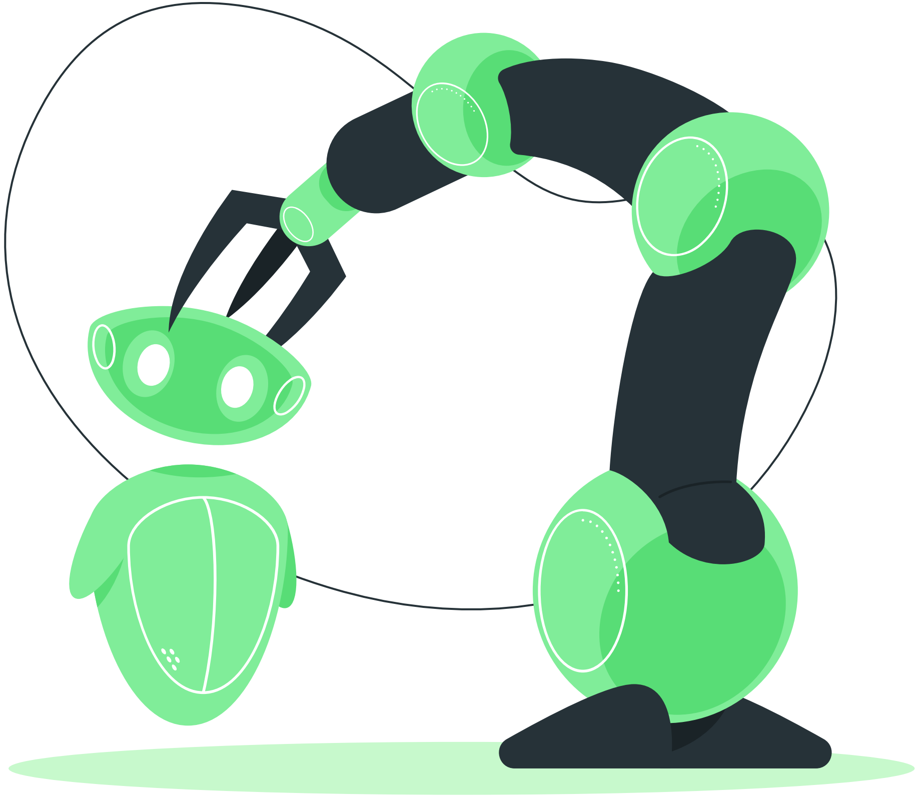
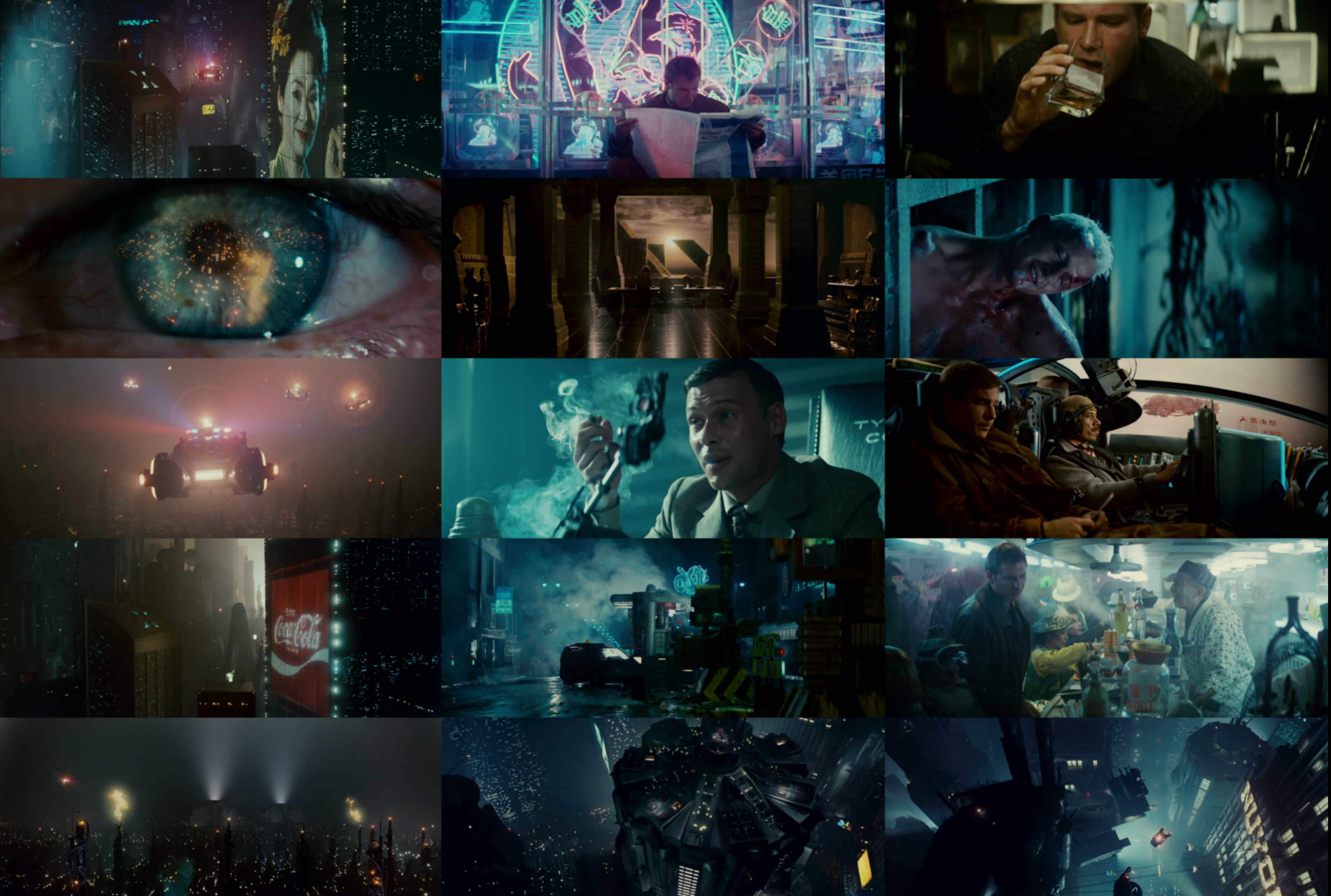
Robot-mediated Referential Communication: To Improve Trust in Human-robot Interaction

Students:
Yigang Qin, Huiqi Zou

Mentors:
Dr Xiaopeng Zhao, Ziming Liu,
Dr Kwai Wong





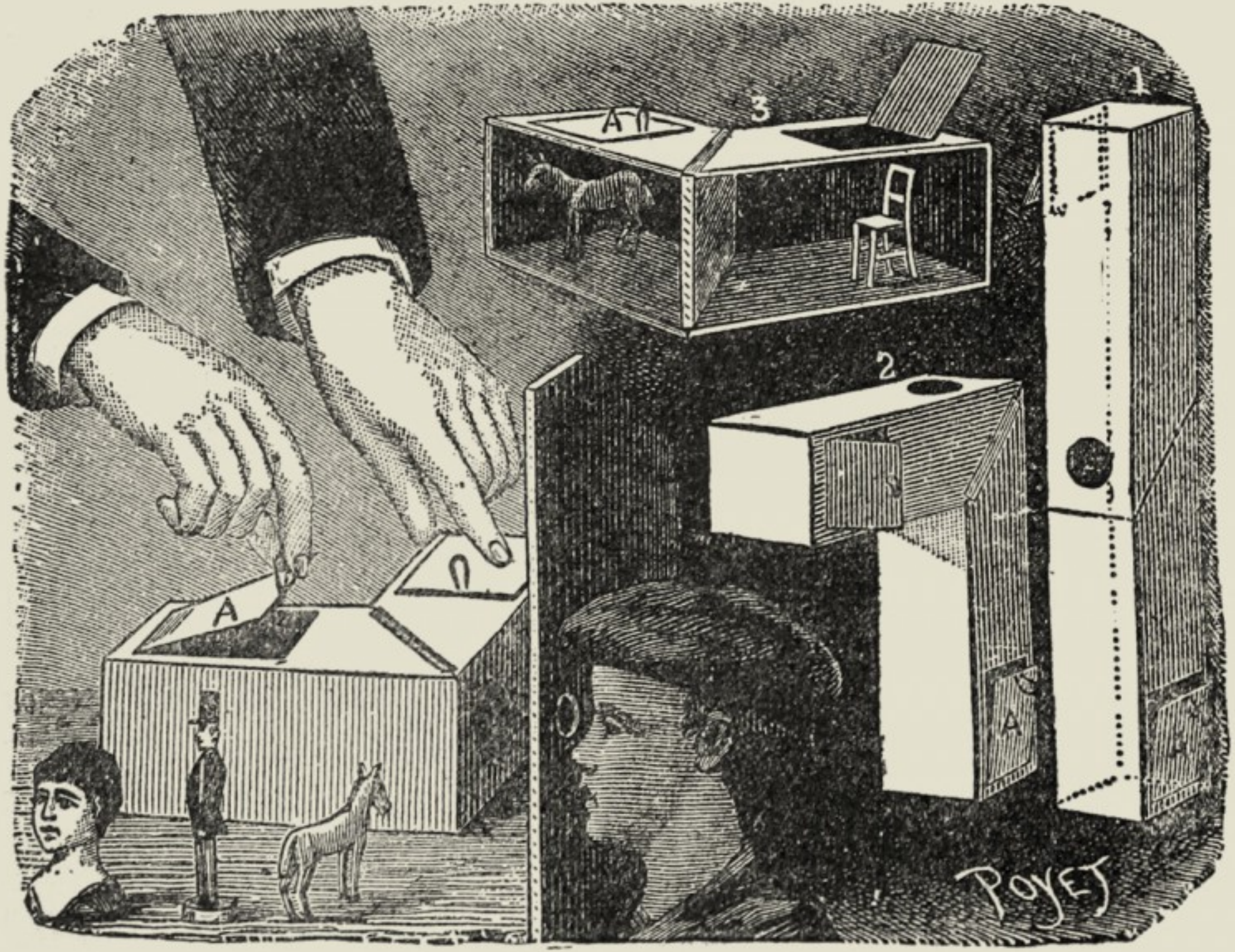
Blade Runner
(1982)
Ridley Scott

Что я им важен...

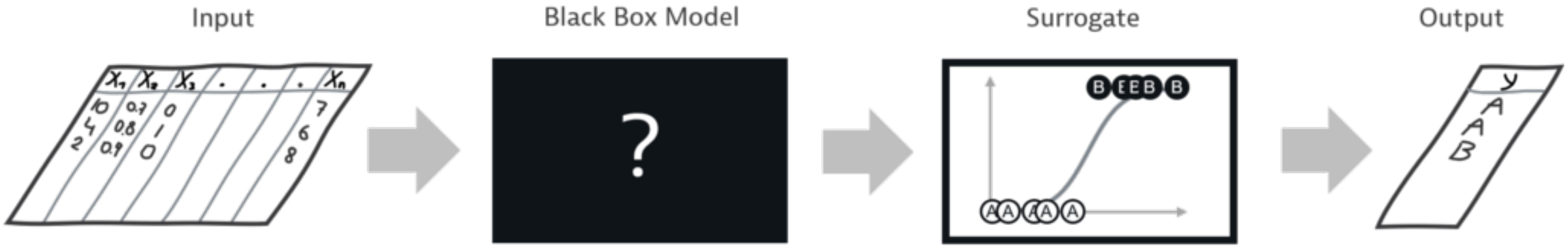
Detroit: Become Human
(2020)

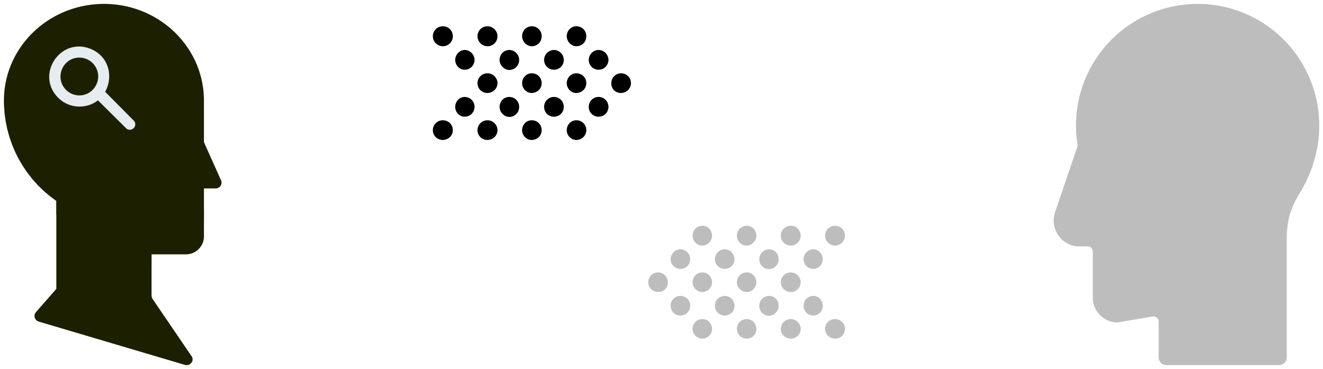
Trust AI robot?

A prosthetic arm displayed at an exhibit about science and technology changing perceptions of humanity.

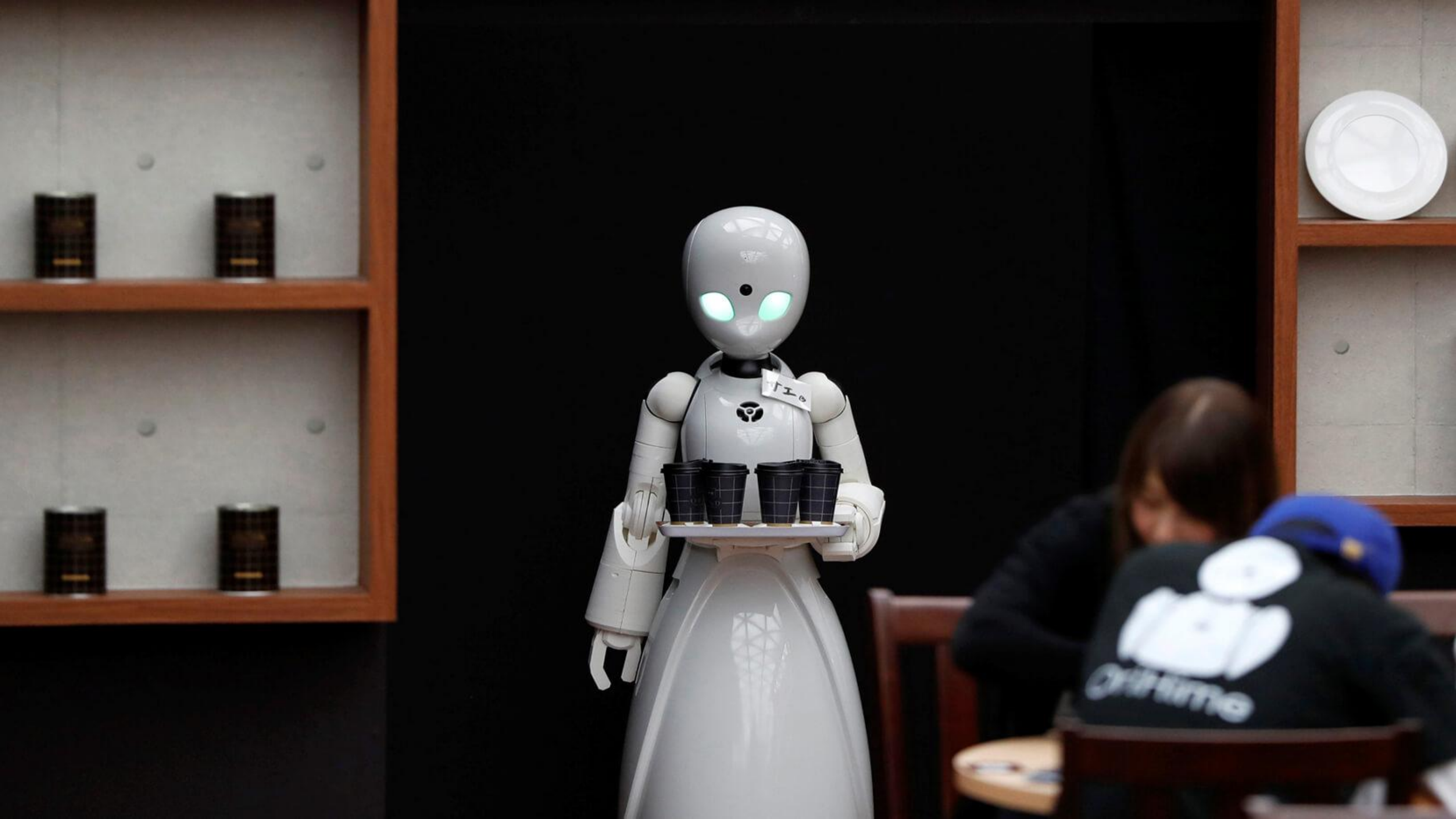


The Black Box Society: The Secret Algorithms That Control Money and Information





Do we know the thinking of an AI model or a robot?


OriHime

Socially Assistive Robot (SAR)

Fairness

Accountability

Transparency

Ethics

Improve human’ s trust in robot-mediated referential communication task

Table of Contents

1 Concept Map

Theory of Mind
Referential Communication
Joint Review

2 Experiment

Experiment Setting

3 System Architecture

Data and System Architecture

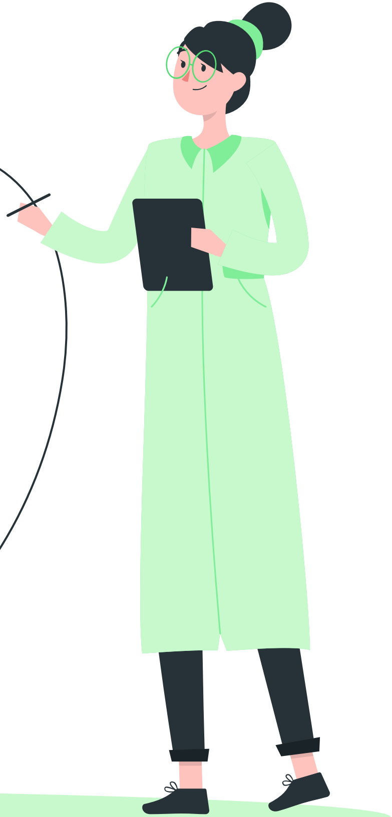
4 Validation

Metrics and Results

1

Concept Map

Theory of Mind, Referential Communication, and Joint Review,



Theory of Mind

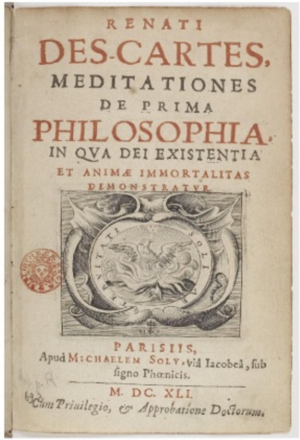
The basic cognitive and social characteristic that enables us to make conjectures about others’ minds through observable, or latent behavioral, and verbal cues.

Theory of Mind

Philosophical root: Philosophy of Mind

René Descartes. *Meditations on First Philosophy*. 1641

The Nature of Human Mind



RENATI
DES-CARTES,
MEDITATIONES
DE PRIMA
PHILOSOPHIA,
IN QVA DEI EXISTENTIA
ET ANIMÆ IMMORTALITAS
DEMONSTRATVR.

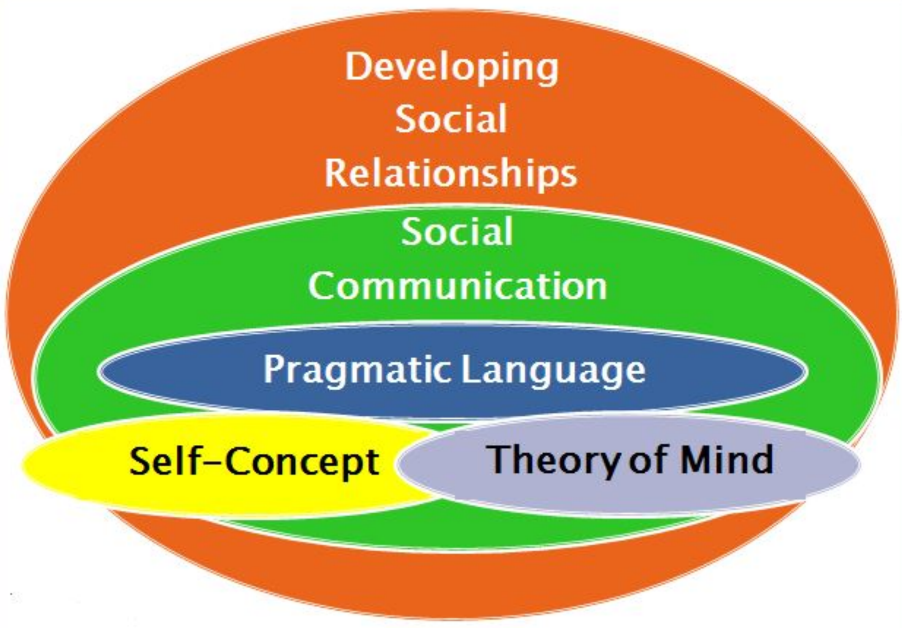
PARISIIS,
Apud MICHAELEM SOLY, viâ Iacobeâ, ſub
ſigno Phœnicis.
M. DC. XLI.
Cum Priuilegio, & Approbatione Doctorum.

Theory of Mind

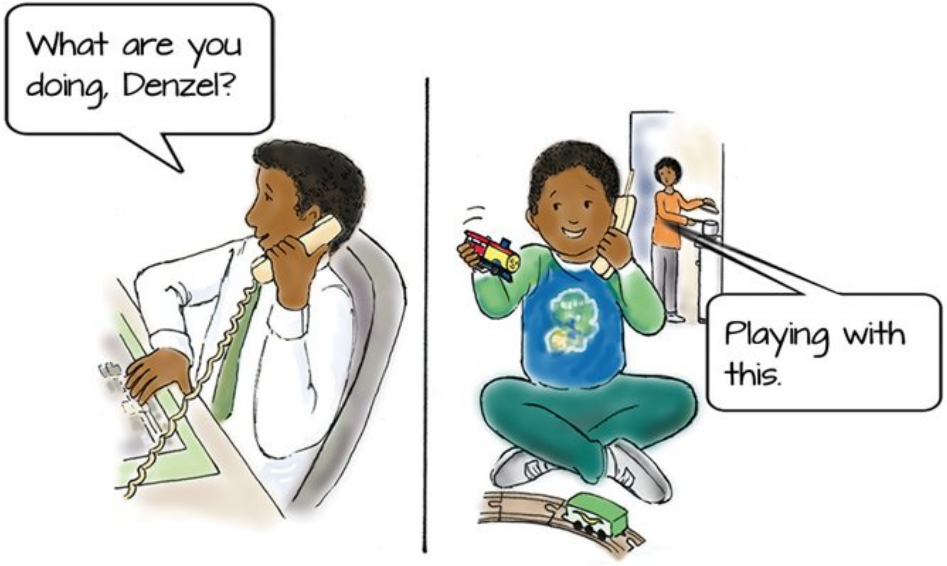
The ability to take someone else’s perspective

> empathy: *the ability to understand and share the feelings of others*

Theory of Mind



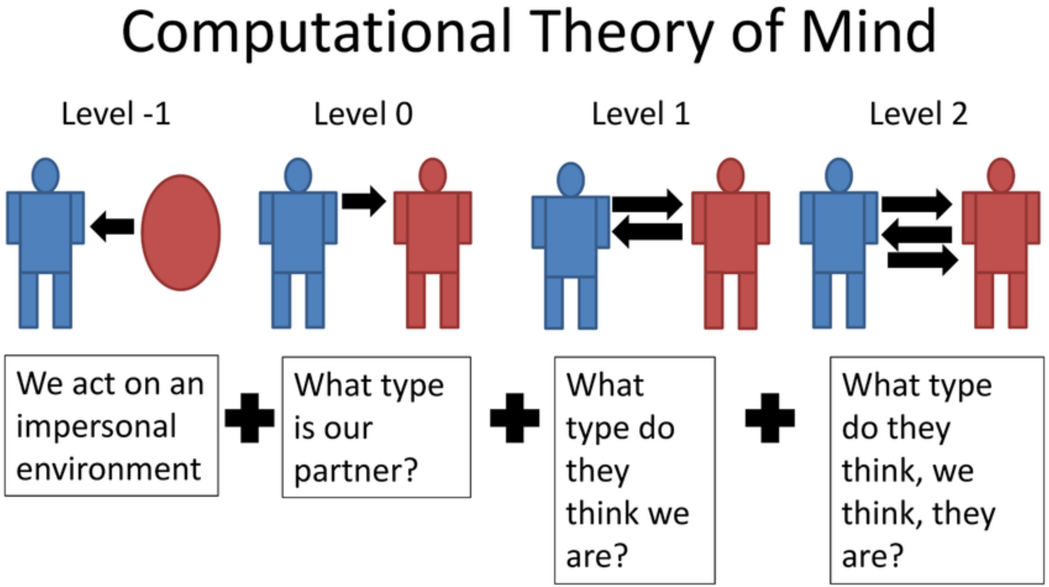
Theory of Mind



Children’s theory of mind in development

- □ Recognizing other’s feelings
- □ Thinking about consequences of actions
- □ Recognizing that someone else may think or feel differently than you do

Theory of Mind



Computational Theory of Mind

We act on an impersonal environment

+

What type is our partner?

+

What type do they think we are?

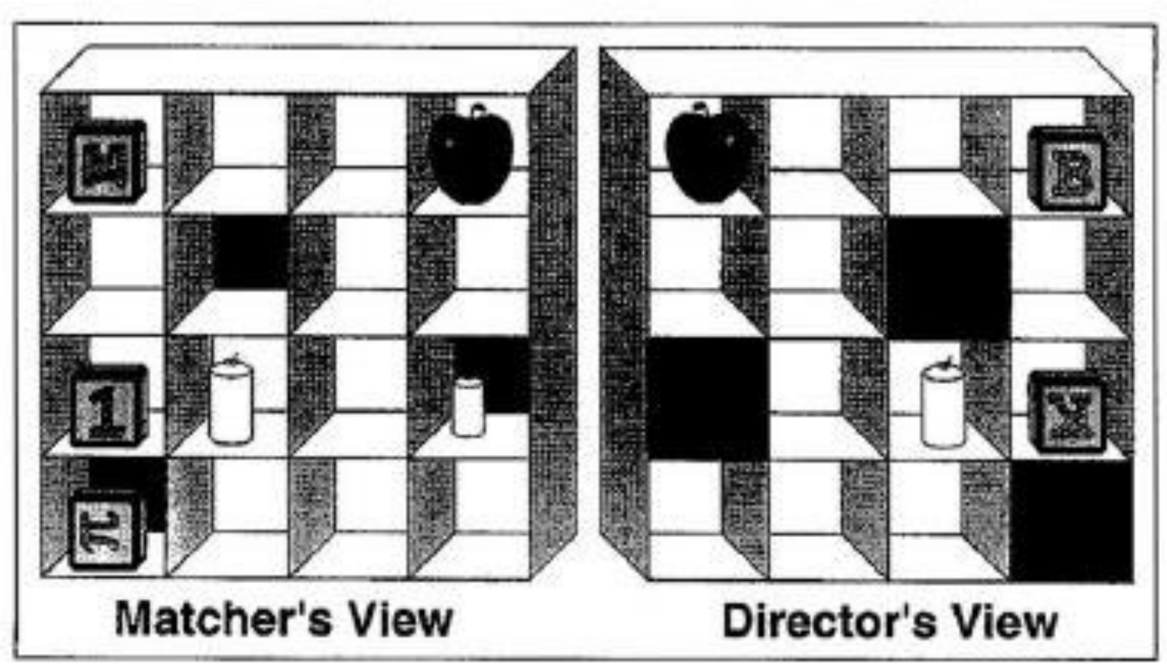
+

What type do they think, we think, they are?

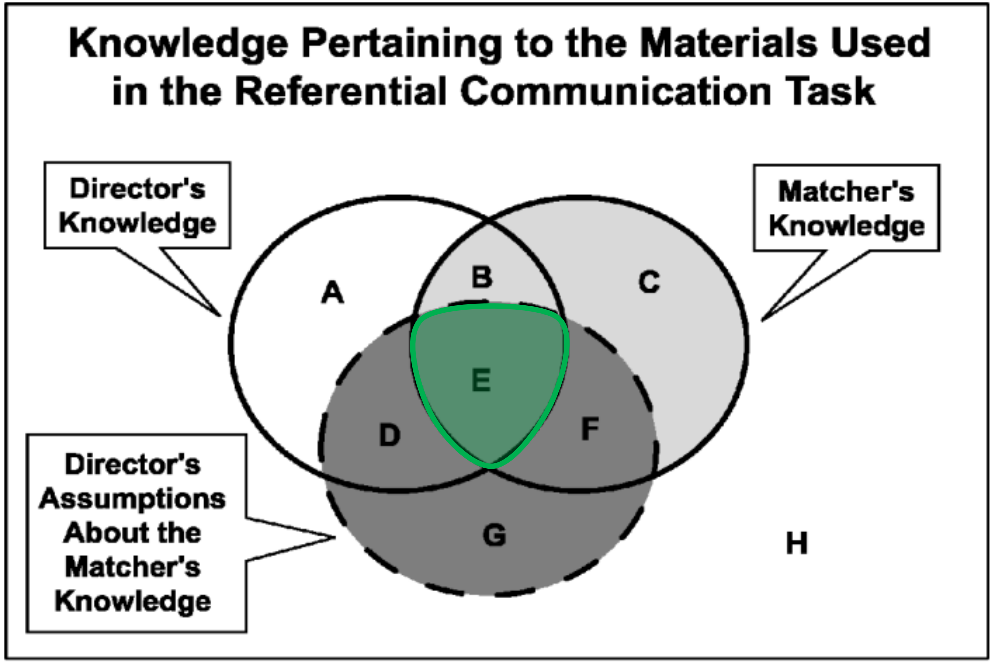
Referential Communication

Communicative action of referring to … something

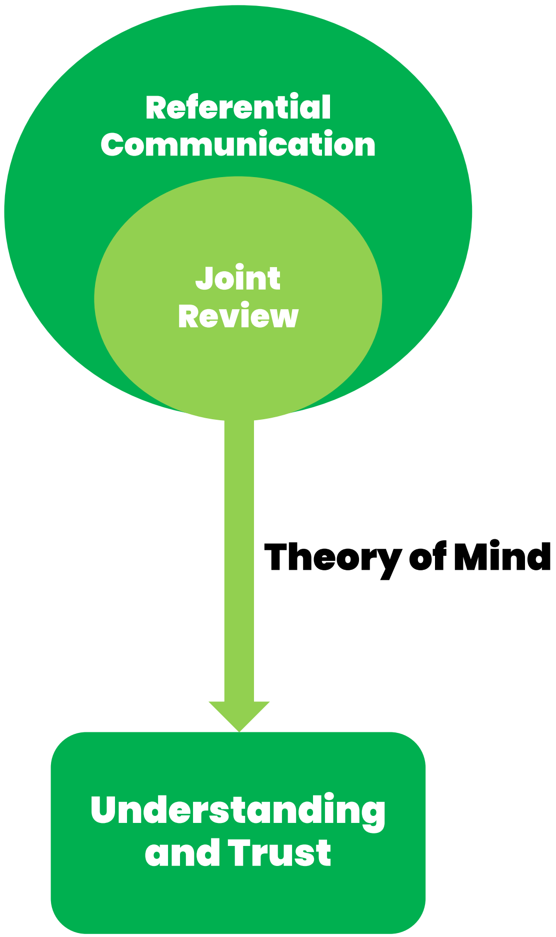
Referential Communication



The most used communicative strategy in Referential Communication Task is Joint Review which is closely related to the Theory of Mind


Knowledge Pertaining to the Materials Used in the Referential Communication Task
Director's Knowledge
Matcher's Knowledge
Director's Assumptions About the Matcher's Knowledge
A
B
C
D
E
F
G
H

A + D:	Director's Unique Knowledge
C + F:	Matcher's Unique Knowledge
D + E:	Director Assumed Shared Knowledge
B + E:	Actual Shared Knowledge
D:	Overestimated Shared Knowledge
E:	Correctly Assumed Shared Knowledge
B:	Underestimated Shared Knowledge
F:	Correctly Assumed Matcher's Knowledge
G:	Overestimated Matcher's Unique Knowledge
H:	Shared Ignorance


Referential Communication

Joint Review

Theory of Mind

Understanding and Trust

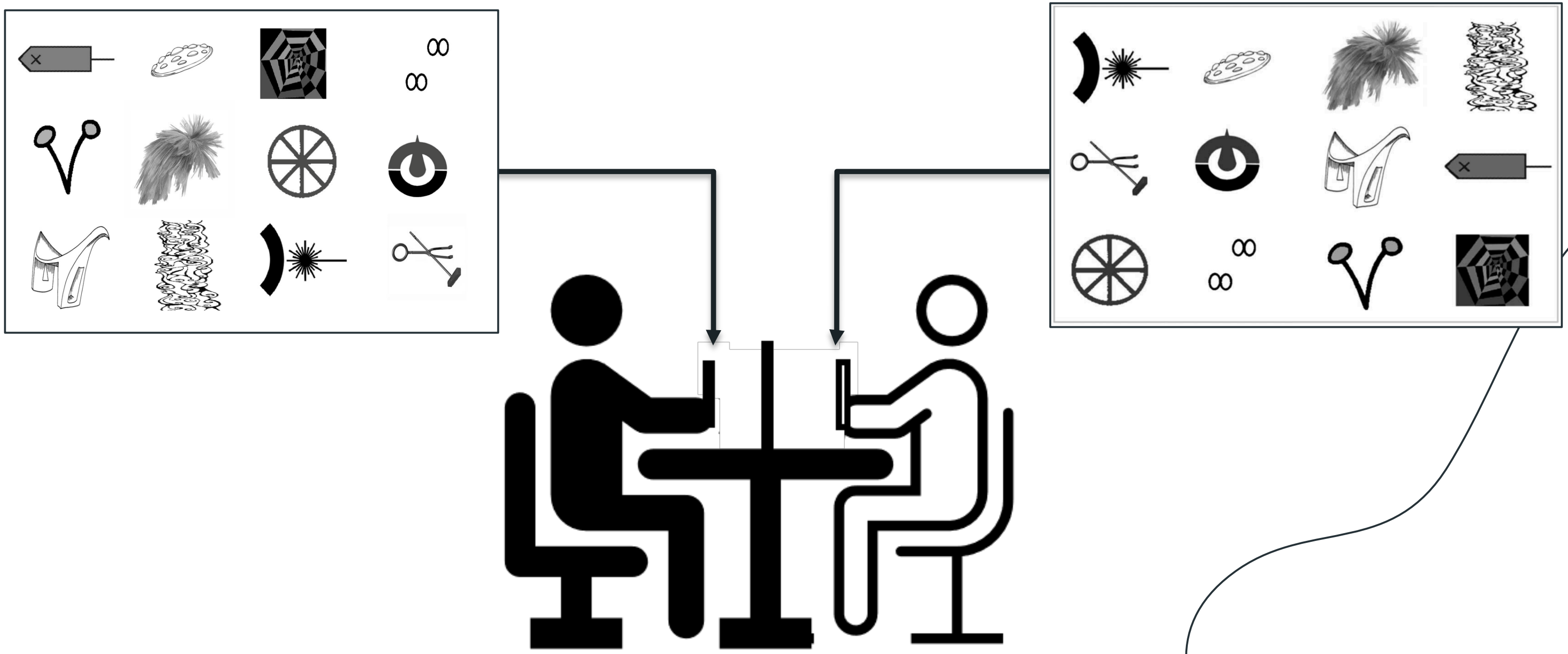
2

Experiment

Referential Communication Task

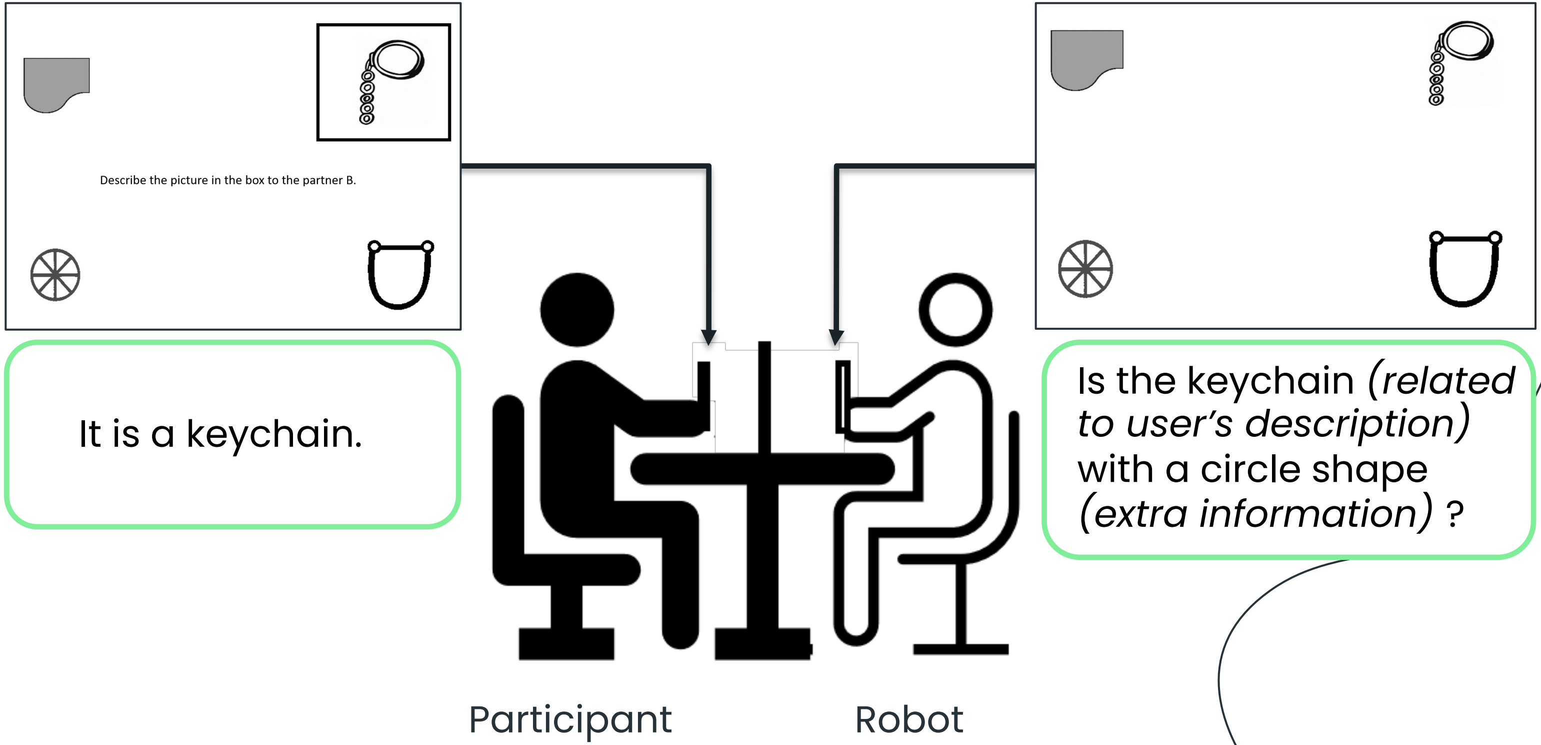
Sorting Phase

Purpose: *Guide the participant in understanding how to communicate with the robot*



Testing Phase

An example that the robot provides extra information relevant to participant's description



Describe the picture in the box to the partner B.

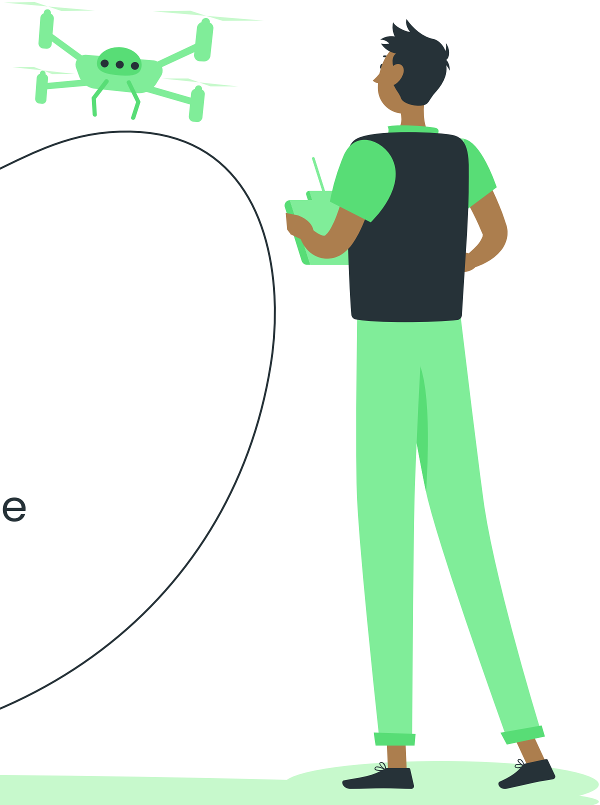
It is a keychain.

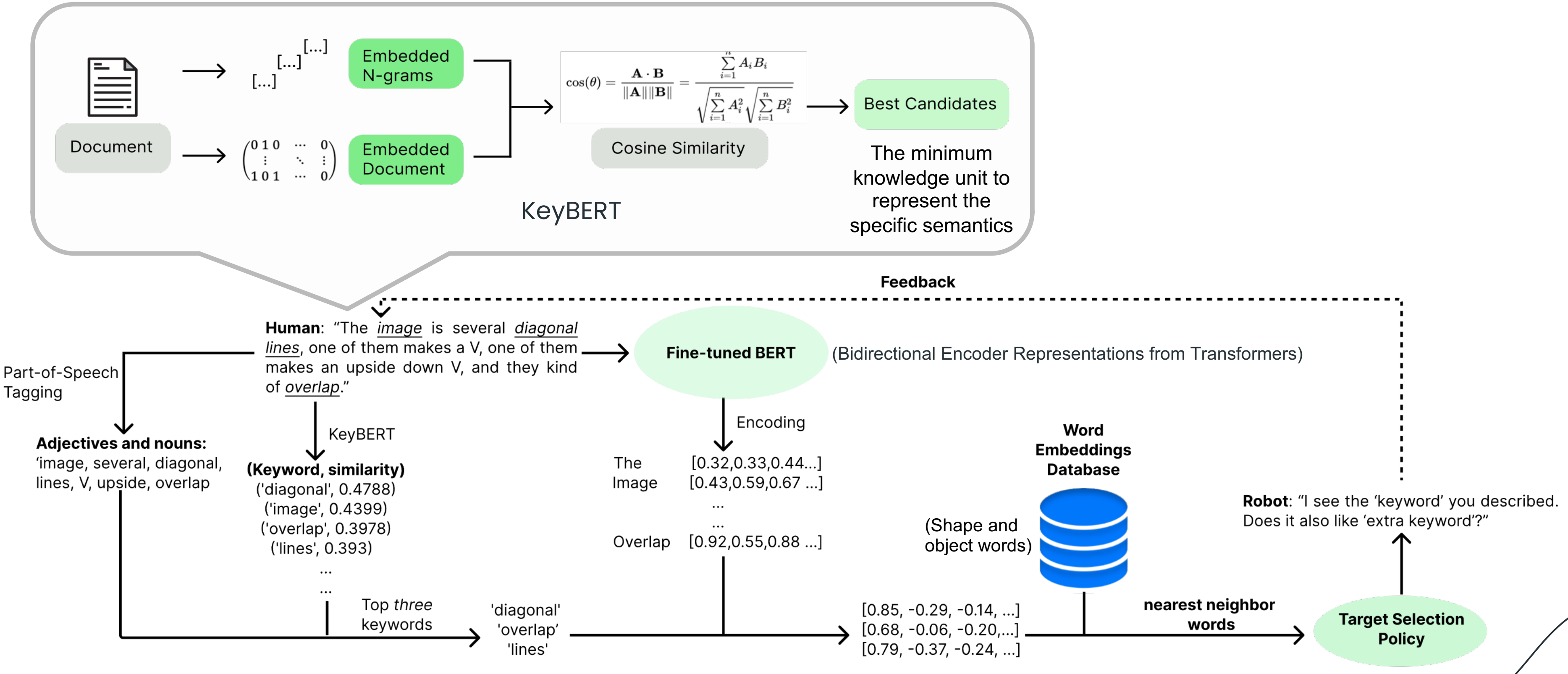
Is the keychain *(related to user's description)* with a circle shape *(extra information)* ?

3

System Architecture

Data and System Architecture





Document

[...] [...] [...]

Embedded N-grams

$$\begin{pmatrix} 0\,1\,0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1\,0\,1 & \cdots & 0 \end{pmatrix}$$

Embedded Document

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine Similarity

Best Candidates

The minimum knowledge unit to represent the specific semantics

KeyBERT

Feedback

Human: "The *image* is several *diagonal lines*, one of them makes a V, one of them makes an upside down V, and they kind of *overlap*."

Fine-tuned BERT

(Bidirectional Encoder Representations from Transformers)

Part-of-Speech Tagging

Adjectives and nouns: 'image, several, diagonal, lines, V, upside, overlap

KeyBERT

(Keyword, similarity)
('diagonal', 0.4788)
('image', 0.4399)
('overlap', 0.3978)
('lines', 0.393)
...
...

Encoding

The [0.32,0.33,0.44...]
Image [0.43,0.59,0.67 ...]
...
...
Overlap [0.92,0.55,0.88 ...]

Word Embeddings Database

(Shape and object words)

Robot: "I see the 'keyword' you described. Does it also like 'extra keyword'?"

Top *three* keywords

'diagonal'
'overlap'
'lines'

[0.85, -0.29, -0.14, ...]
[0.68, -0.06, -0.20,...]
[0.79, -0.37, -0.24, ...]

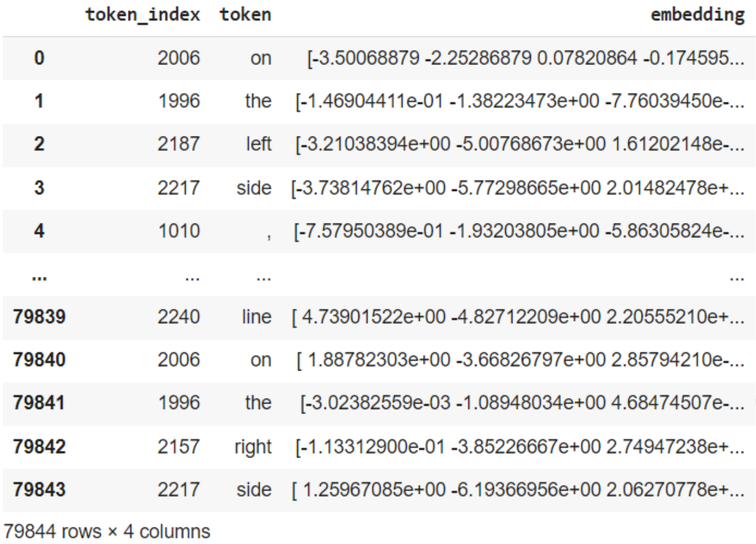
nearest neighbor words

Target Selection Policy

A dialog system that can provide near-human response

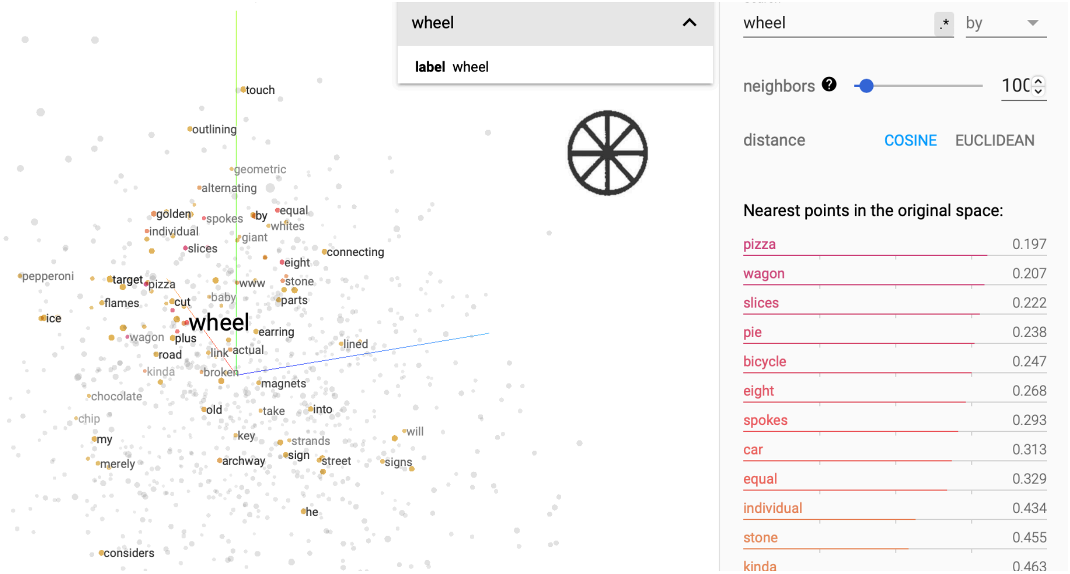
Word Embedding

numerically captures the semantic relations between words



	token_index	token	embedding
0	2006	on	[-3.50068879 -2.25286879 0.07820864 -0.174595...
1	1996	the	[-1.46904411e-01 -1.38223473e+00 -7.76039450e-...
2	2187	left	[-3.21038394e+00 -5.00768673e+00 1.61202148e-...
3	2217	side	[-3.73814762e+00 -5.77298665e+00 2.01482478e+...
4	1010	,	[-7.57950389e-01 -1.93203805e+00 -5.86305824e-...
...	...	...	...
79839	2240	line	[4.73901522e+00 -4.82712209e+00 2.20555210e+...
79840	2006	on	[1.88782303e+00 -3.66826797e+00 2.85794210e-...
79841	1996	the	[-3.02382559e-03 -1.08948034e+00 4.68474507e-...
79842	2157	right	[-1.13312900e-01 -3.85226667e+00 2.74947238e+...
79843	2217	side	[1.25967085e+00 -6.19366956e+00 2.06270778e+...

79844 rows × 4 columns



wheel

label wheel

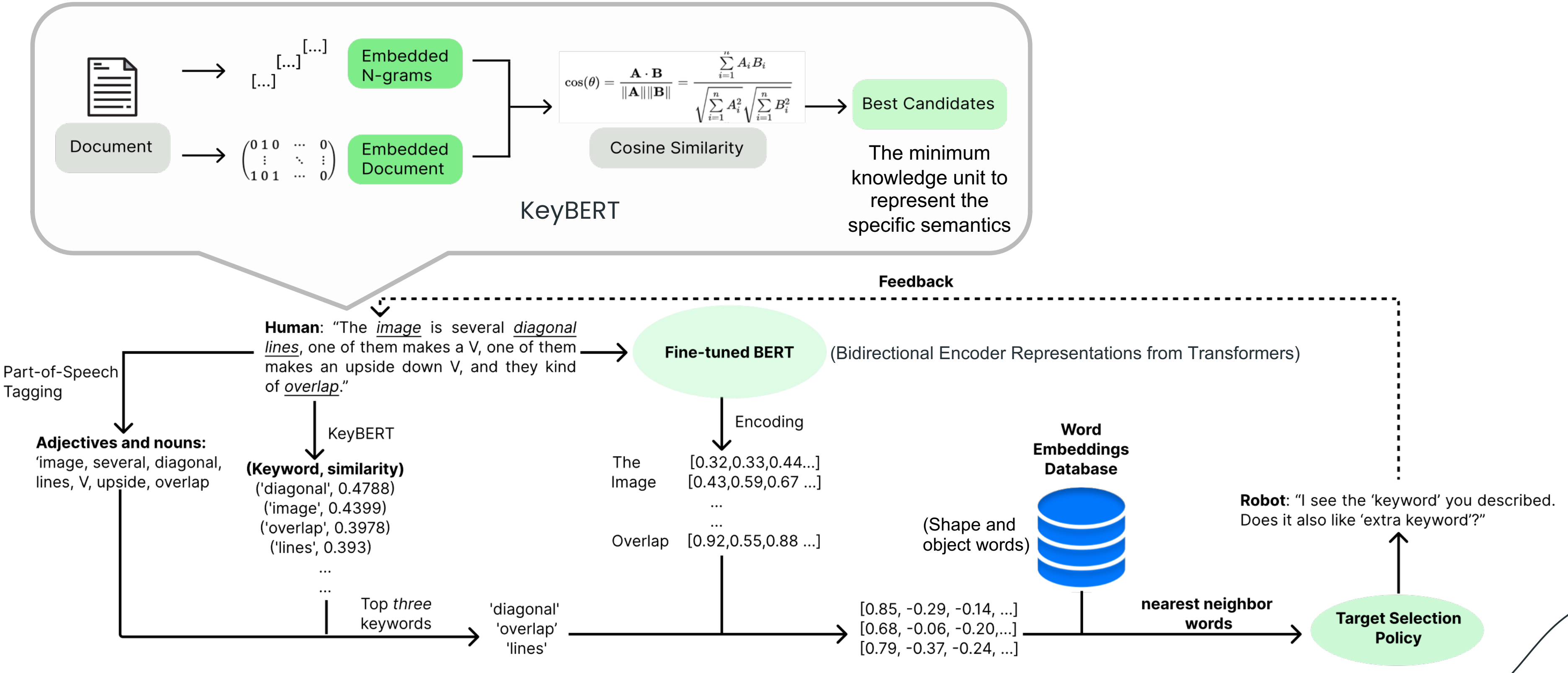
wheel .* by

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

pizza	0.197
wagon	0.207
slices	0.222
pie	0.238
bicycle	0.247
eight	0.268
spokes	0.293
car	0.313
equal	0.329
individual	0.434
stone	0.455
kinda	0.463


Document
[...]
[...]
[...]
Embedded N-grams
Embedded Document
$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$
Cosine Similarity
Best Candidates
The minimum knowledge unit to represent the specific semantics
KeyBERT
Feedback
Part-of-Speech Tagging
KeyBERT
Fine-tuned BERT
(Bidirectional Encoder Representations from Transformers)
Encoding
Word Embeddings Database
(Shape and object words)
Top *three* keywords
nearest neighbor words
Target Selection Policy

Human: “The *image* is several *diagonal lines*, one of them makes a V, one of them makes an upside down V, and they kind of *overlap*.”

Adjectives and nouns: ‘image, several, diagonal, lines, V, upside, overlap

(Keyword, similarity)
('diagonal', 0.4788)
('image', 0.4399)
('overlap', 0.3978)
('lines', 0.393)
...
...

The	[0.32,0.33,0.44...]
Image	[0.43,0.59,0.67 ...]
	...
	...
Overlap	[0.92,0.55,0.88 ...]

'diagonal'
'overlap'
'lines'

[0.85, -0.29, -0.14, ...]
[0.68, -0.06, -0.20,...]
[0.79, -0.37, -0.24, ...]

Robot: “I see the ‘keyword’ you described. Does it also like ‘extra keyword’?”

4

Validation

Metrics and Results

Factor

Token representation

- Use the output features from the last layer
- Sum all the output features from the last four layers

BERT training approach

- Within-task-pre-trained and then fine-tuned (BERT-ITPT-FIT)
- Direct fine-tuned (BERT-FIT)

Information saturation

- Normal situation
- Worst situation (Shape and object words excluded)

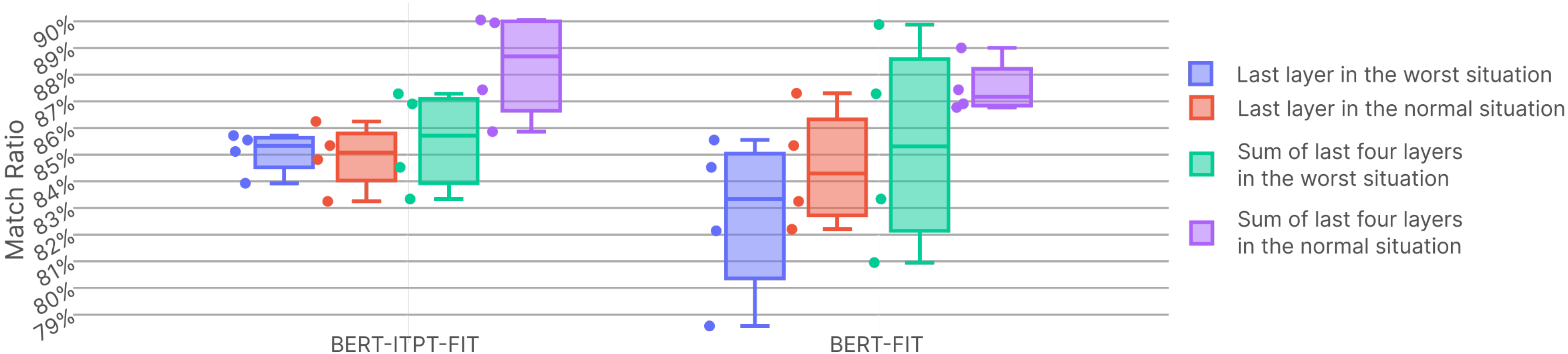
Transcript Classification

Model	Accuracy	Precision	Recall	F1	Subset-model Accuracy
BERT-FIT	.846761 (.027618)	.843257 (.027536)	.864997 (.025090)	.845208 (.027731)	**.818619** (.030962)
BERT-ITPT-FIT	**.850260** (.025637)	**.845776** (.027907)	**.867514** (.023780)	**.849688** (.024957)	.814890 (.029276)

10-fold Cross-validation metrics M (SD) on 48-class-transcript classification

Dialog Simulation

Match: *one of the three extracted words is in the transcripts from the training dataset describing the same target image*



Simulation results for the normal and worst situations

System Features

Understand the users' descriptions

+

Extract keywords for clarification

□

Enhance users' understanding on robot's intention

+

Improve users' trusts towards the robot

References

[1] E. Lunsford, ‘Robots visit Knoxville neuroscience clinic to help improve their artificial intelligence.’, *https://www.wvlt.tv*. https://www.wvlt.tv/2021/08/27/robots-visit-knoxville-neuroscience-clinic-help-improve-their-artificial-intelligence/.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[3] R. Pan, Z. Liu, F. Yuan, M. Zare, X. Zhao, and R. J. Passonneau, ‘A Database of Multimodal Data to Construct a Simulated Dialogue Partner with Varying Degrees of Cognitive Health’, p. 8

[4] M. Grootendorst, ‘KeyBERT: Minimal keyword extraction with BERT.’ Zenodo, 2020. doi: 10.5281/zenodo.4461265.

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

Awesome Questions?