

Theory of Mind and AI

Swati Mishra

Human Centered Artificial Intelligence

Graduate Course - CAS 783

Winter 2025



ENGINEERING

What is Theory of Mind

- Ability for tracking other people's mental states is known as theory of mind.
 - Theory of mind refers to an interconnected set of notions that are combined to explain, predict, and justify the behavior of others
 - The understanding that others have intentions, desires, beliefs, perceptions, and emotions different from one's own and that such intentions, desires, and so forth affect people's actions and behaviors.
-

The Beginnings

THE BEHAVIORAL AND BRAIN SCIENCES (1978), 4, 515–526

Printed in the United States of America

Does the chimpanzee have a theory of mind?

David Premack

*Department of Psychology,
University of Pennsylvania,
Philadelphia, Penna. 19104*

Guy Woodruff

*University of Pennsylvania Primate Facility,
Honey Brook, Penna. 19344*

Experiment

Experiment Design: “Rather than confronting a chimpanzee with an inaccessible object and observe his possible problem solving, we have instead shown him a human actor confronting inaccessible objects, and asked the animal to indicate how he thought the human actor would solve his problem.”



Experiment

Experiment 1: Video Tapes

- Authors made four 30- second videotapes of a human actor in a cage similar to the chimpanzee's struggling to obtain bananas that were inaccessible in one of four different ways.
 - They were attached to the ceiling and out of reach overhead,
 - They were outside the cage wall and horizontally out of reach,
 - They were outside the cage, but with the actor's reach impeded by a box inside the cage, located between him and the bananas;
 - Not only was the actor's reach impeded by a box, but the latter was laden with heavy cement blocks.
-

Experiment

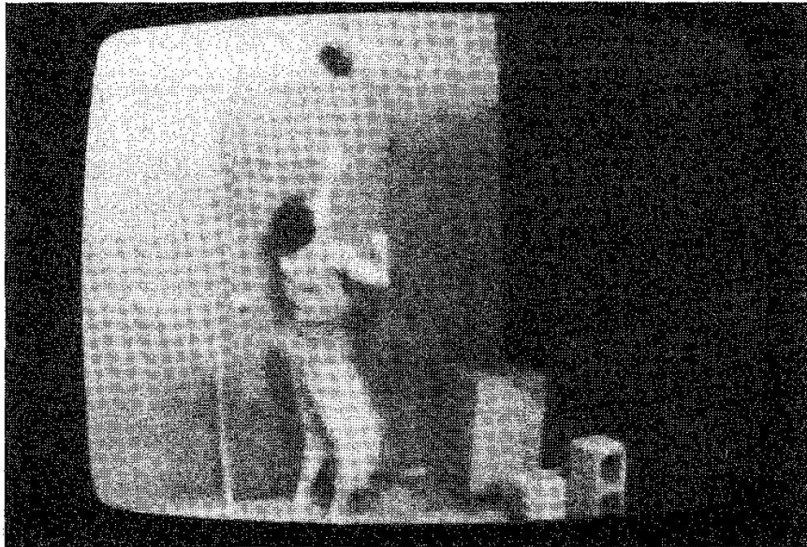
Experiment 1: Photographs

- In addition to the four videotapes, authors took still photographs of the human actor engaged in the behaviors that constituted solutions to the four problems.
 - He was photographed stepping onto a box;
 - Lying on his side and reaching out of the cage with a rod;
 - Moving a box to the side,
 - Removing cement blocks from a box.

Result: Sarah was successful in doing the task in 20 out of 24 trials.

Result Interpretation

- Experiment 1 has three interpretations:
 - Sarah matches physical elements in the correct alternative to the corresponding physical elements in the videotape.



Result Interpretation

- Experiment 1 has three interpretations:
 - Sarah matches physical elements in the correct alternative to the corresponding physical elements in the videotape.
 - **Associationism:** When shown a sequence that one is familiar with. Success comes from the similarity between old and new cases, a similarity that allows generalization to do its job.

How does the animal predict the next event in sequences that are not familiar?

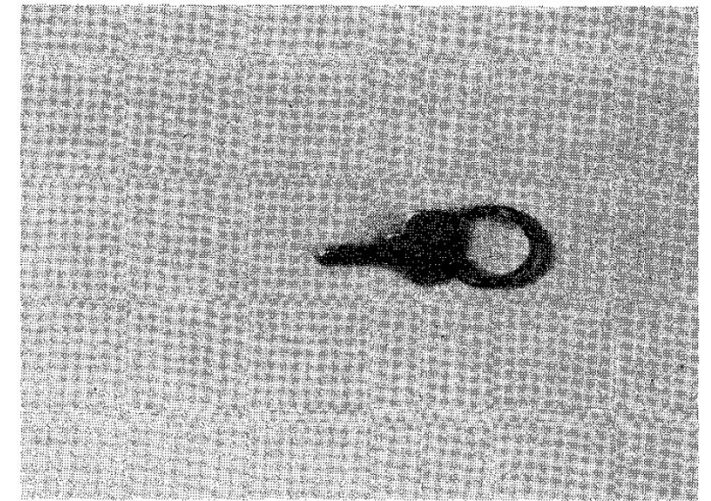
Result Interpretation

- Experiment 1 has three interpretations:
 - Sarah matches physical elements in the correct alternative to the corresponding physical elements in the videotape.
 - **Associationism:** When shown a sequence that one is familiar with. Success comes from the similarity between old and new cases, a similarity that allows generalization to do its job.
 - **Theory of Mind:** They solve the problems by imputing states of mind to the human actors, thus, trying to understand their belief.
 - **Empathy:** They can place themselves in the position of the actor and then use their knowledge to solve the problem.
-

Experiment

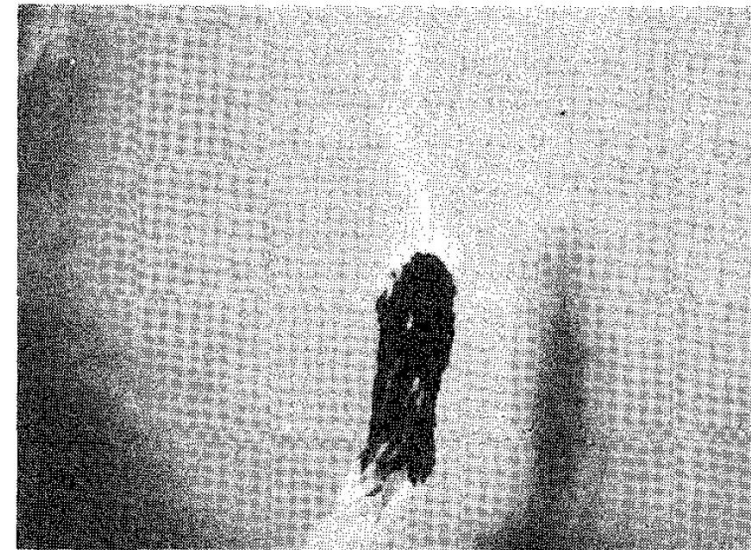
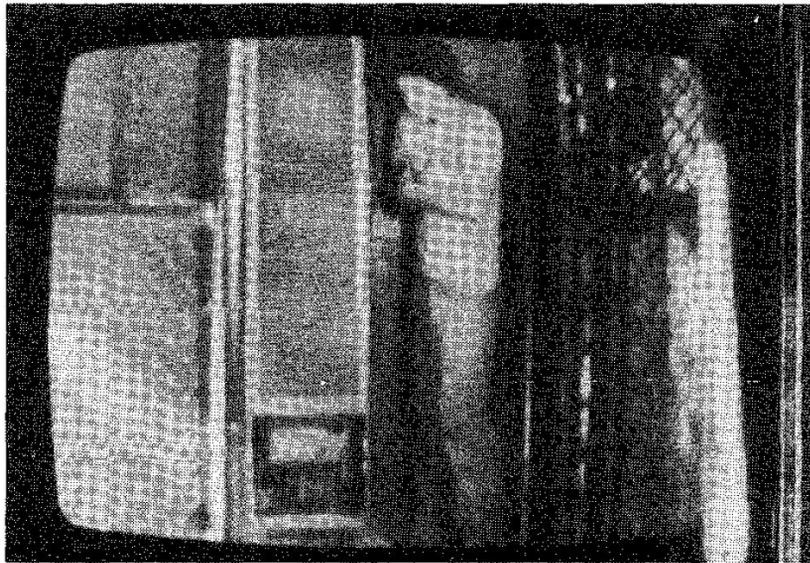
Experiment 2: Video Tapes and Photograph

- Authors made four 30- second videotapes of a human actor:
 - struggling to escape from a locked cage,
 - Trying to fix a malfunctioning heater (as witnessed by an actor who glanced wryly at the heater, even kicked it a little, and at the same time shivered and clasped his arms to his chest),
 - an actor seeking to play an unplugged phonograph,
 - an actor unable to wash down a dirty floor because the hose he held was not properly attached to the faucet.
- The photographs presented a solution.



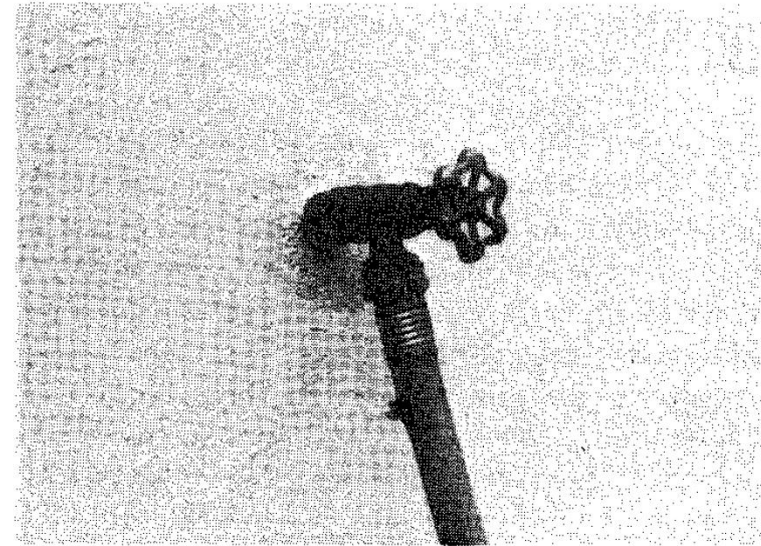
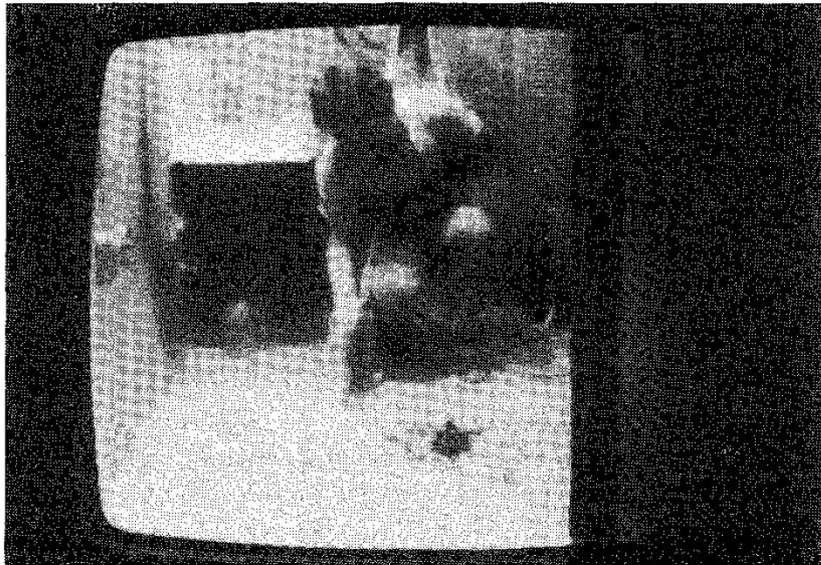
Experiment

Experiment 2: Photographs



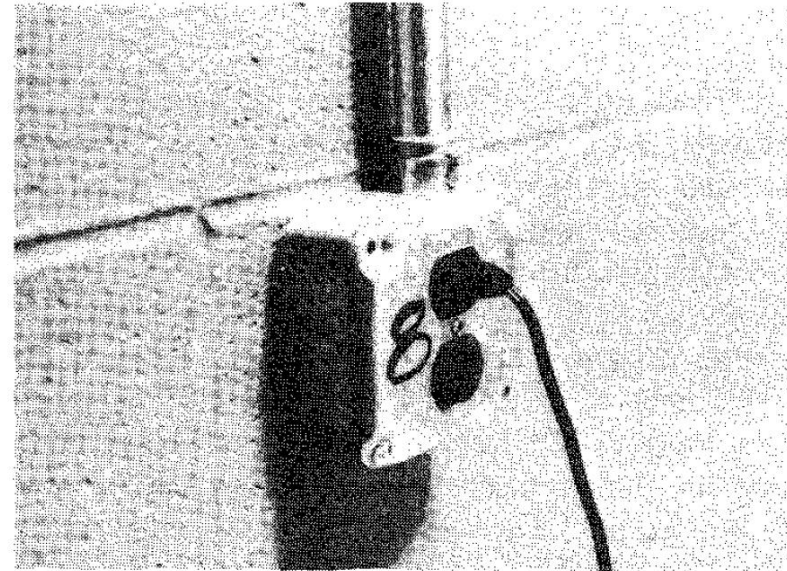
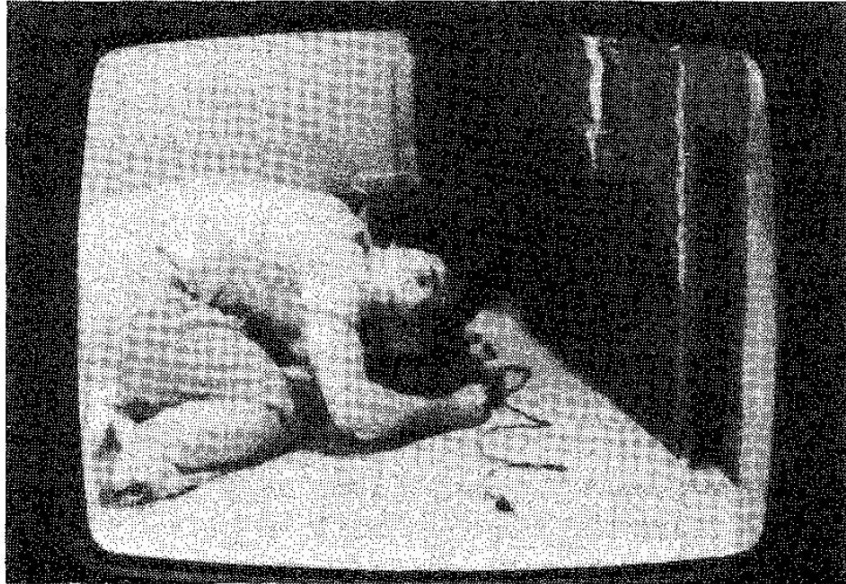
Experiment

Experiment 2: Photographs



Experiment

Experiment 2: Photographs



Findings

- Sarah may be trying to switch between modalities (‘would’/ ’should’) answer questions like:
 - What would a human actor do in this situation?
 - What should I do?
 - What would I like to see him do in this situation?
 - There is some evidence of agent specific knowledge and belief.
 - To have Theory of Mind Sarah must differentiate between:
 - Guess versus Know
 - Pretend versus Real
 - We do not know if :
 - Could Sarah assign different purpose to different agents? (Experiment 3)
 - Could she assign different knowledge and beliefs?
 - We want to find out:
 - What does the agent know about the world?
-

Discussion



Now and Forward...

nature human behaviour



Article








<https://doi.org/10.1038/s41562-024-01882-z>

Testing theory of mind in large language models and humans

Received: 14 August 2023

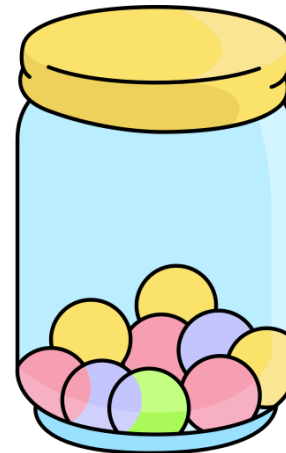
Accepted: 5 April 2024

Published online: 20 May 2024

James W. A. Strachan ¹✉, Dalila Albergo ^{2,3}, Giulia Borghini²,
Oriana Pansardi ^{1,2,4}, Eugenio Scaliti ^{1,2,5,6}, Saurabh Gupta ⁷, Krati Saxena ⁷,
Alessandro Rufo ⁷, Stefano Panzeri ⁸, Guido Manzi ⁷,
Michael S. A. Graziano⁹ & Cristina Becchio ^{1,2}✉

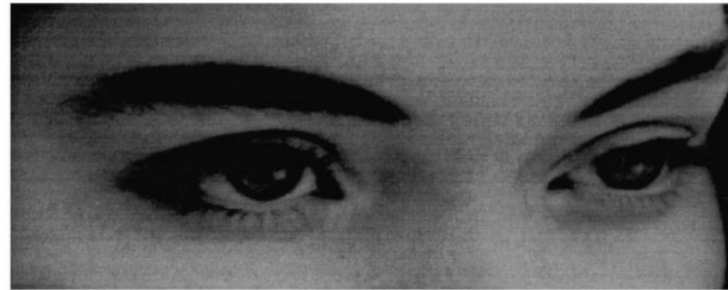
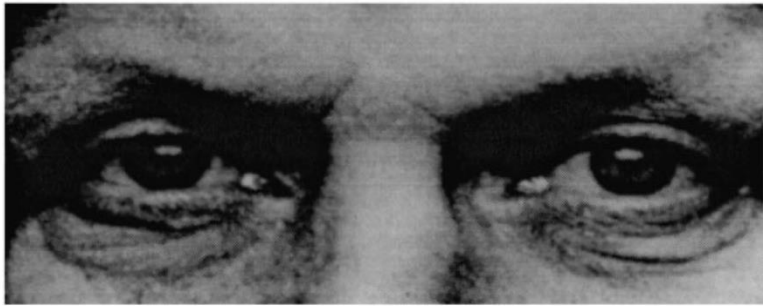
Now and Forward...

- Since then, quite a few tasks have been created to test Theory of Mind.
- Indirect measures of belief attribution using reaction times
- Looking or searching behavior.



Now and Forward...

- Since then, quite a few tasks have been created to test Theory of Mind.
- Indirect measures of belief attribution using reaction times
- Looking or searching behavior.
- Tasks examining the ability to infer mental states from photographs of eyes.



Now and Forward...

- Since then, quite a few tasks have been created to test Theory of Mind.
 - Indirect measures of belief attribution using reaction times
 - Looking or searching behavior.
 - Tasks examining the ability to infer mental states from photographs of eyes
 - Language-based tasks assessing false belief understanding
 - Pragmatic language comprehension.
-

False Belief Task

- First a story is presented of this format:

“Maxi leaves a chocolate in one location (e.g. the drawer) and while they are outside the room the object is transferred to a new location (e.g. the cupboard) by Dave who is in the room.”

- And then a follow up question is asked:

“Where would Maxi look for the chocolate?”

Faux Pas Tasks

- First a story is presented:

“Michael was a very awkward child when he was at high school. He struggled with making friends and spent his time alone writing poetry. However, after he left, he became a lot more confident and sociable. At his ten-year high school reunion he met Amanda, who had been in his English class. Over drinks, she said to him, *“I don't know if you remember this guy from school. He was in my English class. He wrote poetry and he was super awkward. I hope he isn't here tonight.”*

OR

Steve, a scientist, is traveling on a plane with his wife. Suddenly, he is tapped on the shoulder by another scientist. Steve looks up, sees that he knows this man, and says “Oh hi! How nice to run into you! Let me introduce you to my wife, Betsy. Betsy, this is Jeffrey, a good friend of mine from college days”. Betsy says “Oh, hi Jeffrey, pleased to meet you”. The other man replies “Er, my name isn't Jeffrey, it's Mike.”

Faux Pas Tasks

- Then the participants are asked the following questions:
 - In the story did someone say something that they should not have said?' [The correct answer is always 'yes']
 - What did they say that they should not have said? [Correct answer changes for each item]
 - A comprehension question to test understanding of story events [Question changes for every item]
 - A question to test awareness of the speaker's false belief phrased as, 'Did [the speaker] know that [what they said was inappropriate]?' [Question changes for every item. The correct answer is always 'no']
-

Hinting Tasks

- First a story is presented:

Paul has to go to an interview and he's running late. While he is cleaning his shoes, he says to his wife, "Jane, I want to wear that blue shirt but it's very wrinkled."

- And then a follow up question is asked:

What does Paul really mean when he says this?

Paul goes on to say, "It's in the ironing basket."

What does Paul want Jane to do?

Irony Tasks

Onni and Aleksi are in the school canteen. Onni gathers his plate full and eats up fast all the food. “Well, you were hungry,” Aleksi says to him.

The boys need to hurry to the next lesson.

OR

Onni and Aleksi are in the school canteen. Onni gathers a little food on the plate but eats only a small portion of it. “Well, you were hungry,” Aleksi says to him.

The boys need to hurry to the to the next lesson.

Strange Stories

These stories contain different type of vignettes:

- Lie,
- White Lie,
- Joke,
- Pretend,
- Misunderstanding,
- Persuade,
- Appearance/Reality,
- Figure of Speech,
- Sarcasm,
- Forget,
- Double Bluff,
- Contrary Emotions

Having been told each story, participants are asked to account for why the protagonist said what they did, usually with the question:
“Why did X say that?”

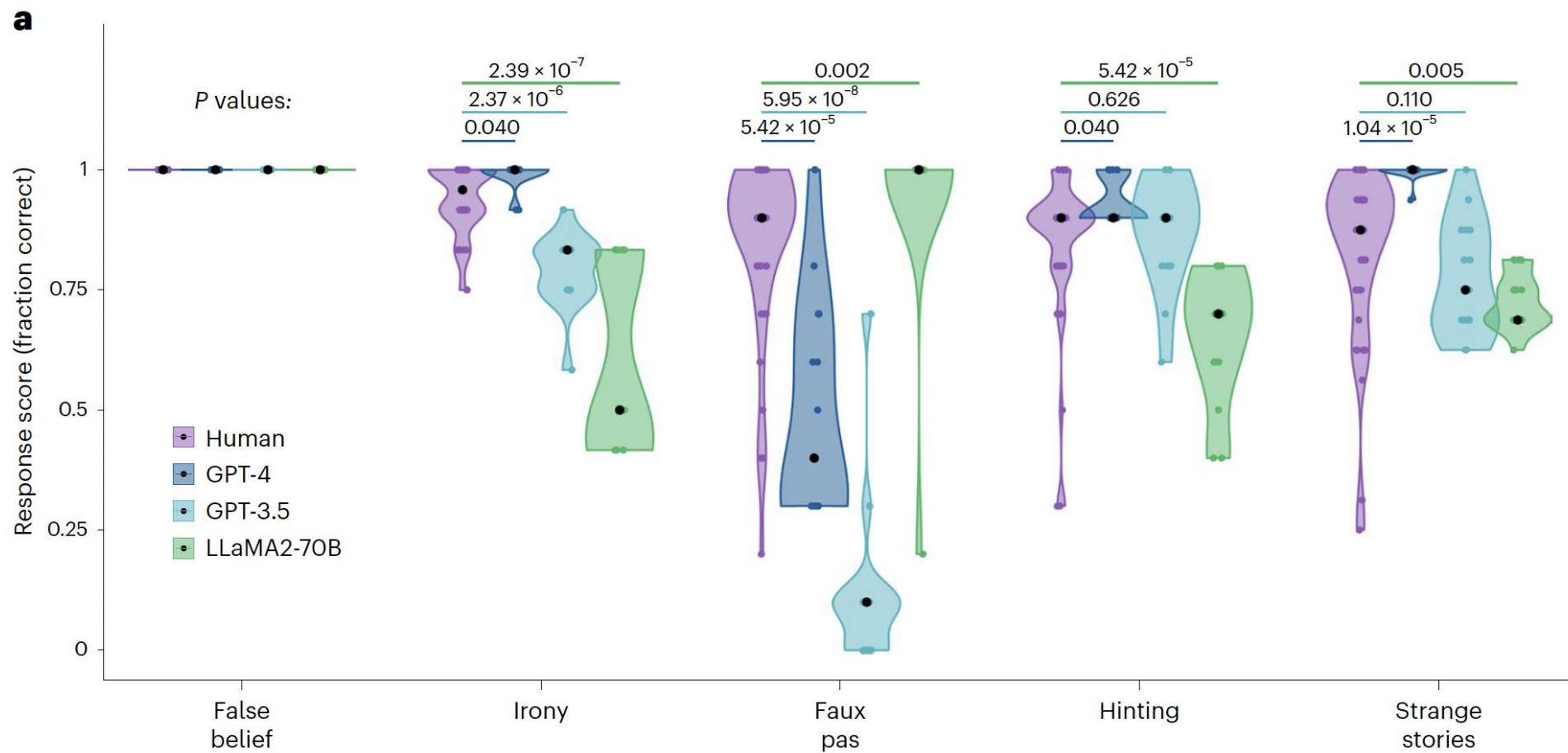
Data Collection

Test	Model	<i>N/n</i>	Items	Dates of data collection
Theory of mind battery	Human	250	7-16	June to July 2023
	GPT-4	75	7-16	April 2023
	GPT-3.5	75	7-16	April 2023
	LLaMA2	75	7-16	October to November 2023

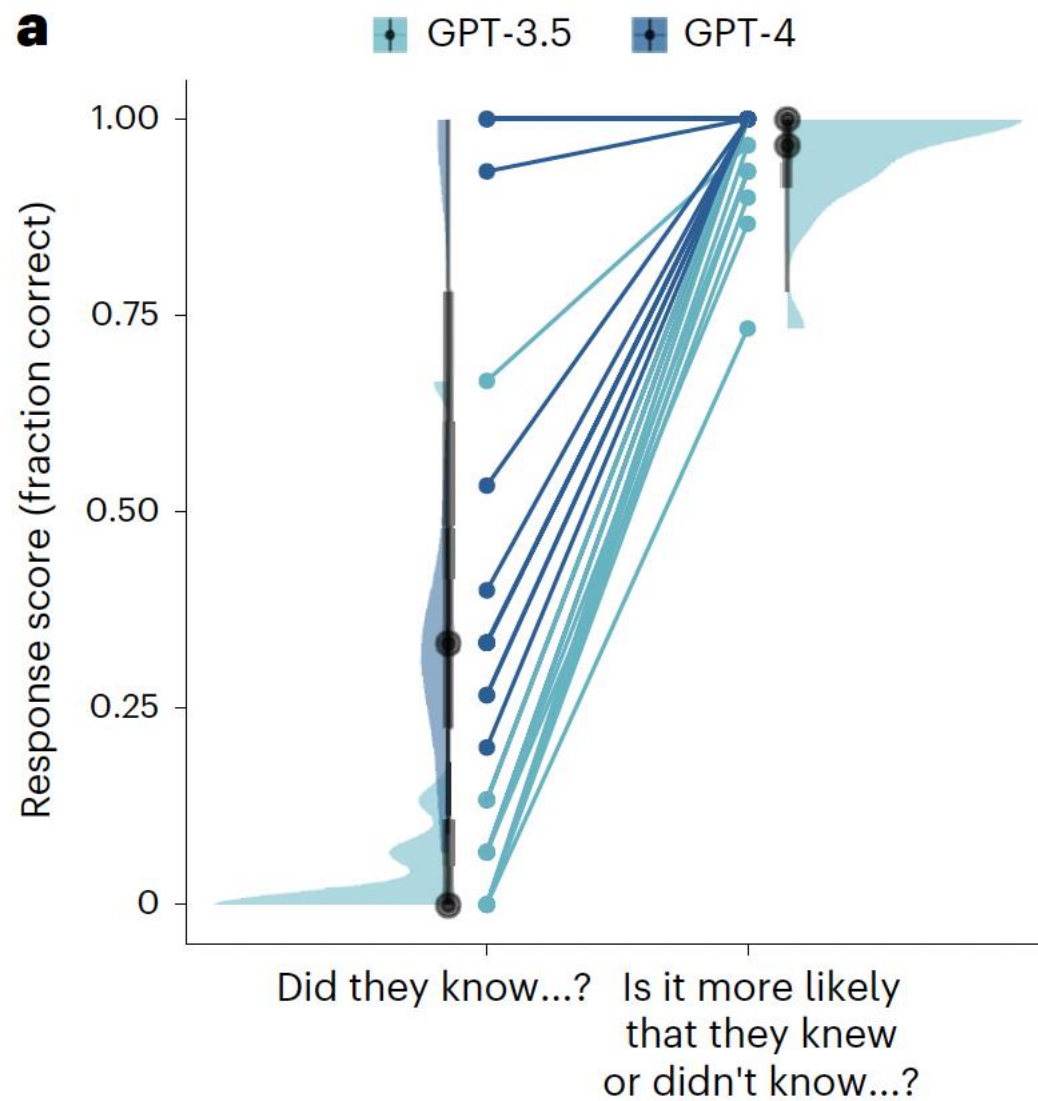
Data Collection

Faux pas likelihood test	GPT-4	15	15	April to May 2023
	GPT-3.5	15	15	April to May 2023
	LLaMA2	15	15	October to November 2023
Belief likelihood test	Human	900	1	November 2023
	GPT-4	270	1	October to November 2023
	GPT-3.5	270	1	October to November 2023
	LLaMA2	270	1	October to November 2023
Item order analysis	GPT-3.5	18	12–15	April to May 2023
False belief perturbations	Human	757	1	November 2023
	GPT-4	225	1	October to November 2023
	GPT-3.5	225	1	October to November 2023
	LLaMA2	225	1	October to November 2023

Findings



Findings



Discussion



Thank you

