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

COGSCI 2024 DYNAMICS OF COGNITION July 24-27 • Rotterdam

Junqi Wang*,[1] Chunhui Zhang*,[1] Jiapeng Li[1,2] Yuxi Ma[1] Lixing Niu[1,3]
Jiaheng Han[1,3] Yujia Peng[1,3,✉] Yixin Zhu[3,✉] Lifeng Fan[1,✉]

[1] Beijing Institute for General Artificial Intelligence, [2] Xi'an Jiaotong University, [3] Peking University

Motivation

- **Evaluating AI Social Intelligence:** Investigate whether large language models can match human-level social intelligence, a key differentiator of human cognition.
- **Benchmarking AI Performance:** Provide a systematic framework to evaluate and compare the social capabilities of AI systems against human performance.
- **Advancing AGI Development:** Identify the current limitations of AI in social intelligence to guide future research and development toward achieving Artificial General Intelligence (AGI).

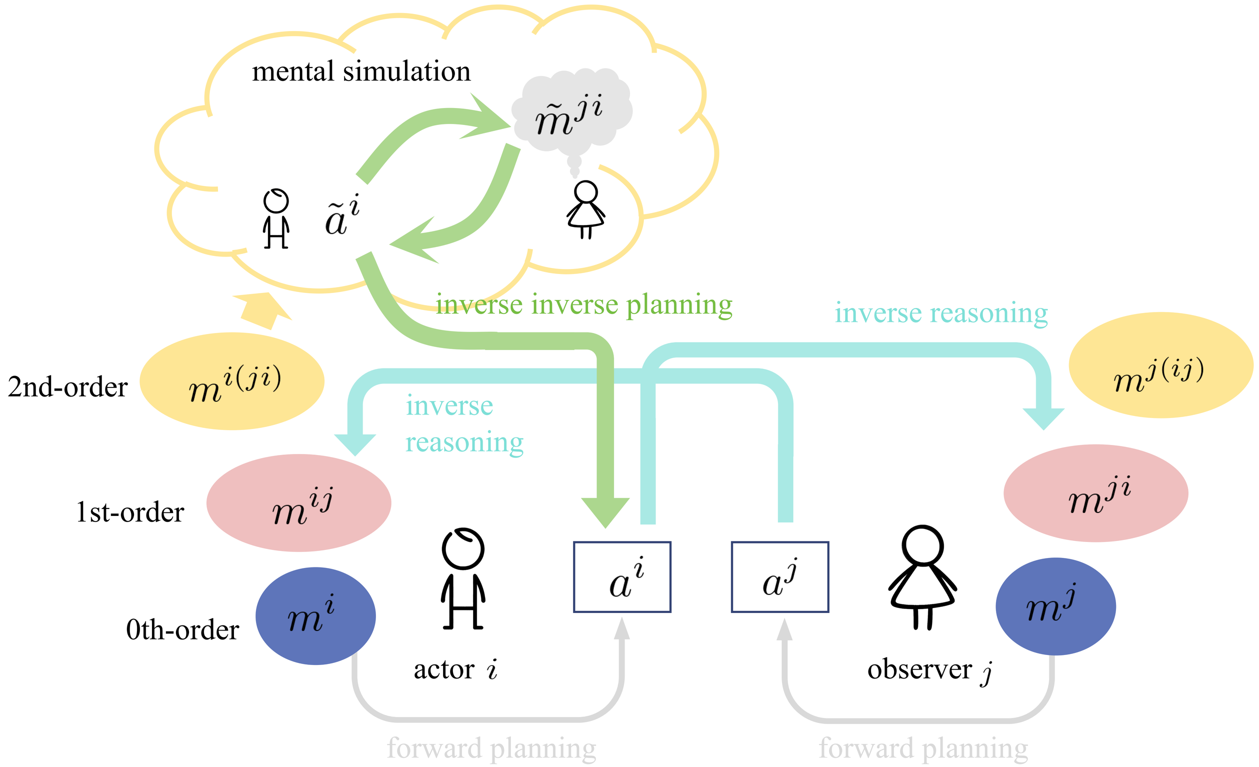


Figure 1. A unified framework of social dynamics.

Contribution

- **Novel Assessment Framework:** Introduce a comprehensive benchmark for evaluating social intelligence through tasks like Inverse Reasoning (IR) and Inverse Inverse Planning (IIP), assessing critical cognitive dimensions.
- **Recursivef Bayesian Inference Model:** Present a computational model capable of interpreting social interactions and capturing the nuances of human social reasoning through recursive Bayesian inference.
- **Empirical Insights:** Provide a detailed analysis comparing human participants, state-of-the-art language models, and the proposed computational model, highlighting the significant gap between AI and human social cognition.

Evaluation Tasks

Two basic tasks, **Inverse Reasoning (IR)** and **Inverse Inverse Planning (IIP)**, are desined to evaluate social intelligence. **IR** tasks requires to infer actor's *preference* on targets from observation, **IIP** asks to plan a path with considering observer's inference of actor's *destination*.

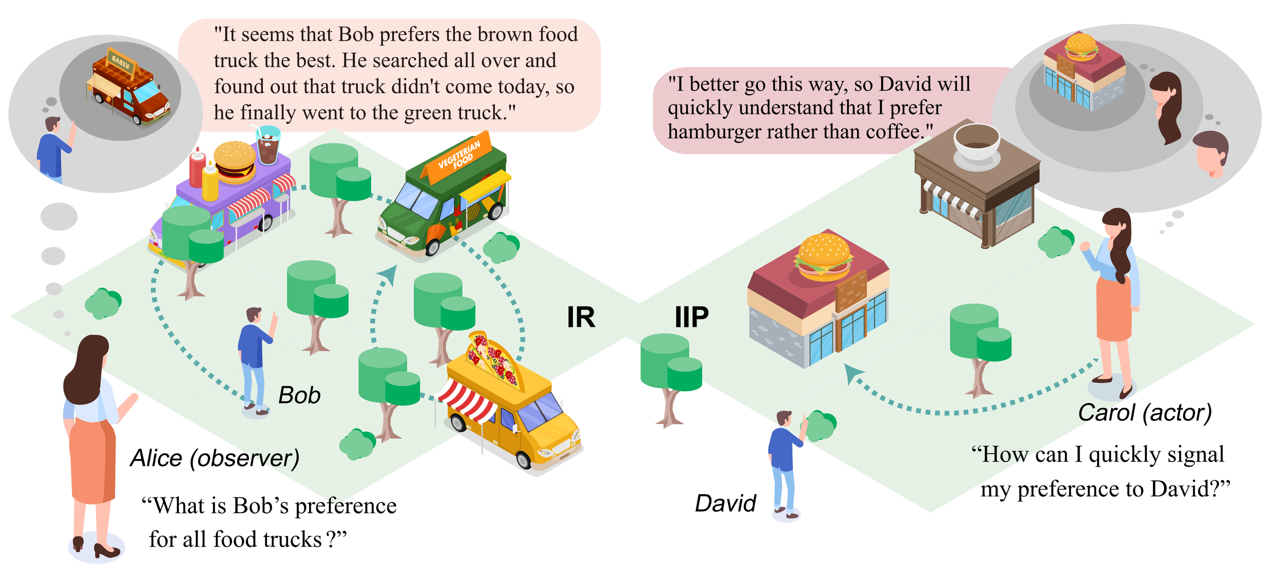
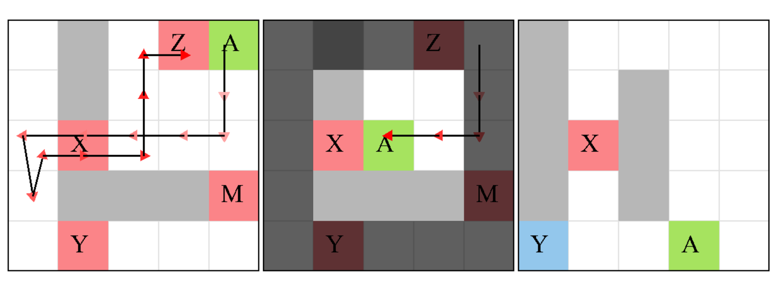
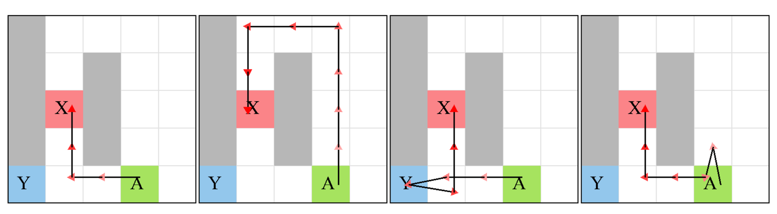


Figure 2. A realistic-style figure showing IR and IIP tasks.

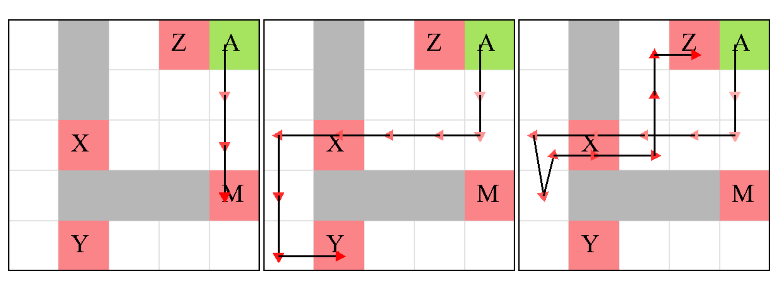


(a) IR task (b) IR perception (c) IIP task

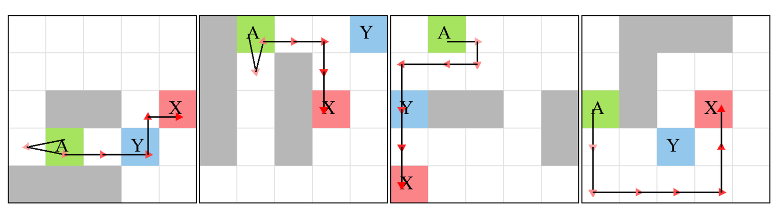


(d) IIP Rte A *Shortest* (e) IIP Rte B *Avoidant* (f) IIP Rte C *Reversed* (g) IIP Rte D *Hybrid*

Figure 3. Input stimuli examples for both tasks.



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



(d) Type I (e) Type II (f) Type III (g) Type IV

Figure 4. (a-c) IR task types. (d-g) IIP task types with *Hybrid* routes.

Computational Model

A general model for ToM is constructed using recursive Bayesian inference. Specific likelihood and priors are constructed for the two tasks.

Algorithm 1: Iterative Bayesian Inference

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

```
Initialize: P_i^0(γ|h) ∝ M(γ, h), k = 0.
for k = 0 to ∞ do
    P^{2k+1}(h|γ) := P^{2k}(γ|h) P_p(h) / P(γ)
    P^{2k+2}(γ|h) := P^{2k+1}(h|γ) P_p(γ) / P(h)
end
return (P_p(γ), P^1(h|γ), P^2(γ|h), ...).
```

- $h \in H$: hypothesis, *preference* in **IR**, and *destination* in **IIP**.
- $\gamma \in \Gamma$: a finite path set on the 5 by 5 grid.
- M: likelihood. Describing a "natural" statistical relation between γ and h.
- In our tasks, M is set to be

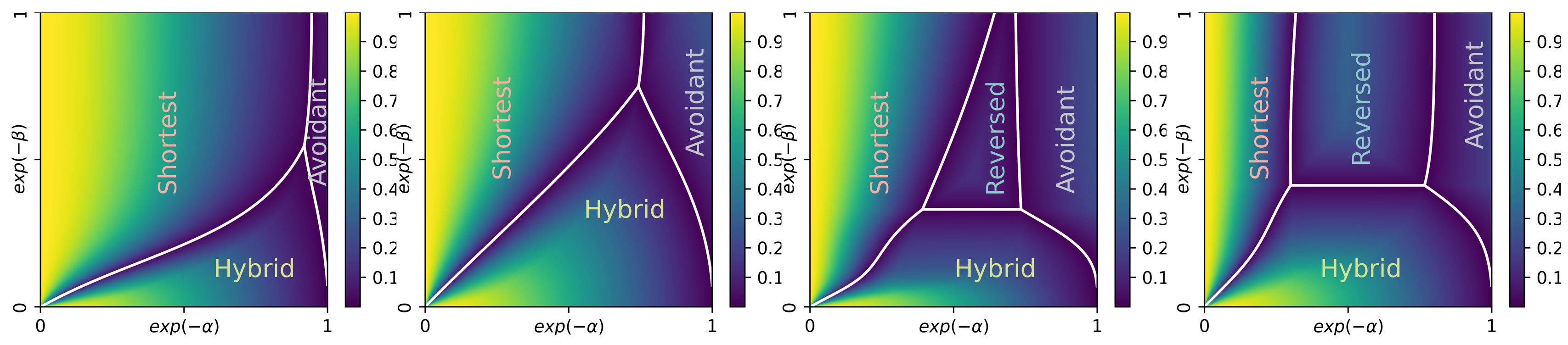
$$M(\gamma, h) \propto \sum_{k=1}^{|\gamma|-1} \varphi(\gamma_{[0:k+1]}, h) e^{-\beta k}, \quad (1)$$

where α, β, φ are numerical and functional parameters.

- $\mathbb{P}_p(\gamma)$, $\mathbb{P}_p(h)$, priors on paths / hypotheses.

Computational Model

Based on the construction, varying two parameters results in various behaviors.



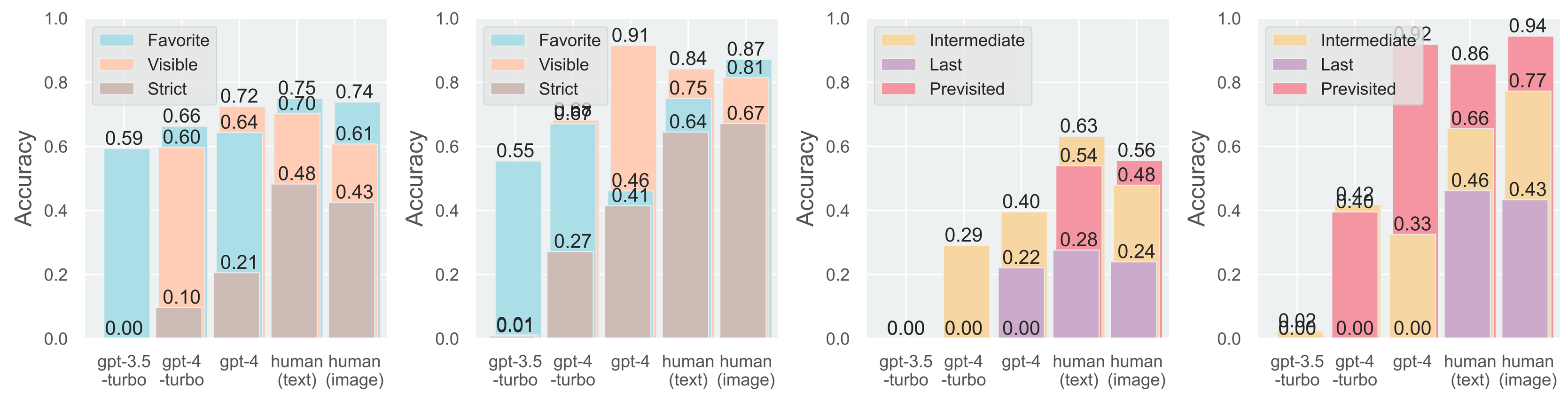
(a) (0.3, 1) (b) (0.99, 1) (c) (0.3, 100) (d) (0.99, 100)

Figure 5. **Model predictions based on posterior probability over parameters $e^{-\alpha}$ and $e^{-\beta}$ on one example (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 $(exp(-\theta), \delta)$.

Experiments

- **Subjects:** GPT 3.5 Turbo, GPT 4 Turbo, GPT 4V, and 75 human subjects.
- **Experiment types:**
  - **Zero-shot** vs. **one-shot** for IR and IIP
  - **Text** vs. **image** for IR and IIP
  - **Bayesian model regression** for IIP
  - **Shortcut analysis** for IR and IIP

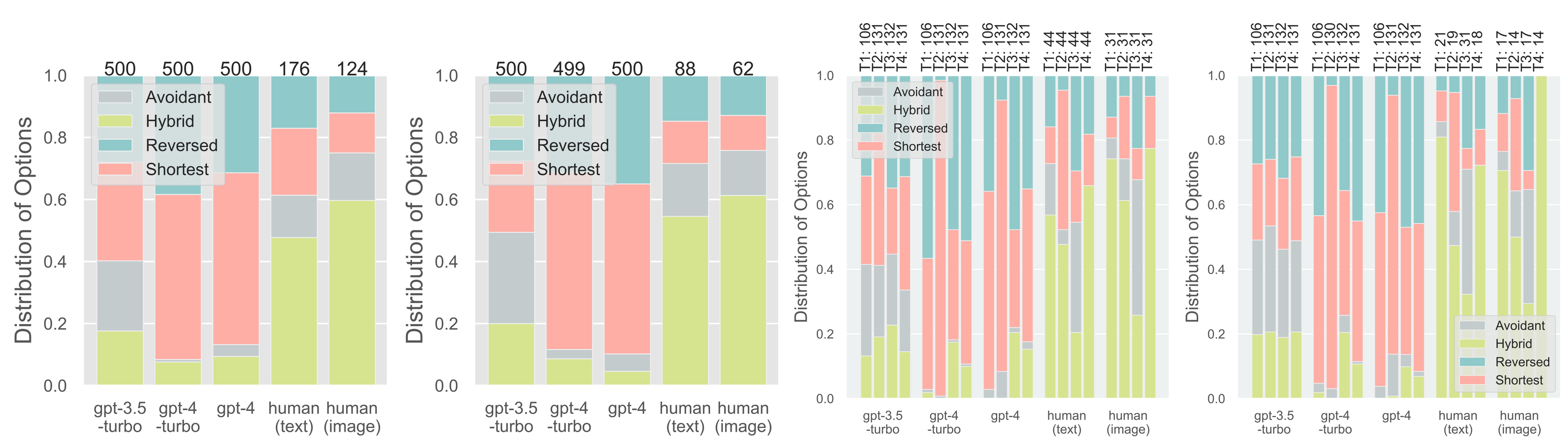
Result of Experiments on IR



(a) Overall Zero-Shot (b) Overall One-Shot (c) Type-specific Zero-Shot (d) Type-specific One-Shot

Figure 6. **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.

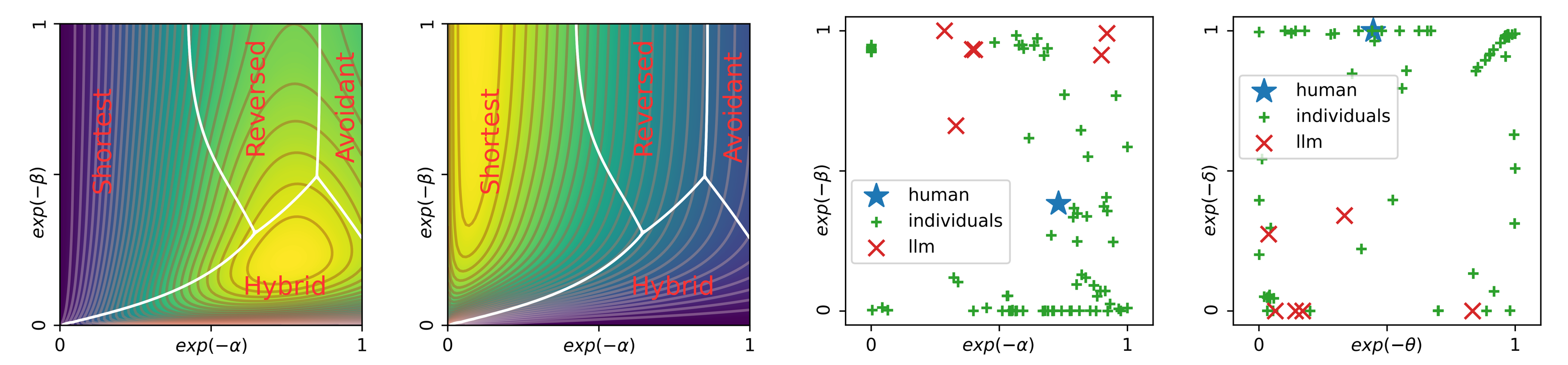
Result of Experiments on IIP



(a) Overall Zero-Shot (b) Overall One-Shot (c) Type-specific Zero-Shot (d) Type-specific One-Shot

Figure 7. **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.

Bayesian Model Likelihood and Regression



(a) Human (b) GPT-4 (c) α-β plane (d) θ-δ plane

Figure 8. **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 5 on the **whole dataset**. (c-d) Regression for human average, LLM and individual humans, mapped onto two planes respectively.

Shortcut Analysis

	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 1. **IR shortcuts analysis on IR accuracy.**

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

Table 2. **IIP path type classification accuracy.**

	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**

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

Conclusion

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