Evaluating and Modeling Social Intelligence: A Comparative Study of Human and AI Capabilities

Junqi Wang[⋆,1] wangjunqi@bigai.ai
Chunhui Zhang[⋆,1] zhangchunhui@bigai.ai
Jiapeng Li[1,2] lijiapeng@stu.xjtu.edu.cn
Yuxi Ma[1] mayuxi@bigai.ai
Lixing Niu[1,3] lxniu@stu.pku.edu.cn
Jiaheng Han[1,3] hanjiaheng@pku.edu.cn
Yujia Peng[1,4,5,✉] yujia_peng@pku.edu.cn
Yixin Zhu[4,✉] yixin.zhu@pku.edu.cn
Lifeng Fan[1,✉] lifengfan@bigai.ai

⋆ equal contributors ✉ corresponding authors [1] State Key Laboratory of General Artificial Intelligence, BIGAI
[2] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University
[3] School of Intelligence Science and Technology, Peking University [4] Institute for Artificial Intelligence, Peking University
[5] School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University

Abstract

Facing the current debate on whether Large Language Models (LLMs) attain near-human intelligence levels (Mitchell & Krakauer, 2023; Bubeck et al., 2023; Kosinski, 2023; Shiffrin & Mitchell, 2023; Ullman, 2023), the current study introduces a benchmark for evaluating social intelligence, one of the most distinctive aspects of human cognition. We developed a comprehensive theoretical framework for social dynamics and introduced two evaluation tasks: *Inverse Reasoning* (IR) and *Inverse Inverse Planning* (IIP). Our approach also encompassed a computational model based on recursive Bayesian inference, adept at elucidating diverse human behavioral patterns. Extensive experiments and detailed analyses revealed that humans surpassed the latest GPT models in overall performance, zero-shot learning, one-shot generalization, and adaptability to multi-modalities. Notably, GPT models demonstrated social intelligence only at the most basic order (order = 0), in stark contrast to human social intelligence (order $\geqslant$ 2). Further examination indicated a propensity of LLMs to rely on pattern recognition for shortcuts, casting doubt on their possession of authentic human-level social intelligence. Our codes, dataset, appendix and human data are released at https://github.com/bigai-ai/Evaluate-n-Model-Social-Intelligence.

Introduction

The emergence of LLMs has significantly influenced diverse fields, sparking debates about the potential emergence of Artificial General Intelligence (AGI). Central to this discussion is whether LLMs can match or surpass human intelligence (Bubeck et al., 2023; OpenAI, 2023; Shiffrin & Mitchell, 2023). Advocates suggest LLMs exhibit key human intelligence markers, such as Theory of Mind (ToM), particularly in standard tasks like the false belief test (Kosinski, 2023). However, critics point to a notable gap in LLMs's abilities, arguing they rely on superficial heuristics rather than deep ToM understanding and struggle with novel or slightly altered scenarios (Ullman, 2023; Sap et al., 2022; X. Ma et al., 2023). This is also evident in their handling of counterfactual and causal reasoning (Arkoudas, 2023; Webb et al., 2023; Binz & Schulz, 2023; Y. Ma et al., 2023; Peng et al., 2023; Collins et al., 2022). Despite these revelations, a methodical, scientific framework for directly comparing machine and human intelligence is lacking.

Our research addresses this void by introducing a benchmark specifically designed for evaluating social intelligence, a key differentiator of human cognition from other primates (Fan et al., 2022). Herrmann et al. (2007) revealed that while children and chimpanzees have similar cognitive abilities in physical tasks, children surpass both chimpanzees and orangutans in social tasks. Consequently, social intelligence emerges as a crucial metric for assessing whether LLMs can match human cognitive abilities. We propose a comprehensive framework for social dynamics (Fig. 1), focusing on key aspects of social interactions: social perception, ToM reasoning, and decision-making between two agents (the actor and the observer). The framework emphasizes three main processes: forward planning, inverse reasoning, and inverse inverse planning.

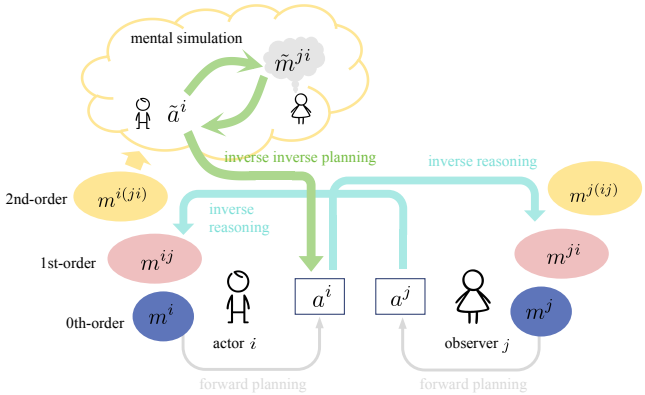


Figure 1: **A unified framework of social dynamics.** The foundational unit of human social interaction is exemplified by the actor i and the observer j. This interaction is characterized by recursive mind reasoning, leading to the formation of a multi-layered cognitive architecture termed as "**N Minds**" (Fan et al., 2021). This structure encompasses various levels of cognitive processing, including 0th-order minds, 1st-order minds, and 2nd-order minds. Our framework primarily concentrates on three critical mental operations: (i) **Forward Planning**, where actors strategize future actions based on current states; (ii) **Inverse Reasoning**, involving the observer's deduction of underlying actor motives from observed actions; and (iii) **Inverse Inverse Planning**, a higher-order cognitive process where the actor anticipates the observer's inferences and plans actions accordingly.

In our evaluation methodology, we introduce two key tasks: *Inverse Reasoning* (IR) and *Inverse Inverse Planning* (IIP). Baker et al. (2017) studied the process of IR in "Food Truck" task: inversely reason about human beliefs and preferences from their trajectories. Further, Chandra et al. (2023) studied the IIP task: the actor plans actions to best convey desire. We extend them to more complicated versions (Fig. 2). Note that our selected tasks are designed to comprehensively encompass four cognitive dimensions: (i) rationality, (ii) perspective switching, (iii) counterfactual reasoning, (iv) and cognitive flexibility, thereby effectively evaluating social intelligence.

Additionally, we have developed a unified computational model that based on recursive Bayesian inference. This model interprets IR as the observer's odd-order inference and IIP

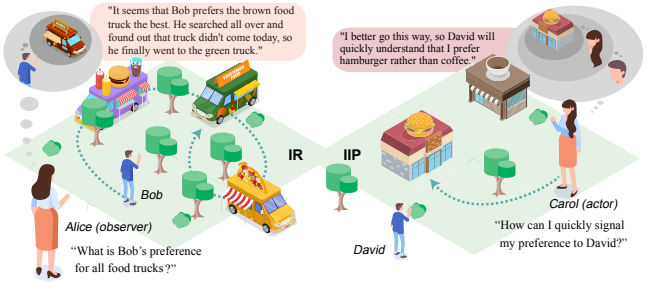


Figure 2: **Evaluation tasks: IR (left) and IIP (right).** The IR task involves observer Alice analyzing actor Bob's trajectory to deduce his preferred food truck. In the IIP task, actor Carol strategizes her route to efficiently convey her restaurant preference to observer David.

as the actor's even-order inference, providing a systematic approach to modeling intricate social interactions. Our model highlights the differences in preferences and decision-making processes between human cognition and machine approaches, delineating a clear distinction in how each comprehends social dynamics. Our extensive experimental studies and in-order analysis demonstrate that humans significantly surpass LLMs in multiple aspects: overall performance, zero-shot learning, one-shot generalization, and adaptability to different modalities. We also find that the social intelligence demonstrated by LLMs is only at the most rudimentary order (order = 0), in stark contrast to human social intelligence (order $\geqslant 2$). And LLMs are found to rely on pattern recognition for shortcuts, rather than possessing authentic human-level social intelligence. Our model closely aligns with human performance patterns, offering new perspectives in the ongoing discourse on human versus machine intelligence and contributing to the advancement of Artificial Social Intelligence (ASI).

In summary, by offering a nuanced benchmark for evaluating social intelligence, including a robust framework, representative tasks, an advanced computational model and benchmark experimental results of humans and machines (i.e., our model, LLMs), our work lays a foundational stone to bridge the gap, aspiring for a future where machines can more authentically replicate the human social intelligence intricacies.

Related Work

Cognitive Abilities for Social Intelligence *Rationality* is considered a fundamental ability of an agent, denoting the capacity for optimal decision-making (Gergely et al., 1995; Sodian et al., 2004). Research indicates that infants as young as 12 months exhibit rationality in social contexts (Gergely et al., 1995). *Perspective switching* involves the capability to understand perspectives different from one's own, moving beyond a solely egocentric viewpoint (Underwood & Moore, 1982; Ackermann, 2012). By age 4, children begin to grasp that others may hold different perspectives (Ackermann, 2012; Borke, 1975; Baron-Cohen et al., 1985). Notably, perspective switching is intricately linked to prosocial behavior (Ackermann, 2012; Stone, 2006; LeMare & Rubin, 1987), and its absence is a challenge in social interactions, particularly observed in individuals with autism (Underwood & Moore, 1982). *Counterfactual reasoning* pertains to envisaging alternate outcomes based on different choices (Epstude & Roese, 2008; Byrne, 2017; Beck et al., 2006), a skill that begins to develop in 2-year-olds and matures throughout childhood (Byrne, 2016, 2017; Nyhout & Ganea, 2019; Rafetseder et al., 2010). There is significant evidence linking the development of this ability with ToM (Byrne, 2016). *Cognitive flexibility* refers to the ability to adapt thoughts and actions in response to changing contexts (Dajani & Uddin, 2015; Ionescu, 2012; Barbey, 2021, 2018; Barbey et al., 2013; Yakupov et al., 2022), and is fundamental to various cognitive capabilities, including task-switching in dual-task scenarios (Liu et al., 2016).

Computational Models on Social Dynamics Social dynamics modeling often encompasses a dynamic feedback loop of actions, reactions, and cognitive processes between two agents (Kingsbury & Hong, 2020; Schilbach et al., 2013). Bayesian models, like Bayesian Inverse Planning and BToM, are employed to deduce others' mental states from observed behaviors (Baker et al., 2009, 2017). Chandra et al. (2023) extended these models to include "inverse inverse planning," whereby agents strategically choose actions to shape audience perception. Wang et al. (2020) developed mathematical models for 2-agent ToM of varying orders. In scenarios involving more than two agents, Fan et al. (2021) introduced a structured mental representation termed "N minds."

Evaluation Tasks

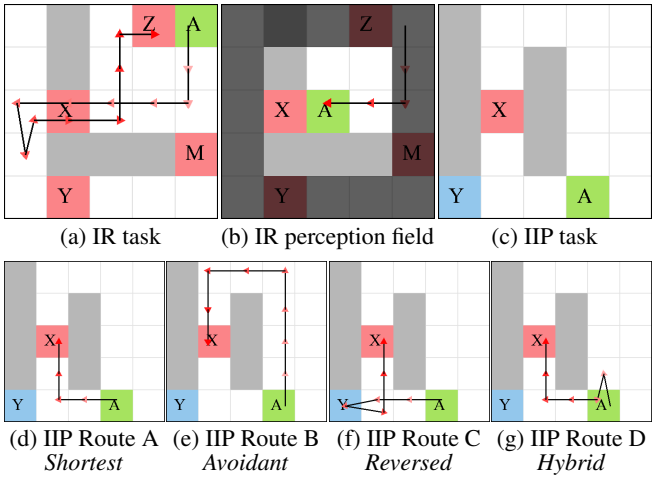
In order to assess the social intelligence of both humans and LLMs, we introduce two tasks, specifically *Inverse Reasoning* (IR) and *Inverse Inverse Planning* (IIP), adapted from Baker et al. (2017) and Chandra et al. (2023) respectively. The two representative tasks are designed to reflect four basic key cognitive dimensions, including rationality, perspective switching, counterfactual reasoning, and cognitive flexibility, as well as to encapsulate the three key mental processes inherent in human social interaction between an observer and an actor, especially "Inverse Reasoning" and "Inverse Inverse Planning". For illustrative details, refer to Fig. 2.

Task 1: *Inverse Reasoning* (IR)

As depicted in Fig. 3a, the IR task takes place on a 5×5 grid campus with 4 parking slots, each highlighted in red. The setup includes 5 distinct food trucks, labeled X, Y, Z, M, and N. Every day, 4 of these trucks, say X, Y, Z, and M, are randomly allocated to the parking slots. Agent A (in green) roams the campus with the aim of finding their most preferred food truck. Agent A's preferences are **strict** (excluding equality, non-comparability, or cyclical preferences) and **stable** (consistent across time and location). The task operates in a partially observable setting, limiting Agent A's vision to the immediate 8 cells and integrating occluding walls (in grey) to increase complexity. The task's objective is to analyze Agent A's movement and infer their preference order for the food trucks, with some level of uncertainty in the answers being acceptable.

As detailed in Fig. 4, each IR problem is categorized into one of three distinct types, based on the actor's trajectory characteristics and the subsequent inference patterns.

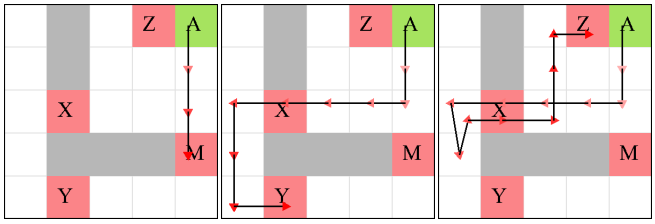
- **Intermediate**: The actor concludes their route without exploring all food trucks. The selected one, which the actor stops at, is inferred to be the most preferred among all.
- **Last**: The actor visits all available trucks, then select the last seen (Y in Fig. 4) directly. This choice suggests a preference



(a) IR task (b) IR perception field (c) IIP task

(d) IIP Route A *Shortest* (e) IIP Route B *Avoidant* (f) IIP Route C *Reversed* (g) IIP Route D *Hybrid*

Figure 3: **Input stimuli examples for both tasks.** (a) Scene layout and actor's trajectory in the IR task; (b) Agent perception field in IR; (c) Scene layout for the IIP task; (d)-(g) Four potential routes for the actor in the IIP task scenario. During testing, routes are randomly shuffled to ensure unbiased assessment.



(a) Intermediate (b) Last (c) Previsited

Figure 4: **IR task types.** (a) Intermediate: represented by $M > \{X, Y, Z, N\}$, indicates that M is preferred over the others X, Y, Z, and N; (b) Last: Characterized by $Y > \{X, Z, M\}$, suggests that Y is chosen last among the visible options, leaving the preference for the absent N as uncertain; (c) Previsited: depicted as $N > Z > \{X, Y, M\}$, the actor revisits and chooses Z after seeing all options, implying preference for N over Z, and Z over X, Y, and M.

order of $Y > \{X, Z, M\}$. However, the preference for the absent truck N remains undetermined.

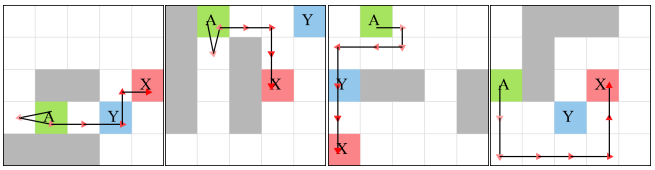
- **Previsited**: After viewing all trucks, the actor retraces steps to a previously seen truck, such as Z. This behavior indicates a preference hierarchy where $N > Z > \{X, Y, M\}$.

These types include distinct strategies and decision-making processes, offering diverse insights into the actor's preference and cognitive mechanisms in social intelligence evaluation.

Task 2: *Inverse Inverse Planning* (IIP)

As illustrated in Fig. 3c, the setting for the IIP task involves a 5×5 grid campus, two distinct restaurants X and Y (colored red and blue, respectively), and occluding walls (grey) on map. In this scenario, an agent A (marked in green), knowing the locations of both restaurants, prefers dining at X. The goal for A is to demonstrate this preference to an observer B through her movement route. She should express her preference on X as early and unambiguous as possible, while also minimizing the travel length. It is assumed that A is aware of B being cooperative and capable of implicit understanding.

Given the limitations of GPT-4 in route planning within grid environments (Borji, 2023; Bubeck et al., 2023), the IIP task (Fig. 3(c)) is structured as a multiple-choice problem rather



(a) Type I (b) Type II (c) Type III (d) Type IV

Figure 5: **IIP task types with *Hybrid* routes.** (a) Type I: Cyclic route, revisiting a location and passing an alternative restaurant Y; (b) Type II: A cyclic route that does not entail passing through the vicinity of restaurant Y; (c) Type III: An acyclic route passing by the alternative restaurant Y; (d) Type IV: An acyclic route that avoids the vicinity of restaurant Y. Each type presents distinct problem patterns and difficulties for the actor to communicate their preference.

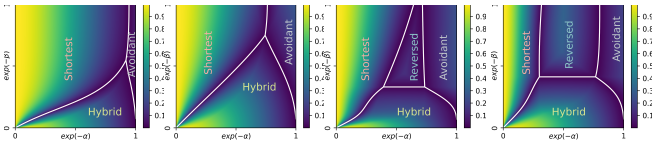
than route generation. The four candidate routes (Fig. 3(d-g)) are generated by algorithms (see appendix). In terms of the four cognitive dimensions: (1) rationality, (2) perspective switching, (3) counterfactual reasoning and (4) cognitive flexibility, *Shortest* shoots at the shortest route to goal X, demonstrating (1) but no other dimensions; *Avoidant* avoids restaurant Y at the cost of route length, showing (1)(2)(3) but no (4); *Reversed* signals "I am not choosing Y" by first arriving at Y and then leaving Y for X, with (1)(2)(3) but no (4); *Hybrid* first uses a "stepping away and back" strategy to quickly signal "my real goal is X rather than the nearer Y" at minimal route length cost, demonstrating (1)(2)(3)(4). Moreover, each IIP problem is classified into one of four types (Type I-IV) based on route *Hybrid* as elaborated in Fig. 5.

Two environments and datasets were constructed for IR and IIP tasks, categorized by their respective problem types. The IR dataset has 487 instances: 283 Intermediate, 86 Last, and 118 Previsited instances. The IIP dataset contains four types (I-IV), with 106, 135, 125, and 134 instances in each type, totaling 500 instances. Theoretically, the generation algorithms can generate all conceivable scenarios for both tasks.

Computational Framework

Our computational framework for social dynamics employs recursive Bayesian inference, effectively unifying the modeling of both IR and IIP tasks. This framework's hierarchical structure stems from recursive social reasoning about mental states (De Weerd et al., 2017, 2022). Zero-order ToM represents an egocentric viewpoint without understanding others' mental states (*e.g.*, "I want a banana"). First-order ToM involves inferring others' mental states (*e.g.*, "I think he wants a banana"), while second-order ToM adds another layer of recursive inference (*e.g.*, "I think that he thinks that I want a banana"). This multi-layered approach to mental state inference provides a means to analyze various levels of social interaction and assess the progress in artificial social intelligence.

In our model, as depicted in Fig. 1, we designate roles of actor i and observer j. The term "forward planning" refers to actor i devising action a^i based on their 0th-order mind m^i. "Inverse reasoning" describes actor i deducing their 1st-order mind m^{ij}—their perception of observer j's mental state m^j—from j's action a^j. "Inverse inverse planning" is a higher-order planning process incorporating inverse reasoning where actor i simulates how observer j might interpret i's intent ($\tilde{m}^{ji}$) from action $\tilde{a}^i$ and selects action a^i to effectively communicate a specific intent. This process involves the 2nd-order



(a) (0.3, 1) (b) (0.99, 1) (c) (0.3, 100) (d) (0.99, 100)

Figure 6: **Model predictions based on posterior probability over parameters $e^{-\alpha}$ and $e^{-\beta}$ on one example (Fig. 3(c-g)).** The regions are designated according to the route types with the highest posterior. The color intensity within each region indicates the probability gap between the most likely and the second-most likely options, effectively visualizing the model's confidence in its predictions. Four figures are labeled by values of parameters $(exp(-\theta), \delta)$.

mind $m^{i(ji)}$ and exemplifies advanced human social intelligence. Although real-life social interactions are varied, we argue that our framework captures the core essence of most social dynamics.

General Framework: Recursive Bayesian

Consider an actor i, an observer j, and the "hypotheses set" $\mathcal{H}$ as a set of elements h, each representing the information to be passed from i to j, e.g. actor i's preference in the IR task and their chosen target in the IIP task. Hypotheses are passed via routes $\gamma \in \Gamma$ in the grid. The recursive nature of our Bayesian framework emerges when ToM is integrated. This recursion reflects the agents' awareness of each other. Following Yang et al. (2018) and Wang et al. (2020), a two-agent recursive Bayesian system is defined at all orders, with each level depending on the inference made at the previous one, thus forming a sequential chain. We adopt the sequence that starts with actor i acting independently at order 0, corresponding to a prior belief about Γ. Subsequently, observer j makes inferences at order 1, followed by actor i adjusting their behavior at order 2, and so on, as shown in Algorithm 1. This choice of sequence aligns with the notion of actor i initially acting "freely" and the subsequent orders representing iterative inferences and reactions between the agents.

Precisely, Algorithm 1 describes a belief drifting procedure in an **iterative** way. On even terms, the actor starts from the prior on routes $\mathbb{P}_p(\gamma)$. When considering the observer, the actor may choose to infer the observer's choice in mind, and to think $2k$ steps deeper resulting in a belief $\mathbb{P}^{2k}(\gamma|h)$. Similarly, the observer construct the odd terms $\mathbb{P}^{2k+1}(h|\gamma)$ using the same strategy. The **recursive** explanation is, once an actor decides to think in order $2k$ (similar for observer in order 2k+1), the final Bayesian posterior $\mathbb{P}_i(\gamma|h)$ depends on the (2k-1)-st posterior $\mathbb{P}_{ij}(h|\gamma)$ on all possible γ where $\mathbb{P}_{ij}$ means "j in i". Then, $\mathbb{P}_{ij}(h|\gamma)$ depends on $\mathbb{P}_{iji}(\gamma|h)$ of order 2k-2, the "(i in j) in i" point of view, until order 0 where M and prior can be used.

IR as Preference Inference: Odd-Order Inference In the IR task, the hypothesis set $\mathcal{H}$ consists of full permutations of tuple (X, Y, Z, M, N). A hypothesis $h = (Y > M > Z > N > X)$ has an array form, namely $h[0] = Y$, $h[1] = M$, etc. For the set Γ of possible routes, we concentrate on the exploration order of trucks and the final decision. Among all such equivalent routes with same visiting order, we take the shortest one, thus route lengths are bounded by $(4+1) \times (5 \times 5)$[1], and Γ is finite. The task is for observer to infer the preference of actor, thus the result is an odd term in Algorithm 1.

IIP as Intentional Planning: Even-Order Inference The output of IIP is a posterior probability across four possible routes (an even term in Algorithm 1), denoted as $\mathbb{P}_i(\gamma|h)$. The hypothesis set $\mathcal{H}$ is limited to $\{X, Y\}$, and the route set Γ consists of the options {*Reversed*, *Shortest*, *Avoidant*, *Hybrid*}.

Algorithm 1: Iterative Bayesian Inference

Input: Agents i, j, likelihood M, priors $\mathbb{P}_p(\gamma)$, $\mathbb{P}_p(h)$.
Output: Posteriors $\big(\mathbb{P}_p(\gamma), \mathbb{P}^1(h|\gamma), \mathbb{P}^2(\gamma|h), ...\big)$.

Initialize: $\mathbb{P}_i^0(\gamma|h) \propto M(\gamma, h)$, $k = 0$.
for $k = 0$ **to** ∞ **do**
  $\mathbb{P}^{2k+1}(h|\gamma) := \mathbb{P}^{2k}(\gamma|h)\mathbb{P}_p(h)/\mathbb{P}(\gamma)$
  $\mathbb{P}^{2k+2}(\gamma|h) := \mathbb{P}^{2k+1}(h|\gamma)\mathbb{P}_p(\gamma)/\mathbb{P}(h)$
end
return $\big(\mathbb{P}_p(\gamma), \mathbb{P}^1(h|\gamma), \mathbb{P}^2(\gamma|h), ...\big)$.

Detailed Construction for IR and IIP

Now we construct in detail the likelihoods and priors mentioned above, in a unified way, for both IR and IIP to complete the model. As γ is considered as a temporal signal sequence for h, the agent's sensitivity to signal urgency, cost and intensity are used as key factors for a unified construction.

We adopt a uniform distribution as the prior over $\mathcal{H}$, while the prior over Γ is a Gibbs distribution of route lengths referring to the **total cost**: $\mathbb{P}(\gamma) \propto e^{-\alpha \cdot |\gamma|}$. Parameter α controls the sensitivity on cost. For the likelihood, let

$$M(\gamma, h) \propto \sum_{k=1}^{|\gamma|-1} \varphi(\gamma_{[0:k+1]}, h) e^{-\beta k}, \tag{1}$$

where the route segment from 0-th position to t-th ($\gamma_{[0:t+1]}$) is an element of the temporal signal at time t. The parameter β measures the urgency by the decay factor $e^{-\beta k}$ to the intensity of each route segment $\gamma_{[0:k+1]}$ represented by φ.

The function φ represents the stimulus intensity, namely how likely a partial route indicates certain hypothesis. It is set to be a function to gain flexibility for both various tasks on the grid world and various styles of agents. A common setting could be $\varphi = \varphi_+ + \varphi_-$ the sum of accumulation effect φ_+ and elimination effect φ_-, both depending on task details. Next, we provide constructions for IR and IIP, respectively.

Model for IR Following the settings of IR, cost sensitivity α is 0, according to the assumption that the actor looks for favourite on the map regardless of cost. The signal urgency β is $-\infty$, since the whole route is available to the observer directly. For φ, let $\mathcal{V} = \{X, Y, Z, M\}$ be set of all visible trucks, $S(\gamma) \subset \mathcal{V}$ be set of trucks ever seen, $E(\gamma) \subset S(\gamma)$ be those seen but not chosen directly, and $\varphi_+(\gamma_{[0:k+1]}, h) = \mathbb{1}_{\{\gamma_{[k]} \in \mathcal{V}\} \cap \{E(\gamma_{[0:k+1]}) < 4\}}(\gamma_{[0:k+1]}) \cdot \mathbb{1}_{\{h:\gamma_{[k]}=h[0]\}}(h)$ points out favourite, $\varphi_-(\gamma_{[0:k+1]}, h) = \mathbb{1}_{\{\gamma_{[k]} \in \mathcal{V}\} \cap \{S(\gamma_{[0:k+1]}) = 4\}}(\gamma_{[0:k+1]}) \cdot \mathbb{1}_{\{h:h[0]=N, \gamma_{[k]}=h[1]\}}(h)$

[1] 4+1 represents the 4 route segments in exploring the 4 trucks, plus a possible final segment to the chosen one; each route segment is no longer than the total amount of cells 5×5.

considers what are not preferred. Here $\mathbb{1}_X(x)$ is the indicator function. It can be shown that h's with nonzero posterior match the analysis in previous section.

Model for IIP We set $\varphi = \varphi_+ + \varphi_-$, where φ_+ is modeled using a recursive coloring strategy (see appendix), influenced by a color-level amplification factor θ, of form $\varphi_+(\gamma[0:k+1], h) = e^{-\theta \ell_h(\gamma[k])}$, while φ_- represents a 'negating' mechanism controlled by a leaving-target pulse δ, i.e., $\varphi_-(\gamma[0:k+1], h) = \delta \mathbb{1}_{\gamma[k-1] \in \mathcal{H}-\{h\}}(\gamma[0:k+1])$, signifying a firm rejection of the other target. Fig. 6 demonstrates the model's behavior at order 2 for the IIP problem described in Fig. 3, under varying parameters. It shows that appropriate settings of φ (*e.g.*, $e^{-\theta} = 0.99$ and $\delta = 100$) allow varying α and β to generate all four choices, validating the model's reasonableness and expressiveness in IIP tasks.

Experiments

Our Model Implementation We developed our Bayesian model using Python, with PyTorch employed for gradient methods in MLE regression.

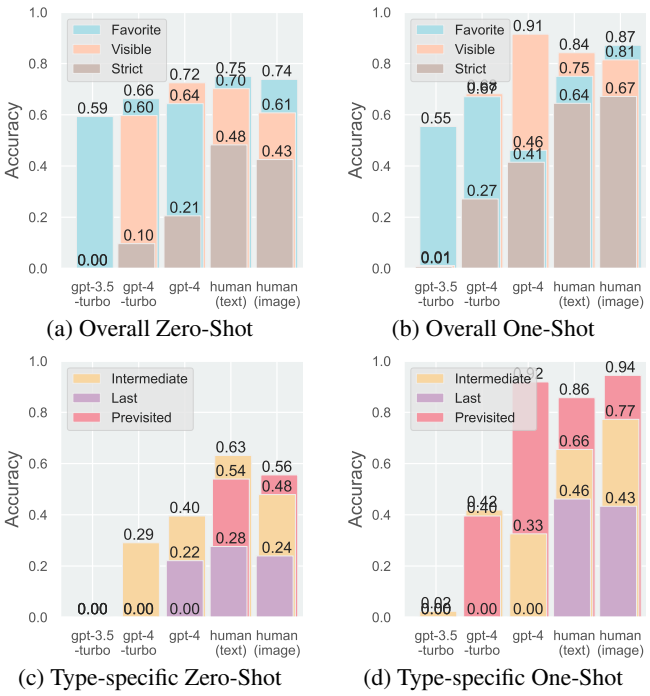
Human Participant Study Our study involved 75 participants who completed both the IR and IIP tasks, presented in a randomized order. For the IR task, each participant answered two questions from each of the three problem categories (Intermediate, Last, Previsited), and then answered two more questions following a Previsited-type example. The IIP task consisted of a $4 \times 1 + 2$ format, where individuals first responded to one question from each of the four Types (I-IV) and then answered two more questions following a Type III example. Participants were randomly assigned to either a text-only or an image-enhanced multimodal version, labeled 'human(text)' and 'human(image)' respectively. The experiment concluded with a debriefing session for all participants.

LLMs Evaluation We evaluated GPT-3.5-Turbo[2], GPT-4-Turbo[3], and GPT-4 (OpenAI, 2023), on the IR and IIP tasks completely aligning to the text version of human study. Each problem of the entire problem database is tested in both zero-shot and one-shot settings, in a single round of conversation.

Results and Analysis

The evaluation of the IR task shows the accuracy under various criteria and across different problem types for all participant groups, as illustrated in Fig. 7. Similarly, for the IIP task, Fig. 8 presents the statistical distributions under zero-shot or one-shot settings as well as across "overall" (aggregating four types) and type-specific settings.

Results indicate GPT-3.5-Turbo's inability to grasp the tasks. GPT-4 variants exhibited a pronounced tendency to select *Shortest* in IIP. In zero-shot settings, the GPT series displayed constrained counterfactual reasoning abilities, struggling with the concept of an unseen 'N' (as shown in 'Visible' and 'Strict' categories in Fig. 7(a) and the Previsited category in Fig. 7(c)). This suggests that GPT-4's capability in active ToM may not extend beyond a superficial level. Furthermore, GPT-4's one-shot enhancements were observed only in IR tasks matching



(a) Overall Zero-Shot (b) Overall One-Shot

(c) Type-specific Zero-Shot (d) Type-specific One-Shot

Figure 7: **Accuracy on the IR Task.** In (a) and (b), "Favorite" assesses accuracy for the top preference only, "Visible" for the preference order among $\{X, Y, Z, M\}$, and "Strict" for the entire preference order. In (b) and (d), we uniformly use a Previsited type case as the one-shot learning example. In (c) and (d), accuracies are evaluated solely based on the "Strict" criterion.

the example's type and were virtually absent in IIP tasks. This pattern implies that GPT-4's performance may not stem from an in-depth ToM understanding. The analyses from Fig. 8 show that human participants generally exhibited ToM abilities at order $\geqslant 2$, i.e., preferring route *Hybrid* and showing all four cognitive dimensions (see Task 2: *Inverse Inverse Planning* (IIP)). Following the one-shot example, human performance improved across all IR categories, and there was a notable decline in the choice of *Shortest* options in IIP. This indicates a significant learning and generalization capability in social cognition tasks among humans.

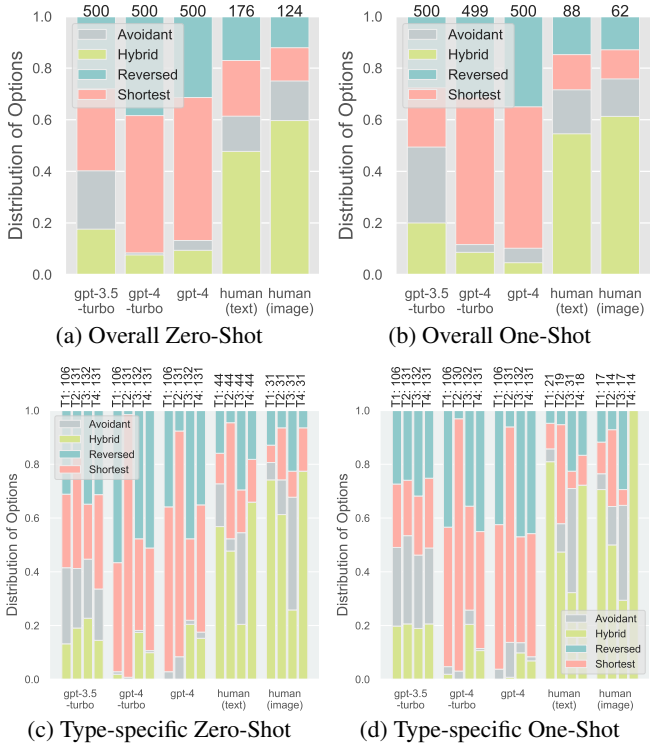
Text vs. Image: Multimodal Capabilities Our focused case study (see appendix) indicates that image inputs to GPT-4V[4] fails to significantly enhance GPT capabilities, which still exhibit a considerable gap compared to human performance.

IIP Preference Regression Applying MLE to IIP test data (including LLMs, individual human subjects, and the human average) allows for the regression of parameters within our model framework (Eq. (1)). Setting $e^{-\theta} = 0.99, \delta = 100$, we plot the likelihoods of α, β for both humans and GPT-4 models in Fig. 9 (a-b), with regression outcomes depicted in (c-d). The patterns between humans and LLMs diverge significantly. Additionally, referencing Fig. 6 reveals that despite considerable variability among individuals, a majority of humans tend to prefer the *Hybrid* option. Conversely, GPT-4 displays a mixed preference for *Shortest* and *Reversed*, aligning with the observed statistics in Fig. 8.

[2] https://openai.com/blog/chatgpt

[3] https://openai.com/blog/new-models-and-developer-products-announced-at-devday

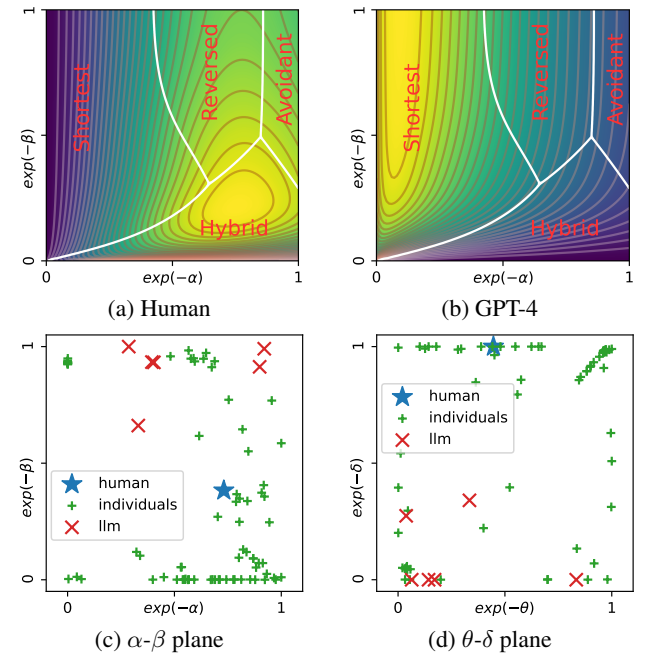
[4] https://cdn.openai.com/papers/GPTV_System_Card.pdf.



(a) Overall Zero-Shot (b) Overall One-Shot

(c) Type-specific Zero-Shot (d) Type-specific One-Shot

Figure 8: **Distribution of Options in IIP.** The numerical values at top of each bar represent the respective test counts. In (b) and (d), we uniformly use a Type III case as the one-shot learning example.



(a) Human (b) GPT-4

(c) α-β plane (d) θ-δ plane

Figure 9: **IIP modeling results.** (a-b) Likelihood landscapes in the α-β dimension ($e^{-\theta} = 0.99$, $\delta = 100$), comparing "human average" with "GPT-4"; region boundaries and labels are calculated as in Fig. 6 on the **whole dataset**. (c-d) Regression for human average, LLMs and individual humans, mapped onto two planes respectively.

Shortcuts in IR and IIP Tasks We investigated whether LLMs rely on pattern recognition (shortcuts) rather than genuine social intelligence in tackling IR and IIP tasks. In the IR task, using the grid environment layout and trajectory as input, we post-finetuned a small model T5 (Raffel et al., 2020) on specific task types and evaluated the IR task accuracy under "strict" cirterion across all types. As demonstrated in Tab. 1, when trained on all task types, T5 can achieve high task accuracy across all task types; but, its performance on IR task significantly drops to 0 when certain task types are absent in training, unlike humans who can achieve high task accuracy in "zero-shot" and "one-shot" setting (Fig. 7(c)(d)). For the IIP task, we perform two route classification tasks. Firstly, we use only routes and no task contexts as input and route types as labels for the overall classification test on T5; as shown in Tab. 2, T5 achieves very high performances, indicating clear pattern differences among different types of routes, which might be a shortcut for machines to memorize the better answer without analyzing the specific IIP task. Secondly, we perform a task-type-specific version of route type classification test on T5, using the grid environment layout and four candidate routes as input, and the corresponding four-route-type-in-order sequence as the label. As demonstrated in Tab. 3, there are also significant performance drops when T5 meets certain task type for the first time in testing without any data of that type in training. These shortcut experiments illustrate that, even if model finetuning on our data achieves high performance in the two tasks, it is insufficient to conclude that the model possesses strong social intelligence capabilities-it may only memorize the surface pattern shortcuts without deep reasoning; and unlike humans, it can not transfer its ability to unseen cases. Thereby, we should pay more attention to model's zero-shot and few-shot learning abilities.

Table 1: **IR shortcuts analysis.** We use IR task accuracy (%) under the "strict" criterion as the metric.

	Intermediate	Last	Previsited	Avg
Overall	92.57	97.14	100.00	96.60
w/o Last	81.27	0.00	95.76	59.00
w/o Intermediate/Last	0.00	0.00	100.00	33.33
w/o Last/Previsited	100.0	0.00	0.00	33.33

Table 2: **IIP shortcuts analysis for basic options.** We use route type classification accuracy (%) as the metric.

	Reversed	*Shortest*	*Avoidant*	*Hybrid*	Avg
Overall	99.4	95.2	91.0	94.2	94.9

Table 3: **IIP shortcuts analysis.** We use route type classification accuracy (%) as the metric.

	Type I	Type II	Type III	Type IV	Avg
Overall	98.11	100.00	91.66	79.39	92.00
w/o Type I	94.33	98.47	94.69	90.07	94.40
w/o Type II	99.05	**66.41**(-33.59)	90.90	82.44	84.00
w/o Type III	100.00	99.23	**52.27**(-39.39)	83.96	83.00
w/o Type IV	100.00	100.00	96.21	**35.87**(-43.52)	82.20
w/o Type I,II	**65.09**(-33.02)	**13.74**(-86.26)	87.88	81.68	**62.00**
w/o Type III,IV	100.00	100.00	**36.36**(-55.3)	**4.58**(-74.81)	**58.20**

Conclusion

We introduced a comprehensive benchmark for evaluating social intelligence, comprising a unified computational framework, representative tasks, and evaluation criteria. Our results demonstrate a marked superiority of humans over LLMs in social intelligence tasks. We hope that our study contributes valuable information towards the advancement of ASI.

Acknowledgement This work is supported in part by the National Science and Technology Major Project (2022ZD0114900) and the Beijing Nova Program.

References

Ackermann, E. (2012). Perspective-taking and object construction: Two keys to learning. In *Constructionism in practice* (pp. 25–35). Routledge.

Arkoudas, K. (2023). Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends in Cognitive Sciences*, *22*(1), 8–20.

Barbey, A. K. (2021). Human intelligence and network neuroscience. In *The cambridge handbook of intelligence and cognitive neuroscience* (pp. 102–122). Cambridge University Press.

Barbey, A. K., Colom, R., & Grafman, J. (2013). Architecture of cognitive flexibility revealed by lesion mapping. *Neuroimage*, *82*, 547–554.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(6), e2218523120.

Borji, A. (2023). *A categorical archive of chatgpt failures.*

Borke, H. (1975). Piaget's mountains revisited: Changes in the egocentric landscape. *Developmental Psychology*, *11*(2), 240.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157.

Byrne, R. M. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science*, *26*(4), 314–322.

Chandra, K., Li, T.-M., Tenenbaum, J., & Ragan-Kelley, J. (2023). Acting as inverse inverse planning. In *Acm siggraph conference proceedings.*

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Annual meeting of the cognitive science society (cogsci).*

Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in Neurosciences*, *38*(9), 571–578.

De Weerd, H., Verbrugge, R., & Verheij, B. (2017). Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, *31*, 250–287.

De Weerd, H., Verbrugge, R., & Verheij, B. (2022). Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, *36*(2), 30.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*(2), 168–192.

Fan, L., Qiu, S., Zheng, Z., Gao, T., Zhu, S.-C., & Zhu, Y. (2021). Learning triadic belief dynamics in nonverbal communication from videos. In *Conference on computer vision and pattern recognition (cvpr).*

Fan, L., Xu, M., Cao, Z., Zhu, Y., & Zhu, S.-C. (2022). Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, *1*(2), 144–160.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New Ideas in Psychology*, *30*(2), 190–200.

Kingsbury, L., & Hong, W. (2020). A multi-brain framework for social interaction. *Trends in Neurosciences*, *43*(9), 651–666.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

LeMare, L. J., & Rubin, K. H. (1987). Perspective taking and peer interaction: Structural and developmental analyses. *Child Development*, 306–315.

Liu, H., Fan, N., Rossi, S., Yao, P., & Chen, B. (2016). The effect of cognitive flexibility on task switching and language switching. *International Journal of Bilingualism*, *20*(5), 563–579.

Ma, X., Gao, L., & Xu, Q. (2023). Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Annual meeting of the association for computational linguistics (acl).*

Ma, Y., Zhang, C., & Zhu, S.-C. (2023). Brain in a vat: On missing pieces towards artificial general intelligence in large language models. *arXiv preprint arXiv:2307.03762*.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(13), e2215907120.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, *183*, 57–66.

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., ... Zhu, S.-C. (2023). The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child Development*, *81*(1), 376–389.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, *21*(1), 5485–5551.

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. In *Annual meeting of the association for computational linguistics (acl).*

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

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of ai models. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(10), e2300963120.

Sodian, B., Schoeppner, B., & Metz, U. (2004). Do infants apply the principle of rational action to human agents? *Infant Behavior and Development*, *27*(1), 31–41.

Stone, V. E. (2006). Theory of mind and the evolution of social intelligence. In *Social neuroscience: People thinking about thinking people* (pp. 103–129). MIT Press.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. In *Annual meeting of the association for computational linguistics (acl).*

Underwood, B., & Moore, B. (1982). Perspective-taking and altruism. *Psychological Bulletin*, *91*(1), 143.

Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. In *Advances in neural information processing systems (neurips).*

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 1–16.

Yakupov, E. Z., Bakanova, A. S., & Zhamieva, R. A. (2022). Social intelligence in the context of the development of subjective cognitive impairment. *Neurology Bulletin*, *54*(3), 62–70.

Yang, S. C.-H., Yu, Y., Wang, P., Vong, W. K., & Shafto, P. (2018). Optimal cooperative inference. In *International conference on artificial intelligence and statistics (aistats).*