Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli

Elad Refoua[a], Zohar Elyoseph[b,c,*], Renata Wacker[d], Isabel Dziobek[d], Iftach Tsafrir[e], Gunther Meinlschmidt[f,g]

[a] *Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel*
[b] *Faculty of Education, University of Haifa, Haifa, Israel*
[c] *Department of Brain Sciences, Faculty of Medicine, Imperial College, London, United Kingdom*
[d] *Humboldt-Universität zu Berlin, Berlin, Germany*
[e] *The Artificial Third Institute, Israel*
[f] *Clinical Psychology and Psychotherapy – Methods and Approaches, Department of Psychology, Trier University, Germany*
[g] *University of Basel and University Hospital Basel, Department of Digital and Blended Psychosomatics and Psychotherapy, Psychosomatic Medicine, Basel, Switzerland*

ARTICLE INFO

Keywords:
Generative artificial intelligence
Social cognition
Theory of mind
Emotion recognition
Multimodal assessment
Generative artificial intelligence and mental health

ABSTRACT

The integration of Generative Artificial Intelligence (GAI) into human social contexts has raised fundamental questions about machines' capacity to understand and respond to complex emotional and social dynamics. While recent studies have demonstrated GAI's promising capabilities in processing static emotional content, the frontier of dynamic social cognition – where multiple modalities converge to create naturalistic social scenarios – remained largely unexplored. This study advances our understanding by examining the social-cognitive capabilities of Google's Gemini 1.5 Pro model through its performance on the Movie for the Assessment of Social Cognition (MASC), a sophisticated instrument designed to evaluate mentalization abilities using dynamic audiovisual stimuli. We compared the model's performance to a human normative sample ($N = 1230$) across varying temperature settings (a parameter controlling the level of randomness in the AI's output, where lower values lead to more deterministic responses and higher values increase variability; set at 0, 0.5, and 1). Results revealed that Gemini 1.5 Pro consistently performed above chance across all conditions (all corrected $ps < 0.001$, Cohen's h range $= 1.17$–1.41) and significantly outperformed the human sample mean ($Z = 2.24$, $p = .025$; Glass's $\Delta = 0.92$, 95 % CI [0.11, 1.72]; Hedges' $g = 0.92$, 95 % CI [0.12, 1.72]). Analysis of error patterns revealed a distribution between hyper-mentalizing (41.0 %; over-attribution of mental states), hypo-mentalizing (46.2 %; under-attribution of mental states), and non-mentalizing (12.8 %; failure to recognize mental states) errors. These findings extend our understanding of artificial social cognition to complex multimodal processing while raising important questions about the nature of machine-based social understanding. The implications span theoretical considerations in artificial Theory of Mind to practical applications in mental health care and social skills training, though careful consideration is warranted regarding the fundamental differences between human and artificial social cognitive processing.

1. Introduction

Integrating Generative Artificial intelligence (GAI) into human life reiterates important questions about the nature of human-machine interactions. As GAI systems become more sophisticated, their ability to understand and respond to human emotions and social cues has become a central topic and concern in both scientific and ethical domains (e.g., Cannarsa, 2021; Elyoseph, Refoua, et al., 2024; Makridakis, 2017). Notably, this ability may substantially contribute to shaping the future of human-machine relationships, also affecting GAI's role in society (e.

[*] Corresponding author. School of Counseling and Human Development, The Faculty of Education, The University of Haifa; Research fellow, Department of Brain Science, Imperial College, London, United Kingdom.
E-mail address: Zohar.j.a@gmail.com (Z. Elyoseph).

https://doi.org/10.1016/j.chbr.2025.100702
Received 3 January 2025; Received in revised form 14 May 2025; Accepted 19 May 2025
Available online 23 May 2025

g., Webster & Ivanov, 2020). One of the most relevant areas where GAI is expected to substantially contribute is human-like communication and interaction (Huang & Rust, 2018).

The potential for GAI to understand and respond to human emotions, intentions, and thoughts necessitates rigorous scientific inquiry before implementation in healthcare and mental health services. Emerging empirical evidence has begun to substantiate these possibilities through systematic investigation. For instance, Stade and colleagues (2024) demonstrated that large language models (LLMs) can effectively analyze linguistic patterns reflecting psychological states, thereby facilitating early detection and monitoring of mental health conditions. This capability has been further validated through research examining suicide risk detection, with deep learning models successfully identifying risk markers in social media content (Ophir et al., 2020). Notably, recent technological advances have extended these capabilities to encompass both textual and visual content analysis for suicide risk assessment (Badian et al., 2023), marking a significant advancement in multimodal risk detection. The applications of GAI in psychological assessment extend beyond risk detection to broader clinical applications, including psychological assessment, experimentation, and practice (Demszky et al., 2023). This expansion of capabilities is particularly evident in clinical populations. For example, Lauderdale and colleagues (2024) demonstrated GAI systems' ability to recognize and assess mental health symptoms and risks in veteran populations, suggesting evidence-based treatment for them. Furthermore, it is suggested that GAI may be employed to analyze psychotherapy sessions directly, offering insights that may not be immediately apparent to human therapists and aiding in the development of personalized interventions tailored to an individual's unique psychological profile (Haber et al., 2024). This integration of GAI and behavioral data could be even more impactful from an Internet of Behavior (IoB) perspective. As IoB extends beyond simple data collection, focusing on understanding and helping modify human behavior through the analysis of digital footprints (Javaid et al., 2021).

Mentalization, a key aspect of social cognition, involves understanding one's own and others' mental states, including emotions, thoughts, and intentions. It encompasses related constructs such as Theory of Mind (ToM), empathy, emotional awareness, and reflective functioning (Fonagy et al., 2018). In humans, impairments in mentalization have been linked to various mental disorders, including schizophrenia (Bora et al., 2009), Autism Spectrum Disorders)ASD) (Lombardo & Baron-Cohen, 2011), and Attention Deficit Hyperactivity Disorder (ADHD) (Pineda-Alhucema et al., 2018). Recent advances in GAI have enabled investigating the performance of GAI systems when conducting social-cognitive tasks. For instance, ChatGPT (OpenAI, 2023) has demonstrated high accuracy in emotion recognition tasks, achieving scores comparable to or exceeding average human performance on the 'Level of Emotional Awareness Scale' (LEAS) (Elyoseph, 2023). Further, GAI has demonstrated substantial abilities in recognizing emotions in facial expressions, as assessed by the 'Reading the Mind in the Eyes Test' (Elyoseph, Refoua, et al., 2024; Refoua et al., 2024). Similarly, ChatGPT-4 was found to perform on par or above human levels on a battery of typical ToM tests (except for faux pas understanding; Strachan et al., 2024). Yet, other earlier studies reported limited social-cognitive capacities of earlier versions of ChatGPT with an accuracy scoring only 10 % above random chance (Sap et al., 2023). Shapira et al. (2023) examined the performance of a variety of LLMs on six different ToM tasks, and concluded that their abilities were not robust, suggesting that the models "rely on shortcuts, heuristics, and spurious correlations, which often lead them astray" (p.8). Taken together, while the question of ToM in GAI is being debated controversially in recent literature (e.g., Mahowald et al., 2024; Perry, 2023), several findings suggest that GAIs can process emotional information from textual and image-based inputs with high accuracy. Yet, to our knowledge, research in the field of GAI-based social cognition has hitherto focused on evaluating these abilities using text-based or static image input only, leaving a significant gap in our understanding of GAI performance in dynamic, multimodal contexts, such as assessments based on audiovisual stimuli of social interaction. Of note, assessment of mentalization abilities based on audiovisual input offers unique advantages over static text or image formats. Video engages multiple dimensions simultaneously, including dynamic visual-spatial cues, linguistic content, and tone of voice, while its temporal aspect adds complexity (Dziobek et al., 2006). Such assessment based on dynamic multimodal material is closer to human experiences during real-world social interactions than static text- or picture-only-based assessments, resulting in higher validity of the test. However, there is a lack of research examining GAI's ability to interpret and respond to the nuanced, context-dependent social cues present in dynamic interactions and how GAI's performance on such tasks compares to that of humans. In 2024, Google made public its Gemini 1.5 Pro model (Google DeepMind, 2024a), showcasing advanced capabilities, including processing input in form of video stimuli. This latest iteration marks a significant leap forward in multimodal GAI, allowing the model to process up to an hour of video, among other complex tasks (Google DeepMind, 2024b).

In this study, we aimed to evaluate the social-cognitive performance of this advanced GAI model on the Movie for the Assessment of Social Cognition (MASC) a video-based social cognition task designed to assess individual performance differences and identify subjects with subtle mentalizing difficulties, showing high sensitivity in detecting impairments in individuals with Asperger syndrome (Dziobek et al., 2006) and other mental disorders, including bipolar disorder (Montag et al., 2010), paranoid schizophrenia (Montag et al., 2011), and borderline personality disorder (Preißler et al., 2010). The MASC's use of a naturalistic social scenario makes it particularly suitable for assessing GAI's ability to interpret complex, real-world social interactions from a third-person-perspective of an observer. We hypothesized that this advanced GAI model would show meaningful performance on the MASC test, though the extent of this performance relative to human levels remained an open empirical question. With the results of the here presented study, we aim to further contribute to the discussion around and help elucidating artificial Theory of Mind in LLMs with relevant potential implications for GAI development and applications in mental health care, social skills training, and the creation of more socially adept GAI systems, including their putative use in IoB innovations.

2. Methods

2.1. Assessment instrument

2.1.1. The MASC

(Dziobek et al., 2006) is a video-based assessment instrument, designed to evaluate mentalizing abilities using naturalistic stimuli of dynamic social interaction. It consists of a video of a 15-min movie depicting four characters (two males and two females) getting together for a dinner party. The video portrays complex social interactions, including instances of friendship, dating, conflicts, and misunderstandings. Throughout the video, viewers are exposed to a variety of social cues including facial expressions, body language, verbal content, and prosody. The characters display a range of emotions and intentions, from subtle to more overt, creating a rich tapestry of social scenarios. The interactions between characters vary in complexity, sometimes involving dyadic exchanges and at other times more complex group dynamics.

When conducting the test, the video is paused multiple times at selected points. At each pause, a question is posed about the characters' feelings, thoughts, or intentions. These in total 45 questions cover different mental state modalities, including emotions, thoughts, and intentions, with varying valence (positive, negative, and neutral). The test incorporates classical social cognition concepts such as first and second-order false beliefs, faux pas, metaphor, sarcasm, and irony. Each item of the MASC is presented in form of a question with a multiple-choice answer format with four options. These options are designed to differentiate between correct mental state inferences and three types of incorrect ones: hyper-mentalizing (representing an over-attribution of mental states beyond what is warranted), hypo-mentalizing (indicating an under-attribution or failure to fully consider mental states), and non-mentalizing (reflecting a complete lack of consideration or failure to recognize mental states). These error types relate to specific social-cognitive deficits in different clinical populations.

In addition to the main questions, the MASC includes several control questions. These questions assess basic comprehension of the plot and non-social aspects of the scenes, ensuring that respondents (or in this case, GAI models) are adequately processing and retaining the content.

Scoring involves assigning one point for correct responses and zero points for incorrect ones. This allows for the derivation of an overall score (maximum 45 points) as well as scores for the different types of mental state inference errors.

2.2. Comparison data

We based our comparisons on data from a doctoral dissertation by McLaren (2023). These data comprised MASC assessments of 1230 undergraduate students (ages 18–25 years, 81 % female) from a large southwestern United States public university, representing diverse ethnic backgrounds: 316 non-Hispanic White, 414 Hispanic White, 151 Black/African American, and 349 Asian/Pacific Islander. Participants were required to have sufficient English fluency and complete all study materials. The mean MASC score of the 1230 participants was 33.19 with a standard deviation (SD) of 5.79. In addition to the overall MASC score, McLaren's (2023) provides a detailed analysis of the types of errors made by the human participants, categorized according to the MASC scoring guidelines into hyper-mentalizing, hypo-mentalizing, and non-mentalizing errors. This error data was originally presented as means and standard deviations for each error type within the specific racial/ethnic subgroups studied (Non-Hispanic White, Hispanic White, Black/African American, Asian/Pacific Islander; see Table 2 in McLaren, 2023). For the purpose of direct comparison with the GAI model in the current study, we calculated the overall distribution of error types across the entire human sample (N = 1230) using weighted averages based on the subgroup sizes and means reported. This yielded an average profile of approximately 5.67 hyper-mentalizing errors (representing 47.0 % of total human errors), 4.20 hypo-mentalizing errors (34.8 % of total human errors), and 2.21 non-mentalizing errors (18.3 % of total human errors) for the human normative sample.

2.3. Generative artificial intelligence tool

This study applied as state-of-the-art LLM Google's Gemini 1.5 Pro (Google DeepMind, 2024a). As of 01–09-2024, Gemini 1.5 Pro was Google's most capable general freely accessible model, featuring multimodal processing capabilities for text, images, audio, and video. As compared to its predecessors, it exhibits improved performance across various tasks, enhanced logical reasoning abilities, and improved factual accuracy, with a context window of up to 2 million tokens (Google DeepMind, 2024b). We accessed the model through the Google AI studio – developer environment (Google DeepMind, 2024c), applying its default configuration, except for the temperature setting, which we varied as part of our experimental design. Temperature in AI models control the randomness of outputs, with lower values producing more deterministic responses and higher values increasing variability. We did not fine-tune or further train the model for this specific task, as our goal was to assess its out-of-the-box performance on the MASC. The model's responses were generated in real-time during the test administration, without any post-processing or manual intervention.

2.4. Verification of Model's naivety to MASC materials

Given the critical importance of ensuring that the AI model's performance reflects genuine social-cognitive processing rather than retrieval of previously learned test-specific information, we undertook several steps to verify, as best as possible, Gemini 1.5 Pro's naivety concerning the MASC materials. This verification is particularly pertinent as the MASC is not an open-source test but is distributed exclusively to researchers under controlled licensing agreements.

First, we reviewed Google's publicly stated policies regarding training data for its large language models. Google's documentation indicates that its training corpora are restricted to data that are either publicly available online or data for which explicit permission has been granted by the owner (Google, 2023; Google Cloud, n.d.-a, n.d.-b). As the MASC materials do not meet these criteria, their inclusion in the model's training data is highly unlikely, aligning with methodological considerations applied in similar GAI assessment research (e.g., Kramer, 2025).

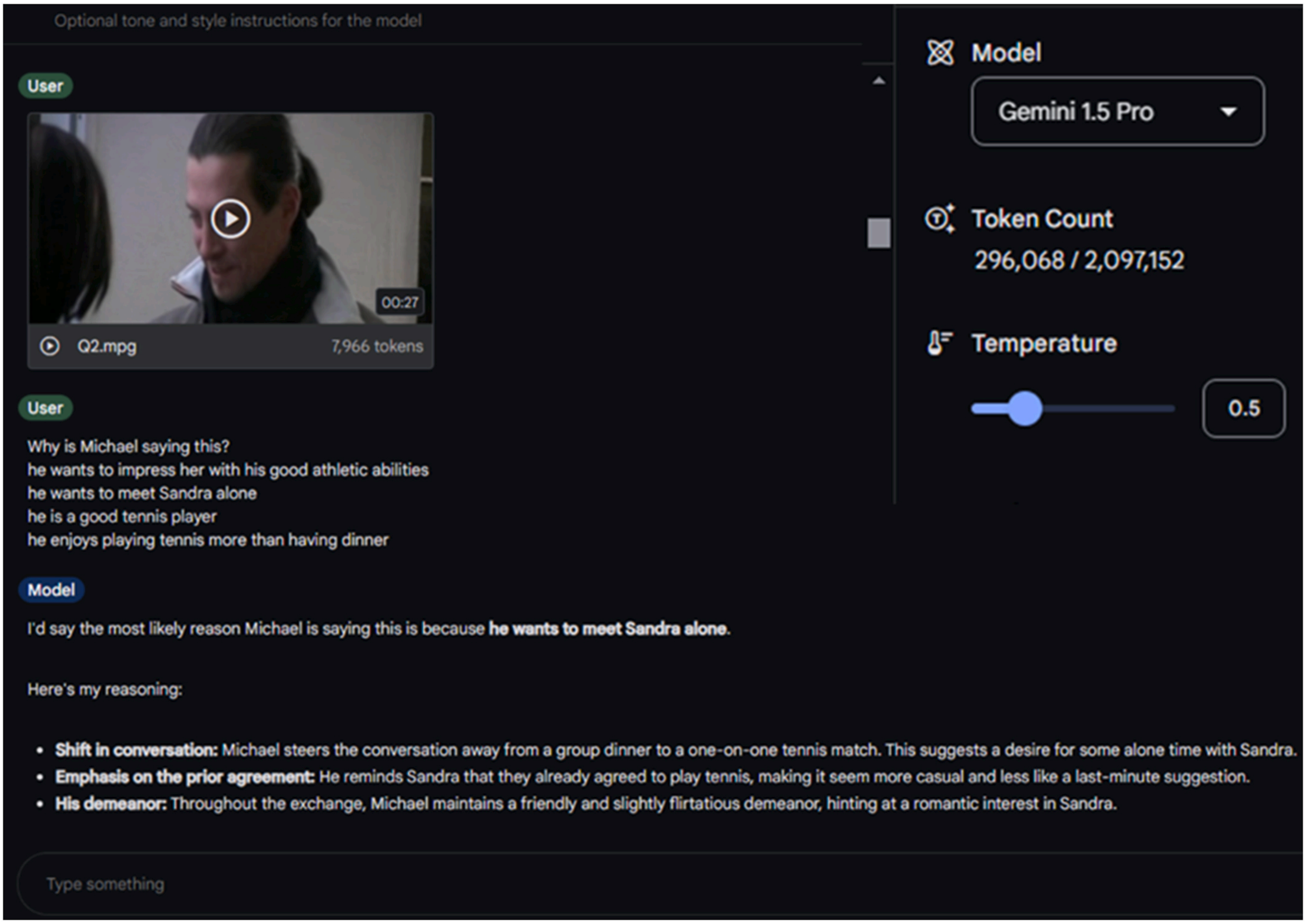
Second, beyond relying on stated policies, we conducted direct empirical probes to assess the model's potential prior knowledge of the MASC materials, framing questions to require specific identification of the test rather than analysis of its content. In separate interactions, without re-presenting the video stimulus in this context, we queried the Gemini 1.5 Pro model regarding its ability to: (i) identify the specific assessment tool (MASC) or the source from which the previously analyzed video originates, (ii) provide the list of specific multiple-choice questions associated with that assessment, and (iii) generate the corresponding correct answers for those questions. Across all these probes, the model consistently failed to provide accurate or specific information that would indicate prior exposure to, or recognition of, the MASC test materials or its specific components.

Furthermore, if the model had been explicitly trained on the MASC questions and their correct answers, performance approaching perfection (i.e., 45/45 correct responses) would be anticipated. However, as detailed in the Results section, the observed performance, while significantly above chance and human average, consistently included errors across all temperature conditions (scores ranged from 36 to 40 out of 45). This imperfection further diminishes the likelihood of direct training data contamination being a primary driver of the model's performance.

Taken together, while absolute certainty regarding the exclusion of specific data from vast training corpora remains challenging, these converging lines of evidence—the MASC's controlled distribution, Google's stated data policies, the model's demonstrated lack of recognition in direct probes, and the non-perfect accuracy observed—substantially mitigate the concern that the model's performance on the MASC was confounded by prior exposure to the test materials. This verification process supports the interpretation of the model's performance as reflecting its intrinsic social-cognitive processing capabilities when presented with novel, dynamic audiovisual social stimuli.

2.5. Procedure

We administered the MASC to the model twice under each of the three temperature conditions: 0, 0.5 and 1, resulting in a total of 6 runs. For each administration, we directly uploaded the MASC video to the Gemini 1.5 Pro model, allowing it to process the actual audiovisual content instead of transcriptions or descriptions. This approach ensured that the model had access to all visual and auditory cues present in the original test material. Each administration of the test was conducted in a single, continuous conversation thread, mimicking the flow of human test-taking. This approach ensured that the model maintained context throughout the test, similar to how a human participant would experience the assessment. The model was presented with the video content and accompanying multiple-choice questions at each pause point, prompting it to select the most appropriate answer from the four options provided. A screenshot of the Gemini 1.5 pro interface is shown in Fig. 1.



Model
Gemini 1.5 Pro

Token Count
296,068 / 2,097,152

Temperature
0.5

Why is Michael saying this?
he wants to impress her with his good athletic abilities
he wants to meet Sandra alone
he is a good tennis player
he enjoys playing tennis more than having dinner

I'd say the most likely reason Michael is saying this is because **he wants to meet Sandra alone.**

Here's my reasoning:

- **Shift in conversation:** Michael steers the conversation away from a group dinner to a one-on-one tennis match. This suggests a desire for some alone time with Sandra.
- **Emphasis on the prior agreement:** He reminds Sandra that they already agreed to play tennis, making it seem more casual and less like a last-minute suggestion.
- **His demeanor:** Throughout the exchange, Michael maintains a friendly and slightly flirtatious demeanor, hinting at a romantic interest in Sandra.

Fig. 1. Interface of the Gemini 1.5 Pro model used during the administration of the MASC test, showcasing the interaction window and selected temperature setting.

2.6. *Statistical analysis*

We conducted the statistical analyses as follows: (1) We applied binomial tests for each of the six model evaluations to determine if performances differed significantly from chance (p = .25), calculating exact Clopper-Pearson 95 % confidence intervals (CIs) for the observed proportions. We applied a Bonferroni correction for multiple comparisons (adjusted α˜.008) to control the familywise error rate. (2) Effect sizes for comparison with chance level were calculated using Cohen's *h* (Cohen, 1988) and Risk Difference (Newcombe, 2006). Corresponding 95 % CIs for these effect sizes were derived from the Clopper-Pearson CIs of the proportions. (3) Percentile rankings were calculated for each run relative to the human normative sample ($N = 1230$; $M = 33.19$, $SD = 5.79$; McLaren, 2023) to contextualize the model's performance. (4) A *z*-test was performed comparing the average score of the model's six evaluations to the human sample mean to assess statistical significance of the overall difference. (5) A Mann-Whitney *U* test was additionally conducted comparing the model's six scores to the human sample distribution as a robust non-parametric verification. (6) To quantify the magnitude of the difference between the AI and human samples, Glass's Δ (Glass et al., 1981) and Hedges' *g*)Hedges, 1981) were calculated. Glass's Δ was selected due to the disparity in variances, using the human sample SD for standardization, while Hedges' *g* provides a bias-corrected estimate based on the pooled standard deviation. 95 % CIs for both Glass's Δ and Hedges' *g* were computed based on the non-central *t*-distribution. (7) Model response consistency between the two runs at each temperature setting was assessed using Cohen's Kappa (κ) and its 95 % CI (Cohen, 1960). (8) The percentage distribution of GAI error types (hyper-, hypo-, and non-mentalizing) across the six evaluations was calculated and compared via a Chi-Square Goodness-of-Fit test ($\chi 2$) to the expected distribution derived from the human normative error profile (McLaren, 2023). All primary statistical tests used a significance level of $p < .05$ (two-tailed), except where Bonferroni correction was applied. Effect size calculations and associated confidence intervals were primarily performed in R (version 4.2.2; R Core Team, 2022), while basic tests were conducted using SPSS (version 29).

3. Results

3.1. *Performance*

Binomial tests indicated that the performance of Gemini 1.5 Pro on

Table 1
Binomial test results for MASC performance comparison between gemini 1.5 pro and chance level (with bonferroni correction).

Run	Score	Proportion	95 % CI [Proportion]		p-value
Temperature 0, Run 1	39	0.87	0.73	0.94	<0.001[a]
Temperature 0, Run 2	37	0.82	0.68	0.92	<0.001[a]
Temperature 0.5, Run 1	39	0.87	0.73	0.94	<0.001[a]
Temperature 0.5, Run 2	40	0.89	0.76	0.96	<0.001[a]
Temperature 1, Run 1	36	0.80	0.65	0.90	<0.001[a]
Temperature 1, Run 2	40	0.89	0.76	0.96	<0.001[a]

Note. CI = Confidence Interval.
[a] p < .05 after Bonferroni correction for multiple comparisons.

the MASC was significantly better than chance guessing (p = .25) across all six evaluations (all corrected *ps* < 0.001; see Table 1). Exact 95 % confidence intervals (CIs) for the proportion correct were calculated using the Clopper-Pearson method (see Table 1). All results remained statistically significant after applying Bonferroni correction for multiple comparisons (adjusted α˜.008).

Gemini 1.5 Pro's performance on the MASC varied across different temperatures (see Fig. 2). At a temperature of 0, the model achieved scores of 39 and 37 out of 45, positioning it at the 84th and 74th percentiles of the human normative sample, equivalent to 0.87 and 0.66 *SDs* above the human normative sample mean. When the temperature was set to 0.5, the model achieved scores of 39 and 40 out of 45, positioning it at the 84th and 88th percentiles (0.87 and 1.17 *SDs* above the human normative sample mean). At a temperature of 1, the model achieved scores of 36 and 40 out of 45, positioning it at the 69th and 88th percentiles (0.48 and 1.17 *SDs* above the human normative sample mean). In all conditions, the LLM correctly answered all control questions, indicating accurate processing of the video content.

We calculated standardized and absolute effect size metrics with 95 % CIs to quantify the magnitude of the model's performance relative to chance level (p = .25; see Table 2). Cohen's *h* ranged from 1.17 to 1.42 across conditions, indicating very large effects (Cohen, 1988) substantially exceeding conventional thresholds. Risk differences ranged from 55.0 % to 63.9 % above chance, demonstrating substantial practical significance (Newcombe, 2006). Corresponding 95 % CIs for these effect sizes are presented in Table 2.

Comparison of the aggregated AI performance (*M* = 38.50, *SD* = 1.64) to the human normative sample (*M* = 33.19, *SD* = 5.79) revealed a statistically significant difference (*Z* = 2.24, *p* = .025), confirmed by a Mann-Whitney test (*U* = 1398.0, *p* = .009). Effect size analyses indicated that the AI model significantly outperformed the human sample average (see Table 3). Specifically, Glass's Δ was 0.92 (95 % CI [0.11, 1.72]), utilizing the human sample's standard deviation, and Hedges' *g* was 0.92 (95 % CI [0.12, 1.72]), using the pooled standard deviation with small-sample correction. Both measures represent large effect sizes.

Analysis of response consistency between the model's two runs at each temperature setting revealed almost perfect agreement, as indicated by Cohen's Kappa values ranging from κ = 0.82 to 0.85 (all *ps* < 0.001). The 95 % confidence intervals further supported substantial to almost perfect agreement across conditions (see Table 4). This indicates a high level of reliability in the model's performance between evaluations, largely independent of the temperature setting.

Table 2
Effect sizes for MASC performance comparison between gemini 1.5 pro and chance level (p = .25).

Run	Cohen's h	95 % CI [h]		Risk Difference (%)	95 % CI [RD]	
Temperature 0, Run 1	1.35	1.01	1.64	61.70	48.20	69.90
Temperature 0, Run 2	1.22	0.89	1.52	57.20	42.90	67.00
Temperature 0.5, Run 1	1.35	1.01	1.64	61.70	48.20	69.90
Temperature 0.5, Run 2	1.42	1.07	1.71	63.90	50.90	71.30
Temperature 1, Run 1	1.17	0.84	1.47	55.00	40.40	65.40
Temperature 1, Run 2	1.42	1.07	1.71	63.90	50.90	71.30

Note. CI = Confidence Interval; RD = Risk Difference (Proportion Correct - 0.25). Risk Difference reported as percentage points above chance. a Confidence intervals derived from the 95 % CI of the proportion correct (Clopper-Pearson method).

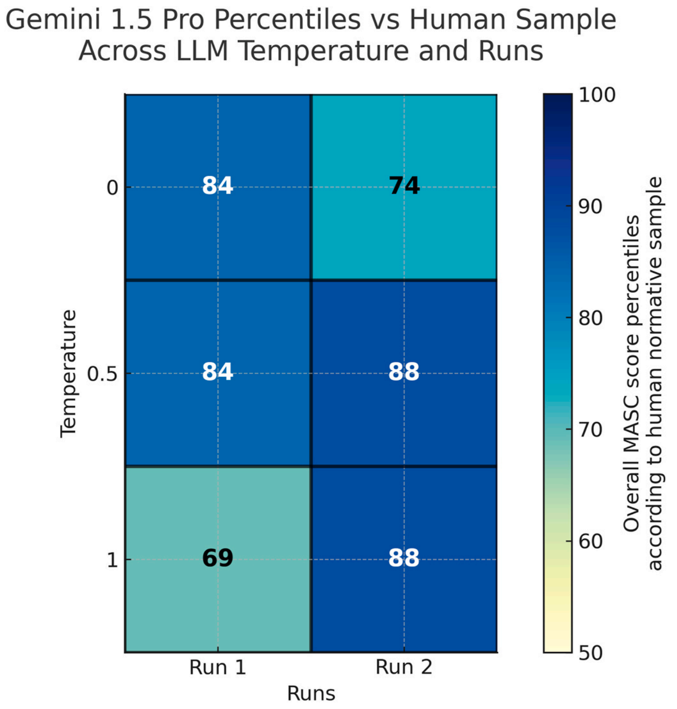


Fig. 2. Heatmap, showing the performance percentiles of the Gemini 1.5 Pro model on the MASC at three temperatures (0, 0.5, 1) and two assessment runs. Abbreviations: LLM, Large Language Model; MASC, Movie for the Assessment of Social Cognition.

Table 3
Effect sizes for MASC performance comparison between gemini 1.5 pro (N = 6) and human sample (N = 1230).

Effect Size Measure	AI Mean (SD)	Human Mean (SD)	Estimate	95 % CI	
Glass's Δ	38.50 (1.64)	33.19 (5.79)	0.92	0.11	1.72
Hedges' g	38.50 (1.64)	33.19 (5.79)	0.92	0.12	1.72

Note. CI = Confidence Interval. Glass's Δ uses the Human sample standard deviation for standardization. Hedges' g uses the pooled standard deviation (5.779) and includes a small-sample bias correction.

Table 4
Inter-run agreement (Cohen's kappa) for gemini 1.5 pro responses.

Temperature Setting	Cohen's κ	95 % CI [κ]	
0	0.82	0.68	0.95
0.5	0.82	0.69	0.95
1	0.85	0.73	0.97

Note. CI = Confidence Interval; κ = Kappa. Agreement calculated between Run 1 and Run 2 for each temperature setting.

3.2. Assessment of errors

The analysis of errors provided insight into the performance shortcomings observed in the Gemini 1.5 Pro model (see Fig. 3). The distribution of errors across all 6 evaluations was as follows: 41.0 % of the errors involved hyper mentalizing, indicating instances where the model attributed more mental states to characters than warranted. 46.2 % of the errors involved hypo mentalizing, indicating instances where the model attributed fewer mental states than necessary. And 12.8 % of the errors involved non-mentalizing, indicating instances where the model failed to recognize or interpret mental states.

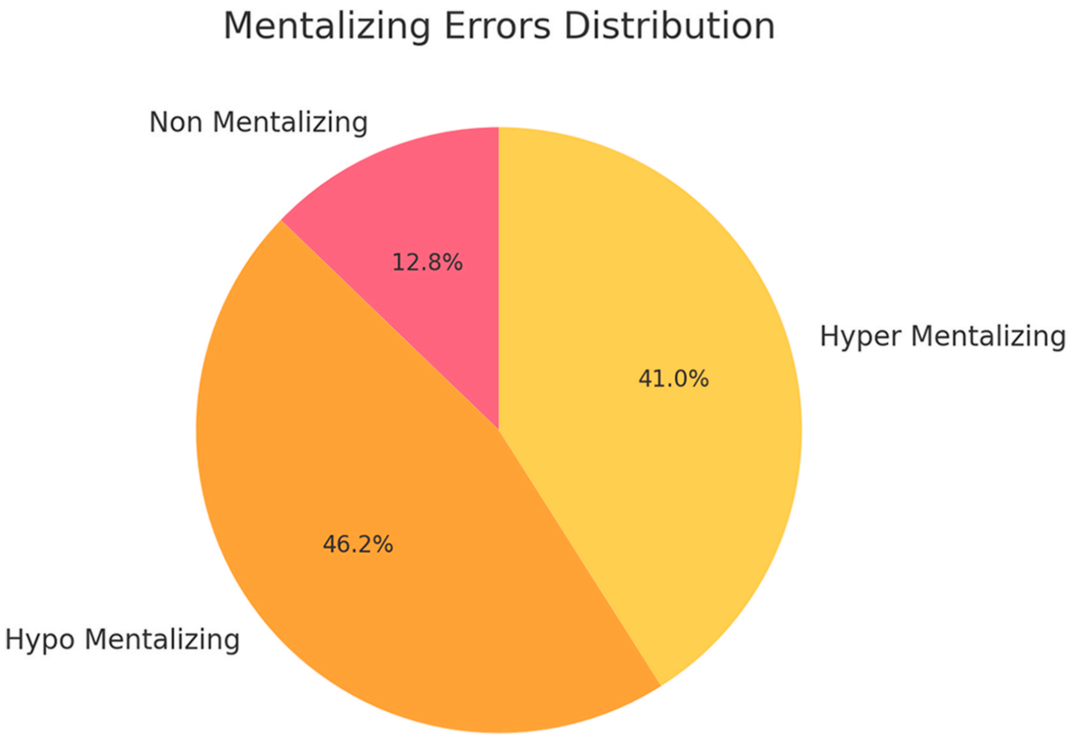


Fig. 3. Pie chart illustrates the distribution of Mentalizing errors made by the Gemini 1.5 Pro model. The errors are categorized into three types: hyper-mentalizing, hypo-mentalizing, and non-mentalizing. Abbreviations: Theory of Mind, TOM.

3.3. Comparison of GAI and human error patterns

The GAI's error distribution and the human normative sample are presented in Table 5. Representative examples from the model's output, illustrating each error type (hyper-mentalizing, hypo-mentalizing, and non-mentalizing), are presented in Appendix A**.** A Chi-Square Goodness-of-Fit test was performed on the observed GAI error frequencies to evaluate whether the overall distribution profiles differed significantly. The test did not yield a statistically significant result **$\chi^2(2, N=39) = 2.36, p = .307$**, indicating that the overall GAI error distribution profile did not significantly deviate from the expected distribution based on the human normative pattern**.**

4. Discussion

The present study evaluated the performance of an advanced GAI system, Gemini 1.5 Pro, on the MASC, a naturalistic assessment of mentalizing abilities using dynamic audiovisual stimuli. Binomial tests indicated that Gemini 1.5 Pro's performances were significantly better than chance across all conditions (all corrected $ps < 0.001$; see Table 1). Large effect sizes were observed when comparing performance to chance level (Cohen's h range = 1.17–1.42; Risk Difference range = 55.0 %–63.9 % above chance; 95 % see Table 2), indicating very large practical significance. Furthermore, Gemini surpassed average human performance across the applied temperature settings. The model's highest scores were achieved at temperatures 0.5 and 1, placing its performances between the 69th and 88th percentiles of a human normative sample. These results indicate its profound capabilities in complex social-emotional interpretation tasks, potentially exceeding typical human performance levels and demonstrating high response reliability across repeated evaluations (κ range = 0.82-0.85; see Table 4).

These findings extend our understanding of GAI's social cognitive capabilities, particularly in complex multimodal processing and theory of mind (ToM). While previous research has demonstrated GAI's proficiency in text-based emotional awareness tasks such as the LEAS (Elyoseph et al., 2023, 2024; Hadar-Shoval et al., 2023) and image-based emotion recognition tests like the RMET (Elyoseph, Refoua, et al., 2024), this study demonstrates GAI's ability to interpret complex social-emotional cues in dynamic, multimodal scenarios approximating real-life interactions. Our findings reveal high performance levels on the MASC, aligning with recent ToM tests (Moghaddam & Honey, 2023; Strachan et al., 2024) and emotion recognition tasks (Elyoseph, Refoua, et al., 2024; Refoua et al., 2024). The magnitude of GAI's performance advantage over the human normative sample represents a large effect size (Glass's Δ = 0.92, 95 % CI [0.11, 1.72]; Hedges' $g = 0.92$, 95 % CI [0.12, 1.72]; see Table 3). An analysis of the pattern of errors was also conducted to provide additional context. A Chi-Square Goodness-of-Fit test comparing the overall GAI and human error distributions did not yield a statistically significant difference indicating that the GAI's overall error profile does not significantly deviates from the human normative pattern. However, it is important to consider that high LLM accuracy can sometimes mask fundamental divergences from human reasoning processes (Sap, 2023; Shapira et al., 2023; Ullman, 2023). Therefore, continued investigation into potential underlying mechanistic differences, possibly reflected in more subtle or task-specific error tendencies, remains a pertinent avenue for future research.

Table 5
Comparison of Error Type Distributions in GAI vs. Human Normative Sample.

Error Type	GAI Model (%)	Human Normative Sample (%)
Hyper-mentalizing	41.0	47.0
Hypo-mentalizing	46.2	34.8
Non-mentalizing	12.8	18.3

Note: Human percentages are weighted averages derived from McLaren (2023).

4.1. Strengths

Our study has several strengths. First, we utilized a standardized, video-based assessment tool (the MASC) that closely mimics real-life social interactions. This approach provides a more ecologically valid measure of social cognition compared to text-based or static image assessments, offering insights into the GAI's performance in interpreting dynamic, multimodal social cues. Second, the MASC is not an open-source test, meaning the GAI model (Gemini 1.5 Pro) was not trained on this specific dataset. This characteristic enhances the validity of our results, as it minimizes the likelihood of the model having had prior exposure to the test material. Third, our study employed real-life large data collected from human participants as a comparison benchmark. This approach allowed for a direct and meaningful comparison between the GAI's performance and actual human subjects, providing a realistic context for interpreting the model's capabilities in social cognition tasks. Fourth, our study utilized Gemini 1.5 Pro, an easily available and commonly used GAI tool. This choice enhanced the practical applicability of our findings, as the study demonstrated the capabilities of an accessible AI system without requiring specialized or custom-developed solutions. This aspect increases the potential for immediate real-world applications and facilitates easier replication and extension of our research by other investigators.

4.2. Limitations

However, our study also has several limitations. First, the use of combined video and audio input modalities limits our understanding of the model's performance with isolated sensory information. This approach may not fully reveal the relative importance of visual versus auditory cues in the GAI's social cognition capabilities. Second, although the MASC assessment materials are distributed under controlled licensing and are not publicly available, minimizing the likelihood of their inclusion in the model's training data according to stated policies (Google, 2023; Google Cloud, n. d.-a, n. d.-b), the possibility of prior exposure cannot be entirely dismissed. We undertook specific verification steps to address this potential confound, including direct empirical probing of the model's recognition and analysis of its performance patterns. These mitigation steps and their outcomes, which suggest the model was indeed naive to the MASC materials, are detailed further in the Methods (see Section 2.4). Third, according to its demographic characteristics, our human comparison sample (McLaren, 2023) is not fully representative of the general population in the US or beyond. As our comparison sample consisted of university students, it is most likely that this sample exhibited above-average performance in emotion recognition tasks. University student populations typically demonstrate higher performance on cognitive tasks compared to general population samples, potentially due to selective educational factors, developmental stage advantages, and cognitive practice effects associated with academic engagement (Henrich et al., 2010). Specifically within social cognition research, educational attainment has been positively associated with theory of mind performance (Tenenberg & Knobelsdorf, 2014), suggesting our comparison sample may represent an elevated benchmark relative to broader population norms. These sampling considerations introduce important contextual parameters for interpreting the comparative positioning of AI performance within the spectrum of human social-cognitive capabilities. Yet, the observation that our GAI model still *outperformed* this sample – that is expected to perform above-average – indicates an even more pronounced superiority of the GAI as compared to human average. Still, future studies with more diverse and representative human comparison samples are warranted to establish fine-grained performance profiles across various demographic segments, including different age cohorts, educational backgrounds, and sociocultural contexts, thereby providing a more comprehensive framework for situating artificial social intelligence within the full spectrum of human capabilities The fourth limitation of this study lies in its assessment of GAI systems' social cognitive abilities solely through observational third-person perspectives, rather than through direct human-AI interactions. This methodological constraint is particularly relevant given that practical applications, especially for individuals with social cognitive impairments, typically involve direct interpersonal engagement. Recent empirical evidence substantiates the significance of this methodological consideration, as demonstrated by Yin et al. (2024) in their investigation of human-AI interactions in emotional support contexts. Their findings reveal a noteworthy paradox: while AI-generated responses surpassed human-generated messages in making recipients feel heard and demonstrated superior emotional detection capabilities, recipients' subjective sense of being understood significantly diminished upon learning of the AI source. This phenomenon further underscores the critical importance of examining GAI performance within authentic interpersonal contexts, particularly when considering therapeutic applications. Fifth, most existing tests, such as the LEAS (Lane et al., 1990), RME (Baron-Cohen et al., 2001), and MASC (Dziobek et al., 2006). Were originally designed to identify clinically relevant impairments in social cognition capacities with clinical samples. Moreover, the multiple-choice format of these tests, while enabling standardized assessment, may not adequately capture the complexity of mental state inference abilities (Oakley et al., 2016; Quesque & Rossetti, 2020). Selecting one out of several given answers allows for pattern recognition strategies rather than demonstrating a genuine understanding of mental states, suggesting that complex social cognition in GAIs should be further tested with ecological open-answer formats. This includes examining the GAI's performance across different languages and cultural variations in social behavior, as the current study was limited to stimulus material from a single cultural context. Finally, the statistical comparison of error patterns, while descriptively suggestive of differences, did not yield a significant result ($p = .307$) and relied on estimated GAI error counts, potentially limiting the strength of conclusions drawn from this specific analysis due to factors like statistical power.

4.3. Critical considerations

While our findings, along with other recent studies, demonstrate GAI models' proficiency in mentalization and social cognition tasks, it is imperative to acknowledge that these systems process and interpret social information through fundamentally different mechanisms than humans, mechanisms that remain largely unknown to our current understanding. The descriptive nature of our results, while valuable, necessitates caution against anthropomorphizing GAI's performance in social-cognitive tasks, among researchers, clinicians, and the general public (Pelau et al., 2021). This cautionary stance is theoretically grounded in fundamental differences between human and artificial social cognition. Human mindreading, as Gallese (2007) demonstrates, cannot be reduced to purely computational processes in dedicated brain modules. Instead, it emerges from an embodied simulation system, fundamentally rooted in the premotor cortex and mirror neuron system, which enables direct experiential understanding of others' actions and intentions.

While standardized instruments like the MASC provide valuable insights, they capture only a subset of the complex abilities required for authentic social understanding. The distinction between successfully completing pre-determined multiple-choice tasks and demonstrating genuine social comprehension in fluid, contextual situations remain substantial. This consideration becomes especially critical in clinical applications, where misinterpreting social cues or mental states could significantly affect patient care and outcomes. Extensive research, including clinical trials, would be necessary to establish GAIs efficacy and safety. One significant concern is that while GAI can foster trust and build rapport with users, it could become a potent instrument in the hands of entities that may prioritize economic gains over genuine support and care (Elyoseph, Refoua, et al., 2024). The risk is particularly acute if GAI systems will intrusively analyze personal conversations, behaviors, and emotions without explicit user consent (Coghlan et al., 2023). Furthermore, any development in this area must prioritize user autonomy, privacy concerns, and the diverse preferences within the targeted communities. Mental health professionals and developers must be vigilant against the influence of epistemic bias in their practice, striking a balance between using GenAI tools and retaining the essential human elements of empathy, intuition, and clinical judgment (Rubin et al., 2024). Additionally, developers and practitioners must consider the risk that people in emotional need may become dependent on or attached to GenAIs in potentially nonadaptive ways, particularly as these systems are designed to foster trust and emotional connection (Munn & Weijers, 2023). This raises fundamental questions about the authenticity of human-AI therapeutic relationships and their potential impact on mental well-being.

4.4. Future directions

Future research in this field should pursue several key directions to further our understanding of GAI's capabilities in social cognition and its potential applications in real-world settings. First, studies should primarily focus on assessing the GAI's performance using only video input. This approach would help disentangle the relevance of visual cues and auditory cues for the model's social cognition abilities. Future research may explore the opportunities and risks associated with integrating such GAI models into real-world IoB systems. Such studies would assess their performance in dynamic, uncontrolled environments, providing insights into the practical applicability of these technologies. Additionally, research into enhancing the transparency and interpretability of these GAI systems will be crucial for their responsible implementation. This line of inquiry could focus on developing and applying methods to explain the GAI's decision-making processes in social cognition tasks, which is expected to be essential for building trust and ensuring ethical use in sensitive applications such as mental health support or social skills training.

Longitudinal studies examining the long-term effects of interaction with socially intelligent GAI on human behavior and social skills may provide valuable insights. These studies could explore how prolonged exposure to GAI systems with advanced social cognition capabilities might influence human social development, particularly in vulnerable populations or those with social cognitive difficulties. In addition, cross-cultural studies would allow evaluating GAI's performance across different languages and cultural contexts, informing about potential cultural biases requiring to be addressed in future developments. Overall, interdisciplinary collaborations between GAI researchers, psychologists, ethicists, and healthcare professionals may facilitate exploring these technologies' potential applications and implications in various domains, particularly in mental health and social support services.

As future research provides greater insight into GAI's social-cognitive capabilities, integrating it with emerging technologies could be a promising avenue for investigation. For example, in smart city environments, GAI with advanced social cognition could enhance public services by better-interpreting citizen needs and behaviors. This capability aligns closely with the emerging field of IoB. In mental health contexts, the integration of GAI systems capable of advanced social cognition with IoB tools could lead to more nuanced and personalized interventions. More specifically, such GAI enhanced IoB devices could be beneficial for individuals with conditions characterized by difficulties in emotional recognition and social interaction, such as autism spectrum disorders (ASD) or alexithymia. The technology's capability to analyze complex social cues and emotional expressions could potentially support and assist subjects suffering from related impairments. Thereby, in the near or mid-term future, GAI systems could facilitate the creation of personalized tools for emotional recognition training, social skills development, and real-time support in social situations. Another field of application may be psychotherapy training, where GAI could assist in analyzing video recordings of therapy sessions, offering insights into therapist interventions and client emotional responses (Fiske et al., 2019; Luyten et al., 2020). This application may contribute to improving the quality of clinical supervision and supporting therapists' skill development. For example, the technology could provide detailed feedback on therapist-client interactions, help identify patterns in therapeutic approaches, and potentially suggest areas for improvement in therapeutic techniques.

4.5. Conclusions

In conclusion, results from our study represent a significant advancement in our understanding of advanced GAI's capabilities in complex social-emotional reasoning. The performance of the GAI under study on the MASC demonstrated a level of social cognitive performance that is in par or even exceeds average human capabilities in certain aspects. These findings indicate relevant opportunities for applications in mental health care, social skills training, and assistive technologies. However, the yet open question whether this study indicates the next frontier of mindreading in GAI should be further investigated. Our work underscores the need for continued research to fully understand the mechanisms, implications and limitations of these abilities, as well as the ethical considerations that must guide their responsible development and application in real-world settings.

CRediT authorship contribution statement

Elad Refoua: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Zohar Elyoseph:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Data curation, Conceptualization. **Renata Wacker:** Writing – review & editing, Methodology. **Isabel Dziobek:** Writing – review & editing, Methodology. **Iftach Tsafrir:** Writing – review & editing, Methodology, Data curation. **Gunther Meinlschmidt:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Generative AI usage declaration

During the preparation of this work, the authors utilized generative artificial intelligence (GAI) tools to assist with language editing and clarity of expression. Specifically, we used Claude, ChatGPT, and Grammarly to help refine the language and improve the overall readability of the text. These AI assistants were employed to enhance the linguistic quality of the manuscript, including grammar, style, and clarity of expression. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interest

Appendix A. Examples of GAI Mentalizing Errors on the MASC

A.1. Example: Hypo-mentalizing Error

- **Context:** A character encounters another character's unexpected dog, prompting a question about the character's feelings.
- **Error Description:** The correct inference identifies a specific negative emotion. The AI's interpretation, however, focused primarily on the element of surprise, reasoning that the character's reaction ("His facial expression and the way he sort of jumps back a bit") primarily suggested he "wasn't expecting a dog to be there". By emphasizing only the unexpectedness based on these cues, the AI potentially overlooked or under-represented the intensity or specific nature of the character's likely primary emotional reaction. This under-attribution reflects a Hypo-mentalizing error.

A.2. Example: Non-mentalizing Error

- **Context:** Following a provocative remark by one character during cooking preparations, another character responds by assigning the first character a specific kitchen task. The question probes the responding character's intention.
- **Error Description:** The correct inference centers on the responding character's internal motivation, specifically a desire for social payback or an emotional reaction to the initial remark. The AI's interpretation acknowledged the context of the "sexist comment" and recognized the character might be "enjoying this little bit of payback" due to a "sly smile", yet its final assessment ultimately prioritized the literal behavioral outcome of "get[ting] him involved in the cooking process". This failure to prioritize the inferred mental state (motivation/payback) over the behavioral action, despite identifying it during reasoning, represents a Non-mentalizing error.

A.3. Example: Hyper-mentalizing Error

- **Context:** During a phone call discussing a social arrangement, one character expresses reluctance, which seems potentially linked to another character expected to be involved. The question probes the speaker's feelings.
- **Error Description:** The correct inference describes a relatively straightforward feeling of reluctance about the arrangement. The AI's interpretation, however, attributed more complex negative interpersonal states, reasoning that the character seemed "a bit exasperated" due to perceived "manipulation" by the caller and because the character anticipated annoyance ("already knows that the other character can be annoying"). This extensive inference of intricate negative states and interpersonal dynamics beyond what was clearly warranted by the immediate interaction exemplifies a Hyper-mentalizing error.

Data availability

Data will be made available on request.

References

Badian, Y., Ophir, Y., Tikochinski, R., Calderon, N., Klomek, A. B., & Reichart, R. (2023). A picture may Be worth a thousand lives: An interpretable artificial intelligence strategy for predictions of suicide risk from social media images. https://doi.org/10.48550/ARXIV.2302.09488.

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. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241–251.

Bora, E., Yücel, M., & Pantelis, C. (2009). Theory of mind impairment: A distinct trait-marker for schizophrenia spectrum disorders and bipolar disorder? *Acta Psychiatrica Scandinavica, 120*(4), 253–264.

Cannarsa, M. (2021). Ethics guidelines for trustworthy AI. In L. A. DiMatteo, A. Janssen, P. Ortolani, F. De Elizalde, M. Cannarsa, & M. Durovic (Eds.), *The cambridge handbook of lawyering in the digital age* (1st ed., pp. 283–297). Cambridge University Press. https://doi.org/10.1017/9781108936040.022.

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital Health, 9*, Article 20552076231183542. https://doi.org/10.1177/20552076231183542

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636.

Elyoseph, Z., Gur, T., Haber, Y., Simon, T., Angert, T., Navon, Y., Tal, A., & Asman, O. (2024). An ethical perspective on the democratization of mental health with generative AI. *JMIR Mental Health, 11*, Article e58011. https://doi.org/10.2196/58011

Elyoseph, Z., Hadar Shoval, D., & Levkovich, I. (2024). Beyond personhood: Ethical paradigms in the generative artificial intelligence era. *The American Journal of Bioethics, 24*(1), 57–59.

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology, 14*, Article 1199058.

Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. *JMIR Mental Health, 11*, Article e54369.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), Article e13216. https://doi.org/10.2196/13216

Fonagy, P., Gergely, G., & Jurist, E. L. (2018). *Affect regulation, mentalization and the development of the self*. Routledge.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Sage Publications.

Google. (2023). Google privacy policy. https://policies.google.com/privacy.

Google Cloud. (n.d.-a). Data governance for generative AI. Google Cloud. https://cloud.google.com/vertex-ai/docs/generative-ai/data-governance.

Google Cloud. (n.d.-b). Overview of Generative AI on Vertex AI. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/overview.

Google DeepMind. (2024a). *Gemini 1.5 pro* [Large language model]. https://deepmind.google/technologies/gemini/.

Google DeepMind. (2024b). Our next-generation model: Gemini 1.5 [Blog post]. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#performance.

Google DeepMind. (2024c). *Google AI studio* [development platform]. https://ai.google.dev/.

Haber, Y., Levkovich, I., Hadar-Shoval, D., & Elyoseph, Z. (2024). The artificial third: A broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health, 11*, Article e54781.

Hadar-Shoval, D., Elyoseph, Z., & Lvovsky, M. (2023). The plasticity of ChatGPT's mentalizing abilities: Personalization for personality structures. *Frontiers in Psychiatry, 14*, Article 1234397.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21*(2), 155–172.

Javaid, M., Haleem, A., Singh, R. P., Rab, S., & Suman, R. (2021). Internet of Behaviours (IoB) and its role in customer services. *Sensors International, 2*, Article 100122.

Kramer, R. S. (2025). Comparing ChatGPT with human judgements of social traits from face photographs. *Computers in Human Behavior: Artificial Humans*. , Article 100156. https://doi.org/10.1016/j.chbah.2025.100156

Lane, R. D., Quinlan, D. M., Schwartz, G. E., Walker, P. A., & Zeitlin, S. B. (1990). The levels of emotional awareness Scale: A cognitive-developmental measure of emotion. *Journal of Personality Assessment, 55*(1–2), 124–134.

Lauderdale, S. A., Schmitt, R., Wuckovich, B., & Desai, H. (2024). *Assessing an AI chatbot's recognition of ptsd symptoms and evidence-based treatments for veterans: A comparative study with ChatGPT-4 and human participants*. https://doi.org/10.13140/RG.2.2.31025.19041

Lombardo, M. V., & Baron-Cohen, S. (2011). The role of the self in mindblindness in autism. *Consciousness and Cognition, 20*(1), 130–140.

Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: State of the art and future directions. *Annual Review of Clinical Psychology, 16*(1), 297–325. https://doi.org/10.1146/annurev-clinpsy-071919-015355

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28*(6), 517–540.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46–60.

McLaren, R. (2023). An investigation of cultural sensitivity and racial bias in the movie for the assessment of social cognition. https://uh-ir.tdl.org/items/3d7991e3-0fa9-4be0-9aaa-043683b4c69a.

Moghaddam, S. R., & Honey, C. J. (2023). Boosting theory-of-mind performance in Large Language models via prompting. https://doi.org/10.48550/arXiv.2304.11490.

Montag, C., Dziobek, I., Richter, I. S., Neuhaus, K., Lehmann, A., Sylla, R., Heekeren, H. R., Heinz, A., & Gallinat, J. (2011). Different aspects of theory of mind in paranoid schizophrenia: Evidence from a video-based assessment. *Psychiatry Research, 186*(2–3), 203–209.

Montag, C., Ehrlich, A., Neuhaus, K., Dziobek, I., Heekeren, H. R., Heinz, A., & Gallinat, J. (2010). Theory of mind impairments in euthymic bipolar patients. *Journal of Affective Disorders, 123*(1–3), 264–269.

Munn, N., & Weijers, D. (2023). Corporate responsibility for the termination of digital friends. *AI & Society, 38*(4), 1501–1502. https://doi.org/10.1007/s00146-021-01276-z

Newcombe, R. G. (2006). A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine, 25*(24), 4235–4240. https://doi.org/10.1002/sim.2683

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the reading the mind in the Eyes test. *Journal of Abnormal Psychology, 125*(6), 818–823.

OpenAI. (2023). ChatGPT [Large language model]. https://chat.openai.com.

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports, 10*(1), Article 16685. https://doi.org/10.1038/s41598-020-73917-0

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*, Article 106855. https://doi.org/10.1016/j.chb.2021.106855

Perry, A. (2023). AI will never convey the essence of human empathy. *Nature Human Behaviour, 7*(11), 1808–1809.

Pineda-Alhucema, W., Aristizabal, E., Escudero-Cabarcas, J., Acosta-López, J. E., & Vélez, J. I. (2018). Executive function and theory of mind in children with ADHD: A systematic review. *Neuropsychology Review, 28*(3), 341–358.

Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social cognition in borderline personality disorder: Evidence for disturbed recognition of the emotions, thoughts, and intentions of others. *Frontiers in Behavioral Neuroscience, 4*.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science, 15*(2), 384–396.

Refoua, E., Meinlschmidt, G., & Elyoseph, Z. (2024). Generative artificial intelligence demonstrates excellent emotion recognition abilities across ethnical boundaries. https://doi.org/10.2139/ssrn.4901183.

Rubin, M., Arnon, H., Huppert, J. D., & Perry, A. (2024). Considering the role of human empathy in AI-driven therapy. *JMIR Mental Health, 11*, Article e56529. https://doi.org/10.2196/56529

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). https://doi.org/10.48550/arXiv.2210.13312.

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). https://doi.org/10.48550/arXiv.2305.14763.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *Npj Mental Health Research, 3*(1), 1–12.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour, 8*(7), 1285–1295.

Tenenberg, J., & Knobelsdorf, M. (2014). Out of our minds: A review of sociocultural cognition theory. *Computer Science Education, 24*(1), 1–24. https://doi.org/10.1080/08993408.2013.869396

Ullman, T. (2023). *Large Language models Fail on trivial Alterations to theory-of-mind tasks* (arXiv:2302.08399). https://doi.org/10.48550/arXiv.2302.08399.

Webster, C., & Ivanov, S. (2020). Robotics, artificial intelligence, and the evolving nature of work. In B. George, & J. Paul (Eds.), *Digital transformation in business and society* (pp. 127–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-08277-2_8.

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences, 121*(14), Article e2319112121.