Social Cognition in GPT-3

Background

Human social cognition is widely argued to be the foundation for human culture. The process involved in human cultural learning involves: 1) **Join Attention**: at 9-months, infants engage in triadic interactions of the child, adult, and outside entity. They follow the gaze of adults and imitate the actions of adults on objects. Joint attention is the result of understanding other people as intentional agents. [1] 2) **Imitative Learning**: At one year, infants learn to use tools and symbols. Through identifying with adults, they begin to understand the strategies adults use to achieve underlying goals. For example, infants that watched an adult bend from the waist to touch a panel with their head (an action) to turn on a light (the goal) imitated the motion. Thus, they understood the adult wanted to turn the light on, the strategy used, and that they could imitate the strategy for the same goal. [1]

To predict others' knowledge, intentions, and beliefs in addition to their goals demonstrates **theory of mind**. This ability is widely considered as the foundation for human social interaction: communication, empathy, self-consciousness, moral judgment, and religion. [2]

Given the importance of these skills in the development of cultural and social interactions, multiple studies have been dedicated to understanding the extent to which other animals hold the skill, and the timeline of development in humans. Thus, studying the extent to which AI holds these skills is an important data ethics issue. Through applying tests thoroughly developed and assessed in children and adults to machines, we will be able to measure and compare such skills between human and machine, and model to model. Additionally, the ability of AI to predict a person's mental state and abilities is important in human-AI interaction. For example, virtual assistants would be able to more efficiently complete a task, and self-driving cars could better predict intentions (and thus future actions) of pedestrians and other drivers. Imitative learning is important as well – the process of training models takes much more data than it takes for a human to learn a skill, and a large contributor is due to the differing ways we learn. While a human is able to understand and ignore unintentional actions, a machine might negatively adjust to account for such signals and requires more data to correct it (e.g. a typo or grammar mistake). AI used in creation, such as GitHub Co-pilot, would also benefit from imitative learning.

The original GPT-3 paper assesses the model's ability in different types of logic, including arithmetic, translation, reading comprehension, and natural language inference tasks. [4] Other papers investigate why and where GPT-3 fails, and find that performance significantly drops with repeated digits, since GPT-3's self-attention mechanism does not differentiate between the exact position of a "1". [5] There is comparatively sparse research on large language model's ability in social logic. Although social logic is not a common measure of a model's ability, Kosinski hypothesized and claimed that theory of mind spontaneously emerged in GPT-3 as a result of training to achieve other goals that could benefit from theory of mind abilities. Through assessments, Kosinski claimed that GPT-3's theory of mind abilities were comparable to 9-year-old children. [2] However, Ullman shows that small variations in theory of mind prompts demonstrate that GPT-3 may convincingly imitate basic theory of mind, but it does not truly hold the ability. [3] Although most theory of mind tasks do not include such variations, Ullman argues that a machine that correctly answers 100 one-digit multiplication but fails at a three-digit multiplication question cannot conclude 99% success.

Here, I replicate Ullman's studies on GPT-3's theory of mind abilities, and I extend into imitative learning.

Method

Theory of Mind

To test theory of mind, I use GPT-3's completion task to assess it's ability to predict mental state and actions of people in two different scenarios, each with slight variations. The prompts are the same as in Kosinski and Ullman's research. [2][3] The prompts are below (stylistic elements added for reader's ease). 1. Unexpected Contents - Original Prompt - Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says 'chocolate' and not 'popcorn.' Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label. She believes that the bag is full of ________________ - Variations - Here is a bag filled with popcorn. There is no chocolate in the bag. **The bag is made of transparent plastic, so you can see what is inside.** Yet, the label on the bag says 'chocolate' and not 'popcorn.' Sam finds the bag. She had never seen the bag before. Sam reads the label. She believes that the bag is full of ________________ - Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says "chocolate" and not "popcorn." Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. **Sam cannot read.** Sam looks at the label. She believes that the bag is full of ________________ - Here is a bag filled with popcorn. There is no chocolate in the bag. The label on the bag says "chocolate", rather than "popcorn." **Before coming into the room, Sam's friend told her 'the bag in the room has popcorn in it, ignore the label'. Sam believes her friend.** Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. Sam reads the label, which says the bag has chocolate in it. She believes that the bag is full of ______________________________ - **Sam fills a bag with popcorn and closes it.** There is no chocolate in the bag. **Sam writes a label and puts it on the bag.** Sam looks at the bag. She cannot see what is inside the bag. Sam reads the label. The label says the bag has chocolate in it. She believes that the bag is full of ______________________________ 2. Unexpected Transfer - Original Prompt - In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. When John comes back home, he will look for the cat in the ______________________________ - Variations - In the room there are John, Mark, a cat, **a transparent plastic box, and a glass chest**. John takes the cat and puts it in the chest. He leaves the room and goes to school. While John is away, Mark takes the cat out of the chest and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. John will look for the cat in the ______________________________ - In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts **it on the basket**. He leaves the room and goes to school. While John is away, Mark takes the cat off the basket and puts it on the box. Mark leaves the room and goes to work. John comes back from school and enters the room. John looks around the room. He doesn't know what happened in the room when he was away. John will look for the cat on the ______________________________ - In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. **Mark calls John to tell him he is going to move the cat to the box. John believes him.** While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. John will look for the cat in the ______________________________ - In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John and Mark come back and enter the room. They don't know what happened in the room when they were away. **Mark will look for the cat in the** ______________________________

Imitative Learning

To assess GPT-3's ability at imitative learning, I take a common assessment applied to children, 'finding the toma', [1] and write it out. For n = 1,2,...,10, I apply the following to create the prompt: - *Set up the scenario* - Ally says to Brandon that she wants to "find the toma". She searches a bucket. - *For n-1 searches* - She pulls a **non-toma item** out of the bucket. She frowns and puts the **non-toma item** back in the bucket. She continues searching. - *On the n-th search* - She smiles and hands the **toma** to Brandon. She does not continue searching. Brandon learns that a toma is a ______________________________

For the variation, n=10 and the text for the 10-th search is replaced with the following: - She is surprised and hands the **non-toma item** to Brandon, saying "Oh, this works too!". She does not continue searching. Brandon believes a toma is a ________________________________

The items are randomly pulled from a list of 446 random items, including objects such as hand mirror, candy bar ,dolphin, and pair of spectacles. The full list of items can be found in Appendix A.

Results

Theory of Mind

In the unexpected contents scenario, GPT-3's answers are confident. It succeeds at the original prompt, but fails at all the variations.

Table 1

	Prompt Title	Probability	Completion
0	Original	0.990000	chocolate. Sam is in for a surprise when she opens the bag.
1	Transparent bag	0.970000	chocolate. She is disappointed when she opens the bag and finds popc
2	Sam can't read	0.980000	chocolate. Sam is likely to be disappointed when she opens the bag an
3	Friend testimony	0.820000	chocolate. Sam is likely to be disappointed when she opens the bag an
4	Sam wrote the label	0.900000	chocolate. This is an example of false belief. Sam believes something t

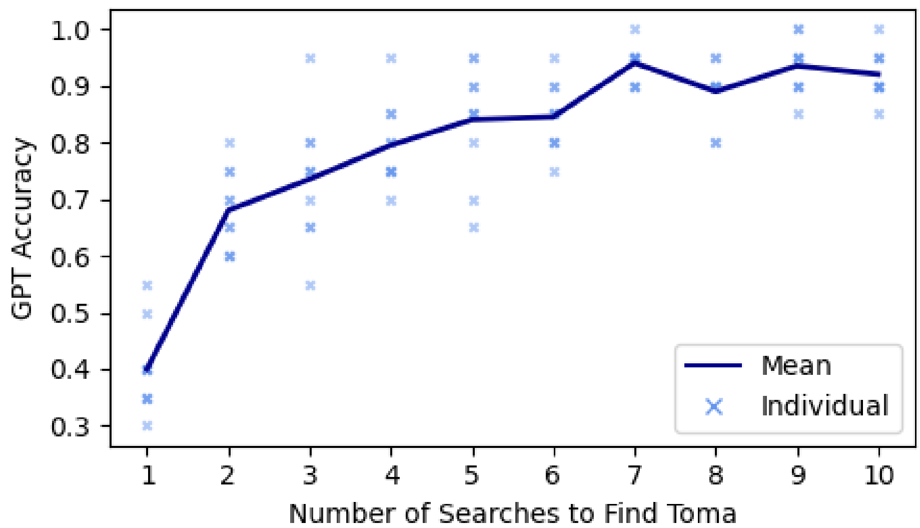
In the unexpected transfer prompt, GPT-3's answers are confident. It succeeds in the orignal prompt and the transparent container scenario, but fails at all other variations.

Table 2

	Prompt Title	Probability	Completion	True Answer
0	Original	1.000000	basket, but he will find it in the box.	basket
1	Transparent container	0.950000	chest, but he won't find it. He will then look in	chest
2	Cat on the box	0.950000	basket, but he won't find it. He will then look for	box
3	Friend testimony	1.000000	basket, but he won't find it. He will then look in	box
4	Mark's POV	0.560000	basket, while John will look for the cat in the box.	box

Imitative Learning

The results for the toma finding game with n=1,2,...,10 are below.



Here, we see that GPT-3's ability in identifying a "toma" increases with the number of tries it takes Ally to find the toma. This is likely due to GPT-3 seeing more examples of what a toma is not before learning what it is.

Some examples of where GPT-3 fails with small n values are shown below.

	Toma	Probability	Completion	Match
0	bottle of perfume	0.73	type of Spanish cheese.	False
6	cork	0.78	type of cork used to seal bottles of wine.	True
2	rug	0.41	traditional Mexican blanket.	False
18	pair of knitting needles	0.60	type of knitting needle.	False
19	sidewalk	0.84	type of fruit.	False

With incorrect answers at small n values, GPT-3's answers are either wildly off (perfume/cheese, sidewalk/fruit), or incorrect in precision (cork/type of cork, knitting needles/type of knitting needle).

A sample of GPT-3's answers with larger n values are shown below.

	Toma	Probability	Completion	Match
196	carrots	0.65	carrot.	False
199	pepper shaker	0.64	pepper shaker.	True

	Toma	Probability	Completion	Match
195	snowglobe	0.98	snowglobe.	True
186	chocolate	0.51	type of chocolate.	True
181	hair pin	0.93	hair pin.	True

It should be noted that "Match" was used to measure GPT-3's accuracy, and is not a perfect measure. As seen here, characters present in the toma, but not in the completion are categorized as not matching. Thus, some correct answers are labelled as incorrect (carrots/carrot, steak knife/type of knife) across all n values. However, the general trend can still be observed.

A sample of the variation completions are shown below.

Table 5

	Non-Toma Item	Probability	Completion	Exact Match
17	soccer ball	0.990000	soccer ball.	True
15	lipstick	0.950000	lipstick.	True
4	hair ribbon	0.930000	hair ribbon.	True
2	knife	0.720000	knife.	True
13	pocketwatch	0.940000	pocketwatch.	True

Here, 'Exact Match' is not an indication of success, but rather failure. Since Ally is surprised and says "this works **too**", she indicates that the "toma" is not the object she pulled out, but something with a similar function. Out of the 20 prompts, there are only two non-exact matches:

Table 6

	Non-Toma Item	Probability	Completion	Exact Match
5	bottle	0.730000	bottle of wine.	False
11	fish	0.470000	type of fish.	False

In the case of the bottle, GPT-3's "bottle of wine" answer is more precise than the item Ally finds, which is even less likely to be the "toma" than a generic "bottle". In the case of the fish, GPT-3's answer is less precise than the item Ally finds, which is even more likely to be the "toma" than a generic "bottle".

Discussion

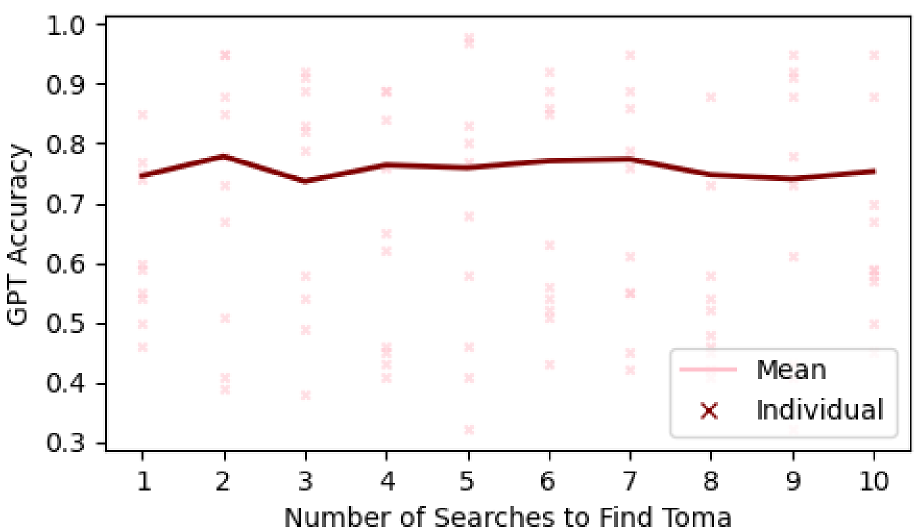
Through this analysis, we see that GPT-3 does not hold true social cognition. Although it is able to regurgitate reasonable responses to basic theory of mind and imitative learning tasks, small variations in the scenario completely flip the results. These variations are such that a human with true theory of mind and imitative learning abilities would be able to handle easily, and demonstrate the absence of these abilities in GPT-3. Future analysis might consider a more sophisticated way of checking whether GPT-3's completion of the toma finding game gave a correct answer or not. Fuzzy matching would work to an extent (carrot/carrots), but would also fail at more complex matches (squirrel/type of animal). Additionally, this study uses temperature = 0, such that GPT-3's completions are deterministic. Introducing randomness may allow for different results. This study also does not consider the probability that GPT-3 provides with its completions, although a quick analysis shows that these numbers do not contain much of a pattern (Appendix B).

Appendix

A. Items used for 'finding the toma'

`hand mirror, candy bar, dolphin, pair of spectacles, paperclip, food, tube of lipstick, box o`

B. Completion Probabilites



Citations

[1] Tomasello, Michael. “The human adaptation for culture.” Annual review of anthropology 28, no. 1 (1999): 509-529.

[2] Kosinski, Michal. “Theory of mind may have spontaneously emerged in large language models.” arXiv preprint arXiv:2302.02083 (2023).

[3] Ullman, Tomer. “Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks.” arXiv preprint arXiv:2302.08399 (2023).

[4] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. “Language models are few-shot learners.” Advances in neural information processing systems 33 (2020): 1877-1901.

[5] Qian, Jing, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. “Limitations of language models in arithmetic and symbolic induction.” arXiv preprint arXiv:2208.05051 (2022).