

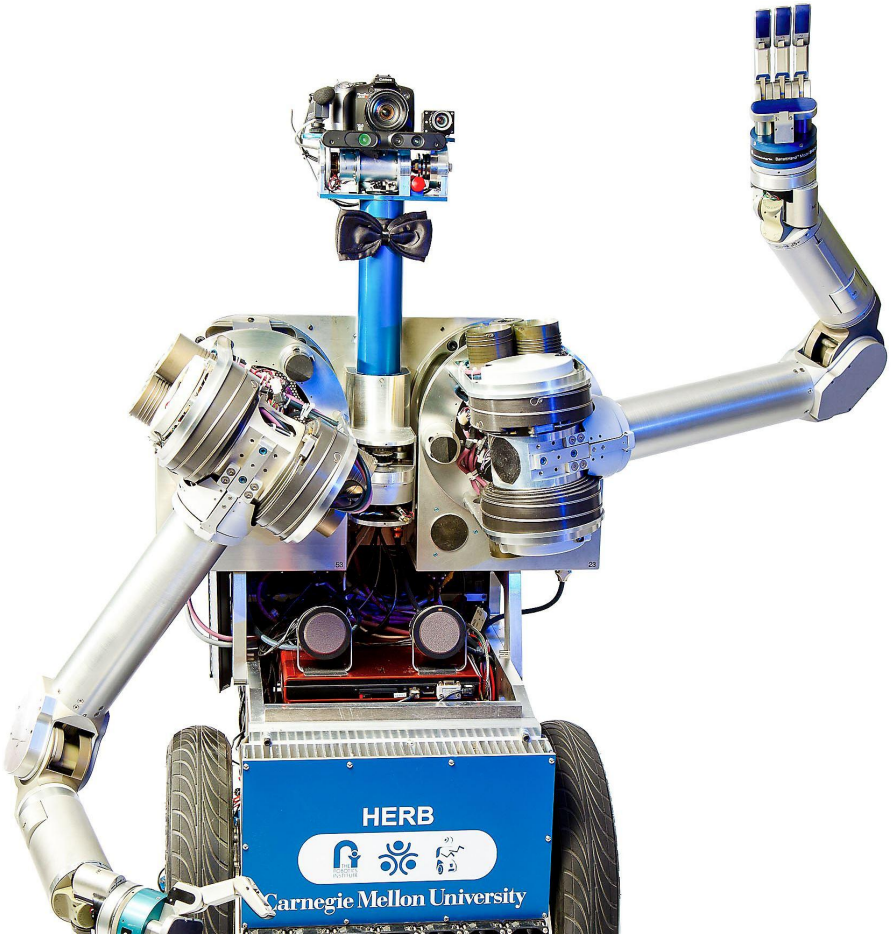

Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation

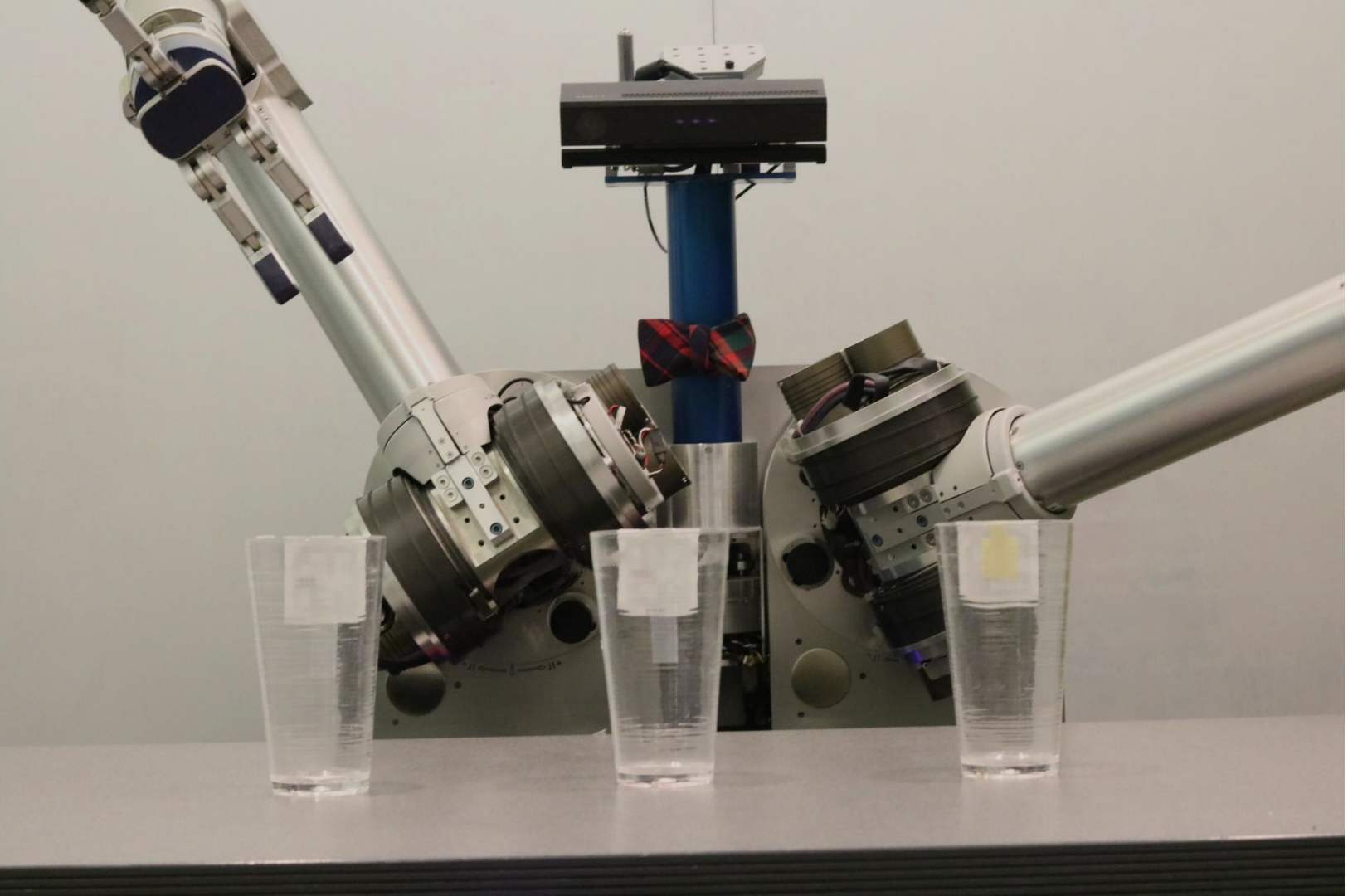
Rosario Scalise, Shen Li
Henny Admoni, Stephanie Rosenthal, Siddhartha S. Srinivasa

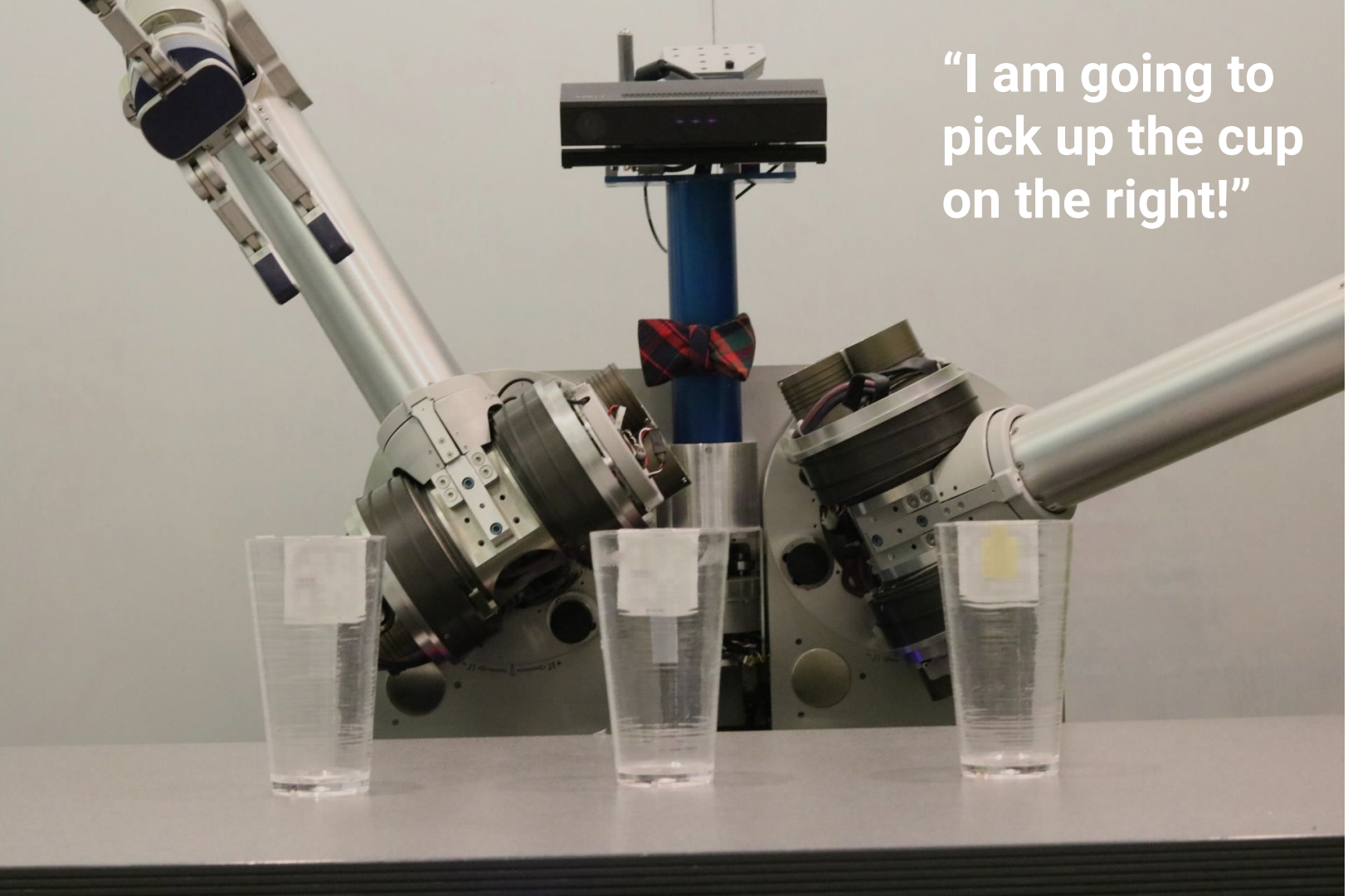
Carnegie Mellon

- Background, why tabletop is important
- Problem: object uniqueness
  - Solution 1: spatial reference
  - Solution 2: perspective
- Study 1
  - Image generation
  - Study design
  - Result
    - Human vs robot
    - Visual search + word frequencies
    - Difficulty
- Study 2
  - Data coding
  - Study design
  - Result
    - Block ambiguity
    - Perspective
- Discussion
  - 3 approaches to give instructions
  - Block ambiguity and perspective ambiguity
  - Neither perspective is the best
  - Future work - interactivity

Herb image courtesy of Pittsburgh Post-Gazette






“I am going to pick up the cup on the right!”

Key Issue: Ambiguity

Question by Jessica Lock from the Noun Project

Key Issue: Ambiguity

As scene complexity increases, so does the difficulty in specifying an object.

Key Issue: Ambiguity

As scene complexity increases, so does the difficulty in specifying an object.

Natural language is inherently ambiguous.

Forms of Ambiguity

Visual Appearance

“Pick up the coffee cup.”

Forms of Ambiguity

Visual Appearance

“Pick up the coffee cup.”

Which one?

Forms of Ambiguity

Perspective

“Pick up the coffee cup on the right.”

Forms of Ambiguity

Perspective

“Pick up the coffee cup on the right.”

Whose right?

Forms of Ambiguity

Proximity

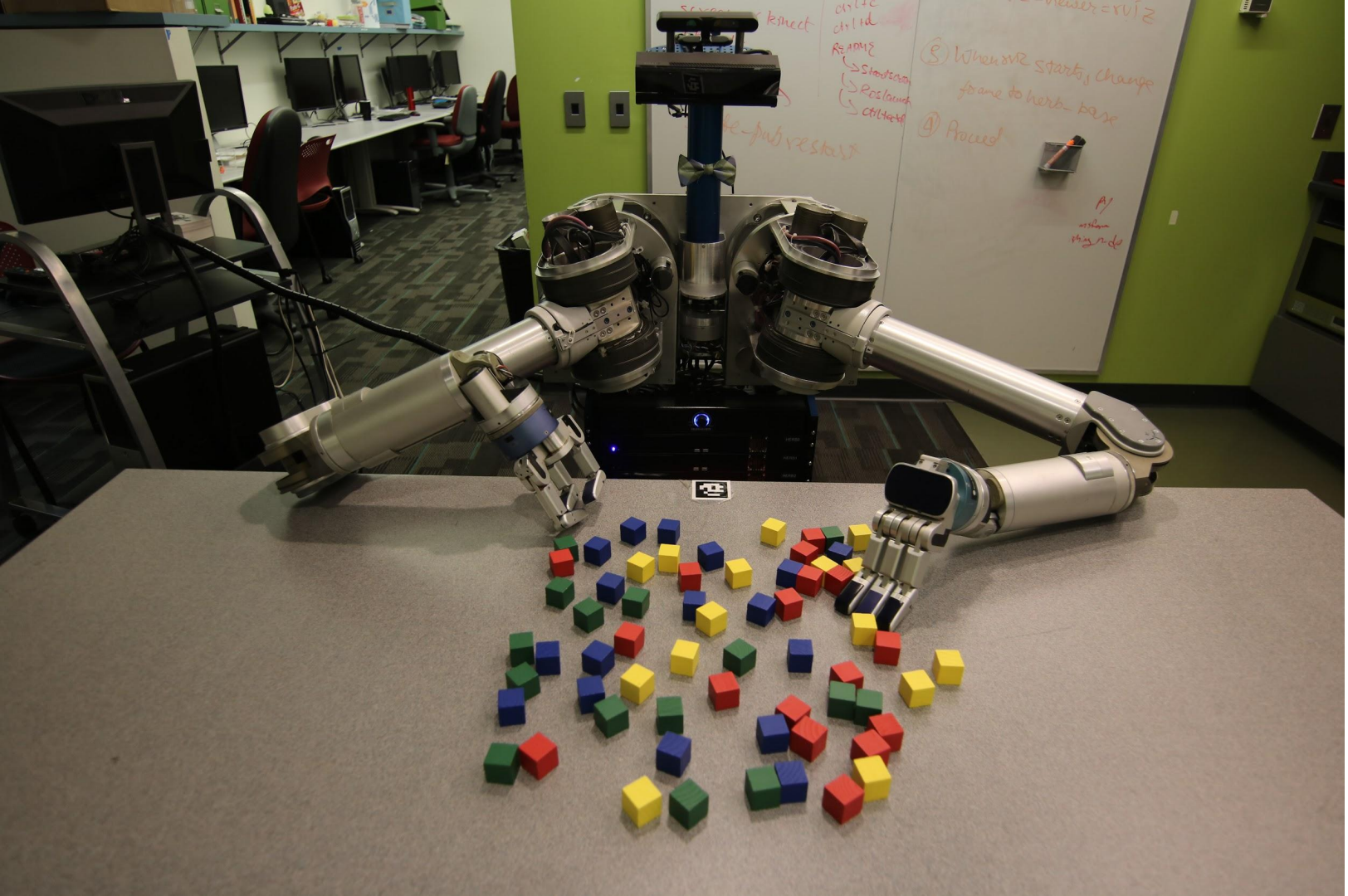
“Pick up the coffee cup next to the donuts.”

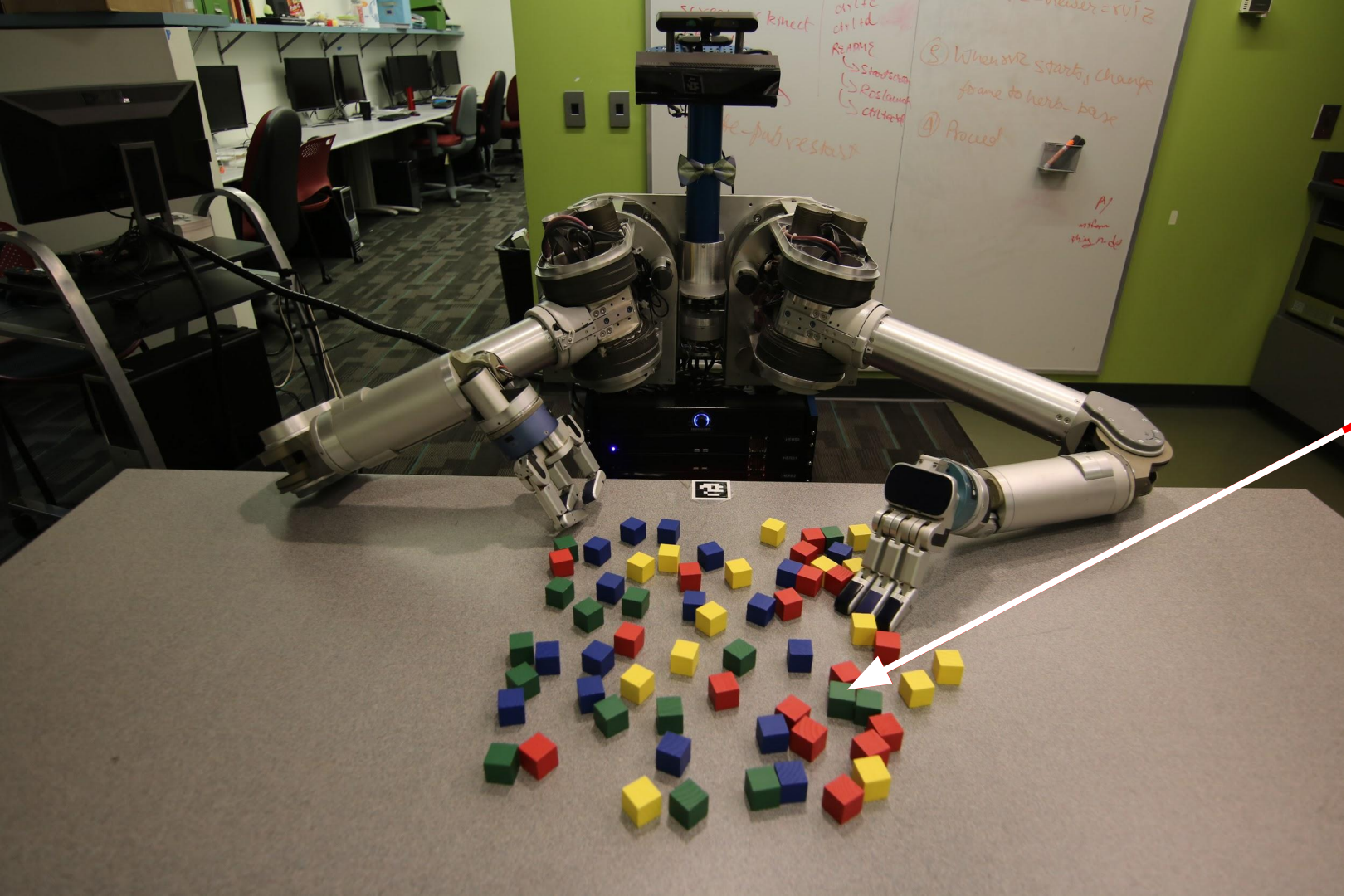
Forms of Ambiguity

Proximity

“Pick up the coffee cup next to the donuts.”

How close is ‘next to’?



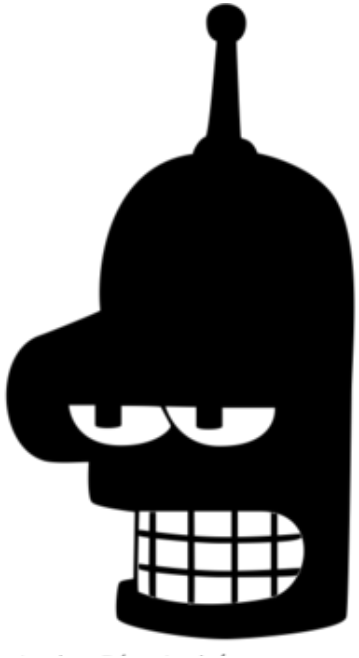


Can you uniquely describe this block?

How can we best *overcome ambiguity* when grounding our references **while** keeping communication natural?

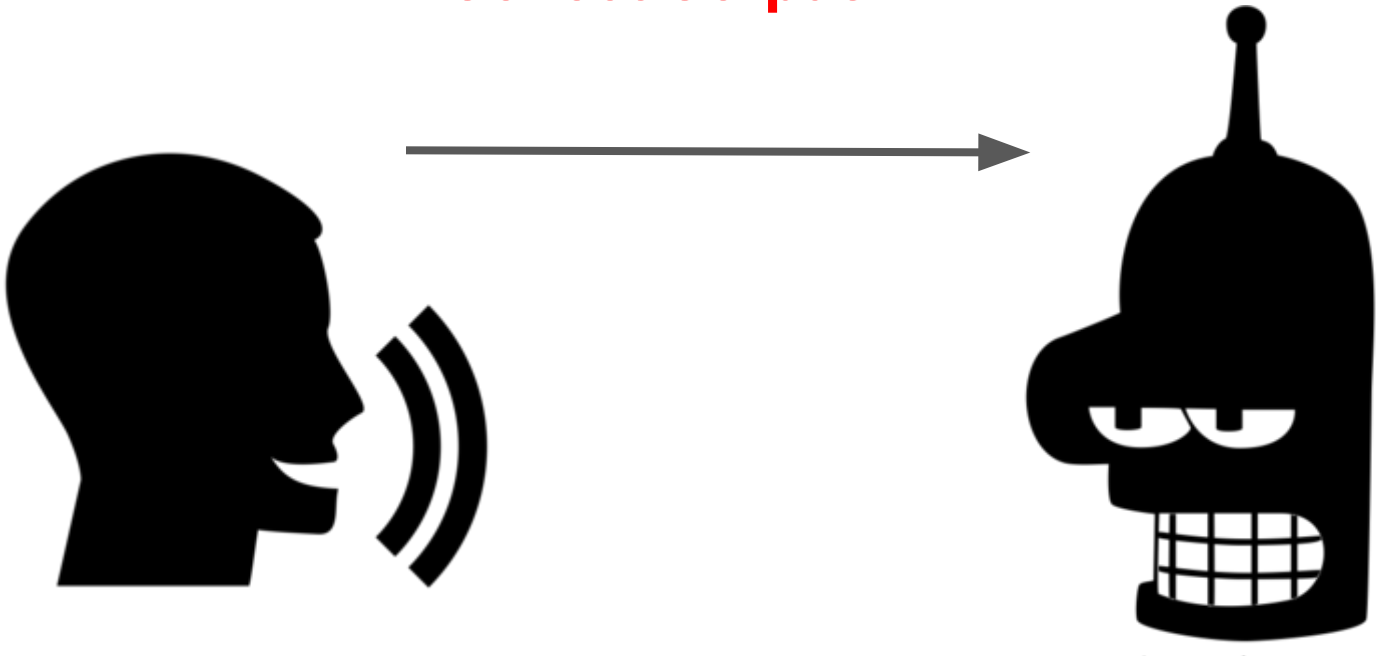
Approach

Learn by **observing** what humans do and **extract best-practices** from the examples that are most successful.



bender by Jordan Díaz Andrés from the Noun Project

Collect Corpus

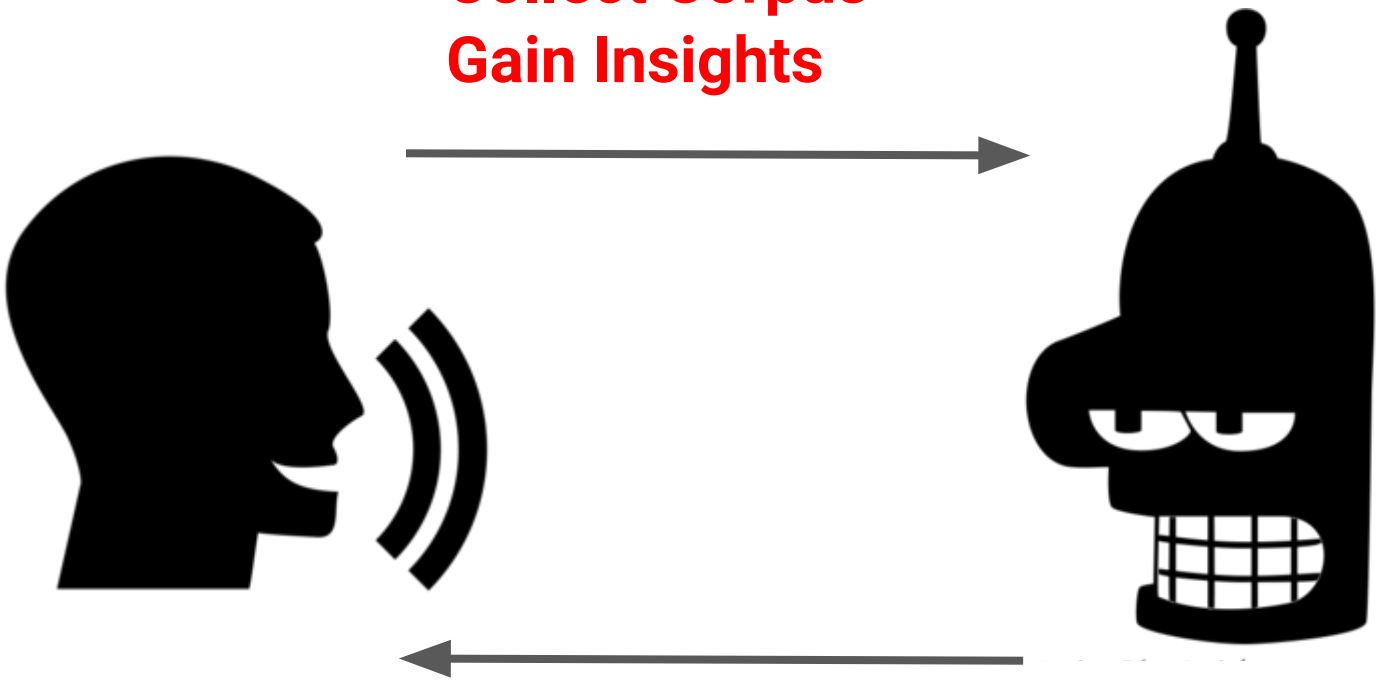


bender by Jordan Díaz Andrés from the Noun Project

Collect Corpus
Gain Insights

bender by Jordan Díaz Andrés from the Noun Project

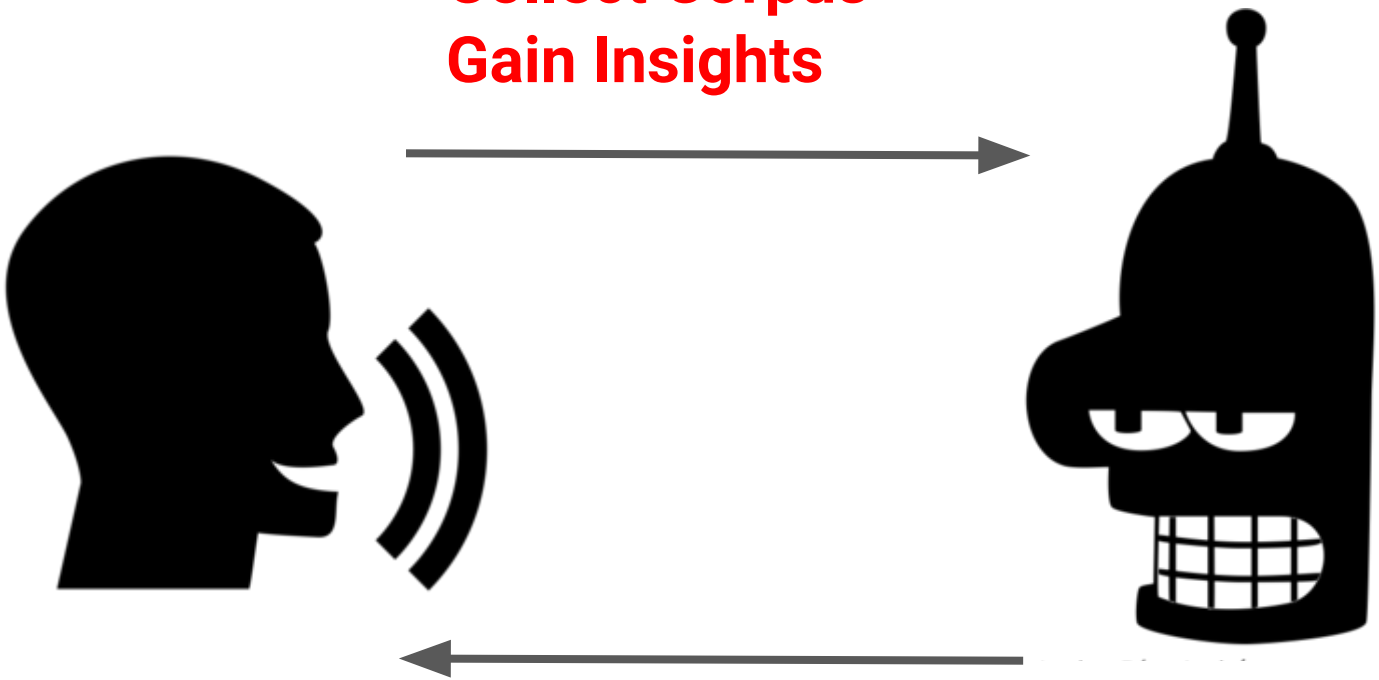
Collect Corpus
Gain Insights



Evaluate Corpus

bender by Jordan Díaz Andrés from the Noun Project

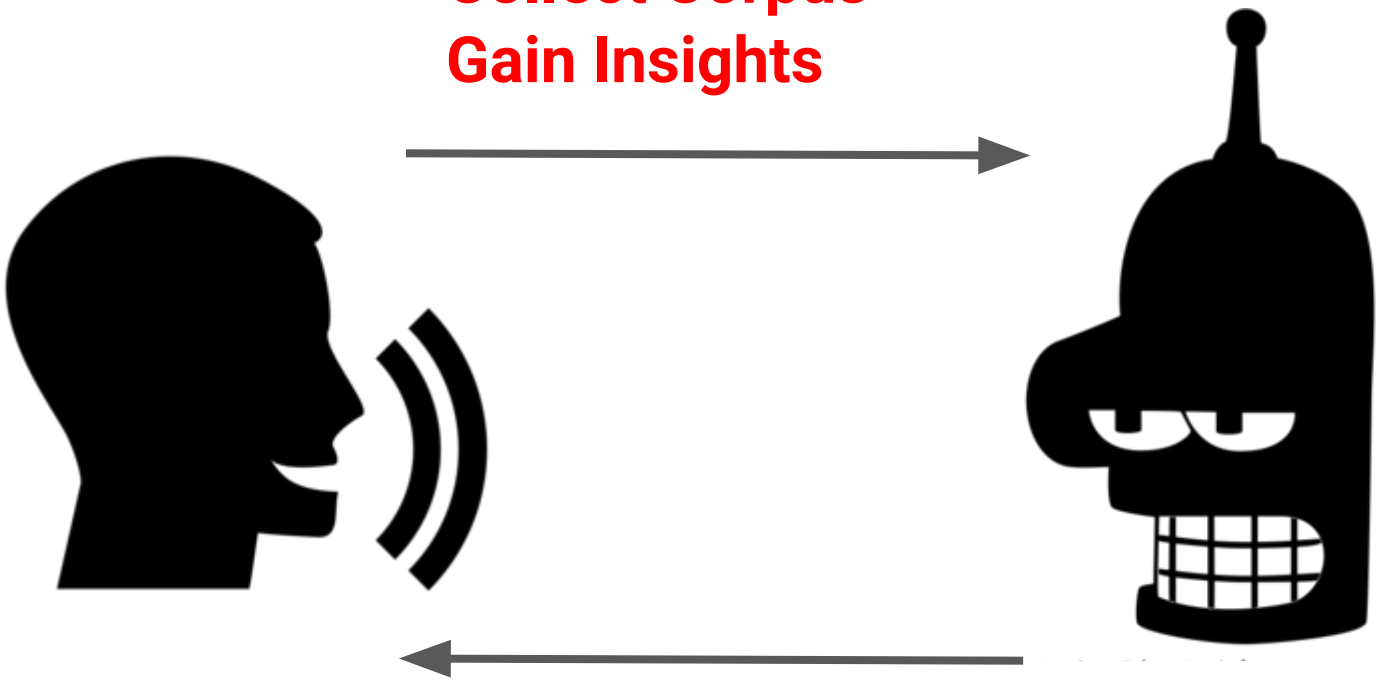
Collect Corpus
Gain Insights



Evaluate Corpus
Extract Guidelines

bender by Jordan Díaz Andrés from the Noun Project

Collect Corpus
Gain Insights

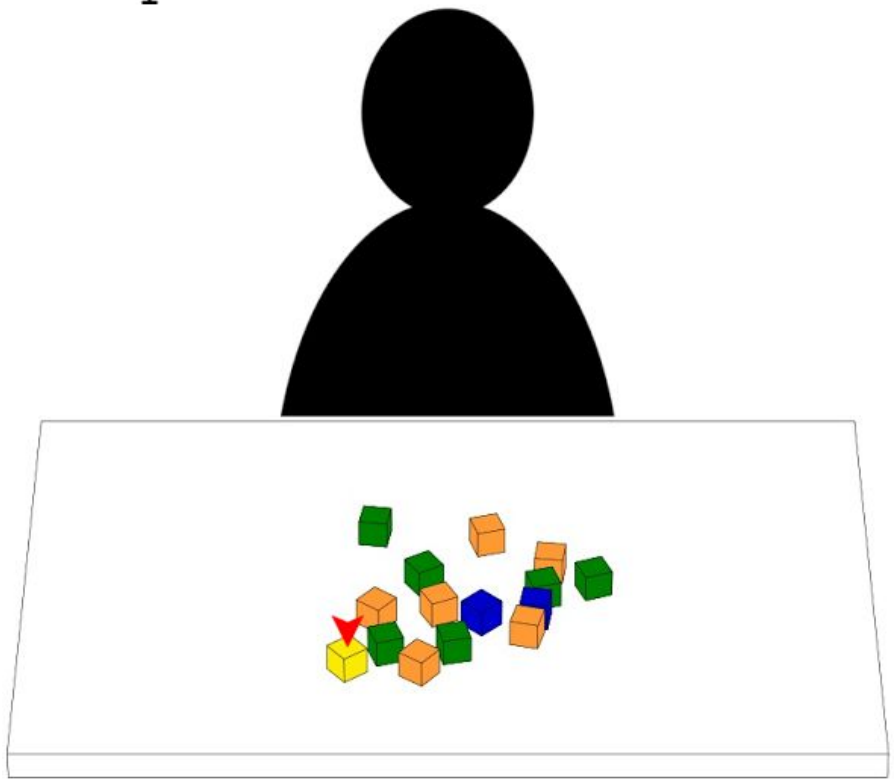


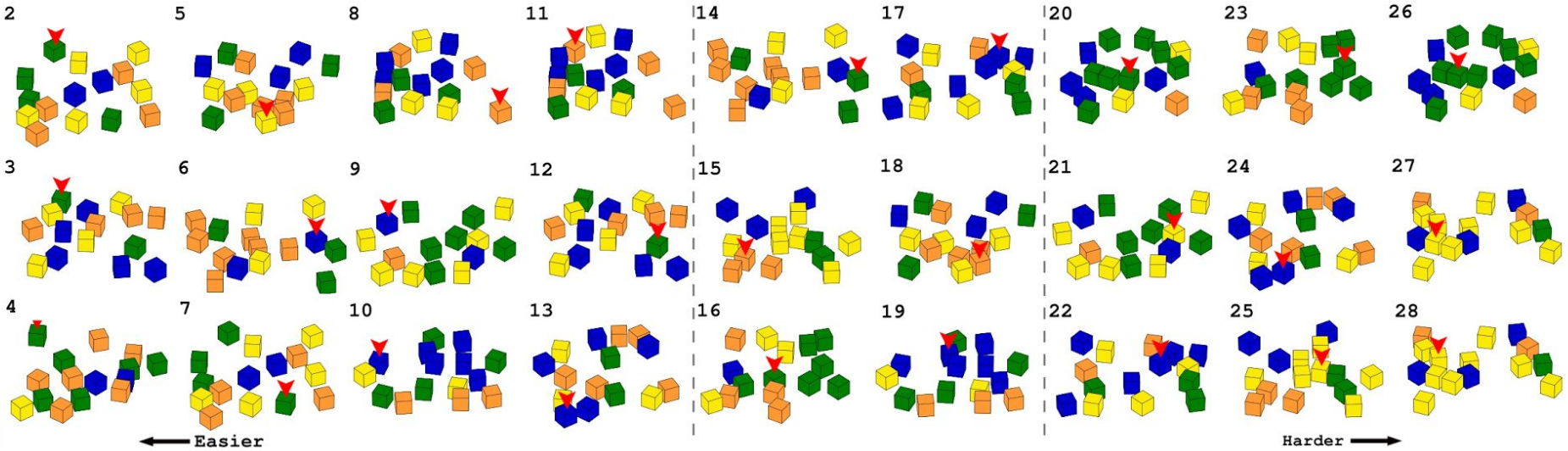
Evaluate Corpus
Extract Guidelines

+ Analysis Tools

bender by Jordan Díaz Andrés from the Noun Project

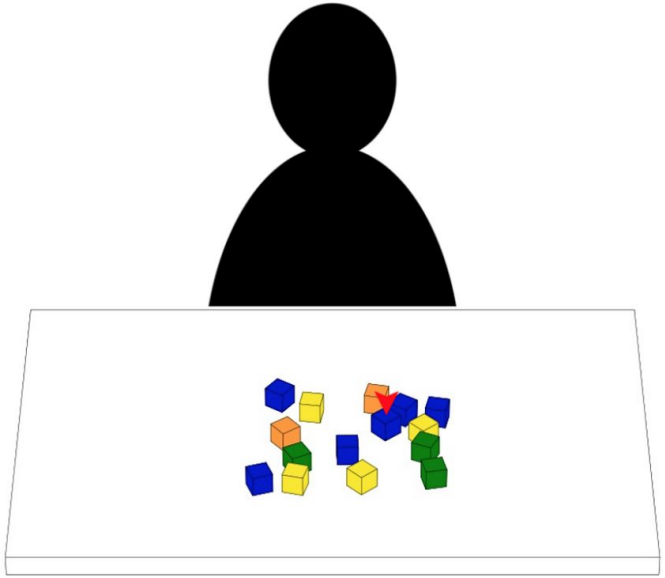
1




2
5
8
11
14
17
20
23
26
3
6
9
12
15
18
21
24
27
4
7
10
13
16
19
22
25
28
Easier
Harder

Study 1 : Collecting Instructions for Corpus

Scenario #1/14



Next

You are facing the table just as it appears in the image, and on the other side of the table is a person represented by the silhouetted figure.
How would you instruct the person to pick up the indicated block? (They cannot see the red arrow):

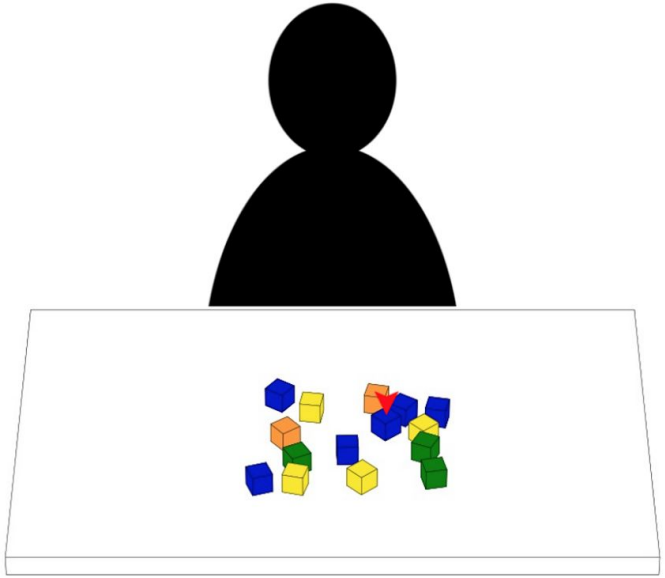
How would you instruct the person represented by the silhouetted figure to pick up the indicated block?.

How difficult did you find it to answer this prompt?

- Very Difficult
- Difficult
- Neither difficult nor easy
- Easy
- Very Easy

Study 1 : Collecting Instructions for Corpus

Scenario #1/14



Next

You are facing the table just as it appears in the image, and on the other side of the table is a**person** represented by the silhouetted figure.
How would you instruct the person to pick up the indicated block? (They cannot see the red arrow):

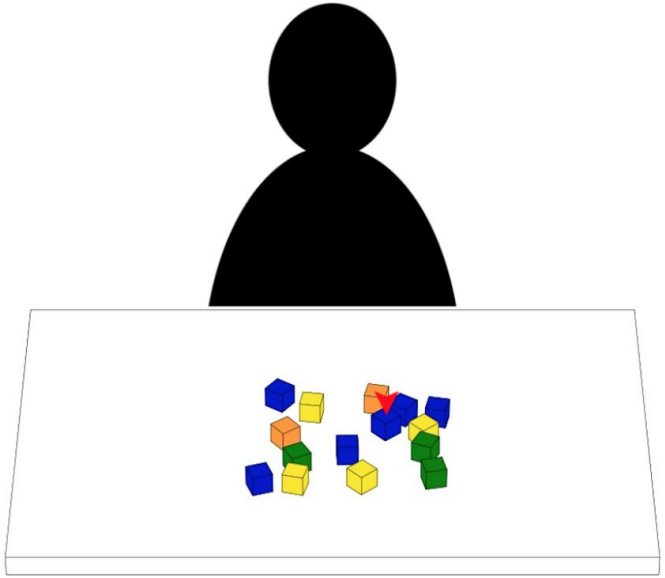
How would you instruct the person represented by the silhouetted figure to pick up the indicated block?.

How difficult did you find it to answer this prompt?

- Very Difficult
- Difficult
- Neither difficult nor easy
- Easy
- Very Easy

Study 1 : Collecting Instructions for Corpus

Scenario #1/14



Next

You are facing the table just as it appears in the image, and on the other side of the table is a **robot** represented by the silhouetted figure.
How would you instruct the person to pick up the indicated block? (They cannot see the red arrow):

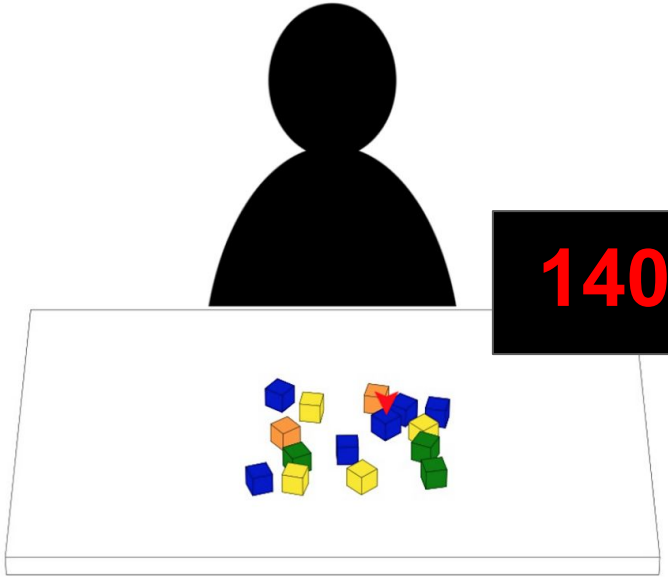
How would you instruct the person represented by the silhouetted figure to pick up the indicated block?.

How difficult did you find it to answer this prompt?

- ☐ Very Difficult
- ☐ Difficult
- ☐ Neither difficult nor easy
- ☐ Easy
- ☐ Very Easy

Study 1 : Collecting Instructions for Corpus

Scenario #1/14



1400 Total

You are facing the table just as it appears in the image, and on the other side of the table is a **robot** represented by the silhouetted figure.
How would you instruct the person to pick up the indicated block? (They cannot see the red arrow):

How would you instruct the person represented by the silhouetted figure to pick up the indicated block?.

How difficult did you find it to answer this prompt?

- Very Difficult
- Difficult
- Neither difficult nor easy
- Easy
- Very Easy

Next

Evaluating

How do we tell how good any specific instruction is?

“Pick up the blue block”

Evaluating

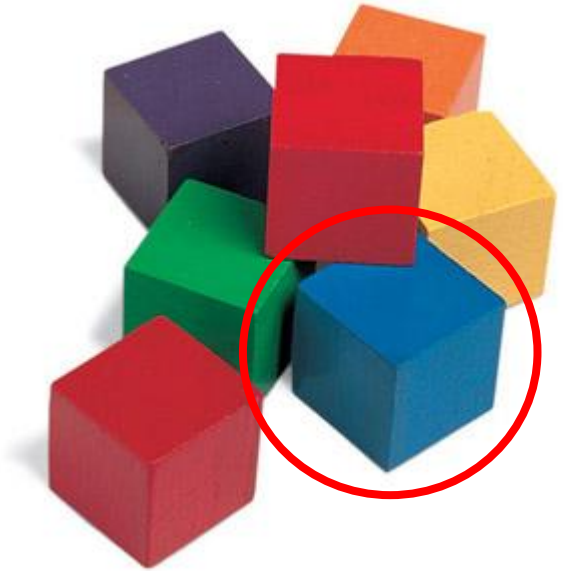
Given an instruction and the stimulus it corresponds to, can people infer the correct block?

“Pick up the blue block”

Evaluating

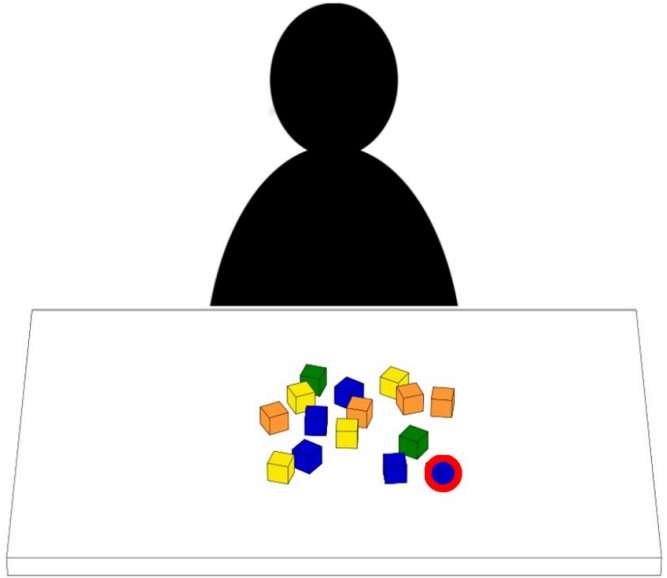
Given an instruction and the stimulus it corresponds to, can people infer the correct block?

“Pick up the blue block”



Study 2 : Corpus Evaluation

Scenario #3/40



Next

You are seated across the table from the silhouetted figure. You have just asked the silhouetted figure to pick up a block. The instructions you gave are shown in the blue text below:

Pick up the green block that is closest to you and on the right.

Which block do you expect the silhouetted figure to pick up? The red circle shows the block, click to show a black and white circle with your selection.
Once you are satisfied with your selection, please click the next button to move on.

Metrics

For each instruction, we calculate:

Metrics

For each instruction, we calculate:

Accuracy: $\dfrac{\text{\# of successful block selections}}{\text{total \# of times instruction is shown}}$

Metrics

For each instruction, we calculate:

Accuracy:

$$\frac{\text{\# of successful block selections}}{\text{total \# of times instruction is shown}}$$

Avg. Completion time: How long it takes to select the indicated block on average

Full investigation and results TBR in:

“Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation”

at IEEE Ro-Man 2016 (Late August)

Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation

Shen Li*, Rosario Scalise*, Henny Admoni, Stephanie Rosenthal, Siddhartha S. Srinivasa

***Abstract*—As humans and robots collaborate together on spatial tasks, they must communicate clearly about the objects they are referencing. Communication is clearer when language is unambiguous which implies the use of spatial references and explicit perspectives. In this work, we contribute two studies to understand how people instruct a partner to identify and pick up objects on a table. We investigate spatial features and perspectives in human spatial references and compare word usage when instructing robots vs. instructing other humans. We then focus our analysis on the clarity of instructions with respect to perspective taking and spatial references. We find that only about 42% of instructions contain perspective-independent spatial references. There is a strong correlation between participants' accuracy in executing instructions and the perspectives that the instructions are given in, as well between accuracy and the number of spatial relations that were required for the instruction. We conclude that sentence complexity (in terms of spatial relations and perspective taking) impacts understanding, and we provide suggestions for automatic generation of spatial references.**

I. Introduction

As people and robots collaborate more frequently on spatial tasks such as furniture assembly [1], warehouse automation [2], or meal serving [3], they need to communicate clearly about objects in their environment. In order to do this, people use a combination of visual features and spatial references. In the sentence “The red cup on the right”, ‘red’ is a visual feature and ‘right’ is a spatial reference.

There is a long line of research in robotics related to communicating about spatial references like ‘furthest to the right’,‘near the back’, and ‘closest’ for navigation task [4]–[10]. However, there are fewer studies involving the communication of spatial references for tabletop or assembly tasks [11]. A common theme in the space of tabletop manipulation tasks is clutter which we view as many potential objects to reason about. See Fig. 1

*Both authors contributed equally.
Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213 {shenli,robo,henny,siddh}@cmu.edu, srosenthal@cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. [Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. Carnegie Mellon is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University. DM-0003432.

This work was (partially) funded by the DARPA SIMPLEX program through ARO contract number 67980-LSDRP, National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), the Office of Naval Research, and the Richard K. Mellon Foundation.

A cluttered table introduces the problem of *object uniqueness* where if there are two objects which are identified in the same manner (e.g. the red cup among two red cups), we are left with an ambiguity. One possible solution to this is to utilize *spatial references* which allow the use of spatial properties to establish a grounding or certainty about the semantic relationship between two entities.

However, even with the use of spatial references, it is still possible to encounter additional ambiguity which originates from the reference frame. Humans often use perspective to resolve this ambiguity as in the example ‘the red cup on your right’. Often times, in tabletop scenarios, the person giving instructions will be situated across the table from their partner and thus will have a different perspective. Therefore, robots that collaborate with humans in tabletop tasks have to both understand and generate *spatial language* and *perspective* when interacting with their human partners. We investigate these key components by collecting a corpus of natural language instructions and analyzing them with our goal of clear communication in mind.

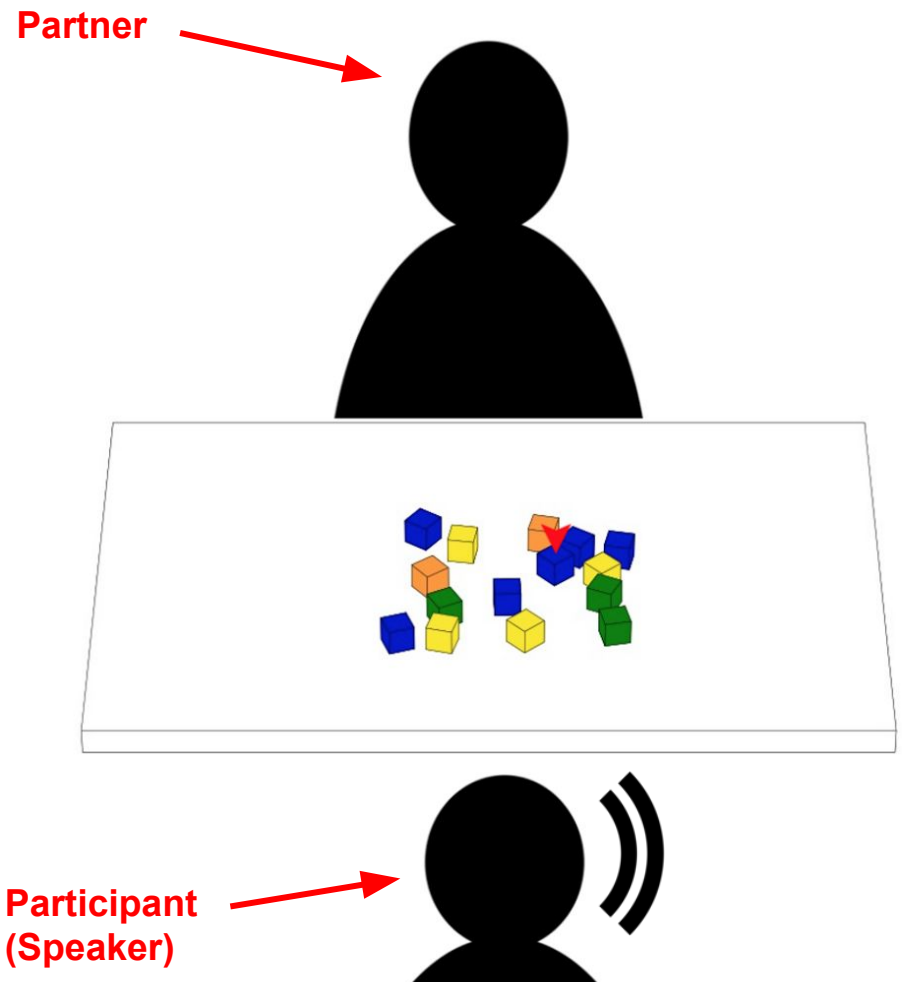
We first conducted a study in which we asked participants to write instructions to either a robot or human partner sitting across the table to pick up an indicated block from the table as shown in Fig. 1. This task raises a perspective problem: does the participant use the partner’s perspective or their own perspective, if any? Blocks were not always uniquely identifiable, and so the task required participants to describe spatial relationships between objects as well. We analyze the instructions from participants for 1) language differences between instructing a human versus a robot partner, 2) trends in language for visual and spatial references, and 3) the perspective(s) participants use when instructing their partners.

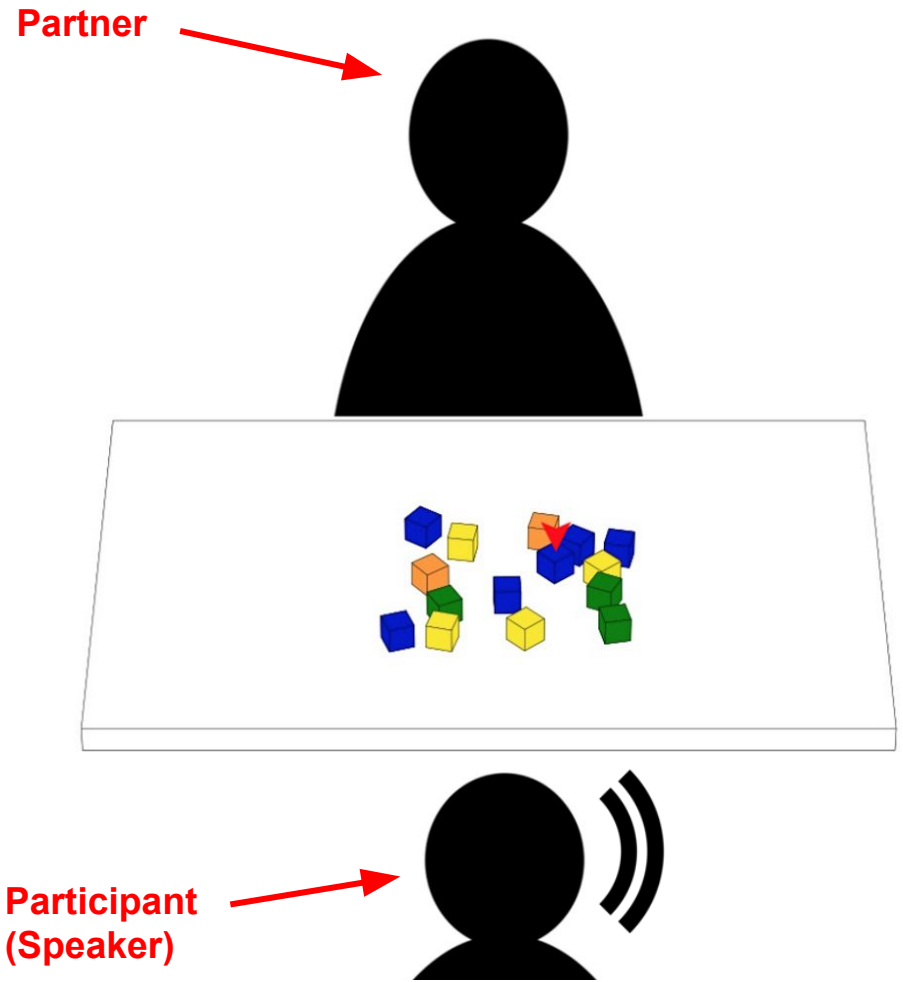
To investigate the effect of perspective, we conducted a second study in which we presented new participants with the instructions from the first study and asked them to select the indicated block. We utilized the correct selection of the indicated block as an objective measure of clarity. In order to establish which instructions contained ambiguities (lack of clarity), we first manually coded the instructions for whether the reference perspective was unknown or explicit (participant’s, partner’s, or neither) and whether there were multiple blocks that could be selected based on the instruction. An unknown perspective implies the instruction is dependent on perspective, but it is not explicitly stated.

Results from the first study show that participants explicitly take the partner’s perspective more frequently when they

Perspectives

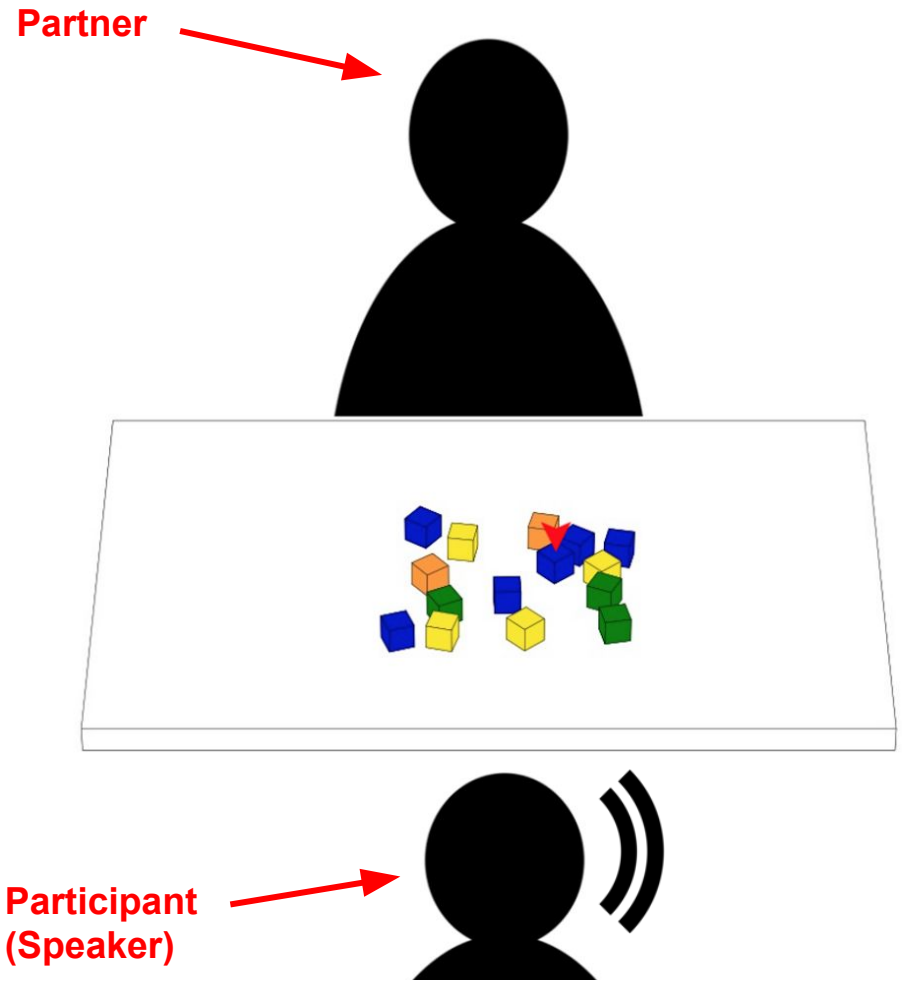
Types of Perspective:





Types of Perspective:

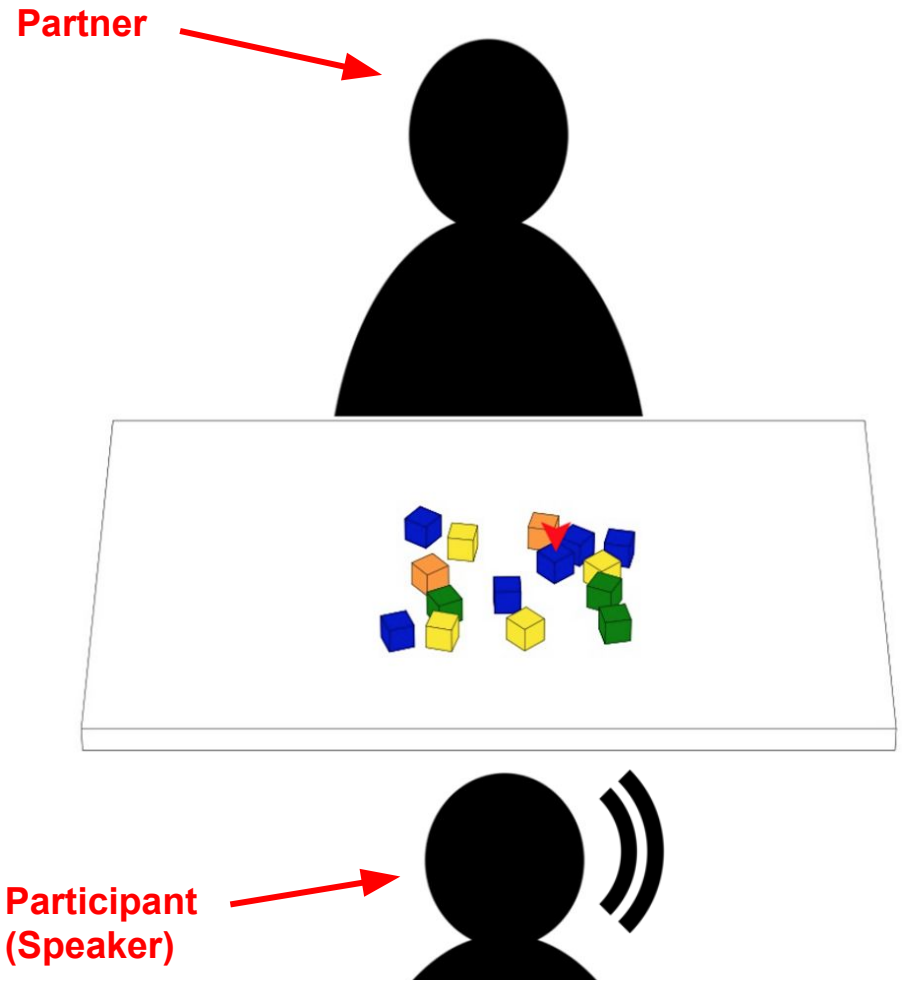
Partner:
"Pick up the blue block on **your** left"



Types of Perspective:

Partner:
"Pick up the blue block on **your** left"

Participant:
"Pick up the blue block on **my** right"

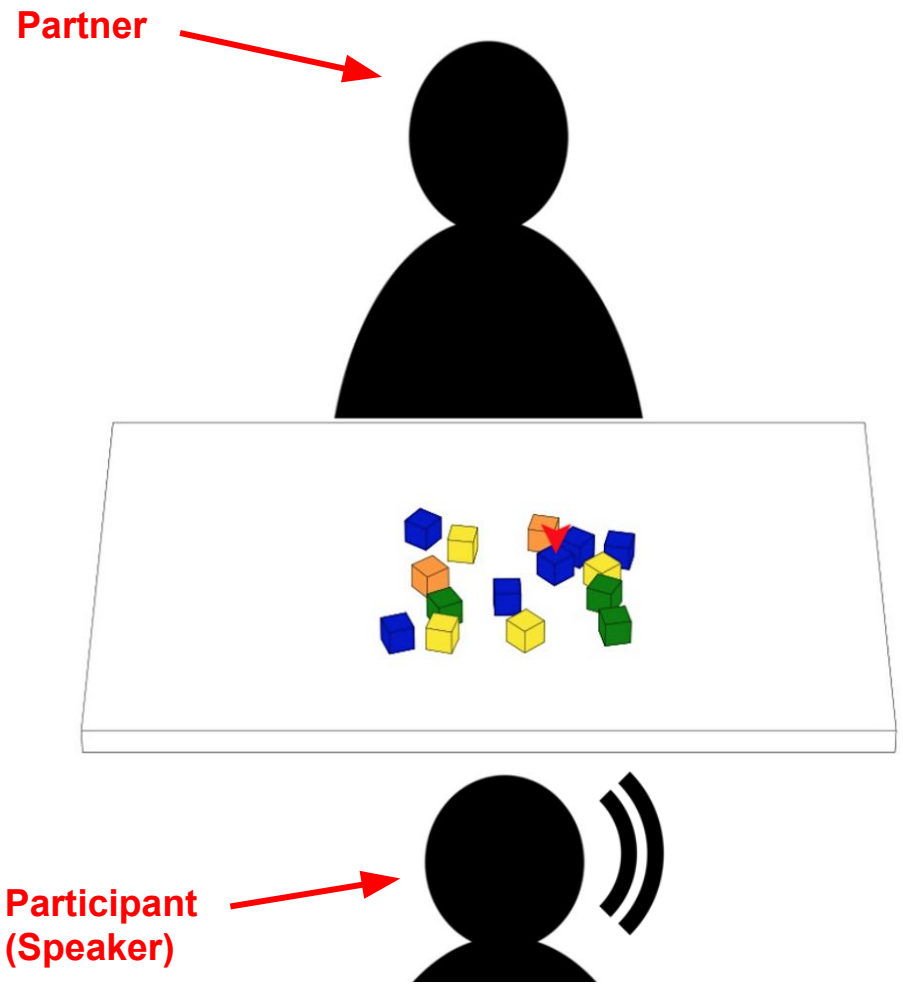


Types of Perspective:

Partner:
"Pick up the blue block on **your** left"

Participant:
"Pick up the blue block on **my** right"

Neither:
"Pick up the blue block **closest** to the orange block."



Types of Perspective:

Partner:
"Pick up the blue block on **your** left"

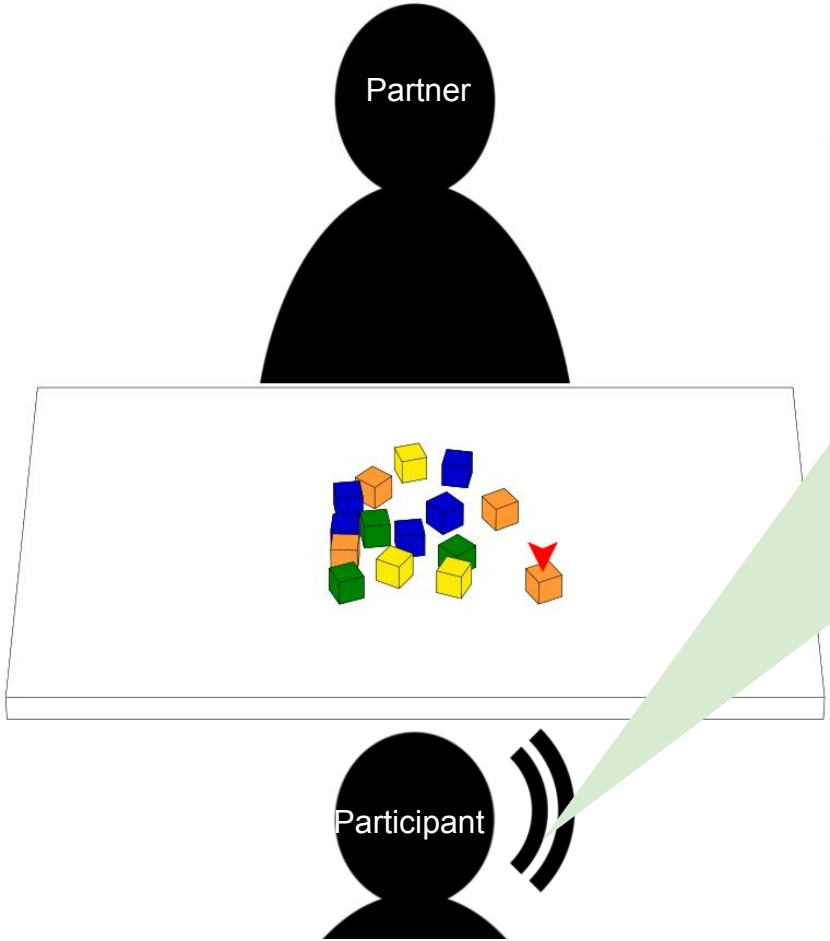
Participant:
"Pick up the blue block on **my** right"

Neither:
"Pick up the blue block **closest** to the orange block."

Unknown:
"Pick up the blue block **to the right** of the orange block."

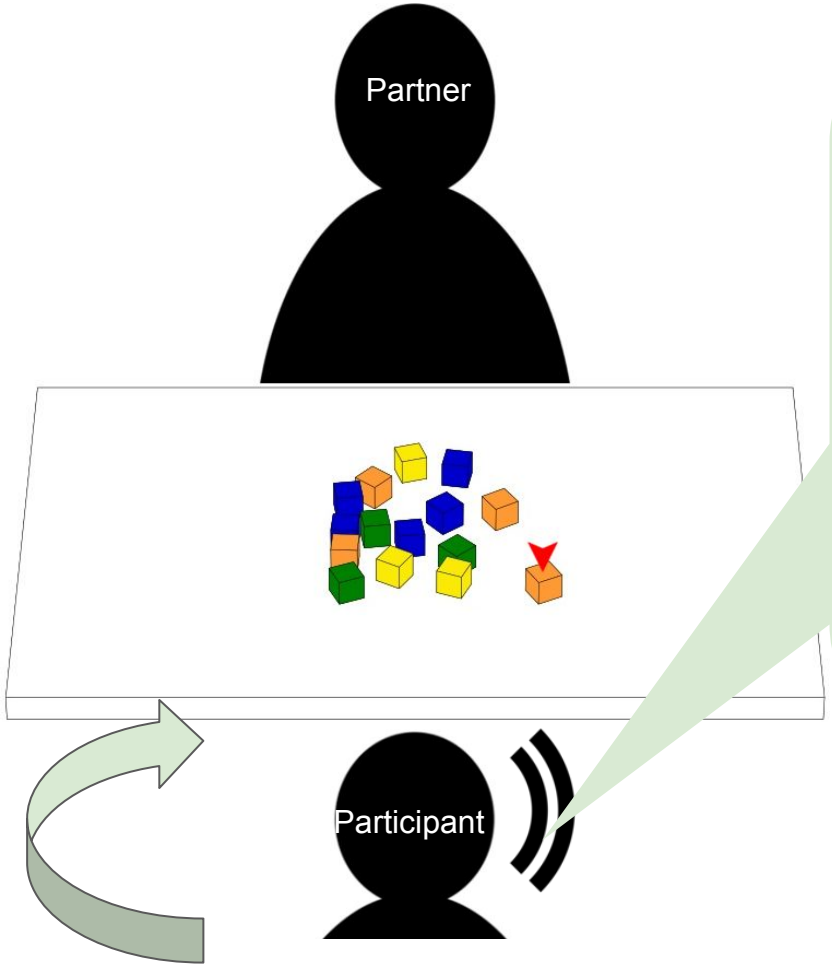
Perspective vs

Accuracy and Completion Time

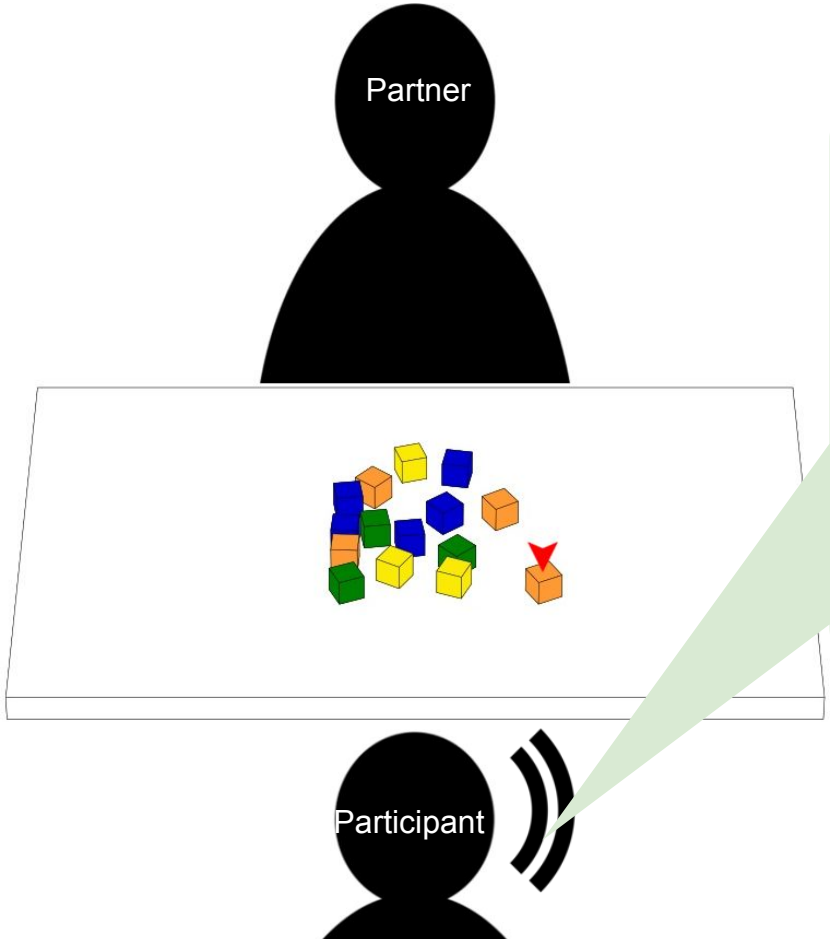


Pick up the box furthest to your left.

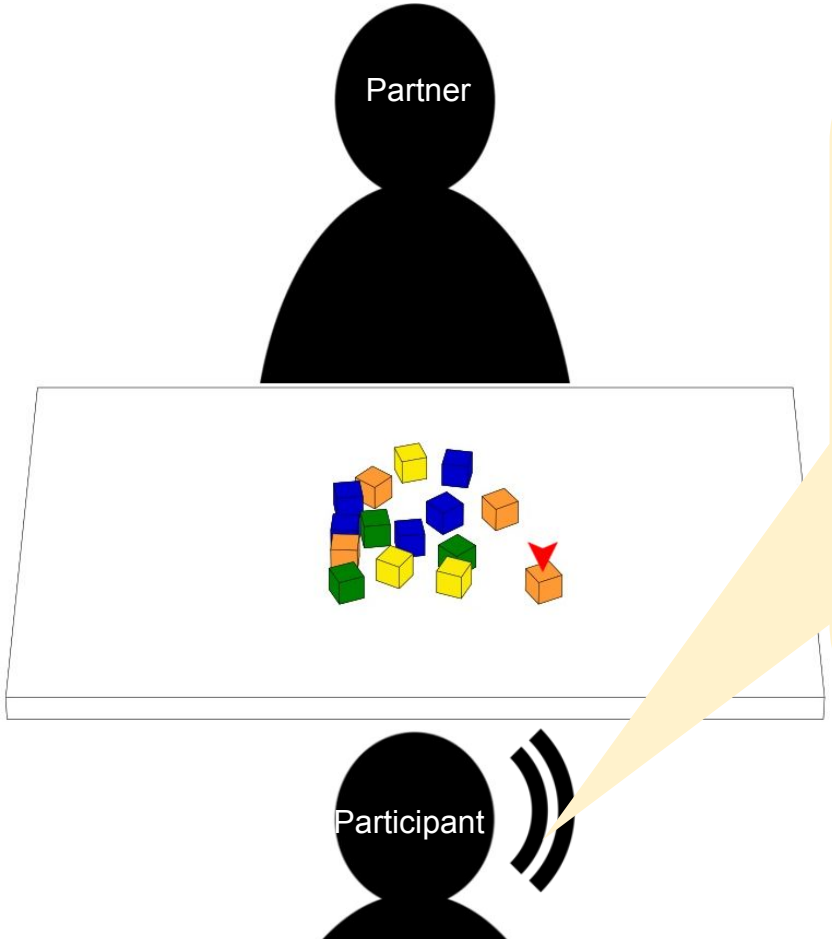
Partner perspective



Pick up the box furthest to **your** left.

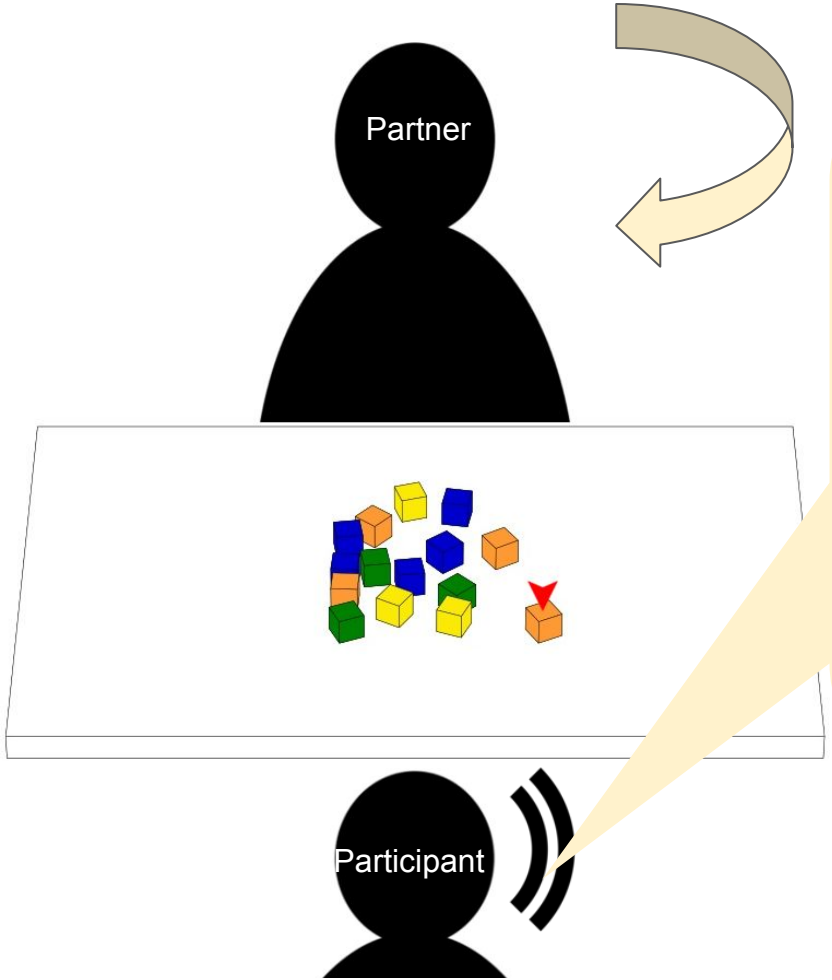


Pick up the box **furthest** to your left.

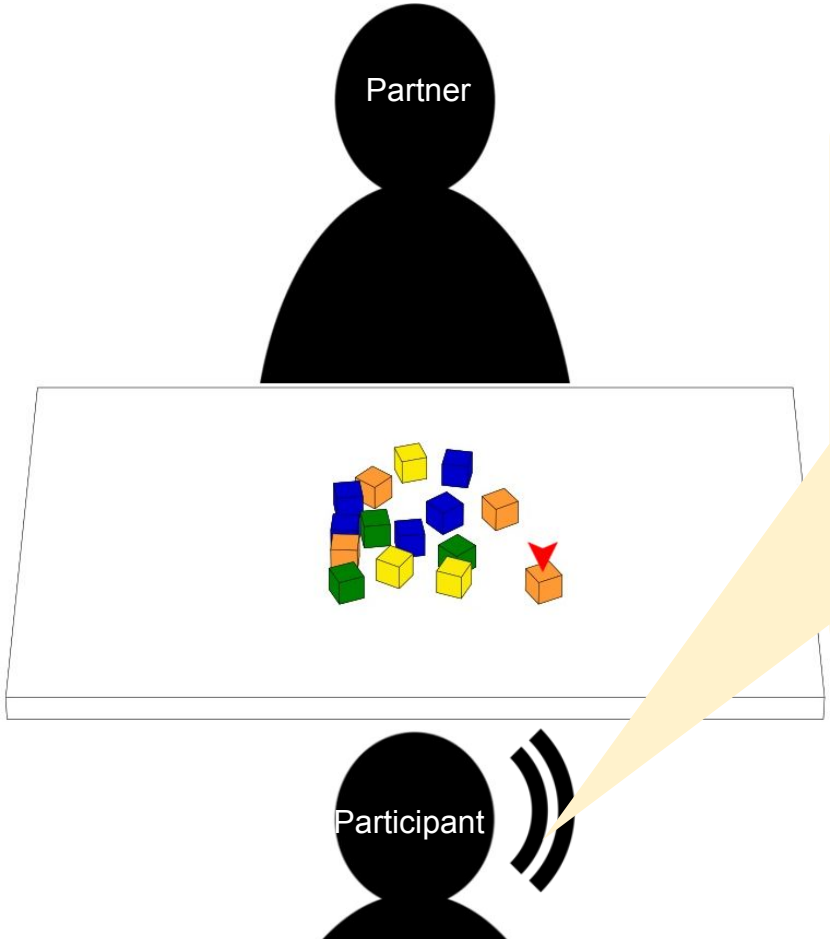


Pick up the orange block closest to my right hand side.

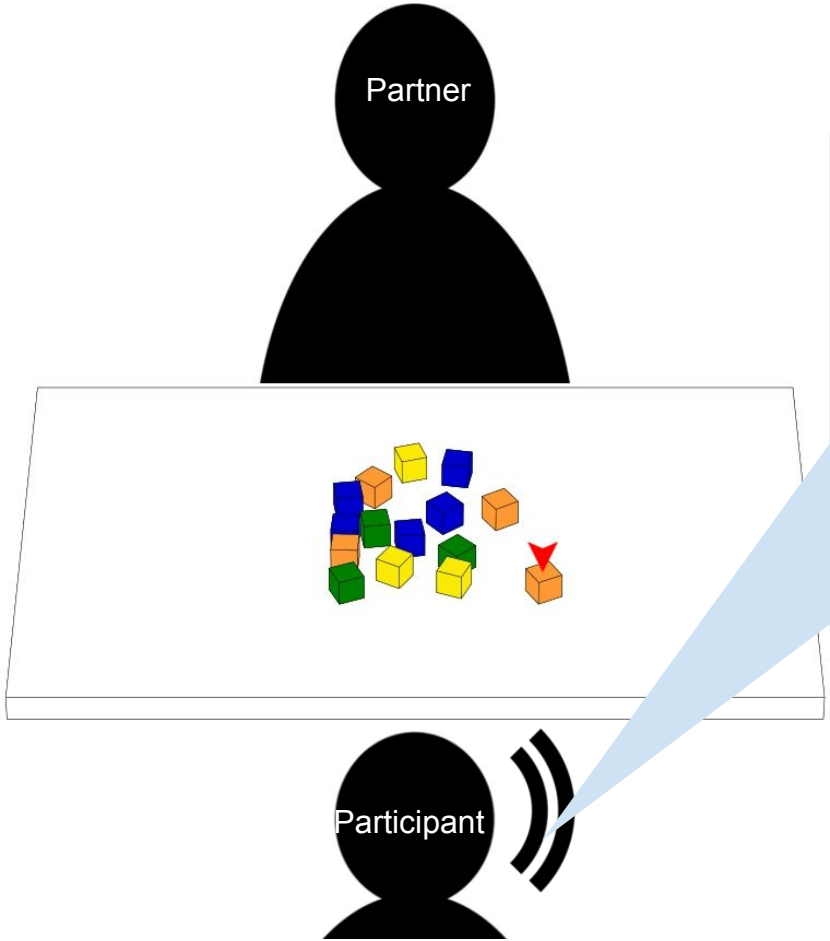
Participant perspective



Pick up the orange block closest to **my** right hand side.

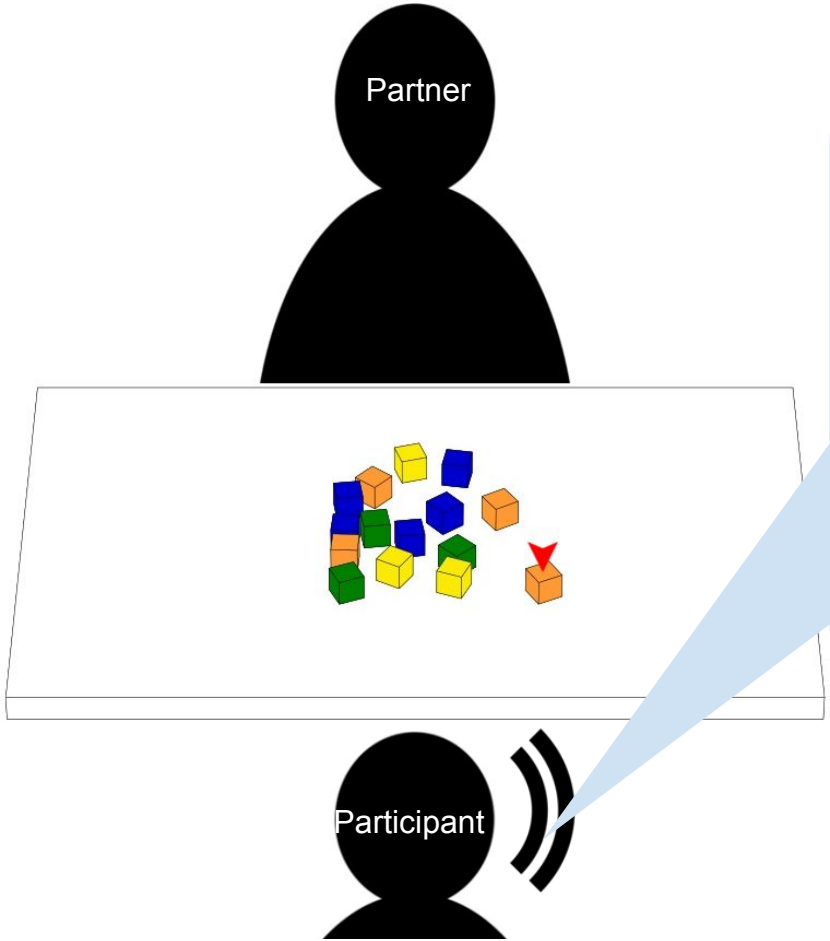


Pick up the orange block **closest** to my right hand side.

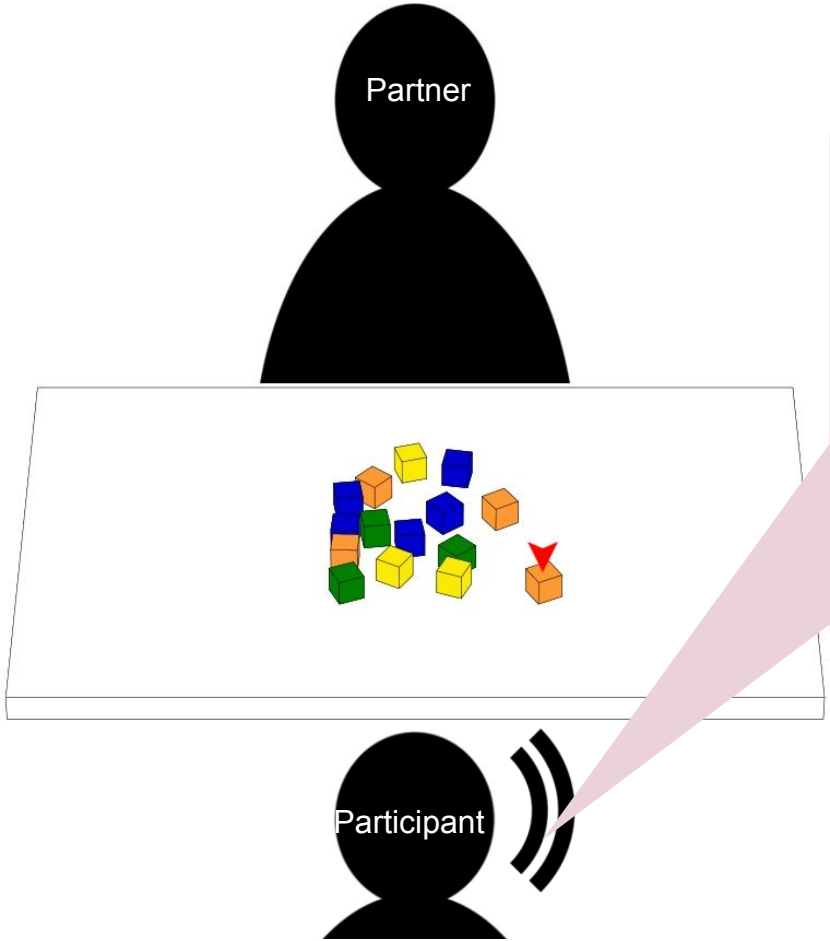


Please pick up the orange block that is closest to me.

Neither perspective

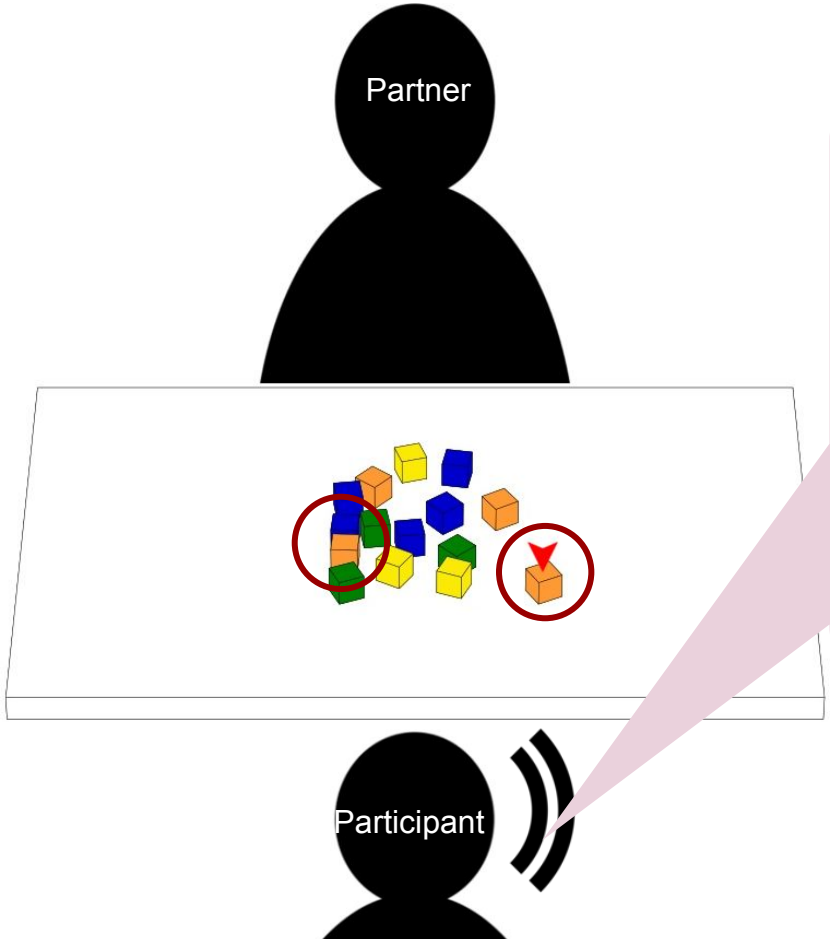


Please pick up the orange block that is **closest to me.**



Pick up the **rightmost** orange block

Right to ???

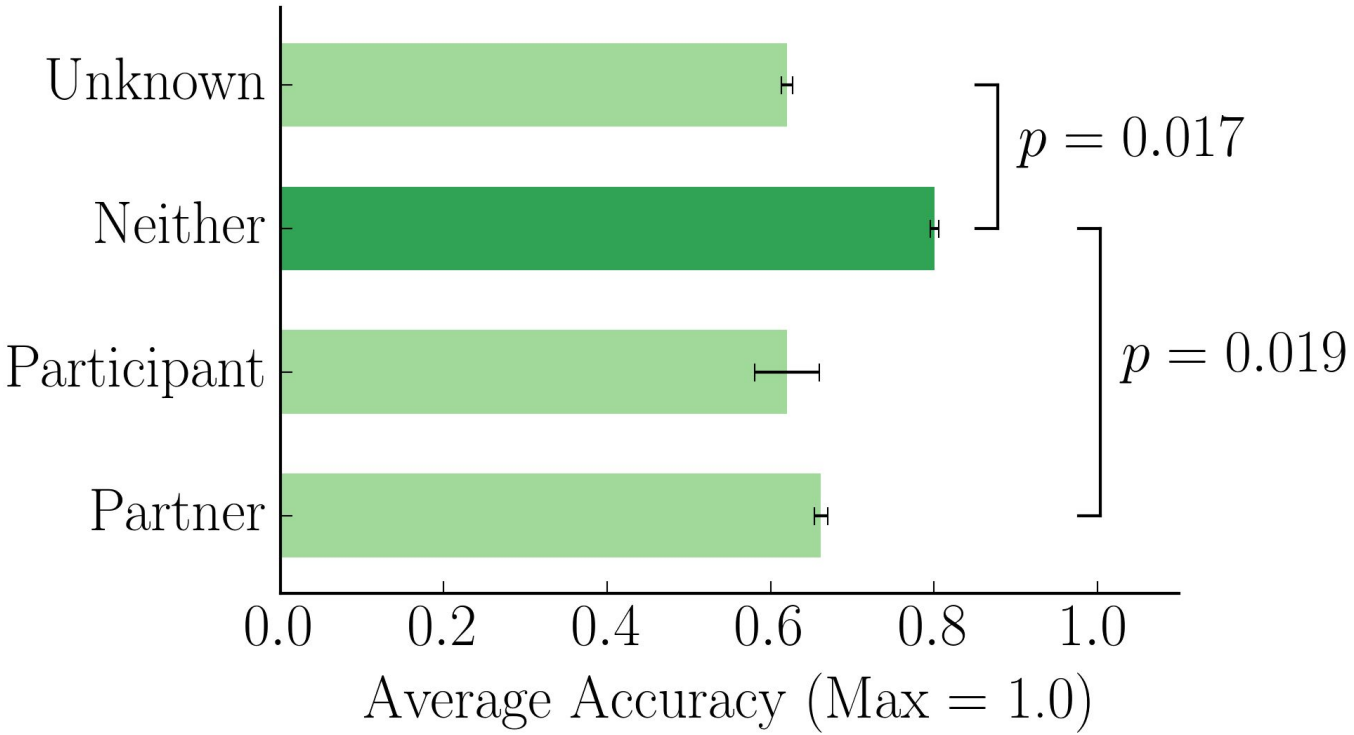


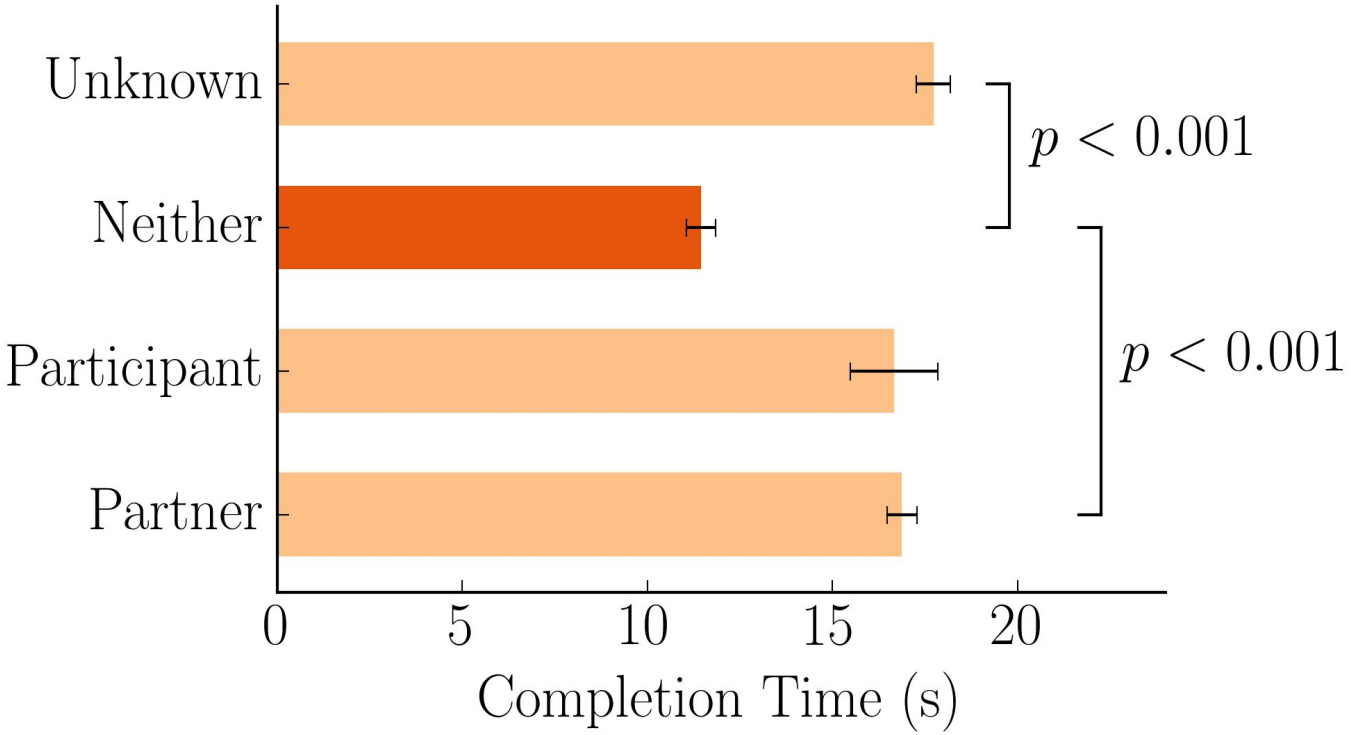
Pick up the **rightmost** orange block

Unknown perspective

Hypothesis:

Neither Perspective is better

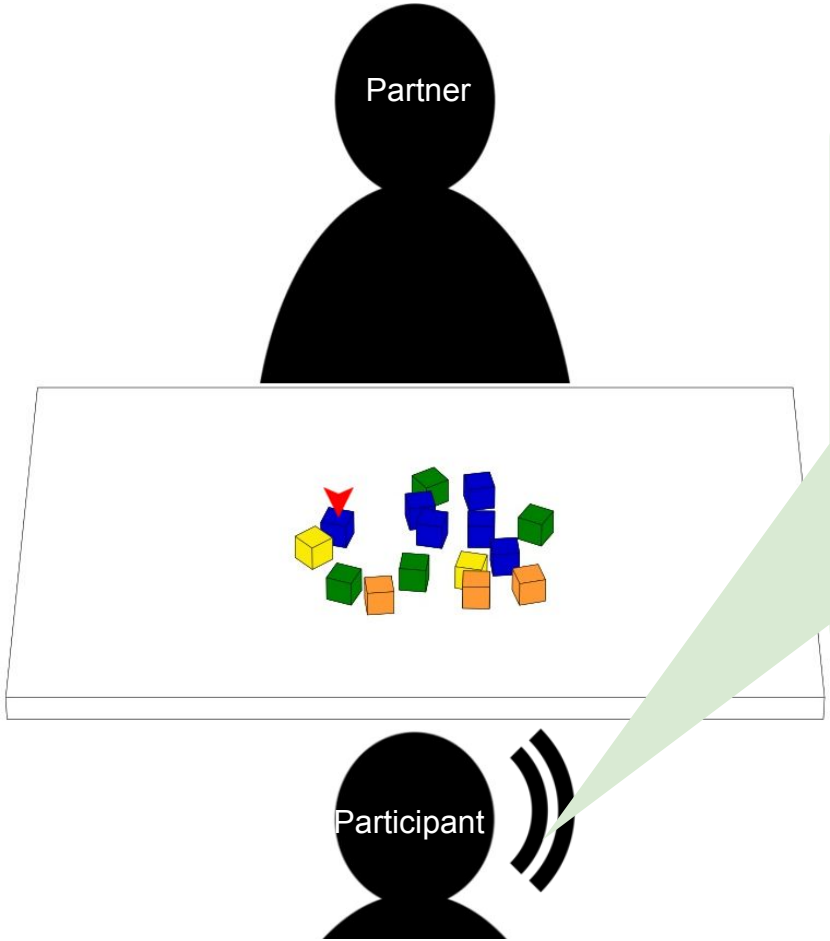

Unknown
Neither
Participant
Partner
$p = 0.017$
$p = 0.019$
0.0
0.2
0.4
0.6
0.8
1.0
Average Accuracy (Max = 1.0)


Unknown
Neither
Participant
Partner
$p < 0.001$
$p < 0.001$
0
5
10
15
20
Completion Time (s)

Result:

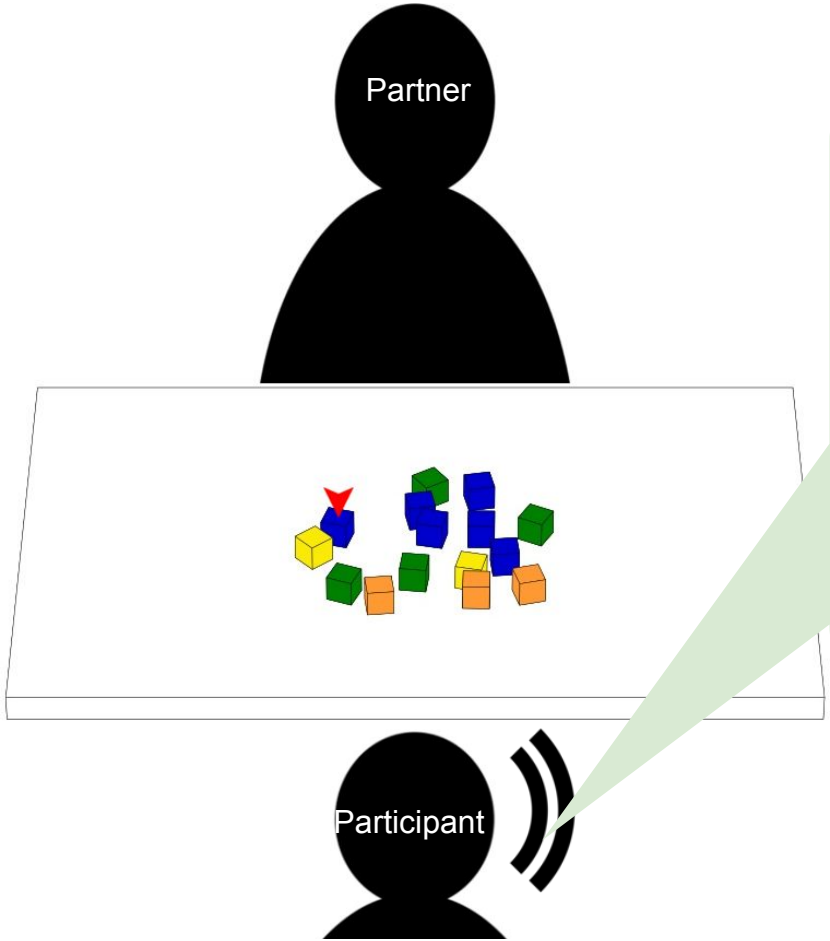
Prefer Neither Perspective

Other Factors



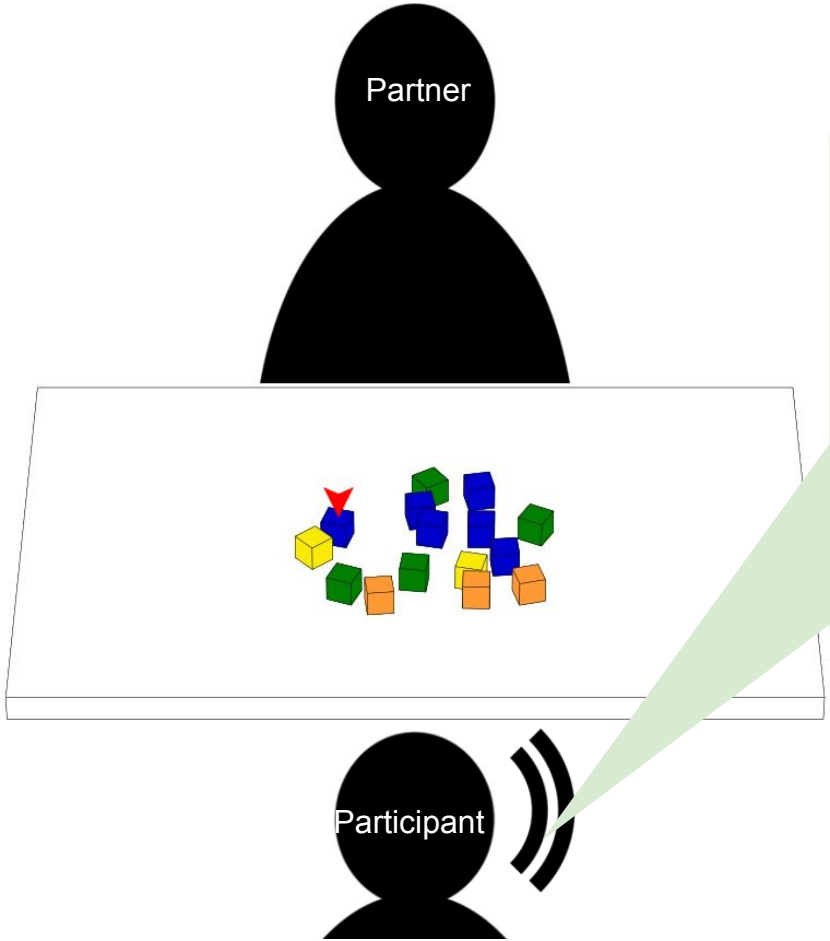
Pick the blue block that is closer to you and right next to the yellow block

Neither perspective



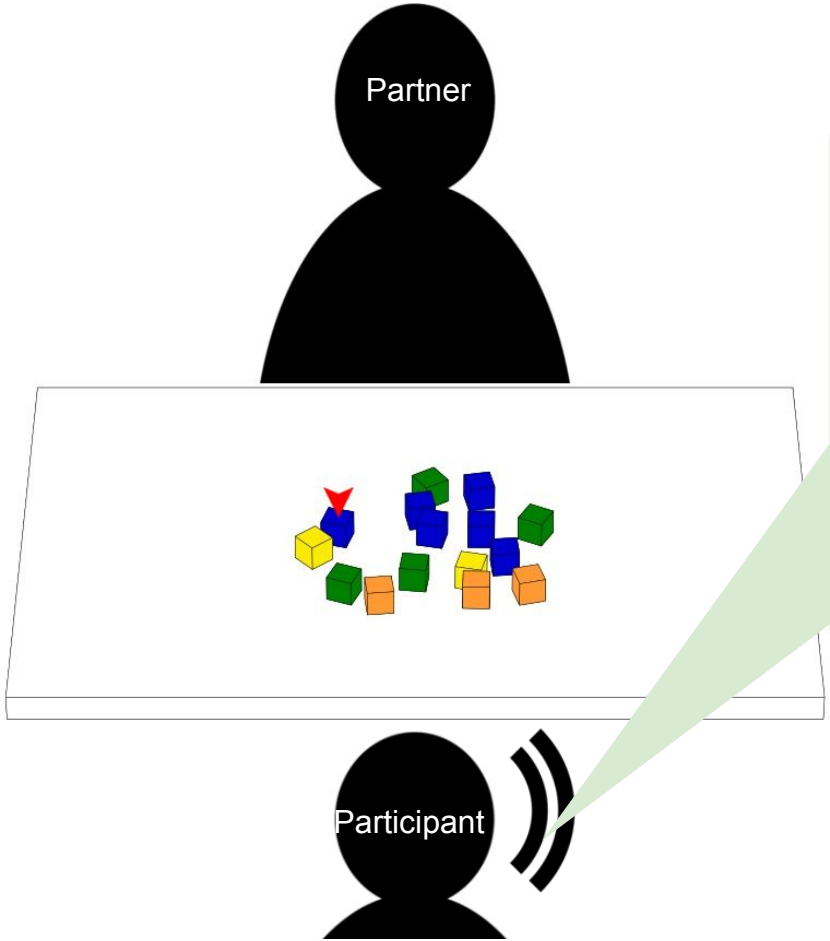
Pick the **blue block** that is closer to you and right **next to the yellow block**

Neither perspective



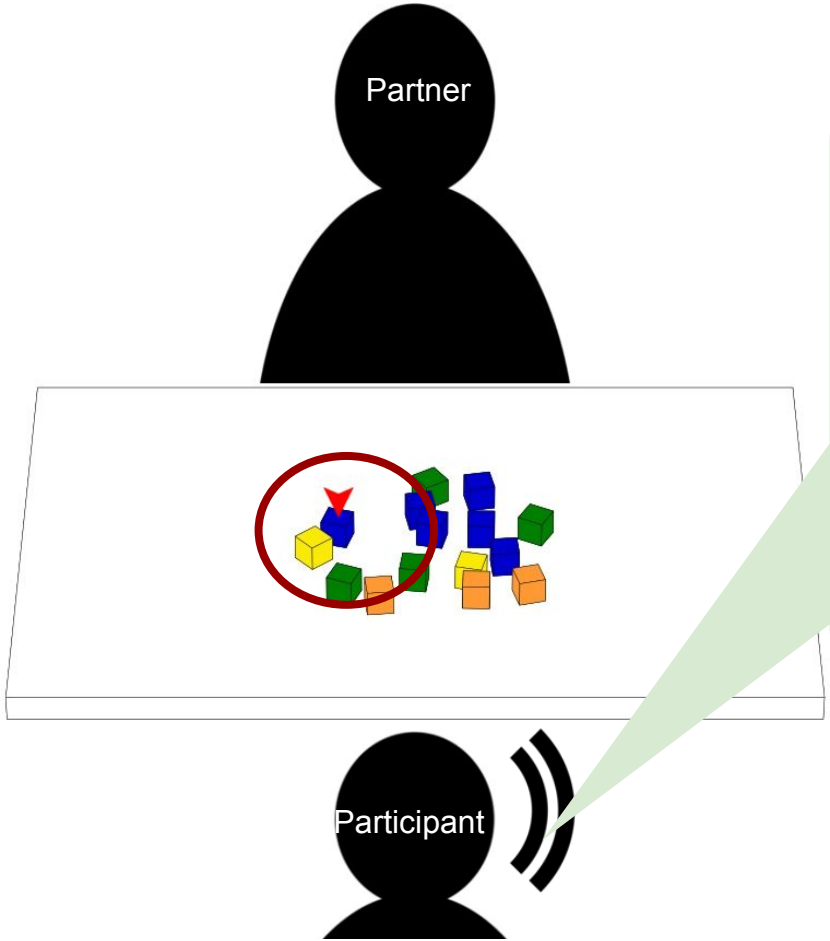
Pick the blue block that is **closer to you** and right next to the yellow block

Neither perspective



Pick up the blue block on your far right.

Partner perspective



Pick up the blue block on **your far right**.

Partner perspective

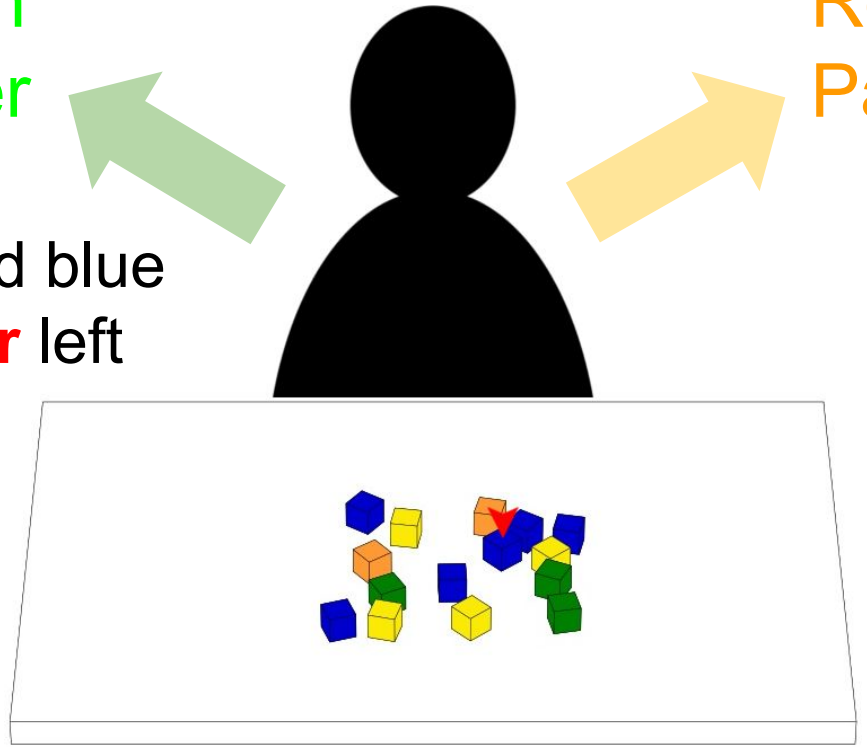
Tradeoff

Robot Partner vs Human Partner

Human Partner

Robot Partner

Pick up the third blue block from **your** left

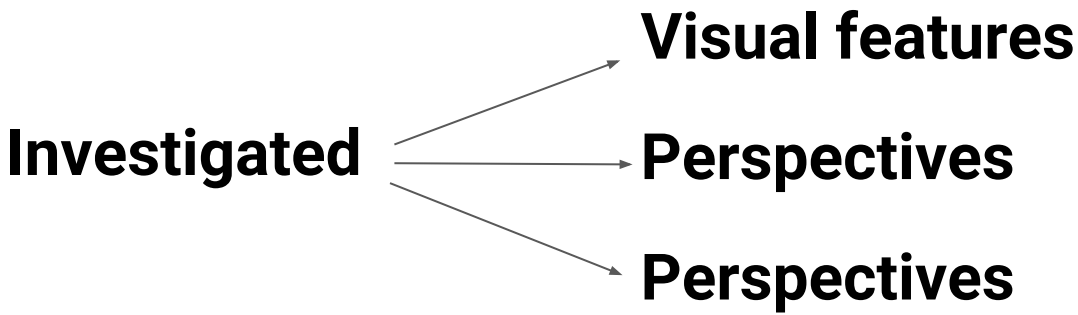



Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation

Rosario Scalise, Shen Li
rscalise@andrew.cmu.edu, shenli@cmu.edu

Thank You!

Learn More @ Poster Session



Dataset will be made available soon!