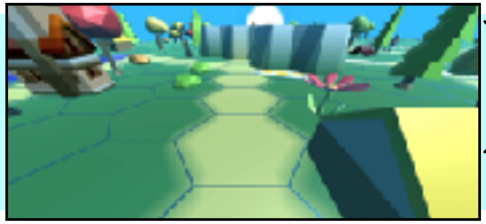
Vision and Language

Berkeley
N L P

slides from: Daniel Fried, Yonatan Bisk, L-P Morency

Situated Instruction Following

f (instruction, ) → actions

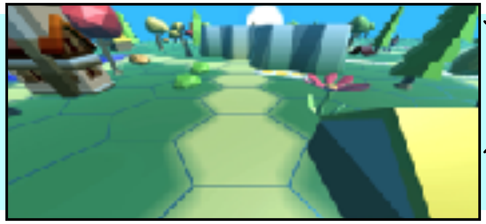
Room to Room, Anderson et al. 2018

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Touchdown, Chen et al. 2018

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right.

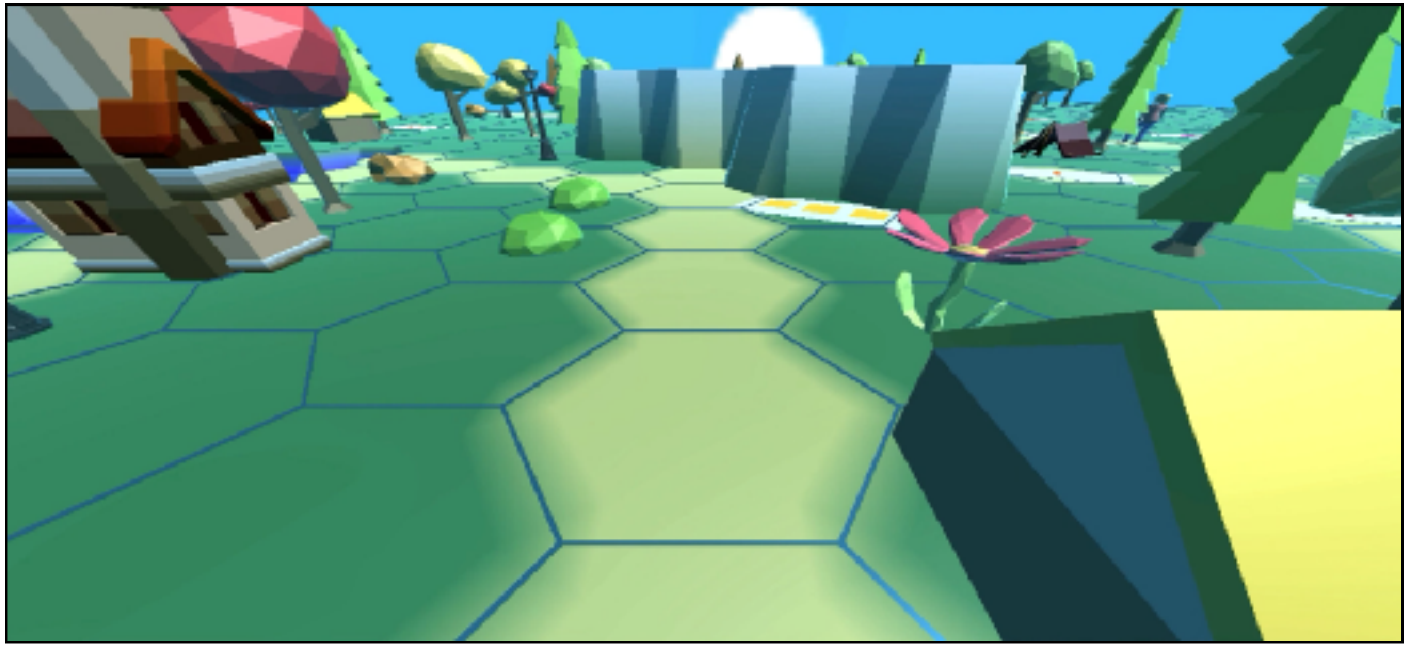
Situated Instruction Following

f (instruction, ) → actions

ALFRED, Shridhar et al. 2020

CerealBar, Suhr et al. 2019

Pick up knife, cut potato, put potato in fridge, remove from fridge, place in the microwave

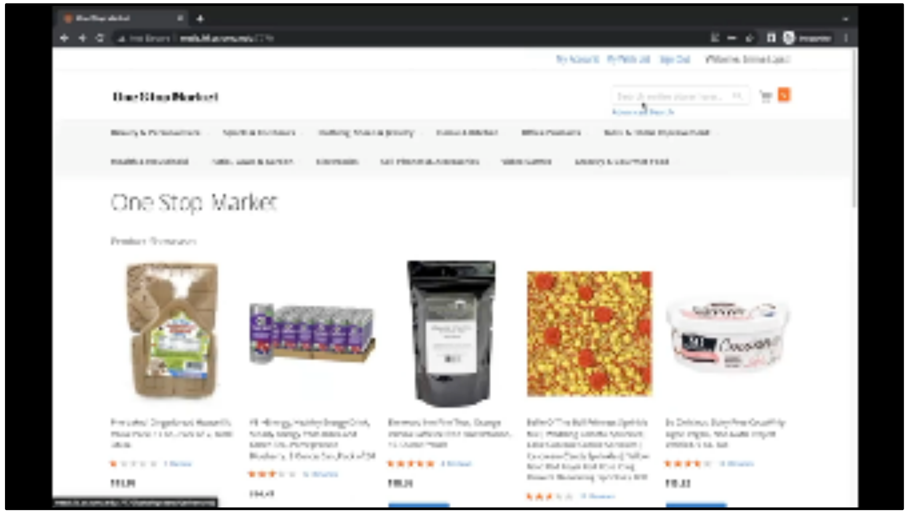


Turn around and get the three red stripes behind you.

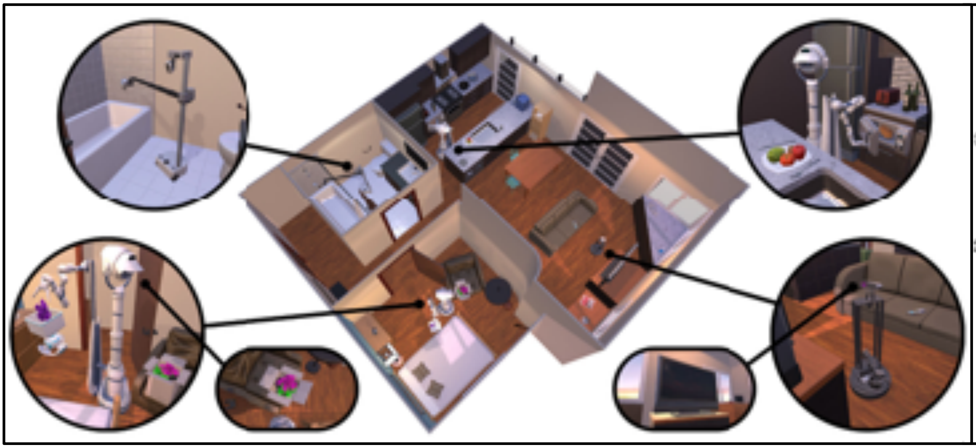
Environments

- 2D or 3D rendered environments
  - Can easily generate new environments on the fly
  - Support manipulable environments
  - Simulation allows for rapid experimentation and evaluation

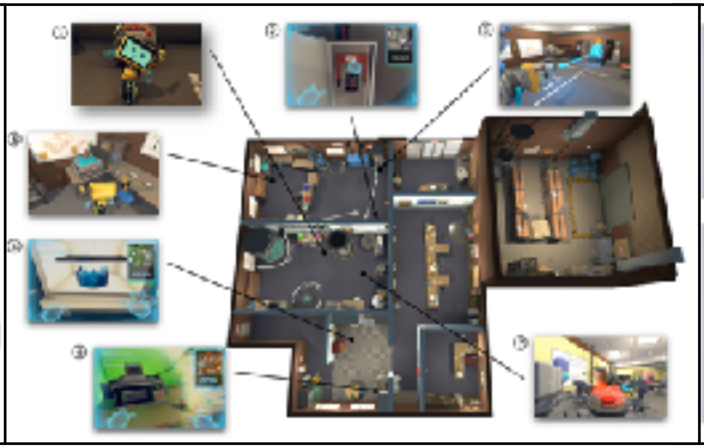
WebArena, Zhou Shuyan et al. 2023



AI2-THOR, Kolve et al. 2022



Alexa Arena, Gao Qiaozi et al. 2023

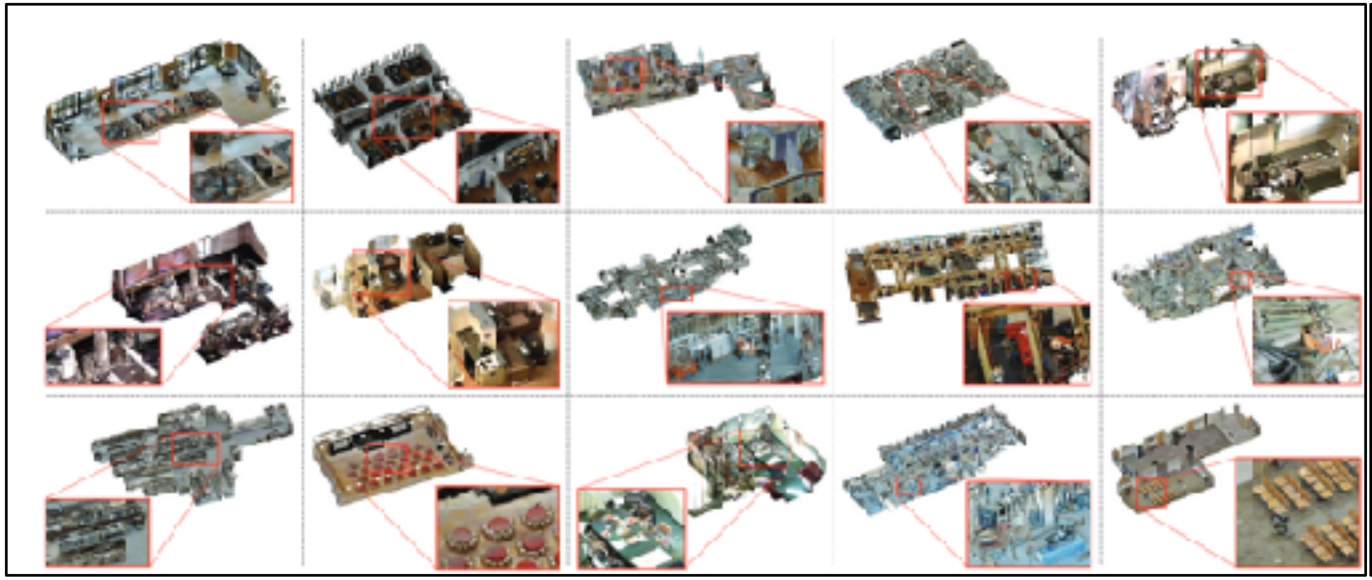


VRKitchen, Gao Xiaofeng et al. 2019

Environments

- 2D or 3D rendered environments
- Photorealistic environments

Gibson Env, Xia Fei et al. 2018

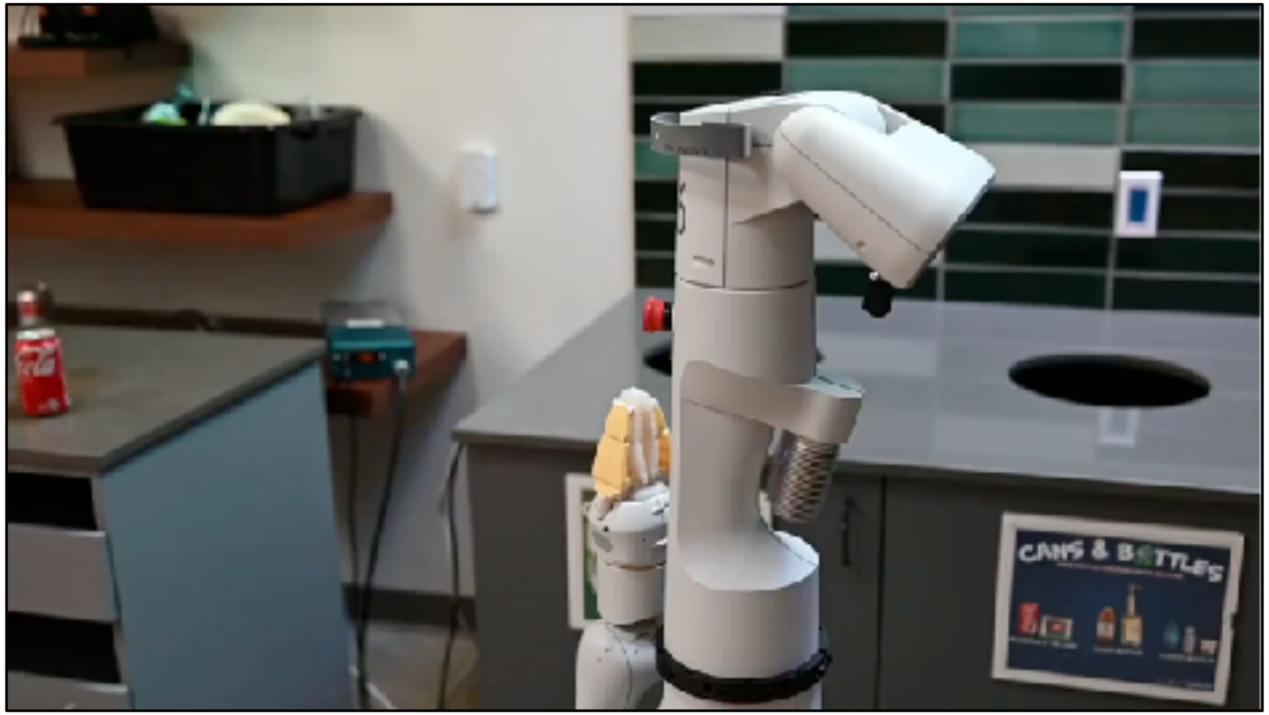


StreetLearn, Mirowski et al. 2019

Environments

- 2D or 3D rendered environments
- Photorealistic environments
- Literal physical embodiment (robotics)

SayCan, Ahn et al. 2022



GRIF, Myers et al. 2023

Place the knife in front of the microwave.

Embodied Agents: Challenges

- Grounding language to perception
- Reasoning about world dynamics
- Grounding language to action
- In collaborative tasks: also reasoning about one’s interlocutor
- Evaluating success

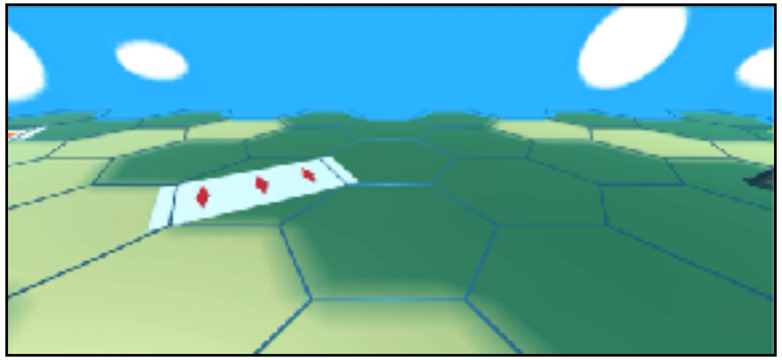
Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

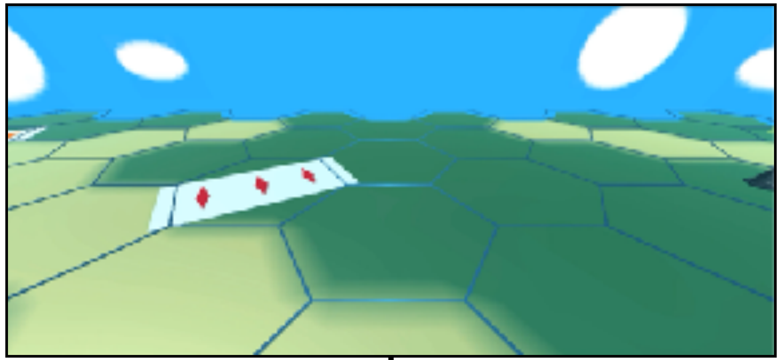
- States $\mathcal{S}$ (and observations $\mathcal{O}$)



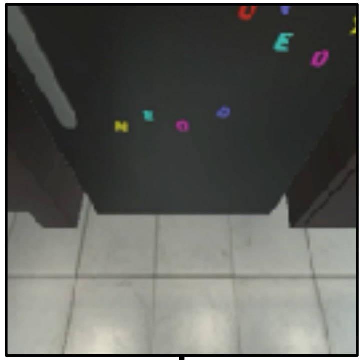
Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

- States $\mathcal{S}$ (and observations $\mathcal{O}$)
- Actions $\mathcal{A}$



LEFT

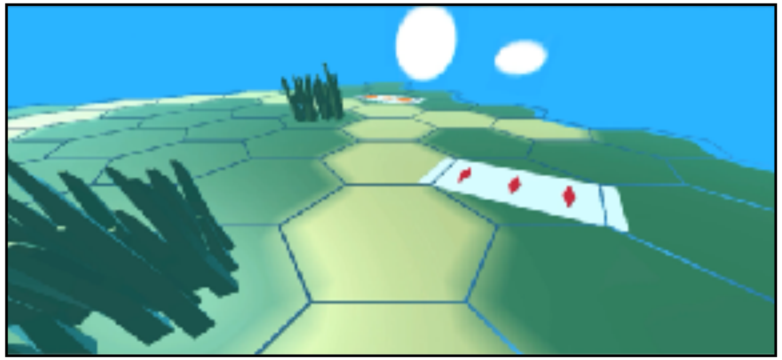


OPEN(FRIDGE)

Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

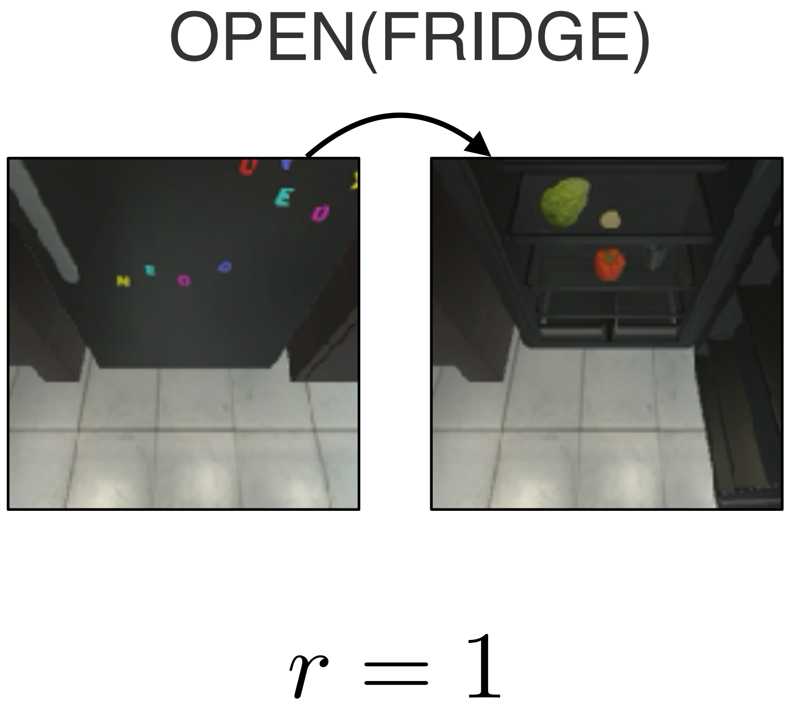
- States $\mathcal{S}$ (and observations $\mathcal{O}$)
- Actions $\mathcal{A}$
- Transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$



Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

- States $\mathcal{S}$ (and observations $\mathcal{O}$)
- Actions $\mathcal{A}$
- Transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta^{\mathcal{S}}$
- Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$



Reasoning about World Dynamics

(Partially observable) Markov decision process formulation of embodied agents

- States $\mathcal{S}$ (and observations $\mathcal{O}$)
- Actions $\mathcal{A}$
- Transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$
- Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

$$\pi : \mathcal{O} \to \Delta^{\mathcal{A}}$$

Reasoning about World Dynamics

- What is your state space?
  - Does it include all information about the environment?
  - Does it include information about the trajectory so far, e.g., previous states and actions?
  - Does it include a natural language instruction?
- Is the environment partially observable?
- What is the action space?
  - Lowest level action space: continuous control
  - Higher level action space: sufficient for simulated environments
- How is the policy implemented?

Embodied Agent Policies

Observation space:

- Previous and current visual observations
- Previous actions
- Instruction

Policy: whatever neural implementation you want

π

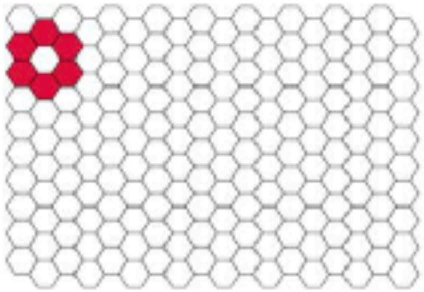
Turn around and get the three red stripes behind you.

Action	Probability
LEFT	64%
RIGHT	2%
FORWARD	28%
BACKWARD	3%
STOP	3%

Grounding Language to Action

- How do we define our action space?
- In many cases, language provides a decent set of abstractions that help us define meaningful higher-level action spaces
- Language can also allude to structured action spaces

1. Make a *red flower*, by coloring in red *all tiles adjacent* to the 2nd tile from the top in the 2nd column from the left.

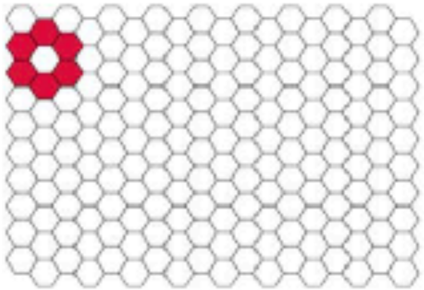


Hexagons, Lachmy et al. 2022

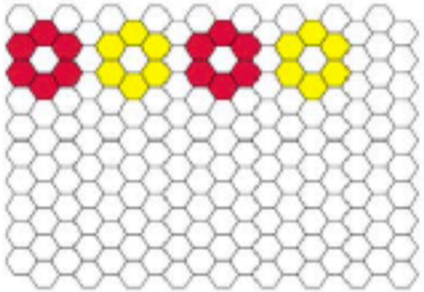
Grounding Language to Action

- How do we define our action space?
- In many cases, language provides a decent set of abstractions that help us define meaningful higher-level action spaces
- Language can also allude to structured action spaces

1. Make a *red flower*, by coloring in red *all tiles adjacent* to the 2nd tile from the top in the 2nd column from the left.



2. *Repeat* this *flower* pattern *across the board* to the right, *alternating* yellow and red, leaving a blank column *between every 2* flowers.

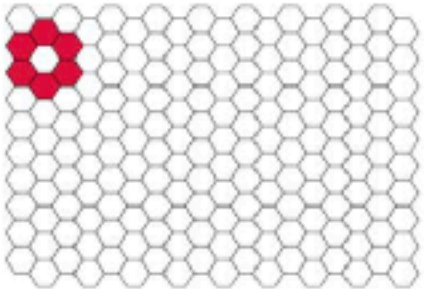


Hexagons, Lachmy et al. 2022

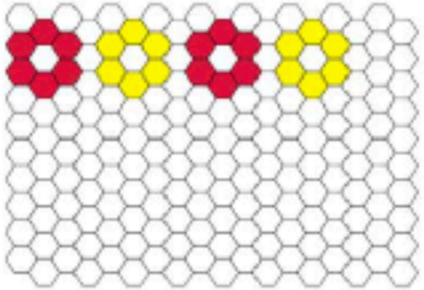
Grounding Language to Action

- How do we define our action space?
- In many cases, language provides a decent set of abstractions that help us define meaningful higher-level action spaces
- Language can also allude to structured action spaces

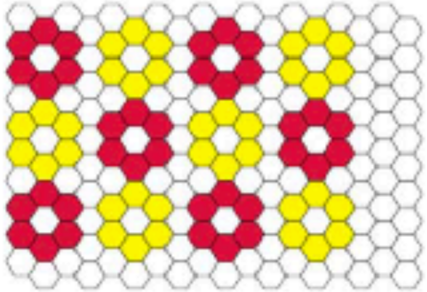
1. Make a *red flower*, by coloring in red *all tiles adjacent* to the 2nd tile from the top in the 2nd column from the left.



2. *Repeat* this *flower* pattern *across the board* to the right, *alternating* yellow and red, leaving a blank column *between every 2* flowers.



3. *Repeat* this *row of flowers* 2 more times, but *reverse* the colors *in each new row.* You should get 6 red flowers and 6 yellow flowers *in total.*



Hexagons, Lachmy et al. 2022

Reasoning about an Interlocutor

- Single instruction following — still could require pragmatic reasoning

Room to Room, Anderson et al. 2018

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Reasoning about an Interlocutor

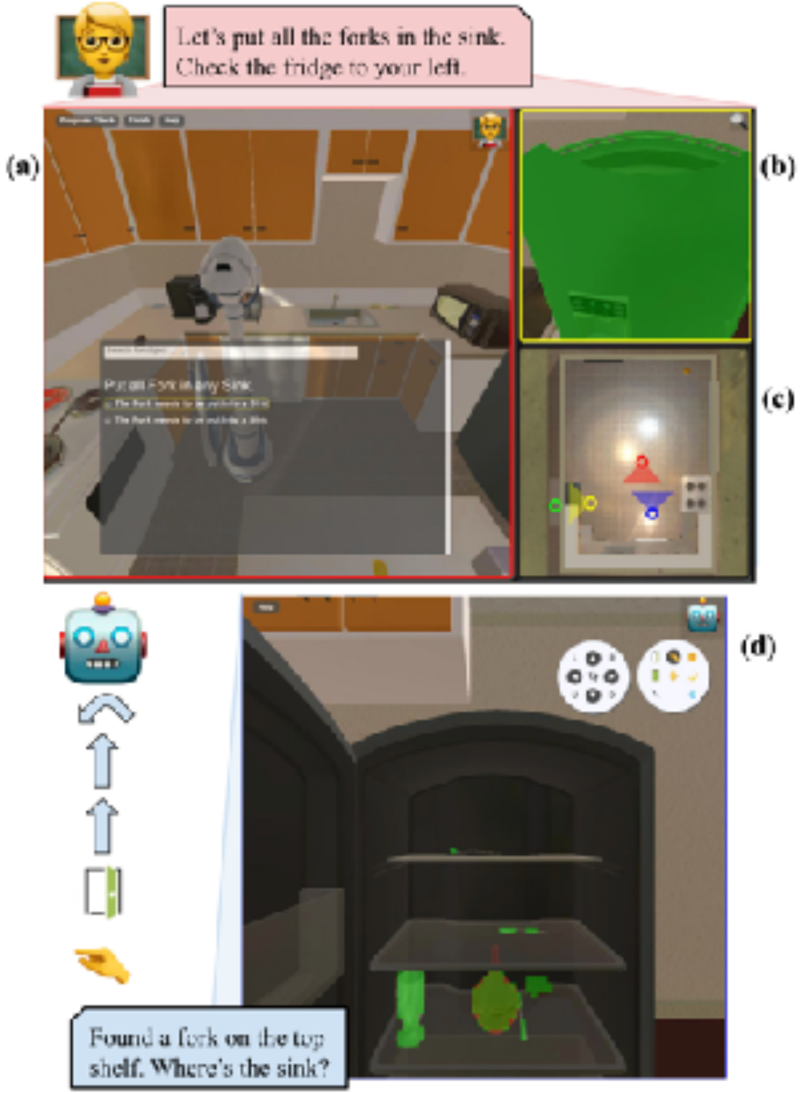
- Single instruction following — still could require pragmatic reasoning
- Following sequences of instructions — user can dynamically instruct the agent according to its current behavior

CerealBar, Suhr et al. 2019

Reasoning about an Interlocutor

- Single instruction following — still could require pragmatic reasoning
- Following sequences of instructions — user can dynamically instruct the agent according to its current behavior
- Bidirectional conversation — agent can ask for clarification or help

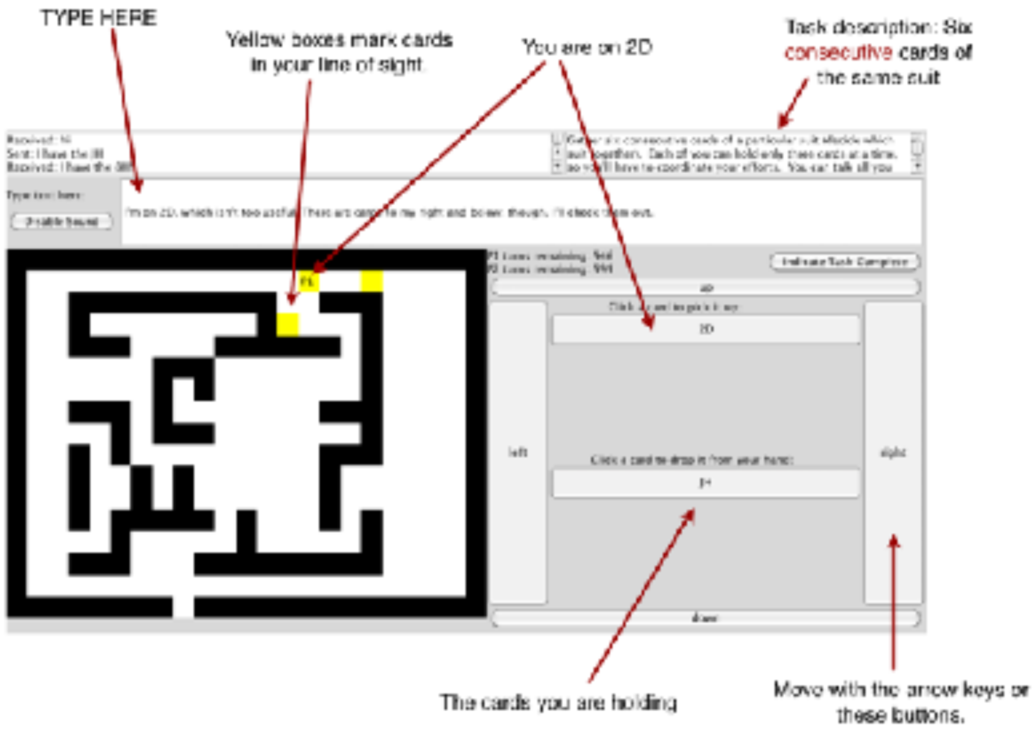
TEACh, Padmakumar et al. 2021



Reasoning about an Interlocutor

- Single instruction following — still could require pragmatic reasoning
- Following sequences of instructions — user can dynamically instruct the agent according to its current behavior
- Bidirectional conversation — agent can ask for clarification or help
- Fully embodied multi-agent conversation — agents can form conventions, negotiate how to solve the task, perform joint planning, etc.

CARDS, Djalali et al. 2011



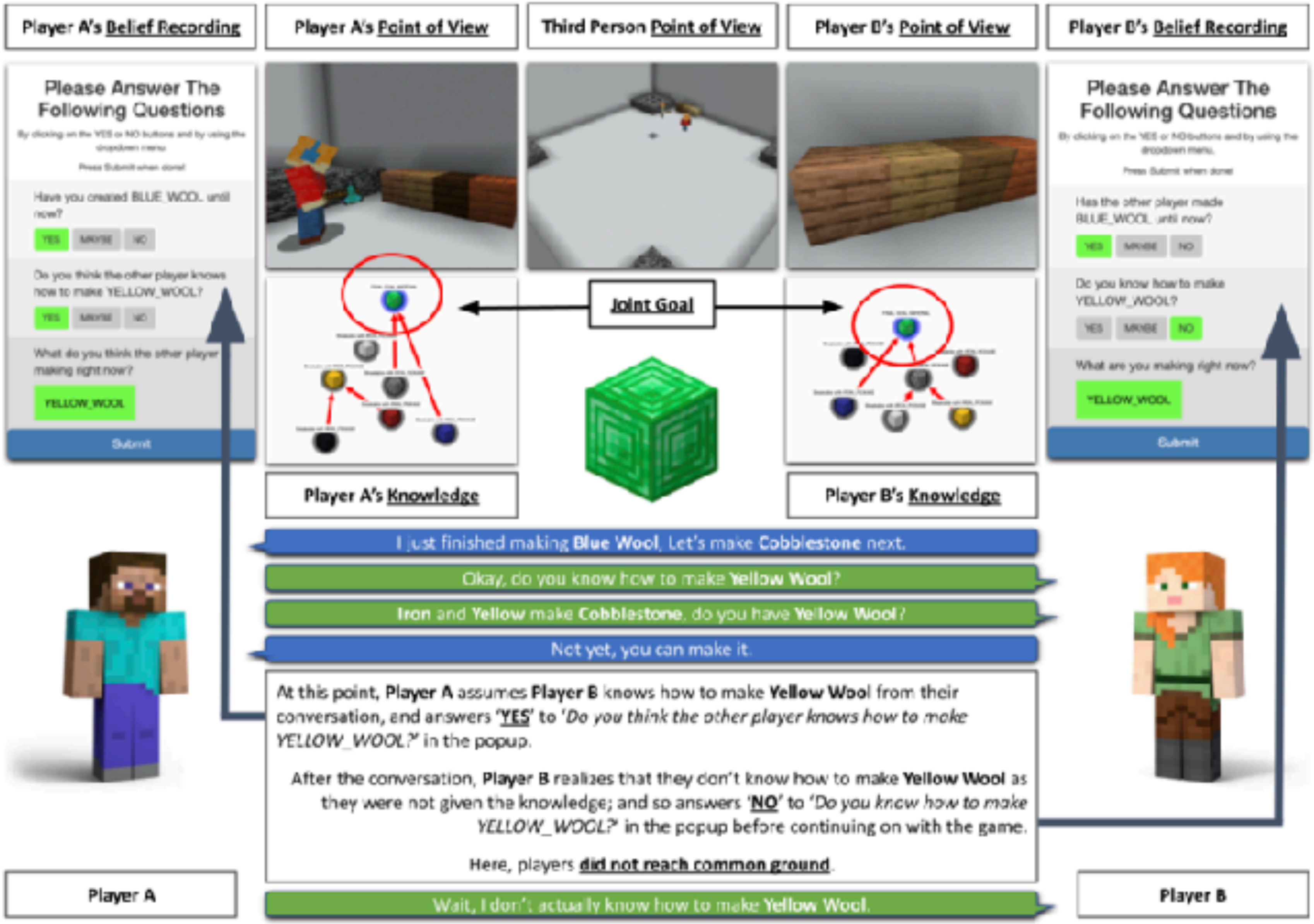
Reasoning about an Interlocutor

- Single instruction following — still could require pragmatic reasoning
- Following sequences of instructions — user can dynamically instruct the agent according to its current behavior
- Bidirectional conversation — agent can ask for clarification or help
- Fully embodied multi-agent conversation — agents can form conventions, negotiate how to solve the task, perform joint planning, etc.

Portal 2 Dialogues

Reasoning about an Interlocutor

- Pragmatic reasoning
- In collaborative tasks: agents need to use language to achieve a shared goal
- Need to model other agent's:
  - Beliefs
  - Goals
  - Observations
  - Knowledge
  - Affordances



MindCraft, Bara et al. 2021

Evaluating Success

- High-level desideratum of language agents: **assist a human user in accomplishing their goal as efficiently as possible.**
- Automatic evaluation
  - Low-level metrics: matching human demonstrations
    - Entire action sequence
    - Action-level accuracy, conditioned on oracle prefix
  - Higher-level metrics: success rate
  - Difficult to define for multi-turn conversation
- Human evaluation
  - When deployed with real users, how effective is the agent?
  - Challenge: human adaptation of expectations, behavior, and language

Learning

- Imitation learning

$$\underbrace{\arg\max_{\theta}}_{\text{}} \; \mathbb{E}_{(o,a)\in\mathcal{D}} \; \pi(a \mid o; \theta)$$

Maximum likelihood objective

Expectation over demonstrations

Policy parameterized with θ

Essentially supervised learning on a dataset of instructions and observations paired with human demonstrations.

Learning

- Imitation learning
- Reinforcement learning

$$\arg\max_{\theta} \mathbb{E}_{\tau\sim\pi_\theta} \mathcal{R}(\tau)$$

Expectation over trajectories sampled from π

Reward achieved by trajectory

$$a_i \sim \pi_\theta(\cdot \mid s_{i-1})$$

$$s_i \sim \mathcal{T}(\cdot \mid s_{i-1}, a_i)$$

$$\mathcal{R}(\tau) = \sum_{i=0}^{|\tau|} \mathcal{R}(s_i, a_i)\gamma^i$$

Learning

- Imitation learning
- Reinforcement learning
- LLM planning methods

SayCan, Ahn et al. 2022

Interaction

- A multi-turn dynamic process where two or more agents respond to one another's actions
- Open language-related questions raised by interaction:
  - How do we reason about other agents?
  - How do we learn language?
  - How do we use language in real-time interaction?
  - How do we coordinate using language?

Reasoning About Other Agents

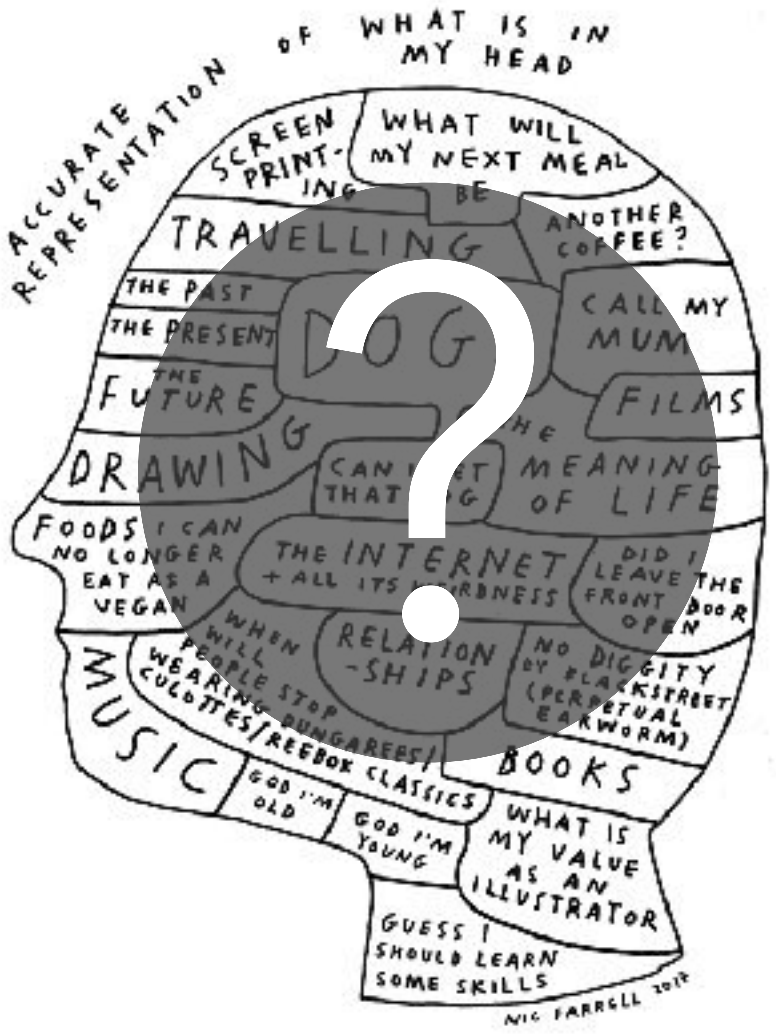
- (Slides from UW CS 447, by Hyunwoo Kim)

Mind

We know we have one

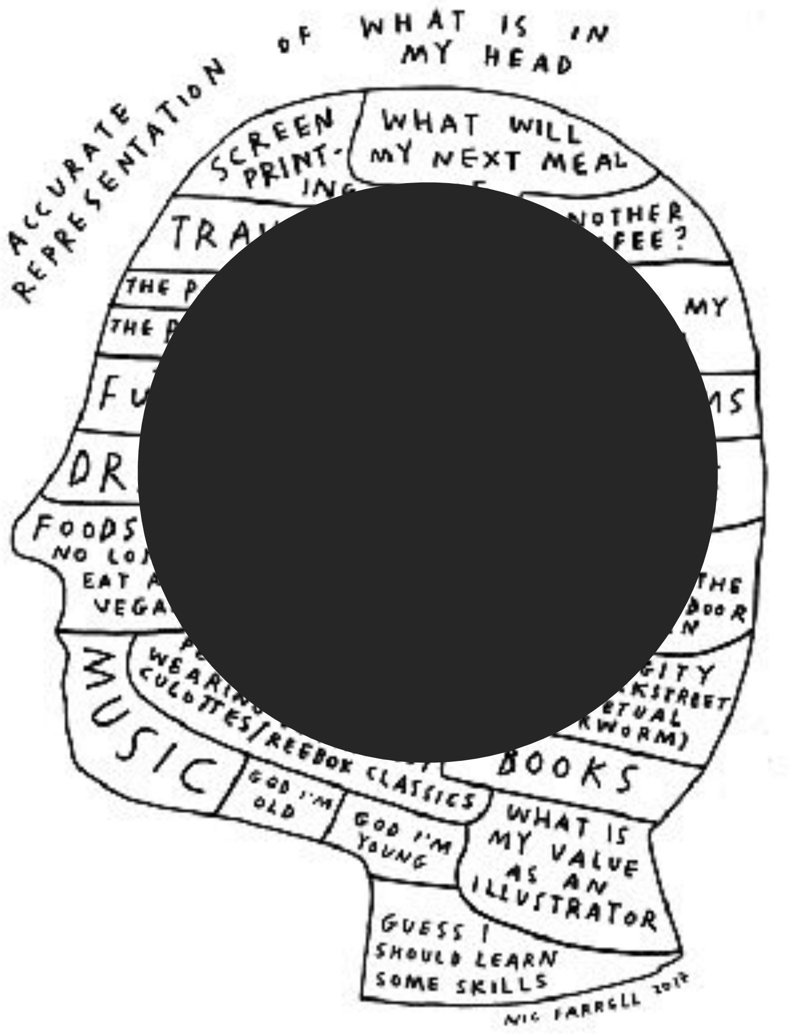
Can others know whether I have one too?

Actually, No.



We can only **presume** that others have one too, based on our observation on me.

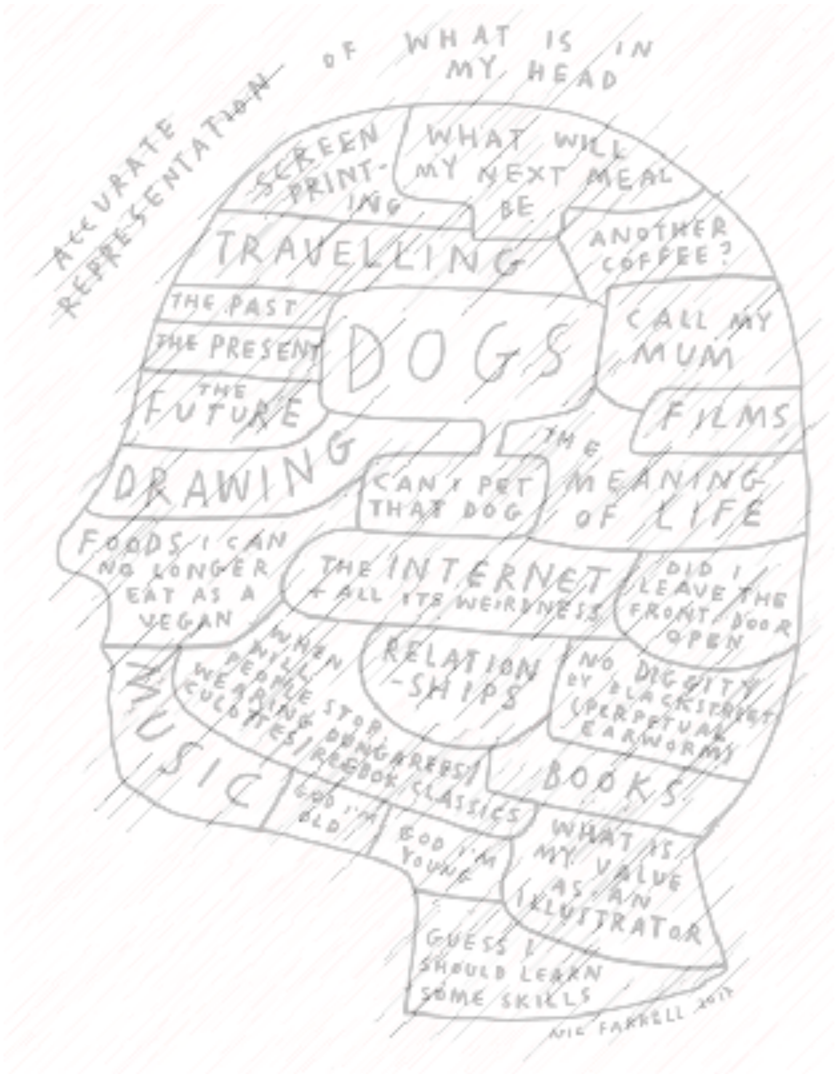
This is the ***Theory of mind*** that **we have**



Theory of Mind

the ability to reason about the mental states **of others**

e.g., desires, beliefs, intentions, etc.



Theory of Mind?

Are we saying machines have a mind?

No, they do not have minds, emotions, or intentions

However, they need social reasoning capabilities

What is theory of mind/social cognition?

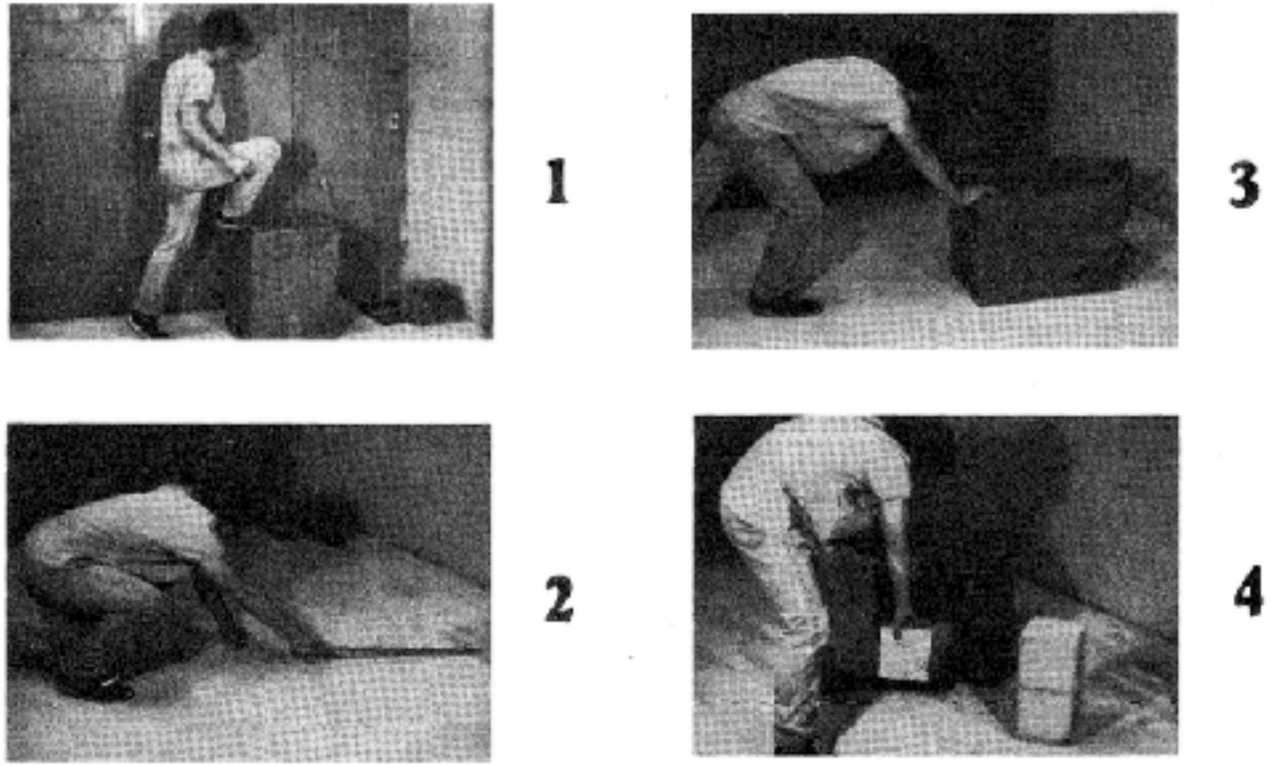
One of the most quintessential human mental function:

Thinking about each other's thoughts

- Our relationship with other people is the most crucial aspect of our lives
- Social cognition takes up a huge part of our reasoning
  - Every minute! Even right now
  - Social factors impacted the evolution of our intelligence

Origin of the term: ToM

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences, 1*(4), 515-526.



Development of ToM

Recognize that others have

1. Diverse desires
2. Diverse beliefs soon after
3. Access to different knowledge bases
4. May have False beliefs
5. Capability of hiding emotions

Development of ToM

Recognize that others have

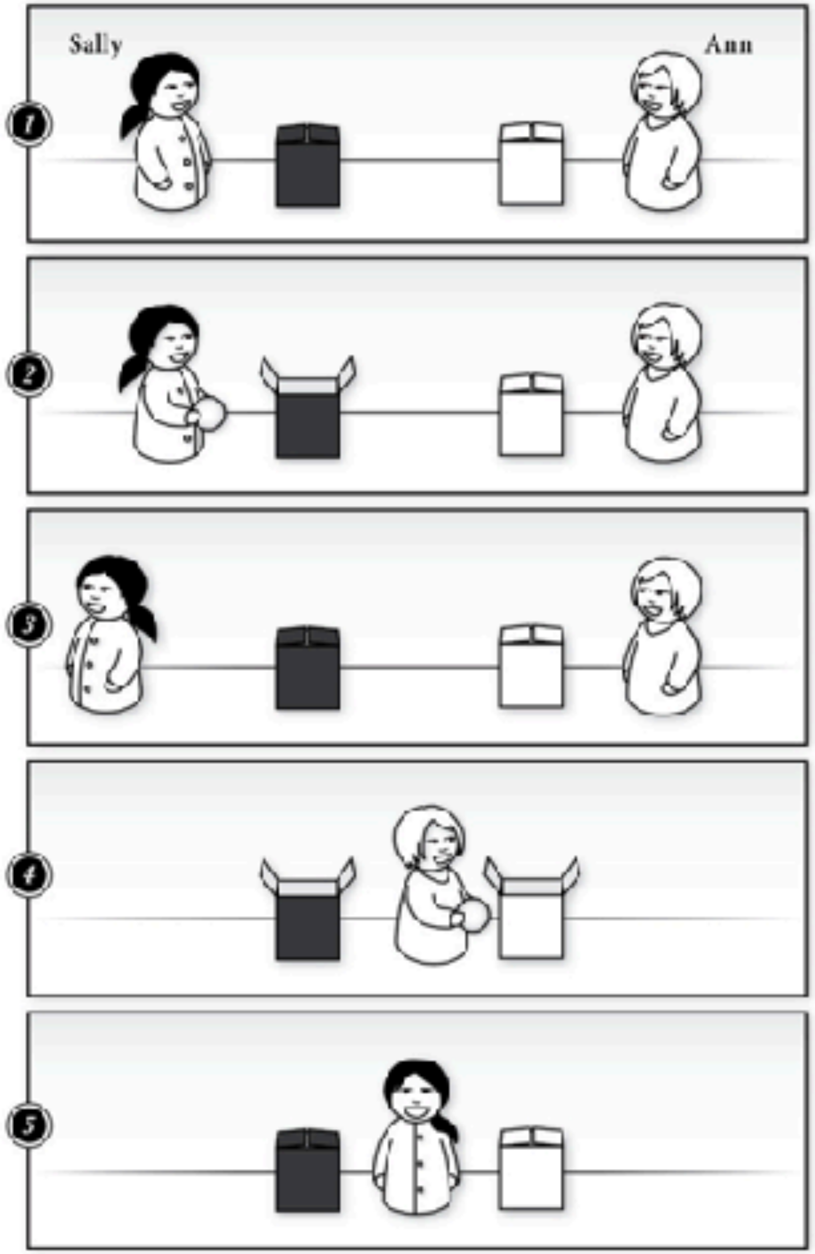
1. Diverse desires
2. Diverse beliefs soon after
3. Access to different knowledge bases
4. **May have False beliefs**
5. Capability of hiding emotions

The Sally-Anne test

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

1. Sally has a black box and Anne has a white box.
2. Sally has a marble. She puts the marble into her box.
3. Sally goes for a walk.
4. Anne takes the marble out of Sally's box and puts into her box.
5. Sally comes back and wants to play with her marble.

Question: Where will Sally look for her marble?



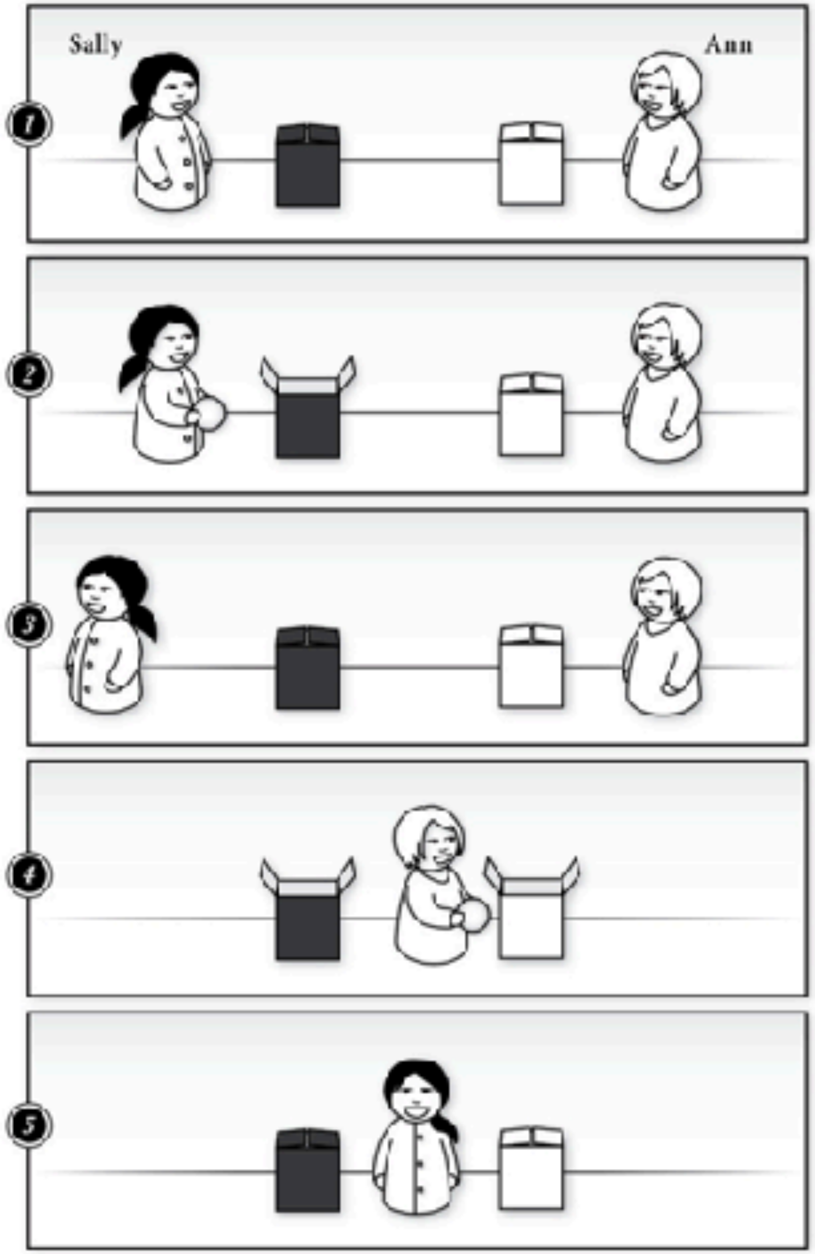
The Sally-Anne test

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Question: Where will Sally look for her marble?

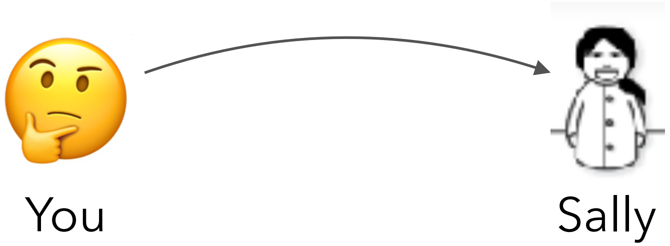
- Before the age of 4: Sally will look for it in Anne's box
- By the age of 4: Sally will look for it in her box

By the age of 4, children begin to understand that others may have ***false beliefs***

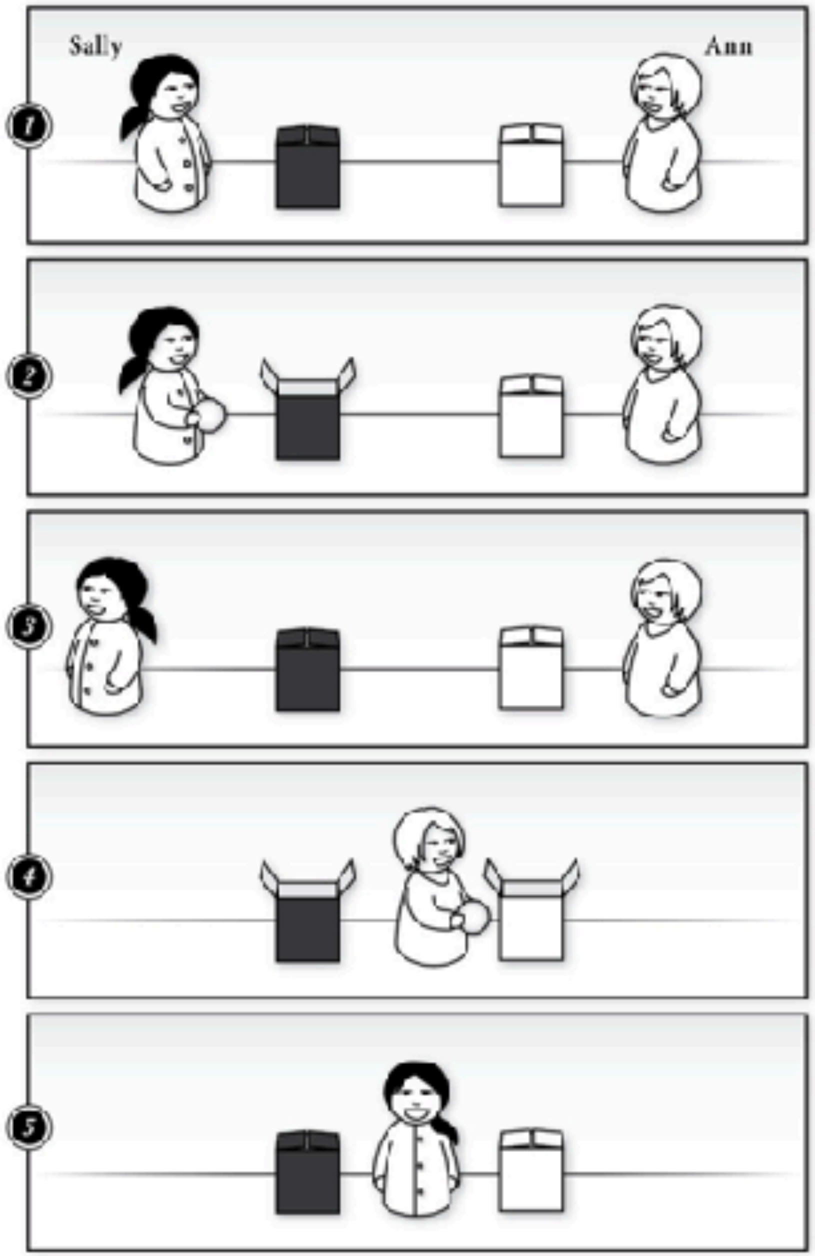


Order of ToM

Where will Sally think her marble is?



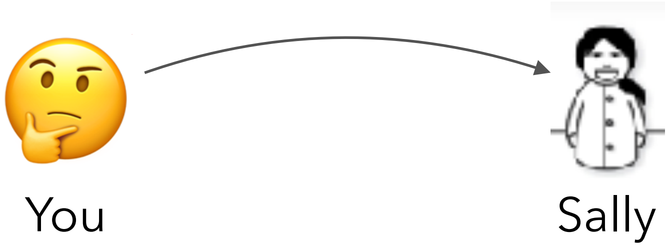
First-order



Order of ToM

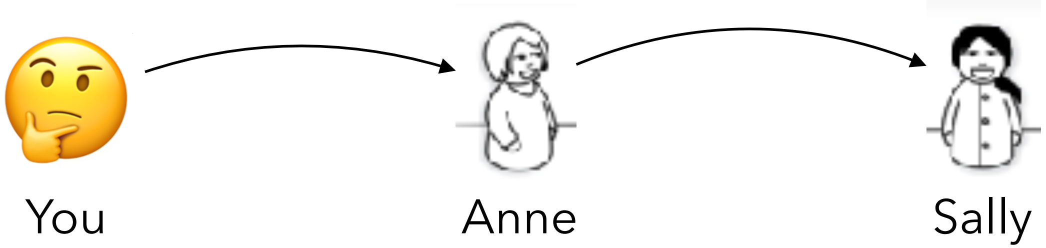
Where will Sally think her marble is?

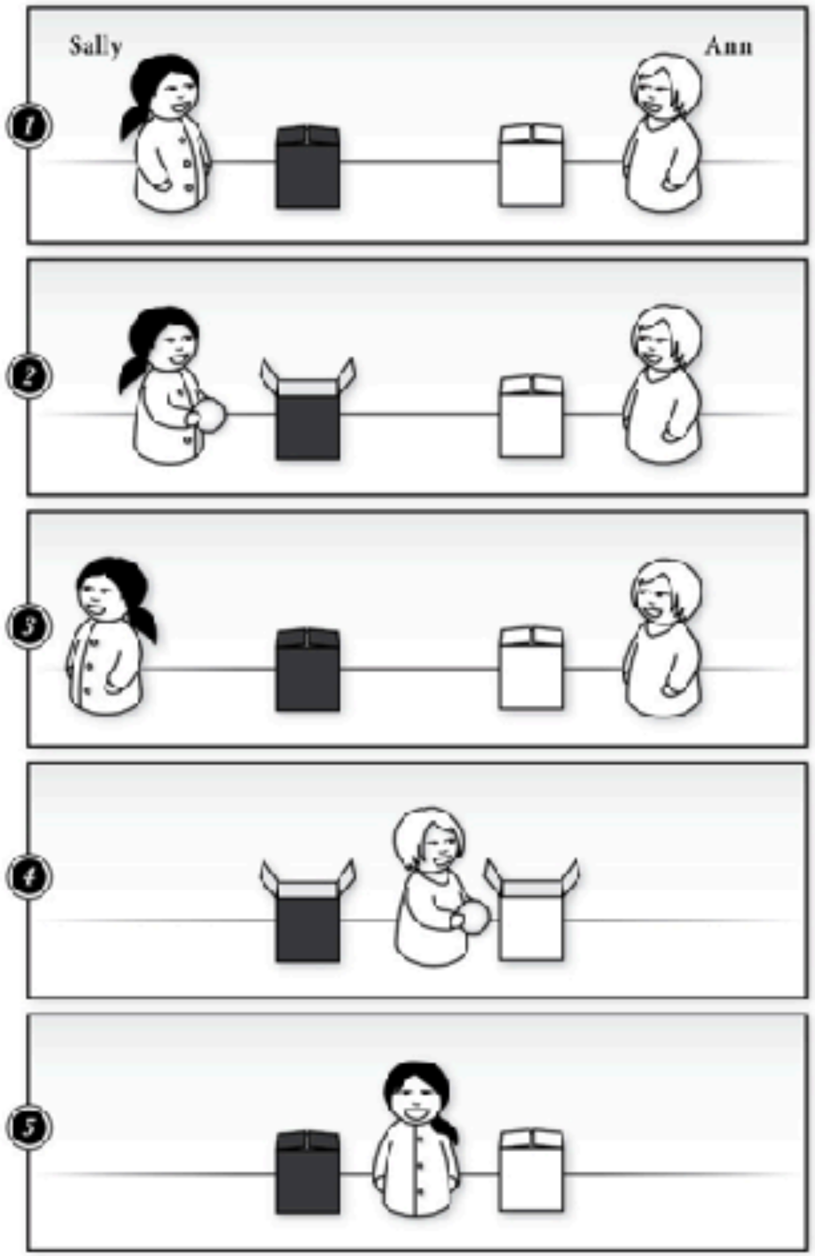
First-order



Where will Anne think
Sally thinks her marble is?

Second-order





Ice cream Van test for testing second-order ToM

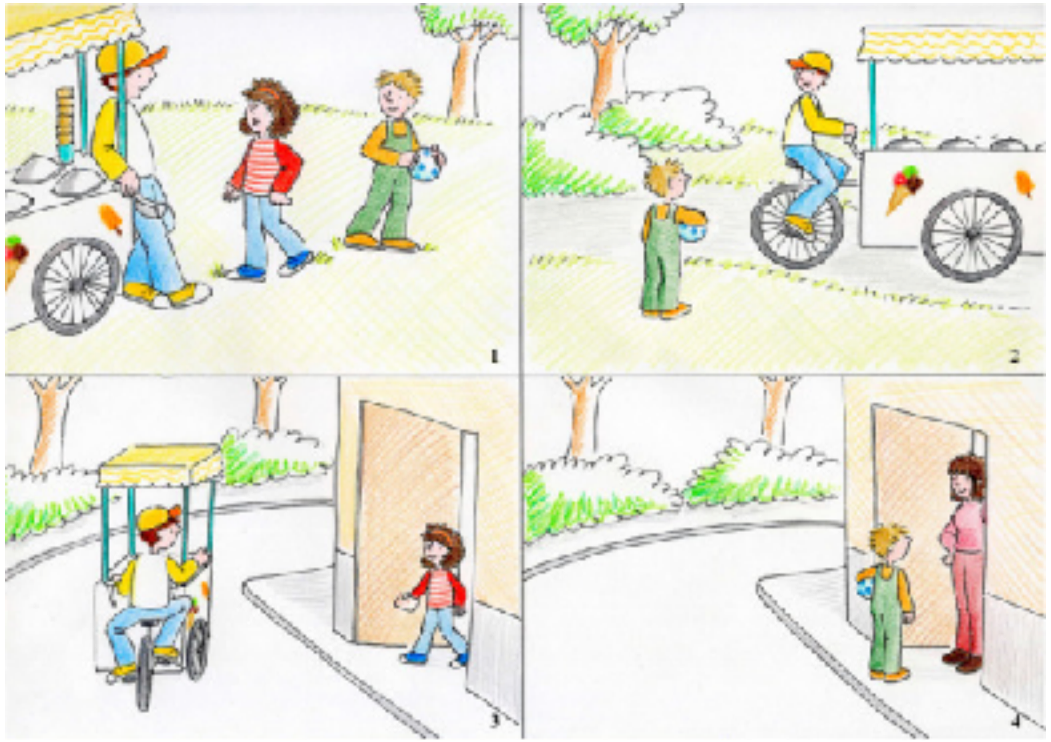
Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology, 39*(3), 437-471.

1. John and Mary sees an ice cream van. Mary wants to buy the ice cream but she forgot her money at home. The ice cream man tells them that he will be here in the park. Mary heads off to her home.

2. The ice cream van leaves and tells John that it will be in front of the church.

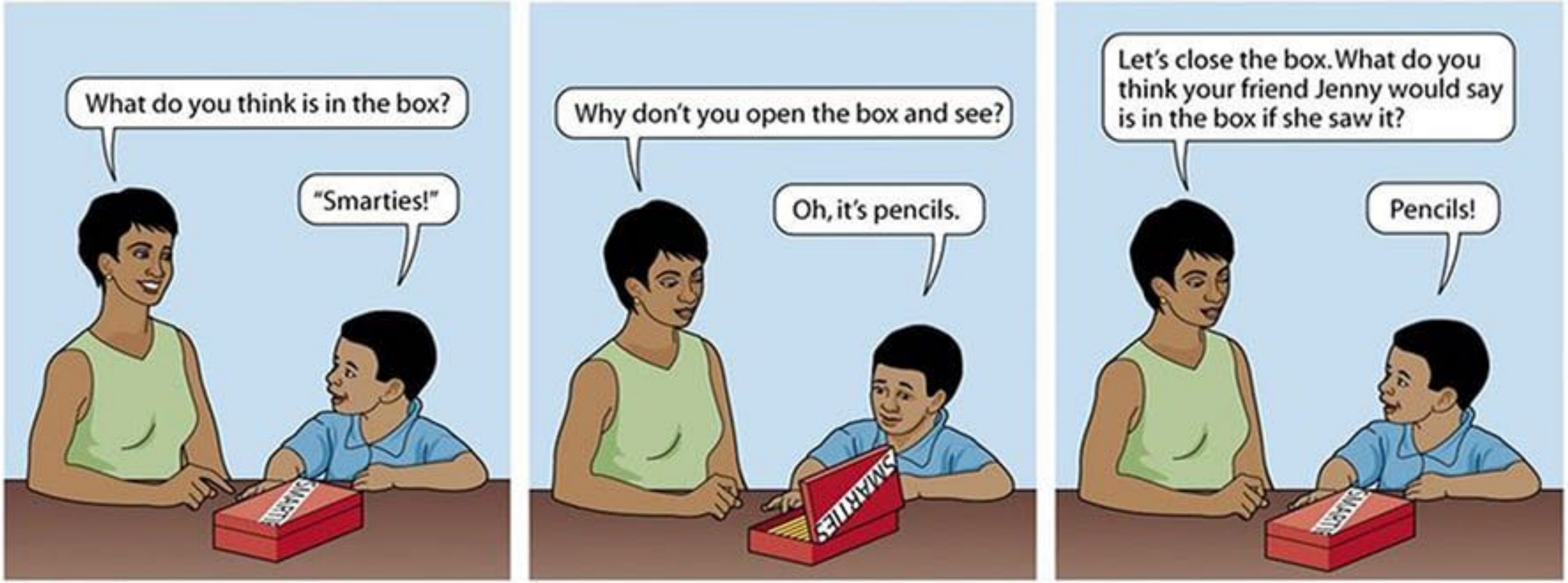
3. When Mary was leaving her house with the money, she coincidently bumps into the ice cream van. The ice cream man tells her he is heading to the church.

4. John later comes to Mary's house and finds out Mary already left to buy the ice cream.



Question: Where does John think Mary has gone to buy the ice cream?

The Smarties task



Gopnik & Astington, 1988

Detecting Faux Pas

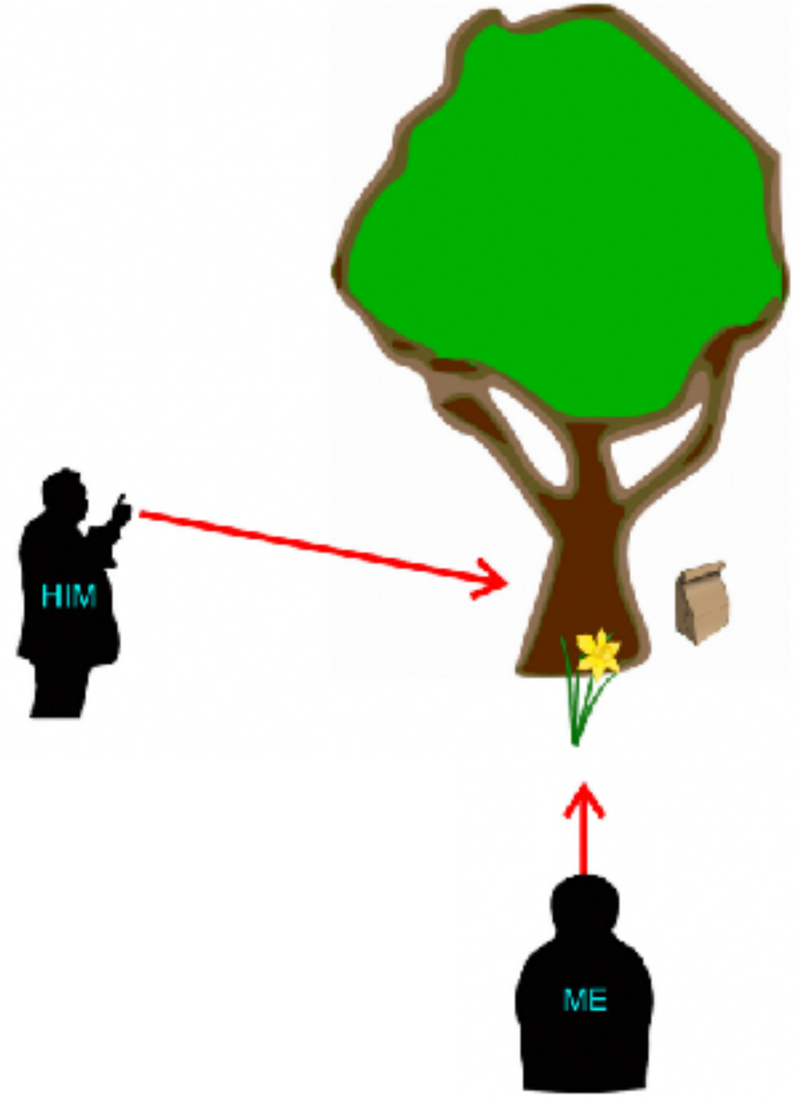
Mrs. West, the teacher, had something to tell her class, "One of the boys in our class, Simon, is very seriously ill" she said. The class were all very sad and were sitting quietly when a little girl, Becky, arrived late. "Have you heard my new joke about sick people?" she asked. The teacher said to her "Sit down and get on with your work."

What did the teacher tell the class at the beginning of the story?
Did Becky know Simon was sick?

Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999

Visuospatial perspective-taking, VPT

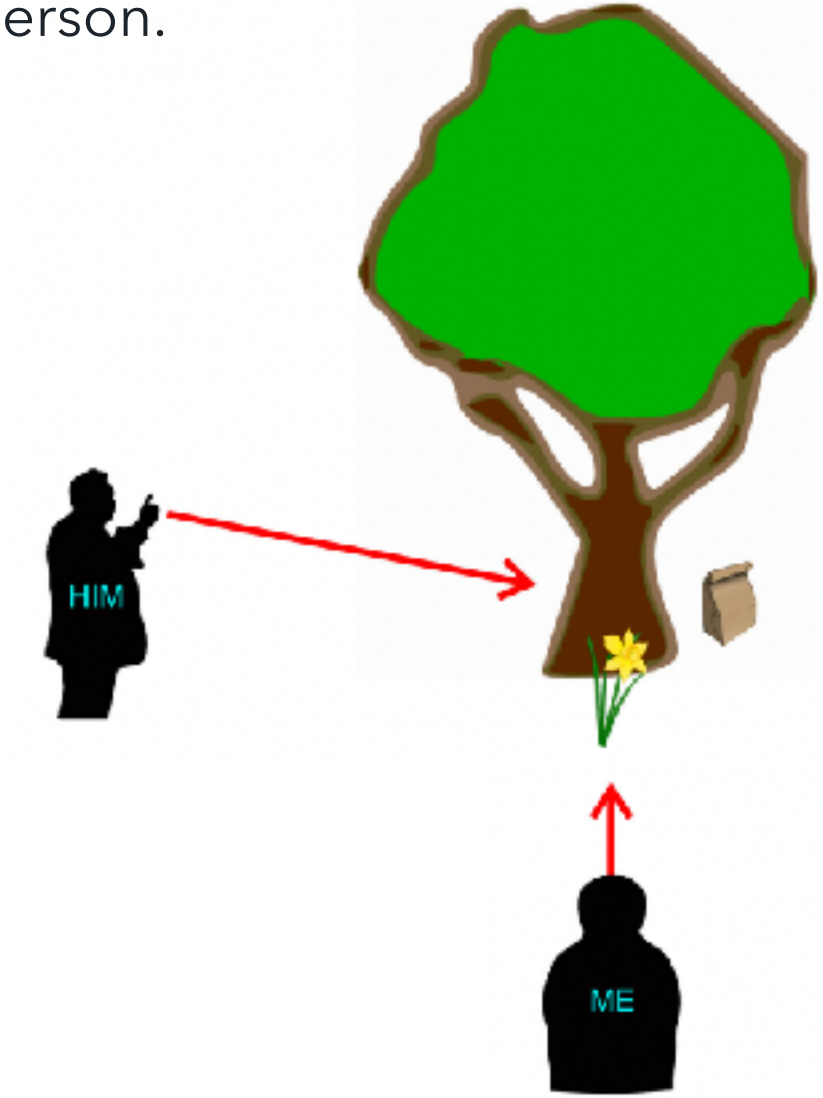
Q: Can he see the bag?



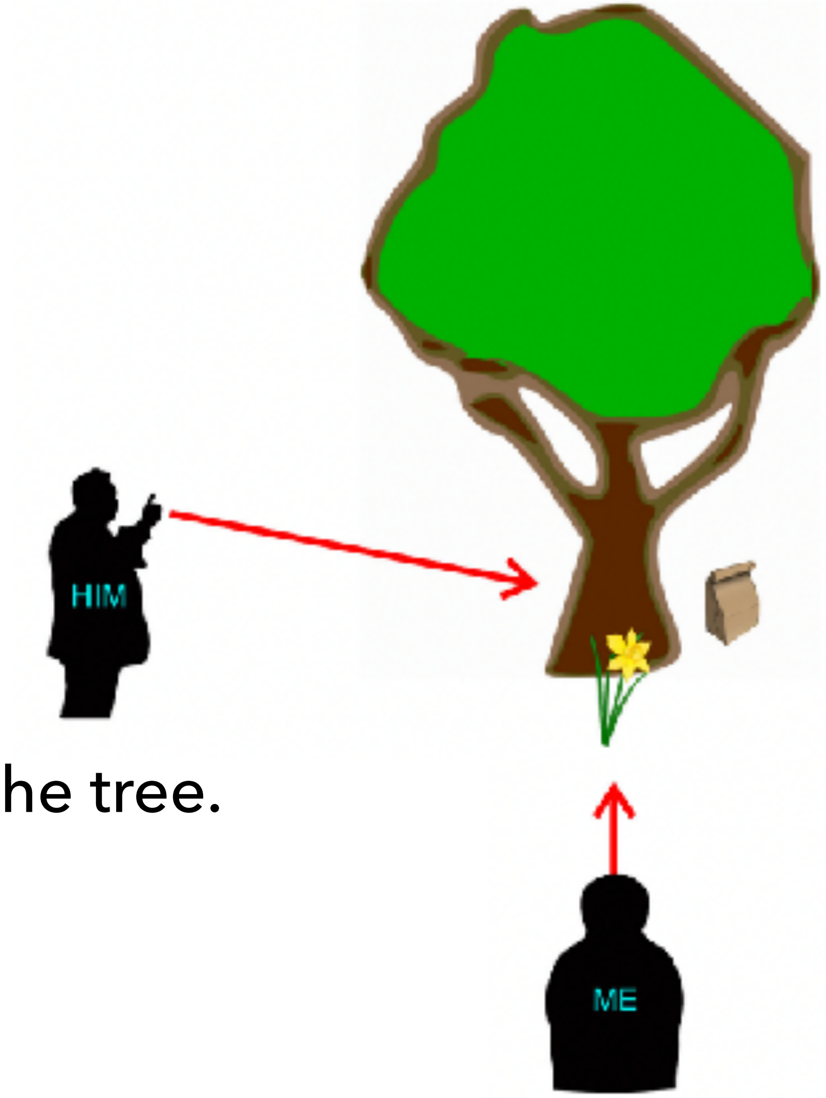
Visuospatial perspective-taking, VPT

Representing what is and what is not visible to another person.

From his perspective, the bag is **NOT** visible.



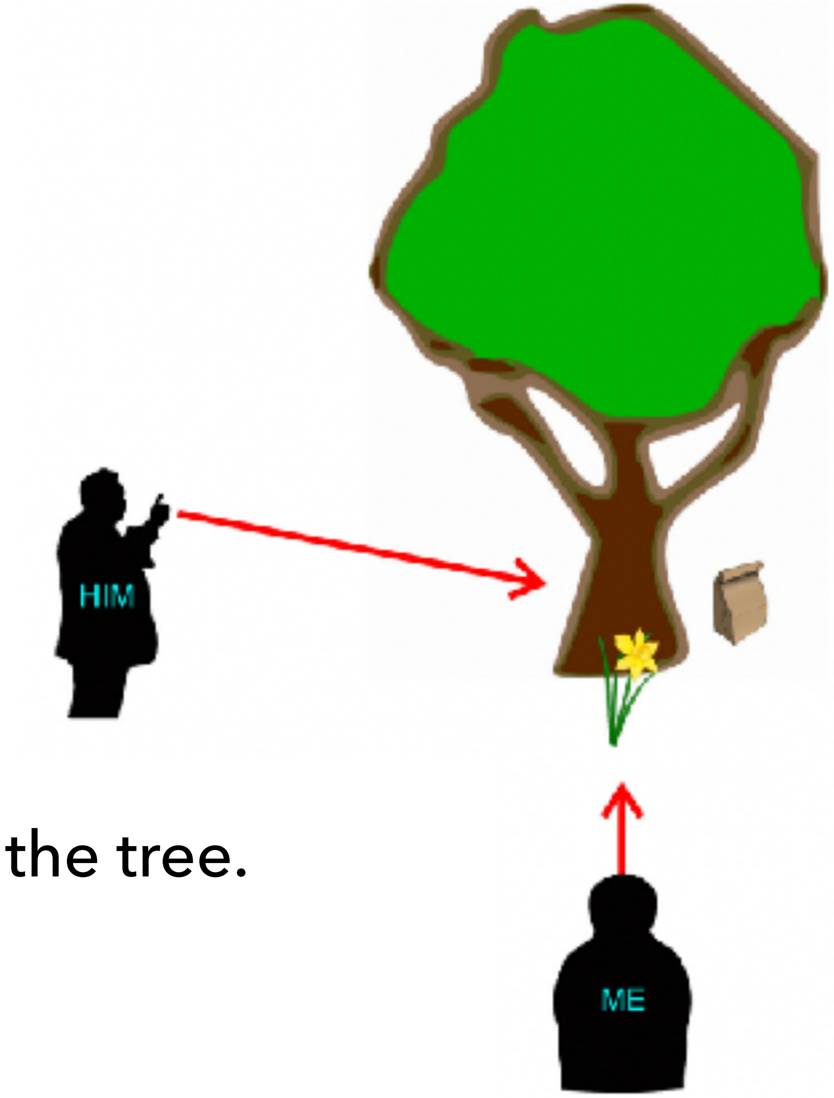
Visuospatial perspective-taking, VPT



From his perspective, the flower is on the **?????** of the tree.

Visuospatial perspective-taking, VPT

Mentally adopting someone else's spatial point of view.

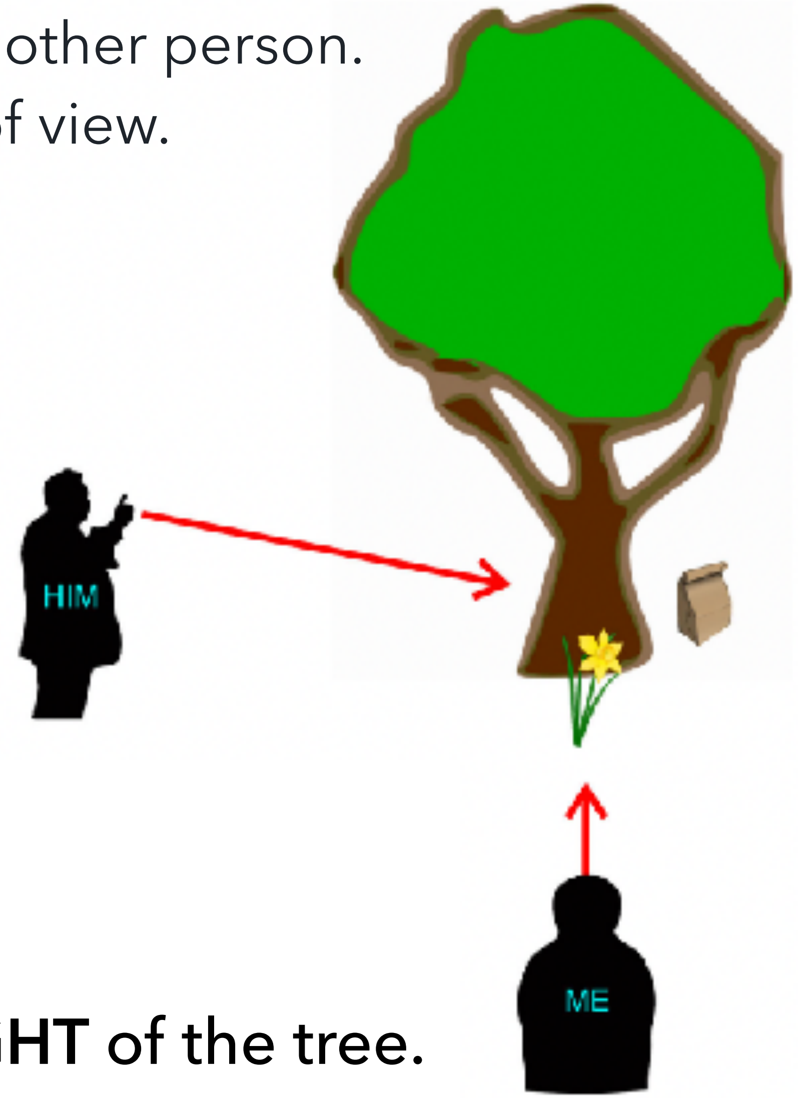


From his perspective, the flower is on the **RIGHT** of the tree.

Visuospatial perspective-taking, VPT

Level 1: Representing what is and what is not visible to another person.

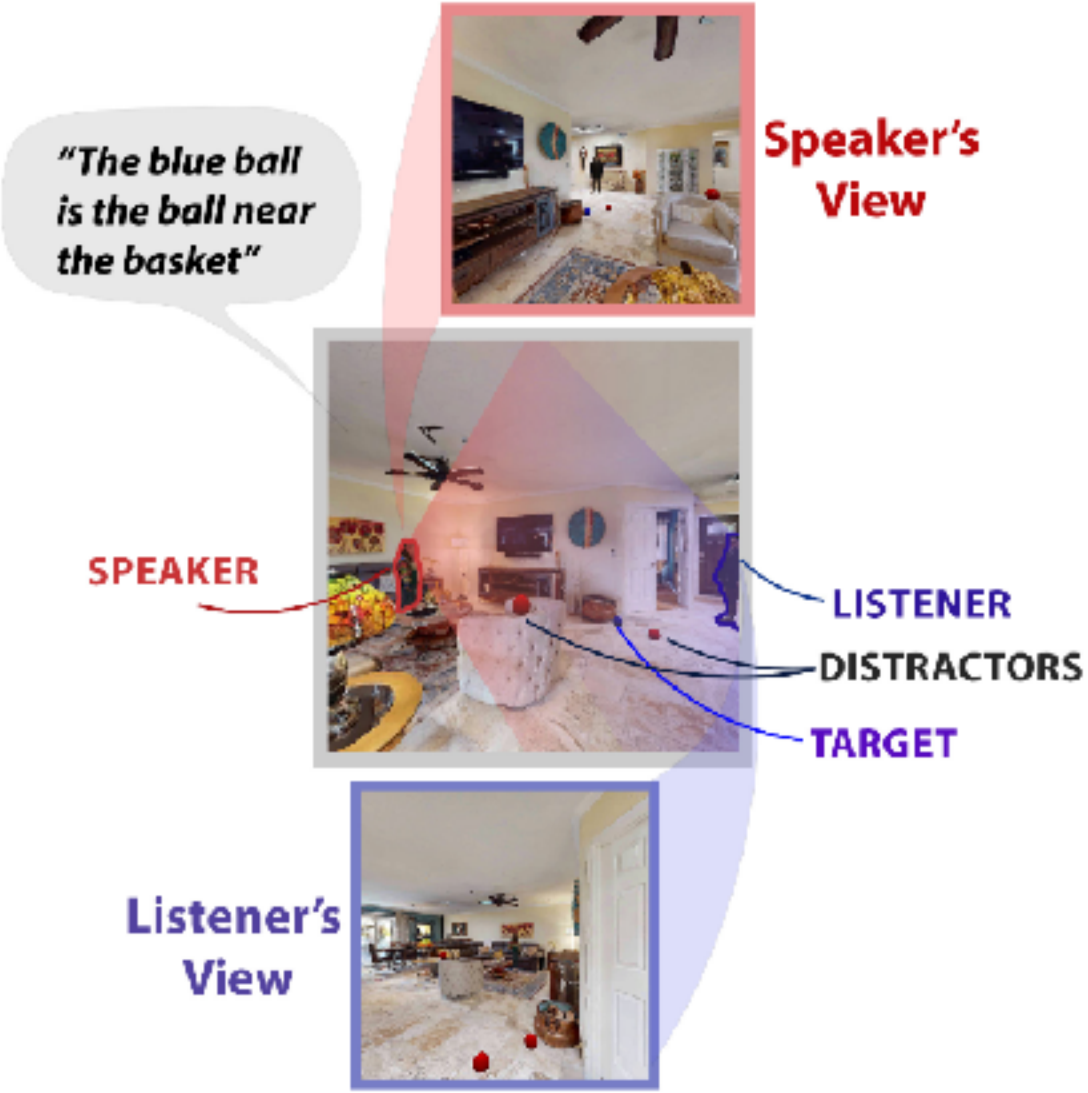
Level 2: Mentally adopting someone else's spatial point of view.



VPT-1: From his perspective, the bag is **NOT** visible.

VPT-2: From his perspective, the flower is on the **RIGHT** of the tree.

Multi-Perspective Reference Games



Tang et al. 2024

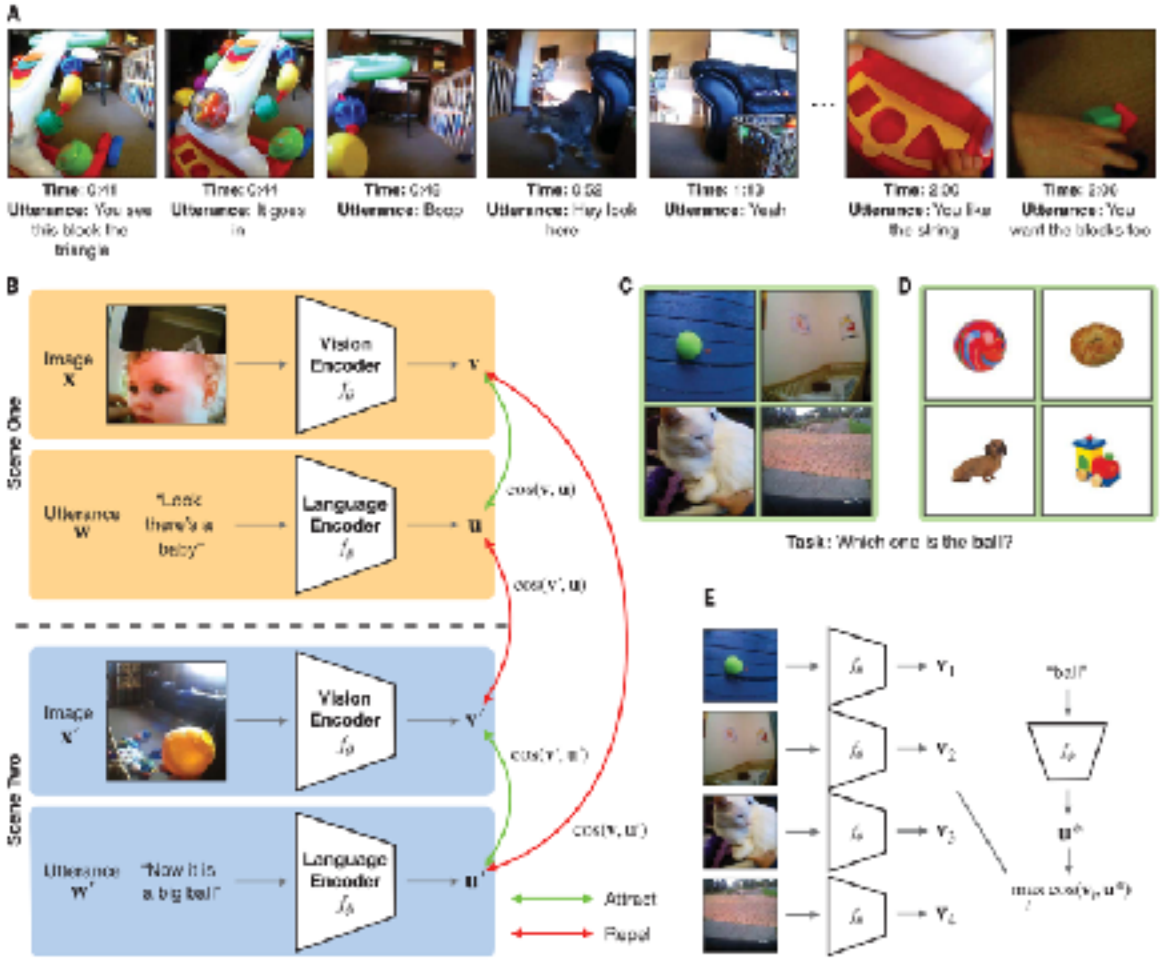
Word Learning through Interaction

- Supervised learning cannot explain how children learn 14,000 words by the age of 6
- What data do children learn from?
  - Situational contexts of when it is used by an adult
  - Linguistic contexts (e.g. syntax) of new words
- Fast mapping:
  - Quickly learning a new lexical item and pairing it with a new abstract concept
- Full mapping:
  - Fully accurate representation of word and meaning
  - Slow process, but children are doing this for ~1,600 words simultaneously

Carey 1978, slides also from CS 294-258 Spring 2024

Word Learning through Interaction

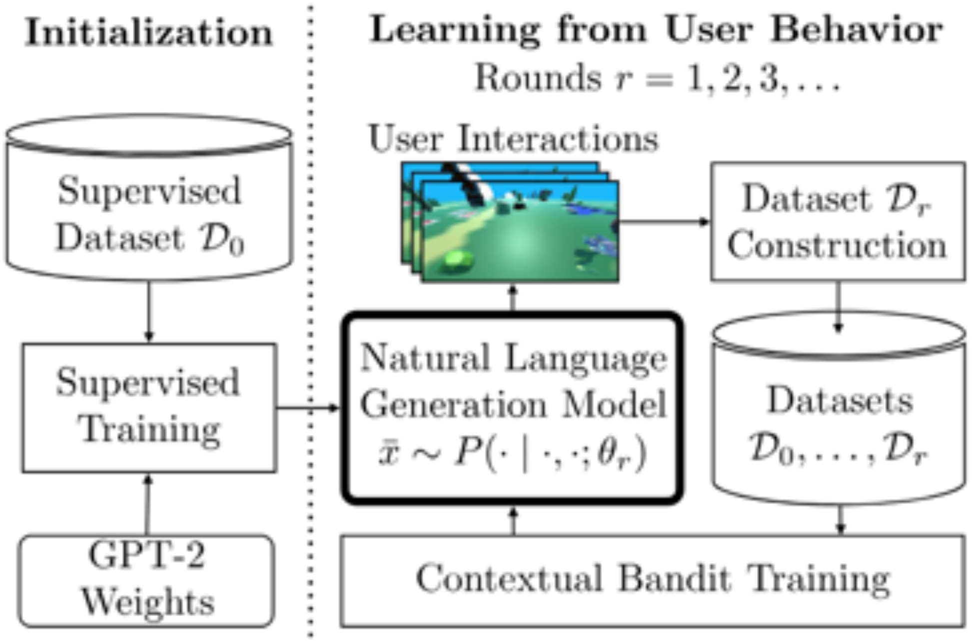
- Principle of mutual exclusivity
- New words have new meanings
- ~1:1 mapping between meaning and form



Vong et al. 2024, slides also from CS 294-258 Spring 2024

Learning through Communicative Success

- Regardless of original speaker intent, a listener's behavioral response tells us something about their language use
- Main principle: learn from observation

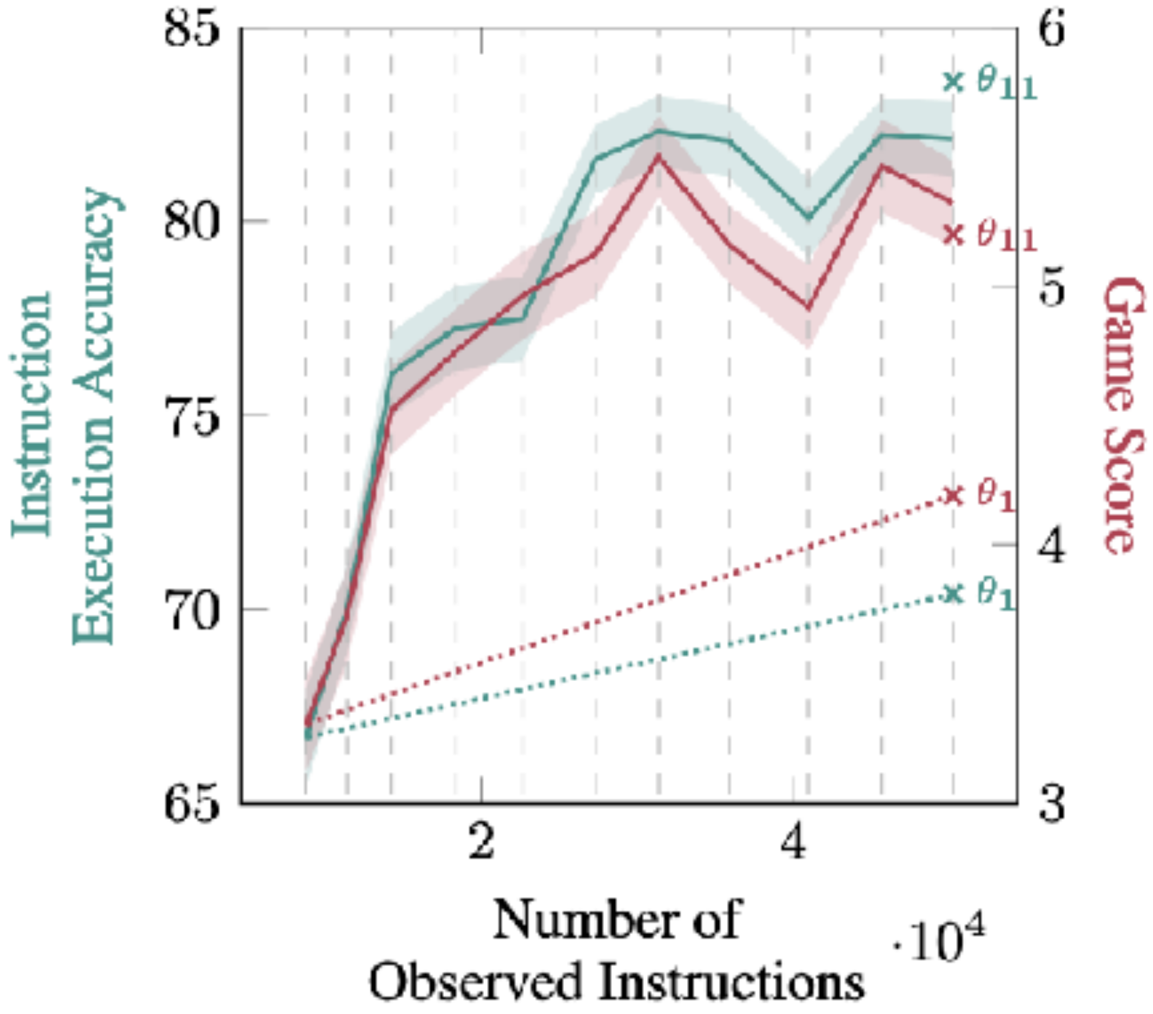


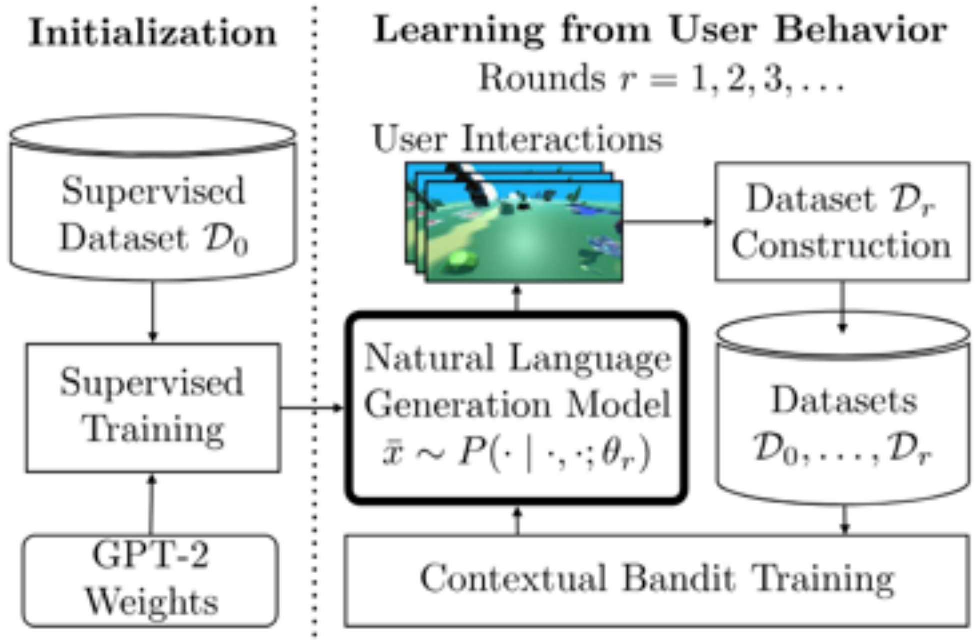
Speaker	Listener Accuracy Val.	Listener Accuracy Test	Avg. Ref. Length
Pre-trained θ	59.7	58.9	61.1
+ Contrastive ($\mathcal{D}_a$)	60.9	–	45.8
+ Contrastive ($\mathcal{D}_h$)	62.1	–	55.7
+ LSO ($\mathcal{D}_a$)	61.5	–	41.7
+ LSO ($\mathcal{D}_h$)	65.6	–	54.6
+ Pos. Only ($\mathcal{D}_a$)	62.1	–	46.7
+ Pos. Only ($\mathcal{D}_h$)	66.0	–	57.2
+ PPL ($\mathcal{D}_a$)	66.7	–	19.8
+ PPL ($\mathcal{D}_h$)	69.2	69.3	15.6
+ Imitation Learning	67.9	68.2	16.8
Human	91.3	90.6	15.8
GPT-4o	66.3	67.1	78.9

$$p_s(x \mid o_s, \mathcal{R}, \hat{t}; \theta') - p_s(x \mid o_s, \mathcal{R}, t; \theta')$$

Kojima et al. 2021,
Tang et al. 2024

Learning from Direct Feedback

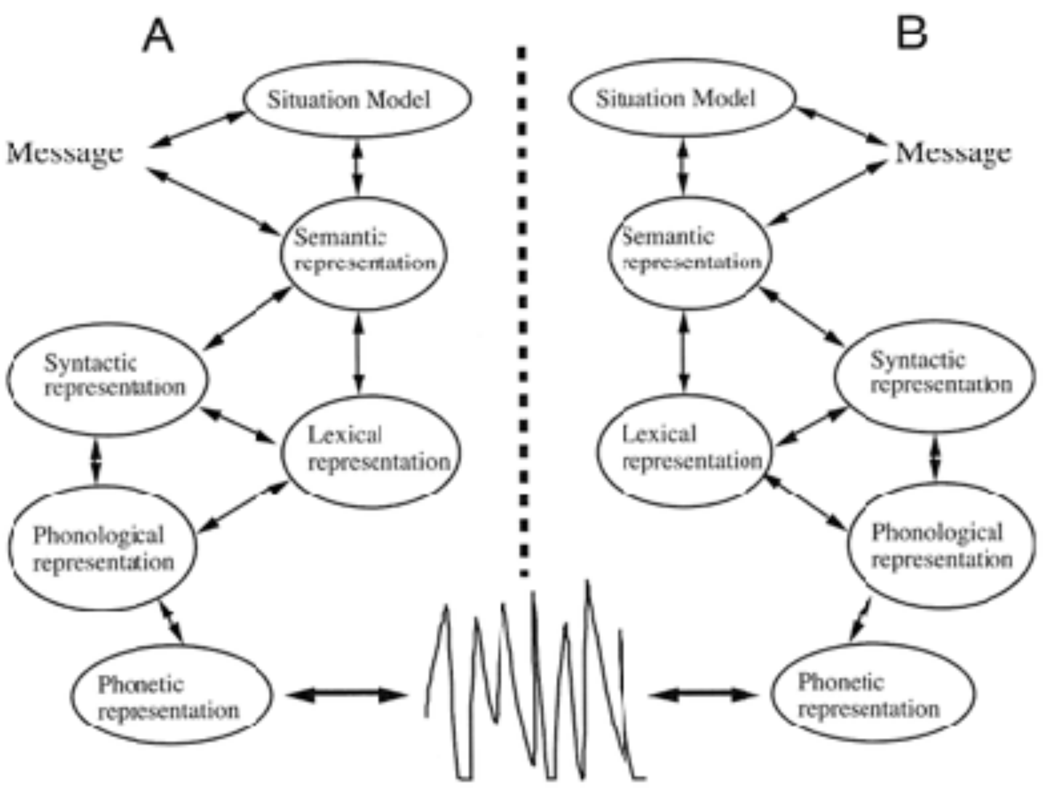


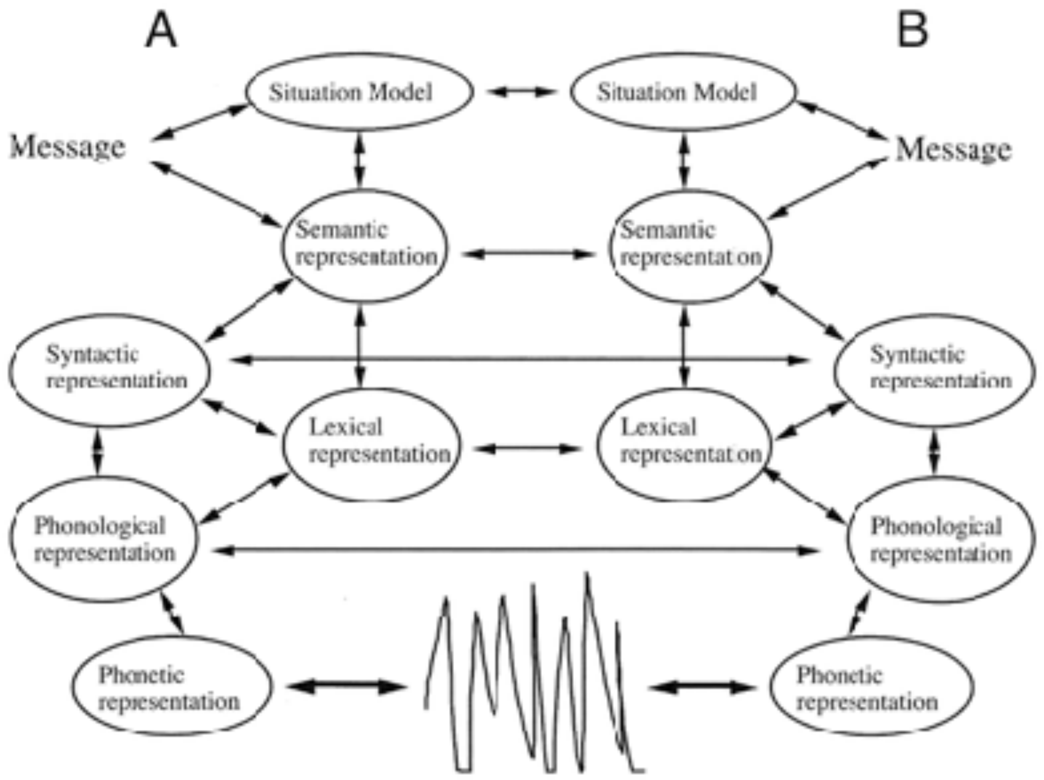


Suhr and Artzi 2023

Language Use in Real-Time Interaction

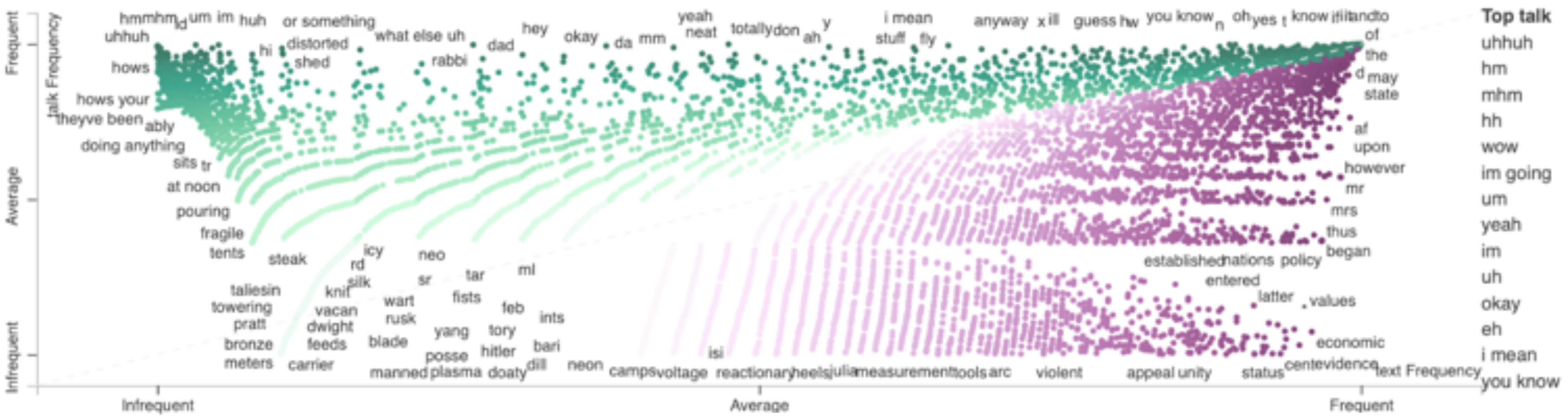
- Turn-taking and backchanneling
- Linguistic alignment

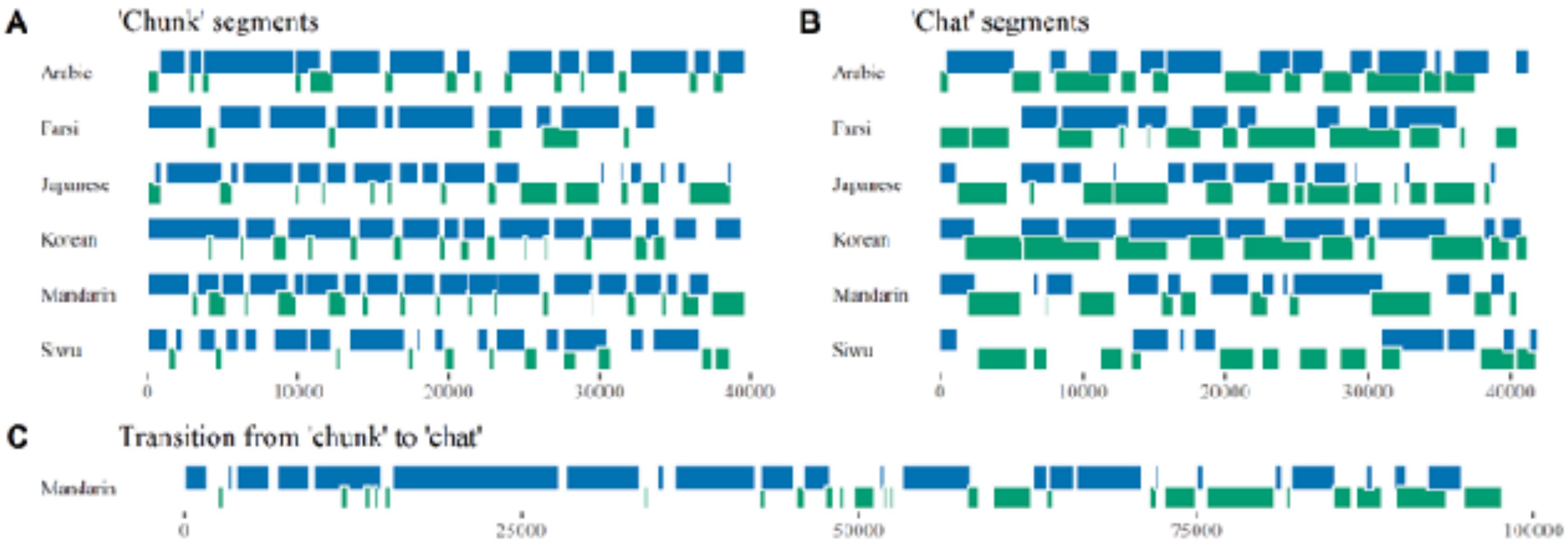




Sacks 1974, Ward and Tsukahara 2000, Pickering and Garrod 2004

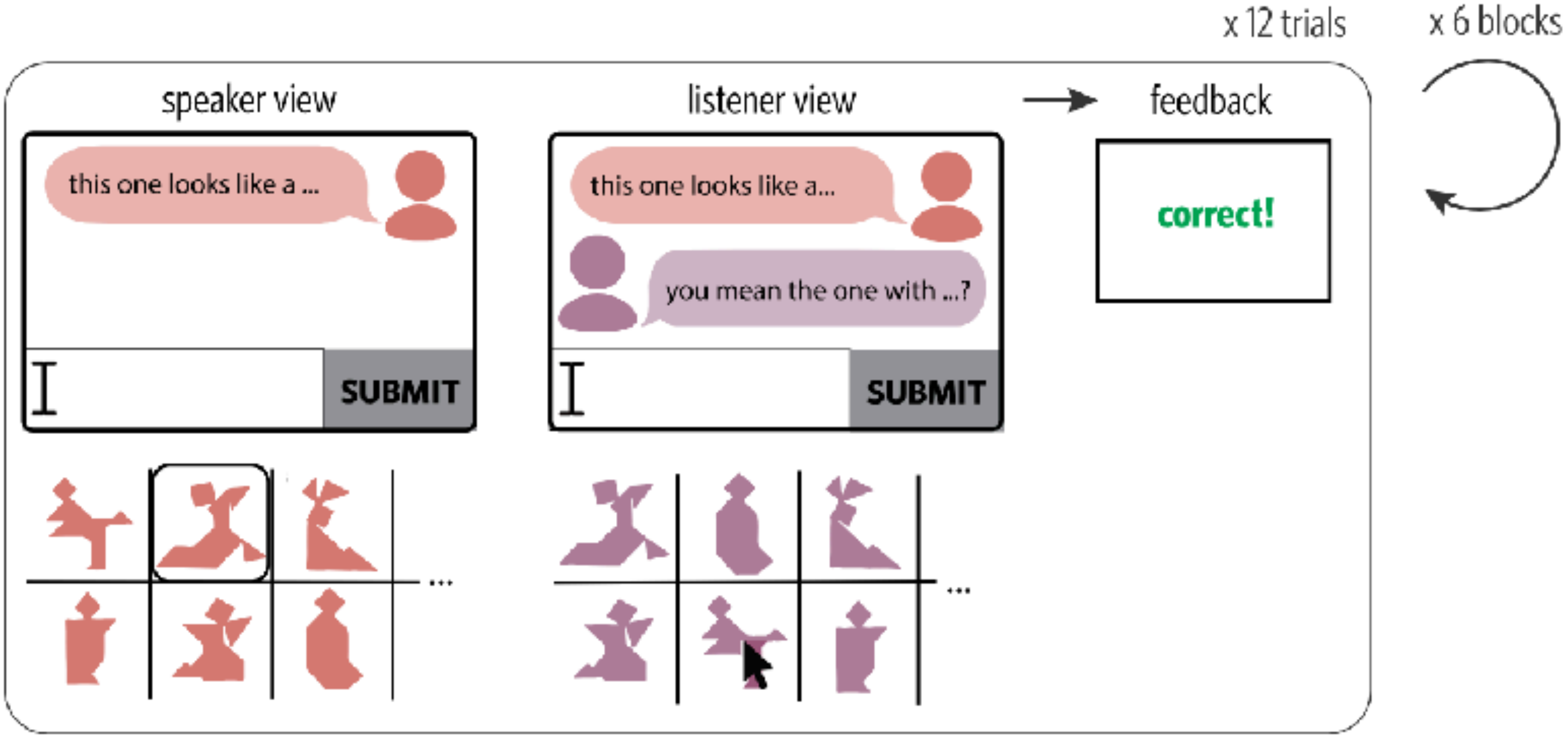
Language Use in Real-Time Interaction





Dingemanse and Liesenfeld 2022

Coordination through Language



Clark and Wilkes-Gibbs 1986, Hawkins et al. 2020

Coordination through Language

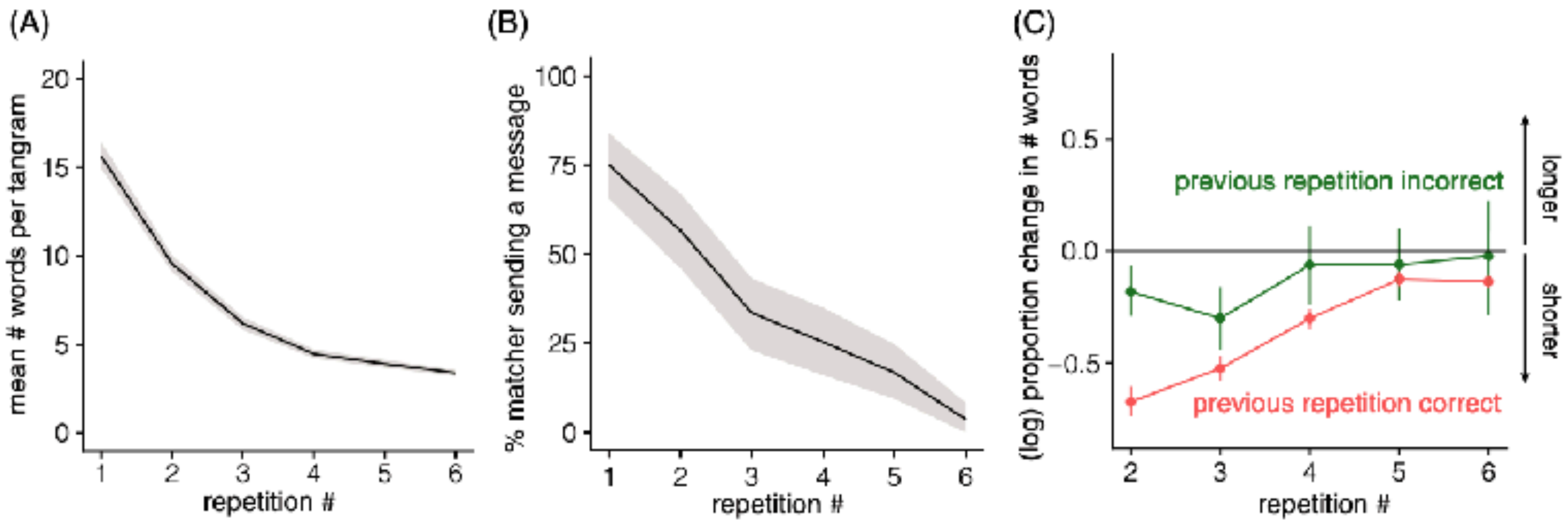
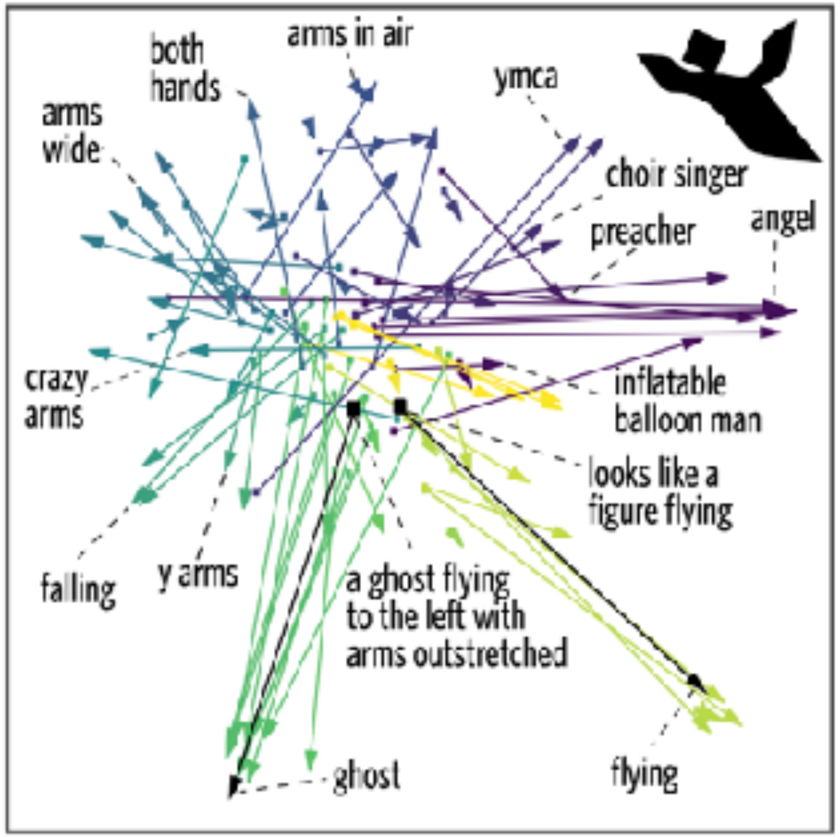


Fig. 2. (A) Directors use fewer words per tangram over time, (B) matchers are less likely to send messages over time, and (C) directors are sensitive to feedback from the matcher's selection, modulating the reduction in message length on the subsequent repetition of a tangram after an error is made.

Clark and Wilkes-Gibbs 1986, Hawkins et al. 2020

Coordination through Language

- Arbitrariness
- Stability



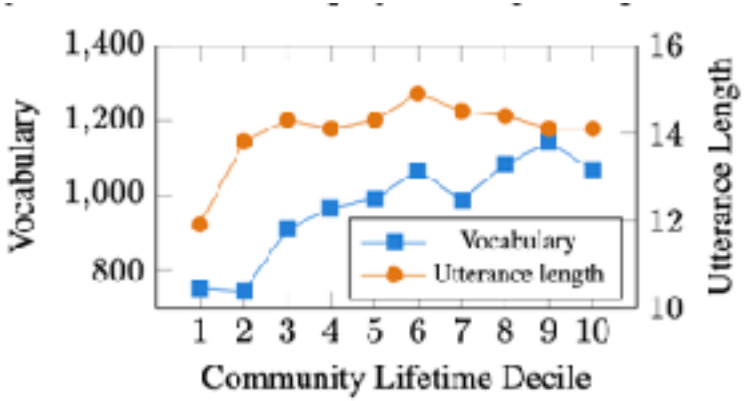


Figure 3: Vocabulary and utterance length over deciles.

Convention formation only happens when there is incentive for it!

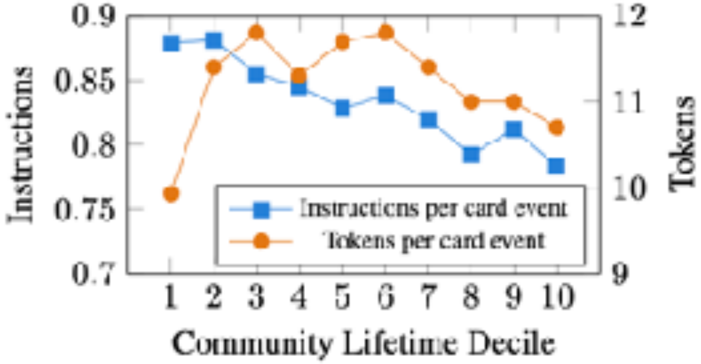


Figure 6: The number of instructions and tokens required for a card event over deciles. Analysis considers only instructions marked complete by the follower.

Clark and Wilkes-Gibbs 1986, Hawkins et al. 2020, Effenberger et al. 2021