# Linguistic scaffolds for policy learning

**Jacob Andreas**

Berkeley → Microsoft Semantic Machines → MIT

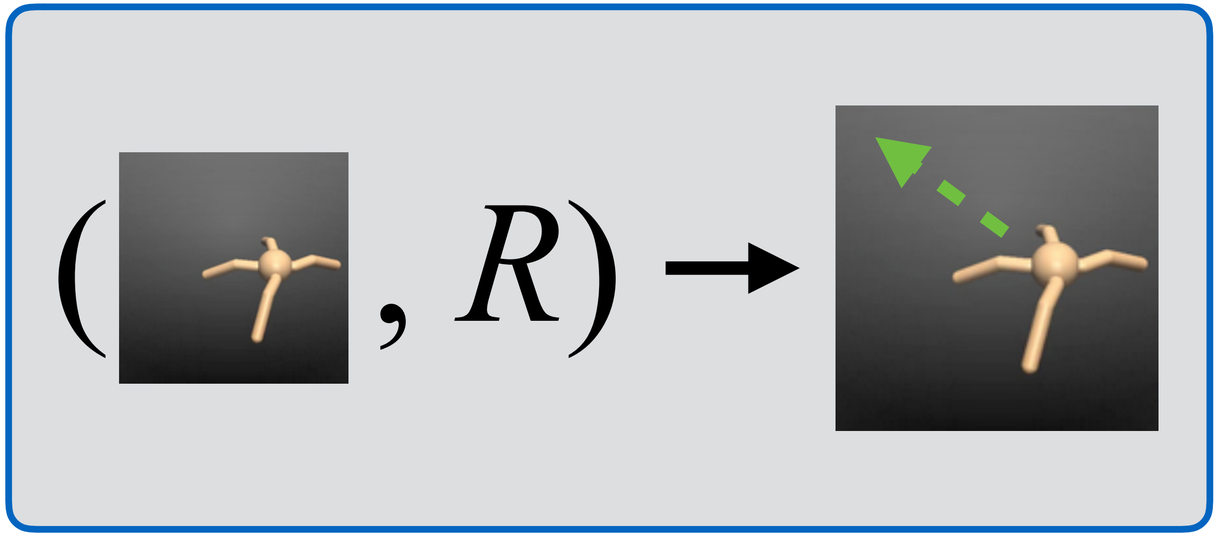
# Linguistic scaffolds for policy learning

# (what can language do for RL?)

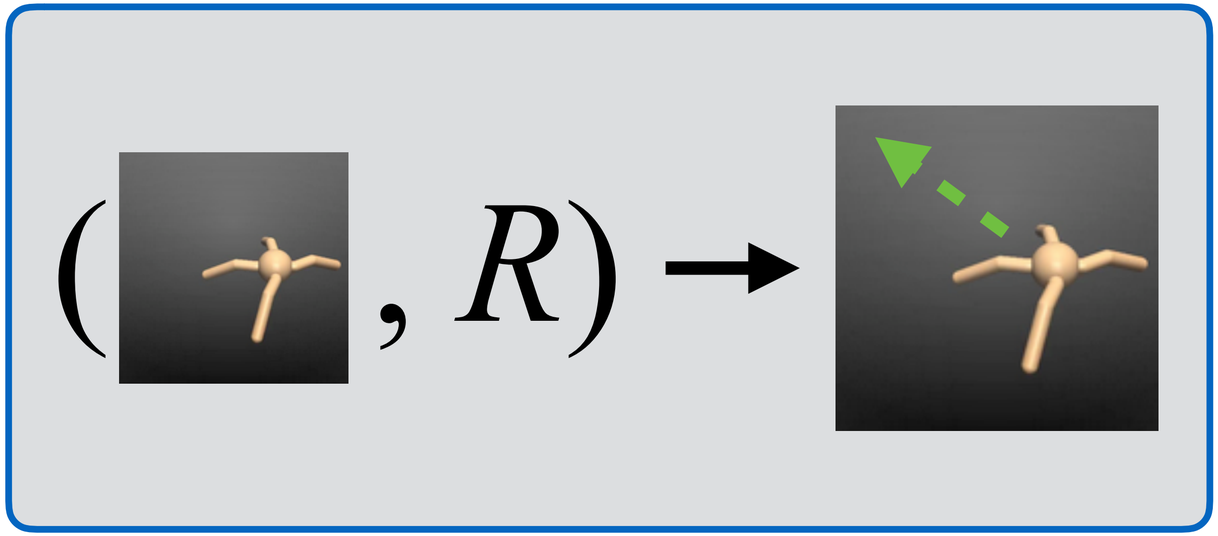
Jacob Andreas

Berkeley → Microsoft Semantic Machines → MIT

# An NLPer's view of RL

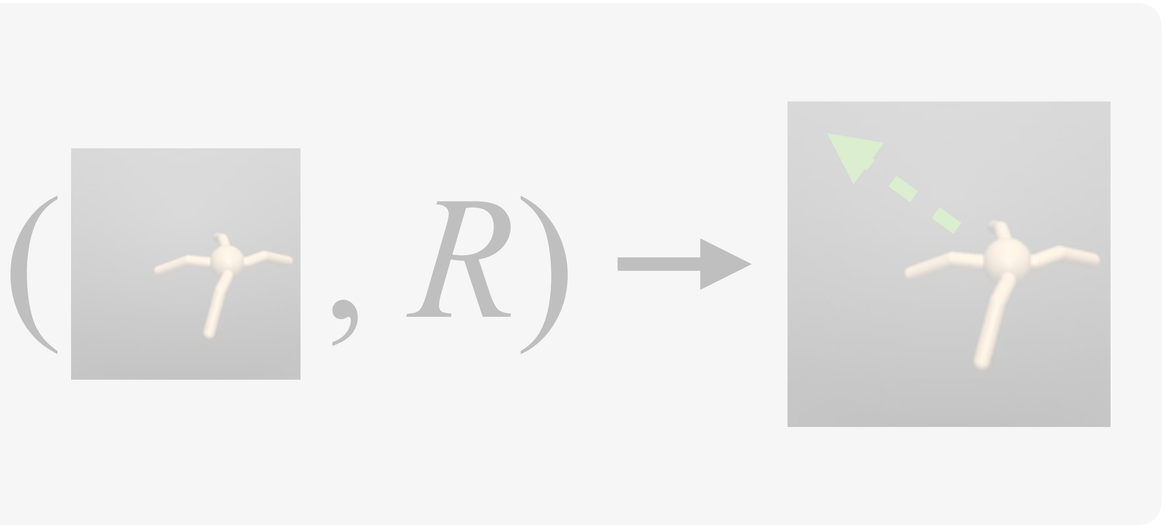


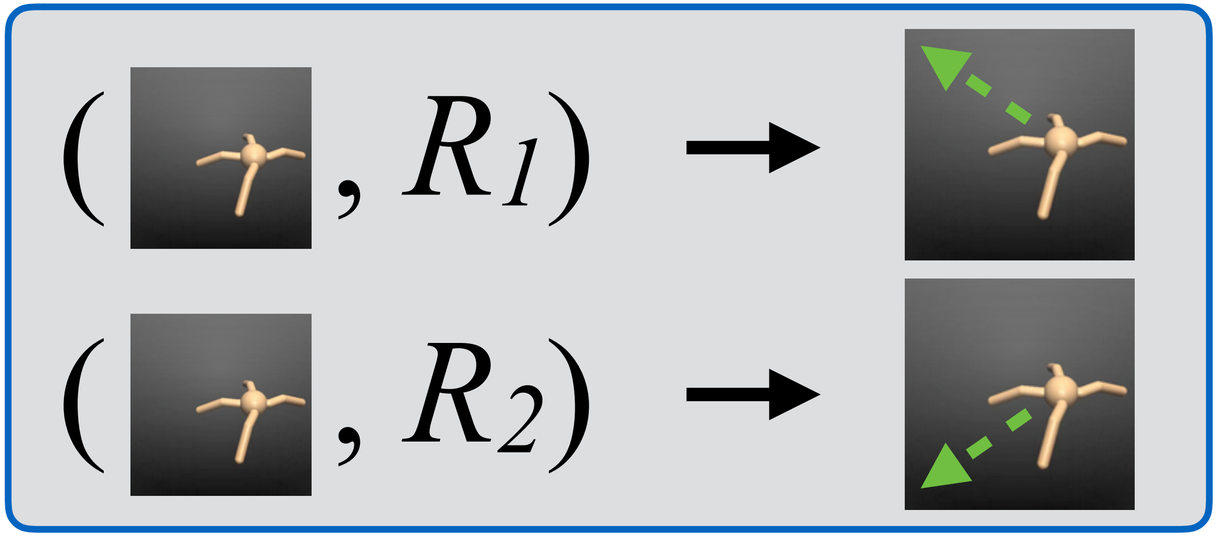
# An NLPer's view of RL




memorize 1 reward fn

# An NLPer's view of RL



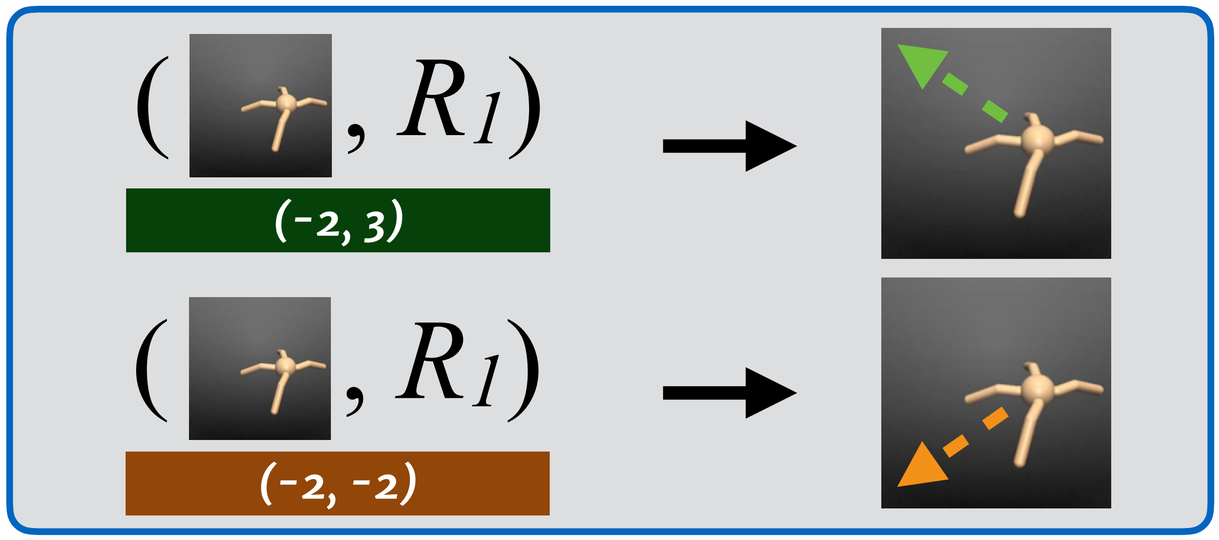




memorize k reward fns

[e.g. Taylor & Stone 09]

# An NLPer's view of RL

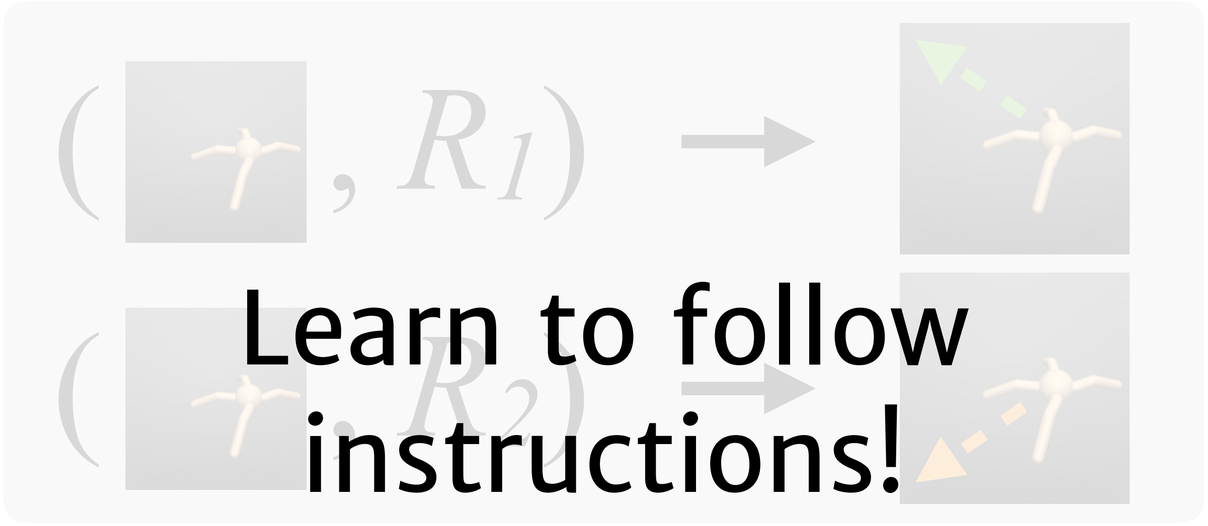
Learn to accomplish new goals!




[e.g. Schaul et al. 15]

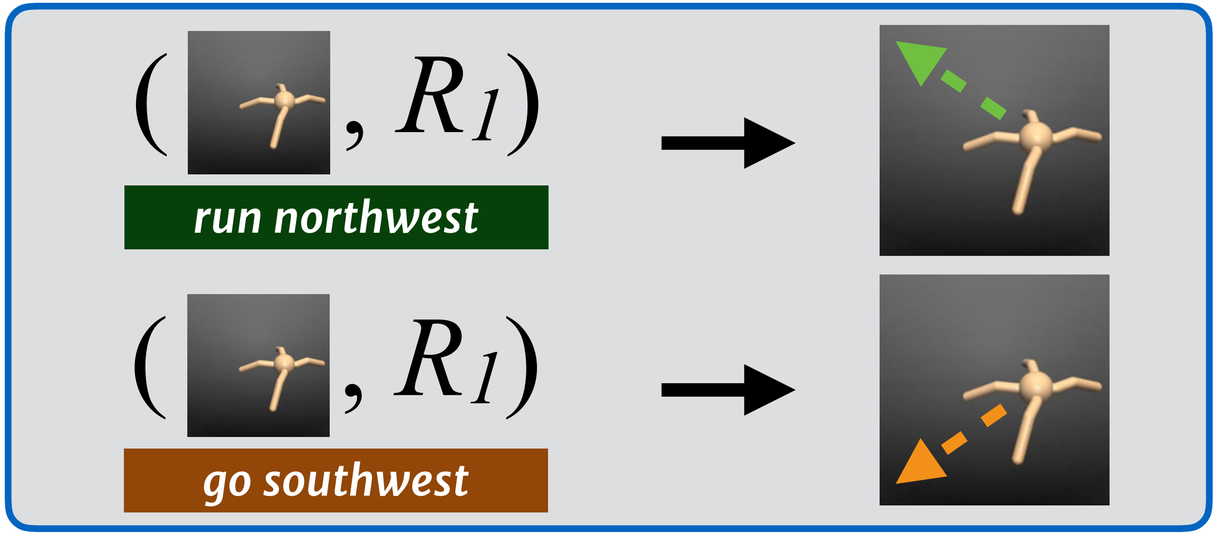
# An NLPer's view of RL



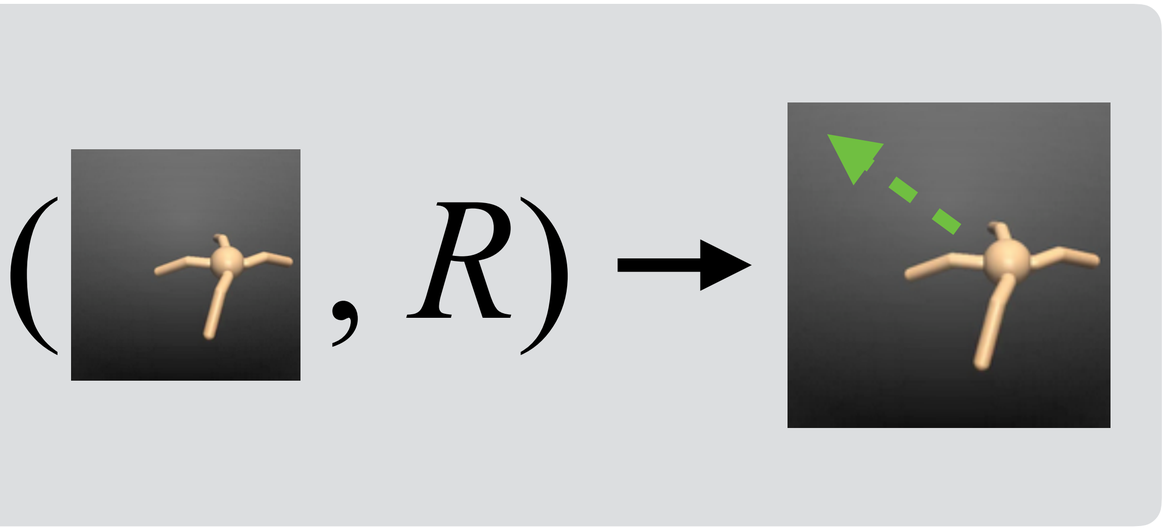


Learn to follow instructions!

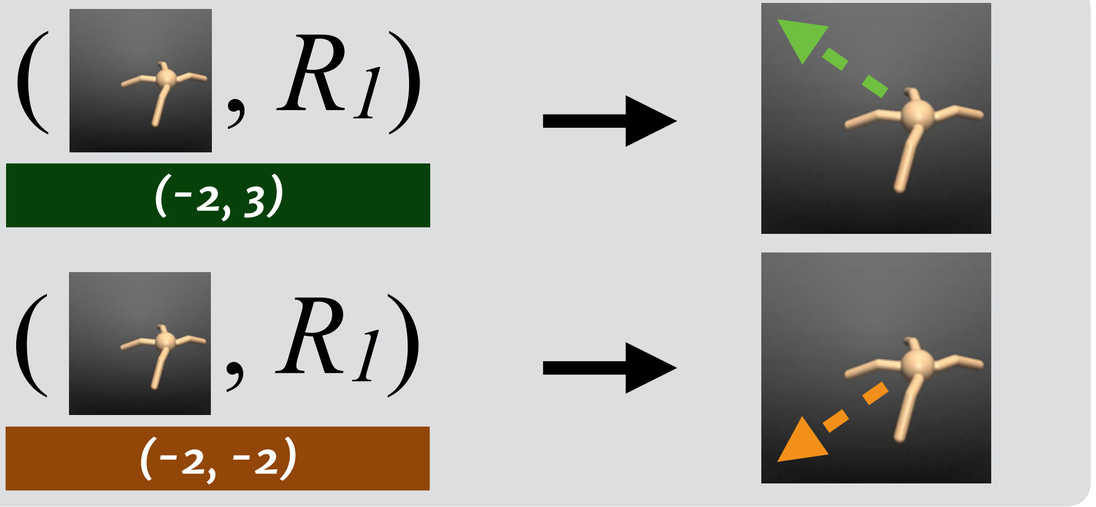


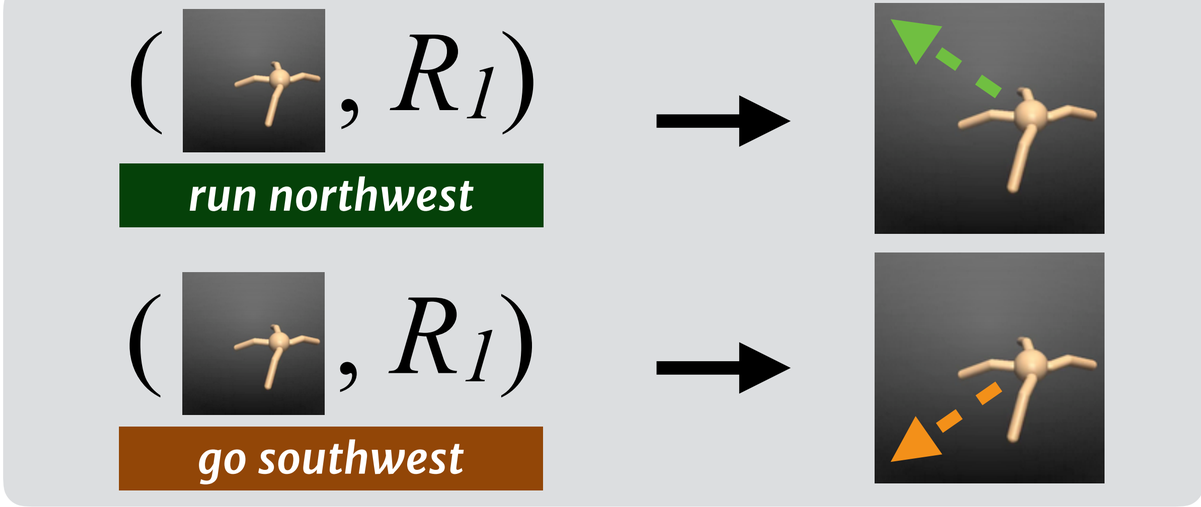



# Instructions as observations

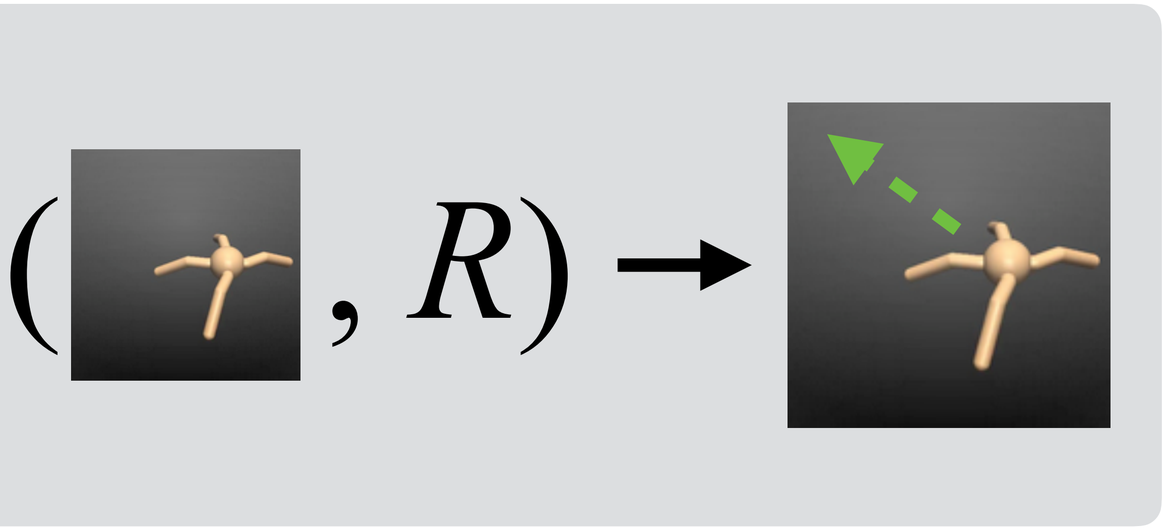




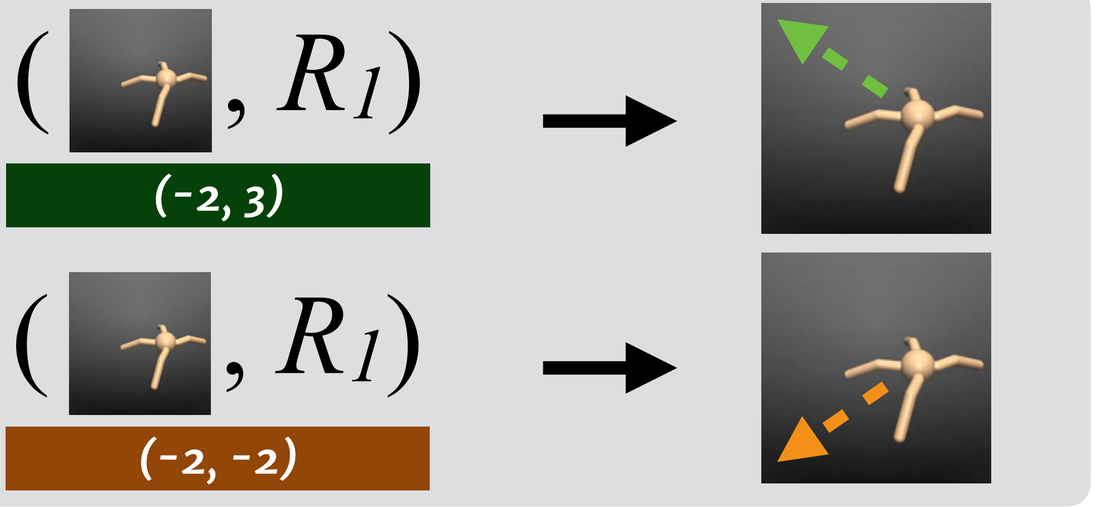


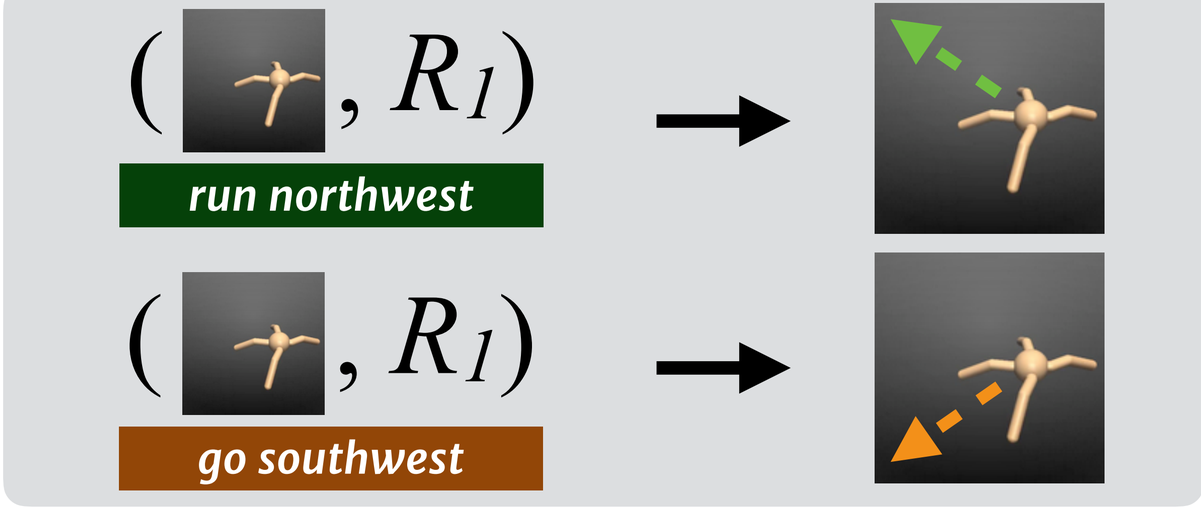



# Instructions as observations



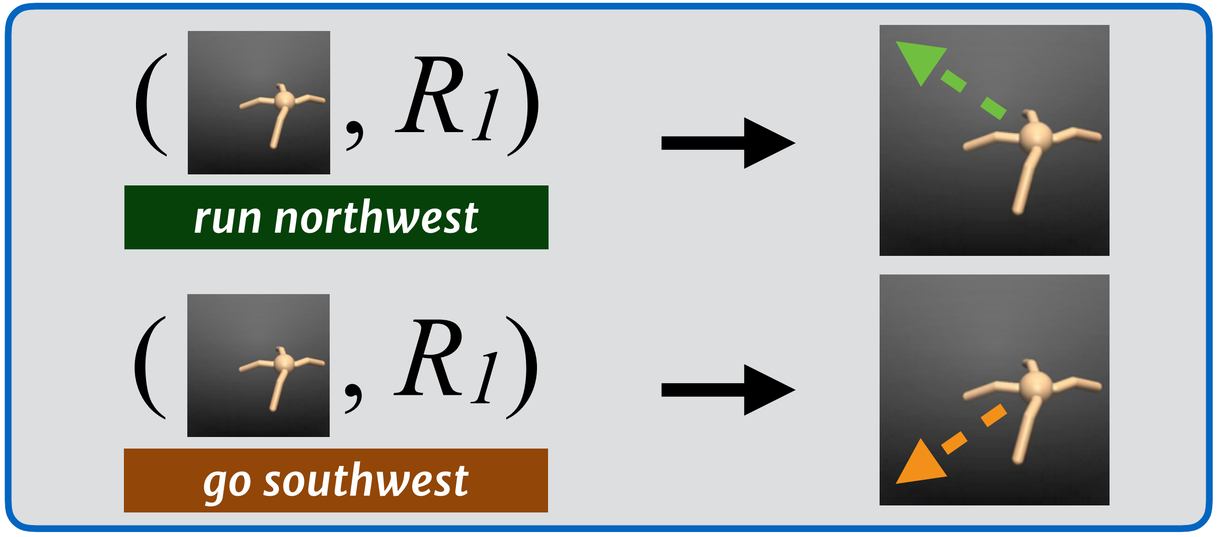




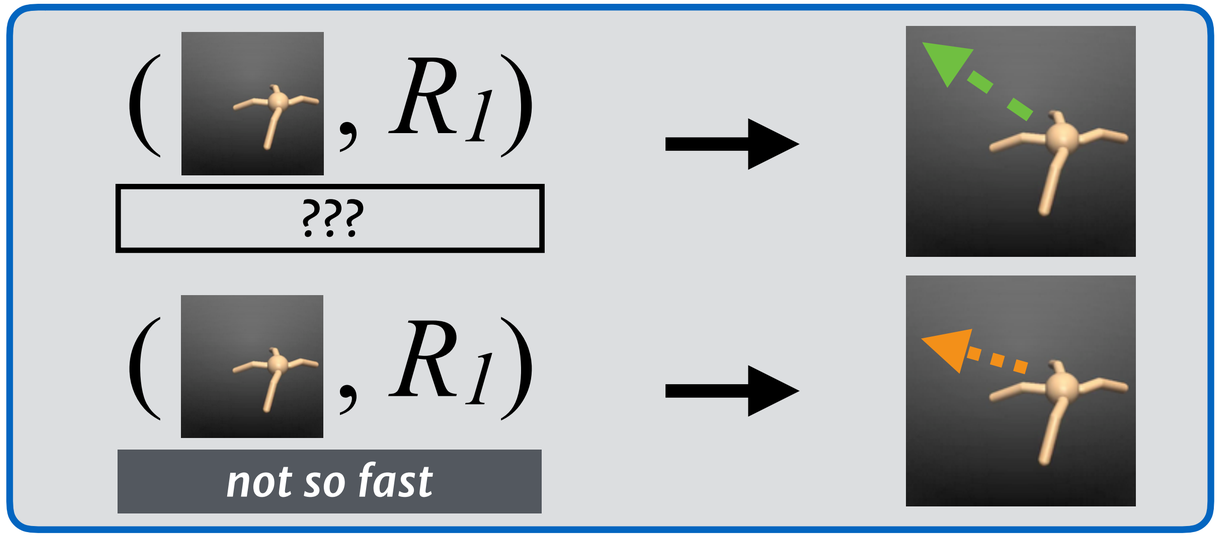
# Beyond observations

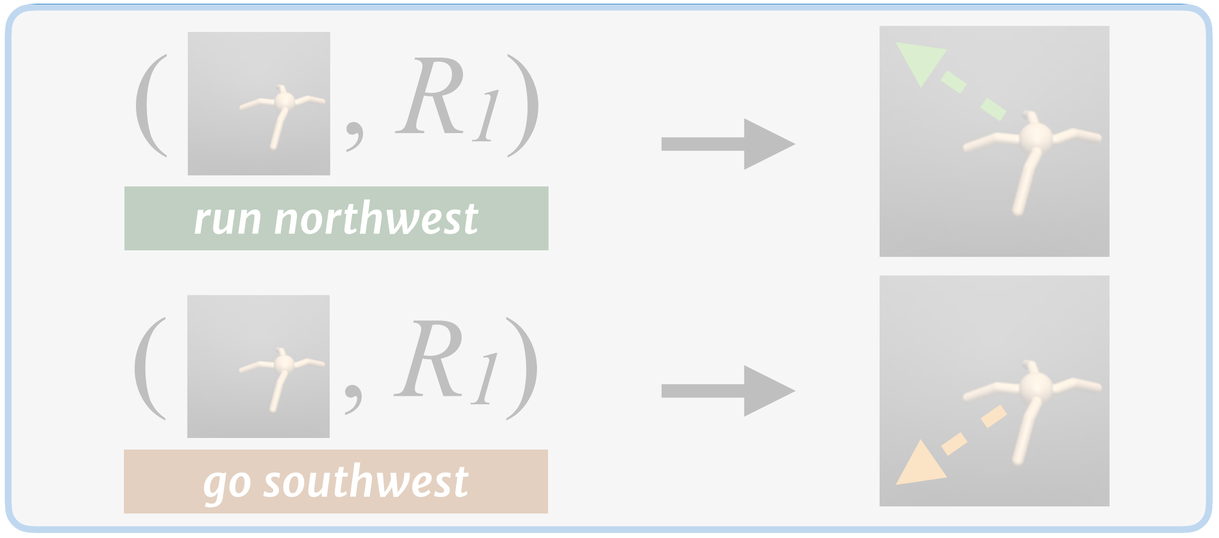
(1) Instructions are moves in a game, not observations of an environment.

# Beyond goals

(2) There's more to language learning than instruction following!



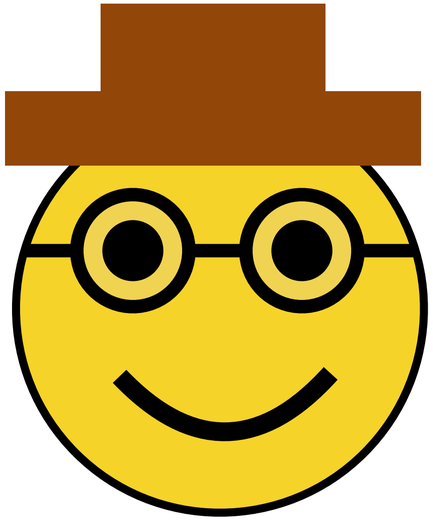
# Language use as gameplay

# Generation & understanding

*Turn right and walk through the kitchen. Go right into the living room and stop by the rug.*

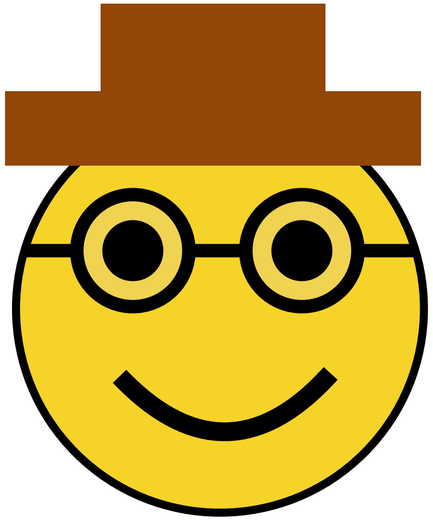
[Anderson et al. 18]

# A reference game



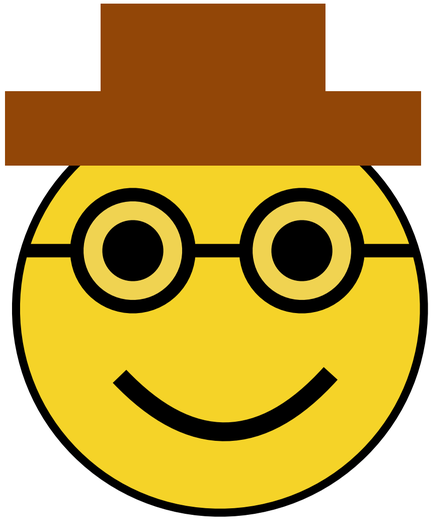
[Frank & Goodman 12]

# “*glasses*"



[Frank & Goodman 12]

# “*glasses*"

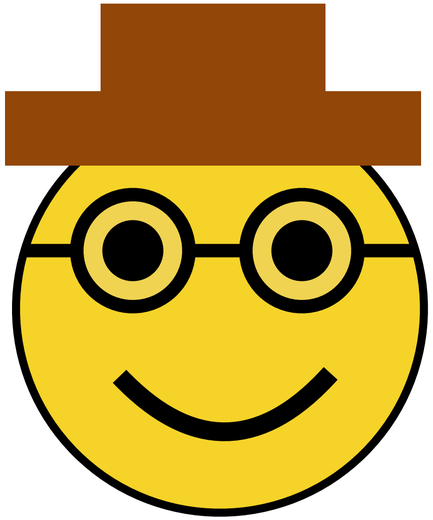


[Frank & Goodman 12]

# “*glasses"*

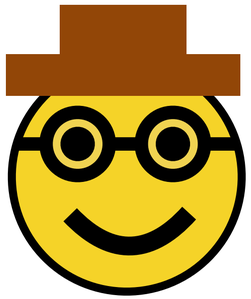
[Frank & Goodman 12]

# “*glasses*"



[Frank & Goodman 12]

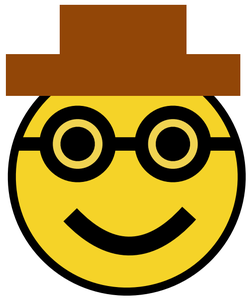
# The rational speech acts model



| | | |
|---|---|---|
| $L_0(\ .\ \mid \textit{glasses})$ | 1/2 | 1/2 |
| $L_0(\ .\ \mid \textit{hat})$ | 0 | 1 |

[Frank & Goodman 12, Degen 13]

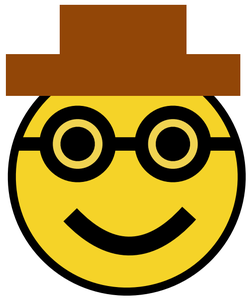
# The rational speech acts model



| | | |
|---|---|---|
| $L_0(\,.\mid glasses)$ | 1/2 | 1/2 |
| $L_0(\,.\mid hat)$ | 0 | 1 |
| $S_1(\,glasses \mid .\,) \propto L_0(\,.\mid glasses)$ | 1 | 1/3 |
| $S_1(\,hat \mid .\,)$ | 0 | 2/3 |

[Frank & Goodman 12, Degen 13]

# The rational speech acts model



| | | |
|---|---|---|
| $L_1(\,.\mid glasses\,) \propto S_1(\,glasses \mid .\,)$ | 3/4 | 1/4 |
| $L_1(\,.\mid hat\,)$ | 0 | 1 |
| $S_1(\,glasses \mid .\,) \propto L_0(\,.\mid glasses)$ | 1 | 1/3 |
| $S_1(\,hat \mid .\,)$ | 0 | 2/3 |

[Frank & Goodman 12, Degen 13]

# Pragmatics

*Q: Do you know what time it is?*

# Pragmatics

*Q: Do you know what time it is?*

*A: Yes*

# Pragmatics

*Q: Do you know what time it is?*

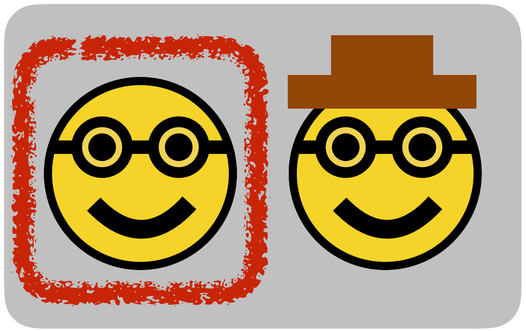
*A: Yes*

*I find his cooking very interesting.*

[Grice 70]

# RSA game tree

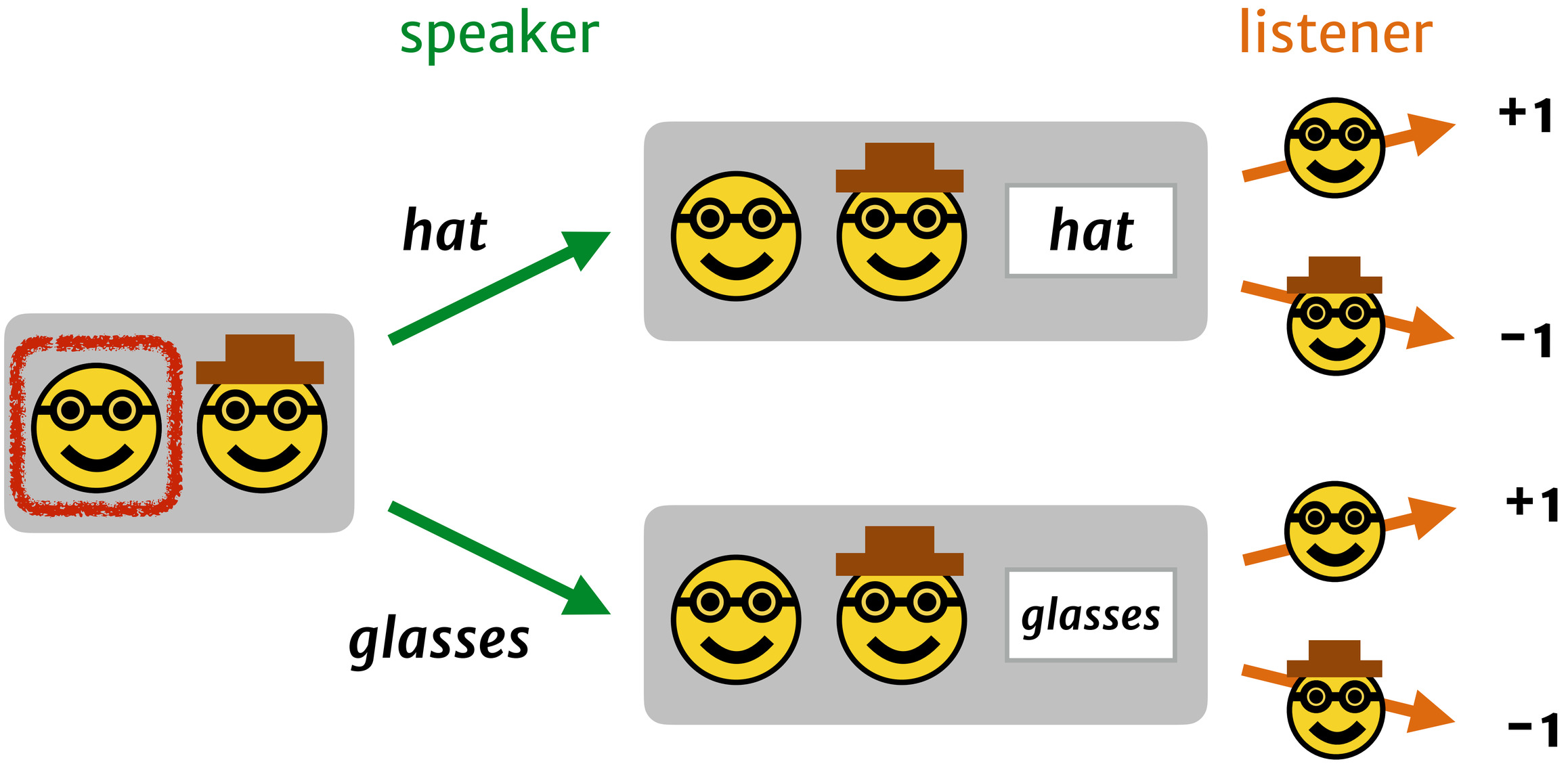
speaker



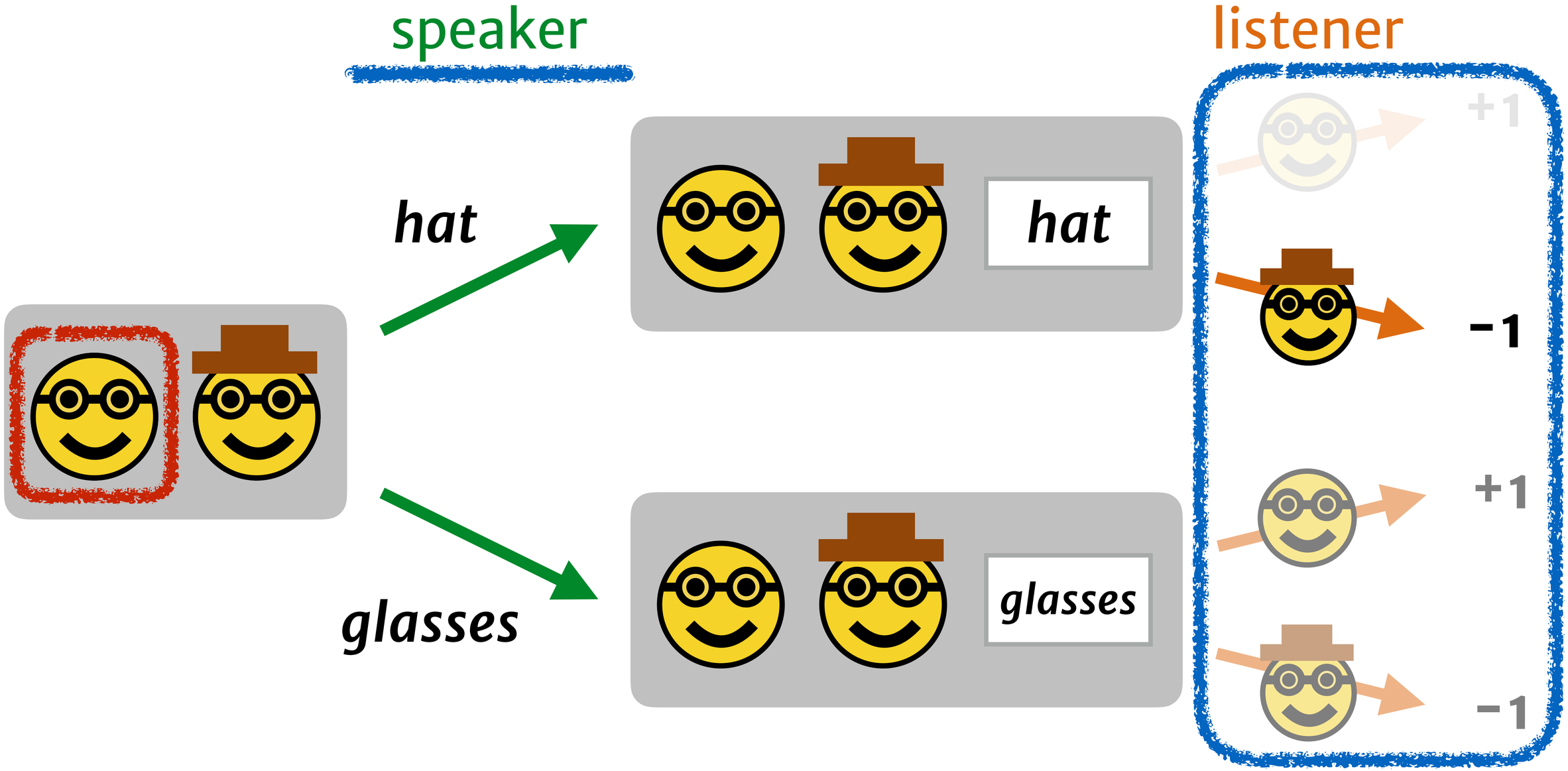
*hat*

*glasses*

# RSA game tree: as speaker



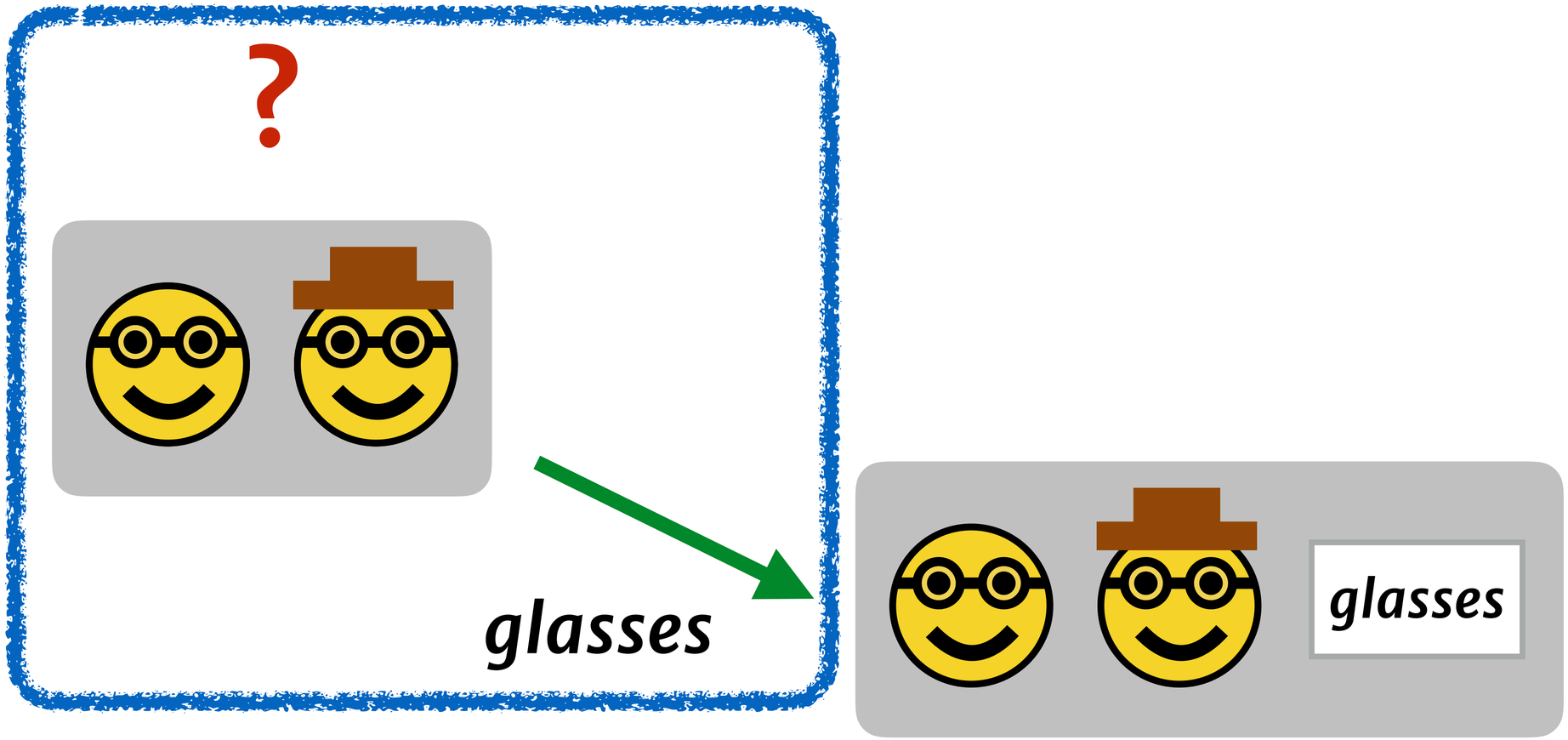
# RSA game tree: as speaker



# RSA game tree: as listener

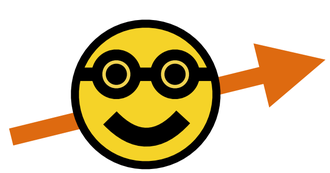
speaker

listener

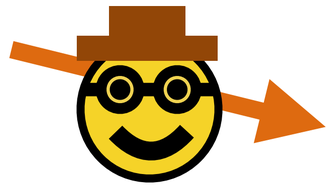








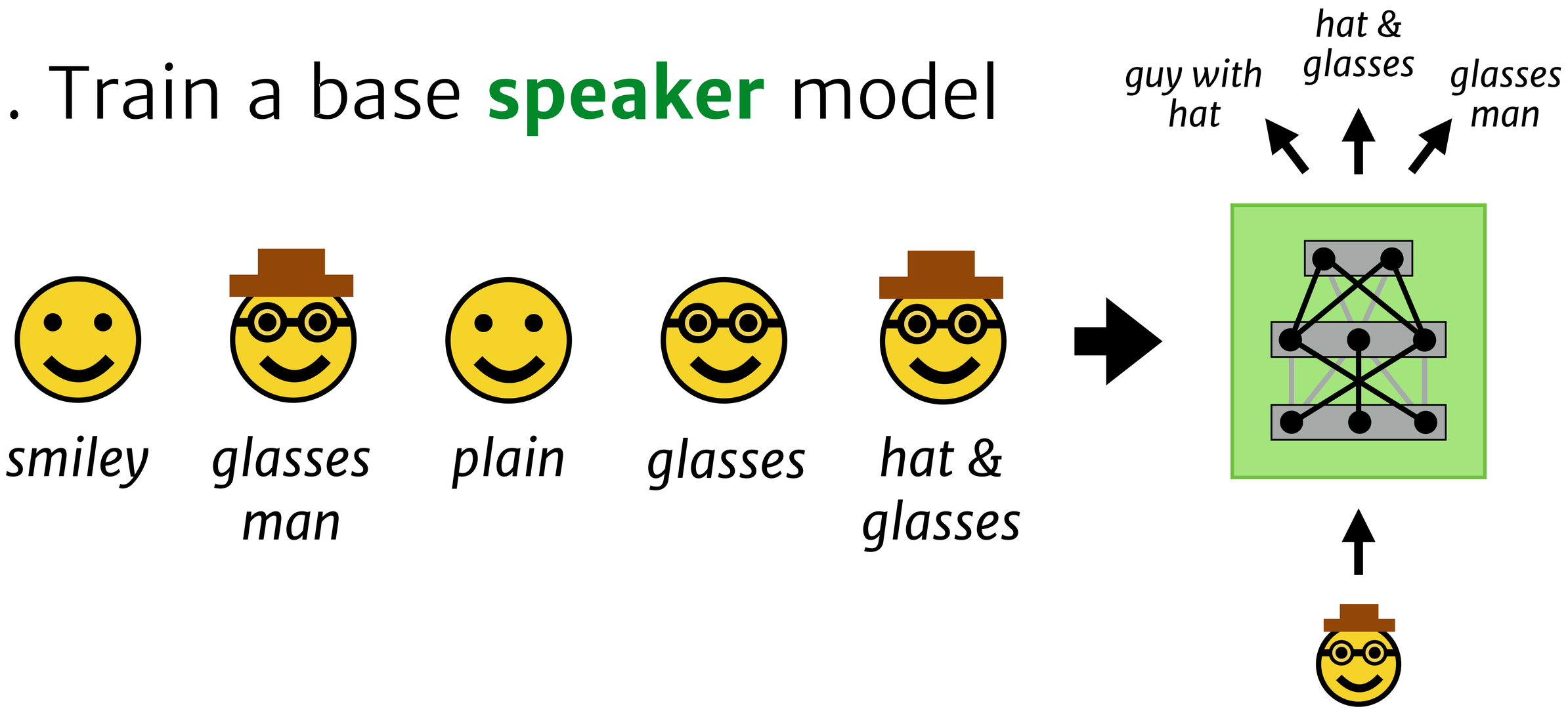
?



?

# A recipe for pragmatic language understanding

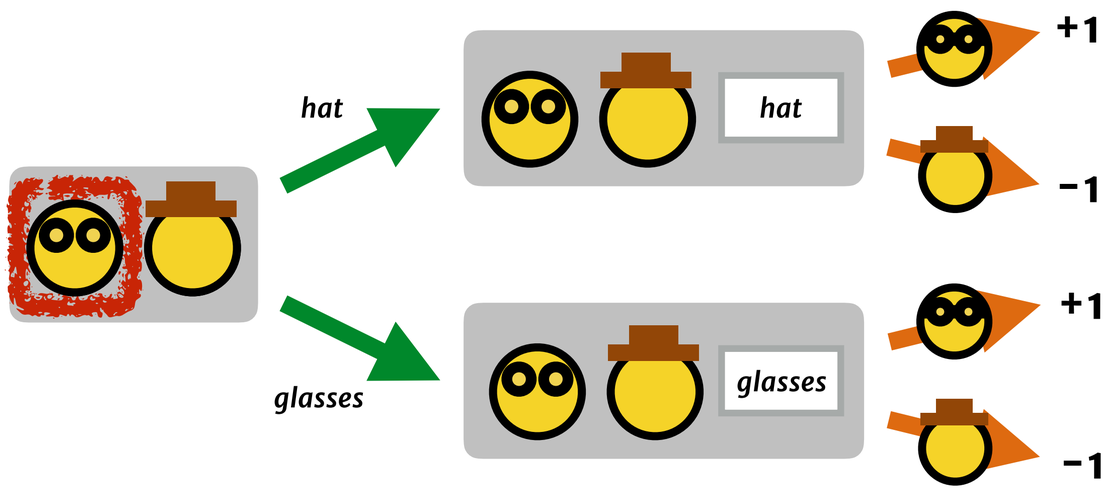
## 1. Train a base **speaker** model



# A recipe for pragmatic language understanding

1. Train a base **speaker** model

2. Solve this POMDP:



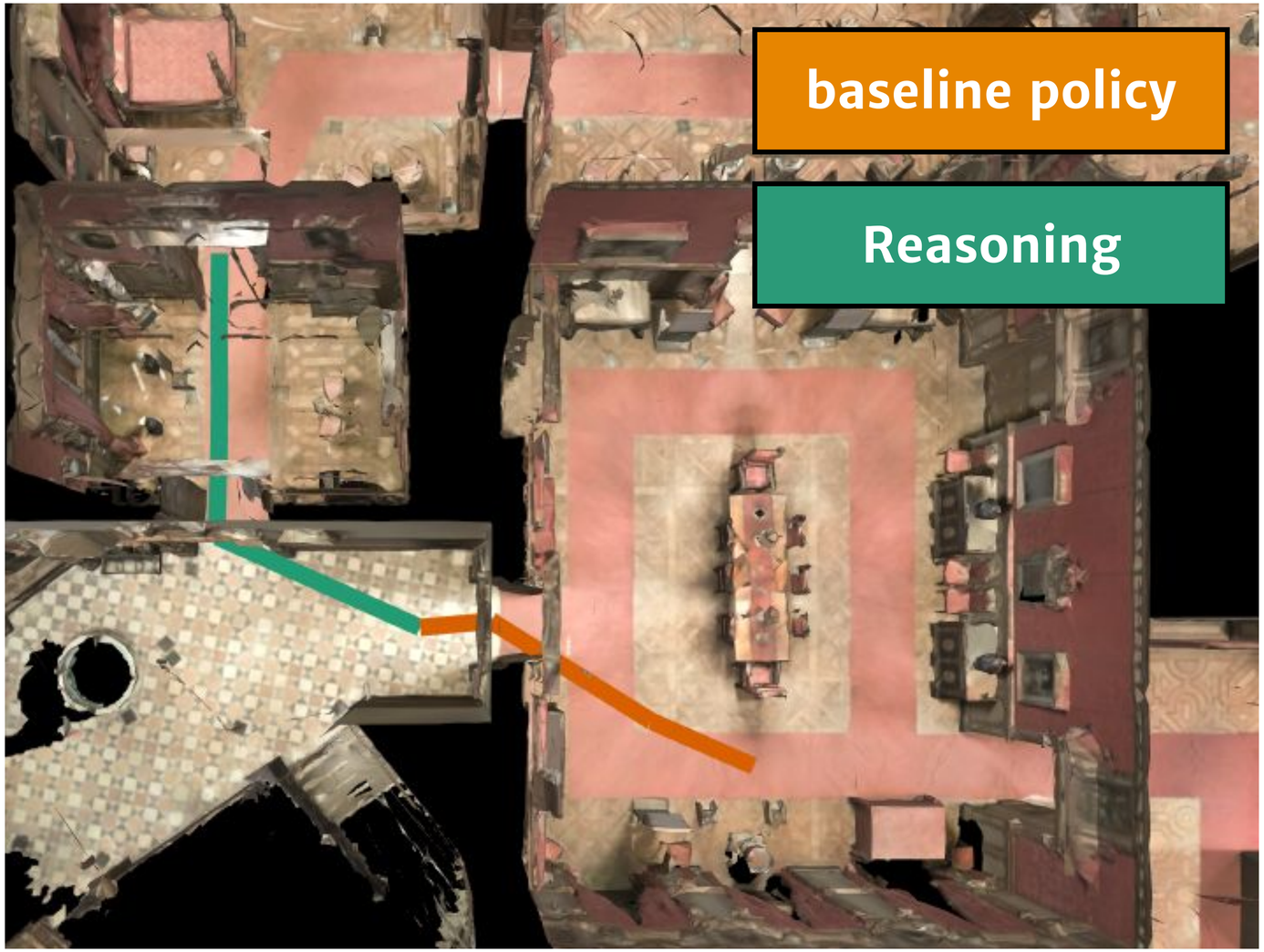

Daniel Fried

Ronghang Hu

Volkan Cirik

Speaker—follower models for vision-and-language navigation. NeurIPS 18.

# Application: instruction following




***human***: *Go through the door on the right and continue straight. Stop in the next room in front of the bed.*

# Application: instruction generation

***seq2seq:*** *Walk past the dining room table and chairs and wait there.*

***reasoning:*** *Walk past the dining room table and chairs and take a right into the living room. Stop once you are on the rug.*

***human:*** *Turn right and walk through the kitchen. Go right into the living room and stop by the rug.*

# Lesson

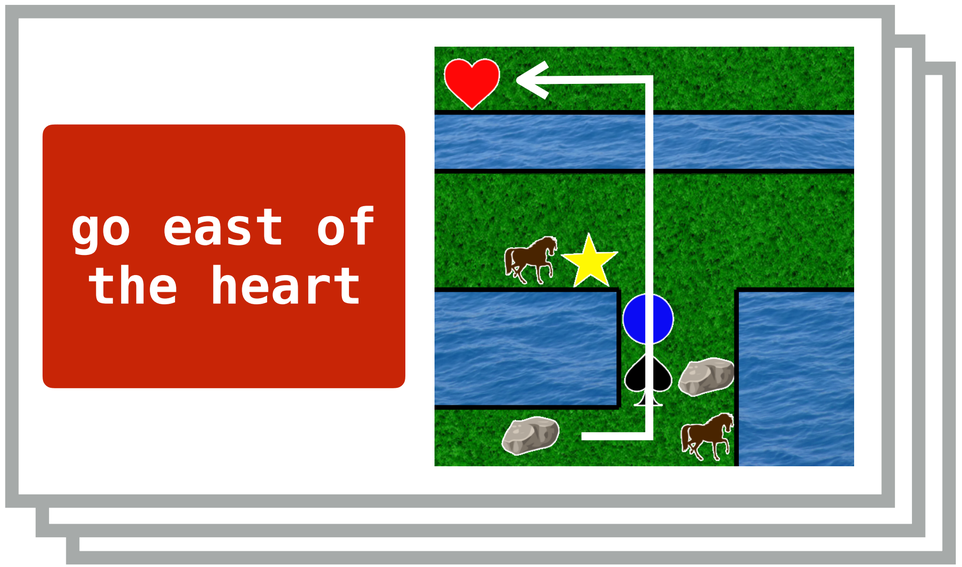
**Utterances are chosen to facilitate correct interpretation in context.**

(This makes the learning problem easier!)

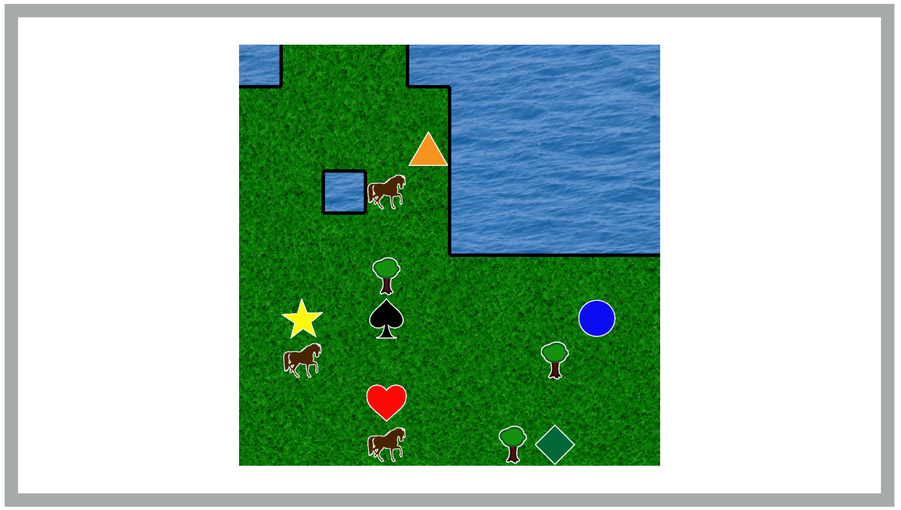
# Language as a scaffold for learning

# What else is an instruction follower good for?

## Language learning

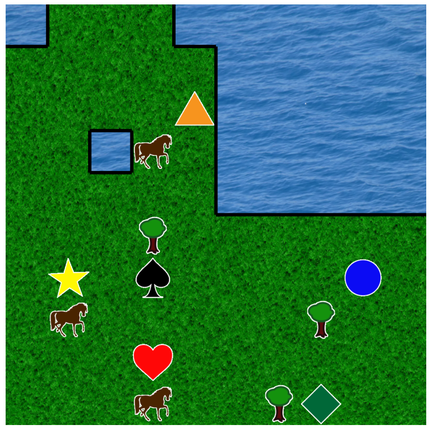



## Reinforcement learning



Learning with latent language.
A, Klein & Levine. NAACL 18.

# Pretraining via language learning



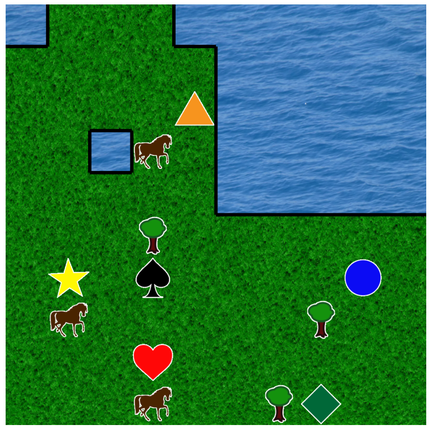
$$\pi(\ ;\eta,\ )$$

NORTH

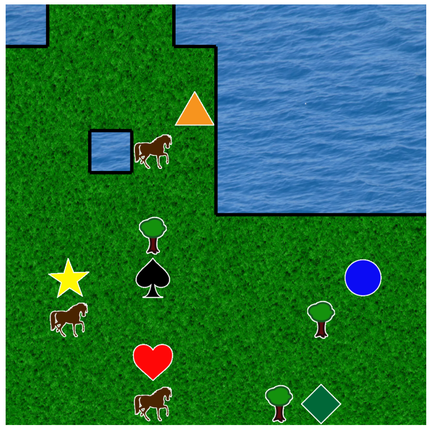
go east of the heart

[Branavan et al., 09]

# (Standard) reinforcement learning



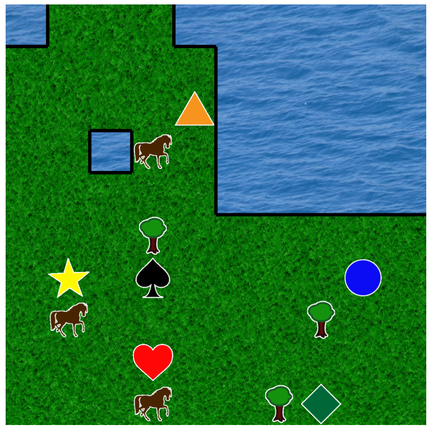
# Concept learning



$$R(\pi(\quad ; \eta, \quad ), \rightarrow \texttt{NORTH,…}$$

**find the horse**

# Concept learning

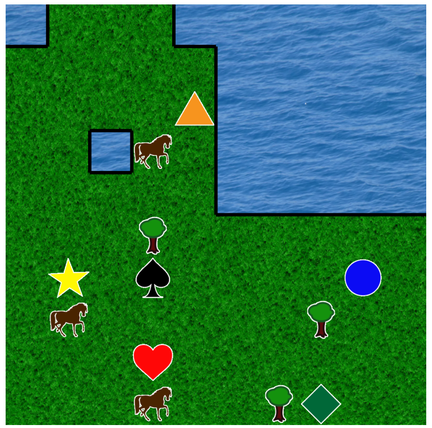


$R(\pi($ $; \eta,$ $),$ → **NORTH,…**

**find the horse**

-0.52

# Concept learning

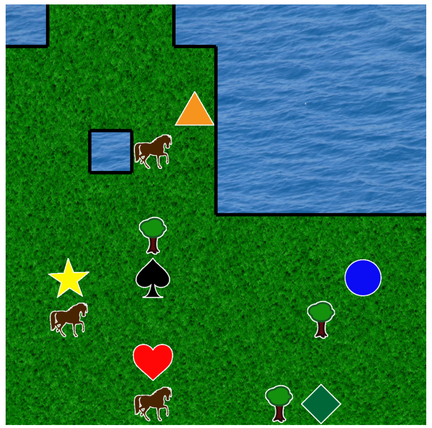


$$R(\pi(\quad ; \eta, \quad ), \rightarrow \textbf{SOUTH,…}$$

find the horse -0.52

left of heart 0.33

# Concept learning

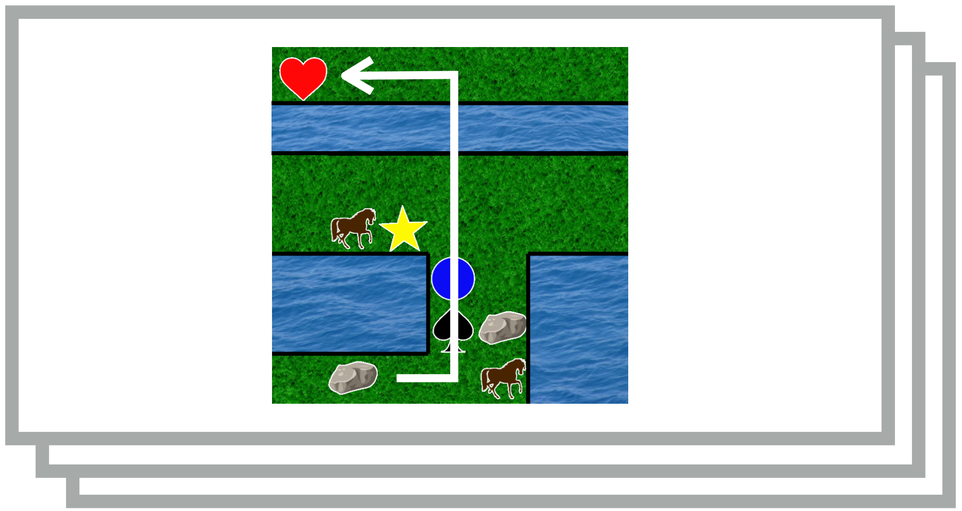


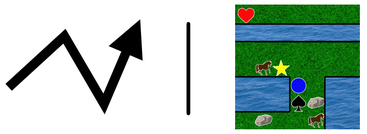
$$R(\pi(\quad ; \eta, \quad ), \rightarrow \texttt{SOUTH,…}$$

| | |
|---|---|
| find the horse | -0.52 |
| left of the heart | 0.33 |
| **heart east side** | 0.95 |

# As multitask learning

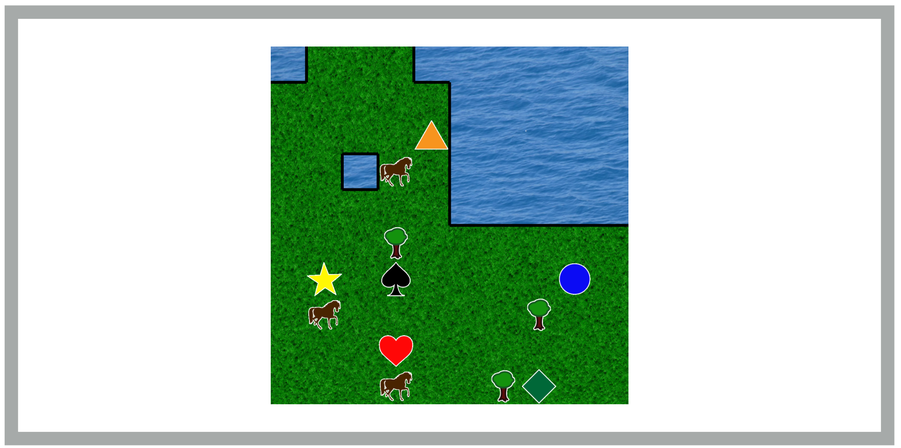
## Language learning

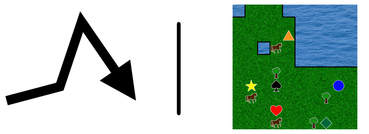


$$\arg\min_{\eta} R(\pi(\text{  } ; \eta, \text{  }))$$

go east of the heart

## Reinforcement learning



$$\arg\min_{\eta} R(\pi(\text{  } ; \eta, \text{}))$$

???

find the triangle

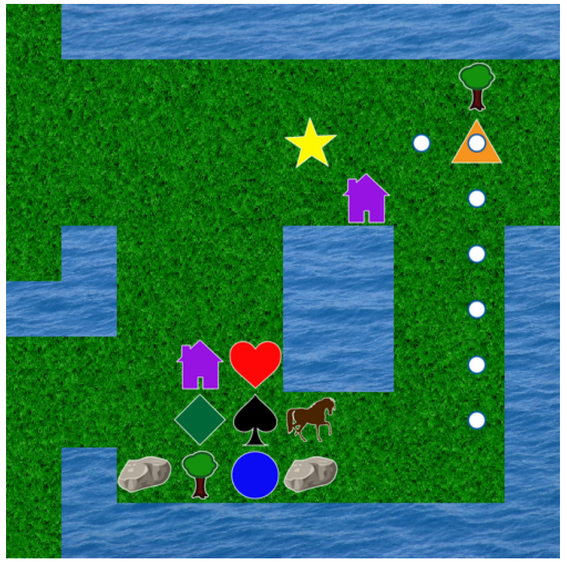
[Caruana 97]

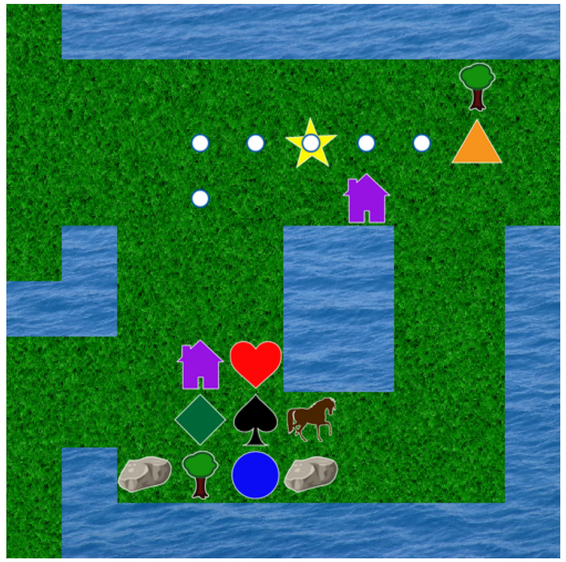
# As a language game...

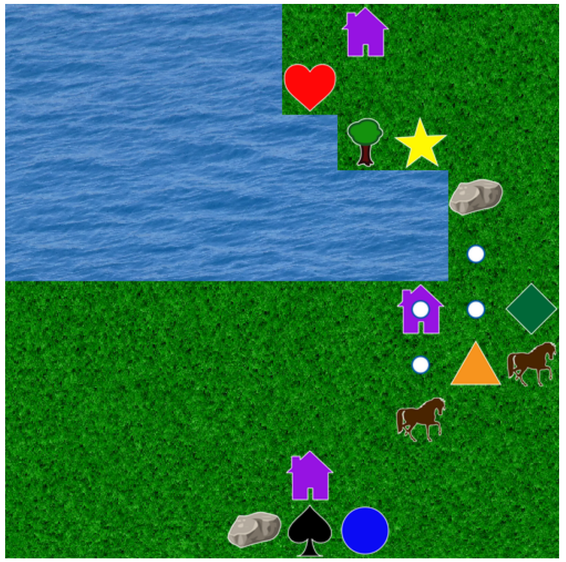
$$\arg\min \qquad \pi \circ R$$


**speaker model** **listener loss**

# Results



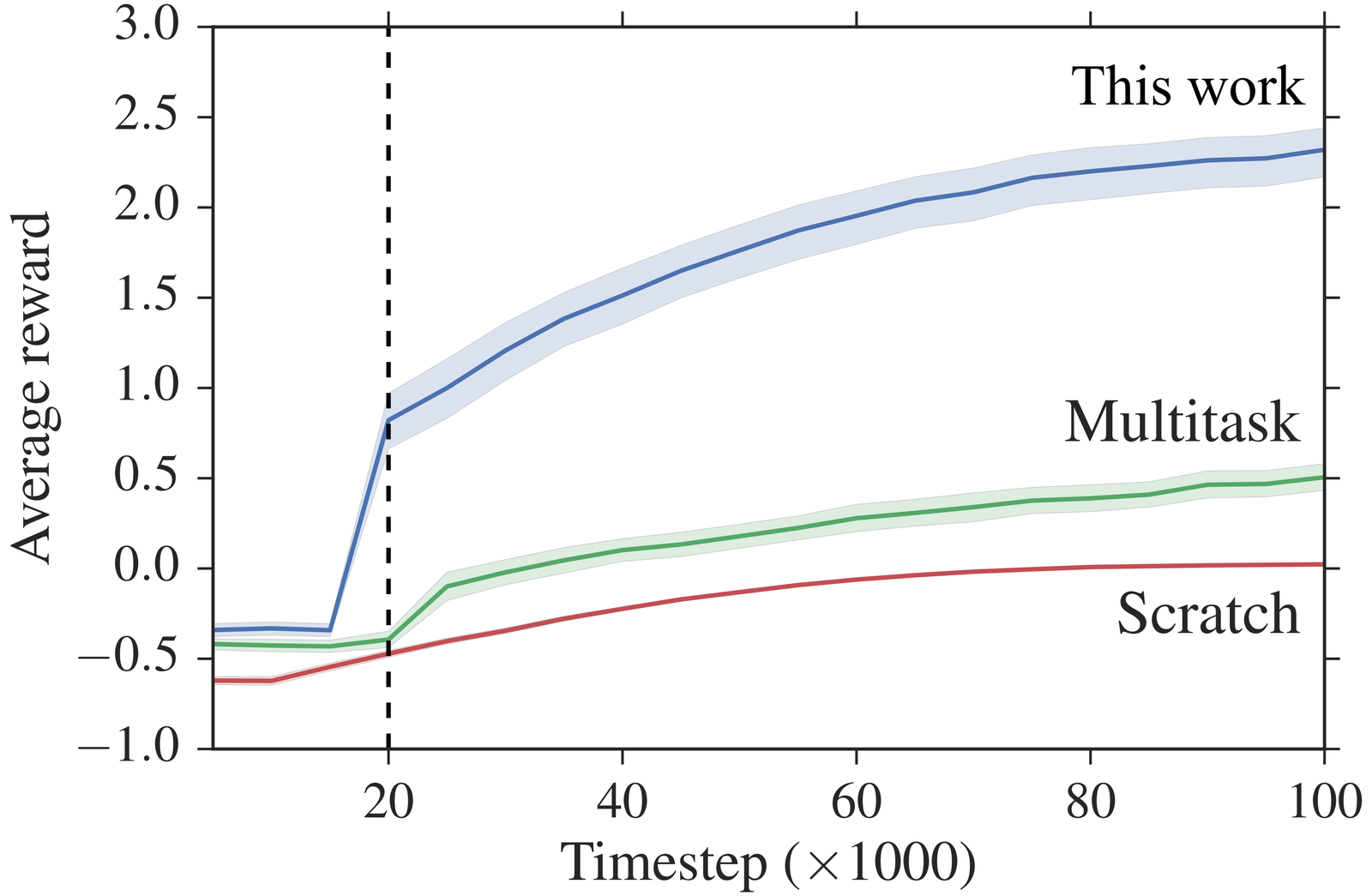




True description: `reach cell on left of triangle`

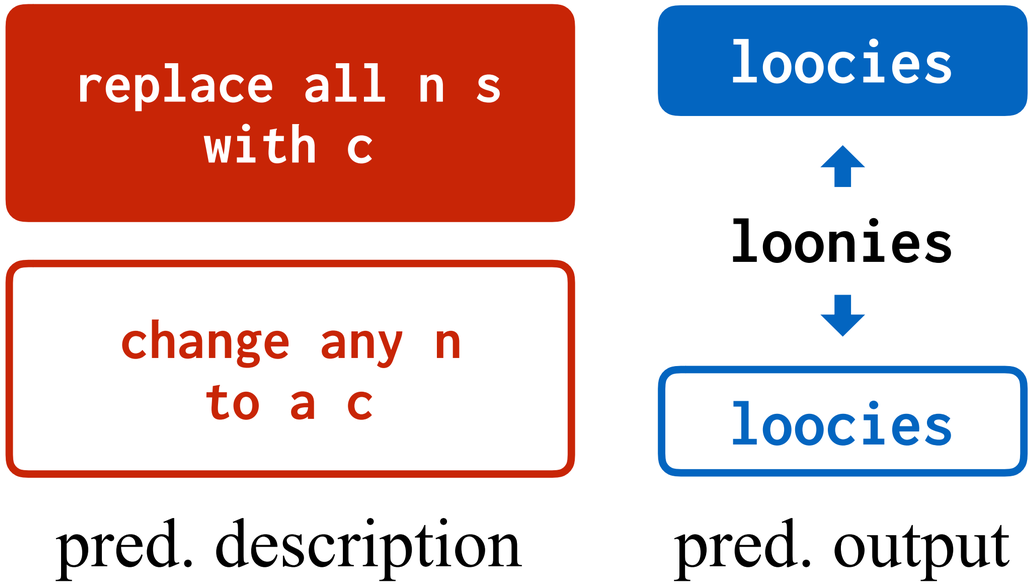
Pred description: `reach square left of triangle`

# Results: RL

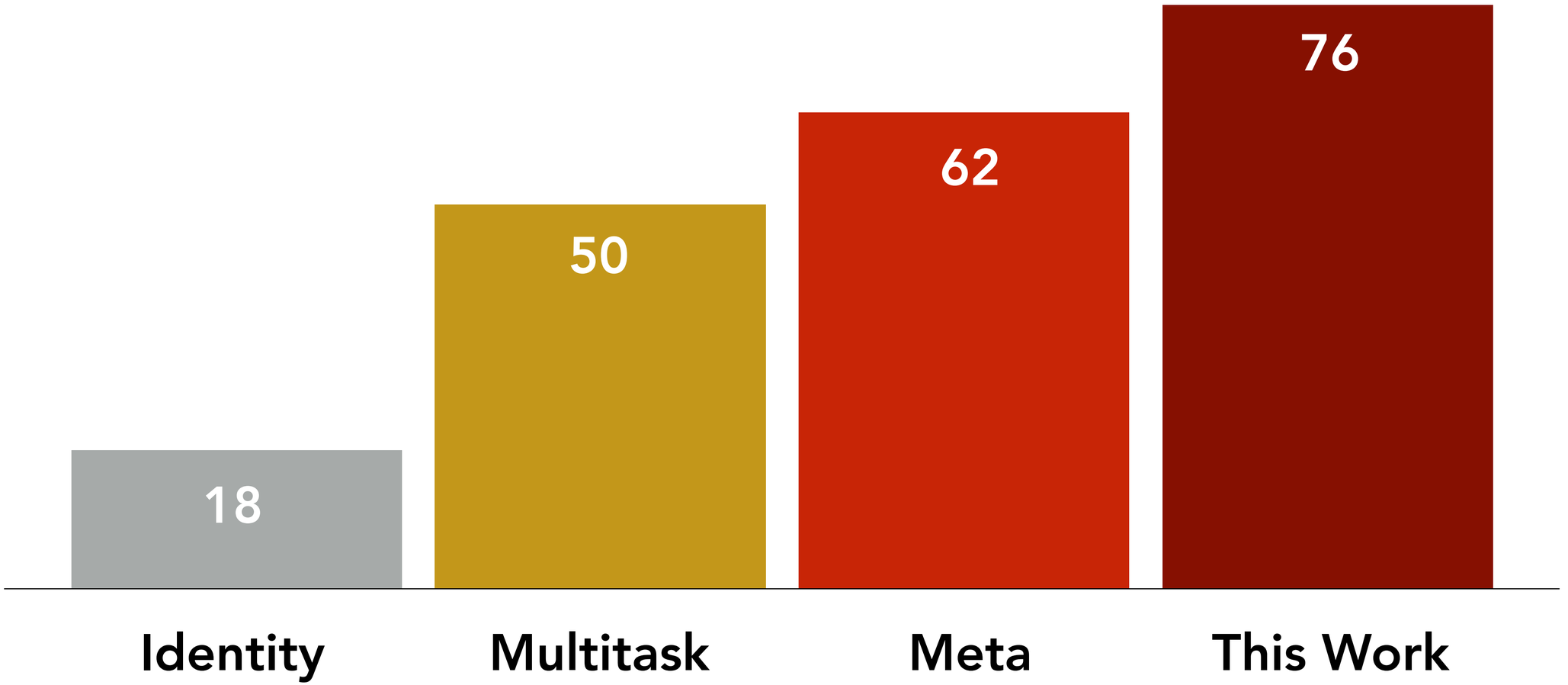

3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Average reward
20
40
60
80
100
Timestep (×1000)
This work
Multitask
Scratch

# Results

| examples | | |
|---|---|---|
| emboldens | | emboldecs |
| kisses | | kisses |
| loneliness | ➡ | locelicess |
| vein | | veic |
| dogtrot | | dogtrot |

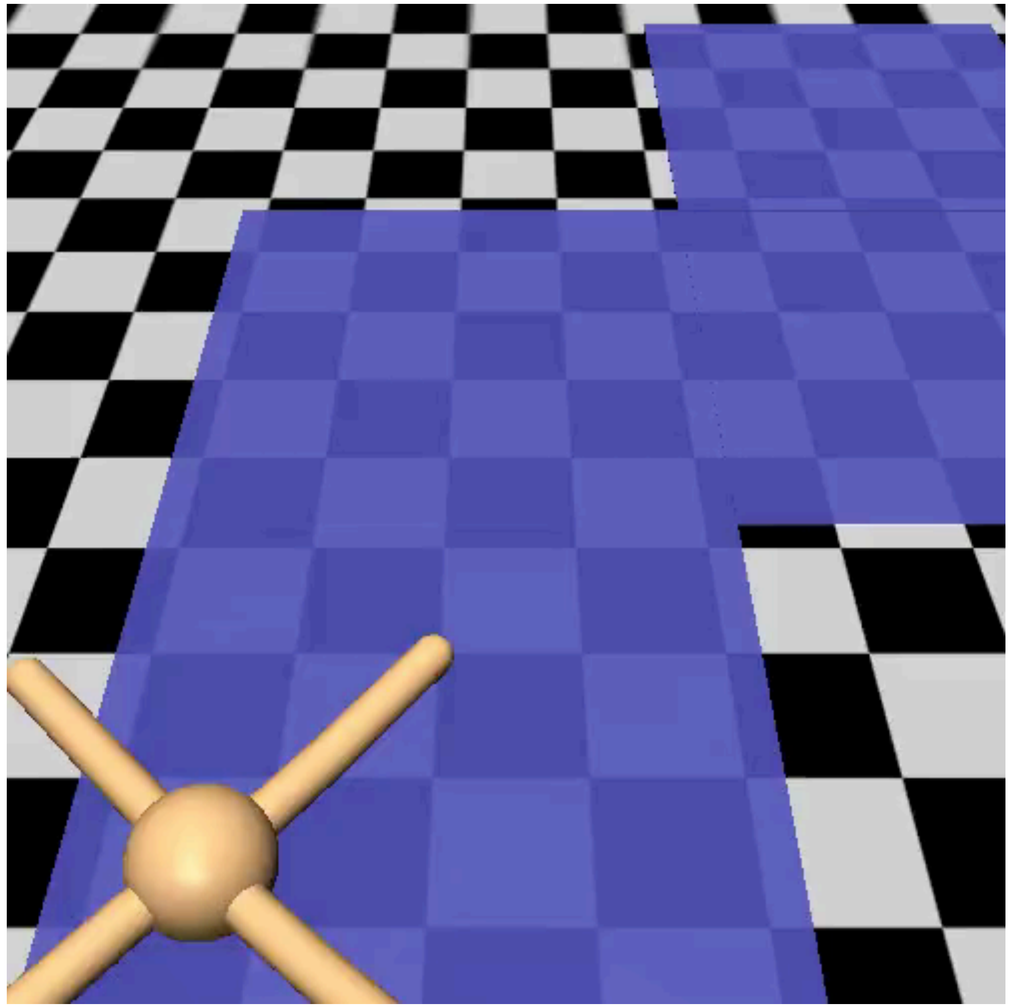


# Results: programming by demonstration



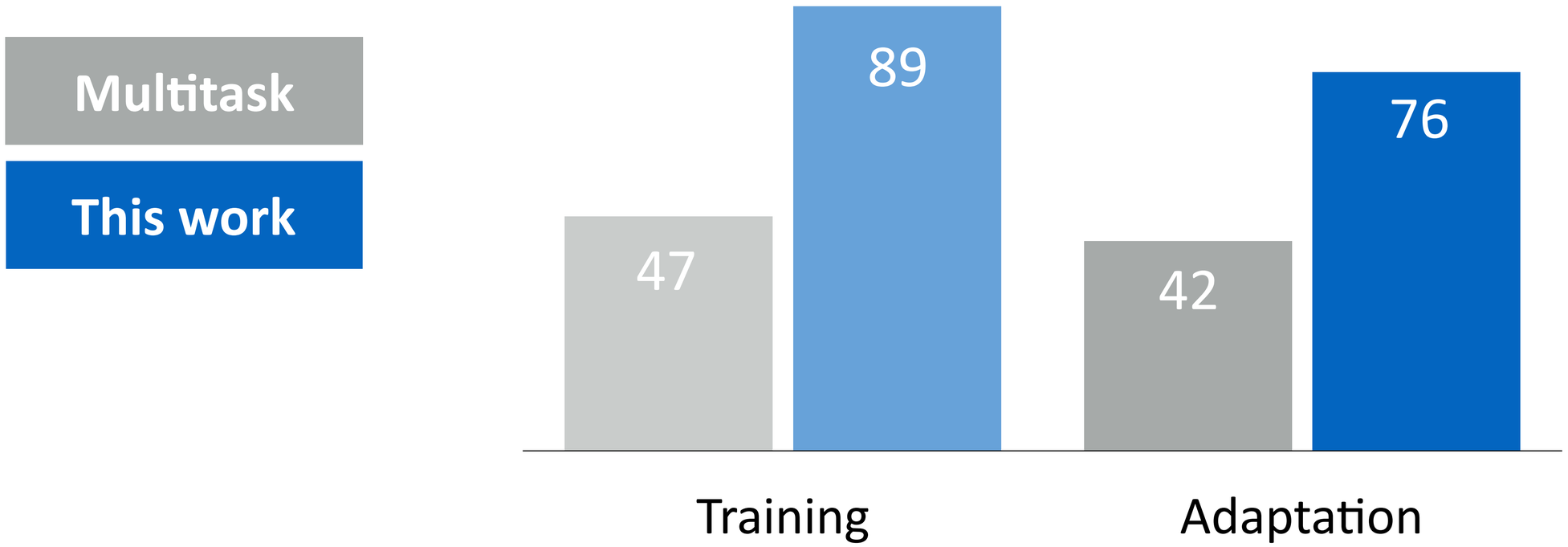
# Results: locomotion

**north, east, north**



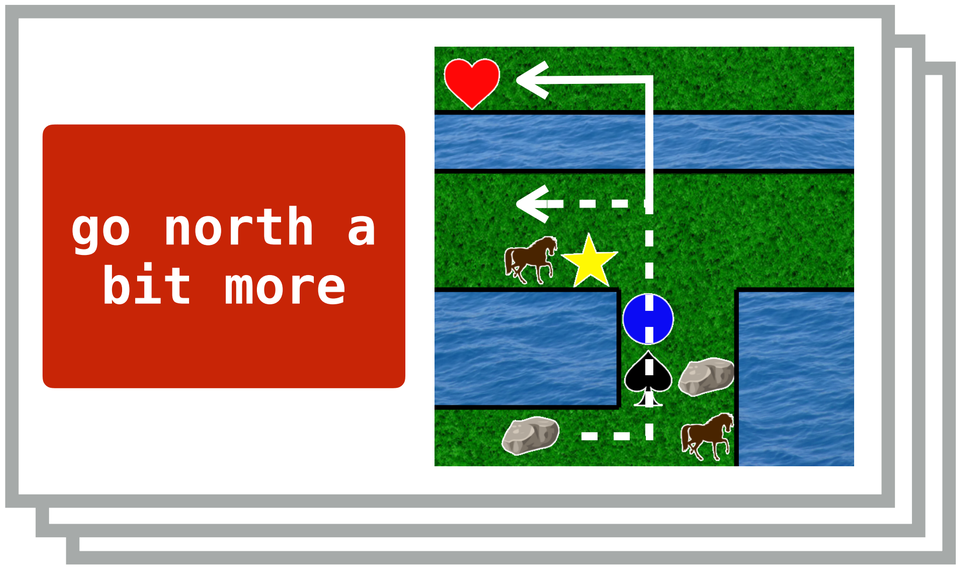
Modular multitask reinforcement learning with policy sketches. A, Klein & Levine. ICML 2017

# Generalization

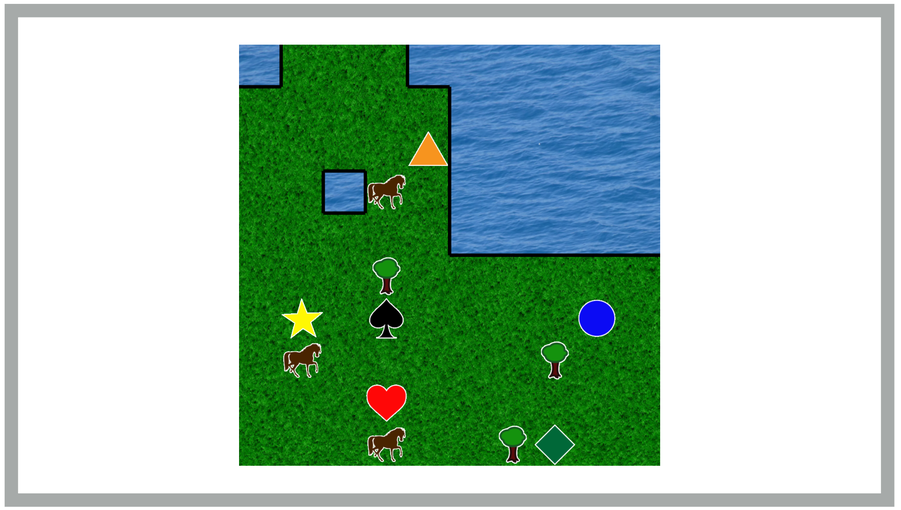


# Learning with corrections

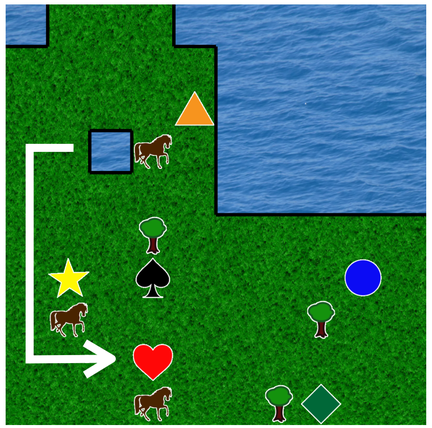
## Language learning




## Reinforcement learning



# Pretraining by learning to correct



$$\pi(\ ;\eta,\ )$$

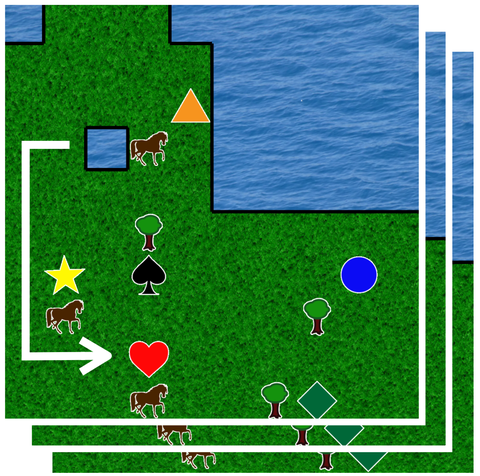
NORTH

further east

JD Co-Reyes

Guiding policies with language via meta-learning. ICLR 19.

# Pretraining by learning to correct

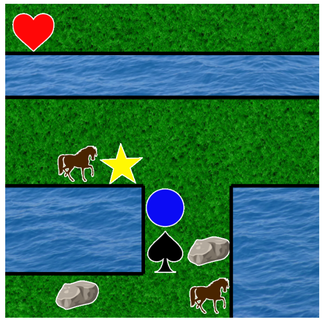


$$\pi(\ ;\eta,\ )$$

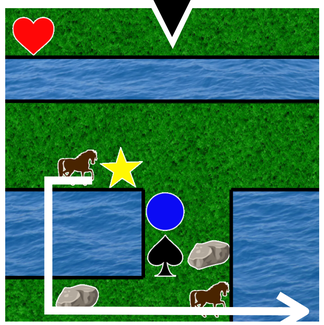
NORTH

further east

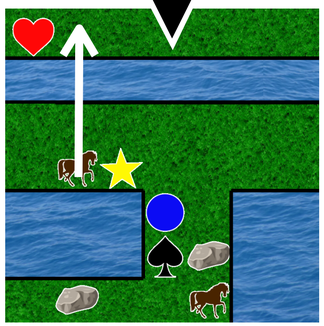
# Learning from corrections



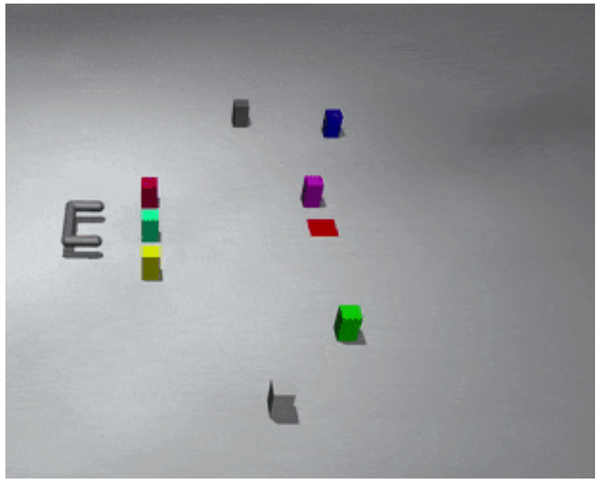
$\pi(\cdot\ ;\ \eta,$ [ ] $)$ → **WEST,…**



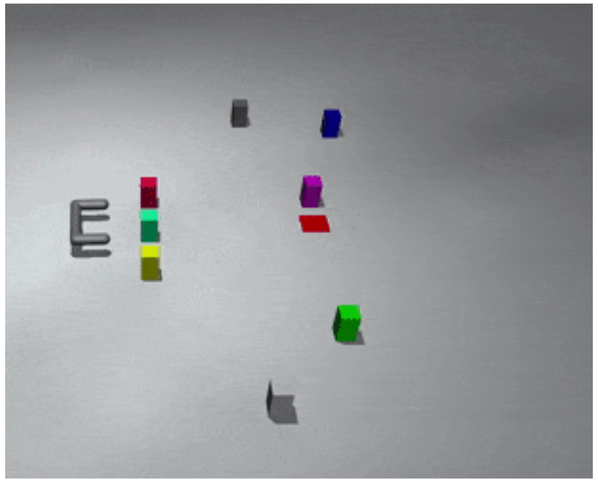
$\pi(\cdot\ ;\ \eta,$ **go to top** $)$ → **NORTH,…**



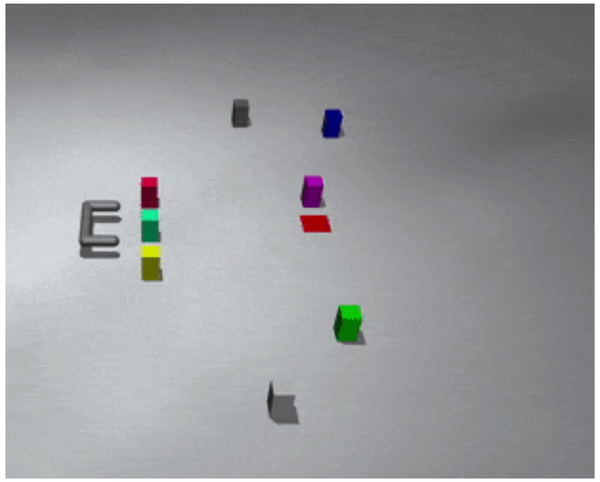
$\pi(\cdot\ ;\ \eta,$ **further west** $)$ → **NORTH,…**



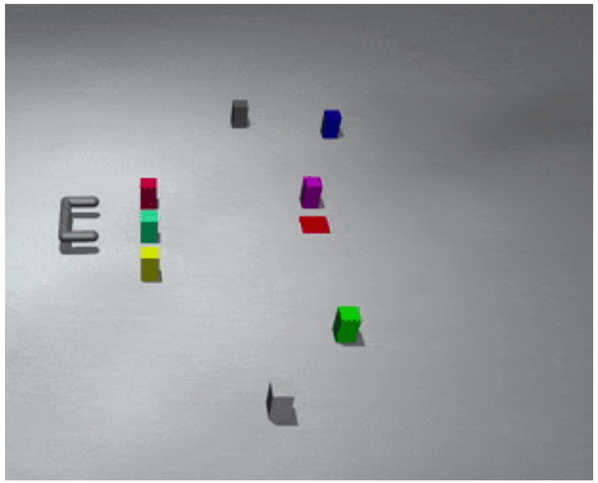
*Touch cyan block.*



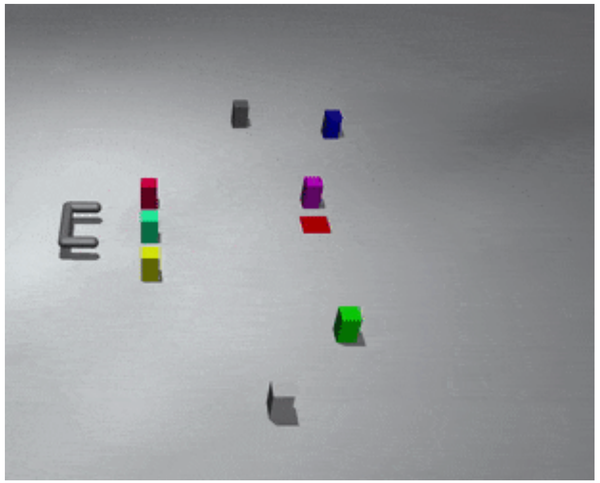
*Move closer to magenta block.*

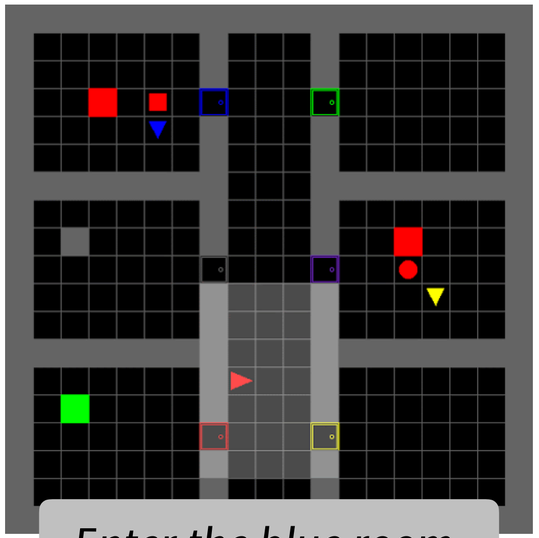


*Move a lot up.*

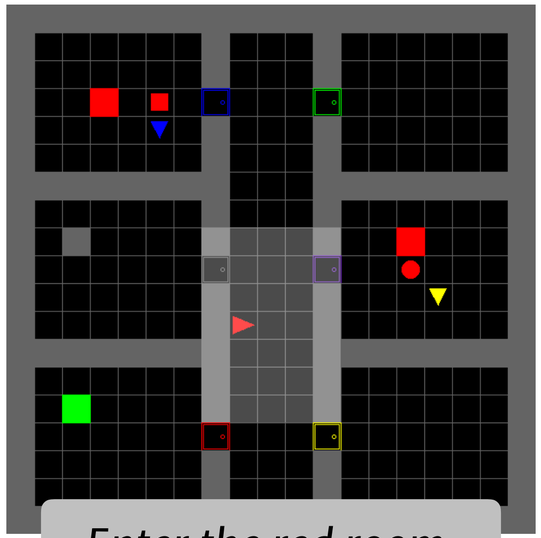


*Move a little up.*

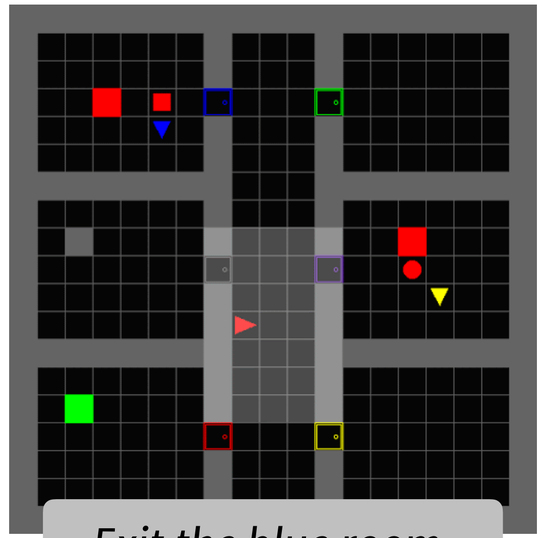




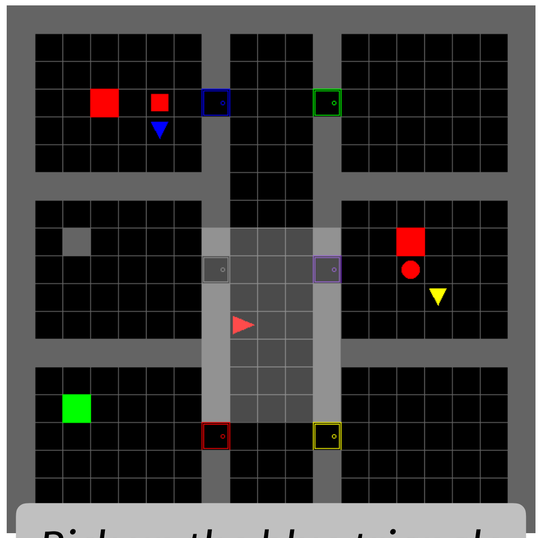
Enter the blue room.



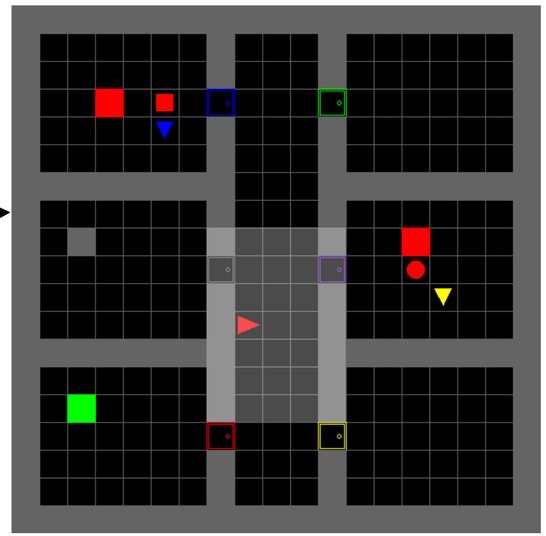
Enter the red room.



Exit the blue room.



Pick up the blue triangle



# Lesson

**Language is useful as side information, not just a goal specification.**

Use it with / instead of instructions as
a representational bottleneck
or interactive advice

# So what comes next?

# What comes next?

Challenges for the field:

# What comes next?

Challenges for the field:

- **huge** datasets

# What comes next?

Challenges for the field:

- **huge** datasets
- with **fake** annotations

# What comes next?

Challenges for the field:

- **huge** datasets
- with **fake** annotations
- that look **very little like natural language**

# What comes next?

Challenges for the field:

- **huge** datasets → Learn to make do without an annotation for every rollout!
- with **fake** annotations
- that look **very little like natural language**

# What comes next?

Challenges for the field:

- **huge** datasets → Learn to make do without an annotation for every rollout!
- with **fake** annotations → Learn to generalize from fake strings to real ones!
- that look **very little like natural language**

# What comes next?

Challenges for the field:

- **huge** datasets → Learn to make do without an annotation for every rollout!
- with **fake** annotations → Learn to generalize from fake strings to real ones!
- that look **very little like natural language**

  → Pay attention to human evals (or scope claims accordingly)!

# **Learn more:** Luketina *et al.*, *A survey of reinforcement learning informed by natural language*

`https://arxiv.org/abs/1906.03926`