



# Do conversational agents have a theory of mind? A single case study of ChatGPT with the Hinting, False Beliefs and False Photographs, and Strange Stories paradigms

Eric Brunet-Gouet, Nathan Vidal, Paul Roux

## ► To cite this version:

Eric Brunet-Gouet, Nathan Vidal, Paul Roux. Do conversational agents have a theory of mind? A single case study of ChatGPT with the Hinting, False Beliefs and False Photographs, and Strange Stories paradigms. 2023. hal-03991530v1

**HAL Id: hal-03991530**

**<https://hal.science/hal-03991530v1>**

Preprint submitted on 15 Feb 2023 (v1), last revised 21 Jun 2023 (v2)

**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 4.0 International License

# Do conversational agents have a theory of mind? A single case study of ChatGPT with the Hinting, False Beliefs and False Photographs, and Strange Stories paradigms.

Authors: Eric Brunet-Gouet<sup>1,2\*</sup>, Nathan Vidal<sup>2</sup>, Paul Roux<sup>1,2</sup>

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

<sup>2</sup> Université Paris-Saclay, Université Versailles Saint-Quentin-En-Yvelines, DisAP-DevPsy-CESP, INSERM UMR1018, 94807 Villejuif, France

## \* Correspondence:

Eric Brunet-Gouet

Centre Hospitalier de Versailles, 177 route de Versailles, 78150 Le Chesnay, France

[ebrunet@ght78sud.fr](mailto:ebrunet@ght78sud.fr)

DOI : 10.5281/zenodo.7637476

**Keywords:** conversational agents, ChatGPT, theory-of-mind, indirect speech, False Beliefs

## Abstract

In this short report we consider the possible manifestation of theory-of-mind skills by the recently proposed OpenAI's ChatGPT conversational agent. To tap into these skills, we used an indirect speech understanding task, the hinting task, and 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. Our results show that the artificial model has quite limited performances on the Hinting task when either original scoring or revised SCOPE's rating scales are used. To better understand this limitation, 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. In addition, the use of a False Belief/False Photographs paradigm to assess belief attribution skills demonstrates that ChatGPT keeps track of successive physical states of the world and may refer to a character's erroneous expectations about the world. 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 it. These findings suggest that ChatGPT may answer about a character's intentions or beliefs when the question focuses on these mental states, but does not use such references spontaneously on a regular basis. This may guide AI designers to improve inference models by privileging mental states concepts in order to help chatbots having more natural conversations. This work offers an illustration of the possible application of psychological constructs and paradigms to a cognitive entity of a radically new nature, which

44 leads to a reflection on the experimental methods that should in the future propose evaluation  
45 tools designed to allow the comparison of human performances and strategies with those of the  
46 machine.

# 1 Introduction

Having a theory of mind can be conceived as the capacity to understand and predict the behavior of others or of oneself by using mental concepts such as intentions, beliefs, knowledge, etc. In the present work, we address the possibility of applying these concepts and some available paradigms to new artificial intelligence (AI) technologies, considered as a cognitive agent, 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 if an individual has a theory-of-mind (ToM) it is necessary to analyze his/her behavior, interactions with congeners in experimental conditions that require the use of these skills. These different conditions are associated with either the absence of language or with non-proficient communication skills, and lead to the design of varieties of nonverbal paradigms. 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). Precisely about these notions, authors have proposed to test the theory of mind through the comprehension of indirect language in the sense that it requires to make assumptions about the underlying intentions.

It was shown that patients with schizophrenia suffering from communication and/or social cognition disorders present deficits in the comprehension of indirect speech. Widely replicated and validated experimental paradigms have been proposed to measure the deficit of patients and to measure their pragmatic skills. 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.

Older than the hinting task, the attribution of epistemic mental states, namely beliefs and knowledge, could be tested through the so-called False Belief 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 one had left the scene.

The task is to figure out that the second character 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 actual state at a specific moment in history. The False Photographs task is one of them and includes a description very close to the False Beliefs paradigm but is not based on the erroneous mental state of the character and provides a task structure of similar complexity (Zaitchik 1990). 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 children with autism (Happé 1994).

In the field of artificial intelligence, the development of conversational agents has recently made striking progress allowing these models to produce responses that compete with human responses. These agents are disembodied and "purely linguistic" by nature. If computer linguistic agents have more and more sophisticated communication capacities, their internal architectures do not necessarily include models of mental states. In what follows, we investigate whether methods of assessments of pathological ToM in humans can be applied to a conversational agent. We have chosen to focus on the recently proposed ChatGPT model<sup>1</sup>, 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-Brizuela 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' differences with human experts has been recently questioned (Guo et al. 2023, 202). The use of deep learning techniques based on training over massive databases, makes it particularly complicated to answer the question of the existence, at least implicitly, 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 fact that they may 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 neurosciences' objectives of understanding brain computations and discovering neural codes.

According to Mortensen's investigation of ChatGPT's social skills, the chatbot did not consider it had a ToM while it knew some about the concept (2023). To provide further investigations about its skills, like a single clinical case study, we confronted this artificial agent with several measures of ToM. Using standardized ratings issued from psychopathological research, we report some quantitative measures of ChatGPT performances and some insights to improve them by adding different cues or specific instructions. Following an exploratory approach, when it appeared relevant, we slightly modified the tasks in order to investigate the AI's skills

---

<sup>1</sup> 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>

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 alter the accuracy of the responses.

## 2 Methods

### 2.1 Hinting task

Free online ChatGPT (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.

Example: 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!"

Invariably, a question is asked "What does the character [here George] 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 answers (here "Can I have a drink" and/or "Can I have a few minutes to settle down after my journey before we start talking business") is rewarded by 2 points (HINTING 1 scores).

In the absence of a correct answer, a second version (HINTING 2) with an additional cue is tested:

Example (following): 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?

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 be used. Three scoring systems 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 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 is simply the number of correct answers to the HINTING 1 and the HINTING 2 questions, allowing us to compare the scores with the two conditions described below.

In order 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. 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.

Example: 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?

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". 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 understood the term "communicative intention" was verified with a direct interrogation<sup>2</sup>.

Example: 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?

## 2.2 False beliefs and False photographs task

The version proposed here of the False Belief and False Photograph tasks 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

---

<sup>2</sup> The concept of communicative intention is accurately defined by ChatGPT when asked for:

*Question* (Jan 30th): What is a communicative intention?

*ChatGPT*: Communicative intention refers to the purpose or goal that a speaker has in mind when producing an utterance or message. It refers to the intended message or information that the speaker wants to convey to the listener. Communicative intentions can vary widely, from providing information, making requests, giving advice, making a joke, persuading someone, etc. Understanding the communicative intention behind an utterance is crucial in effective communication and in building social relationships.

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.

Example: 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 ?

- In the EXPECT condition, the question is about where Jenny expects to find her chocolate, introducing an interrogation about the mental state.

Example: 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?

Finally, the task of the False Photographs (PHOTO) highlights an inconsistency between the state of the real world at the end of the story and that described by a photograph taken before Alan's chocolate transfer sequence.

Example: Alan and Jenny are in the kitchen. Jenny puts her chocolate inside the cupboard. Then, she takes a photograph of 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 looks at the photograph she took. Where are the chocolates in the photograph?

## 2.3 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). 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 mental state attributions and the second on physical states only.



Example: (MENTAL) 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?

## 3 Results

Please note that all the results of the evaluations are available in the Supplementary Material.

### 3.1 Hinting task

In all cases and conditions, ChatGPT 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, ChatGPT 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."). These utterances give, at least, 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. The scores obtained by ChatGPT given the conditions and the rating systems are reported in Table 1. Answers and quotations are reported in Supplementary Material.

ChatGPT 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 Corcoran's and Klein's rating systems show that the overall performances (HINTING 1+2) remain low respectively at 10/20 and 7 to 9/20. Let's remember that Corcoran's and Klein's control subjects performed at 17.9/20 (SD=2) and 16/10 (SD=2.5), respectively, and that the patients with schizophrenia scored 15.7/20 (SD=3.4) and 13.7/20 (SD=3.4). In all cases, ChatGPT performed worse than human subjects.

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.

## 3.2 False beliefs and False photographs task

The conversational agent exhibits perfect performances at the False Beliefs task in the EXPECT condition and at the False Photographs task (see Table 2). 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.

## 3.3 Strange stories mental and physical task

ChatGPT answered quite correctly to both MENTAL and PHYSICAL conditions in the Strange Stories task (see Table 3) when the scoring system described in White et al. (2009) was used.

# 4 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 extremely sophisticated model of conversational agent, OpenAI's ChatGPT, 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 we would have done for a human subject. We used simple scoring grids already defined in clinical research to observe the quality of the responses.

The first observation of the experiment is the amazing quality of the answers given by ChatGPT. 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 as a way to avoid answering, an obsessive indecision or a smokescreen strategy.

Second, the quantification of the quality of intentional inferences in the hinting task with the criteria defined by Corcoran et al. (1995) or Klein et al. (2020) demonstrates a weak use of this logic by ChatGPT. The scores obtained in both the initial and secondary questions do not compete with human performances even in psychopathological conditions. This result could

322 totally exclude ChatGPT from being considered as having intentional mental states inferential  
323 skills.

324  
325 However, a more careful analysis of the responses shows that when the model receives  
326 additional cues and more focused questions about the intention of the characters it significantly  
327 improves its responses. For 8 or 9 questions out of 10, in the HINTING 2 set, the correct answer  
328 is provided which is largely favorable even if it only results in a modest score with the classical  
329 score formula.

330  
331 Additional experiments with the modified hinting task provide further insight into the factors  
332 that can help ChatGPT to respond correctly. First of all, adding (COMMUNICATIVE  
333 condition) a cue prior to the question and directing the model to consider the most likely  
334 communicative intention does not seem to improve the responses (score from 2 to 4/10).

335  
336 A greater gain (scores of 6 or 8/10) although intermediate with the HINTING 2 condition is  
337 found when it is clearly specified that the question is about what the character issuing the  
338 indirect message wants their interlocutor to do (INJUNCTION), in the absence of any other  
339 cue. This result suggests that this key question directs the answer to the mentalist concept and  
340 allows the inference of an intentional mental state. In this case, ChatGPT uses “X wants Y to”,  
341 “X asks Y to”, or “X would like Y to” to specify the intention. Thus, we find that some  
342 questions allow the language model to provide responses that encompass the intentions  
343 normally presented in simple life situations. However, it has a heavy tendency to stay “strictly  
344 focused on the given question” as noted by Guo et al. (2023).

345  
346 Regarding the False Belief paradigm, the results are also contrasted and depend directly on the  
347 questions asked. We notice that ChatGPT has real capacities to exploit complex sequences of  
348 information indicating successive actions and even taking into account representations of a  
349 changing reality. The False Photographs task allows the model to demonstrate this clearly  
350 because it brings, in certain answers, logical reasoning about the changes of objects’ state (i.e.  
351 position of the chocolate) contrasting with the static state of the photograph:

352  
353       Example: ChatGPT: “The chocolates would be in the cupboard on the photograph that  
354       Jenny took, as that is where she put them before taking the photograph and Alan moved  
355       them after she took it”).

356  
357 Similarly, we note in the case of False Beliefs an ability to take into account the knowledge of  
358 one of the characters to anticipate what she expects. Jenny will expect to find her chocolate in  
359 the cupboard, as that is where she put it before leaving the kitchen. Surprisingly, sometimes  
360 emotional elements appear concerning the possible reaction of the character in the absence of  
361 the object in the expected place.

362  
363       Example: ChatGPT: “She may be surprised or confused to find it in the fridge or on the  
364       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 it 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 due to impressive performances. 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 8 sentences) and explicitly dealing with mental states such as the characters' beliefs urges ChatGPT to use theory of mind concepts. However, as discussed below, 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.

All these results raise deeper questions in philosophy of mind that go beyond the scope of our article. It appears from the experiments shown here that, under certain prompting conditions, ChatGPT responds by referring to some mental concepts. According to this perspective it would be possible to say that it uses knowledge of intentional logic regularities to produce a verbal description of certain expected behaviors. Again, we don't say that this form of intentional stance (Dennett 1987) is modularized or even explicitly introduced in the design of ChatGPT, by analogy to the question of the neural embedding of theory of mind in humans, i.e. the so-called social brain (Brothers 1990). Non-domain-specific and implicit computational mechanisms may be at work as part of the encoding and decoding mechanisms of the transformer. These have been set up by the lengthy training procedure which brought into light regularities at a lexical and semantic levels, but without a modular implementation of a stable and independent representations of characters' mental states. In that sense, it is unlikely that ChatGPT January 2023 version, while responding appropriately to first-order intentions and beliefs, would be able to generalize this stance to social situations it never met, and be considered as a full theory-of-mind agent.

## 5 Limitations of the study

The present study is obviously insufficient to clarify the question of the existence of a theory of mind in a computer 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 may produce changes in measured performances. Based on that, we conducted the assessment on two successive versions (December 2022 and January 2023), and we found only slight differences that may be partly related to some stochastic parts in the model.

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. 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!<sup>3</sup> 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.<sup>4</sup> 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?

---

<sup>3</sup> *Question* (Jan 06th version): What do you know about the hinting task?

*ChatGPT*: "The hinting task is a task that involves providing hints or clues to help a user or system solve a problem or accomplish a goal. This can include providing suggestions for the next steps in a task, providing additional information to help clarify a question or problem, or providing feedback on the progress of a task. The specific details of a hinting task can vary depending on the application or context in which it is being used.

*Question*: Do you know the hinting task designed to study schizophrenic patients?

*ChatGPT*: I am not aware of any specific hinting task that is specifically designed to study schizophrenia in patients, but it is possible that such a task could be developed for research purposes. Schizophrenia is a complex mental disorder that affects how a person thinks, feels, and behaves. It can cause symptoms such as hallucinations, delusions, and disordered thinking. Research on schizophrenia often involves studying the cognitive processes and brain function of patients with the disorder. A hinting task could potentially be used as a tool to study these processes and to better understand how they are affected by the disorder.

<sup>4</sup> *Question*: What do you know about the Strange Stories task for studying autism?

*ChatGPT*: The Strange Stories task is a well-known task used in research on autism spectrum disorder (ASD). It was first introduced by Simon Baron-Cohen and his colleagues in the early 2000s. The task involves presenting individuals with short written scenarios, referred to as "strange stories", which have unusual or unexpected events. Participants are then asked questions about the stories to assess their ability to understand mental states (e.g. beliefs, desires, intentions) and theory of mind. Studies using the Strange Stories task have shown that individuals with ASD tend to have difficulties understanding mental states and exhibit reduced theory of mind compared to neurotypical individuals. As a result, the task is commonly used as a diagnostic tool for assessing theory of mind and social cognition in individuals with 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 by 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/False Photograph test and Strange Stories. 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 cognitive 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.

As shown by the results, 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 focuses on people's intentions, the conversational model produces 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

exist but seem not to be favored by the model. In addition, some False Belief attribution and complex dynamic representation of the world exists. Yet, sometimes these skills are not manifested and concrete answers without taking into account epistemic mental states are prioritized. In addition, further investigation of higher order mental states attributions (i.e., beliefs about another one's beliefs) should be conducted.

Whatever the technology in use, if ChatGPT was a human patient suffering from poor pragmatic skills, one would recommend training it on corpuses containing numerous dialogues or conversations between two or more characters based on hints and beliefs. In addition, it would be interesting rewarding answers involving two agents and containing mental state terms like intentional (X wants that Y, X wishes that Y, etc.) or epistemic verbs (X thinks that Y, to believe, to know, etc.). Based on such a reinforcement, it could be that the algorithm produces more natural answers and adopts preferentially an intentional stance like humans do.

## 7 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.

## 8 Author contributions

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

## 9 Funding

The Centre Hospitalier de Versailles provided financial support for the publication.

## 10 Acknowledgments

The authors would like to thank deepl.com AI for their suggestions in some of the translations.

## 498 Tables

499 Table 1: ChatGPT scores at the hinting task and modified versions.

Conditions	ChatGPT version	Rating system	Rater 1	Rater 2
HINTING 1+2 trial 1	dec 15	Corcoran rating system (max = 20)	10	10
HINTING 1+2 trial 1	dec 15	Klein rating system (max = 20)	7	9
HINTING 1 trial 1	dec 15	Number of correct answers (max = 10)	2	-
trial 2	jan 09	Number of correct answers (max = 10)	4	-
trial 3	jan 09	Number of correct answers (max = 10)	2	-
trial 4	jan 30	Number of correct answers (max = 10)	1	-
HINTING 2 trial 1	dec 15	Number of correct answers (max = 10)	8	-
trial 2	jan 09	Number of correct answers (max = 10)	8	-
trial 3	jan 09	Number of correct answers (max = 10)	9	-
trial 4	jan 30	Number of correct answers (max = 10)	8	-
INJUNCTION trial 1	dec 15	Number of correct answers (max = 10)	6	-
trial 2	jan 09	Number of correct answers (max = 10)	7	-
trial 3	jan 09	Number of correct answers (max = 10)	6	-
trial 4	jan 30	Number of correct answers (max = 10)	8	-
COMMUNICATIVE trial 1	dec 15	Number of correct answers (max = 10)	4	-
trial 2	jan 09	Number of correct answers (max = 10)	2	-
trial 3	jan 09	Number of correct answers (max = 10)	4	-
trial 4	jan 30	Number of correct answers (max = 10)	4	-

500

501



502 Table 2: ChatGPT scores at the False Belief tasks and the False Photographs task. Note that the  
503 tests were performed twice and resulted in the same ratings.

Conditions	ChatGPT version	Scores
False Beliefs (LOOK) “Where will Jenny look to find it ?” trial 1	jan 09	0 on 2
trial 2	jan 09	0 on 2
False Beliefs (EXPECT) “Where will Jenny expect to find it ?” trial 1	jan 09	2 on 2
trial 2	jan 09	2 on 2
False Photographs (PHOTO) trial 1	jan 09	2 on 2
trial 2	jan 09	2 on 2

504  
505

506 Table 3: ChatGPT scores at both sets of stories from the Strange Stories task. Each condition  
507 consists of eight stories, scoring a maximum of two points.

Conditions	ChatGPT version	Scores
Mental stories (MENTAL)	jan 30	13 on 16
Physical stories (PHYSICAL)	jan 30	12 on 16

508

## 11 References

- Baillargeon, Renée, Rose M. Scott, and Zijing 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.
- Brothers, Leslie. 1990. "The Social Brain: A Project for Integrating Primate Behavior and Neuropsychology in a New Domain." *Concepts in Neuroscience* 1: 27–51.
- 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.
- Dennett, Daniel Clement. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.
- 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 Handicapped, and Normal Children and Adults." *Journal of Autism and Developmental Disorders* 24 (2): 129–54. <https://doi.org/10.1007/BF02172093>.
- 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>.
- Mortensen, Ditte. 2023. "ChatGPT Is Constantly Tricked by People, but Who's Fooling Who?" *Prototypr.io* (blog). February 2023. <https://prototypr.io/post/chatgpt-is-constantly-tricked-by-people-but-whos-fooling-who>.
- 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.
- Sperber, Dan, and Deirdre Wilson. 1986. *Relevance : Communication and Cognition*. Oxford: Basil Blackwell.
- 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).
- Zaitchik, D. 1990. "When Representations Conflict with Reality: The Preschooler's Problem with False Beliefs and 'False' Photographs." *Cognition* 35 (1): 41–68.