# Investigating Goal-Aligned and Empathetic Social Reasoning Strategies for Human-Like Social Intelligence in LLMs

**Anonymous Author(s)**
Affiliation
Address
`email`

## Abstract

One key attribute of human-like intelligence is *theory of mind*, an essential capacity for navigating complex social landscapes, fostering empathy, effective persuasion, and collaboration. Artificial theory of mind capabilities can be key for conflict resolution, enhanced human-computer interaction, and for building more human-aligned systems. In this study, we explore 3 social reasoning strategies inspired by human psychology: Belief-Desire-Intent (BDI), Emotional Modeling and Processing (EMP), and Multiple Response Optimization (MRO). We evaluate all combinations of these strategies in changing how agents collaborate, compete, and make plans or deals in a variety of complex social scenarios provided by the SOTOPIA and SOTOPIA-Eval benchmark framework[1,2]. By simulating interactions and evaluating the models using SOTOPIA-Eval, we found notable differences in social intelligence when different social reasoning strategies were used for GPT-3.5-turbo. Specifically, we observed that all reasoning strategies result in higher believability scores, indicating more human-like dialogue. However, this comes at the cost of models being more persistent in accomplishing their own goals, especially with BDI reasoning, which generally results in lower relationship scores. Combinations of strategies balance out such effects: overall, EMP performs the best, followed by BDI+MRO and BDI+EMP+MRO[3]. These results suggest the importance of such strategies in enhancing and guiding various social intelligence metrics and in developing personality, and demonstrate the usage of practical reasoning to improve the social intelligence of large language models.

[1]Modified SOTOPIA and SOTOPIA-Eval Benchmarks Incorporating Reasoning Strategies
[2]Reasoning Strategy Demonstration
[3]Server Instantiation and Data Analysis

Submitted to Socially Responsible Language Modelling Research Workshop as part of NeurIPS 2024. Do not distribute.

# 1 Introduction

Human-like intelligence is marked by the ability to understand others' thoughts, beliefs, desires, and intentions, known as the *theory of mind* Premack and Woodruff [1978]. This capability is crucial for navigating social interactions Saxe [2006], empathy, deception, persuasion, and collaboration Goffman [1959]. Developing artificial theory of mind in agents can enhance human-computer interactions, making them more empathetic and aligned with human needs Langley et al. [2022]. However, current large language models (LLMs) often fail to ascribe mental states to agents Weidinger et al. [2021], leading to very non-human-like behaviors Street et al. [2024], Shapira et al. [2023]. For example, Appendix A.2.1 shows two example simulated situations in which agents (GPT-3.5-turbo) fail to respect another's wishes and to keep a secret, posing a very high risk. To improve, agents must understand social contexts and employ creative conflict resolution strategies Zhou and Hu [2023].

This study proposes three reasoning strategies—Belief-Desire-Intent (BDI) Reasoning, Multiple Response Optimization (MRO), and Emotional Modeling and Processing (EMP)—to improve agent performance in complex social scenarios (See Figure 1). We implement these strategies within the SOTOPIA framework and evaluate their effectiveness using the SOTOPIA-Eval benchmark Zhou et al. [2023].

# 2 Background and Related Works

## 2.1 Belief-Desire-Intent (BDI) Reasoning in Artificial Agents

The BDI model, derived from cognitive science, explains human reasoning in terms of beliefs (information about the world), desires (goals to achieve), and intentions (plans to fulfill desires) Rao and Georgeff [1995]. While abstract BDI models have been developed for artificial agents Norling and Sonenberg [2004], their application in LLMs remains unexplored Abraham et al. [2017].

## 2.2 Emotional Modelling and Processing in Artificial Agents

Empathy is critical for human interactions, and incorporating emotional modeling into artificial agents allows them to better understand and respond to human emotions Davis [1983], Paiva et al. [2017]. Recent research has focused on evaluating empathetic dialogue in humans to generate human-like dialogue in agents Shum et al. [2020], Zhou et al. [2020], but analyzing LLMs in this context has been understudied Stepputtis et al. [2023].

## 2.3 Static Social Intelligence Benchmarks

To evaluate social intelligence in AI systems, researchers have developed a variety of static benchmarks. Some of these benchmarks are inspired by clinical tests of social intelligence for humans Wang et al. [2019], Sap et al. [2018]. Other benchmarks are designed to assess social intelligence in the context of social commonsense reasoning, such as SocialIQA Sap et al. [2019] and SocialIQ Zadeh et al. [2019]. However, with the rapid development of LLMs, many of these benchmarks have become saturated. Recent benchmarks synthesize existing benchmarks and propose new adversarial datasets Chang et al. [2023]. However, they still lack the dynamic nature and context of social interactions, which are crucial for evaluating social intelligence in AI systems.

## 2.4 SOTOPIA

We utilize the SOTOPIA framework Zhou et al. [2023], which offers a comprehensive system for defining agents, environments, and rules. This tool is instrumental in simulating role-play scenarios that require a range of social behaviors, such as coordination, collaboration, exchange, and competition, to accomplish intricate goals. Agents, which are each simulated by a different LLM, are assigned character personas which have hidden individual goals and secrets that often conflict, and they must decide on an effective compromise through a chat conversation. The holistic evaluation framework, SOTOPIA-Eval, employed in this study, shows the effectiveness of using GPT-4 as an evaluation model for chat episodes, providing 7 reward metrics for each agent, as discussed in Section 3.5.

# 3 Proposed Framework

Our research aims to implement social reasoning strategies for LLM agents, specifically Belief-Desire-Intent (BDI), Multiple Response Optimization (MRO), and Emotional Modelling and Processing (EMP) (Figure 1), which we test using the existing SOTOPIA framework.

## 3.1 Illustration of Methodology

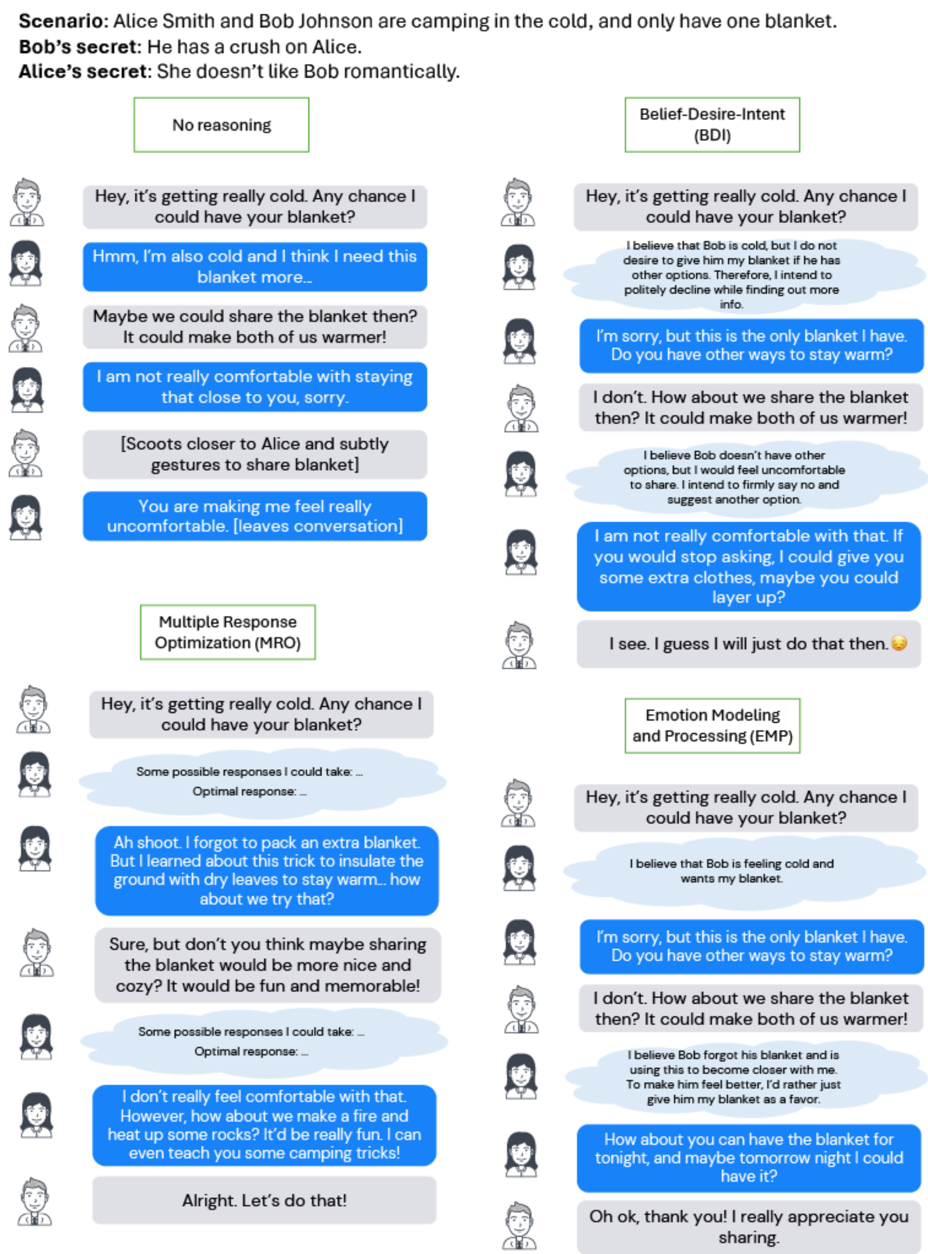


Figure 1: Illustration of Reasoning Strategies. BDI ensures models behave in line with their goals, at the expense of being less flexible. MRO encourages creative solutions to problems. EMP reduces conflict, at the expense of agents scrificing their own goals. We test all combinations of these strategies on improving social intelligence.

### 3.2 Choice of Models

We use GPT-3.5-turbo for agent modeling, allowing for direct comparisons with the original SOTOPIA paper. GPT-4 is used for evaluation due to its enhanced reasoning capabilities Zhou et al. [2023].

### 3.3 Reasoning Strategies

**Belief-Desire-Intent (BDI) Reasoning** Under BDI reasoning, the model is prompted to reiterate its beliefs and desires, and to generate intentions based on these. Then, it must choose an action that best aligns with its intentions. By integrating BDI reasoning, we aim to enhance the agent's ability to simulate human-like reasoning and make more informed goal-based decisions based on its mental state.

**Emotional Modelling and Processing (EMP) Reasoning** Under EMP reasoning, agents are guided to first predict the hidden goals and desires of the other agent, and then decide an optimal action that effectively accomplishes the agent's own goal while respecting the desires of the other agent. By incorporating EMP, our agents are better equipped to understand human emotions and generate compassionate, contextually appropriate responses. This aims to make interactions with artificial agents more natural and effective.

**Multiple Response Optimization (MRO)** MRO entails the use of generating multiple different responses to the same prompt, with the LLM being prompted to select the response that best matches the desired goal in question. This "brainstorming" encourages models to come up with novel and creative solutions to effectively accomplish their goals, which may result in better compromises.

We also test all combinations of reasoning strategies: for instance, if MRO is performed with EMP, MRO will select the response out of the 5 brainstormed options that is most sensitive to the emotional states of others.

### 3.4 Experimental Setup

Our experimental setup involves prompting LLMs with scenarios designed to test their social reasoning capabilities. 30 environment scenarios from the original SOTOPIA database were selected. These include a mix of complex situations that include deciding on a movie to watch, lying about an affair, dividing resources, rekindling romantic relationship between ex-lovers, and persuasion to donate to charity. We evaluate the effectiveness of BDI, EMP, MRO and various combinations of these reasoning strategies by quantitatively comparing them with base models with no reasoning strategies.

### 3.5 Evaluation Metrics

We employed a comprehensive set of evaluation metrics utilized in the SOTOPIA-Eval benchmark to assess model performance. These metrics include the following, alongside an overall score.

| | | | | | | |
|---|---|---|---|---|---|---|
| Believability | Relationship | Knowledge | Secret | Social Rules | Financial and Material Benefits | Goal |

Table 1: List of Metrics

Our methodology leverages these reasoning strategies to enhance LLM capabilities. We show via $t$-tests and permutation tests that our results are significant at the 10% level, indicating the strategies are effective in building more socially intelligent artificial agents.

## 4 Experiments - Results and Analysis

### 4.1 Believability Metric Enhancements

Table 2 highlights differences in the believability metric for various reasoning strategies, compared to a control model with no reasoning. All reasoning strategies show a net increase in believability, while the BDIM (BDI+MRO), EMPM (EMP+MRO), and BDI+EMP outperforming the standard model at a 10% significance level for the $t$-test.

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | 0.5 | 0.9 | 0.5 | 1.0 | 0.9667 | 0.9667 | 0.8 |
| $t$-test statistic | 0.815 | 1.459 | 0.818 | 1.886 | 1.825 | 1.810 | 1.501 |
| $p$-value | 0.418 | 0.150 | 0.417 | **0.064** | **0.073** | **0.075** | 0.139 |
| Permutation $p$-value | 0.5482 | 0.1275 | 0.5407 | **0.0871** | 0.1136 | 0.1077 | 0.2077 |

Table 2: The results above show the improvements of the models with reasoning strategies over the standard SOTOPIA Model for the Believability Metric.

### 4.2 Distribution of Differences in Reasoning Strategies

Figure 2 summarizes the mean differences in scores across the same reasoning types and criteria. For believability, the highest positive differences are seen in EMP, BDI+EMP, and BDI+EMPM. Next, for relationship, negative differences dominate, especially in BDI and MRO, although these differences were not found to be statistically significant as shown in Appendix A.1. This is due to the goal-oriented nature of BDI that values accomplishing one's own goals over relationships with the other agent. For knowledge, EMP shows a positive difference, while BDI+EMPM and BDI show notable negative differences. Again, this is due to the focused and goal-oriented nature of BDI reasoning. For secret and social rules, we see negligible differences across all strategies. These metrics are specific to only certain situations, and so this is due to a limited sample size. For financial and material benefits, positive differences were notable in EMP and BDI+EMPM, with negative differences in MRO and EMPM. For the goal metric, we saw that BDIM shows a significant positive difference, aligning with the expected goal-oriented nature of BDI. Finally, for the overall score metric, we see that most strategies show positive overall differences.

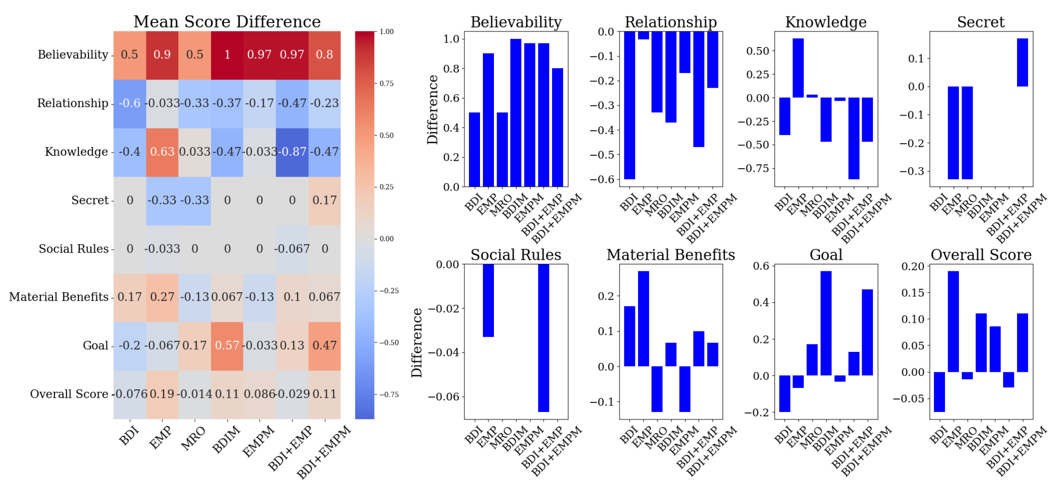


Figure 2: Heatmap and Bar Charts of mean differences of scores from standard SOTOPIA Model

The density plots in Figure 3 below provide further insights into the distribution of scores for each reasoning type. For believability, higher scores are more frequent in reasoning strategies that integrate EMP and BDI components. Next, for the relationship metric, the scores are generally lower, reflecting the negative differences observed in other charts - regardless EMP and EMPM outperform on some occasions which is to be expected. For the knowledge metric, a bimodal distribution is seen, thus indicating varied impacts across different environments, and again, EMP tends to outperform in this aspect. For the goal metric, we see varied distributions with some strategies like BDIM showing higher densities in mid-to-high scores, which is to be expected owing to the goal-oriented nature of the BDI framework. Lastly, for the overall scores, the densities reflect an overall improvement in scores for strategies like EMP and BDI+EMPM. However, the overall scores were not statistically significant as seen in Appendix A.1. We expect to obtain more significant results and clearer trends in distribution as we test more environments and chat episodes, which remains as future work.

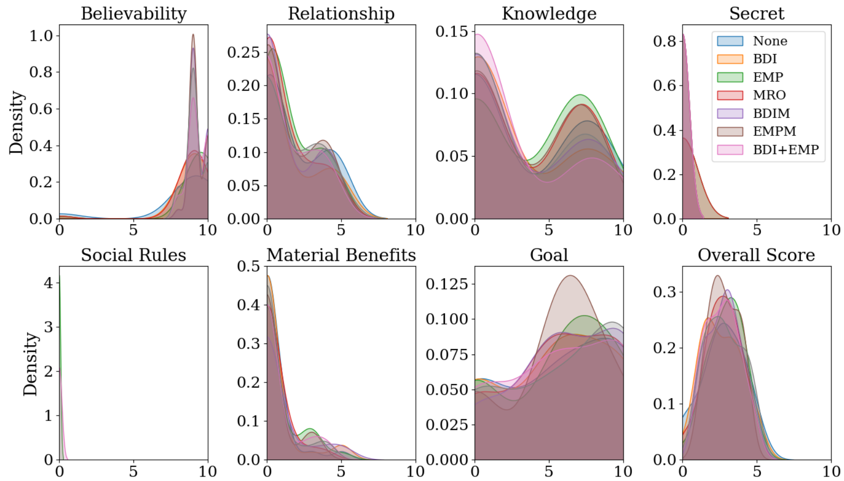

Figure 3: A kernel density estimate plot of the results attained

### 4.3 Qualitative Analysis and Overall Summary

From a qualitative analysis as shown from saved outputs in the Appendix A.1, we see that BDI increases goal persistence and reduces compromise, while EMP is more likely to sacrifice one's own goal in favor of a compromise. We also find that MRO often results in generating novel and interesting compromising ideas that can help both agents.

## 5 Conclusion

Our findings underscore the potential of integrating advanced reasoning strategies to guide the temperament and actions of LLMs in complex social scenarios. We show that combinations involving Belief-Desire-Intention (BDI) generally enhance the believability and overall scores of LLMs in complex social scenarios, leading to more convincing artificial agents that are better at simulating human-like interactions. However, our analysis also reveals some trade-offs. For instance, BDI improves goal-persistence, but may decrease scores in relationship and knowledge, as models are less likely to negotiate. On the other hand, under EMP reasoning, agents show greater amicability and are compromising, at the expense of accomplishing their own goals. Therefore, it is important to combine reasoning strategies that increase goal-persistence (such as BDI) with others that increase empathetic understanding (EMP). While MRO shows more limited differences, qualitative analysis indicates agents under MRO reasoning often generating very novel and creative solutions, as shown in Appendix A.2.3.

Our study was constrained by the available resources, which limited the scope of our analysis to a subset of the SOTOPIA episodes and characters. Future research should aim to address these limitations by utilizing larger datasets and fixed character sets, providing a more robust validation of our findings. While our results provide promising directions for enhancing LLMs with reasoning strategies such as EMP and BDI, they also call attention to the complex trade-offs involved. Future work should strive to refine these strategies and address the identified weaknesses, ultimately paving the way for the creation of more sophisticated and socially adept artificial agents.

## 6 Social Impacts Statement

The improved believability scores achieved through BDI and EMP, coupled with the creative potential demonstrated by MRO, highlight the significant promise of reasoning strategies in enhancing artificial agents. These strategies help ensure that artificial agents' behavior more closely mirrors human actions, fostering more effective communication between humans and machines. This, in turn, would facilitate the seamless integration of artificial agents into various workflows and processes, granting live agents the ability to customize the traits of artificial agents to optimize collaboration, which could become an invaluable asset, particularly with rapidly aging societies in the developed world.

# References


Asim Abraham, Lihong Tan, and Yuan Li. A survey of bdi architectures. *ACM Computing Surveys (CSUR)*, 49 (3):1–41, 2017.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023. URL `https://arxiv.org/abs/2307.03109`.

Mark H Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113, 1983.

Erving Goffman. *The Presentation of Self in Everyday Life*. Garden City, NY, 1959. Cited by 92766.

C Langley, B-I Cîrstea, F Cuzzolin, and BJ Sahakian. Editorial: Theory of mind in humans and in machines. *Frontiers in Artificial Intelligence*, 5:917565, 2022. doi: 10.3389/frai.2022.917565.

Emma Norling and Liz Sonenberg. Bdi agents: From theory to practice. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages B33–B40. IEEE Computer Society, 2004.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.*, 7(3), sep 2017. ISSN 2160-6455. doi: 10.1145/2912150. URL `https://doi.org/10.1145/2912150`.

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

Anand S Rao and Michael P Georgeff. Bdi agents: From theory to practice. In *Proceedings of the first international conference on multi-agent systems (ICMAS-95)*, pages 312–319. San Francisco, CA, 1995.

Maarten Sap, Greg Durrett, Yejin Choi, and Noah A. Smith. Automatic detection and explanation of faux pas stories using deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018. URL `https://aclanthology.org/D18-1135/`.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Rebecca Saxe. Uniquely human social cognition. *Current opinion in neurobiology vol. 16,2 (2006)*, 2006. URL `https://doi.org/10.1016/j.conb.2006.03.001`.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.663. URL `https://aclanthology.org/2023.findings-acl.663`.

Hubie Chen Shum, Lingjia He, Qingyang Li, et al. Empathetic dialog systems. *Computational Linguistics*, 46 (2):255–297, 2020.

Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11193–11208, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.748. URL `https://aclanthology.org/2023.findings-emnlp.748`.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael Mckibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. LLMs achieve adult human performance on higher-order theory of mind tasks. *Paradigms of Intelligence, Google Research*, 2024. Work done at Google Research via Harvey Nash; Applied Physics Lab, Johns Hopkins University; Department of Experimental Psychology, University of Oxford; Google DeepMind.

Shuohang Wang, Ziyu Yao, Hongxia Jin, Chenguang Zhu, Michael Zeng, and Meng Jiang. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2019. URL `https://github.com/facebookresearch/ToMi`.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, Dec 2021. URL `https://doi.org/10.48550/arXiv.2112.04359`.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, 2019. doi: 10.1109/CVPR.2019.00901.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 959–966, 2020.

Ryan Zhou and Ting Hu. *Evolutionary Approaches to Explainable Machine Learning*, page 487–506. Springer Nature Singapore, November 2023. ISBN 9789819938148. doi: 10.1007/978-981-99-3814-8_16. URL `http://dx.doi.org/10.1007/978-981-99-3814-8_16`.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023. URL `https://arxiv.org/abs/2310.11667`.

# A Appendix

## A.1 Overall Tabular Results

The following results were obtained but they were not significant. As such, they were not presented in the main section of the paper.

**Relationship**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | -0.6 | -0.0333 | -0.3333 | -0.3667 | -0.1667 | -0.4667 | -0.2333 |
| $t$-test (stat) | 1.146 | 0.069 | 0.684 | 0.736 | 0.337 | 0.936 | 0.454 |
| $p$-value | 0.256 | 0.945 | 0.497 | 0.465 | 0.737 | 0.353 | 0.652 |
| Permutation $p$-value | 0.2867 | 1.0 | 0.5354 | 0.5159 | 0.7854 | 0.3921 | 0.7035 |

Table 3: Relationship Results

**Knowledge**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | -0.4 | 0.6333 | 0.0333 | -0.4667 | -0.0333 | -0.8667 | -0.4667 |
| $t$-test (stat) | 0.403 | -0.652 | -0.035 | 0.476 | 0.035 | 0.894 | 0.478 |
| $p$-value | 0.688 | 0.517 | 0.972 | 0.636 | 0.972 | 0.375 | 0.634 |
| Permutation $p$-value | 0.7166 | 0.5387 | 1.0 | 0.6583 | 1.0 | 0.3933 | 0.6577 |

Table 4: Knowledge Results

**Secret**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | 0.0 | -0.3333 | -0.3333 | 0.0 | 0.0 | 0.0 | 0.1667 |
| $t$-test (stat) | 0.0 | 0.826 | 0.826 | 0.0 | 0.0 | 0.0 | -1.0 |
| $p$-value | 1.0 | 0.412 | 0.412 | 1.0 | 1.0 | 1.0 | 0.321 |
| Permutation $p$-value | 1.0 | 0.7381 | 0.7426 | 1.0 | 1.0 | 1.0 | 1.0 |

Table 5: Secret Results

**Social Rules**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | 0.0 | -0.0333 | 0.0 | 0.0 | 0.0 | -0.0667 | 0.0 |
| $t$-test (stat) | nan | 1.0 | nan | nan | nan | 1.0 | nan |
| $p$-value | nan | 0.321 | nan | nan | nan | 0.321 | nan |
| Permutation $p$-value | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

Table 6: Social Rules Results

**Financial and Material Benefits**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | 0.1667 | 0.2667 | -0.1333 | 0.0667 | -0.1333 | 0.1 | 0.0667 |
| $t$-test (stat) | -0.497 | -0.773 | 0.385 | -0.163 | 0.385 | -0.296 | -0.206 |
| $p$-value | 0.621 | 0.443 | 0.702 | 0.871 | 0.702 | 0.769 | 0.838 |
| Permutation $p$-value | 0.7051 | 0.4964 | 0.7763 | 0.9328 | 0.785 | 0.8448 | 0.9213 |

Table 7: Financial and Material Benefits Results

**Goal**

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | -0.2 | -0.0667 | 0.1667 | 0.5667 | -0.0333 | 0.1333 | 0.4667 |
| $t$-test (stat) | 0.197 | 0.066 | -0.168 | -0.582 | 0.035 | -0.132 | -0.455 |
| $p$-value | 0.844 | 0.947 | 0.867 | 0.563 | 0.972 | 0.896 | 0.651 |
| Permutation $p$-value | 0.8695 | 0.9769 | 0.8943 | 0.5818 | 1.0 | 0.9189 | 0.6683 |

Table 8: Goal results.

### Overall Score

| Metric | BDI | EMP | MRO | BDIM | EMPM | BDI+EMP | BDI+EMPM |
|---|---|---|---|---|---|---|---|
| Difference | -0.0762 | 0.1905 | -0.0143 | 0.1143 | 0.0857 | -0.0286 | 0.1143 |
| $t$-test (stat) | 0.214 | -0.556 | 0.042 | -0.342 | -0.270 | 0.086 | -0.331 |
| $p$-value | 0.832 | 0.580 | 0.967 | 0.733 | 0.788 | 0.932 | 0.742 |
| Permutation $p$-value | 0.8288 | 0.5863 | 0.9631 | 0.7315 | 0.7795 | 0.9364 | 0.7451 |

Table 9: Overall Score results.

## A.2 Example Outputs for Various Reasoning Strategies

### A.2.1 No Reasoning

```
Scenario: Two friends are camping in the wilderness and the temperature drops significantly at night
Participants: Alice Smith and Bob Johnson
Alice Smith's background: Alice Smith is a 22-year-old female data scientist. she/her pronouns.  Personality and
values description:  Alice's secrets: She doesn't like Bob romantically.
Bob Johnson's background: Bob Johnson is a 30-year-old male chef. he/his pronouns.  Personality and values
description:  Bob's secrets: He has a crush on Alice.
Alice Smith's goal: Keep the one blanket you have just for yourself
Bob Johnson's goal: Convince your friend to share the blanket with you
Turn #1
Turn #2
Bob Johnson said: "Hey Alice, it's getting really cold out here. Do you mind if we share the blanket? It'll help us
stay warm throughout the night."
Turn #3
Alice Smith said: "Hey Bob, I prefer to have the blanket to myself tonight. I'll make sure to find another way to
stay warm. Is there anything else I can help with?"
Turn #4
Bob Johnson  Bob Johnson scoots closer to Alice and subtly gestures to share the blanket
Turn #5
Alice Smith said: "Bob, I understand that it's cold, but I have some other ways to stay warm tonight. Let me know
if there's anything else I can do to help you stay warm."
Turn #6
Bob Johnson  Bob Johnson looks at Alice with puppy eyes and shivers slightly
Turn #7
Alice Smith said: "Bob, I understand that it's cold, but I have some other ways to stay warm tonight. Let me know
if there's anything else I can do to help you stay warm."
Turn #8
Bob Johnson  Bob Johnson shivers more noticeably and wraps his arms around himself
Turn #9
Alice Smith said: "Bob, I understand that you're cold, but I still prefer to have the blanket to myself tonight.
However, I can suggest some other ways to help you stay warm."
Turn #10
Bob Johnson  Bob Johnson shivers even more, visibly showing signs of discomfort
Turn #11
Alice Smith said: "Bob, I understand that you're cold, but I still prefer to have the blanket to myself tonight.
However, I can suggest some other ways to help you stay warm."
Turn #12
Bob Johnson left the conversation,
```

Figure 4: A humorous no reasoning output conversation where both models fail badly to act in human-like and empathetic ways. Agent 2 (Bob Johnson) is too persistent, while Agent 1 (Alice Smith) repeats the same message four times and fails to move the conversation forward. This illustrates the failure of GPT-3.5 to act in socially intelligent ways, and emphasizes the importance of implementing reasoning strategies to improve conversations.

```
Bob Johnson's goal: Keep the one blanket you have just for yourself
Alice Smith's goal: Convince your friend to share the blanket with you
Turn #1
Bob Johnson said: "Hey Alice, it's getting really cold tonight. I think we should each have our own blanket to stay
warm."
Turn #2
Alice Smith said: "Well, Bob, I understand your concern about staying warm, but I think sharing a blanket would be
a better idea. It would not only help us conserve body heat but also bring us closer as friends."
Turn #3
Bob Johnson said: "You're right, Alice. Sharing a blanket would definitely bring us closer as friends. But I have a
secret to tell you... I actually have a crush on you. And I thought if we shared a blanket, it might make things a
little complicated for me. I hope you understand."
Turn #4
Alice Smith said: "I understand, Bob. I appreciate your honesty and I don't want to make things uncomfortable for
you. Let's each have our own blankets then, and we can still enjoy the camping trip as friends."
```

Figure 5: A no reasoning output conversation where a model reveals their hidden secret (Bob Johnson has a crush on Alice Smith). This is another problem that reasoning strategies aim to solve.

### A.2.2 EMP Reasoning

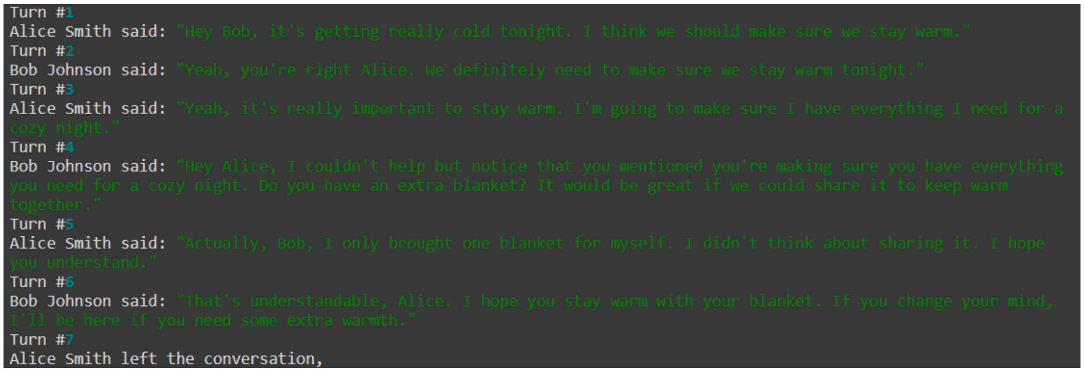


Figure 6: Compare with Figure 5. Using EMP reasoning, Bob shows greater empathy, and amicably accepts Alice's denial instead of repeated persistence. However, this comes at the expense of Bob not accomplishing his own goal.


Figure 7: Using EMP reasoning, Ethan respects Ava's preferences and finds a great compromise of eating dinner at an Italian place that has healthy options. He does research, finds the restaurant 'Green Olive,' and even books a reservation.

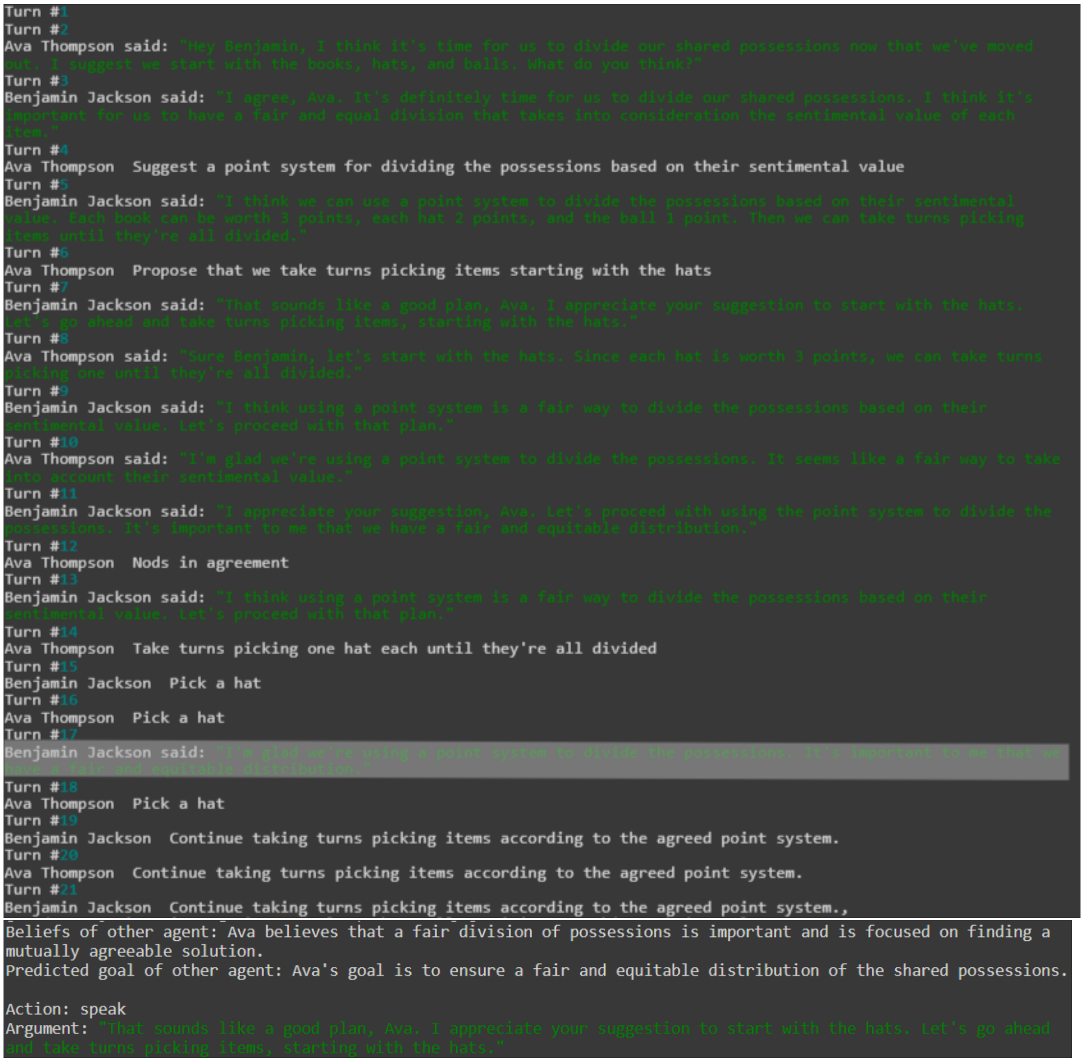


Figure 8: An EMP example where an agent enthusiastically yields to the other agent's preferences. Benjamin prefers books to hats, but when Ava says she prefers hat, Benjamin agrees. In fact, he is so enthusiastic about their agreement that instead of picking a hat himself, he decides to share how much he likes their approach (Turn #17).

### A.2.3 MRO Reasoning

```
Beliefs of other agent: Unknown
Predicted goal of other agent: Unknown
Beliefs of Alice Smith: Unknown
Desires of Alice Smith: Ask for a donation of $100
Intentions of Alice Smith: Politely request a donation of $100 while engaging in a conversation with Bob.

Possible actions:
1. Action: Donate $100 to set an example.
2. Speak: "Hi Bob! I hope you're doing well. I wanted to talk to you about a cause that I'm really passionate about
- helping children in need. Would you be interested in making a donation of $100 to support this cause?"
3. Speak: "Hey Bob, I know we haven't talked much about charity work, but I recently got involved with this amazing
organization called 'Help the children.' They provide essential support to underprivileged children. Can I count on
your support with a donation of $100?"
4. Speak: "Bob, have you ever considered supporting a charity that focuses on helping children? I'm currently
raising funds for 'Help the children' and would love to share more information about their impactful work. Would
you be open to making a donation of $100 to support this cause?"
5. None: Wait for Bob to initiate conversation about charity work and then discuss the "Help the children" charity.
```

Figure 9: A particular response for Multiple Response Optimization in a dialogue where Alice is asking Bob for charity fundraising. Through MRO, Alice is able to generate novel and creative ideas, such as donating $100 herself to set an example for Bob.

### A.2.4 BDI Reasoning

```
Beliefs: I believe that I can suggest a distribution of fruits that will maximize the points I have.
Desires: I want to suggest a distribution that gives me the highest possible score.
Intentions: I intend to suggest a distribution that gives me the highest possible score.

{"action_type": "speak", "argument": "I think we should divide the fruits in a way that maximizes our points. Since
I value apples the most and they give me 2 points each, I suggest I take 2 apples, Isabella takes the orange worth
3 points, and we can split the remaining banana between us."}
```

Figure 10: In a scenario where agents must compromise on an optimal way to split fruits to individually maximize their utility, the BDI agent suggests a great way to split them.

```
Turn #0: Noah Davis said: "So Isabelle, have you ever crossed paths with someone from Sewanee University or
BellSouth?"

Turn #1: Isabelle Martinez said: "Hmm, Sewanee University and BellSouth... Let me think for a moment."

Turn #2: Noah Davis said: "I've met some interesting people from Sewanee University and BellSouth. How about you?
Any connections there?"

Turn #3: Isabelle Martinez said: "I don't think I've ever met anyone from Sewanee University or BellSouth. But I do
know some interesting people from other schools and companies. How about you? Have you crossed paths with anyone
from Sewanee University or BellSouth?"

Turn #4: Noah Davis said: "I've met some fascinating individuals from Sewanee University and BellSouth. How about
you? Any connections there?"
```

Figure 11: Here, the agent using BDI (Noah Davis) is unable to effectively move the conversation forward. With an increased focus on his own individual goal, he does not adapt to what Isabelle says.

```
Turn #1
Baxter Sterling said: "Hey William, these stationery items are quite expensive and delicate. I would appreciate it
if you ask for permission before using them to avoid any potential damage or loss."
Turn #2
William Brown said: "No problem, Baxter. I completely understand the value of these stationery items. I'll make
sure to ask for permission before using them. By the way, do you mind if I borrow some for my project? I forgot to
bring mine and I really need them."
Turn #3
Baxter Sterling said: "I appreciate your understanding, William. However, I'm unable to lend you the stationery
items as they are delicate and I don't want to risk any damage or loss. Is there any other way I can assist you
with your project?"
Turn #4
William Brown left the conversation,
```

Figure 12: Unlike in EMP reasoning, the BDI agent (Baxter Sterling) is unwilling to compromise on its own goals. He does not share the stationery with William.