Assessing mentalization in humans and large language models

Aamir Sohail [1,2,3*], **Xintong Zhong** [4], **Arkady Konovalov** [1,2], **Patricia L. Lockwood** [1,2,3,5,6], **Lei Zhang** [1,2,3,5*]

[1] *Centre for Human Brain Health, University of Birmingham, Birmingham, UK.*
[2] *School of Psychology, University of Birmingham, Birmingham, UK.*
[3] *Institute for Mental Health, University of Birmingham, Birmingham, UK.*
[4] *Department of Psychology, Virginia Tech, Blacksburg, USA.*
[5] *Centre for Developmental Science, University of Birmingham, Birmingham, UK.*
[6] *Department of Experimental Psychology, University of Oxford, Oxford, UK.*

*** Corresponding author emails:** axs2210@student.bham.ac.uk; l.zhang.13@bham.ac.uk

Abstract

Mentalization - the ability to infer others' beliefs and intentions to guide one's own choices - is a key cognitive function underlying human social interactions. Large language models (LLMs) demonstrate behaviour consistent with humans on theory-of-mind tasks, yet whether these models can guide adaptive behaviour through mentalization is unknown. Here we use two economic games with cognitive computational modeling to uncover the latent strategies underlying mentalization in LLMs. We tested individual LLM agents across four model families, DeepSeek, GPT-4.1, GPT-5 and Gemini 2.0 Flash (N = 2,099), against opponents of varying sophistication and examined whether a prompting strategy designed to elicit strategic reasoning improved performance. We benchmarked results against human participants (N = 251) as a comparative measure. Across both games, LLMs showed clear behavioural and computational signatures of mentalizing that differed markedly by model provider and size. Strategic prompting generally improved performance by inducing more sophisticated reasoning, yet the extent of the benefit differed across the two tasks. Last, GPT-5 agents flexibly adapted their recursive depth of reasoning to increasingly sophisticated opponents, demonstrating superior performance to human participants. Collectively, we demonstrate different capacities for mentalization across LLMs, and highlight cognitive computational modeling as a formal method for assessing comparative intelligence across humans and machines.

Keywords: *mentalization, large language models, artificial intelligence, computational modeling*

Main

Mentalization is a cognitive process associated with interpreting the behaviour of others as the result of latent mental states, beliefs and emotions [1–3]. In humans, mentalization shapes decision-making in social contexts by predicting the actions of others and adjusting one's own behaviour accordingly [4–10]. Evidence also has suggested the presence of mentalization in a select range

of non-human animal species, including dogs, corvids, chimpanzees and gorillas [11–20] reflecting an evolved capacity for social intelligence.

Beyond humans and other animals, a topic of significant recent development concerns whether generative artificial intelligence (AI) systems, such as large language models (LLMs) demonstrate behaviour consistent with mentalizing. Researchers have turned to experimental methods applied within the fields of psychology, cognitive science and neuroscience to uncover features of behaviour and reasoning abilities in LLMs [21–30]. By converting behavioural tasks to text which are submitted as prompts, assessments of learning and decision-making - typically used among human participants - can be directly applied to LLMs, providing deeper computational insights beyond superficial measures of choice [31–37].

Despite being a type of neural network trained upon large quantities of text-based data [38], LLMs demonstrate emergent properties synonymous with reasoning and decision-making in humans [21, 31, 39–48]. In humans, mentalization is conventionally assessed using role-based tasks, stories necessitating the inference of fictional characters' beliefs [49, 50]. Large language models are similarly able to correctly interpret mental states and suggest appropriate actions, often performing at or above human level in such tasks [51–57]. However, LLMs fail at simple modifications to these assessments [58] and struggle in theory-of-mind (ToM) tasks necessitating complex reasoning and planning [55, 59, 60], casting doubt as to whether successes reflect genuine mentalization or shallow heuristics [61–63]. As a result, it has been recently argued that the role-based tasks commonly used may not be suitable for assessing complex cognitive processes [61], by only describing superficial elements of behaviour [64–69] which may be solved without requiring the explicit, human-like simulation of mental states [70]. Therefore, whilst the observed choices between human and LLMs may be identical, this does not necessarily reflect the same inferential or cognitive process [65, 66, 71, 72].

Contrasting with story-based assessments, economic games present a formal methodology for assessing strategic social behaviour by examining distinct behavioural and computational mechanisms [73–79]. In these games, participants are directly part of the interaction rather than a third-party observer, and are often tasked with outplaying an opponent [e.g., the Inspection Game; [80]]. Importantly, this context allows researchers to test recursive mentalization at different levels by experimentally manipulating the reasoning depth of the opponent. Under this framework, a player may attempt to estimate the opponent's historical choice distribution and form first-order beliefs, or – if anticipating that their actions are being tracked - form second-order beliefs and modify their strategy according to her opponent's reaction [77, 81, 82]. Accordingly, this reasoning process can also reach higher levels. Conventionally used to assess behaviour among human participants, economic games have recently been employed to objectively measure specific biases and patterns of decision-making in LLMs [21, 31, 39–46, 48, 83–92], which demonstrate emergent properties of deliberative thought. Furthermore, the performance of LLMs on these tasks is often improved using targeted prompting strategies – such as chain-of-thought (CoT) – [93–95], designed to elicit step-by-step reasoning indicative of human decision-making [40, 48, 96–100].

In humans, choice behaviour can be formally analysed using cognitive computational modeling, a theory-driven quantification ascribing latent cognitive processes and computational parameters

to observed actions [101–105]. In the context of dynamic social behaviour, these parameters can additionally be assessed trial-by-trial to uncover temporal patterns of cognition. Concerning mentalization, computational models have highlighted the capacity for recursive mentalization in humans [80, 106] who are able to dynamically adapt their recursive depth to the inferred strategy of the opponent, i.e., *adaptive* mentalization [107]. Cognitive models also offer a mechanistic formalization of behaviour across species [108, 109], solely requiring choice data to infer upon cognitive processes. Subsequently, computational modeling has recently been endorsed in AI and LLM research as a formal approach to infer the internal processes of AI systems from observed choices [110, 111]. For example, this approach has uncovered latent parameters underlying learning biases in LLMs, by fitting reinforcement learning (RL) models [112–114]. To date, a comparative assessment of mentalization across humans and LLMs using strategic games has yet to be undertaken, despite a growing interest towards developing 'human-like' machine intelligence [26, 115].

In the current study, we employ two commonly used economic games – the inspection game and rock-paper-scissors - to uncover the behavioural signatures and latent parameters of recursive mentalization in large language models. We include several popular LLMs that were subjected to the experimental tasks as participants (total N = 2,099), and examine the impact of a simple prompting strategy – Social Chain-of-Thought (SCoT) – previously shown to improve LLM performance in economic games [40]. We also compare their performance to human participants (total N = 251; mean age = 24.48 $\pm$ 3.81 years; range 18–36) across both tasks, analysing differences in recursive mentalization across humans and LLMs. In doing so, we move beyond conventional descriptive assessments and provide a more formal understanding of machine intelligence.

Results

For our study, we selected two repeated-interaction games commonly used to assess strategic play in humans, the Inspection game and rock-paper-scissors (RPS). To assess recursive mentalization, we employed the inspection game [80, 106], a repeated non-cooperative game where two players assume the roles of an "employee" and an "employer" and make one of two choices simultaneously before receiving a payoff (Fig. 1a). On each trial, the employee chooses whether to "work" or "shirk," and the employer chooses whether to "inspect" or "not inspect" the employee. In our version, the participants (i.e., human/LLM) always played the role of the employee. Subsequently, the payoff matrix is specified such that participants earn points by adapting to the opponent's predicted choice. Participants played against a computer algorithm (designated as the employer, operating at level $k = 2$, second-order belief) designed to make adaptive choices to participants' expectations over the course of the task. Therefore, the total payoff received reflects the participants' capacity for recursive reasoning. We also tested participants' ability for *adaptive* mentalization; the ability to flexibly adapt the level of recursive mentalization to opponents of different sophistication. For this, we employed a zero-sum RPS game previously used with human participants [107]. In contrast to the inspection game, which is played against a fixed algorithm across all trials, participants in the RPS game played against an algorithm programmed to make choices in accordance with a specific reasoning level k, where $k \in \{0, 1, 2\}$. On each trial, participants received a single point

for winning, lost a single point for losing and received no points in the case of a tie. Participants played a block of 40 trials against each *k*-level (a total of 120 trials across 3 blocks) with blocks randomised and counterbalanced across participants (Fig. 1b).

Data were collected in both human participants (inspection game: n = 67; RPS: n = 184) and LLMs (inspection game: all groups n $\geq$ 49; RPS: all groups n $\geq$ 188; see Methods). Human participants in both tasks were provided with instructions that they would be playing against a real other person, but were in fact playing against the respective computer algorithm. Human data for the inspection game were collected in person, and online for the RPS game (see Methods). Tasks were appropriately re-coded into a text-based format for API inference in LLMs (see Methods/SI) (Fig. 1c). Converted tasks retained identical payoff rules, opponent configurations and trial lengths, however, as LLMs demonstrate bias in the conventional RPS game [116], choices were re-coded in the RPS game to letters J/Q/Z that did not carry semantic meaning for LLMs; an established protocol commonly used to control for text biases [40]. To uncover the possible effects of strategic prompting on mentalization, LLMs were also prompted with Social Chain-of-Thought (SCoT) [40], a prompting scheme where LLMs are specifically asked to make a prediction on the opponents action in each trial before making their choice. The following LLMs were selected for participation in the respective tasks. Inspection game: Gemini, DeepSeek, GPT-4.1, GPT-5; RPS: DeepSeek, GPT-5 (Table 1). The selections were made considering practical constraints with monetary costs for running the study and the availability of specific models at the time of data collection (4/8/2025 – 14/10/2025). Where possible, LLMs were tested under both conditions (conventional prompt + SCoT), with the deliberate exception of GPT-5 which natively supports deliberative reasoning [117].

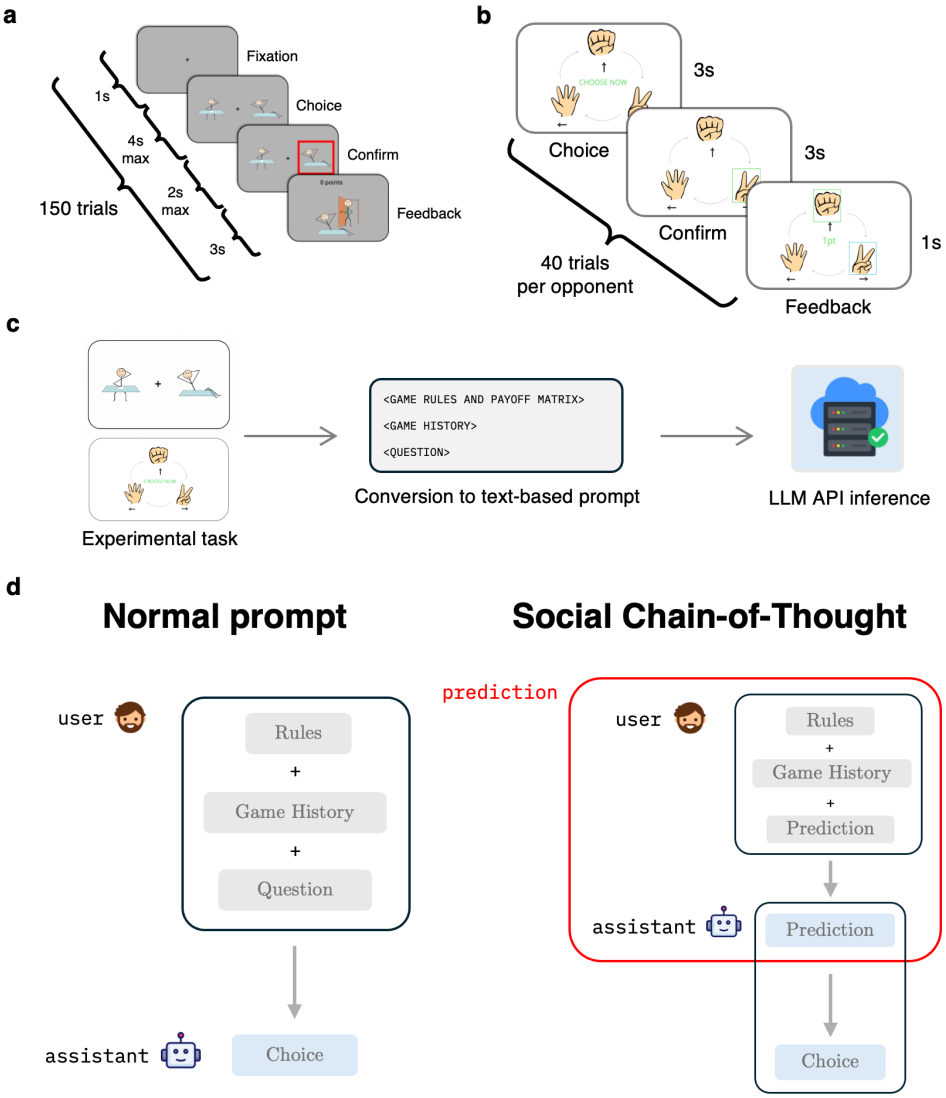


Figure 1: **Experimental design for humans and large language models. (a)** For the inspection game, on each trial human participants were firstly presented with a fixation cross (1 second) and had a maximum of 4 seconds to make their choice, followed by the choice confirmation (max 2 seconds). They were then provided feedback based on the choice of the opponent (3 seconds). Across a single block of 150 trials, humans played against a sophisticated algorithm designed to engage in the 'Influence play' signature of reasoning. **(b)** In the RPS task, humans on each trial had a maximum of 3 seconds to make their choice, after which they received feedback for 1 second. Participants played 3 blocks of 40 trials against an adaptive algorithm which made decisions reflecting different levels of mentalization, from zero-order (k=0) to second-order (k=2). The block order was counterbalanced and randomised across participants. **(c)** For large language model (LLM) participants, each task was converted into a text-based prompt, matching the original task where possible. These were then submitted to the LLMs for API inference. **(d)** Prompting conditions across both tasks. In the normal prompt, the task rules and choice history were presented along with a statement asking for the agent's choice. In the SCoT condition, the agent was instead asked to predict the opponent's choice. This prediction was then used to select the corresponding rational choice. Icons from Flaticons.com.

Table 1: **An overview of the algorithms and large language models used in the study.** *Note.* CHASE = Cognitive Hierarchy Assessment; EWA = Experience Weighted Attraction; ME = Mixed-equilibrium; RL = Reinforcement Learning; ToMk = Theory-of-Mind *k*.

Inspection game		**Rock paper scissors**	
Participants	Human participants DeepSeek / DeepSeek-SCoT Gemini / Gemini-SCoT GPT-4.1 / GPT-4.1-SCoT GPT-5	**Participants**	Human participants DeepSeek / DeepSeek-SCoT GPT-5
Models	Influence play Fictitious play RL Mixed-equilibrium (ME)	**Models**	CHASE Reward learner Self-tuning EWA Full EWA Fictitious play ToMk

Assessing mentalization using the inspection game

The LLMs altogether consisted of seven groups: gemini-2.0-flash (Gemini), gemini-2.0-flash with SCoT (Gemini-SCoT), deepseek-chat (DeepSeek), deepseek-chat with SCoT (DeepSeek-SCoT), gpt-4.1-2025-04-14 (GPT-4.1), gpt-4.1-2025-04-14 with SCoT (GPT-4.1-SCoT), and gpt-5-2025-08-07 (GPT-5). We first tested how performance on the inspection game differed across each LLM (all n > 48) and prompting strategy compared with the human group (n = 67). As LLM groups met normality assumptions but presented unequal variance, Welch's independent samples t-tests were used, with Bonferroni correction applied across the seven comparisons representing the LLM model type $\times$ prompting strategy combination ($\alpha = 0.007$). Altogether, most combinations earned significantly different payoffs from humans. Specifically, Gemini (M_{payoff} = 14.2; $t_{(112.01)}$ = -36.6, p = 3.58×10^{-64}, g = -6.40, 95% CI [-7.29, -5.49]), Gemini-SCoT (M_{payoff} = 22.4; $t_{(109.42)}$ = -13.7, p = 1.59×10^{-25}, g = -2.37, 95% CI [-2.84, -1.89]), DeepSeek (M_{payoff} = 18.0; $t_{(110.59)}$ = -26.5, p = 1.16×10^{-49}, g = -4.59, 95% CI [-5.28, -3.89]), DeepSeek-SCoT (M_{payoff} = 25.1; $t_{(106.59)}$ = -6.04, p = 2.32×10^{-8}, g = -1.04, 95% CI [-1.42, -0.65]) and GPT-4.1 (M_{payoff} = 18.5; $t_{(95.42)}$ = -27.2, p = 8.7×10^{-47}, g = -4.57, 95% CI [-5.26, -3.88]), all earned significantly less than humans. The only exceptions were GPT-4.1-SCoT (M_{payoff} = 27.8; $t_{(114.78)}$ = 1.74, p = 0.085, g = 0.31, 95% CI [-0.06, 0.68]) and GPT-5 (M_{payoff} = 26.6; $t_{(112.18)}$ = -1.55, p = 0.123, g = -0.26, 95% CI [-0.57, 0.05]), which did not significantly differ from humans. We next sought to understand how the choice of LLM and the prompting strategy jointly affected performance. For the three models tested under both strategies (DeepSeek, Gemini, GPT-4.1), we fitted a 2 (prompt) $\times$ 3 (model) factorial ANOVA to mean payoff per trial. As cell variances were unequal (Levene's $F_{(5, 293)}$ = 3.79, p = 0.002), we report Type-III Wald F-tests. We found significant main effects of model ($F_{(2, 293)}$ = 286.95, p = 9.62×10^{-70}, partial η^2 = 0.68, 95% CI [0.62, 0.72]), and prompt ($F_{(1, 293)}$ = 2501.89, p = 1.57×10^{-145}, partial η^2 = 0.90, 95% CI [0.88, 0.91]), with a significant model $\times$ prompt interaction ($F_{(2, 293)}$ = 14.96, p = 6.53×10^{-7}, partial η^2 = 0.09, 95% CI [0.04, 0.16]). Whilst every model earned signifi-

cantly higher payoffs under SCoT (Fig. 2a), the magnitude of this benefit varied, being smallest for DeepSeek (Hedges' g = —5.29, 95% CI [—6.12, —4.45], $p = 1.95 \times 10^{-46}$), intermediate for Gemini (g = —5.77, [—6.67, —4.87], $p = 3.64 \times 10^{-49}$) and largest for GPT-4.1 (g = —6.31, [—7.28, —5.35], $p = 5.26 \times 10^{-46}$). GPT-5, tested under normal prompting alone, outperformed all other normally-prompted models ($F_{(3,\ 121.62)} = 821$, $p = 1.61 \times 10^{-80}$, $\omega^2 = 0.92$, 95% CI [0.90, 0.93]; Games–Howell all $p < 0.001$ except DeepSeek vs. GPT-4.1, $p = 0.17$). Therefore, whilst both model and SCoT prompting impacted task performance, where SCoT benefited every model, the size of that benefit was model-dependent.

We then examined whether the predictions for the opponent's choice made by each model under SCoT prompting differed from chance (50%) (Fig. 2b). One-sample t-tests revealed significant differences in predictive accuracy, with GPT-4.1-SCoT predicting the opponent's choice significantly above chance level ($t_{(49)} = 10.98$, $p = 8.33 \times 10^{-15}$, d = 1.55), Gemini-SCoT predicting significantly below chance ($t_{(49)} = -13.32$, $p = 6.56 \times 10^{-18}$, d = -1.88), while DeepSeek-SCoT did not differ significantly from chance ($t_{(49)} = 0.63$, p = 0.53, d = 0.09) (Fig. 2b). Furthermore, examining switch frequency - trials where the subject changed choices from the previous trial - revealed differences across groups (Fig. 2c). Across groups, the mean switch frequency also significantly correlated with the mean payoff (r = 0.735, 95% CI [0.063, 0.948], $t_{(6)} = 2.657$, p = 0.038, $R^2 = 0.540$) showing that strategic adaptation was crucial for performance against the opponent (Fig. 2d). Social-Chain-of-Thought therefore presents a simple, robust prompting method for improving performance among LLMs in the inspection game by enhancing predictive reasoning towards the actions of the opponent. However, this improvement was scaled to the model's performance under normal prompting.

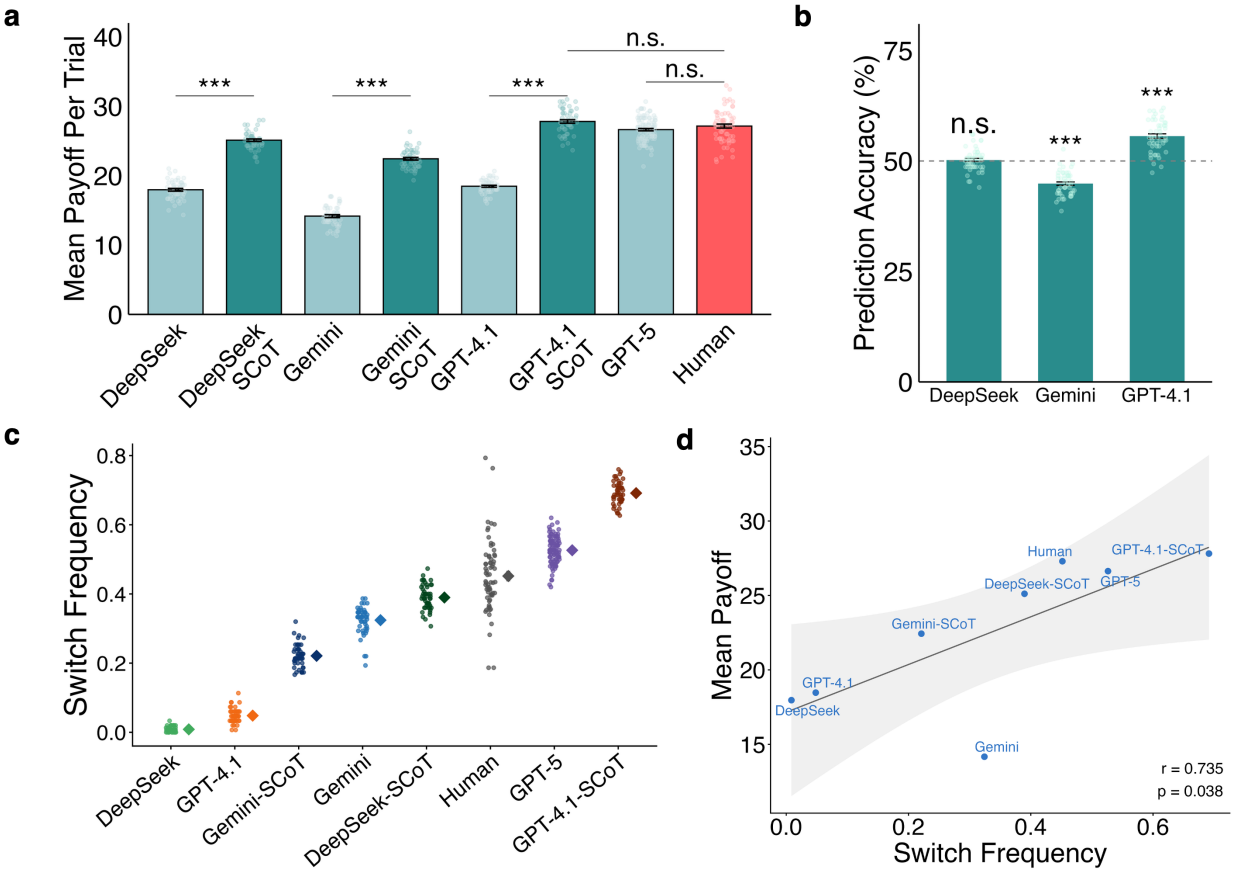


Figure 2: **Behavioural results for humans and LLMs in the inspection game. (a)** Social Chain-of-Thought (SCoT) robustly improved performance on the task across LLMs, as measured by the mean payoff achieved per trial. Human participants did not significantly differ in their performance on the inspection game with GPT-5 and GPT-4.1-SCoT agents. **(b)** Among SCoT-prompted LLMs, GPT-4.1-SCoT predicted the opponent's choice significantly above chance, Gemini-SCoT significantly below chance, whilst DeepSeek-SCoT did not deviate significantly from chance level in its predictions. **(c)** Switch frequencies varied across human and LLM groups, with the greatest between-subject spread observed among humans. **(d)** Mean payoff and switch frequency are positively correlated across LLMs and humans. Asterisks represent significant differences (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

Having established a behavioural difference between SCoT and normally-prompted LLM groups, we next tested whether these behavioural changes manifested computationally through altered reasoning mechanisms following the established computational framework of this task [80]. To do so, we fit a computational model ('influence play') that formalizes both how players react to their opponent's past choices (first-order beliefs) and how they anticipate the influence of their own choice on the opponent's behaviour (i.e., second-order beliefs). We also fit two simpler complementary models: a reinforcement learning (RL; zero-order beliefs) model that learns action values from payoffs without representing the opponent, and a model representing first-order mentalization ('fictitious play'; first-order beliefs). We assessed the models both through their predictive accuracy, by assessing which model best predicts observed choices, and

their generative accuracy, by determining whether the winning model reproduces the observed behaviour.

To determine predictive accuracy, we fitted each model to the behavioural data of participants within each group under the hierarchical Bayesian framework [79, 118], in which each participant's parameters were drawn from a group-level distribution. We compared the three cognitive models alongside a mixed-equilibrium (ME) strategy [106] under which choices feature no trial-by-trial learning, using the leave-one-out information criterion (LOOIC) where a lower score indicates better fit (see Methods).

Model comparison (Fig. 3a) indicated that human participants employed second-order mentalization, with the influence play model providing the best fit to the data (LOOIC = 12236; ΔELPD vs. mixed-equilibrium reference = +60.0, SE = 25.6). Among LLMs, SCoT prompting consistently shifted the winning model toward more sophisticated opponent-tracking. Specifically, DeepSeek shifted from reinforcement learning (LOOIC = 708; ΔELPD = +99.5, SE = 25.1) to fictitious play (LOOIC = 5387; ΔELPD = +2306.8, SE = 47.7), while GPT-4.1 shifted from reinforcement learning (LOOIC = 2557; ΔELPD = +147.7, SE = 29.0) to influence play (LOOIC = 8829; ΔELPD = +721.1, SE = 34.1). For Gemini, no learning model outperformed the mixed-equilibrium reference under normal prompting, whereas SCoT prompting produced a clear influence play winner (LOOIC = 5589; ΔELPD = +1269.5, SE = 51.2). Finally, GPT-5 was best fit by the influence play model (LOOIC = 16782; ΔELPD = +1152.6, SE = 39.8), demonstrating an ability for sophisticated reasoning without being explicitly cued to do so (see Appendix for full model fits).

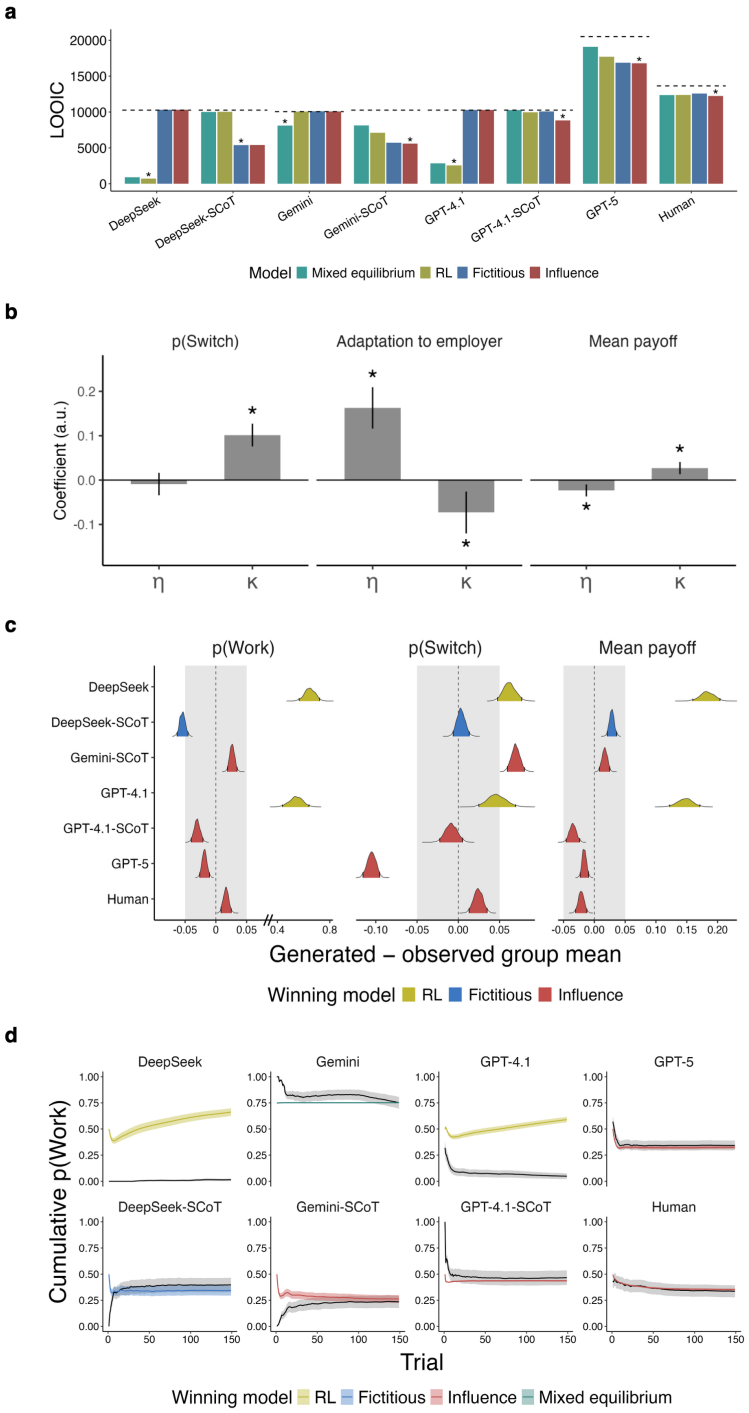
We next tested the direct mapping of the influence-play model's parameters with behaviour by regressing model-free behavioural measures on the fitted first-order learning rate (η) and second-order weight (κ) within each group (see Table S3 for the complete per-group coefficients). Following previous results among human participants [106], we anticipated three relationships if the parameters map onto behaviour as intended. Namely, a higher second-order weight (κ) should produce more switching, a higher first-order learning rate (η) should produce stronger tracking of the employer's most recent action, and a higher κ should produce a higher mean payoff. This pattern was indeed realised in the human group (Fig. 3b), where all three predicted relationships held and were statistically significant. The parameter representing second-order beliefs (κ) was strongly associated with switching (b = 0.101, $t_{(64)}$ = 7.76, 95% CI [0.075, 0.128], p = 8.5×10^{-11}) and task performance (b = 0.027, $t_{(64)}$ = 3.86, 95% CI [0.013, 0.041], p = 2.7×10^{-4}), whilst the first-order learning rate (η) was associated with adaptation to the employer's last move (b = 0.163, $t_{(64)}$ = 6.85, 95% CI [0.115, 0.210], p = 3.4×10^{-9}). However, across the LLM groups, this mapping was weaker and only partial (see Table S3). For LLM groups, comprising agents running the same model under a fixed prompt, the between-subject variance was minimal, reflecting homogeneity among parameter estimates. Therefore, whilst the switching and opponent-tracking relationships often reached significance in individual LLM groups, this was a likely consequence of less variation in the parameter estimates. Conversely, the human group represented a heterogeneous sample which carried genuine between-subject variance compared to LLM groups. Ultimately these results remain interpretable only for human participants, and reflect challenges concerning the homogeneity of parameter estimates among LLM agents.

Finally, we assessed the generative fit of each winning model by simulating choices from its own generated history and comparing the behaviour to several choice patterns from the observed data: the probability of working, switch rate and mean payoff (Table 2; see Methods for full details). To do so, for each model, we took the posterior-predictive distribution of the discrepancy between the simulated and observed statistic, testing for the percentage of the resulting distribution which fell within ± 0.05 of the shared [0,1] range using a Region of Practical Equivalence (ROPE). Following recommended guidelines [119, 120], we accepted a ROPE statistic when more than 97.5% of the distribution fell within the range.

Building on earlier results, the influence play model also demonstrated strong generative accuracy within human subjects, reproducing all three statistics within the ROPE (Fig. 3c). On-the-other-hand, the LLM groups differed along prompting condition. Where SCoT prompting was applied, the winning model - influence play for Gemini and GPT-4.1, and fictitious play for DeepSeek - mostly produced simulated behaviour within the ROPE. Under normal prompting, by contrast, the groups whose selected model was reinforcement learning (DeepSeek and GPT-4.1) generally failed to accurately reproduce their observed behaviour. Plotting the cumulative probability of work, p(Work), across trials (Fig. 3d) revealed that the RL model was found to slowly drift the agent's probability of working upwards, while the observed rate among these two groups remained near zero, signalling its poor generative fit and underlying choice homogeneity. Finally, GPT-5, an influence player without any Chain-of-Thought cue, demonstrated a mixed accuracy, able to generatively recover the level and payoff but not the switch rate.

Table 2: **Generative fit of the winning models for the inspection game.** *obs* is the observed statistic and *gen* is the generated posterior-predictive mean with its 95% predictive interval in brackets. Bold marks statistics accepted under the ROPE equivalence test where more than 97.5% of the discrepancy distribution fell within ± 0.05. Gemini under normal prompting had no learning model that beat the mixed-equilibrium reference and was therefore not included.

Group	Winning model	p(Work) obs	p(Work) gen	p(Switch) obs	p(Switch) gen	Payoff obs	Payoff gen
Human	Influence play	**0.336**	**0.352** [.34, .36]	**0.455**	**0.479** [.47, .49]	**0.545**	**0.524** [.51, .53]
DeepSeek	RL	0.014	0.662 [.58, .74]	0.004	0.066 [.05, .08]	0.359	0.542 [.52, .56]
DeepSeek-SCoT	Fictitious play	0.396	0.342 [.33, .35]	**0.393**	**0.396** [.39, .41]	**0.502**	**0.530** [.52, .54]
Gemini-SCoT	Influence play	**0.234**	**0.261** [.25, .27]	0.223	0.292 [.28, .30]	**0.448**	**0.465** [.46, .47]
GPT-4.1	RL	0.048	0.591 [.49, .69]	0.044	0.091 [.07, .11]	0.371	0.518 [.49, .54]
GPT-4.1-SCoT	Influence play	**0.466**	**0.436** [.43, .45]	**0.696**	**0.687** [.67, .70]	**0.556**	**0.521** [.51, .53]
GPT-5	Influence play	**0.343**	**0.324** [.32, .33]	0.531	0.426 [.42, .44]	**0.533**	**0.516** [.51, .52]



227

Figure 3: **Predictive and generative fit for the inspection game. (a)** Model comparison across groups using the leave-one-out information criterion (LOOIC); winning models are highlighted. The dashed line above each group indicates the LOOIC from chance-level responding. **(b)** Standardised coefficients from the within-group regression of each winning model's parameters on choice behaviour from human participants. **(c)** Posterior-predictive discrepancy (generated minus observed group mean) for each group's winning model, with the full ROPE-shaded discrepancy distribution. **(d)** The rolling probability of working for the generated against the observed data.

Taken together, human strategic learning in the inspection game was robustly well-explained by second-order mentalizing, where the influence play model displayed strong predictive and generative accuracy, and its parameters mapped strongly onto behaviour. Amongst LLM groups, Social Chain-of-Thought prompting robustly moved LLMs towards strategic mentalization from model-free reinforcement learning or non-learning. These models were generatively adequate for strategic mentalization and inadequate for RL, with the latter reflecting rigid patterns of choice behaviour.

Assessing adaptive mentalization using rock-paper-scissors

To assess *adaptive* mentalization, we firstly examined whether the mean payoff differed across humans (n = 172 retained for behavioural analyses) and LLMs (all n > 187) in the RPS game, when playing against algorithms programmed to reason at different k levels (k = 0, 1, or 2) (Fig. 4a). Separating groups by prompting strategy and reasoning level of the opponent, we ran a linear mixed-effects model with random intercepts for participants, given the unbalanced design where humans played against all three bot levels while LLM participants played against a single bot level. LLMs consisted of three groups: DeepSeek, DeepSeek-SCoT, and GPT-5. Two models included in the previous task (Gemini-Flash-2.0 and GPT-4.1) were not included due to technical issues and monetary constraints respectively.

Doing so revealed significant main effects of group ($F_{(3, 1363.2)} = 768.3$, $p = 2.31\times10^{-292}$, $\eta^2_p = 0.63$) and bot level ($F_{(2, 2096.8)} = 439.8$, $p = 3.25\times10^{-160}$, $\eta^2_p = 0.30$), as well as a significant interaction ($F_{(6, 1908.3)} = 306.3$, $p = 3.5\times10^{-275}$, $\eta^2_p = 0.49$). Overall, GPT-5 performed best, earning significantly higher payoffs than humans ($t_{(661)} = 25.5$, $p < 1\times10^{-10}$, d = 1.58, 95% CI [1.45, 1.71]), and both DeepSeek variants (vs. DeepSeek: $t_{(2228)} = 40.7$, $p < 1\times10^{-10}$, d = 2.44, 95% CI [2.32, 2.56]; vs. DeepSeek-SCoT: $t_{(2228)} = 42.2$, $p < 1\times10^{-10}$, d = 2.53, 95% CI [2.40, 2.65]). GPT-5 also outperformed all other groups at each bot level (all $p < 0.001$). Conversely, both DeepSeek ($t_{(661)} = 14.0$, $p < 1\times10^{-10}$, d = 0.86, 95% CI [0.74, 0.99]) and DeepSeek-SCoT ($t_{(659)} = 15.4$, $p < 1\times10^{-10}$, d = 0.95, 95% CI [0.83, 1.07]) earned significantly less than humans. The two DeepSeek groups also did not differ overall ($t_{(2228)} = 1.44$, $p = 0.477$, d = 0.09, 95% CI [-0.03, 0.20]).

Performance also strongly differed across groups in response to increasing opponent sophistication. Whilst humans showed a modest performance decline with increasing sophistication (β = -0.032, SE = 0.007, $t_{(686)} = -4.45$, $p = 9.96\times10^{-6}$, d = -0.24), DeepSeek performed significantly worse (β = -0.233, SE = 0.007, $t_{(2224)} = -33.22$, $p = 7.73\times10^{-197}$, d = -1.73), whilst DeepSeek-SCoT also performed significantly worse but less severely (β = -0.116, SE = 0.007, $t_{(2224)} = -16.53$, $p = 5.29\times10^{-58}$, d = -0.86). In contrast, GPT-5 was the only group to show improved performance with increasing bot sophistication (β = 0.057, SE = 0.007, $t_{(2224)} = 8.11$, $p = 8.4\times10^{-16}$, d = 0.42), a pattern significantly different from all other groups (all $p < 0.001$). SCoT prompting did not significantly change DeepSeek's overall payoff ($p = 0.477$), but attenuated its performance decline against more sophisticated opponents (difference in slopes: $t_{(2224)} = 11.82$, $p < 0.001$, d = 0.87).

To uncover latent unobserved mechanisms of adaptive mentalization, we fit computational models representing specific decision-making strategies. This included the 'Cognitive Hierarchy Assessment' (CHASE) model, previously developed and validated among human participants to infer moment-by-moment adaptive changes in mentalization strategy from observed choice behaviour

[107]. The CHASE model dynamically tracks a player's changing belief about the level of cognitive sophistication of the opponent, determining the appropriate mentalization strategy on each trial. Additional computational models representing less sophisticated and fixed strategies were also fitted separately across all groups (see Methods). Conversely to humans (n = 184 retained for modeling) who played against all three opponent levels in sequence, LLMs played separate games of a single block against each k-level opponent. To match the existing protocol and to facilitate efficient parameter estimation, individual LLM instances were combined to create LLM participants consisting 3 blocks of 40 trials each, representing each programmed bot-level k, where $k \in \{0, 1, 2\}$. This approach is validated theoretically by the models - which do not model memory effects across blocks - and practically by a permutation analysis across individual LLM instances, revealing no significant differences in parameter estimation (see Supplementary Information section 2). To determine the winning model, we applied Random effects Bayesian model selection (RFX-BMS) [121] which revealed different winning models across groups. Specifically, for DeepSeek, CHASE was the clear winner (PXP = 1.00), substantially outperforming all alternatives. In contrast, DeepSeek with the SCoT prompting condition was best characterized by Fictitious play (PXP = 0.99), a model where agents estimate the probabilities with which the other player chooses their actions and then best respond to them [80]. On the other hand, GPT-5 showed a unanimous preference for the Theory-of-Mind k (ToMk) model [122] (PXP = 1.00), which assumes that individuals reason about what their opponent is likely to do and may consider multiple layers of strategic thinking. Finally, human participants were best described by CHASE (PXP = 1.00), consistent with previous results. Good-to-excellent parameter recovery further validated the winning model for each group (Supplementary Figure 4). Notably, CHASE provided the absolute best fit across all datasets, as measured by the negative log-likelihood, but was penalised for model complexity (See Appendix for exact model fits).

To assess the degree to which LLMs and human players were able to correctly adapt their behaviour to the perceived recursive level of the opponent, we examined belief distributions from the CHASE model, representing the trial-by-trial estimate of the probability that the opponent is operating at each reasoning level, for each subject individually (see Methods). Although ToMk provided the best overall fit for GPT-5, CHASE fitted choices equally well in raw log-likelihood (paired $t_{(187)} = 0.70$, $p = 0.48$). We therefore used CHASE as a common space for interpreting trial-level reasoning, a mechanism not possible among other models. When doing so, we specifically restricted trials to only those where beliefs exceeded a threshold of 0.5, reflecting a dominant belief, following similar analyses performed previously [107] (Fig. 4b). Of note, doing so removed almost half of trials (45.4%) where no dominant belief was expressed. DeepSeek-SCoT agents were determined to not form adaptive beliefs when fitted with the CHASE model, with participants almost exclusively estimated to operate at $\kappa<2$, and were therefore not suitable for analysis.

To formally test whether accuracy differed jointly by model and opponent sophistication, we ran a mixed-design repeated-measures ANOVA on each subject's percentage of correctly-inferred trials, where subjects missing data for any opponent level were excluded (DeepSeek n=97, GPT-5 n=150, Human n=126). As Mauchly's test indicated a violation of sphericity, Greenhouse-Geisser-corrected p-values were calculated for within-subject effects. Doing so revealed significant main

effects of model ($F_{(2,370)}$ = 224.91, p = 1.20×10^{-64}) and opponent level ($F_{(2,740)}$ = 251.68, p = 2.10×10^{-77}, ε = 0.917), and a significant model $\times$ opponent-level interaction ($F_{(4,740)}$ = 64.55, p = 3.51×10^{-43}), confirming that the effect of opponent sophistication on adaptive mentalization differed by group.

Next, we sought to determine whether agents' were correctly able to adapt to the opponent, using a repeated-measures Friedman test across the four k-levels for each group and opponent-level combination. Where this test was significant, we subsequently ran one-sided Wilcoxon signed-rank tests, with an a priori hypothesis for a higher percentage of trials for the k-level one higher than the opponent, indicating adaptive mentalization. Each class of tests were Holm-corrected (α=0.05). Friedman tests firstly confirmed non-uniform belief distributions across the four candidate levels (all $p_{holm} < 0.001$). Signed-rank tests subsequently revealed that GPT-5 agents adapted appropriately across all opponent levels, with trials for k+1 significantly higher than every alternative level (all $p_{holm} \leq 8.3\times10^{-21}$). For DeepSeek, this held against the simplest opponent (k=0; all three comparisons $p_{holm} < 5.3\times10^{-18}$) and partially against the intermediate opponent (k=1; correct level k=2 only exceeded k=0 and k=3, both $p_{holm} = 1.81\times10^{-9}$; but not k=1, $p_{holm} \approx 1$), but not the most sophisticated opponent (k=2; all three p_{holm} = 1.0). Humans showed the same pattern with strong adaptation towards k=0 (all three $p_{holm} \leq 1.1\times10^{-3}$), partial adaptation against k=1 (correct level k=2 exceeded k=3, $p_{holm} = 1.74\times10^{-14}$, but not k=0 or k=1, both p_{holm} = 0.339), but not against the most sophisticated opponent (k=2; all $p_{holm} \approx 1.0$).

To relate adaptation to performance, we then correlated each subject's proportion of trials where they adapted correctly with their mean score in the task for each group. We specifically restricted these analyses to agents deemed to adapt to the opponent (i.e., an estimated κ>1; DeepSeek n=139, GPT-5 n=182, Human n=130). Following the prior analyses, we similarly defined 'correct' trials where a definitive belief (> 0.5) was held at the adaptive level (Fig. 4c). We also performed the same tests using the single highest belief, as excluded trials were incorrect by definition in the former method, resulting in many agents with 0% accuracy. We used Spearman-rank correlations with Holm-correction across the 6 tests. Although none of the correlations survived correction, the trend was generally positive across groups and analyses. Restricting trials to dominant beliefs revealed a positive trend for DeepSeek (Spearman ρ = 0.180, p_{adj} = 0.2032), humans (Spearman ρ = 0.163, p_{adj} = 0.2567), and GPT-5 (Spearman ρ = 0.120, p_{adj} = 0.3171). Similarly, using the winner-take-all approach revealed similar positive trends for DeepSeek (ρ = 0.119, p_{adj} = 0.3257) and humans (ρ = 0.175, p_{adj} = 0.2331) but a slight negative trend for GPT-5 (ρ = $-$0.073, p_{adj} = 0.3279). Altogether, these analyses highlight distinct patterns of adaptive mentalization across groups, which positively corresponds with improved task performance, albeit weakly.

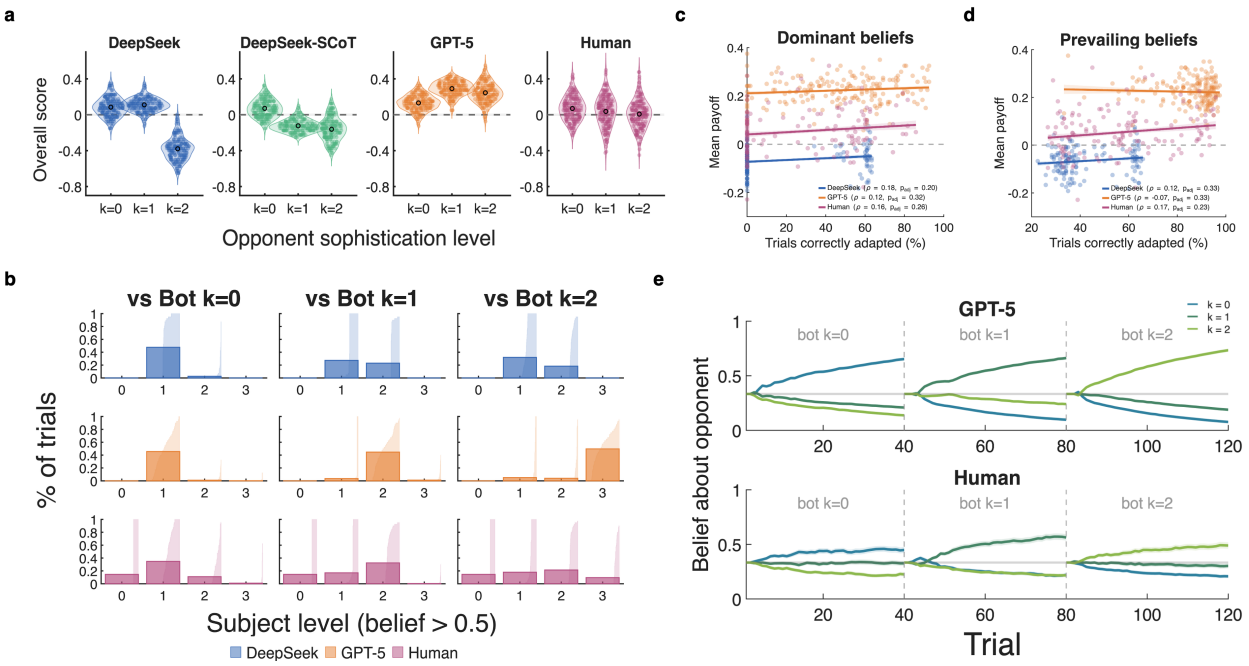


Figure 4: **Signatures of adaptive mentalization in humans and large language models. (a)** Behavioural performance on the task, determined by the mean score, split by group, revealed differences between humans and LLMs. Whilst both DeepSeek groups and humans perform worse with increasing bot sophistication, GPT-5 conversely improves. **(b)** The inferred level of mentalization among agents plotted as a proportion of trials per subject in each group. Reflecting the behavioural differences in performance, DeepSeek and human participants adapt well to low-adaptive reasoners, but struggle with opponents capable of higher-order mentalization. Conversely, GPT-5 participants were able to adapt well to all reasoning levels. For visualization, prevailing beliefs (>0.5) were taken directly from the CHASE model and adjusted to reflect the k-level of the agent by adjusting one level higher. DeepSeek-SCoT participants were excluded from the plot, being almost exclusively estimated to have a value of k<2 when fit with the CHASE model. **(c)** Correlations between the percentage of trials where agents correctly adapted to the inferred mentalising level of the opponent, defined using a dominant belief. **(d)** Correlations between the percentage of trials where agents correctly adapted to the inferred mentalising level of the opponent, defined using a prevailing belief. **(e)** CHASE-inferred beliefs about opponent sophistication across play. For each agent group (top: GPT-5, n = 174; bottom: humans, n = 80), lines show the model's mean posterior belief that the opponent is reasoning at each level across the 120 trials (κ = 3 participants only). A grey horizontal line denotes chance level (1/3); shaded bands are $\pm$1 SEM.

Finally, we analysed whether agents could correctly infer their opponent's sophistication by examining the CHASE model's posterior belief in the true opponent level on the final trial of each block. As less than 2% of DeepSeek agents met this criteria, we only conducted these analyses where κ=3 was the estimated level, across GPT-5 (n = 174) and human participants (n = 80) (Fig. 4e), where human participants missing a final belief for k=0 (two missing) and k=1 (one missing) opponents and were excluded. One sample, two-tailed t-tests against chance (1/3) revealed that both GPT-5 (k=0: $t_{(173)} = 22.30$, $p = 9.18\times10^{-53}$, $d = 1.69$; k=1: $t_{(173)} = 23.38$, $p = 1.99\times10^{-55}$, $d = 1.77$; k=2: $t_{(173)} = 29.72$, $p = 7.3\times10^{-70}$, $d = 2.25$) and humans (k=0: $t_{(77)} = 4.23$, $p = 6.45\times10^{-5}$, $d = 0.48$; k=1: $t_{(78)} = 7.86$, $p = 1.74\times10^{-11}$, $d = 0.88$; k=2: $t_{(79)} = 5.35$, $p = 8.52\times10^{-7}$, $d = 0.60$) held

above-chance correct beliefs at every sophistication level. A linear mixed model (belief ~ agent $\times$ level + (1 | subject)) revealed a significant agent $\times$ opponent interaction ($F_{(2, 505.0)} = 19.79$, $p = 5.33\times10^{-9}$, $\eta^2_p = 0.07$) with main effects for agent ($F_{(1, 418.7)} = 48.38$, $p = 1.36\times10^{-11}$, $\eta^2_p = 0.10$) and level ($F_{(2, 504.2)} = 22.30$, $p = 5.22\times10^{-10}$, $\eta^2_p = 0.08$). This highlights the trend where GPT-5's final belief also increased with opponent depth, whereas humans showed a weaker, inverted-U pattern peaking against the k=1 opponent. Comparing across both groups, GPT-5's beliefs were significantly more accurate than humans' at every level (Welch's t; k=0: $t_{(121.2)} = 6.63$, $p = 1\times10^{-9}$, $d = 0.99$; k=1: $t_{(115.5)} = 3.05$, $p = 0.003$, $d = 0.47$; k=2: $t_{(113.9)} = 7.60$, $p = 9.14\times10^{-12}$, $d = 1.18$), with the advantage widest against the most sophisticated (k=2) opponent. All comparisons survived Bonferroni correction, with 6 comparisons ($\alpha = 0.0083$) for tests against chance, and 3 comparisons ($\alpha = 0.0167$) for group comparisons. Therefore, while both humans and GPT-5 tracked opponent sophistication above chance, GPT-5 did so more accurately and, unlike humans, increasingly so as opponents grew more sophisticated.

Discussion

A growing interest concerns whether artificial intelligence systems understand the thoughts and beliefs of others, and use this information to guide their own actions. Previous studies have assessed theory-of-mind in LLMs by measuring choice accuracy in response to story-based prompts, obscuring the latent mechanisms underlying observed behaviour. Here we employ two behavioural tasks previously validated in human participants, the inspection game and rock-paper-scissors (RPS). These tasks provide a normative assessment of machine intelligence [65], and allow the examination of behaviour at specific depths of mentalizing [69].

Across two strategic interaction games, humans demonstrated both recursive and adaptive mentalization, replicating previous results and demonstrating the robustness of the tasks used [80, 106, 107]. On the other hand, LLMs demonstrated varying mentalization strategies across the two strategic games, with significant differences observed across model providers. Furthermore, LLMs prompted using Social Chain-of-Thought (SCoT) demonstrated robust improvements in performance, reflecting more sophisticated reasoning. Yet, groups differed in their generative accuracy, limiting interpretations towards latent processes of mentalization. In the RPS game, GPT-5 agents demonstrated efficient adaptive mentalization, whilst DeepSeek agents appropriately adapted to bots playing at the lowest level but struggled to adapt to complex reasoners. Compared to prior studies assessing mentalization using story-based one-shot tasks or binary choice paradigms, our approach leverages trial-level computational modeling to determine the specific latent mechanisms underlying mentalization. These models can be subject to rigorous validation by determining predictive and generative accuracy, allowing for the direct mapping of mechanism to behavior.

Cognitive computational modeling, as implemented in our study, offers a normative tool for quantifying latent (social) cognitive processes and tracking how they change in response to targeted interventions [79, 123–126]. Previous work has examined differences in mentalization across several species of non-human primates [14]. We use these principles to determine whether

similar behavioural outcomes across LLMs and humans reflect shared latent computations of mentalization. Our results demonstrate that this framework is also sufficiently sensitive to detect meaningful differences in mentalizing behaviour across model providers, and captures the effect of strategic prompting, a manipulation comparable to language-based psychotherapy in humans [127]. Prompting strategies designed to improve ToM in LLMs should ultimately be evaluated against process-level measures, as relying on surface-level patterns may obscure latent changes in social cognition [65, 70]. Yet, whilst providing a deeper level of understanding than assessing choices, computational modeling still obscures the mechanistic components of LLM which must instead be inferred from their observed behaviour [110]. This is further mired by the inferential properties of LLMs, which as pseudo-statistical machines utilize traditional stochastic processes and human-like abductive reasoning to reach conclusions [128, 129]. Despite the emergence of frameworks and best practices for determining computational equivalence between humans and LLMs [23, 66, 72, 110, 130, 131], doing so currently remains a challenging task. This is evidenced by the opaque definition of theory-of-mind in machines, which requires an interdisciplinary approach, one incorporating theories from cognitive psychology, neuroscience and artificial intelligence [132]. Looking ahead, a comprehensive assessment of mentalization in LLMs will likely require a multi-modal approach involving multiple levels of explanation [23, 133, 134].

The observed differences in task performance may be partly attributed to the model size, where GPT models may constitute significantly more parameters and are trained upon larger quantities of data than the DeepSeek and Gemini models used in the tasks. Large language models are based on the transformer architecture of neural networks, pretrained on a large text-based corpus. Subsequently, the neural scaling law demonstrates that their performance generally improves as the model size and training data increase [135]. Larger transformer models also demonstrate emerging properties absent in smaller models [136] including in-context learning [137] and multi-step reasoning [138]. Subsequently, model size in LLMs has been shown to robustly correlate with performance across benchmarks assessing economic and rational choice [112, 139–142]. Larger models often may possess longer context histories, capable of accurately retrieving and processing more information when making choices. This is supported by comprehension checks for the tasks used in the current study which report differences with the retrieval accuracy of task-related information across model sizes (see Supplementary Information section 3).

Modeling analyses from the RPS task revealed DeepSeek was unable to appropriately calibrate its recursive level of thinking to the opponent, an ability not significantly improved by SCoT prompting. On the other hand, GPT-5 agents were able to accurately form beliefs across all recursive levels. These results follow recent evidence suggesting that smaller LLMs generally are able to accurately infer others' perceptions [143], but struggle to appropriately form and update their beliefs about other agents [144]. Furthermore, whilst LLMs are generally inconsistent in how they update their beliefs [145], larger models learn to update their beliefs about propositions more consistently [144]. Unlike human-based belief updating which is bounded by data and computational constraints [146], LLMs are able to burden more computationally intensive processes of reasoning over extended periods. Taken together, GPT-5 agents may be able to more accurately predict the actions of complex reasoners by using stored choice histories of the opponent together

with deliberative reasoning. Strategic prompting is widely used for evoking complex, human-like cognitive operations among LLMs [47, 147–150] including theory-of-mind (ToM) [151–155], where performance is improved by appropriate prompting which elevate ToM reasoning [153]. However, prompting-induced improvements in mentalization are context-dependent [153] and task performance can more generally be improved by providing strategy-specific information within prompts [67, 156]. This altogether suggests that the SCoT prompt may be inappropriately vague for smaller models given the specific task demands of the RPS game. Prompts including additional task or strategy-specific information could potentially elicit adaptive mentalization in models less capable of sophisticated reasoning.

We note that a mechanistic explanation of mentalization is ultimately not possible due to the closed-sourced nature of the models used. However, specific activity patterns of hidden embeddings across several open-sourced LLMs are found to correspond with their ability to represent another's perspective, where the percentage of embeddings displaying significant selectivity positively correlates with model size [157]. Furthermore, recent mechanistic work has suggested that social reasoning - including ToM - is specifically encoded in extremely sparse, localised parameter patterns tied to contextual processing [158]. An open question therefore remains whether larger closed-source models also demonstrate specific patterns of activity corresponding to mentalization depth, and whether parameter-level signatures of mentalization exist within them.

Our study has theoretical and practical implications across the fields of artificial intelligence and machine psychology. As LLMs are increasingly employed in different sectors, their ability to accurately interpret and appropriately respond to human mental states becomes critical [159]. Large language models as chatbots are widely used as therapeutic tools [160–164], capable of building artificial relationships [165–167] and providing emotional support [168–171], roles necessitating accurate mentalization of the user's preferences, feelings and potential actions. Yet, LLMs often provide inappropriate or harmful responses, a consequence of not accurately interpreting subtle text-based cues [172]. Indeed, AI systems are often criticized for not being able to demonstrate affective and motivational empathy, mentalization-derived components essential to human-based therapy [173–175]. The capacity for LLMs to mentalize also influences user perception and interactions. Humans frequently anthropomorphise AI tools, attributing mental states based on their own expectations and the agents' observed behaviour [176–178], which in turn predicts trust and social closeness towards these systems [168, 179]. Experimentally portraying LLMs intentionally as social entities has also been shown to increase attributions of warmth, empathy, and competence among users [180], traits which can be enhanced by inducing a theory-of-mind in LLMs [181]. Finally, instilling the capacity for complex 'human-like' cognition among LLMs can also augment and complement human thinking and decision-making [182]. Human intelligence strongly differs from machine intelligence, where humans build causal models of the world, and efficiently apply relevant prior learned information and knowledge to new tasks and situations [26]. Conversely, generative AI systems such as LLMs are statistical inference machines based on pattern recognition. Building machines that learn and think more strongly like humans - including a capacity for mentalization - can lead to more knowledgeable, reliable and trustworthy systems [115].

We acknowledge several practical limitations of our experiments. The analytical choices used

for LLM configuration, including model parameter settings and prompt configurations, significantly impact their accuracy for replicating human data [183–186]. Therefore, our response patterns may not be observed using different model providers, parameters or prompts. Additionally, the fixed choice history of 10 trials may obscure the effect of different context windows on choice behaviour [43]. We also specifically focus on a specific group of model providers and do not include any open-source models in our analyses, a recommended practice for understanding human cognition with LLMs [187]. Acknowledging these points, we ultimately follow existing guidelines for evaluating the cognitive abilities of LLMs [110, 133], aiming to test specific hypotheses concerning the computational form of the target process - mentalization - and the effect of a single additional prompting scheme previously validated. We also follow the recommended practice of 'machine validation' [188], where statistical patterns and strategies uniquely exploited by machines are accounted for. In our primary analyses, we generated the experiment text without the use of LLMs and place models in the role of an active participant akin to human participants. Yet, LLMs may nevertheless exploit lexical patterns present in the experimental prompt when making choices [65]. Subsequently, we attempt to control for various confounds including choice order and composition in our primary analyses and formally investigate their effect in our supplementary analyses, reporting varying effects of these confounds by model and task (see Supplementary Information section 3). Finally, our experimental design does not entirely dissociate a machine's knowledge (i.e., 'competence') from it's application (i.e., 'performance') [189]. Whilst we aimed to similarly limit both humans and machines by providing identical task instructions and ensuring a thorough understanding of the task rules and objectives, the outcome history was only provided to LLMs and not humans. We acknowledge that a similar trial history made available to humans, or conversely, burdening machines with the equivalent human limitations, would strengthen our claims.

Researchers are increasingly turning to methods rooted in psychology and cognitive science to understand the behaviour of large language models [190]. Yet, how these systems actively represent others remains unclear. Our findings highlight key differences in the ability to recursively mentalize across LLMs, evident through observed behaviour and the underlying computations. Artificial intelligence (AI) continues to develop at an alarming rate, with some forecasting that digital computation may eventually supersede human intelligence [191]. Simultaneously, building AI systems with advanced cognition stands to improve their application across a range of social domains, yet increases risks toward malicious use [192, 193]. Within this complex, rapidly evolving space, understanding the cognition of these systems is necessary to appropriately scale the development of AI technologies.

Methods

Experimental procedure

Humans

Inspection game

The inspection game is a repeated non-cooperative game where players may implement strategies corresponding with different levels of recursive thinking [80, 106]. In the game, two players assume the roles of an "employee" and an "employer" and make one of two choices simultaneously. On each trial, the employee chooses whether to "work" or "shirk," and the employer chooses whether to "inspect" or "not inspect" the employee. In our version (Figure 1 a), the participant always played the role of the employee.

Each trial began with a fixation cross for 1 second, after which the options of "work" or "shirk" were presented. Participants had up to 4 seconds to indicate their choice by pressing the respective arrow key, and the chosen option was highlighted with a red square for up to 2 seconds. The choice of their opponent, whether to "inspect" or "not inspect", was then presented on the screen for 3 seconds, alongside the outcome of that trial. The task consisted of a single block of 150 trials. Points from the task were added and converted to a bonus payment in GBP at the end of the session.

Participants were told they were playing against a real opponent, but were actually playing against a computer algorithm based on the influence learning model fitted to real human data from previous data [106]. The artificial opponent's responses were generated by the fitted influence learning model, which would both (1) track player choices and respond optimally to their frequency and (2) use its own past choices to form player expectations, before (3) introducing randomness to the choices based on these two computations (used model parameters: the first-order belief updating η = 0.3; second-order belief updating κ = 0.1; inverse temperature β = 1.5). We used MATLAB 2012a (MathWorks) and the Cogent 2000 v125 toolbox (www.vislab.ucl.ac.uk/Cogent/) to present the task.

Rock-paper-scissors

To assess adaptive mentalization, we employed a zero-sum rock-paper-scissors (RPS) game. In contrast to the inspection game, which is played against a fixed algorithm across all trials, participants in the RPS game played against an algorithm programmed to make choices in accordance with a specific reasoning level *k*, where $k \in \{0, 1, 2\}$. Subsequently, this experimental design is suitable for assessing whether humans and LLMs are capable of adapting their behaviour to the *k*-level of the opponent [107, 194].

Participants were provided instructions that they will be playing a simple repeated Rock-Paper-Scissors game with three different opponents (Figure 1 b). They were instructed that on each trial, they must make a choice between the three options, Rock (using the 'Right' arrow key), Paper (using the 'Up' arrow key) and Scissors (using the 'Left' arrow key), and that both players will be making the choices at the same time. Participants had up to 3 seconds to make their choice, after

which they received feedback on the opponent's choice as well as the reward outcome (+1 for win, -1 for loss, 0 for tie). Participants were told that they would play 40 trials against a specific opponent, after which they would be matched to a new opponent. In reality, participants played against a computer algorithm designed to make choices corresponding to different levels of mentalization, k (where $k \in 0, 1, 2$). Opponents were randomised and counterbalanced across participants.

Large language models

Inspection game

Large language models completed a translated version of the inspection game presented to humans. To keep the information as similar to that presented to the human sample, roles, choice options and the payoff matrix were left unchanged. However, the reward metric was changed from 'points' to 'cents', incentivizing the LLMs to perform well [195, 196]. Large language models also completed 150 trials playing the role of the employee against the same algorithmic opponent as the employer. In all prompting scenarios (see Supplementary Information section 1), models were presented with the game rules, payoff, choice history, and choice question.

Rock-paper-scissors

Replicating the experimental procedure used in humans, large language models also completed 40 trials of a repeated game with the same algorithmic opponent programmed to make choices at a given recursive reasoning level. However, in contrast to human participants, who played against all 3 levels, each LLM played only a single block against a specific level. Whilst the payoff matrix was left unchanged, choice options were recoded to prevent potential choice bias with numbers (i.e., '1' selected more often than '2' or '3') or the conventional choices ('Rock', 'Paper', 'Scissors' / 'R', 'P', 'S') previously highlighted [116]. Instead, three single letters ('J', 'Q', 'Z') were selected as approximately equally uncommon letters in the English corpora. In all prompting scenarios, models were presented with the game rules, choice history, and choice question.

Both tasks were programmed using custom code integrating the oTree {otree==5.11.1} experimental platform [197] with the Python package botex {botex==0.2.0} [198] and LiteLLM {litellm==1.73.1} for LLM API calls. Whilst in previous studies the choice history has included all previous trials, a rolling choice history of 10 trials was chosen to represent an approximation of working memory among human players, whilst accounting for time and monetary costs with practically running the experiments. No time limits were enforced during both tasks.

Participants

Humans

Inspection game

Human participants were recruited through Oxford University databases, social media, email lists, and adverts in local newspapers. A priori, we excluded individuals who studied psychology, had a history of neurological or psychiatric disorders, or had abnormal vision. Participants received

£10 per hour and a bonus of up to £5 based on their task performance (including several different tasks). All participants provided written informed consent. The Oxford University Medical Sciences Inter Divisional Research Ethics Committee and National Health Service Ethics approved the study. For behavioural analyses, only those with no more than 5 missing trials were included, leaving 67 participants (25 males, mean age = 23.3 $\pm$4.21). All of these participants' data were also included for modeling.

Rock-paper-scissors

Participants were recruited from the general population on Prolific, restricted to UK residents and native English speakers as part of a larger study. Before completing the RPS task, they filled out demographic questionnaires and questions using verbal mentalization tasks.

Participants used their own devices to access the study on Qualtrics and Cognition and were paid £18 for an anticipated 120 minutes of work (median time: 117 minutes). The study was approved by the University of Birmingham STEM Ethics Committee (ERN 22-1192, ERN 09-719F, ERN 2311).

A total of 185 human participants completed the original experiment. For behavioural analyses, only those with less than 5 missing trials were included, leaving 172 participants (73 males, 97 females, 2 missing; mean age = 24.95 $\pm$3.57 years). For modeling analyses, a single participant was excluded for not having valid parameter estimates, leaving 184 human participants (76 males, 106 females, 2 missing; mean age = 24.91 $\pm$3.57 years) for all modeling analyses.

Large language models

The following LLMs were selected for participation in the respective tasks. Inspection game: Gemini, DeepSeek, GPT-4.1, GPT-5; RPS: DeepSeek, GPT-5. These models were selected based upon time and monetary costs when running the study, and from primary comprehension checks that assessed a wider set of candidate models (see Supplementary Information section 3). Given the commonly reported issues with model response formats and content [199], several steps were made to ensure reliability in model outputs. Firstly, API calls were made to model providers using LiteLLM, unifying model communication with a common interface. Under this framework, structured outputs in JSON format were strictly enforced across all models using the Instructor {instructor==v1.6.0} Python library where not natively supported. Secondly, strict Pydantic schemas defined the required response structure and data types. Finally, model responses were programmatically limited to those defined as response options within the prompt (i.e., 'WORK'/'SHIRK' or 'J'/'Q'/'Z').

Following a power analysis (see Power calculations), the sample size per LLM per experiment consisted of:

- Inspection game: GPT-4.1 (50 standard prompt, 50 SCoT), DeepSeek (50 standard prompt, 50 SCoT), Gemini (49 standard prompt, 50 SCoT), GPT-5 (100 Normal)
- RPS (Individual LLM instances): DeepSeek (566 standard prompt, 568 SCoT), GPT-5 (566 standard prompt)

In modeling analyses for the RPS task, individual runs were combined within each LLM group, with remainder runs discarded, resulting in the following sample sizes: DeepSeek (188 standard prompt, 189 SCoT), GPT-5 (188 Normal).

Prompting conditions

A previous study [40] demonstrated the effect of a simple prompting strategy - 'Social Chain-of-Thought' (SCoT) - for improving LLM performance in economic games by asking the model to form a prediction for the opponent's choice before making a decision. To test whether this ability also extends to the current study, two prompting conditions were tested. Firstly, in the standard prompt condition, models were simply asked to make a choice between the two options, with no explicit prompt for engaging in deliberate thought. In the SCoT condition, models were additionally prompted on each trial to predict the choice of the opponent before making their choice.

Parameters

A default *temperature* parameter of 0 was set for Gemini, DeepSeek, GPT-4.1 whilst a temperature of 1 was set for GPT-5 in accordance with the restricted parameter values provided by OpenAI. A setting of 'medium' was chosen for the *reasoning_effort* parameter with GPT-5, providing detailed reasoning without explicit instruction, and considering monetary restrictions for the study. Token lengths for model responses through the *max_tokens* parameter was not limited. LLM data was collected from the periods 04/08/2025 - 17/09/2025 and 22/09/2025 - 14/10/2025 for the inspection and RPS games respectively.

Power calculations

Inspection game

Comparisons between model and human performance by strategy

Comparisons between model $\times$ strategy combinations and human participants required seven pairwise tests. Using Welch's two-tailed t-tests with Bonferroni correction (α = 0.007) and assuming large effects (d = 0.8), power analyses indicated approximately 36 participants per LLM group to achieve 80% power. This was met with our sample of at least 49 agents per group, and an allocation ratio of 1.34 reflecting the 67 participants in the human group.

Within-model task effects of strategic prompting

Following previous results, we expected to observe a significantly higher score across all conditions in both tasks when implementing the SCoT prompt compared to the baseline condition [40]. This covers our analyses for calculating group differences for choice selection, task performance and prediction accuracy. For the inspection game, this required three comparisons. For Welch's two-tailed t-tests with Bonferroni correction (α = 0.017), assuming a large effect size (d = 0.8), power analyses indicated a required sample size of 35 per LLM group to achieve 80% power. This was met with our sample of at least 49 agents per group, and an allocation ratio of 1.0.

Across-model task effects of strategic prompting

Differences in task performance were also expected across LLMs, across models and prompting strategies. To justify our sample size for the 2 (prompt) $\times$ 3 (model) factorial ANOVA, we ran a power analysis for a fixed-effects ANOVA with six groups. Expecting a medium effect size (f = 0.25, α = 0.05), we required a total sample size of N = 158 to achieve 80% power, whilst the main effect of prompt required N = 128. Both were exceeded by our sample of at least 49 agents per cell. For the separate one-way comparison of the four normally-prompted models, a fixed-effects, one-way ANOVA with four groups and a medium-to-large effect size (0.25 < f < 0.4) required a total of 76 < N < 180 (19 < N < 45 per group) to achieve 80% power (α = 0.05), which was also met.

Rock-paper-scissors

Within-human effects of bot level on performance

To examine the effect of bot level on human performance, a one-way repeated measures ANOVA was conducted. Expecting a medium effect size (f = 0.25), power analyses indicated a required sample size of 28 participants to achieve 80% power (α = 0.05) with three repeated measurements (bot levels 0, 1, and 2). This was exceeded with our sample of 172 human participants, which was sufficient to detect 80% power with a small effect size (f = 0.15).

Cross-group comparisons of performance across humans and LLMs

Comparisons of task performance across humans and LLMs were conducted using a linear mixed model with group and bot level as fixed effects and subject as a random effect. For the main effect of group, expecting a large effect size (f = 0.40), power analyses indicated a required sample size of 13 participants per group to achieve 80% power (α = 0.05). For the main effect of bot level, expecting a medium effect size (f = 0.25), power analyses indicated a total sample size of 28 participants to achieve 80% power. For the group $\times$ bot level interaction, expecting a medium effect size (f = 0.25), power analyses indicated a required sample size of 10 participants per group to achieve 80% power. In all tests, the calculated sample sizes were substantially exceeded with our collected data.

Within-group pairwise comparisons of bot level effects

To examine differences in performance between bot levels within each group, paired t-tests with Bonferroni corrections were conducted. This required three pairwise comparisons per group (bot level 0 vs 1, 0 vs 2, and 1 vs 2). For one-tailed paired t-tests with Bonferroni correction (α = 0.017), assuming a medium effect size (dz = 0.50), power analyses indicated a required sample size of 38 per group to achieve 80% power. This was met or exceeded across all groups with our sample sizes of 172 human participants and 566-568 participants per LLM group.

Choice probability distributions

To test whether choice frequencies for rock, paper, and scissors deviated from a uniform distribution (1/3 probability for each choice), chi-square goodness-of-fit tests were conducted for each

group. For human participants, where we expected minimal deviation from uniformity, assuming a small effect size (w = 0.10), power analyses indicated a required sample size of 964 trials to achieve 80% power (α = 0.05, df = 2). For LLM groups, where we expected small to medium deviations from uniformity, assuming a small-medium effect size (w = 0.20), power analyses indicated a required sample size of 241 trials to achieve 80% power. Additionally, chi-square tests on first choice selections required a sample size of 39 participants per group to detect a large effect size (w = 0.50) with 80% power. In all cases, the required sample sizes were substantially exceeded with our collected data.

Power calculations were computed using GPower 3.1 Software [200]. Effect sizes were labelled following guidance [201] suggesting values of d/dz = 0.2/0.5/0.8 (small/medium/large) for t-tests, f = 0.10/0.25/0.40 for ANOVA effects, and w = 0.10/0.30/0.50 for chi-square tests.

Behavioural analyses

Rock-paper-scissors

Linear mixed model specification for task performance (payoff)

Linear mixed-effects models (LMMs) were used to assess differences in mean payoff across model strategies and bot sophistication levels.

Two complementary LMMs were fitted to the data:

LMM1: for ANOVA, main effects, interaction effects, and simple effects comparisons

```
Payoff ~ ModelStrategy * BotLevel + (1|SubjectID)
```

LMM2: for within-group comparisons (i.e., across bot-levels)

```
Payoff ~ ModelStrategy * BotLevelNumeric + (1|SubjectID)
```

In the models, each LLM subject was coded as per data collection, only completing a single block. Models were subsequently fit to 172 human participants, 566 DeepSeek participants, 568 DeepSeek-SCoT participants and 566 GPT-5 participants. All analyses were conducted in R using the *lme4* package for model fitting, *emmeans* for marginal means and contrasts, and *effectsize* for effect size calculations.

Computational modeling

Inspection game

Candidate models

Following prior studies [80, 106], we fit a series of computational models representing participants' possible strategies across the game. In the simplest learning strategy, players may simply choose the action in the recent past which provided the most reward through reinforcement learning (RL). However, players may also anticipate that the opponent will form a prediction of their choice and counter it accordingly. This approach, termed 'fictitious play', represents a strategy corresponding to first-order recursive thinking. Players may adopt a still more sophisticated approach by assuming the opponent adopts a first-order approach themselves and countering this choice accordingly.

Termed 'influence play', this strategy represents second-order recursive thinking. Finally, to test whether participants optimised by removing predictability rather than by learning, we included a mixed-equilibrium (ME) reference model in which choices follow a Bernoulli process with a single per-subject rate and no trial-by-trial updating [106]. For the inspection game, the mixed-strategy Nash equilibrium for the employee is P(Work) = 0.5.

Reinforcement learning. The reinforcement learning (RL) model suggests that participants do not pay attention to the opponent's choices but simply learn from their previous actions and rewards. This model modifies the value of each choice for each trial *t* in response to a prediction error δ weighted by a learning rate η:

$$V_t^{\text{Work}} = V_{t-1}^{\text{Work}} + \eta \cdot \delta_{t-1} \tag{1}$$

The prediction error quantifies the difference between the predicted and the actual outcome, reflecting the choice and whether it was subsequently rewarded:

$$\delta_{t-1} = R_{t-1} - V_{t-1}^{\text{Work}} \tag{2}$$

Fictitious play. The fictitious play model assumes that the participant instead monitors their opponent's probability of selecting an action (specifically, the probability of choosing "Not Inspect") based on the opponent's previous play history:

$$P_t^{\text{Not Inspect}} = P_{t-1}^{\text{Not Inspect}} + \eta \cdot \delta_{t-1}^{P} \tag{3}$$

This first-order belief is subsequently used to select the action that maximises the expected outcome. The prediction error δ_{t-1}^{P} is updated:

$$\delta_{t-1}^{P} = P_{t-1} - P_{t-1}^{\text{Not Inspect}} \tag{4}$$

Here P_{t-1} represents the observed action taken by the opponent; 1 if the employer chose "Not Inspect" and 0 if the opponent chose "Inspect". The participant's choice value in the model is subsequently determined based on the payoff matrix:

$$V_t^{\text{Work}} = 50 - 100 \cdot P_t^{\text{Not Inspect}} \tag{5}$$

Influence learning. The influence learning model integrates both first- and second-order beliefs, suggesting that participants' decisions are jointly influenced by the opponent's past choices and expectations based on observing the participant's past choices:

$$P_t^{\text{Not Inspect}} = P_{t-1}^{\text{Not Inspect}} + \eta \cdot \delta_{t-1}^{P} + \kappa \cdot \delta_{t-1}^{Q} \tag{6}$$

The second-order belief in the influence learning model is weighted by κ, and the prediction error of the inferred opponent's belief about the participant's action is calculated as:

$$\delta_{t-1}^{Q} = Q_{t-1} - Q_{t-1}^{\text{Work}} \tag{7}$$

where Q_{t-1} is the participant's action at the trial, recorded as 1 when the participant chooses "Work" and 0 when the participant chooses "Shirk". Q_{t-1}^{Work} is the opponent's inferred participant's working probability:

$$Q_{t-1}^{\text{Work}} = \frac{1}{5} - \frac{1}{5\beta} \cdot \log\left(\frac{1 - P_{t-1}^{\text{Not Inspect}}}{P_{t-1}^{\text{Not Inspect}}}\right) \tag{8}$$

$P_t^{\text{Not Inspect}}$ is derived by combining the first-order belief update $\eta \cdot \delta_{t-1}^{P}$ and the second-order belief update $\kappa \cdot \delta_{t-1}^{Q}$, and is then used to update the choice value V_t^{Work}. Importantly, if the weight κ for the second-order belief is equal to 0, the model reduces to the fictitious play model.

Mixed-equilibrium reference. The mixed-equilibrium (ME) model [106] provides a non-learning baseline. Each participant's choices are modelled as a Bernoulli process with a single rate δ_s (the probability of working), with no dependence on trial history:

$$Q_t^s \sim \text{Bernoulli}(\delta_s) \tag{9}$$

Because the model is memoryless, its likelihood is scored on the same observation set as the learning models (see below), so that the LOO comparison against ME is valid.

Predictive fit: model fitting and comparison

All models were implemented in JAGS (v4.3.2) [202] and fitted via the {rjags} and {runjags} interfaces in R v4.4.1 (2024-06-14). Predictive model comparison was conducted using approximate leave-one-out cross-validation (LOO-CV), with the pointwise log-likelihood evaluated in R and the LOO Information Criterion (LOO-IC) computed via the {loo} v2.7.0 package [203].

We used a hierarchical Bayesian structure for all groups, in which each participant's parameters were drawn from a group-level distribution described by a mean and a between-subject variance. The human group represents a heterogeneous sample, and its group-level distribution was estimated under weakly-informative priors. However, the LLM groups comprise near-identical agents running the same model under a fixed prompt. LLM and human models differed only in the group-level mean hyperprior, for which we placed a weakly-informative logit-normal prior (median 0.5) in place of the human Beta(1, 1). The between-subject variance, freely estimated under the same prior as the human group, returned near-zero values for every LLM group. An initial attempt to fit the LLM groups under the same diffuse priors as the human group produced unstable group-level estimates, consistent with this homogeneity.

Parameterisation. The learning-rate parameters were bounded to the unit interval. The first-order learning rate (η), the second-order weight (κ) and the inverse temperature (β) were each drawn from a Beta distribution truncated to $[0.001, 0.999]$. For the human group, the group-level Beta distributions were parameterised by a mean and precision, with the mean drawn from Beta$(1, 1)$ and the precision from Gamma$(1, 0.1)$. For the LLM groups, the same beta mean/precision structure was used, but the group-level mean was given a weakly-informative logit-normal prior (median 0.5) with the precision again from Gamma$(1, 0.1)$. The ME model used the same Beta mean/precision form for its per-subject rate, with the precision drawn from

Uniform$(0.001, 100)$.

Decision rule. For the influence and fictitious play models, the belief q_t that the employer would not inspect was updated trial-by-trial and mapped to a probability of working via a logistic decision rule:

$$P(\text{Work})_t = \sigma(15\beta\,(2 - 4q_t)) \tag{10}$$

where the boundary-separation coefficient is pinned at 15. For the RL model, the probability of working was derived from the difference in action values:

$$P(\text{Work})_t = \sigma(15\beta\,(V_t^{\text{Work}} - V_t^{\text{Shirk}})) \tag{11}$$

Convergence and comparison. All models were estimated using 3 independent Markov chains. Convergence was assessed using the Gelman-Rubin statistic ($\widehat{R}$) [204] and the effective sample size (ESS), with $\widehat{R} < 1.01$ and ESS > 400 for all reported parameters. Models were compared within each group using LOO-CV, which approximates the expected log pointwise predictive density (ELPD) for new data [203], where lower LOO-IC values indicate a better out-of-sample predictive accuracy.

Model–behaviour fit: parameter–behaviour correspondence

To assess the construct validity of the fitted parameters we regressed three model-free behavioural summaries on the fitted η and κ. Parameters were taken from the influence-play fit for every group where it presented as the winning model. The predictors were the fitted η and κ, which were log-transformed and then z-scored across the pooled analysis set, so that coefficients are comparable across groups.

For each group we fitted three separate ordinary-least-squares (OLS) linear regressions using the `lm()` function from the R package {stats} v4.4.1, one per behavioural outcome, each entering η and κ as predictors:

- Trial-to-trial switching: $p(\text{Switch})_i \sim \eta_i + \kappa_i$
- Lagged tracking of the employer's previous move: $r(\text{employer action}_{t-1}, \text{choice}_t)_i \sim \eta_i + \kappa_i$
- Task performance: mean payoff$_i \sim \eta_i + \kappa_i$

As DeepSeek-SCoT's winning model was fictitious play, we ran the same three regressions for DeepSeek-SCoT using its own winning fictitious-play fit, with η as the sole predictor. Because this places DeepSeek-SCoT's η on a different z-scale from the pooled set used for the other groups, its coefficients are not directly comparable and are marked in Table S3.

Coefficients are reported as standardised slopes with their standard errors and two-sided p-values. Given the substantial differences in parameter variance, we additionally computed the scale-free partial correlation which expresses each effect within the group on a bounded [−1, 1] scale.

Generative fit: posterior predictive check

For the posterior predictive check, we drew 2,000 samples from each group's winning-model posterior, using each draw's per-participant parameter values to simulate one full replicate dataset - the same number of participants and trials as observed - by running the model forward. For each statistic - the overall probability of working, the switch rate and the mean payoff - we computed the group-level mean of that statistic across all simulated participants, yielding a 2,000-value distribution of simulated group means. Subtracting the single fixed observed group mean from every draw gave a discrepancy distribution which centred on zero when the model reproduces the observed level and is shifted away from zero when it does not.

We examined each statistic with the region-of-practical-equivalence (ROPE) equivalence test [119], in the full-posterior form following recommended guidelines [120]. Subsequently, we computed the percentage of the distribution lying within a ROPE of ± 0.05, and accepted a statistic as practically reproduced when this exceeded 97.5%, rejected it below 2.5%, and left it undecided in between. The ROPE half-width was fixed at 0.05 on the shared [0,1] scale on which all three statistics lie. The percentage in ROPE per statistic is given in Table S4.

Rock-paper-scissors

Candidate models

Fitted models were identical to a previous study investigating adaptive mentalization among human participants [107], which we only briefly summarise below. We refer readers to the original publication for a more comprehensive description and formulation of the models.

Cognitive Hierarchy Assessment (CHASE) model

The CHASE model works on the premise that there is an action that non-strategic players (defined as k = 0) would choose, whereas strategic players add a limited number of recursive reasoning steps by iteratively best-responding to that action. Importantly, adaptive agents infer the level of recursive reasoning of the opponent by integrating evidence for the different levels over time.

More formally, non-strategic play (k = 0) is governed by an updating rule which maps the history of the game to attractions for each action at each trial. In this context, attractions are a noisy representation of historical action frequencies with an exponential memory decay. These attractions are linked to the observed behaviour, governing the extent to which agents tend to stick to the historical action frequencies or act randomly.

While a level-0 agent solely responds to their own action frequencies, higher-level agents try to predict what the opponent will play, or predicting what the opponent thinks they will play, etc. Specifically, an agent with sophistication k assumes that the sophistication of the other agent is exactly one level lower (i.e. k - 1) and therefore applies k steps of recursive reasoning. While a k = 1 agent is certain that the opponent must be k = 0, higher-level agents face uncertainty about which level the opponent is most likely playing. To adapt to different strategic players, adaptive agents form and update beliefs about the opponent's level of sophistication. Agents then form an

integrated prediction over the most likely opponent action, weighted by the belief distribution over the opponent's level, and noisily best respond to it.

In total, the resulting model is characterized by 5 parameters: α, β, γ, λ and κ (regulating the speed of updating attractions, recursive reasoning noise, sensitivity to evidence for opponent's level, loss aversion, and depth of mentalization ability, respectively).

Alternative models

Reinforcement learning. A simple non-strategic learning rule is to repeat the actions that were successful in the past. Specifically, the model updates a value for each action at a learning rate α and selects actions via a softmax over these values with inverse temperature β, learning only from its own payoffs with no representation of the opponent.

Fictitious play. A more sophisticated approach than simple reinforcement learning is fictitious play, where agents try to estimate the probabilities with which the other player chooses her actions, and then best respond to them. This agent corresponds to a CHASE model with κ = 1 (i.e. equivalent to k = 1). Both fictitious play and reinforcement learning are fully specified by two free parameters, a learning rate α and a temperature parameter β.

Experience-weighted attraction (EWA). EWA [205] is a hybrid model combining elements of reinforcement learning and fictitious play. In brief, this is achieved by introducing a parameter δ that quantifies the extent to which an agent also learns from foregone payoffs (i.e. the payoffs that would have resulted when choosing different actions). Our formulation of EWA has three free parameters (δ, □ , ρ) that govern the updating process, where we manually set the initial values for attractions and experience-equivalents that are free parameters in the original formulation.

Self-tuning EWA. This is a simplified version of the EWA [206] that fixes some parameter values to empirical values and replaces others with functions of experience. As a result, only the behavioural temperature parameter β is estimated.

Theory-of-Mind k (ToMk). Similar to CHASE, the ToMk model [122] also involves simulating what opponents of increasing sophistication would play, but uses a heuristic confidence-updating mechanism to form a response, rather than a Bayesian update on which the CHASE model is based.

Specifically, the level-0 agent is defined as noisily responding to recency-weighted past action frequencies of the opponent. The level-1 agent simulates the opponent's level-0 behavior, computes their most likely action and then responds combining the best response to this opponent prediction with the expected value from their 'own' level-0 strategy. Finally, higher levels simulate this response process from the opponent's perspective and then recursively integrate higher-level responses with a confidence for each level.

The ToMk model consists three free parameters: a learning rate α governing both the frequency and confidence updates, an inverse temperature β controlling choice stochasticity, and an integer κ setting the highest level of recursive reasoning the agent performs.

Model fitting and comparison

To fit the models to our participants' behaviour, we used maximum likelihood estimation by combining an initial grid-search with the *fminunc* function in MATLAB. We fitted one set of parameters per subject (i.e. across opponents) for each model, but reset all relevant prior belief variables to uniform upon facing a new opponent where appropriate (e.g. beliefs, attractions, etc) (Fig S2). We computed negative log-likelihood (negLL), Akiake Information Criterion (AIC) and Bayesian information Criterion (BIC) scores for model comparison metrics whilst accounting for differences in the number of free parameters. We used random effects Bayesian model comparison to ensure that the comparison is not overly sensitive to outliers, using the VBA toolbox [207]. We computed the winning model through the protected exceedance probability (PXP), quantifying the likelihood that a model is expressed more frequently than all other candidates, accounting for the possibility of chance differences.

Data and code availability

Data and code for modeling and analysis are available on request.

Acknowledgements

A.S. was funded by the MRC AIM and AIM iCASE Grant (MR/W007002/1). L.Z. was supported by the Wellcome Trust (228268/Z/23/Z), the Royal Society (IES\R3\243253), and the Birmingham–Nanjing University Joint Research and Education Centre pump-priming funding. The computations described in this paper were partially performed using the Birmingham Environment for Academic Research (BEAR). P.L.L. was supported by a Medical Research Council Fellowship (MR/P014097/1 and MR/P014097/2), a Jacobs Foundation Research Fellowship, a Leverhulme Prize (PLP-2021-196), a Wellcome Trust/Royal Society Sir Henry Dale Fellowship (223264/Z/21/Z) and a UKRI EPSRC Frontiers Research Guarantee/ERC Starting Grant (EP/X020215/1). We thank Ian Apperly, Robert Lee, Ayat Abdurahman, Daniel Drew and Luca Hargitai for their help with data collection, and Niklas Buergi for helpful discussions regarding the modelling analyses.

Author contributions

A.S.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration.
X.Z.: Investigation, Data Curation, Writing – Review & Editing
A.K.: Resources, Writing – Review & Editing, Supervision
P.L.L.: Resources, Writing – Review & Editing, Supervision, Funding acquisition
L.Z.: Conceptualization, Methodology, Validation, Resources, Writing – Review & Editing, Supervision, Project administration, Funding acquisition

Competing interests

The authors declare no competing interests.

Declaration of generative AI

In the preparation of this manuscript, agentic workflows using Claude Code (models Sonnet, Opus and Fable) were used to screen text supplied by the authors, and for AI-assisted programming. However, all programming outputs were verified and screened by the lead author, and no original text was generated by the agents. Use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgment of generative AI in academic research [208, 209].

References

1. Fonagy, P., Gergely, G. & Jurist, E. L. *Affect Regulation, Mentalization and the Development of the Self*. (Routledge, 2018). DOI: 10.4324/9780429471643
2. Fonagy, P. & Allison, E. What is mentalization?: The concept and its foundations in developmental research. *Minding the Child* (2012).
3. Freeman, C. What is mentalizing? An overview. *British Journal of Psychotherapy* **32,** 189–201 (2016). DOI: 10.1111/bjp.12220
4. Apperly, I. A. What is 'theory of mind'? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology* **65,** 825–839 (2012). DOI: 10.1080/17470218.2012.676055
5. Apperly, I. A. & Butterfill, S. A. Do humans have two systems to track beliefs and belief-like states? *Psychological Review* **116,** 953–970 (2009). DOI: 10.1037/a0016923
6. Frith, C. D. & Frith, U. in *The Neural Basis of Mentalizing* (eds. Gilead, M. & Ochsner, K. N.) 17–45 (Springer International Publishing, 2021). DOI: 10.1007/978-3-030-51890-5_2
7. Kliemann, D. & Adolphs, R. The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology* **24,** 1–6 (2018). DOI: 10.1016/j.copsyc.2018.02.015
8. Luyten, P. & Fonagy, P. The neurobiology of mentalizing. *Personality Disorders: Theory, Research, and Treatment* **6,** 366–379 (2015). DOI: 10.1037/per0000117
9. Schurz, M. *et al.* Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin* **147,** 293–327 (2021). DOI: 10.1037/bul0000303
10. Tamir, D. I. & Thornton, M. A. Modeling the predictive social mind. *Trends in Cognitive Sciences* **22,** 201–212 (2018). DOI: 10.1016/j.tics.2017.12.005
11. Berke, M. D., Horschler, D. J., Royka, A., Santos, L. R. & Jara-Ettinger, J. What primates know about other minds and when they use it: A computational approach to comparative theory of mind. 2023.08.02.551487 (2025). DOI: 10.1101/2023.08.02.551487
12. Catala, A., Mang, B., Wallis, L. & Huber, L. Dogs demonstrate perspective taking based on geometrical gaze following in a guesser–knower task. *Animal Cognition* **20,** 581–589 (2017). DOI: 10.1007/s10071-017-1082-x
13. de Waal, F. B. M. Apes know what others believe. *Science* **354,** 39–40 (2016). DOI: 10.1126/science.aai8851
14. Devaine, M. *et al.* Reading wild minds: A computational assay of theory of mind sophistication across seven primate species. *PLOS Computational Biology* **13,** e1005833 (2017). DOI: 10.1371/journal.pcbi.1005833
15. Keefner, A. Corvids infer the mental states of conspecifics. *Biology & Philosophy* **31,** 267–281 (2016). DOI: 10.1007/s10539-015-9509-8
16. Krupenye, C. & Call, J. Theory of mind in animals: Current and future directions. *WIREs Cognitive Science* **10,** e1503 (2019). DOI: 10.1002/wcs.1503

17. Maginnity, M. E. & Grace, R. C. Visual perspective taking by dogs (canis familiaris) in a guesser–knower task: Evidence for a canine theory of mind? *Animal Cognition* **17,** 1375–1392 (2014). DOI: 10.1007/s10071-014-0773-9
18. Miller, R., Claisse, E., Timulak, A. & Clayton, N. S. Development of cognition in corvids. 2026.02.27.708529 (2026). DOI: 10.64898/2026.02.27.708529
19. Royka, A. & Santos, L. R. Theory of mind in the wild. *Current Opinion in Behavioral Sciences* **45,** 101137 (2022). DOI: 10.1016/j.cobeha.2022.101137
20. Taylor, A. H. Corvid cognition. *WIREs Cognitive Science* **5,** 361–372 (2014). DOI: 10.1002/wcs.1286
21. Brady, O., Nulty, P., Zhang, L., Ward, T. E. & McGovern, D. P. Dual-process theory and decision-making in large language models. *Nature Reviews Psychology* **4,** 777–792 (2025). DOI: 10.1038/s44159-025-00506-1
22. Demszky, D. *et al.* Using large language models in psychology. *Nature Reviews Psychology* **2,** 688–701 (2023). DOI: 10.1038/s44159-023-00241-5
23. Frank, M. C. & Goodman, N. D. Cognitive modeling using artificial intelligence. (2025). DOI: 10.1146/annurev-psych-030625-040748
24. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95,** 245–258 (2017). DOI: 10.1016/j.neuron.2017.06.011
25. Jackson, M. O. *et al.* AI behavioral science. (2026). DOI: 10.48550/arXiv.2509.13323
26. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40,** e253 (2017). DOI: 10.1017/S0140525X16001837
27. Mathis, M. W. Leveraging insights from neuroscience to build adaptive artificial intelligence. *Nature Neuroscience* **29,** 13–24 (2026). DOI: 10.1038/s41593-025-02169-w
28. Sadeh, S. & Clopath, C. The emergence of NeuroAI: Bridging neuroscience and artificial intelligence. *Nature Reviews Neuroscience* **26,** 583–584 (2025). DOI: 10.1038/s41583-025-00954-x
29. Voudouris, K., Cheke, L. & Schulz, E. Bringing comparative cognition approaches to AI systems. *Nature Reviews Psychology* **4,** 363–364 (2025). DOI: 10.1038/s44159-025-00456-8
30. Zador, A. *et al.* Catalyzing next-generation artificial intelligence through NeuroAI. *Nature Communications* **14,** 1597 (2023). DOI: 10.1038/s41467-023-37180-x
31. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* **120,** e2218523120 (2023). DOI: 10.1073/pnas.2218523120
32. Crockett, M. J. & Messeri, L. AI surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences* **30,** 203–215 (2026). DOI: 10.1016/j.tics.2025.09.012
33. Hagendorff, T. *et al.* Machine psychology. (2024). DOI: 10.48550/arXiv.2303.13988
34. Huang, J. *et al.* On the reliability of psychological scales on large language models. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds. Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 6152–6173 (Association for Computational Linguistics, 2024). DOI: 10.18653/v1/2024.emnlp-main.354

35. Löhn, L., Kiehne, N., Ljapunov, A. & Balke, W.-T. Is machine psychology here? On requirements for using human psychological tests on large language models. in *Proceedings of the 17th International Natural Language Generation Conference* (eds. Mahamood, S., Minh, N. L. & Ippolito, D.) 230–242 (Association for Computational Linguistics, 2024). DOI: 10.18653/v1/2024.inlg-main.19

36. Peereboom, S., Schwabe, I. & Kleinberg, B. Cognitive phantoms in large language models through the lens of latent variables. *Computers in Human Behavior: Artificial Humans* **4,** 100161 (2025). DOI: 10.1016/j.chbah.2025.100161

37. Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D. & Hawkins, R. D. Using LLMs to advance the cognitive science of collectives. *Nature Computational Science* **5,** 704–707 (2025). DOI: 10.1038/s43588-025-00848-z

38. Vaswani, A. *et al.* Attention is all you need. in *Advances in Neural Information Processing Systems* **30,** (Curran Associates, Inc., 2017).

39. Abdulhai, M. *et al.* Moral foundations of large language models. (2023). DOI: 10.48550/arXiv.2310.15337

40. Akata, E. *et al.* Playing repeated games with large language models. *Nature Human Behaviour* 1–11 (2025). DOI: 10.1038/s41562-025-02172-y

41. Cheung, V., Maier, M. & Lieder, F. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences* **122,** e2412015122 (2025). DOI: 10.1073/pnas.2412015122

42. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science* **5,** 627–634 (2025). DOI: 10.1038/s43588-025-00840-7

43. Fontana, N., Pierri, F. & Aiello, L. M. Nicer than humans: How do large language models behave in the prisoner's dilemma? *Proceedings of the International AAAI Conference on Web and Social Media* **19,** 522–535 (2025). DOI: 10.1609/icwsm.v19i1.35829

44. Lampinen, A. K. *et al.* Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* **3,** pgae233 (2024). DOI: 10.1093/pnasnexus/pgae233

45. Nie, A. *et al.* MoCa: Measuring human-language model alignment on causal and moral judgment tasks. (2023). DOI: 10.48550/arXiv.2310.19677

46. Sun, L. *et al.* Large language models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity* **57,** 101870 (2025). DOI: 10.1016/j.tsc.2025.101870

47. Wang, Y. *et al.* Strategic Chain-of-Thought: Guiding accurate reasoning in LLMs through strategy elicitation. (2024). DOI: 10.48550/arXiv.2409.03271

48. Yax, N., Anlló, H. & Palminteri, S. Studying and improving reasoning in humans and machines. *Communications Psychology* **2,** 51 (2024). DOI: 10.1038/s44271-024-00091-8

49. Wellman, H. M., Cross, D. & Watson, J. Meta-analysis of Theory-of-Mind Development: The truth about false belief. *Child Development* **72,** 655–684 (2001). DOI: 10.1111/1467-8624.00304

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

51. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with GPT-4. (2023). DOI: 10.48550/arXiv.2303.12712

52. Bujka, Z., Lukacs, A., Vedres, P. & Babarczy, A. Do large language models possess a theory of mind? A comparative evaluation using the strange stories paradigm. (2025).

53. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* **121,** e2405460121 (2024). DOI: 10.1073/pnas.2405460121

54. Liu, Y., Pretty, E. J., Huang, J. & Sugawara, S. TactfulToM: Do LLMs have the theory of mind ability to understand white lies? in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (eds. Christodoulopoulos, C., Chakraborty, T., Rose, C. & Peng, V.) 25043–25061 (Association for Computational Linguistics, 2025). DOI: 10.18653/v1/2025.emnlp-main.1272

55. Sarıtaş, K., Tezören, K. & Durmazkeser, Y. A systematic review on the evaluation of large language models in theory of mind tasks. (2025). DOI: 10.48550/arXiv.2502.08796

56. Strachan, J. W. A. *et al.* Testing theory of mind in large language models and humans. *Nature Human Behaviour* **8,** 1285–1295 (2024). DOI: 10.1038/s41562-024-01882-z

57. Zhou, P. *et al.* How FaR are large language models from agents with Theory-of-Mind? (2023). DOI: 10.48550/arXiv.2310.03051

58. Ullman, T. Large language models fail on trivial alterations to Theory-of-Mind Tasks. (2023). DOI: 10.48550/arXiv.2302.08399

59. Attanasio, M. *et al.* Does ChatGPT have a typical or atypical theory of mind? *Frontiers in Psychology* **15,** 1488172 (2024). DOI: 10.3389/fpsyg.2024.1488172

60. Moore, J. *et al.* Do large language models have a planning theory of mind? Evidence from MindGames: A multi-step persuasion task. (2025). DOI: 10.48550/arXiv.2507.16196

61. Marchetti, A., Manzi, F., Riva, G., Gaggioli, A. & Massaro, D. Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. *Cyberpsychology, Behavior, and Social Networking* **28,** 505–514 (2025). DOI: 10.1089/cyber.2024.0536

62. Pang, D. K. F., Pang, S. K. Y., Broeker, M. D. & Hibble, A. Do large language models have a theory of mind? *Proceedings of the National Academy of Sciences* **122,** e2507080122 (2025). DOI: 10.1073/pnas.2507080122

63. Shapira, N. *et al.* Clever hans or neural theory of mind? Stress testing social reasoning in large language models. (2023). DOI: 10.48550/arXiv.2305.14763

64. Holtzman, A., West, P. & Zettlemoyer, L. Generative models as a complex systems science: How can we make sense of large language model behavior? *Journal of Social Computing* **6,** 75–94 (2025). DOI: 10.23919/JSC.2025.0009

65. Hu, J., Sosa, F. & Ullman, T. Re-evaluating theory of mind evaluation in large language models. (2025). DOI: 10.48550/arXiv.2502.21098

66. Ku, A. *et al.* Levels of analysis for large language models. (2025). DOI: 10.48550/arXiv.2503.13401
67. Mondorf, P. & Plank, B. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. (2024). DOI: 10.48550/arXiv.2404.01869
68. Riemer, M. *et al.* Position: Theory of mind benchmarks are broken for large language models. (2025). DOI: 10.48550/arXiv.2412.19726
69. Wagner, E., Alon, N., Barnby, J. M. & Abend, O. Mind your theory: Theory of mind goes deeper than reasoning. (2025). DOI: 10.48550/arXiv.2412.13631
70. Lu, Y.-L., Zhang, C., Song, J., Fan, L. & Wang, W. Do theory of Mind Benchmarks Need Explicit Human-like Reasoning in language models? (2025). DOI: 10.48550/arXiv.2504.01698
71. Blank, I. A. What are large language models supposed to model? *Trends in Cognitive Sciences* **27,** 987–989 (2023). DOI: 10.1016/j.tics.2023.08.006
72. Connell, L. & Lynott, D. What can language models tell us about human cognition? *Current Directions in Psychological Science* **33,** 181–189 (2024). DOI: 10.1177/09637214241242746
73. Camerer, C. F. Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences* **7,** 225–231 (2003). DOI: 10.1016/S1364-6613(03)00094-9
74. Camerer, C. F. Does strategy research need game theory? *Strategic Management Journal* **12,** 137–152 (1991). DOI: 10.1002/smj.4250121010
75. Colman, A. M. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences* **26,** 139–153 (2003). DOI: 10.1017/S0140525X03000050
76. Griessinger, T. & Coricelli, G. The neuroeconomics of strategic interaction. *Current Opinion in Behavioral Sciences* **3,** 73–79 (2015). DOI: 10.1016/j.cobeha.2015.01.012
77. Konovalov, A. in *Neuroeconomics: Core Topics and Current Directions* (eds. Smith, D. V., Lockwood, P. L. & Fareri, D. S.) 451–465 (Springer Nature Switzerland, 2026). DOI: 10.1007/978-3-032-02925-6_25
78. Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M. & Gläscher, J. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia* **146,** 107488 (2020). DOI: 10.1016/j.neuropsychologia.2020.107488
79. Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J. & Lamm, C. Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience* **15,** 695–707 (2020). DOI: 10.1093/scan/nsaa089
80. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. Neural correlates of mentalizing-related computations during strategic interactions in humans. *PNAS Proceedings of the National Academy of Sciences of the United States of America* **105,** 6741–6746 (2008). DOI: 10.1073/pnas.0711099105
81. Camerer, C. F., Ho, T.-H. & Chong, J.-K. A cognitive hierarchy model of games*. *The Quarterly Journal of Economics* **119,** 861–898 (2004). DOI: 10.1162/0033553041502225

82. Devaine, M., Hollard, G. & Daunizeau, J. The social bayesian brain: Does mentalizing make a difference when we learn? *PLOS Computational Biology* **10,** e1003992 (2014). DOI: 10.1371/journal.pcbi.1003992

83. Brookins, P. & DeBacker, J. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin* **44,** 25–37 (2024).

84. Feng, X. *et al.* A survey on large language model-based social agents in game-theoretic scenarios. (2025). DOI: 10.48550/arXiv.2412.03920

85. Jia, J., Yuan, Z., Pan, J., McNamara, P. E. & Chen, D. LLM strategic reasoning: Agentic study through behavioral game theory. (2025). DOI: 10.48550/arXiv.2502.20432

86. Kirshner, S., Pan, Y. & Wu, J. X. Pro-social when simple and cold-hearted when complex: How task difficulty shapes LLM behavior. (2025). DOI: 10.2139/ssrn.5266963

87. Kitadai, A., Rico Lugo, S. D., Tsurusaki, Y., Fukasawa, Y. & Nishino, N. Can AI with High Reasoning Ability Replicate Human-like Decision Making in economic experiments? *Group Decision and Negotiation* **34,** 1303–1326 (2025). DOI: 10.1007/s10726-025-09946-9

88. Lorè, N. & Heydari, B. Strategic behavior of large language models: Game structure vs. Contextual framing. (2023). DOI: 10.48550/arXiv.2309.05898

89. Piedrahita, D. G. *et al.* Corrupted by reasoning: Reasoning language models become free-riders in public goods games. (2025). DOI: 10.48550/arXiv.2506.23276

90. Roberts, J., Moore, K. & Fisher, D. Do large language models learn human-like strategic preferences? in *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)* (eds. Kamalloo, E. et al.) 97–108 (Association for Computational Linguistics, 2025). DOI: 10.18653/v1/2025.realm-1.8

91. Sun, H., Wu, Y., Cheng, Y. & Chu, X. Game theory meets large language models: A systematic survey. (2025). DOI: 10.48550/arXiv.2502.09053

92. Zhang, Y. *et al.* LLM as a mastermind: A survey of strategic reasoning with large language models. (2024). DOI: 10.48550/arXiv.2404.01230

93. Chen, B., Zhang, Z., Langrené, N. & Zhu, S. Unleashing the potential of prompt engineering for large language models. *Patterns* **6,** (2025). DOI: 10.1016/j.patter.2025.101260

94. Lin, Z. How to write effective prompts for large language models. *Nature Human Behaviour* **8,** 611–615 (2024). DOI: 10.1038/s41562-024-01847-2

95. Wei, D., Tsheringla, S., McPartland, J. C. & Allsop, A. Z. A. S. A. Combinatorial approaches for treating neuropsychiatric social impairment. *Philosophical Transactions of the Royal Society B: Biological Sciences* **377,** 20210051 (2022). DOI: 10.1098/rstb.2021.0051

96. Anglin, K. L., Milan, S., Hernandez, B. & Ventura, C. Improving alignment between human and machine codes: An empirical assessment of prompt engineering for construct identification in psychology. (2025). DOI: 10.48550/arXiv.2512.03818

97. Jiamin, O., Emile, E., Vincent, B., Paulina, P. & Yuli, S. Social preferences with unstable interactive reasoning: Large language models in economic trust games. (2025). DOI: 10.48550/arXiv.2505.17053

98. Ma, J. Can machines think like humans? A behavioral evaluation of LLM-agents in dictator games. (2024). DOI: 10.31219/osf.io/arvhx

99. Phelps, S. & Russell, Y. I. The machine psychology of cooperation: Can GPT models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity* **6,** 015018 (2025). DOI: 10.1088/2632-072X/ada711

100. Schoenegger, P., Jones, C. R., Tetlock, P. E. & Mellers, B. Prompt engineering large language models' forecasting capabilities. (2025). DOI: 10.48550/arXiv.2506.01578

101. Farrell, S. & Lewandowsky, S. *Computational Modeling of Cognition and Behavior*. (Cambridge University Press, 2018).

102. Farrell, S. & Lewandowsky, S. Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science* **19,** 329–335 (2010). DOI: 10.1177/0963721410386677

103. Guest, O. & Martin, A. E. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science* **16,** 789–802 (2021). DOI: 10.1177/1745691620970585

104. Hackel, L. M. & Amodio, D. M. Computational neuroscience approaches to social cognition. *Current Opinion in Psychology* **24,** 92–97 (2018). DOI: 10.1016/j.copsyc.2018.09.001

105. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8,** e49547 (2019). DOI: 10.7554/eLife.49547

106. Hill, C. A. *et al.* A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience* **20,** 1142–1149 (2017). DOI: 10.1038/nn.4602

107. Buergi, N., Aydogan, G., Konovalov, A. & Ruff, C. C. A neural signature of adaptive mentalization. *Nature Neuroscience* 1–11 (2026). DOI: 10.1038/s41593-026-02219-x

108. Redish, A. D. *et al.* Computational validity: Using computation to translate behaviours across species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **377,** 20200525 (2021). DOI: 10.1098/rstb.2020.0525

109. Robbins, T. W. & Cardinal, R. N. Computational psychopharmacology: A translational and pragmatic approach. *Psychopharmacology* **236,** 2295–2305 (2019). DOI: 10.1007/s00213-019-05302-3

110. Palminteri, S. & Wu, C. *Beyond Computational Functionalism: The Behavioral Inference Principle for Machine Consciousness*. (2025). DOI: 10.31234/osf.io/s7ptu_v3

111. Taschereau-Dumouchel, V., Hwang, J. S., Lau, H. & LeDoux, J. E. The ethical impasse of current consciousness science. *Neuron* **0,** (2026). DOI: 10.1016/j.neuron.2026.04.007

112. Coda-Forno, J., Binz, M., Wang, J. X. & Schulz, E. CogBench: A large language model walks into a psychology lab. (2024). DOI: 10.48550/arXiv.2402.18225

113. Hayes, W. M., Yax, N. & Palminteri, S. Relative value encoding in large language models: A multi-task, multi-model investigation. *Open Mind* **9,** 709–725 (2025). DOI: 10.1162/opmi_a_00209

114. Schubert, J. A., Jagadish, A. K., Binz, M. & Schulz, E. In-context learning agents are asymmetric belief updaters. (2024). DOI: 10.48550/arXiv.2402.03969

115. Collins, K. M. *et al.* Building machines that learn and think with people. *Nature Human Behaviour* **8,** 1851–1863 (2024). DOI: 10.1038/s41562-024-01991-9

116. Vidler, A. & Walsh, T. Playing games with large language models: Randomness and strategy. (2025). DOI: 10.48550/arXiv.2503.02582

117. Singh, A. *et al.* OpenAI GPT-5 system card. (2026). DOI: 10.48550/arXiv.2601.03267

118. Sohail, A. & Zhang, L. BayesCog: A freely available course in bayesian statistics and hierarchical bayesian modeling for psychological science. *BayesCog* (2025). DOI: 10.31234/osf.io/ua5ng_v1

119. Kruschke, J. K. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science* **1,** 270–280 (2018). DOI: 10.1177/2515245918771304

120. Makowski, D., Ben-Shachar, M. S., Chen, S. H. A. & Lüdecke, D. Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology* **10,** (2019). DOI: 10.3389/fpsyg.2019.02767

121. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies — revisited. *NeuroImage* **84,** 971–985 (2014). DOI: 10.1016/j.neuroimage.2013.08.065

122. Weerd, H. de, Diepgrond, D. & Verbrugge, R. Estimating the use of higher-order theory of mind using computational agents. *The B.E. Journal of Theoretical Economics* **18,** (2018). DOI: 10.1515/bejte-2016-0184

123. Lockwood, P. L. & Klein-Flügge, M. C. Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience* **16,** 761–771 (2021). DOI: 10.1093/scan/nsaa040

124. Moutoussis, M., Shahar, N., Hauser, T. U. & Dolan, R. J. Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry (Cambridge, Mass.)* **2,** 50–73 (2018). DOI: 10.1162/CPSY_a_00014

125. Nair, A., Rutledge, R. B. & Mason, L. Under the hood: Using computational psychiatry to make psychological therapies more mechanism-focused. *Frontiers in Psychiatry* **11,** (2020). DOI: 10.3389/fpsyt.2020.00140

126. Sohail, A. & Zhang, L. Informing the treatment of social anxiety disorder with computational and neuroimaging data. *Psychoradiology* **4,** kkae010 (2024). DOI: 10.1093/psyrad/kkae010

127. Ben-Zion, Z. *et al.* Assessing and alleviating state anxiety in large language models. *npj Digital Medicine* **8,** 132 (2025). DOI: 10.1038/s41746-025-01512-6

128. Floridi, L., Morley, J., Novelli, C. & Watson, D. What kind of reasoning (if any) is an LLM actually doing? On the stochastic nature and abductive appearance of large language models. (2025). DOI: 10.48550/arXiv.2512.10080

129. Yetman, C. C. Representation in large language models. (2025). DOI: 10.48550/arXiv.2501.00885

130. Beckmann, P. & Queloz, M. Mechanistic indicators of understanding in large language models. (2025). DOI: 10.48550/arXiv.2507.08017
131. Butlin, P. *et al.* Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences* **0,** (2025). DOI: 10.1016/j.tics.2025.10.011
132. Cuzzolin, F. A formal definition and meta-model for a machine theory of mind. (2026). DOI: 10.48550/arXiv.2606.03471
133. Ivanova, A. A. How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour* **9,** 230–233 (2025). DOI: 10.1038/s41562-024-02096-z
134. Wang, R., Luo, Y., Wang, Y., Zhang, L. & Wu, H. Towards naturalistic social neuroscience: A multi-level framework integrating real-world phenotyping, neurobiology, and computational mechanisms. *Cognitive, Affective, & Behavioral Neuroscience* **26,** 1420–1436 (2026). DOI: 10.3758/s13415-026-01413-5
135. Kaplan, J. *et al.* Scaling laws for neural language models. (2020). DOI: 10.48550/arXiv.2001.08361
136. Berti, L., Giorgi, F. & Kasneci, G. Emergent abilities in large language models: A survey. (2025). DOI: 10.48550/arXiv.2503.05788
137. Wang, X. *et al.* Do larger language models generalize better? A scaling law for implicit reasoning at pretraining time. (2025). DOI: 10.48550/arXiv.2504.03635
138. Ruan, Y., Maddison, C. J. & Hashimoto, T. Observational scaling laws and the predictability of language model performance. (2024). DOI: 10.48550/arXiv.2405.10938
139. Mina, M., Ruiz-Fernández, V., Falcão, J., Vasquez-Reina, L. & Gonzalez-Agirre, A. Cognitive biases, task complexity, and result interpretability in large language models. in *Proceedings of the 31st International Conference on Computational Linguistics* (eds. Rambow, O. et al.) 1767–1784 (Association for Computational Linguistics, 2025).
140. Momentè, F. *et al.* Triangulating LLM progress through benchmarks, games, and cognitive tests. (2025). DOI: 10.48550/arXiv.2502.14359
141. Raman, N. *et al.* STEER: Assessing the economic rationality of large language models. (2024). DOI: 10.48550/arXiv.2402.09552
142. Zhou, Z. *et al.* Rationality check! Benchmarking the rationality of large language models. (2025). DOI: 10.48550/arXiv.2509.14546
143. Jung, C. *et al.* Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. (2024). DOI: 10.48550/arXiv.2407.06004
144. Imran, S. *et al.* Are LLM belief updates consistent with bayes' theorem? (2025). DOI: 10.48550/ARXIV.2507.17951
145. Pal, A., Kitanovski, T., Liang, A., Potti, A. & Goldblum, M. Incoherent beliefs & inconsistent actions in large language models. (2025). DOI: 10.48550/arXiv.2511.13240
146. Zhu, J.-Q. & Griffiths, T. L. Computation-limited bayesian updating: A resource-rational analysis of approximate bayesian inference. *Psychological Review* **133,** 619–635 (2026). DOI: 10.1037/rev0000573

147. Ebouky, B., Bartezzaghi, A. & Rigotti, M. Eliciting reasoning in language models with cognitive tools. *arXiv.org* (2025).
148. Geng, H., Xu, B. & Li, P. UPAR: A kantian-inspired prompting framework for enhancing large language model capabilities. (2023). DOI: 10.48550/arXiv.2310.01441
149. Kramer, O. & Baumann, J. Unlocking structured thinking in language models with cognitive prompting. *arXiv.org* (2024).
150. Patil, A. & Jadon, A. Advancing reasoning in large language models: Promising methods and approaches. *arXiv.org* (2025).
151. Chen, R., Jiang, W., Qin, C. & Tan, C. Theory of mind in large language models: Assessment and enhancement. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T.) 31539–31558 (Association for Computational Linguistics, 2025). DOI: 10.18653/v1/2025.acl-long.1522
152. Lin, Z., Chan, C., Song, Y. & Liu, X. Constrained reasoning chains for Enhancing Theory-of-Mind in large language models. (2024). DOI: 10.48550/arXiv.2409.13490
153. Moghaddam, S. R. & Honey, C. J. Boosting Theory-of-Mind Performance in large language models via prompting. (2023). DOI: 10.48550/arXiv.2304.11490
154. Sarangi, S., Elgarf, M. & Salam, H. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. (2025). DOI: 10.48550/arXiv.2501.09056
155. Wilf, A., Lee, S. S., Liang, P. P. & Morency, L.-P. Think twice: Perspective-taking improves large language models' Theory-of-Mind Capabilities. (2023). DOI: 10.48550/arXiv.2311.10227
156. Zhang, Y., Wisniewski, G., Tomeh, N. & Charnois, T. Reasoning strategies in large language models: Can they follow, prefer, and optimize? (2025). DOI: 10.48550/arXiv.2507.11423
157. Jamali, M., Williams, Z. M. & Cai, J. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. (2023). DOI: 10.48550/arXiv.2309.01660
158. Wu, Y. *et al.* How large language models encode theory-of-mind: A study on sparse parameter patterns. *npj Artificial Intelligence* **1,** 20 (2025). DOI: 10.1038/s44387-025-00031-9
159. Amirizaniani, M. Mind over machine: Evaluating theory of mind reasoning in LLMs and humans. in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* 1068–1070 (Association for Computing Machinery, 2025). DOI: 10.1145/3701551.3707417
160. Casu, M., Triscari, S., Battiato, S., Guarnera, L. & Caponnetto, P. AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Applied Sciences* **14,** 5889 (2024). DOI: 10.3390/app14135889
161. Farzan, M., Ebrahimi, H., Pourali, M. & Sabeti, F. Artificial intelligence-powered cognitive behavioral therapy chatbots, a systematic review. *Iranian Journal of Psychiatry* **20,** 102–110 (2025). DOI: 10.18502/ijps.v20i1.17395

162. Hua, Y. *et al.* Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry* **24,** 383–394 (2025). DOI: 10.1002/wps.21352

163. Rollwage, M. *et al.* A cognitive layer architecture to support large-language model performance in psychotherapy interactions. *Nature Medicine* 1–9 (2026). DOI: 10.1038/s41591-026-04278-w

164. Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H. & Samtani, S. Improving Workplace Well-being in modern organizations: A review of Large Language Model-based Mental Health Chatbots. *ACM Trans. Manage. Inf. Syst.* **16,** 3:1–3:26 (2025). DOI: 10.1145/3701041

165. Chu, M. D., Gerard, P., Pawar, K., Bickham, C. & Lerman, K. Illusions of intimacy: How emotional dynamics shape human-AI relationships. (2025). DOI: 10.48550/arXiv.2505.11649

166. Hong, Y., Choi, J., Kim, M. & Kim, B. Can LLMs and humans be friends? Uncovering factors affecting human-AI intimacy formation. (2025). DOI: 10.48550/arXiv.2505.24658

167. Jones, M., Griffioen, N., Neumayer, C. & Shklovski, I. Artificial intimacy: Exploring normativity and Personalization Through Fine-tuning LLM Chatbots. in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* 1–16 (Association for Computing Machinery, 2025). DOI: 10.1145/3706598.3713728

168. Folk, D., Heine, S. J. & Dunn, E. Individual differences in anthropomorphism help explain social connection to AI companions. *Scientific Reports* **15,** 36548 (2025). DOI: 10.1038/s41598-025-19212-2

169. Li, J. *et al.* AI-exhibited Personality Traits Can Shape Human Self-concept through conversations. (2026). DOI: 10.1145/3772318.3790654

170. Meng, J., Zhang, R., Qin, J., Lee, Y.-J. & Lee, Y.-C. AI-mediated social support: The prospect of human–AI collaboration. *Journal of Computer-Mediated Communication* **30,** zmaf013 (2025). DOI: 10.1093/jcmc/zmaf013

171. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R. & Yang, D. The rise of AI companions: How human-chatbot relationships influence well-being. (2025). DOI: 10.48550/arXiv.2506.12605

172. Adrian, O. ChatGPT: More than a million users show signs of mental health distress and mania each week, internal data suggest. *BMJ : British Medical Journal (Online)* **391,** NaN–NaN (2025). DOI: 10.1136/bmj.r2290

173. Gabriels, K. & Goffin, K. Therapy chatbots and emotional complexity: Do therapy chatbots really empathise? *Current Opinion in Psychology* **68,** 102263 (2026). DOI: 10.1016/j.copsyc.2025.102263

174. Perry, A. AI will never convey the essence of human empathy. *Nature Human Behaviour* **7,** 1808–1809 (2023). DOI: 10.1038/s41562-023-01675-w

175. Yirmiya, K. & Fonagy, P. Mentalizing without a mind: Psychotherapeutic potential of generative AI. *Journal of Medical Internet Research* **27,** e79156 (2025). DOI: 10.2196/79156

176. Cohn, M. *et al.* Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 1–15 (Association for Computing Machinery, 2024). DOI: 10.1145/3613905.3650818

177. Colombatto, C. & Fleming, S. M. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness* **2024,** niae013 (2024). DOI: 10.1093/nc/niae013

178. Inie, N., Druga, S., Zukerman, P. & Bender, E. M. From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* 2322–2347 (Association for Computing Machinery, 2024). DOI: 10.1145/3630106.3659040

179. Colombatto, C., Birch, J. & Fleming, S. M. The influence of mental state attributions on trust in large language models. *Communications Psychology* **3,** 84 (2025). DOI: 10.1038/s44271-025-00262-1

180. Yao, X. & Xi, Y. From assistants to digital beings: Exploring anthropomorphism, humanness perception, and AI anxiety in large-language-model chatbots. *Social Science Computer Review* (2025). DOI: 10.1177/08944393251354976

181. Schlesener, E. A., Ziolkowski, M., Wong, S. K., Westmoreland, B. & Babu, S. V. 'Am i understood?': How the interplay between embodiment and theory of Mind Behavior Affects LLM-based Conversational Agents on perceived trust, anthropomorphism, presence, usability, and user experience. *ACM Trans. Interact. Intell. Syst.* (2025). DOI: 10.1145/3774779

182. Gonzalez, C. & Heidari, H. A cognitive approach to human–AI complementarity in dynamic decision-making. *Nature Reviews Psychology* **4,** 808–822 (2025). DOI: 10.1038/s44159-025-00499-x

183. Cummins, J. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. (2025). DOI: 10.48550/arXiv.2509.13397

184. Eigner, E. & Händler, T. Determinants of LLM-assisted Decision-Making. (2024). DOI: 10.48550/arXiv.2402.17385

185. Gui, G. & Toubia, O. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *SSRN Electronic Journal* (2023). DOI: 10.2139/ssrn.4650172

186. Loya, M., Sinha, D. & Futrell, R. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. in *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds. Bouamor, H., Pino, J. & Bali, K.) 3711–3716 (Association for Computational Linguistics, 2023). DOI: 10.18653/v1/2023.findings-emnlp.241

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

188. Soubki, A. & Rambow, O. Machine theory of mind needs machine validation. in *Findings of the Association for Computational Linguistics: ACL 2025* (eds. Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T.) 18495–18505 (Association for Computational Linguistics, 2025). DOI: 10.18653/v1/2025.findings-acl.951

189. Firestone, C. Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences* **117,** 26562–26571 (2020). DOI: 10.1073/pnas.1905334117

190. Wulff, D. U. & Mata, R. Addressing longstanding challenges in cognitive science with language models. *Trends in Cognitive Sciences* (2026). DOI: 10.1016/j.tics.2026.06.016

191. Hinton, G. E. Will digital intelligence replace biological intelligence? (2024).

192. Bengio, Y. *et al.* International AI safety report 2026. (2026). DOI: 10.48550/arXiv.2602.21012

193. Bengio, Y. *et al.* Managing extreme AI risks amid rapid progress. *Science* **384,** 842–845 (2024). DOI: 10.1126/science.adn0117

194. Jiang, Y., Wu, H.-T., Mi, Q. & Zhu, L. Neurocomputations of strategic behavior: From iterated to novel interactions. *WIREs Cognitive Science* **13,** e1598 (2022). DOI: 10.1002/wcs.1598

195. Todasco, M. Going all-in on LLM accuracy: Fake prediction markets, real confidence signals. (2025). DOI: 10.17605/OSF.IO/DC24T

196. Wang, S. *et al.* When experimental economics meets large language models: Evidence-based Tactics. (2025). DOI: 10.48550/arXiv.2505.21371

197. Chen, D. L., Schonger, M. & Wickens, C. oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* **9,** 88–97 (2016). DOI: 10.1016/j.jbef.2015.12.001

198. Edossa, F. W., Gassen, J. & Maas, V. S. Using large language models to explore contextualization effects in economics-based accounting experiments. (2024). DOI: 10.2139/ssrn.4891763

199. Liu, Y. *et al.* Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. *Information Processing & Management* **61,** 103809 (2024). DOI: 10.1016/j.ipm.2024.103809

200. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* **41,** 1149–1160 (2009). DOI: 10.3758/BRM.41.4.1149

201. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 2013). DOI: 10.4324/9780203771587

202. Plummer, M. JAGS: A program for analysis of bayesian graphical models using gibbs sampling. in *Proceedings of the 3rd international workshop on distributed statistical computing* (eds. Hornik, K., Leisch, F. & Zeileis, A.) **124,** 1–10 (2003).

203. Vehtari, A., Gelman, A. & Gabry, J. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27,** 1413–1432 (2017). DOI: 10.1007/s11222-016-9696-4

204. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472 (1992). DOI: 10.1214/ss/1177011136

205. Camerer, C. & Hua Ho, T. Experience-weighted attraction learning in normal form games. *Econometrica* **67,** 827–874 (1999). DOI: 10.1111/1468-0262.00054

206. Ho, T. H., Camerer, C. F. & Chong, J.-K. Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory* **133,** 177–198 (2007). DOI: 10.1016/j.jet.2005.12.008

207. Daunizeau, J., Adam, V. & Rigoux, L. VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLOS Computational Biology* **10,** e1003441 (2014). DOI: 10.1371/journal.pcbi.1003441

208. Hosseini, M., Gordijn, B., Kaebnick, G. E. & Holmes, K. Disclosing generative AI use for writing assistance should be voluntary. *Research Ethics* **21,** 728–735 (2025). DOI: 10.1177/17470161251345499

209. Porsdam Mann, S. *et al.* Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nature Machine Intelligence* **6,** 1272–1274 (2024). DOI: 10.1038/s42256-024-00922-7