



HAL
open science

Can a conversational agent pass theory-of-mind tasks? A case study of ChatGPT with the Hinting, False Beliefs, and Strange Stories paradigms.

Eric Brunet-Gouet, Nathan Vidal, Paul Roux

► **To cite this version:**

Eric Brunet-Gouet, Nathan Vidal, Paul Roux. Can a conversational agent pass theory-of-mind tasks? A case study of ChatGPT with the Hinting, False Beliefs, and Strange Stories paradigms.. Human and Artificial Rationalities, Lecture Notes in Computer Science, 14522, Springer Nature Switzerland, pp.107-126, 2024, Lecture Notes in Computer Science, 978-3-031-55245-8. 10.1007/978-3-031-55245-8_7. hal-03991530v2

HAL Id: hal-03991530

<https://hal.science/hal-03991530v2>

Submitted on 21 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

Can a conversational agent pass theory-of-mind tasks? A case study of ChatGPT with the Hinting, False Beliefs, and Strange Stories paradigms.

Eric Brunet-Gouet^{1, 2}[0000-0002-3784-7817], Nathan Vidal²[0009-0009-1396-470X]

and Paul Roux^{1, 2}[0000-0003-0321-4189]

¹ Centre Hospitalier de Versailles, Service Hospitalo-Universitaire de Psychiatrie d'Adultes et d'Addictologie, Le Chesnay, France

² Université Paris-Saclay, Université Versailles Saint-Quentin-En-Yvelines, DisAP-DevPsy-CESP, INSERM UMR1018, 94807 Villejuif, France
ebrunet@ght78sud.fr

Abstract. We investigate the possibility that the recently proposed OpenAI's ChatGPT conversational agent could be examined with classical theory-of-mind paradigms. We used an indirect speech understanding task, the hinting task, a new text version of a False Belief/False Photographs paradigm, and the Strange Stories paradigm. The hinting task is usually used to assess individuals with autism or schizophrenia by requesting them to infer hidden intentions from short conversations involving two characters. In a first experiment, ChatGPT 3.5 exhibits quite limited performances on the Hinting task when either original scoring or revised rating scales are used. We introduced slightly modified versions of the hinting task in which either cues about the presence of a communicative intention were added or a specific question about the character's intentions were asked. Only the latter demonstrated enhanced performances. No dissociation between the conditions was found. The Strange Stories were associated with correct performances but we could not be sure that the algorithm had no prior knowledge of the test. In the second experiment, the most recent version of ChatGPT (4-0314) exhibited better performances in the Hinting task, although they did not match the average scores of healthy subjects. In addition, the model could solve first and second order False Beliefs tests but failed on items with reference to a physical property like object visibility or more complex inferences. This work offers an illustration of the possible application of psychological constructs and paradigms to a conversational agent of a radically new nature.

Keywords: large language model, ChatGPT, theory-of-mind, indirect speech, False Beliefs

DOI: 10.5281/zenodo.8009748

Version 2. The manuscript has evolved since the first deposit on Zenodo. Title has changed and a second experiment was made. Please now refer to this version.

1 Introduction

Having a theory of mind is conceived as the capacity of an individual to impute mental states to himself or to others (either to conspecifics or to other species as well), such as intentions, beliefs, or knowledge (Premack and Woodruff 1978). In the present work, we address the possibility of applying these concepts and some available paradigms to new artificial intelligence (AI) technologies, a conversational agent based on a Large Language Model (LLM), and highlight their potential contributions in the addressed areas. Historically, the concept was coined to investigate animals like chimpanzees (Premack and Woodruff 1978), and was successful to study early child development (Baillargeon, Scott, and He 2010) and pathological conditions like autism (Baron-Cohen, Leslie, and Frith 1985). To determine whether an individual without language or with non-proficient communication skills has a theory-of-mind (ToM), the preferred approach is to observe his/her behavior when he/she interacts with others (animals or humans) in experimental conditions that require the use of these skills. However, the existence of articulate language and the ability to conduct conversations and process complex requests makes it seemingly easy to detect theory of mind abilities in the individual. Firstly, the use of mental state terms or concepts (i.e., “I believe”, “she thinks”, “he wants”, “they seek to” etc.) can be considered as a proof of the conceptual capacity to attribute volitional or epistemic mental states to another one. Secondly, language is part of a pragmatic context of communication. It has long been conceptualized that our ability to communicate is largely based on the building of a shared knowledge with the interlocutor and that deciphering indirect or metaphorical language requires the understanding of the communicative intention of interlocutors (Sperber and Wilson 1986). Authors have proposed to test the theory of mind through the comprehension of indirect language in the sense that pragmatics could be understood as a sub-module of ToM, however the intricacy of the two constructs does not mean that they should be confused theoretically and empirically (see discussion in Bosco, Tirassa, and Gabbatore 2018).

Cognitive neuropsychologists theorized that patients with schizophrenia suffer from communication and/or social cognition disorders that could be reflected by deficits in the comprehension of indirect speech (Frith 1992; Hardy-Baylé, Sarfati, and Passerieux 2003). Experimental paradigms have been proposed to measure the deficit of patients and to measure their pragmatic skills (Bazin et al. 2005; Langdon et al. 2002; Mazza et al. 2008). Here, we consider the *Hinting Task* introduced by Corcoran et al (1995). This task was designed to test the ability of subjects to infer the real intentions behind indirect speech utterances. The task consists of ten short stories presenting an interaction between two characters in which one character drops a very obvious hint. Allowing to detect deficits in schizophrenia, the task was selected by the SCOPE initiative to create a social cognition assessment battery (Pinkham, Harvey, and Penn 2018). More recently norms in healthy subjects and schizophrenic patients were published with both the original scoring system and a new one proposed to improve psychometric properties by lowering ceiling effects (Klein et al. 2020). In the following study, we will consider both criteria to score a subject’s performance.

Before the creation of the hinting task, the attribution of epistemic mental states, namely beliefs and knowledge, could be tested through the so-called False Belief paradigm also referred to as the *Unexpected Transfer* paradigm (Baron-Cohen, Leslie, and Frith 1985; Wimmer 1983). In its classical version, two dolls, Ann and Sally, are involved in a short sequence in which one of them moves an object after the other doll has left the scene. The task is to figure out that the second doll will look for the object in the place she left it and not where it actually is. This task was regularly posed as a gold standard to attest the presence or deficiency of a first-order theory of mind (i.e., inferences about another's mental state) and different versions have been proposed in order to distinguish a capacity for mentalization from a simple memorization of the world's state at a specific moment of the sequence. However, as Perner and Wimmer discussed, understanding the behavior of others is not based solely on inferring their representations in terms of their true or erroneous beliefs about the world: it is also necessary to take into account what others think about the thoughts of others, the so-called second-order beliefs (1985). The paradigm of false beliefs of the second order is an interesting complexification in developmental psychology because it is supposed to be acquired later, around 5-6 years, in children than the ability to understand the first order, which would be evident between the ages of 3 and 5 (Miller 2009).

Last, we focused in this work on Happé's *Strange Stories* task which is another paradigm allowing to test mental and physical inference skills from short texts. This task was initially developed for studying pragmatics and ToM in children with autism (Happé 1994) and was revised with a control condition not involving mental state attribution (White et al. 2009). It consists of unambiguous situations in which untrue utterances are made by a character for different reasons (i.e., lie, sarcasm, double-bluff, etc.) depending on the context, the emotional content, or the relationship between the characters. The rationale for this task was to provide diverse real-life situations compared to the first- or second-order false-belief tasks, for which autistic patients could possibly present non-ToM strategies to pass the test. Interestingly, such a paradigm highlighted reduced performances of autistic children in the mentalizing conditions (White et al. 2009). Note also that unlike the hinting task, this task does not target a stereotyped conversation situation in pragmatical terms.

In the field of artificial intelligence, the development of conversational agents has recently made striking progress, allowing these models to produce responses that resemble human responses. These agents are disembodied and "purely linguistic" by nature, and trained over extremely large text corpuses essentially from internet databases. If computer linguistic agents have increasingly sophisticated communication capacities, their internal architectures do not necessarily include models of mental states. This raises a fundamental doubt about the ability of these "stochastic parrots" according to the denomination of Bender et al. (2021), to take into account the communicative intent and share knowledge during a verbal exchange, because they have not been trained in such a situation: "Text generated by a LM [Language Model] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind.

It can't have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that."

In what follows, we investigate whether methods of assessments of pathological ToM in humans can be applied to a conversational agent. Usability of psychological constructs, like the Big Five Personality Model, with LLM was discussed by Pellert et al. who argued for adopting a new psychometric perspective (2022). We chose to focus on the recently proposed ChatGPT model¹, which is the state of the art in this field. It is based on multiple steps fine-tuning of a transformer-based architecture with reinforcement learning for human feedback and the use of large-scale conversation datasets (see reviews of generative models in (see reviews of generative models in Gozalo-Brihueza and Garrido-Merchan 2023)). This model brought a lot of attention because it has extremely impressive skills to produce sound and well-formulated answers to a wide range of questions: its similarities with human experts has recently been investigated (Guo et al. 2023 p. 202). The use of deep learning techniques based on training over massive databases, makes it particularly complicated to answer the question of the existence, even implicit, of mental concepts represented within very complex, poorly interpretable structures. It is interesting to note that the interpretability of deep-learning models, i.e., the way they can be understood by humans, is a complex question with legal, societal and ethical consequences (Carvalho, Pereira, and Cardoso 2019), and that it may be considered as sharing some logics with neuroscientists' objectives of understanding brain computations and discovering neural codes.

Recent studies reported experimental evidence of the manifestation of ToM skills in LLM. Through a conversation with the ChatGPT, Mortensen reported that the chatbot did not consider it had a ToM while it knew some about the concept (2023). Other authors assessed ChatGPT-3 with both a Chinese and an English version of Sally and Ann False Belief paradigm and introduced subtle variations in the phrasing of the test (Dou 2023). They reported quite dramatic differences in performances but acknowledged that ChatGPT could score like a human subject at the test under certain conditions. Yet, the reasons of these discrepancies remain unclear. Kosinski investigated successive versions of this model (GPT-3 "davinci-001", "davinci-002", GPT-3.5 and GPT-4) with a False Belief test (2023). He reported a continuous improvement of the performances in the False Belief test that reached nearly a 95% success rate for GPT-4, yielding similar results than a 7-year-old child.

Building on Kosinski's positive findings of GPT's ToM skills, Ullman proposed to challenge GPT-3.5 with slightly modified versions of ToM tasks in order to find out subtle conditions that elicit failure (2023). The unfavorable conditions included statements about visibility of the objects, uninformative labels placed on the objects, or information given by a trusted person, which can't be taken into account properly by the

¹ OpenAI, ChatGPT: Optimizing language models for dialogue, <https://openai.com/blog/chatgpt/> (November 30, 2022). To access the web-based chat: <https://chat.openai.com/chat>

LLM. Errors in the slightly modified versions of TOM tasks highlighted negligence or incoherence in the hierarchization or the integration of information by GPT-3.5, leading the author to question the existence of a theory of mind in this model. Finally, another study based on the SOCIALIQA database and a progressive test procedure with an increasing number of examples (k-shot probing), brought into light a limited capacity of GPT-3 to process mental states compared with factual questions (Sap et al. 2023).

To further investigate ChatGPT’s skills, we conducted our study like a single clinical case study and tested with several evaluations of ToM abilities, i.e., the Hinting task, False Beliefs paradigm and the Strange Stories. Adult patients with clinical conditions like schizophrenia exhibit impaired performances in the hinting task, which impair their real-life functioning (Pinkham, Harvey, and Penn 2018), therefore we used their scores and that of healthy subjects, taken from the psychometric study of (Klein et al. 2020), to provide a clinically relevant range of values. This enabled us to place ChatGPT’s scores on a standardized scale. Following an exploratory approach, whenever it appeared relevant, we slightly modified the tasks in order to investigate the AI’s skills by measuring performance improvement when specific cues were added. In these cases, we investigated changes in the wording of the questions to see the extent to which their focus on the characters’ intentional states or expectations might have altered the accuracy of the responses. Two experiments were conducted and investigated two successive versions of ChatGPT (i.e., 3.5 and 4).

2 Experiment #1

2.1 Methods

Hinting task. Free online OpenAI’s ChatGPT-3.5 (Dec 15th 2023 version) was interrogated with a series of questions from the hinting task. For a subpart of the following procedure subsequent test sessions were conducted with the Jan 09th and 30th versions. In its original version, this task designed to assess patients with schizophrenia consists of ten situations described by three sentences involving two characters (Corcoran, Mercer, and Frith 1995). In each situation, one protagonist says something that may be indirectly interpreted as a request toward the other character (see an example in Table 1). Invariably, a question is asked “What does the character really mean when he says this?”, in order to test the chatbot or the human subject’s capability to infer a communicative intention. The expected answer is rewarded by two points (HINTING 1 scores). In the absence of a correct answer, a second version (HINTING 2) with an additional cue and a question focusing on the character’s request to his/her interlocutor is proposed (see Table 1). If respondents clarify the subjects’ intention (“George wants Angela to get him or offer to get him a drink.”), they score one-point.

The text responses were rated independently by two of the authors (EBG and NV) according to specific and validated criteria (HINTING 1+2). Between each sentence the chatbot’s history was erased so that no context could interfere. Three scoring sys-

tems were used in order to compare the model’s skills with those of healthy or individuals with schizophrenia, and with the models’ itself. The first one is the original version described by Corcoran et al. (1995), the second one is a revised version from the SCOPE initiative by Klein et al. (Klein et al. 2020). The third scoring system (False/True) corresponds to the number of correct answers that are defined as the correct understanding of the intentional meaning of the sentence, allowing us to compare the scores with the two following conditions, for which no validated scoring system is available.

To investigate ChatGPT’s errors when it deals with intentional situations, two other versions of the sentences were used. These modifications were made in an exploratory approach based on the observed errors in order to reveal the sensitivity of the model to the pragmatic aspects of the questions:

1. Character’s Intention version (INJUNCTION): the three sentences as described above (HINTING 1) are kept and the question from the second part of the task (HINTING 2) is added without the additional cue (see Table 1). This experimental condition thus clarifies the question by indicating that one character expects something from the other. However, no information is added to the context. Thus, performances can be compared with that of HINTING 1.
2. Communicative intention version (COMMUNICATIVE): In this version the question of the original version (HINTING 1) was modified and indicated explicitly that a character likely had a “communicative intention” (see Table 1). This version allowed us to test whether the chatbot could use this concept to urge an inference about the character’s intention. The fact that ChatGPT had some conceptual knowledge about “communicative intention” was tested with a direct interrogation (see Supplementary Material 1.1).

False beliefs task. The version proposed here of the False Belief task was designed specifically to test the capabilities of ChatGPT. The test was performed twice to check the answer’s coherence. It involves a description of a classic initial situation with two protagonists in a kitchen, Alan and Jenny, the latter placing her chocolate in the cupboard. The object transfer sequence was made more complex than in the classical versions and was declined in several versions (two False Beliefs situations) in order to make sure that the model does not respond randomly based on the probability of the usual presence of chocolates in cupboards.

The question is asked in two ways. Indeed, it appeared depending on the formulation, ChatGPT could answer very differently to the questions leading either to success or failure:

- In the LOOK condition, the question is about where Jenny will look for the chocolate when she returns (see Table 1).
- In the EXPECT condition, the question is about where Jenny expects to find her chocolate, introducing an interrogation about the mental state (see Table 1).

Strange stories mental and physical task. In this task, the ChatGPT is required to interpret short vignettes and is asked to explain why a character says something that is not literally true (White et al. 2009; Happé 1994). To succeed it has to attribute mental states such as desires, beliefs or intentions, and sometimes higher order mental states such as one character's belief about what another character knows. Scoring system ranges from 0 to 2 points for each story depending on the quality of the interpretation. In addition, two conditions of equal difficulty are presented: MENTAL and PHYSICAL, the former focusing on pragmatics and mental state attributions and the second on physical states only (see Table 1).

1.1 Results

Please note that all the results of the evaluations of this article are available in the Supplementary Material. The data were reported as Google Sheets and processed with Jamovi (www.jamovi.org).

Hinting task. ChatGPT 3.5 performs poorly at the first question of the task (HINTING 1) with a number of correct answers between 1 and 4 on a total of 10 questions. However, this score is largely improved in the second part of the task when a cue and a question about the character's demands is asked (HINTING 2). Indeed, more questions are answered correctly, raising the score to 8 or 9 on 10. Even if this improvement is significant, the use of validated rating systems by two raters shows that the overall performances (HINTING 1+2) remain at 10 or 13/20 (Corcoran system) and 8 to 9/20 (Klein system). Based on the work of (Klein et al. 2020), let's remind that healthy subjects ($n=286$) performed at 17.9/20 ($SD=2$) and 16/20 ($SD=2.5$), with both scoring systems respectively, and that the patients with chronic schizophrenia ($n=375$) scored 15.7/20 ($SD=3.4$) and 13.7/20 ($SD=3.4$), at their first evaluation. In all cases, ChatGPT performed worse than human subjects with z-scores at -2.4 and -2.8 for Corcoran and Klein's scores respectively (those values correspond to the best performance of ChatGPT measured in the present study compared with the normal distribution from the aforementioned article).

Compared to the HINTING 2 condition, INJUNCTION allowed ChatGPT to give 6 or 8 correct answers out of 10, while COMMUNICATIVE only reached a score of 4/10. The scores obtained by ChatGPT given the conditions are reported in Table 2.

In all cases and conditions, ChatGPT 3.5 was able to give a detailed answer from one to five sentences. In the majority of the cases, it provided several hypotheses, which often included the correct answer. In some cases, the model qualified its answers by expressing doubts ("It is difficult to say for certain", "It is not clear from the information provided...") or even incertitude ("It's not possible for me to accurately determine the true meaning behind Rebecca's statement."). Answers and quotations are reported in Supplementary Tables.

False beliefs task. The conversational agent exhibits perfect performances at the False Beliefs task in the EXPECT condition (100% of good response, see Table 3). On the contrary the LOOK condition is failed at the exception of one item when the chocolate is returned to its initial place. We scored this answer as zero because the justification was wrong, leading to a total of 0% of good response.

Strange stories mental and physical task. ChatGPT answered quite correctly to both MENTAL (13/16, 81% of good response) and PHYSICAL (12/16, 75% of good response) conditions in the Strange Stories task (see Table 4) when the scoring system described in White et al. (2009) was used.

2.2 Discussion

The first observation of this experiment is the amazing quality of the answers given by ChatGPT-3.5. They are correctly organized, and present a dialectical effort to discuss several hypotheses. Of importance, characters' names and roles are preserved in the answers. Our results concur with Guo et al.'s findings that this model provides "organized" with clear logic" answers which tend to be "long" and "detailed" (2023). However, such answers would not match the natural answers of human subjects unless they were urged to discuss all hypotheses and their probabilities. These rather long and hesitant answers could appear to a clinician as a way to avoid answering, an obsessive indecision or a smokescreen strategy. In some case the model moderates its own conclusions using formulations like "it is not clear that". These utterances give the appearance of the existence of some kind of metacognitive judgment. It also gives to the reader the impression that one preponderant answer strategy is based on a reformulation of the questions that does not require much inference skills. However, in many cases it appears that ChatGPT is able to conclude and even produces some intentional hypotheses as shown below.

Second, the Hinting Task highlighted the poor intentional inferences of ChatGPT-3.5 according to the criteria defined by Corcoran et al. (1995) or Klein et al. (2020). The scores obtained in both the initial and secondary questions do not compete with human performances even in psychopathological conditions. This result could totally exclude ChatGPT-3.5 from being considered as having intentional mental states inferential skills. However, a more careful analysis of the responses shows that when the model receives additional cues and more focused questions about the intention of the characters, its responses are significantly improved. Eight to nine good responses out of 10 were given in the HINTING 2 condition, which is likely to indicate that ChatGPT-3.5 has the ability to make intentional inferences, even if those scores are considered weak with the classical score formula.

Additional experiments with the modified hinting task provide further insight into the factors that can help ChatGPT to respond correctly. First of all, adding (COMMUNICATIVE condition) a cue prior to the question and directing the model to

consider the most likely communicative intention does not seem to improve the responses (score from 2 to 4/10).

A greater gain (scores of 6 or 8/10), although weaker than with the HINTING 2 condition, is found when it is clearly specified that the task requires a statement about what the character wants the other character to do (INJUNCTION). This result suggests that the injunction directs the answer to the mentalist concept and allows the inference of an intentional mental state. In this case, ChatGPT uses “X wants Y to”, “X asks Y to”, or “X would like Y to” to specify the intention. Thus, we find that some indications in addition to the question allow the language model to provide responses that encompass the intentions normally presented in simple life situations. However, it has a heavy tendency to stay “strictly focused on the given question” as noted by Guo et al. (2023).

Regarding the False Belief paradigm in Experiment #1, the results are also contrasted and depend directly on the questions asked. This is in line with Dou’s report of ChatGPT-3’s variable performances in a Ann and Sally False Belief test depending on complex combinations of phrasing and language (2023), and with the fact that alteration of ToM tasks can induce errors of ChatGPT-3.5 (Ullman 2023). We notice that ChatGPT has real capacities to exploit complex sequences of information indicating successive actions and even taking into account representations of a changing reality.

We note in the case of False Beliefs an ability to take into account the knowledge of one of the characters to anticipate what she expects. Jenny will expect to find her chocolate in the cupboard, as that is where she put it before leaving the kitchen. Surprisingly, sometimes emotional elements are given to describe the possible reaction of the character to the location of the object.

Example: ChatGPT-3.5: “She may be surprised or confused to find it in the fridge or on the table”.

Let us note the failure of the model in the LOOK condition. In our case, this type of question seems to induce the necessity for the model to answer to the final position of the object and not to the position anticipated by Jenny. One explanation could be that ChatGPT is trying to answer the literal question: where will Jenny look for the object is a question corresponding to the final location of the object. But ChatGPT does not infer that this question is about Jenny’s expectations (her mistaken beliefs) that will guide her search strategy. However, such interpretation is contradicted by Mortensen’s report of a correct answer in another version of Ann and Sally’s false belief paradigm, ending with a look question (“Where will Sally look for her marble?”) (2023). As indicated above, our paradigm was, on purpose, made more complex, with several successive moves of the chocolate in order to dampen the success of a probability answer. It appears that subtle modifications of the formulation and the situation produce changes that are difficult to anticipate, sometimes allowing ChatGPT to either produce answers whose argumentation indicates that it has taken into account the mental state of the character or answers that are incompatible with a theory of mind.

Finally, the strange stories test also seems to bring positive results regarding the inference capacities of the model. Such a level of performance is unexpected given the complexity of the stories (their length) compared to the simple situations used in the hinting task. It is possible, however, that these longer, more detailed contexts (often eight sentences) that are explicitly dealing with mental states such as the characters' beliefs, urges ChatGPT to use theory of mind concepts. However, as discussed in the limitations of the study, ChatGPT could have been exposed to the data on this task (logics, situations, etc.) during the learning procedure. It is therefore difficult for us to integrate the results of the Strange Stories task in our conclusions. This emphasizes the needs for the creation of large-scale test corpora and the design of theory of mind task generators to test models on unpublished sets of situations or stories.

3 Experiment #2

Taking into account the results from experiment #1 and some of the discrepancies in the ratings, the assessments were conducted again using the up-to-date GPT-4-0314 model through its API. As shown with different paradigms by Ullman (2023) and Dou (2023) simple modifications of the false belief task could greatly alter GPT-3 performance. To extend our investigations with GPT-4 and determine if the main conclusions are replicated, we wanted to see if the new version brought performance improvements to the Hinting Task and to the False Belief Task. We implemented supplementary first and second-order mental states conditions to the tasks to challenge the model with more complicated inferences. We added information about whether the character can or cannot see the object or the action as proposed by Ullman (2023). Finally, we tested whether the implicit attribution of mental states like true or false beliefs could be used to feed deductive reasoning.

3.1 Methods

In Table 5 the different conditions are described. Code is provided in Supplementary Material 2.

The HINTING 1+2 was scored with Klein's method (Corcoran, Mercer, and Frith 1995; Klein et al. 2020) by four independent raters among whom EBG and NV. Each input was tested three times to check the stochastic variability in the answers (Pellert et al. 2022).

The False beliefs was also assessed in a revised version taking into account problems encountered in Experiment #1: for instance, the initial position where Jenny places the chocolate changed in order to check that other answers than "the cupboard" could be given. The false belief task was also revised in its phrasing to assess several possible situations concerning first and second order beliefs varying the number of characters. This allowed us to test the model's flexibility when it infers characters' expectations.

The use of second-order mental states ("X thinks that Y will look in Z") aims at increasing the difficulty of the task according to theory of mind literature. Table 5 describes the conditions and provides examples. In the condition Visibility, we introduce the physical information about the fact that the object is left visible on the table, in order to prevent false belief to be used. In the Same vs. Different places conditions, two characters' beliefs attributions are required implicitly to infer if the character will look for the object at the same or a different place.

3.2 Results

Regarding the Hinting Task ratings, the percent overall agreement between raters was 84.44% and the fixed-marginal Fleiss's kappa reached 0.69 (95% CI = [0.54, 0.85]) (4 raters, 3 levels, 30 cases) corresponding to a substantial agreement. Across the four raters and the three repetitions, the mean ratings of the Hinting Task using Klein's system was 13.4 (SD=1.3, range: [11 15]). These performances are equivalent to those found in adult patients with chronic schizophrenia (z-score=-0.08) but slightly inferior to the healthy controls group (z-score=-1.04) (Klein et al., 2020). Among 120 ratings, the 0, 1, and 2 values frequencies were respectively 2.5%, 60.8% and 36.7%.

Interrater reliability of the False Beliefs ratings (45 items, 4 raters) reached the maximum agreement value of 100%. The conversational agent achieved a maximum performance in 1st and 2nd order False Beliefs tests but failed at the 1st order Visibility and the 2nd order Different Place conditions (see Table 6).

3.3 Discussion

For the Hinting Task and the False Beliefs task, the more recent version of ChatGPT-4 accessed through its API demonstrates a higher level of performance than in Experiment #1. The text responses are also more concise, consisting of one or two sentences, and provide an acceptable answer to the question, allowing an easier evaluation of their correctness.

The Hinting Task elicits clearer answers and is no longer based on contradictory arguments as reported above. When we compare the performance of ChatGPT-4 to the scores distribution of a control subjects group (Klein et al. 2020), we only observe a trend towards lower scores (-2.6 points, z-score=-1.04). Moreover, the model reaches average performances of patients with schizophrenia. Even if performances were improved compared to Experiment #1, the error patterns, with a majority of items scoring 1, could reveal that the additional cue and the request question are still helpful to succeed in the task. In the previous experiment, we argued that the indication to focus on the character's demand could be an incentive to adopt an intentional reading. The results of the Experiment #2 are in line with this hypothesis.

The False Belief task brings a very high level of performance and contrary to experiment #1, the use of the formulation "Where will [the character] look to find the chocolate?" does not seem to cause any difficulty. We also note that several changes of formulation, successive transfers of the object (up to three locations), the increase of the number of characters from two to three are not criteria that systematically penalize the model. Nevertheless, the robustness of the model with regards to these complexity parameters would benefit from being confirmed in other studies.

However, as reported by Ullman, the model does not seem to be able to handle practical physical information such as the notion of visibility of objects (Ullman 2023). Indeed, the only condition of the first-order false beliefs leading to a failure consists in indicating that the supposedly hidden object is placed in evidence in the room, so that the character does not need to use his belief to guide the search. Providing this information was intended to encourage the LLM to consider 1) the physical reality of the scene which requires practical knowledge of people's environment, 2) the consequences of this scenery on the character's visual perception, and 3) the primacy of visual perception in feeding the cognitive system and updating beliefs. It is possible that the learning corpus of this recent version of ChatGPT, that is entirely text-based and does not rely on physical immersion or bodily embodiment, did not allow it to be trained for one or more of these steps.

The order of false belief attribution was similarly investigated with the introduction of a second order formulation such as "Where [first character] thinks that [second character] will look for the chocolate?" While this increment in the ToM orders corresponds to an increase in the complexity of the task, ChatGPT-4 still succeeds. Sap et al. reported a slight decrease of performances of GPT-3-DAVINCI in a second order False Beliefs task by comparison with the first order one, at least when the number of examples in the k-shot language probing procedure varies from 2 to 8 (2023). The only failures are found when the implicit formulation of the Different place condition was used. Here again, this may indicate the sensibility of the model to unusual and complex story structures (Ullman 2023; Dou 2023).

4 General discussion

Detecting and evaluating manifestations of intentional reasoning and more generally of a theory of mind in an individual is a complex task that has given rise to numerous research and methodological proposals. In this paper, we investigate the ability of a recent and sophisticated model of conversational agent, OpenAI's two versions of ChatGPT (3.5 and 4), to use intentional reasoning to understand ambiguous language. To do so, we simply confronted the chatbot with items from well-known tasks that are standard in research on mentalization. If this simple implementation of a single case methodology cannot by itself solve the question of the presence of a form of theory of mind in natural language processing models, the analysis of the results reveals some intriguing and important methodological questions for the future. It should also be

noted that our approach was deliberately observational, as in clinical investigations of single cases. Indeed, we do not use, in our analysis, any knowledge of the information processing mechanisms or computations used by ChatGPT to generate responses. We considered ChatGPT like a human subject. We used simple scoring grids already defined in clinical research to observe the quality of the responses.

As shown by the results, the 3.5 version of ChatGPT does not exhibit spontaneous and reliable use of theory of mind inference to find the most likely interpretation of hints in simple conversations. However, in certain conditions, when the question focused on people's intentions, the conversational model produced correct interpretations with explicit and even well-argued references to mental states and improves its scores. Arguably, some capabilities to “infer” first-order mental states existed but were not to be favored by the model. In addition, some False Belief attribution and complex dynamic representation of the world existed.

An interesting result of the second experiment with ChatGPT-4-0314 is the improved performance and the better quality of the answers. Thus, for both the Hinting task and the false beliefs, the number of correct answers is increased. In particular, the model can handle complex False Belief tasks including a formulation where a second order mental state is questioned. Moreover, its ability to take into account indirect speech, without reaching that of an average-performing adult, reveals a capacity to infer the underlying intentions as soon as the questions are clarified. When the questions do not focus on the character's requests toward another character, the answers do not focus regularly to the hidden intention. The analysis of the errors shows the sensitivity of the model to the test conditions as it is unable to process additional information of a practical nature (the visibility of an object for example), or complex formulations (false beliefs).

5 Limitations

The present study is obviously insufficient to clarify the question of the existence of a theory of mind in a numeric human-like model. The small number of items makes it difficult to draw a conclusion without being able to produce a statistical inference by repeating test trials. A longer procedure would be needed to test the model studied on multiple occasions. We note, however, that the small number of trials corresponds to the usual conditions for the use of tests in clinical practice. It is interesting that the sophistication of ChatGPT allowed us to consider it as a usual human subject.

Let's also note that successive versions of ChatGPT produce changes in measured performances as reported by Kosinski (2023). Based on that, we conducted the assessment on two successive versions (ChatGPT-3.5 December 2022 and January 2023), and we found only slight differences that may be partly related to some stochastic parts in the model. An additional experiment with a newer version ChatGPT-4 in April 2023 brought strong evidence of improved skills. Taken together, the results demonstrate a

difficulty posed by these evolving models in providing comparable data and replication of experiments. It is thus crucial that LLM developers make available a versioning of their models.

Finally, we are not aware of the possible contamination of answers by the succession of questions we asked. Eliminating the question history does not guarantee a perfect reset of the response system (at least if we consider ChatGPT as a human being). Moreover, ChatGPT's large learning base could include the data and knowledge gathered about the tasks we used, due to the overwhelming literature concerning these paradigms. A direct questioning of ChatGPT about the hinting task does not seem to reveal any particular knowledge of the task or its authors while it acknowledges that it could be useful (see Supplementary material 1.2). Concerning the False Belief tests, Kosinski acknowledges that they could have been seen by the model during the training phase (2023), and thus designs a novel test paradigm. In the present study, we also designed from scratch the tests we used. The drawback of this procedure is that our False Beliefs tests lack validation and prior psychometric knowledge. On the contrary, the Happé's Strange Stories task is clearly referred to by ChatGPT as a well-known paradigm used to study social cognition deficits in autism (see Supplementary material 1.3). When investigating ChatGPT's knowledge on these tasks, it provides intriguing answers:

Question: Can you give an example of the Strange Stories task used to study autism?

ChatGPT: Sure! Here's an example of a Strange Stories task scenario: "John is looking for his keys. He thinks they are in the kitchen, but they are not there. He goes to the living room and finds them on the sofa." After presenting this scenario to the participant, the researcher might ask questions like:

Where does John think his keys are?

Where do the keys actually end up being?

Where does John find his keys?

The answers to these questions can provide insight into the participant's ability to understand John's mental states and beliefs, which is an important aspect of theory of mind and social cognition. This information can then be used to better understand an individual's strengths and weaknesses in these areas, and potentially inform intervention plans for individuals with autism.

While not being able to source in the literature the example given by ChatGPT and ensure its accuracy, it is not possible to exclude a deeper knowledge of this test by the model who may have been challenged with it during its training. For the experimenter, if we consider the design of tasks that can be used by both humans and machines, it will be necessary to guarantee their confidentiality, particularly since the AIs may be trained on any available data, even articles in scientific literature describing these same tasks.

6 Conclusion

This study challenged a recently designed conversational agent, ChatGPT, with the hinting task, a classical intention reading task used to assess social cognitive skills in schizophrenia, and with a new False Beliefs test and the Strange Stories task. Adopting a radical, blind to technological considerations about the models' design, cognitive psycho-pathological and clinical perspective, we used this task as in a single case study and proposed revised versions in order to characterize the AI performances. The main finding is that some paradigms classically designed for human study are applicable in conditions quite close to their use in clinical or research settings. The psychological constructs that are attached to them (theory of mind, attribution of beliefs or intentions, communicative intention, etc.) seem to be implementable in research even though the conversational agent being studied is of a radically new nature. However, the results concur with Mortensen's note that "In a way interaction with ChatGPT is unlike any other interaction we know as humans" (2023). These new cognitive agents of an artificial nature are likely to challenge the double-dissociations of performance found in humans that underlie neuropsychological constructs as we know them, which could lead to profound revisions of conceptual frameworks in neuroscience.

7 Ethical statement and conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. They acknowledge that the emergence of LLMs and their deployment by industries or even governments pursuing economic or political goals raises broad ethical questions about their use in academic research. The implementation of moral rules, of mechanisms that mimic human behavior and might deceive people, of content neutrality is a research topic in itself that cannot be left to the designers of these models and should be handled by open, independent and peer-evaluated research, as proposed in the present study.

8 Author contributions

EBG and NV participated in the assessments. EBG wrote the first draft of the manuscript which was corrected and commented by all the authors.

9 Acknowledgments

We are grateful to Clarissa Montgomery and Augustin Chartouny for their kind participation in the assessments. The authors would like to thank deepl.com AI for their suggestions in some of the translations.

Table 1. List of experimental conditions used in Experiment #1.

Experiment	Task	Conditions names	Number of items	Rating system	Associated construct	Rationale
#1	Hinting	HINTING1 and HINTING2	10 + 10	(Klein et al. 2020) + (Corcoran, Mercer, and Frith 1995) + False/true	Attribution of intentions to others, inferring indirect speech meaning	Two step validated task testing pragmatic language understanding in situations where a character produces indirect speech. In reference to ToM skills, the task requires the attribution of hidden and of communicative intentions.
<p><i>Ex. HINTING1: George arrives in Angela's office after a long and hot journey down the motorway. Angela immediately begins to talk about some business ideas. George interrupts Angela saying: "My, my! It was a long, hot journey down that motorway!" What does George really mean when he says this?</i></p> <p><i>Ex. HINTING2: George arrives in Angela's office after a long and hot journey down the motorway. Angela immediately begins to talk about some business ideas. George interrupts Angela saying: "My, my! It was a long, hot journey down that motorway!" George goes on to say: "I'm parched!" What does George want Angela to do?</i></p>						
#1	Hinting	INJUNCTION	10	False/true	idem	The same as HINTING2 but without the additional cue. This allows to test the specific effect of the question focusing on the character's demands.
<p><i>Ex. George arrives in Angela's office after a long and hot journey down the motorway. Angela immediately begins to talk about some business ideas. George interrupts Angela saying: "My, my! It was a long, hot journey down that motorway!" What does George want Angela to do?</i></p>						
#1	Hinting	COMMUNICATIVE	10	False/true	idem	The same as HINTING1 but with a focus on the concept of "communicative intention".
<p><i>Ex. George arrives in Angela's office after a long and hot journey down the motorway. Angela immediately begins to talk about some business ideas. George interrupts Angela saying: "My, my! It was a long, hot journey down that motorway!" Considering George's most likely communicative intention, what does he really mean when he says this?</i></p>						
#1	False Beliefs	LOOK	2	False/true	First order false belief attribution	This test follows the unexpected object transfer paradigm allowing to test the capability to infer a first-order false belief. The question focuses on the character's search strategy.
<p><i>Ex. Alan and Jenny are in the kitchen. Jenny puts her chocolate inside the cupboard. Then, she leaves the kitchen. Alan moves the chocolate to the table, and then he moves it to the fridge, and moves it again to the table. Later, Jenny comes back to the kitchen to eat her chocolate. Where will Jenny look to find it?</i></p>						
#1	False Beliefs	EXPECT	2	False/true	First order false belief attribution	Same as in LOOK, but with the focus on the subject's expectations (i.e., her mental state).
<p><i>Ex. Alan and Jenny are in the kitchen. Jenny puts her chocolate inside the cupboard. Then, she leaves the kitchen. Alan moves the chocolate to the table, and then he moves it to the fridge, and moves it again to the table. Later, Jenny comes back to the kitchen to eat her chocolate. Where will Jenny expect to find it?</i></p>						
#1	Strange stories	MENTAL	16	False/true	Complex social situations understanding	The task depicts various situations requiring the understanding of involving understanding of double bluff, white lie, persuasion, and misunderstanding, etc.
<p><i>Ex. Simon is a big liar. Simon's brother Jim knows this, he knows that Simon never tells the truth! Now yesterday Simon stole Jim's ping-pong paddle, and Jim knows Simon has hidden it somewhere, though he can't find it. He's very cross. So he finds Simon and he says, "Where is my ping-pong paddle? You must have hidden it either in the cupboard or under your bed, because I've looked everywhere else. Where is it, in the cupboard or under your bed?" Simon tells him the paddle is under his bed. Why will Jim look in the cupboard for the paddle?</i></p>						
#1	Strange stories	PHYSICAL	16	False/true	Complex situation understanding	This task is the control task corresponding to MENTAL, not requiring mental state attribution.
<p><i>Ex. Two enemy powers have been at war for a very long time. Each army has won several battles, but now the outcome could go either way. The forces are equally matched. However, the Blue army is stronger than the Yellow army in foot soldiers and artillery. But the Yellow army is stronger than the Blue Army in air power. On the day of the final battle, which will decide the outcome of the war, there is heavy fog over the mountains where the fighting is about to occur. Low-lying clouds hang above the soldiers. By the end of the day the Blue army has won. Why did the Blue army win?</i></p>						

Table 2. Raw scores (scored as true or false) at the hinting task. Answers were obtained and scored at four different epochs. Maximum score is 10.

Conditions	ChatGPT version	dec 15	jan 09	jan 09	jan 30	ChatGPT's maximum score
HINTING 1	dec 15	2	4	2	1	4
HINTING 2	dec 15	8	8	9	8	9
INJONCTI ON	dec 15	6	7	6	8	8
COMMUN ICATIVE	dec 15	4	2	4	4	4

Table 3. ChatGPT scores at the False Belief tasks and the False Photographs task. Note that the tests were performed twice and resulted in the same ratings. Maximum value is 2.

Conditions	jan 09	jan 09
False Beliefs (LOOK) “Where will Jenny look to find it ?”	0	0
False Beliefs (EXPECT) “Where will Jenny expect to find it ?”	2	2

Table 4. ChatGPT scores at both sets of stories from the Strange Stories task. Each condition consists of eight stories, scoring a maximum of two points. Maximum score is 16.

Conditions	ChatGPT version	Scores
Mental stories (MENTAL)	jan 30	13
Physical stories (PHYSICAL)	jan 30	12

Table 5. List of experimental conditions used in Experiment #2.

Experiment	Task	Conditions names	Number of items	Rating system	Associated construct	Rationale
#2	Hinting	HINTING1 and HINTING2	10 + 10	(Klein et al. 2020)	See table #XYZ	See Table 1.
#2	False Beliefs 1 st order	Move vs. No move	2 + 2	False/True	First order ToM	As in the LOOK condition of experiment #1.
<p><i>Ex. Move: Alan and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it inside the fridge. Later, Jenny comes back to the kitchen. Where will Jenny look to find the chocolate?</i></p> <p><i>Ex. No move: Alan and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it inside the cupboard. Later, Jenny comes back to the kitchen. Where will Jenny look to find the chocolate?</i></p>						
#2	False Beliefs 1 st order	Visibility	1	False/True	First order ToM + pragmatic reasoning	As in (Ullman 2023), an additional information about the object's visibility should prevent the character from using his erroneous belief about its location.
<p><i>Ex. Alan and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and puts it on the table which is the first place one can see when entering the room. Later, Jenny comes back to the kitchen. Where will Jenny look to find the chocolate?</i></p>						
#2	False Beliefs 1 st order	Trio	2	False/True	First order ToM	The Move condition is made more complex adding a third character.
<p><i>Ex. Alan, John and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it in the fridge. Then, he leaves the kitchen. Then, John takes the chocolate and hides it in the trashcan. Later, Jenny and Alan come back to the kitchen. Where will Jenny look to find the chocolate?</i></p>						
#2	False Beliefs 1 st order	Trio Different place vs. Same place	1 + 1	False/True	First order ToM	The 1 st order prediction about both characters behavior is used implicitly as an input into a deductive mechanism.
<p><i>Ex. Different place: Alan, John and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it in the fridge. Then, he leaves the kitchen. Then, John takes the chocolate and hides it in the trashcan. Later, Jenny and Alan come back to the kitchen. Will Alan and Jenny look for the chocolate at the same place?</i></p>						
#2	False Beliefs 1 st order	Trio	4	False/True	Second order ToM	These conditions contrast the second order beliefs of one character about the false beliefs of another character having participated in the successive moves of the object.
<p><i>Ex. Alan, John and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it in the fridge. Then, he leaves the kitchen. Then, John takes the chocolate and hides it in the trashcan. Later, Jenny and Alan come back to the kitchen. Where Alan thinks that Jenny will look for the chocolate?</i></p>						
#2	False Beliefs 2 nd order	Trio different place vs. Same place	1 + 1	False/True	Second order ToM	The 2 nd order prediction about two characters' false beliefs is used implicitly as an input into a deductive mechanism.
<p><i>Ex. Different place: Alan, John and Jenny are in the kitchen. Jenny hides her chocolate inside the cupboard. Then, she leaves the kitchen. Then, Alan takes the chocolate and hides it in the fridge. Then, he leaves the kitchen. Then, John takes the chocolate and hides it in the trashcan. Later, Jenny and Alan come back to the kitchen. Does John think that Alan and Jenny will look for the chocolate at the same place?</i></p>						

Table 6. ChatGPT-4-0314 assessment with the revised version of the False Beliefs test, allowing to test its performance on 1st and 2nd order false beliefs inferences. Mean scores of the four raters and three repetitions are provided.

Order of mental state	Condition	Number of items	Mean Scores
1st Order	Move vs. no move	6	6
	duo_visibility	3	0
	Trio	6	6
	Same vs. Different place	6	6
2nd Order	Trio	12	12
	Trio Different place	3	0
	Trio Same place	3	3

10 References

Baillargeon, Renée, Rose M. Scott, and Zijiang He. 2010. “False-Belief Understanding in Infants.” *Trends in Cognitive Sciences* 14 (3): 110–18. <https://doi.org/10.1016/j.tics.2009.12.006>.

Baron-Cohen, S., A. M. Leslie, and U. Frith. 1985. “Does the Autistic Child Have a ‘Theory of Mind’?” *Cognition* 21 (1): 37–46.

Bazin, N., Y. Sarfati, F. Lefrere, C. Passerieux, and M. C. Hardy-Bayle. 2005. “Scale for the Evaluation of Communication Disorders in Patients with Schizophrenia: A Validation Study.” *Schizophrenia Research* 77 (1): 75–84.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. <https://doi.org/10.1145/3442188.3445922>.

Bosco, Francesca M., Maurizio Tirassa, and Iliaria Gabbatore. 2018. “Why Pragmatics and Theory of Mind Do Not (Completely) Overlap.” *Frontiers in Psychology* 9 (August): 1453. <https://doi.org/10.3389/fpsyg.2018.01453>.

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. 2019. “Machine Learning Interpretability: A Survey on Methods and Metrics.” *Electronics* 8 (8): 832. <https://doi.org/10.3390/electronics8080832>.

Corcoran, R., G. Mercer, and C. D. Frith. 1995. “Schizophrenia, Symptomatology and Social Inference: Investigating ‘Theory of Mind’ in People with Schizophrenia.” *Schizophrenia Research* 17 (1): 5–13.

Dou, Zenan. 2023. “Exploring GPT-3 Model’s Capability in Passing the Sally-Anne Test A Preliminary Study in Two Languages.” Preprint. Open Science Framework. <https://doi.org/10.31219/osf.io/8r3ma>.

Frith, C. D. 1992. *The Cognitive Neuropsychology of Schizophrenia*. Hove, UK: Laurence Erlbaum Associates Publishers.

Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchan. 2023. “ChatGPT Is Not All You Need. A State of the Art Review of Large Generative AI Models.” <https://doi.org/10.48550/ARXIV.2301.04655>.

Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. “How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.” <https://doi.org/10.48550/ARXIV.2301.07597>.

Happé, Francesca G. E. 1994. “An Advanced Test of Theory of Mind: Understanding of Story Characters’ Thoughts and Feelings by Able Autistic, Mentally Hand-

icapped, and Normal Children and Adults.” *Journal of Autism and Developmental Disorders* 24 (2): 129–54. <https://doi.org/10.1007/BF02172093>.

Hardy-Baylé, M. C., Y. Sarfati, and C. Passerieux. 2003. “The Cognitive Basis of Disorganization Symptomatology in Schizophrenia and Its Clinical Correlates: Toward a Pathogenetic Approach to Disorganization.” *Schizophrenia Bulletin* 29 (3): 459–71.

Klein, Hans S., Cassi R. Springfield, Emily Bass, Kelsey Ludwig, David L. Penn, Philip D. Harvey, and Amy E. Pinkham. 2020. “Measuring Mentalizing: A Comparison of Scoring Methods for the Hinting Task.” *International Journal of Methods in Psychiatric Research* 29 (2). <https://doi.org/10.1002/mpr.1827>.

Kosinski, Michal. 2023. “Theory of Mind May Have Spontaneously Emerged in Large Language Models.” arXiv. <http://arxiv.org/abs/2302.02083>.

Langdon, R., M. Coltheart, P. B. Ward, and S. V. Catts. 2002. “Disturbed Communication in Schizophrenia: The Role of Poor Pragmatics and Poor Mind-Reading.” *Psychological Medicine* 32 (7): 1273–84.

Mazza, Monica, Vittorio Di Michele, Rocco Pollice, Massimo Casacchia, and Rita Roncone. 2008. “Pragmatic Language and Theory of Mind Deficits in People with Schizophrenia and Their Relatives.” *Psychopathology* 41 (4): 254–63. <https://doi.org/10.1159/000128324>.

Miller, Scott A. 2009. “Children’s Understanding of Second-Order Mental States.” *Psychological Bulletin* 135 (5): 749–73. <https://doi.org/10.1037/a0016854>.

Mortensen, Ditte. 2023. “ChatGPT Is Constantly Tricked by People, but Who’s Fooling Who?” *Protypypr.io* (blog). February 2023. <https://protypypr.io/post/chatgpt-is-constantly-tricked-by-people-but-whos-fooling-who>.

Pellert, Max, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2022. “AI Psychometrics: Using Psychometric Inventories to Obtain Psychological Profiles of Large Language Models.” Preprint. PsyArXiv. <https://osf.io/jv5dt>.

Perner, Josef, and Heinz Wimmer. 1985. “‘John Thinks That Mary Thinks That...’ Attribution of Second-Order Beliefs by 5- to 10-Year-Old Children.” *Journal of Experimental Child Psychology* 39 (3): 437–71. [https://doi.org/10.1016/0022-0965\(85\)90051-7](https://doi.org/10.1016/0022-0965(85)90051-7).

Pinkham, Amy E, Philip D Harvey, and David L Penn. 2018. “Social Cognition Psychometric Evaluation: Results of the Final Validation Study.” *Schizophrenia Bulletin* 44 (4): 737–48. <https://doi.org/10.1093/schbul/sbx117>.

Premack, D., and G. Woodruff. 1978. “Does the Chimpanzee Have a Theory of Mind?” *The Behavioral and Brain Sciences*. 1: 515–26.

Sap, Maarten, Ronan LeBras, Daniel Fried, and Yejin Choi. 2023. “Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs.” arXiv.

<http://arxiv.org/abs/2210.13312>.

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

Ullman, Tomer. 2023. “Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks.” arXiv. <http://arxiv.org/abs/2302.08399>.

White, Sarah, Elisabeth Hill, Francesca Happé, and Uta Frith. 2009. “Revisiting the Strange Stories: Revealing Mentalizing Impairments in Autism.” *Child Development* 80 (4): 1097–1117. <https://doi.org/10.1111/j.1467-8624.2009.01319.x>.

Wimmer, H. 1983. “Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception.” *Cognition* 13 (1): 103–28. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5).