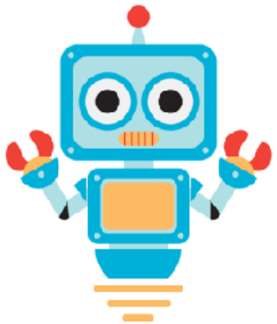
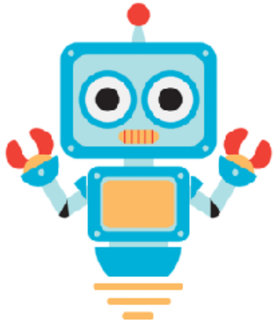
It Takes Two to Tango: Towards Theory of AI's Mind



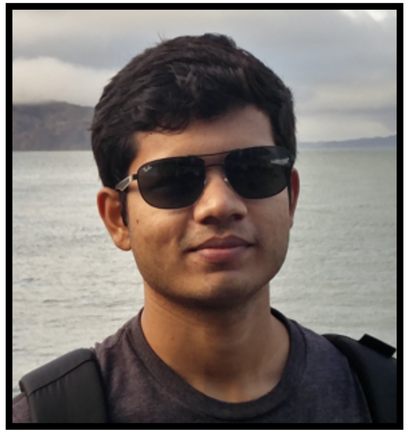
It Takes Two to Tango: Towards Theory of AI's Mind



Arjun Chandrasekaran*
Georgia Tech

Deshraj Yadav*
Georgia Tech

Prithvijit Chattopadhyay*
Georgia Tech



Viraj Prabhu*
Georgia Tech

Devi Parikh
Georgia Tech

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

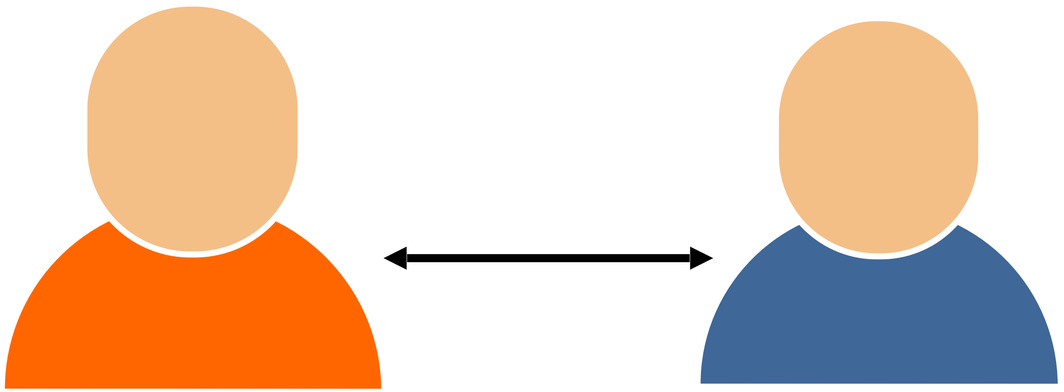
Theory of Mind (ToM)

Theory of Mind (ToM)

- Ability to attribute mental states to others

Theory of Mind (ToM)

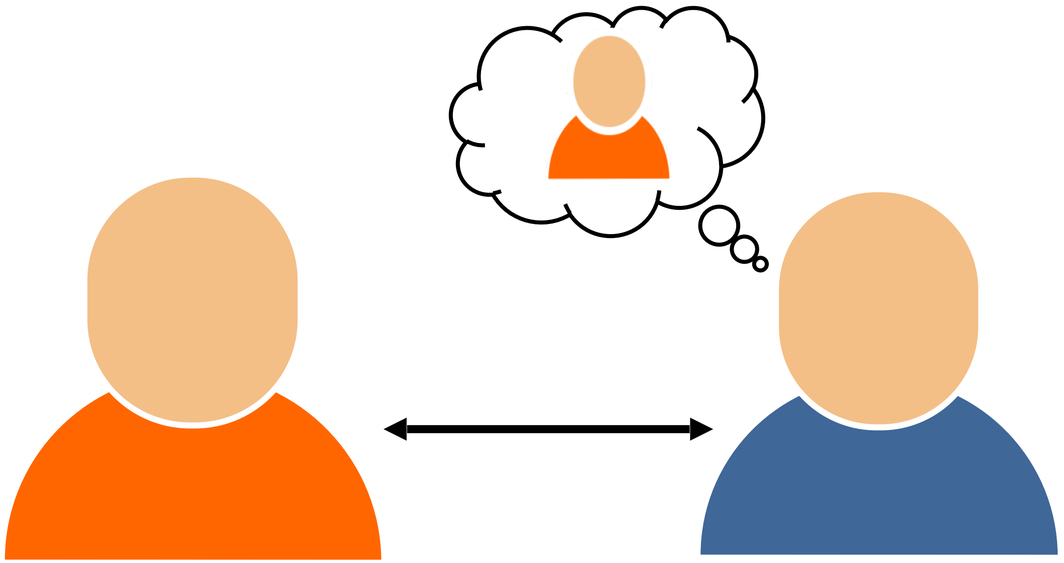
- Ability to attribute mental states to others



Humans attribute mental states to fellow humans

Theory of Mind (ToM)

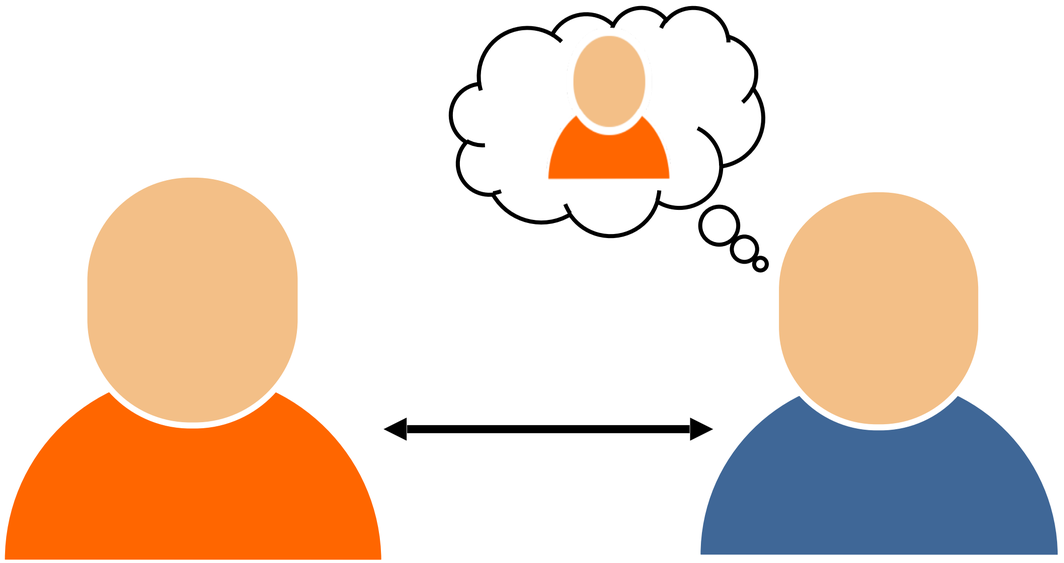
- Ability to attribute mental states to others



Humans attribute mental states to fellow humans

Theory of Mind (ToM)

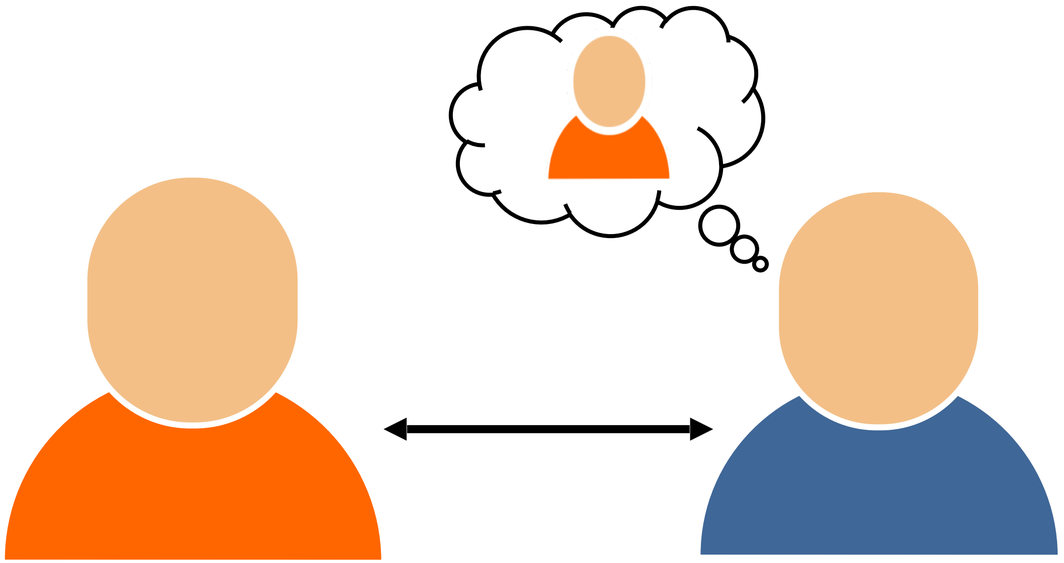
- Ability to attribute mental states to others



Humans attribute mental states to fellow humans → Make reasonable inferences about their behavior

Theory of Mind (ToM)

- Ability to attribute mental states to others



Humans attribute mental states to fellow humans → Make reasonable inferences about their behavior

- Crucial for collaborative team-performance

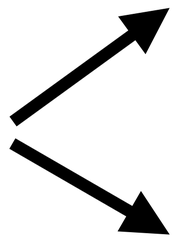
Motivation

Motivation

- Traditional AI Research

Motivation

- Traditional AI Research

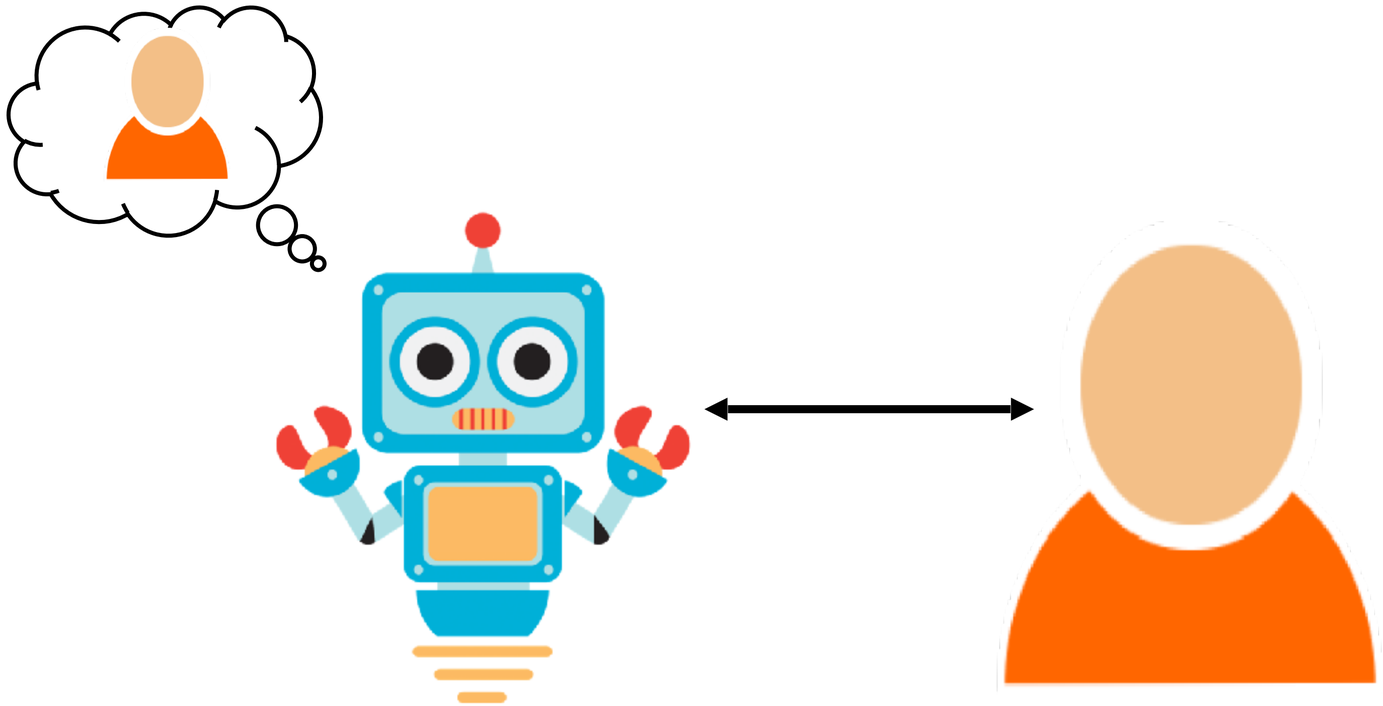


AI more accurate

AI more *human*-like

Motivation

- Traditional AI Research
  - AI more accurate
  - AI more *human*-like



Motivation

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

Motivation

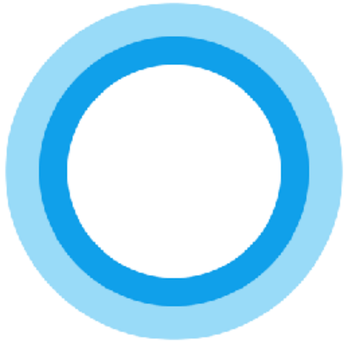
- As AI progresses ⟶ rise of collaborative work with AI agents

Siri

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

Siri

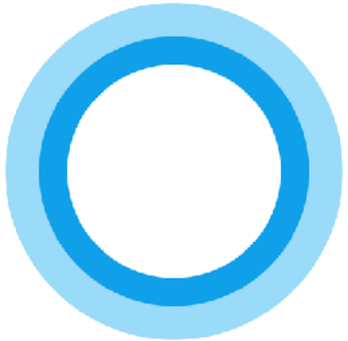


Cortana

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

Siri



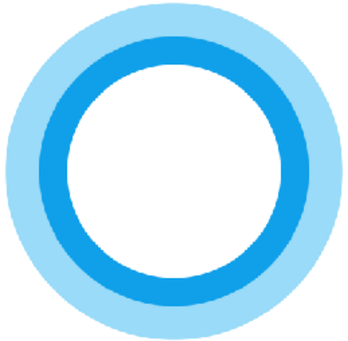
Cortana

Google Assistant

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

Siri



Cortana

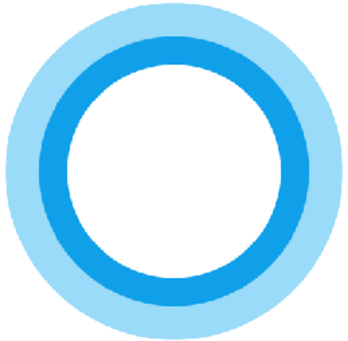
Google Assistant

⟵ Personal Assistants

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

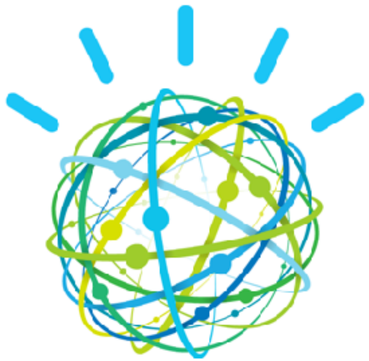
Siri



Cortana

Google Assistant

⟵ Personal Assistants

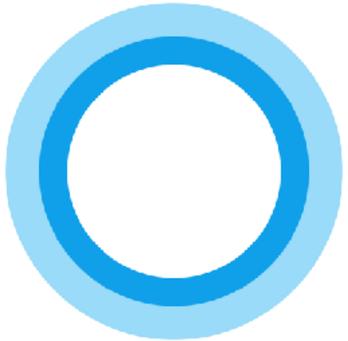


IBM Watson

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

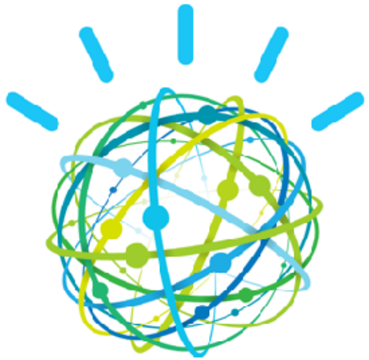
Siri



Cortana

Google Assistant

Personal Assistants



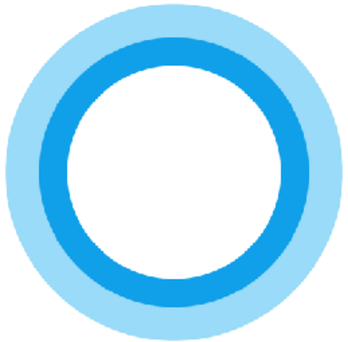
IBM Watson

Self-driving Cars

Motivation

- As AI progresses ⟶ rise of collaborative work with AI agents

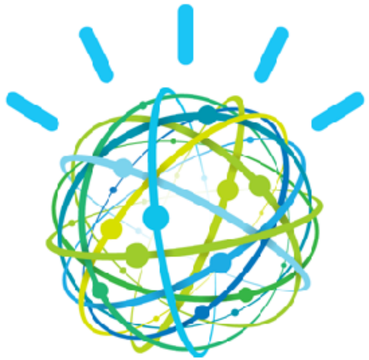
Siri



Cortana

Google Assistant

⟵ Personal Assistants



IBM Watson

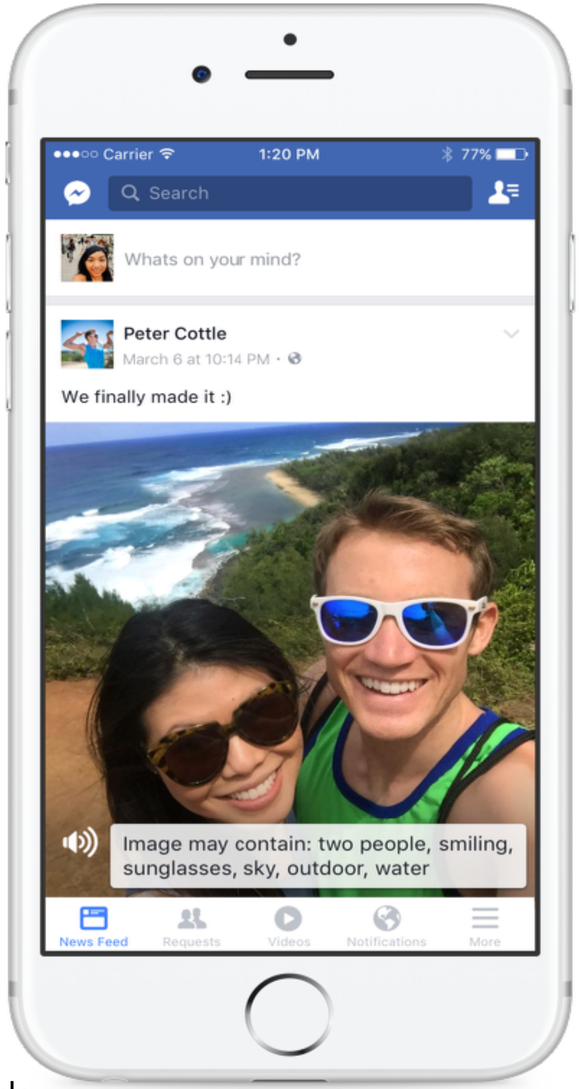
Self-driving Cars

⟵ Sensitive Applications

Motivation

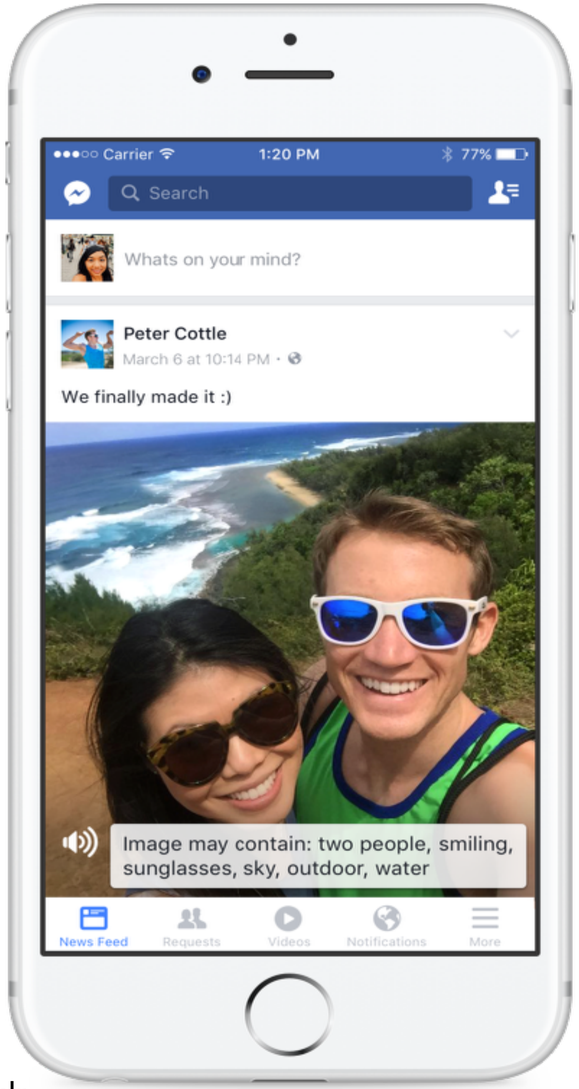
Motivation

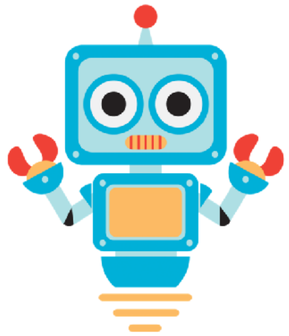
- Aiding visually impaired users



Motivation

- Aiding visually impaired users

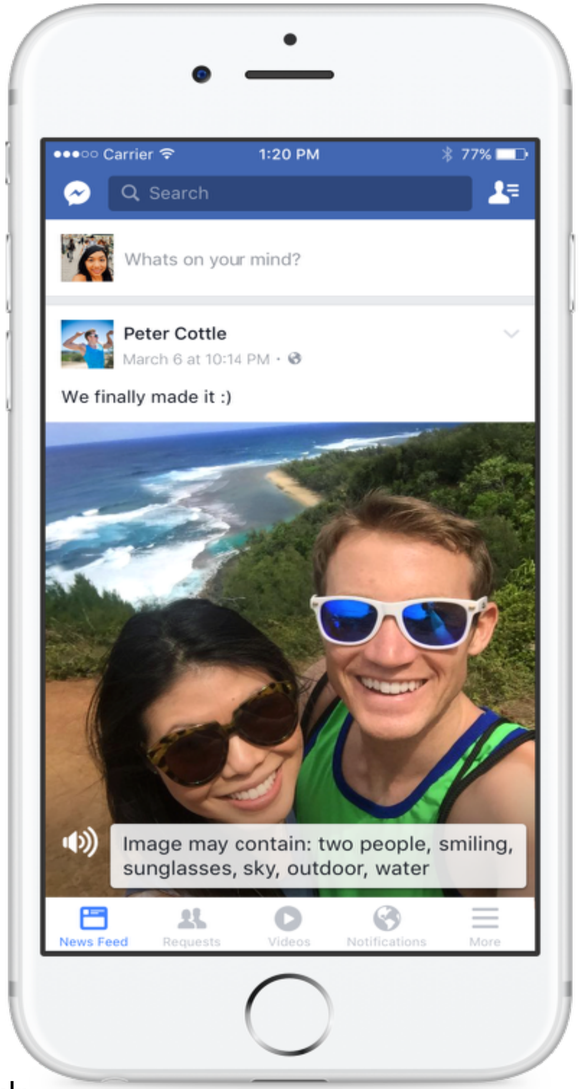


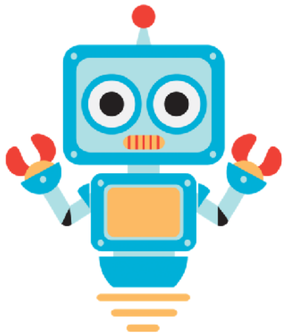


Peter just uploaded a picture from his vacation in Hawaii

Motivation

- Aiding visually impaired users



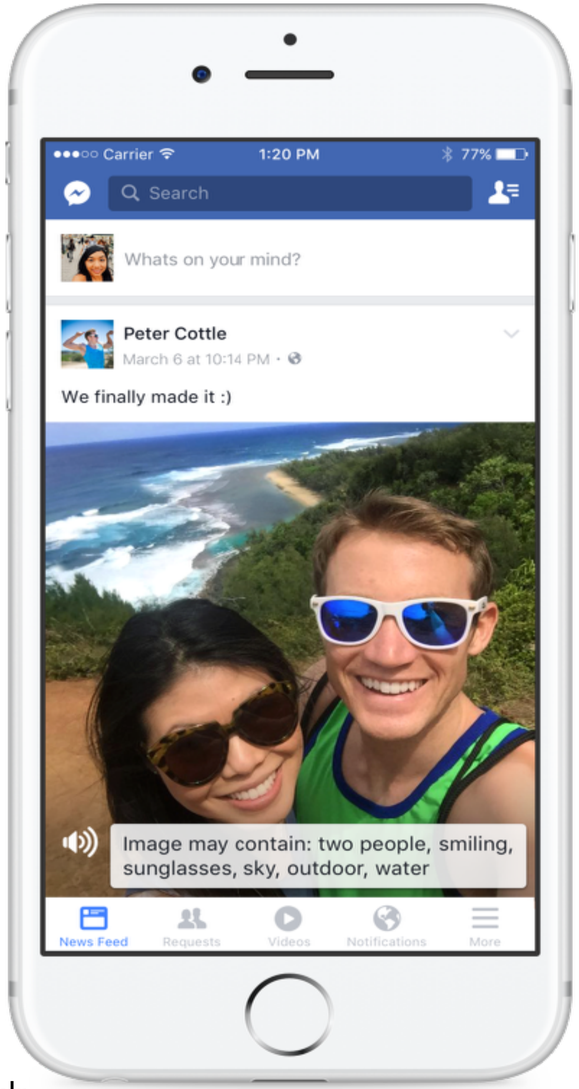


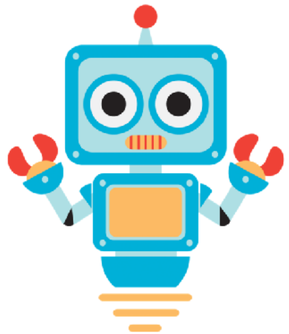
Peter just uploaded a picture from his vacation in Hawaii

Great, is he at the beach?

Motivation

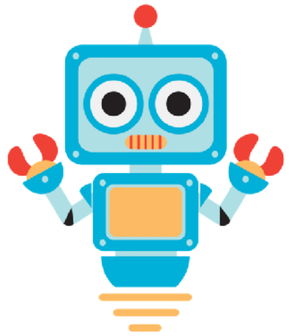
- Aiding visually impaired users





Peter just uploaded a picture from his vacation in Hawaii

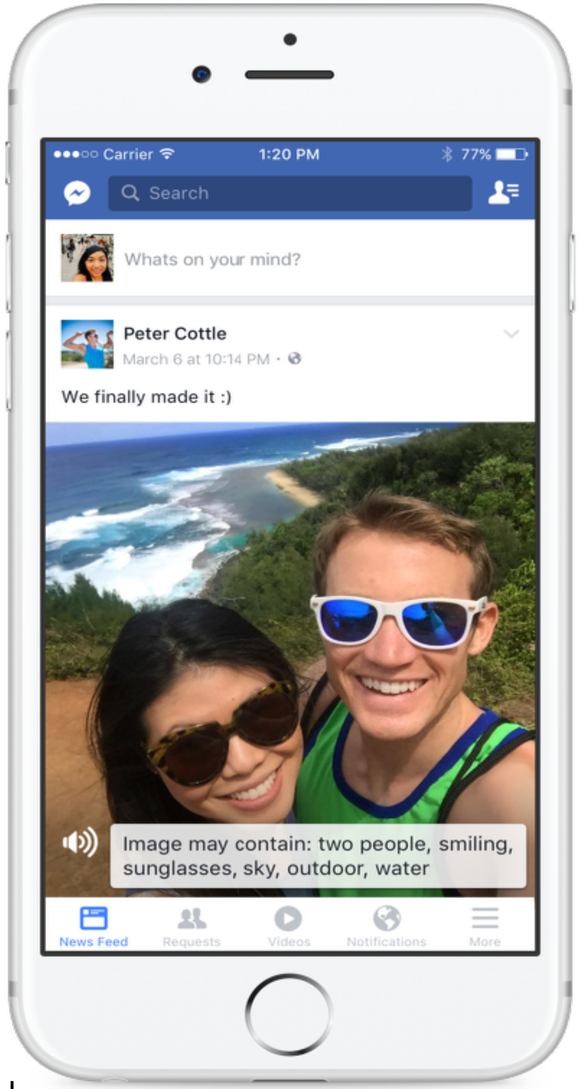
Great, is he at the beach?

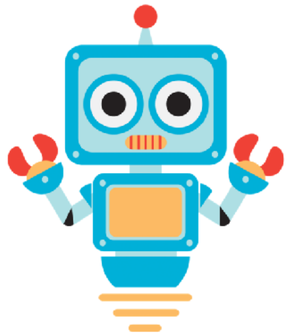


No, on a mountain

Motivation

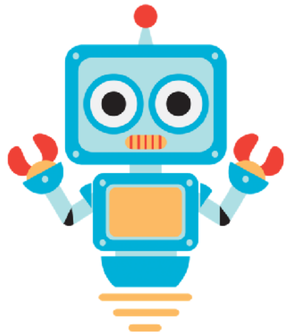
- Aiding visually impaired users





Peter just uploaded a picture from his vacation in Hawaii

Great, is he at the beach?



No, on a mountain

.....

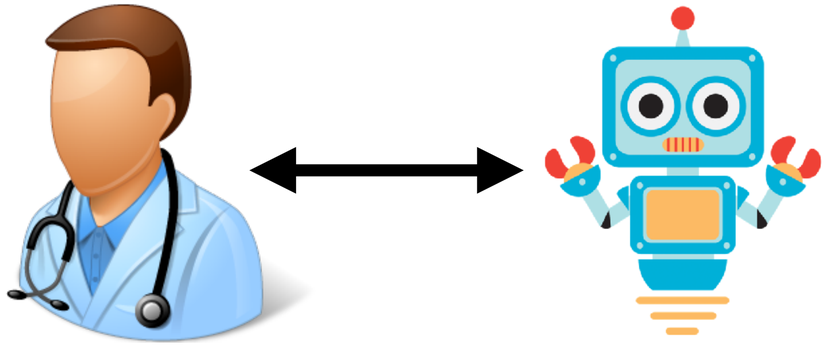
Motivation

Motivation

- Human-AI teams in healthcare

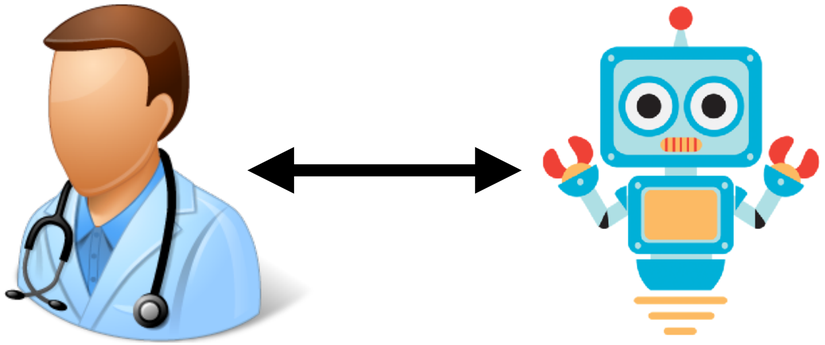
Motivation

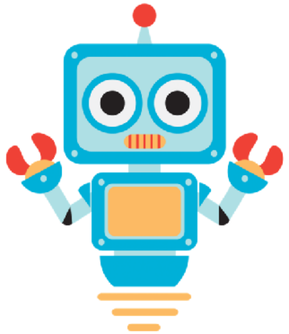
- Human-AI teams in healthcare



Motivation

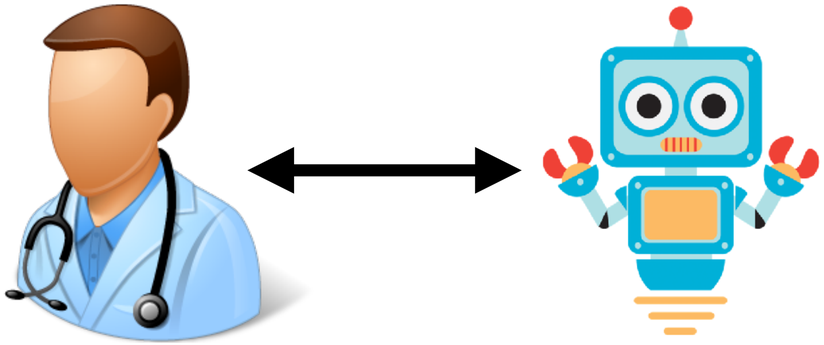
- Human-AI teams in healthcare

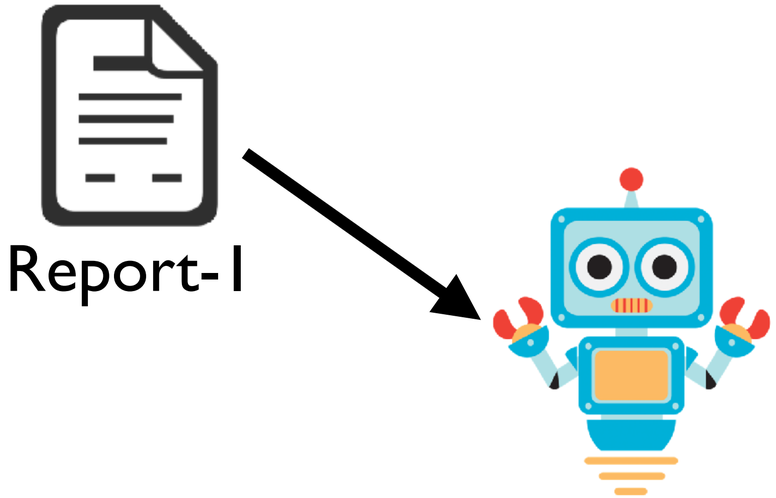




Motivation

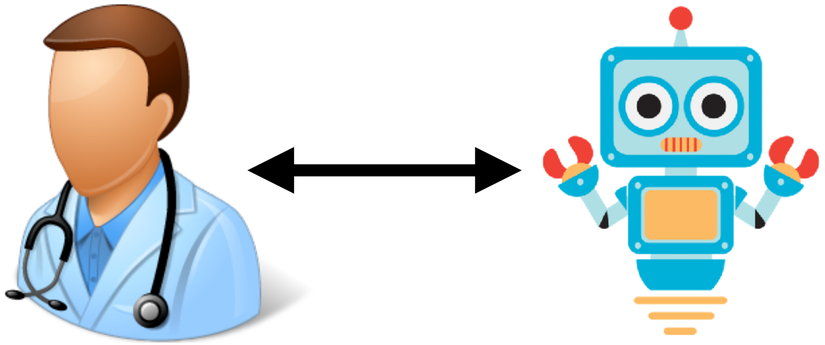
- Human-AI teams in healthcare

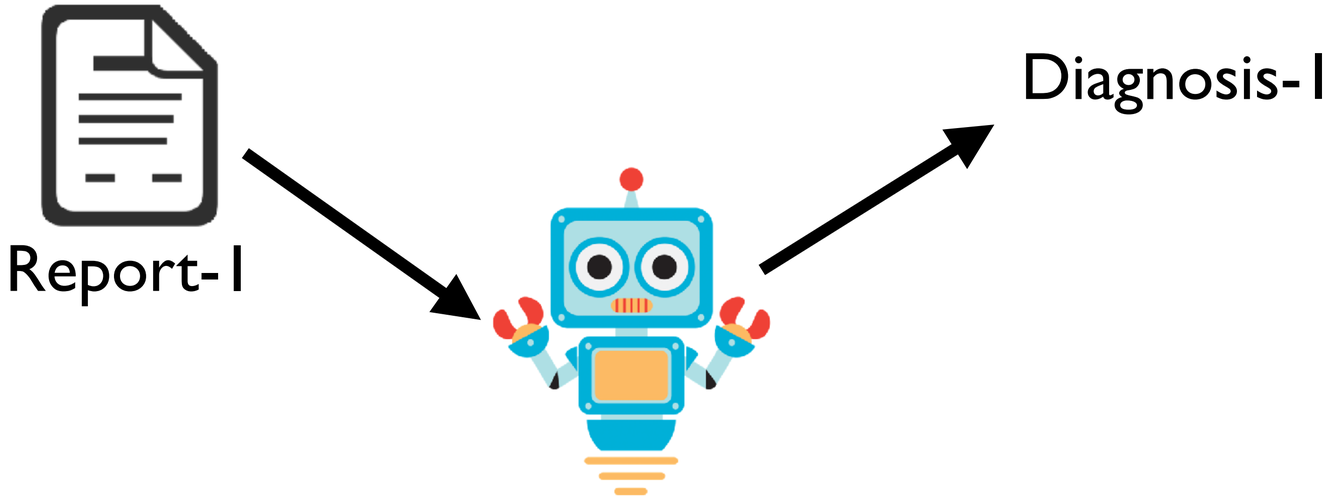




Motivation

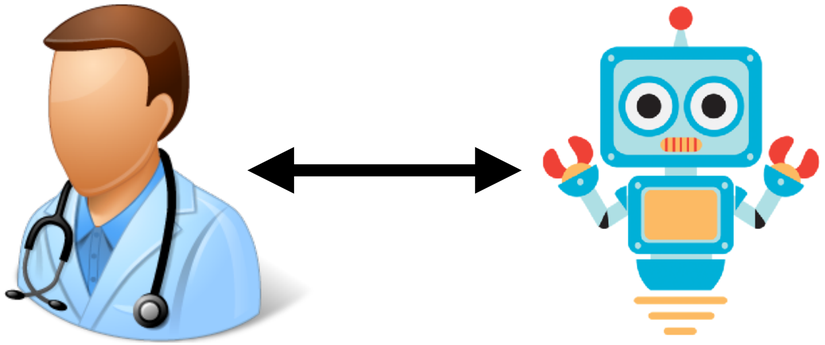
- Human-AI teams in healthcare

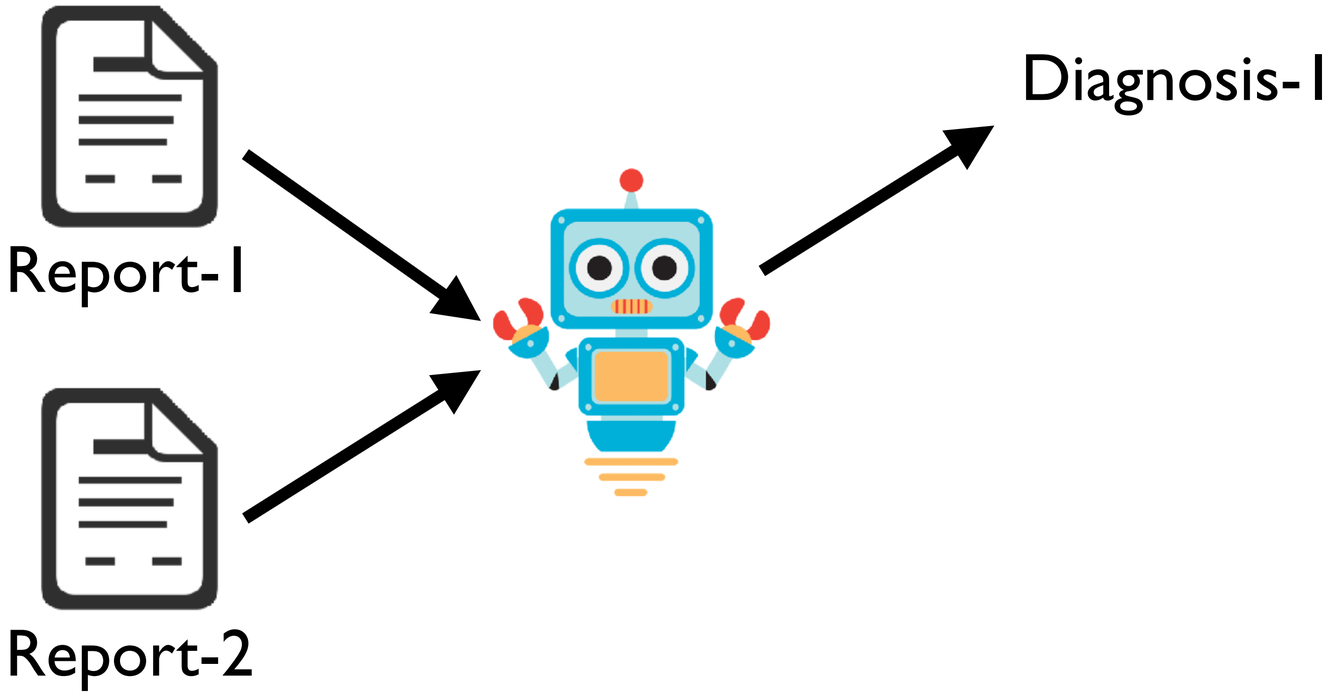




Motivation

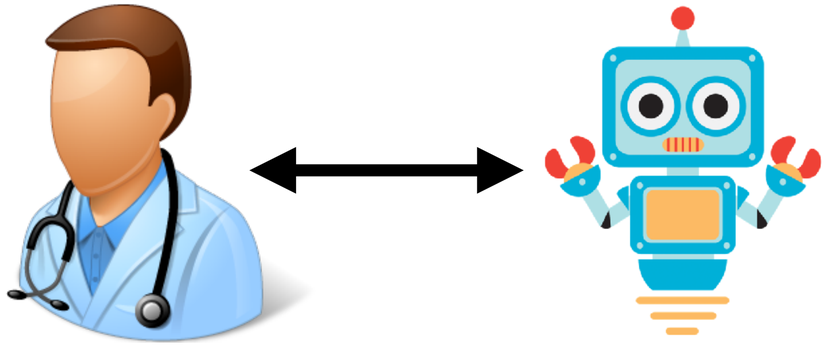
- Human-AI teams in healthcare

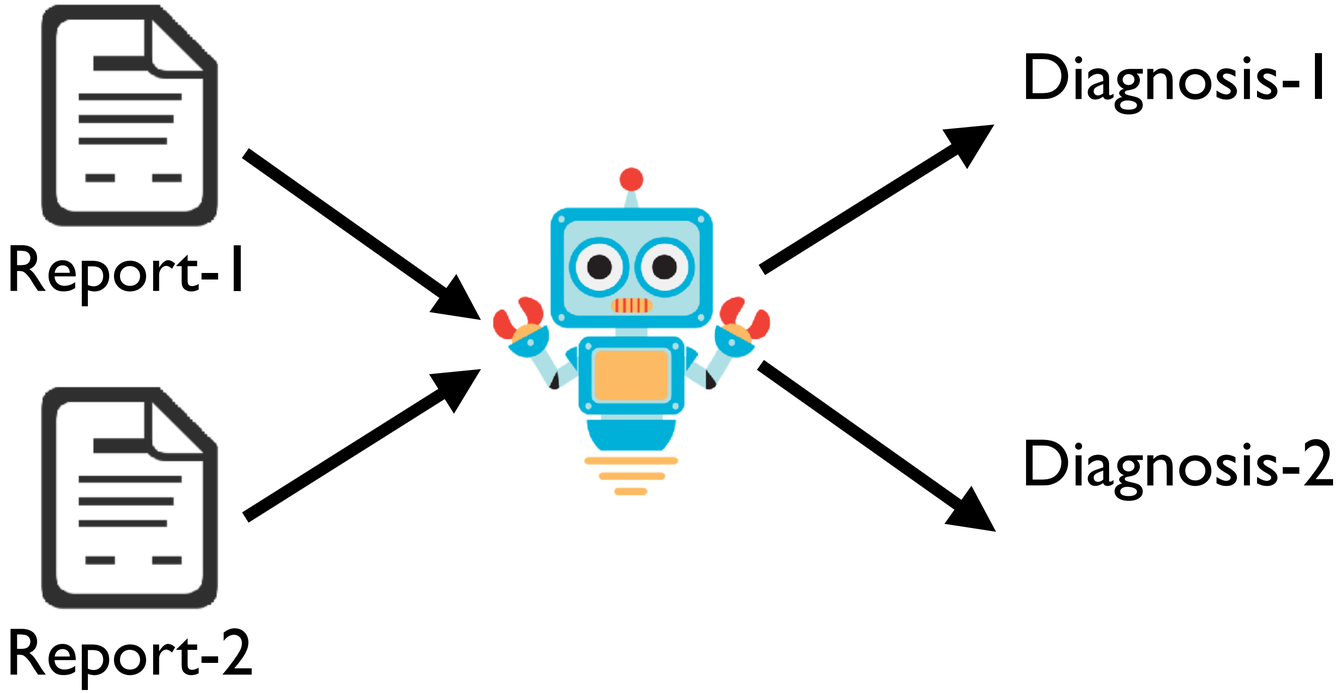




Motivation

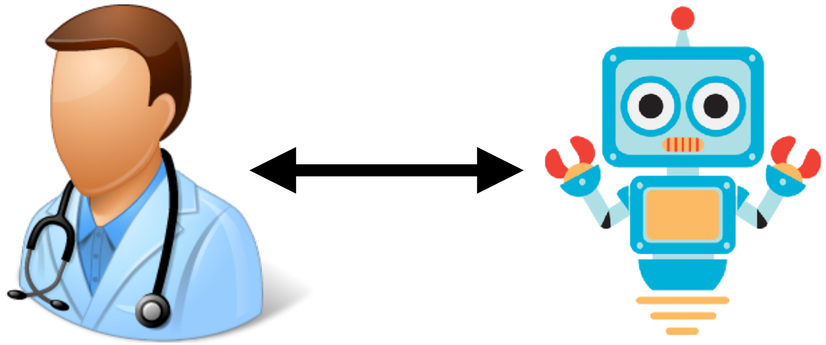
- Human-AI teams in healthcare

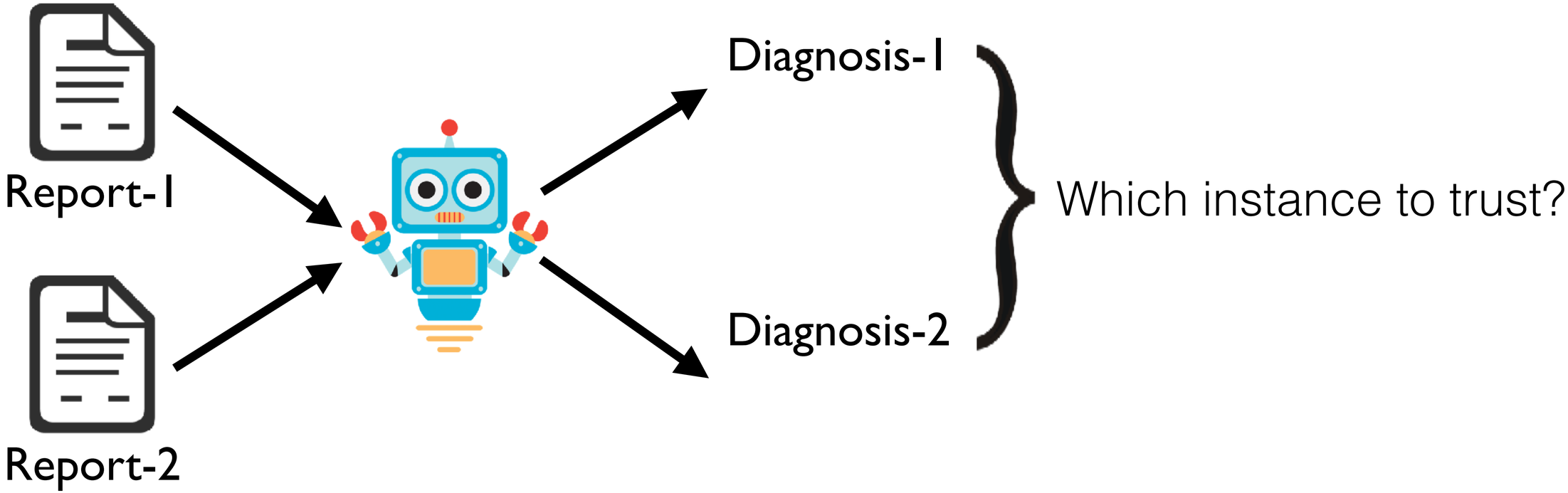




Motivation

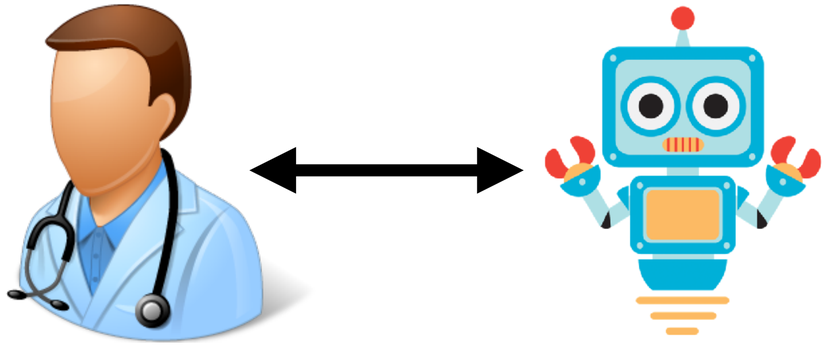
- Human-AI teams in healthcare

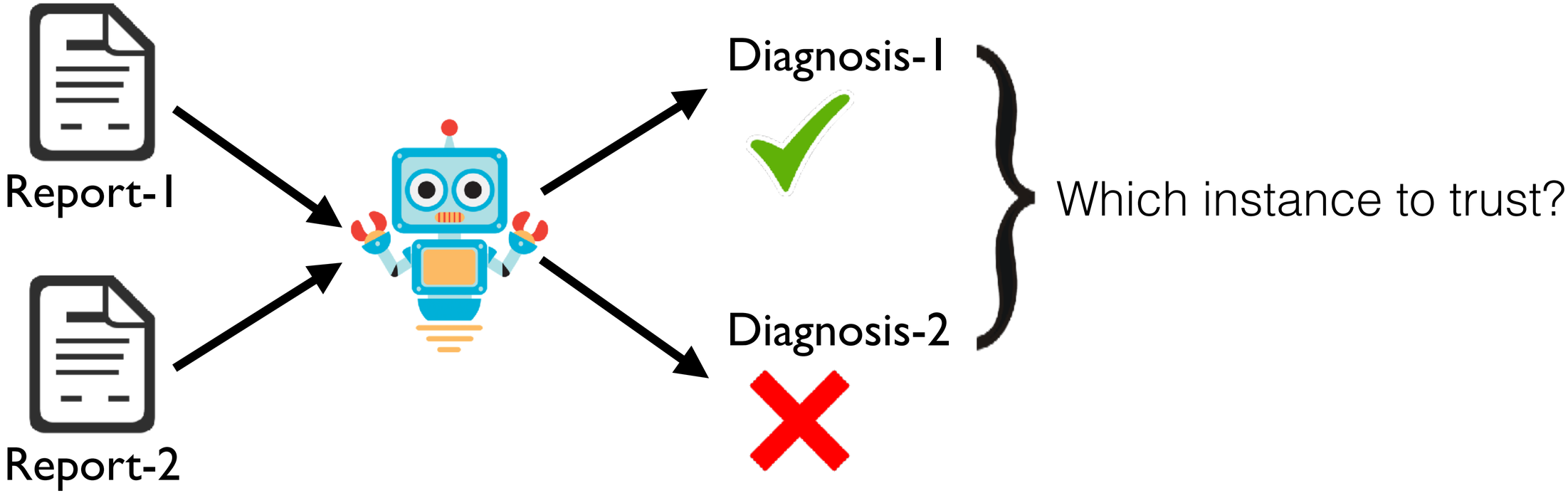




Motivation

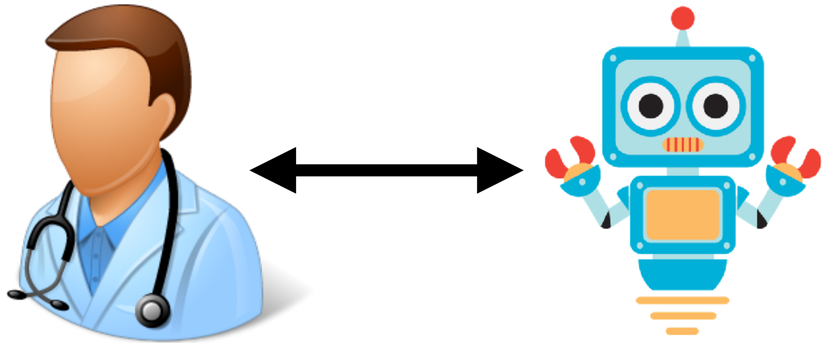
- Human-AI teams in healthcare

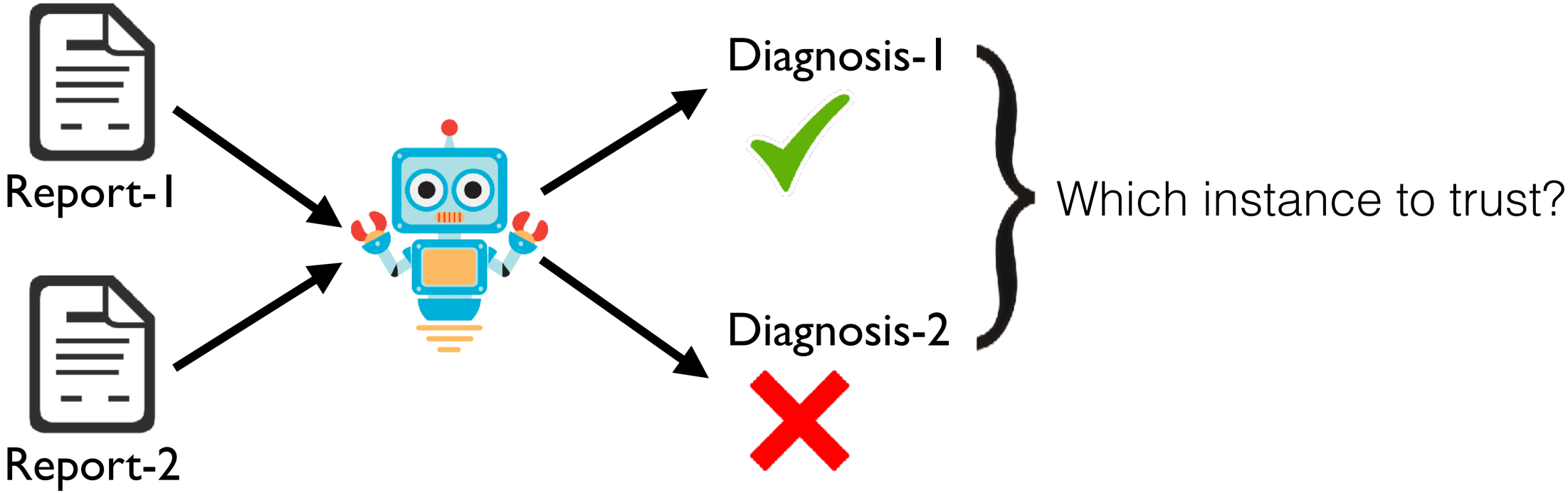




Motivation

- Human-AI teams in healthcare

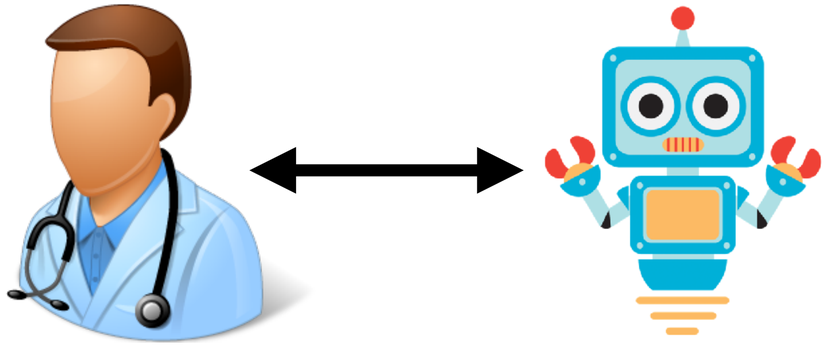


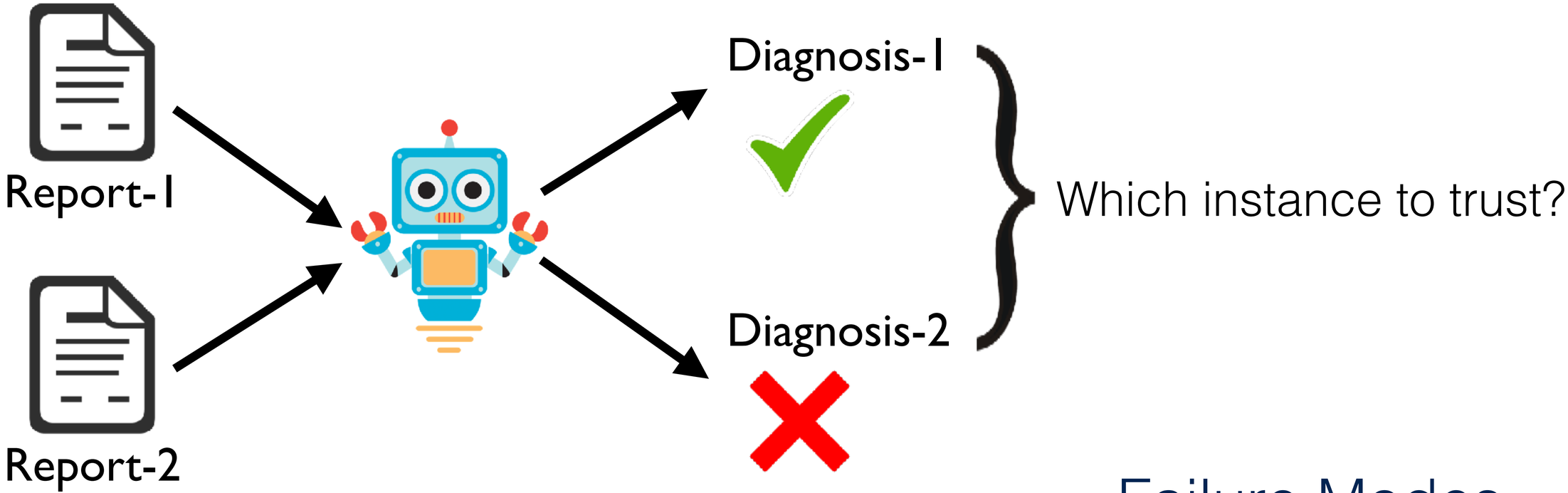


- Critical for human to have a sense of AI's

Motivation

- Human-AI teams in healthcare



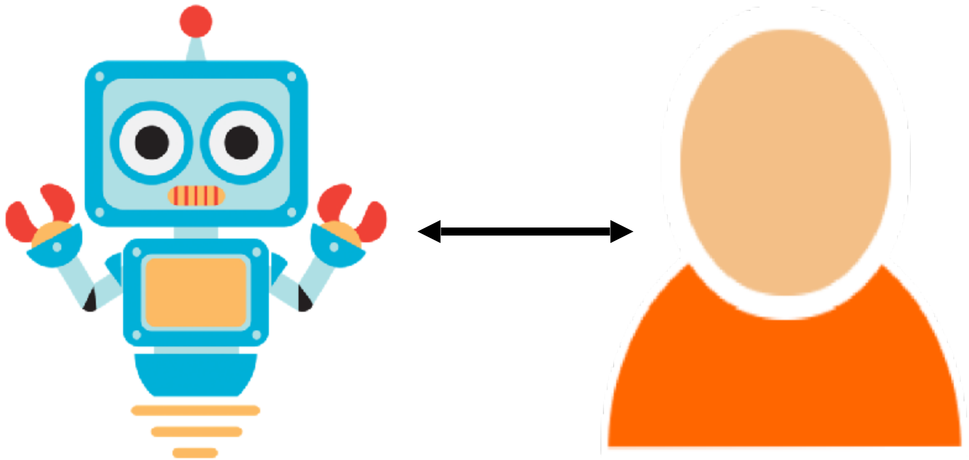


- Critical for human to have a sense of AI's
  - Failure Modes
  - Behavior

Theory of AI's Mind (ToAIM)

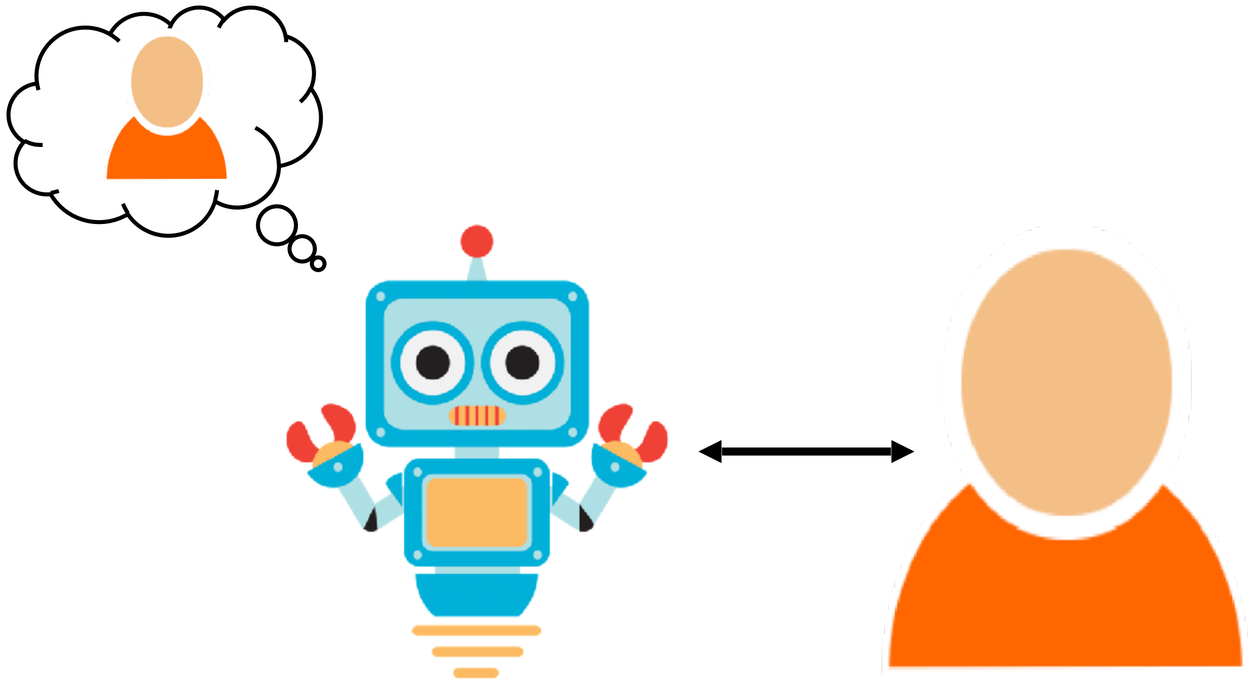
Theory of AI's Mind (ToAIM)

- For human-AI teams to be effective, humans must *also* develop theory of AI's mind



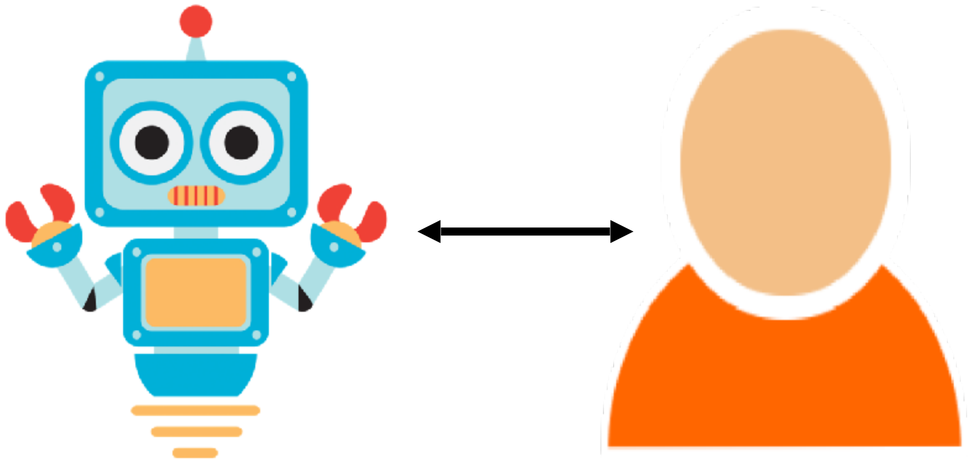
Theory of AI's Mind (ToAIM)

- For human-AI teams to be effective, humans must *also* develop theory of AI's mind



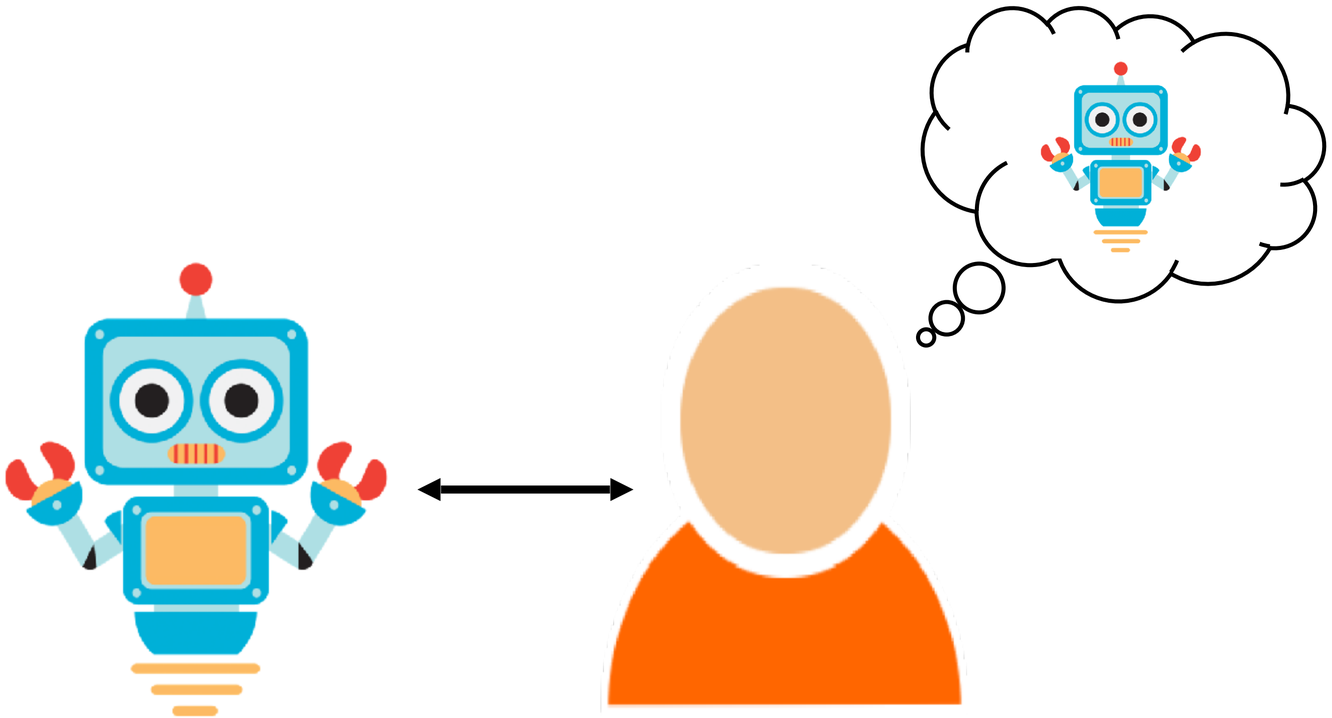
Theory of AI's Mind (ToAIM)

- For human-AI teams to be effective, humans must *also* develop theory of AI's mind



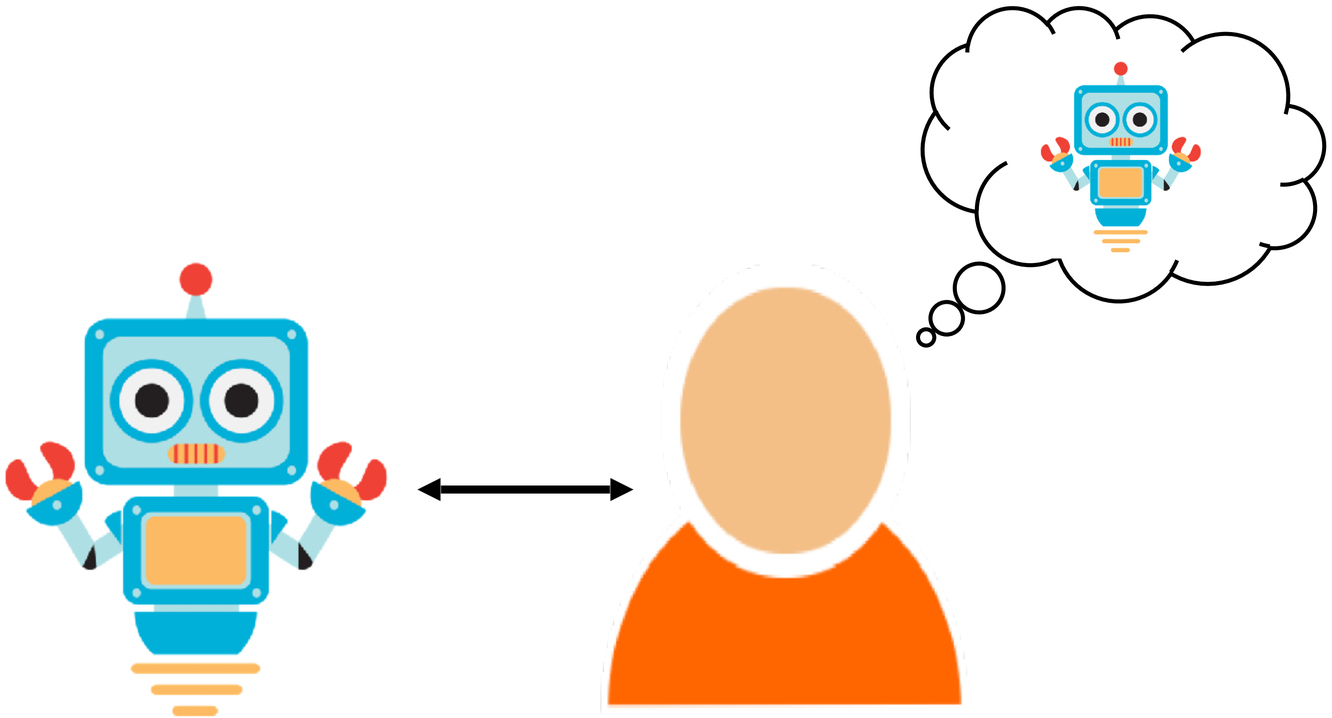
Theory of AI's Mind (ToAIM)

- For human-AI teams to be effective, humans must *also* develop theory of AI's mind



Theory of AI's Mind (ToAIM)

- For human-AI teams to be effective, humans must *also* develop theory of AI's mind
- Predict success, failure and responses
- Approximate a neural network!



Outline

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

AI Agent

AI Agent

- Vicki : A VQA Model

  Initial Scope: Visual Question Answering (VQA)

AI Agent

- Vicki : A VQA Model

  Initial Scope: Visual Question Answering (VQA)

AI Agent

- Vicki : A VQA Model

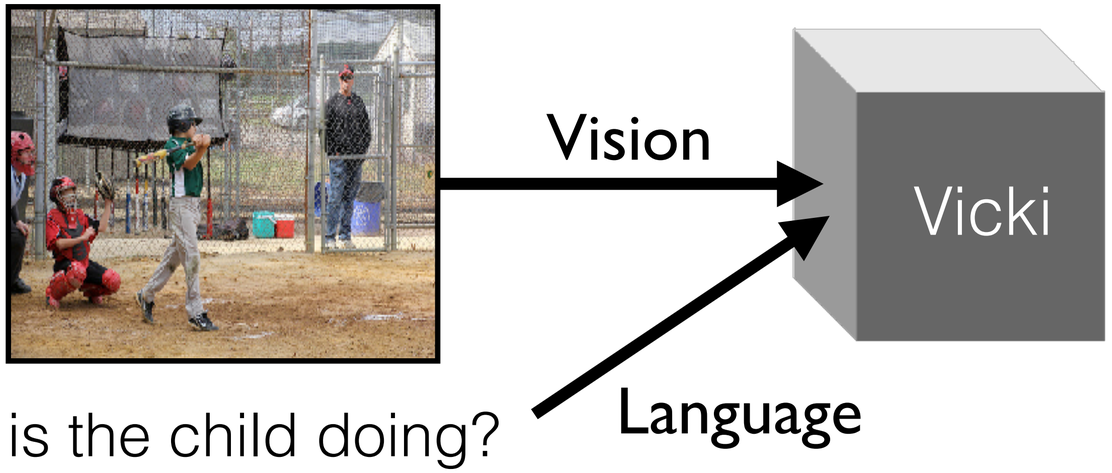
Initial Scope: Visual Question Answering (VQA)

What is the child doing?

AI Agent

- Vicki : A VQA Model

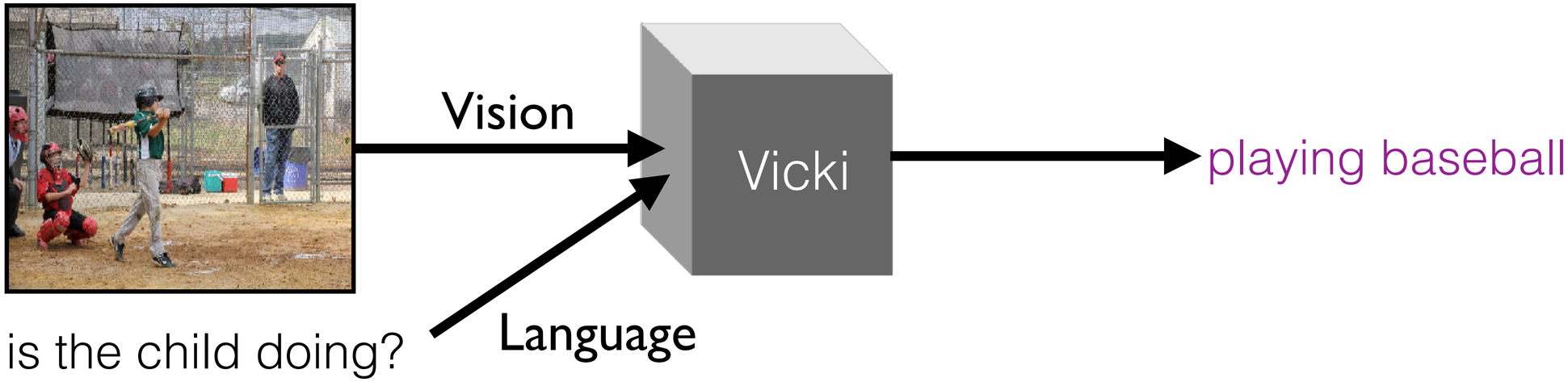
Initial Scope: Visual Question Answering (VQA)



AI Agent

- Vicki : A VQA Model

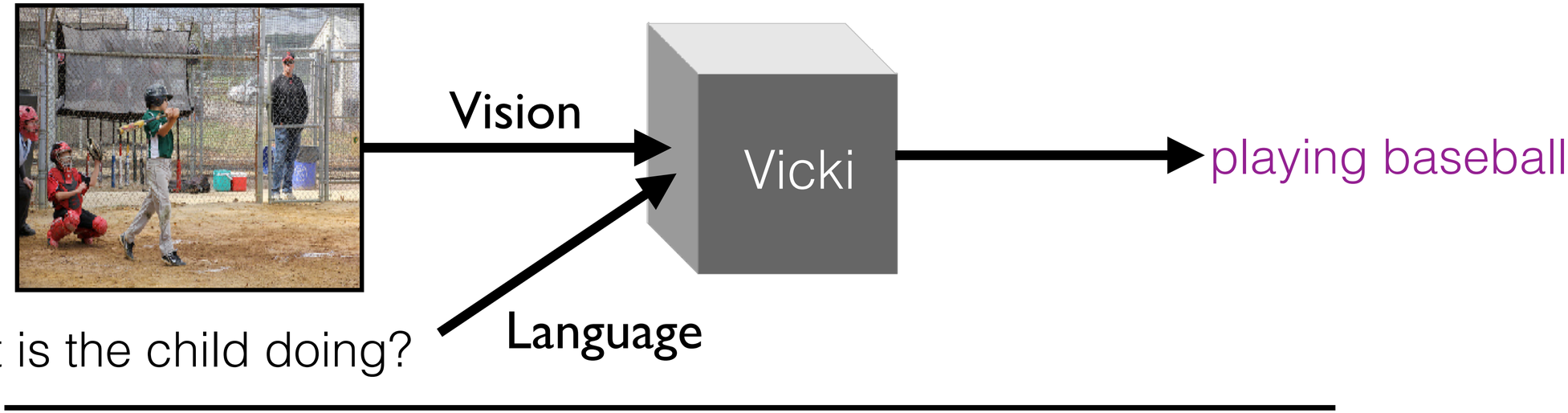
Initial Scope: Visual Question Answering (VQA)



AI Agent

- Vicki : A VQA Model

Initial Scope: Visual Question Answering (VQA)



VQA

Dataset

AI Agent

- Vicki : A VQA Model

Initial Scope: Visual Question Answering (VQA)

Vision

Vicki

playing baseball

What is the child doing?

Language

VQA

Dataset

248349 QI-pairs
in train-split

AI Agent

- Vicki : A VQA Model

Initial Scope: Visual Question Answering (VQA)

Vision

Vicki

playing baseball

What is the child doing?

Language

Vicki can answer only from the top-1k answers

VQA

Dataset

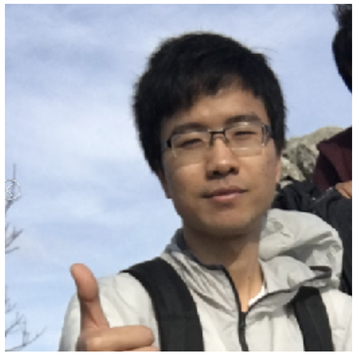
248349 QI-pairs in train-split

AI Agent

AI Agent

- VQA model by Lu, Yang et al. NIPS 2016

Jiasen Lu

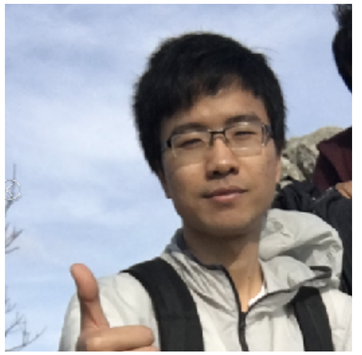


Jianwei Yang

AI Agent

- VQA model by Lu, Yang et al. NIPS 2016
- Hierarchical co-Attention model

Jiasen Lu

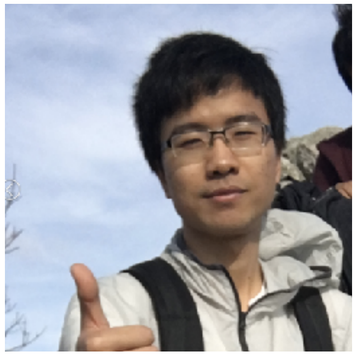


Jianwei Yang

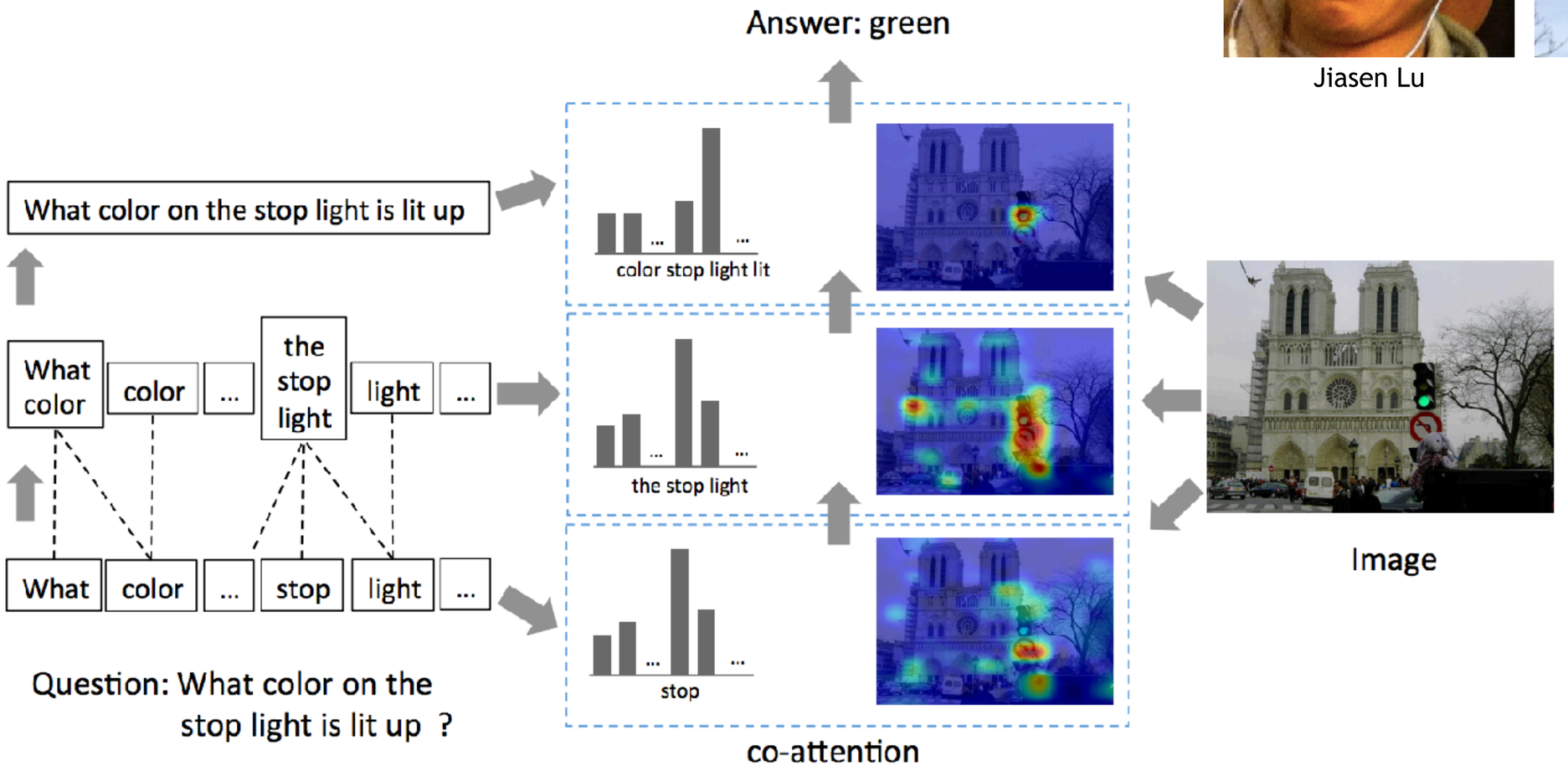
AI Agent

- VQA model by Lu, Yang et al. NIPS 2016
- Hierarchical co-Attention model

Jiasen Lu



Jianwei Yang



Vicki’s Quirks

Vicki’s Quirks

- Imperfect vision

Vicki’s Quirks

- Imperfect vision
- Limited capability to understand language

Vicki’s Quirks

- Imperfect vision
- Limited capability to understand language
- Can’t reason about common-sense

Vicki’s Quirks

- Imperfect vision
- Limited capability to understand language
- Can’t reason about common-sense
- Limited vocabulary

Vicki’s Quirks

- Imperfect vision
- Limited capability to understand language
- Can’t reason about common-sense
- Limited vocabulary
- Doesn’t understand question-image relevance

Vicki's Quirks

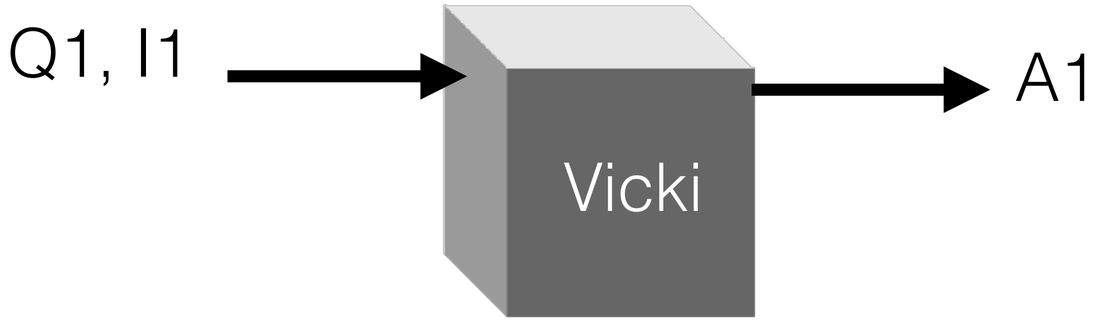
- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases

Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases

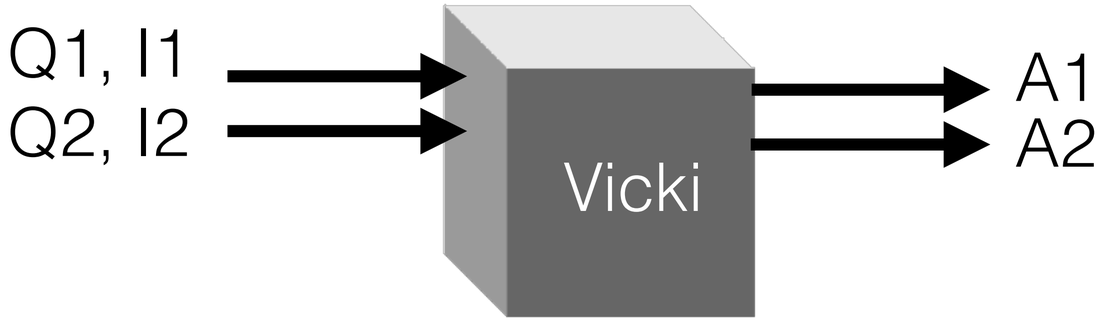
Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases



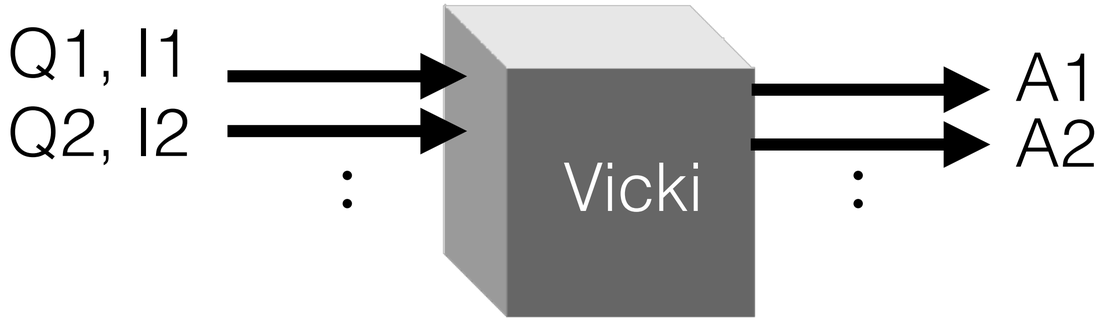
Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases



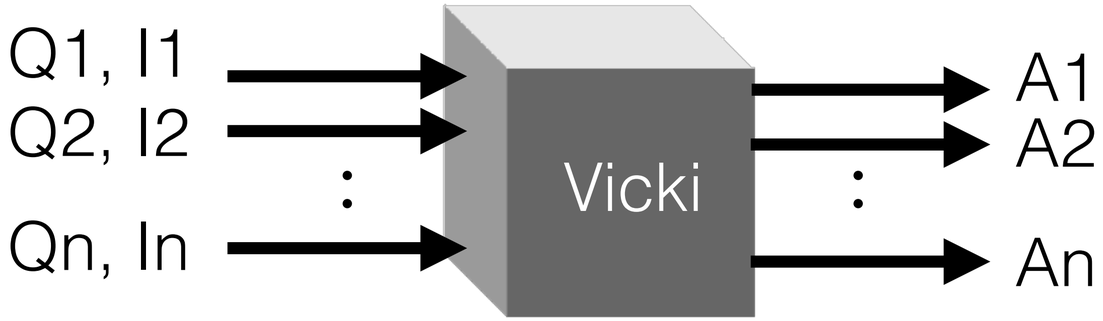
Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases



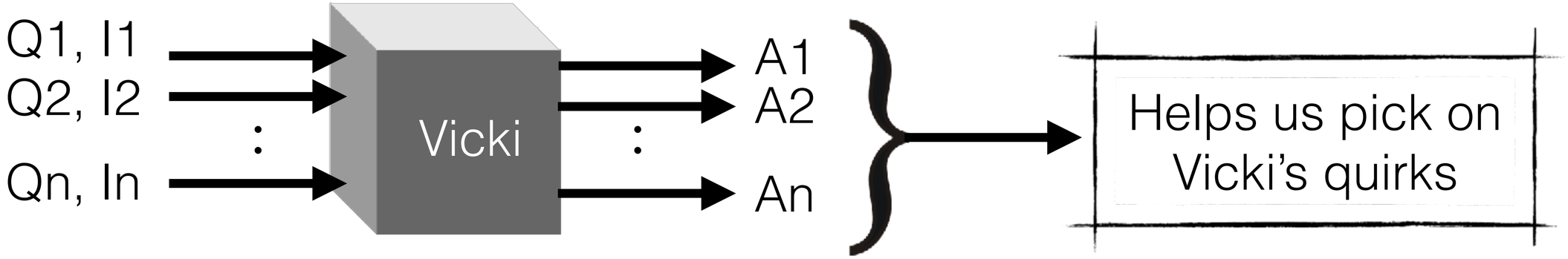
Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases



Vicki's Quirks

- Imperfect vision
- Limited capability to understand language
- Can't reason about common-sense
- Limited vocabulary
- Doesn't understand question-image relevance
- Heavily influenced by dataset biases



Vicki’s Quirks

Vicki’s Quirks

What color is the grass? Blue

Vicki's Quirks

What color is the grass? Blue

What are the people doing? Eating

Vicki's Quirks

What color is the grass? Blue

What are the people doing? Eating

How many people are there? 4

What is the man holding? Fire Hydrant

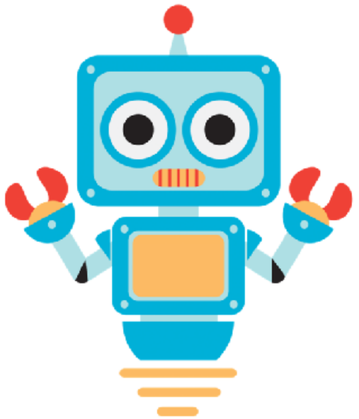
ToAIM

ToAIM

- To study/evaluate ToAIM → Large-scale experiments on MTurk

ToAIM

- To study/evaluate ToAIM → Large-scale experiments on MTurk

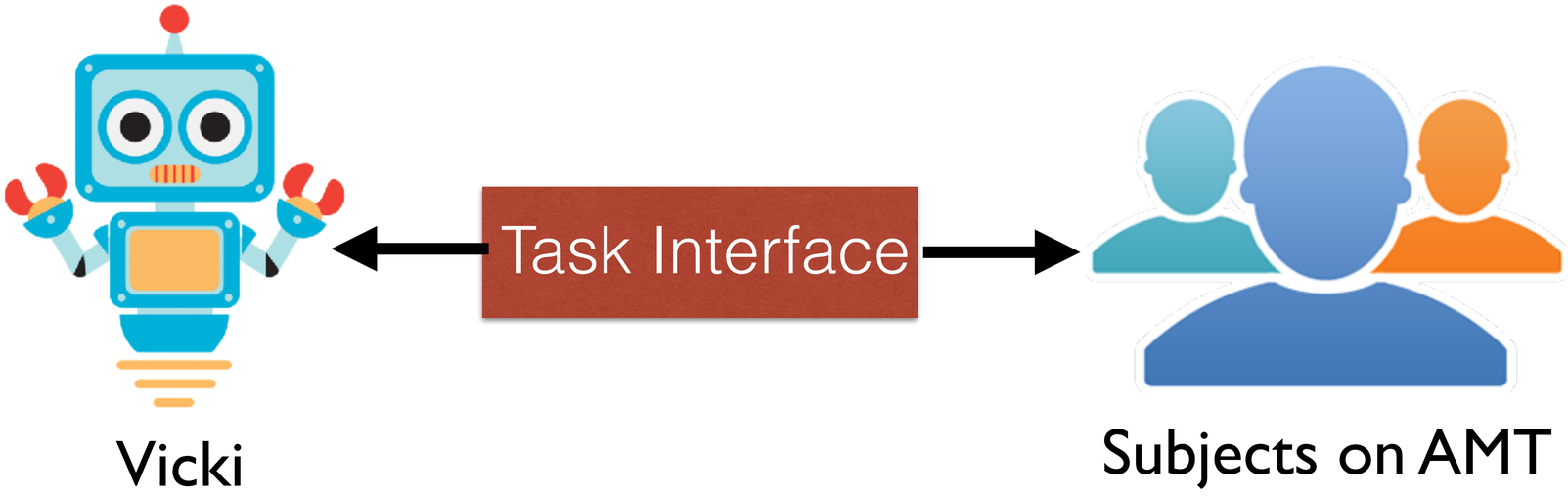


Vicki

Subjects on AMT

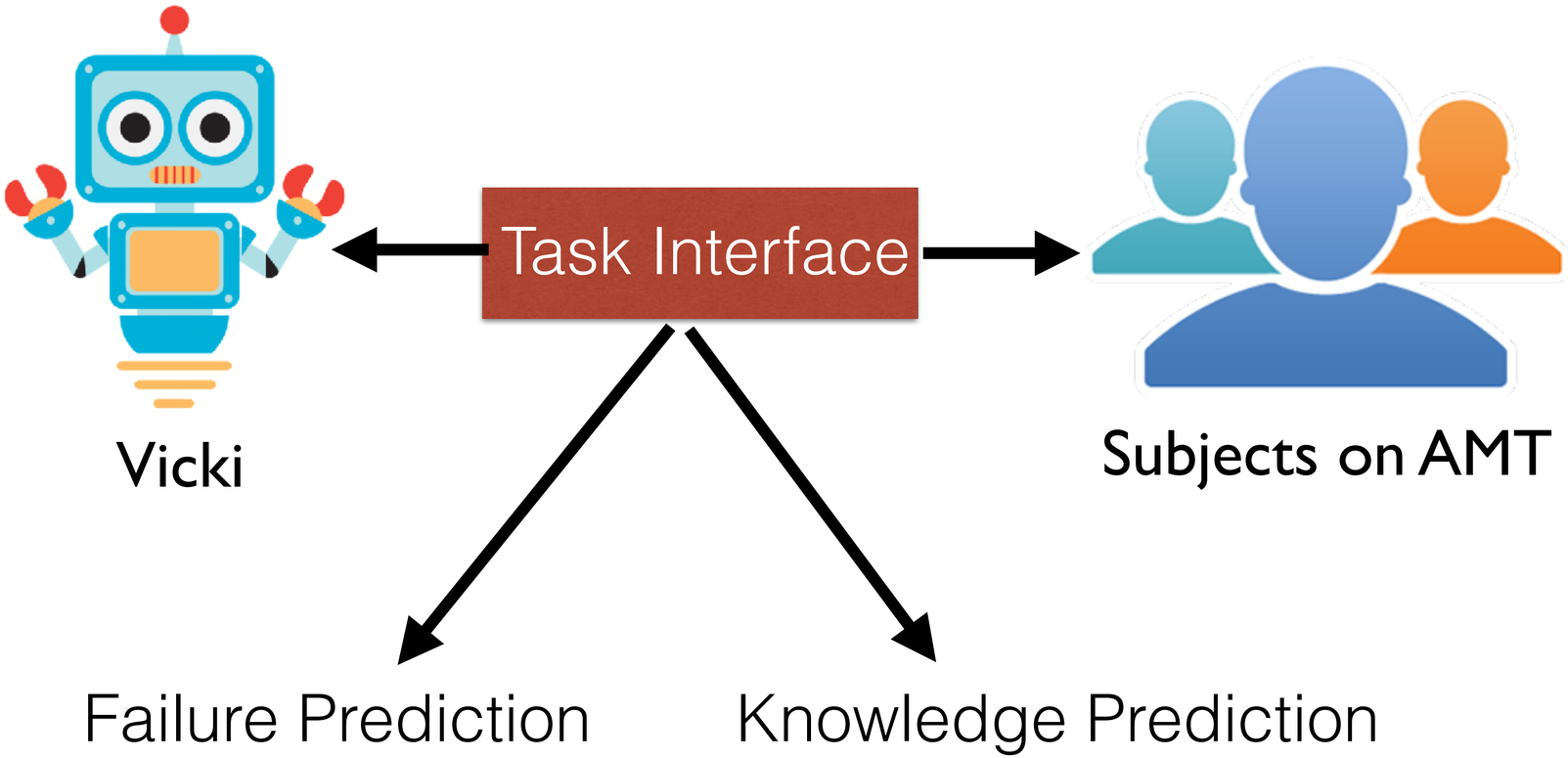
ToAIM

- To study/evaluate ToAIM → Large-scale experiments on MTurk



ToAIM

- To study/evaluate ToAIM → Large-scale experiments on MTurk



ToAIM

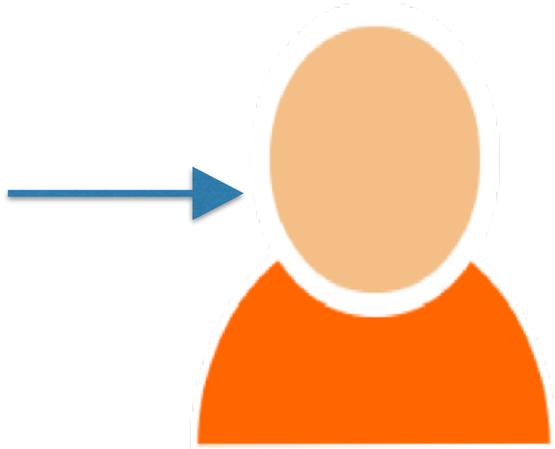
ToAIM

- Failure Prediction

ToAIM

- Failure Prediction

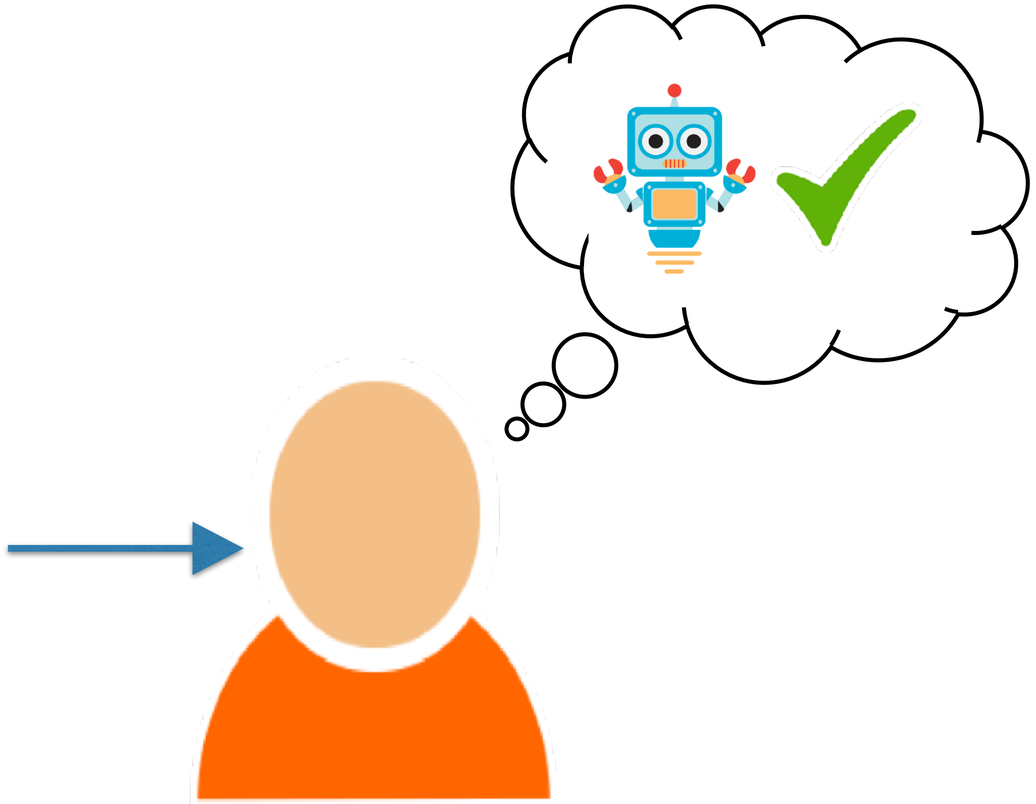
How many people are there?



ToAIM

- Failure Prediction

How many people are there?

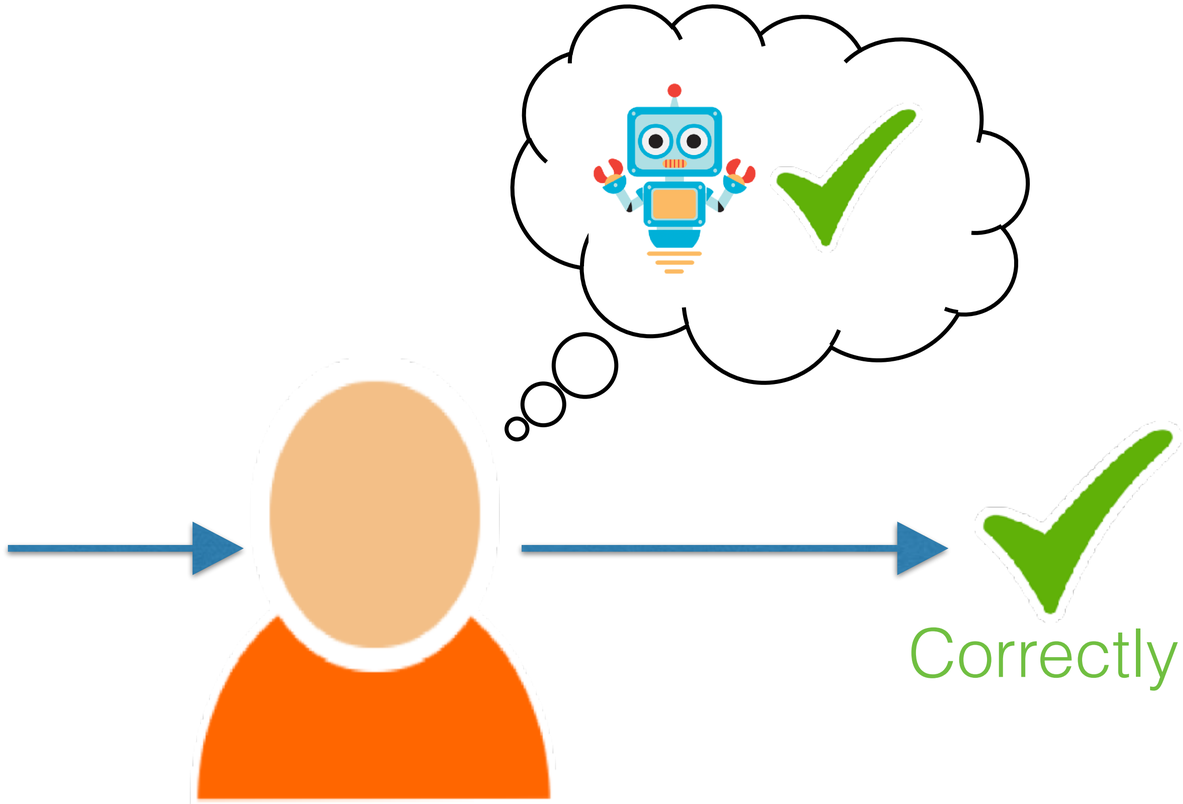


Subject thinks Vicki will answer correctly

ToAIM

- Failure Prediction

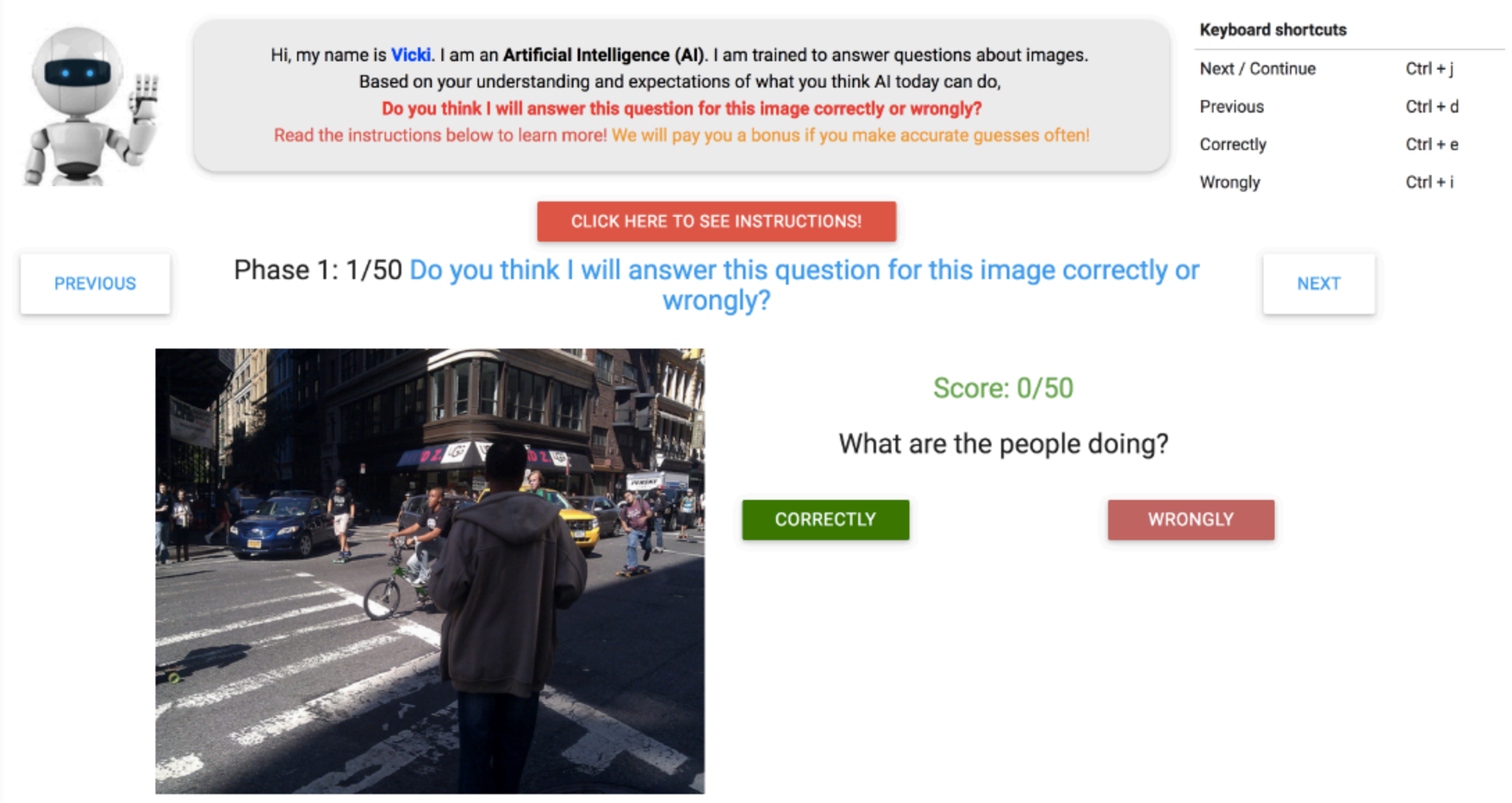
How many people are there?



Subject thinks Vicki will answer correctly

ToAIM

- Failure Prediction



ToAIM

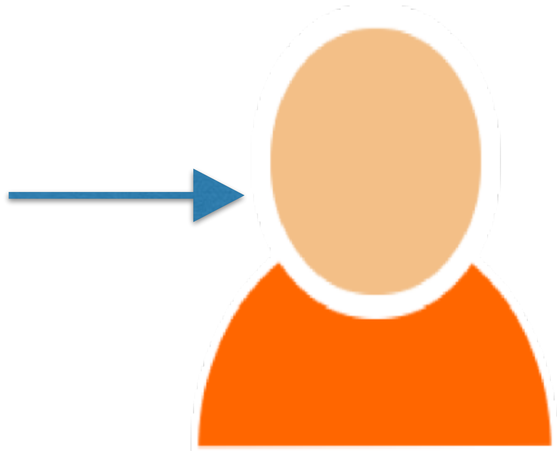
ToAIM

- Knowledge Prediction

ToAIM

- Knowledge Prediction

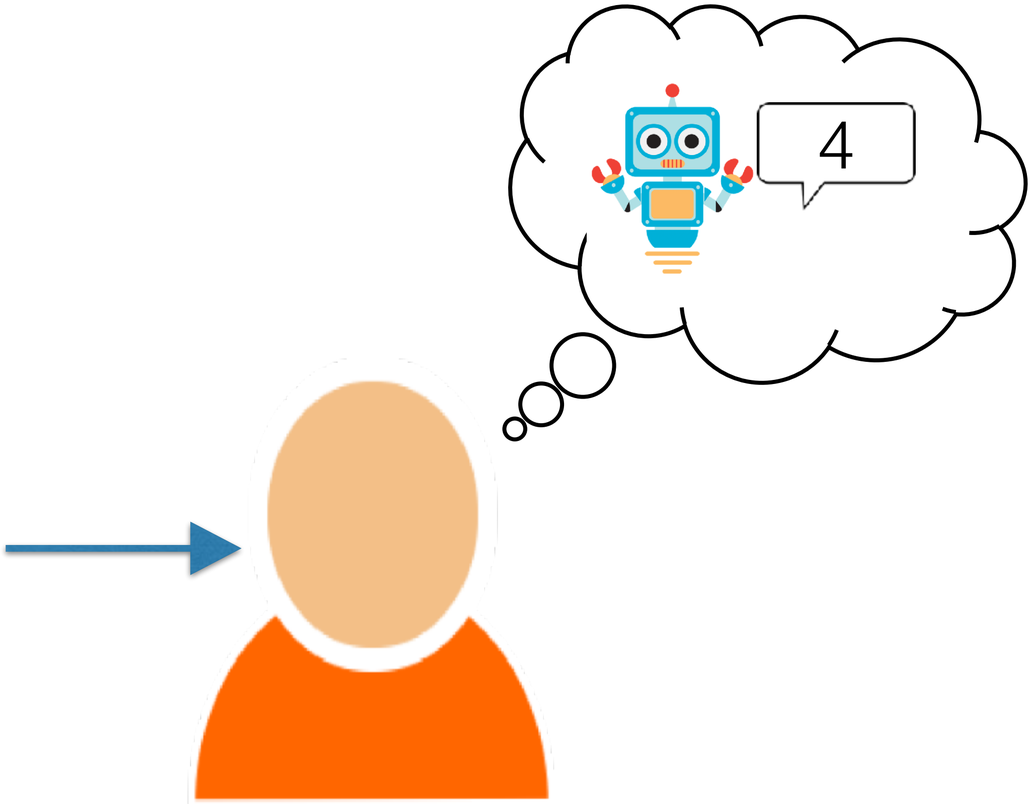
How many people are there?



ToAIM

- Knowledge Prediction

How many people are there?

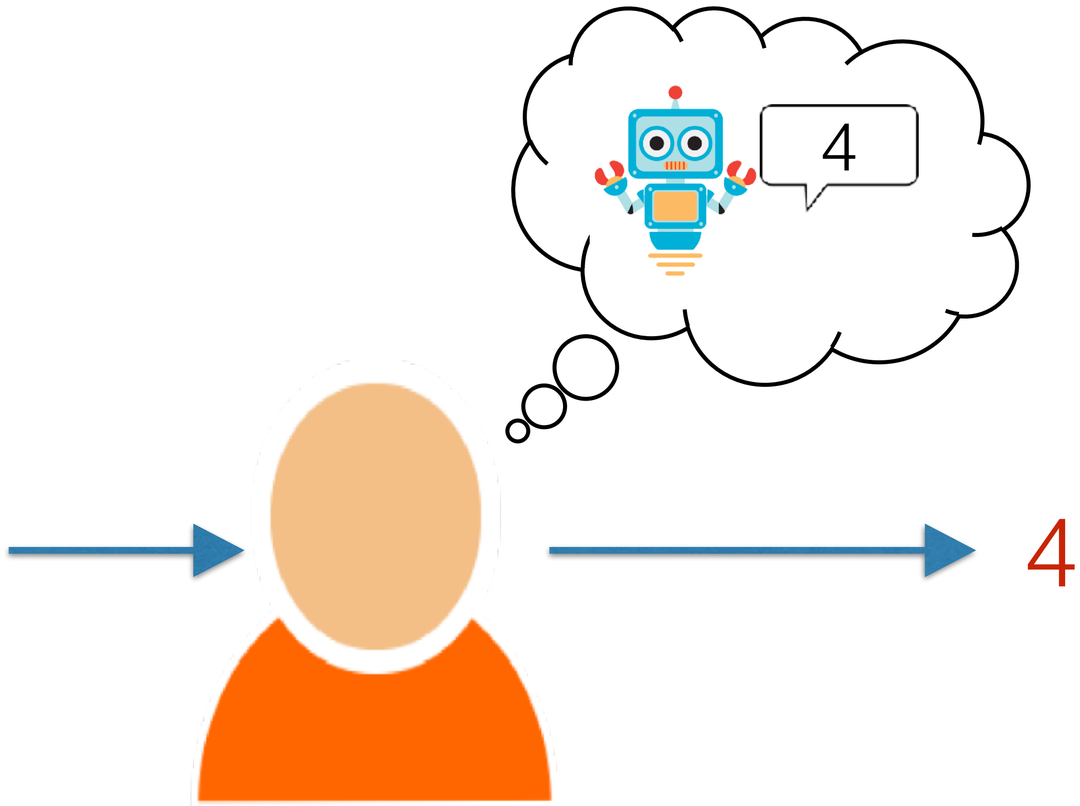


Subject thinks Vicki will answer 4

ToAIM

- Knowledge Prediction

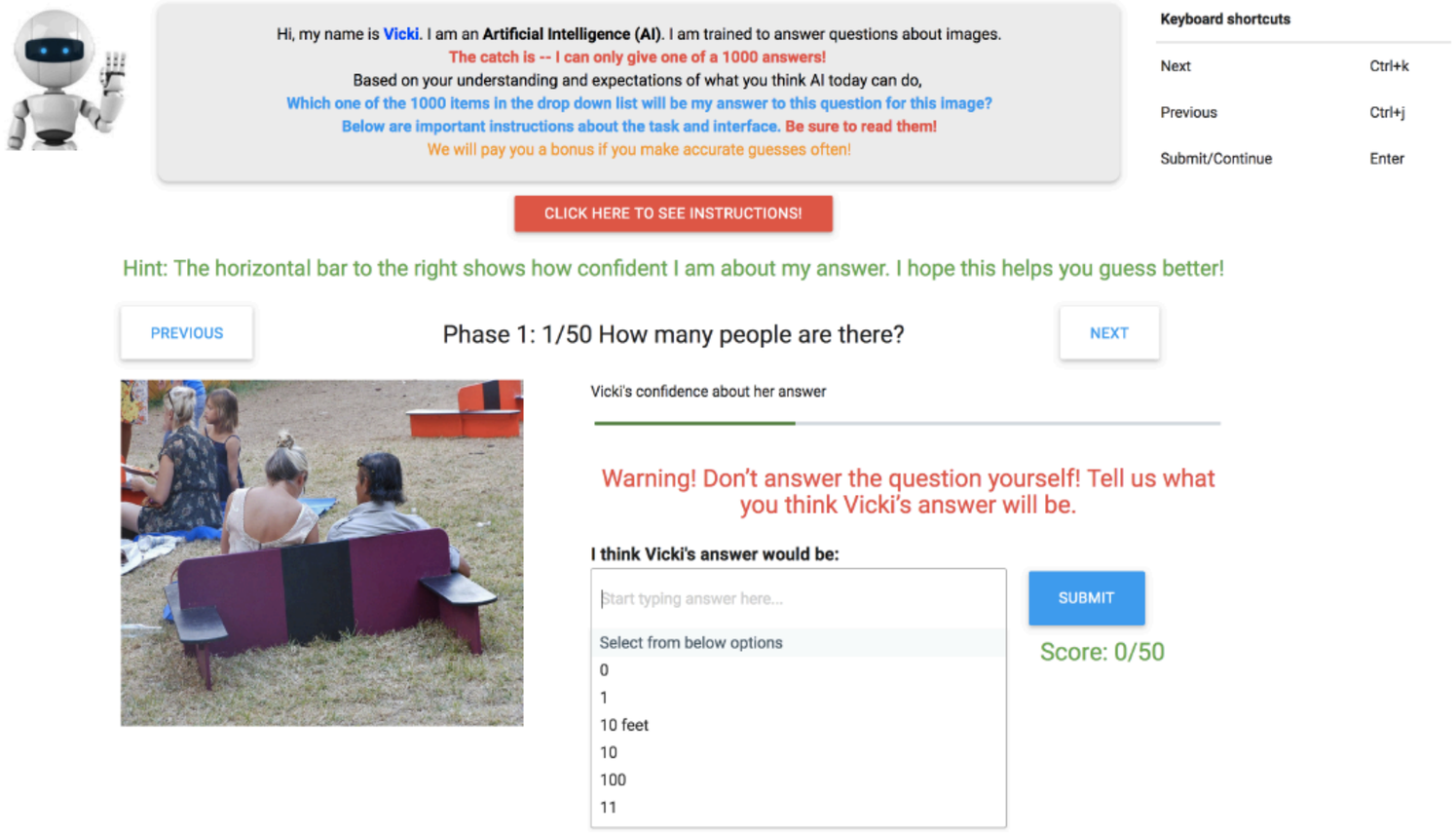
How many people are there?



Subject thinks Vicki will answer 4

ToAIM

- Knowledge Prediction



ToAIM

ToAIM

- We evaluate the role of

ToAIM

- We evaluate the role of
  - Training
  - Explanation Modalities

ToAIM

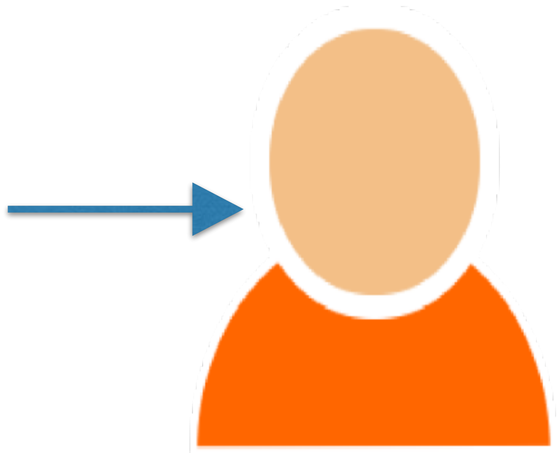
ToAIM

- Training via Instant Feedback

ToAIM

- Training via Instant Feedback

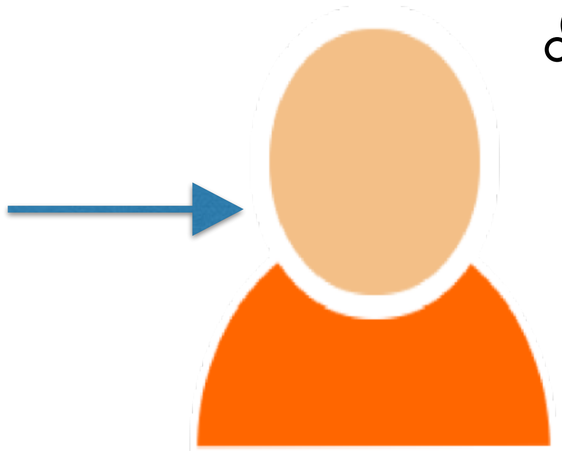
How many people are there?

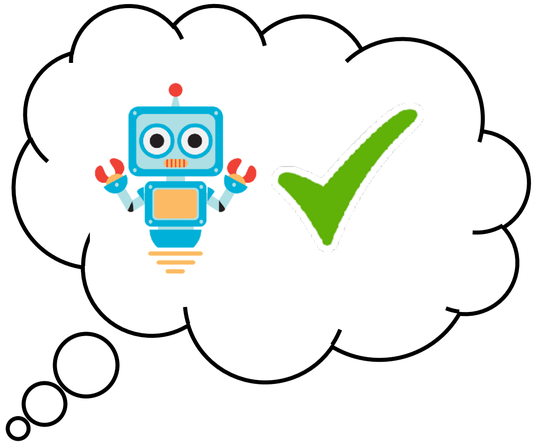


ToAIM

- Training via Instant Feedback

How many people are there?



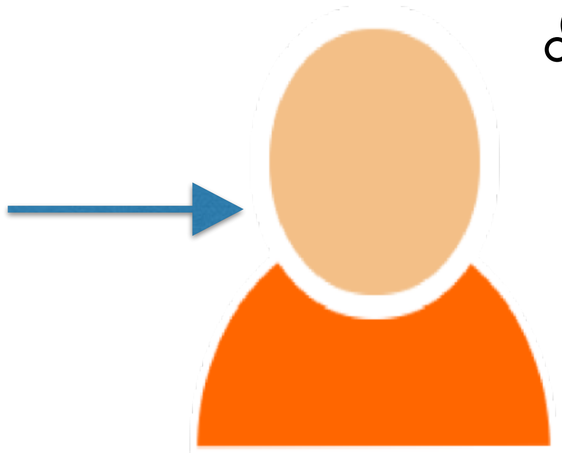


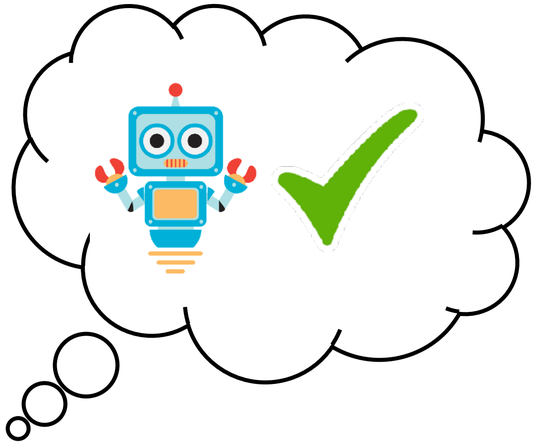
FP

ToAIM

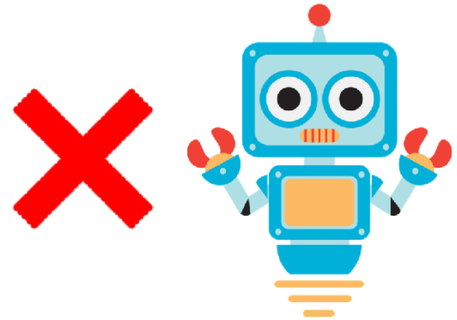
- Training via Instant Feedback

How many people are there?





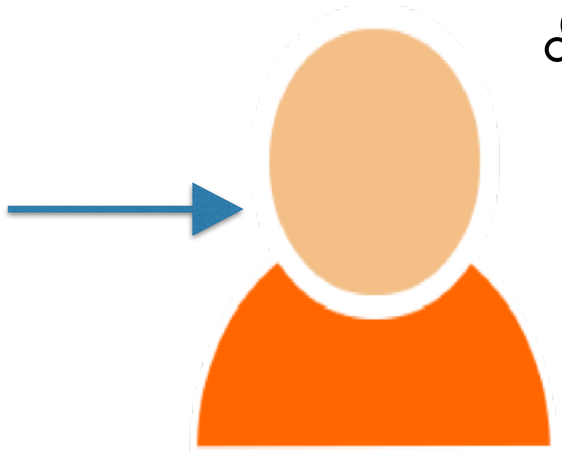
FP

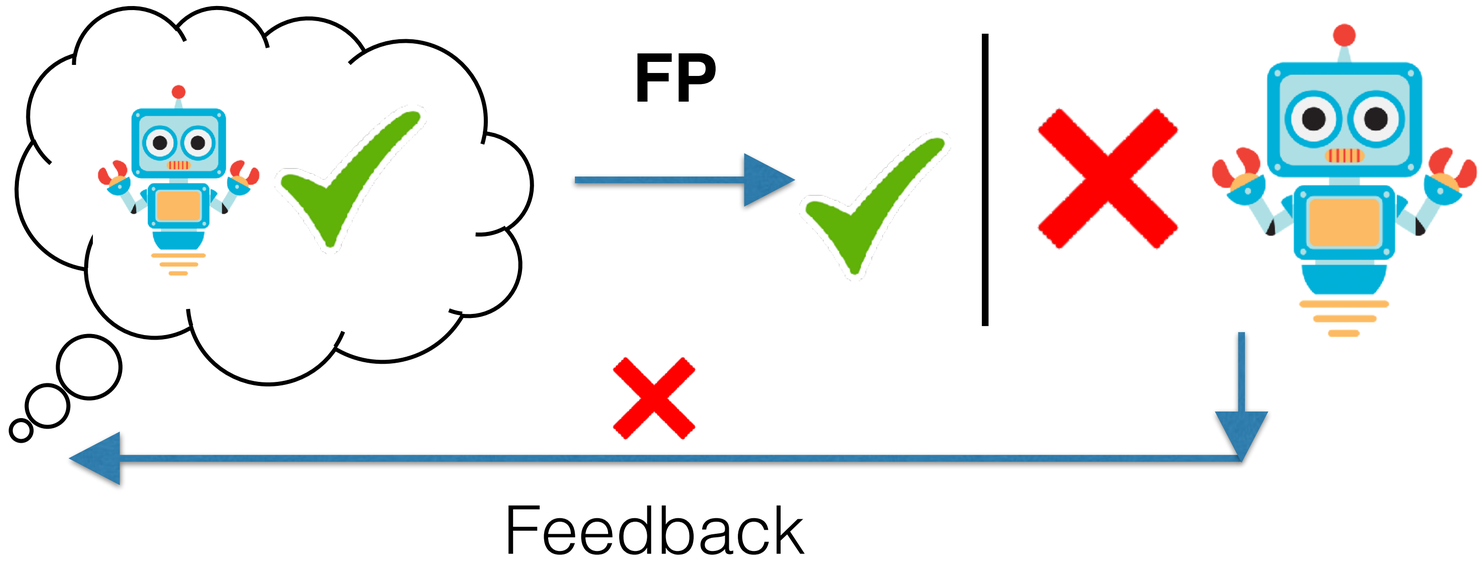


ToAIM

- Training via Instant Feedback

How many people are there?

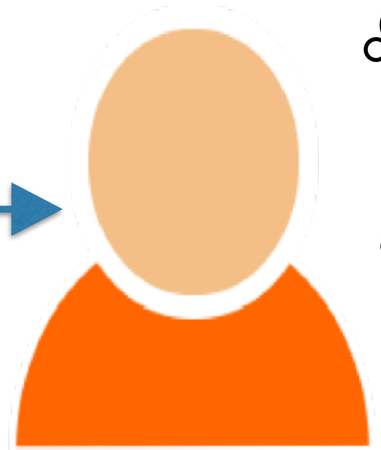


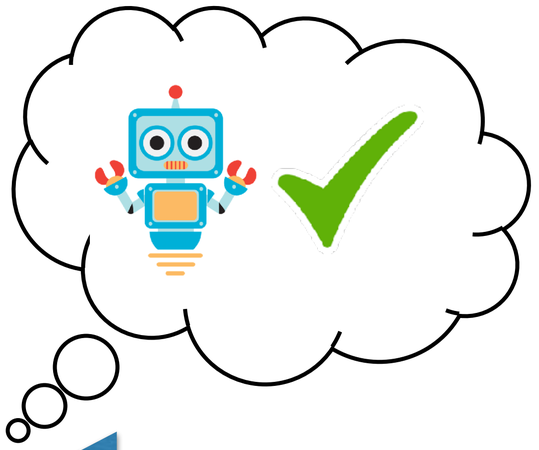


ToAIM

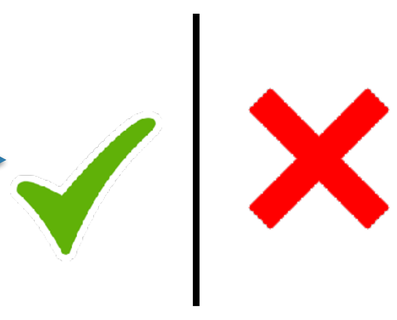
- Training via Instant Feedback

How many people are there?

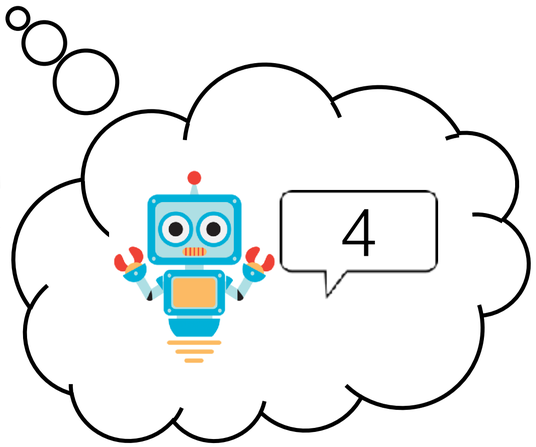




FP



Feedback



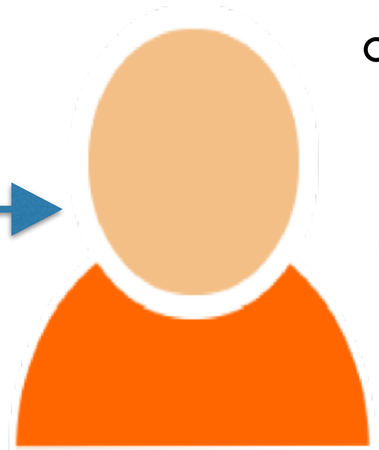
4

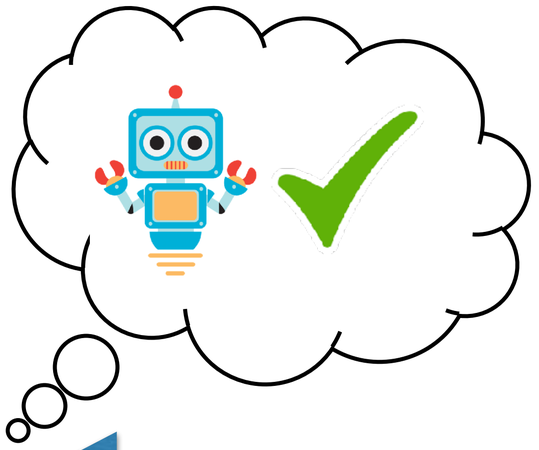
KP

ToAIM

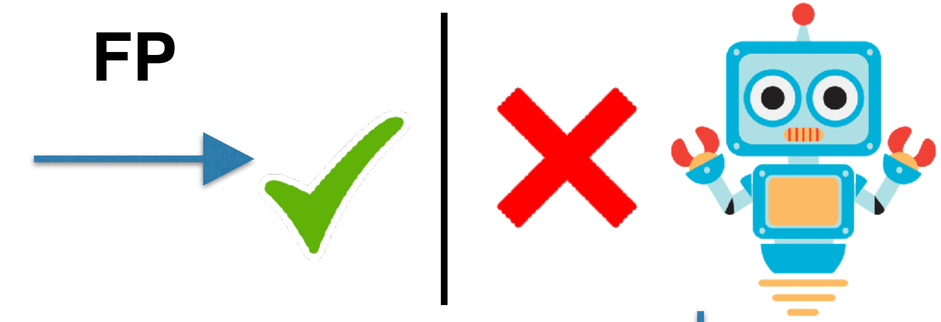
- Training via Instant Feedback

How many people are there?

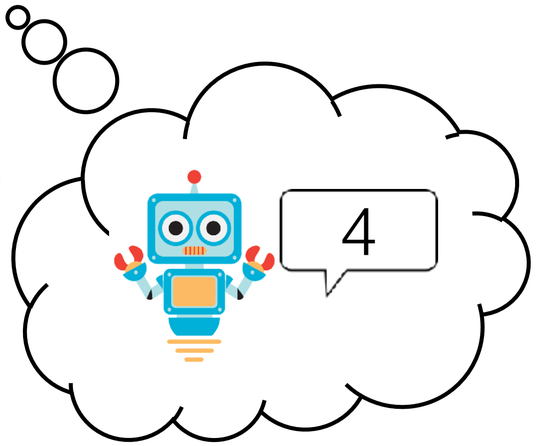




FP



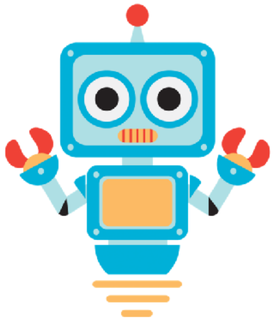
Feedback



KP

4

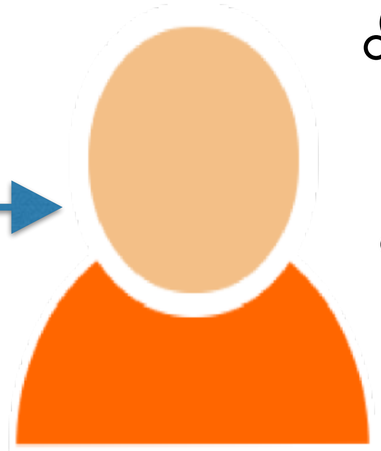
3

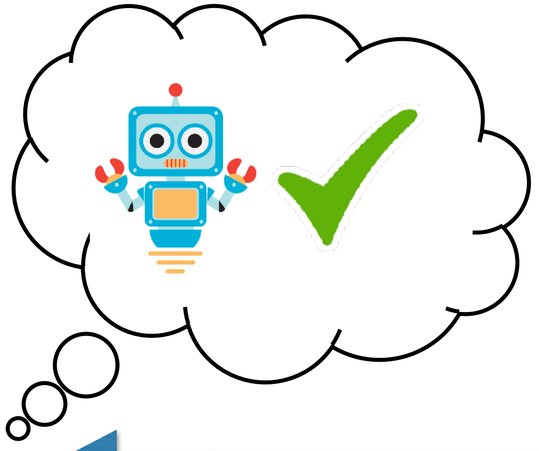


ToAIM

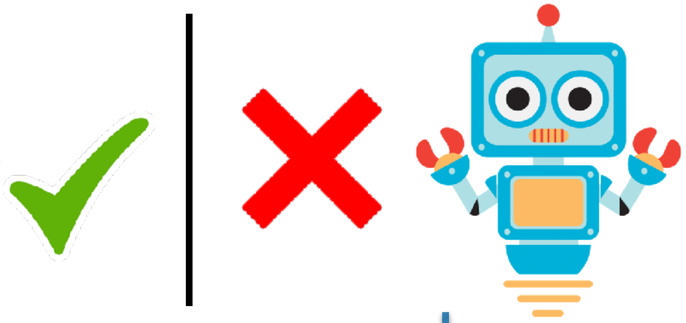
- Training via Instant Feedback

How many people are there?



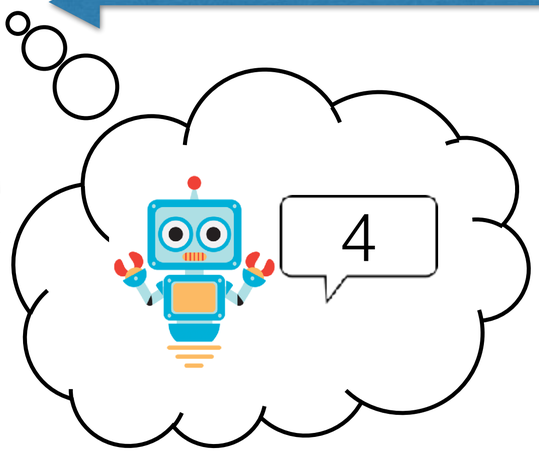


FP



Feedback

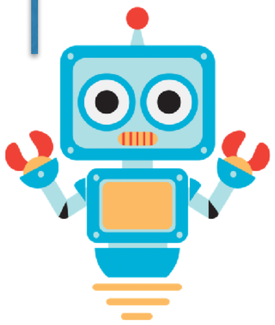
3



4

KP

3



ToAIM

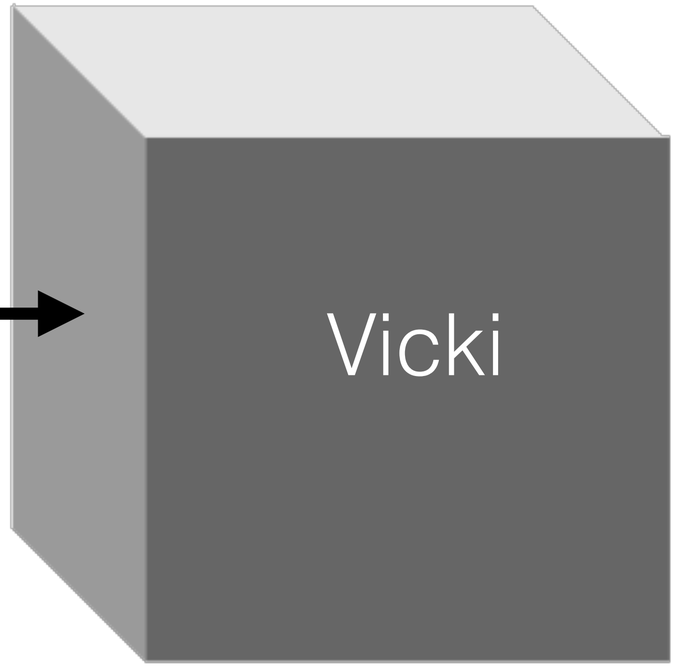
ToAIM

- Explanation Modalities

ToAIM

- Explanation Modalities

What is the child doing?



ToAIM

- Explanation Modalities
- Access to internal states of a model for a prediction

What is the child doing?

Vicki

playing baseball

Access to internal states of a model

ToAIM

- Explanation Modalities

ToAIM

- Explanation Modalities
- Explicit question-image attention maps from HieCoAtt

How many people are there ?

QI-Attention

ToAIM

- Explanation Modalities
- Explicit question-image attention maps from HieCoAtt

How many people are there ?

QI-Attention

Which words in the question to listen to?

ToAIM

- Explanation Modalities
- Explicit question-image attention maps from HieCoAtt

How many people are there ?

QI-Attention

Which words in the question to listen to?

Which regions in the image are important?

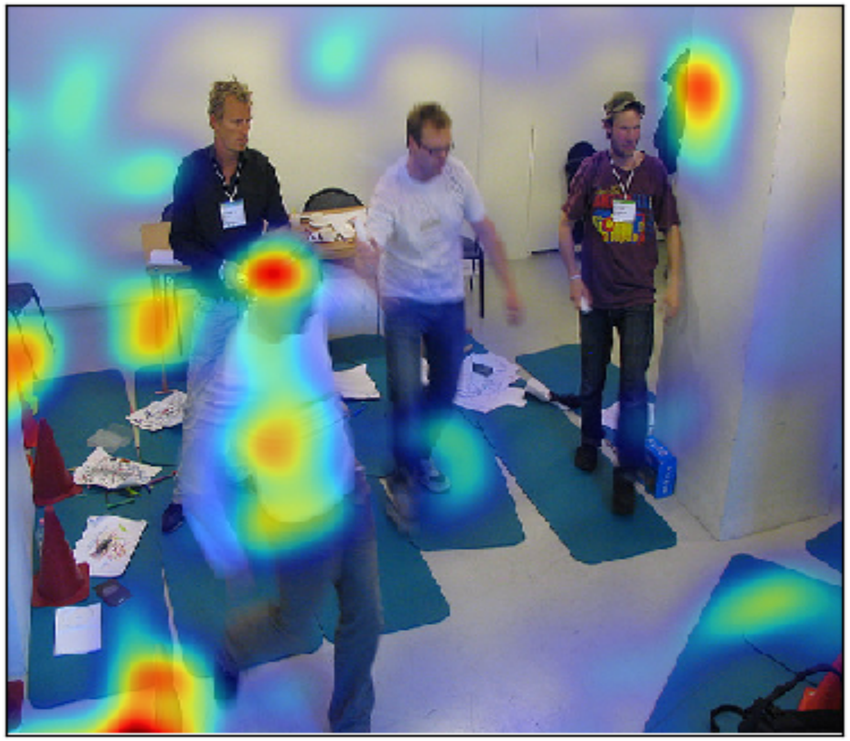
ToAIM

- Explanation Modalities

ToAIM

- Explanation Modalities
- Grad-CAM: implicit attention mechanism

How many people are there?

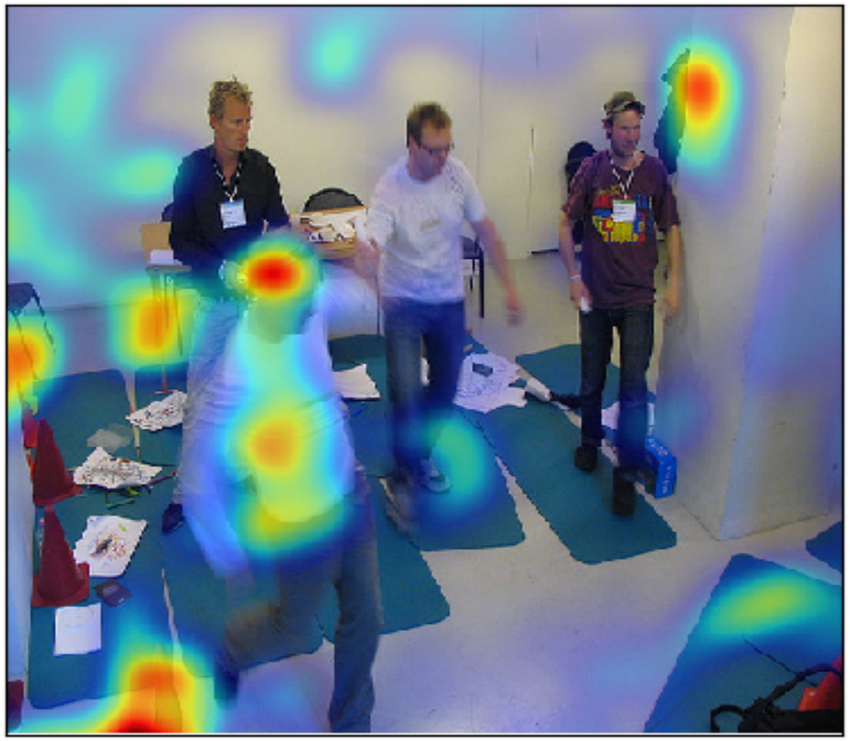


Grad-CAM

ToAIM

- Explanation Modalities
- Grad-CAM: implicit attention mechanism

How many people are there?



Grad-CAM

Attention visualization corresponding to Vicki's most confident answer

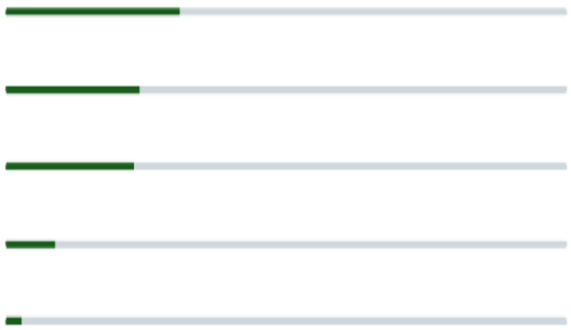
ToAIM

- Explanation Modalities

ToAIM

- Explanation Modalities
- Vicki’s confidence in top-5 answers

How many people are there?

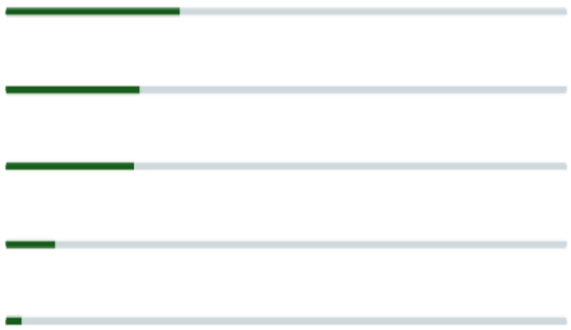


Top-5 answer confidence

ToAIM

- Explanation Modalities
- Vicki’s confidence in top-5 answers

How many people are there?



Vicki’s confidence in the top-5 answers without revealing the answers

Top-5 answer confidence

ToAIM

ToAIM

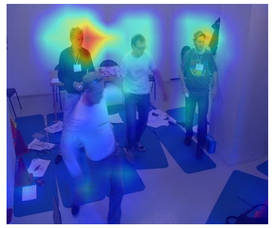
- Training + Explanation Modalities

ToAIM

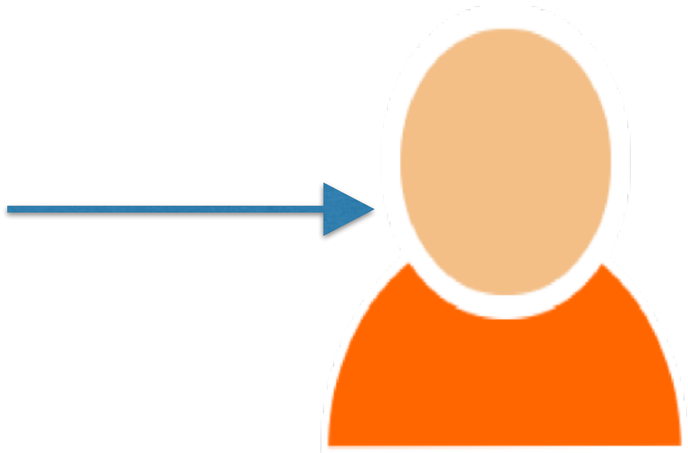
- Training + Explanation Modalities

How many people are there?

+



How many people are there?

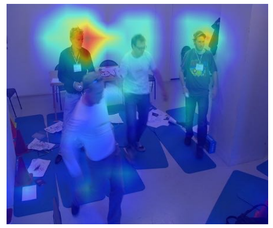


ToAIM

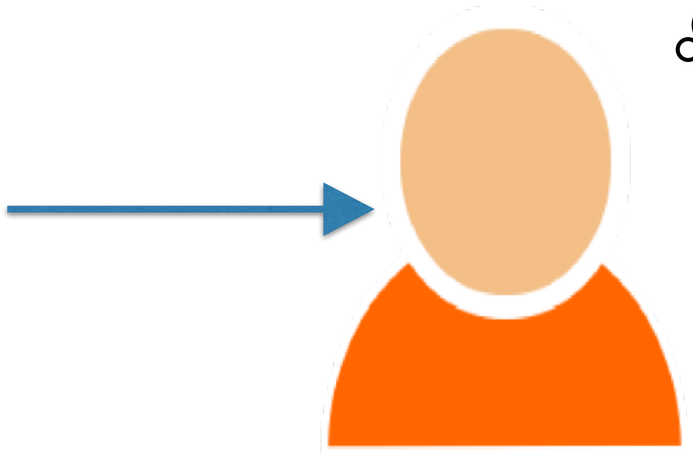
- Training + Explanation Modalities

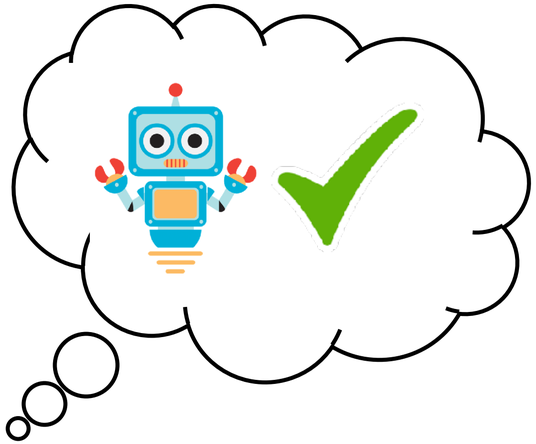
How many people are there?

+



How many people are there?



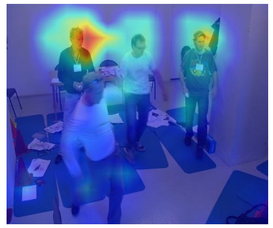


ToAIM

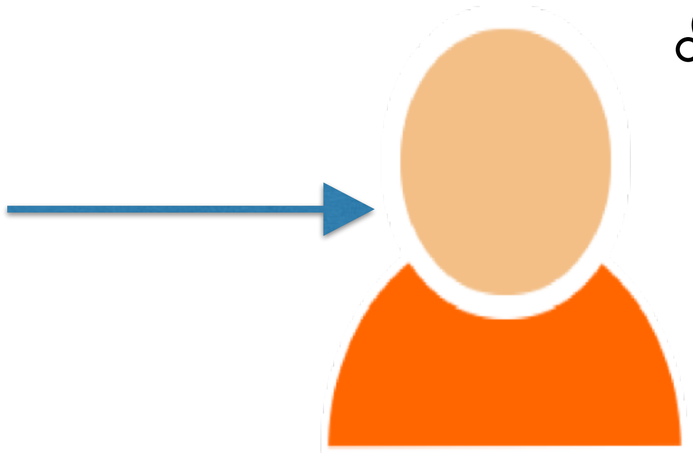
- Training + Explanation Modalities

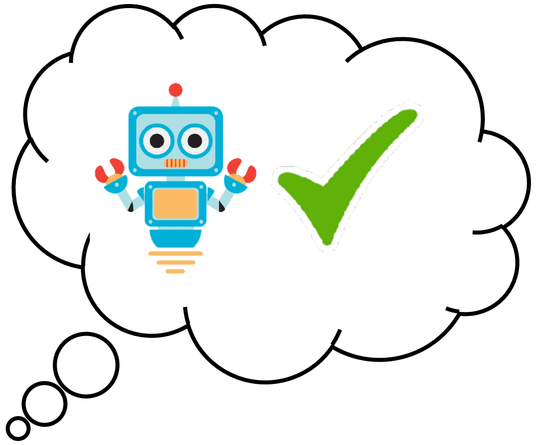
How many people are there?

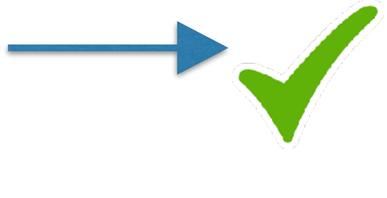
+

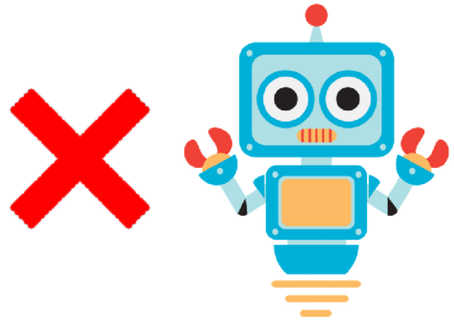


How many people are there?







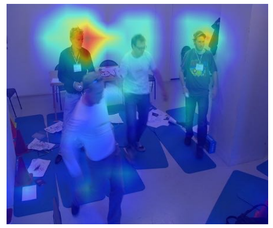


ToAIM

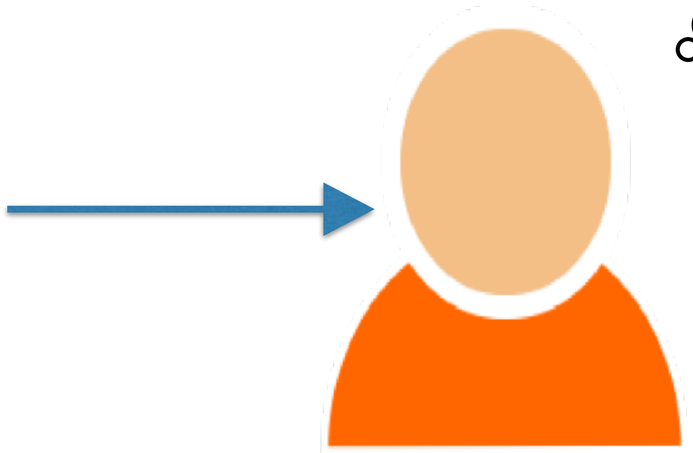
- Training + Explanation Modalities

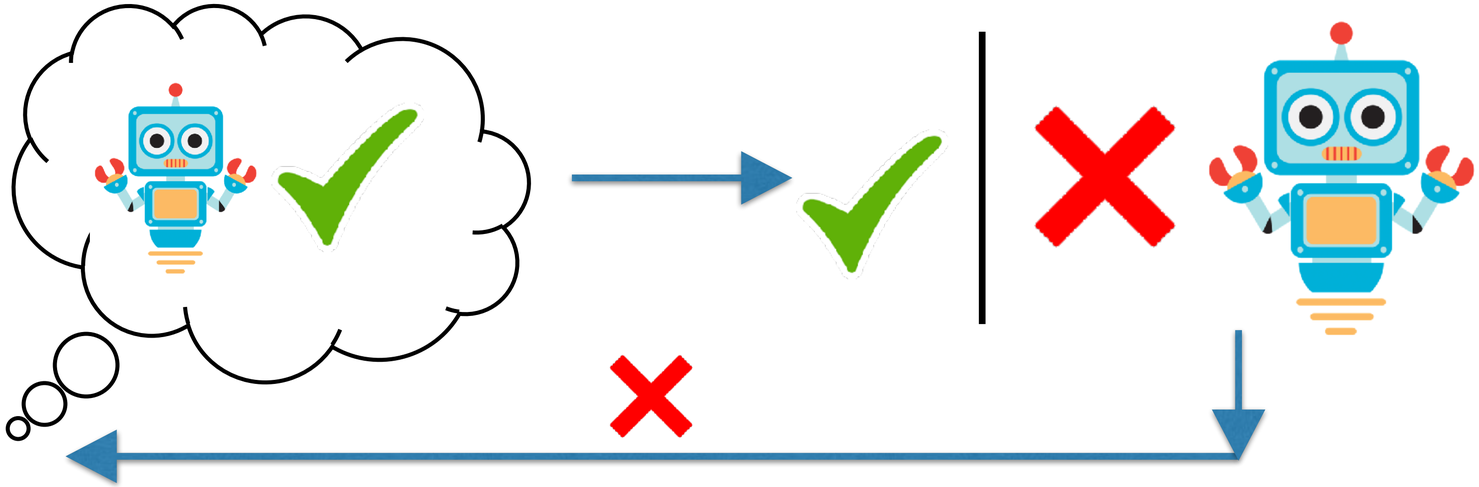
How many people are there?

+



How many people are there?





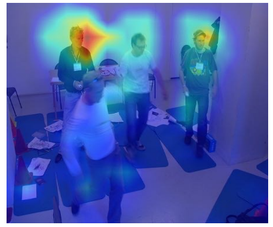
Feedback

ToAIM

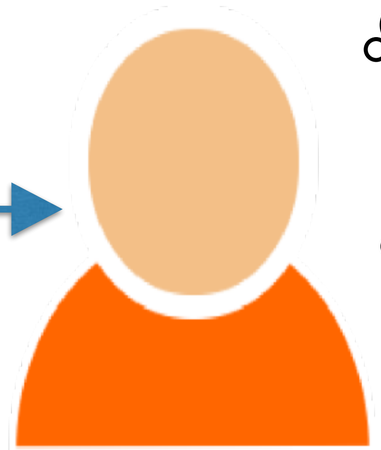
- Training + Explanation Modalities

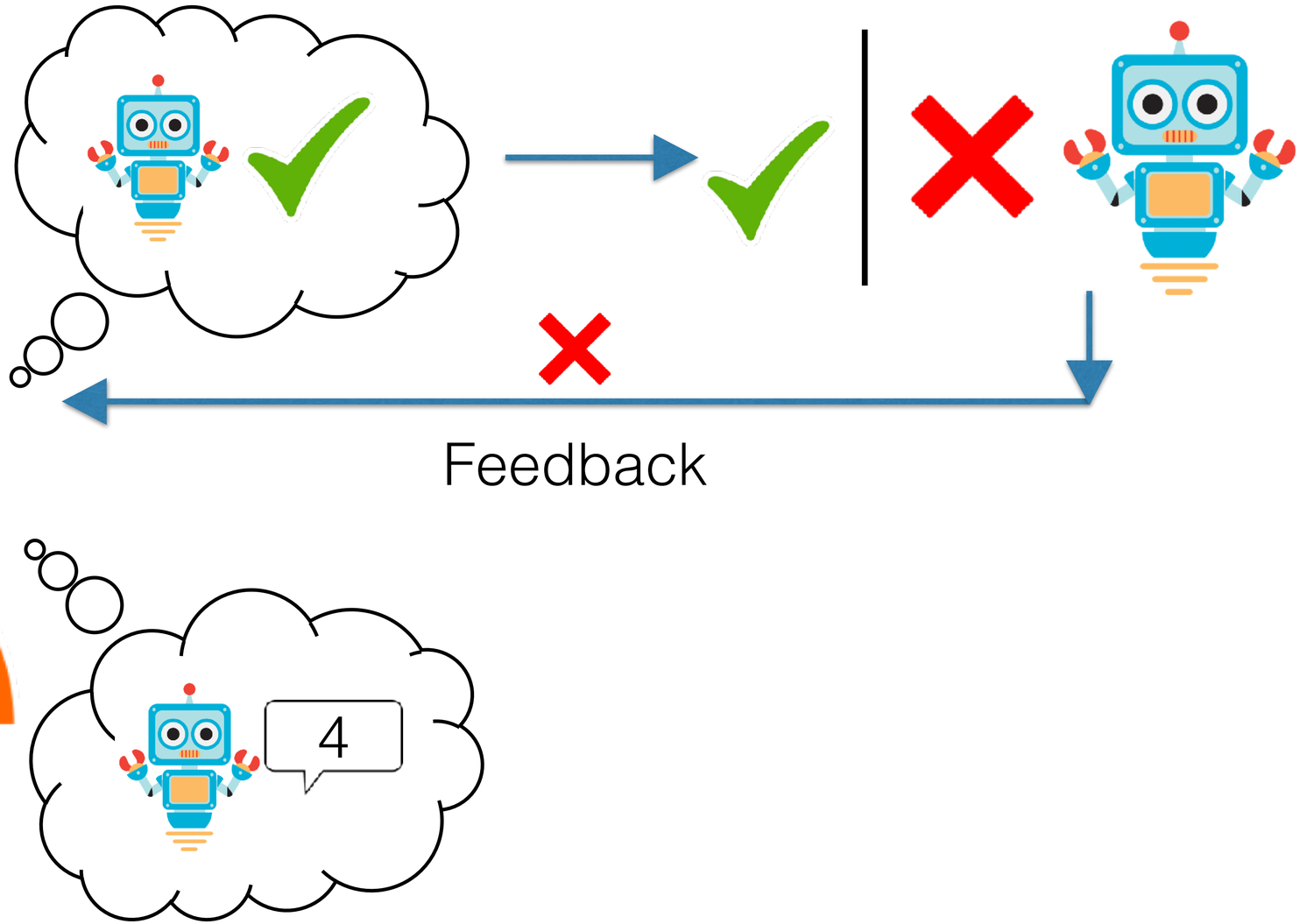
How many people are there?

+



How many people are there?





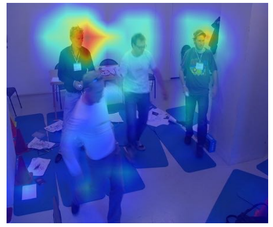
Feedback

ToAIM

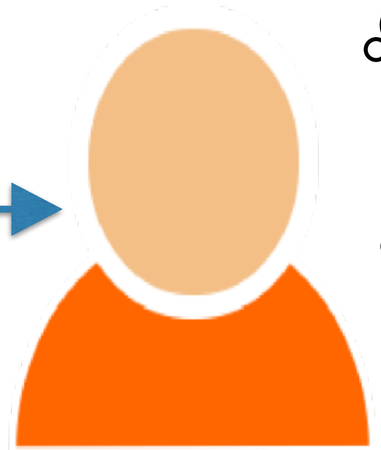
- Training + Explanation Modalities

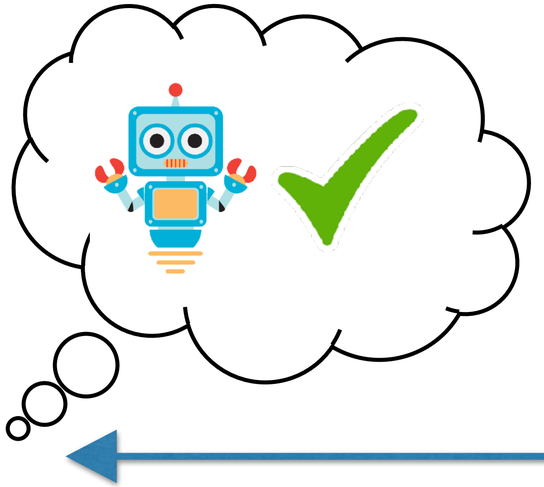
How many people are there?

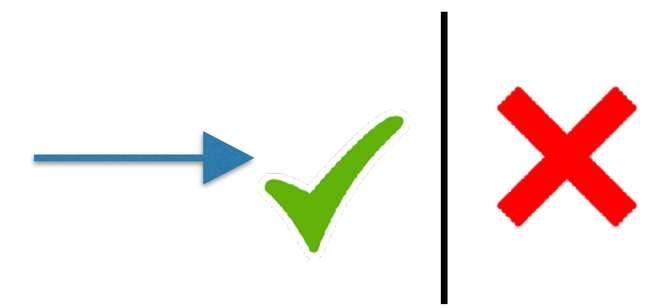
+



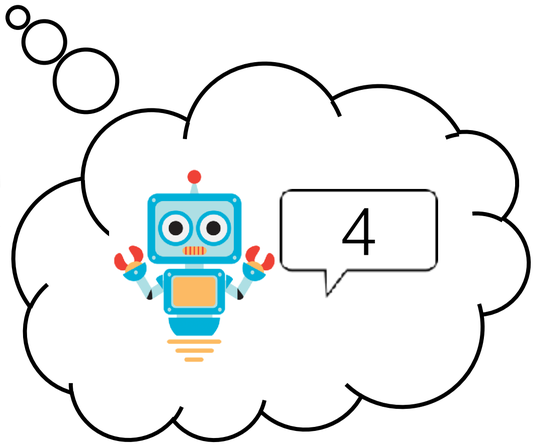
How many people are there?



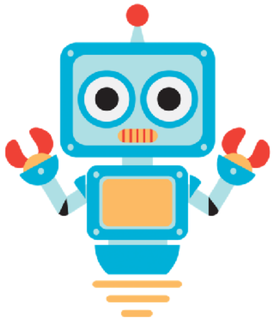




Feedback



4 | 3

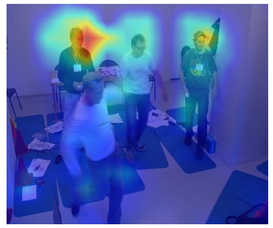


ToAIM

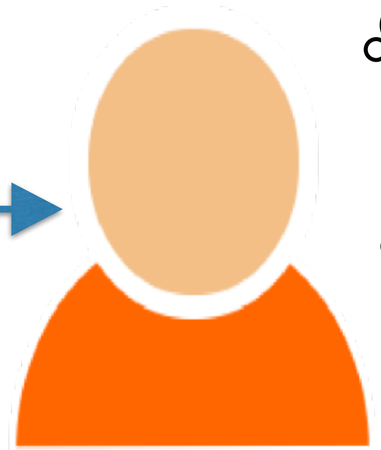
- Training + Explanation Modalities

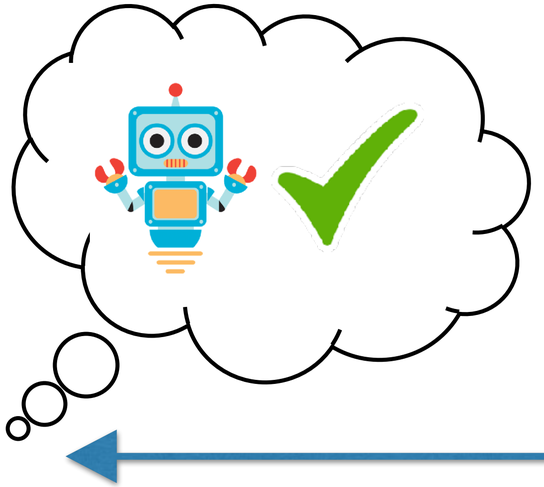
How many people are there?

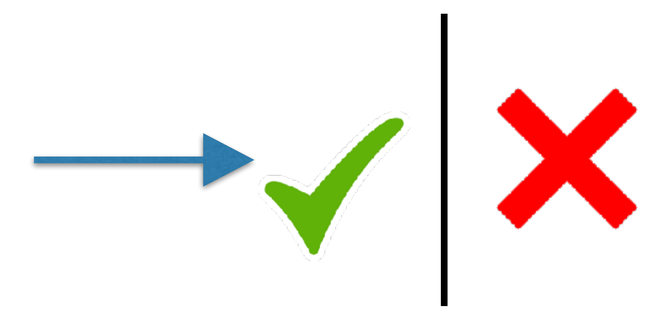
+

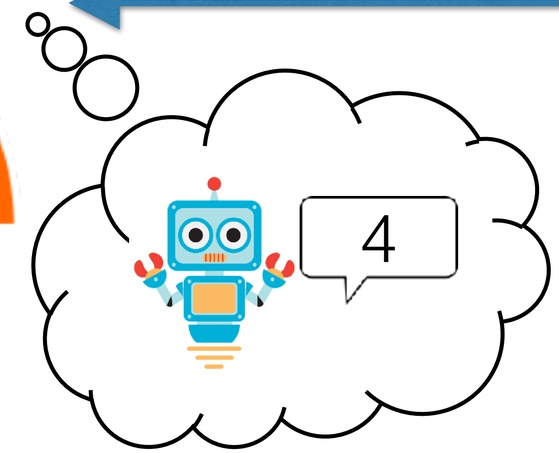


How many people are there?









Feedback

3

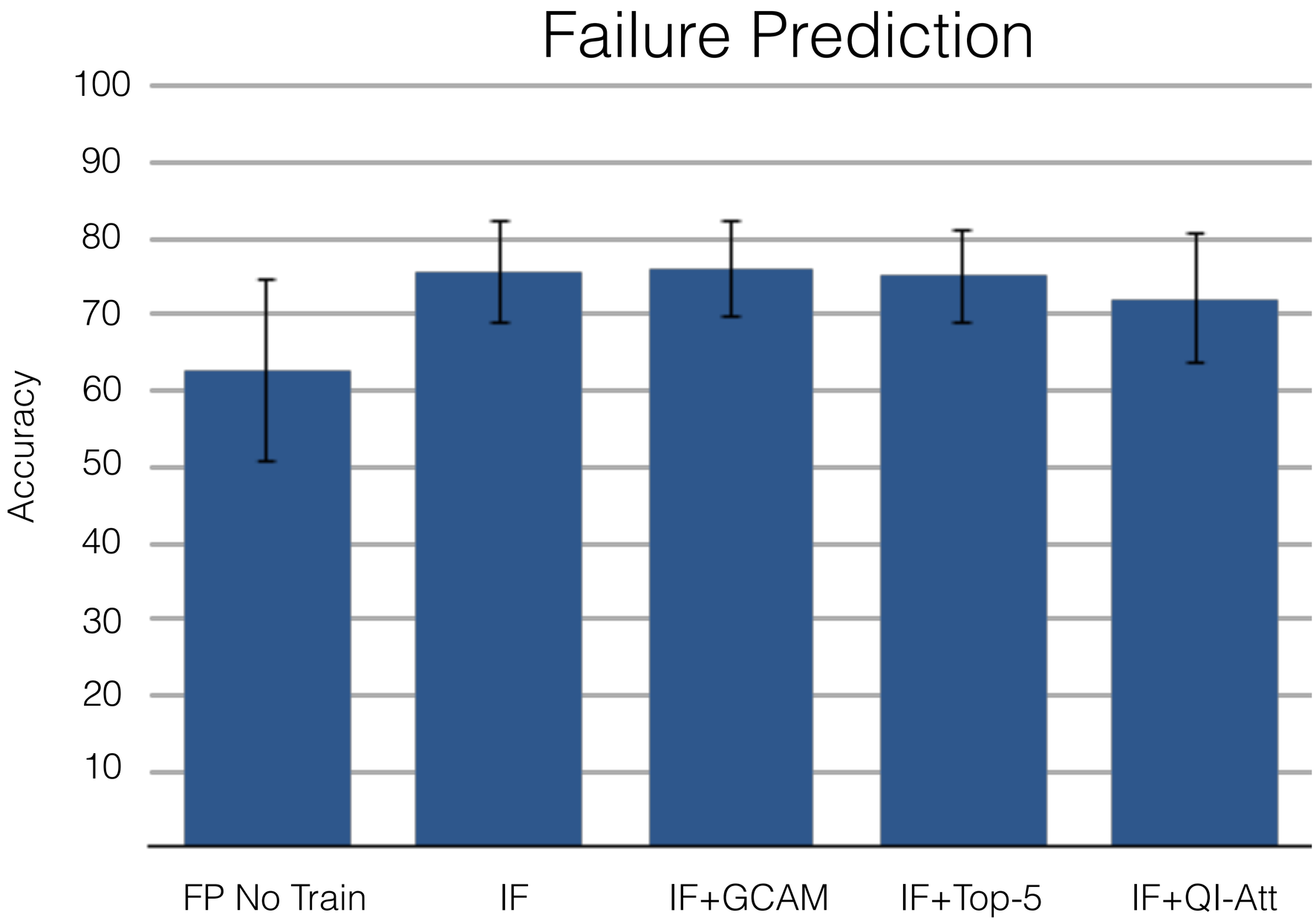
4 | 3

Experimental Results

Experimental Results

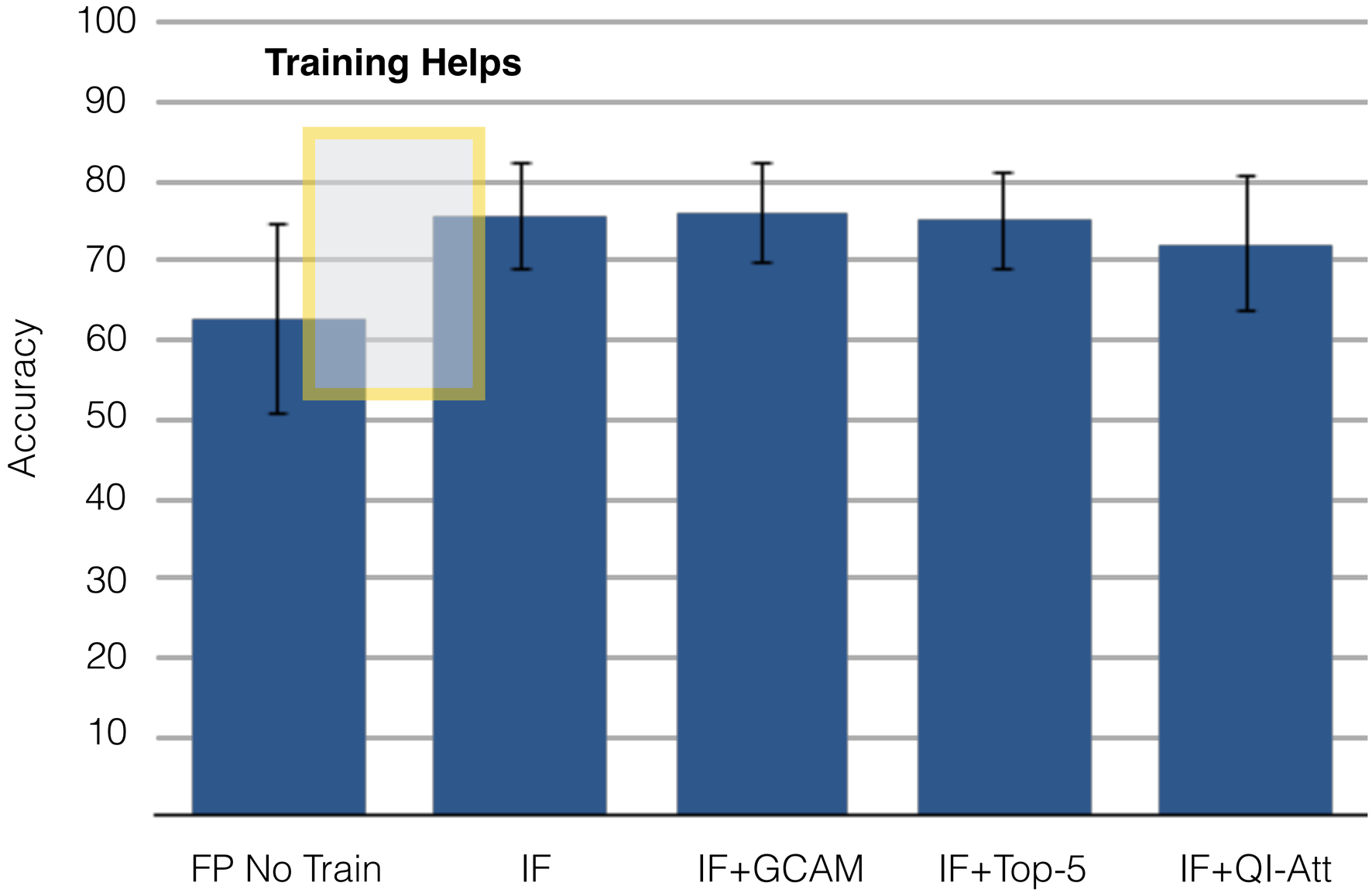
Failure Prediction

Experimental Results



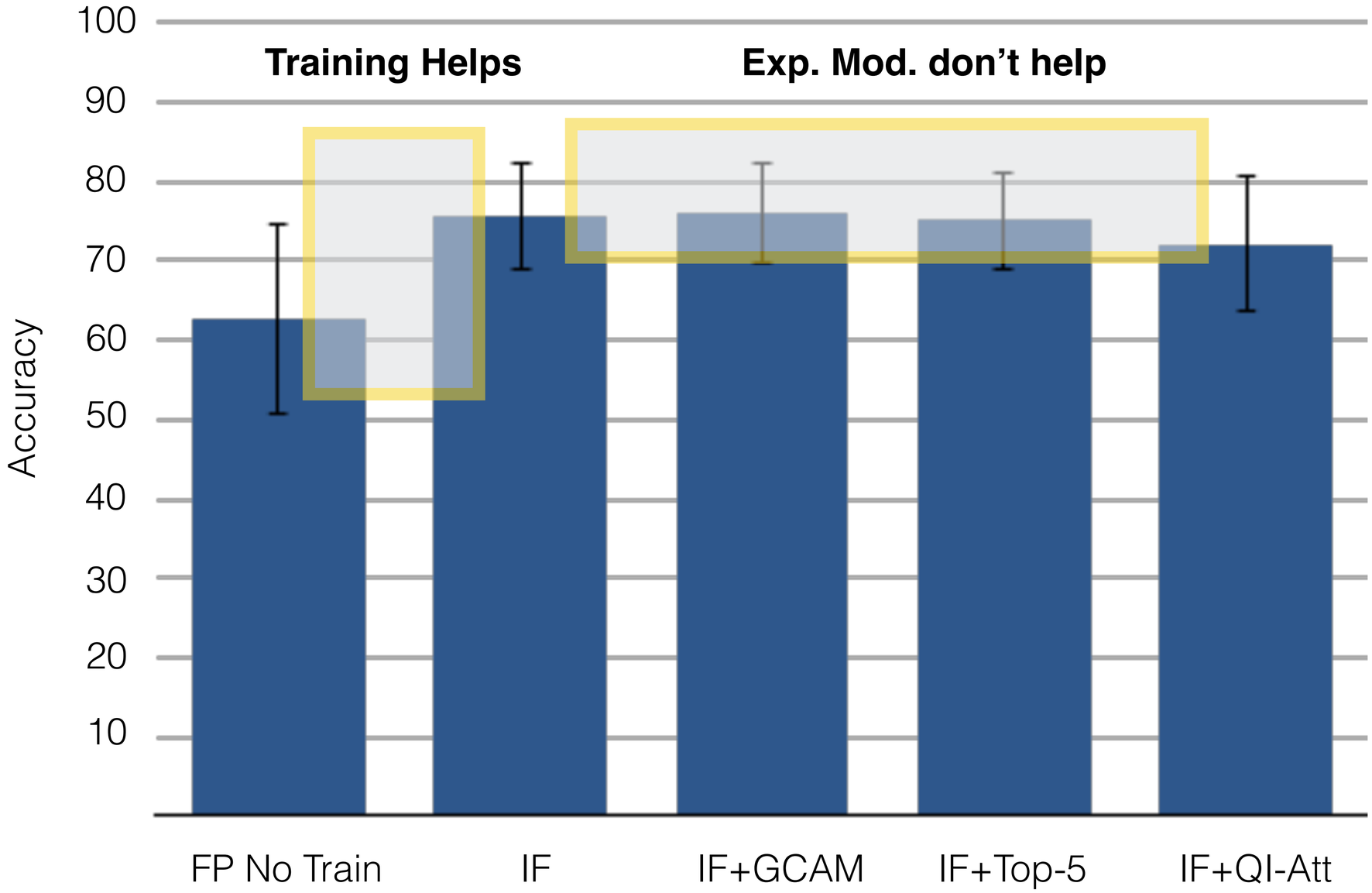
Experimental Results

Failure Prediction



Experimental Results

Failure Prediction



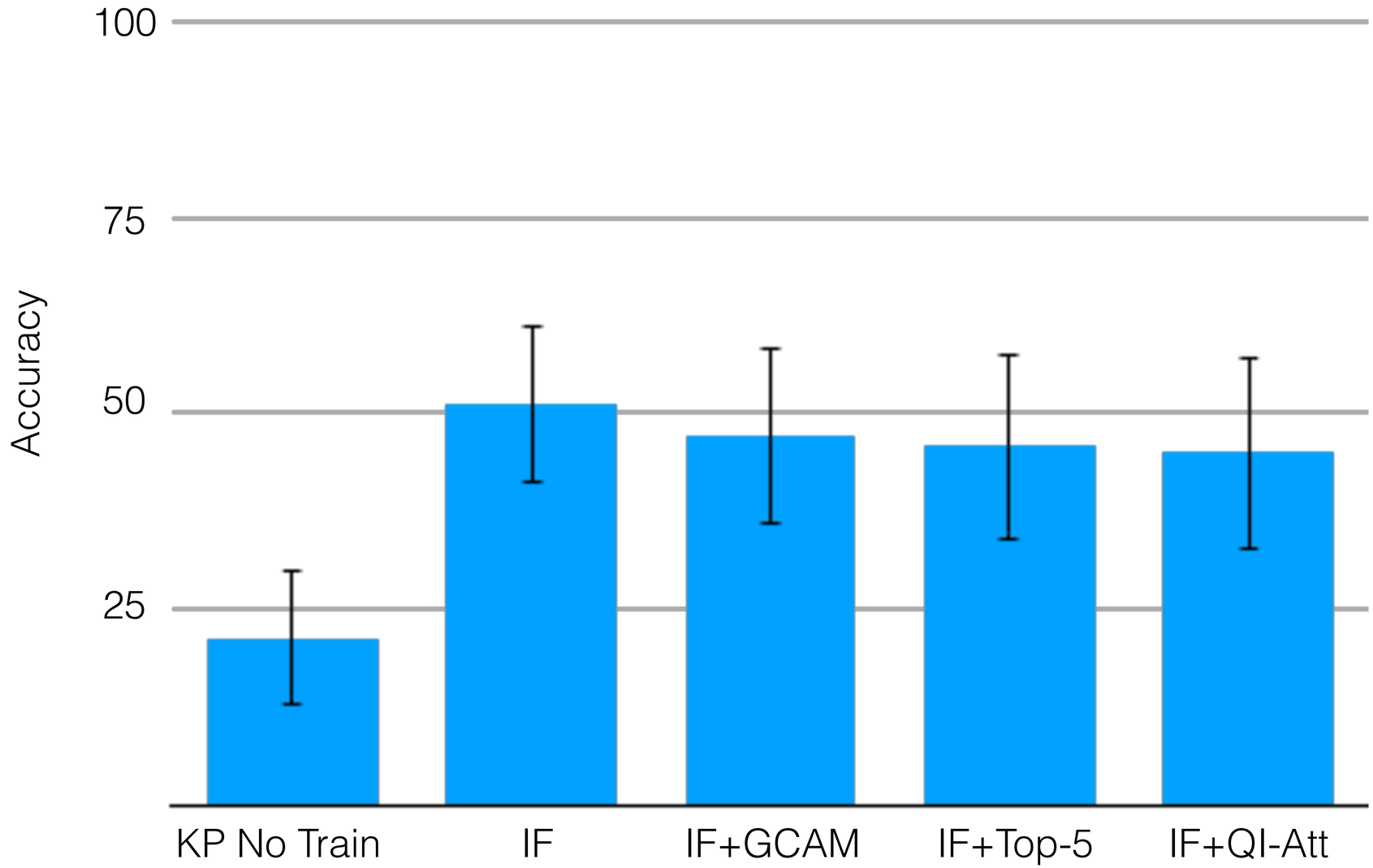
Experimental Results

Experimental Results

Knowledge Prediction

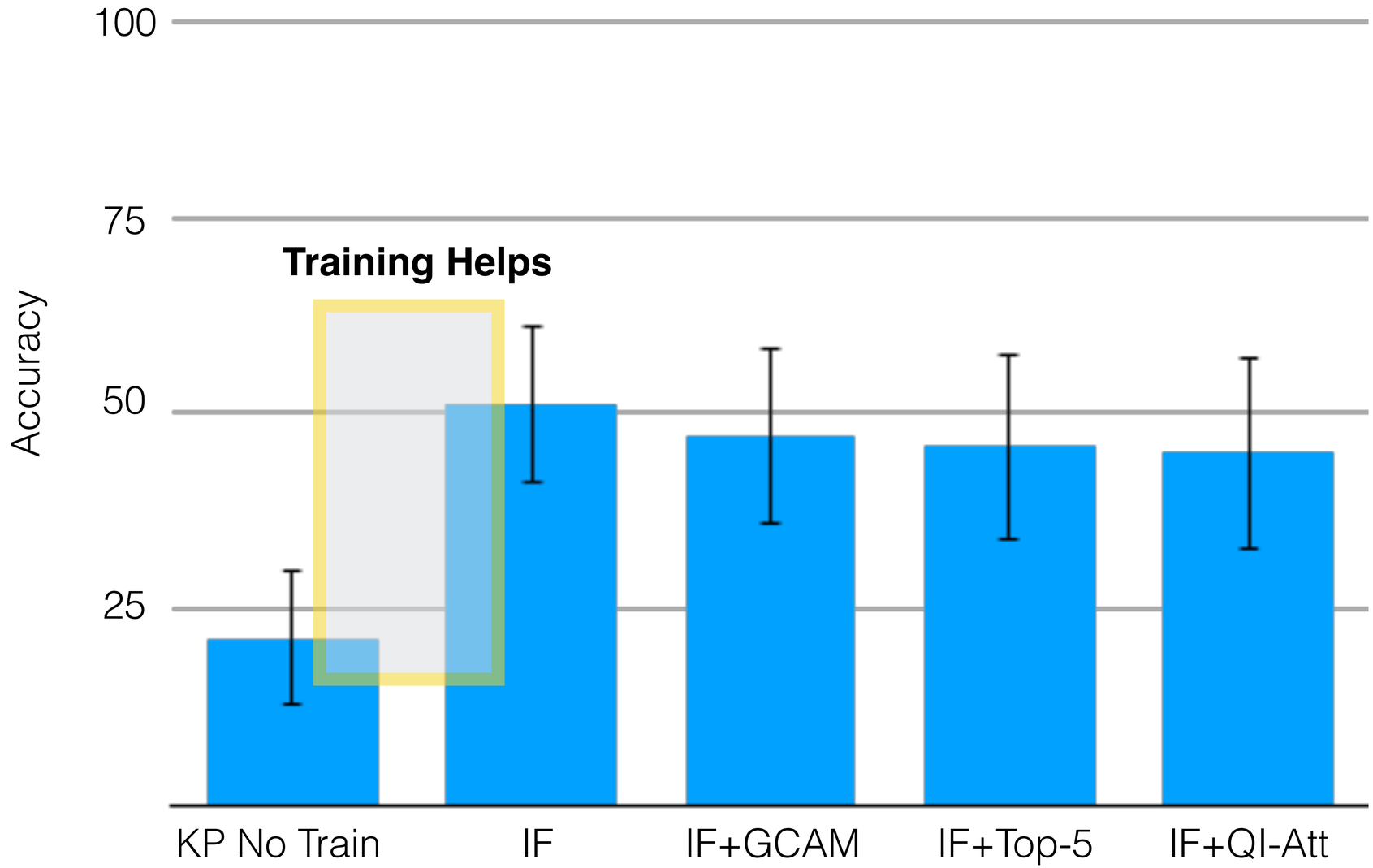
Experimental Results

Knowledge Prediction



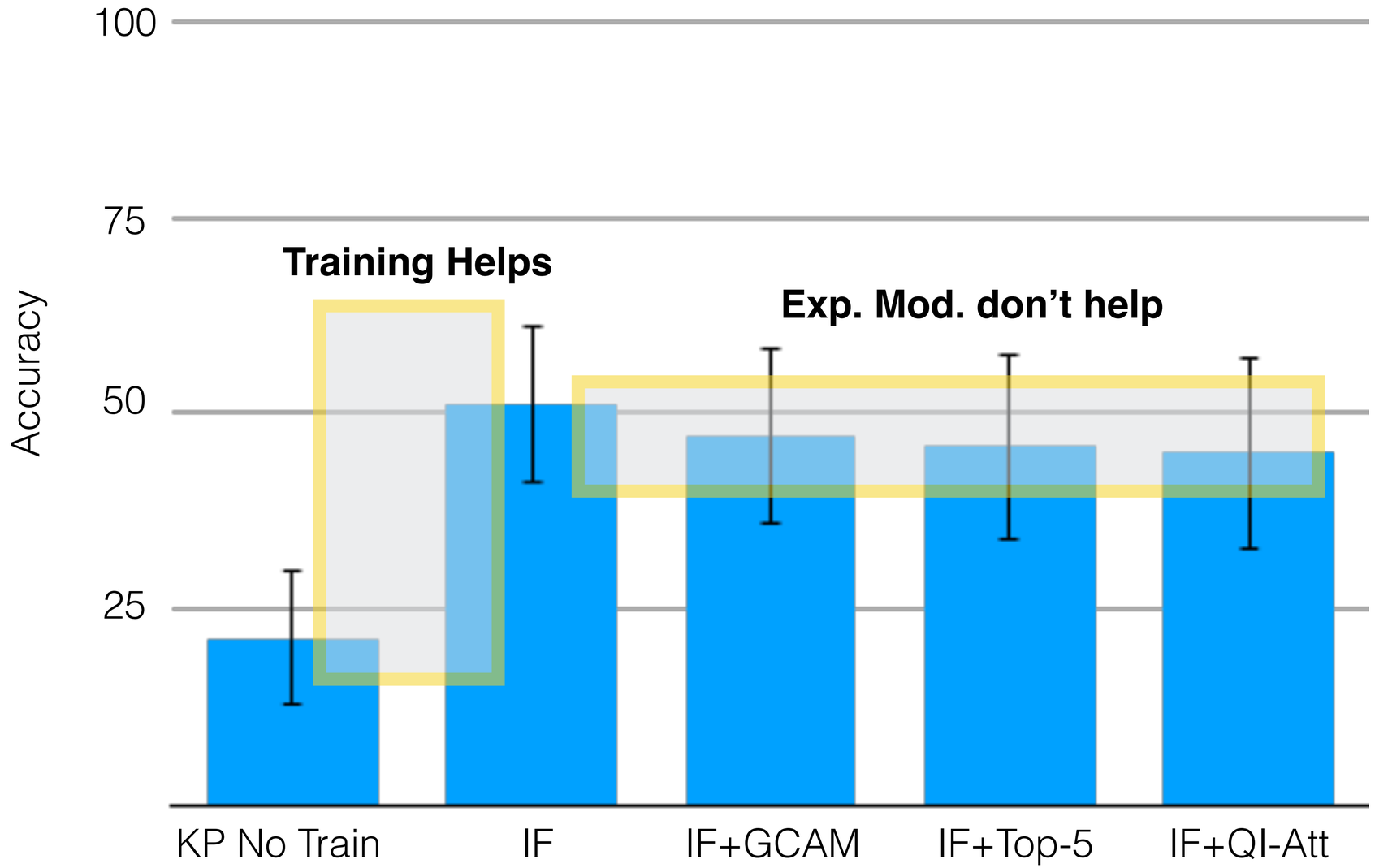
Experimental Results

Knowledge Prediction



Experimental Results

Knowledge Prediction



Outline

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Take-away messages

Take-away messages

- Advocate a research agenda towards ToAIM

Take-away messages

- Advocate a research agenda towards ToAIM
- Lay people successfully form ToM with a few (50) examples

Take-away messages

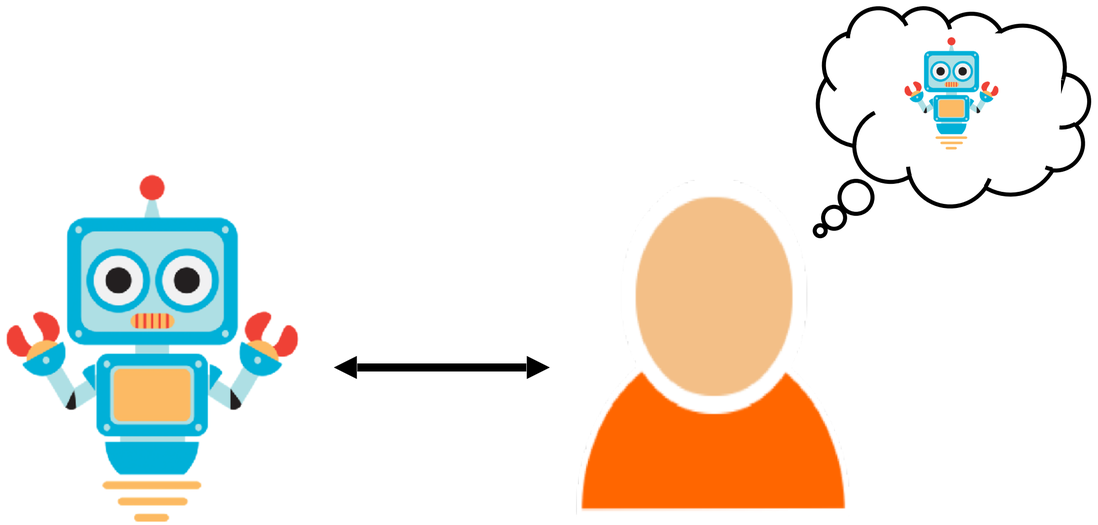
- Advocate a research agenda towards ToAIM
- Lay people successfully form ToM with a few (50) examples
- Existing explanation modalities don't help in predicting what AI will do

Take-away messages

- Advocate a research agenda towards ToAIM
- Lay people successfully form ToM with a few (50) examples
- Existing explanation modalities don't help in predicting what AI will do
- Novel evaluation protocol for explanation modalities

Take-away messages

- Advocate a research agenda towards ToAIM
- Lay people successfully form ToM with a few (50) examples
- Existing explanation modalities don't help in predicting what AI will do
- Novel evaluation protocol for explanation modalities



ArXiv

Computer Science > Computer Vision and Pattern Recognition

It Takes Two to Tango: Towards Theory of AI's Mind

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, Devi Parikh

(Submitted on 3 Apr 2017)

Theory of Mind is the ability to attribute mental states (beliefs, intents, knowledge, perspectives, etc.) to others and recognize that these mental states may differ from one's own. Theory of Mind is critical to effective communication and to teams demonstrating higher collective performance. To effectively leverage the progress in Artificial Intelligence (AI) to make our lives more productive, it is important for humans and AI to work well together in a team. Traditionally, there has been much emphasis on research to make AI more accurate, and (to a lesser extent) on having it better understand human intentions, tendencies, beliefs, and contexts. The latter involves making AI more human-like and having it develop a theory of our minds.

In this work, we argue that for human-AI teams to be effective, humans must also develop a theory of AI's mind - get to know its strengths, weaknesses, beliefs, and quirks. We instantiate these ideas within the domain of Visual Question Answering (VQA). We find that using just a few examples(50), lay people can be trained to better predict responses and oncoming failures of a complex VQA model. Surprisingly, we find that having access to the model's internal states - its confidence in its top-k predictions, explicit or implicit attention maps which highlight regions in the image (and words in the question) the model is looking at (and listening to) while answering a question about an image - do not help people better predict its behavior

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL)
Cite as: arXiv:1704.00717 [cs.CV]
(or arXiv:1704.00717v1 [cs.CV] for this version)

Outline

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Outline

Theory of AI's mind (ToAIM): Motivation

Theory of AI's mind (ToAIM): Experimental Setup and Results

Theory of AI's mind (ToAIM): Take-away messages

Theory of AI's mind (ToAIM): Ongoing work - Human-AI Games

Ongoing Work

Ongoing Work

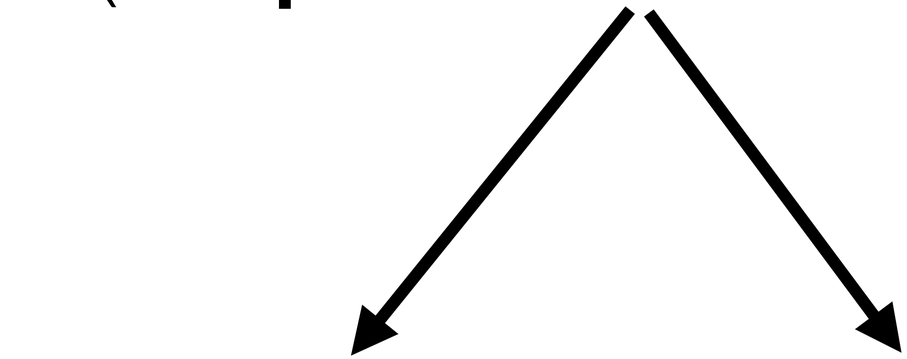
- Evaluate human-AI collaborative performance

Ongoing Work

- Evaluate human-AI collaborative performance
- Goal-driven tasks (**cooperative human-AI games**)

Ongoing Work

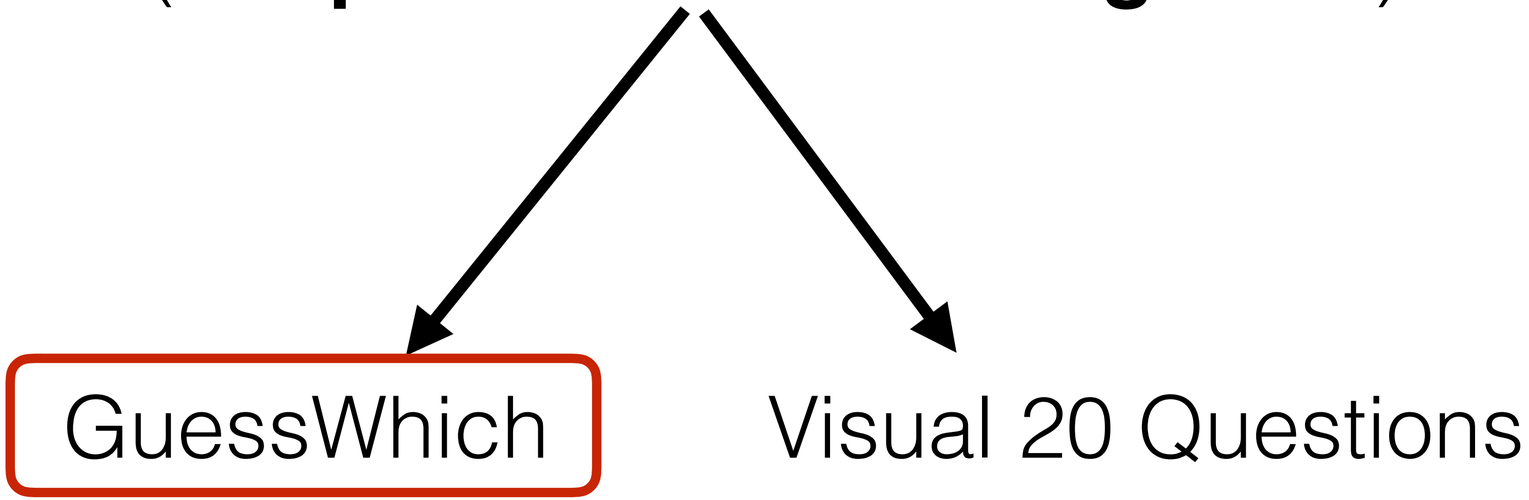
- Evaluate human-AI collaborative performance
- Goal-driven tasks (**cooperative human-AI games**)



GuessWhich Visual 20 Questions

Ongoing Work

- Evaluate human-AI collaborative performance
- Goal-driven tasks (**cooperative human-AI games**)



GuessWhich

GuessWhich

- **Players**: 2 Agents - Questioner - **Q** and Answerer - **A**

GuessWhich

- **Players**: 2 Agents - Questioner - **Q** and Answerer - **A**
- **Parameters**:

GuessWhich

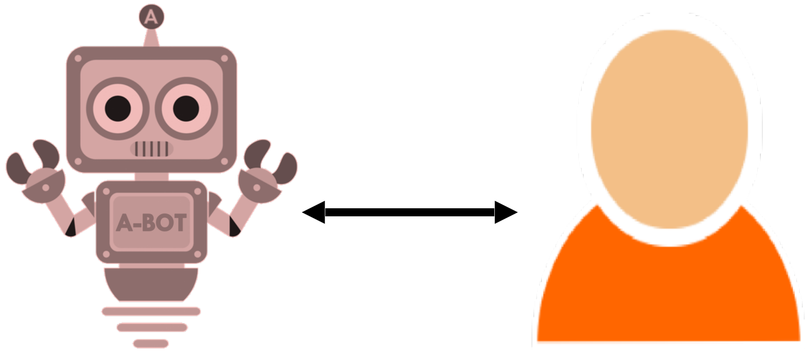
- **Players**: 2 Agents - Questioner - **Q** and Answerer - **A**
- **Parameters**:
    - Fixed pool of images 

GuessWhich

- **Players**: 2 Agents - Questioner - **Q** and Answerer - **A**
- **Parameters**:
    - Fixed pool of images 
    - Fixed number of rounds of dialog (10)

GuessWhich

- **Players**: 2 Agents - Questioner - **Q** and Answerer - **A**
- **Parameters**:
    - Fixed pool of images 
    - Fixed number of rounds of dialog (10)
- In our implementation, Human (Questioner) and AI (Answerer)



GuessWhich

GuessWhich

A selects an image from

GuessWhich

A selects an image from 

GuessWhich

A selects an image from 

GuessWhich

A selects an image from  unknown to **Q**

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

Q asks questions so as to locate 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

Q asks questions so as to locate  from 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

Q asks questions so as to locate  from 

A answers **Q**'s questions according to 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

Q asks questions so as to locate  from 

A answers **Q**'s questions according to 

Q makes a guess about 

GuessWhich

A selects an image from  unknown to **Q**

A and **Q** are given a caption to get a idea about 

Q makes a guess about 

Q asks questions so as to locate  from 

A answers **Q**'s questions according to 

Q makes a guess about  after every round of dialog

GuessWhich

GuessWhich

- **Human-AI Team**: Human (Questioner) and AI (Answerer)

GuessWhich

- **Human-AI Team**: Human (Questioner) and AI (Answerer)
- Image retrieval setting. Metrics: mean-rank & mean-reciprocal rank

GuessWhich

- **Human-AI Team**: Human (Questioner) and AI (Answerer)
- Image retrieval setting. Metrics: mean-rank & mean-reciprocal rank
- Our AI Agent ⟶ Visual Conversation Agents

GuessWhich

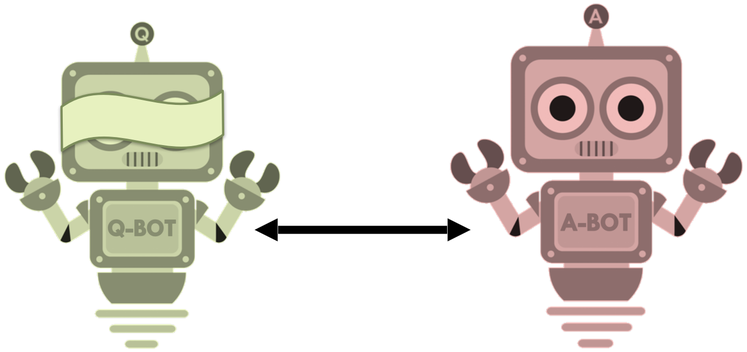
- **Human-AI Team**: Human (Questioner) and AI (Answerer)
- Image retrieval setting. Metrics: mean-rank & mean-reciprocal rank
- Our AI Agent ⟶ Visual Conversation Agents
  - Das & Kottur et al. ICCV 2017

Abhishek Das

Satwik Kottur

GuessWhich

- **Human-AI Team**: Human (Questioner) and AI (Answerer)
- Image retrieval setting. Metrics: mean-rank & mean-reciprocal rank
- Our AI Agent ⟶ Visual Conversation Agents
  - Das & Kottur et al. ICCV 2017



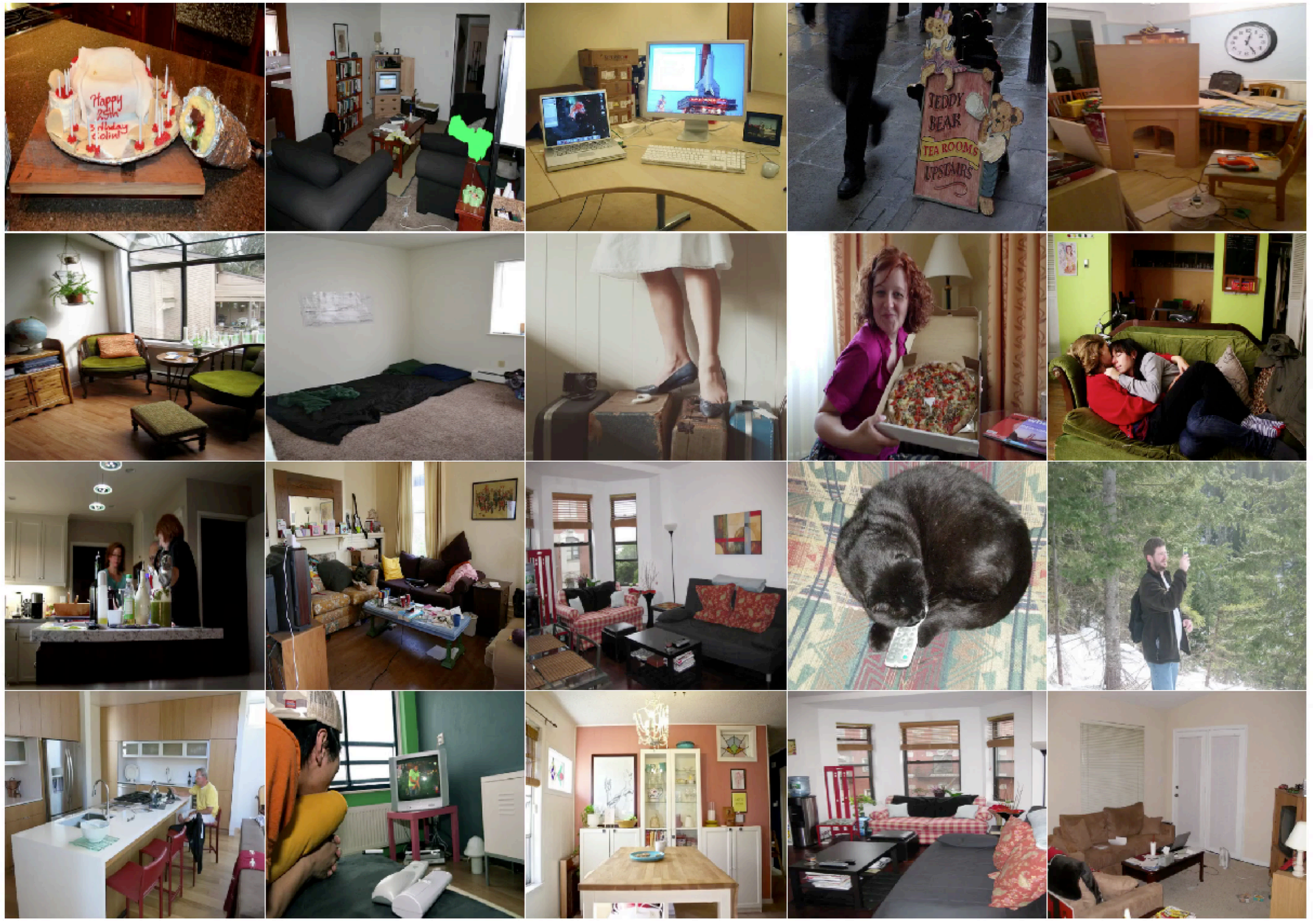
Questioner Answerer

Abhishek Das

Satwik Kottur

GuessWhich Game

Total points earned: **0** Possible points for this game: **200**



Hi, my name is Abot. I am an Artificial Intelligence. I have been assigned one of these images as the target image. I am not allowed to show you the image, but as a start, I will describe the image to you in a sentence. You can then ask me follow up questions about it. When ready, submit one of the images on the left as your best guess. I will try to describe the image and answer your questions, but I am not perfect. I make quite a few mistakes. I hope we can work together to find the image! Let's do this! Note: My knowledge of English is limited. Sometimes if I don't know the right word, I say UNK. You will win points based on how accurately you are able to guess.

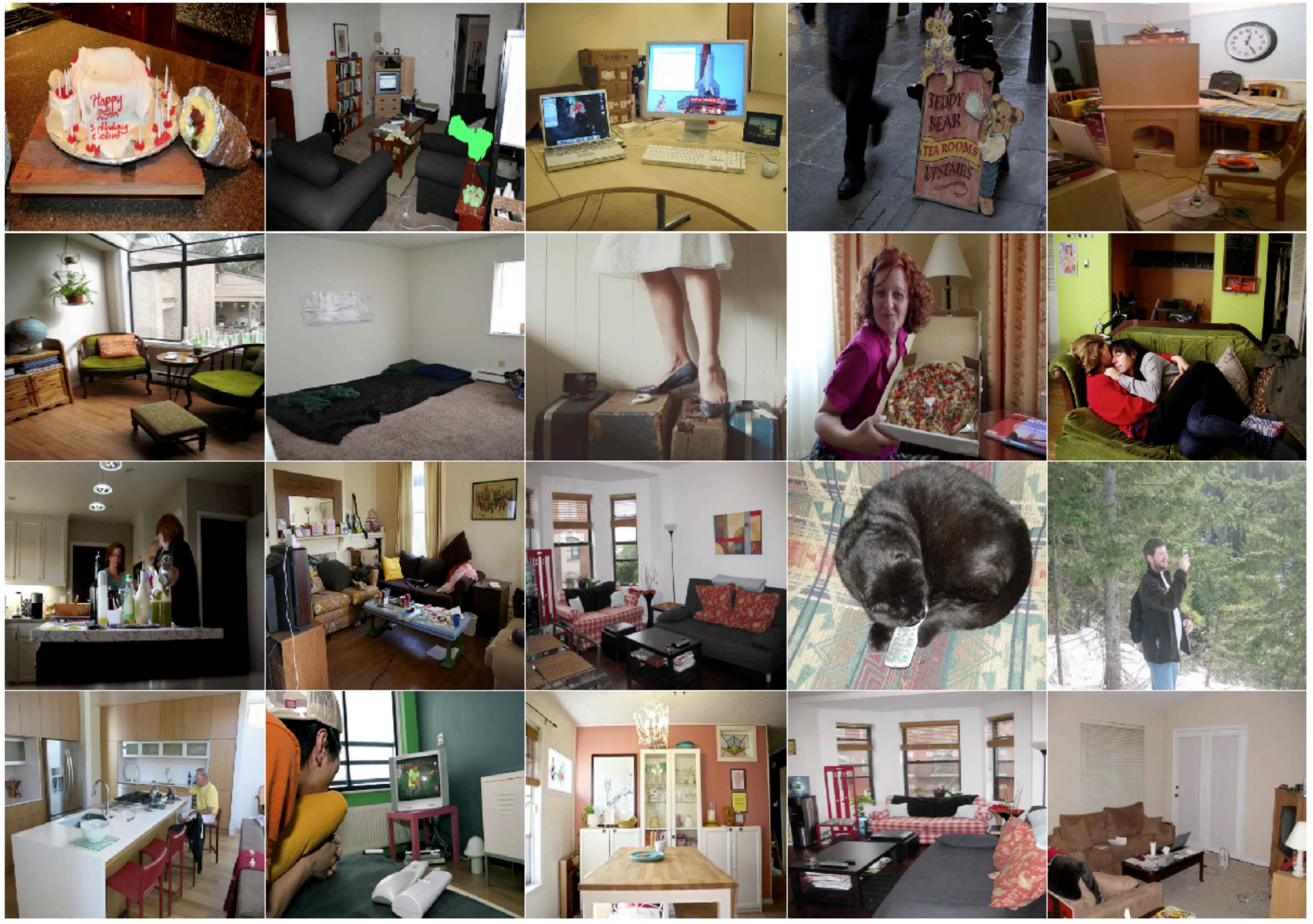
Target image description: a man sitting on a couch with a laptop

Based on your understanding of the image description, pick the image that you think is the most relevant.

Click on the image that is your best guess

GuessWhich Game

Total points earned: **0** Possible points for this game: **200**



Hi, my name is Abot. I am an Artificial Intelligence. I have been assigned one of these images as the target image. I am not allowed to show you the image, but as a start, I will describe the image to you in a sentence. You can then ask me follow up questions about it. When ready, submit one of the images on the left as your best guess. I will try to describe the image and answer your questions, but I am not perfect. I make quite a few mistakes. I hope we can work together to find the image! Let's do this! Note: My knowledge of English is limited. Sometimes if I don't know the right word, I say UNK. You will win points based on how accurately you are able to guess.

Target image description: a man sitting on a couch with a laptop

Based on your understanding of the image description, pick the image that you think is the most relevant.

Click on the image that is your best guess

That’s all folks!
Questions?

- Interface videos: https://deshraj.github.io/TOAIM/
- Interfaces: https://github.com/deshraj/TOAIM/tree/master/Assets/Interfaces
- Guess-Which: http://gw.cloudcv.org/