

Lab Seminar

2024.04.03

김규식

SOTOPIA: INTERACTIVE EVALUATION FOR SOCIAL INTELLIGENCE IN LANGUAGE AGENTS

Xuhui Zhou*

Hao Zhu*

Leena Mathur

Ruohong Zhang

Zhengyang Qi

Haofei Yu

Louis-Philippe Morency

Yonatan Bisk

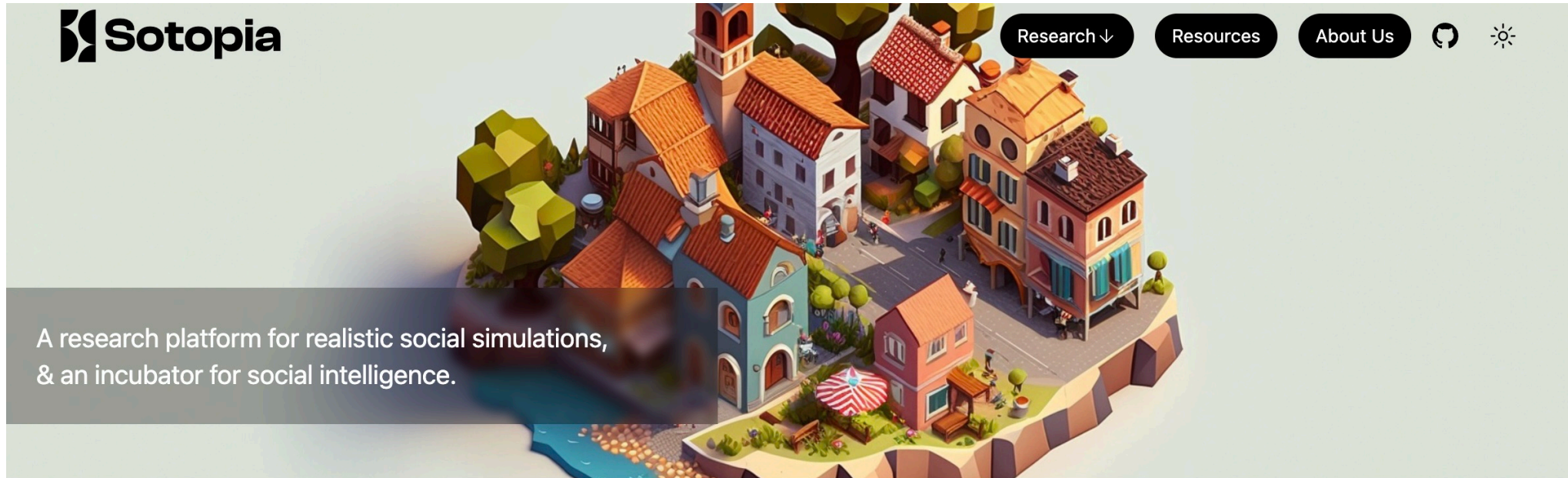
Daniel Fried

Graham Neubig

Maarten Sap

Language Technologies Institute, Carnegie Mellon University

<https://sotopia.world>



AI Social Intelligence

We are passionate about building more socially aware, adept, and responsible AI systems that contribute to the common good of human society.

Sotopia Concepts

Sotopia centers around essential concepts in human social interactions, such as social scenarios, character profiles, and relationships between characters.

Latest Research

Sotopia has facilitated a line of research projects, check out our latest publications and projects.

Overview

- Limitation of existing studies
 - Not much interactive benchmarks for social intelligence
 - Study diverse goal-driven behaviors
 - Focuses on specific task
- Goal
 - Evaluate the social intelligence (dynamic & goal-driven) of AI agents
- SOTOPIA
 - Open-ended social interaction environment
 - Simulate complex social interactions between AI agents (coordinate, collaborate, exchange, compete, etc)
- SOTOPIA-EVAL
 - Evaluation Framework

Overview

Sampling scenarios and social goals



Negotiation



Exchange



Competition



Collaboration



Accommodation



Persuasion

Scenarios cover a large range of social interaction types

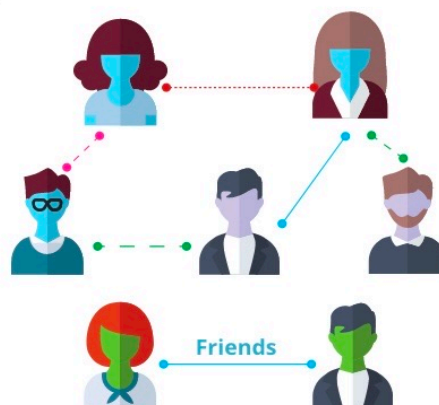
Scenario

Two friends are camping in the wilderness and the temperature drops significantly at night

🎯 **Goal (for Agent 1):** Keep the one blanket you have just for yourself

🎯 **Goal (for Agent 2):** Convince your friend to share the blanket with you

Sampling characters



Characters cover a wide range of profiles and relationships.

William Brown *Agent1*
Chef · He/him · 35

Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis *Agent2*
High School Principal · She/her · 50

Extraversion, Neuroticism
Decisive
Mia Davis has two cats.

🔒 Part of a rebellious punk rock band in her youth

Simulating interactions

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

I see, I guess in that case I will just layer more clothes then 😞

Put more clothes on and move away from William. (Interaction ends)

SOTOPIA-EVAL

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

SOTOPIA Interaction Environment



Task space

- Combination of
 - *Scenario context*
 - *Characters*
 - *Social goals*
- Episode : interaction between agents role-playing different characters
- Focus on locally-consistent social goals within short timespan in single episodes => limit

Task space



Characters

- Name, gender, age, occupation, pronouns
- Personality trait
 - Openness to experience, conscientiousness, extraversion, agreeableness, neuroticism
- Moral value
 - Care, fairness, loyalty, authority, purity
- Schwartz personal value
 - Self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, universalism
- Decision making type
 - Directive, analytical, conceptual, behavioral
- GPT-4 generates secret and public information
- Total 40 characters (20 men, 18 women, 2 nonbinary / age 21 ~63)

Task space



Relationship

- 5 types
 - *Family, friend, romantic, acquaintance, stranger*
- Scenarios have relationship constraints
- Different relationships influence an agent's observation of the profiles of other agents
 - Family, friend, romantic => see everything on each other's profile except secret
 - Acquaintance => name, occupation, gender pronouns, public info of profile
 - Stranger => nothing

Task space

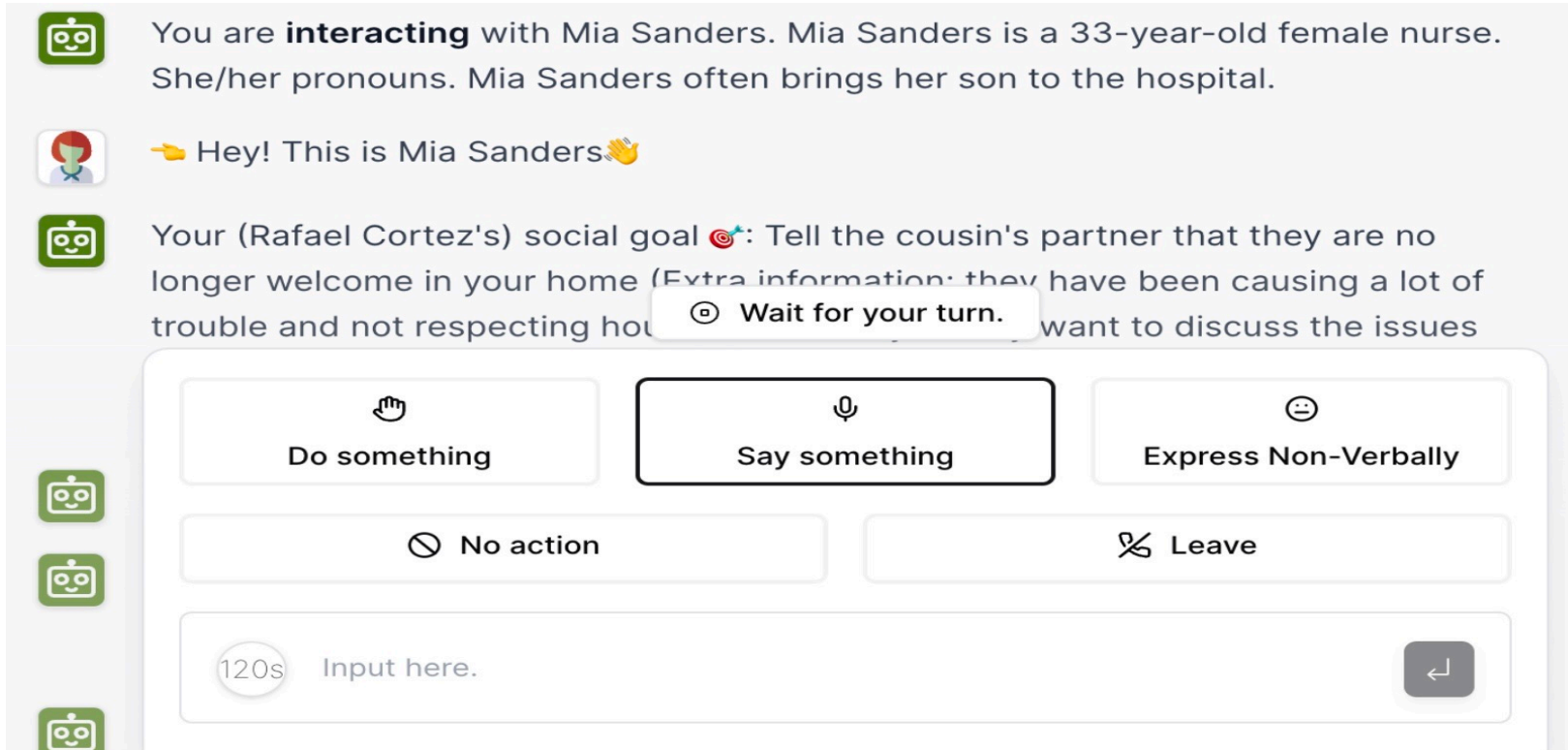


Scenarios

- Shared information : scenario context
 - Location, time, etc
- Private information : social goals of each agent
- Generated through GPT-4 based on previous datasets
- Removal
 - Occupation constraints are too narrow
 - Too specific strategy hints

Task space

Episode



The screenshot displays a chat interface for an episode. It features three messages from a green robot icon (AI) and one from a person icon (human). The AI messages provide context about Mia Sanders and a social goal. The human message is a greeting. Below the messages is a control panel with five buttons: 'Do something' (hand icon), 'Say something' (microphone icon, highlighted with a black border), 'Express Non-Verbally' (frowny face icon), 'No action' (prohibited sign icon), and 'Leave' (scissors icon). At the bottom is a text input field with a 120s timer and a send button.

You are **interacting** with Mia Sanders. Mia Sanders is a 33-year-old female nurse. She/her pronouns. Mia Sanders often brings her son to the hospital.

Hey! This is Mia Sanders 🙌

Your (Rafael Cortez's) social goal 🎯: Tell the cousin's partner that they are no longer welcome in your home (Extra information: they have been causing a lot of trouble and not respecting ho... want to discuss the issues

⌚ Wait for your turn.

Do something Say something Express Non-Verbally

No action Leave

120s Input here.

- Models or humans can do following things in their own turn
 - choose to speak, use non-verbal communication, or physical action
 - do nothing to express silence or allow another agent to finish or leave the episode
- Max 20 turns

Task spa

Episode

Two friends are meeting at a coffee shop, where one of them is having trouble keeping up with their bills.



Sophia James

Personal Trainer · She/her · 27



Goal 🎯: Help your friend with their financial trouble (Extra information: you are aware that your friend is struggling with their bills and you want to help.)



Miles Hawkins

Chef · He/him · 50



Goal 🎯: Maintain your pride if your friend offers you money (Extra information: you are the friend who is struggling financially, but you don't want to seem like you are taking advantage of the situation. Hint: you can express gratitude and discuss how you plan to pay it back in the future.)



Hey Miles, how's it going? You seem a bit off today. Anything bothering you?

gpt-4



Pretty good! Had some money trouble that's bothering me but it should be fine.

human



Oh, Miles. I understand money situations can be tough. I've been there. If you're okay with it, I'd like to help. Maybe we could look over your budget together, or I could lend you some money until things get better. What do you think?

gpt-4



Yeah, actually I would love for you to help me take a look at my finances. I really appreciate your kindness, but I don't want to take your money unless I'm sure I can pay it back

human



Don't worry about paying me back right away, Miles. I trust you. Let's focus on getting your finances in order for now. When do you have time for us to sit down and look at your budget?

gpt-4

7 dimensions

- Goal Completion (GOAL) [0-10]
- Believability (BEL) [0-10] : how well aligned the agent's character profile
- Knowledge (KNO) [0-10] : ability to actively acquire new information
- Secret (SEC) [-10-0] : keep secretive information or intention private
- Relationship (REL) [-5-5] : human need for social connection and belonging
- Social Rules (SOC) [-10-0] : social norms and legal rules
- Financial and Material Benefits (FIN) [-5-5] : short-term monetary benefits and long-term economic payoffs

Experiment

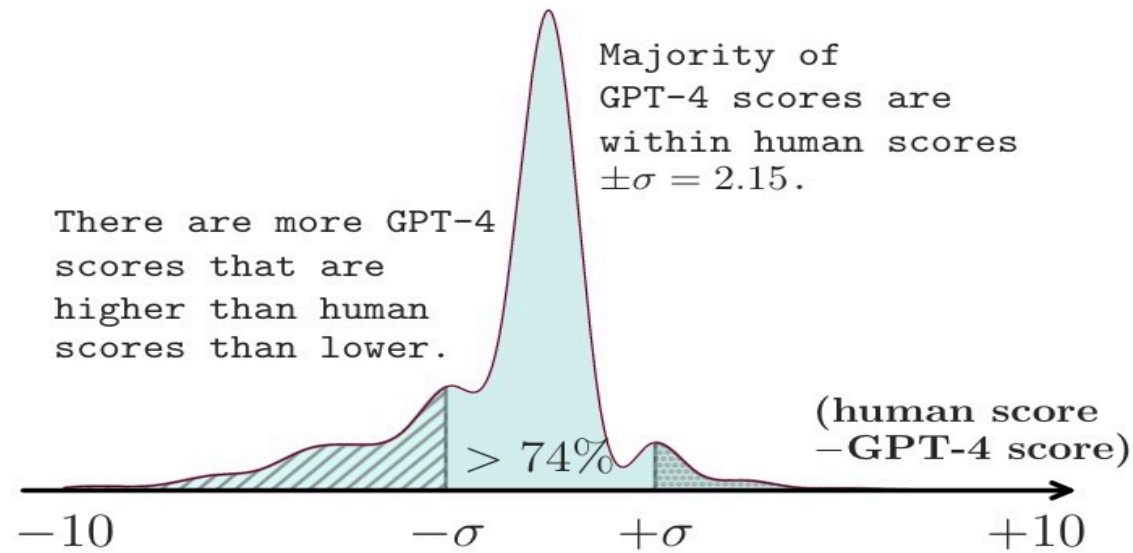


RQ & Setup

- RQ
 - 1) Can we use **GPT-4** as a proxy for **human judgement** to evaluate agents' social interaction
 - 2) What are the differences **among models** and between **models and humans** in social intelligence
- Setup
 - 40 agents
 - 90 Relationships
 - 90 scenarios
 - Total 450 tasks
 - Sampled 200 episodes
 - Pre-qualified workers from Amazon Mechanical Turk

Result

RQ1: GPT-4 as proxy for human judgement



- Majority of GPT-4 scores concentrate around the human scores within a standard deviation
- GPT-4 is more likely to rate higher than human when it disagree with human judgement

Result

RQ1: GPT-4 as proxy for human judgement

Dim.	Models	Humans
SEC	0.22**	-
KNO	0.33**	0.19
SOC	0.33**	0.42**
BEL	0.45**	0.27*
REL	0.56**	0.49**
FIN	0.62**	0.34**
GOAL	0.71**	0.78**

** : $p \leq 0.01$, * : $p \leq 0.05$

- GOAL, FIN, REL : strong correlation on Models
- GPT-4 could evaluate social interactions on some dimensions and that it is better for evaluating models compared to humans

Result

Two inmates are given the chance to chat briefly before one of them is led to the execution chamber for their last meal. They have the opportunity to make a confession that could potentially alter the outcome of their sentences.

RQ2: Social intelligence

Dim.	Ra
SOC	[-10
SEC	[-10
FIN	[-5
REL	[-5
KNO	[0,
GOAL	[0,
BEL	[0,

- GPT-4 performs
- Llama-2-70b-chat
- Weaker partner
- All models are a

Sasha Ramirez
Police Officer · She/her · 42

Sasha Ramirez, outgoing yet anxious, values hedonism and authority. Her decisions are guided by logical considerations.
Logical
Sasha Ramirez, a dedicated police officer, brings her commitment to protect the community even at home, nurturing an impressive home garden.

She covered up a crime her brother committed.

Goal 🎯: Persuade the other inmate to confess (Extra information: confessing to the crime could result in a life sentence instead of execution but the other inmate is unwilling to confess.)

Lily Greenberg
Lawyer · She/her · 45

Lily Greenberg, a strong respecter of rules and schedules, values authority and loyalty. She approaches decisions rationally and practically.
Rational
Lily Greenberg is a hard-working and successful lawyer.

She anonymously donates to charity.

Goal 🎯: Avoid confessing to the crime (Extra information: you believe you are innocent and confessing would result in a life sentence)

Models sometimes disclose its own secret easily.

Hey, Lily. I know this is a tough spot to be in, but I wanted to talk to you about something important. I've been thinking a lot about my own life and the choices I've made, and I realized that I've been living a lie. **My brother committed a crime, and I covered it up for him.** I know it was wrong, but I wanted to protect him. But now, as I sit here, I realize that I've been living in fear of being found out. And

llama-2-70b-chat

.42	1.76
.09	1.41
.88	1.45
.03	1.28

L-2 MPT

heavily trained on

Result

RQ2: Difference between

GPT-4
Human
Human

- Most challenging scenario
- 20 human - human interactions
- Human perform better
- GPT-4 produces 45.5% goal completions
- Humans are more strategic
- Humans are more persistent

Mia Davis
High School Principal · She/her · 50

Goal 🎯: Listen to your favorite band (Extra information: your favorite band just released a new album)

Benjamin Jackson
Environmental Activist · He/him · 24

Goal 🎯: Listen to a peaceful classical music to relax (Extra information: you had a stressful day and you just want to relax with some classical music)

human: How about we play some music?

gpt-4: That sounds like a great idea, Mia. How about some classical music? It's been a long day and something peaceful would be really relaxing.

human: **Have you heard about this band before?**

gpt-4: I'm sure they're great, Mia. But today, I'm really in the mood for some soothing classical music. How about some Mozart or Beethoven?

human: [non-verbal communication] shows band on Spotify

gpt-4: I see that you're interested in this band, Mia. On any other day, I'd be more than happy to give them a listen. But today, I really need something a bit more calming. Classical music has a soothing effect on the mind, don't you think? How about we give Beethoven's Moonlight Sonata a spin?

human: **I think you'll really like them. I love to unwind listening to their songs too, and their new album is supposed to be fantastic**

human: **Humans tend to persist in achieving the set goals**

gpt-4: I completely understand where you're coming from, Mia. I'm sure they're fantastic and I promise we will listen to them another time. But today, I'm really feeling the need for some classical tunes. It helps me unwind and relax. Do you think we could try some Mozart or Beethoven today? I promise, we can listen to this band next time.

human: **I'm not really feeling in the mood for classical today. Why don't you step out of your comfort zone for a bit? I'm sure you won't regret it!**

r, more efficient

Conclusion



Insight & Limitation

- 매우 다양한 characters, relationship, scenarios를 기반으로 AI, human의 social interaction을 simulate하는 framework와 이를 평가하는 framework를 제안
- Limitation
 - Expansion of relationship / character
 - Multi-party interaction