UNIVERSITY
OF TRENTO

Decoding Minds

Mentalistic Inference in Autism Spectrum Disorders and ChatGPT Models

Doctoral Candidate: Dalila Albergo

Supervisor: Prof. Stefano Panzeri

Co-supervisor: Prof. Cristina Becchio

Doctoral programme in Cognitive and Brain Sciences

XXXV Cycle

Academic Year 2023-2024

A Simona

e all'eterot(r)opia del girasole

"La bellezza non è nascosta per nessuno."

Albergo, S. (2020). *La percezione dell'esperienza estetica in soggetti non vedenti.*

[Tesi di Laurea, Università di Bologna]

Contents

Abstract

Mentalistic inference, the process of deducing others' mental states from behaviour, is a key element of social interactions, especially when challenges arise. Just by observing an action or listening to a verbal description of it, adults and infants are able to make robust and rapid inferences about an agent's intentions, desires, and beliefs. This thesis considers perspectives from Autism Spectrum Disorders (ASDs) and large language models, specifically GPT models.

Individuals with ASDs struggle to read intentions from movements, but the mechanisms underlying these difficulties remain unknown. In a set of experiments (Chapter 2), we employed combined motion tracking, psychophysics, and computational analyses to examine intention reading in ASDs with single-trial resolution. Single-trial analyses revealed that challenges in intention reading arise from both differences in kinematics between typically developing individuals and those with ASD, and a diminished sensitivity in reading intentions to variations in movement kinematics. This aligns with the idea that internal readout models are tuned to specific action kinematics, supporting the role of sensorimotor processes in shaping cognitive understanding and emphasizing motor resonance, a key aspect of embodied cognition. Targeted trainings may enhance and improve this ability.

In a second set of experiments, we compared Theory of Mind, a core feature of mentalistic inference, in GPT models and a large sample of human participants. We found that GPT models exhibited human-level abilities in detecting indirect requests, false beliefs, and misdirection, but failed on faux pas. Rigorous hypothesis testing enabled us to show that this failure was apparent and was linked to a cautious approach in drawing conclusions rather than from an inference deficit.

Collectively, the results presented in this thesis suggest that the convergence of insights from clinical research and advancements in technology is essential for fostering a more inclusive understanding of mentalistic inferences.

Keywords

Mentalistic Inference, Action Observation, Theory of Mind, Autism Spectrum Disorders, ChatGPT

Synopsis

This dissertation is organized into four distinct chapters.

Chapter 1 introduces the concept of mentalistic inference. Furthermore, this chapter develops the theoretical background underpinning the studies conducted, centring around action observation in individuals with Autism Spectrum Disorders and the examination of Theory of Mind in both human and artificial agents.

Chapters 2 and 3 each present a specific study in detail. These chapters follow the structural format found in the research articles where the studies have been originally published. Within each chapter, the sections provide a comprehensive examination of the methodologies, results, and implications of the respective studies.

Chapter 4 delves into a discussion that synthesizes common themes emerging from the findings presented in Chapters 2 and 3. This chapter acts as a unifying platform, facilitating a deeper understanding of the overarching concepts and connections between the individual studies. Moreover, Chapter 4 offers insights into the broader implications of the research, emphasizing the significance of the collective findings and laying the foundations for future directions.

Chapter 1. General Introduction

1.1. Mentalistic Inference

Mentalistic inference refers to the process of making deductions or drawing conclusions about someone's mental states, such as thoughts, beliefs, intentions, desires, and emotions, based on observable behaviours, verbal expressions, or other available information (Frith & Frith, 1999, 2006; Frith & Wolpert, 2004; Hamlin et al., 2013; Heyes & Frith, 2014). In everyday life, people often engage in mentalistic inference to understand and predict the behaviour of others. For example, if someone is smiling and laughing, we might infer that they are happy. If someone is frowning and avoiding eye contact, we might infer that they are upset or uncomfortable. These inferences are based on our understanding of typical patterns of behaviour associated with certain mental states.

Thus, the study of mentalistic inference is crucial for understanding how humans navigate social interactions and interpret the actions of others (Baron-Cohen et al., 2013; Becchio et al., 2012; Frith & Frith, 2006). It is also relevant in fields such as artificial intelligence, where researchers aim to develop machines that can understand and respond to human mental states (Brooks & Szafir, 2019; El Kaliouby & Robinson, 2004).

Accurately inferring someone's mental state can be challenging, as individuals may express their thoughts, emotions and intentions in diverse ways, and context plays a significant role in interpretation. To provide a more detailed example, imagine strolling along a sidewalk where the simple act of waving a hand can take on different interpretations. In this context, someone observing the gesture might perceive it not as an attempt to hail a taxi but rather as a friendly greeting directed towards a taxi driver. The nuances of social cues in such situations

highlight the complexity of human communication and the potential for diverse interpretations based on contextual factors (Figure 1).

Figure 1. Example of interaction failure in interpreting a social cue (generated by Dall-E 3; adapted from cartoon by Bestie, unknown year).

Misunderstandings may occur due to differences in perceived mental states, with observers lacking insight into the waver's intentions. Factors like (i) misinterpreting hand gesture kinematics or (ii) assuming a friendly wave is an attempt to hail a taxi may contribute to interaction failures. This highlights how ingrained patterns impact decoding actions, potentially leading to miscommunication when context deviates from expectations and emphases the need for context-aware interpretation, in cases in which movement kinematics is not informative (Koul et al., 2019), to avoid communication breakdowns.

The prior example, where misinterpretations might arise from difficulties in perceiving kinematic patterns or mental states, may provide a context for our investigation.

Our studies seek to contribute insights into the nuances of human social interaction, exploring mentalistic inference in two different forms: intention reading from movement kinematics and the application of Theory of Mind principles. In the first study, we examined how observers can infer others' intentions based on the variations in their movement kinematics; in the second study, we examined how agents can infer others' mental states in brief stories presented as written input.

1.2. Studies included in the Dissertation

The studies presented in this dissertation explore two aspects of mentalistic inference. The first study (Chapter 2) investigates intention reading from movement kinematics in children with Autism Spectrum Disorders (ASD) and typically developing (TD) children. The second study (Chapter 3) investigates Theory of Mind in GPT models and human participants.

1.2.1. Action observation in Autism Spectrum Disorders

ASD is traditionally characterized by three core symptoms: (i) impairments in verbal and nonverbal communication, (ii) restricted and repetitive interests, and (iii) deficits in social interaction across multiple contexts (American Psychiatric Association, 2013). However, frequently observed clinical manifestations, not encompassed in diagnostic criteria, pertain to movement - both in terms of executing movements and deriving meanings from others' actions (Gowen & Hamilton, 2013; Kilroy et al., 2019).

Motor difficulties and atypical movement patterns are commonly observed in ASD at different levels, including motor coordination and motor planning. Motor coordination may be impaired both at the level of fine (e.g., manual dexterity) and gross motor skills (e.g., balance and walking; Lidstone et al., 2020; Stins & Emck, 2018; Valagussa et al., 2018). Difficulties in motor planning, as the process of transforming a present state into a series of motor commands, include organizing motor knowledge and longer reaction times when planning movements (Gowen & Hamilton, 2013). Interestingly, a growing body of research is investigating the connection between motor skills and social communication abilities within the ASD population, leading to defining ASD as *pathology of intersubjectivity* (Fuchs, 2015). Clinical evidence suggests that inadequate motor skills are linked to heightened challenges in social communication among children with ASD (Dziuk et al., 2007; Green et al., 2009; Hirata et al., 2014).

Additionally, children with ASD encounter significant difficulties in grasping others' intentions when relying solely on motor cues (Boria et al., 2009; Castelli et al., 2002; Cossu et al., 2012; Edey et al., 2016). The accurate perception and interpretation of other people's mental states are pivotal in social interaction, and proficiency in understanding the purpose behind others' actions, and consequently responding appropriately, establishes the foundation for social reciprocity. The challenges in social interaction experienced by individuals with ASD may be associated with difficulties in processing and comprehending biological motion information, a deficit that appears to manifest early in their developmental stages (Kaiser et al., 2010; McPartland et al., 2011; Pavlova, 2012).

A recent study delved into the processes of mental inference, particularly within the realms of Theory of Mind (ToM) and the discernment of deceptive intentions from body movements and prior knowledge about mental states (Cristiano et al., 2023). Participants viewed videos of object-lifting actions with truthful or deceptive intents, accompanied by mentalistic priors. Transcranial magnetic stimulation probed the ToM network's role in mental inference. Results showed heightened sensitivity to negative priors, emphasizing mental inference's broader role in integrating short-term mental states with behaviour, especially in discerning deceptive intentions. The study also highlighted mental inference's impact on processing truthful actions in a social and mentalistic context, contributing to understanding others' intentions. These findings deepen comprehension of mental inference's intricate involvement in social interactions, emphasizing its role in interpreting intentions.

However, the underlying mechanisms of these impairments remain unclear: are they related to a general difficulty that children with ASD face in reading others' intentions through movement kinematics, or are they related to dissimilarities in movement kinematics between typically developing and children with ASD? Our study aimed at disentangling these two hypotheses by combining motion tracking, psychophysics, and single-trial computational analyses.

1.2.2. Theory of Mind in Human and Artificial Agents

Theory of Mind (ToM) can be described as the capacity to ascribe mental states to the self and others, understanding that others hold beliefs, intentions, desires that may diverge from one's own, and recognizing that these can explain their actions and behaviours (Frith & Frith, 1999; Premack & Woodruff, 1978).

An impaired ability to attribute mental states to others is proposed to stem from a flawed development of ToM modules (Frith, 1989; Karmiloff-Smith et al., 1995). The concept of ToM has been developed especially as a core cognitive feature in ASD, which seem to manifest early, potentially emerging by the end of the first year of life, especially when considering joint attention deficits. Moreover, these deficits appear to be pervasive, demonstrating universality when assessed at the appropriate developmental stage or when using sensitive, age-appropriate tests, even in older, high-functioning individuals (Baron-Cohen, 2001).

Various facets of ToM have been examined and delineated in both the typical and atypical development of children. Some of them involve simple tasks, including distinguishing between mental and physical states, understanding the brain's functions, making distinction between appearance and reality (Baron-Cohen, 1989), and recognizing mental state words and producing them in spontaneous speech (Baron-Cohen, 1989; Tager-Flusberg, 1992). Others involve more complex tasks, for example spontaneously producing pretend play (e.g., Leslie, 1987; Lewis & Boucher, 1988), understanding beliefs as causes of emotions (Baron-Cohen et al., 1993), inferring others' mental states from their gaze (e.g., Leekam et al., 1997), monitoring one's own intentions (Phillips et al., 1998), produce and understand deception (e.g., Sodian & Frith, 1992) and imagining and drawing non existing objects (Scott & Baron-Cohen, 1996).

The wealth of ToM research has led to the development of numerous measurements. These include the Hinting Task, False Belief tasks, the Faux Pas test, Strange Stories, and Irony

comprehension tests. In Hinting Tasks, participants are typically presented with scenarios or situations where one person needs to convey information to another person without explicitly stating it. Instead, they provide hints or indirect cues, and the participant's task is to interpret these hints and grasp the intended message (Corcoran, 2003). False Belief tasks investigates the awareness that distinct individuals may hold varying thoughts about constant situations and are mentioned as first-order if they solely entail deducing the mental state of one individual or second-order if take into account embedded mental states (Baron-Cohen et al., 1985). Faux Pas tests pertains to pragmatics and escribes a situation where a character unintentionally makes a remark that offends the listener due to the speaker's lack of knowledge or failure to recall crucial information (Baron-Cohen et al., 1999). The Strange Stories provide a method for assessing more sophisticated mentalistic inference skills, including the ability to reason about misdirection, manipulation, deception, and misunderstandings, as well as second- or higher-order mental states. These stories are well-suited for evaluating the advanced cognitive abilities of higher-functioning children and adults (Happé, 1994; White et al., 2009). Finally, Irony comprehension tests refer to figurative speech, which necessitates an understanding of the speaker's intentions beyond a literal interpretation of words (e.g., Winner et al., 1998).

Understanding and developing a ToM is crucial for both human social interactions and the creation of socially intelligent artificial agents (Cuzzolin et al., 2020). Artificial agents, similar to young children learning from interactions, need to acquire ToM capabilities to navigate diverse social situations effectively. Constructing an artificial ToM involves either developing agent models or allowing learning through interactions (Hofstede, 2019). However, existing approaches often overlook internal mental states, hindering artificial agents from perceiving and responding to social signals effectively. Current agent-modelling methods focus on behaviours reproduction but lack consideration for internal states, hindering their ability to interpret verbal and non-verbal cues in human interactions (Oguntola et al., 2021). Robust assessment methods are crucial to

evaluate artificial ToM accuracy, potentially drawing inspiration from human social intelligence assessments and requiring nuanced evaluations of artificial ToM development (Zadeh et al., 2019).

In our study, to investigate ToM in GPT models, we adopted a comprehensive battery of psychological tests, ranging from less cognitively demanding tasks like understanding indirect requests, to more demanding tasks like recognizing complex mental states such as misdirection or irony. Each model was exposed to multiple items on each test across independent sessions, and their performance was compared to a large sample of human participants. This approach helped us to better explore the variability of GPT models' skills in social reasoning.

Chapter 2. Intersecting kinematic encoding and readout of intention in autism[1]

2.1. Introduction

The ability to intuit what others are thinking or wanting from observing their behaviour - mind reading - is key to social interaction. Much like print reading, mind reading involves the derivation of meaning from signs (Heyes & Frith, 2014). In print reading, the signs are marks on paper. In mind reading, the signs are movement traces (Becchio et al., 2018). Individuals with autism spectrum disorders (ASDs) have difficulty inferring the mental states of others, including their intentions, from their body movements (e.g., Boria et al., 2009; Castelli et al., 2002; Edey et al., 2016). However, the computational bases of these difficulties are unknown.

One proposal is that such difficulties reflect a specific deficit in mind reading (Heyes & Frith, 2014). Typically developing (TD) observers read intention by extracting and processing subtle intention-related kinematic variations (about 3% of the total variance) out of trial-to-trial variations unrelated to intention (Patri et al., 2020). Individuals with ASD would have difficulty reading intention, possibly due to an overall lower sensitivity to single-trial kinematics or to a deficit in identifying or processing intention-informative variations in single-trial movement

[1] This chapter was published in 2022 by Montobbio N., Cavallo A., Albergo D., Ansuini C., Battaglia F., Podda J., Nobili L., Panzeri S., Becchio C. as a research article on the Proceedings of the National Academy of Sciences (PNAS). It is retrievable at https://doi.org/10.1073/pnas.2114648119.
To enhance clarity and logical coherence of this dissertation, the paragraphs' order and the figures' numbers have been changed without altering the content of the published paper.

kinematics. This accords with the view that individuals with ASD have difficulty sampling relevant and irrelevant variability (Van de Cruys et al., 2014) and therefore, get lost in incidental, trial-to-trial variations (Lawson et al., 2017). This hypothesis predicts a general impairment in intention reading in autism.

Alternatively, difficulties in attributing intentions to actions could be rooted in kinematic (dis-) similarities between typical and autistic kinematics (Cook, 2016). This hypothesis is based on the view that the same internal models used during action execution serve as the basis for action perception, prediction, and inference during action observation (Hommel et al., 2001; Wolpert et al., 2003). Because individuals with ASD move differently compared with TD individuals - in particular, they differ in the way they prospectively control their intentional actions (Cavallo et al., 2018, 2021; Cook et al., 2013) - this hypothesis makes the distinctive prediction that observers with ASD, with autistic internal models, should be less accurate in predicting the actions performed by TD individuals relative to those performed by individuals with ASD. Conversely, TD observers, with typical models, should be less accurate in predicting the actions performed by individuals with ASD relative to those performed by TD individuals. From this perspective, ASD difficulties would not reflect an individual intention reading failure but rather, would arise from reciprocal difficulties in social interaction (Cook, 2016; Schilbach, 2016).

Previous work has shown a TD advantage for TD actions (Casartelli et al., 2020; Edey et al., 2016) but no ASD advantage for ASD actions (Edey et al., 2016). This has been interpreted as evidence that TD observers' models are tuned to typical actions, whereas ASD observers' models are tuned (or possibly untuned) comparably with both TD and ASD actions (Edey et al., 2016). However, the advantage TD observers show for TD actions is not necessarily indicative of kinematic similarity and might instead reflect the higher informativeness of TD kinematics relative to ASD kinematics: that is, the fact that TD actions encode more intention information compared with ASD actions (Cavallo et al., 2018). Conversely, the lack of advantage of ASD observers for

ASD actions might reflect the lower informativeness and higher variability of ASD kinematics relative to TD kinematics (Cavallo et al., 2021; Edey et al., 2016). Thus, previous studies cannot rule out the possibility that group differences in intention reading relate to differences in how intention information is encoded in TD and ASD kinematics. Moreover, because intention reading was computed as the average response across individual trials with variable kinematics, these studies cannot determine the readout computations that inform intention inferences in TD and ASD observers: what information TD and ASD observers read in TD and ASD kinematics and how.

This study aimed to move beyond these limitations by combining accurate recording of movement kinematics and psychophysical measures of intention discrimination with a specifically designed analytic framework. This framework allowed us to link kinematic encoding - how intention information is encoded in TD and ASD movement kinematics during action execution - and kinematic readout - how TD and ASD observers read intention information encoded in TD and ASD visual kinematics during action observation - at the single-subject, single-trial level. In a two-by-two factorial design, TD and ASD children observed actions performed by TD and ASD children. Using a kinematic encoding model, we first quantified the intention information in TD and ASD single-movement kinematics and determined the set of kinematic features that encode this information in TD and ASD actions. Then, we developed a kinematic readout model to quantify how and how well TD and ASD observers read the intention information encoded in TD and ASD actions. Finally, adapting methods developed in the studies by Panzeri and Patri (Panzeri et al., 2017; Patri et al., 2020), we examined how kinematic encoding and readout intersect at the single-trial level across observer groups and observed actions. This approach allowed us to move beyond representations averaged over trials and participants and test alternative hypotheses regarding the origin of difficulties in intention reading in ASD.

2.2. Materials and Methods

The research protocol was approved by the local ethics committee (ASL3 Genovese) and complied with the principles of the revised Helsinki Declaration (World Medical Association, 2013). Written informed consent was obtained from the parents of the children prior to participation in the experiment.

2.2.1. Participants

We report the results of 35 ASD children (29 males) without accompanying intellectual impairment and 35 TD children (29 males). Groups were matched for gender, age (TD mean ± SD = 9.8 ± 1.1 y; ASD mean ± SD = 10.2 ± 1.4 y; $t_{(68)} = 1.345$, $p = 0.183$), and full-scale IQ as measured by the Wechsler Scale of Intelligence (Wechsler, 2012; TD mean ± SD = 104 ± 10.2; ASD mean ± SD = 99.5 ± 11.1; $t_{(68)} = 1.749$, $p = 0.085$). Children with ASD were diagnosed according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013). The Autism Diagnostic Observation Schedule (ADOS, Lord et al., 2012) and the Autism Diagnostic Interview-Revised (ADI-R, Lord et al., 1994) were administered by two experienced professionals. Autistic traits were assessed in 19 ASD children and 16 TD children using the SRS (Constantino, 2013) and were more prevalent in ASD compared with TD children (TD mean ± SD = 50.25 ± 10.42; ASD mean ± SD = 83.0 ± 16.05; $p < 0.001$), with some overlap between the two groups in the moderate-range score (Appendix I, Figure S3A). All children had normal or corrected-to-normal vision and were screened for exclusion criteria (pharmacological treatment, epilepsy, and any other neurological and psychiatric conditions). All but three of the children (two in the ASD group and one in the TD group) were right handed according to the Edinburgh Handedness Inventory (Oldfield, 1971).

More comprehensive tables with the clinical and demographic information of both groups can be found in Appendix I (Table S1 and Table S2).

2.2.2. Experimental Design and Procedures

Stimuli. Stimuli were selected from a dataset of 940 grasping actions obtained by recording 20 TD and 20 ASD children performing grasp-to-pour and grasp-to-place actions. For grasp-to-pour trials, a glass (height = 10 cm; diameter = 6.5 cm) was placed 19 cm from the bottle. Participants were instructed to reach for the bottle, lift it, and pour some water into the glass. A co-experimenter refilled the bottle on each trial. For grasp-to-place trials, a box (height = 6 cm; diameter = 10 cm) was placed 19 cm from the bottle. Participants were instructed to reach for the bottle, lift it, and place it in the box. Detailed procedures and the apparatus are described in the study by Cavallo and colleagues (Cavallo et al., 2021). Briefly, reach-to-grasp movements were tracked using a near-infrared camera motion capture system with six optical cameras (frame rate, 100 Hz; Vicon System) and simultaneously filmed from a lateral viewpoint using a video camera fully synchronized with the optical cameras (Vicon Vue; 100 frames/s, resolution 1,280 x 720). As previously described (Cavallo et al., 2021), the child's right hand was outfitted with retroreflective hemispheric markers (6.5 mm in diameter) placed on the metacarpal joint and the tip of the index and the little finger, the trapezium bone and the tip of the thumb, the radial aspect of the wrist, and the hand dorsum. This marker set allowed us to finely track both the distal (hand shape) and proximal (transport) components of the motions. The kinematic data were run through a 6-Hz low-pass Butterworth filter. Based on previous studies investigating prospective action control in TD and ASD children (Cavallo et al., 2018, 2021), we extracted 15 kinematic variables (Appendix I, Figure S2):

- WV defined as the module of the velocity of the wrist marker (millimetres per second);
- WH defined as the z component of the wrist marker (millimetres);
- GA defined as the distance between the marker placed on the thumb tip and the one placed on the tip of the index finger (millimetres);
- TX, TY, and TZ defined as x, y, and z coordinates of the tip of the thumb (millimetres);
- IX, IY, and IZ defined as x, y, and z coordinates of the tip of the index (millimetres);

- FPX, FPY, and FPZ (finger plane) defined as x, y, and z components of the thumb-index plane (i.e., the three-dimensional components of the vector that is orthogonal to the plane; this plane provides information about the abduction/adduction movement of the thumb and index finger independent of the effects of wrist rotation and of finger flexion/extension);
- DPX, DPY, and DPZ (dorsum plane) defined as x, y, and z components of the radius-phalanx plane (this plane provides information about the abduction, adduction, and rotation of the hand dorsum independent of the effects of wrist rotation).

Custom software (MATLAB; MathWorks Inc.) was used to extract the selected variables. Each variable was calculated at intervals of 10% of the movement duration from reach onset to reach offset.

Selection of action stimuli. From the dataset of grasping actions, we selected, for each group, 50 grasping actions (grasp *to place*, n = 25; grasp *to pour*, n = 25) according to the following criteria: 1) minimized within-intention distance (using the metric reported in Cavallo et al., 2016) and 2) mean duration of movements not significantly different between intentions.

Video clips that corresponded to the selected reach-to-grasp actions were used as stimuli for the intention discrimination task. Each video clip began with reach onset and ended with the contact between the hand and the bottle. To allow participants sufficient time to focus on the reach onset, static frames ranging in duration from 160 to 800 ms (in 160-ms increments) were randomly added to the beginning of each video.

Intention discrimination task. Participants were seated in front of a 24-inch computer monitor (resolution 1,280 × 720; 100 Hz) at a viewing distance of 50 cm. The task structure conformed to a one-interval forced choice task with binary choice (*to place* vs. *to pour*). Each trial began with the presentation of a white central fixation cross for 1,000 ms. Then, a video clip showing the reach-to-grasp action was presented. After the video (followed by a waiting window of 80 ms), a screen prompted participants to indicate the action (*to place* or *to pour*) that would

follow the observed grasp (5,000 ms). For half of the participants, the Italian word *mettere* (*to place*) on the left prompted a button press with the index finger of the left hand, and the word *versare* (*to pour*) on the right prompted a button press with the index finger of the right hand. The position of the two words was counterbalanced across participants. Participants completed two sessions in which they observed reach-to-grasp actions performed by TD children and ASD children in counterbalanced order. Each session consisted of 50 experimental trials performed in three blocks (10, 20, and 20 trials), with a 2-min break between each block. Participants received no feedback either during the practice block or during the experimental blocks.

The task was revised and administered by a clinically experienced experimenter. During the experiment, the experimenter positioned herself behind the child. The experimenter was the same for all participants. Participants were introduced to the stimuli and were given both written and verbal instructions. Practice trials (n = 20) were included before the experimental session to familiarize the child with the task and ensure that they had understood the task. Participants were instructed to respond as accurately and quickly as possible during the presentation of the prompt screen. If they did not comply with the instructions (e.g., they responded during video presentation), the experimenter would ask them to repeat practice trials. We verified that the number of participants who repeated the practice did not differ between groups (six in the ASD group and six in the TD group). The proportion of responses during video presentation or within the first 100 ms of the response window was generally low (TD observers, mean ± SEM = 0.021 ± 0.007; ASD observers, mean ± SEM = 0.015 ± 0.004) and did not differ between groups ($t_{(68)}$ = 0.752, p = 0.455). Stimuli presentation, timing, and randomization were controlled using E-prime V2.0 software (Pittsburgh, PA, USA).

Eye movement data. Gaze direction was measured with an infrared eye tracker (SMI RED500; SensoMotoric Instruments). The eye tracker suffered a fatal technical failure before testing was completed; moreover, calibration of the eye tracker was unsuccessful in some

participants. Therefore, eye-tracking data are available for 35 TD participants and 23 ASD participants. For each observer and each trial, we extracted the sequence of spatial position coordinates (scan path). We computed the fraction of time in which the scan path was within the screen in each trial and then averaged this value across trials for each observer. This fraction was overall high (TD observers, mean ± SEM = 0.91 ± 0.01; ASD observers, mean ± SEM = 0.88 ± 0.03) and did not differ between groups ($t_{(56)} = 0.860$, $p = 0.393$).

2.2.3. Quantification and Statistical Analysis

Data preprocessing. Trials for which participants provided a response during the waiting window or within the first 100 ms of the response window were discarded from analyses (< 2% of trials). We verified that the pattern of results and their significance remained similar even when all trials were included.

Mixed effects models to assess statistical differences in intention discrimination performance, response bias, and model performance. We used mixed effects models to assess the significance of differences in intention discrimination performance (Figure 2D), response bias (Appendix I, Figure S1B), and encoding and readout model performance (Figure 3C and Figure 4B, respectively) compared with chance and across observer groups and observed actions. We used logistic regression with single-trial accuracy and response as the dependent variable to assess differences in intention discrimination performance and response bias and linear regression, with the fraction of correct predictions of each video across cross-validation repetitions as the dependent variable, to assess differences in encoding and readout model performance. The chance-level null hypothesis distribution for encoding and readout model performance was created by fitting the model after randomly permuting across trials the observer's choice labels.

To determine the fixed and random effects to include in the model, we applied a model selection procedure that started from the model with the most complex structure to arrive at a

model that included only the significant predictors. We first selected the random effects structure of the model by keeping the full fixed effects structure and using the Bayesian Information Criterion (BIC; Schwarz, 1978). The BIC rewards model fit and penalizes model complexity. We then retained the optimal random effects structure and selected the best fixed effects structure by conducting likelihood ratio tests between models differing only by the presence or absence of one predictor (Agresti, 2007). Model selection results are reported in Appendix I, Table S3. CIs for model coefficients and statistical comparisons for the effects are reported in Appendix I, Table S4. We performed model fitting using the R package *lme4* (Bates et al., 2015). We performed comparisons across levels of the selected models using the *glht* command from the R package *multcomp* (Hothorn et al., 2008). The multcomp package estimates the value and SE of each effect, from which a z value (to calculate two-sided *p*-values) is computed. The results are reported in Appendix I, Table S4 along with CIs for the estimates of the regression coefficients and for the SD of random effects of the selected models, computed using the bootstrap option in the R function *confint*.

Quantifying and assessing the significance of individual task performance. Because there was no significant response bias, we quantified individual performance on the intention discrimination task as the fraction of correct intention choices. For each participant, we assessed the significance of discrimination performance against chance using a binomial test separately for TD observed actions and ASD observed actions (Appendix I, Table S7).

Single-trial kinematic vector. To quantify single-trial kinematics, the 15 kinematic variables of interest were averaged, for each grasping action, over 10 epochs (t), each spanning 10% of the normalized movement time (0% to 10%, 10% to 20%, etc., of the movement duration from reach onset to reach offset). Next, for each trial and epoch, we created a 15-dimensional, single-trial time-dependent kinematic vector, $\vec{K}(t)$, whose entries, for each trial, were the 15 kinematic variables averaged over that time epoch. We used this kinematic vector for all logistic

regressions (see below). We verified that increasing the number of kinematic features by considering the x, y, and z component of all the markers used to compute the kinematic variables of interest (3 x 6 retroreflective markers) did not improve the performance of the kinematic encoding model. This observation held true for both TD and ASD actions and even when using a finer time windowing (25 or 50 movement epochs rather than 10 movement epochs, as in the analyses reported in the main text; $p < 0.02$ for all movement epochs number). These control analyses suggest that our kinematic encoding model provided adequate spatial and temporal resolution to capture intention-related variations in TD and ASD kinematics.

Logistic regression models of kinematic encoding and readout. To determine the dependence of intention (kinematic encoding model) and intention choice (kinematic readout model) on kinematics over time, we used a logistic regression to estimate the single-trial cumulative probability $y(t)$ (i.e., the cumulative evidence) in favour of the intention *to place* as function of the time-dependent kinematic vector in that trial up to time t. Specifically, we modelled $y(t)$ as a sigmoid transformation of the sum of two terms: a linear transformation of the kinematic vector $\vec{K}(t)$, describing the evidence provided by the single-trial kinematic vector at the current time epoch (t) and a drift term, describing the contribution of the cumulated evidence $y(t-1)$ provided by the kinematic vectors up to the previous time epoch $(t-1)$. More precisely, the equation of the logistic model was as follows:

$$P([y(0) = 'to\ pour']) = P([y(0) = 'to\ place']) = \frac{1}{2}$$

$$P\left([y(t) = 'to\ pour'] \,\middle|\, \vec{K}(1), \dots, \vec{K}(t)\right) = \sigma\left(\vec{K}(t) \cdot \overrightarrow{\beta} + w \cdot \left(y(t-1) - \frac{1}{2}\right) + \beta_0\right)$$

$$P\left([y(t) = 'to\ place'] \,\middle|\, \vec{K}(1), \dots, \vec{K}(t)\right) = 1 - P\left([y(t) = 'to\ pour'] \,\middle|\, \vec{K}(1), \dots, \vec{K}(t)\right)$$

where σ is the sigmoid function, $\overrightarrow{\beta}$ is the vector containing the values of the regression coefficients of each kinematic variable, w is a coefficient weighting the accumulation of information over time, and β_0 is a kinematic-independent bias term. The value of $y(t)$ computed at reach offset provides

the final probability of intention (kinematic encoding model) or intention choice (kinematic readout model) associated with the kinematics of the whole trial. In this model, a single regression coefficient is assigned to each variable, meaning that the contribution of each kinematic variable is weighted equally across all time epochs. More complex models with different regression weights assigned to each variable at different epochs as in the study by Patri and colleagues (Patri et al., 2020) yielded no better performance ($p > 0.09$ for all observer groups and observed actions), confirming that, despite its simplicity, our model fit well both intention encoding and readout.

Training logistic regression models. Training and evaluation were performed in a similar manner for encoding and readout models. Each model was trained on a set of 50 trials. We z scored the single-trial kinematic vectors within each model to avoid penalizing predictors with larger value ranges. We trained the models by minimizing the negative binomial log likelihood with L^2 penalty via stochastic gradient descent with adaptive moment estimation (Adam; Kingma & Ba, 2014). The training was marginally improved by a data augmentation scheme based on small random deformations over the time dimension (Appendix I, *Data augmentation procedure for training the logistic regressions* has full details). The parameter λ, which controls the strength of the L^2 regularization term, was set to 0.05 for all models. A cross-validation approach for tuning this hyperparameter was also tested and yielded similar results. The kinematic encoding and readout models were implemented using *Python/PyTorch* (Paszke et al., 2019).

Kinematic encoding model. The kinematic encoding model expressed the probability that a grasping action was performed with the intent *to pour* as a function of the kinematic vector of that action. We trained separate encoding models for TD and ASD actions. We used the encoding model to quantify the intention information encoded in movement kinematics (Figure 3C) and to identify the kinematic variables that carry intention information in TD and ASD movement kinematics (Figure 3D).

Kinematic readout model. The kinematic readout model expressed the probability of intention choice in each trial as a function of the kinematic vector measured in that trial. We trained the readout model separately for each observer in each session.

Evaluation of model performance. We evaluated the performance of the encoding and readout models by repeated fivefold cross-validation (50 random splits; Kim, 2009). We computed the most likely value of Y for each trial by taking the *argmax* over Y of $P(Y|\vec{K})$ in the equation of the logistic model. Model performance was computed as the fraction of correct trials averaged over folds and random splits.

Statistics on the proportion of readers. We used a binomial test to establish whether the number of readers and the fraction of good readers were statistically significant in each group. To assess the significance of differences between observer groups and observed actions in the proportion of readers and in the proportion of good readers, we used a nonparametric permutation test.

Estimate of CIs of model coefficients. For all kinematic encoding and readout models, we obtained estimates and 95% CIs for the regression coefficients from a bootstrap distribution obtained by fitting the models to data randomly sampled with replacement from the original training data.

Classification of individual kinematic variables as informative for encoding or readout. We assessed the informativeness of individual variables (Figure 3D) by testing whether the corresponding encoding coefficients were significantly different from zero. We retained as informative those variables whose encoding coefficients (absolute value) were greater than the 95th percentile of the null hypothesis values obtained when training the kinematic encoding models with permuted trial labels. A similar procedure was used to determine the number of observers who read each variable during action observation (shown in Figure 6B).

Computation of discrimination performance and confidence predicted by the kinematic readout model. In Figure 4C, we used the kinematic readout model to estimate the intention discrimination performance of individual participants. Using the equation of the logistic model, the intention choice predicted as most likely by the readout model was computed for each trial and compared with the actual intention choice. The individual intention discrimination performance was obtained by averaging the probability of correct choice across all trials for a given participant. For the analysis in Figure 4D, we computed the confidence of the single-trial model prediction as the deviation of the estimated probability of *to pour* from chance (0.5).

Computation of overlap and alignment. We computed two indices of intersection between encoding and readout: overlap and alignment. The index quantifying the overlap in kinematic space between encoding and readout was computed by taking the elementwise absolute value of β_{enc} and β_{read} and computing the normalized scalar product of the resulting vectors:

$$overlap(\beta_{enc}, \beta_{read}) = \frac{\langle abs(\beta_{enc}), abs(\beta_{read}) \rangle}{\|\beta_{enc}\| \, \|\beta_{read}\|} \in [0,1]$$

The overlap index measures the amount of weight common to the two vectors, regardless of the sign of the coefficients.

The index quantifying the alignment of encoding and readout in kinematic space was computed as the normalized scalar product between the encoding and readout vectors:

$$alignment(\beta_{enc}, \beta_{read}) = \frac{\langle \beta_{enc}, \beta_{read} \rangle}{\|\beta_{enc}\| \, \|\beta_{read}\|} \in [-1,1]$$

Note that the absolute value of the alignment index is bounded from above by the value of overlap. Alignment values close to zero can be found either with low overlap values (when the variables with nonzero weights differ between encoding and readout) or with high overlap values (when the two models select the same variables with nonzero weights, but the signs of the weights are inconsistent).

In Figure 7 and in Appendix I, Table S6, the statistics of the overlap index and the alignment index were computed over the set of observers who were classified as readers when observing either TD or ASD actions.

Permutation test to assess the significance of overlap and alignment. To assess the significance of the overlap and alignment indices, we compared them with a null hypothesis distribution obtained by recomputing their values after random permutation (n = 10^5 random permutations) of the entries of the encoding vectors.

Conventions for p-values. Appendix I, Table S3-Table S5Table S report details of logistic mixed effects models' statistical tests, linear mixed effects models' statistical tests, and nonparametric permutation tests. Reported *p*-values are two sided and Holm-Bonferroni corrected. In the figures, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ns indicates $p > 0.05$. Following standard notation, asterisks above bars indicate significance of difference from chance of an individual quantity, and asterisks above brackets indicate significance of difference between two quantities.

Statistical significance of correlations. The significance of correlation values was assessed using the *scipy.stats* Python module, with two-sided parametric Student statistics for Pearson correlation and two-sided permutation distribution for Spearman correlation (Best & Roberts, 1975). We used the *SciPy* package (Virtanen et al., 2020). Significance values are shown in Figure 4D.

2.3. Results

Eight- to 13-y-old ASD children (n = 35) and age- and intelligence quotient (IQ)- matched TD children (n = 35) watched a hand reaching for a bottle and judged on the intention of the observed grasp (see *2.2. Materials and Methods*). To capture natural movement variability, we selected 100 representative reach-to-grasp actions (50 ASD actions and 50 TD actions) from a large action dataset obtained by tracking and simultaneously filming TD and ASD children reaching for a bottle with the intent to place or pour (Cavallo et al., 2018). In a within-subjects counterbalanced order, participants watched videos of reach-to-grasp actions performed by TD children and ASD children (Figure 3 A-C and Appendix I, Figure S1A). All statistical comparisons' *p*-values are reported graphically in Figure 2-Figure 7 and numerically in Appendix I, Table S3-Table S5Table S. Effect sizes are reported in Appendix I, Table S3-Table S5Table S.

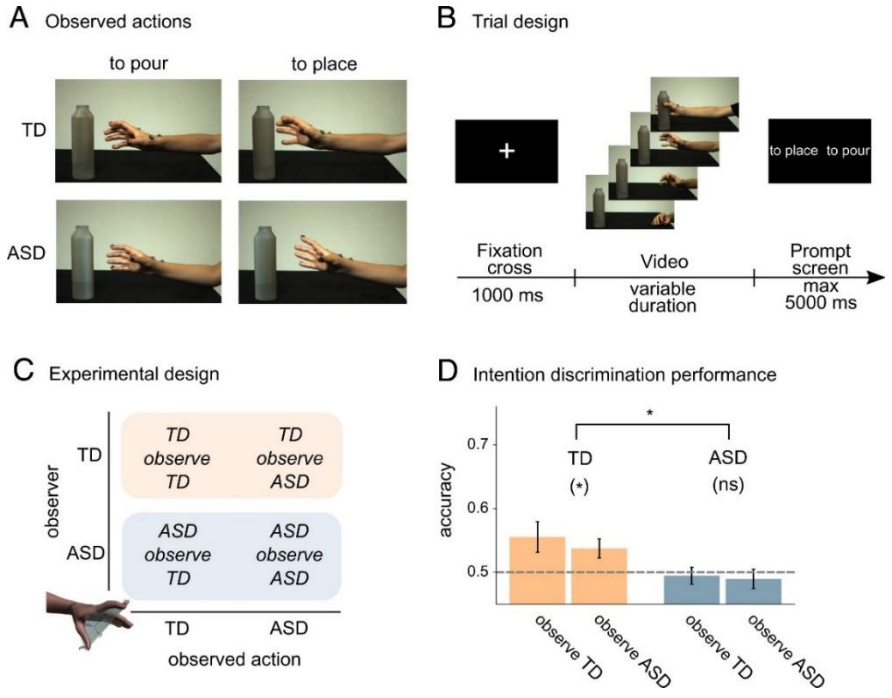

Figure 2. Experimental design and results of intention discrimination. (A) Example video frames of grasp-to-pour and grasp-to-place actions produced by TD and ASD children. Each video began with reach onset and ended with the contact between the hand and the bottle. (B) Trial design of the intention discrimination task. (C) Schematic representation of the experimental design. (D) Trial-averaged intention discrimination performance (fraction correct) for each observer group and observed action. Histograms represent mean ± SEM across participants. ns indicates $p > 0.05$. * $p < 0.05$.

Trial-Averaged Intention Discrimination in TD and ASD Observers. We used logistic mixed effects models to test statistically whether average intention discrimination performance, computed as the fraction of correct intention choices, differed from chance and across observer groups (TD, ASD) and observed actions (TD, ASD). We found a significant main effect of observer group, indicating that ASD observers were poorer at discriminating intention than TD observers (Figure 2D and Appendix I, Table S4). Neither the main effect of observed action nor the interaction between observer group and observed action reached significance (Appendix I, Table S3). Intention discrimination performance was above chance for TD observers but not for ASD observers (Figure 2D and Appendix I, Table S6Table S). Additional analyses conducted to explore the relationship between intention discrimination performance and autistic traits (in a subset of participants for whom autistic trait quantification was available; see *2.2. Materials and Methods*) revealed that TD observers with higher Social Responsiveness Scale (SRS) scores were poorer at discriminating intention than TD observers with lower SRS scores (Appendix I, Figure S3). For both TD and ASD observers, control analyses revealed no effect of IQ on trial-averaged performance (Appendix I, Table S5).

Kinematic Encoding and Readout of Intention Information at the Single-Subject, Single-Trial Level. The above results capture trial-averaged differences between groups. However, they do not quantify what information individual TD and ASD observers read in TD and ASD kinematics and how. To do so, we developed an analytic framework to directly model how information encoded in movement kinematics is read out with single-subject, single-trial resolution. Our formalism was inspired by recent mathematical advances in linking information encoding and readout in a neural population to inform single-trial behaviour choices (Panzeri et al., 2017; Valente et al., 2021). Here we adapted this formalism to investigate how information is coded in movement kinematics (rather than in a neural population).

Kinematic Encoding of Intention Information in TD and ASD Actions. The first step was to determine kinematic encoding: that is, how intention information is encoded in trial-to-trial variations in movement kinematics of TD and ASD actions. Figure 3A shows the temporal profile of two kinematic variables, wrist height (WH) and grip aperture (GA), under the intention *"to pour"* and *"to place"* during TD and ASD reach-to-grasp movements. Each line represents a single reach-to-grasp act. Consistent with previous reports (Cavallo et al., 2016; Patri et al., 2020), individual movement traces showed a large variability across trials and individuals. To isolate the variability that conveys intention information from the trial-to-trial variability unrelated to intention, we developed a single-trial kinematic encoding model based on logistic regression. We represented single-trial kinematics as a time-dependent vector in the multidimensional space of values of 15 intention-sensitive kinematic variables (par. 2.2. Materials and Methods). The kinematic encoding model computed, separately for TD and ASD actions, the probability that a reach-to-grasp movement was performed with a given intention (*to pour*) as a logistic regression of the time-dependent kinematic vector, with a drift term for modelling the accumulation of evidence over time (Figure 3B and par. 2.2. Materials and Methods). Across trials, model performance for TD actions, measured as the fraction of action intentions correctly predicted by the model, was above 95%. For ASD action, model performance was lower but still above 90% (Figure 3C and Appendix I, Table S6). This suggests that, despite the large variability across trials and individuals, both TD and ASD actions exhibited a consistent pattern of intention modulation.

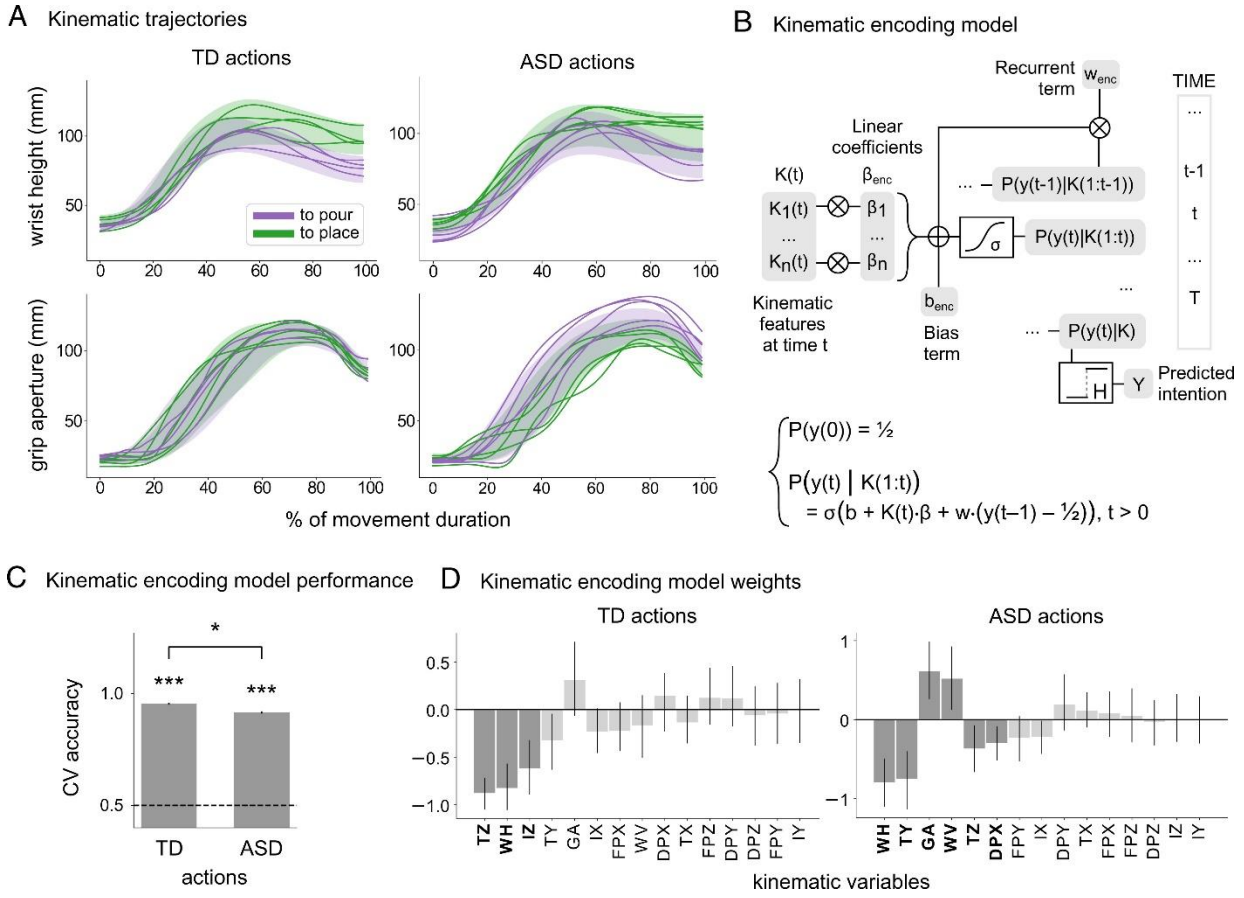

Figure 3. Encoding of intention information in movement kinematics. (A) Time course of WH and GA for reach-to-grasp actions performed by TD and ASD children with the intention *to pour* or *to place*. Coloured curves show representative trajectories for each intention, and coloured areas show one SD across executed trials. (B) Block diagram and equation of the kinematic encoding model used to quantify intention information in movement kinematics. σ is the sigmoid function, β is the vector containing the values of the regression coefficients of each kinematic variable, and w is a coefficient weighting the accumulation of information over time. (C) Cross-validated (CV) performance of kinematic encoding models trained on TD actions and ASD actions quantified as the fraction of trial correctly predicted. Histograms represent mean ± SEM across folds. * $p < 0.05$; *** $p < 0.001$. (D) Contribution (weight) of each kinematic variable to the kinematic encoding of intention in TD and ASD movement kinematics as measured by the regression coefficient of the variable in the logistic regression. A positive (negative) weight is assigned to a variable distributed across trials with higher (lower) values for grasp-to-pour actions compared with grasp-to-place actions. Dark bars indicate variables carrying intention information. Error bars indicate 95% CIs computed by bootstrapping. DPY, y dorsum plane; DPZ, z dorsum plane; FPX, x finger plane; FPY, y finger plane; FPZ, z finger plane; IX, x index; IY, y index; TX, x thumb.

Figure 3D visualizes the contribution (weight) of each kinematic variable to the encoding of intention information in TD and ASD kinematics, as measured by the regression coefficient of the variable in the logistic regression. A positive (negative) encoding weight is assigned to a

variable distributed across trials, with higher (lower) values for grasp-to-pour actions compared with grasp-to-place actions. For example, WH is generally higher for the grasp-to-place action and is, therefore, negatively weighted for both TD and ASD actions (Figure 3A). TD and ASD actions exhibited partially different patterns of intention encoding. For TD actions, intention information was encoded in WH and in the displacement of the thumb (z thumb [TZ]) and index finger (z index [IZ]) along the z axis (Appendix I, Figure S2). For ASD actions, intention information was distributed across a larger set of variables. Some variables carrying intention information in ASD kinematics also carried intention information in TD kinematics (WH, TZ). Other variables informative in ASD kinematics (wrist velocity [WV], GA, y thumb [TY], and x dorsum plane [DPX]) did not carry intention information in TD kinematics. Consistent with previous work demonstrating differences in the way that TD and ASD prospectively control their actions (Cavallo et al., 2021), these results demonstrate differences in the kinematic encoding of *to pour* and *to place* intentions in TD and ASD actions.

Kinematic Readout of Intention Information in TD and ASD Observers. Having determined how intention information is encoded in the single-trial kinematics of TD and ASD actions, we next fitted single-trial intention choices to a kinematic readout model to investigate how TD and ASD observers read such information from observing TD and ASD actions. The kinematic readout model computed the probability of intention choice (*to pour*) in each trial as a logistic regression of the time-dependent kinematic vector for that trial (Figure 4A).

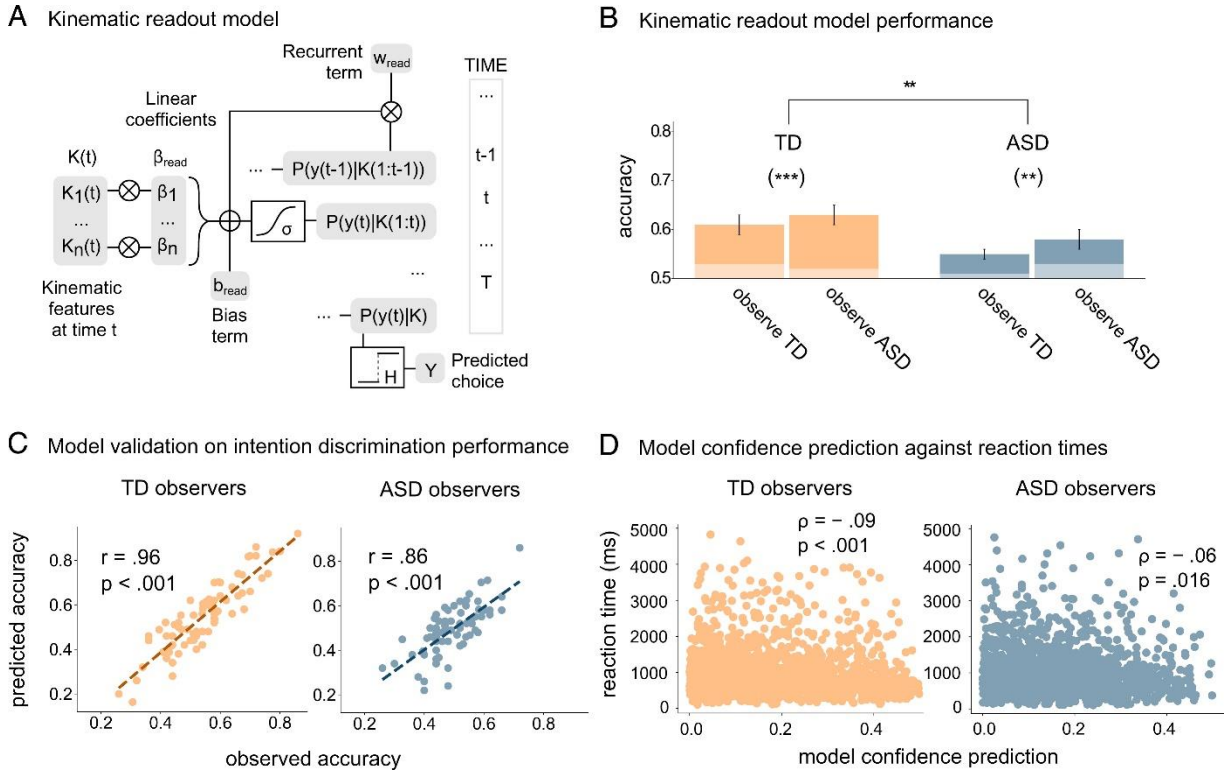

Figure 4. Readout of intention from single-trial kinematics. (A) Block diagram and equation of the model used to quantify kinematic readout of intention. (B) Cross-validated performance of the kinematic readout models quantified as the fraction of correctly predicted intention choices. Histograms represent mean ± SEM across participants for each observer group and observed action. The light sub-bars represent chance-level performance quantified as the mean of the null hypothesis distribution of cross-validated model performance. ** $p < 0.01$; *** $p < 0.001$ the asterisk above brackets shows the significance of the difference between the TD and the ASD group accuracy; the asterisks above bars (in parentheses) show the significance against chance. (C) Scatterplots of the relationship between the observed intention discrimination performance and the performance predicted by the kinematic readout model across individual participants for TD and ASD observers separately. Pearson's correlation coefficients (r) and their significance values (p) are reported. (D) Scatterplots of the relationship between trial-level reaction times and model prediction confidence computed as the deviation of the estimated probability of *to pour* from chance. Spearman's correlation coefficients (ρ) and their significance values (p) are reported.

Across trials and conditions, kinematic readout model performance, measured as the fraction of intention choices correctly predicted by the model, was significantly above chance (Figure 4B and Appendix I, Table S6). The strong correlation between observed intention discrimination performance and performance predicted by the readout model confirmed that our kinematic readout model was able to capture intention discrimination performance at the individual level (Figure 4C). Although reaction times were not used to fit the model parameters,

we also found weak, but significant, negative trial-to-trial relationship between reaction time and model prediction confidence (Figure 4D). This suggests that observers were slightly faster to judge intention on trials that were classified with greater confidence by the model. Taken together, these analyses suggest that our kinematic readout model provided a plausible description of how well observers discriminated intention from single-trial kinematics.

Sensitivity of Intention Readout to Movement Kinematics. Having verified the ability of the kinematic readout model to capture statistical dependencies between intention choices and single-trial variations in movement kinematics, we used it to test the hypothesis that poor intention discrimination in ASD (Figure 2D) reflects an overall reduced sensitivity of intention readout to single-trial variations in visual kinematics. One concrete way to assess this is to measure how well the kinematic readout model predicts single-trial intention choices (regardless of whether the predicted intention choices are correct or incorrect). If intention readout in ASD is not sensitive to single-trial variations in movement kinematics, the kinematic readout model should be at chance in predicting ASD intention choices. As shown in Figure 3B, this was not the case. Sensitivity of intention readout as measured by kinematic readout model performance was lower in ASD compared with TD (Appendix I, Table S4) but still significantly above chance for both observer groups and observed actions (Appendix I, Table S6). This suggests that lower sensitivity of the intention readout to the single-trial kinematics cannot fully account for ASD failure to discriminate intention apparent in Figure 2D.

Identifying Readers. Readout patterns are variable across observers. We next used the kinematic readout model to parse this heterogeneity and identify, at the individual level, observers whose intention readout was sensitive to single-trial variations in movement kinematics (hereinafter readers). To this end, for each observer, we computed the intention choices predicted by the kinematic readout model (regardless of whether the predicted choices were correct or incorrect) and compared the obtained value with a null distribution of randomly permuted choices.

The distribution of readers and non-readers in each group is shown in Figure 5 as a function of intention discrimination performance and readout strength (defined as individual model performance, z scored with the null hypothesis accuracy on randomly permuted choices). For both TD and ASD observed actions, the proportion of readers was higher in the TD group (20 of 35) than in the ASD group (11 of 35). In both groups, the proportion of readers exceeded the proportion expected by chance for both TD actions and ASD actions, with no significant difference between observed actions (Appendix I, Table S7).

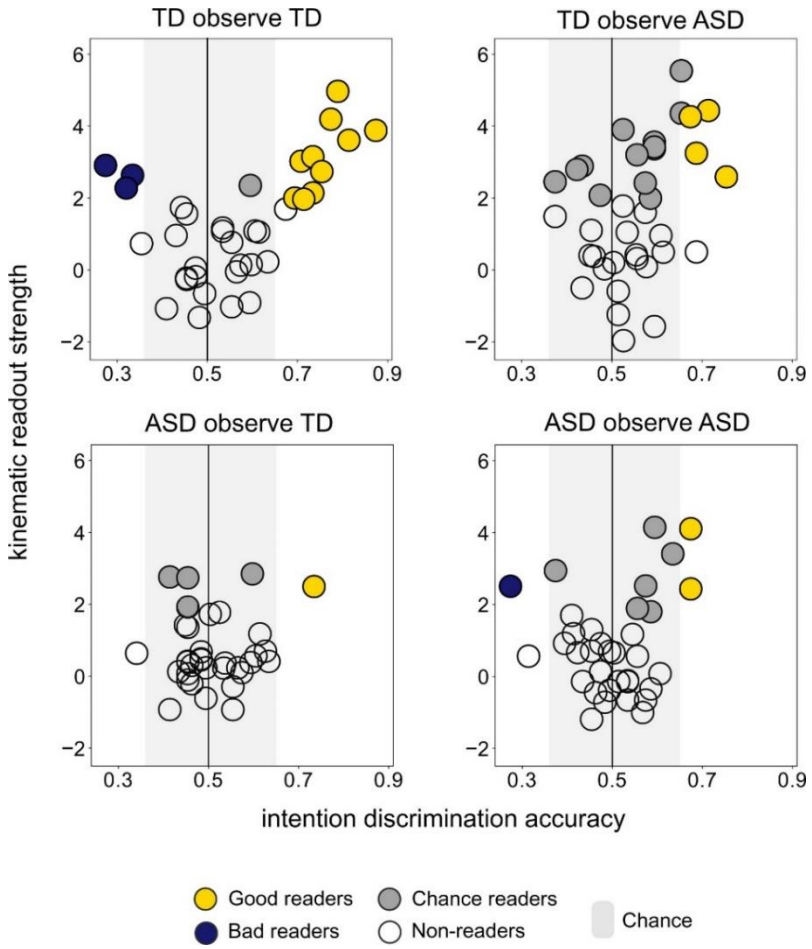

Figure 5. Distribution of readers and non-readers as a function of intention discrimination accuracy and readout strength. Readout strength is quantified for each observer as the z-scored readout model performance. The larger the readout strength, the stronger the sensitivity of intention readout to movement kinematics. Readers (observers whose intention readout is sensitive to single-trial movement kinematics) are represented as filled circles; good readers (intention discrimination accuracy above chance level) are represented as yellow circles, bad readers (intention discrimination accuracy below chance level) are represented as blue circles, and chance readers (intention discrimination accuracy at chance level) are represented as grey circles. Non-readers are shown as open circles.

Readers Good (and Bad) at Reading TD and ASD Actions. The notion of *reader* is agnostic with respect to intention discrimination performance - observers might read movement kinematics (as measured by kinematic readout model performance) and still perform at chance or even below chance. For example, readers would perform at chance if they read variations that do not encode intention information; they would perform below chance if they read variations that encode intention information but do not read the encoded information correctly: for instance, they interpret a decrease in WH, encoding the intention *to pour*, as indicative of *to place*. To look at the relationship between readout and intention discrimination performance, we used a binomial test to stratify readers based on their ability to discriminate intention (Appendix I, Table S7).

As shown in Figure 5, readers with intention discrimination above chance (*good readers*) in the TD group outnumbered good readers in the ASD group. In the TD group, the proportion of good readers was significantly higher for TD actions (10 observers of 14) compared with ASD actions (4 observers of 17). This increase was partially offset by the presence of three TD bad readers for TD actions. For both TD and ASD actions, the proportion of good readers was higher than expected by chance. In the ASD group, the proportion of good readers for both TD (1 observer of 5) and ASD actions (2 observers of 9) did not differ from that expected by chance, with no difference between observed actions (Appendix I, Table S7). Although the small sample of subgroups urges caution in interpretation, these results suggest a predominance of good readers among TD observing TD actions.

Intersecting Kinematic Encoding and Readout. Our results so far reveal differences in the ability to read out intention information across observers and observed actions. However, these analyses do not identify the specific features that are read out, whether readers read informative variables or noninformative variables, and how well they read the encoded information. To address this issue, we examined how specific features were read by TD and ASD readers during observation of TD and ASD actions.

We computed the contribution (weight) of each kinematic variable to the intention readout as the variable regression coefficient in the readout logistic regression. A positive (negative) weight is assigned to a variable distributed across trials with higher (lower) values for the intention choice *to pour* compared with *to place*. We examined, separately for each observer group and observed action, the overlap in the distribution of readout weights relative to encoding weights: whether readout weights were assigned to intention-informative variables. For variables carrying intention information, we also examined whether the signs of the readout weights correctly aligned with the signs of the encoding weights. A positive (negative) readout weight assigned to a variable with a positive (negative) encoding weight would indicate correct alignment; a positive (negative) readout weight assigned to a variable with a negative (positive) encoding weight would indicate incorrect alignment. For example, an increase in WH encodes *to place*, and thus, WH is assigned a negative encoding weight (Figure 3A and D). An incorrectly aligned positive readout weight would incorrectly interpret an increase in WH as signalling *to pour*.

Figure 6A visualizes the overlap and alignment of readout weights relative to encoding weights across kinematic variables (averaged across observers). To provide a complementary visualization of the interindividual reproducibility of readout, Figure 6B shows the number of readers who read a given variable in each condition. For TDs observing TD actions, comparison of the distribution of readout weights relative to encoding weights revealed a near-perfect overlap - the three variables that are read out more and by more observers (WH, TZ, and IZ) are also the three variables that encode intention information in TD kinematics (Figure 6A). Filled bars indicate that the readout weights mostly aligned to the encoding weight correctly. Also, as shown in Figure 6B, most observers correctly interpreted the intention information encoded in these variables. Although IZ does not carry intention information in ASD kinematics (Figure 3D), WH, TZ, and IZ were also the three variables most frequently read by TD observers in ASD actions. Variables such as GA and WV, which encode intention information in ASD kinematics but not in

TD kinematics, were only read out - mostly incorrectly - by a limited fraction of TD observers who observed ASD actions.

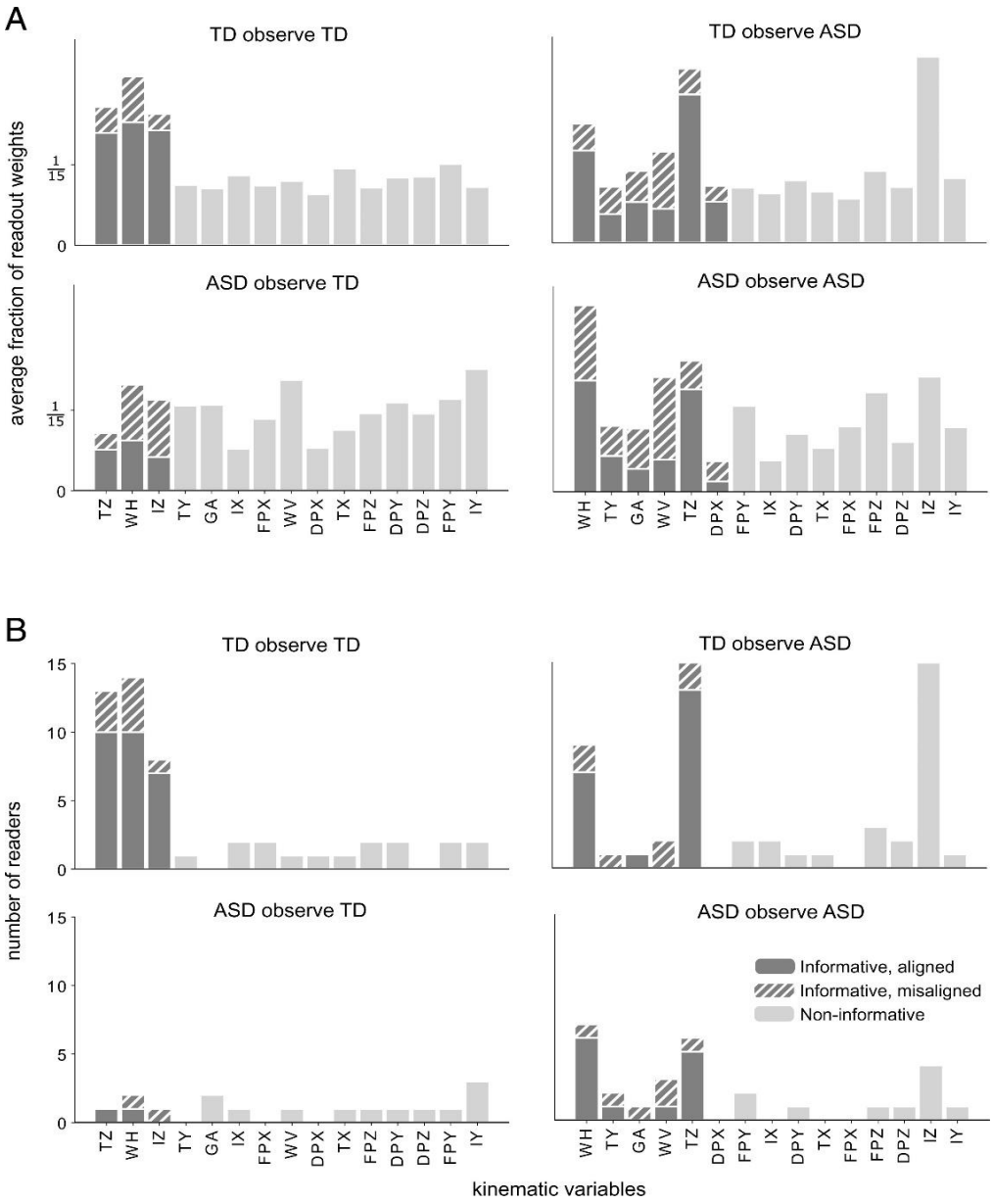

Figure 6. Distribution and sign of readout weights across kinematic variables. (A) Bar graphs of the average fraction of kinematic readout weights across readers for each observer group and observed action. Kinematic variables are ordered from left to right from most informative to least informative. Variables carrying intention information are shown as dark bars. Filled bars indicate correct alignment of kinematic readout weights relative to encoding weights. Striped bars indicate incorrect alignment. (B) Bar plots of the number of readers who read each kinematic variable. A given variable is read by a reader if the corresponding readout weight is significantly different from zero. The order of variables and colour coding are the same as in A. DPY, y dorsum plane; DPZ, z dorsum plane; FPX, x finger plane; FPY, y finger plane; FPZ, z finger plane; IX, x index; IY, y index; TX, x thumb.

For ASD readers, the readout weights showed greater (although not perfect) overlap with the encoding weights during observation of ASD actions compared with TD actions. Specifically, ASD observers consistently read out two variables - WH and TZ - of the six variables encoding intention information in ASD actions. While WH and TZ also carry intention information in TD kinematics, ASD observers assigned little readout weight to these variables (or other informative variables) when observing TD actions. Diagonal striped bars indicate that, regardless of the observed actions (TD vs. ASD), information was misread in most variables.

We computed two indices that quantitatively summarized the above results across all variables. The first index quantified the overlap in the distribution of readout and encoding weights as the normalized scalar product between the absolute values of the encoding and readout vectors. The second index quantified the alignment of the readout weights relative to the encoding weights as the normalized scalar product between the encoding and readout vectors (Figure 7A). A reader who is good at both identifying informative features and interpreting their information would have both high overlap and high alignment. A reader who is good at identifying informative features but not good at interpreting their information would have high overlap but low alignment. Consistent with the intuition conveyed by Figure 6, the results showed a significant overlap in the distribution of readout and encoding weights for TD readers observing TD actions (but not ASD actions) and for ASD readers observing ASD actions (but not TD actions). TD readers showed significant alignment across both TD and ASD actions. In contrast, alignment was not significant for ASD readers for either action (Figure 7B and Appendix I, Table S6).

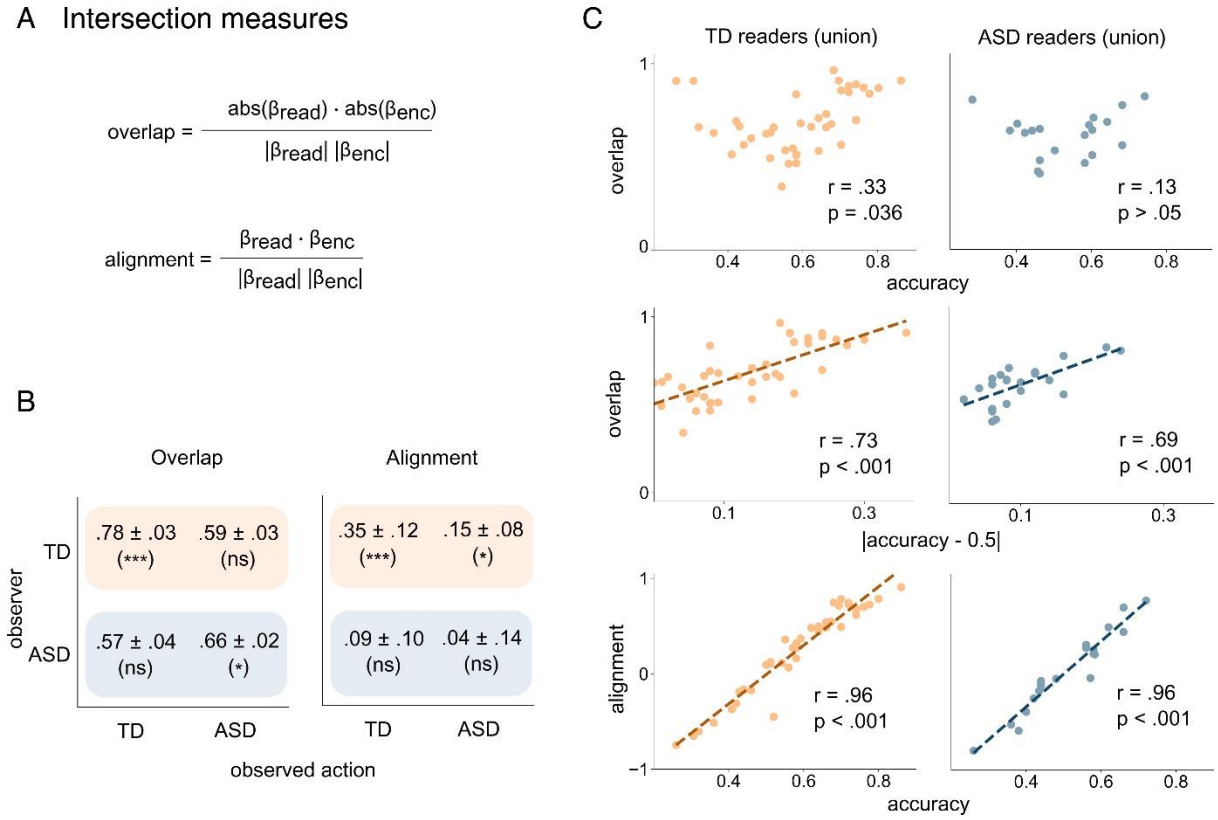

Figure 7. Indices of the intersection between kinematic intention encoding and readout. (A) Overlap and alignment indices. (B) Average values of the overlap and alignment indices, reported as mean ± SEM across readers, for each observer group and observed action. (C) Relationship between overlap and intention discrimination accuracy (Top), overlap and deviation of intention discrimination accuracy from the 0.5-chance level (Middle), and alignment and intention discrimination accuracy (Bottom). Pearson's correlation coefficients (r) and their significance values (p) are reported for significant linear relationships. ns indicates $p > 0.05$. * $p < 0.05$; *** $p < 0.001$.

Figure 7C illustrates the correlation of overlap and alignment with individual intention discrimination performance separately for TD and ASD readers. Overlap did not correlate with individual intention discrimination. However, we found a significant positive correlation between overlap and deviation of individual intention discrimination performance from chance (defined as the absolute value of the difference between task performance and the 0.5-chance level). Alignment correlated positively with individual intention discrimination for both TD and ASD readers. This indicates that, in readers, individual intention discrimination depended not only on the selection of informative features but also, on their correct interpretation. In other words, the

(in-)ability of readers to discriminate intentions was related to their (in-)ability to correctly interpret the intention information extracted from informative features.

2.4. Discussion

Many current perspectives on action reading in autism are based on the quantification of average intention discrimination across repeats of observed actions (Casartelli et al., 2020; Edey et al., 2016). However, kinematics is variable across trials and individuals (Latash, 2012). Trial-averaged analyses may obscure how intention information is encoded and read out in single-trial kinematics. Here, we have developed an analytic approach that enabled us to reveal intention readout computations with single-trial resolution.

By applying this approach, we were able to uncover that single-trial intention choices in ASD systematically reflected trial-to-trial variations in visual kinematics. This is demonstrated by the finding of a lower but still significant sensitivity of intention readout to single-trial kinematics (as measured by kinematic readout model performance) in ASD compared with TD. Corroborating this finding, the proportion of ASD observers who read trial-to-trial variations in movement kinematics (ASD readers, about one-third of ASD observers), although lower than the proportion of TD readers (about two-thirds of TD observers), exceeded the proportion of readers expected by chance for both TD and ASD actions. These findings indicate that while the average intention discrimination in ASD was at chance, single-trial intention choices by a sizeable proportion of individual observers were not random.

A second implication of our results is that for both TD and ASD readers kinematic similarity was important for identifying variations that carry intention-related information. Unlike in print reading, where all marks on paper encode meaning, in mind reading, readers must first extract, from trial-to-trial variations, those variations that encode intention information. Our single-trial results show that TD readers were able to extract such variations during observation of TD actions but not ASD actions. Conversely, ASD readers were able to extract intention-informative variations during the observation of ASD actions but not TD actions. This *same group* advantage is consistent with the principle that internal readout models (or codes) of TD observers

are tuned to typical actions and internal readout models of ASD observers are tuned to autistic actions (Cook, 2016).

What are the exact tuning properties of typical and autistic models? Are internal readout models *feature based*, such that TD (ASD) readers assign more weight to those individual features that encode intention information in TD (ASD) movement kinematics? Or is visual kinematics more likely to be processed as a perceived whole, such that similar to face processing (Calder, 2011). changes in configural information (i.e., relationship between individual features) influence the identification of individual features?

Our kinematic readout model results provide an initial opportunity to answer these questions. If feature identification is integrated into the overall kinematic configuration, then TD intention-informative features should be weighted less when presented in the context of ASD visual kinematics than in TD visual kinematics. Conversely, ASD intention-informative features should be weighted less when presented in the context of TD visual kinematics compared with ASD visual kinematics. Consistent with this prediction, ASD readers weighted less ASD intention-informative features when observing TD actions compared with ASD actions. Configural effects in the TD readout were less clear. In contrast to ASD readers, TD readers appeared to weight TD intention-informative features equally in TD and ASD visual kinematics. In particular, IZ - a feature that carries intention information in TD visual kinematics but not in ASD visual kinematics - was weighted similarly during observation of TD and ASD actions. Combined, these data may indicate a difference in the properties of TD and ASD internal readout models, with ASD internal models being more sensitive to the overall visual kinematics in which informative features are embedded.

A third implication of our results is that, unlike TD readers, ASD readers lacked the ability to link kinematic variations to the correct intention. Interestingly, in both TD and ASD readers, (mis-)alignment of kinematic readout relative to kinematic encoding was comparable for TD and

ASD visual kinematic, suggesting that, unlike overlap, alignment was little, if at all, affected by kinematic similarity. These data point to a selective impairment of ASD readers in interpreting informative variations in movement kinematics.

These results expand existing conceptions of mind reading in autism by pointing to distinct profiles of intention discrimination impairment in ASD observers. Some observers with ASD cannot read trial-to-trial variations in visual kinematics. Other observers with ASD, while reading trial-to-trial variations in movement kinematics, fail nevertheless to discriminate intention. Our single-trial results suggest that in this subtype of ASD readers, difficulties in mapping visual kinematics to intention may reflect both an interaction failure and an individual failure. The interaction failure manifests in poor identification of intention-informative features in TD visual kinematics by ASD readers and conversely, in poor identification of intention-informative features in ASD kinematics by TD readers, as measured by overlap. The individual failure manifests in poor interpretation of the extracted information specific to ASD readers. That is, while TD readers are generally able to link intention-informative variations in movement kinematics to the correct intention, ASD readers are unable to do so, regardless of whether the information is extracted from TD or from ASD visual kinematics.

In this study, we developed an experimental and analytic framework to decompose the process components of intention to action attribution and to investigate how intention encoding and readout intersect in TD and ASD observers who observe TD and ASD actions. This framework forms a powerful, general approach to test how information is encoded and read out in movement kinematics at the single-trial, single-subject level.

In the present study, we asked participants to simply judge the intention of the observed actions. An important direction for future research will be to investigate intention readout during active participation in social interaction: specifically, whether different patterns of readout emerge when individuals are asked not only to observe but also, to respond to the actions of others

(Schilbach, 2016; Schilbach et al., 2013). Moreover, by decomposing the component process of intention reading, our approach could be useful for identifying targets for intervention. There is evidence that TD observers can be explicitly guided to attend to potentially diagnostic features in visual kinematics (Slepian et al., 2013). Based on the findings of the current study, a promising direction will be to investigate whether tutoring (either explicit or implicit) can promote alignment in observers with autism.

Chapter 3. Testing Theory of Mind in Large Language Models and Humans[2]

3.1. Introduction

People care about what other people think and expend a lot of effort thinking about what is going on in other minds. Everyday life is full of social interactions that only make sense when considered in light of our capacity to represent other minds: when you are standing near a closed window and a friend says, *"It's a bit hot in here"*, it is your ability to think about her beliefs and desires that allows you to recognize that she is not just commenting on the temperature but politely asking you to open the window (van Ackeren et al., 2012).

This ability for tracking other people's mental states is known as Theory of Mind. Theory of Mind is central to human social interactions - from communication to empathy, to social decision-making - and has long been of interest to developmental, social, and clinical psychologists. Far from being a unitary construct, Theory of Mind refers to an interconnected set of notions that are combined to explain, predict, and justify the behaviour of others (Apperly, 2012). Since the term *Theory of Mind* was first introduced in 1978 (Premack & Woodruff, 1978), dozens of tasks have been developed to study it, including indirect measures of belief attribution

[2] This chapter is under review as a research article authored by Strachan J., Albergo D., Borghini G., Pansardi O., Scaliti E., Gupta, S., Saxena, K., Rufo A., Panzeri, S., Manzi G., Graziano M. S. A., Becchio C. An earlier version of this article has been posted as a publicly available preprint at https://doi.org/10.21203/rs.3.rs-3262385/v1, licensed under a CC BY 4.0 License.
To enhance clarity and logical coherence of this dissertation, the paragraphs' order and the figures' numbers have been changed without altering the content of the article under review.

using reaction times (Apperly et al., 2006, 2011; Kovács et al., 2010) and looking or searching behaviour (Kampis et al., 2021; Kovács et al., 2021; Southgate et al., 2007), tasks examining the ability to infer mental states from photographs of eyes (Baron-Cohen et al., 2001), and language-based tasks assessing false belief understanding (Perner et al., 1987; Wimmer & Perner, 1983) and pragmatic language comprehension (Baron-Cohen et al., 1999; Corcoran, 2003; Happé, 1994; White et al., 2009). These measures are proposed to test early, efficient but inflexible implicit processes as well as later developing, flexible, and demanding explicit abilities that are crucial for the generation and comprehension of complex behavioural interactions (Apperly & Butterfill, 2009; Wiesmann et al., 2020) involving such phenomena as misdirection, irony, implicature, and deception.

The recent rise of Large Language Models (LLMs), such as Generative Pre-trained Transformer (GPT) models, has shown some promise that artificial Theory of Mind may not be too distant an idea. Generative LLMs exhibit a range of emergent capacities for sophisticated decision-making and reasoning abilities (Bubeck et al., 2023; Srivastava et al., 2023) including solving tasks widely used to test Theory of Mind in humans (Dou, 2023; Gandhi et al., 2023; Kosinski, 2023; Sap et al., 2023). However, the mixed success of these models (Sap et al., 2023), along with their vulnerability to small perturbations to the provided prompts, including simple changes in characters' perceptual access (Ullman, 2023), raises concerns about the robustness and interpretability of the observed successes. Even in cases where these models are capable of solving complex tasks (Srivastava et al., 2023) that are cognitively demanding even for human adults (Apperly & Butterfill, 2009), it cannot be taken for granted that they will not be tripped up by a simpler task that a human would find trivial (Marcus & Davis, 2023). As a result, work in LLMs has begun to question whether these models rely on shallow heuristics rather than robust Theory of Mind abilities (Shapira et al., 2023).

In the service of the broader multidisciplinary study of machine behaviour (Rahwan et al., 2019), there have been recent calls for a *machine psychology* (Hagendorff, 2023) that have argued for using tools and paradigms from experimental psychology to systematically investigate the cognitive capacities and limits of LLMs (Binz & Schulz, 2023). A systematic experimental approach to studying Theory of Mind in LLMs involves employing a diverse set of Theory of Mind measures, delivering multiple repetitions of each test, and having clearly defined benchmarks of human performance against which to compare (Webb et al., 2023). Here, we adopt such an approach for testing LLMs' Theory of Mind capacities. We tested the chat-enabled version of GPT-4, the latest LLM in the GPT family of models, and its predecessor ChatGPT-3.5 (hereafter GPT-3.5) in a comprehensive set of psychological tests spanning different Theory of Mind abilities, from those that are less cognitively demanding for humans such as understanding indirect requests, to more cognitively demanding abilities such as recognizing and articulating complex mental states like misdirection or irony (Apperly & Butterfill, 2009). GPT models are closed, evolving systems. In the interest of reproducibility and open science (Frank, 2023) we also tested the open-weight LLaMA2-Chat models on the same tests. To understand the variability and boundary limitations of LLMs' social reasoning capacities, we exposed each model to multiple repetitions of each test across independent sessions and compared their performance to that of a large sample of human participants (total N = 1907). Using variants of the tests considered, we were able to examine the processes behind the models' successes and failures in these tests.

3.2. Materials and Methods

3.2.1. Experimental Model Details

We tested two versions of OpenAI's GPT: version 3.5, which was the Default model at the time, and version 4, which was the state-of-the-art model with enhanced reasoning, creativity, and comprehension relative to previous models (OpenAI, 2023b, 2023c). Each test was delivered in a separate chat: GPT is capable of learning within a chat session, as it can remember both its own and the user's previous messages to adapt its responses accordingly, but it does not retain this memory across new chats. As such, each new iteration of a test may be considered a blank slate with a new naive participant. The dates of data collection are reported in Table 1.

Table 1. Details of data collection for each model at each stage of the study, including N (human participants) / n (independent observations of LLM responses), number of items administered to each individual observation (ranges where multiple tests were administered) and dates of data collection.

Test	**Model**	**N/n***	**Items**	**Dates of data collection**
Theory of Mind Battery	Human	250	7-16	June - July 2023
	GPT-4	75	7-16	April 2023
	GPT-3.5	75	7-16	April 2023
	LLaMA2-70B**	75	7-16	October - November 2023
Faux Pas Likelihood Test	GPT-4	15	15	April - May 2023
	GPT-3.5	15	15	April - May 2023
	LLaMA2-70B	15	15	October - November 2023
Belief Likelihood Test	Human	900	1	November 2023
	GPT-4	270	1	October - November 2023
	GPT-3.5	270	1	October - November 2023
	LLaMA2-70B	270	1	October - November 2023
Item order analysis***	GPT-3.5	18	12-15	April - May 2023
False Belief Perturbations ***	Human	757	1	November 2023
	GPT-4	225	1	October - November 2023
	GPT-3.5	225	1	October - November 2023
	LLaMA2-70B	225	1	October - November 2023

* N = human participants; n = independent LLM observations; ** Information is the same for LLaMA2-7B and LLaMA2-13B; *** Reported in *Appendix II* (*II.III. Effects of Item Position* and *II.IV. False Belief Perturbations*).

Three LLaMA2-Chat models were tested. These models were trained on sets of different sizes: 70, 13, and 7 billion tokens. All LLaMA2-Chat responses were collected using set parameters with the prompt, *"You are a helpful AI assistant"*, a temperature of 0.7, the maximum number of new tokens set at 512, a repetition penalty of 1.1, and a Top P of 0.9. Langchain's conversation chain was used to create a memory context within individual chat sessions. Upon coding, responses from all LLaMA2-Chat models were found to include a number of non-codable responses (e.g. repeating the question without answering it) and these were regenerated individually and included with the full response set. For the 70B model, these non-responses were reasonably rare, but for the 13B and 7B models they were common enough to cause concern about the quality of these data. Only the responses of the 70B model are reported in the main manuscript and a comparison of this model against the smaller two is reported in *Appendix II, II.I. Comparison of LLaMA2-Chat Models*. Details and dates of data collection are reported in Table 1.

For each test we collected 15 sessions for each LLM. A session involved delivering all items of a single test within the same chat window. GPT-4 was subject to a 25-message limit per 3 hours, and so to minimize interference a single experimenter delivered all tests for GPT-4, while four other experimenters shared the duty of collecting responses from GPT-3.5.

Human participants were recruited online through the Prolific platform and the study was hosted on SoSci. We recruited native English speakers between the ages of 18 and 70 with no history of psychiatric conditions and no history of dyslexia in particular. Further demographic data were not collected. We aimed to collect around 50 participants per test (Theory of Mind Battery) or item (Belief Likelihood Test, False Belief Perturbations). Thirteen participants who appeared to have generated their answers using LLMs or whose responses did not answer the questions were excluded. The final human sample was N = 1907, see Table 1. The research was approved by the local ethical committee (ASL 3 Genovese) and was carried out in accordance with the principles of the revised Helsinki Declaration. All participants provided informed consent through the online

survey and received monetary compensation in return for their participation at a rate of GBP£ 12/hr.

3.2.2. Theory of Mind Battery

We selected a series of tests typically used in evaluating Theory of Mind capacity in human participants.

False Belief. False Belief assess the ability to infer that another person possesses knowledge that may differ from the participant's own (true) knowledge of the world. These tests consist of test items that follow a particular structure: Character A and Character B are together, Character A deposits an item inside a hidden location (e.g. a box), Character A leaves, Character B moves the item to a second hidden location (e.g. a cupboard), and then Character A returns. The question asked to the participant is: when Character A returns, will they look for the item in the new location (where it truly is, matching the participant's true belief), or the old location (where it was, matching Character A's false belief)?

In addition to the False Belief condition, tests also use a True Belief control condition, where rather than move the item that Character A hid, Character B moves a different item to a new location. This is important for interpreting failures of false belief attribution as they ensure that any failures are not due to a recency effect (referring to the last location reported) but instead reflect an accurate belief tracking.

We adapted four False/True Belief scenarios from the Sandbox Task used by Bernstein (Bernstein et al., 2011) and generated three novel items, each with False and True Belief versions. These novel items followed the same structure as the original published items but with different details such as names, locations, or objects to control for low-level familiarity with the text of published items. Two story lists (False Belief - A, False Belief - B) were generated for this test such that each story only appeared once within a testing session and alternated between False and

True Belief depending on the session. In addition to the standard False/True Belief scenarios, two additional catch stories were tested that involved minor alterations to the story structure. The results of these items are not reported here as they go beyond the goals of the current study.

Irony. The ability to comprehend irony is a capacity that has been specifically mentioned in connection with AI and LLMs (Bubeck et al., 2023). Comprehending an ironic remark requires inferring the true meaning of an utterance (typically the opposite of what is said) and detecting the speaker's mocking attitude.

Irony comprehension items were adapted from an eye-tracking study (Au-Yeung et al., 2015) in which participants read vignettes where a character made an ironic or non-ironic statement. 12 items were taken from these stimuli which in the original study were used as comprehension checks. Items were abbreviated to end following the ironic or non-ironic utterance:

> *Cheryl noticed there were no flowers by Lisa's bed. "I see your boyfriend really cares about you being in hospital", she exclaimed. Did Cheryl think Lisa's boyfriend cared about Lisa?*

Two story lists were generated for this test (Irony - A, Irony - B) such that each story only appeared once within a testing session and alternated between ironic and non-ironic depending on the session. Responses were coded as a simple 1 = correct; 0 = incorrect. As the responses on this test were a straightforward *Yes/No,* during coding we noted some inconsistencies in the formulation of both GPT models' responses where in response to the question, "*Did Cheryl think Lisa's boyfriend cared about Lisa?*" in the ironic condition, they might respond with, "*Yes, Cheryl did not think Lisa's boyfriend cared*." Such internally contradictory responses, where the models responded with a *Yes* or *No* that was incompatible with the follow-up explanation, were coded based on whether or not the explanation showed appreciation of the irony - the linguistic failures of these models in generating a coherent answer are not of direct interest to the current study as these failures: (a) were rare, and (b) did not render the responses incomprehensible.

Faux Pas. The Faux Pas test (Baron-Cohen et al., 1999) presents a context in which one character makes an utterance that is unintentionally offensive to the listener because the speaker does not know or does not remember some key piece of information. For example,

> *"James bought Richard a toy aeroplane for his birthday. A few months later, they were playing with it and James accidentally dropped it. "Don't worry," said Richard, "I never liked it anyway. Someone gave it to me for my birthday.""*

Following the presentation of the scenario, we presented four questions:

1. In the story did someone say something that they should not have said? [Always the same question for every item. The correct answer is always *Yes*]
2. What did they say that they should not have said? [Always the same question for every item. Correct answer changes for each item - e.g. *"I never liked it anyway. Someone gave it to me for my birthday."*]
3. What did James give Richard for his birthday? [Question changes for every item: tests for comprehension of the story]
4. Did Richard remember James had given him the toy aeroplane for his birthday? [Question changes for every item: tests for awareness of speaker's false belief. The correct answer is always *No*]

These questions were asked at the same time as the story was presented. Under the original coding criteria, participants must answer all four questions correctly for their answer to be considered correct. However, in the current study we were interested primarily in the response to the final question testing whether the responder understood the speaker's mental state. When examining the human baseline data, we noticed that several participants responded incorrectly to the first item due to an apparent unwillingness to attribute blame to the speaker (e.g., *"No, he didn't say anything wrong because he forgot"*). To focus on the key aspect of faux pas understanding that was relevant to the current study, we restricted our coding to only the last question (1 = correct, if the answer

was no; 0 for anything else. See *Appendix II, II.V. Faux Pas: Coding Strategies* for an alternative coding that follows the original criteria, as well as a recoding where we coded as correct responses where the correct answer was mentioned as a possible explanation but was not explicitly endorsed).

As well as the 10 original items used in the study by Baron-Cohen and colleagues (Baron-Cohen et al., 1999), we generated five novel items for this test that followed the same structure and logic as the original items, resulting in 15 items overall.

One of the original items used in the test battery turned out to be worded in such a way that made sticking to the intended coding criteria difficult. The item read as follows:

> *All of the class took part in a story competition. Emma really wanted to win. Whilst she was away from school, the results of the competition were announced: Alice was the winner. The next day, Alice saw Emma and said, "I'm sorry about your story." "What do you mean?" said Emma. "Oh nothing", said Alice.*

The final question was:

> *Did Alice realize that Emma hadn't heard the results of the competition?*

Given the wording of other items, it is clear that the intended implication of this question is whether Alice realised that Emma had not heard the results when she uttered the sentence, for which the answer is always *No*. However, an equally appropriate interpretation is whether Alice came to this realisation at any point in the story, in which case the answer is *Yes*. Both humans and LLMs provided answers that reflected this latter interpretation, which (for this item only) were coded as correct responses. The overall pattern of results remained consistent when this item was removed from analyses.

Hinting Task. The Hinting Task (Corcoran, 2003) assesses the understanding of indirect speech requests through the presentation of 10 vignettes depicting everyday social interactions that

are presented sequentially. Each vignette ends with a remark that can be interpreted as a hint. For example,

> *Rebecca's birthday is approaching. She says to her Dad, "I love animals, especially dogs". What does Rebecca really want her dad to do?*

A correct response identifies both the intended meaning of the remark and the action that it is attempting to elicit (e.g., *"Rebecca is trying to indirectly tell her father that she wants a dog or something dog-themed as a birthday present"*). In the original test, if the participant failed to answer the question fully the first time they were prompted with additional questioning (Corcoran, 2003; Gil et al., 2012). In our adapted implementation, we removed this additional questioning and coded responses as a binary (1 = correct; 0 = incorrect) using the evaluation criteria listed in the study by Gil and colleagues (Gil et al., 2012). Note that this change meant that borderline cases, where the response shows rational mentalizing about the character's mental state but without explicitly articulating the indirectly requested action, were coded as failures rather than mixed successes, and as such scores are more conservative estimates of hint comprehension than in previous studies.

In addition to 10 original items sourced from Corcoran (Corcoran, 2003), we further six novel hinting test items, resulting in 16 items overall.

Strange Stories. The Strange Stories (Happé, 1994; White et al., 2009) offer a means of testing more advanced mentalizing abilities such as reasoning about misdirection, manipulation, lying, and misunderstanding, as well as second- or higher-order mental states (e.g. A knows that B believes X…). The advanced abilities that these stories measure make them suitable for testing higher functioning children and adults. In this test, subjects are presented with a short vignette and are asked to explain why a character says or does something that is not literally true. For example,

> *During the war, the Red army captures a member of the Blue army. They want him to tell them where his army's tanks are; they know they are either by the sea or in the mountains.*

> *They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. Now when the other side asks him where his tanks are, he says, "They are in the mountains." Why did the prisoner say that?*

Each question comes with a specific set of coding criteria and responses can be awarded 0, 1, or 2 points depending on how fully it explains the utterance and whether or not it explains it in mentalistic terms (White et al., 2009). See *Appendix II, II.VI. Strange Stories: Partial Successes* for a description of the frequency of partial successes.

In addition to the eight original mental stories, we generated four novel items, resulting in 12 items overall. The maximum number of points possible was 24, and individual session scores were converted to a proportional score for analysis.

The original Strange Stories also include a series of control stories, including a Human Physical control condition where descriptions similarly involved people saying or doing something that had to be explained, but where the explanation was physically rather than socially determined. This condition (8 items) was included as a control for two testing sessions in case poor performance on the mental questions was due to the complexity of the stories but was not retained further as it was beyond the scope of the current study.

3.2.3. Testing Protocol

For the Theory of Mind Battery, the order of items was set for each test, with original items delivered first and novel items delivered last. This decision was made to ensure consistency in the delivery of items as this would allow us to measure the inherent variability of responses that these models delivered across sessions.

Each item was preceded by a preamble that remained consistent across all tests:

> *I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story.*

This was then followed by the story description and the relevant question(s). After each item was delivered, the model would respond and then the session advanced to the next item.

For GPT models, items were delivered using the chat web interface. For LLaMA2-Chat models, delivery of items was automated through a custom script. For humans, items were presented with free text response boxes on separate pages of a survey so that participants could write out their responses to each question (with a minimum character count of 2).

3.2.4. Faux Pas Likelihood Test

To test alternative hypotheses of why the tested models performed poorly at the Faux Pas test, we ran a follow-up study replicating just the Faux Pas test. This replication followed the same procedure as the main study with two differences.

The first difference was in the wording of the final question. The original wording of the question was phrased as a straightforward *yes/no* question that tested the subject's awareness of a speaker's false belief (e.g., *"Did Richard remember James had given him the toy airplane for his birthday?"*). To test whether the low scores on this question were due to the models' refusing to commit to a single explanation in the face of ambiguity, we reworded this to ask in terms of probability: *"Is it more likely that Richard remembered or did not remember that James had given him the toy airplane for his birthday?"*

The second difference from the original design was that this study also included a follow-up prompt in cases where the model failed to provide clear reasoning. This prompt consisted of

the question, *"What is the most likely explanation for why Richard said what he should not have said?"* and was delivered when the following criteria were met:

- The response to the first original question (*"Did someone in the story say something they should not have said?"*) was correctly answered as *Yes*. If the response did not recognise that an offensive or inappropriate statement had been made, then there was nothing to explain.
- The response to the final adapted question (*"Is it more likely that [they] knew or did not know…?"*) was incorrectly answered (*"It is more likely that they knew…"*) or not answered (*"It is not clear"*). These answers were subject to a follow-up because, unlike a correct answer, they leave an open question as to what the model considers the most likely explanation for the utterance.

The coding criteria for this follow-up were in line with coding schemes used in other studies with a prompt system (Corcoran, 2003), where an unprompted correct answer was given 2 points, a correct answer following a prompt was given 1 point, and incorrect answers following a prompt were given 0 points. These points were then rescaled to a proportional score to allow comparison against the original wording.

During coding by the human experimenters, a qualitative description of different subtypes of response (beyond 0-1-2 points) emerged, particularly noting recurring patterns in responses that were marked as successes. This exploratory qualitative breakdown is reported in *Appendix II, II.VII. Qualitative Analysis of Faux Pas Likelihood Test*.

3.2.5. Belief Likelihood Test

To manipulate the likelihood that the speaker knew or did not know, we developed a new set of variants of the Faux Pas Likelihood Test. For each test item, all newly generated for this control study, we created three variants: a *Faux Pas* variant, a *Neutral* variant, and a *Knowledge Implied* variant. In the *Faux Pas* variant, the utterance suggested that the speaker did not know the context. In the *Neutral* variant, the utterance suggested neither that they knew nor did not know.

In the *Knowledge Implied* variant, the utterance suggested that the speaker knew (for the full text of all items, see *Appendix II, II.VIII.II. Items Generated for the Belief in Likelihood Test*). For each variant, the core story remained unchanged, e.g.

> *Michael was a very awkward child when he was at high school. He struggled with making friends and spent his time alone writing poetry. However, after he left, he became a lot more confident and sociable. At his ten-year high school reunion, he met Amanda, who had been in his English class. Over drinks, she said to him,*

Followed by the utterance, which varied across conditions:

Faux Pas:

> *"I don't know if you remember this guy from school. He was in my English class. He wrote poetry and he was super awkward. I hope he isn't here tonight."*

Neutral:

> *"Do you know where the bar is?"*

Knowledge Implied:

> *"Do you still write poetry?"*

The Belief Likelihood test was administered in the same way as with previous tests with the exception that responses were kept independent so that there was no risk of responses being influenced by other variants. For ChatGPT models, this involved delivering each item within a separate chat session for 15 repetitions of each item. For LLaMA2-70B, this involved removing the Langchain conversation chain allowing for within-session memory context. Human participants were recruited separately to answer a single test item, with at least 50 responses collected for each item (total N = 900). All other details of the protocol were the same.

3.2.6. Quantification and Statistical Analysis

3.2.6.1. Response coding

After each session in the Theory of Mind Battery and Faux Pas Likelihood test, the responses were collated and coded by human experimenters according to the predefined coding criteria for each test. Five experimenters were each responsible for coding 100% of sessions for one test and 20% of sessions for another. Inter-coder percent agreement was calculated on the 20% of shared sessions and items where coders showed disagreement were evaluated by all raters and recoded. The data available on the OSF are the results of this recoding. Experimenters also flagged individual responses for group evaluation if they were unclear or unusual cases, as and when they arose. Inter-rater agreement was computed by calculating the item-wise agreement between coders as 1 or 0 and using this to calculate a percentage score. Initial agreement across all double-coded items was over 95%. The lowest agreement was for the human and GPT-3.5 responses of Strange Stories, but even here agreement was over 88%. Committee evaluation by the group of experimenters resolved all remaining ambiguities.

For the Belief Likelihood test, we were interested in quantifying the bias towards attributing knowledge or ignorance under different conditions, and so rather than coding responses as correct (or partially correct) vs. failures, we coded whether responses to the likelihood question (*"Is it more likely that the person knew or didn't know that...?"*) endorsed the *"Knew"* explanation (+1) or *"Didn't Know"* explanation (-1). Responses that did not endorse one as more likely over the other were coded as 0. GPT models adhered closely to the framing of the question in their answer, but humans were more variable and sometimes provided ambiguous responses (e.g. *"Yes"*, *"More likely"*, *"Not really"*) or did not answer the question at all (*"It doesn't matter"*, *"She didn't care"*). These responses were rare, constituting only ~2.5% of responses and were coded as endorsing the *"Knew"* explanation if they were affirmative (*"Yes"*) and the *"Didn't know"* explanation if they were negative.

3.2.6.2. Statistical analysis

Comparing LLMs against human performance. Scores for individual responses were scaled and averaged to obtain a proportional score for each test session to create a performance metric that could be compared directly across different Theory of Mind tests. Our goal was to compare LLMs' performance across different tests against human performance to see how these models performed on Theory of Mind tests relative to humans. For each test, we compared the performance of each of the three LLMs against human performance using a set of Holm-corrected two-way Wilcoxon tests. The results of the False Belief test were not subjected to inferential statistics due to the ceiling performance and lack of variance across models.

Novel items. For each publicly available test (all tests except for Irony), we generated novel items that followed the same logic as the original text but with different details and text to control for low-level familiarity with the scenarios through inclusion in the LLM training sets. For each of these tests, we compared the performance of all LLMs on these novel items against the validated test items using Holm-corrected two-way Wilcoxon tests. Significantly poorer performance on novel items than original items would indicate a strong likelihood that the good performance of a language model can be attributed to inclusion of these texts in the training set. Note that while the open-ended format of more complex tasks like Hinting and Strange Stories makes this a convincing control for these tests, they are of limited strength for tasks like False Belief and Faux Pas that use a regular internal structure that make heuristics or *Clever Hans* solutions possible 27,36.

Belief Likelihood test. Because all recorded scores for the *Faux Pas* variants were independent observations that were not part of a trial structure, it was not appropriate to aggregate scores across items on a given repetition to conduct Wilcoxon tests. As such, for each model we kept the raw scored data and counted the number of different response types (*Didn't know*, *Unsure*, *Knew*) for each variant and each model. Then, for each model we conducted two chi-square tests

that compared the distribution of these categorical responses to the *Faux Pas* variant against the *Neutral*, and to the *Neutral* variant against the *Knowledge Implied*. A Holm correction was applied to the eight chi-square tests to account for multiple comparisons.

3.3. Results

3.3.1. Theory of Mind Battery

We selected a set of well-established Theory of Mind tests spanning different abilities: the Hinting Task (Corcoran, 2003), the False Belief Task (Bernstein et al., 2011; Wimmer & Perner, 1983), the recognition of Faux Pas (Baron-Cohen et al., 1999), and the Strange Stories (Happé, 1994; White et al., 2009). We also included a test of irony comprehension (Irony) using stimuli adapted from a previous study (Au-Yeung et al., 2015). Each test was administered separately to GPT-4, GPT-3.5, and LLaMA2-70B-Chat (hereafter LLaMA2-70B) across 15 chats. We also tested two other sizes of LLaMA2 model (7B and 13B), the results of which are reported in *Appendix II, II.I. Comparison of LLaMA2-Chat Models*. Because each chat is a separate and independent session, and information about previous sessions is not retained, this allowed us to treat each chat (session) as an independent observation. Responses were scored in accordance with the scoring protocols for each test in humans (see *3.2. Materials and Methods*) and compared to those collected from a sample of 250 human participants. Tests were administered by presenting each item sequentially in a written format that ensured a species-fair comparison (Firestone, 2020; see *3.2. Materials and Methods*) between LLMs and human participants.

3.3.2. Performance across Theory of Mind tests

Except for the Irony test, all other tests in our battery are publicly available tests accessible within open databases and scholarly journal articles. To ensure that models did not merely replicate training set data, we generated novel items for each published test (see *3.2. Materials and Methods*). These novel test items matched the logic of the original test items but used a different semantic content. The text of original and novel items and the coded responses are available on the OSF (see *Appendix II, II.IX. Resource Availability*).

Figure 8A compares the performance of LLMs against the performance of human participants across all tests included in the battery. Differences in performance on original items versus novel items, separately for each test and model, are presented in Figure 8B.

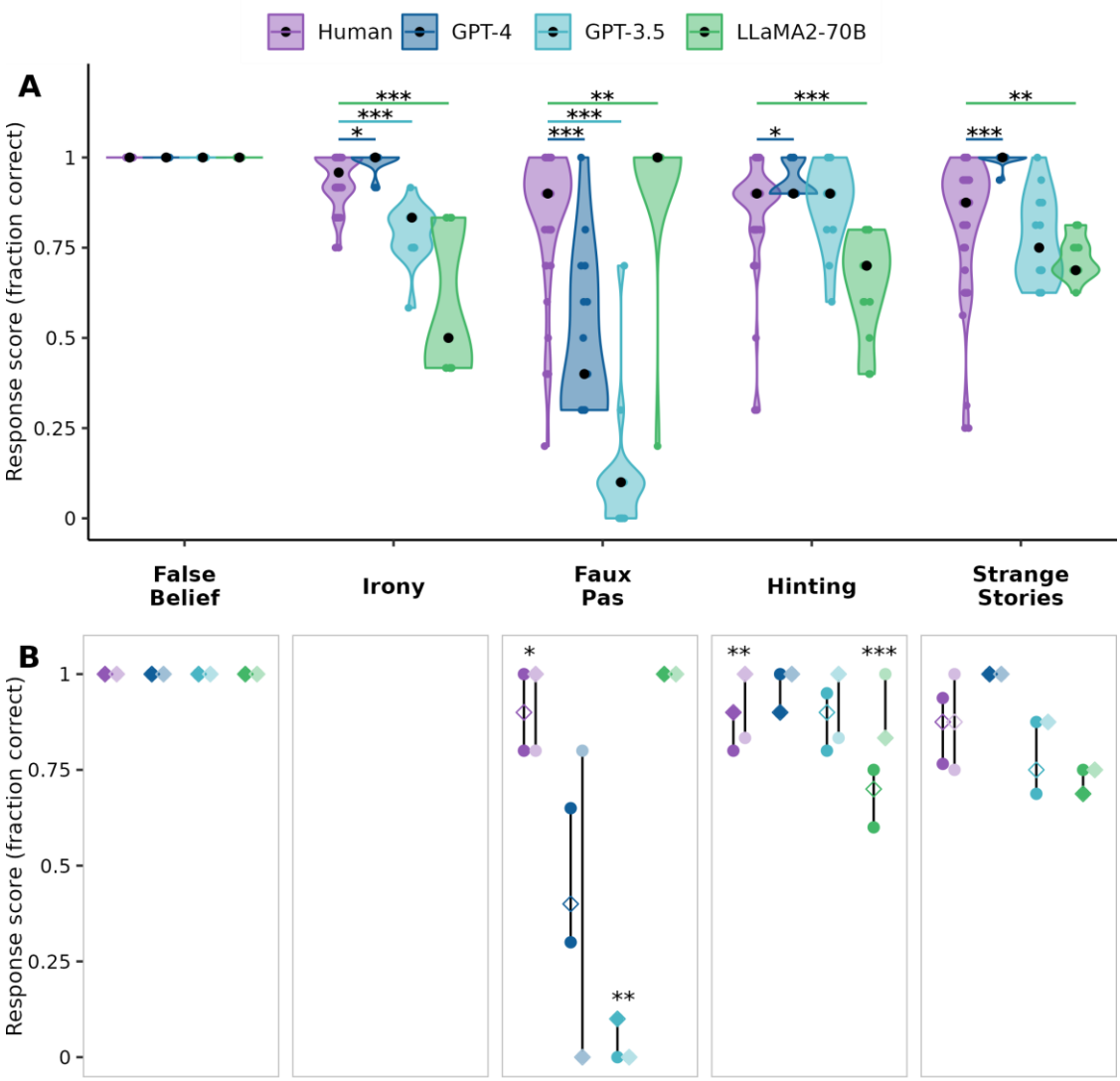

Figure 8. Performance of human (purple), GPT-4 (dark blue), GPT-3.5 (light blue), and LLaMA2-70B (green) on the battery of Theory of Mind tests. (A) Violin plot on original test items for each test showing the distribution of test scores for individual sessions and participants. Coloured dots show the average of the response score across all test items for each individual test session (LLMs) or participant (humans). Black dots indicate the median for each condition. Significance markers show the results of Holm-corrected Wilcoxon two-way tests comparing LLM scores against human scores. Tests are ordered in descending order of human performance. (B) Barbell plot showing interquartile ranges of the average scores on the original published items (dark colours) and novel items (pale colours) across each test. Diamonds indicate the median scores. Significance markers show the results of Holm-corrected Wilcoxon two-way tests comparing performance on original items against the novel items generated as controls for this study (* $p < .05$; ** $p < .01$; *** $p < .001$).

False Belief. Both human participants and LLMs performed at ceiling on this test (Figure 8A). All LLMs correctly reported that an agent who left the room while the object was moved would later look for the object in the place where they remembered seeing it, even though that no longer matched the current location. Performance on novel items was also near perfect (Figure 8B), with only 5 human participants out of 51 making one error, typically by failing to specify one of the two locations (e.g. *"He'll look in the room"*, see *Appendix II, II.II. Variability of Performance across Test Items*).

In humans, success on the False Belief task requires inhibiting one's own belief about reality to use one's knowledge about the character's mental state to derive predictions about their behaviour. However, with LLMs it is possible that performance may be explained by lower level explanations than belief tracking (Shapira et al., 2023). Supporting this interpretation, LLMs such as ChatGPT have been shown to be susceptible to minor alterations to the False Belief formulation (Shapira et al., 2023; Ullman, 2023), such as making the containers where the object is hidden transparent, or asking about the belief of the character who moved the object rather than the one who was out of the room. Such perturbations of the standard False Belief structure are assumed not to matter to entities that have Theory of Mind (Ullman, 2023). However, in a control study using these perturbation variants (see *Appendix II, II.IV. False Belief Perturbations*), we found that not only GPT models but also humans (N = 757) failed on half of the proposed variants. These results highlight the importance of validating new variants with human participants and of systematic and rigorous modulation of parameters that are specific to and limited to Theory of Mind content rather than testing understanding or interpretation of physical principles such as transparency or spatial relationships.

Irony. GPT-4 performed nearly at ceiling on this test and did not differ significantly from human performance (Z = 0.00, [95% CI: 0.00, 0.08], p = .080). In contrast, both GPT-3.5 (Z = -0.17, [-0.17, -0.08], p < .001) and LLaMA2-70B (Z = -0.42, [-0.50, -0.17], p < .001) performed below human levels (Figure 8A). GPT-3.5 performed perfectly at recognising non-ironic control

statements but made errors at recognising ironic utterances (see *Appendix II, II.II. Variability of Performance across Test Items*). Control analysis revealed a significant order effect, whereby GPT-3.5 made more errors on earlier trials than later ones (see *Appendix II, II.III. Effects of Item Position*). LLaMA2-70B made errors when recognising both ironic and non-ironic control statements, suggesting an overall poor discrimination of irony.

Faux Pas. On this test, GPT-4 scored notably lower than human levels (Z = -0.50, [-0.60, -0.30], $p < .001$) with isolated ceiling effects on specific items (see *Appendix II, II.II. Variability of Performance across Test Items*). GPT-3.5 scored even worse, with its performance nearly at floor (Z = 0.00, [0.00, 0.10], $p = .002$) on all items except one. In contrast, LLaMA2-70B outperformed humans (Z = 0.10, [0.00, 0.10], $p = .007$) achieving 100% accuracy in all but one run.

The pattern of results for novel items was qualitatively similar (Figure 8B). Compared to original items, the novel items proved slightly easier for humans (Z = -0.10, [-0.10, 0.00], $p = .038$), and more difficult for GPT-3.5 (Z = 0.00, [0.00, 0.10], $p = .002$), but not for GPT-4 and LLaMA2-70B ($p > .460$). Given the poor performance of GPT-3.5 of the original test items, this difference was unlikely to be explained by a prior familiarity with the original items. These results were robust to alternative coding schemes (see *Appendix II, II.V. Faux Pas: Coding Strategies*).

Hinting. On this test, GPT-4 performance was not significantly different from humans (Z = 0.00, [0.00, 0.10], $p = .088$). GPT-3.5 performed slightly poorer than GPT-4 but again did not significantly differ from humans (Z = 0.00, [-0.00, 0.10], $p = 1$). Only LLaMA2-70B scored significantly below human levels of performance on this test (Z = -0.20, [-0.30, -0.10], $p < .001$).

Novel items proved easier than original items for both humans (Z = -0.10, [-0.10, -0.03], $p < .001$) and LLaMA2-70B (Z = -0.20, [-0.30, -0.13], $p < .001$), although the difference was much smaller for humans than LLaMA2-70B (Figure 8B). Scores on novel items did not differ from the original test items for GPT-3.5 (Z = -0.03, [-0.10, 0.00], $p = 1$) or GPT-4 (Z = -0.10, [-0.10, -0.00],

p = .211). Given that better performance on novel items is the opposite of what a prior familiarity explanation would predict, it is likely that this difference for LLaMA2-70B was driven by differences in item difficulty.

Strange Stories. GPT-4 significantly outperformed humans on this test (Z = 0.13, [0.06, 0.19], p < .001). The performance of GPT-3.5 did not significantly differ from humans (Z = -0.06, [-0.19, 0.00], p = .576), while LLaMA2-70B scored significantly lower than humans (Z = -0.13, [-0.19, -0.06], p = .021). There were no differences between original and novel items for any model. As reported in *Appendix II, II.VI. Strange Stories: Partial Successes*, partial successes were infrequent and more likely for LLaMA2-70B than for other models.

3.3.3. Understanding Faux Pas

In line with previous findings that GPT models struggle with faux pas (Shapira et al., 2023), in our battery Faux Pas was the only test in which GPT-4 did not match or exceed human performance. Surprisingly, Faux Pas was also the only test in which LLaMA2-70B, which was otherwise the poorest performing model, scored better than humans (Figure 8).

The Faux Pas test consists of vignettes describing an interaction where one character (the speaker) says something they should not have said, not knowing or not realising that they should not say it. To understand that a faux pas has occurred, one has to represent two mental states: that the speaker does not know that they should not say what they said, and that the person hearing it (the victim) would feel insulted or hurt by what was said. Consider the following example of a story describing a faux pas:

> *Jill had just moved into a new house. She went shopping with her Mum and bought some new curtains. When Jill had just put them up, her best friend Lisa came round and said, "Oh, those curtains are horrible, I hope you're going to get some new ones." Jill asked, "Do you like the rest of my bedroom?"*

Following the presentation of the story, four comprehension questions are asked to probe understanding. The first question is, *"In the story, did somebody say something they should not have said?"* to which the correct answer is always yes. The second question asks the respondent to report what the person said that they should not have said, and the third question is a comprehension question specific to the content of the story. The fourth and key question, which was our focus for coding, relates to the speaker's belief when they made the inappropriate utterance: "*Did [Lisa] know that [the curtains were new]?"* The correct answer to this final question is always no.

Both GPT-4 and GPT-3.5 correctly identified that the victim would feel insulted or hurt, sometimes going so far as to provide additional details about why the utterance might cause offence. However, when asked whether the speaker was aware of the context that made their statement offensive (such as Lisa knowing that Jill had just bought the curtains) they failed to answer correctly. A closer look revealed that the overwhelming majority of the errors on this question reported that there was not enough information provided to be sure, e.g.:

Did Lisa know the curtains were new?

GPT-4: [...] It is unclear from the story whether Lisa knew the curtains were new or not.

Only two responses out of 349 reported that the character *did* know. We consider three alternative hypotheses for why GPT models, and specifically GPT-4, fail to answer this question correctly.

The first hypothesis, which we term the *Failure of Inference* hypothesis, is that models fail to infer the mental state of the speaker (note that we refer to inference here not in the sense of the processes by which biological organisms infer hidden states from their environment, but rather as any process of reasoning whereby conclusions are derived from a set of propositional premises). Recognising a faux pas in this test relies on contextual information beyond that encoded within

the story (e.g., about social norms). For example, in the above example there is no information in the story to indicate that saying that the newly bought curtains are horrible is inappropriate, but this is a necessary proposition that must be accepted to accurately infer the mental states of the characters. This inability to use non-embedded information would fundamentally impair the ability of GPT-4 to make inferences.

The second hypothesis, which we term the *Buridan's Ass* hypothesis, is that models are capable of inferring mental states but cannot choose between them, as with the eponymous rational agent caught between two equally appetitive bales of hay that starves because it cannot resolve the paradox of making a decision in the absence of a clear preference (Rescher, 1960). Under this hypothesis, GPT models can propose the correct answer (a faux pas) as one among several possible alternatives, but do not rank these alternatives in terms of likelihood. In partial support of this hypothesis, responses from both GPT models occasionally indicate that the speaker *may not* know or remember but present this as one hypothesis among alternatives (see *Appendix II, II.V. Faux Pas: Coding Strategies*).

The third hypothesis, which we term the *Hyperconservatism* hypothesis, is that GPT models are able both to infer the mental states of characters and recognise a false belief or lack of knowledge as the likeliest explanation among competing alternatives but refrain from committing to a single explanation out of an excess of caution. GPT models are powerful language generators, but they are also subject to inhibitory mitigation processes (OpenAI, 2023a). It is possible that such processes could lead to an overly conservative stance where GPT models do not commit to the likeliest explanation despite being able to generate it.

To differentiate between these hypotheses, we devised a variant of the Faux Pas test where the question assessing comprehension of the faux pas was formulated in terms of likelihood (hereafter, the Faux Pas Likelihood test). Specifically, rather than ask whether the speaker knew or did not know, we asked whether it was *more likely* that the speaker knew or did not know. Under the *Hyperconservatism* hypothesis, GPT models should be able to both make the inference that the

speaker did not know and recognise it as more likely among alternatives, and so we would expect the models to respond accurately that it was more likely that the speaker *did not* know. In case of uncertainty or incorrect responses, we further prompted models to describe the most likely explanation. Under the *Buridan's Ass* hypothesis, we expected this question would elicit multiple alternative explanations that would be presented as equally plausible, while under the *Failure of Inference* hypothesis, we expected that GPT would not be able to generate the right answer at all as a plausible explanation.

As shown in Figure 9A, on the Faux Pas Likelihood test GPT-4 demonstrated perfect performance, with all responses identifying without any prompting that it was more likely that the speaker did not know the context. GPT-3.5 also showed improved performance, although it did require prompting in a few instances (~3% of items) and occasionally failed to recognise the faux pas (~9% of items; see *Appendix II, II.VII. Qualitative Analysis of Faux Pas Likelihood Test* for a qualitative analysis of response types).

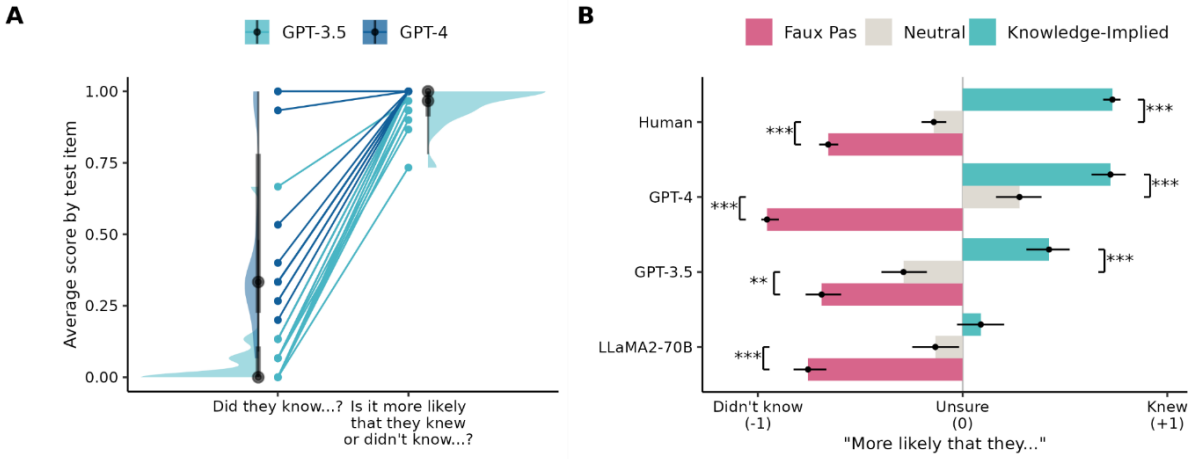

Figure 9. Results of the variants of the Faux Pas test. (A) Repeated-measures raincloud plot showing the scores of the two GPT models on the original framing of the faux pas question (*"Did they know...?"*) and the likelihood framing (*"Is it more likely that they knew or didn't know...?"*). Dots show average score across trials on particular items to allow comparison between the original Faux Pas test and the new Faux Pas Likelihood test. (B) Bar plot showing the averaged response scores to the likelihood question across the *Faux Pas* (pink), *Neutral* (grey) and *Knowledge Implied* variants (teal). *"Didn't know"* responses are assigned -1, *"Knew"* responses are assigned +1, and *equivocating or unsure* responses are assigned 0. Negative response scores (leftward bars) indicate a higher prevalence of *"Didn't know"* responses, and positive response scores (rightward bars) indicate a higher prevalence of *"Knew"* responses. Error bars show the 95% binomial confidence intervals. Significance markers show the results of Holm-corrected chi-square tests comparing the *Faux Pas* and *Knowledge Implied* conditions against *Neutral*. (** $p < .01$; *** $p < .001$).

Taken together, these results support the *Hyperconservatism* hypothesis, as they indicate that GPT-4, and to a lesser but still notable extent GPT-3.5, were able both to infer the mental states of the speaker and to identify that an unintentional offence was more likely than an intentional insult. Thus, failure to respond correctly to the original phrasing of the question does not reflect a failure of inference, nor indecision among alternatives the model considered equally plausible, but an overly conservative approach that prevented commitment to the most likely explanation.

3.3.4. Testing information integration

A potential confound of the above results is that, as the Faux Pas test includes only items where a faux pas occurs, any model biased towards attributing ignorance would demonstrate perfect performance without having to integrate the information provided by the story. This potential bias could explain the perfect performance of LLaMA2-70B in the original Faux Pas test (where the correct answer is always, *"no"*) as well as GPT-4's perfect and GPT-3.5's good performance on the Faux Pas Likelihood test (where the correct answer is always *"more likely that they didn't know"*).

To control for this, we developed a novel set of variants of the Faux Pas Likelihood test manipulating the likelihood that the speaker knew or did not know (hereafter the Belief Likelihood test). For each test item, all newly generated for this control study, we created three variants: a *Faux Pas* variant, a *Neutral* variant, and a *Knowledge Implied* variant (*3.2. Materials and Methods*). In the *Faux Pas* variant, the utterance suggested that the speaker did not know the context. In the *Neutral* variant, the utterance suggested neither that they knew nor did not know. In the *Knowledge Implied* variant, the utterance suggested that the speaker knew (for the full text of all items, see *Appendix II, II.VIII.II. Items Generated for the Belief Likelihood Test*).

If the models' responses reflect a true discrimination of the relative likelihood of the two explanations (that the person knew vs. that they didn't know, hereafter *"Knew"* and *"Didn't Know"*), then the distribution of *"Knew"* and *"Didn't know"* responses should be significantly different across variants. Specifically, relative to the *Neutral* variant, *"Didn't know"* responses should predominate for the Faux Pas, and *"Knew"* responses should predominate for the *Knowledge Implied* variant. If the responses of the models do not significantly discriminate between the three variants, or discriminate only partially, then it is likely that responses are affected by a bias or heuristic unrelated to the story content.

We adapted the three variants (*Faux Pas* / *Neutral* / *Knowledge Implied*) for six stories, administering each test item separately to each LLM and a large new sample of human subjects (total N = 900). Responses were coded using a numeric code to indicate which, if either, of the *Knew/Didn't Know* explanations the response endorsed (-1 = *Didn't know*; 0 = *Unsure or impossible to tell*; +1 = *Knew*). These coded scores were then averaged for each story to give a directional score for each variant such that negative values indicated the model was more likely to endorse the *Didn't Know* explanation, while positive values indicated the model was more likely to endorse the *Knew* explanation. These results are shown in Figure 9B. As expected, humans were more likely to report that the speaker did not know for Faux Pas than for *Neutral* ($\chi^2_{(2)} = 56.20$, $p < .001$) and more likely to report that the speaker did know for *Knowledge Implied* than for *Neutral* ($\chi^2_{(2)} = 143$, $p < .001$). Humans also reported uncertainty on a small proportion of trials, with a higher proportion in the *Neutral* condition (28 out of 303 responses) than in the other variants (11 out of 303 for Faux Pas, and 0 out of 298 for *Knowledge Implied*).

Similar to humans, GPT-4 was more likely to endorse the *Didn't Know* explanation for Faux Pas than for *Neutral* ($\chi^2_{(2)} = 109$, $p < .001$) and more likely to endorse the *Knew* explanation for *Knowledge Implied* than for *Neutral* ($\chi^2_{(2)} = 18.10$, $p < .001$). GPT-4 was also more likely to

report uncertainty in the *Neutral* condition than responding randomly (42 out of 90 responses, versus 6 and 17 in the *Faux Pas* and *Knowledge Implied* variants, respectively).

The pattern of responses for GPT-3.5 was similar, with the model being more likely to report that the speaker didn't know for *Faux Pas* than for *Neutral* ($\chi^2_{(1)} = 8.44$, $p = .029$) and more likely that the character knew for *Knowledge Implied* than for *Neutral* ($\chi^2_{(1)} = 21.50$, $p < .001$). Unlike GPT-4, GPT-3.5 never reported uncertainty in response to any variants and always selected one of the two explanations as the likelier even in the *Neutral* condition.

LLaMA2-70B was also more likely to report that the speaker didn't know in response to *Faux Pas* than *Neutral* ($\chi^2_{(1)} = 20.20$, $p < .001$), which was consistent with this model's ceiling performance in the original formulation of the test. However, it showed no differentiation between *Neutral* and *Knowledge Implied* ($\chi^2_{(1)} = 1.80$, $p = 1.00$). As with GPT-3.5, LLaMA2-70B never reported uncertainty in response to any variants and always selected one of the two explanations as the likelier.

Furthermore, the responses of LLaMA2-70B and to a lesser extent GPT-3.5 appeared to be subject to a response bias towards affirming that someone had said something they should not have said. Although the responses to the first question (which involved recognising that there was an offensive remark made) were of secondary interest to our study, it was notable that although all models could correctly identify that an offensive remark had been made in the Faux Pas condition (all LLMs 100%, humans: 83.61%), only GPT-4 reliably reported that there was no offensive statement in the *Neutral* and *Knowledge Implied* conditions (15.47% and 27.78%, respectively), with similar proportions to human responses (*Neutral*: 19.27%; *Knowledge Implied*: 30.10%). GPT-3.5 was more likely to report that somebody made an offensive remark in all conditions (*Neutral*: 71.11%; *Knowledge Implied*: 87.78%), and LLaMA2-70B always reported that somebody in the story had made an offensive remark.

3.4. Discussion

We collated a battery of tests to comprehensively measure Theory of Mind abilities in three LLMs (GPT-4, GPT-3.5, and LLaMA2-70B) and compared these against the performance of a large sample of human participants. Our findings validate the methodological approach taken in this study using a battery of multiple tests spanning Theory of Mind abilities, exposing language models to multiple sessions and variations in both structure and content, and implementing procedures to ensure a fair, non-superficial comparison between human and non-human minds (Firestone, 2020). This approach enabled us to reveal the existence and operation of specific mechanisms in artificial minds that would have remained hidden using a single Theory of Mind test, or a single run of each test.

Both GPT models exhibited impressive abilities to reason about beliefs, intentions, and non-literal utterances, with GPT-4 exceeding human levels in the Strange Stories. Both GPT-4 and GPT-3.5 failed only on the Faux Pas test. Conversely, LLaMA2-70B, which was otherwise the poorest performing model, outperformed humans on the Faux Pas. Understanding a faux pas involves two aspects: recognising that one person (the victim) feels insulted or upset and understanding that another person (the speaker) holds a mistaken belief or lacks some relevant knowledge. To examine the nature of models' successes and failures on this test, we developed and tested new variants of the Faux Pas test in a set of control experiments.

Our first control experiment using a likelihood framing of the belief question (Faux Pas Likelihood Test), showed that GPT-4, and to a lesser extent GPT-3.5, were able to infer the mental state of both the victim and the speaker and recognise that the most likely explanation involved the speaker not knowing or remembering the relevant knowledge that made their statement inappropriate. Despite this, both models consistently provided an incorrect response (at least when compared against human responses) when asked whether the speaker knew or remembered this knowledge, claiming that there was insufficient information provided. In line with the

Hyperconservatism hypothesis, these findings imply that while GPT models can recognise unintentional offence as the most likely explanation, they avoid committing to it. This finding is consistent with longitudinal evidence that GPT models have become more reluctant to answer opinion questions over time (Chen et al., 2023).

Further supporting that GPT's failures at recognising faux pas were due to an excess of caution in answering the belief question rather than a failure of inference, a second experiment using the Belief Likelihood test showed that GPT responses integrated information in the story to accurately interpret the speaker's mental state. When the utterance suggested that the speaker knew, GPT responses acknowledged the higher likelihood of the *"Knew"* explanation. LLaMA2-70B, on the other hand, did not differentiate between scenarios where the speaker was implied to know and when there was no information one way or another, raising the concern that the perfect performance of LLaMA2-70B on this task may be an illusion of understanding.

The pattern of failures and successes of GPT models on the Faux Pas test and its variants may be the result of their underlying architecture. In addition to transformers (generative algorithms that produce text output), GPT models also include mitigation measures to improve factuality and avoid overreliance (OpenAI, 2023a). These measures include training to reduce hallucinations, the propensity of GPT models to produce nonsensical content or fabricate details that are not true in relation to the provided content. Failure on the Faux Pas test may be an exercise of caution driven by these mitigation measures, as passing the test requires committing to an explanation that lacks full evidence. This caution can also explain differences between tasks: both the Faux Pas and Hinting tests require speculation to generate correct answers from incomplete information. However, while the Hinting task allows for open-ended generation of text in ways to which LLMs are well suited, answering the Faux Pas test requires going beyond this speculation to commit to a conclusion.

The cautionary epistemic stance of GPT models introduces a fundamental difference in the way that humans and GPT models react to social uncertainty (FeldmanHall & Shenhav, 2019). In humans thinking is, first and last, for the sake of doing (Fiske, 1992; James et al., 1981). Humans generally find uncertainty in social environments to be aversive and will incur additional performance costs in order to reduce it (Plate et al., 2023). Theory of Mind often plays a central role in reducing uncertainty, as the ability to reason about mental states - in combination with information about context, past experience, and knowledge of social norms - allows people to reduce uncertainty and commit to likely hypotheses, allowing for successful navigation of the social environment as active agents (Bonnefon & Rahwan, 2020; Frith & Frith, 2006). GPT models, on the other hand, exercise caution despite having access to the cognitive (or cognition-analogous) tools needed to reduce uncertainty, and so avoid engaging with uncertain options. The dissociation we describe between speculative reasoning and commitment mirrors recent evidence that while GPT models demonstrate sophisticated and accurate reasoning about belief states, they struggle to translate this reasoning into strategic decisions and actions (Zhou et al., 2023).

These findings highlight a dissociation between *competence* and *performance* (Firestone, 2020), suggesting that GPT models have the competence for mentalistic inferences but may refrain from using it under uncertain circumstances. Such a distinction can be difficult to capture with quantitative approaches that code only for target response features, as machine failures and successes are the result of non-human-like processes (Binz & Schulz, 2023; see *Appendix II, II.VII. Qualitative Analysis of Faux Pas Likelihood Test* for a preliminary qualitative breakdown of how GPT models' successes on the new version of the Faux Pas test may not necessarily reflect perfect or human-like reasoning).

While LLMs are designed to emulate human-like responses, this does not mean that this analogy extends to the underlying cognition giving rise to those responses (Bonnefon & Rahwan, 2020). In this context, our findings imply a difference in how humans and GPT models trade off

the costs associated with social uncertainty against the costs associated with prolonged deliberation (Hanks et al., 2011). This difference is perhaps not surprising considering that resolving uncertainty is a priority for brains adapted to deal with embodied decisions, such as deciding whether to approach or avoid, fight or flight, cooperate or defect. GPT models and other LLMs do not operate within an environment and do not need to resolve competition between action choices, so may have limited advantages in narrowing the future prediction space.

The dis-embodied cognition of GPT models can explain failures in recognising faux pas, but they may also underlie their success on other tests. One example is the False Belief test, one of the most widely used tools to date for testing the social cognitive abilities of LLMs (Brunet-Gouet et al., 2023; Bubeck et al., 2023; Dou, 2023; Kosinski, 2023; Sap et al., 2023; Ullman, 2023). In this test, participants are presented with a story where a character's belief about the world (the location of the item) differs from the participant's own belief. The challenge in these stories is not remembering where the character last saw the item, but rather in reconciling the incongruence between conflicting mental states. This is challenging for minds that have their own perspective, their own sense of self, and their own ability to track out-of-sight objects. However, if a mind does not have its own self-perspective because it does not need to navigate a body through an environment, as with GPT (Yiu et al., 2023), then tracking the belief of a character in a story does not pose the same challenge.

An important direction for future research will be to examine the impact of these mechanisms on second-person, real-time human-machine interactions (Redcay & Schilbach, 2019; Schilbach et al., 2013). Failure of commitment by GPT models, for example, may lead to negative affect in human conversational partners. However, it may also foster curiosity (FeldmanHall & Shenhav, 2019). Understanding how GPTs' mentalistic inferences (or their absences) influence human social cognition in dynamically unfolding social interactions is an open challenge for future work. As artificial intelligence continues to evolve it becomes increasingly important to heed calls

for open science and open access to these models (Frank, 2023). Allowing researchers direct access to the parameters, data, and documentation used to construct models can allow for targeted probing and experimentation into the key parameters affecting social reasoning. As such, these models can not only serve to accelerate the development of future AI technologies, but also serve as models of human cognition. This highlights the value of a methodology that uses tools from human cognitive psychology to compare powerful but black-boxed systems like ChatGPT with human data and open models like LLaMA2-Chat, a practice that will be increasingly pivotal in future studies.

Chapter 4. General Discussion

The objective of this chapter is to guide the reader through the theoretical and methodological implications of the findings presented in Chapter 2 and 3.

The primary aim of this dissertation was to explore mentalistic inference, the ability to deduct other people's mental states, specifically in the forms of reading intentions from movement kinematics and the application of Theory of Mind principles. The experimental samples, ASD children and GPT models, were compared to control samples, TD children and human agents.

We found that both TD and ASD children are sensitive to single-trial kinematics; the same group advantage suggests that internal readout models are tuned to the way we move. However, a selective impairment of ASD readers leads to misinterpreting informative variations in movement kinematics.

We also found that GPT models exhibit human-level performance across various ToM tests, validating their understanding of mental states and social dynamics. However, they struggle with the Faux Pas and their caution leads to avoid commitment to explanations in uncertain situations, possibly influenced by the models' mitigation measures.

In the following paragraphs, these results are contextualized within the frameworks of relevant theoretical constructs, providing insights that can inform and guide future research and applications.

4.1. Same-Group Advantage and Faulty Intention Reading in Autism Spectrum Disorders

In our first study, for both TD and ASD readers, the identification of intention-related variations relied on kinematic similarity. Our findings from single-trial analyses (see *Chapter 2, Figure 7B*) indicate that TD readers successfully identified intention-related variations when observing TD actions, but not when observing ASD actions. Conversely, ASD readers could identify intention-informative variations when observing ASD actions but not when observing TD actions. This same-group advantage aligns with the principle that internal readout models of TDs are finely tuned to typical actions, whereas internal readout models of ASDs are specifically tuned to autistic actions. These results, together with the evidence of impairments and atypical movement patterns being reported in motor coordination and planning in ASD children (see *Chapter 1, 1.2.1. Action observation in Autism Spectrum Disorders* for more details), suggest that internal readout models are tuned to the way we move. This linkage underscores the significance of sensorimotor processes and bodily experiences in shaping cognitive understanding, providing additional support for the overarching framework of embodied cognition in the perception and interpretation of actions.

Over time, researchers have delved into the concept of embodied cognition in the context of action, uncovering interference effects on self-generated actions (Grafton, 2009). These effects, supported by experiments involving the observation of limb postures or actions, suggest that knowledge about action is embodied, implying a form of simulation during action observation action (Craighero et al., 1999, 2002; Hamilton et al., 2004). The debate extends to whether mental or motor simulation serves as a means for decoding action understanding, with considerations for various levels of abstraction (Jacob & Jeannerod, 2005). Simulation can involve highly abstract mental inference, particularly prevalent in novel situations (Brass et al., 2007), or aligning perceived actions with internal models, including the more robust direct matching of kinematically

identical movements. Real-world limitations lead to a need for another form of simulation, comparing analogous movements that may not be identical. For example, observing actions from different perspectives can be simulated at a similar level of abstraction (Anquetil & Jeannerod, 2007).

When we observe someone else's actions, we not only recognize the observed actions but also simulate the corresponding motor and sensory experiences internally (Cook, 2016). Motor resonance is a key aspect of embodied simulation and refers to a phenomenon in which observing or perceiving someone else's actions activates corresponding motor representations in the observer's own motor system and plays a crucial role in understanding and imitating the actions of others, which is essential for social learning, empathy, and communication. Additionally, motor resonance is thought to play a role in comprehending the intentions and emotions of others through the internal simulation of their actions. (e.g., Brown & Marsden, 2001) and extensive evidence supports the inherent connection and reciprocal influence between the motor and visual systems in these contexts (Press & Cook, 2015).

Interestingly, theoretical frameworks examining the interplay between the visual and motor systems propose that individuals who closely resemble each other in their execution of actions are more prone to experiencing motor resonance when observing each other's actions (Friston et al., 2011; Kilner et al., 2007; Rizzolatti & Craighero, 2004; Wolpert et al., 2003), enhancing action perception (Casile & Giese, 2006) and prediction (Aglioti et al., 2008). The link between these notions and the spread motor atypicalities in the ASD domain children (see *Chapter 1, 1.2.1. Action observation in Autism Spectrum Disorders* for more details) is supported by previous studies finding that kinematic ASD atypicalities were positively correlated with symptom severity and the more atypical an ASD participant's kinematics, the less likely they were to classify TD movements as natural. These reinforce the link between the execution of atypical movements and biased perception of natural movements in individuals with ASD (Cook et al., 2013). Furthermore,

earlier research posited challenges for children with ASD in translating visual information from observed actions into a motor execution program; these anomalies are proposed to play a role in, and are interconnected with, the challenges individuals with ASD face in social interactions (Becchio & Castiello, 2012).

Despite having a same-group advantage as described above, our findings also indicate that ASD children exhibit a specific difficulty in interpreting variations in movement kinematics that carry informative cues. This selective impairment suggests that, among ASD readers, there is a notable challenge in accurately perceiving and understanding the subtle changes in movement patterns that convey meaningful information. This may contribute to a broader comprehension issue related to the interpretation of subtle cues in movement, impacting the overall understanding of dynamic and informative aspects of social interactions.

Individuals who struggle to deal with unexpected or unknown situations are often described as having an intolerance of uncertainty. On the other hand, individuals with ASD frequently express a preference for certainty and the lack of fixed reference points may lead them to encounter heightened levels of anxiety. This has an impact on daily lives, as suggested by research showing significant correlations between anxiety and intolerance of uncertainty indexes (Jenkinson et al., 2020).

Bayesian and predictive coding theories of perception and cognition may frame these results (Knill & Richards, 1996; Rao & Ballard, 1999). These theories propose that perception results from a dynamic interplay between bottom-up sensory signals and top-down internal models based on prior knowledge. External information, conveyed through bottom-up signals, moves from specific to general levels in cognitive processing, while internal models influence perception in a top-down manner. Bayesian reasoning encourages individuals to express their beliefs in terms of probabilities and to update these probabilities as new evidence becomes available; it provides a framework for updating beliefs in a rational and probabilistic manner as new information becomes

available. In Bayesian terms, updating beliefs involves adjusting probabilities based on new evidence.

The *Imbalance Hypothesis* framework (Brock, 2012; Lawson et al., 2014; Pellicano & Burr, 2012; Van de Cruys et al., 2014), arising from a Bayesian perspective, suggests that there is an imbalance between the influence of priors/predictions and likelihoods/prediction errors in autistic perception and cognition. Particularly, the *Imbalance Hypothesis* posits that the brain in ASD prioritize bottom-up information over top-down knowledge (Chrysaitis & Seriès, 2023).

Within this framework, the *High Inflexible Precision to Prediction Errors in Autism (HIPPEA) theory* (Van de Cruys et al., 2014) suggests that individuals with ASD have an inflexible high precision setting of prediction errors, leading to the development of very precise predictions, and creating a cycle of more prediction errors. Palmer and colleagues dispute the idea of a consistently high and inflexible precision setting, proposing that precision setting in ASD is context-dependent; differences in the ability to predict uncertainty in the environment are more consistent findings. In unpredictable environments, individuals with ASD tend to prioritize prediction errors and sensory information, challenging the notion of a rigidly high precision setting (Palmer et al., 2017). From a *HIPPEA* perspective, this selective impairment that we find in our study could be attributed to an inflexible high precision setting, where ASD individuals struggle to adapt their predictions to the dynamic and subtle differences in kinematic features entangled in the movements. The rigid precision setting may hinder their ability to accurately perceive and elaborate the informative aspects of social interactions, leading to difficulties in understanding the variations of movement kinematics.

On the other hand, considering the context-dependent precision setting proposed, we might interpret the results as indicative of challenges in predicting uncertainty specifically related to movement cues. In unpredictable environments, individuals with ASD may prioritize prediction errors and sensory information, contributing to difficulties in accurately perceiving and understanding the variations in movement kinematics. Findings in our study can be discussed

within a nuanced framework that considers both the inflexible high precision setting suggested by the *HIPPEA* theory and the context-dependent precision setting proposed. This approach allows for a more comprehensive understanding of the challenges faced by individuals with ASD in interpreting subtle variations in movement kinematics and contributes to the broader comprehension related to social interactions in this specific population.

4.2. GPT models' Failures in the Faux Pas Test

The findings from our study on GPT models show that an artificial agent has the capability to discern the social implications conveyed within a narrative presented in textual form, performing with human-like proficiency in the False Belief test, Irony comprehension test, and Hinting task. Remarkably, in certain instances, the artificial agent surpasses human performance, particularly evident in scenarios involving the Strange Stories. This was true for all tests except for the Faux Pas test. On this test, both GPT models failed.

The embodied cognition theory offers an interesting perspective to interpret this failure. The aim of the Faux Pas test is to assess the ability to recognize and understand social instances where a speaker unintentionally makes a social blunder, and the listener must grasp the implicit meaning or unintended offense. In this test, the mentalistic explanation necessitates complex computations, specifically involving second-order representations of mental states where one agent's goal depends on another agent's goal.

The model's performance could be hindered by its limited ability to grasp the full context of a faux pas, including the broader context in which the statement or action occurs. Human participants draw on their extensive real-world experiences and cultural knowledge to interpret such scenarios accurately, which may be beyond the model's capabilities. Thus, the Faux Pas test exposes a notable challenge for artificial intelligence, as exemplified by the performance of GPT. This deficiency in the model's capability to match human performance in this specific test can be elucidated through the perspective of embodied cognition theory.

The *embodied cognition* framework, mentioned above, emerged in the field of artificial intelligence (R. A. Brooks, 1995) and suggesting that cognitive processes are grounded and rooted in the body and shaped by its interaction with the surrounding world. According to this view, human cognition has profound origins in the processing of sensorimotor experiences and understanding the mind necessitates examining its association with a physical body actively

engaging with the environment it is nestled in (Wilson, 2002). Moreover, *Cognition is for action* (Smith & Semin, 2004; Wilson, 2002) and the mind is conceptualized in action-oriented terms, characterized by inner structures that function as operators influencing the external world through their role in shaping actions (Clark, 1998). Cognition is embodied action because it is contingent upon the types of experiences derived from being in a body with diverse sensorimotor skills which are intricately woven into a broader cultural, psychological and biological context and because sensory and motor processes, along with perception and action, are inseparable components in the realm of lived cognition (Varela et al., 2017). Cognition being embodied means that it emerges from the interaction of the body with the external world. This viewpoint underscores the impact of experiences rooted in the body's unique perceptual and motor abilities, which collectively shape the framework for reasoning, memory, emotion, language, and all aspects of mental life (Thelen et al., 2001).

When considered within a social framework, the concept of Socially Situated Cognition becomes relevant (Semin & Smith, 2013; Smith & Semin, 2004). This perspective posits that communication is inherently action-oriented, aiming to achieve specific objectives in the social realm. This becomes evident when examining various speech acts, such as requests, commands, persuasive endeavours, or questions, where the speaker explicitly seeks to influence the recipient' beliefs, attitudes, or behaviours. Even when the objective is not overtly apparent, other forms of speech still carry implications for social action. Conversations frequently entail the exchange of information that can enhance adaptive action in subsequent situations. Thus, communication is not seen as an impartial, objective translation or expression of internal representations (Smith & Semin, 2004). In the realm of language understanding, these theories suggest that comprehension is not solely a product of abstract mental processes but is profoundly influenced by sensory experiences, motor functions, and the contextual environment in which language is employed.

Contextual understanding and the assimilation of real-world experiences emerge as pivotal elements in explaining GPT's failure in the Faux Pas test. Human cognition, enriched by sensory

and motor engagement, often relies on facial expressions, body language, and emotional tone for comprehension. GPT, lacking sensory experiences and physical interactions, encounters challenges in capturing these embodied aspects of cognition.

In this framework, the temporal dynamics inherent in human experiences also play a significant role. While human experiences unfold over time, allowing for the development of a temporal context, GPT processes text in an instantaneous manner (Konvalinka et al., 2023). Faux pas situations often involve a sequence of events and reactions, and a temporal understanding is crucial for interpreting the impact of a statement or action and GPT may struggle to capture these temporal dynamics. As in illustration, consider the example item reported in paragraph 3.3.2:

> *Kim helped her Mum make an apple pie for her uncle when he came to visit. She carried it out of the kitchen. "I made it just for you," said Kim. "Mmm", replied Uncle Tom, "That looks lovely. I love pies, except for apple, of course!"*

Grasping the temporal dynamics may be crucial in shedding light on the character's inadvertent social misstep. Uncle Tom's remark stemmed from a specific circumstance - Kim offering him a homemade apple pie. His lack of awareness becomes evident as he expresses a dislike for apple pies, oblivious to the fact that the one crafted by Kim indeed contains apples. Had he been attuned to the context, he might have chosen his words more thoughtfully. Notably, the narrative does not suggest that he is incapable of recognizing the pie as an apple pie; rather, it highlights his temporary reality in making an unintentionally insensitive comment.

Moreover, the concept of embodied simulation becomes relevant when considering the human ability to draw upon a rich repository of real-world experiences for understanding appropriate behaviour and social norms. Human interpretation of situations, like recognizing and responding to faux pas, relies on the nuanced understanding developed through embodied experiences in diverse cultural and social contexts. Embodied simulation posits that mental representations are often grounded in the simulation of sensory and motor experiences associated with specific situations. In contrast, GPT, primarily trained on textual data, may lack the breadth

of experiential inputs that humans naturally integrate into their understanding. The limitations of GPT in this regard highlight the difference between human cognition, which is deeply intertwined with embodied experiences, and artificial intelligence models that may not possess the same embodied simulation capabilities.

GPT models not only lack the necessity for physical interaction in the real world and the urgency to make immediate decisions to take action but are also not embedded in a social context that inherently includes uncertainty. Social contexts are permeated by uncertainty, whose reduction may involve behavioural means, such as trying novel approaches, or cognitive strategies, such as envisioning potential outcomes. In interpersonal interactions, the presence of significant uncertainty is commonplace, encompassing unknown emotional states, past experiences, and future behaviours of others (Plate et al., 2023).

Our *Hyperconservatism hypothesis* suggests that GPT models can deduce characters' mental states but tend to refrain from committing to a single explanation due to excessive caution. This cautious approach to knowledge sets humans and GPT models apart in responding to social uncertainty (FeldmanHall & Shenhav, 2019; Plate et al., 2023). While humans actively seek to reduce uncertainty using tools like Theory of Mind, GPT models exercise caution and avoid engaging with uncertain options. A modified Faux Pas test supports the *Hyperconservatism hypothesis*, demonstrating that GPT models can infer the speaker's mental states, particularly recognizing unintentional offenses. Failures in responding correctly are attributed to an excessively conservative approach rather than an inability to consider plausible alternatives.

4.3. Future directions

Building on the research findings from the studies presented above, this paragraph aims to contribute to set a possible path for further investigations.

4.3.1. Training Interventions

The difficulties found in our study in interpreting variations in movement kinematics among ASD children highlights the need for targeted interventions to enhance their social cognition skills. An effective approach may involve implementing interventions aimed at developing intention reading abilities, focusing on the subtle changes intrinsic to the dynamic movement patterns.

We designed a protocol to test training interventions for the enhancement of intention discrimination abilities. Building upon the insights gained from the findings presented in the first study in this dissertation (see *Chapter 2*), the compelling need for a personalized approach becomes evident. Thus, the design described below has the aim of being a pilot study to better inform more appropriate variations needed to be explored further. The experimental design is composed of 3 parts: pre-test, training, and post-test. In the pre-test phase, 2 blocks of 100 trials, we ask participants to observe reach-to-grasp actions performed either with the intentions *to drink* from the bottle or *to pour* some water from it. Similar to the study discussed previously, their task is to indicate which intention the person in the video is going to pursue: *to drink* or *to pour*. In the training phase, 1 block of 100 trials, after their response, participants are exposed to the full video, including the drinking, or pouring phase; this kind of feedback, supposed to induce perceptual learning was chose both to give visual feedback of the response and to expose participants to the full kinematics of the movement. In the post-test, in 2 blocks of 100 trials, participants completed the same task as in the pre-test. The most representative videos for the two intentions were selected as stimuli, obtaining 2 datasets of 100 videos (50 to-drink and 50 to-pour) with comparable

kinematic properties: one dataset was used for pre- and post-test and the other dataset was used for training. Videos were obtained with the same procedure of the study in Chapter 2 but in this case they all showed reaching actions performed by neurotypical adults. In this pilot, we collected data from 10 neurotypical adults (4 males, mean age ± SD = 24.3 ± 2.75 y).

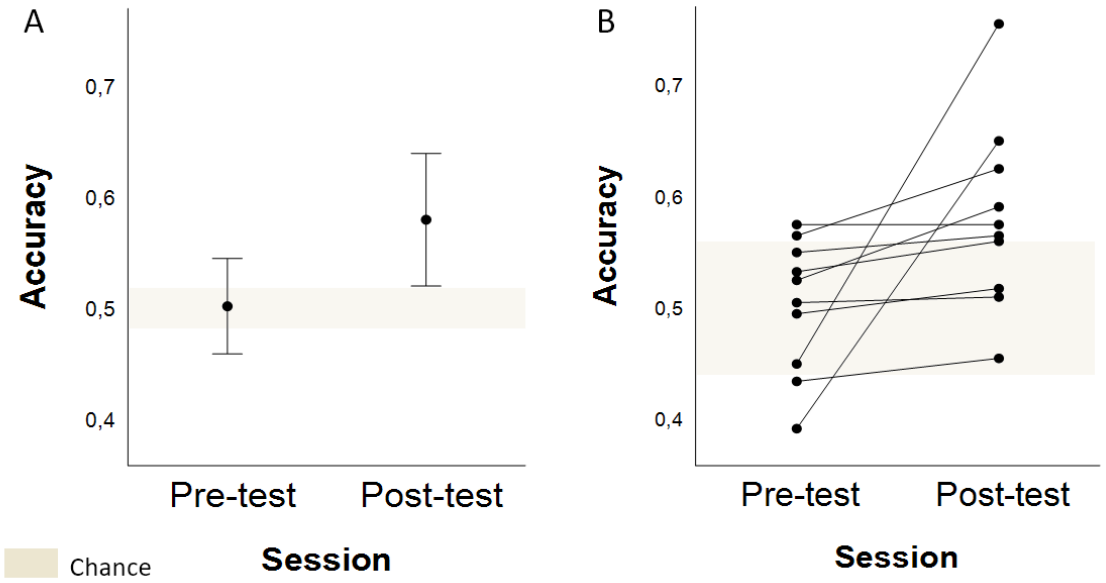

Figure 10. Results of the pilot study investigating the effect of a training protocol on our intention reading paradigm. (A) Group-averaged intention discrimination performance. (B) Single-subject data points. The beige area depicts chance, resulting from binomial tests, adapted with the number of trials performed.

Results at the group level show that, on average, participants performed at chance in the pre-test and above chance at the post-test (Figure 10A). However, when looking at the single-subject performances (Figure 10B), we can appreciate the inter-individual differences: at the pre-test, 2 participants perform below chance (bad readers), 2 perform above chance (good readers) and 8 at chance; at the post-test, 1 of the bad readers performs at chance, the other bad reader becomes a good reader, 4 chance-performers become good readers and the 2 good readers keep being good readers.

These results confirm the need for a personalized approach, specifically tailored according to the intention reading skills assessed at pre-test. First, it would be crucial to differentiate between chance readers and non-readers; this distinction would facilitate the development of customized strategies aimed to improve specific mechanism. Second, based on individual skills, various training methods can be explored. For instance, to enhance the tuning of the observers' internal models with the kinematic patterns of the observed actions, a motor training may be effective on the performance in terms of deviance from chance, potentially leading to an increased overlap index. Alternatively, a visual training, like the one proposed in this pilot study, could improve performance in terms of accuracy, possibly resulting in an increased alignment index (see *Computation of overlap and alignment* in *2.2.3. Quantification and Statistical Analysis*). Third, when applying this paradigm to the ASD population, it would be beneficial to examine potential correlations between learning abilities and ASD phenotypes, based on communication and interactions skills. Finally, it would be necessary to compare the results with a control sample to evaluate the effectiveness of the intervention against any spontaneous performance improvements that could be attributed to factors such as repeated exposure to stimuli.

4.3.2. Towards More Ecological Paradigms

In the present study we used videos (Chapter 2) and textual input (Chapter 3) to prompt mentalistic inference. However, in real contexts, we are used to face environments rich of multi-sensorial and multi-dimensional information. We foresee the possibility to integrate our paradigms and export them in more ecological experimental settings.

The latest advancements in GPT-4 introducing the possibility to incorporate image inputs in addition to textual data, on one hand, and the attempts being conducted to integrate Chat-GPT with talking, drawing and editing features, on the other hand (Wu et al., 2023), suggest that an integration of models generating language and analysing visual inputs will be possible in the near future. In this landscape, a potential advancement in our paradigm may include combining visual

and textual contexts as inputs to investigate ToM in artificial intelligence. This approach aims to explore environments characterized by intricate details and foster a more equitable balance between bottom-up and top-down attention mechanisms. Notably, this method dispenses with the necessity for textual prompts to direct responders' attention towards specific details mentioned in the text. Consequently, this development is poised to generate more authentic scenes and entrust the artificial agent with the responsibility of directing attention towards the salient aspects of the stimuli.

An additional step in this direction involves leveraging this approach for simulating human behaviour and replicating intention discrimination tasks in artificial agents. This methodology holds potential for offering feedback to inform the refinement of our training protocols. For instance, conducting multiple simulations at high speeds with minimal resource requirements may allow us to test various iterations of intention-discrimination training. The validation of training can be achieved by assessing the proficiency of an artificial agent in relying on accurate kinematic features to yield correct responses, thereby serving as a litmus test.

The significance of these simulations lies not only in the meticulous adjustment of processes and mechanisms governing decision-making but also, and perhaps more crucially, in fine-tuning individual subject variations that capture unique idiosyncrasies. Consequently, this approach may exhibit adaptability to diverse clinical populations. Furthermore, it capitalizes on the recent advancements in GPT releases - by now, only available for textual inputs -, enabling users to partially train the model using their own data.

This approach may be used to assists clinicians in making accurate and timely diagnoses, leveraging data from various sources for a comprehensive understanding of individuals' behaviour. AIs may contribute to personalized assessments by learning from diverse cases and adapting evaluation criteria based on individual differences. This fine-tuned approach may enhance clinicians' understanding of strengths and challenges of the single child, leading to more effective intervention plans.

4.4. Conclusions

The ubiquity of interactions is essential when trying to understand what is necessary to achieve a smooth communication and which aspects make an interaction effective. Here, different models of relationships were considered, articulated through the observation or the description of actions performed by others.

The findings in our studies underscore the impact of sensorimotor experiences on understanding mentalistic inferences. The first study, revealing a same-group advantage in intention reading based on kinematic similarity, emphasises the significance of sensorimotor processes, in particular motor resonance, in interpreting actions. The second study, exposing GPT models' failure in the Faux Pas test, underlies its lack of embodied experiences and limited ability to grasp contextual nuances in social scenarios. These insights align with the broader framework that cognition, whether in humans or artificial agents, is deeply rooted in the body's interaction with the environment, emphasizing the crucial role of sensory and motor processes in shaping mental life and understanding social actions. *"We must perceive in order to move, but we must also move in order to perceive"* (Gibson, 2014), and given that perception can be seen as a sensorimotor experience (O'Regan & Noë, 2001), an agent can generate valuable data by having an active, self-controlled, sensing body that enables the creation or elicitation of suitable inputs.

Moreover, understanding others may not necessarily be the foremost goal in social interactions. Rather, the focus might be on predicting what the other agent is going to do next, thus ensuring a smooth flow of the interaction (Snyder & Cantor, 1998). Shifting the focus from individual agents to the interaction itself alters the perspective, leading to the attribution of failure to transition from being solely on one side or the other to being cantered (Bolis et al., 2017). In these contexts, examining interactions instead of isolated individual agents becomes fundamental. When failures are viewed in the context of the interplay between entities, the need for interventions targeting the whole social context becomes evident, potentially addressing the dynamics and

factors influencing the interaction, aiming to enhance cooperation, communication, and overall system functionality.

Within this framework, deepening the concept of human variability becomes crucial to tailor personalized interventions. Human interactions inherently exhibit diversity due to the variability in individuals' behaviours, perspectives, communication styles, beliefs, and desires. Recognizing that individuals differ in their cognitive processes, perceptions, and interpretations underscores the complexity of interactions. Investigations must acknowledge and accommodate this variability to be effective. A one-size-fits-all approach may fall short in addressing the nuanced aspects of human interactions.

In the realm of mentalistic inference, exploring the domains of human and artificial agents illuminates the multifaceted nature of human cognition, shedding light on which are the essential features required for smooth interaction. Understanding how individuals navigate social interactions may help to simplify the complexity of cognitive processes underpinning these mechanisms. Artificial intelligence, apparently striving to emulate human-like cognition, grapples with challenges in responding to but not comprehending the intricacies of mentalistic inference. As we advance technologies, it becomes crucial to integrate insights from clinical research to foster a more inclusive and adaptable approach to human-machine interactions. In the convergence of these domains, mentalistic inferences serves not only as a bridge between neurodiversity and technology but also as a reminder of the rich complexity of human cognition.

Appendix I. Supplementary Material for Chapter 2

I.I. Supplementary Methods

Data augmentation procedure for training the logistic regressions. The training was enhanced by a data-augmentation scheme based on small random deformations in the time dimension. Warping transformations are a well-established augmentation method for image datasets (Shorten & Khoshgoftaar, 2019). Here, we adapted this approach by applying the following simple transformation to the time variable t:

$$g_\delta(t) = 10 \cdot \left(\frac{t}{10}\right)^\delta$$

where δ is drawn from a uniform distribution over $\left[\frac{1}{1.5}, 1.5\right]$, independently for each kinematic vector.

We then assigned values to the 10 new artificially generated time points by interpolating the kinematic data between the original time points. This yields a warped version of the input data, with either the first ($\delta > 1$) or the last part ($\delta < 1$) of the movement more represented in the binning. Increasing the training set by this data augmentation procedure increased the cross-validated test-set accuracy of kinematic encoding models compared to encoding models without data augmentation (from .874 ± .005 to .935 ± .003, mean ± SEM across all cross-validation folds, $p < .001$, two-sided t test). For kinematic readout models, the increase in cross-validated test-set accuracy was marginal but still significant across observer groups and observed actions (from .588 ± .010 to .592 ± .010, mean ± SEM across all participants and action sets, $p < .001$, two-sided t-test).

I.II. Supplementary Figures and Tables

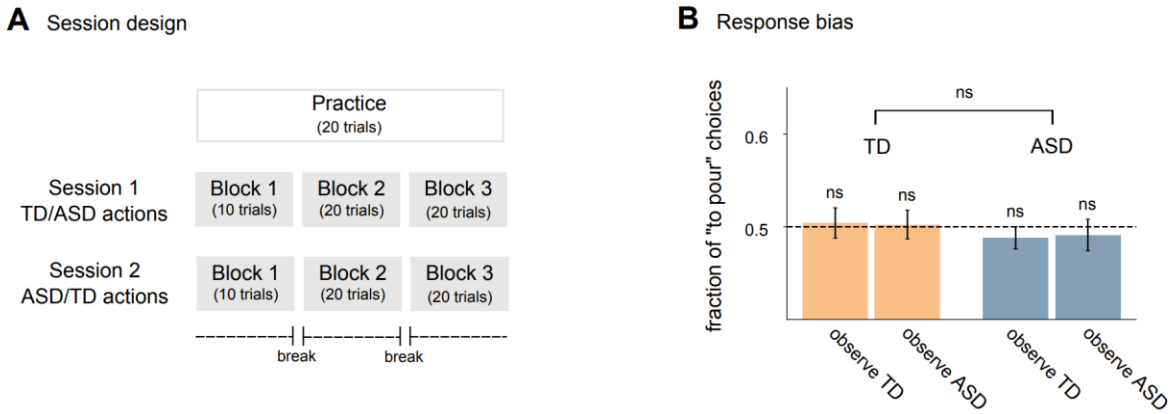

Figure S1. Session design and response bias. (A) After completing 20 practice trials, participants completed two sessions in which they observed reach-to-grasp actions performed by TD children and ASD children in counterbalanced order. Each session consisted of 50 experimental trials performed in three blocks (10, 20 and 20 trials), with a two-minute break between each block. (B) Fraction of '*to pour*' choices in the intention discrimination task. Results are reported as mean ± SEM across participants.

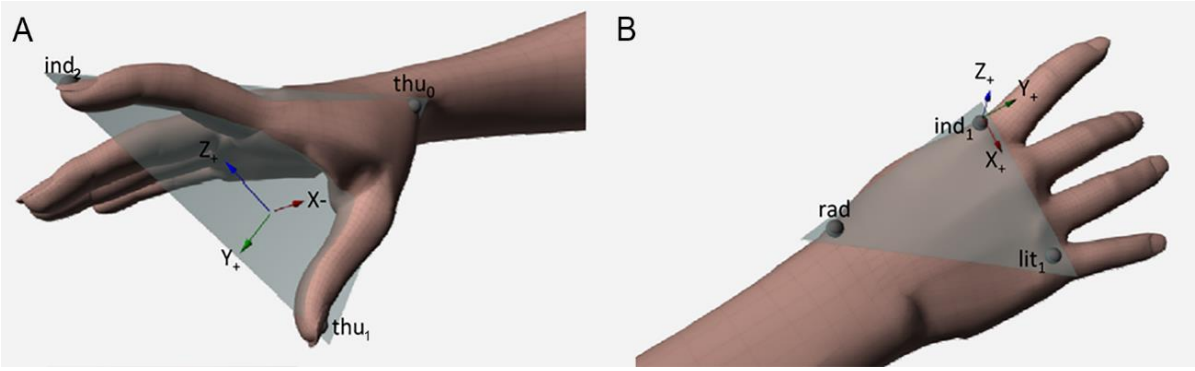

Figure S2. Lateral and frontal view of hand model. (A) Finger plane defined as the plane passing through markers 'thu0', 'thu1' and 'ind2'. (B) Dorsum plane defined as the plane passing through markers 'rad', 'ind1' and 'lit1'. X-, y- and z-thumb are defined as x-, y-, and z-coordinate of 'thu1' marker with respect to 'ind1' at reach-onset. X-, y- and z-index are defined as x-, y-, and z-coordinate of 'ind2' marker with respect to 'ind1' at reach-onset. Grip aperture is defined as the distance between 'thu1' and 'ind2'. Wrist velocity and wrist height are computed as the module of velocity and the z-component of the 'rad' marker. Additional markers (not used to compute the variables of interest) were placed on the tip of the little finger and at the centre of the hand dorsum.

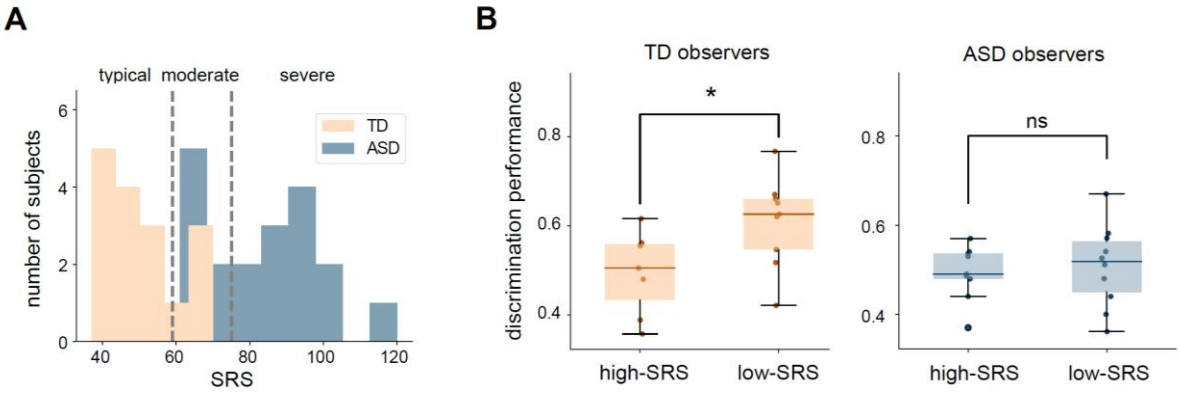

Figure S3. Relationship of intention discrimination performance to SRS. (A) Autistic traits were assessed in 19 ASD children and 16 TD children using the Social Responsiveness Scale (SRS) and were more prevalent in ASD compared to TD children (TD observers, mean ± SD = 50.0 ± 10.1; ASD observers, mean ± SD = 83.0 ± 15.6; $p < .001$), with some overlap between the two groups in the moderate range score. (B) To explore the relationship between intention discrimination performance and autistic traits in the non-clinical sample, we performed a median split on the TD data, such that TD observers were divided into high-SRS and low-SRS groups. Comparing intention discrimination performance between high-SRS and low-SRS groups revealed that intention discrimination performance was significantly lower in the low-SRS TD group compared to the high-SRS TD group ($p = .036$). The relationship between behavioural performance and autistic traits was therefore replicated in our non-clinical sample, between high-SRS and low-SRS participants. The median split on ASD data (such that ASD observers were divided into high-SRS and low-SRS) did not reveal a statistically significant difference.

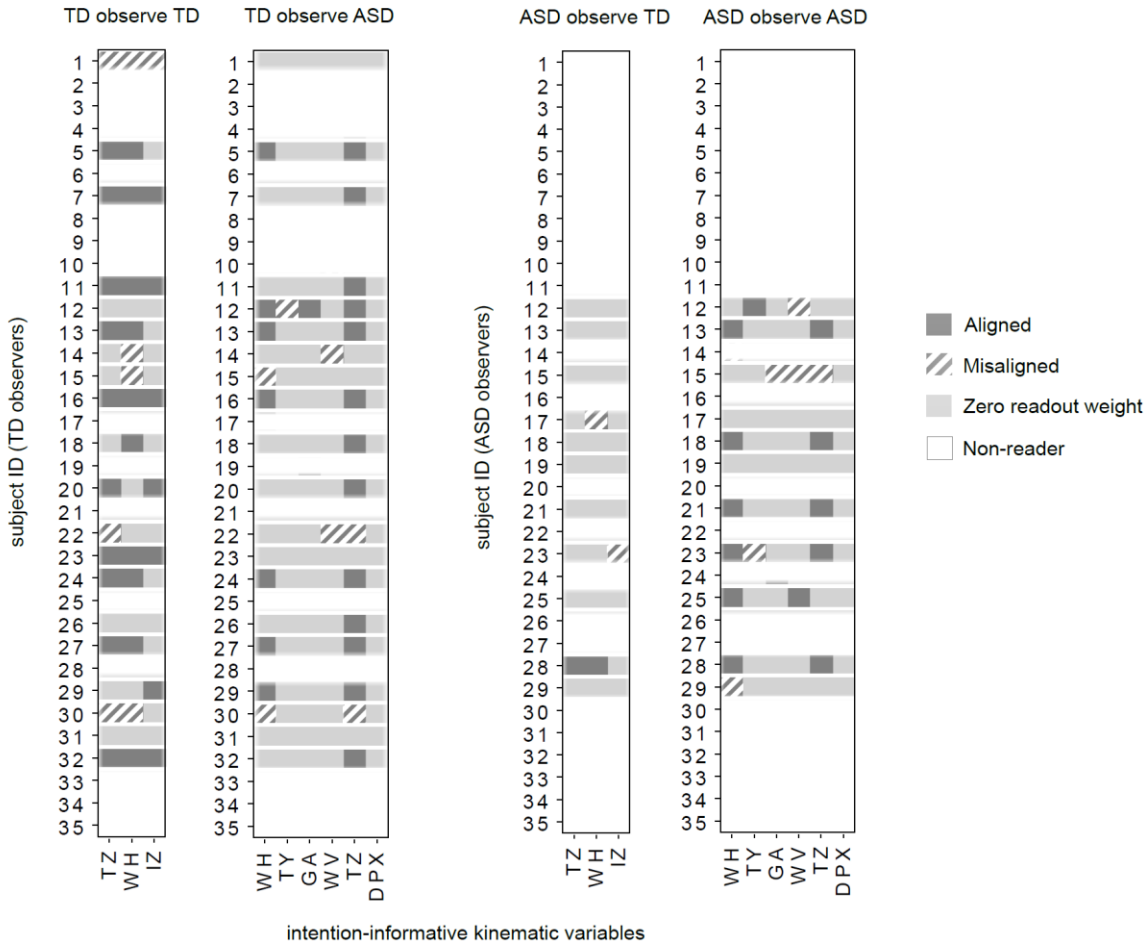

Figure S4. Readout weights assigned by individual observers to intention-informative kinematic features. We used the kinematic readout model to identify individual observers whose intention readout was sensitive to single-trial variations in movement kinematics (readers). We examined, separately for each reader, whether kinematic readout weights were assigned to intention informative kinematic features. For kinematic variables carrying intention information, we also examined whether the sign of the readout weight correctly aligned with the sign of the encoding weight.

Table S1. Clinical and demographic information of the TD group (N = 35). IQ: Intelligence Quotient (Wechsler, 2012). SRS: Social Responsiveness Scale (Constantino, 2013). ADOS: Autism Diagnostic Observation Schedule; ADOS_SA: Social Affection domain; ADOS_RRB: Restricted and Repetitive Behaviors domain (Lord et al., 2012). ADI-R: Autism Diagnostic Interview-Revised; ADI-R_A: Qualitative abnormalities in reciprocal social interaction subscale; ADI-R_B: Qualitative abnormalities in communication subscale; ADI-R_C: Restricted, repetitive, and stereotyped behavior patterns subscale; ADI-R_D: Developmental anomalies at or before 36 months subscale (Lord et al., 1994).

Subject	Sex	Age	IQ	SRS	ADOS	ADOS_SA	ADOS_RRB	ADI-R_A	ADI-R_B	ADI-R_C	ADI-R_D
S01	M	10	82		7	6	1	14	3	6	4
S02	M	9	119		8	6	2	12	9	7	2
S03	F	11	113		8	6	2	12	9	7	2
S04	M	11	89		9	8	1	18	19	3	1
S05	M	11	98	85	8	6	2	10	5	7	2
S06	F	9	85	97	10	8	2	11	8	5	5
S07	M	12	110		9	8	1	12	11	6	3
S08	M	11	88		9	8	1	10	8	5	2
S09	M	9	112		8	6	2	10	11	5	2
S10	M	8	100		8	6	2	11	9	4	4
S11	M	13	88		8	7	1	11	9	3	2
S12	M	10	92		8	7	1	20	15	10	4
S13	M	11	110		8	6	2	10	10	8	1
S14	M	9	107	81	8	7	1	10	8	5	1
S15	M	9	81		9	8	1	10	7	6	2
S16	M	11	104		8	7	1	12	8	3	2
S17	M	11	98		8	7	1	10	9	4	1
S18	M	11	97		8	7	1	8	17	5	1
S19	M	13	102		13	11	2	9	8	3	1
S20	M	11	91	94	8	7	1	11	11	5	3
S21	M	12	101	84	13	11	2	22	16	5	1
S22	M	9	87	64	12	9	3	16	11	7	1
S23	M	10	123	67	11	8	3	10	8	6	2
S24	F	10	96	95	9	8	1	13	8	4	1
S25	M	12	110	85	14	12	2	8	5	7	0
S26	M	9	86	65	8	6	2	10	8	6	5
S27	M	10	99	69	10	9	1	12	9	4	2
S28	F	9	103	82	9	7	2	11	10	4	2
S29	M	8	119	62	9	7	2	10	8	4	1
S30	F	8	101	98	8	6	2	15	8	5	1
S31	M	8	90	61	9	7	2	10	8	4	1
S32	M	11	92	93	18	13	5	18	18	6	5
S33	M	10	92	74	8	6	2	11	8	5	1
S34	M	11	109	101	13	10	3	13	9	5	2
S35	F	10	109	120	12	10	2	15	12	4	2

Table S2. Clinical and demographic information of the TD group (N = 35). IQ: Intelligence Quotient (Wechsler, 2012). SRS: Social Responsiveness Scale (Constantino, 2013).

Subject	Sex	Age	IQ	SRS
S40	F	12	114	
S41	M	10	104	
S42	F	9	115	
S43	M	10	104	
S44	M	12	111	
S45	M	12	89	
S46	M	12	100	
S47	M	9	94	
S48	F	10	115	
S49	M	11	109	
S50	F	10	99	
S51	M	9	110	
S52	M	9	112	
S53	M	9	102	
S54	M	9	83	
S55	M	10	105	
S56	M	9	106	
S57	M	9	106	59
S58	M	10	101	41
S59	M	9	114	47
S60	F	9	103	
S61	M	9	119	
S62	M	10	113	37
S63	F	12	119	48
S64	M	9	106	39
S65	M	9	102	54
S66	M	10	115	38
S67	M	8	106	53
S68	M	10	99	48
S69	M	10	95	65
S70	M	8	77	41
S71	M	10	106	44
S72	M	10	81	65
S73	M	10	108	55
S74	M	9	97	70

Table S3. Mixed Effects Model selection. We used Mixed Effects Models to assess the effects of Observer Group and Observed Action on intention discrimination performance and response bias (with Logistic Mixed Effects Models), and average readout model performance across CV repetitions (with Linear Mixed Effects Models). For intention discrimination performance, we also tested for an effect of session block and order to rule out significant learning effects. The notation Observer Group×Observed Action indicates both main effects of Observer Group and Observed Action and their interaction. Retained models are highlighted in bold.

Intention discrimination performance

Random effects structure selection (Fixed effects: Observer Group×Observed Action, Block, Order)					
Model	Random effects	df	BIC	Deviance	
$\mathbf{m_0^a}$	Subject (intercept and Observed action slope)	9	9506.6	9418.2	(singular fit)
$\mathbf{m_1^a}$	**Subject (intercept)**	**7**	**9498.2**	**9427.5**	
$\mathbf{m_2^a}$	null	6	9568.6	9506.7	
Fixed effects structure selection (Random effects: Subject intercept)					
Model	Fixed effects	df	BIC	Deviance	LRT
$\mathbf{m_1^a}$	Observer Group×Observed Action, Block, Order	7	9498.2	9427.5	
$\mathbf{m_3^a}$	Observer Group×Observed Action, Order	6	9480.8	9427.8	vs. m_1^a: $p > .05$
$\mathbf{m_4^a}$	Observer Group×Observed Action, Block	6	9489.3	9427.5	vs. m_1^a: $p > .05$
$\mathbf{m_5^a}$	Observer Group×Observed Action	5	9542.4	9507.0	vs. m_4^a, m_3^a : $p > .05$
$\mathbf{m_6^a}$	Observer Group, Observed Action	4	9463.6	9428.3	vs. m_5^a: $p > .05$
$\mathbf{m_7^a}$	Observed Action	3	9461.2	9434.7	vs. m_6^a: $p < .05$
$\mathbf{m_8^a}$	**Observer Group**	**3**	**9455.9**	**9429.4**	**vs. $\mathbf{m_6^a}$: $\boldsymbol{p > .05}$**
$\mathbf{m_9^a}$	null	2	9453.5	9435.8	vs. m_8^a: $p < .05$

Response bias

Random effects structure selection (Fixed effects: Observer Group×Observed Action)					
Model	Random effects	df	BIC	Deviance	
$\mathbf{m_0^b}$	Subject (intercept and Observed Action slope)	7	9551.0	9489.1	(singular fit)
$\mathbf{m_1^b}$	**Subject (intercept)**	**5**	**9534.6**	**9490.4**	
$\mathbf{m_2^b}$	null	4	9576.3	9540.9	
Fixed effects structure selection (Random effects: Subject (intercept))					
Model	Fixed effects	df	BIC	Deviance	LRT
$\mathbf{m_1^b}$	Observer Group×Observed Action	5	9534.6	9490.4	
$\mathbf{m_3^b}$	Observer Group, Observed Action	4	9526.5	9491.2	vs. m_1^b: $p > .05$
$\mathbf{m_4^b}$	Observed Action	3	9517.7	9491.2	vs. m_3^b: $p > .05$
$\mathbf{m_5^b}$	Observer Group	3	9517.8	9491.3	vs. m_3^b: $p > .05$
$\mathbf{m_6^b}$	**null**	**2**	**9509.0**	**9491.3**	**vs. $\mathbf{m_4^b}$, $\mathbf{m_5^b}$: $\boldsymbol{p > .05}$**

Readout model performance

Random effects structure selection (Fixed effects: Observer Group×Observed Action)					
Model	Random effects	df	BIC	Deviance	
$\mathbf{m_0^r}$	**Subject (intercept and Condition slope)**	**7**	**38250**	**38180**	
$\mathbf{m_1^r}$	Subject (intercept)	5	41105	41052	
Fixed effects structure selection (Random effects: Subject (intercept and Condition slope))					
Model	Fixed effects	df	BIC	Deviance	LRT
$\mathbf{m_0^r}$	Observer Group×Observed Action	7	38250	38180	
$\mathbf{m_2^r}$	Observer Group, Observed Action	6	38242	38180	vs. m_0^r: $p > .05$
$\mathbf{m_3^r}$	Observed Action	5	38239	38186	vs. m_2^r: $p < .05$
$\mathbf{m_4^r}$	**Observer Group**	**5**	**38235**	**38182**	**vs. $\mathbf{m_2^r}$: $\boldsymbol{p > .05}$**
$\mathbf{m_5^r}$	null	4	38232	38188	vs. m_4^r: $p < .05$

Table S4. Comparisons (significance) and coefficient values (effect size) for the retained Mixed Effects Models for intention discrimination performance, response bias and kinematic readout model performance. For the comparisons, we report the z-value computed with the Mixed Effects Model and the two-sided *p*-value computed from the z-test. All *p*-values are Holm-Bonferroni corrected for the number of comparisons listed for each entry. For the coefficient values, we report bootstrap confidence intervals for the estimates of regression coefficients and for the standard deviation of the random effects of the selected models computed using the R function *confint*. Confidence Interval (CI) values report the 95% (2.5% to 97.5%) confidence interval. Significant comparisons and effect sizes are highlighted in bold.

Intention discrimination performance (retained model: m_8^a)

Comparisons (significance)			Coefficient values (effect size)		
	z	*p*		Estimate	95% CI
TD Group – chance	**3.134**	**.003**	Global intercept	–.033	[–.161, .096]
ASD Group – chance	–.528	.597	**Observer Group**	**.233**	**[.055, .414]**
TD Group – ASD Group	**2.594**	**.009**	**Random intercept**	**.315**	**[.227, .379]**

Response bias (retained model: m_6^b)

Comparisons (significance)			Coefficient values (effect size)		
	z	*p*		Estimate	95% CI
All – chance	–.045	.964	Global intercept	–.002	[–.081, .074]
			Random intercept	.270	[.192, .330]

Readout model performance (retained model: m_4^r)

Comparisons (significance)			Coefficient values (effect size)		
	z	***p***		Estimate	95% CI
TD Group – chance	**5.605**	**<.001**	**Global intercept**	**27.959**	**[26.379, 29.754]**
ASD Group – chance	**2.422**	**.016**	**Observer Group**	**3.098**	**[.486, 5.384]**
TD Group – ASD Group	**2.470**	**.014**	**Random intercept**	**6.330**	**[5.362, 7.344]**
			Random slope	**5.560**	**[4.708, 6.776]**
			corr(Random intercept, Random slope)	**–.641**	**[–.766, –.477]**

Table S5. Relationship of intention discrimination performance to IQ. Participants in the TD and ASD groups were matched for IQ, however, it is still possible that within each group, intention discrimination performance was influenced by IQ. To rule out this possibility, as a first pass, we included Full Scale IQ as measured by the Wechsler Scale of Intelligence (WISC-IV) as fixed effect in Logistic Mixed Effects Models. Using a Likelihood Ratio Test (LRT), we verified that IQ scores had no significant effect on intention discrimination performance. As a second-pass analysis, we used a stepwise regression method to iteratively examine the potential explanatory role of each the four WISC-IV indices (Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index and Processing Speed Index) in intention discrimination performance. Estimates and standard errors (SE) of regression coefficients are reported for each predictor, as well as two-sided *p*-values obtained from t-tests for linear hypotheses. Stepwise regression did not reveal any significant relationship between discrimination performance and WISC-IV indices. The reported *p*-values were not corrected for multiple comparisons because none of the predictors had a raw *p*-value over the threshold for significant discrimination performance.

Fixed effects structure selection including Full Scale IQ

Model	Fixed effects	BIC	Deviance	LRT
$\boldsymbol{m_8^{a\times IQ}}$	Observer Group × IQ	9470.9	9426.7	
$\boldsymbol{m_8^{a,IQ}}$	Observer Group, IQ	9463.3	9428.0	vs. $m_8^{a\times IQ}$: p = .255
$\boldsymbol{m_8^{a}}$	Observer Group	9455.9	9429.4	vs. $m_8^{a,IQ}$: p = .241

Stepwise linear regression of discrimination performance against WISC-IV subscales

TD group	Coefficient ($\times 10^{-3}$)	SE ($\times 10^{-3}$)	*p*
Verbal Comprehension (VC)	1.00	1.92	.605
Perceptual Reasoning (PR)	1.70	1.58	.291
Working Memory (WM)	-0.50	1.59	.755
Processing Speed (PS)	.1.62	1.27	.211
ASD group	Coefficient ($\times 10^{-3}$)	SE ($\times 10^{-3}$)	*p*
Verbal Comprehension (VC)	0.18	0.88	.839
Perceptual Reasoning (PR)	0.20	0.92	.826
Working Memory (WM)	0.09	1.03	.934
Processing Speed (PS)	0.24	0.86	.780

Table S6. Summary of permutation tests. For each entry, we first report the value of the quantity to be tested (mean ± SEM). All *p*-values are computed comparing this value to the null-hypothesis distribution computed on permuted data. All *p*-values are Holm-Bonferroni corrected for the number of comparisons listed for each entry. For reference only, the z-scores of the tested values with respect to the mean and standard deviation of the null-hypothesis distribution are also reported.

Encoding models cross-validated accuracy against chance

	CV accuracy	***z* score**	***p***
TD actions	.955	4.841	<.001
ASD actions	.916	4.705	<.001

Readout models cross-validated accuracy against chance

	CV accuracy	***z* score (readout strength)**	***p***
TD observe TD	.614±.020	7.502	<.001
TD observe ASD	.626±.023	8.392	<.001
ASD observe TD	.553±.014	3.226	.003
ASD observe ASD	.580±.018	4.008	<.001

Overlap index against chance

	Overlap	***z* score**	***p***
TD observe TD	.783±.031	8.079	<.001
TD observe ASD	.592±.025	.970	.332
ASD observe TD	.573±.037	–.156	.332
ASD observe ASD	.655±.024	2.753	.010

Alignment index against chance

	Alignment	***z* score**	***p***
TD observe TD	.353±.120	6.078	<.001
TD observe ASD	.147±.078	2.544	.031
ASD observe TD	.086±.101	1.089	.556
ASD observe ASD	.038±.144	.524	.556

Table S7. Summary of binomial and non-parametric tests for proportions. For each entry, we report the value of the proportion to be tested. For comparisons against chance, the *p*-values are computed comparing this value to a null-hypothesis binomial distribution. For two-proportion comparisons, the *p*-values are computed comparing the value of the proportion to be tested to the null-hypothesis distribution computed on permuted data. For reference only, z-scores of the tested values with respect to the mean and standard deviation of the null-hypothesis distribution are also reported.

Readers / observers: between groups and against chance

	Proportion	***z***	***p***
TD Group – chance	20/35	14.154	<.001
ASD Group – chance	11/35	7.174	<.001
TD Group – ASD Group	20/35 – 11/35	2.153	.034

TD readers / TD observers: between actions and against chance

	Proportion	***z***	***p***
TD Actions – chance	14/35	9.501	<.001
ASD Actions – chance	17/35	11.827	<.001
TD Actions – ASD Actions	14/35 – 17/35	.713	.631

ASD readers / ASD observers: between actions and against chance

	Proportion	***z***	***p***
TD Actions – chance	5/35	2.521	.029
ASD Actions – chance	9/35	5.623	<.001
TD Actions – ASD Actions	5/35 – 9/35	1.187	.252

Good readers / readers: between groups and against chance

	Proportion	***z***	***p***
TD Group – chance	12/20	11.286	<.001
ASD Group – chance	3/11	3.389	.015
TD Group – ASD Group	12/20 – 3/11	3.166	.001

TD good readers / TD readers: between actions and against chance

	Proportion	***z***	***p***
TD Actions – chance	10/14	11.404	<.001
ASD Actions – chance	4/17	3.505	.009
TD Actions – ASD Actions	10/14 – 4/17	2.290	.018

ASD good readers / ASD readers: between actions and against chance

	Proportion	***z***	***p***
TD Actions – chance	1/5	1.539	.226
ASD Actions – chance	2/9	2.371	.071
TD Actions – ASD Actions	1/5 – 2/9	.554	.632

I.III. Data Availability

This study did not generate new unique reagents or materials. The code supporting the main results of this study is described in par. 2.2. Materials and Methods and has been deposited in GitHub (https://github.com/noemimontobbio/ASD_encoding_readout).

I.IV. Acknowledgments

We thank all participants and their families for their efforts to participate in the study. This research was supported by EnTimeMent EU H2020 FETPROACT Grant 824160 and by NIH BRAIN Initiative Grant R01NS109961. DiNOGMI contributed to this work within the framework of the DiNOGMI Department of Excellence MIUR 2018 to 2022 (*legge 232/2016*).

Appendix II. Supplementary Material for Chapter 3

II.I. Comparison of LLaMA2-Chat Models

We collected data on the full Theory of Mind battery for three LLaMA2-Chat models, sized at 7 billion (7B), 13 billion (13B), and 70 billion (70B) parameters. Performance of the three LLaMA2-Chat models is shown in Figure S5. Numerical values for statistical comparisons are reported in Table S8.

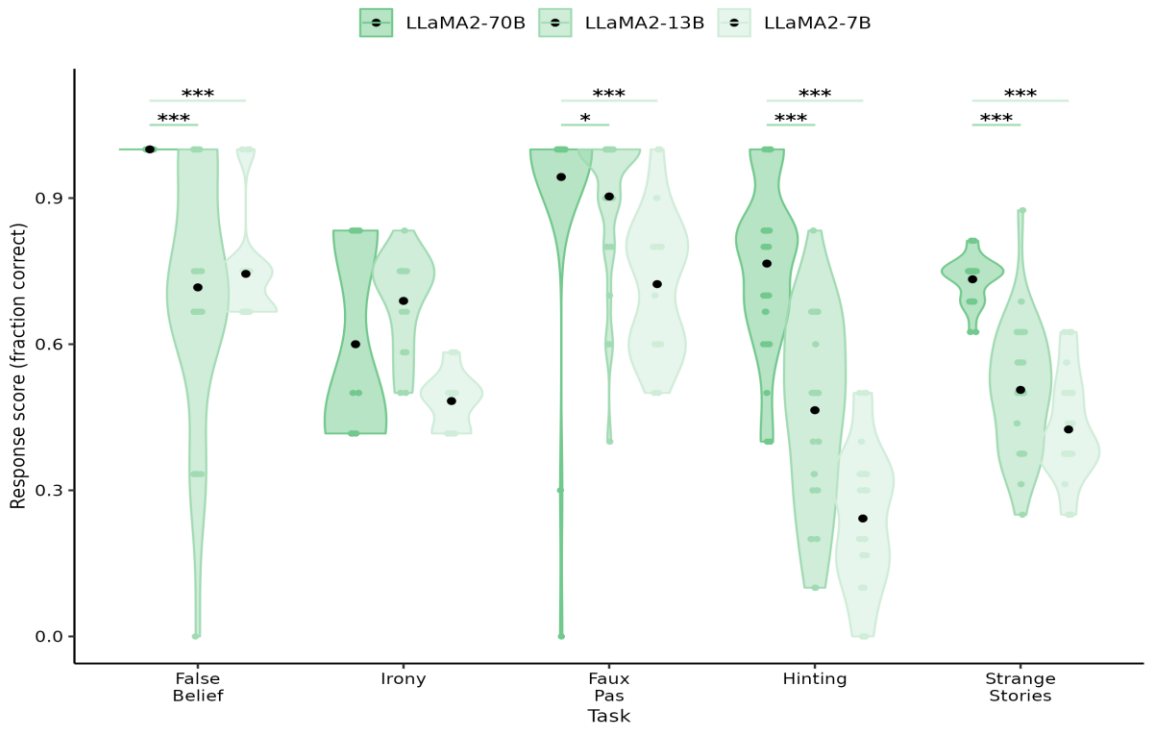

Figure S5. Violin plot: performance of three sizes of LLaMA2-Chat models.

Unlike the other models, both LLaMA2-7B and LLaMA2-13B provided a large proportion of non-answer responses (e.g., *"Sure! Go ahead and tell me the story"* despite the story and questions already being posed, or answers that cut off halfway through answering the question).

In total, approximately 22% of LLaMA2-7B responses and 10% of LLaMA2-13B responses were non-answers. To obtain a complete dataset, whenever we encountered a non-answer, we regenerated a response until we obtained a codable response. LLaMA2-70B significantly outperform the smaller models in all tests except in the Irony comprehension test.

Table S8. Pairwise comparisons (corrected Wilcoxon tests) of three LLaMA2-Chat models across tests in the Theory of Mind Battery.

Task	**Estimate**	**Model 1**	**Model 2**	**Statistic**	**CI (low)**	**CI (high)**	**p (corr.)**	**p < .05**
False Belief	0.29	LLaMA2-70B	LLaMA2-13B	225.00	0.29	0.33	0.00	***
False Belief	0.29	LLaMA2-70B	LLaMA2-7B	225.00	0.29	0.29	0.00	***
False Belief	-0.00	LLaMA2-13B	LLaMA2-7B	83.00	-0.04	0.00	0.54	
Irony	-0.08	LLaMA2-70B	LLaMA2-13B	90.00	-0.25	0.08	0.70	
Irony	0.00	LLaMA2-70B	LLaMA2-7B	132.00	-0.08	0.33	0.70	
Irony	0.25	LLaMA2-13B	LLaMA2-7B	211.00	0.17	0.25	0.00	***
Faux Pas	0.05	LLaMA2-70B	LLaMA2-13B	175.00	0.00	0.15	0.01	*
Faux Pas	0.25	LLaMA2-70B	LLaMA2-7B	210.00	0.20	0.30	0.00	***
Faux Pas	0.20	LLaMA2-13B	LLaMA2-7B	197.50	0.10	0.25	0.00	**
Hinting	0.30	LLaMA2-70B	LLaMA2-13B	225.00	0.22	0.37	0.00	***
Hinting	0.53	LLaMA2-70B	LLaMA2-7B	225.00	0.45	0.60	0.00	***
Hinting	0.23	LLaMA2-13B	LLaMA2-7B	207.00	0.13	0.32	0.00	***
Strange Stories	0.25	LLaMA2-70B	LLaMA2-13B	214.50	0.19	0.28	0.00	***
Strange Stories	0.31	LLaMA2-70B	LLaMA2-7B	225.00	0.28	0.34	0.00	***
Strange Stories	0.06	LLaMA2-13B	LLaMA2-7B	174.00	0.03	0.12	0.04	*

II.II. Variability of Performance across Test Items

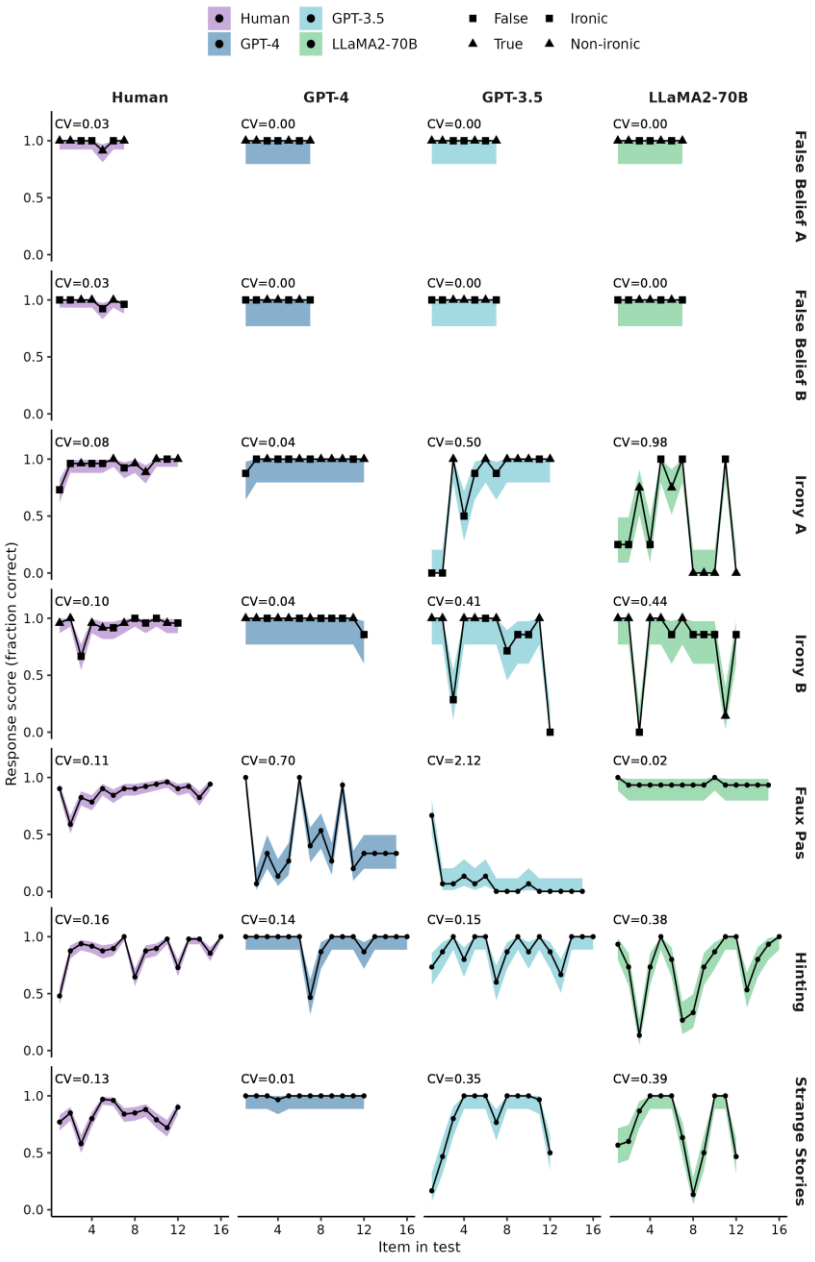

Figure S6. Means and 68% binomial confidence intervals across trials of the four experimental models on each item within the test. We report 68% Cis because they correspond to 1 standard deviation for Gaussian distributions. For each model and each task, the coefficient of variation (CV) is shown above the plot. A and B for the False Belief and Irony tasks are separated out in this figure into the two set lists used in the study on which the order or items remained the same but the trial state (False/True Belief; Ironic/Non-ironic) varied from trial to trial. For these two tests, trial state is denoted by point shape. Human responses to individual items on tests can be variable, as different people bring different intuitions or priors that affect their interpretation of particular stories.

Figure S6 shows a breakdown of individual item performance for all models across all tests included in our Theory of Mind Battery. Comparing LLMs and human item-wise performance revealed no systematic patterns where humans and LLMs systematically failed on the same items within a test.

To quantify the relative variability of human and model response scores across items, we computed, for each test and experimental model, the Coefficient of Variation (CV), that is, the ratio between the standard deviation of the mean response scores across items and the grand mean across items of the response scores. This analysis shows that while human responses were variable on some tests, there was low relative variability across test items within each test. For GPT-4, the CV in item-wise performance was also low on all tests except for the Faux Pas. GPT-3.5 and LLaMA2-70B showed higher CVs. Specifically, GPT-3.5 showed higher CV in item-wise performance on Irony, Faux Pas, and Strange Stories. LLaMA2-70B showed higher CV on Irony, Hinting, and Strange Stories.

II.III. Effects of Item Position

In the Theory of Mind Battery, each chat with LLMs was a separate and independent session, ruling out between-session order effects. However, since all models remember previous messages within an individual chat session, this introduces the potential for order effects driven by the position of an item within the session.

To test for order effects at the item level, we fit a binary logistic regression (quasibinomial for Strange Stories) to individual item scores on the original test items using item position as a predictor for each model on each test. Due to perfect performance in the False Belief test, this test was not included in this analysis. Results are shown in Table S9.

GPT-4 and LLaMA2-70B did not show any effects of item position across any test. GPT-3.5 showed significant item order effects on response scores for the Faux Pas, Strange Stories, and the Irony tests, but not for Hinting. For the Faux Pas test, the slop of the effect was negative such that GPT-3.5 performed worse on later items than on earlier ones, while for the Strange Stories and Irony tests, the slope was positive indicating that the model performed better on later than on earlier items.

Table S9. Effect of item position for each test.

Task	Model	Est	SE	Statistic	p	p (corr.)	p < .05
Irony	GPT-4	0.00	0.21	0.00	1.00	1.00	
Irony	GPT-3.5	0.18	0.06	3.13	0.00	0.02	*
Irony	LLaMA2-70B	-0.07	0.04	-1.58	0.11	0.80	
Faux Pas	GPT-4	0.08	0.06	1.42	0.16	0.94	
Faux Pas	GPT-3.5	-0.49	0.13	-3.72	0.00	0.00	**
Faux Pas	LLaMA2-70B	0.00	0.13	0.00	1.00	1.00	
Hinting	GPT-4	-0.24	0.13	-1.86	0.06	0.51	

Task	Model	Est	SE	Statistic	p	p (corr.)	p < .05
Hinting	GPT-3.5	0.05	0.09	0.55	0.58	1.00	
Hinting	LLaMA2-70B	-0.03	0.06	-0.54	0.59	1.00	
Strange Stories	GPT-4	0.10	0.44	0.22	0.83	1.00	
Strange Stories	GPT-3.5	0.78	0.17	4.68	0.00	0.00	***
Strange Stories	LLaMA2-70B	-0.15	0.07	-2.18	0.03	0.28	

These effects could indicate that item ordering influenced GPT-3.5's performance. However, because in the original testing protocols items were presented in the fixed order prescribed by the original validated version of each test (see *3.2. Materials and Methods*), they could also reflect difficulties related to specific items and their distribution within a given session. To isolate order effects from other item-specific effects, we collected another set of data with GPT-3.5 presenting items in a randomised order for each session on the Faux Pas, the Strange Stories, and the Irony Comprehension tests. To determine how many follow-up samples we need to collect, we conducted a power analysis using the learning effects identified with GPT-3.5. The most conservative effect size to use for estimating required sample size was for the Irony test. As such, we fit a power curve to estimate the number of necessary trials using the *powerSim* package in R that runs a number of simulations (n = 1000) over a range of sample sizes to estimate statistical power. This analysis indicated that 12 sessions would be sufficient to provide 80% power. The testing was identical to the protocol used for the Theory of Mind Battery with the exception that all items were presented in a randomised order and that for the Irony test only ironic items were included.

Fitting a (quasi-)binomial logistic regression to predict scores as a function of trial position revealed an order effect for the Irony test, whereby GPT-3.5 made more errors on earlier trials than later ones. In contrast, errors in Faux Pas and Strange Stories did not exhibit an order effect.

The results of this randomised order dataset are shown in Figure S7 and Table S10.

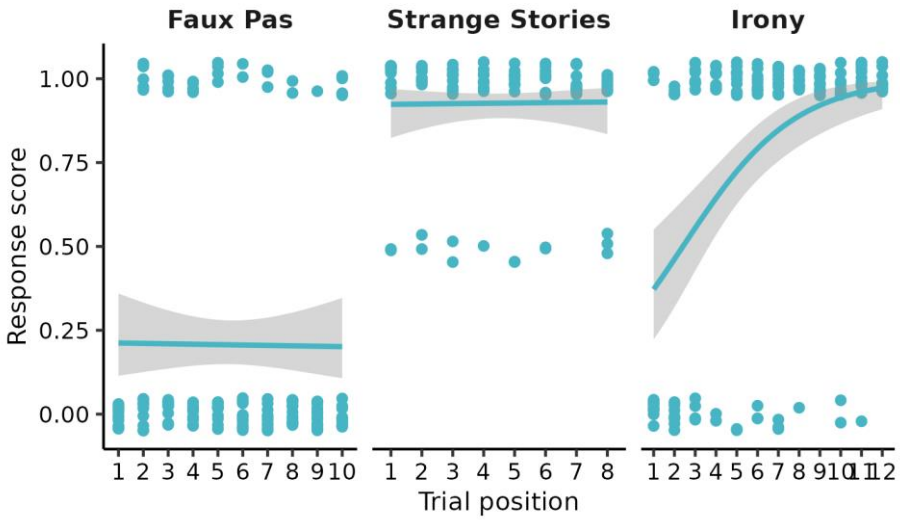

Figure S7. Effect of item position on randomized order items.

Table S10. Effect item position of shuffled data.

Task	Est	SE	Statistic	p	p (corr.)	p < .05
Faux Pas	-0.01	0.07	-0.11	0.92	1.00	
Irony	0.37	0.08	4.76	0.00	0.00	*
Strange Stories	0.01	0.12	0.12	0.90	1.00	

II.IV. False Belief Perturbations (adapted from Ullman 2023)

In humans, success on the False Belief task requires inhibiting one's own belief about reality to use one's knowledge about the character's mental state to derive predictions about their behaviour. However, with LLMs it is possible that performance on the False Belief task may be explained by lower level explanations than belief tracking. Supporting this interpretation, LLMs such as ChatGPT have been shown to be susceptible to minor alterations to the False Belief formulation (Ullman, 2023; Shapira et al., 2023) such as making the containers where the object is hidden transparent, or asking about the belief of the character who moved the object rather than the one who was out of the room. Such perturbations of the False Belief structure are assumed not to matter to entities that have Theory of Mind (Ullman, 2023). However, such an assumption has not been tested in humans. To compare the effects of these perturbations in humans and LLMs, we collected a new dataset of responses using five perturbations of three different False Belief stories. We also included a standard False Belief variant to replicate the models' ceiling performance. The perturbations, adapted from Ullman (2023), we used as follows:

- False Belief. The standard formulation as presented in the test battery. Example:
  In the kitchen there are Lucy, Mia, a carton of orange juice, a fridge, and a cupboard. Lucy puts the carton of orange juice in the cupboard. She then leaves the kitchen and goes to school. While Lucy is away, Mia takes the orange juice out of the cupboard and puts it in the fridge. Mia leaves the room and goes to work. Lucy comes back from school and enters the kitchen. She doesn't know what happened in the kitchen when she was away. When Lucy comes back home, where will she look for the orange juice?

- Transparent. The containers in which the object was hidden were made of transparent plastic or glass so that the actor would not have to open them to see what was inside them.
  Rather than a cupboard and a fridge, the containers are a transparent plastic box and a glass-fronted cabinet such that the juice can be readily seen inside.

- Preposition. The preposition in the stories was changed such that the object was no longer obscured (e.g. “in the box” became “on the box”).
  Rather than putting the juice carton in the cupboard, Lucy puts it on the cupboard, and similarly with Mia moving it to on the fridge. When Lucy enters the room, the story describes her looking around, so she should be able to see the juice immediately.
- Testimony. The agent who moved the object (the Mover) told the target character that they were going to move the object.
  All details are as in the original, except that Mia calls/texts Lucy to tell her that she is going to move the juice and Lucy believes her.
- Mover. The question asked about the belief of the Mover rather than the character who was out of the room.
  All details are as in the original except that the question asks where Mia will look for the juice.

We adapted three False Belief stories to generate variants for each, resulting in 15 new stories (for the full text, see *Appendix II, II.VIII.I. False Belief Perturbations*, below). To control for any cross-influence between variants, we elected to test each item separately in a different chat for each LLM (n = 15 repetitions per item), and with a new sample of ~50 humans (total N = 757). The results of these variants are shown in

Figure S8.

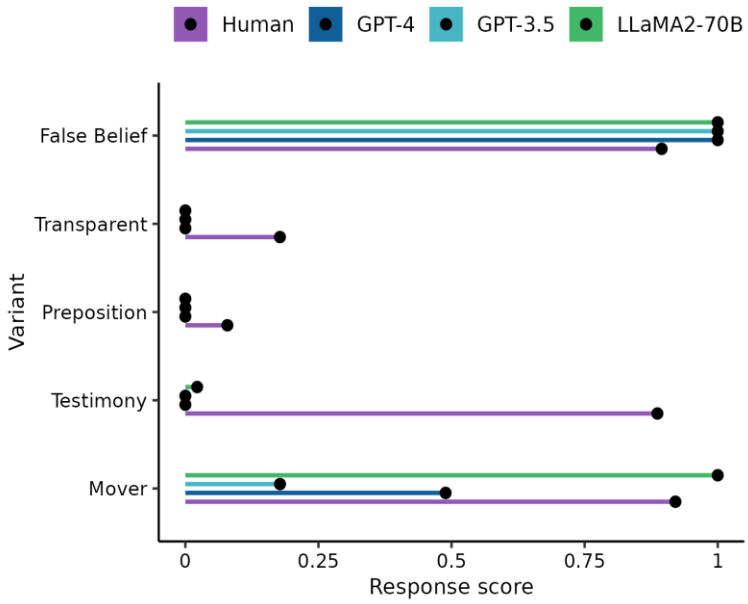

Figure S8. Performance of LLMs and humans across perturbations of the False Belief task.

We replicated the poor performance of GPT models observed by previous studies (e.g. Ullman, 2023), with both GPT-3.5 and GPT-4 failing on Transparent, Preposition and Testimony perturbations. LLaMA2-70B performed similarly poorly, although it consistently passed the Mover variant. Contradicting the assumption that these perturbations do not affect entities that have a Theory of Mind, humans also failed on Transparent and Preposition perturbations. Similar to LLMs, when the story involved transparent containers or changes to prepositions, humans were also likely to report that a character would look for the object where they left it.

It is worth noting that these control variants present diverse challenges that go beyond tracking mental states, and may involve understanding physical properties, relationships between objects, and spatial reasoning capabilities. They also differ in terms of the type of belief updating: the variants where humans performed 'poorly' according to the intuitions proposed by Ullman are those where the character's belief can only be updated after they return to the room, while other variants where performance is more successful involve manipulations of belief states that exist

prior to returning. These results highlight the need for rigorous investigation that includes human validation and systematic manipulation of factors that are relevant to Theory of Mind.

II.V. Faux Pas: Coding Strategies

The Faux Pas task consists of vignettes describing an interaction where a speaker says something they should not have said, not knowing or not realising that they should not have said it. To understand that a faux pas has occurred, one must recognise this lack of knowledge or realisation. The coding strategy reported in the main manuscript focusses on this element by coding responses on the basis of how participants (LLMs or humans) respond to the fourth comprehension question: *"Did [the speaker] know/realise/remember [the information that made their statement inappropriate]?"* To be coded as correct, the response to this question has to commit to the correct answer (*"No"*). We focused on this question because this was the key question about mental states that determined the interpretation of the faux pas.

This methodological choice of coding strategy is important, and it does reflect a departure from the strategy described in Baron-Cohen et al. (1999), where participants must answer all four comprehension questions correctly to pass the test. Here, we report the results where the same responses were coded with this strategy. Furthermore, we also consider an alternative coding strategy. We adopted a strict strategy where responses to the final question that equivocated or expressed an uncertainty were not marked as correct. As an exploratory analysis, we recoded responses where the correct answer was mentioned as a plausible alternative but was not explicitly endorsed as correct rather than incorrect, to see if the poor performance of GPT was driven by our penalising uncertainty.

Four-question coding. The results of the four-question coding scheme were consistent with those reported in the main manuscript (Figure S9).

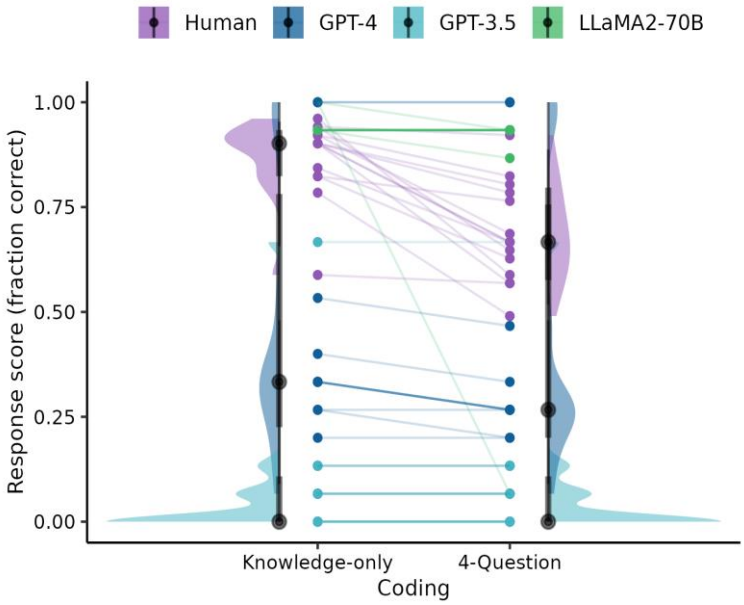

Figure S9. Four-question coding strategy. Side-by-side comparison of Human and LLM performance on the Faux Pas test using the knowledge-only coding criteria (*"Did they know...?"* question only) and the four-question coding criteria (all four questions coded as correct). Individual data points and lines show data aggregated by test item across sessions. Density plots are not shown for LLaMA2-70B as there was too little variability.

The performance of LLMs was largely unchanged under the four-question coding scheme. For humans, the scores were significantly lower under the four-question coding scheme than under the knowledge-only scheme. Upon examination of the responses, this was driven by responses to the first comprehension question: *"In the story, did someone say something they should not have said?"* The goal of this question is to ensure that participants recognise that the speaker's utterance could cause hurt or offence to the victim, and as such responses were marked correct only if participants responded, *"Yes"*. However, a sizeable minority of human participants appeared to interpret this question as one of moral judgement, and used the speaker's lack of knowledge as justification for why they were not "in the wrong" for saying what they did (e.g. *"no he didn't say anything wrong because he didn't know"*). Furthermore, despite answering no to the first question, human participants could frequently identify the offensive statement when prompted (*"Nothing*

'wrong', but if you're asking the question, probably that he doesn't like apple") and reliably recognised that the speaker was not aware of the context.

To verify whether this reduction of the human scores affected our conclusions, we compared human and GPT responses under the four-question coding scheme. As shown in Table S11, despite higher error rates under the four-question coding scheme than the knowledge-only, humans still performed significantly better at the task than both GPT models, and LLaMA2-70B continued to perform better than humans overall.

Table S11. Comparison of LLMs against humans under four-question coding.

Estimate	Reference	Model	Statistic	CI (low)	CI (high)	p (corr.)	p < .05
0.27	Human	GPT-4	595.00	0.13	0.60	0.00	**
0.67	Human	GPT-3.5	745.50	0.47	0.80	0.00	***
-0.13	Human	LLaMA2-70B	205.50	-0.33	-0.00	0.01	**

Alternative Coding Scheme. The uncertainty of GPT-3.5 and GPT-4 in answering the Faux Pas questions was frequently attributed to the answer not being present or directly mentioned within the story (*"It is not clear from the story whether [they] knew"*). Responses to some items indicated that GPT models could consider the correct answer as plausible but did not consider it more plausible than other alternatives (*"it could be that [they] did not know, or that [they] knew and were just expressing an opinion"*). The coding criteria for this task were strict such that responses to the two-alternative question, *"Did [the Speaker] know...?"* were only coded as correct if they committed to the answer *'No'*. It is possible that this strict *Commit* coding approach penalized the performance of GPT models. In order to control for this, we recoded the responses of both GPT models and LLaMA2-70B the original Faux Pas task to mark as correct any responses that acknowledge consideration of the correct answer (*'No, the Speaker did not know/remember the context'*), even if they did not commit to it (e.g., *'The Speaker might not have remembered the context, or they might have remembered'* would have been marked incorrect under the first

(*Commit*) coding scheme and correct under the new (*Consider*) one). As shown in Figure S10, this recoding resulted in marginal improvements in score that did not significantly affect the overall task performance.

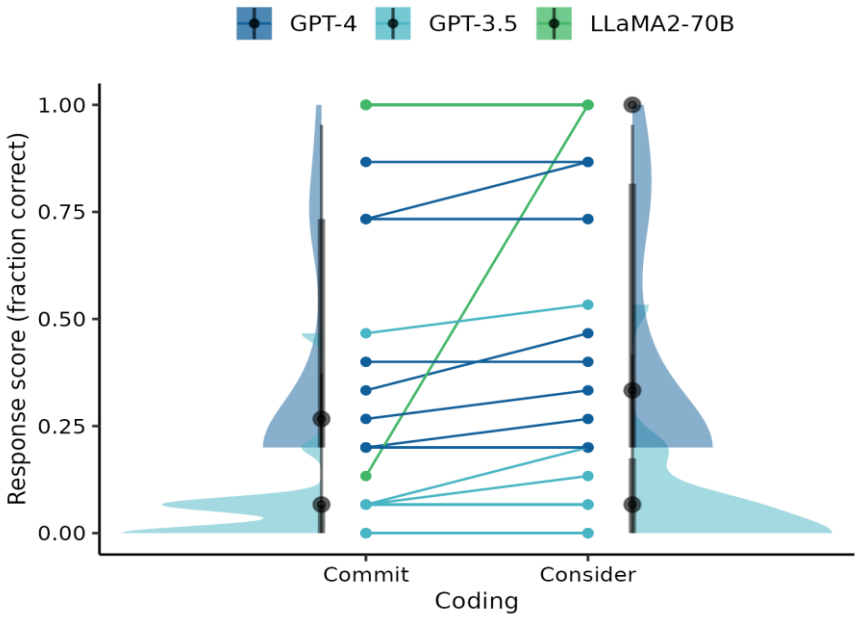

Figure S10. Alternative coding strategy. Side-by-side comparison of GPT performance on the Faux Pas test using the strict coding criteria ("*Did they know...?"* answer only accepted if *"No"* actively endorsed) and the new alternative coding (coded as correct if *"No"* was considered a viable option but not actively endorsed). Individual data points and lines show data aggregated by test item across sessions. Density plots are not shown for LLaMA2-70B as there was too little variability.

II.VI. Strange Stories: Partial Successes

Unlike other tasks, Strange Stories uses a three-level scoring system rather than a binary correct/incorrect judgement. As such, while the session-level responses of other tasks can be inferred from their aggregated scores, the Strange Stories have two ways that responses can lose points: Responses that fail to understand or explain the story in a meaningful way are coded as failures, while explaining the events of a story in non-mentalistic terms are rated as partial successes. As an example, consider the following story:

> *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?*

Examples of each kind of answer:

- Failure: *Jim will not look in the cupboard for the paddle because Simon has told him that the paddle is under his bed.*
- Partial Success: *Jim will look in the cupboard for the paddle because Simon lied about where it was hidden, claiming that it was under his bed when it was actually somewhere else. Therefore, Jim cannot trust Simon's answer about where he hid the paddle and needs to check both places to find it.* [This is only a partial success because it does not recognise that Jim will use his knowledge of Simon's untrustworthiness to reason about where the paddle actually is].
- Full Success: *Jim will look in the cupboard for the paddle because he knows that Simon is a big liar and never tells the truth. Since Simon said the paddle is under his bed, Jim believes*

the opposite must be true, so he will look in the cupboard instead.

As shown in Figure S11, breaking down different response types revealed that partial successes were infrequent, and were more likely for LLaMA2-70B than any other models.

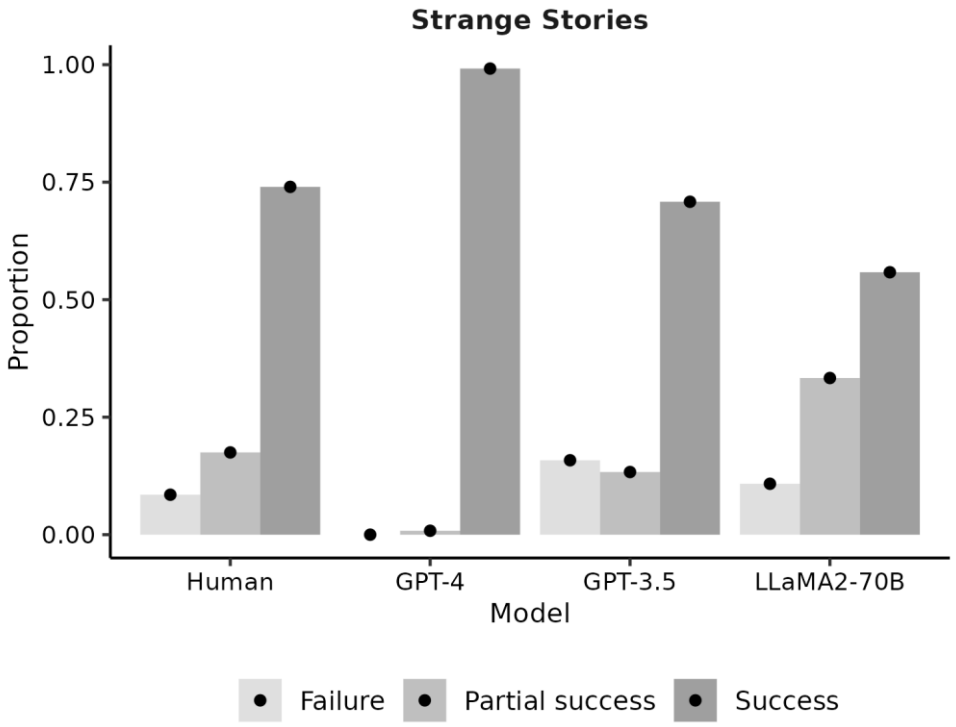

Figure S11. Counts of different response types on the Strange Stories test: failure, mixed success, and full success.

II.VII. Qualitative Analysis of Faux Pas Likelihood Test

The Faux Pas Likelihood test, which presented stories with a final question framed in terms of likelihood rather than an objective property of the character's mind, resulted in much better performance of GPT models than the original framing of the question. However, while both models scored very high, explanations spontaneously added by GPT models suggest that successes did not all follow from correct inference. To quantify this intuition, we refined the coding criteria to differentiate between different subtypes of successes, listed below with examples of possible responses fitting the criteria:

- Full explanation: the response provides an explanation that goes beyond the limits of what is described in the story to infer the cause of an offensive utterance using hypothetical reasoning or counterfactuals.

  Example: *It is more likely that Alice did not realize that Emma hadn't heard the results of the competition. If she had realized, she would have been more careful not to reveal the outcome of the competition inadvertently.*

- Event explanation: the response provides an explanation that provides a rational or adequate explanation, but it does so by drawing only on information included within the story without or with only minimal additional inference.

  Example: *It is more likely that Becky did not know that Simon was sick before making her comment, as she arrived late to class after the news was announced.*

- No explanation: the response provides the correct answer (that it is more likely they the speaker did not know) without any additional explanation. Alternatively, the response may reproduce details from the story that do not actually explain the behaviour.

  Example: *It is less likely that Richard remembered that James had given him the toy aeroplane for his birthday, as he made a comment that suggests he did not appreciate the gift, even*

though it was given to him by James.

- Tautological explanation: the response provides the correct answer but with an explanation that assumes the conclusion it is trying to explain.
  Example: *It is more likely that Lisa did not know that the curtains were new, as she made the insensitive comment without realizing that Jill had just purchased them.*

We also coded mixed successes and failures according to different subtypes:

- Mixed success following an uncertain response: the first response provided was uncertain or non-committal (e.g. *It is not clear*...) but the model provided the correct answer when prompted to provide the most likely explanation.
- Mixed success following an incorrect response: in the initial response the model correctly identified that something inappropriate was said but claimed that the speaker understood the context, leaving an open question as to why they said what they said. Following a follow-up prompt, the model provided an explanation that gave some acknowledgement that the offense was nonetheless unintentional.
- Failure to recognise the offence: the response does not report that the key statement was offensive or inappropriate, meaning that a follow-up to understand why the speaker might have said it is unnecessary.
- Failure with recognition of offence: after being prompted (either following an uncertain or incorrect response), the model still does not provide the correct explanation for the faux pas.

The counts of these different kinds of responses are shown in Figure S12. As shown in Figure S12A, the pattern of response for successes was similar for GPT-4 and GPT-3.5. Full and complete explanations involving hypotheticals or subjunctive clauses were rare: more often, the models would provide explanations that restipulated the events or facts related in the narrative. The most frequent elaboration for both models, however, was to present tautologies or circular descriptions as though they were explanations.

Mixed successes and failures were rare and exclusively seen in responses from GPT-3.5, the most common type of failure being failure to recognise a statement as offensive.

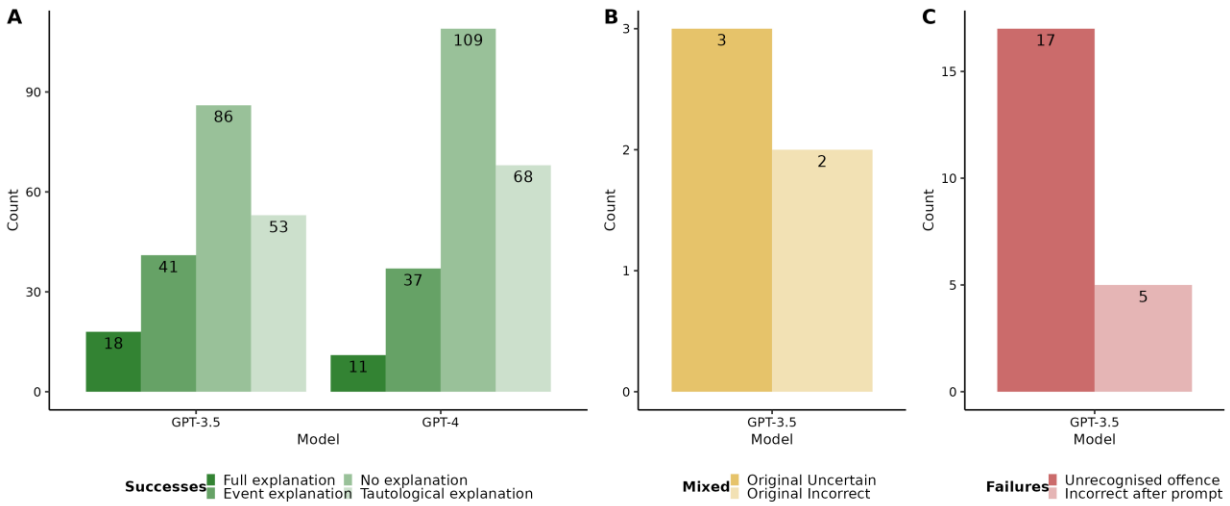

Figure S12. Qualitative breakdown of response types on Faux Pas Likelihood test. Barplots show counts of the different response types on the adapted Faux Pas Likelihood test. The figures show the number of four identified types of successes, two types of mixed successes, and two types of failures identified through manual coding of the responses.

II.VIII. Age and ToM

Information about the human sample collected for the study is reported in Table S12.

Table S12. Sample sizes and age details for each ToM task and the additional tasks used in the study.

Task	**N**	**Age range**	**Age mean ± sd**
False Belief	51	21-62	37.20 ± 11.14
Irony	50	23-59	37.08 ± 10.26
Faux Pas	51	19-64	39.51 ± 11.68
Hinting	48	21-70	38.44 ± 12.21
Strange Stories	50	21-69	34.86 ± 11.99
Belief Likelihood Test	849*	18-70	41.27 ± 12.98
False Belief Perturbations	755**	18-70	41.73 ± 13.16

*demographic data from 51 participants were not collected due to technical issues (original N = 900)

**demographic data from 2 participants were not collected due to technical issues (original N = 757)

To assess a possible relationship between age and the ToM tasks performance, rank correlations using the Spearman's ρ and the Kendall's τ were performed. Performance or age were considered outlier values if they were more than 1.5*Interquartile Ranges above the third quartile or below the first quartile of the data. A total of 12 participants were removed from the analyses. As reported in Table S13, no significant correlations between age and performance were found.

Table S13. Statistics for the Spearman's ρ and the Kendall's τ to assess correlation between age and ToM performance.

Task	***ρ***	***p-value****	***τ***	***p-value****
Irony	-0.07	0.6	-0.062	0.6
Faux Pas	-0.16	0.3	-0.130	0.3
Hinting	-0.11	0.9	-0.007	1.0
Strange Stories	0.15	0.3	0.110	0.3

*p-values were corrected to account for ties in the data

One difference with respect to prevous studies reporting a relationship between age and ToM (Slessor et al., 2007; Henry et al., 2013), is that our dataset was not collected to explore age-related differences in the human sample. The age range of our participants was thus both younger and narrower compared to those in earlier studies.

II.IX. Full Text Items

II.IX.I. False Belief Perturbations (adapted from Ullman, 2023)

Variant	John's cat	Lucy's juice	Grace's biscuits
	I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story.	*I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story.*	*I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story.*
False Belief	In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. When John comes back home, where will he look for the cat?	In the kitchen there are Lucy, Mia, a carton of orange juice, a fridge, and a cupboard. Lucy puts the carton of orange juice in the cupboard. She then leaves the kitchen and goes to school. While Lucy is away, Mia takes the orange juice out of the cupboard and puts it in the fridge. Mia leaves the room and goes to work. Lucy comes back from school and enters the kitchen. She doesn't know what happened in the kitchen when she was away. When Lucy comes back home, where will she look for the orange juice?	In the living room there are Grace, her grandmother, some chocolate biscuits, a metal tin, and a ceramic jar. Whenever Grace visits her grandmother, she always gets a chocolate biscuit from where they are stored in the metal tin. Today, she gets a biscuit and then leaves. While Grace is gone, her grandmother takes the chocolate biscuits out of the metal tin and puts them into the ceramic jar. Grace comes back for a visit and enters the living room. She doesn't know what happened in the living room when she was away. When Grace comes to visit, where will she look for the chocolate biscuits?
Transparent	In the room there are John, Mark, a cat, **a transparent plastic box,** and a **glass chest**. John takes the cat and puts it in the **chest**. He leaves the room and goes to school. While John is away, Mark takes the cat out of the **chest** and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. When John comes back home, where will he look for the cat?	In the kitchen there are Lucy, Mia, a carton of orange juice, a **transparent plastic box**, and a **glass-fronted cabinet** Lucy puts the carton of orange juice in the **transparent plastic box**. She then leaves the kitchen and goes to school. While Lucy is away, Mia takes the orange juice out of the **box** and puts it in the **glass-fronted cabinet**. Mia leaves the room and goes to work. Lucy comes back from school and enters the kitchen. She doesn't know what happened in the kitchen when she was away. When Lucy comes back home, where will she look for the orange juice?	In the living room there are Grace, her grandmother, some chocolate biscuits, a **clear plastic container**, and a **glass jar**. Whenever Grace visits her grandmother, she always gets a chocolate biscuit from where they are stored in the **clear plastic container**. Today, she gets a biscuit and then leaves. While Grace is gone, her grandmother takes the chocolate biscuits out of the **clear plastic container** and puts them into the **glass jar**. Grace comes back for a visit and enters the living room. She doesn't know what happened in the living room when she was away. When Grace comes to visit, where will she look for the chocolate biscuits?

Variant	John's cat	Lucy's juice	Grace's biscuits
Preposition	In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it **on** the basket. He leaves the room and goes to school. While John is away, Mark takes the cat **off** the basket and puts it **on** the box. Mark leaves the room and goes to work. John comes back from school and enters the room. **John looks around the room.** He doesn't know what happened in the room when he was away. When John comes back home, where will he look for the cat?	In the kitchen there are Lucy, Mia, a carton of orange juice, a fridge, and a cupboard. Lucy puts the carton of orange juice **on** the cupboard. She then leaves the kitchen and goes to school. While Lucy is away, Mia takes the orange juice **off of** the cupboard and puts it **on** the fridge. Mia leaves the room and goes to work. Lucy comes back from school and enters the kitchen. **Lucy looks around the kitchen.** She doesn't know what happened in the kitchen when she was away. When Lucy comes back home, where will she look for the orange juice?	In the living room there are Grace, her grandmother, some chocolate biscuits, **a metal tray**, and **a ceramic plate**. Whenever Grace visits her grandmother, she always gets a chocolate biscuit from where they are stored **on the metal tray**. Today, she gets a biscuit and then leaves. While Grace is gone, her grandmother takes the chocolate biscuits **off of the metal tray** and puts them **onto the ceramic plate**. Grace comes back for a visit and enters the living room. **Grace looks around the living room.** She doesn't know what happened in the living room when she was away. When Grace comes to visit, where will she look for the chocolate biscuits?
Testimony	In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. **Mark calls John to tell him he is going to move the cat to the box. John believes him.** While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John comes back from school and enters the room. He doesn't know what happened in the room when he was away. When John comes back home, where will he look for the cat?	In the kitchen there are Lucy, Mia, a carton of orange juice, a fridge, and a cupboard. Lucy puts the carton of orange juice in the cupboard. She then leaves the kitchen and goes to school. **Mia texts Lucy to tell her that she is going to move the orange juice to the fridge. Lucy believes her.** While Lucy is away, Mia takes the orange juice out of the cupboard and puts it in the fridge. Mia leaves the room and goes to work. Lucy comes back from school and enters the kitchen. She doesn't know what happened in the kitchen when she was away. When Lucy comes back home, where will she look for the orange juice?	In the living room there are Grace, her grandmother, some chocolate biscuits, a metal tin, and a ceramic jar. Whenever Grace visits her grandmother, she always gets a chocolate biscuit from the metal tin. Today, **she gets a biscuit and her grandmother tells her that she is going to move the chocolate biscuits from the metal tin to the ceramic jar before Grace next visits. Grace believes her grandmother, and she leaves.** While Grace is gone, her grandmother takes the chocolate biscuits out of the metal tin and puts them into the ceramic jar. Grace comes back for a visit and enters the living room. She doesn't know what happened in the living room when she was away. When Grace comes to visit, where will she look for the chocolate biscuits?

Variant	John's cat	Lucy's juice	Grace's biscuits
Mover	In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. **John and Mark** come back and enter the room. **They** don't know what happened in the room when they were away. When **Mark** comes back home, where will he look for the cat?	In the kitchen there are Lucy, Mia, a carton of orange juice, a fridge, and a cupboard. Lucy puts the carton of orange juice in the cupboard. She then leaves the kitchen and goes to school. While Lucy is away, Mia takes the orange juice out of the cupboard and puts it in the fridge. Mia leaves the room and goes to work. **Lucy and Mia** come back from school and enter the kitchen. **They** don't know what happened in the kitchen when they were away. When **Mia** comes back home, where will she look for the orange juice?	In the living room there are Grace, her grandmother, some chocolate biscuits, a metal tin, and a ceramic jar. Whenever Grace visits her grandmother, she always gets a chocolate biscuit from where they are stored in the metal tin. Today, she gets a biscuit and then leaves. While Grace is gone, her grandmother takes the chocolate biscuits out of the metal tin and puts them into the ceramic jar. The grandmother then leaves to go shopping. **Grace and her grandmother come back and enter the living room. They** don't know what happened in the living room when they were away. When **the grandmother** comes back home, where will she look for the chocolate biscuits?

II.IX.II. Items Generated for the Belief Likelihood Test

Story	Faux Pas	Neutral	Knowledge Implied
	I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story	*I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story*	*I am going to tell you a short story about some people. At the end of this story a person will say or do something. When I've finished telling it I will ask you some questions about what happened in the story*
1	Michael was a very awkward child when he was at high school. He struggled with making friends and spent his time alone writing poetry. However, after he left, he became a lot more confident and sociable. At his ten-year high school reunion, he met Amanda, who had been in his English class. Over drinks, she said to him, '**I don't know if you remember this guy from school. He was in my English class. He wrote poetry and he was super awkward. I hope he isn't here tonight.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where were Michael and Amanda? Is it more likely that Amanda knew or didn't know that Michael was the boy from her English class?	Michael was a very awkward child when he was at high school. He struggled with making friends and spent his time alone writing poetry. However, after he left, he became a lot more confident and sociable. At his ten-year high school reunion, he met Amanda, who had been in his English class. Over drinks, she said to him, '**Do you know where the bar is?**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where were Michael and Amanda? Is it more likely that Amanda knew or didn't know that Michael was the boy from her English class?	Michael was a very awkward child when he was at high school. He struggled with making friends and spent his time alone writing poetry. However, after he left, he became a lot more confident and sociable. At his ten-year high school reunion, he met Amanda, who had been in his English class. Over drinks, she said to him, '**Do you still write poetry?**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where were Michael and Amanda? Is it more likely that Amanda knew or didn't know that Michael was the boy from her English class?

Story	Faux Pas	Neutral	Knowledge Implied
2	Laura painted a picture of Olivia, who decided to hang it in her living room at home. A couple of months later, Olivia invited Laura to her place. While the two friends chatted over a cup of tea in the living room, Olivia's son came in and said, '**Laura, you should help my mum choose which paintings to hang in the house, as you can see, she has no good taste at all!**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Olivia hang Laura's painting? Is it more likely that Olivia's son knew or didn't know that Laura painted the painting?	Laura painted a picture of Olivia, who decided to hang it in her living room at home. A couple of months later, Olivia invited Laura to her place. While the two friends chatted over a cup of tea in the living room, Olivia's son came in and said, '**I'm looking forward to your party next week, Laura.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Olivia hang Laura's painting? Is it more likely that Olivia's son knew or didn't know that Laura painted the painting?	Laura painted a picture of Olivia, who decided to hang it in her living room at home. A couple of months later, Olivia invited Laura to her place. While the two friends chatted over a cup of tea in the living room, Olivia's son came in and said, '**I'd love to have a portrait of myself to hang in my room.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Olivia hang Laura's painting? Is it more likely that Olivia's son knew or didn't know that Laura painted the painting?
3	Jeremy had been saving up for months to buy his dream car: a convertible sports car that was painted a very vivid green. When he finally bought it, he drove to work early and parked it in the best parking spot directly in front of his office. Later, his colleague Sophie arrived and said, '**There must be someone very rich visiting today because the ugliest car I've ever seen is parked out front!**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where was Jeremy's car parked? Is it more likely that Sophie knew or didn't know that the car was Jeremy's?	Jeremy had been saving up for months to buy his dream car: a convertible sports car that was painted a very vivid green. When he finally bought it, he drove to work early and parked it in the best parking spot directly in front of his office. Later, his colleague Sophie arrived and said, '**My alarm clock broke this morning. I need to buy a new one after work.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where was Jeremy's car parked? Is it more likely that Sophie knew or didn't know that the car was Jeremy's?	Jeremy had been saving up for months to buy his dream car: a convertible sports car that was painted a very vivid green. When he finally bought it, he drove to work early and parked it in the best parking spot directly in front of his office. Later, his colleague Sophie arrived and said, '**Hey. I'd love to go for a ride after work.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where was Jeremy's car parked? Is it more likely that Sophie knew or didn't know that the car was Jeremy's?

Story	Faux Pas	Neutral	Knowledge Implied
4	Jim was happy because his parents just opened a new restaurant near where he worked. One day, at work, he decided to invite his colleagues to the restaurant. His colleague Ann said, '**I'd love to go out to dinner, but I strongly advise against that restaurant, I went the other night, and the food was terrible.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Jim invite his colleagues? Is it more likely that Ann knew or didn't know that the new restaurant belonged to Jim's parents?	Jim was happy because his parents just opened a new restaurant near where he worked. One day, at work, he decided to invite his colleagues to the restaurant. His colleague Ann said, '**I'm sorry, but I can't go tonight. I'm going to the cinema with my sister.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Jim invite his colleagues? Is it more likely that Ann knew or didn't know that the new restaurant belonged to Jim's parents?	Jim was happy because his parents just opened a new restaurant near where he worked. One day, at work, he decided to invite his colleagues to the restaurant. His colleague Ann said, '**Oh I'm not sure. I don't want your parents to assume we're dating.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Where did Jim invite his colleagues? Is it more likely that Ann knew or didn't know that the new restaurant belonged to Jim's parents?
5	Jane was talking to her sister on Christmas Eve about how her gift for their brother Matt: a big box of sweets from a recent vacation. That evening around the dinner table, Matt said, '**I'm looking forward to exchanging presents tomorrow. I just hope nobody got something unimaginative like sweets!**' In the story did someone say something that they should not have said? What did they say that they should not have said? What gift had Jane bought her brother for Christmas? Is it more likely that Matt knew or didn't know that Jane had bought him sweets?	Jane was talking to her sister on Christmas Eve about how her gift for their brother Matt: a big box of sweets from a recent vacation. That evening around the dinner table, Matt said, '**I hope everybody remembered their Christmas jumpers for tomorrow.**' In the story did someone say something that they should not have said? What did they say that they should not have said? What gift had Jane bought her brother for Christmas? Is it more likely that Matt knew or didn't know that Jane had bought him sweets?	Jane was talking to her sister on Christmas Eve about how her gift for their brother Matt: a big box of sweets from a recent vacation. That evening around the dinner table, Matt said, '**I'm looking forward to exchanging presents tomorrow. I can't wait to indulge my sweet tooth.**' In the story did someone say something that they should not have said? What did they say that they should not have said? What gift had Jane bought her brother for Christmas? Is it more likely that Matt knew or didn't know that Jane had bought him sweets?

Story	Faux Pas	Neutral	Knowledge Implied
6	Gareth was the singer in a band who had one well-known hit decades ago but no further success. He went into a café where Emma was working. "Good morning, how can I help you today?" asked Emma. Gareth was about to reply when his song came on the radio. Emma quickly turned the radio off and said, '**Not that song again. I hate it.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Who was working at the café? Is it more likely that Emma knew or didn't know that Gareth wrote the song on the radio?	Gareth was the singer in a band who had one well-known hit decades ago but no further success. He went into a café where Emma was working. "Good morning, how can I help you today?" asked Emma. Gareth was about to reply when his song came on the radio. Emma quickly turned the radio off and said, '**We have a special deal on if you want a pastry with any cold drink.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Who was working at the café? Is it more likely that Emma knew or didn't know that Gareth wrote the song on the radio?	Gareth was the singer in a band who had one well-known hit decades ago but no further success. He went into a café where Emma was working. "Good morning, how can I help you today?" asked Emma. Gareth was about to reply when his song came on the radio. Emma quickly turned the radio off and said, '**Oh, what funny timing! I love this song.**' In the story did someone say something that they should not have said? What did they say that they should not have said? Who was working at the café? Is it more likely that Emma knew or didn't know that Gareth wrote the song on the radio?

II.X. Resource Availability

All resources are available on a repository stored on the Open Science Framework (OSF) at the following link: https://osf.io/fwj6v/?view_only=0b3619922c9142e9a0653f258156f1a5

This repository contains all test items, data, and code reported in this study. Test items and data are available in an Excel file that includes the text of every item delivered in each test, the full text responses of each model to each item, and the code assigned to each response. This file is available in the repository under *'raw_text_data/Theory of Mind Battery Responses.xlsx'* at the following URL: https://osf.io/dbn92?view_only=0b3619922c9142e9a0653f258156f1a5

The code used for all analysis is included as a Markdown file in the same repository. The data used by this file is available as a number of CSV files under *'scored_data/'* in the repository, and all materials necessary for replicating the analysis can be downloaded as a single .zip file within the main repository titled *'Full R Project Code.zip'* at the following URL: https://osf.io/j3vhq?view_only=0b3619922c9142e9a0653f258156f1a5

II.XI. Acknowledgments

This work is supported by the European Commission through Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01). The first author was supported by a Humboldt Research Fellowship for Experienced Researchers provided by the Alexander von Humboldt Foundation.

References

Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, *11*(9), Articolo 9. https://doi.org/10.1038/nn.2182

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed). Wiley-Interscience.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Anquetil, T., & Jeannerod, M. (2007). Simulated actions in the first and in the third person perspectives share common representations. *Brain Research*, *1130*, 125–129. https://doi.org/10.1016/j.brainres.2006.10.091

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844. https://doi.org/10.1111/j.1467-9280.2006.01791.x

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental continuity in theory of mind: Speed and accuracy of belief–desire reasoning in children and adults. *Child Development*, *82*(5), 1691–1703. https://doi.org/10.1111/j.1467-8624.2011.01635.x

Au-Yeung, S. K., Kaakinen, J. K., Liversedge, S. P., & Benson, V. (2015). Processing of Written Irony in Autism Spectrum Disorder: An Eye-Movement Study. *Autism Research: Official Journal of the International Society for Autism Research*, *8*(6), 749–760. https://doi.org/10.1002/aur.1490

Baron-Cohen, S. (1989). The Autistic Child's Theory of Mind: A Case of Specific Developmental Delay. *Journal of Child Psychology and Psychiatry*, *30*(2), 285–297. https://doi.org/10.1111/j.1469-7610.1989.tb00241.x

Baron-Cohen, S. (2001). Theory of mind in normal development and autism. *Prisme*. https://www.semanticscholar.org/paper/Theory-of-mind-in-normal-development-and-autism-Baron-Cohen/ec44403f5ae8e9ddb101e62dac8e68e13a884e5b

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Lombardo, M., & Tager-Flusberg, H. (2013). *Understanding Other Minds: Perspectives from developmental social neuroscience*. OUP Oxford.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. https://doi.org/10.1023/a:1023035012436

Baron-Cohen, S., Spitz, A., & Cross, P. (1993). Do children with autism recognise surprise? A research note. *Cognition and Emotion*, *7*(6), 507–516. https://doi.org/10.1080/02699939308409202

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Becchio, C., & Castiello, U. (2012). Visuomotor resonance in autism spectrum disorders. *Frontiers in Integrative Neuroscience*, *6*. https://www.frontiersin.org/articles/10.3389/fnint.2012.00110

Becchio, C., Cavallo, A., Begliomini, C., Sartori, L., Feltrin, G., & Castiello, U. (2012). Social grasping: From mirroring to mentalizing. *NeuroImage*, *61*(1), 240–248. https://doi.org/10.1016/j.neuroimage.2012.03.013

Becchio, C., Koul, A., Ansuini, C., Bertone, C., & Cavallo, A. (2018). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of Life Reviews*, *24*, 67–80. https://doi.org/10.1016/j.plrev.2017.10.002

Bernstein, D. M., Thornton, W. L., & Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Experimental Aging Research*, *37*(5), 481–502. https://doi.org/10.1080/0361073X.2011.619466

Best, D. J., & Roberts, D. E. (1975). Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. *Applied Statistics, 24*(3), 377. https://doi.org/10.2307/2347111

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond Autism: Introducing the Dialectical Misattunement Hypothesis and a Bayesian Account of Intersubjectivity. *Psychopathology*, *50*(6), 355–372. https://doi.org/10.1159/000484353

Bonnefon, J.-F., & Rahwan, I. (2020). Machine Thinking, Fast and Slow. *Trends in Cognitive Sciences*, *24*(12), 1019–1027. https://doi.org/10.1016/j.tics.2020.09.007

Boria, S., Fabbri-Destro, M., Cattaneo, L., Sparaci, L., Sinigaglia, C., Santelli, E., Cossu, G., & Rizzolatti, G. (2009). Intention understanding in autism. *PloS One*, *4*(5), e5596. https://doi.org/10.1371/journal.pone.0005596

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating Action Understanding: Inferential Processes versus Action Simulation. *Current Biology*, *17*(24), 2117–2121. https://doi.org/10.1016/j.cub.2007.11.057

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: Comment on Pellicano and Burr. *Trends in Cognitive Sciences*, *16*(12), 573–574. https://doi.org/10.1016/j.tics.2012.10.005

Brooks, C., & Szafir, D. (2019). *Building Second-Order Mental Models for Human-Robot Interaction* (arXiv:1909.06508). arXiv. https://doi.org/10.48550/arXiv.1909.06508

Brooks, R. A. (1995). Intelligence Without Reason. In *The Artificial Life Route to Artificial Intelligence*. Routledge.

Brown, P., & Marsden, J. F. (2001). Book Review: Cortical Network Resonance and Motor Activity in Humans. *The Neuroscientist*, *7*(6), 518–526. https://doi.org/10.1177/107385840100700608

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023). *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.* https://doi.org/10.5281/zenodo.7637476

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Calder, A. J. (2011). *The Oxford handbook of face perception*. Oxford University Press.

Casartelli, L., Federici, A., Fumagalli, L., Cesareo, A., Nicoli, M., Ronconi, L., Vitale, A., Molteni, M., Rizzolatti, G., & Sinigaglia, C. (2020). Neurotypical individuals fail to understand action vitality form in children with autism spectrum disorder. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27712–27718. https://doi.org/10.1073/pnas.2011311117

Casile, A., & Giese, M. A. (2006). Nonvisual Motor Training Influences Biological Motion Perception. *Current Biology*, *16*(1), 69–74. https://doi.org/10.1016/j.cub.2005.10.071

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain: A Journal of Neurology*, *125*(Pt 8), 1839–1849. https://doi.org/10.1093/brain/awf189

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., & Becchio, C. (2016). Decoding intentions from movement kinematics. *Scientific Reports*, *6*, 37036. https://doi.org/10.1038/srep37036

Cavallo, A., Romeo, L., Ansuini, C., Battaglia, F., Nobili, L., Pontil, M., Panzeri, S., & Becchio, C. (2021). Identifying the signature of prospective motor control in children with autism. *Scientific Reports*, *11*(1), 3165. https://doi.org/10.1038/s41598-021-82374-2

Cavallo, A., Romeo, L., Ansuini, C., Podda, J., Battaglia, F., Veneselli, E., Pontil, M., & Becchio, C. (2018). Prospective motor control obeys to idiosyncratic strategies in autism. *Scientific Reports*, *8*(1), 13717. https://doi.org/10.1038/s41598-018-31479-2

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Chrysaitis, N. A., & Seriès, P. (2023). 10 years of Bayesian theories of autism: A comprehensive review. *Neuroscience & Biobehavioral Reviews*, *145*, 105022. https://doi.org/10.1016/j.neubiorev.2022.105022

Clark, A. (1998). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Constantino, J. N. (2013). Social Responsiveness Scale. In F. R. Volkmar, *Encyclopedia of Autism Spectrum Disorders* (pp. 2919–2929). Springer. https://doi.org/10.1007/978-1-4419-1698-3_296

Cook, J. L. (2016). From movement kinematics to social cognition: The case of autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1693), 20150372. https://doi.org/10.1098/rstb.2015.0372

Cook, J. L., Blakemore, S.-J., & Press, C. (2013). Atypical basic movement kinematics in autism spectrum conditions. *Brain: A Journal of Neurology*, *136*(Pt 9), 2816–2824. https://doi.org/10.1093/brain/awt208

Corcoran, R. (2003). Inductive reasoning and the understanding of intention in schizophrenia. *Cognitive Neuropsychiatry*, *8*(3), 223–235. https://doi.org/10.1080/13546800244000319

Cossu, G., Boria, S., Copioli, C., Bracceschi, R., Giuberti, V., Santelli, E., & Gallese, V. (2012). Motor Representation of Actions in Children with Autism. *PLOS ONE*, *7*(9), e44779. https://doi.org/10.1371/journal.pone.0044779

Craighero, L., Bello, A., Fadiga, L., & Rizzolatti, G. (2002). Hand action preparation influences the responses to hand pictures. *Neuropsychologia*, *40*(5), 492–502. https://doi.org/10.1016/S0028-3932(01)00134-8

Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: A motor-visual attentional effect. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1673–1692. https://doi.org/10.1037/0096-1523.25.6.1673

Cristiano, A., Finisguerra, A., Urgesi, C., Avenanti, A., & Tidoni, E. (2023). Functional role of the theory of mind network in integrating mentalistic prior information with action kinematics during action observation. *Cortex*, *166*, 107–120. https://doi.org/10.1016/j.cortex.2023.05.009

Cuzzolin, F., Morelli, A., Cîrstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: Theory of mind in AI. *Psychological Medicine*, *50*(7), 1057–1061. https://doi.org/10.1017/S0033291720000835

Dou, Z. (2023). *Exploring GPT-3 Model's Capability in Passing the Sally-Anne Test a Preliminary Study in Two Languages*. OSF Preprints. https://doi.org/10.31219/osf.io/8r3ma

Dziuk, M. A., Larson, J. C. G., Apostu, A., Mahone, E. M., Denckla, M. B., & Mostofsky, S. H. (2007). Dyspraxia in autism: Association with motor, social, and communicative deficits. *Developmental Medicine & Child Neurology*, *49*(10), 734–739. https://doi.org/10.1111/j.1469-8749.2007.00734.x

Edey, R., Cook, J., Brewer, R., Johnson, M. H., Bird, G., & Press, C. (2016). Interaction takes two: Typical adults exhibit mind-blindness towards those with autism spectrum disorder. *Journal of Abnormal Psychology*, *125*(7), 879–885. https://doi.org/10.1037/abn0000199

El Kaliouby, R., & Robinson, P. (2004). Mind reading machines: Automated inference of cognitive mental states from video. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, *1*, 682–688 vol.1. https://doi.org/10.1109/ICSMC.2004.1398380

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Firestone, C. (2020). Performance vs. Competence in human-machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, *63*(6), 877–889. https://doi.org/10.1037//0022-3514.63.6.877

Frank, M. C. (2023). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour*, *7*(11), 1825–1827. https://doi.org/10.1038/s41562-023-01732-4

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1), 137–160. https://doi.org/10.1007/s00422-011-0424-z

Frith, C. D., & Frith, U. (1999). Interacting Minds—A Biological Basis. *Science*, *286*(5445), 1692–1695. https://doi.org/10.1126/science.286.5445.1692

Frith, C. D., & Wolpert, D. (2004). *The Neuroscience of Social Interaction: Decoding, Influencing, and Imitating the Actions of Others*. Oxford University Press UK.

Frith, C., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, U. (1989). Autism and "Theory of Mind". In C. Gillberg, *Diagnosis and Treatment of Autism* (pp. 33–52). Springer US. https://doi.org/10.1007/978-1-4899-0882-7_4

Fuchs, T. (2015). Pathologies of Intersubjectivity in Autism and Schizophrenia. *Journal of Consciousness Studies*, *22*(1–2), 191–214.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). *Understanding Social Reasoning in Language Models with Language Models* (arXiv:2306.15448). arXiv. https://doi.org/10.48550/arXiv.2306.15448

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press. https://doi.org/10.4324/9781315740218

Gil, D., Fernández-Modamio, M., Bengochea, R., & Arrieta, M. (2012). [Adaptation of the Hinting Task theory of the mind test to Spanish]. *Revista De Psiquiatria Y Salud Mental*, *5*(2), 79–88. https://doi.org/10.1016/j.rpsm.2011.11.004

Gowen, E., & Hamilton, A. (2013). Motor Abilities in Autism: A Review Using a Computational Context. *Journal of Autism and Developmental Disorders*, *43*(2), 323–344. https://doi.org/10.1007/s10803-012-1574-0

Grafton, S. T. (2009). Embodied Cognition and the Simulation of Action to Understand Others. *Annals of the New York Academy of Sciences*, *1156*(1), 97–117. https://doi.org/10.1111/j.1749-6632.2009.04425.x

Green, D., Charman, T., Pickles, A., Chandler, S., Loucas, T., Simonoff, E., & Baird, G. (2009). Impairment in movement skills of children with autistic spectrum disorders.

Developmental Medicine & Child Neurology, *51*(4), 311–316. https://doi.org/10.1111/j.1469-8749.2008.03242.x

Hagendorff, T. (2023). *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods* (arXiv:2303.13988). arXiv. https://doi.org/10.48550/arXiv.2303.13988

Hamilton, A., Wolpert, D., & Frith, U. (2004). Your Own Action Influences How You Perceive Another Person's Action. *Current Biology*, *14*(6), 493–498. https://doi.org/10.1016/j.cub.2004.03.007

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226. https://doi.org/10.1111/desc.12017

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(17), 6339–6352. https://doi.org/10.1523/JNEUROSCI.5613-10.2011

Happé, F. G. (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), 129–154. https://doi.org/10.1007/BF02172093

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging*, *28*(3), 826–839. https://doi.org/10.1037/a0030677

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. https://doi.org/10.1126/science.1243091

Hirata, S., Okuzumi, H., Kitajima, Y., Hosobuchi, T., Nakai, A., & Kokubun, M. (2014). Relationship between motor skill and social impairment in children with autism spectrum disorders. *International Journal of Developmental Disabilities*, *60*(4), 251–256. https://doi.org/10.1179/2047387713Y.0000000033

Hofstede, G. J. (2019). GRASP agents: Social first, intelligent later. *AI & SOCIETY*, *34*(3), 535–543. https://doi.org/10.1007/s00146-017-0783-7

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *The Behavioral and Brain Sciences*, *24*(5), 849–878; discussion 878-937. https://doi.org/10.1017/s0140525x01000103

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9*(1), 21–25. https://doi.org/10.1016/j.tics.2004.11.003

James, W., Burkhardt, F., Bowers, F., Skrupskelis, I. K., & James, W. (1981). *The principles of psychology*. Harvard University Press.

Jenkinson, R., Milne, E., & Thompson, A. (2020). The relationship between intolerance of uncertainty and anxiety in autism: A systematic literature review and meta-analysis. *Autism*, *24*(8), 1933–1944. https://doi.org/10.1177/1362361320932437

Kaiser, M. D., Hudac, C. M., Shultz, S., Lee, S. M., Cheung, C., Berken, A. M., Deen, B., Pitskel, N. B., Sugrue, D. R., Voos, A. C., Saulnier, C. A., Ventola, P., Wolf, J. M., Klin, A., Vander Wyk, B. C., & Pelphrey, K. A. (2010). Neural signatures of autism. *Proceedings of the National Academy of Sciences*, *107*(49), 21223–21228. https://doi.org/10.1073/pnas.1010412107

Kampis, D., Kármán, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science*, *8*(8), 210190. https://doi.org/10.1098/rsos.210190

Karmiloff-Smith, A., Klima, E., Bellugi, U., Grant, J., & Baron-Cohen, S. (1995). Is There a Social Module? Language, Face Processing, and Theory of Mind in Individuals with Williams Syndrome. *Journal of Cognitive Neuroscience*, *7*(2), 196–208. https://doi.org/10.1162/jocn.1995.7.2.196

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

Kilroy, E., Cermak, S. A., & Aziz-Zadeh, L. (2019). A Review of Functional and Structural Neurobiology of the Action Observation Network in Autism Spectrum Disorder and Developmental Coordination Disorder. *Brain Sciences*, *9*(4), Articolo 4. https://doi.org/10.3390/brainsci9040075

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv.Org. https://arxiv.org/abs/1412.6980v9

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Konvalinka, I., Kompatsiari, K., & Li, Q. (2023). The fine-grained temporal dynamics of social timing: A window into sociality of embodied social agents. Comment on «The evolution of social timing» by L. Verga, S. A. Kotz, & A. Ravignani. *Physics of Life Reviews*, *47*, 95–98. https://doi.org/10.1016/j.plrev.2023.09.017

Kosinski, M. (2023). *Theory of Mind Might Have Spontaneously Emerged in Large Language Models* (arXiv:2302.02083). arXiv. https://doi.org/10.48550/arXiv.2302.02083

Koul, A., Soriano, M., Tversky, B., Becchio, C., & Cavallo, A. (2019). The kinematics that you do not expect: Integrating prior information and kinematics to understand intentions. *Cognition*, *182*, 213–219. https://doi.org/10.1016/j.cognition.2018.10.006

Kovács, Á. M., Téglás, E., & Csibra, G. (2021). Can infants adopt underspecified contents into attributed beliefs? Representational prerequisites of theory of mind. *Cognition*, *213*, 104640. https://doi.org/10.1016/j.cognition.2021.104640

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.)*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental Brain Research*, *217*(1), 1–5. https://doi.org/10.1007/s00221-012-3000-4

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293–1299. https://doi.org/10.1038/nn.4615

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*, 302. https://doi.org/10.3389/fnhum.2014.00302

Leekam, S., Baron-Cohen, S., Perrett, D., Milders, M., & Brown, S. (1997). Eye-direction detection: A dissociation between geometric and joint attention skills in autism. *British Journal of Developmental Psychology*, *15*(1), 77–95. https://doi.org/10.1111/j.2044-835X.1997.tb00726.x

Leslie, A. M. (1987). Pretense and representation: The origins of «theory of mind». *Psychological Review*, *94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Lewis, V., & Boucher, J. (1988). Spontaneous, instructed and elicited play in relatively able autistic children. *British Journal of Developmental Psychology*, *6*(4), 325–339. https://doi.org/10.1111/j.2044-835X.1988.tb01105.x

Lidstone, D. E., Miah, F. Z., Poston, B., Beasley, J. F., & Dufek, J. S. (2020). Manual dexterity in children with autism spectrum disorder: A cross-syndrome approach. *Research in Autism Spectrum Disorders*, *73*, 101546. https://doi.org/10.1016/j.rasd.2020.101546

Lord, C., DiLavore, P. C., & Gotham, K. (2012). *Autism diagnostic observation schedule: Western Psychological Services Torrance*. CA.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. https://doi.org/10.1007/BF02172145

Marcus, G., & Davis, E. (2023). How Not to Test GPT-3 [Substack newsletter]. *Marcus on AI*. https://garymarcus.substack.com/p/how-not-to-test-gpt-3

McPartland, J. C., Coffman, M., & Pelphrey, K. A. (2011). Recent Advances in Understanding the Neural Bases of Autism Spectrum Disorder. *Current opinion in pediatrics*, *23*(6), 628–632. https://doi.org/10.1097/MOP.0b013e32834cb9c9

Oguntola, I., Hughes, D., & Sycara, K. (2021). Deep Interpretable Models of Theory of Mind. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 657–664. https://doi.org/10.1109/RO-MAN50785.2021.9515505

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

OpenAI. (2023a). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2023b). *ChatGPT version 3.5* [Software].

OpenAI. (2023c). *ChatGPT version 4* [Software].

O'Regan, J. K., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, *129*(1), 79–103. https://doi.org/10.1023/A:1012699224677

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological Bulletin*, *143*(5), 521–542. https://doi.org/10.1037/bul0000097

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention, and Behavior. *Neuron*, *93*(3), 491–507. https://doi.org/10.1016/j.neuron.2016.12.036

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. https://doi.org/10.48550/ARXIV.1912.01703

Patri, J.-F., Cavallo, A., Pullar, K., Soriano, M., Valente, M., Koul, A., Avenanti, A., Panzeri, S., & Becchio, C. (2020). Transient Disruption of the Inferior Parietal Lobule Impairs the Ability to Attribute Intention to Action. *Current Biology*, *30*(23), 4594-4605.e7. https://doi.org/10.1016/j.cub.2020.08.104

Pavlova, M. A. (2012). Biological Motion Processing as a Hallmark of Social Cognition. *Cerebral Cortex*, *22*(5), 981–995. https://doi.org/10.1093/cercor/bhr156

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, *16*(3), 337–348. https://doi.org/10.1111/j.2044-835X.1998.tb00756.x

Plate, R. C., Ham, H., & Jenkins, A. C. (2023). When uncertainty in social contexts increases exploration and decreases obtained rewards. *Journal of Experimental Psychology. General*, *152*(9), 2463–2478. https://doi.org/10.1037/xge0001410

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Press, C., & Cook, R. (2015). Beyond action-specific simulation: Domain-general motor contributions to perception. *Trends in Cognitive Sciences*, *19*(4), 176–178. https://doi.org/10.1016/j.tics.2015.01.006

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. «Sandy», … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Articolo 1. https://doi.org/10.1038/4580

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews. Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Rescher, N. (1960). Choice Without Preference. A Study of the History and of the Logic of the Problem of Buridan's Ass. *Kant Studien*, *51*(1–4), 142–175. https://doi.org/10.1515/kant.1960.51.1-4.142

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience*, *27*(1), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2023). *Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs* (arXiv:2210.13312). arXiv. https://doi.org/10.48550/arXiv.2210.13312

Schilbach, L. (2016). Towards a second-person neuropsychiatry. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1686), 20150081. https://doi.org/10.1098/rstb.2015.0081

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scott, F. J., & Baron-Cohen, S. (1996). Imagining Real and Unreal Things: Evidence of a Dissociation in Autism. *Journal of Cognitive Neuroscience*, *8*(4), 371–382. https://doi.org/10.1162/jocn.1996.8.4.371

Semin, G. R., & Smith, E. R. (2013). Socially Situated Cognition in Perspective. *Social Cognition*, *31*(2), 125–146. https://doi.org/10.1521/soco.2013.31.2.125

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (arXiv:2305.14763). arXiv. https://doi.org/10.48550/arXiv.2305.14763

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Slepian, M. L., Young, S. G., Rutchick, A. M., & Ambady, N. (2013). Quality of professional players' poker hands is perceived accurately from arm motions. *Psychological Science*, *24*(11), 2335–2338. https://doi.org/10.1177/0956797613487384

Slessor, G., Phillips, L. H., & Bull, R. (2007). Exploring the specificity of age-related differences in theory of mind tasks. *Psychology and Aging*, *22*(3), 639–643. https://doi.org/10.1037/0882-7974.22.3.639

Smith, E. R., & Semin, G. R. (2004). Socially Situated Cognition: Cognition in its Social Context. In *Advances in Experimental Social Psychology* (Vol. 36, pp. 53–117). Academic Press. https://doi.org/10.1016/S0065-2601(04)36002-8

Snyder, M., & Cantor, N. (1998). Understanding personality and social behavior: A functionalist strategy. In *The handbook of social psychology, Vols. 1-2, 4th ed* (pp. 635–679). McGraw-Hill.

Sodian, B., & Frith, U. (1992). Deception and Sabotage in Autistic, Retarded and Normal Children. *Journal of Child Psychology and Psychiatry*, *33*(3), 591–605. https://doi.org/10.1111/j.1469-7610.1992.tb00893.x

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (arXiv:2206.04615). arXiv. http://arxiv.org/abs/2206.04615

Stins, J. F., & Emck, C. (2018). Balance Performance in Autism: A Brief Overview. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00901

Tager-Flusberg, H. (1992). Autistic Children's Talk about Psychological States: Deficits in the Early Acquisition of a Theory of Mind. *Child Development*, *63*(1), 161–172. https://doi.org/10.1111/j.1467-8624.1992.tb03604.x

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, *24*(1), 1–34. https://doi.org/10.1017/S0140525X01003910

Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (arXiv:2302.08399). arXiv. https://doi.org/10.48550/arXiv.2302.08399

Valagussa, G., Trentin, L., Signori, A., & Grossi, E. (2018). Toe Walking Assessment in Autism Spectrum Disorder Subjects: A Systematic Review. *Autism Research*, *11*(10), 1404–1415. https://doi.org/10.1002/aur.2009

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S.-A. (2012). Pragmatics in action: Indirect requests engage theory of mind areas and the cortical motor network. *Journal of Cognitive Neuroscience*, *24*(11), 2237–2247. https://doi.org/10.1162/jocn_a_00274

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675. https://doi.org/10.1037/a0037665

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The Embodied Mind, revised edition: Cognitive Science and Human Experience*. MIT Press.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), Articolo 3. https://doi.org/10.1038/s41592-019-0686-2

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Wechsler, D. (2012). *Wechsler Intelligence Scale for Children, Fourth Edition*. https://doi.org/10.1037/t15174-000

White, S., Hill, E., Happé, F., & Frith, U. (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

Wiesmann, C. G., Friederici, A. D., Singer, T., & Steinbeis, N. (2020). Two systems for thinking about others' thoughts in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(12), 6928–6935. https://doi.org/10.1073/pnas.1916725117

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Winner, E., Brownell, H., Happé, F., Blum, A., & Pincus, D. (1998). Distinguishing Lies from Jokes: Theory of Mind Deficits and Discourse Interpretation in Right Hemisphere Brain-Damaged Patients. *Brain and Language*, *62*(1), 89–106. https://doi.org/10.1006/brln.1997.1889

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1431), 593–602. https://doi.org/10.1098/rstb.2002.1238

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). *Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models* (arXiv:2303.04671). arXiv. https://doi.org/10.48550/arXiv.2303.04671

Yiu, E., Kosoy, E., & Gopnik, A. (2023). *Imitation versus Innovation: What children can do that large language and language-and-vision models cannot (yet)?* PsyArXiv. https://doi.org/10.31234/osf.io/kt9es

Zadeh, A., Chan, M., Liang, P. P., Tong, E., & Morency, L.-P. (2019). *Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence*. 8807–8817. https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., Upadhyay, S., & Faruqui, M. (2023). *How FaR Are Large*

Language Models From Agents with Theory-of-Mind? (arXiv:2310.03051). arXiv. https://doi.org/10.48550/arXiv.2310.03051

"La mia sensazione era chiara: riuscivo a percepire il suo benessere,

che intanto aveva raggiunto anche me,

era una specie di estasi, non c'era spazio per niente altro,

solo per il piacere dato dalla musica, dal potersi muovere liberamente,

dal godimento dell'istante.

(...)

Da spettatrice, io stessa ho goduto del suo momento,

ho in un certo senso invidiato la profondità di quel godimento, di quella libertà,

e ho pensato di aver riconosciuto, evidente davanti a me,

una piena e potente esperienza estetica, fuori dal mio corpo, padrona di quello di un altro."

Albergo, S. (2020). *La percezione dell'esperienza estetica in soggetti non vedenti.*

[Tesi di Laurea, Università di Bologna]