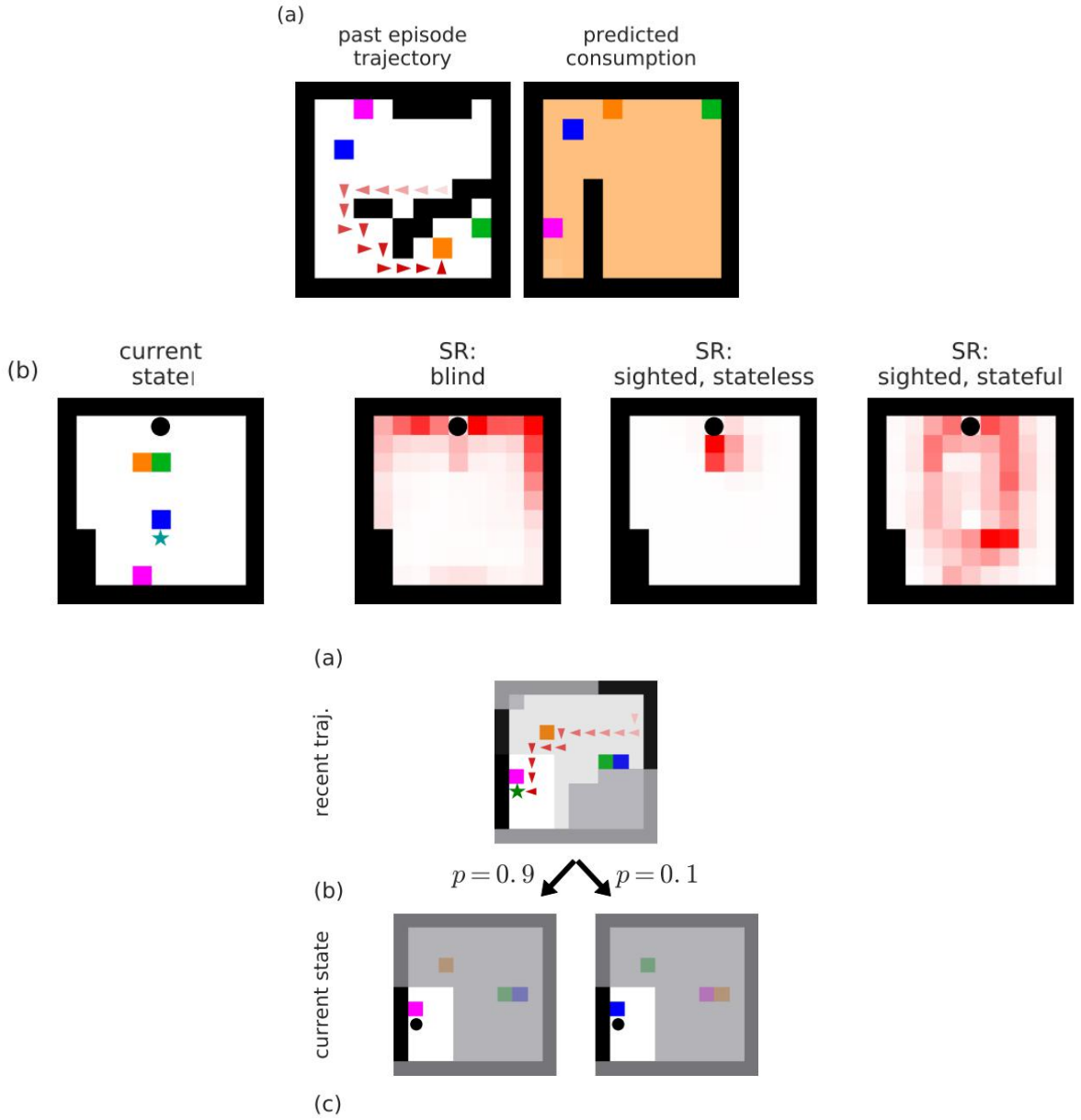


Rational quantitative attribution of beliefs, desires and percepts in human mentalizing nature 2017

Methods

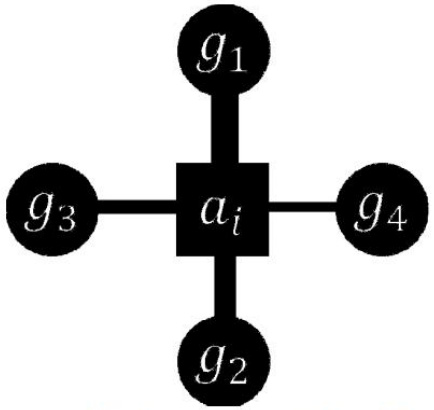
数据驱动：需要3200万个训练示例来完成六个月大婴儿的目标推断。如果人类以类似的方式学习心理理论，婴儿每天需要接受近17.5万次有标签的目标训练。



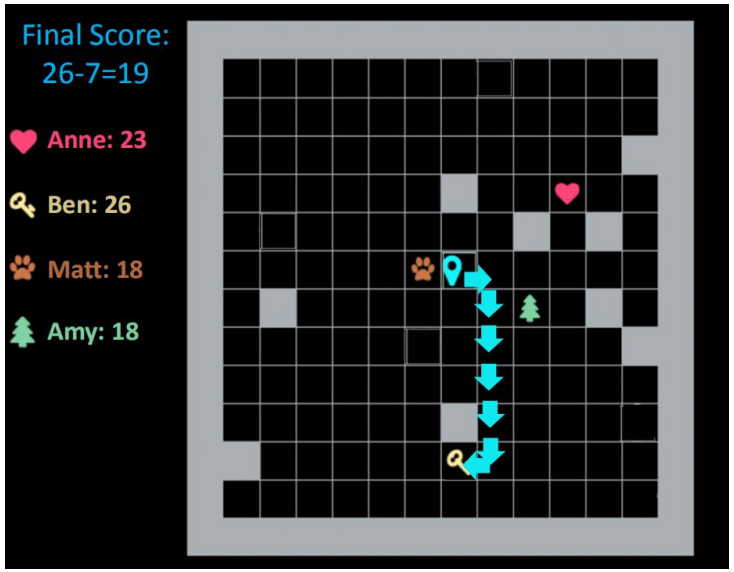


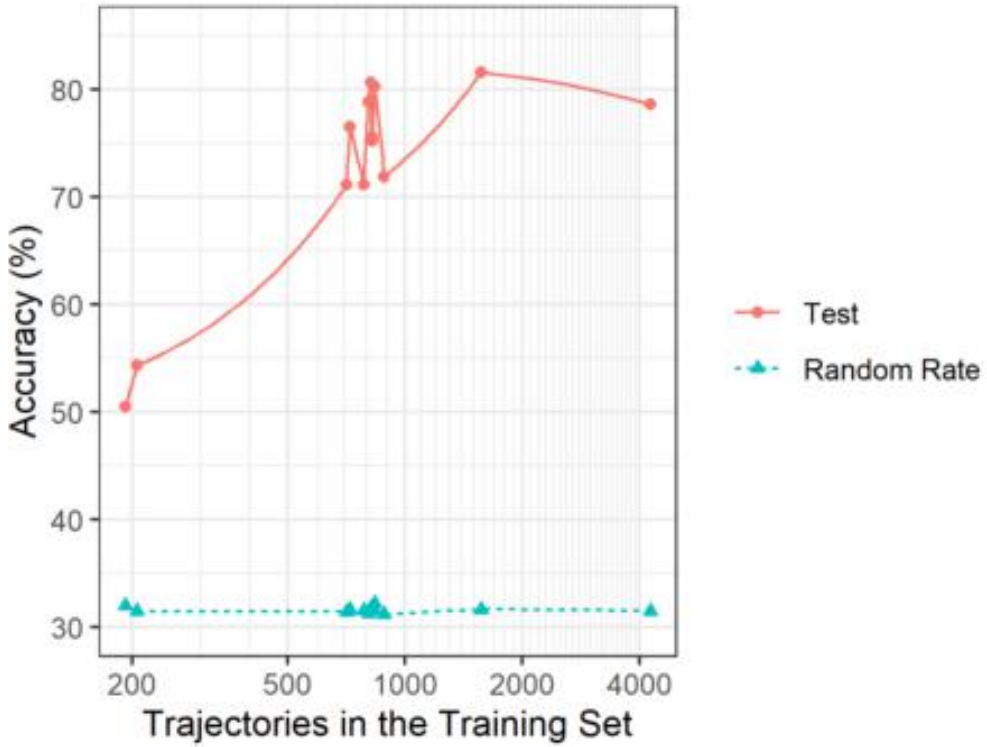
Machine Theory of Mind . PMLR, 2018

Methods



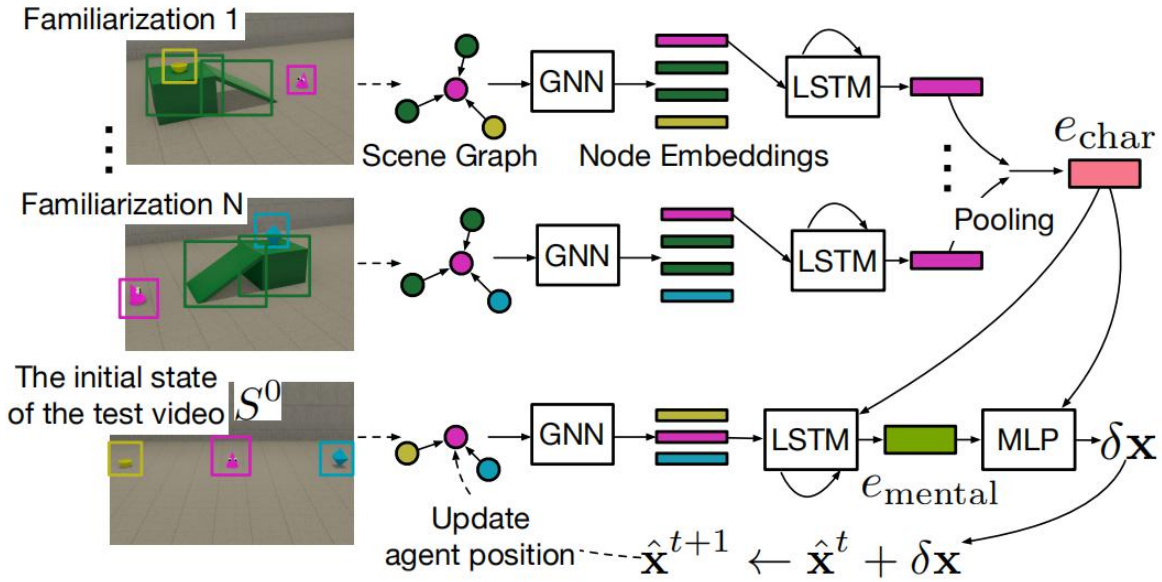
Target (g_s)	Social Support Score (u_s)
g_1	21
g_2	17
g_3	10
g_4	7





Using Machine Theory of Mind to Learn Agent Social Network Structures from Observed Interactive Behaviors with Targets IEEE, 2018

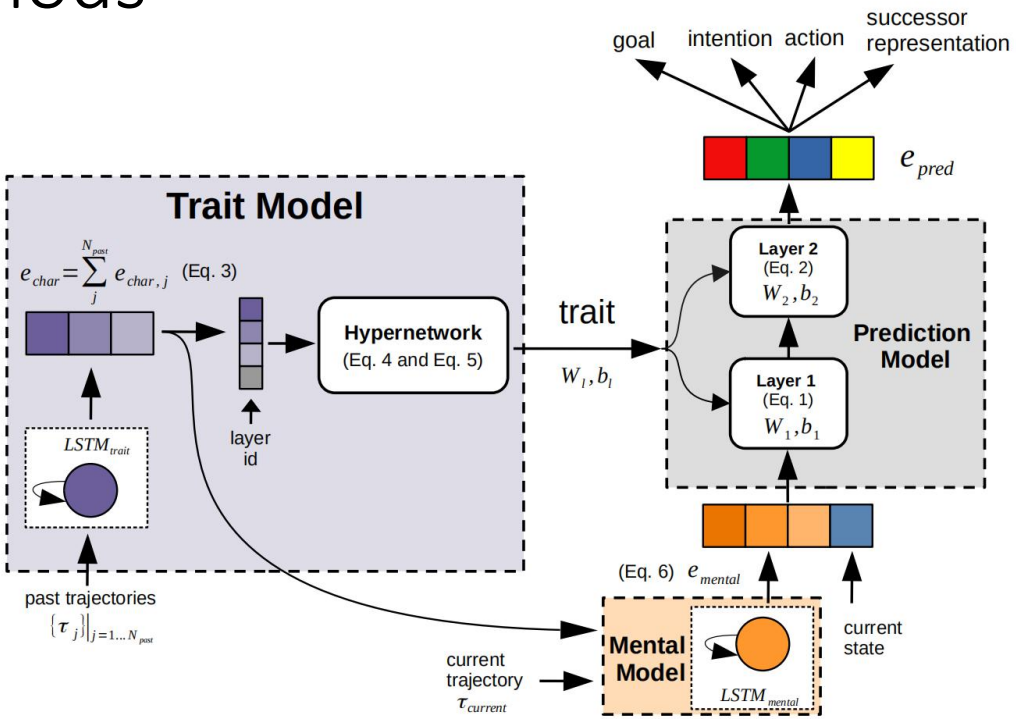
Methods

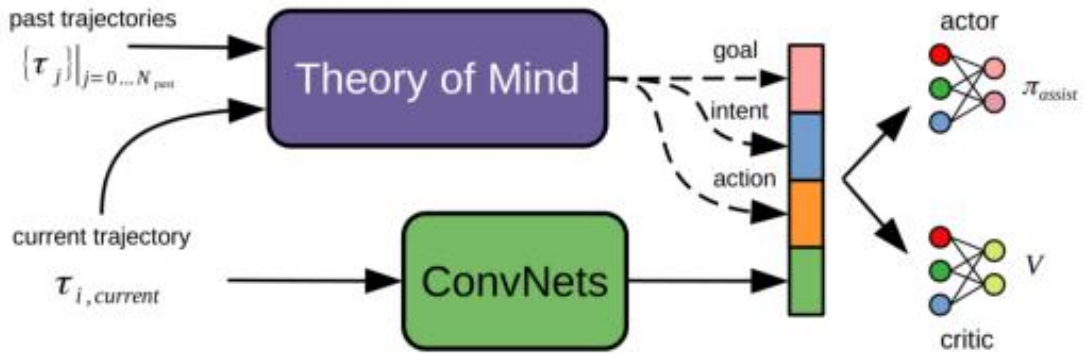


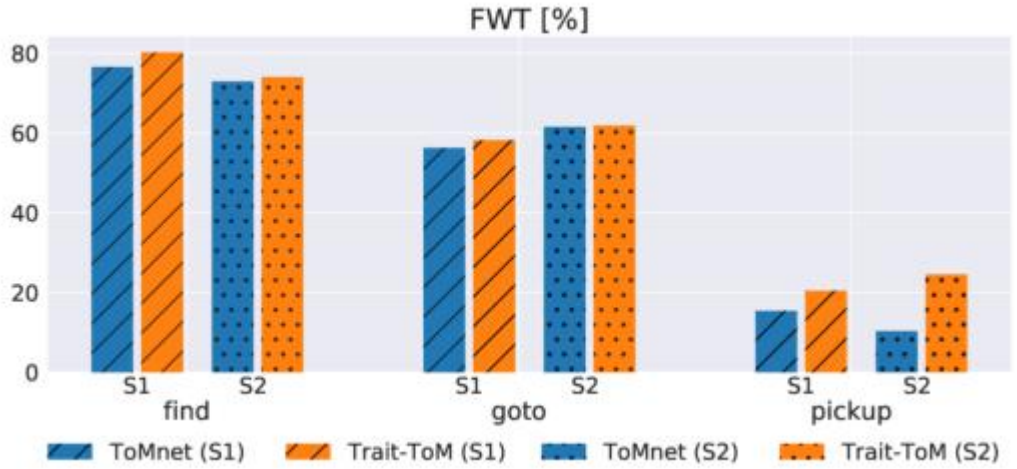
Condition	Method	Goal Preferences					Action Efficiency						Unobs.			Cost-Reward			All
		1.1	1.2	1.3	1.4	All	2.1	2.2	2.3	2.4	2.5	All	3.1	3.2	All	4.1	4.2	All	
	Human	.95	.95	.92	.97	.95	.87	.93	.86	.95	.94	.91	.88	.94	.92	.82	.91	.87	.91
All	ToMnet-G	.57	1.0	.67	1.0	.84	.95	1.0	.95	1.0	1.0	.98	.93	.87	.89	.82	.97	.89	.90
All	BIPaCK	.97	1.0	1.0	1.0	.99	1.0	1.0	.85	1.0	1.0	.97	.93	.88	.90	.90	1.0	.95	.96
G1	ToMnet-G	.50	.90	.63	.88	.75	.90	.75	.45	.90	.05	.66	.58	.77	.69	.48	.48	.48	.65
G1	BIPaCK	.93	1.0	1.0	1.0	.98	1.0	1.0	.80	1.0	1.0	.97	.93	.82	.86	.88	1.0	.94	.94
G2	ToMnet-G	.37	.95	.63	.88	.71	.35	.60	.75	.68	.85	.65	.63	.80	.73	.55	.95	.75	.71
G2	BIPaCK	.93	1.0	1.0	1.0	.98	1.0	1.0	.75	1.0	.95	.95	.88	.85	.87	.83	1.0	.92	.94

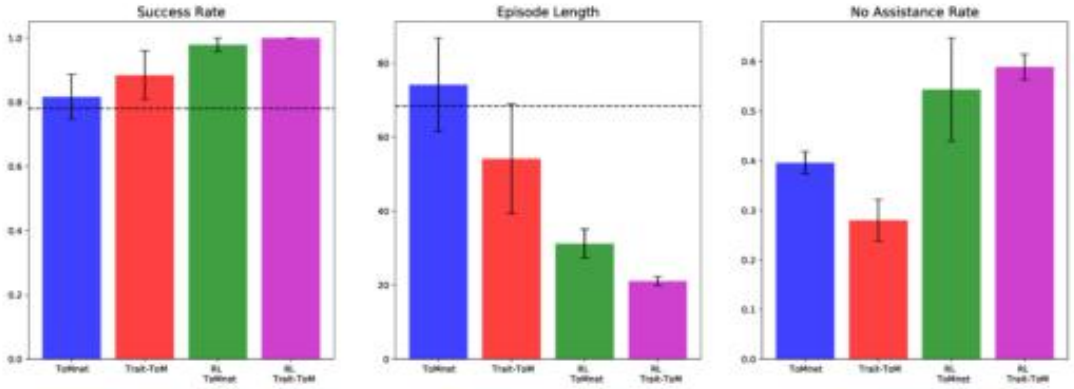
AGENT: A Benchmark for Core Psychological Reasoning. PMLR 2021

Methods









Learning Theory of Mind via Dynamic Traits Aribution. AAMAS 2022

Summary

- 模型和数据集不匹配，现有提出的数据集还为得到利用
- 现有提出的模型尚未在统一的数据集上得到验证，无法进行比较
- 对于深度学习的方法，如何提高模型的可解释性
- 对于贝叶斯方法和逆强化学习方法，如何减低计算的复杂性