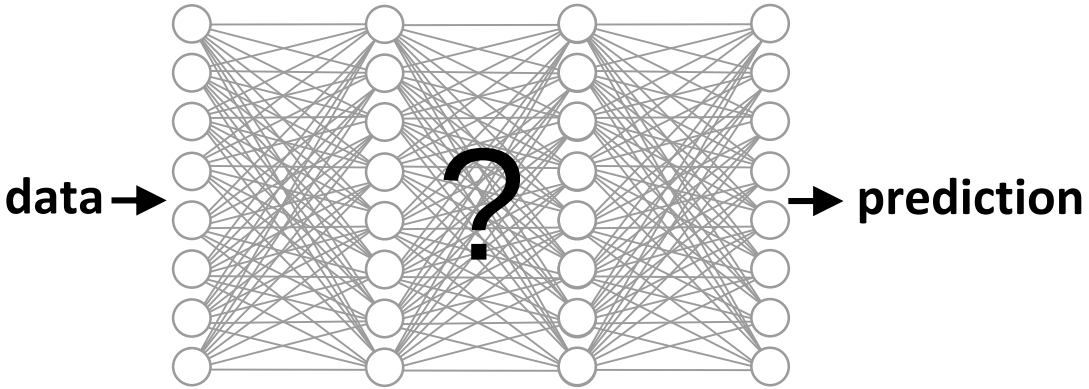
Deeply Explainable AI

Darrell, Abbeel, Dragan, Klein, Griffiths, Canny, Saenko, Akata, Hoogs

Current deep neural networks (DNNs) are "**black boxes**"



- Do not expose their **decision making** process
- Do not provide their **confidence** in their predictions
- Not clear whether they can be **trusted** and/or **corrected**

How do we make

Deep Learning

more **explainable** and **trustworthy**?

Overview

Challenge Problems (Kate)

Explicit and Implicit Explanation Models (Kate)

Learning How to Explain (Zeynep)

Modeling the User (Trevor)

Program Schedule (Trevor)

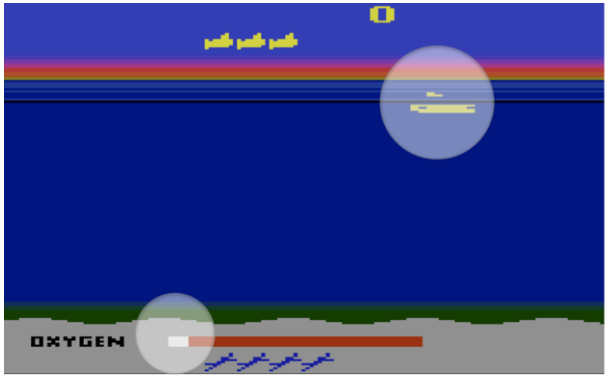
Challenge Problems

Autonomy Challenge:
control of autonomous vehicles

Can we explain the agent's behavior?

- Demonstration-based control
- Reinforcement-learned control
- In simulated environments

Control Explanation: A left banked turn is engaged to avoid crashing into canyon



Textual Explanation: Submarine is going up to surface to replenish oxygen

Attenton Expalnation: Agent decides to go up after looking at oxygen bar and current position

Route Explanation: Avoid buildings, Avoid being seen by people

Analytics Challenge:
Multimedia **E**vent **Q**uestion **A**nswering (MEQA)

Can we explain the system's answers?

- Answering questions about images / video with associated audio / text
- Interactive UI enables natural language dialogue with system

Data collection

- Extend VQA, MovieQA datasets

Q: Can these people arrest someone?
A: Yes
Explanation: … because they are Vancouver police

Q: What is he doing?
A: Juggling
Explanation:
…because he has two balls in his hands while two are in the air

[Park, Hendricks, Akata, Schiele, Darrell, Rohrbach; in submission]

Analytics Challenge:
Multimedia **E**vent **Q**uestion **A**nswering (MEQA)

Can we explain the system's answers?

- Answering questions about images / video with associated audio / text
- Interactive UI enables natural language dialogue with system

Data collection

- Extend VQA, MovieQA datasets
- Extend Berkeley DeepDrive dataset (100K hours of driving video, GPS/sensor IMU data) with text

User: "why did you turn left"?
To avoid the traffic delay near the shopping mall.

User: "why did we stop?"
I can't tell, is this a shadow or a pothole?
User: "how do you know how to drive through this intersection?"
I learned from 236 prior driving exemplars transiting in this direction at this time of day in the past month.

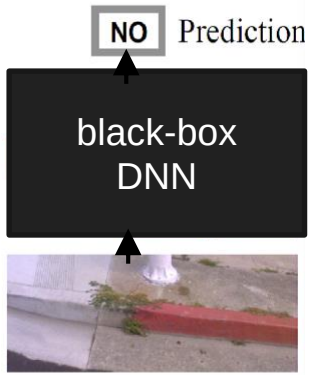
Explicit Explanation Models

Explicit Explanation Models

Explain higher-level reasoning in DNNs

- Explainable decision path for multi-task, control and planning
- Provide structure and intermediate state

Q: Can you park here?

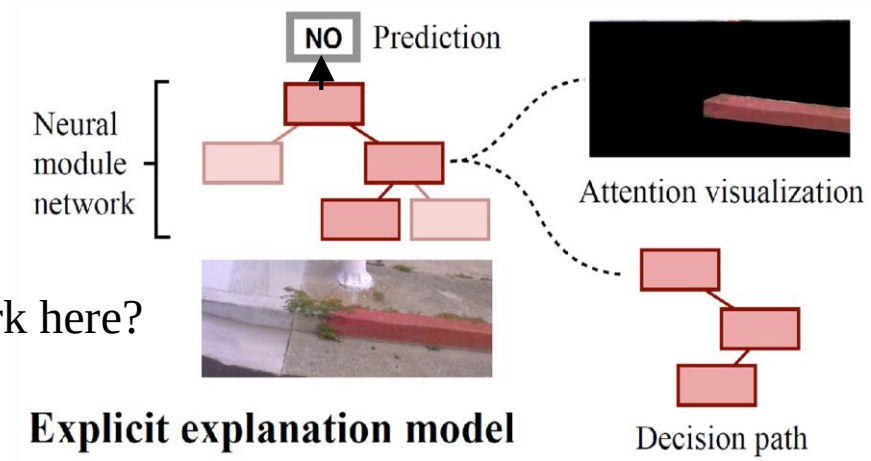


Explicit Explanation Models

Explain higher-level reasoning in DNNs

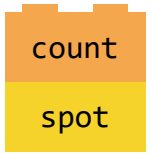
- Explainable decision path for multi-task, control and planning
- Provide structure and intermediate state

Q: Can you park here?

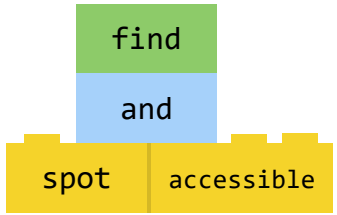


Neural module networks

How many parking spots are there?



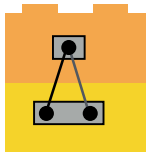
Where is the accessible parking spot?



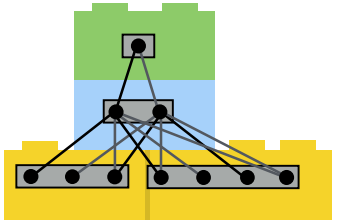
[Andreas, Rohrbach, Klein, Darrell; CVPR 2016, EMNLP 2016]

Neural module networks

How many parking spots are there?



Where is the accessible parking spot?

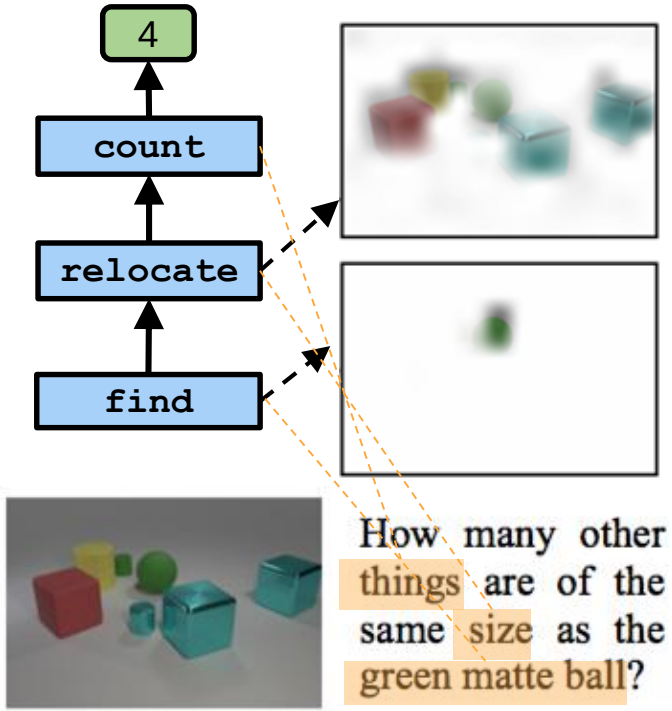


[Andreas, Rohrbach, Klein, Darrell; CVPR 2016, EMNLP 2016]

Neural module networks

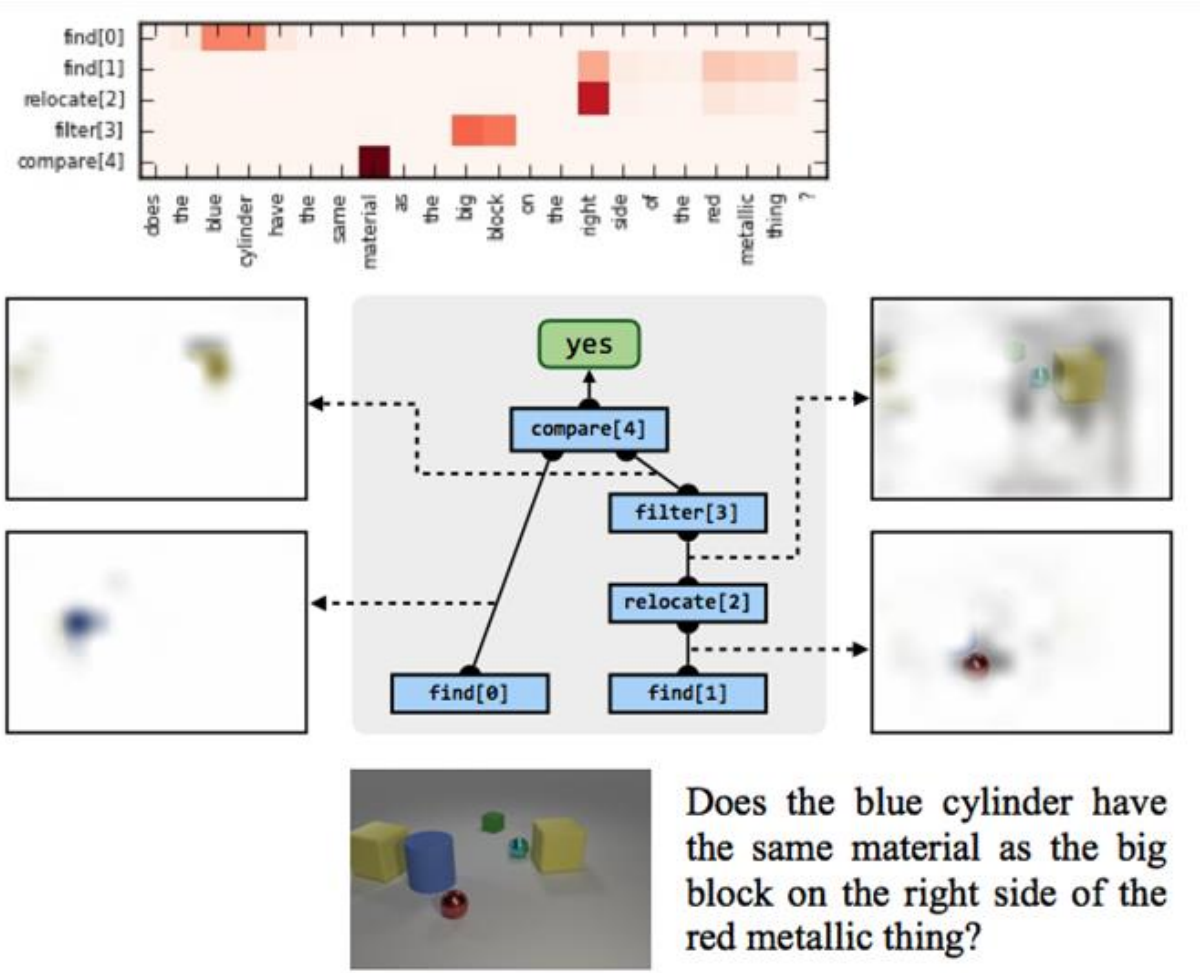
Explain answers via

- exposing modules used for prediction
- showing the chain of reasoning
- visualizing text and image attention



[Hu, Andreas, Rohrbach, Darrell, Saenko; in submission]

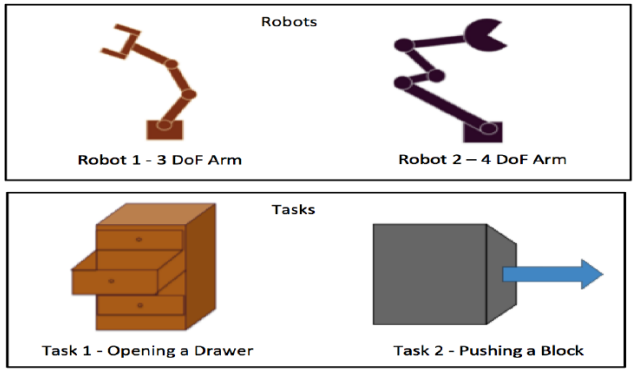
Neural module networks

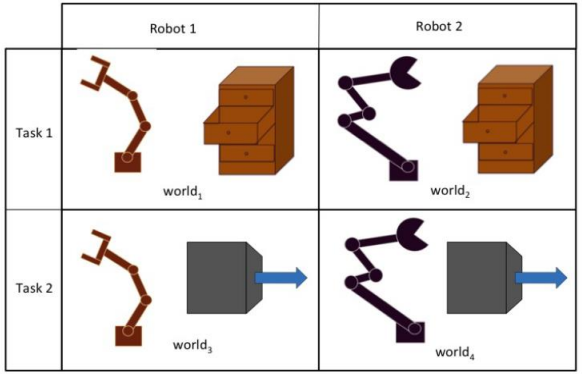


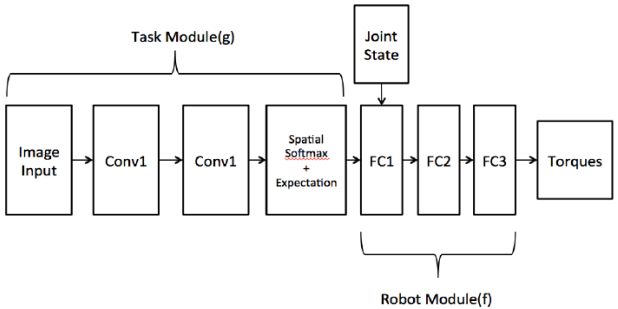
[Hu, Andreas, Rohrbach, Darrell, Saenko; in submission]

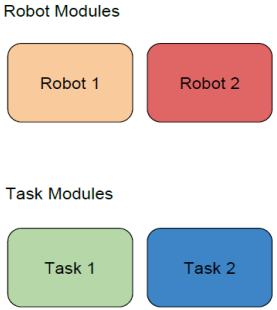
Learning Modular Neural Network Policies for Multi-Task and Multi-Robot Transfer

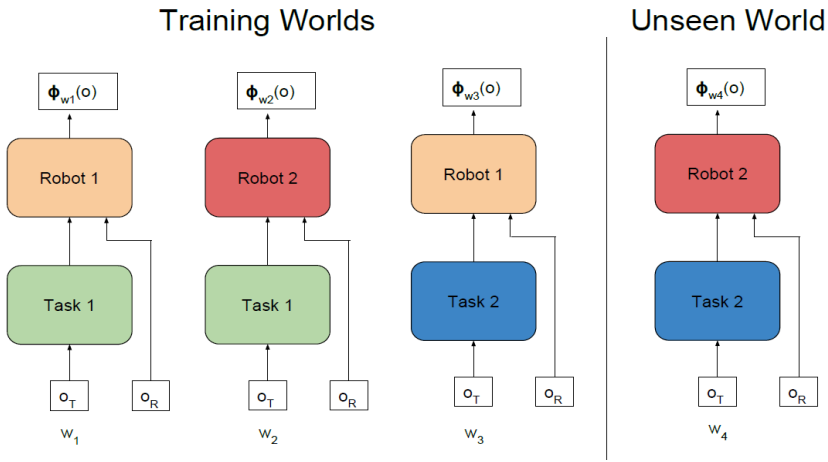
Coline Devin*[1] Abhishek Gupta*[1] Trevor Darrell[1] Pieter Abbeel[1] Sergey Levine[1]











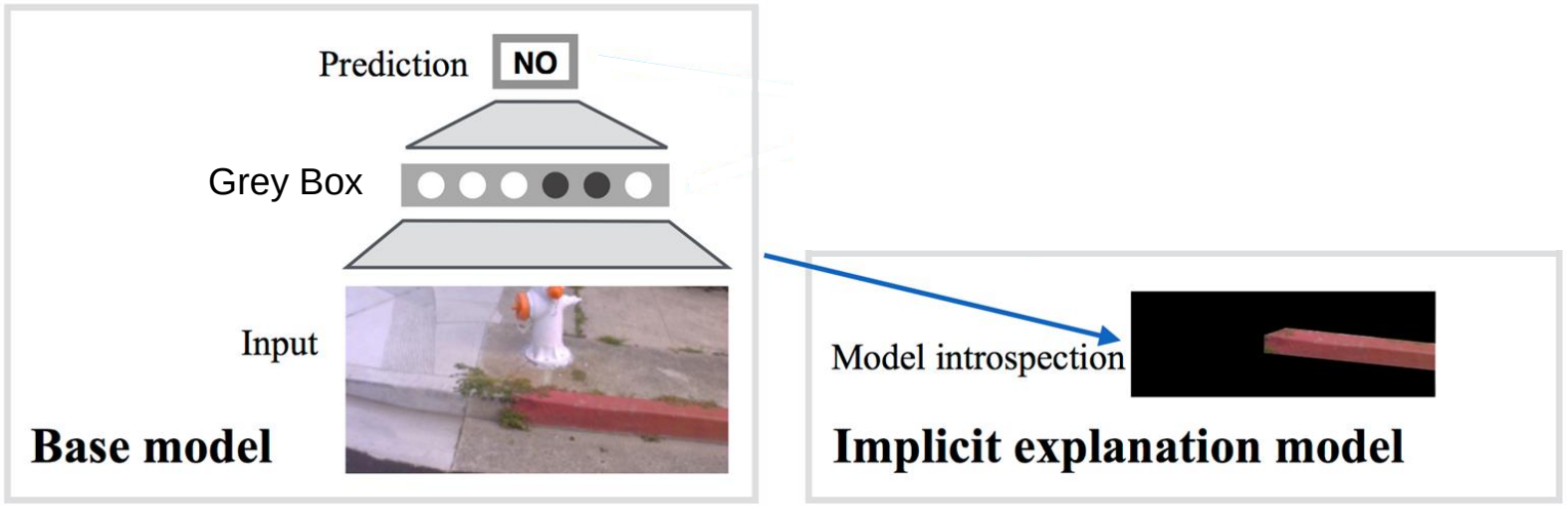
Implicit
Explanation Models

Implicit Explanation Models

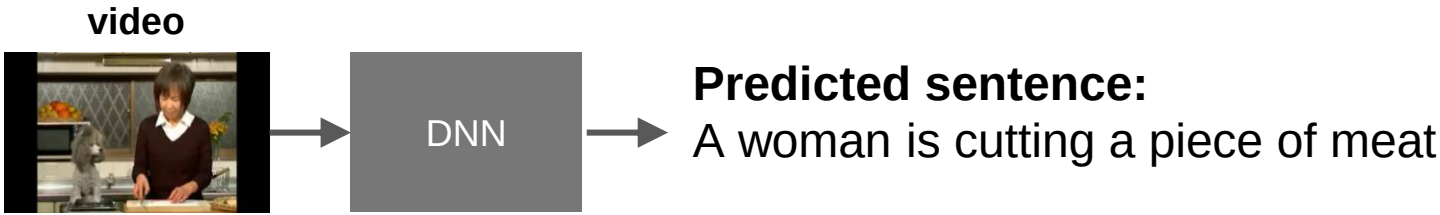
Recover visualizations or exemplars from black/grey box DNN

No explicit internal meaning representation is needed

Q: Can you park here?

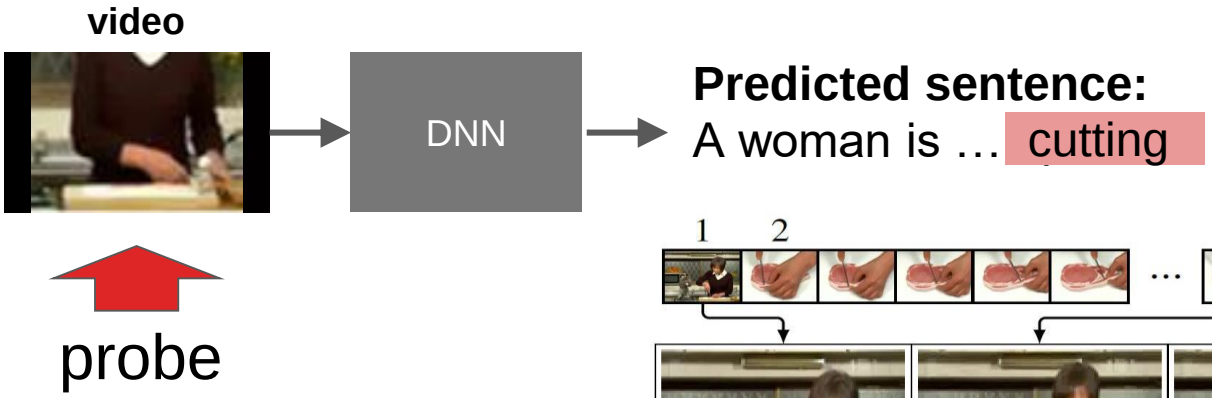


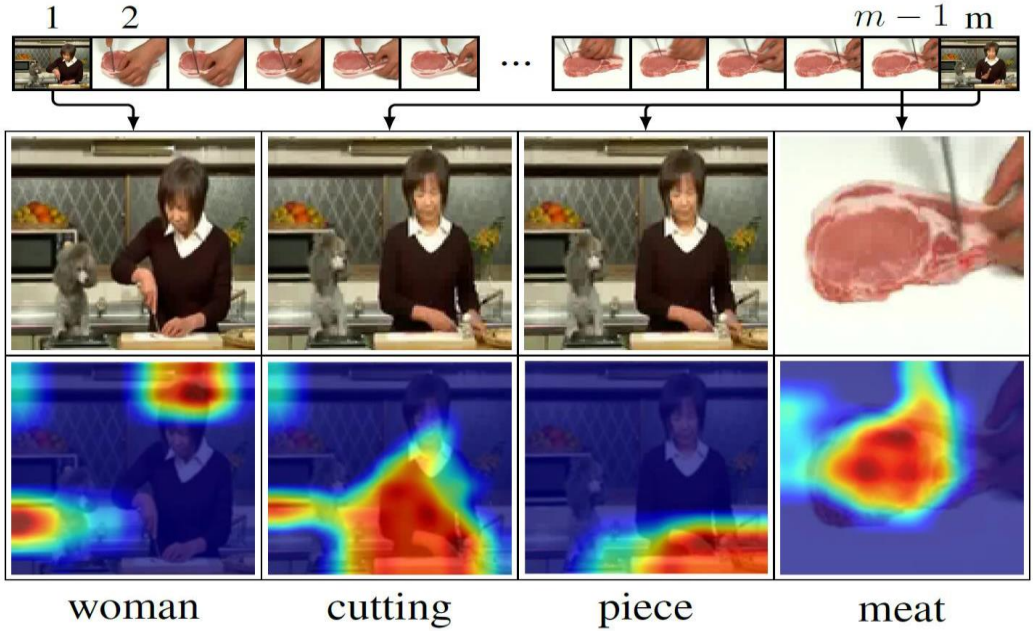
Implicit explanation via Top-down Saliency



[Ramanishka, Das, Zhang, Saenko; CVPR17] [project page]

Implicit explanation via Top-down Saliency

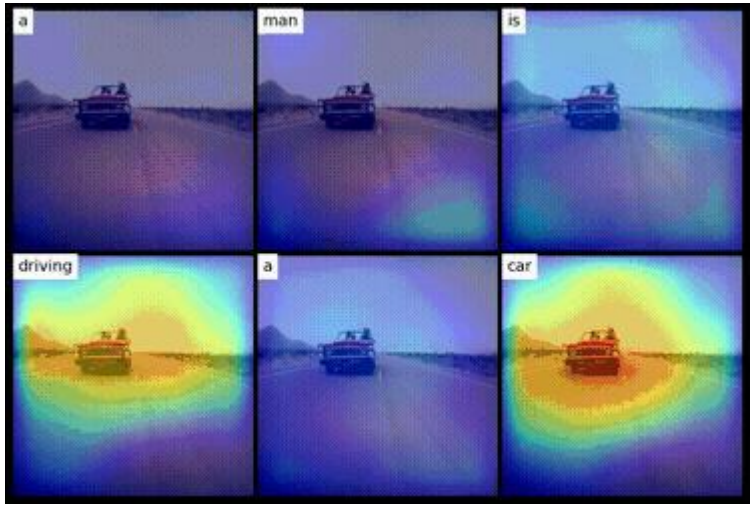




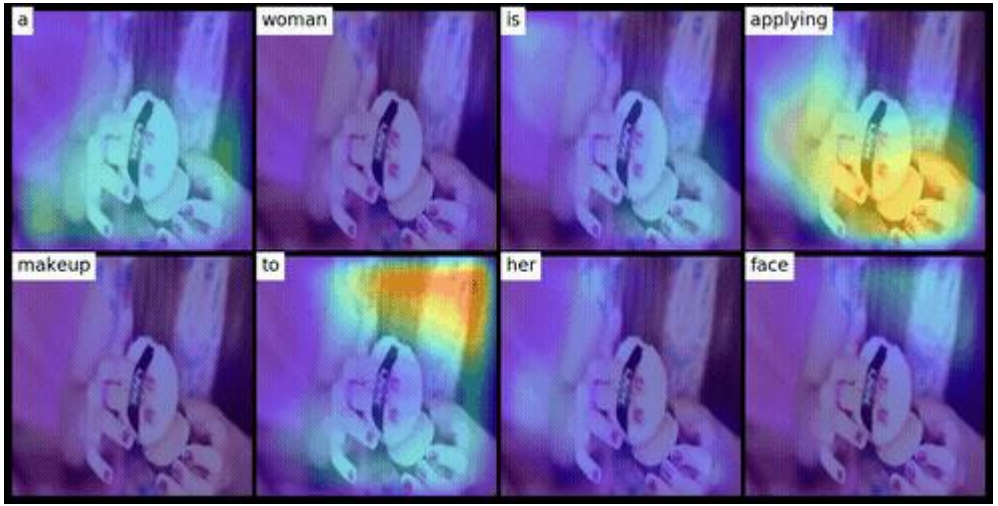
[Ramanishka, Das, Zhang, Saenko; CVPR17] [project page]

Implicit Explanations via Top-down Saliency

Prediction: A man is driving a car



Prediction: A woman is applying makeup to her face



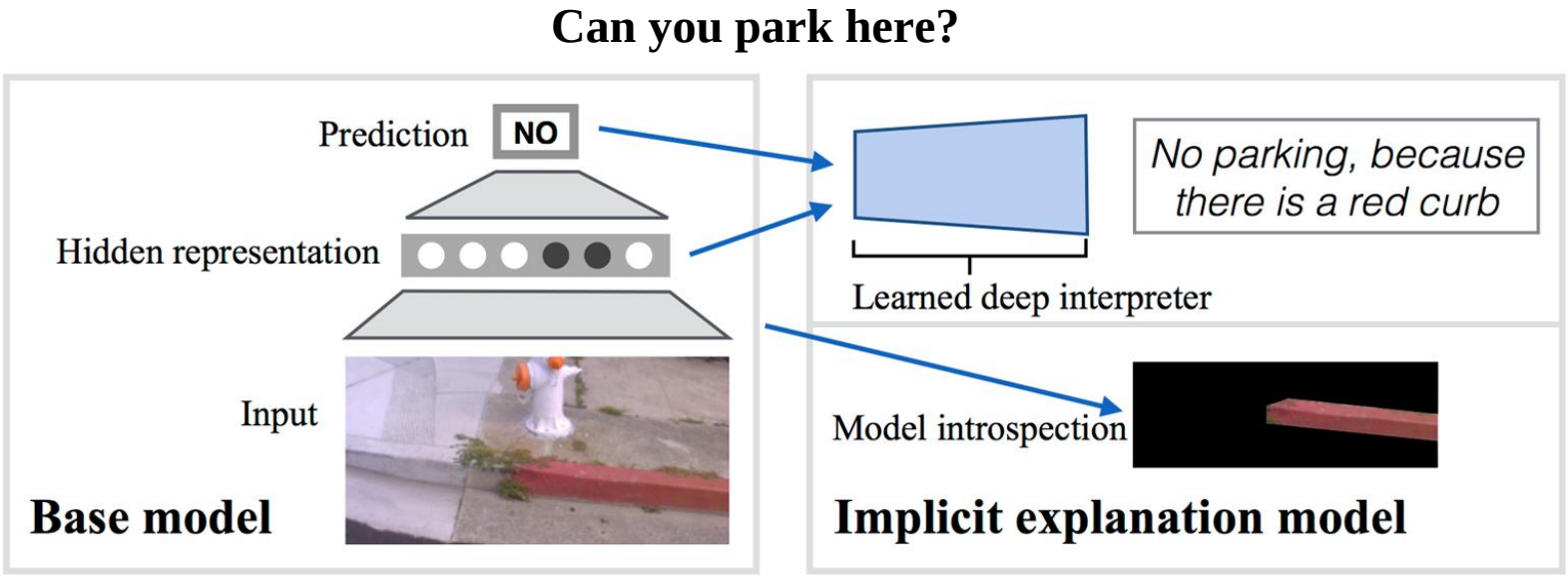
[Ramanishka, Das, Zhang, Saenko; CVPR17] [project page]

Learning how to explain (or, Talking to the User)

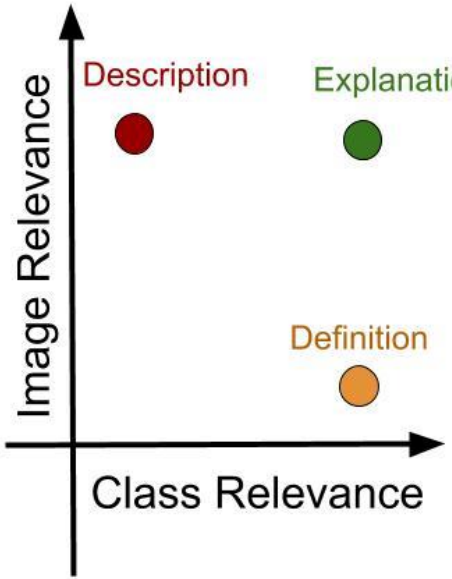
Implicit Model for Textual Explanations

Translate DNN hidden state into

- visualizations and exemplars
- human-interpretable language



Textual Explanations



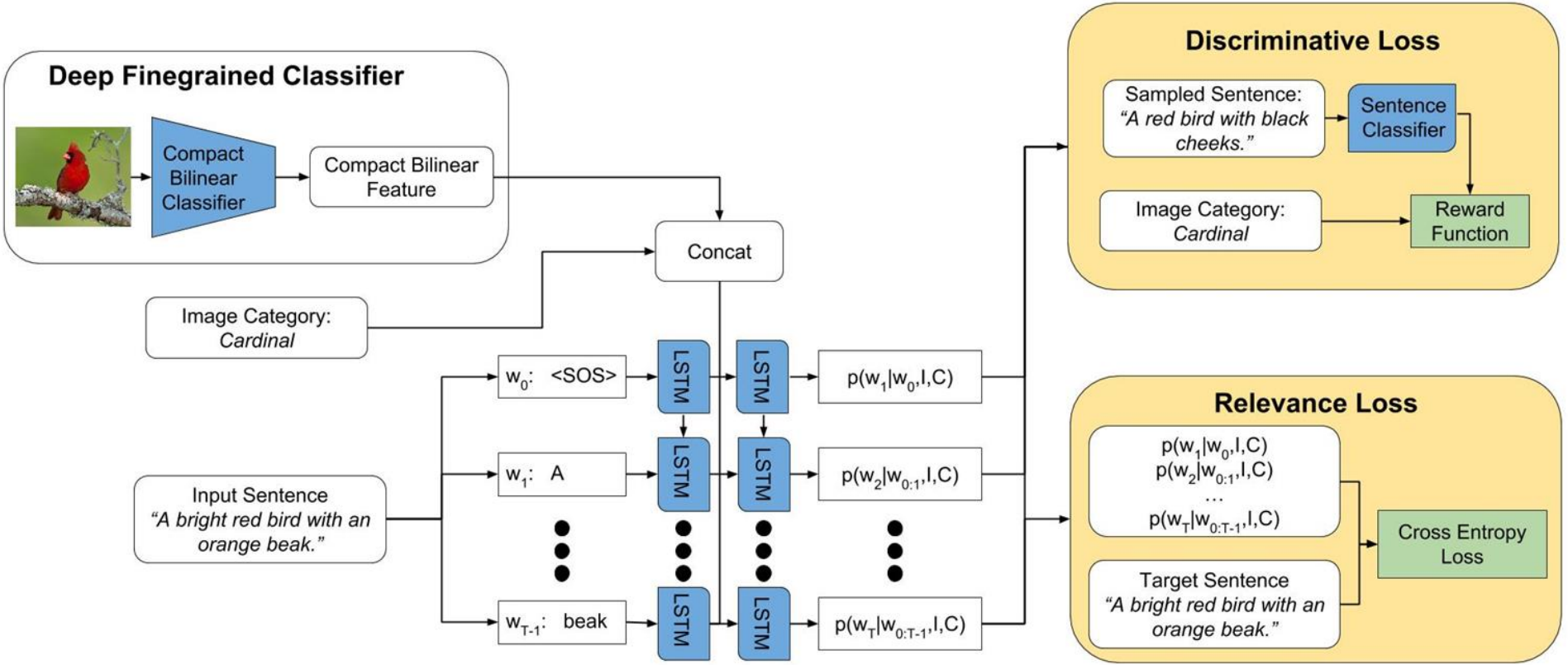
Description: This is a bird with a white neck in the water.
Definition: *Western Grebe* has a yellow pointy beak and a white neck.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Description: This is a large flying bird with black wings.
Definition: *Laysan Albatross* is a seabird with a hooked yellow beak.
Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Description: This is a bird with a white neck in the water.
Definition: *Laysan Albatross* is a seabird with a hooked yellow beak.
Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

[Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell; ECCV 2016]

Generating Textual Explanations



Definition vs. Explanation: Qualitative Results

*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a red spot on its head.

Explanation: this is a black and white bird with a **red spot** on its crown.

*This is a **Downy Woodpecker** because...*

Definition: this bird has a white breast black wings and a red spot on its head.

Explanation: this is a white bird with a black wing and a **black and white** striped head.

Definition vs Explanation: Failure Cases

Correct class: Laysan Albatross, **Predicted class:** Cactus Wren

Explanation: ...this is a **brown and white spotted** bird with a long pointed beak.
Cactus Wren Definition: This bird has a long thin beak with a **brown body** and black spotted feathers.
Laysan Albatross Definition: This bird has a **white head and breast** a grey back and an orange beak.

Correct class: Laysan Albatross, **Predicted class:** Laysan Albatross

Explanation: ...this bird has a **white head and breast** with a long hooked bill.
Laysan Albatross Definition: This bird has a **white head and breast** a grey back and an orange beak.

How to Couple Visual and Textual Data?

CNN

spoon
plate
bowl
table
food
corn
…

Is this going to be a feast?

LSTM

person
Is?
feast
going to be
…

Yes

- **All elements can interact**
- **Multiplicative interaction**

How to Couple Visual and Textual Data?

CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Concatenation

FC

RELU

FC

Yes

☑ **All elements can interact**

? **Multiplicative interaction**

\- **Difficult to learn output classification**

How to Couple Visual and Textual Data?

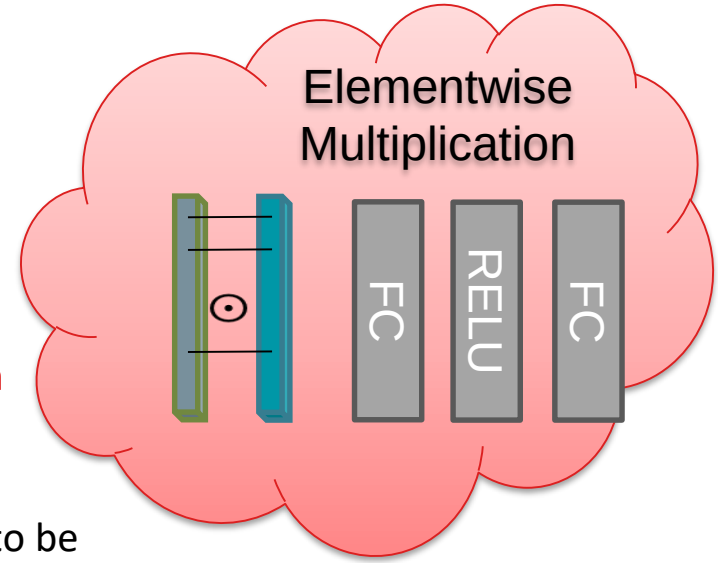
CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...



Yes

❓ **All elements can interact**

☑ **Multiplicative interaction**

- Difficult to learn input embedding

How to Couple Visual and Textual Data?

CNN

spoon
plate
bowl
table
food
corn
…
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
…

Outer Product / Bilinear Pooling

FC

Yes

- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ? **High #activations & computation**
- ? **High #parameters**

[Lin, RoyChowhudry, Maji; ICCV 2015]

How to Couple Visual and Textual Data?

CNN

spoon
plate
bowl
table
food
corn
…
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
…

Compact Bilinear Pooling

MCB

FC

Yes

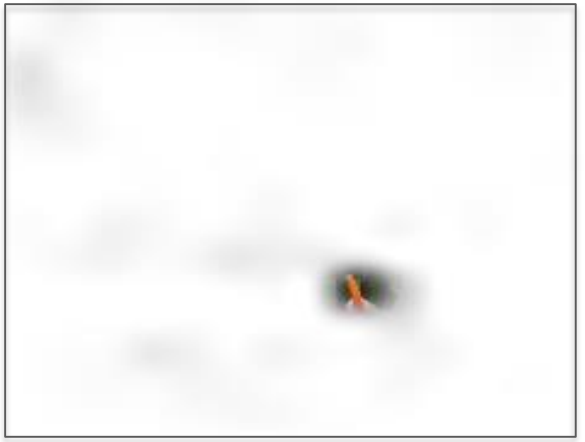
- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☑ **Low #activations & computation**
- ☑ **Low #parameters**

[Zhang, Shelhamer, Gao, Darrell; ICLR workshop 2016]
[Gao, Beijbom, Zhang, Darrell, CVPR 2016]

Attention Visualizations

What is the woman feeding the giraffe?

Carrot



Attention Visualizations

What color is her shirt?

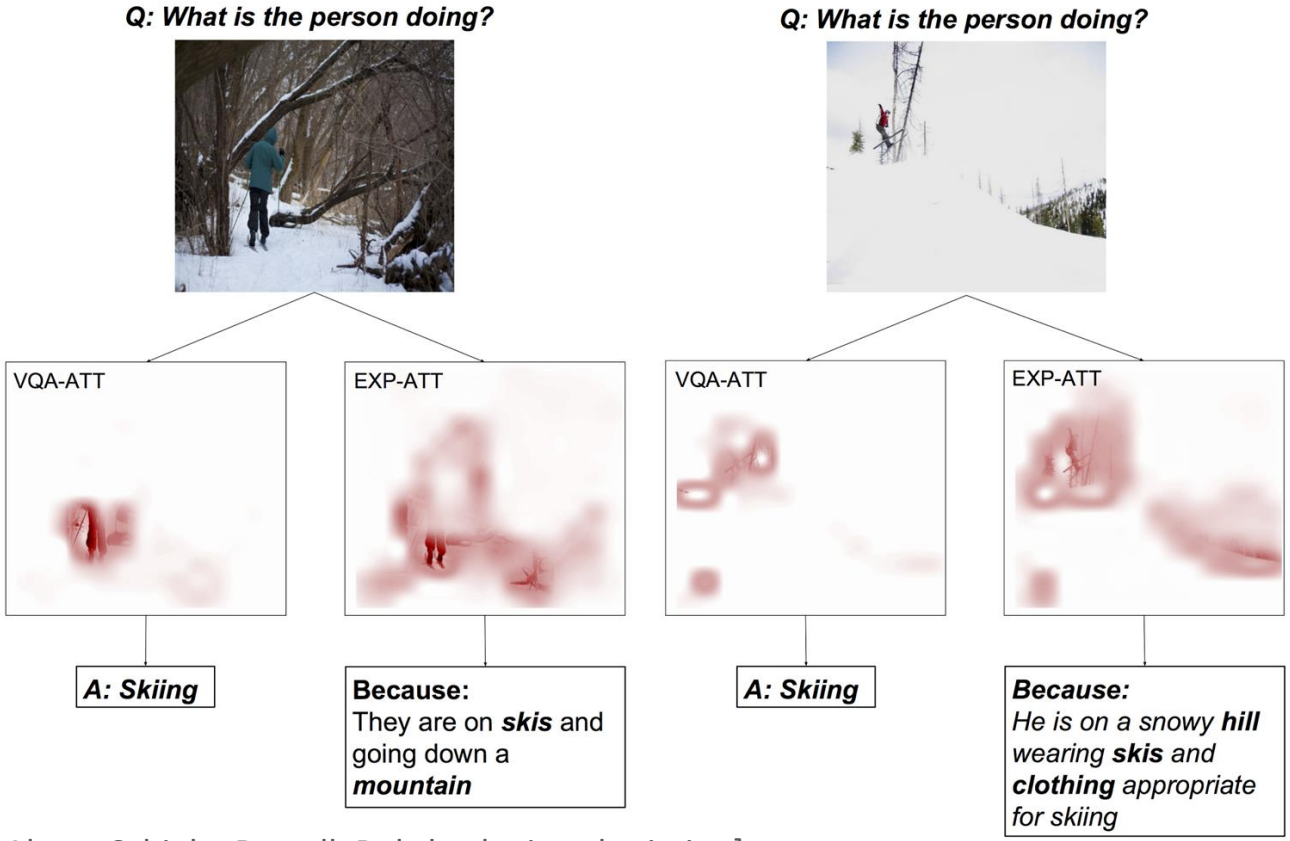
Purple

Attention Visualizations

What is her hairstyle for the picture?

Ponytail

Justifying Decisions and Pointing to the Evidence

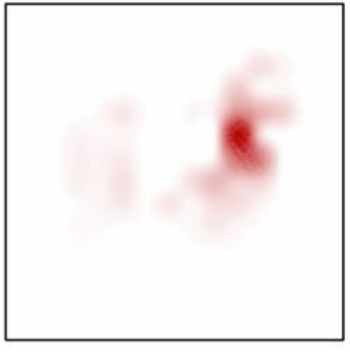


[Park, Hendricks, Akata, Schiele, Darrell, Rohrbach; in submission]

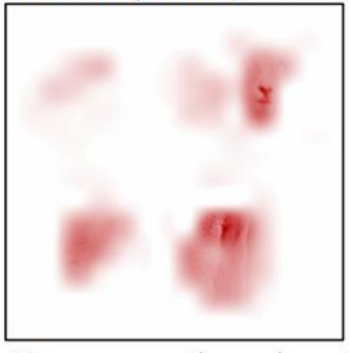
Attentive Explanations of VQA

Q: What type of animal is this?

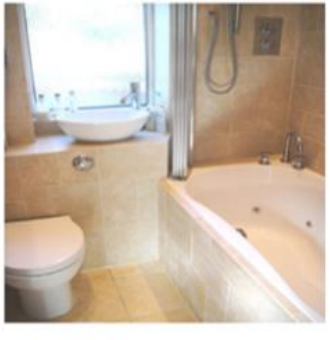
A: Sheep



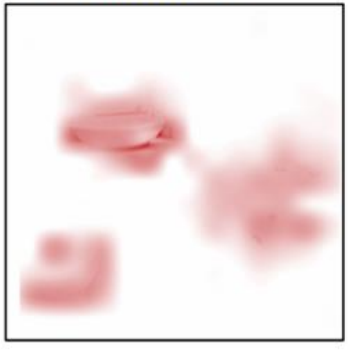
Because... it has four legs and long fluffy hair



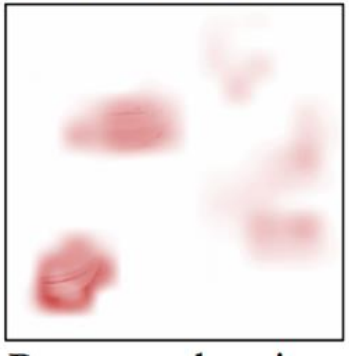
Q: What room is this?



A: Bathroom

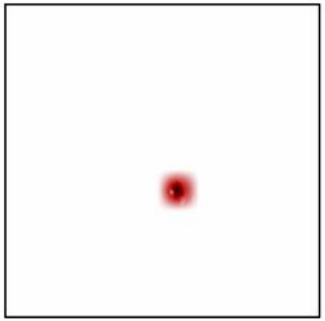


Because... there is a toilet and sink in the room

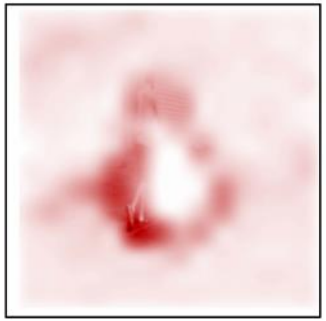


Attentive Explanations of Activities

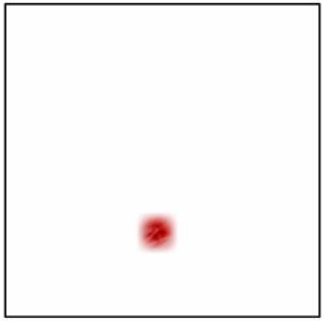
I can see that he is mowing lawn



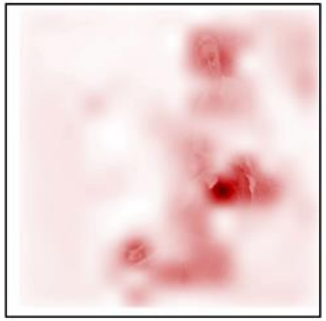
Because... he is pushing a lawn mower over a grassy lawn



I can see that he is mowing lawn



Because... he is kneeling i the grass next to a lawn mower



Modeling the User

Reflexive and Rational Models

Reflexive Agents

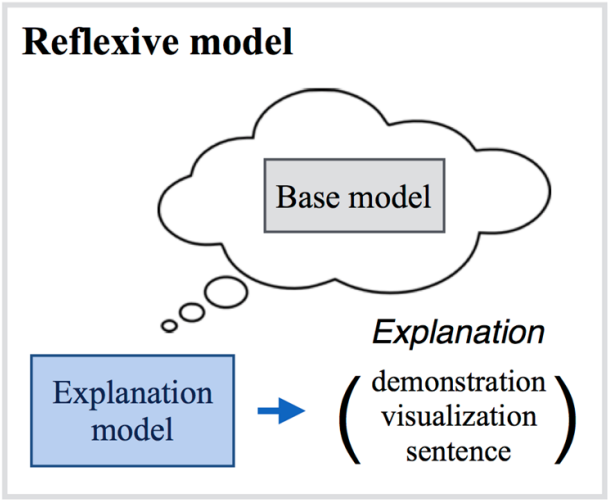
from examples by humans

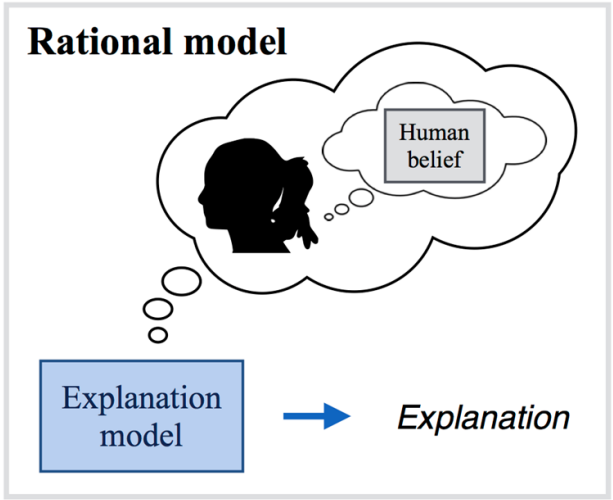
independent of user state.

Rational Agents

how the system makes predictions

why a mistake was made in a scenario



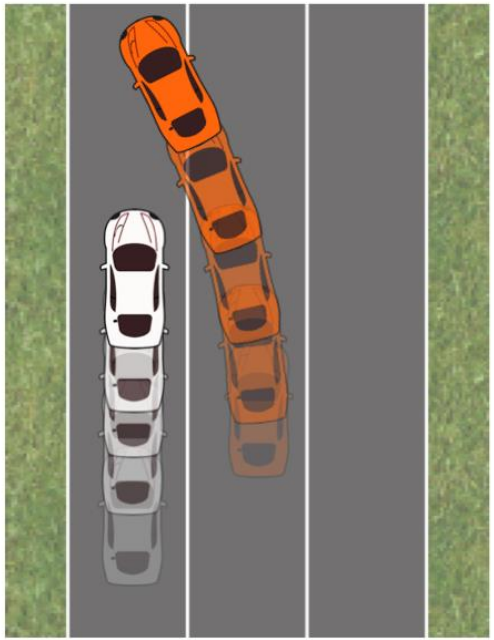


Explainable Models for Dynamic Policies

Dynamic policies

provide example state/action sequences

help user understand model behavior.



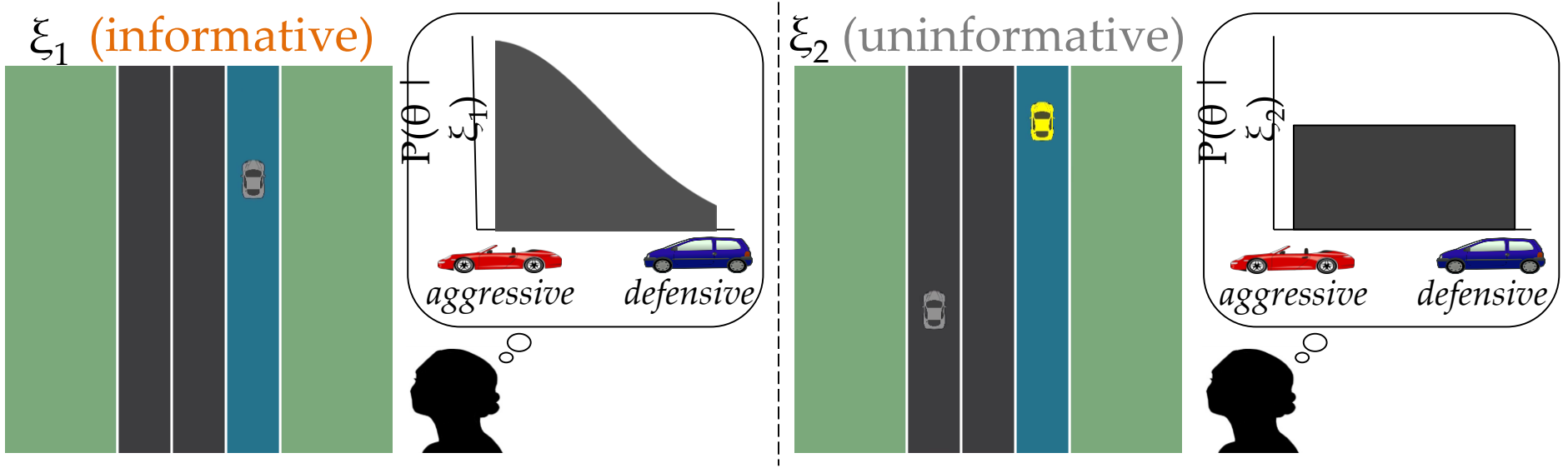
(a) Explain learned policy via example rollouts

How can we help users better anticipate what a robot will do?

Key insight:
Users need to understand the *tradeoffs* that a robot makes

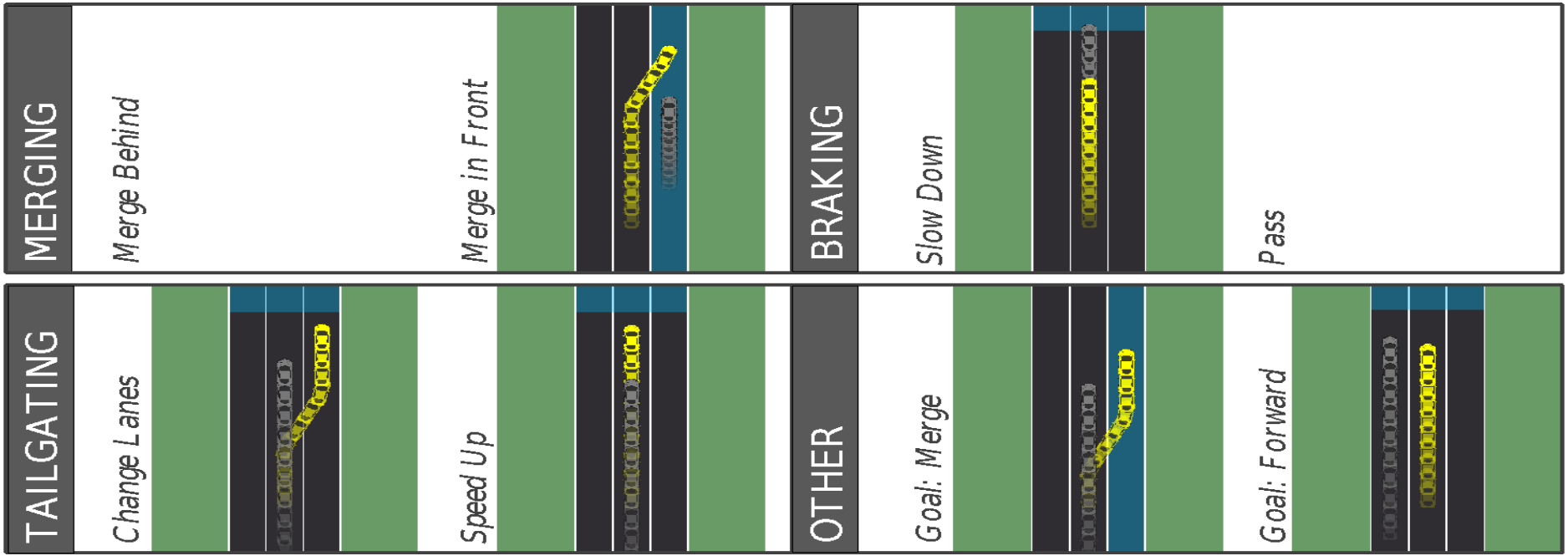
Robots Inevitably Communicate via Their *Behavior*

Humans naturally reason about others' utility functions*



*Jara-Ettinger, J., et al. Trends in Cognitive Sciences (2016)

Experiment Setup: Environments



User Study: Noise Models

Current Test Environment

Please watch all four video clips below. One video clip shows Carl driving, and the other three show imposter cars that look the same as Carl but drive differently. (Click on the video to start it.)

Which video clip do you think shows Carl driving?

○ Video 1 ○ Video 2 ○ Video 3 ○ Video 4

To what extent do you agree or disagree with the following statements?

The car in **Video 1** drives in a similar way as Carl.

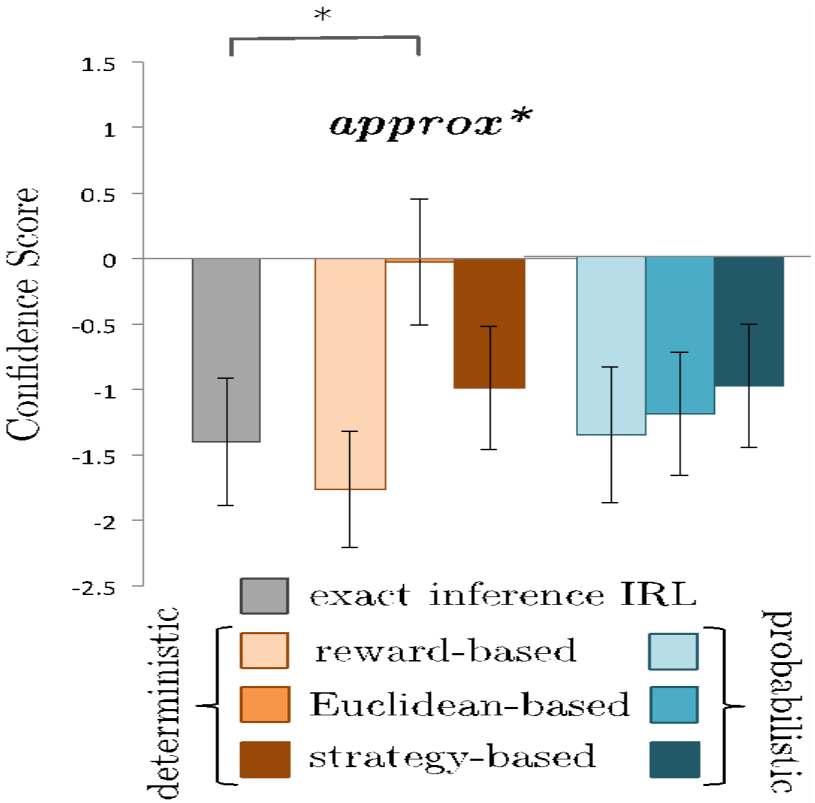
○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neither agree nor disagree ○ Somewhat agree ○ Agree ○ Strongly agree

The car in **Video 2** drives in a similar way as Carl.

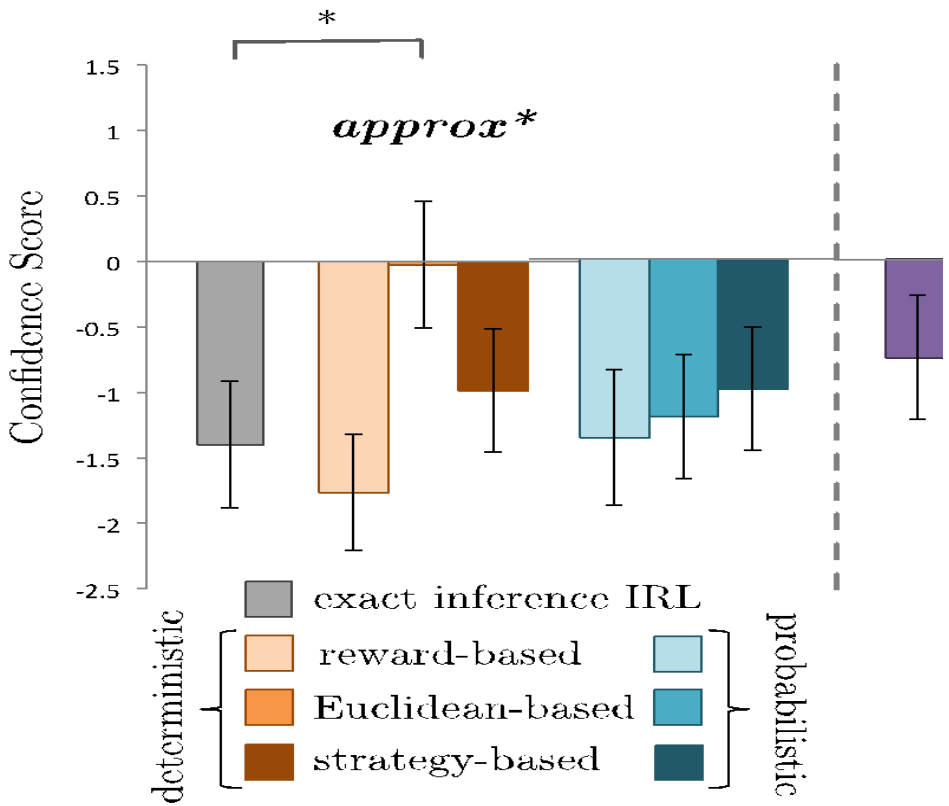
○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neither agree nor disagree ○ Somewhat agree ○ Agree ○ Strongly agree

• • •

User Study: Noise Models



User Study: Noise Models

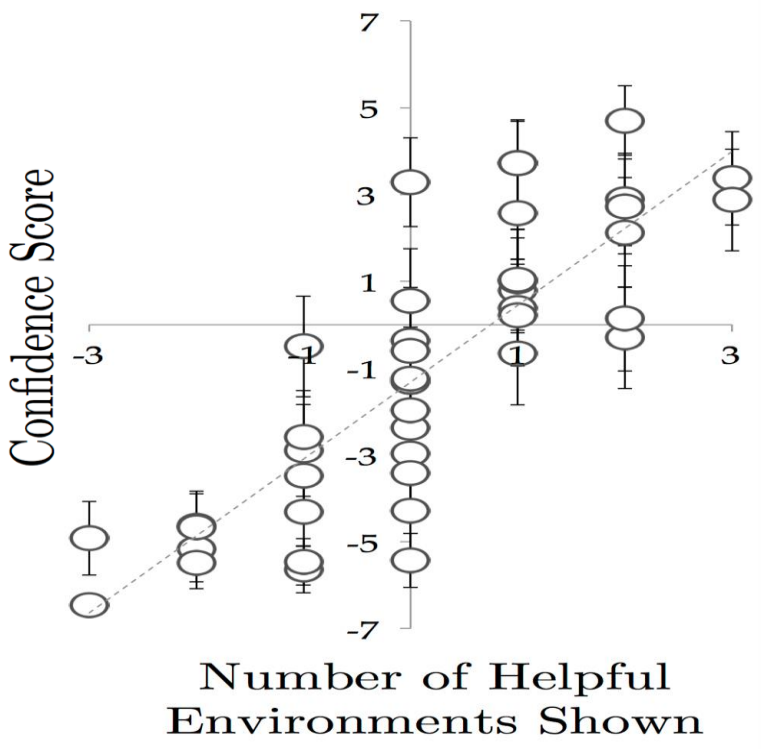


Coverage Matters

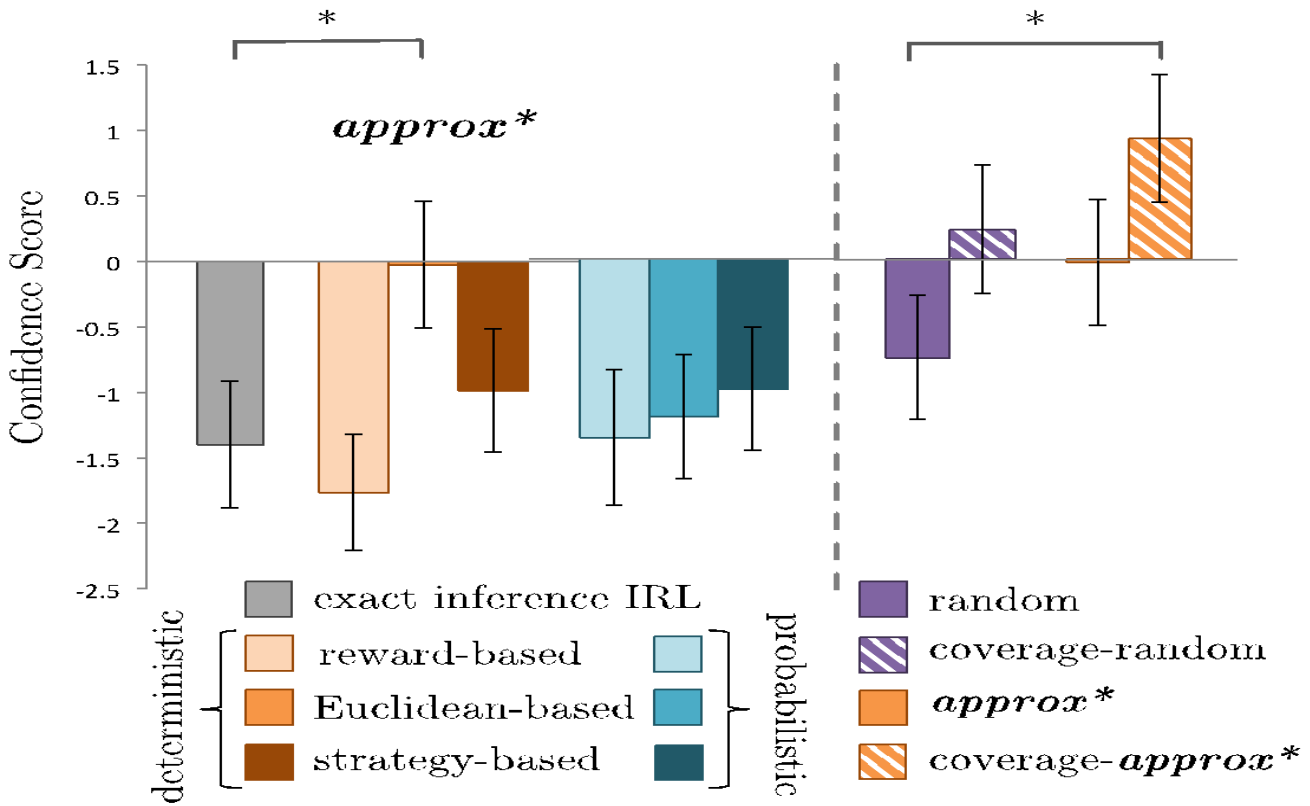
Given x examples shown in strategy A and y from B:

 # helpful environments shown in A = $\begin{cases} x, & \text{if } x > 0. \\ -y, & \text{otherwise.} \end{cases}$

Pearson's $r = 0.83$



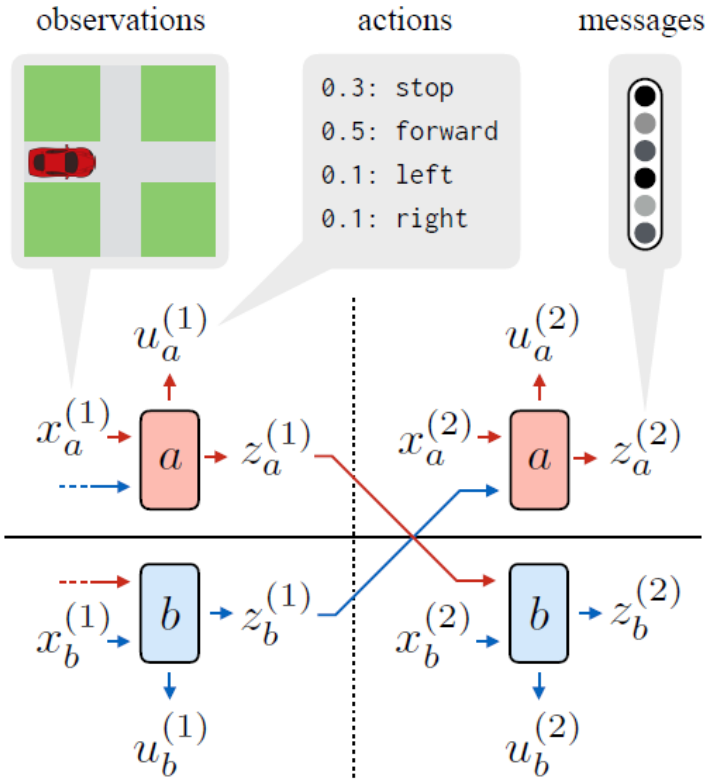
User Study: Coverage



What is Neuralese?

Idea: induce communication protocols for multiagent tasks

Example: navigating a contested intersection



Translating Neuralese

- Idea: infer a mapping between neuralese and natural language
  - (1) Have neural system learn a code for a task
  - (2) Have humans do the same task, communicating in NL
  - (3) Compute mapping from neuralese to NL using *belief matching*

- Belief matching
  - Q: How do you know x in neuralese means the same as y in NL?
  - A: If they induce the optimally similar distributions over belief states

$$q(z, z') = \mathbb{E}\left[\mathcal{D}_{\mathrm{KL}}(\beta(z, X_b) \,||\, \beta(z', X_b)) \mid z, z'\right]$$

  - (Applies more generally to knowing two messages mean the same thing)

Examples

Task: Distinguish the starred image

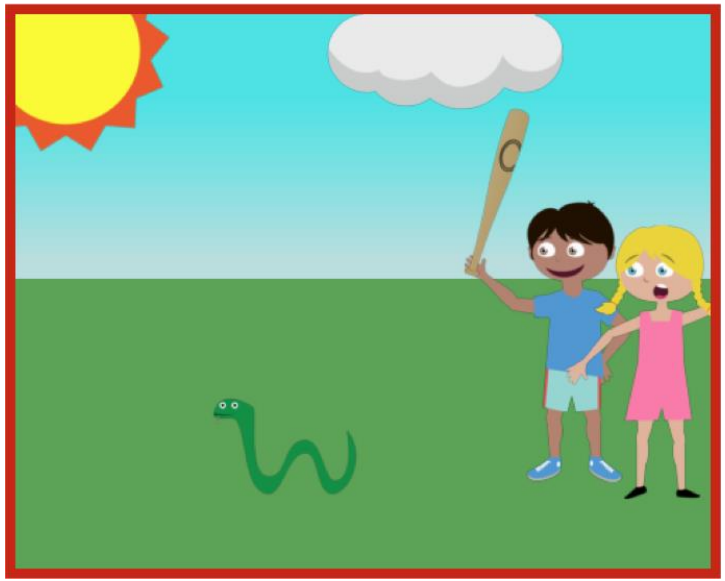
large bird, black wings, black crown

Task: Avoid collision with invisible car

you first, following, going down

The reference game

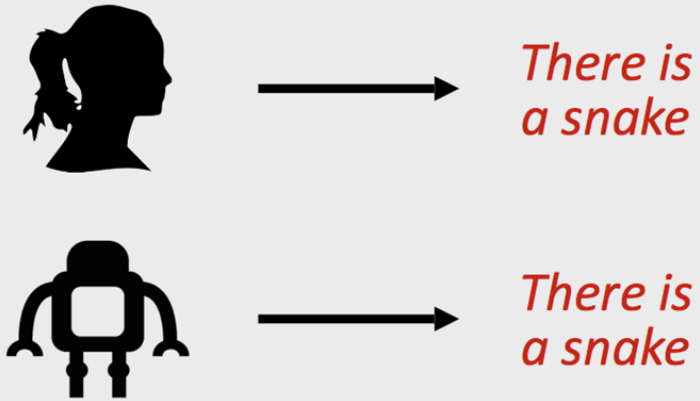
Mike is holding a baseball bat



The reference game

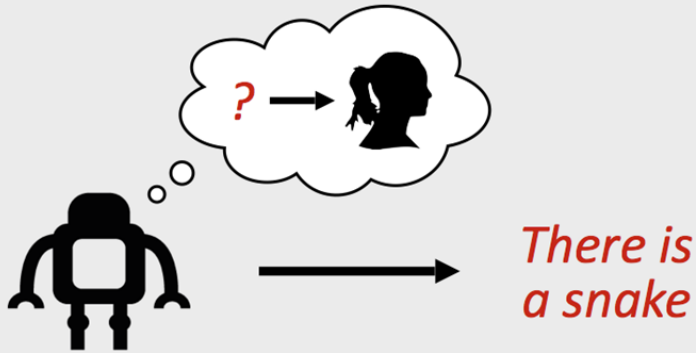
DIRECT APPROACH:

Imitate successful human play

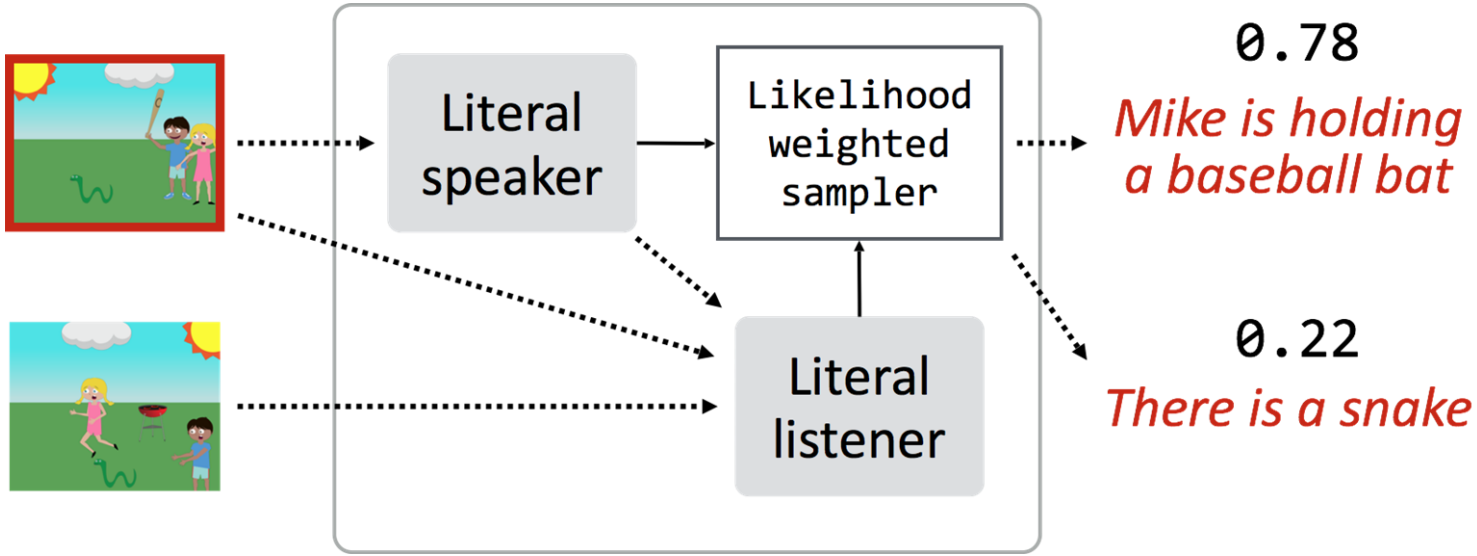


DERIVED APPROACH:

Reason about listener beliefs



The reference game



Program Schedule

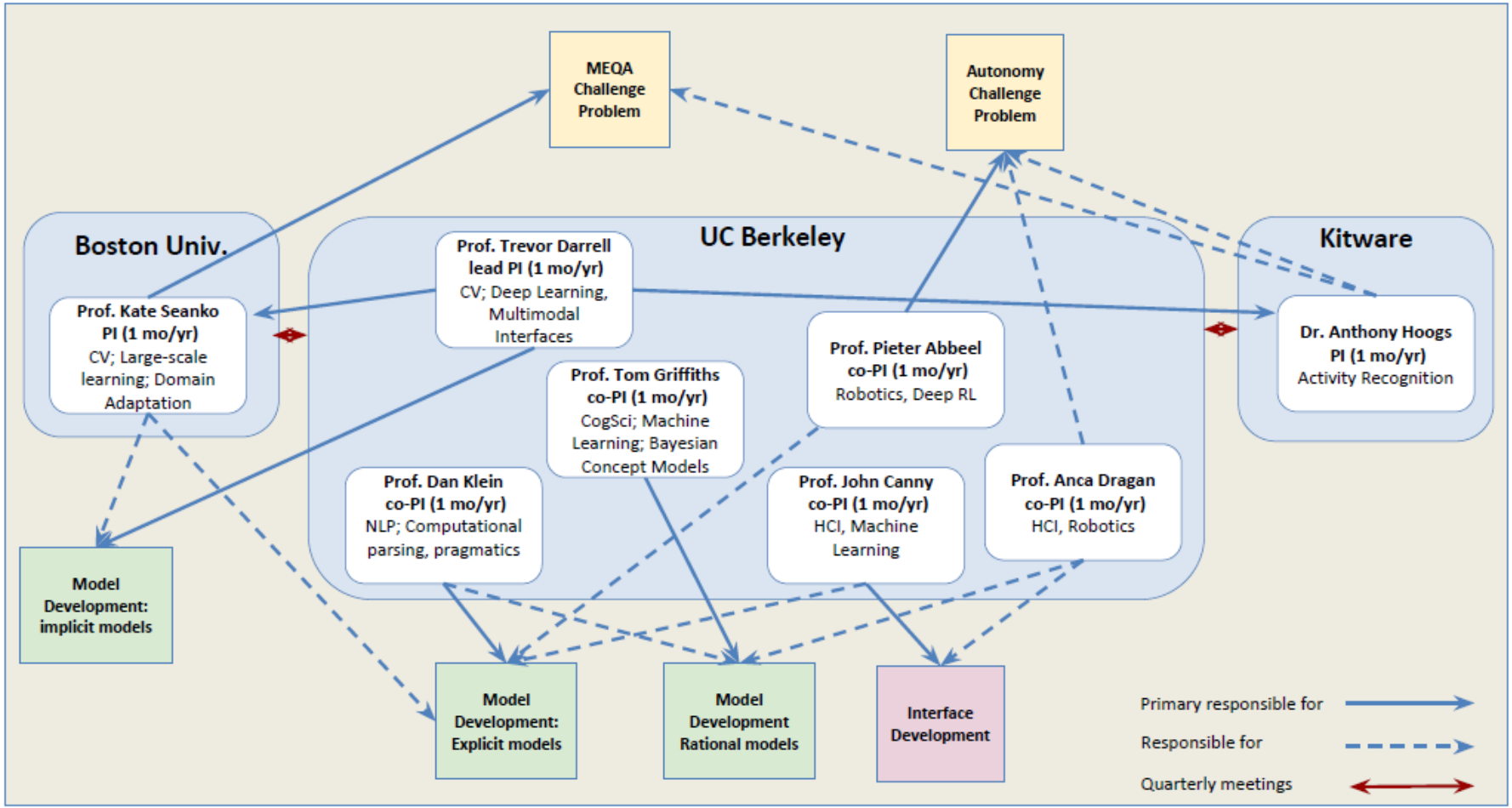
Program Schedule

Phase I

- algorithmic development and detailed specification
- definition, and initial evaluation of the above challenge areas and datasets

Phase II

- performance improvements
- system demonstrations
- in-situ user-oriented evaluations
- preparation for transition


MEQA Challenge Problem
Autonomy Challenge Problem
Boston Univ.
Prof. Kate Seanko
PI (1 mo/yr)
CV; Large-scale learning; Domain Adaptation
UC Berkeley
Prof. Trevor Darrell
lead PI (1 mo/yr)
CV; Deep Learning, Multimodal Interfaces
Prof. Tom Griffiths
co-PI (1 mo/yr)
CogSci; Machine Learning; Bayesian Concept Models
Prof. Dan Klein
co-PI (1 mo/yr)
NLP; Computational parsing, pragmatics
Prof. Pieter Abbeel
co-PI (1 mo/yr)
Robotics, Deep RL
Prof. John Canny
co-PI (1 mo/yr)
HCI, Machine Learning
Prof. Anca Dragan
co-PI (1 mo/yr)
HCI, Robotics
Kitware
Dr. Anthony Hoogs
PI (1 mo/yr)
Activity Recognition
Model Development: implicit models
Model Development: Explicit models
Model Development Rational models
Interface Development
Primary responsible for
Responsible for
Quarterly meetings

Thank you!